YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

PhD PROGRAMME

DATA MINING WITH EMPHASIS ON EXPLORATORY ANALYSIS

AYE AYE WIN

JUNE, 2015

CERTIFICATION

I hereby certify that the contents of this dissertation are wholly my own work unless otherwise referenced or acknowledged. Information from sources is referenced with original comments and ideas from the writer herself.

Aye Aye Win

PhD- Ah-1

June, 2015

ABSTRACT

Data mining is an analytical tool that is used in solving critical decision making problems by analyzing enormous amount of data in order to discover relationships and unknown patterns among variables in the data. This study focused on the investigation of the application of data mining techniques based on the tuberculosis (TB) diagnosis data set. The required data were organized from 659 TB suspected patients who came to the Union Tuberculosis Institute (UTI), Yangon during September and October 2013. This study attempted to predict whether a TB suspect has TB or not through the classification models by using decision tree method under the data mining techniques. The classification task with five different algorithms was made using decision tree method. It was found that the decision tree model of Algorithm I was found to be less accurate which used original data without preprocessing. The other four models which have performed preprocessing task revealed a better prediction having the same accuracy. Thus, this study proved that the decision tree method did not need the use of variable aggregation and feature reduction. The findings indicated that Active Specific Lung Lesion variable is the best predictor for making diagnosis about the present or absence of TB. The categorical value 'Yes' on Active Specific Lung Lesion is the most significant predictor of TB. Besides, the results obtained from decision tree method were compared with the results from logistic regression method. It was able to show that the accuracy of prediction for existence of TB disease or not is the same in two methods. It has also been observed that decision tree technique can provide classification rules which can identify the symptoms of TB. Therefore, decision tree method is found to be advantageous for the complex problems to make correct decisions according to the application used in this dissertation. Moreover, an alternative decision tree model was constructed without including X-ray result (Active Specific Lung Lesion variable). Even though the results from this was less accurate model using only patient's symptoms, the rules of this model were useful for people who had not undergone medical check-up at clinics in order to the predict the present of TB. The classification rules provided by the decision tree model (without X-ray results) revealed that there is a better advantageous for the healthcare centers which have no X-ray machine since these rules can be used to make the efficient prediction for diagnosis. By using these rules (in Appendix D), the field workers should encourage the patient who has high likelihood of TB positive to go to the nearest healthcare center where X-ray machine, advanced technologies for diagnosis and expert technicians has.

ACKNOWLEDGEMENTS

First and foremost, I express my appreciation to Professor Dr. Daw Khin Naing Oo, Rector, Yangon University of Economics for her permission to write my dissertation. My deepest appreciation and gratitude goes to Professor Dr. Tun Aung, Pro-Rector, Yangon University of Economics for his guidance to submit this thesis.

I am greatly thankful to Professor Dr. Thet Lwin (Retired), Professor Dr. Daw Khin San Myint (Retired), Professor Daw Htar Htar (Retired), Professor Daw Mya Mya Win (Retired) and Professor U Ngwe Soe (Retired) for their invaluable guidance and suggestions. I am very grateful to U Nyan Lin (Technical Advisor, UNDP), Professor Dr. Daw San Kyi (Retired) and Professor Dr. Win Tun, Acting Rector, Monywa University of Economics for their guidance and supervision while preparing my dissertation. I would like to express my special thanks to my supervisor Professor Dr. Lay Kyi, Pro-Rector (Retired), Yangon University of Economics for his close supervision and guidance. I am really indebted to Professor Dr. Daw Khin May Than, Head of Department of Statistics, Yangon University of Economics for her valuable advice and help. I am greatly thankful to Professor Dr. Daw Maw Maw Khin, Professor Dr. Daw Mya Thandar, and Associate Professor Daw Aye Aye Than for their suggestions. Special acknowledgements go to Dr. Daw Tin Mi Mi Khing, Union Tuberculosis Institute, Yangon, for her permission to use tuberculosis diagnosis data set. I would like to express my special thank to teacher Daw Than Than Myint, Lecturer, Defence Services Academy (DSA) for her valuable advice and help.

I sincerely thank all my teachers and all my colleagues from the Yangon University of Economics for their encouragements on my study.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation		Page
TB	Tuberculosis Disease	2
MDR-TB	Multidrug Resistant TB	3
AIDS	Acquired Immune Deficiency Syndrome	3
WHO	World Health Organization	3
HIV	Human Immunodeficiency Virus	3
NTP	National Tuberculosis Programme	4
DOTS	Directly Observed Treatment, Short Course	4
UTI	Union Tuberculosis Institute	5
AI	Artificial Intelligence	6
CRISP-DM	Cross-Industry Standard for Data Mining	6
SEMMA	Sample, Explore, Modify, Model, Assess	6
OLS	Ordinary Least Squares	31
CART	Classification and Regression Tree	31
PTB	Pulmonary Tuberculosis	34
RPTB	Retroviral Pulmonary Tuberculosis	34
DMT	Data Mining Techniques	34
ID3	Iterative Dichotomiser 3	37
CHAID	Chi-squared Automatic Interaction Detection	37
TP	True Positive	45
TN	True Negative	45
FP	False Positive	45
FN	False Negative	45
ESR	Erythrocyte Sedimentation Rate	48
AFB	Acid Fast Bacillus	48

CHAPTER 1

INTRODUCTION

Since the late 1980s, data mining has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making. Generally, data mining (sometimes called data or knowledge discovery) is the process of mechanisms and techniques to extract hidden information from data. Technically, data mining is the process of finding correlations or patterns among dozens of fields (variables) in large relational database[1].

The successful organizations have the ability to manage and plan for their future activities and to retrieve facts that happen in current situation and problems. In today's competitive environment, organizations require comprehensive business analysis support that is easy to understand. Not only do organizations require improved data viewing, but they also require the ability to find out information or knowledge in many different ways in which it will provide the unique insight required to make decisions. Therefore, data have now become central and even vital to an organization's survival.

In current situation, there are large amount of data and complex structure of data in many applications. Therefore, people encounter many problems in order to extract hidden and valuable knowledge from large amount of data for their organization. Data sets are often inaccurate, incomplete, and/or have redundant or insufficient information. As a consequence, the analysis of data and using existing data for correct prediction of state of nature for use in similar problems has been an important and challenging research area for many years. Data can be analyzed in various ways. The types of information obtained from data mining include associations, sequences, classifications, clusters and forecasts. Among these types, classification of information is an important part of decision making tasks. Many decision making tasks are instances of classification problem or can be formulated into a classification problem which includes prediction and forecasting problems, diagnosis or pattern recognition. Classification problems can be solved either by statistical method or data mining method.

The combination of data mining and statistical analysis is the search for valuable information from large volumes of data. It is now widely used in healthcare industry. The healthcare industry generates a great deal of information regarding disease data that can be collected and analyzed or mined to determine the hidden patterns. Those extracted

1. http:// www.Anderson.uda.edu./ faculty/ Jason.frand/ teacher/ technologies/ palace/ datamining.htm

patterns are used to interpret the new or existing data into useful information. Medical data mining has great potential for exploring the hidden pattern in the data sets of the medical domain. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently.

Classification is one of the major tasks in the data mining field. Among the available options in the data mining field, the most popular models in medicine are logistic regression, artificial neural network, and decision tree. Since the medical domain classification problem is highly nonlinear in nature, it is difficult to develop a comprehensive model to take into account all the independent variables using conventional statistical modeling techniques. Furthermore, for some historical information involved in millions of data it is quite difficult to analyze. In this situation, the problem is how to reduce the variables to a minimal number that can completely predict the response (outcome) variable. This study describes an approach of using multiple statistical analysis methods for data mining and the purpose of this study is to address the data mining algorithms to predict tuberculosis (TB) diagnosis by using patients' records.

1.1 Rationale of the Study

Data analysis for predicting and forecasting based on decisional information are applicable in various areas especially in business, healthcare and bioinformatics. The ability to understand and to accurately predict the data value can lead to substantial improvement in the overall aspect of the organization. Although conventional regression and time series analysis can perform the prediction, recently, there has been a growing interest in using data mining methods to analyze and model for prediction. It becomes important to know which data set will be applicable for particular organization or business and which technique can be used in order to select useful data from a vast amount of data storage or database. For such reasons, it is needed to examine how data mining methods can be used to identify critical predictors in order to predict response value.

The healthcare environment is still 'information rich' but 'knowledge poor'. The medical data set which includes patient records is difficult to analyze because it consists of huge volume and heterogeneity, temporality of data and high frequency of missing values. These data need to be collected in an organized form. These collected data can then be integrated to form which are ready to analyze for data mining. Applying data

mining technique to patient's attribute (data sets) is useful to build pattern or model that can be used to make correct diagnosis (prediction) for a particular type of disease.

Tuberculosis, which a few decades ago, was considered to be almost under control, has once again become a serious world-wide problem because of AIDS. It is one of the leading causes of infectious disease mortality in the world, with recorded over two million deaths annually and it is estimated that one-third of the world's population is latently infected. It is an infectious disease caused by the bacillus *Mycobacterium tuberculosis*. This *bacterium* widely exists in humans, cattle, sheep and birds. All of the organs in the body can be affected by tuberculosis. But most of the tuberculosis cases occur in lungs. It typically affects the lungs (pulmonary TB) but can affect other sites as well (extra-pulmonary TB). Lung tuberculosis can be seen on very wide age range. From new born babies to old people, everybody can be affected by this disease. Symptoms are cough, fatigue, exhaustion, anorexia, night sweating, fever (not exceeding 37.5 centigrade degree), cavities and hemoptysis on advanced cases. The disease spreads in the air when people who are sick with pulmonary TB expel bacteria, for example, by coughing.

Globally, 3.7% of new cases and 20% of previously treated cases are estimated to have Multidrug Resistant TB (MDR-TB). In 2011, there were an estimated 8.7 million incident cases of (globally, equivalent to 125 cases per 100,000 population) and 1.4 million people died from TB. Geographically, the burden of TB is highest in Asia and Africa. India and China combined have almost 40% of the world's TB cases; the South-East Asia and Western Pacific Regions of which they are a part account for 60%. The African Region has approximately one quarter of the world's cases, and the highest rates of cases and deaths relative to population[2].

Myanmar is one of the 22 TB high burden countries that account for 80% of all new TB cases arising each year, and the 27 countries that account for 85% of the global MDR-TB burden. Moreover, Myanmar is included in the 41 global priority countries for TB/HIV due to a high and growing HIV prevalence. A nationwide TB prevalence survey which was conducted in 2009-2010, revealed that the prevalence of TB in Myanmar was two to three times higher than previously estimated. These estimates were based on the latest nationwide smear-positive TB prevalence survey conducted in 1994. Moreover, in 2006, a TB prevalence study was carried out in Yangon division, reporting an incidence rate which was 2.3 times higher than the currently estimated rate[3].

According to the report of National Tuberculosis Programme (NTP) in 2014, the observed prevalence of smear-positive TB was 171 per 100,000 popultion and that of

2. World Health Organization (WHO), 2012. Global Tuberculosis Report.
3. World Health Organization (WHO), 2012. Review of the National Tuberculosis Programme of Myanmar

bacteriologically positive TB 434 per 100,000 population and there was 43 per 100,000 population died from TB disease. In 2009, 134,023 TB cases were notified (all new and retreatment cases) corresponding to a case notification rate of 220 (all forms of TB) per 100,000 population. In the same year, 41,389 new smear-positive cases were reported or 70 cases per 100,000 population. The proportion of new smear-positive cases out of all pulmonary cases was 30.9% and the proportion of extra-pulmonary cases out of all TB cases was 23.6%. Out of all new and re-treatment cases in 2009, 4.8% was re-treatment cases. Male to female ratio was 2:1 in new smear positive cases. The most affected age group was between 25-54 years which represents the most active socio-economic age group. In addition, states showed a significantly higher prevalence than regions, which may be related to access to TB services. The TB prevalence was also higher in urban area (especially Yangon) than rural ones.

The first nationwide drug resistant survey was carried out in 2002 showing 3.9% MDR-TB among new cases and 15.5% MDR-TB among re-treatment cases. The second nationwide drug resistant TB survey was conducted in 2007 and showing 4.2% MDR-TB among new cases and 10.0% MDR-TB among re-treatment cases. These preliminary data indicate that MDR-TB transmission was still ongoing but that the production of drug resistant cases has leveled off, which was probably a result of Myanmar's successful DOTS (Directly Observed Treatment, Short Course) programme. Despite limited resources for TB control, the NTP of Myanmar has delivered excellent basic TB control services in 314 out of 325 townships (95% administrative DOTS coverage) during the last few years.

The underlying research problem that necessitated this study is the existence of high death rate of TB at a national level. As stated by the experts of the hospital, some of the laboratory results of a TB suspected patient do not indicate clearly the bacteria for TB. Hence, by assuming the patients' disease can be TB, this patient started the therapy for the disease. After some weeks it may be discovered that it is wrongly diagnosed. This leads to delay the control program of the disease, and because of such kind of problems lots of patients die.

The amount of data stored in medical databases increases exponentially with time. The technological advancement resulted in the management of huge computerized data acquisition and storage of databases contains hidden knowledge that can be important and useful for decision making. It is impossible and time consuming to unravel this knowledge. Moreover, improper conclusions ultimately affect decision making.

Consequently, a need to use more efficient techniques and to have or provide knowledge in a comprehensive form as well as to arrive at better results has developed from both the owner and users of the data bases. This has led to the exploration of a new field of research called data mining.

However, the problem is that all those previous studies were conducted by using a very small proportion of the database. Besides, in those studies, data analysis was conducted by using simple statistical techniques (such as regression and verification techniques). Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world databases, these techniques have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases. Thus, this study investigated how tuberculosis can be diagnosed using the background history of the patients data available in Union Tuberculosis Institute (UTI) by applying the methods of data mining technology.

The study also provides an overview of the information discovery techniques and highlights some important statistical techniques used in data mining for application to healthcare, including cluster discovery methods, logistic regression and decision tree analysis. The purpose of this study is to address the data mining algorithms to predict tuberculosis diagnosis using patients' records. A comparative study for the performance of the prediction of some data preprocessing algorithms was carried out in this study.

1.2 Objectives of the Study

In data mining, the choice of technique meant to be used in analyzing a data set depends on the understanding the nature of data by the analysts. The data preprocessing is of crucial importance for data mining and it usually starts with data exploration phase in order to perform data understanding. Since most of the existing data sets may have different format and contain missing values, inaccurate data, redundant data or insufficient information, it is necessary for data miners to explore data as a first step of data preprocessing. Moreover, there is a wealth of data available within the healthcare system. However, there is a lack of effective analysis tools to discover hidden relationship and trends in data. In this situation, the problem is how to reduce the variables to a minimal number that can completely predict the response (outcome) variable and which technique does the best for medical data sets. For this reason, the following objectives are set in this study:

1. To identify algorithms which can be applied in order to extract hidden and useful information from existing databases or secondary data
2. To examine the essence of data preprocessing and data exploration phase in data mining
3. To find the classification rules which are useful for medical field workers in order to diagnose tuberculosis based on the best classification algorithm using medical symptoms of the patient.

1.3 Background

Data mining is an analytical tool that is used in solving critical decisions by analyzing the enormous amount of data in order to discover relationships and unknown patterns in the data. Data mining method is an algorithm designed to analyze data or to extract patterns from data. Most data mining methods are based on concept from machine learning, pattern recognition and statistics. Researchers from different branches of Mathematics, Statistics, Marketing and Artificial Intelligence[4] (AI) will use different terminologies. Where a statistician sees dependent variables, and artificial intelligence researcher sees features and attributes, others see records and fields (Berry, M. J. A. and Linoff, G. S., 2004).

The primary goal of data mining is to extract knowledge from data to support the decision-making process. In order to apply successfully, the data mining solution must be viewed as a process rather than a set of tools or techniques. Many data mining process methodologies are available. However, the various steps do not differ much from one methodology to the other. Some standard processes are CRISP-DM and SEMMA. CRISP-DM stands for Cross-Industry Standard for Data Mining, is an industry standard process consisting of sequence of steps that are usually involved in a data mining study. The other SEMMA is developed by SAS Institute. The acronym SEMMA stands for *s*ample, *e*xplore, *m*odify, *m*odel, *a*ssess. While each step of either approach is not needed in every analysis, this process provides a good coverage of the steps needed, starting with data exploration, data collection, data processing, analysis, inferences drawn, and implementation (Olson, D. L. and Delen, D., 2008).

Data preparation or preprocessing is critical to construct successful implementation of data mining. The purpose of data preparation is to decide data set (data unit and data field) that is ready to use for data modeling phase. Once the data resources available are identified, they are needed to be selected, cleaned, built into the form

4. Artificial Intelligence is the branch of computer science concerned with making computer behave like human.

desired and transformed (Olson, D. L. and Delen, D., 2008). Preprocessing of data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation. Data preprocessing is an important issue for both data warehousing and data mining, as real world data tend to be incomplete, redundant, and inconsistent.

The essential steps for data mining process are exploration stage and data modeling stage. Exploration helps refine and redirect the discovery process. If visual exploration does not reveal clear trends, one can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. Modeling techniques in data mining include artificial neural networks, decision trees, rough set analysis, support vector machines, and statistical models.

Final step for data mining is post-processing which includes evaluation of model performance. This is where the analyst evaluates the usefulness and the reliability of findings from the data mining process. In this final step of the data mining process, analyst assesses the models to estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set put aside (and not used during the model building) during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Predictive accuracy, computational speed, robustness, scalability, and interpretability are five criteria for the evaluation of classification and prediction methods.

1.4 Method of Study

In this study, an exploratory research through data mining techniques was conducted in order to set rules using the data set of medical field. This study focuses on classification (diagnosis) of a particular disease for patient; existence or non-existence. The medical data set used in this study is tuberculosis diagnosis data set which was obtained from Latha and Aung San townships, UTI in Myanmar. This data set includes patient's records which are organized by symptoms of TB suspected patients and it contains information about (659) patients who came to UTI for their medical check-up. Each of the patient record consists of (34) different variables: one dependent variable (outcome: TB or Non-TB) and (33) independent variables.

In this study, two data mining techniques: decision tree for logic solution and logistic regression for predictive data mining were used to develop the classification

models using a medical data set. First, data preprocessing on this data set is performed by using data cleaning methods, data transformation methods, missing value detection methods and variable reduction methods which include clustering and chi-squares test. After performing the preprocessing of data, the resulted data set is split into two subsets: training and testing samples[5]. Training sample is used for constructing the model and the testing sample is used to compute its quality of prediction. The most common technique used is called "cross-validation", which is used to measure the accuracy of the two prediction models for performance purpose. It splits the data into ten subsets, called folds. In the first step, it reserves the first fold for testing: it uses data from folds 2-10 for constructing the model and tests it on the first fold. Then it reserves the second fold for testing; folds 1 and 3-10 are used for constructing the model and fold 2 for testing it. This process goes on for all folds, totaling 10 times. At each step one fold is held out for testing and others are used for training. During testing, test samples are supplied to the model, having their class labels "hidden" and then their predicted class labels assigned by the model are compared with their corresponding original class labels to calculate prediction accuracy.

1.5 Scope and Limitations of the Study

The data from various sources such as insurance, micro-finance, marketing and healthcare can be analyzed by using data mining methods in order to find something new in the existing data and to quickly pull out usable information. Cancer diagnosis data under the medical field and loan data of micro-finance were tried to use in this study. If cancer diagnosis data would be obtained for this study, cancer disease which is ambiguous to classify can be nearly predicted based on the record of cancer suspected patients by using data mining methods. For micro-finance organizations, data mining methods can help to decide the borrowers' ability to payback their loan based on their profiles. Although there had been trials to obtain these types of data, it was impossible to organize them due to the nature of organizations and sources. Thus, this study focused on the tuberculosis diagnosis in medical field only.

To cover all states and regions in the entire country is impossible because of the limited time frame for this research and having communication problems. For this reason, the research focused on one region of the whole country, which is Yangon region. The choice of this region is due to the fact that Yangon has large number of records on tuberculosis suspected patients and high rate of occurrence of tuberculosis disease in

5. Separating data into training and testing sets is an important part of evaluating data mining models. More than 50% of the data set is used for developing model and which is called training data set because this task is performed to train model in data mining approach. After a model has been processed by using the training set, the model can be test by making predictions against the test set.

Myanmar. Therefore, this region will provide a good source of data for this study. In this study, information on tuberculosis suspected patients were obtained from Latha and Aung San Townships under UTI in Yangon, Myanmar. The cross-section data on tuberculosis diagnosis were organized from the representative patients who came to UTI during the period of 1st September to 31st October, 2013.

1.6 Organization of the Study

This study consists of six chapters. Chapter 1 is the introductory chapter in which rationale of the study, objectives of the study, background, method of study, scope and limitations of the study, and organization of the study are presented. The rest of the dissertation is organized as follows: Chapter 2 provides conceptual background and some works on data mining. In Chapter 3, importance of data preprocessing and the classification methods are reviewed. Chapter 4 describes an introduction of the data set used and the data preprocessing methodology employed for this study. In Chapter 5, classification models are built and comparisons of the performance of the different models are discussed. Based on the analysis in Chapter 5, conclusion is drawn and presented in Chapter 6.

CHAPTER 2

CONCEPTUAL BACKGROUND

The term, 'Data' (in singular, datum) comes from the Latin word which means 'those that are given'. Data are any facts, numbers, or text that can be processed into useful information or knowledge. In current situation, there are large amount of data and complex structure of data in many applications such as sales and marketing, healthcare/ medical diagnosis, supply chain management, process control, bioinformatics and astronomy. Therefore, people encounter many problems in order to extract hidden and valuable knowledge from large amount of data for their organization. Data sets are often inaccurate, incomplete, and/or have redundant or insufficient information.

2.1 Data Mining

Data mining is a process used by organization to generate useful information from raw data. Data mining aims to reveal knowledge about the data under consideration. This knowledge takes the form of patterns within the data that embody the understanding of the data. Patterns are also referred to as structures, models and relationships. Data mining has been defined from different perspectives by individual authors as follows:

"Data Mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amount of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques." (Gartner group, 2000)

"Data Mining is a new type of exploratory and predictive data analysis whose purpose is to delineate systematic relations between variables where there are no (or incomplete) a priori expectations as to the nature of these relations." (Luan, J., 2002)

"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." (Berry, M. J. A. and Linoff, G. S. 2004)

"Data Mining is the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners." (Murthy, K. I., 2010).

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression

analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability.

In the 1960s, statisticians used terms like “Data Fishing” or “Data Dredging” to refer to what they considered the bad practice of analyzing data without a priori hypothesis. The term “Data Mining” appeared around 1990 in the database community. At the beginning of the century, there was a phrase “database mining”, trademarked by HNC[9], a San Diego-based company (now merged into (FICO[10]) to pitch their Data Mining Workstation; researchers consequently turned to “data mining”. Other terms used include Data Archeology, Information Harvesting, Information Discovery, Knowledge Extraction etc. Gregory Piatetsky-Shapiro (1989) coined the term “Knowledge Discovery in Databases for the first workshop on the topic “data mining” and this term became more popular in Artificial Intelligence (AI) and Machine Learning Community. However, the term data mining became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

As data sets have grown in size and complexity, direct “hands-on” data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision tree (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and discovered by algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

There are slightly different underlying approaches where statistical methods foremost are designed for hypothesis testing while data mining methods are more focused on searching for the best among all possible hypotheses (Witten, I. H. and Frank, E., 2005). Traditional statistical analysis involves an approach that is usually directed, in that a specific set of expected outcomes exists. This approach is referred to as supervised (hypothesis development and testing). Data mining is, in some way, an extension of statistics, with a few artificial intelligence (AI) and machine learning twists thrown in. Statistics is at the core of data mining- helping to distinguish between random noise and significant findings, and providing a theory for estimating probabilities of predictions, etc (Murthy, I. K., 2010). Classical statistical approaches are fundamental to data mining. Automated AI ways are also used. However, systematic exploration through classical

6. HNC, known as a “neural network” company, Software Inc. is San Diego's largest software company and develops predictive software solutions for business-to-consumer service companies. These solutions allow companies to make more intelligent and profitable decisions and are marketed to industries- including financial, insurance, retail, telecommunications and the Internet.
7. HNC and DARPA, beginning in 1998, have worked on developing “cortronic neural networks,” which would allow machines to interpret aural and visual stimuli to think like humans. The cortronic concept was developed by HNC Software’s chief scientist and co-founder, Robert Hecht-Nielsen. HNC merged with the Minneapolis-based Fair Isaac Corporation (FICO), a computer analysis and

statistical methods is still the basis of data mining. Data mining covers the entire process of data analysis, including data cleaning, preparation and visualization of the results, as well as how to produce predictions in real-time, etc. By using the combination of statistics and data mining, the user enables to make effective utilization of the available information, to gain a better understanding of the past, and predict the future through better decision making.

Both statistics and data mining are concerned with learning from data and transformation of those data into useful information, data mining helps to find out the patterns and associations between the variables in the data and statistics helps to get process that data to get useful information. Data mining and statistics will inevitably grow toward each other in the coming times because data mining will not become knowledge discovery without statistical thinking, statistics will not be able to succeed on massive and complex data sets without data mining approaches (Murthy, I. K., 2010). Data mining is most useful for prediction and scoring but not for casual statistical analysis (Luan, J., 2002).

Data mining method is an algorithm designed to analyze data or to extract patterns from data. Most data mining methods are based on concepts from machine learning, pattern recognition and statistics (Berry, M. J. A. and Linoff, G. S., 2004). There are variety of mining methods and techniques including clustering, decision trees, neural network, rule induction, etc. Hence the main problem is that there are still no established criteria for deciding which data mining methods and techniques to use in which circumstances.

2.2 Exploratory Data Analysis

Exploratory studies fall under the category of inductive research. Exploratory research is an important mechanism of generating knowledge, when the problem under investigation is from a new research area and when access to detailed qualitative or quantitative data is available. Traditionally, most of the researches have been dominated by deductive research. Existing theories from various disciplines are used to develop hypothesis. Data are then collected through methodologies like a survey or a lab experiment to test the hypotheses. One of the reasons why researchers rely on such confirmatory research is the lack of large size data for conducting exploratory research. For exploratory research, large datasets are needed to examine patterns and to test the models (Raja, U. 2006).

The term exploratory data analysis was first used in the psychological and behavioral sciences. In brief, exploratory data analysis emphasizes flexible searching for clues and evidence, whereas confirmatory data analysis stresses evaluation of the evidence. It is concerned with both: the exploratory approach is focused on making the data analysis in a stepwise manner, evaluating at each step the appropriateness of the model and the data, and if necessary modifying the model and/ or the data basis. At each step, new insight is gained in terms of correlations between objects or variables outlying samples or the effects of preprocessing or numerous other important conditions necessary to reach valid conclusions. The exploratory approach lets the results from the iterative exploratory procedure help the analyst to define and find the combinations of analysis conditions that provide the optimal understanding of data (Andersson, C. A., 2000).

The comprehensive book on exploratory data analysis by Tukey (1977) has the following dictum “Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone- as the first step.” and in the same reference, the necessity for confirmatory analysis is also stressed. Tukey compares the task of doing exploratory analysis with that of detective looking for clues and hints to be able to find the truth.

Exploratory data analysis is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. The exploratory data analysis approach does not impose deterministic or probabilistic models on the data. On the contrary, it allows the data to suggest admissible models that best fit the data. Exploratory data analysis techniques do not share in that rigor or formality. These techniques make up for that lack of rigor by being very suggestive, inductive, and insightful about what the appropriate model should be. These techniques are also subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions. The exploratory data analysis approach often makes use of all the available data. In this sense there is no corresponding loss of information. Exploratory data analysis needs both the data mining and the application of statistical tools for data discovery. Since the essence of data preparation for data mining is exploratory in nature so the combination of statistical tools with the data mining proving to be of worth in exploratory data analysis (Murthy, I. K., 2010).

2.3 Data Mining Process

Although many data mining process methodologies are available, the various steps do not differ much from one methodology to other. Some standard processes are CRISP-DM and SEMMA. CRISP-DM stands for Cross-Industry Standard for Data Mining, is an industry standard process consisting of sequence of steps that are usually involved in a data mining study. The other SEMMA is developed by SAS Institute. (Olson, D. L. and Delen, D., 2008).

2.3.1 CRISP-DM Methodology

The CRISP-DM project began in mid-1997 and was funded in part by the European commission. This model consists of six phases intended as a cyclical process as in Figure (2.1) (Olson, D. L. and Delen, D. 2008). These six phases are:

1. Business Understanding

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

2. Data Understanding

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

3. Data Preparation

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

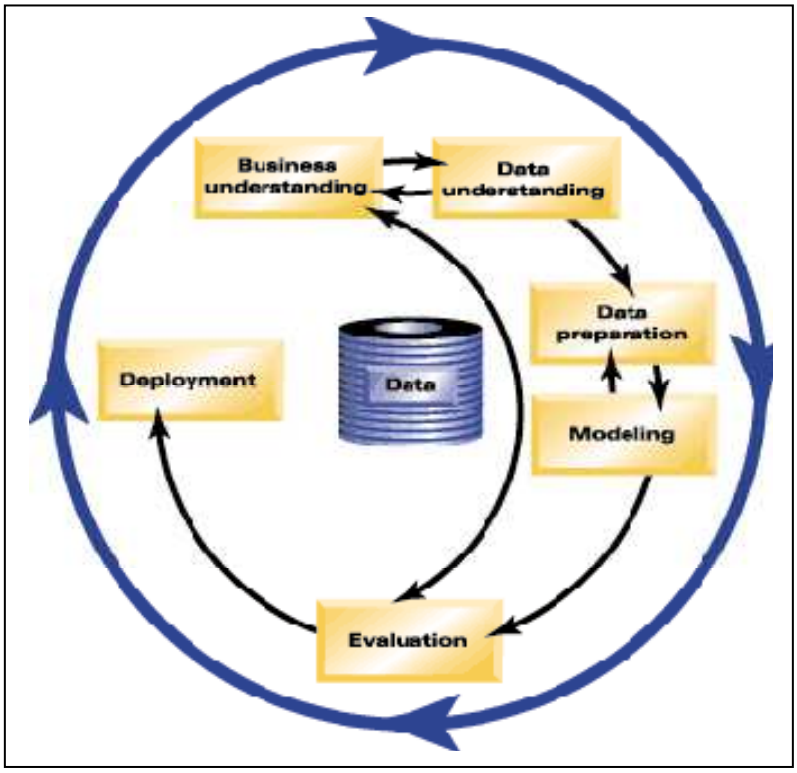


Figure (2.1) The CRISP-DM Model

Source: Olson D. L. and Delen D., 2008

4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques such as clustering, regression trees, decision tree and neural networks for the same data mining problem type.

5. Evaluation

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

6. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.3.2 SEMMA Methodology

The SEMMA analysis cycle guides the analyst through the process of exploring the data using visual and statistical techniques, transforming data to uncover the most

significant predictive variables, modeling the variables to predict outcomes, and assessing the model's accuracy by testing it with new data. A pictorial representation of SEMMA is given in Figure (2.2) (Olson, D. L. and Delen, D., 2008). This model has five steps.

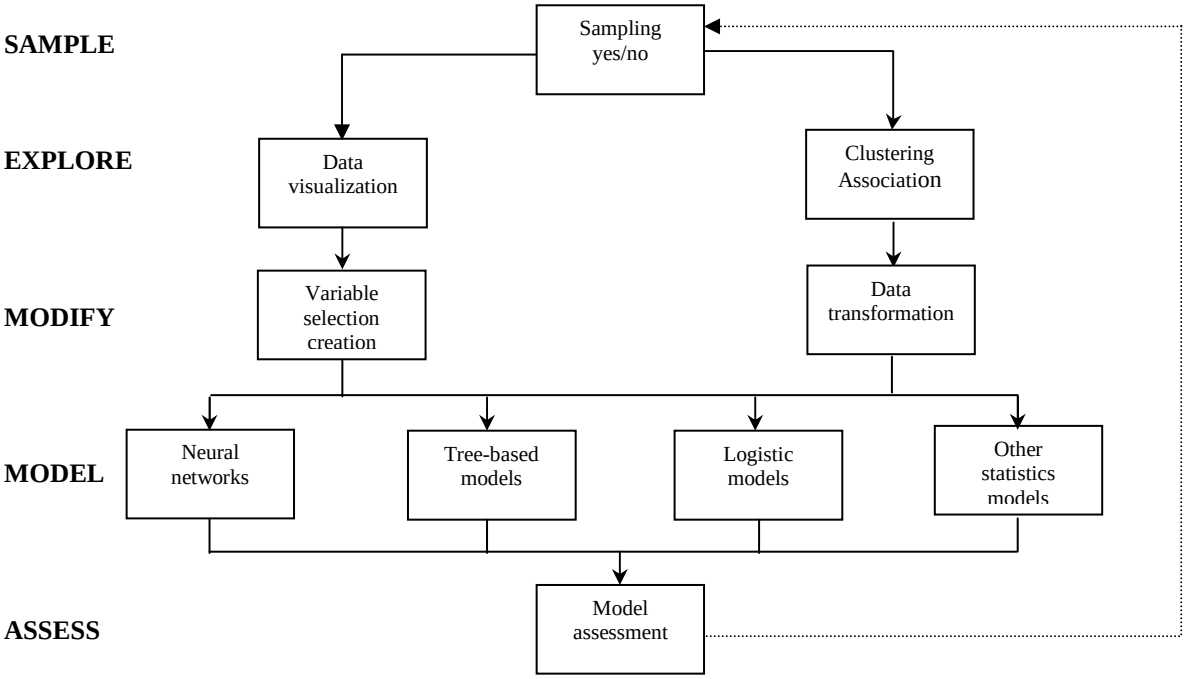


Figure (2.2) The SEMMA Model

Source: Olson D. L. and Delen D., 2008

1. Sample

This is where a portion of a large data set (big enough to contain the significant information yet small enough to manipulate quickly) is extracted.

The standard process of data mining is to take this large set of data and divide it, using a portion of the data (the training set[3]) for development of the model (no matter what modeling technique is used), and reserving a portion of the data (the test set) for testing the model that's built. In some applications a third split of data (validation set) is used to estimate parameters from the data. By dividing the data and using part of it for model development, and testing it on a separate set of data, a more convincing test of model accuracy is obtained. This idea of splitting the data into components is often carried to additional levels in the practice of data mining. Further portions of the data can be used to refine the model.

2. Explore

This is where the analyst searched for unanticipated trends and anomalies in order to gain a better understanding of the data set. After sampling data, the next step is to explore them visually or numerically for inherent trends or groupings. Exploration helps refine and redirect the discovery process.

3. Modify

This is where the analyst creates, selects, and transforms the variables upon which to focus the model construction process. Based on the discoveries in the exploration phase, one may need to manipulate data to include information such as the grouping of data unit and significant subgroups, or to introduce new variables. It may also be necessary to look for outliers and reduce the number of variables, to narrow them down to the most significant ones.

4. Model

This is where the analyst searches for a variable combination that reliably predicts a desired outcome. Modeling techniques in data mining include artificial neural networks, decision trees, rough set analysis, support vector machines, logistic models, and memory-based reasoning.

5. Assess

This is where the analyst evaluates the usefulness and the reliability of findings from the data mining process. In this final step of the data mining process user assesses the models to estimate how well it performs.

The following table illustrates the comparison of two models.

Table (2.1)

Comparison of CRISP-DM and SEMMA

CRISP	**SEMMA**	**Description**
Business Understanding	Assumes well-defined question	Goals are defined Develop tools to better utilize problem report
Data understanding	Sample Explore	Looked at data in problem reports
Data preparation	Modify data	Data pre-processing: Data field selection Data cleaning Data transformation
Modeling	Model	Data modeling
Evaluation Deployment	Assess	Analyzing results

Source: Olson D. L. and Delen D., 2008

The SEMMA approach is completely compatible with the CRISP-DM approach and SEMMA approach was used in this study.

2.4 Data Mining Tasks

Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result for various tasks. The data mining tasks typically fall into the general categories listed below (Jackson J., 2002).

1. Data Summarization
2. Segmentation
3. Classification
4. Prediction
5. Dependency analysis

Data Summarization gives the user an overview of the structure of the data and is generally carried out in the early stages of a project. This type of initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information.

Segmentation separates the data into interesting and meaningful sub-groups or classes. In this case, the analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization. Automatic clustering techniques can detect previously unsuspected and hidden structures in data that allow segmentation.

Classification assumes that a set of objects—characterized by some attributes or features—belong to different classes. The class label is a discrete qualitative identifier (large, medium, or small). The objective is to build classification models that assign the correct class to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling.

Prediction is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem type is also known as regression, and if the prediction deals with time series data, then it is often called forecasting.

Dependency analysis deals with finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of an item given information on other data items. Dependency analysis

has close connections with classification and prediction because the dependencies are implicitly used for the formulation of predictive models.

These tasks can be further divided into two major categories: predictive and descriptive tasks. Classification and prediction are predictive in their nature, while data summarization, segmentation, clustering and dependency analysis can be seen as descriptive.

2.5 Data Mining Techniques

Data mining combines techniques from machine learning, statistics, pattern recognition, database theory, and visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases. The selection of data mining techniques mainly depends on the type of data used for mining and the expected outcome of the mining process. The domain experts play a significant role in the selection of technique and algorithm for data mining.

Some of the commonly used statistical analysis techniques and machine learning algorithms which are used for performing data mining tasks which are descriptive tasks and predictive tasks are presented below:

Descriptive and Visualization Techniques include simple descriptive statistics such as: averages and measures of variation, counts and percentages, and cross-tabs and simple correlations. They are useful for understanding the structure of the data. Visualization is primarily a discovery technique and is useful for interpreting large amounts of data; visualization tools include histograms, box plots, scatter diagrams, and multi-dimensional surface plots (Jackson, J., 2002).

Cluster Analysis seeks to organize information about variables so that relatively homogeneous groups, or "clusters," can be formed. The goal of clustering is to identify clusters of records that exhibit similar behaviors or characteristics hidden in the data. The clusters may be mutually exclusive and exhaustive or may consist of a richer representation such as hierarchical or overlapping categories (Guo, L., ASA, 2002). In clustering, there is no pre-classified data and no distinction between independent and dependent variables. Clustering can be said as identification of similar classes of objects. By using clustering techniques, overall distribution pattern and correlations among data attributes (features or variables) can be discovered. Clustering can be used as preprocessing approach for other tasks such as attribute subset selection and classification (Ramageri B. M., 2010).

Correlation refers to any of a broad class of statistical relationships. Correlation is a statistical method used to assess a possible linear association between two or more variables. It is simple both to calculate and to interpret. Correlation is measured by a statistic called the correlation coefficient, which represents the strength of the linear association between the variables in question. It is a dimensionless quantity that takes a value in the range −1 to +1.

Neural Networks is a class of systems modeled after the human brain. As the human brain consists of millions of neurons that are inter-connected by synapses, neural network is formed from large numbers of simulated neurons, connected to each other in a manner similar to brain neurons. As in the human brain, the strength of neuron inter-connections may change (or be changed by the learning algorithm) in response to a presented stimulus or an obtained output, which enables the network to "learn". A disadvantage of neural network is that building the initial neural network model can be especially time-intensive because input processing almost always means that raw data must be transformed. Variable screening and selection requires large amounts of the analysts' time and skill. Also, for the user without a technical background, figuring out how neural networks operate is far from obvious.

Case-Based Reasoning is a technology that tries to solve a given problem by making direct use of past experiences and solutions. A case is usually a specific problem that was encountered and solved previously. Given a particular new problem, this technique examines the set of stored cases and finds similar ones. If similar cases exist, their solution is applied to the new problem, and the problem is added to the case base for future reference. A disadvantage of this method is that the solutions included in the case database may not be optimal in any sense because they are limited to what was actually done in the past, not necessarily what should have been done under similar circumstances. Therefore, using them may simply perpetuate earlier mistakes.

Genetic Algorithms operates through procedures modeled upon the evolutionary biological processes of selection, reproduction, mutation, and survival of the fittest to search for very good solutions to prediction and classification problems. It is used in data mining to formulate hypotheses about dependencies between variables in the form of association rules or some other internal formalism. A disadvantage of this technique is that the solutions are difficult to explain. Also, they do not provide interpretive statistical measures that enable the user to understand why the procedure arrived at a particular solution.

Decision Trees are like those used in decision analysis where each non-terminal node represents a test or decision on the data item considered. Depending on the outcome of the test, one chooses a certain branch. To classify a particular data item, one would start at the root node and follow the assertions down until a terminal node (or leaf) is reached; at that point, a decision is made. Decision tree can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules. A disadvantage of decision tree is that trees should never be used with small data sets.

Association Rules are statements about relationships between the attributes of a known group of entities and one or more aspects of those entities that enable predictions to be made about aspects of other entities who are not in the group, but who possess the same attributes. Association rule mining finds interesting associations and/or correlation relationships among large set of data items. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis (Ramageri B. M., 2002). Association rules shows attributed value conditions that occur frequently together in a given dataset. It is useful for determining correlations between attributes of a relation and have applications in marketing, financial, and retail sectors. It's typical application is market-basket analysis, where the technique is applied to analyze point-of sales transaction data to identify product affinities. A retail store is usually interested in what items sell together; therefore it can determine what items to display together for effective marketing. Association rule is also known as link analysis and it is often applied in conjunction with database segmentation or clustering (Guo, L., ASA, 2010).

Discriminant Analysis is used to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no natural ordering on the groups. It is a technique for classifying a set of observations into two or more predefined classes. The purpose is to determine the class of an observation based on a set of variables known as predictors or input variables (analogous to independent variables in regression). The model is built based on a set of observations for which the classes are known. This set of observations is sometimes referred to as the training set. Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions. These discriminant functions are used to predict the class of a new observation with unknown class.

Factor Analysis is useful for understanding the underlying reasons for the correlations among a group of variables. The main applications of factor analytic techniques are to

reduce the number of variables and to detect structure in the relationships among variables; that is to classify variables. Therefore, factor analysis can be applied as a data reduction or structure detection method. In an exploratory factor analysis, the goal is to explore or search for a factor structure (Jackson J., 2002).

Principal Component Analysis allows the analyst to use a reduced number of variables in ensuring analysis and can be used to eliminate the number of variables, though with some loss of information. However, the elimination of some of the original variables should not be a primary objective when using principal component analysis.

It also allows figuring correlations between variables which will permit the model builder to group those variables or to choose the variable that will represent best. The calculation of pairwise correlation coefficients can be used as a technique to exclude the variables that have a relatively low and insignificant correlation with the target variable. Principal component analysis is very useful when there are many independent variables and when those independent variables are highly correlated between them. Principal Component Analysis is a linear algebra technique for continuous attributes that finds new attributes (principal components) characterized by: (1) Being linear combinations of the original attributes, (2) Orthogonal to each other and (3) Capture the maximum amount of variation in the data (Salame E. J., 2011).

Regression Analysis is a statistical tool that uses the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from the other variable(s) (independent variables). The variable which is necessary to estimate is referred to as dependent, while the variable used in the model to predict the dependent variable is called independent. However, the number of potential independent variables may be unlimited and the model is referred to as multiple regression if it involves more than one independent variables. Logistic regression is used when the response variable is qualitative outcome. Although logistic regression finds a "best fitting" equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, that is, it maximizes the probability of obtaining the observed results given the fitted regression coefficients. Because logistic regression does not make any assumptions about the distribution for the independent variables, it is more robust to violations of the normality assumption (Jackson J., 2002).

Table (2.2) is a matrix that summarizes the data mining analysis tasks and the techniques used for performing these tasks.

Table (2.2)
Data Mining Tasks and Techniques

Data Mining and Statistical Techniques	Data Summarization	Segmen-tation	Classifi-cation	Prediction	Dependency Analysis
Descriptive and visualization	✓	✓			✓
Cluster Analysis		✓			
Correlation Analysis					✓
Neural Networks		✓	✓	✓	
Case-Based Reasoning					✓
Genetic Algorithms			✓	✓	✓
Decision Trees			✓	✓	
Association Rules					✓
Discriminant Analysis			✓		
Factor Analysis			✓		✓
Principle Component Analysis					✓
Regression Analysis				✓	✓

Source: Own Compilation

2.6 Uses of Predictive Data Mining

Even if the number of data mining tasks and the name of the tasks vary slightly, they all cover the same concept. These tasks can further be divided into two major categories: predictive and descriptive tasks (Berry, M. J. A. and Linoff, G. S., 2004). The ultimate goal of data mining is prediction and predictive data mining is the most common type of data mining. Data mining has a wide range of applications, including manufacturing, finance, telecommunications, bioinformatics, and neuronal studies. In manufacturing, data mining methods are widely implemented to predict the outcome of manufacturing process such as defective parts. In finance, credit analysis and prediction of loan payments are always critical to the business of financial firms. Data mining methods assist the firms to evaluate risk of their customers and also identify the important factors and eliminate irrelevant factors. In telecommunications, data mining aids providers to facilitate their churn detection activities and analyze fraudulent patterns and any unusual activities. Data mining methods also support researchers to better understand the biological process (molecular patterns, DNA and protein sequences). Data mining methods have been applied to discover gene expression patterns and identify complex disease genes. The following are the some examples of how data mining can be used in specific application area.

Credit Scoring: This problem, also called credit evaluation, is an attempt to classify applicants for credit into either "good" or "bad" risk classes. In this case, customers apply to a bank for a loan or credit card. These customers supply the bank with information which includes age, income, employment history, education, bank accounts, existing debts, etc. The bank does further background checks to establish credit history of customer. Based on this information, the bank must decide whether to make the loan or issue the credit card. The bank has a large database of existing and past customers. Some of these defaulted on loans; others frequently made late payments etc. An outcome variable *"Status"* is defined, taking value *"good"* or *"default"*. Each of the past customers is scored with a value for status. Background information is available for all the past customers. Using data mining techniques, a risk prediction model can be built by taking as input the background information, and outputs a risk estimate (probability of default) for a prospective customer.

Churn Prediction: When a customer switches to another provider, this can be called *"churn"*. Examples are cell-phone service and credit card providers. Based on customer information and usage patterns, the probability of churn and the retention probability (as a function of time) can be predicted. This information can be used to evaluate prospective customers to decide on acceptance and present customers to decide on intervention strategy.

Healthcare: The first one is treatment effectiveness in healthcare. Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. In such a case, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. Another one is to determine whether the patient has a heart condition. Based on the attributes which consists of age, sex, chest pain type, blood pressure, cholesterol, level of fasting blood sugar < 120, resting ECG, maximum heart rate, induced angina, old peak, slope, number of colored vessels etc., a heart condition can be decided.

Market Basket Analysis: In a supermarket, suppose as a manager, he may like to learn more about the buying habits of the customers. In this case, the problem is how to decide groups or sets of items customers are likely to purchase on a given trip to the store. To answer this question, market basket analysis from association rule mining may be performed on the retail data of customer transaction (transaction dataset include

transaction ID, and items from customer who bought more than 1 item) at store. The result may be used to plan marketing or advertising strategies as well as catalog design different store layouts. In one of the strategy, items that are frequently purchased together can be placed in close proximity in order to further encourage the sale of such items together. Market Basket Analysis can help retailers to plan which items to put on sale at reduced prices.

Customer Relationship Management: A company has collected data showing how much of their product consumers buy. For each consumer, the company has demographic and economic information which include age, gender, education, hobbies, income and occupation. Since the company has a limited budget, company's managers want to determine how to use the demographic data to predict which people are the most likely buyers of product thus manager can focus advertising on that group. Some data mining techniques (decision tree, support vector machine) can be used for this type of analysis because it show which combination of attributes best predict the purchase of the product.

2.7 Importance of Preprocessing in Data Mining

It is unrealistic to expect that data will be perfect after they have been extracted. Since good models usually need good data, a thorough cleaning of the data is an important step to improve the quality of data mining methods.

The data preprocessing is critical to construct successful implementation of data mining. Some selected data sets may have different formats and contain missing values because they are chosen from different data sources. The purpose of data preprocessing is to decide data set (data unit and data field), that is, ready to use for data modeling phase. There are many statistical methods and visualization tools that can be used to preprocess the selected data. Common statistics, such as maximum, minimum, mean, and mode can be readily used to aggregate or smooth the data, while scatter plots and box plots are usually used to filter outliers. More advanced techniques (including regression analysis, cluster analysis, decision tree, or hierarchical analysis) may be applied in data preprocessing depending on the requirements for the quality of the selected data. Because data preprocessing is detailed and tedious, it demands a great deal of time. In some cases, data preprocessing could take over 50% of the time of the entire data mining process. Shortening data preprocessing time can reduce much of the total computation time in data mining. Once the data resources available are identified, they are needed to be selected, cleaned, built into the form desired and transformed (Olson, D. L. and Delen, D., 2008).

2.7.1 Data Selection

Deciding on data to be used for analysis is essential in data preparation for data mining. Criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a data table (Jackson, J., 2002). When there is a reduction in the number of columns, there is variable or feature reduction and when there is reduction in the number of rows, the sample points or records are reduced.

The inclusion of variables in a model is not an easy task. Some criteria for variable selection considered by Salame (2011) are:

1. Is the variable legal?
2. Is it reasonable and factual?
3. Is it easily interpreted?
4. Is it difficult to manipulate?

The selected variables for the relevant data should be independent of each other. Variable independence means that the variables do not contain overlapping information. A careful selection of independent variables can make it easier for data mining algorithms to quickly discover useful knowledge patterns.

Data analysts need to guard against multicollinearity, a condition where some of the predictor variables are correlated with each other. Multicollinearity leads to instability in the solution space, leading to possible incoherent results, such as in multiple regression, where a multicollinear set of predictors can result in a regression that is significant overall, even when none of the individual variables are significant. Even if such instability is avoided, inclusion of variables that are highly correlated tends to overemphasize a particular component of the model, since the component is essentially being double counted. The use of too many predictor variables to model a relationship with a response variable can unnecessarily complicate the interpretation of the analysis and violates the principle of parsimony: that one should consider keeping the number of predictors to a size that could easily be interpreted. Also, retaining too many variables may lead to over-fitting, in which the generality of the findings is hindered because the new data do not behave the same as the training data for all the variables (Larose, D. T., 2005).

Further, analysis solely at the variable level might miss the fundamental underlying relationships among predictors. For example, several predictors might fall

naturally into a single group (a factor or a component) that addresses a single aspect of the data.

In order to perform data selection, dimension reduction methods are used to reduce variables or features. Dimension reduction methods have the goal of using the correlation structure among the predictor variables to accomplish the following:

- To reduce the number of predictor components
- To help ensure that these components are independent
- To provide a framework for interpretability of the results

Data reduction methods include simple tabulation, aggregation, clustering, factor analysis, principal component analysis, discriminant analysis and correlation analysis.

2.7.2 Data Cleaning

The purpose of data preprocessing is to clean selected data for better quality. Some selected data may have different formats because they are chosen from different data sources. Cleaning involves identification of missing, inconsistent, or mistaken values. Some entries are clearly invalid, caused by either human error or the technical errors. Those errors that are correctable are corrected. If all errors detected for a report are not corrected, that report is discarded from the study. In general, data cleaning means to filter, aggregate, and fill in missing values. By filtering data, the selected data are examined for outliers and redundancies. Redundant data are the same information recorded in several different ways. By aggregating data, data dimensions are reduced to obtain aggregated information. Note that although an aggregated data set has a small volume, the information will remain. Missing data can also be a particularly pernicious problem. Especially when the data set is small or the number of missing fields is large, not all records with a missing field can be deleted from the sample. By smoothing data, *missing* values of the selected data are found and new or reasonable values then added. These added values could be the average (mean) number of the variable (for continuous variable) or the mode (for categorical variable). A *missing* value often causes no solution when a data-mining algorithm is applied to discover the knowledge patterns (Olson, D. L. and Delen, D., 2008).

Six methods were suggested by Han and Kamber (2006) to fill the missing values:

1. Ignore the record
2. Fill the missing value manually
3. Use global constant

4. Replace the missing value with the mean
5. Replace the missing value with the mean of all samples of that category
6. Use the most likely value through the help of regression

2.7.3 Data Transformation

In data preparation, data transformation is to use simple mathematical formulations to convert different measurements of selected and cleaned data into unified numerical scale for the purpose of data analysis. In terms of representation of data, data transformation may be used to (1) transform from numerical to numerical scales, and (2) recode categorical data to numerical scales. One reason for transformation is to eliminate differences in variable scales (Olson, D. L. and Delen, D., 2008).

As always when performing simple linear regression, the first thing an analyst should do is to construct a scatter plot of the response versus the predictor to see if the relationship between the two variables is indeed linear. If the relationship is not linear, it would not be appropriate to model the relationship between two variables using a linear approximation such as simple linear regression. Such a model would lead to erroneous estimates and incorrect inference. Thus, an analyst applies transformations to all of the numerical variables that require it, to induce linear relationship between two variables. The analyst may choose from the transformations such as the natural log transformation and the square root transformation. For the variables which contain only positive values, the natural log transformation can be applied. However, for the variables that contained zero values as well as positive values, the square root transformation can be applied, since ln(x) is undefined for $x = 0$ (Larose, D. T., 2005).

Many data mining techniques are sensitive to the scale of the variables (attributes).The data values have to be transformed in order to generate new analytical variables and to fix skewed variable distribution. This is easily handled by normalizing (equation 2.1). If the maximum (max) and minimum (min) values are not known, standardizing (equation 2.2) can be used to transform all variables so they get zero mean and unit variance. Below, X is the variable to be transformed x′ the new value of x, μ is the mean and σ is the standard deviation:

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{2.1}$$

$$x' = \frac{x - \mu}{\sigma} \tag{2.2}$$

The result of successful data preprocessing for data mining is to improve the quality of the data which will help in improving “the accuracy and efficiency of the subsequent mining process”. In fact, analyst judgment is critical to successful implementation of data mining. Proper selection of data to include in searches is critical. Data transformation also is often required. Too many variables produce too much output, while too few can overlook key relationships in the data. Fundamental understanding of statistical concepts is mandatory for successful data mining (Olson, D. L. and Delen, D., 2008).

2.8 Feature (Dimensionality) Reduction Methods

Feature reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction (Verbeek, J.J. 2004). In order to improve the efficiency, the noisy (such as outliers, incorrect format and inconsistent) and redundant data may be removed and minimize the execution time, it is needed to reduce the number of variables in the original data set.

The entire process of model building for classification begins with collection of evidence acquired from various data sources or warehouses. In the ideal situation, the data should be of low-dimensionality, independent and discriminative so that its values are very similar to characteristics in the same class but very different in features from different classes. Raw data hardly satisfies these conditions and therefore a set of procedures called feature selection and feature extraction is required to provide a relevant input for classification system.

The terminology used in the area of feature reduction methods varies throughout the literature. In this study, the term feature reduction is used as a general term comprising both, feature selection and feature extraction methods. The term feature selection is reduction of the dimensionality by selecting attributes that are a subset of the old, original variables and the term feature extraction is reduction of the dimensionality by using low-rank approximation techniques to create new attributes that are a combination of the old, original variables. In other literature, feature extraction is sometimes also referred to as feature transformation, dimensionality reduction or feature construction.

As the dimensionality of data increases, many types of data analysis and classification problems become significantly harder. This can lead to problems for both supervised and unsupervised learning. Feature extraction and feature (subset) selection

methods are two types of techniques for reducing the attribute space. While in feature selection a subset of the original attributes is extracted, feature extraction in general produces linear combinations of the original attribute set. In both approaches, the goal is to select a low dimensional subset of the attribute space that covers most of the information of the original data. During the last years, feature selection and feature extraction techniques have become a real prerequisite for data mining applications (Janecek, A., 2009).

Feature selection has been an active research topic in recent times. This is an issue of great concern to researchers in pattern recognition, statistics as well as data mining areas. It is due to the fact that the data sets being generated are of high dimensionality. The rationale behind feature selection is to choose a subset of input variables by removing features with little or no predictive information. This is considered as a preprocessing stage to data mining with the objectives of increasing learning accuracy, improving the performance of predictors, providing faster and more cost-effective predictors, and providing better understanding of the underlying process that generated the data. It is intuitively reasonable from these objectives to think that large number of features is not informative because they are either irrelevant or redundant to the prediction model (Danso, S. O., 2006). In medical data mining, feature selection methods have been widely used to find attribute value that are most associated with a disease or subtype of certain disease.

Feature extraction refers to algorithms and techniques which create new attributes as (often linear) combinations of the original attributes in order to reduce the dimensionality of a data set. Rather than selecting a subset of the features, these techniques involve some type of feature transformation and aim at reducing the dimension such that the representation is as faithful as possible to the original data set, but with a lower dimension and removed redundancy. Because the new attributes are combinations of the original ones, the transformation process is also referred to as feature construction or feature transformation. This process of constructing new features can be followed by or combined with a feature subset selection process- the original feature set is first extended by the newly constructed features and then a subset of features is selected. Adding newly computed features to the original attributes can increase the classification results achieved with these feature sets more than replacing the original attributes with the newly computed features (Janecek, A., 2009).

Feature (dimension) reduction techniques are used to find compact representation of data by mapping each point to a lower dimensional continuous vector. A representation of the data in fewer dimensions can be advantageous for further processing of the data. However the reduction of the number of dimensions should not result in loss of information relevant to the task at hand. Thus, there is a trade-off between the advantages of the reduced dimensionality and the loss of information incurred by the dimension reduction (Verbeek, J.J, 2004).

2.9 Some Works on Data Mining

There are several studies for comparing the performance of model developed by using modeling techniques. One of the studies examined the performance of ordinary least squares (OLS) model and neural network model to see which model does a better job to predict the changes in stock prices (Tjung, L. C., Kwon,O., Tseng, K. C. and Bradly, J., 2011). They also identified critical predictors to forecast stock prices to increase forecasting accuracy for the professionals in the markets. OLS is a linear model that has relatively high forecasting error to forecast a non-linear environment in the stock markets. OLS model can only trace one independent variable at a time. On the other hand, neural network model has a high precision, improving prediction in non linear setting, and addressing problems with a great deal of complexity. Therefore they concluded that neural network does a better job compared to OLS model. For further direction, recommendations were made to include more techniques to find the best model for the financial forecasting purpose.

Kao and Chih (2001) stated that the classification and regression tree (CART) and the analytical neural networks provide an alternative to logistic regression, especially when the relationship between dependent and independent attributes are non linear. Their decision to use CART is based on a previous study that stated that CART is essentially non-parametric.

Delen et al. (2004) compared two data mining methods (neural network and decision trees) and a statistical one (logistic regression) with respect to their applications in medicine. The experiments proved that the decision trees are the most accurate predictors, with the neural networks to be the second and logistic regression to be the third. The authors focused also on traditional methods of medical prognosis as a method which encompasses estimations of potential complications and recurrence of the disease. The traditional statistical methods: Kaplan-Meier test or Cox-Propositional hazard models

are being gradually replaced by the data mining techniques and knowledge discovery. Finally, the authors mentioned several problems and issues that may arise while mining the breast cancer data. First of all it is the heterogeneity of the data that constitutes a problem to data mining algorithms. The data can also be incomplete, redundant, inconsistent and imprecise. Thus preparation of the data may require more data reduction than in case of the data from other types of sources. The authors concluded that the data mining, being a powerful tool, still requires a human to assess the results in terms of relevance, applicability and importance of the extracted knowledge.

Another study was conducted by Razi and Athappilly (2005). They performed a three-way comparison of prediction accuracy involving non-linear regression, neural networks and CART models. The prediction errors of these three models are compared where the dependent variable is continuous and predictor variables are all categorical. They recommended that neural networks and CART models produced better prediction accuracy than non linear regression model. However, neither neural networks nor CART model showed any clear advantage of one over the other.

Chang and Liou (2007) conducted on the investigation of the application of artificial intelligence and data mining techniques to the prediction models of breast cancer. The artificial neural networks, decision tree, logistic regression, and genetic algorithm were used for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. The authors used 699 records which acquired from the breast cancer patients at the University of Wisconsin, 9 predictor variables, and 1 outcome variable for the data analysis. Their results revealed that the accuracies of logistic regression model were 0.9434, the decision tree model 0.9434, the neural network model 0.9502, the genetic algorithm model 0.9878. The accuracy of the genetic algorithm was significantly higher than the average predicted accuracy of 0.9612. The predicted outcome of the logistic regression model was higher than that of the neural networks model but no significant difference was observed. The average predicted accuracy of the decision tree model was 0.9435 which was the lowest of all 4 predictive models. The authors indicated that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification, the expression and complexity of the classification rule. Their results showed that the genetic algorithm was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible.

Paliwal and Kumar (2009) did a thorough review of the application of neural networks. Ninety-six studies were compared neural networks with regression analysis, logistic regression, and discriminant analysis applied in the field of accounting and finance, health and medicine, engineering and manufacturing, marketing, and general applications. A summary of the performance of neural networks is stated as follows: multilayered feed forward neural network outperformed in about 58% of the cases, and in the rest of the cases it performed equivalently to the traditional statistical methods (24%) and in (18%) of the cases traditional statistical methods outperformed. Additionally they mentioned that the statistical methods are based on assumptions and consequently the validity of their performance will be essential.

Ansari and Soni (2011) provided a survey of current techniques of knowledge discovery in databases using data mining techniques that are used in today's medical research particularly in Heart Disease Prediction. Firstly, they concluded that the outcome of predictive data mining technique on the same data set reveals that decision tree outperforms. Sometimes Bayesian classification is having similar accuracy as of decision tree but other predictive methods like k-nearest neighbor's algorithm (k-NN), neural networks, classification based on clustering are not performing well. The second conclusion is that the accuracy of the decision tree and Bayesian's classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Shouman et al. (2011) investigated the application of a range of techniques to different types of decision trees seeking better performance in heart disease diagnosis. They also stated that decision tree is one of the successful data mining techniques used in the diagnosis of heart disease. The authors systematically tested combination of discretization, decision tree type and voting to identify a more robust, more accurate method. They have concluded that the supervised discretization methods do not show any enhancement in the decision tree accuracy either with or without voting. They indicated that nine voting with equal frequency discretization and gain ratio decision tree can enhance the accuracy of the diagnosis of heart disease. The authors also conducted the calculation for the sensitivity, specificity and accuracy in order to evaluate the performance of the alternative decision trees. They have proposed a model that outperforms J4.8 decision tree[8] and Bagging algorithm in the diagnosis of heart disease patients.

8. J4.8 decision tree is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the Weka project team.

Asha et al. (2012) presented a methodology for the automated detection and classification of Tuberculosis. Their methodology was based on clustering and classification that classifies TB into two categories, Pulmonary Tuberculosis (PTB) and retroviral PTB (RPTB) that is those with Human Immunodeficiency Virus (HIV) Infection. Initially, the authors used k-means clustering to group the TB data into two clusters and assigned classes to clusters. Their methodology was evaluated using 700 raw TB data obtained from city hospital. They concluded that the best obtained accuracy was 98.7% from support vector machine compared to other classifiers. Furthermore, they recommended that the proposed approach helps doctors in their diagnosis decisions and also in their treatment planning procedures for different categories.

Liao et al. (2012) presented a review of literature concerned with data mining techniques (DMT) and its applications, from 2000 to 2011. They concluded that development of DMT is tending to become more expertise-oriented and that the development of DMT applications is more problem-centered. The authors also suggested that different social science methodologies, such as psychology, cognitive science and human behavior might use DMT as an alternative methodology. They concluded that integration of qualitative and scientific method and the integration of studies of DMT methodologies will increase understanding of the subject. Moreover, the authors recommended that the ability to continually change and provide new understanding is the principle advantage of DMT methodologies, and will be at the core of DMT applications, in future.

Durairaj and Ranjani (2013) focused on the comparison of a variety of techniques, approaches and different tools and its impact on the healthcare sector. Their aims were to make detailed study report of different types of data mining application in healthcare sector and to reduce the complexity of the study of the healthcare data transactions. The authors also presented a comparative study of different data mining applications, techniques and different methodologies applied for extracting knowledge from database generated in the healthcare industry. They indicated that developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise. It was observed from the study that a combination of more than one data mining techniques than a single technique for diagnosing or predicting disease in healthcare sector could yield more promising results. They also conducted the comparison study and they concluded that the interesting result that data mining techniques in all the

healthcare applications give a more encouraging level of accuracy like 97.77% for cancer prediction.

Saumya et al. (2014) presented an overview of the current research being carried out using the data mining techniques for prognosis of cancers. The goal of the authors was to identify the well-performing data mining algorithms used on medical databases in order to predict survivability of cancer patients. The authors identified the algorithms: decision trees, support vector machine, artificial neural networks, Naïve Bayes and fuzzy rules. They showed that it is very difficult to name a single data mining algorithm as the best for predicting survivability of all the cancer types and for all the different contexts. The authors concluded that depending on certain situations, sometimes some algorithms perform better than others, but there are cases when a combination of algorithms results in more effective survival prediction.

CHAPTER 3

SOME PREDICTIVE METHODS AND MODEL BUILDING

Prediction is a data analyzing technique which determines important data classes or may construct models which can predict future data trend. Both classification and regression are used for prediction. While the classification predicts the categorical values, the regression is used in the prediction of numeric and continuous values (Han, J. and Kamber, M., 2006). Since data used in this study are categorical values, the classification techniques are applied for developing models.

Classification is most commonly applied as data mining technique, which employs a set of pre-classified examples to develop a model that maps a data item into one of several predefined classes. Once developed, the model is used to classify a new instance into one of the classes. Classification task is to determine the unknown class of new instances (diagnoses for newly arrived patients). The method for doing for this is called classifier, the problem of finding a good classifier is called classification problem. The data classification process involves learning and classification. The process of constructing the classifier is called classifier induction from data or also called learning from data. In learning, the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy of the classification (algorithm) rules. If the accuracy is acceptable, the rule can be applied to the new data (novel data) tuples. The part of data used for learning is often called learning data or training data. In statistics terminology, the task is to predict, and the statistical term for a classifier is predictive model. Sometimes it the term modeling is also used (Demsar, J., 2010).

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted into classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and k-nearest neighbor classification (Han, J. and Kamber, M., 2006). This

section discusses the concepts and principles of the decision tree method and logistic regression method in carrying out classification prediction.

3.1 Decision Tree

Decision tree learning is one of the most widely used and practical methods for classification. In this method, tree models can be represented as a set of if-then rules that improve human readability. Decision trees are very simple to understand and interpret by domain experts. A decision tree represents a hierarchical segmentation of the data. The initial data set constitute the root node which is partitioned to two or more segments based on a series of simple rules. Each resulting segment is further divided into sub segments and so on until no further division is possible and this partitioning process performs as recursive partitioning. The hierarchy constitutes the tree and the segments and sub segments constitute the nodes. Moreover, the nodes that are not further partitioned, is defined as terminal nodes or leaf.

Decision-makers prefer less complex decision trees, since the tree complexity has a crucial effect on its accuracy. The tree complexity is explicitly controlled by the stopping criteria used and the pruning method employed. Usually the tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used (Rokach, L. and Maimon, O., 2006).

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology (Han, J. and Kamber, M., 2006).

3.1.1 Decision Tree Inducers

Decision tree inducers are algorithms that automatically construct a decision tree from a given data set. Typically the goal is to find the optimal decision tree by minimizing the generalization error. There are various top-down decision tree inducers such as ID3, C4.5, CART and CHAID. Some consist of two conceptual phases: growing

and pruning (C4.5 and CART). Other inducers perform only the growing phase. A typical algorithmic framework for top-down inducing of a decision tree can be constructed by using growing and pruning. Note that these algorithms are greedy by nature and construct the decision tree in a top-down, recursive manner (also known as 'divide and conquer'). In each iteration, the algorithm considers the partition of the training set using the outcome of a discrete function of the input attributes. The selection of the most appropriate function is made according to some splitting measures. After the selection of an appropriate split, each node further subdivides the training set into smaller subsets, until no split gains sufficient splitting measure or stopping criteria is satisfied (Rokach, L. and Maimon, O., 2006).

3.1.2 Splitting Criteria

In most of the cases, the discrete splitting functions are univariate. Univariate means that an internal node is split according to the value of a single attribute. Consequently, the inducer searches for the best attribute upon which to split. There are various univariate criteria. These criteria can be characterized in different ways as follows (Rokach, L. and Maimon, O., 2006):

1. According to the origin of the measure: information theory, dependence, and distance.
2. According to the measure structure: impurity based criteria, normalized impurity based criteria and binary criteria.

The most common criteria are impurity- based, information gain, gini index, likelihood-ratio chi-squared statistics, normalized impurity based criteria, gain ratio, distance measure and binary criteria etc. Many of the researchers point out that in most of the cases, the choice of splitting criteria will not make much difference on the tree performance (Rokach, L. and Maimon, O., 2006).

3.1.3 Stopping Criteria

The growing phase continues until a stopping criterion is triggered. The following conditions are common stopping rules:

1. All instances in the training set belong to single value of y.
2. The maximum tree depth has been reached.
3. The number of cases in the terminal node is less than the minimum number of cases for parent nodes.

4. If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes.
5. The best splitting criteria is not greater than a certain threshold.

3.1.4 Pruning Methods

Employing tightly stopping criteria tends to create small and under-fitted decision trees. On the other hand, using loosely stopping criteria tends to generate large decision trees that over-fitted to the training set. Pruning methods were developed for solving this dilemma. According to this methodology, a loosely stopping criterion is used, letting the decision tree to over-fit the training set. Then the over-fitted tree is cut back into a smaller tree by removing sub-branches that are not contributing to the generalization of accuracy. Employing pruning methods can improve the generalization performance of decision tree, especially in noisy domains (Rokach, L. and Maimon, O., 2006).

There are various techniques for pruning decision trees. The most popular techniques are cost-complexity pruning, reduced error pruning, minimum error pruning, pessimistic pruning, error-based pruning, optimal pruning and minimum description length pruning etc.

A number of different inducers may be used for building decision tree including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), Quest and C5.0. In this study, numerous models were built by using the SPSS software 20 and the default method CHAID inducer was used.

CHAID (Chisquare–Automatic–Interaction–Detection) was originally designed to handle nominal attributes only. For each input attribute a_i, CHAID finds the pair of values in V_i that is least significantly different with respect to the target attribute. The significant difference is measured by the p value obtained from a statistical test. The statistical test used depends on the type of target attribute. If the target attribute is continuous, an F test is used. If it is nominal, then a Pearson chi–squared test is used. If it is ordinal, then a likelihood–ratio test is used.

For each selected pair, CHAID checks if the p value obtained is greater than a certain merge threshold. If the answer is positive, it merges the values and searches for an additional potential pair to be merged. The process is repeated until no significant pairs are found. The best input attribute to be used for splitting the current node is then selected, such that each child node is made of a group of homogeneous values of the selected attribute. Note that no split is performed if the adjusted p value of the best input

attribute is not less than a certain split threshold. This procedure also stops when one of the following conditions is fulfilled:

1. Maximum tree depth is reached.
2. Minimum number of cases in node for being a parent is reached, so it cannot be split any further.
3. Minimum number of cases in node for being a child node is reached.

CHAID handles missing values by treating them all as a single valid category. CHAID does not perform pruning (Rokach, L. and Maimon, O., 2006). CHAID limits itself to categorical variables. Continuous variables need to be changed into ranges or classes. However, one benefit of CHAID is its ability to stop the split before overfitting occurs (Luan, J. 2002).

3.1.5 Advantages and Weaknesses of Decision Trees

Decision trees have several advantages. The tree model does not provide evidence of causality but provides an explanation on how it determined the estimated probability. It can handle all types of variables; and collinearity does not affect the performance of the tree model.

The ***advantages*** of decision trees can be summarized as follows :

1. They are good for classification and prediction
2. They are useful for variable selection
3. They do not require transformation of the variables
4. They are robust to outliers
5. They use nonlinear and non-parametric relationship between the predictors and target variable which allow a wide range of relationships
6. They are useful when classes can be divided through vertical and horizontal splitting of the predictor space
7. They can handle missing values
8. They generate transparent rules useful in managerial applications

The ***weaknesses*** of decision trees can be summarized as follows:

1. They are sensitive to changes in the data
2. The splits are done on single predictor rather than on combinations of predictors
3. They have a lower performance when the best split of the predictor space is diagonal
4. They require a large data set

Decision Tree models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions. These are also useful for exploring data to gaining sight into the relationships of a large number of candidate input variable to the target variable. Because decision trees combine both data exploration and modeling, they are a powerful first step in modeling process even when building the final model using some other techniques (Danso, S.O.,2006).

3.2 Logistic Regression

Logistic regression is helpful when it is needed to predict a categorical variable from a set of predictor variables. Binary logistic regression is similar to linear regression except that it is used when the dependent variable is dichotomous. It is also useful when some of the independent variables are dichotomous and some are continuous. There are fewer assumptions for logistic regression than for multiple regression and for discriminant analysis; this technique has become popular, especially in health related fields. Binary logistic regression assumes that the dependent or outcome variable is dichotomous and, like most other statistics, that the outcomes are independent and mutually exclusive; that is a single case can only be represented once and must be in one group or the other. Moreover, logistic regression requires large samples to be accurate (Leech, N. L., Barrett, K. C. and Morgan, G. A., 2005). According to Leech, Barrett and Morgan, there should be a minimum of 20 cases per predictor, with a minimum of 60 total cases.

Logistic regression refers to methods for describing the relationship between a categorical response variable and a set of predictor variables (Larose, D. T., 2005). Binary logistic regression is more useful when it is needed to model the event probability for a categorical response variable with two outcomes. Using the binary logistic regression procedure, the doctor can determine whether the patient is more likely to be present or absent for particular disease, company manager can estimate the probability that a particular customer will churn and the loan officer can assess the risk of expending credit to a particular customer.

Logistic regression assumes that the relationship between the predictor and the response is nonlinear. In this study, the logistic regression model adopted aims to find the relationship between a dichotomous dependent variable and a set of categorical and continuous attributes. The model and equations used are based on the work of Larose, D. T. (2005). In linear regression, the response variable is considered to be a random

variable $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$ with conditional mean $\pi(x) = E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$. The conditional mean for logistic regression takes on a different form from that of linear regression. Specifically,

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_k x_k}} \quad (3.1)$$

(or)

$$\pi(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_k x_k)}} \quad (3.2)$$

The minimum for $\pi(x)$ is obtained at $\lim_{a\to-\infty}[e^a/1+e^a] = 0$, and the maximum for $\pi(x)$ is obtained at $\lim_{a\to\infty}[e^a/1+e^a] = 1$. Thus, $\pi(x)$ may be interpreted as the probability that the positive outcome is present for records with $X = x$, and $1 - \pi(x)$ may be interpreted as the probability, with $0 \leq \pi(x) \leq 1$. That is $\pi(x)$ may be interpreted as the probability that the positive outcome is present for record with X= x and 1- $\pi(x)$ may be interpreted as the probability that the positive outcome is absent for such records.

The useful transformation for logistic regression is the logit transformation, as follows:

$$g(x) = ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum \beta_i x_i \quad (3.3)$$

From the equation (3.1), the odds can be derived:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_k x_k} = e^{\beta_0+\sum \beta_i x_i} \quad (3.4)$$

Odds may be defined as the probability that an event occurs divided by the probability that the event does not occur. Note that when the event is more likely than not to occur, odds > 1; when the event is less likely than not to occur, odds < 1; and when the event is just likely as not to occur, odds = 1. Note also that the concept of odds differs from the concept of probability, since probability range from zero to one, and odds can range from zero to infinity. Odds indicate how much more likely it is that an event occurred compared to its not occurring (Larose, D. T., 2005).

The probability of the event can be estimated as:

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1+ e^{\hat{g}(x)}} \quad (3.5)$$

The beta parameters are estimated using the method of maximum likelihood estimator and not the least squares method because the relationship between the predictors and the response is nonlinear.

3.3 Model Building

Model building of data mining refers to a series of activities such as deciding on the type of data mining operations (or tasks), selecting the data mining algorithms and mining the data. First, the type of the data mining operation such as data summarization, segmentation classification, prediction, and dependency analysis or association rule discovery, must be chosen. Based on the operations chosen for the application, an appropriate data mining technique is then selected based on the nature of the knowledge to be mined. The next step is selecting a particular algorithm within the data mining technique chosen. Choosing a data mining algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining technique with the overall objective of data mining. After an appropriate algorithm is selected, the data are finally mined using the algorithm to extract novel patterns hidden in databases.

Modeling techniques in data mining include neural networks, case-based reasoning, genetic algorithms, decision trees, association rule, rough set analysis and support vector machines. Each type of model has particular strengths, and is appropriate within specific data mining situations. Model derived from the same sample data can be very different from one algorithm to another. As a result, different models represent the knowledge learned in different formats as well. For example, decision tree represents the knowledge in tree structure, instance-based algorithms, such as nearest neighbor, use the instances themselves to represent what is learned, Naïve Bayes methods represents knowledge in the form of probabilistic summaries. In this study, predictive modeling is applied to the data set available although there are several modeling techniques for the same data mining problem. Predictive modeling permits the value of one variable to be predicted from the known values of other variables.

After the data set is preprocessed, it is analyzed using data mining tool. There are varieties of tools available for data mining such as Weka, Orange, R, SAS and SPSS. Among these tools, SPSS was selected since the package has good GUI interface. In

addition, this package allows for preparation, transformation and modeling algorithms on a dataset. In this study, SPSS software was employed to build decision trees model and logistic regression model. Analysis of the classification models was made in terms of detailed accuracy of the classifier on the training data set as tested on the tested data based on a confusion matrix of each model.

3.4 Evaluation of Model Performance

This section discusses how to measure the quality of the obtained models, either to choose the optimal model or to check the quality of the model to decide whether it is good enough to be used. The purpose of assessing the performance of model is to determine how well the model will behave if it is used in practice. The performance of the data mining model can be measured with the use of the confusion matrix[9].

Classifier Accuracy Measure

Evaluation is the key to making real progress in data mining (Witten, I. H. and Frank, E., 2005). To evaluate performance of classification algorithms, one way is to split samples into two sets, training samples and test samples. Training samples are used to build a learning model while test samples are used to evaluate the accuracy of the model. During validation, test samples are supplied to the model, having their class labels "hidden", and then their predicted class labels assigned by the model are compared with their corresponding original class labels to calculate prediction accuracy. For this purpose, re-sampling techniques in which only a (randomly) chosen subset of data is used for constructing the model and the remaining part is used to compute its quality are used. This is repeated multiple times and in the end the average quality of the model over all trials is reported. The most common used technique used is called cross-validation. If two labels (actual and predicated) of a test sample are same, then the prediction to this sample is counted as a *success*; otherwise, it is an *error* (Witten, I. H. and Frank, E., 2005).

In classification problems, the primary source of performance measurement is a confusion matrix or coincidence matrix (Olson, D. L. and Delen, D., 2008). Table (3.1) shows a confusion matrix for two-class classification problem.

9. For a binary problem, the confusion matrix is a two-dimensional square matrix. The row indexes of a confusion matrix correspond to actual values observed and used for model testing; the column indexes correspond to predicted values produced by applying the model to the test data. For any pair of actual/predicted indexes, the value indicates the number of records classified in that pairing. A confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models.

Table (3.1)

The Confusion Matrix

Actual Class	Predicted Class	
	A	B
A	TP: True Positive	FN: False Negative
B	FP: False Positive	TN: True Negative

The *true positive* (*TP*) and *true negative* (*TN*) are correct classifications in samples of each class, respectively. A *false positive* (*FP*) is when a class B sample is incorrectly predicted as a class A sample; a *false negative* (*FN*) is when a class A sample is predicted as a class B sample. Then each element of a confusion matrix shows the number of test samples for which the actual class is the row and the predicted class is the column.

Some measures of evaluating performance have to be introduced. One common measure is accuracy defined as correct classified instances divided by total number of instances. An often used performance evaluation term is *error rate*, which is defined as the proportion of errors made over a whole set of test samples. Error rate is a measurement of overall performance of a classification algorithm (also known as a classifier); however, a lower error rate does not necessarily imply better performance on a target task. For example, there are 10 samples in class A and 90 samples in class B. If TP = 5 and TN = 85, then FP = 5, FN = 5and error rate is only 10%. However, in class A, there are only 50% samples are correctly classified. To more impartially evaluate the classification results, some other evaluation metrics are constructed:

1. True positive rate (TP rate) = TP/ (TP + FN), also known as *sensitivity* or *recall*, which measures the proportion of samples in class A that are correctly classified as class A.
2. True negative rate (TN rate) = TN/ (FP + TN), also known as *specificity*, which measures the proportion of samples in class B that are correctly classified as class B.
3. False positive rate (FP rate) =FP/ (FP + TN) = 1- specificity.
4. False negative rate (FN rate) =FN/ (TP + FN) = 1- sensitivity.
5. Positive predictive value (PPV) =TP/ (TP + FP), also known as *precision*, which measures the proportion of the claimed class A samples are indeed class A samples.
6. Accuracy = (TP + TN)/(TP+TN+FP+FN) = (TP + TN)/ Total Classes

If the number of samples for training and testing is limited, a standard way of predicting the error rate of a learning technique is to use stratified k-fold cross validation. In k-fold cross validation, first, a full data set is divided randomly into k disjoint subsets

of approximately equal size, in each of which the class is represented in approximately the sample proportions as in the full data set (Witten, I. H. and Frank, E., 2005). Then the above process of training and testing will be repeated k times on the k data subsets. In each iteration, (1) one of the subsets is held out in turn, (2) the classifier is trained on the remaining (k – 1) subsets to build classification model, (3) the classification error of this iteration is calculated by testing the classification model on the holdout set. Finally, the k numbers of errors are added up to yield an overall error estimate. Obviously, at the end of cross validation, every sample has been used exactly once for testing.

A widely used selection for k is 10. Empirical studies showed that 10 seen to be an optimal number of folds (that optimize the time it takes to complete the test while minimizing the bias and variance associated with the validation process). In 10-fold cross-validation, the entire dataset is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as original dataset. Each fold is used once to test the performance of the classifier that is generated from the combined data of the nine folds, leading to 10 independent performance estimates (Delen, D.,Walker, G. and Kadam, A., 2004).

CHAPTER 4

DESCRIPTIONS AND PREPROCESSING OF TUBERCULOSIS DIAGNOSIS DATA SET

In this study, the secondary data of medical domain were used to develop classification models in order to diagnose for diseases. The data set consists of data instances (patient records). Each instance is described by features or values of variables (in statistics) or attributes (in machine learning). These features can be numerical (continuous) or discrete (symbolic, ordinal or nominal). Often there is a certain variable that has to be predicted. Statisticians call it 'dependent variable', in medicine it can be called the term 'outcome' and in machine learning it will be called 'class' or class attribute (Demsar, J., 2010).

4.1 Source of Data

This study focuses on classification (diagnosis) of a particular disease for a patient; existence or non-existence. The medical data set used in this study is tuberculosis diagnosis data set which was obtained from Latha and Aung San Townships, UTI in Myanmar. This medical data set is based on the records of clinic attendants who have been registered during the period of two months (1st September to 31st October, 2013) by UTI in Myanmar. The data set includes information on 659 patients who were examined at a clinic. Each of those records consists of 34 different variables: one dependent variable (outcome: TB or Non-TB) and 33 independent variables (predictors). The full list of variables is as follows:

Table (4.1)

Variables and its Domain in Tuberculosis Diagnosis

Variables	***Domain***
1. Gender	Female = 0, Male=1
2. Age	Age in years
3. Weight	Weight in kilogram
4. Smoking[10]	None=0, Little(<5) = 1, Moderate(5-10) = 2, Very Much(11+) = 3
5. Alcohol	No = 0, Yes =1
6. BCG Vaccine	No = 0, Yes =1

10. None = 0 (a person is a non-smoker), Little =1 (a person smokes less than 5 cigarettes per day), Moderate = 2 (a person smokes between 5 and 10 cigarettes per day), Very Much = 3 (a person smokes more than 10 cigarettes per day)

Table (4.1) Variables and Its Domain in Tuberculosis Diagnosis (continued)

7. Malaise	No = 0, Yes =1
8. Arthralgia	No = 0, Yes =1
9. Exhaustion	No = 0, Yes =1
10. Unwillingness for work	No = 0, Yes =1
11. Loss of Appetite	No = 0, Yes =1
12. Loss in Weight	No = 0, Yes =1
13. Sweating at Night	No = 0, Yes =1
14. Chest Pain	No = 0, Yes =1
15. Back Pain	No = 0, Yes =1
16. Coughing	No = 0, Yes =1, With Mucous = 2
17. Hemoptysis	No = 0, Yes =1
18. Fever	Normal =0, High =1, Subfebrile =2
19. Migration	No = 0, Yes =1
20. Diabetes	No = 0, Yes =1
21. ESR	No = 0, Yes =1
22. Haematocrit	Normal =0, Low = 1, High =2
23. Haemoglobin	Normal =0, Low = 1, High =2
24. Leucocytes	Normal =0, Low = 1, High =2
25. No. of Leukocytes Type	Normal =0, Low = 1, High =2
26. Active Specific Lung Lesion	No = 0, Yes =1
27. Calcific Tissue	No = 0, Yes =1
28. Cavity	No = 0, Yes =1
29. Pneumonic Infiltration	No = 0, Yes =1
30. Pleural Effusion	No = 0, Yes =1
31. HIV	Negative = 0, Positive =1
32. Sputum AFB[11]	Negative = 0, Positive =1
33. Gxpert	Negative = 0, Positive =1
34. Outcome	Non-TB = 0, TB = 1

11. Sputum stain for mycobacteria is laboratory test performed on a sample of the patient's sputum (phlegm). It is also known as Acid Fast Bacillus stain (AFB) or tuberculosis (TB) smears.

4.2 Nature of the Tuberculosis Diagnosis Data Set

The data were in a hard copy format with 659 records and 38 variables. For the purpose of the analysis in this study, one variable will be chosen as the dependent variable which is to be predicted. The dependent variable is outcome or class variable. It takes the value of 'one' in the case of existence TB and 'zero' in the case of non-TB. Thirty-one of the 33 predictors are categorical or discrete variables and 2 variables (age and weight) have the numeric value.

The records of patients were collected by healthcare workers. The categorical value for the variable 1 to 16, 19, 20 and 31 are obtained by discussing with patient. The categorical value for variable 18 can be measured by thermometer. The result of categorical value for the variable 21 to 25 was obtained by performing blood testing. The result of categorical value for the variable 26, 28, 29 and 30 can be recorded by doctor's decision with the use of X-ray result. For the result of categorical value for the variable 27 (Calcific Tissue), Ultrasound examination is done. The result of the variable 32 (Sputum AFB) is obtained from which sample of sputum is tested in laboratory to diagnose the TB disease.

In the records of the data set of 659 patients, there are 425 patients (64%) with TB and the remaining 234 patients (36%) without TB. In this data set, there are many instances containing more than one missing (unavailable) attribute value which is denoted by "?".

4.3 Data Presentation

The main source of the data used to undertake this study was patients' real data taken from the UTI, Yangon. Among 659 records, all records which have missing value on basic attributes (such as gender, age, and weight) were excluded for this study. The remaining data set has 601 records and some records retain missing value on some variables. First, all the data were encoded in an Excel format. After the data have been encoded, the entire data set was put in one file having many records. Next, preprocessing techniques were applied to make it appropriate for the mining purpose.

Exploring on Some Attributes

Distribution of TB Disease

Table (4.2) shows the distribution of TB disease for the people who come to UTI for their medical check-up.

Table (4.2)

Distribution of TB Disease

Outcome	Frequency (Number of Patients)	Percent
Non-TB	210	34.9%
TB	391	65.1%
Total	601	100.0%

Source: Tuberculosis Diagnosis Data Set, UTI

In the records of the data set of 601patients, there are 391 patients (65%) with TB and the remaining 210 patients (35%) without TB. The distribution of TB disease was described as pie chart in Figure (4.1).

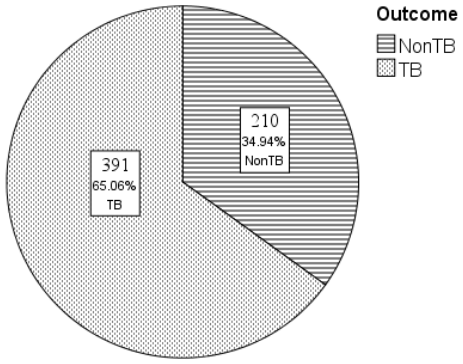


Figure (4.1) Pie Chart for TB Disease

Source: Table (4.2)

Gender

Both Table (4.3) and Figure (4.2) show the number of persons who belong to each class (TB or Non-TB) by gender.

Table (4.3)

Number of TB Suspected Patients by Gender

Gender	Outcome				Total
	Non-TB	%	TB	%	
Female	80	38%	135	35%	215
Male	130	62%	256	65%	386
Total	210	100%	391	100%	601

Source: Tuberculosis Diagnosis Data Set, UTI

According to Table (4.3), among all the patients who were examined at clinic, male is more than female. Among 391 TB positive patients, 256 (65 %) are male and 135

(35%) are female. It was also found that male is more likely to suffer from TB than female. Because male have more social activities than female, they are more likely to be affected by TB.

According to Figure (4.2a), among the 601 TB suspected patients, 22% are female TB positive, 13% are female TB negative, 43% are male TB positive and 22% are male TB negative. As shown in Figure (4.2b), the rate of TB disease of male exceeds that of female. If a person was male, the probability of existence of TB would be 0.6631. The probability of existence of TB would be 0.6279 for female.

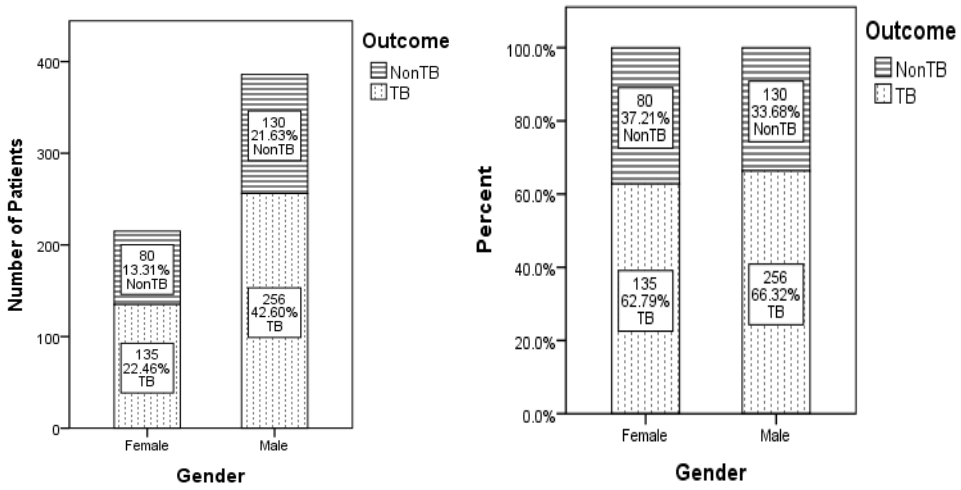


Figure (4.2) Bar Chart for TB Disease by Gender

Source: Table (4.3)

Age

Table (4.4) illustrates the number of patients who have TB disease or not by their age group.

Table (4.4)

Number of TB Suspected Patients by Age Group

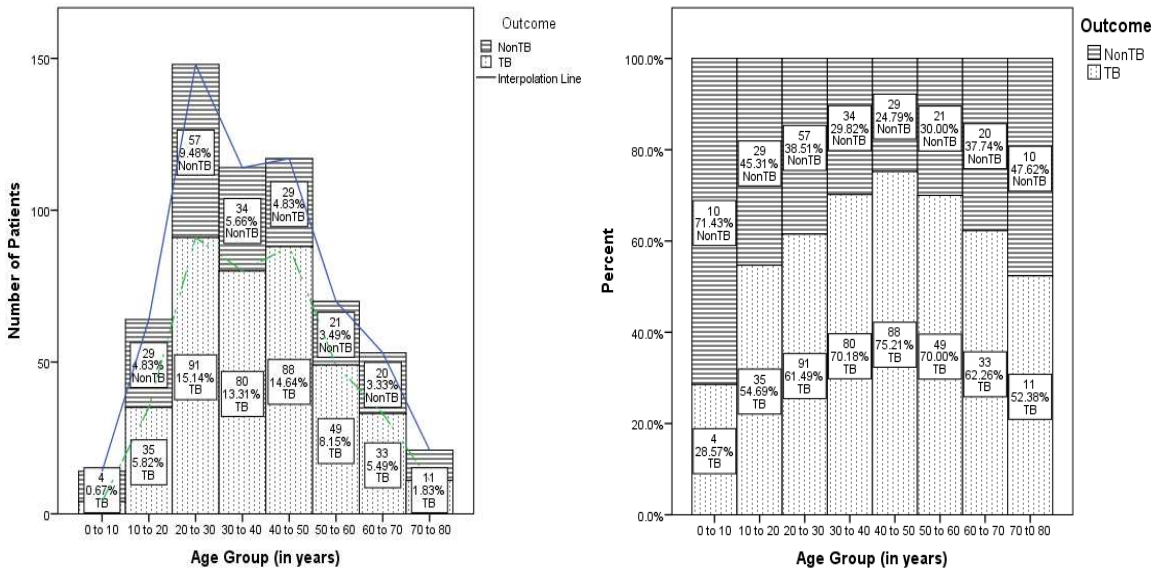
Age (years)	Outcome				Total
	Non-TB	%	TB	%	
0 to 10	10	5%	4	1%	14
10 to 20	29	14%	35	9%	64
20 to 30	57	27%	91	23%	148
30 to 40	34	16%	80	20%	114
40 to 50	29	14%	88	23%	117
50 to 60	21	10%	49	13%	70
60 to 70	20	9%	33	8%	53
70 to 80	10	5%	11	3%	21
Total	210	100%	391	100%	601

Source: Tuberculosis Diagnosis Data Set, UTI

According to Table (4.4), almost the people who were examined at clinic are at the age between 21years and 50 years. It was also noted that TB disease mostly occurred at the age between 21 to 50 years. It was found that 23 % of TB positive is aged between 21 to 30 years and 41 to 50 years and 20% of TB positive is aged between 31 to 40 years. Among the TB positive patients, the incidence of smear positive TB case is low at the age of 1 to 10 years and old-aged adult 71 to 80 years.

The bar chart for Table (4.4) was drawn in Figure (4.3). As shown in Figure (4.3b), the highest rate of existence of TB disease was between the age group of 41to 50 years. The probability of existence of TB disease in this age group is 0.75 and thus there is 75 people have TB positive per 100 people among age group of 41 to 50 years. The second highest rate occurs at the age group of 31 to 40 years and 51 to 60 years. The probability of existence of TB disease in these age groups is 0.70 and it means that there is 70 people have TB positive per 100 people among age group of 31 to 40 years and 51 to 60 years.

According to the age group data, it could be concluded that the incidence of smear positive TB case is high among 21 to 50 years old. Because people at the age of 21 to 50 are the major workforce of the country, they work with many people from different societies. Thus, they are infected by TB disease from other people who have TB disease in their societies.



(a) **(b)**

Figure (4.3) Bar Chart for TB Disease by Age Group

Source: Table (4.4)

Smoking Habit

Table (4.5) shows the distribution of number of TB suspected patients by their smoking habit and it is illustrated as bar chart in Figure (4.4).

Table (4.5)

Number of TB Suspected Patients by Smoking Habit

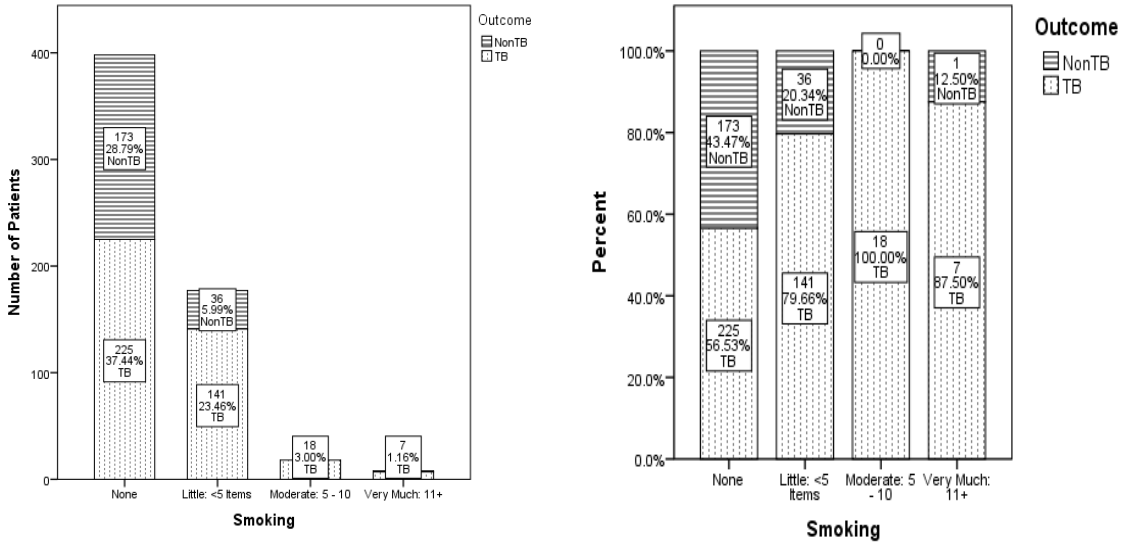
Smoking (no. of cigarettes)	Outcome				Total
	Non-TB	%	TB	%	
None	173	82.5%	225	58%	398
Little: < 5 Items	36	17.0%	141	36%	177
Moderate: 5- 10	0	0%	18	5%	18
Very Much: 11+	1	0.5%	7	1%	8
Total	210	100.0%	391	100%	601

Source: Tuberculosis Diagnosis Data Set, UTI

Table (4.5) shows that most of the people who were examined at clinic are do not smoke. It indicates that 58% (225 patients out of 391 TB positive patients) are non-smokers, 42% (166 out of 391 TB positive patients) are smokers.

According to Figure (4.4), the probability of existence of TB for non-smoker people is 0.5653. It can also be seen that there is hundred percent of occurrence of TB

disease for people who smoked moderately and eighty-eight percent of occurrence of TB disease for people who smoked heavily.



(a) (b)

Figure (4.4) Bar Chart for TB Disease by Smoking Habit

Source: Table (4.5)

Drinking Habit

Table (4.6) shows the distribution of number of TB suspected patients by their drinking habit and it is illustrated as bar chart in Figure (4.5).

Table (4.6)

Number of TB Suspected Patients by Drinking Habit

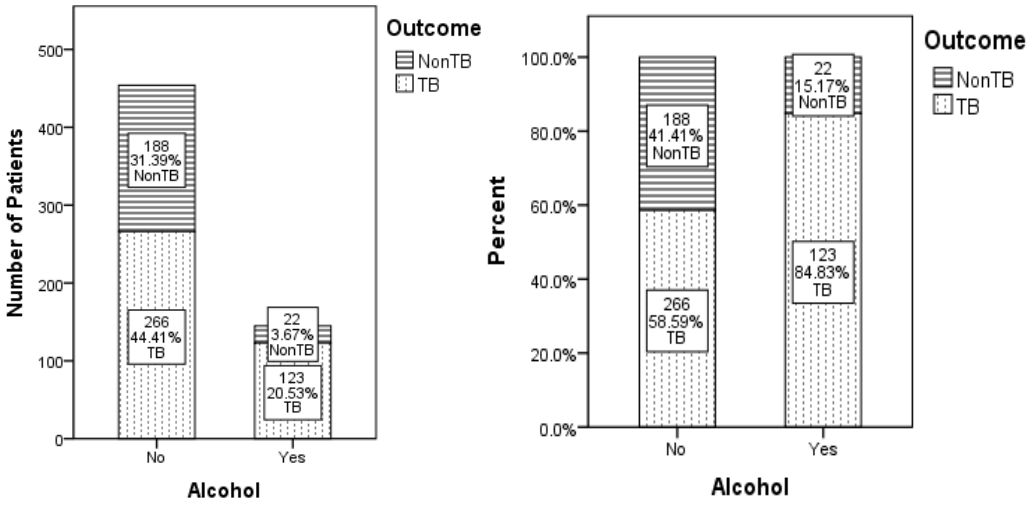
Alcohol	Outcome				Total
	Non-TB	%	TB	%	
No	188	90%	266	68%	454
Yes	22	10%	123	32%	145
Missing	0	0%	2	0%	2
Total	210	100%	391	100%	601

Source: Tuberculosis Diagnosis Data Set, UTI

As shown in Table (4.6), among the 391 TB positive patients, 266 patients (68%) are non-alcoholic and 123 patients (32%) are alcoholic. It was also found that among 210 TB negative patients, 188 (90%) are non-alcoholic and 22 (10%) are alcoholic.

According to Figure (4.5), it can be seen that the probability of occurrence of TB disease for drinkers is greater than that of people who do not drink alcohol. The

probability of existence of TB for alcoholic is 0.85 and the probability of TB positive for non-alcoholic is 0.59.



(a) **(b)**

Figure (4.5) Bar Chart for TB Disease by Drinking Habit

Source: Table (4.6)

4.4 Data Preprocessing

The data preprocessing phase covers all activities to construct the final data set from the initial raw data. Once the data resources available are identified, these data need to be selected, cleaned, build into the form desired, and transformed. Data selection, data cleaning and data transformation in preprocessing of data modeling need to occur in this phase and data exploration at a greater depth can be applied during this phase. Based on the discoveries in the exploration phase, one may need to manipulate data to include information such as the grouping of data unit and significant subgroups, or to introduce new variables. It may also be necessary to reduce the number of variables to narrow them down to the most significant ones (Olson, D. L. and Delen, D., 2008). This study was made use of the MS-Excel application and SPSS software package in order to perform data preprocessing process. Figure (4.6) shows the flow of data preprocessing for this analysis.

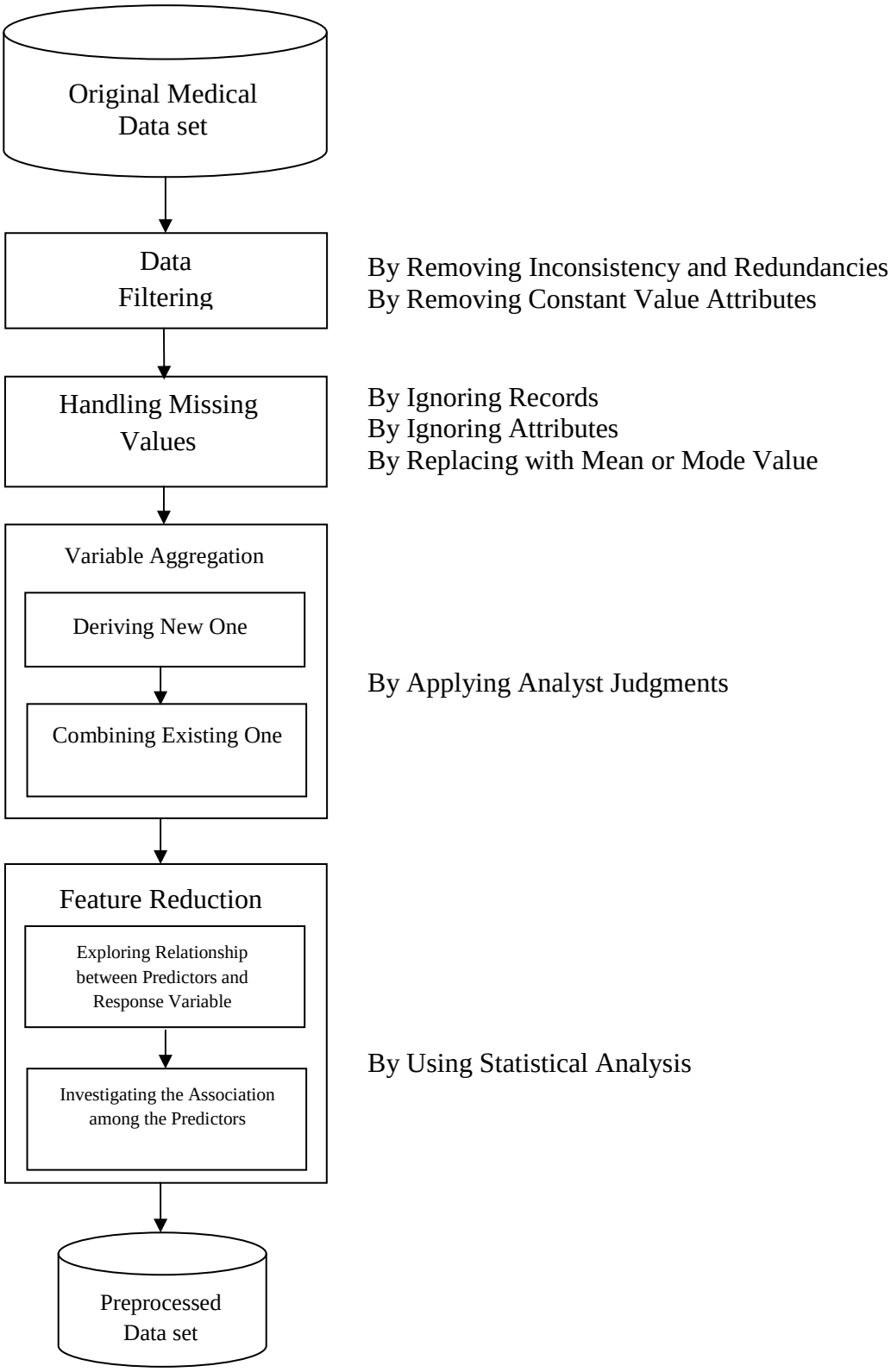


Figure (4.6) Flow of Data Preprocessing
Source: Own Compilation

4.4.1 Data Filtering

The process of removing unnecessary information, inconsistent information and irrelevant information through the removal of observations (records or attributes) is called data filtering. In the tuberculosis diagnosis data set, some records (instances) were removed because they have inconsistency and irrelevant information. For example, a person who is 10 years old but his or her weight is over 50 kg. This condition is impossible in real cases. Again, it is also necessary to check that there is constant value attribute for some variables. In such a situation, these variables should be removed from the data set. In this data set, migration variable has constant value attribute because more than 86% of the records are missing value attributes. Moreover, about 60% of the records have missing value and 30% of the records have negative attribute value (0) on diabetes variable. Therefore, these two variables were removed for developing classification model. Again, 10 variables (Haematocrit, Haemoglobin, Leucocytes, No. of Leukocytes Type, Calcific Tissue, Cavity, Pneumonic Infiltration, Pleural Effusion, HIV, and Gxpert) were removed from the data set because more than 90% of the records are missing for those variables. After performing data filtering, the data set remains 601 records with 20 predictors. Remaining variables are Gender, Age, Weight, Smoking, Alcohol, BCG Vanccine, Malaise, Arthralgia, Exhaustion, Unwillingness for work, Loss of Appetite, Loss in Weight, Sweating at Night, Chest Pain, Back Pain, Coughing, Hemoptysis, Fever, Active Specific Lung Lesion, Sputum AFB. The remaining variables used for developing classification model are shown in Table (4.7) together with variables removed which are shaded in gray color.

Table (4.7)

Prediction Variables Used for Model Building

1. Gender	12. Loss in Weight	24. Leucocytes
2. Age	13. Sweating at Night	25. No. of Leukocytes Type
3. Weight	14. Chest Pain	26.Active Specific Lung Lesion
4. Smoking	15. Back Pain	
5. Alcohol	16. Coughing	27. Calcific Tissue
6. BCG Vanccine	17. Hemoptysis	28. Cavity
7. Malaise	18. Fever	29. Pneumonic Infiltration
8. Arthralgia	19. Migration	30. Pleural Effusion
9. Exhaustion	20. Diabetes	31. HIV
10.Unwillingness for work	21. ESR	32. Sputum AFB
	22. Haematocrit	33. Gxpert
11. Loss of Appetite	23. Haemoglobin	34. Outcome

4.4.2 Handling Missing Values

A lot of missing values make it especially difficult to build good models and draw any medical conclusions. In the tuberculosis diagnosis data set, there are many instances containing more than one missing (unavailable) attribute value which is denoted by "?". Table (4.8) describes the count of missing value for each variable.

Previously, every attribute or variable which consists of more than 30% of records with missing value has been ignored. Now, all records or instances which contain more than 30% missing attributes value were ignored. In the tuberculosis diagnosis data set, there are two records which have more than 30% missing attributes value. Therefore, the data set remains 599 records with 20 predictors. After that, remaining missing values were replaced by the mean value (in Age, missing value are replaced with the mean age) or represented by the mode (in Smoking habit, missing value are replaced with the mode value).

Table (4.8)

Missing Value Analysis

Variable	N	Missing	
		Count	Percent
Age	601	0	.0
Weight	601	0	.0
Gender	601	0	.0
Smoking	601	0	.0
Alcohol	599	2	.3
BCG_Vaccine	593	8	1.3
Malaise	596	5	.8
Arthralgia	596	5	.8
Exhaustion	597	4	.7
Unwillingnes_for_Work	592	9	1.5
Loss_of_Appetite	597	4	.7
Loss_in_Weight	599	2	.3
Sweating_at_Nights_	588	13	2.2
Chest_Pain	599	2	.3
Back_Pain	599	2	.3
Coughing	599	2	.3
Hemoptysis	592	9	1.5
Fever	592	9	1.5
Active_Specific_Lung_Lesion	579	22	3.7
Sputum_AFB	578	23	3.8
Outcome	601	0	.0

Source: Tuberculosis Diagnosis Data Set, UTI

4.4.3 Variable Aggregation

Analyst judgment is critical to decide including significant variables (Olson, D. L. and Delen, D., 2008). Too many variables produce too much output, while too few can overlook key relationships in the data. Rule of thumb in statistics is that the sample size must be at least 30 times the number of variables (Demsar, J., 2010). Therefore, it is needed to reduce some variables. Variable aggregation can be performed by analyst judgment and it can be divided into two sub-activities: deriving new one and combining some into existing one.

Deriving New Variables

In this phase, the resulted data set contains 599 records together with 20 predictors. It is needed to combine two or more variables into new one by using the aggregate/ integrate rule in order to get relevant and sufficient variables. In this case, new variable Weight_Condition was derived by combining Gender, Age and Weight variable. Weight_Condition variable is categorical variable and its attribute values were underweight, normal, and overweight. These values were determined by the rule that is, what specified weight should fall in any specified age range by gender.

Combining Some into Existing One

Some variables are needed to reduce from the data set because of having overlap meaning. In the tuberculosis diagnosis data set, there is no existing one that is derived from combining some variables.

After performing variable aggregation, the variables: Gender, Age and Weight were removed from this data set and new variable: Weight_Condition variable was added for continuous study instead of three variables. Therefore, the remaining data set consists of 18 predictors and this data set was used for Algorithm IV.

4.4.4 Feature (Dimensionality) Reduction

In data mining, a problem in classification and prediction is to find ways to reduce the dimensionality of the variable or feature space to overcome the risk of over-fitting. Data over-fitting happens when the number of variables is large and the number of training sample is comparatively small. In such situation, a decision function can perform well on classifying training data, but does poorly on the test sample.

There are several methods that aim at solving these problems by reducing the number of features and thus the dimensionality of the data. The general objectives for reducing the number of features are to avoid over-fitting and thus to improve the prediction performance of classification algorithms, reduce the computational cost of the training and prediction phase, and also provide a better understanding of the achieved results. Dimensionality reduction is one popular technique to remove irrelevant and redundant attributes. Feature (dimensionality) reduction techniques can be categorized mainly into feature extraction and feature selection. In feature extraction approach, features are projected into a new space with lower dimensionality. On the other hand, the

feature selection approach aims to select a small subset of features that minimize redundancy and maximize relevance to the target (class label).

Exploring the Relationship between the Predictors and Response

To create useful models, it is critical to identify the initial set of variables that will be used in the data mining process. Selection of too few variables can result in an incomplete analysis and may result in excluding critical factors from the final model. On the other hand, inclusion of irrelevant variables can adversely affect the model building process and can affect the correct identification of significant factors. The identification of initial set of variables for use in the data mining process requires domain knowledge and a deep understanding of the data set.

In this section, the investigation of variable-by-variable association between the predictors and the target variable ‘outcome’ was performed. In order to conduct variable selection, test of independence (chi-square test) was used. After conducting variable selection using chi-square test, 16 predictors was selected from 18 predictors and the variables selected are Smoking, Alcohol, BCG Vaccine, Malaise, Arthralgia, Exhaustion, Unwillingness for work, Loss of Appetite, Loss in Weight, Sweating at Night, Chest Pain, Back Pain, Coughing, Fever, Active Specific Lung Lesion and Sputum AFB. The variables reduced are Hemoptysis and Weight_Condition because they are less significant than other variables.

Investigating the Association among the Predictors

The selected variables for relevant data should be independent of each other. Variable independence means that the variables do not contain overlapping information. A careful selection of independent variables can make it easier for data mining algorithms to quickly discover useful knowledge patterns. In this section, clustering analysis was used to group features (variables) in order to reduce features. The following table illustrates the output of cluster membership for 16 predictors.

Table (4.9)

Cluster Membership

Predictors	10 Clusters
Smoking	1
Alcohol	2
BCG Vaccine	3
Malaise	4
Arthralgia	4
Exhaustion	4
Unwillingnes for Work	4
Loss of Appetite	5
Loss in Weight	5
Sweating at Nights	6
Chest Pain	7
Back Pain	7
Coughing	8
Fever	9
Active Specific Lung Lesion	5
Sputum AFB	10

Table (4.9) shows the homogeneous group of variables for 16 predictors and it can be illustrated by dendrogram in Figure (4.7).

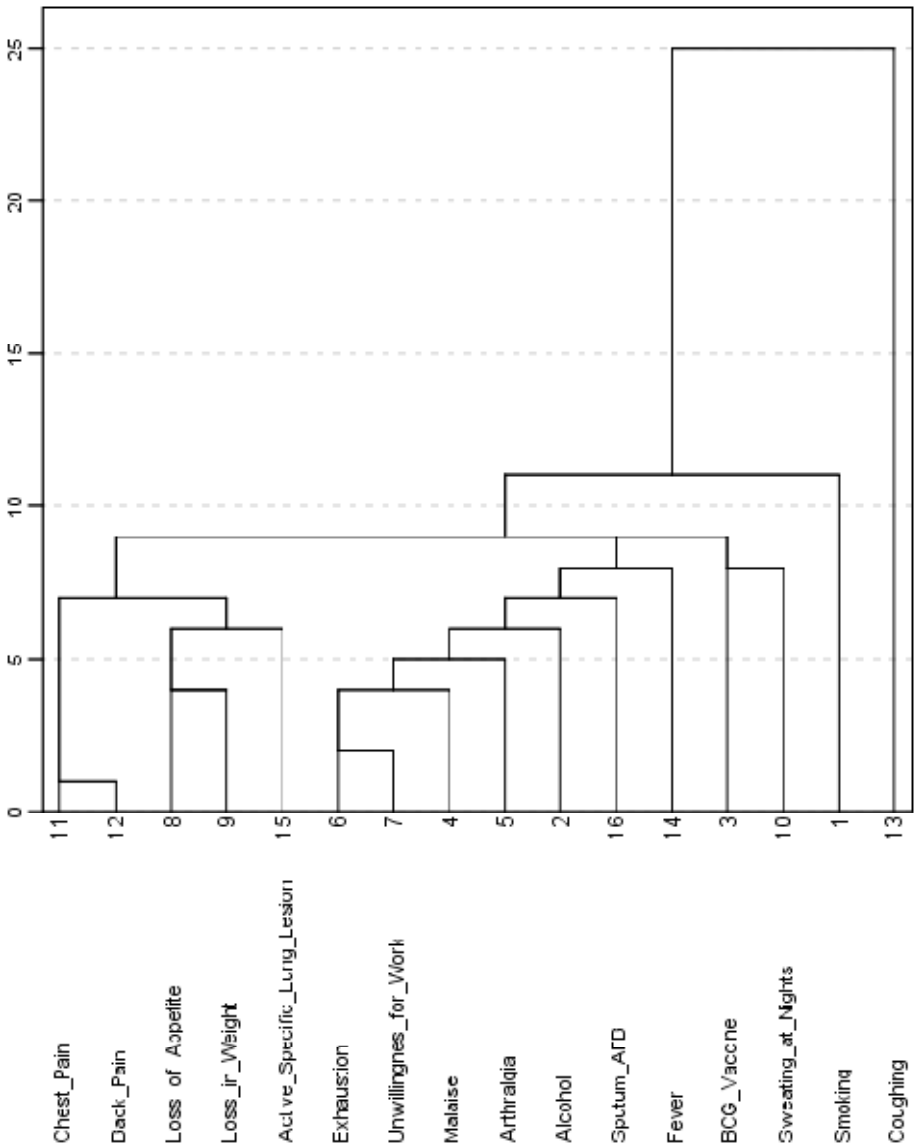


Figure (4.7) Dendrogram for Clustering 16 Predictors

Cluster 4, 5 and 7 have more than one variable. That is four variables (Malaise, Arthralgia and Exhaustion and Unwillingness for Work) have the same meaning or same direction because they fall in cluster 4. Also cluster 5 and 7 consist of two variables and three variables respectively. In such a case, the most significant one was extracted for each cluster by using logistic regression. Exhaustion for cluster 4, Back Pain for cluster 5 and Active Specific Lung Lesion for cluster 7 were selected for next algorithm. Therefore, the final data set consists of 10 predictors (Smoking, Alcohol, BCG Vanccine, Exhaustion, Sweating at Night, Back Pain, Coughing, Fever, Active Specific Lung Lesion and Sputum AFB) and this data set was used for Algorithm V in order to construct decision tree model.

CHAPTER 5

CLASSIFICATION MODELS AND EVALUATION OF CLASSIFICATION MODELS

Model building is the process of searching for a good model from some candidate models. In this study, the tuberculosis diagnosis data set was used for developing decision tree models and logistic regression model, and a comparison of their accuracy was made. The main aim of this study is to apply different algorithms in order to predict existence of mycobacterium tuberculosis bacteria on patients. In this section, the resulted models of each algorithm are presented and performances of each model are compared.

5.1 Algorithms for Tuberculosis Diagnosis

The followings are the different algorithms used for developing decision tree model on tuberculosis diagnosis data set. In order to examine the importance of preprocessing in data mining, the different algorithms are employed in model developing. Moreover, it can be investigated that decision tree can serve as for data exploration.

Algorithm I

Step 1: Use No-preprocessed Data Set Step 2: Run Decision Tree Analysis

Algorithm II

```
Step 1: Perform data filtering
        LOOP: For each record
                IF record is inconsistent
                        THEN remove it
        END LOOP
        LOOP: For each variable
                Count same attribute value
                Count missing value
                IF count of same value attribute > 75%
                        THEN remove it
                IF count of missing > 30%
                        THEN remove it
        END LOOP
Step 2: Fill missing value for each variable with mode value of all records
        LOOP: For each variable
                Find mode value
                IF attribute value is missing
                        THEN insert mode value
        END LOOP
Step 3: Run Decision Tree Analysis
```

Algorithm III

```
Step 1: Perform data filtering
        LOOP: For each record
                IF record is inconsistent
                        THEN remove it
        END LOOP
        LOOP: For each variable
                Count same attribute value
                Count missing value
                IF count of same value attribute > 75%
                        THEN remove it
                IF count of missing > 30%
                        THEN remove it
        END LOOP
Step 2: Perform record clustering for specified variable
        LOOP: For each variable
                Cluster the records by homogeneity
                Count number of records in same cluster
                Find mode value of each cluster
        END LOOP
Step 3: Fill missing value for each variable with cluster mode value
        LOOP: For each variable
                IF record is missing
                        THEN Check cluster number
                         Insert same cluster mode value
        END LOOP
Step 4: Run Decision Tree Analysis
```

Algorithm IV

Step 1: Perform data filtering (procedures same as in Algorithm II)
Step 2: Fill missing value for each variable with mode value of all records
Step 3: Perform variable aggregation
Step 4: Run Decision Tree Analysis

Algorithm V

Step 1: Perform data filtering
Step 2: Fill missing value for each variable with mode value of all records
Step 3: Perform variable aggregation
Step 4: Perform feature reduction
Step 5: Run Decision Tree Analysis

Initially, 38 attributes were identified in tuberculosis diagnosis data set, out of which 4 attributes (Sr. No., Name, Township, and Date) were discarded as the attributes were not relevant to the research in hand. Original data set (no preprocessed data set) which has 34 variables is used for Algorithm I. After performing for data filtering and missing value handling, the resulted data set consists of 21 variables (20 predictor variables and 1 dependent variable). This data set was used for Algorithm II and III and it contains 599 records of which 390 records were with TB and remaining 209 records were without TB. After performing data filtering, missing value handling and variable aggregation, the resulted data set consists of 19 variables (18 predictor variables and 1 dependent variable). Therefore, eighteen predictors were used for Algorithm IV. After data preprocessing which includes data filtering, missing value handling variable aggregation and feature reduction, the resulted data set consists of 11 variables (10 predictor variables and 1 dependent variable). This final data set was used for Algorithm V.

5.2 Different Classification Models for Tuberculosis Diagnosis

Classification task is to determine the unknown class of new instants (diagnose for newly arrived patients). In this subsection, classification models were derived with different algorithms by using decision tree method. This study made use of the SPSS software package to build decision tree models. This software partitions the data set prepared for analysis in training and test facts where training facts are used to train and

build the models, and test facts are used to test the performance of the model. By default, the software automatically sets 10-fold cross-validation for testing purposes. For this study, the default method CHAID (Chi-squared Automatic Interaction Decision) was used. At each step, CHAID choose the independent (predictor) variables that have the strongest the interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different with respect to the dependent variable. In this study, numerous models were built, and performance of the model was tested to check its efficiency and effectiveness. Finally, the confusion matrix was used to evaluate the accuracy and performance of the models built with the decision tree method.

5.2.1 Decision Tree Model of Algorithm I

For Algorithm I, the original tuberculosis diagnosis data set which consists of 659 records with 33 predictors was used in order to produce results of classification model. In this Algorithm, the classification model was conducted by using decision tree analysis directly without preprocessing on data set.

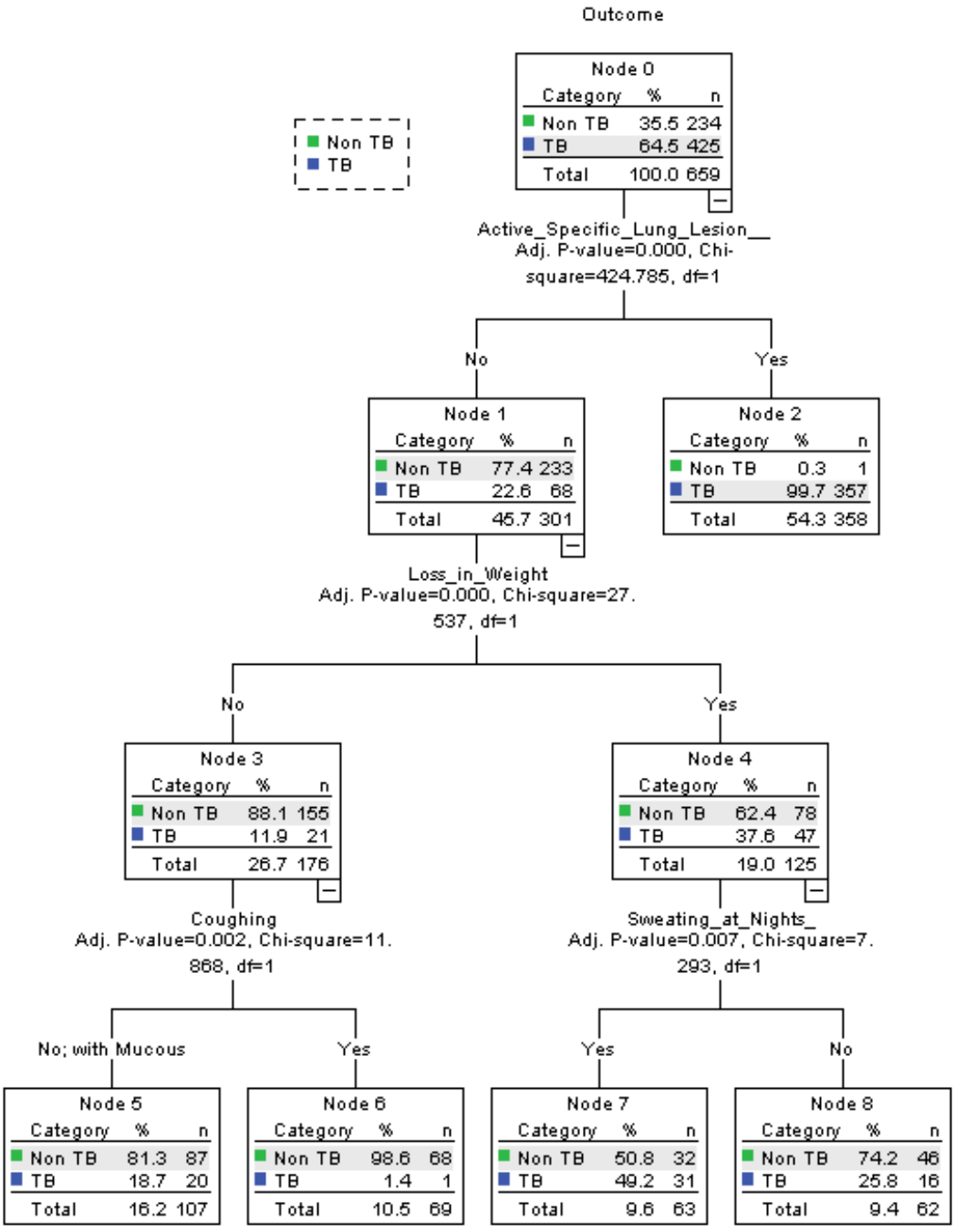


Figure (5.1) Decision Tree for Tuberculosis Diagnosis using Algorithm I with 33 Predictors

Although all thirty-three independent variables which are included in original data set were specified for algorithm I, only four were included in the decision tree model and these are Active Specific Lung Lesion, Loss in Weight, Coughing and Sweating at Nights. Twenty-nine variables did not make a significant contribution to the model; therefore, they were automatically dropped from the tree model of Algorithm I. As shown in Figure (5.1), Active Specific Lung Lesion variable is the best predictor of TB

diagnostic. The categorical value ‘Yes’ of Active Specific Lung Lesion is only significant predictor of TB disease. Of the TB suspected patients in this categorical value, 99.7% have TB disease. Since there are no child nodes below it, this is considered a terminal node. For the categorical value ‘No’ on Active Specific Lung Lesion, the next best predictor is Loss in Weight. For the categorical value ‘No’ on Loss in Weight, the next best predictor is Sweating at Nights and for the value ‘No’ on that, the next best predictor is Coughing. Table (5.1) describes a set of rules which are generated from the decision tree classification model of Algorithm I.

Table (5.1)

Classification Rules and Likelihood of TB using Algorithm I

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has the categorical value ‘Yes’ on Active Specific Lung Lesion	99.7%
2	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she has Loss in Weight AND he/she feels Sweating at Nights	49.2%
3	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she has Loss in Weight AND he/she does not feel Sweating at Nights	25.8%
4	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘No’ or ‘With Mucous’ on Coughing	18.7%
5	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘Yes’ on Coughing	1.4%

5.2.2 Decision Tree Model of Algorithm II

In Algorithm II, reduced data set which had been conducted for missing value handling was used. This data set consists of 599 records with 20 predictors. Although 20 independent were specified for Algorithm II, but only three were included in the tree model and these are Active Specific Lung Lesion, Loss in Weight and Coughing. Seventeen predictors did not make a significant contribution to the model; therefore, these variables were automatically dropped from the tree model of Algorithm II.

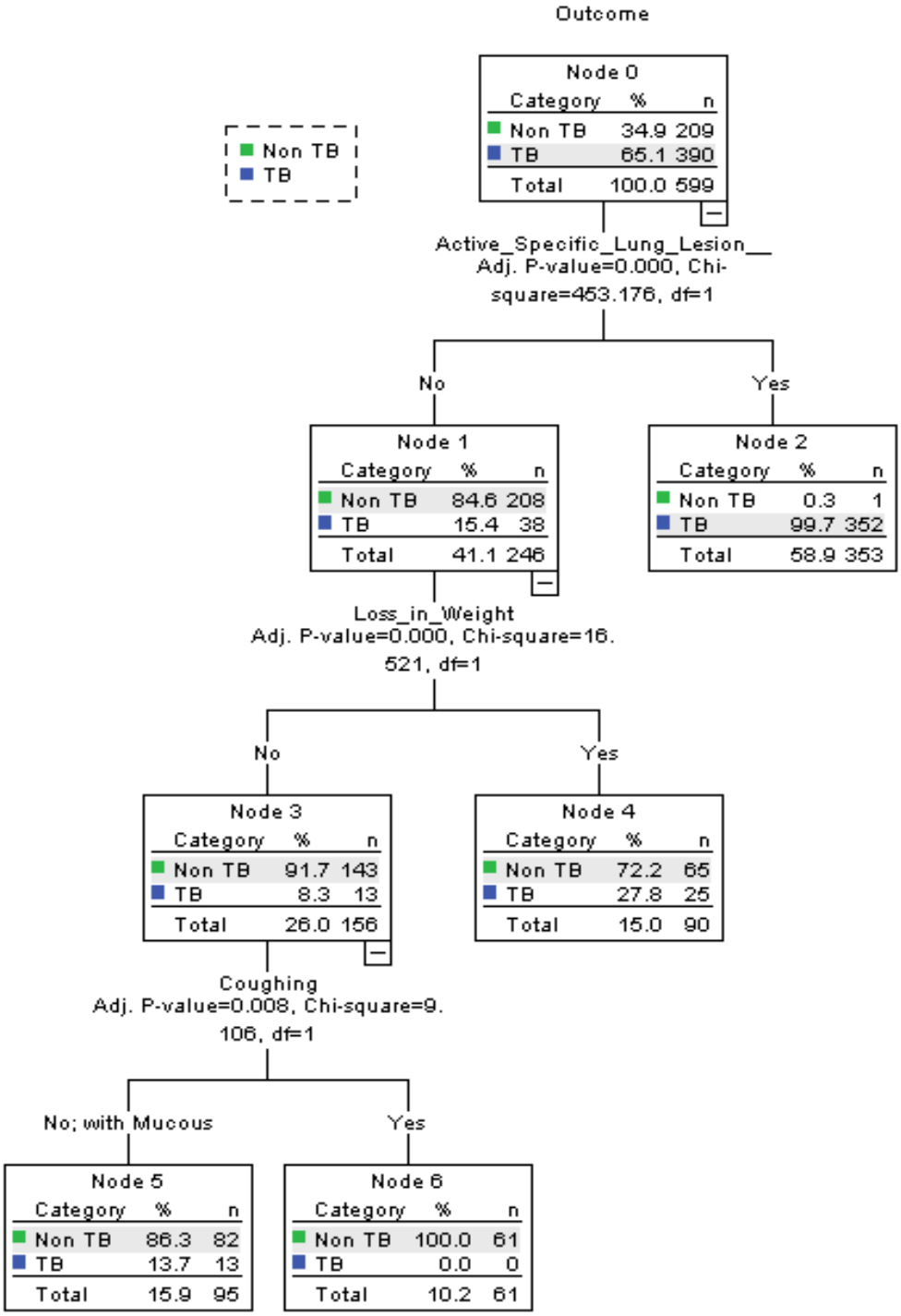


Figure (5.2) Decision Tree for Tuberculosis Diagnosis using Algorithm II with 20 Predictors

As shown in Figure (5.2), Active Specific Lung Lesion variable is the best predictor of TB diagnostic. The categorical value 'Yes' on Active Specific Lung Lesion is only significant predictor of TB disease. Of the TB suspected patients in this categorical value, 99.7% have TB disease. This value is same as in tree model of Algorithm I but one more predictor (Sweating at Nights) was dropped in this tree model.

Some of the rules of Algorithm II for diagnosis of TB disease are summarized in Table (5.2). For this Algorithm, the data set which had been reduced by performing both data filtering and missing value handling was used.

Table (5.2)

Classification Rules and Likelihood of TB using Algorithm II

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has the categorical value ‘Yes’ on Active Specific Lung Lesion	99.7%
2	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she has Loss in Weight	27.8%
3	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘No’ or ‘With Mucous’ on Coughing	13.7%
4	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘Yes’ on Coughing	0%

5.2.3 Decision Tree Model of Algorithm III

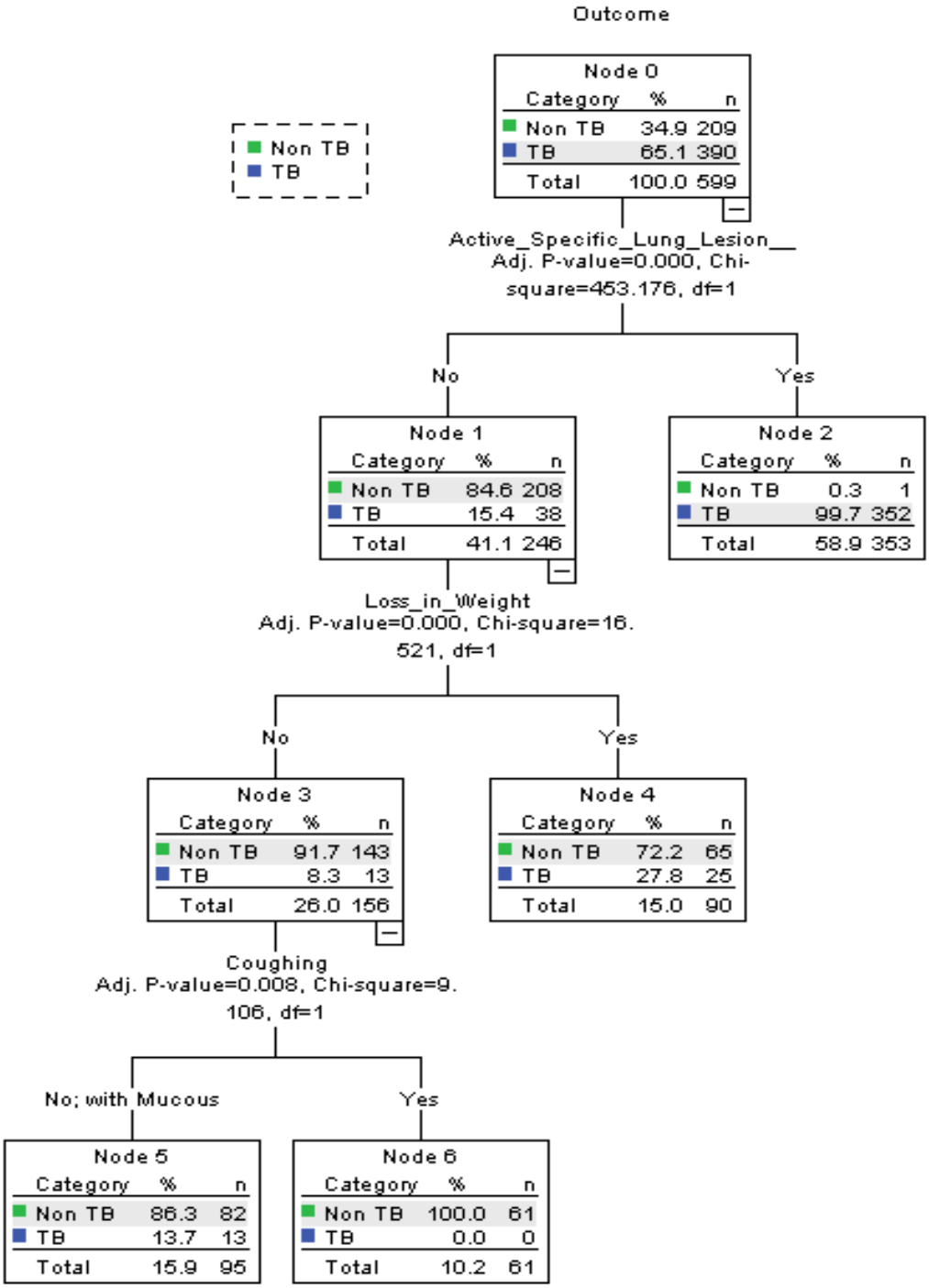


Figure (5.3) Decision Tree for Tuberculosis Diagnosis using Algorithm III with 20 Predictors

The data set used for the Algorithm III is the same as in Algorithm II. The main difference in Algorithm II and III is the type of handling missing value. As it can be seen in Figure (5.2) and Figure (5.3), the resulted models of decision tree have the same value for both the number of leaves and size of the tree. This means the models built by two

Algorithms were exactly identical even if it differs in procedures in algorithm. According to Algorithm I, II and III, the most relevant node for the construction of the tree was Active Specific Lung Lesion although the number of predictors is different. The followings are few of the rules which were discovered between attributes by Algorithm III. It can be seen that these rules are the same as in Algorithm II.

Table (5.3)

Classification Rules and Likelihood of TB using Algorithm III

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has the categorical value ‘Yes’ on Active Specific Lung Lesion	99.7%
2	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she has Loss in Weight	27.8%
3	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘No’ or ‘With Mucous’ on Coughing	13.7%
4	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘Yes’ on Coughing	0%

5.2.4 Decision Tree Model of Algorithm IV

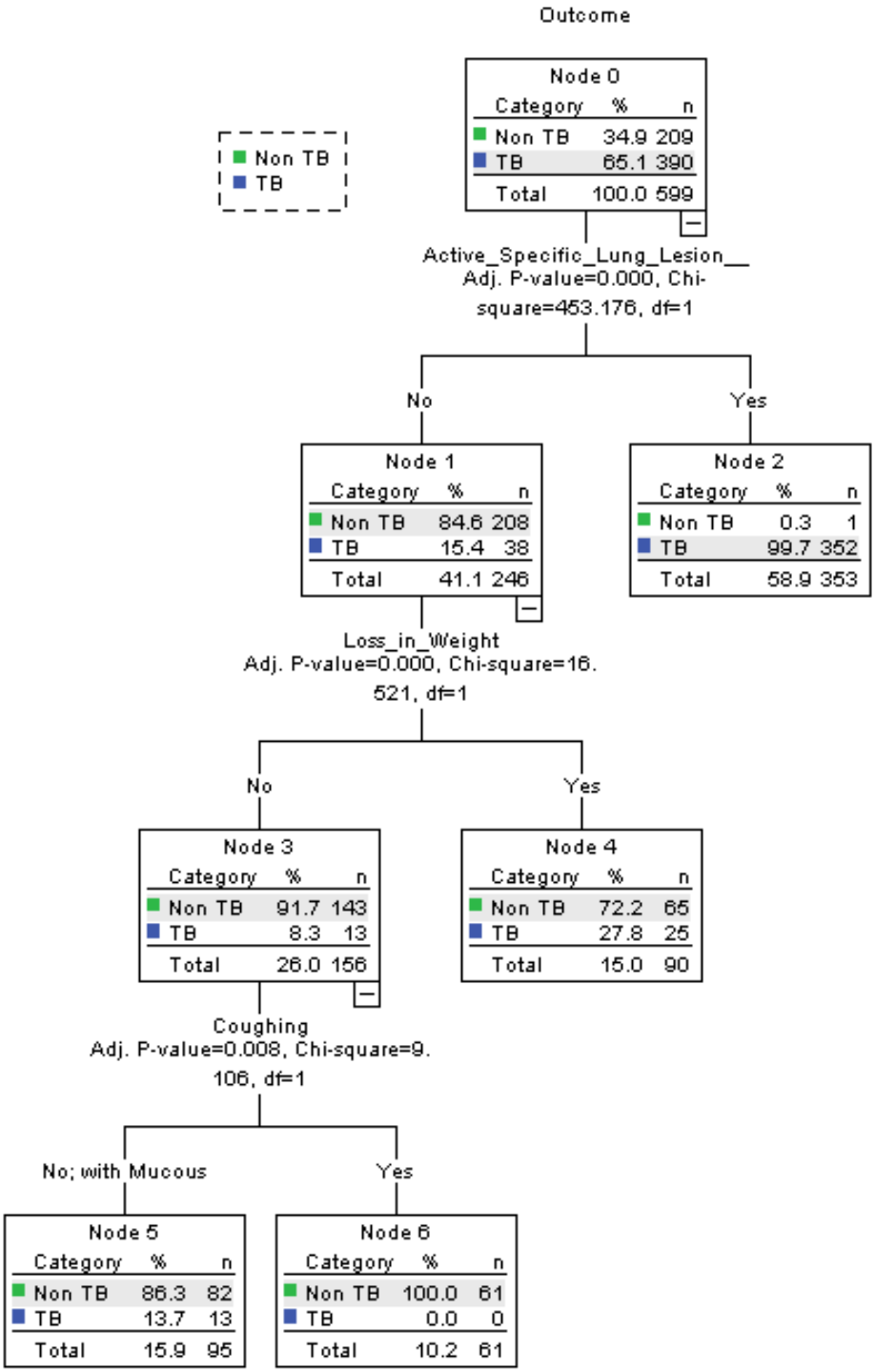


Figure (5.4) Decision Tree for Tuberculosis Diagnosis using Algorithm IV with 18 Predictors

Algorithm IV used tuberculosis diagnosis data set which consists of 599 records with 18 predictors. This data set was resulted by performing both handling missing and

variable aggregation. The resulted classification model of this algorithm is exactly same as Algorithms II and III. As shown in Figure (5.4), the most relevant node for the construction of the tree was Active Specific Lung Lesion. As it can be seen from figure (4.2), Figure (5.3) and Figure (5.4), the resulted models of decision tree have the same value for both the number of leaves and the size of the tree. This means the models built by three Algorithms were exactly identical even if these Algorithms differ in procedures and in the use of number of predictors for Algorithm.

Some of the rules of Algorithm IV for diagnosis of TB disease are summarized in Table (5.4). The resulted rules are exactly same as in Algorithm II and III.

Table (5.4)

Classification Rules and Likelihood of TB using Algorithm IV

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has the categorical value ‘Yes’ on Active Specific Lung Lesion	99.7%
2	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she has Loss in Weight	27.8%
3	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘No’ or ‘With Mucous’ on Coughing	13.7%
4	IF a person has the categorical value ‘No’ on Active Specific Lung Lesion AND he/she does not have Loss in Weight AND he/she has the categorical value ‘Yes’ on Coughing	0%

5.2.5 Decision Tree Model of Algorithm V

This Algorithm used tuberculosis data set which consists of 599 records with 10 predictors. This data set was resulted from performing handling missing, variable aggregation and feature reduction. The resulted classification model of this Algorithm is different from Algorithm I, II, III and IV but the most relevant node is exactly same as in previous Algorithms. Although 10 predictors were specified for Algorithm V, only three were included in the decision tree model.

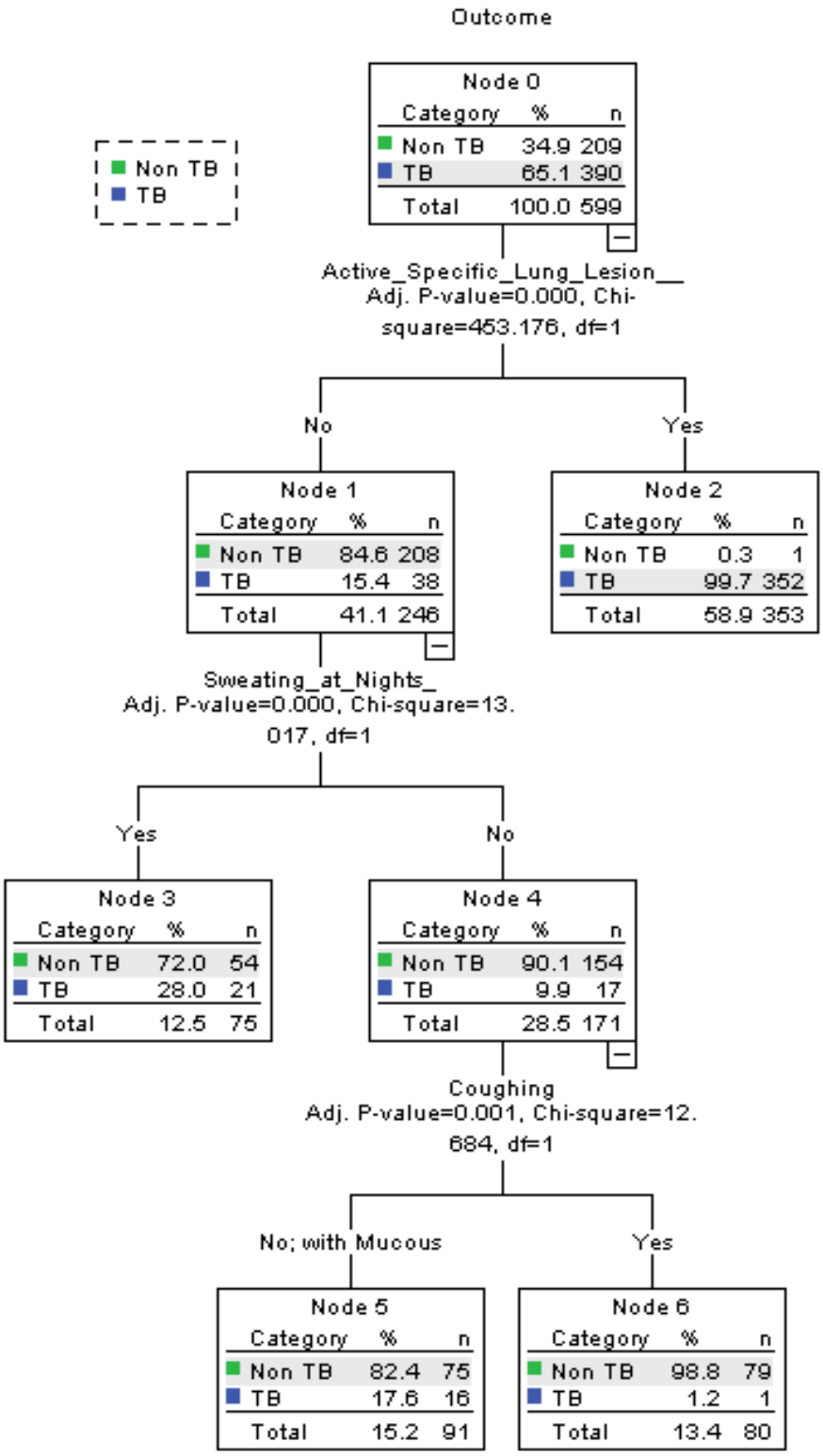


Figure (5.5) Decision Tree for Tuberculosis Diagnosis using Algorithm V with 10 Predictors

As shown in Figure (5.5), the most relevant node for the construction of the tree is Active Specific Lung Lesion. The categorical value 'Yes' of Active Specific Lung Lesion is only significant predictor of TB disease. Of the TB suspected patients in this

categorical value, 99.7% have TB disease. Since there are no child nodes below it, this is considered a terminal node. For the categorical value 'No' on Active Specific Lung Lesion, the second best predictor is Sweating at Night. The inclusion of second best predictor in this model was totally different from the model of Algorithm II, III and IV. Some of the rules of Algorithm V for diagnosis of TB disease are summarized in Table (5.5). The rules of this Algorithm are distinct from those of previous Algorithms.

Table (5.5)

Classification Rules and Likelihood of TB using Algorithm V

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has the categorical value 'Yes' on Active Specific Lung Lesion	99.7%
2	IF a person has the categorical value 'No' on Active Specific Lung Lesion AND he/she feels Sweating at Night	28.0%
3	IF a person has the categorical value 'No' on Active Specific Lung Lesion AND he/she does not feel Sweating at Night AND he/she has the categorical value 'With Mucous' OR 'No' on Coughing	17.6%
4	IF a person has the categorical value 'No' on Active Specific Lung Lesion AND he/she does not feel Sweating at Night AND he/she has the categorical value 'Yes' on Coughing	1.2%

5.3 Performance Evaluation for Classification Models

The developed model is evaluated for measuring predictability performance. The purpose of assessing the performance of the model is to determine how well the model will behave if used in practice. In this study, the models were evaluated based on the accuracy measures discussed in Section (2.5). In order to validate the prediction results of the comparison of the different algorithms, the 10-fold cross validation which can be implemented by SPSS software was used. A single prediction has the four different possible outcomes which can be shown as a confusion matrix. The followings are the confusion matrix for each algorithm respectively.

In these matrices, True Positive (TP) is the number of person who is correctly classified as a TB patient when he/ she is actually TB patient. True Negative (TN) is the number of person who is correctly classified as a Non-TB patient when he/ she is actually TB negative patient. A False Positive (FP) occurs when the outcome is incorrectly

predicted as TB patient (or TB positive) when it is actually Non-TB (or TB negative). A False Negative (FN) occurs when the outcome is incorrectly predicted as Non-TB patient (or TB negative) when it is actually TB (or TB positive).

Table (5.6)
Confusion Matrix for Algorithm I

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	425	TP 357	FN 68	Sensitivity 84.0%
Non-TB	234	FP 1	TN 233	Specificity 99.6%
Total	659	45.7%	54.3%	Overall Accuracy 89.5%

$$\text{Classification Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{357+233}{659} = 89.5\%$$

A measure which may be more important than classification accuracy is that: what proportion of TB positive patient will recognize as TB positive by the classifiers. This is call sensitivity of the model.

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{357}{425} = 84.0\%$$

Sensitivity is 84%, which means that the model fails to detect the disease in 16% of the cases.

Specificity is the proportion of TB negative (healthy) instances which were recognized as TB negative (healthy).

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{233}{234} = 99.6\%$$

Specificity is 99.6%, which means that this model has the possibility of 0.4% wrongly diagnosing patient without TB disease, and giving TB treatment unnecessarily.

Table (5.7)
Confusion Matrix for Algorithm II

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	390	TP 352	FN 38	Sensitivity 90.3%
Non-TB	209	FP 1	TN 208	Specificity 99.5%
Total	599	41.1%	58.9%	Overall Accuracy 93.5%

As shown in Table (5.7), classification accuracy of the Algorithm II is 93.5%, sensitivity of the model is 90.3% and specificity of the model is 99.5%.

Table (5.8)
Confusion Matrix for Algorithm III

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	390	TP 352	FN 38	Sensitivity 90.3%
Non-TB	209	FP 1	TN 208	Specificity 99.5%
Total	599	41.1%	58.9%	Overall Accuracy 93.5%

Table (5.8) shows the performance result on the tuberculosis diagnosis data set using Algorithm III in order to develop classification model. It can be seen that the performance of this model is the same as in Algorithm II. Therefore, not only the decision rules of Algorithm II and III, but also the performance measures are exactly the same. Classification accuracy of this algorithm is 93.5%, sensitivity of the model is 90.3% and specificity of the model is 99.5%.

Table (5.9)
Confusion Matrix for Algorithm IV

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	390	TP 352	FN 38	Sensitivity 90.3%
Non-TB	209	FP 1	TN 208	Specificity 99.5%
Total		41.1%	58.9%	Overall Accuracy 93.5%

Table (5.9) shows the performance result on the tuberculosis diagnosis data set using algorithm IV in order to develop classification model. It can be seen that the performance of this model is the same as in Algorithm III. Therefore, not only the decision rules of Algorithm II, III and IV, but also the performance measures are exactly the same. Classification accuracy of this Algorithm is also 93.5%, sensitivity of the model is 90.3% and specificity of the model is 99.5%.

Table (5.10)

Confusion Matrix for Algorithm V

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	390	TP 352	FN 38	Sensitivity 90.3%
Non-TB	209	FP 1	TN 208	Specificity 99.5%
Total	599	41.1%	58.9%	Overall Accuracy 93.5%

Table (5.10) shows the performance measures of Algorithm V. It can be seen that the classification accuracy, sensitivity and specificity of this model are exactly the same as in the models of Algorithms II, III and IV although the resulted models and rules of Algorithm V are extremely distinct.

5.4 Comparisons and Discussions

Table (5.11) illustrates some measure of evaluating performance of the classification models which are developed by different algorithms on the tuberculosis diagnosis data set.

Table (5.11)

Performance Results from the Five Algorithms

Classifier	Sensitivity[12]	Specificity[12]	Accuracy	Precision	Error Rate
Algorithm I	84.0%	99.6%	89.5%	99.7%	10.5%
Algorithm II	90.3%	99.5%	93.5%	99.7%	6.5%
Algorithm III	90.3%	99.5%	93.5%	99.7%	6.5%
Algorithm IV	90.3%	99.5%	93.5%	99.7%	6.5%
Algorithm V	90.3%	99.5%	93.5%	99.7%	6.5%

As it can be seen in Table (5.11), the model of Algorithm I achieves accuracy of 89.5% at 84.0% sensitivity and 99.6% specificity on data set and that of Algorithm II achieve accuracy of 93.5% at 90.3% sensitivity and 99.5% specificity on data set. Therefore, the accuracy of Algorithm II is better than that of Algorithm I although these two algorithms are performed on the same data set but distinct in number of records and predictors. It is noted that data filtering and missing value handling provide more accuracy of model performance, and original data set without preprocessing can give less accuracy. Since Algorithms II, III and IV give the same performance, different approach

12. Specificity and Sensitivity are mentioned in page 24 and 74.

on missing value handling and variable aggregation do not affect for classification model. Although the rules and tree model of Algorithm V are different from others (except algorithm I), the accuracies are the same. It can be assumed that the decision tree method can serve as variable selection because of the same rules and same accuracies in Algorithms II, III and IV. Moreover, Algorithm II is the best algorithm since it has minimum step and same accuracy in comparison of other algorithms.

5.5 Logistic Regression Model for Tuberculosis Diagnosis

Logistic regression was conducted to assess whether the three predictor variables- Active Specific Lung Lesion, Loss in Weight and Coughing significantly predicted whether or not a patient has TB disease. In the data set, the patients who have TB disease are represented by a '1' in the dependent variable and the patients who do not have TB disease are represented by a '0' in the dependent variable. Let $\pi(x)$ be the probability of TB positive of patient which is denoted as follows:

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

The significant variables which were extracted by decision tree technique based on Algorithm II were used for developing logistic regression model. The reason for doing so is that decision tree technique can be used as data exploration for other modeling techniques. This study is multiple logistic regression case, since more than one predictor variable are used to classify the response variable. In this study, dependent/ outcome variable is dichotomous (TB or Non-TB) and predictor variables have categorical value in which two predictors (Active Specific Lung Lesion and Loss in Weight) are dichotomous and another one (Coughing) is a trichotomous predictor. Thus, this trichotomous predictor is needed to code using indicator (dummy) variables. Then this indicator values to two new indicator variables: Coughing_Mucous and Coughing_Only. Each record is assigned to it a value of zero or 1 for each of Coughing_Mucous and Coughing_Only. If a patient has the categorical value 'No' on Coughing, Coughing_Mucous=0 and Coughing_Only =0, if a patient has the categorical value 'Yes' on Coughing, Coughing_Mucous=0 and Coughing_Only =1, and if a patient has the categorical value 'with Mucous' on Coughing, Coughing_Mucous=1 and Coughing_Only =0.

Omnibus Tests

There is needed to examine whether a relationship between TB disease and the following set of predictors: Active Specific Lung Lesion, Loss in Weight, Coughing_Mucous and Coughing_Only. To determine whether the model or equation built with all four variables considered together is significant or not, the omnibus test was used based on logistic regression technique.

Table (5.12)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	583.520	4	$1x10^{-7}$
	Block	583.520	4	$1x10^{-7}$
	Model	583.520	4	$1x10^{-7}$

The results of logistic regression analysis are provided in Table (5.12) and Table (5.13). According to the Chi-square statistic and p-value, the model developed with all four variables is highly significant. Therefore, the overall model is useful for classifying TB disease.

Table (5.13)

Results of Logistic Regression of Tuberculosis Diagnosis

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Loss_in_Weight	1.632	0.390	17.490	1	$2x10^{-5}$	5.112
Active_Specific_Lung_Lesion	7.696	1.053	53.372	1	$1x10^{-7}$	2200.196
Coughing_Mucous	0.315	0.478	0.434	1	$5x10^{-1}$	1.370
Coughing_Only	-1.687	0.486	12.057	1	$1x10^{-3}$	0.185
Constant	-1.995	0.314	40.449	1	$1x10^{-6}$	0.136

a. Variable(s) entered on step 1: Loss_in_Weight, Active_Specific_Lung_Lesion, Coughing_Mucous, Coughing_Only.

However, not all variables contained in the model need necessarily be useful. The *p*-values for the (Wald) *z*-statistics for each of the predictors were examined. All *p*-values are small except one, indicating that there is evidence that each predictor belongs in the model, except Coughing_Mucous. The Wald *z*-statistic for Coughing_Mucous is 0.51,

with a large *p*-value of 0.51, indicating that this variable is not useful for classifying TB disease. Therefore, now Coughing_Mucous is omitted from the model and the logistic regression is again proceeded to run with the remaining variables. The results are shown in Table (5.14) and (5.15).

Table (5.14)

Omnibus Tests of Model Coefficients After Omitting Coughing_Mucous

		Chi-square	df	Sig.
Step 1	Step	583.093	3	$1x10^{-7}$
	Block	583.093	3	$1x10^{-7}$
	Model	583.093	3	$1x10^{-7}$

It can be seen that the omission of Coughing_Mucous has barely affected the remaining analysis. All remaining variables are considered significant and retained in the model.

Table (5.15)

Results of Logistic Regression of Tuberculosis Diagnosis After Omitting Coughing_Mucous

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Loss_in_Weight	1.667	0.387	18.545	1	$1x10^{-5}$	5.295
Active_Specific_Lung_Lesion	7.734	1.051	54.170	1	$1x10^{-7}$	2284.933
Coughing_Only	-1.787	0.461	15.038	1	$1x10^{-4}$	0.167
Constant	-1.925	0.291	43.638	1	$1x10^{-7}$	0.146

a. Variable(s) entered on step 1: Loss_in_Weight, Active_Specific_Lung_Lesion, Coughing_Only.

Table (5. 15) provides the estimated logit:

$$\hat{g}(x) = -1.925 + 1.667(\text{Loss in Weight}) + 7.734\ (\text{Active Specific Lung Lesion}) - 1.787\ (\text{Coughing\_Only})$$

The above equation can be used to calculate the probability of present of TB positive for each patient who has distinct combination of symptoms: Loss in Weight,

Active Specific Lung Lesion and Coughing_Only. The probability of five patients suspected to be experiencing different symptoms were found to be as follows:

Patient 1: Loss in Weight 'Yes', Active Specific Lung Lesion 'Yes' and Coughing_only 'No'

The logit : $\hat{g}(x) = -1.925 + 1.667(1) + 7.734(1) - 1.787(0)$

$= 7.476$

The probability that patient 1 will have TB positive is

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{7.476}}{1 + e^{7.476}}$$

$\hat{\pi}(x) = 0.9994$

Thus, the probability of a patient who has categorical value 'Yes' on Loss in Weight, 'Yes' on Active Specific Lung Lesion and 'No' on Coughing_Only is 99.94%.

Patient 2: Loss in Weight 'Yes', Active Specific Lung Lesion 'Yes' and Coughing_only 'Yes'

The logit: $\hat{g}(x) = -1.925 + 1.667(1) + 7.734(1) - 1.787(1)$

$= 5.689$

The probability that patient 2 will have TB positive is

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{5.689}}{1 + e^{5.689}}$$

$\hat{\pi}(x) = 0.997$

Thus, the probability of a patient who has categorical value 'Yes' on Loss in Weight, 'Yes' on Active Specific Lung Lesion and 'Yes' on Coughing_Only is 99.7%.

Patient 3: Loss in Weight 'No', Active Specific Lung Lesion 'Yes' and Coughing_only 'No'

The logit: $\hat{g}(x) = -1.925 + 1.667(0) + 7.734(1) - 1.787(0)$

$= 5.809$

The probability that patient 3 will have TB positive is

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{5.809}}{1 + e^{5.809}}$$

$\hat{\pi}(x) = 0.9970$

Thus, the probability of a patient who has categorical value 'No' on Loss in Weight, 'Yes' on Active Specific Lung Lesion and 'No' on Coughing_Only is 99.70%

Patient 4: Loss in Weight 'No', Active Specific Lung Lesion 'Yes' and Coughing_only 'Yes'

The logit: $\hat{g}(x) = -\ 1.925 + 1.667(0) + 7.734\ (1) - 1.787\ (1)$

$= 4.022$

The probability that patient 4 will have TB positive is

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{4.022}}{1 + e^{4.022}}$$

$$\hat{\pi}(x) = 0.9824$$

Thus, the probability of a patient who has categorical value 'No' on Loss in Weight, 'Yes' on Active Specific Lung Lesion and 'Yes' on Coughing_Only is 98.24%.

Patient 5: Loss in Weight 'Yes', Active Specific Lung Lesion 'No' and Coughing_only 'Yes'

The logit: $\hat{g}(x) = -\ 1.925 + 1.667(1) + 7.734\ (0) - 1.787\ (1)$

$= -2.045$

The probability that patient 5 will have TB positive is

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{-2.045}}{1 + e^{-2.045}}$$

$$\hat{\pi}(x) = 0.1146$$

Thus, the probability of a patient who has categorical value 'Yes' on Loss in Weight, 'No' on Active Specific Lung Lesion and 'Yes' on Coughing_Only is 11.46%.

The classification rules together with the likelihood of TB positive were extracted from the results of logistic regression model and the summarized results are presented in Table (5.16).

Table (5.16)

Classification Rules and Likelihood of TB using Logistic Regression

Rule	Symptoms			Likelihood of TB positive
	Weight in Loss	Active Specific Lung Lesion	Coughing_Only	
1	Yes	Yes	No	99.94%
2	Yes	Yes	Yes	99.70%
3	No	Yes	No	99.70%
4	No	Yes	Yes	98.24%
5	Yes	No	Yes	11.46%

Table (5.17) illustrates the confusion matrix for logistic regression model.

Table (5.17)

Confusion Matrix for Logistic Regression Model

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	390	TP 352	FN 38	Sensitivity 90.3%
Non-TB	209	FP 1	TN 208	Specificity 99.5%
Total	599	41.1%	58.9%	Overall Accuracy 93.5%

As shown in Table (5.17), classification accuracy of logistic regression model is 93.5%, sensitivity of the model is 90.3% and specificity of the model is 99.5%. It can be seen that the performance of logistic regression model is the same as Algorithm II of the decision tree technique. Although the performance of predictability does not differ in two techniques; decision tree and logistic regression, output of the logistic regression cannot reveal explicitly for field workers. It is needed to perform for calculation to obtain the likelihood values in logistic regression method. And it can be noted that logistic regression is not as easy as decision tree to extract quick meaningful information from the results. It is time consuming and needs the analyst to put more effort for conducting interpretation for the output of logistic regression. Thus, the decision tree can make it easier for data mining algorithm to discover useful knowledge.

5.6 Decision Tree Model and Classification Rules for Tuberculosis Diagnosis without X-ray and Laboratory Results

In the previous subsection, it can be noted that the resulted models reveal that the most significant variable is Active Specific Lung Lesion. That is, Active Specific Lung Lesion is the best predictor for deciding the result of TB disease (existence or not). Therefore, these models are not applicable for the situations in which there is no X-ray result. In real situations, there can be many regions in which X-ray result and laboratory result cannot be obtained urgently. In this case, it is needed to predict TB disease based on only patient's symptoms without reviewing X-ray. Furthermore, the result of laboratory test for Sputum AFB cannot be obtained in such a region. In this subsection, the decision tree model was constructed by Algorithm II without X-ray result and laboratory result. The reasons for the choice of this algorithm are that the accuracies of four algorithms II, III, IV and V are the same and the resulted rules are the same. It can be noted that the resulted model of algorithm which performs variable selection task and that of the algorithm which does not perform variable selection task are the same. Therefore variable selection task is not essential for developing decision tree models. Furthermore, since there are 599 records with 20 predictors in Algorithm II, the data set used is reasonable for the rule of thumb in which the sample size must be at least 30 times of the variables. By removing Active Specific Lung Lesion variable and Sputum AFB variable the remaining data set includes 18 predictors which was used for developing alternative decision tree model.

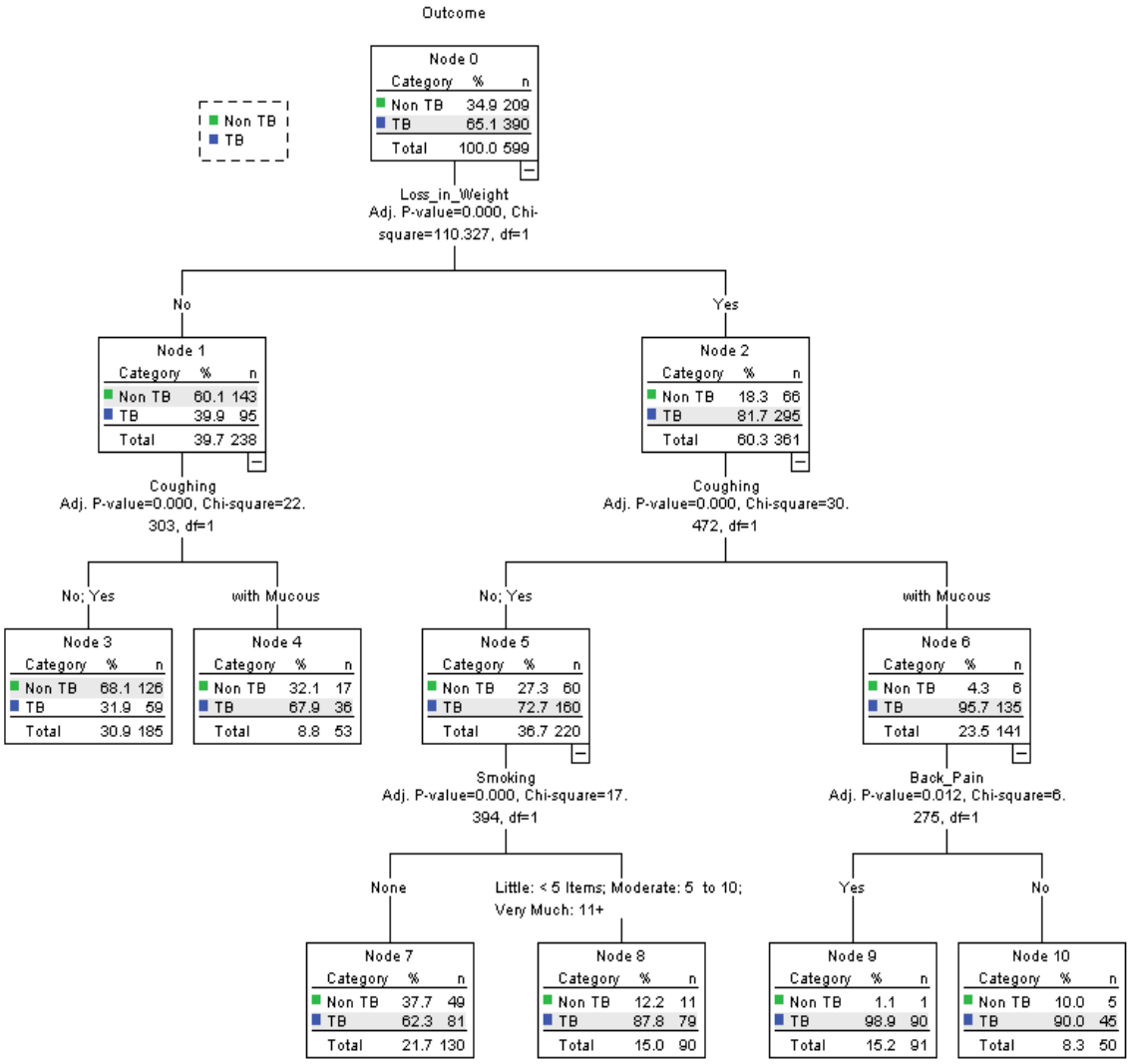


Figure (5.6) Decision Tree for Tuberculosis Diagnosis without X-ray and Laboratory Results with 18 Predictors

The resulted classification model of this Algorithm is extremely different from the models in the previous subsection. Although 18 predictors were specified, only four were included in the tree model and they are Loss in Weight, Coughing, Smoking and Back Pain. Fourteen predictors did not make a significant contribution to the model; therefore, these variables were automatically dropped from the tree model. As shown in Figure (5.6), the most relevant node for the construction of the tree was Loss in Weight variable. Some of the rules are summarized in Table (5.18) for TB diagnosis without X-ray and laboratory results.

Table (5.18)

Classification Rules and Likelihood of TB (without X-ray and Laboratory Results)

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has Loss in Weight AND he/she has the categorical value 'With Mucous' on Coughing AND he/she feels Back-Pain	98.9%
2	IF a person has Loss in Weight AND he/she has the categorical value 'With Mucous' on Coughing AND he/she does not feel Back-Pain	90.0%
3	IF a person has Loss in Weight AND he/she has the categorical value 'No' or 'Yes' on Coughing AND he/she is a Smoker	87.8%
4	IF a person does not have Loss in Weight AND he/she has the categorical value 'With Mucous' on Coughing	67.9%
5	IF a person has Loss in Weight AND he/she has the categorical value 'No' or 'Yes' on Coughing AND he/she does not smoke	62.3%
6	IF a person does not have Loss in Weight AND he/she has the categorical value 'No' or 'Yes' on Coughing	31.9%

According to the Table (5.18), the resulted rules are extremely distinct from previous algorithms. These rules are applicable for field workers for performing prevention and cure of TB disease in rural areas because X-ray results cannot be obtained in most of the rural areas. By using these rules, field workers can decide to whether the TB suspected patient needs to take further medical check-up or not in order to diagnose TB disease.

Table (5.19)

Confusion Matrix (Without X-ray and Laboratory Results)

Outcome	Actual	Predicted		
		TB	Non-TB	Correct %
TB	390	TP 331	FN 59	Sensitivity 84.9%
Non-TB	209	FP 83	TN 126	Specificity 60.3%
Total		69.1%	30.9%	Overall Accuracy 76.3%

Table (5.19) shows the performance result on the tuberculosis data set using 18 predictors without X-ray and Laboratory results in order to develop the classification model. It can be seen that the performance of this model is less than that of the models in previous models which were conducted with the variables obtained from X-ray and Laboratory result. Therefore, not only the resulted rules of this model are different from previous models, but the performance measures are also the extremely distinct. Classification accuracy of this algorithm is 76.3%, sensitivity of the model is 84.9% and specificity of the model is 60.3%. Although there is less accuracy of this model, it can be constructed by using predictors which are only patients' symptoms without viewing X-ray and laboratory result. This analysis produced classification rules (in Appendix D) which are useful for medical field workers in order to diagnose tuberculosis using symptoms of the patient. Moreover, these resulted rules can also be used for deciding whether it is necessary to perform medical check-up or not in order to diagnose TB disease. If the likelihood of TB positive is high which is revealed by the resulted rules, medical check-up would be needed.

CHAPTER 6

CONCLUSION

Data mining is an analytical tool that is used in solving critical decisions by analyzing enormous amounts of data in order to discover relationships among the variables and unknown patterns in the data. It is a process used by organization to generate useful information from raw data. The types of information obtained from data mining include association, sequences, classification, clusters, and forecasts. Classification is one of the major tasks in the data mining field. In current situation, there are large amounts of data and complex structure of data in many applications viz; sales and marketing, healthcare/ medical diagnosis, supply chain management, process control, bioinformatics and astronomy. The healthcare environment is still 'information rich' but 'knowledge poor'. The medical data set which includes patients' records is difficult to analyze because of its characteristics consisting of huge volume and heterogeneity, irrelevant data and high frequency of null value. In such a case, applying data mining techniques to patient's attribute is useful to build pattern or model that can be used to make diagnosis for a particular disease.

This study attempted to develop the model in order to classify the existence or non-existence of a particular disease for a patient by using data mining techniques. The data used in this study were obtained from the medical profiles of patients in Latha and Aung San townships under UTI, Yangon in Myanmar. The profile included characteristics, behavior, symptoms and medical test results of patients who came to UTI for their medical check-up during the period from September 2013 to October 2013.

Based on the data set which included 34 different variables: one dependent variable (outcome: TB or Non-TB) and 33 independent variables, data mining methods were employed to analyze the data. The classification task was made using decision tree technique, and different decision tree models were conducted by different algorithms. Before the development of classification models, data preprocessing techniques were conducted with the objectives of filtering the insufficient data from the original data, to estimate missing value and to reduce some variables which are overlapping.

6.1 Findings

The purpose of this study is to access the effectiveness of data mining algorithms in predicting the presence of tuberculosis and to compare their performance regarding predictability. By using five algorithms, decision tree models were developed in order to predict TB disease. It has been found that the accuracy of a model is the same as any other model except Algorithm I. Among these models developed in this study, Algorithm I which was directly employed on original data set without performing data preprocessing has the least accuracy of prediction. It indicates that the data preprocessing and exploration stage are critical to construct successful implementation of data mining. The findings indicate that the attribute value 'Yes' on Active Specific Lung Lesion affects the TB disease. The findings did not show a great deal of predictive accuracy in the four Algorithms (excluding Algorithm I). It was noted that the accuracy of classification models of Algorithm II and that of III which did not perform variable aggregation and that of Algorithm IV and V which performed variable aggregation do not differ from each other. Thus, it can be concluded that variable aggregation is not essential for developing decision tree models.

Moreover, this study has shown that the selected significant variables in the classification models of Algorithm II, III and IV are the same. The significant variables of these models are Active Specific Lung Lesion, Loss in Weight and Coughing. According to the results, Active Specific Lung Lesion variable can be seen at the top of the all three decision tree models. Active Specific Lung Lesion has been found to be the most significant variable and it can be also known that 99.7% of the patients have TB positive if the categorical value was 'Yes' on Active Specific Lung Lesion variable. When categorical value of Active Specific Lung Lesion was 'No', the second most significant variable was Loss in Weight. Thus, the likelihood of TB positive has been found to be 27.8% when the categorical value of Active Specific Lung Lesion was 'No' and that of Loss in Weight was 'Yes'. The last significant variable for these three models has been found to be Coughing and if the categorical value of Coughing was 'No' or 'with Mucous', the likelihood of TB positive was13.7% while there was 'No' on both Active Specific Lung Lesion and Loss in Weight variables.

The Algorithm V covered data filtering, missing value handling and variable aggregation. The data set used for this algorithm consists of 10 predictors. It has also revealed that the best predictor was Active Specific Lung Lesion and 99.7% of the patients have TB positive when the categorical value of this variable was 'Yes'. The

second best predictor was Sweating at Night. Thus, when the categorical value of Active Specific Lung Lesion was ‘No’ and that of Sweating at Night was ‘Yes’, the likelihood of TB positive would be 28.0%. The last significant variable in this model was Coughing. If the categorical value of Coughing was ‘No’ or ‘with Mucous’, together with that of ‘No’ on Active Specific Lung Lesion and Sweating at Night, then the likelihood of TB positive would be 17.6%.

Moreover, in order to explore the significant symptoms of TB positive without X-ray results (or excluding the attribute value of Active Specific Lung Lesion), the alternative decision tree model was developed based on the Algorithm II. For the suspected patients who cannot be examined with the result of X-ray, it can be predicted whether the patient is TB positive or not based on the symptoms; Loss in Weight, Coughing, Smoking and Back Pain. Especially in rural areas, these significant symptoms which were provided by alternative decision tree model can be applicable to diagnose TB disease. Based on the findings of this study, it was clearly defined and verified the most significant variable which can predict TB disease through decision tree technique based on the empirical data.

In addition, logistic regression model for TB diagnosis was also developed based on the selected variables which were extracted by using decision tree technique. The reason for doing so is decision tree technique can be used as data exploration for other modeling techniques. Although the performance of predictability does not differ in two techniques; decision tree and logistic regression, output of the logistic regression cannot reveal explicitly for field workers. And it can be noted that the result of logistic regression is difficult to extract quick meaningful information. Thus, decision tree can make it easier for data mining algorithm to discover useful knowledge.

In conclusion, the data preprocessing is critical for prediction purpose when developing the model in data mining on secondary data (or existing data set). This is because the accuracy of the decision tree model which is developed by using preprocessed data is more accurate than the accuracy of the model which is achieved by using original data set without preprocessing. Furthermore, the accuracy of each developed decision tree models is the same, on the data set used achieved after performing the feature (dimensionality) reduction or not. Therefore, it can be noted that decision tree method can serve as feature reduction or variable selection method itself. Moreover, decision trees are powerful first step in modeling process even when building

the final model using some other techniques since this technique combines both data exploration and modeling.

It has also been observed that decision tree technique can provide the classification rules which can identify the symptoms of TB positive. In addition, the classification rules which are resulted from decision tree models showed an interesting pattern. Among the resulted classification rules for TB diagnosis from different decision tree models, the alternative decision tree model based on the patient's symptoms only (without X-ray result) revealed that there is a better advantageous for the healthcare centers which have no X-ray machines since these rules can be used to make the efficient prediction for diagnosis. According to the resulted rules, the probability of getting TB disease for a person who has *Loss in Weight*, with Mucous on *Coughing* and feel *Back-pain*, is 98.9%. Since this likelihood of TB positive is high, it is necessary to seriously encourage in order to performing medical check-up urgently at the place where there is modern equipments and techniques. Similarly, these actions should be taken by the following patients:

(i) A person who has *Loss in Weight*, with Mucous on *Coughing* and does not feel *Back-pain* (90.0% TB positive)

(ii) A person who has *Loss in Weight*, with Mucous on *Coughing* and is *Smoker* (87.8% TB positive)

(iii) A person who does not have *Loss in Weight*, with Mucous on *Coughing* (67.9% TB positive)

(iv) A person who has *Loss in Weight*, without Mucous on *Coughing* and does not *Smoke* (62.3% TB positive)

Because these patients have high likelihood of TB positive, they have to confirm for TB disease through X-ray and to take treatments if necessary. The alternative decision tree model also indicated that the probability of TB positive for a person who has not *Loss in Weight*, and without Mucous on *Coughing* is 31.9%. Thus, this person does not need to take urgently medical check-up because the likelihood of TB positive is low.

The field workers should encourage the following patients to go to the closed place where there is X-ray machine, advanced technology for diagnosis and expert technicians.

(i) A patient (with high likelihood of TB positive) lives in rural area where there is no medical facility and modern diagnostic equipment.

(ii) A patient (with high likelihood of TB positive) is inability to take medical check-up due to the time, high charge for transportation and expensive for medical check-up.

6.2 Recommendations

Based on this empirical study, it is strongly recommended as follows:

1. Among the statistical models, the decision tree models developed in this study are better for the classification tasks on the categorical data.
2. This study indicates that decision technique cannot reduce the performance of prediction without variable aggregation and feature reduction.
3. This research was undertaken for tuberculosis diagnosis; hence it is recommended that a data mining concept might be applied for diagnosis of other types of diseases in our country as well.
4. The developed decision tree models could have been used to predict whether a TB suspected patient has TB disease or not and to support the healthcare workers in order to diagnose tuberculosis disease based on the classification rules.
5. The developed database can be used as a baseline for the hospital especially for TB clinic case term workers to encode easily their future data concerning the detail information of the tuberculosis patients.
6. This study was limited to study TB suspected patients in UTI at Yangon. If the large amount of data on TB suspected patients and the information on TB suspected patients at Mandalay can be obtained, the performance of the model might be increased.

There are two types of error in prediction on the tuberculosis diagnosis data set by using decision tree method. First is committing the error when a person was classified as TB positive patient but he or she did not suffer actually from TB disease (It is called FP: False Positive). The second one is committing the error when a person was classified as a person who had not TB disease but he or she was infected by TB disease (It is called FN: False Negative). Both two types of error are important for medical field. As a consequence from these, if a person had taken TB treatments without actually suffering from TB disease, he or she would feel seriously side-effects. If a person had not taken TB treatments while he or she was suffering TB disease, life might be lost.

In current situation, TB disease can be diagnosed in the simple way through X-ray in the medical field. Thus, the advantages of decision tree models developed in this study cannot be obviously revealed. If the decision tree model for the case of cancer diagnosis

can be constructed based on the records of cancer suspected patients, there will be more benefits in the medical field. Cancer disease is ambiguous to classify correctly whether it is cancer or not. Therefore, decision tree model under the data mining techniques can give efficient advantages for the complex problems to make correct decisions in any applications. Moreover, the acquisition of data for data mining techniques will be larger and larger; it would increase the accuracy of prediction on most significant factors.

REFERENCES

1. Andersson, C. A. (2000), Exploratory Multivariate Data Analysis with Applications in Food Technology, *PhD Dissertation*, The Royal Veterinary and Agricultural University, Denmark.

2. Ansari, U. and Soni, S. (2011), Predictive Data Mining For Medical Diagnosis: An Overview of Heart Disease Prediction, *International Journal of Computer Application*, Vol. 17, No.8, pp. 43-48.

3. Asha, T., Natarajan, S. and Murthy, K. N. B. (2012), A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification, Bangalore Institute of Technology, Karnataka, India.

4. Berry, M. J. A. and Linoff, G. (2004), *Data Mining Techniques: for Marketing, Sales and Customer Relationship Management*, Wiley, New York, Chichester.

5. Chang, W. P. and Liou, D. M. (2007), Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data, Institute of Public Health, National Yang-Ming University, Tainwan.

6. Danso, S. O. (2006), An Exploration of Classification Prediction Techniques in Data Mining: The Insurance Domain, *M.Sc Dissertation*, Bournemouth University.

7. Delen D., Walker G., Kadam A. (2004), Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods, *Artificial Intelligence in Medicine*, Elsevier Inc., San Frencisco, pp. 1-15.

8. Demsar, J. (2010), Data Mining- Material for Lectures on Data Mining at Kyoto University, Department of Health Informatics in July 2010, University of Kyoto.

9. Durairaj, M. and Ranjani, V. (2013), Data Mining Applications in Healthcare Sector: A Study, *International Journal of Scientific & Technology Research*, Vol. 2, Issue 10, pp. 29-39.

10. Gartner Group, The Gartner Group CRM Glossary.[http:// www. gartnerweb. com/ public/ static/ hotc].

11. Guo, L. (2002), ASA, Applying Data Mining Techniques in Property/ Casual Insurance, University of Central Florida, *The CAS Committee on Management Data and Information*, pp.1-25.

12. Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, Second Edition, Elsevier Inc., San Frencisco.

13. Holbrey, R. (2006), Dimension Reduction Algorithms for Data Mining and Visualization, University of Leeds, Edinburgh.

14. Jackson, J. (2002), Data Mining: A Conceptual Overview, *Communications of the Association for Information Systems*, Vol. 8, pp.267-296.

15. Janecek, A. (2009), Effcient Feature Reduction and Classification Methods, *Ph.D Dissertation*. University Wein.

16. Jolliffe, I.T. (2002), *Principal Component Analysis*, Second Edition, Springer-Verlag, New York.

17. Kao, L. J., and Chih C. C. (2001), Mining the Customer Credit by using the Neural Networks Model with Classification and Regression Trees Approach, Vancouver.

18. Larose, D. T. (2005), *Data Mining Methods and Models*, John Wiley & Sons, Inc., Hoboken, New Jersey, Canada.

19. Leech, N. L., Barrett, K. C. and Morgan, G. A. (2005), *SPSS for Intermediate Stastics: Use and Interpretation*, Second Edition, Lawrence Erlbaum Associates, New Jersey, London.

20. Liao, S. H., Chu, P.H. and Hsiao, P. Y. (2012), Data Mining Techniques and Applications- A Decade Review from 2000 to 2011, *Expert Systems with Applications*, Elsevier Ltd., Vol. 39, pp. 11303-11311.

21. Liu, H. (2004), Effective Use of Data Mining Technologies on Biological and Clinical Data, National University of Singapore.

22. Luan, J. (2002), Data Mining and Its Applications in Higher Education, Spring, Wiley Periodicals, Inc., Cabrillo College in Aptos, California.

23. Murthy, K. I. (2010), Data Mining- Statistics Applications: A key to Managerial Decision Making, College of Management and Economics Studies, University of Petroleum and Energy Studies, Dehradun.

24. Olson, D. L. and Delen, D. (2008), *Advanced Data Mining Techniques*, Springer, USA.

25. Paliwal, M. and Kumar, U. A. (2009), Neural Networks and Statistical Techniques: A Review of Applications, *Expert Systems with Applications*, Vol. 36, Issue 1, pp. 2-17.

26. Raja, U. (2006), *Open Source Software Development and Maintenance: An Exploratory Analysis*, Texas A & M University, Pakistan.

27. Ramageri, B. M. (2010), Data Mining Techniques and Applications, *Indian Journal of Computer Science and Engineering*, Vol. 1, No. 4, pp. 301-305.

28. Razi, M. A. and Athappilly, K. (2005), A Comparative Predictive analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Trees (CART) Models, *Expert System with Application*, Elsevier Inc., Vol. 29, pp. 65-74.

29. Rokach, L. and Maimon, O. (2006), *Data Mining and Knowledge Discovery Handbook*. Tel-Aviv University, Italy.

30. Salame, E. J. (2011), Applying Data Mining Techniques to Evaluate Application for Agricultural Loans, *Ph.D Dissertation*. University of Nebraska, Lincoln.

31. Saumya, T. M. D., Rupasinghe, T. and Abeysinghe, P. (2014), A Literature Review in Data Mining Models Used for Survivability Prediction of Cancer Patients, *11-th Informational Conference on Business Management*, University of Kelaniya, Srilanka, pp. 144-152.

32. Shouman, M., Turner, T. and Stocker, R. (2011), Using Decision Tree for Diagnosing Heart Disease Patients, *Conference in Research and Practice in Information Technology (CRPIT)*, Australian Computer Society Inc., Vol. 121, pp. 23-29.

33. Timm, N. H. (2002), *Applied Multivariate Analysis*. Springer-Verlag, New York.

34. Tjung, L. C., Kwon, O., Tseng, K. C. and Broadley, J.(2011), Forecasting Financial Stocks using Data Mining, Califonia State University Fresno, pp. 1-25.

35. Tukey, J. W. (1977), *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Califonia.

36. Two Crows Corporation. (1999), Introduction to Data Mining and Knowledge Discovery, Third Edition, [online]. www. Twocrows.com/ intro-dm.pdf.

37. User Guide (2003), Insightful Miner 3 User's Guide, Seattle, *Insightful Corporation*: Washington.

38. Verbeek, J.J. (2004), Mixture Model for Clustering and Dimension Reduction. *ASCI dissertation series number 107*, Technology Foundation STW, Dutch Ministry of Economic Affairs.

39. Witten, I. H. and Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition., Morgan Kaukmann, San Franciso.

40. WHO (2012), Global Tuberculosis Report.

41. WHO (2012), Review of the Tuberculosis Programme of Myanmar.

42. http:// www.Anderson.uda.edu./ faculty/ Jason.frand/ teacher/ technologies/ palace/ dataming.htm, August 2011.

43. http://www.bookpump.com, August 2011.

44. http://www.citeulike.org, August 2011.

45. http://www.dissertation.com, August 2011.

46. http:// www. itl.nist.gov/ div898/handbook/eda/, September 2012.

47. http://www.marketingprofs.com/articles/2010, September 2012.

48. http:// www. statoo.com/en/datamining, September 2012

49. http://www.statsoft.com/textbook/data-mining-techniques, September 2012.

50. http:// www. Ics. Uci. Edu/ ~mlearning/ MLRepository.html, February 2013.

APPENDICES

Appendix A

A 1 χ^2 Statistical Measure

χ^2 measure evaluates features individually by measuring the χ^2-statistic with respect to the class. Different from the preceding methods, χ^2 measure can only handle features with discrete values. χ^2 measure of a feature *f* with discrete values is defined as

$$\chi^2(f) = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad \text{(A 1)}$$

where *k* is the number of classes, A_{ij} is the number of samples with i^{th} value of *f* in j^{th} class, E_{ij} is the expected frequency of A_{ij} and

$$E_{ij} = R_i * C_j / n \quad \text{(A 2)}$$

R_i is the number of samples having the i^{th} value of *f*, C_j is the number of samples in the j^{th} class, and *n* is the total number of samples. A feature f_j is considered to be more relevant than a feature f_l (l ≠ j) if χ^2 (f_j) > χ^2 (f_l). Obviously, the worst χ^2 value is 0 if the feature has only one value. The degree of freedom of the χ^2-statistic measure is (*w*-1)*(*k*-1). To apply χ^2 measure to numeric features, a discretization preprocessing has to be taken.

A 2 Cluster Analysis

Cluster analysis is one of the basic techniques that are often applied in analyzing large data sets. Originating from the area of statistics, most cluster analysis algorithms have originally been developed for relatively small data sets. In recent years, the clustering algorithms have been extended to efficiently work on large datasets, and some of them even allow the clustering of the high-dimensional feature vectors. Clustering is one of the most useful tasks in data mining process for discovering group and identifying interesting distributions and patterns in the underlying data. Clustering studies have no dependent variables (Guo, L., 2002). Cluster analysis can be used by the medical/ health care industry to improve predictive accuracy by segmenting databases into more homogeneous groups. Then the data of each group can be explored, analyzed, and modeled. The clustering is not only used to group instances (records) but also used to group features (variables) in order to extract features. It is often applied to data sets in

which a class is nonexistent or irrelevant from the perspective of knowledge. Moreover, the method can also be used for classification purposes to determine a missing value of an attribute of an instance by assigning it to one of the cluster. Each cluster groups instances that are in some way similar or related to each other.

There are many clustering methods in the literature. These methods can be categorized broadly into: partitioning methods, hierarchical methods, and density-based methods. The partitioning methods use a distance-based metric to cluster the points based on their similarity. To initiate a cluster analysis one constructs a proximity matrix. The proximity matrix represents the strength of the relationship between pairs of rows in $Y'_{p\times n}$ or the data matrix $\mathbf{Y}_{n\times p}$. Algorithms designed to perform cluster analysis are usually divided into two broad classes called hierarchical and nonhierarchical clustering methods. Hierarchical clustering helps in data visualization and summarization. Generally speaking, hierarchical methods generate a sequence of cluster solutions beginning with clusters containing a single object and combines objects until all objects form a single cluster; such methods are called agglomerative hierarchical methods. Other hierarchical methods begin with a single cluster and split objects successively to form clusters with single objects; these methods are called diversive hierarchical methods. In both the agglomerative and diversive processes, a tree diagram, or dendogram, is created as a map of the process. The agglomerative hierarchical procedures fall into three broad categories: Linkage, Centroid, and Error Variance methods. Among these procedures, only linkage algorithms may be used to cluster either objects (items) or variables. The other two methods can be used to cluster only objects. Nonhierarchical methods may only be used to cluster items (Timm, N. H., 2002).

For clustering variables, proximity measures for clustering rows of $\mathbf{Y}_{n\times p}$ may also be used to cluster rows of $\mathbf{Y}_{p\times n}$, or variables. When clustering variables, one is likely to standardize the variables and use some measure of association for clustering.

Some popular and widely used data mining clustering techniques such as hierarchical and k-means clustering techniques are statistical techniques and can be applied on high dimensional datasets.

Appendix B

Appendix Tables

Table (B 1)
Clustering for Alcohol

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Alcohol	No	348	43	6
	Yes	107	19	12

Table (B 2)
Clustering for BCG_Vaccine

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
BCG_Vaccine	No	252	1	1
	Yes	278	10	2

Table (B 3)
Clustering for Malaise

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Malaise	No	344	4	5
	Yes	183	8	3

Table (B 4)
Clustering for Arthralgia

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Arthralgia	No	470	9	8
	Yes	57	3	0

Table (B 5)

Clustering for Exhaustion

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Exhaustion	No	360	3	5
	Yes	168	9	3

Table (B 6)

Clustering for Unwillingnes_for_Work

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Unwillingnes_for_Work	No	356	16	5
	Yes	138	24	6

Table (B 7)

Clustering for Unwillingnes_for_Work

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Loss_of_Appetite	No	280	11	3
	Yes	235	14	0

Table (B 8)

Clustering for Loss_in_Weight

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Loss_in_Weight	No	225	4	0
	Yes	331	2	3

Table (B 9)
Clustering for Sweating_at_Nights

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Sweating_at_Nights	No	317	14	0
	Yes	228	11	2

Table (B 10)
Clustering for Chest_Pain

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Chest_Pain	No	261	16	3
	Yes	211	41	5

Table (B 11)

Clustering for Back_Pain

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Back_Pain	No	259	18	3
	Yes	213	39	5

Table (B 12)

Clustering for Coughing

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Coughing	No	157	3	1
	Yes	189	2	2
	with Mucous	170	4	12

Table (B 13)

Clustering for Hemoptysis

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Hemoptysis	No	475	5	8
	Yes	59	2	1

Table (B 14)

Clustering for Fever

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Fever	Normal	443	5	13
	High	64	0	7
	Subfebrille	35	1	2

Table (B 15)

Clustering for Active_Specific_Lung_Lesion

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Active_Specific_Lung_Lesion	No	235	1	1
	Yes	307	2	10

Table (B 16)

Clustering for Sputum_AFB

		Average Linkage (Between Groups)		
		1	2	3
		Count	Count	Count
Sputum_AFB	Negative	393	1	6
	Positive	148	2	5

Table (B 17)

Chi-Square Tests for Relationship between Predictors and Outcome

No.	Predictors	Pearson Chi-Square Value	df	Sig- value
1	Smoking Habit	39.805	3	0.000
2	BCG_Vaccine	18.670	1	0.000
3	Alcohol	32.746	1	0.000
4	Malaise	57.678	1	0.000
5	Arthraligia	5.723	1	0.017
6	Exhaustion	62.365	1	0.000
7	Unwillingnes_for_Work	51.504	1	0.000
8	Loss_of_Appetite	33.571	1	0.000
9	Loss_in_Weight	110.327	1	0.000
10	Sweating_at_Nights	30.115	1	0.000
11	Chest_Pain	49.256	1	0.000
12	Back_Pain	49.256	1	0.000
13	Coughing	68.628	2	0.000
14	Hemoptysis	.854	1	0.355
15	Fever	36.841	2	0.000
16	Sputum_AFB	100.796	1	0.000
17	Weight_Condition	1.496	2	0.473
18	Active_Specific_Lung_Lesion	453.176	1	0.000

Appendix C

Output of Classification Rules of Decision Tree Models

C 1 Decision Rules of Algorithm I

/* Node 5 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight != "Yes") AND (Coughing != "Yes")
THEN
Node = 5
Prediction = 0
Probability = 0.813084

/* Node 6 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight != "Yes") AND (Coughing = "Yes")
THEN
Node = 6
Prediction = 0
Probability = 0.985507

/* Node 7 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight = "Yes") AND (Sweating_at_Nights_ != "No")
THEN
Node = 7
Prediction = 0
Probability = 0.507937

/* Node 8 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight = "Yes") AND (Sweating_at_Nights_ = "No")
THEN
Node = 8
Prediction = 0

Probability = 0.741935

/* Node 2 */.
IF (Active_Specific_Lung_Lesion__ != "No")
THEN
Node = 2
Prediction = 1
Probability = 0.997207

C 2 Decision Rules of Algorithm II, III, IV

/* Node 5 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight != "Yes") AND (Coughing != "Yes")
THEN
Node = 5
Prediction = 0
Probability = 0.863158

/* Node 6 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight != "Yes") AND (Coughing = "Yes")
THEN
Node = 6
Prediction = 0
Probability = 1.000000

/* Node 4 */.
IF (Active_Specific_Lung_Lesion__ = "No") AND (Loss_in_Weight = "Yes")
THEN
Node = 4
Prediction = 0
Probability = 0.722222

```
/* Node 2 */.
IF (Active_Specific_Lung_Lesion__ != "No")
THEN
Node = 2
Prediction = 1
Probability = 0.997167
```

C 3 Decision Rules of Algorithm V

```
/* Node 3 */.
IF (Active_Specific_Lung_Lesion = "No")  AND  (Sweating_at_Nights = "Yes")
THEN
Node = 3
Prediction = 0
Probability = 0.720000

/* Node 5 */.
IF (Active_Specific_Lung_Lesion = "No")  AND  (Sweating_at_Nights != "Yes")  AND  (Coughing != "Yes")
THEN
Node = 5
Prediction = 0
Probability = 0.824176

/* Node 6 */.
IF (Active_Specific_Lung_Lesion = "No")  AND  (Sweating_at_Nights != "Yes")  AND  (Coughing = "Yes")
THEN
Node = 6
Prediction = 0
Probability = 0.987500
```

```
/* Node 2 */.
IF (Active_Specific_Lung_Lesion != "No")
THEN
Node = 2
Prediction = 1
Probability = 0.997167
```

C 4 Decision Rules of Decision Tree without X-ray Results

```
/* Node 3 */.
IF (Loss_in_Weight = "No")  AND  (Coughing != "with Mucous")
THEN
Node = 3
Prediction = 0
Probability = 0.681081

/* Node 4 */.
IF (Loss_in_Weight = "No")  AND  (Coughing = "with Mucous")
THEN
Node = 4
Prediction = 1
Probability = 0.679245

/* Node 7 */.
IF (Loss_in_Weight != "No")  AND  (Coughing != "with Mucous")  AND  (Smoking != "Little: < 5 Items"  AND  Smoking !=
"Moderate: 5  to 10"  AND  Smoking != "Very Much: 11+")
THEN
Node = 7
Prediction = 1
Probability = 0.623077
```

/* Node 8 */.
IF (Loss_in_Weight != "No") AND (Coughing != "with Mucous") AND (Smoking = "Little: < 5 Items" OR Smoking = "Moderate: 5 to 10" OR Smoking = "Very Much: 11+")
THEN
Node = 8
Prediction = 1
Probability = 0.877778

/* Node 9 */.
IF (Loss_in_Weight != "No") AND (Coughing = "with Mucous") AND (Back_Pain != "No")
THEN
Node = 9
Prediction = 1
Probability = 0.989011

/* Node 10 */.
IF (Loss_in_Weight != "No") AND (Coughing = "with Mucous") AND (Back_Pain = "No")
THEN
Node = 10
Prediction = 1
Probability = 0.900000

Appendix D

Summarized Classification Rules for Tuberculosis Diagnosis using Variables Clustering (Without X-ray Results)

Rule	Descriptions	Likelihood of TB Positive
1	IF a person has the categorical value 'Yes' on *Loss in Weight* AND he/she has the categorical value 'with Mucous' on *Coughing* AND he/she feels *Back-pain*	98.9%
2	IF a person has the categorical value 'with Mucous' on *Coughing* AND he/she feels *Chest-pain* AND he/she has the categorical value 'Yes' on *Loss of Appetite*	98.4%
3	IF a person has the categorical value 'with Mucous' on *Coughing* AND he/she feels *Back-pain*	95.7%
4	IF a person has the categorical value 'Yes' on *Loss in Weight* AND he/she has the categorical value 'with Mucous' on *Coughing*	95.7%
5	IF a person has the categorical value 'Yes' on *Loss in Weight* AND he/she has the categorical value 'with Mucous' on *Coughing* AND he/she does not feel *Back-pain*	90.9%
6	IF a person has the categorical value 'Yes' or 'No' on *Coughing* AND he/she feels *Sweating at Nights* AND he/she is a *Smoker*	90.7%
7	IF a person has the categorical value 'with Mucous' on *Coughing* AND he/she feels *Chest-pain* AND he/she has the categorical value 'No' on *Loss of Appetite*	89.7%
8	IF a person has the categorical value 'Yes' on *Loss in Weight* AND he/she has the categorical value 'No' or 'Yes' on *Coughing* AND he/she is a *Smoker*	87.8%
9	IF a person has the categorical value 'Yes' or 'No' on *Coughing* AND he/she feels *Malaise*	78.4%%
10	IF a person has the categorical value 'with Mucous' on *Coughing* AND he/she does not feel *Chest-pain*	78.4%
11	IF a person has the categorical value 'with Mucous' on *Coughing* AND he/she does not feel *Back-pain*	77.2%

12	IF a person has the categorical value 'No' on *Loss in Weight* AND he/she has the categorical value 'with Mucous' on *Coughing*	67.9%
13	IF a person has the categorical value 'Yes' or 'No' on Coughing AND he/she does not feel *Malaise* AND he/she feels *Sweating at Nights*	66.0%
14	IF a person has the categorical value 'Yes' or 'No' on *Coughing* AND he/she feels *Sweating at Nights* AND he/she is a *non-Smoker*	62.5%
15	IF a person has the categorical value 'Yes' on *Loss in Weight* AND he/she has the categorical value 'No' or 'Yes' on *Coughing* AND he/she is a *non- Smoker*	62.3%
16	IF a person has the categorical value 'Yes' or 'No' on *Coughing* AND he/she does not feel *Sweating at Nights* AND he/ she has the categorical value 'Yes' on *Loss of Appetite*	60.6%
17	IF a person has the categorical value 'Yes' or 'No' on *Coughing* AND he/she does not feel *Malaise* AND he/she does not feel *Sweating at Nights*	33.8%
18	IF a person has the categorical value 'No' on *Loss in Weight* AND he/she has the categorical value 'No' or 'Yes' on *Coughing*	31.9%
19	IF a person has the categorical value 'Yes' or 'No' on *Coughing* AND he/she does not feel *Sweating at Nights* AND he/ she has the categorical value 'No' on *Loss of Appetite*	31.1%

Appendix E

Patient's Profile

No. ( ) **Patient's Name ……..** **Township…….** **Date ……..**

1. Gender	2. Age	3. Weight	4. Smoking Addiction	5. Alcohol Addiction	6. BCG Vaccine	7. Malaise	8.Arthralgia	9. Exhaustion (Tiredness)	10.Unwillingness for Work
○ Male ○ Female	……Years	…….(Kgm)	○ None ○ Little(<5 items) ○ Moderate(6-10 items) ○ Very Much(11+)	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes

11.Loss of Appetite	12.Loss in Weight	13.Sweating at Nights	14. Chest Pain	15. Back Pain	16. Coughing	17. Hemoptysis	18. Fever	19. Migrate	20. Diabetes
○ No ○ Yes	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes	○ No ○ Yes ○ with Mucous	○ No ○ Yes	○ Normal ○ High ○ Subfebrille	○ No ○ Yes ( )	○ No ○ Yes

21. ESR	22. Haematocrit	23. Haemoglobin	24. Leucocyte	25. Number of Leucocyte Type	26. Active Specific Lung Lesion	27. Calcific Tissue	28. Cavity
○ Normal ○ Moderate ○ High	○ Normal ○ Low ○ High	○ Normal(10-15) ○ Low (<10) ○ High (>15)	○ Normal ○ Low ○ High	○ Normal ○ Lymphocytic Dense ○ Macrophage Dense	○ No ○ Yes	○ No ○ Yes	○No ○Yes

29. Pneumonic Infiltration	30. Pleural Effusion	31. HIV	32. Sputum AFB	33. Gxpert			
○ No ○ Yes	○ No ○ Yes	○ Negative ○ Positive	○ Negative ○ positive	○ Negative ○ Positive			

Outcome: ○TB ○Non-TB