An Integrated Supervised and Unsupervised Learning Approach to Predict the Outcome of Tuberculosis Treatment Course

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2011

Sharareh RostamNiakanKalhori

School of Computer Science

TABLE OF CONTENTS

The Final Word Count: 44297

LIST OF TABLE

TABLE OF FIGURE

ABSTRACT

The University of Manchester, Faculty of Engineering and Physical Sciences
Abstract of thesis submitted by Sharareh RostamNiakanKalhori for the degree of Doctor of Philosophy in Informatics.
Entitled: Integrated Supervised and Unsupervised Learning Method to Predict the Outcome of Tuberculosis Treatment Course

Tuberculosis (TB) is an infectious disease which is a global public health problem with over 9 million new cases annually. Tuberculosis treatment, with patient supervision and support is an element of the global plan to stop TB designed by the World Health Organization in 2006. The plan requires prediction of patient treatment course destination. The prediction outcome can be used to determine how intensive the level of supplying services and supports in frame of DOTS therapy should be. No predictive model for the outcome has been developed yet and only limited reports of influential factors for considered outcome are available.

To fill this gap, this thesis develops a machine learning approach to predict the outcome of tuberculosis treatment course, which includes, firstly, data of 6,450 Iranian TB patients under DOTS (directly observed treatment, short course) therapy were analysed to initially diagnose the significant predictors by correlation analysis; secondly, these significant features were applied to find the best classification approach from six examined algorithms including decision tree, Bayesian network, logistic regression, multilayer perceptron, radial basis function, and support vector machine; thirdly, the prediction accuracy of these existing techniques was improved by proposing and developing a new integrated method of k-mean clustering and classification algorithms. Finally, a cluster-based simplified decision tree (CSDT) was developed through an innovative hierarchical clustering and classification algorithm. CSDT was built by k-mean partitioning and the decision tree learning. This innovative method not only improves the prediction accuracy significantly but also leads to a much simpler and interpretative decision tree.

The main results of this study included, firstly, finding seventeen significantly correlated features which were: age, sex, weight, nationality, area of residency, current stay in prison, low body weight, TB type, treatment category, length of disease, TB case type, recent TB infection, diabetic or HIV positive, and social risk factors like history of imprisonment, IV drug usage, and unprotected sex ($P \leq 0.048$); secondly, the results by applying and comparing six applied supervised machine learning tools on the testing set revealed that decision trees gave the best prediction accuracy (74.21%) compared with other methods; thirdly, by using testing set, the new integrated approach to combine the clustering and classification approach leads to the prediction accuracy improvement for all applied classifiers; the most and least improvement for prediction accuracy were shown by logistic regression (10%) and support vector machine (4%) respectively. Finally, by applying the proposed and developed CSDT, cluster-based simplified decision trees were optioned, which reduced the size of the resulting decision tree and further improved the prediction accuracy.

Data type and having normal distribution have created an opportunity for the decision tree to outperform other algorithms. Pre-learning by k-mean clustering to relocate the objects and put similar cases in the same group can improve the classification accuracy. The compatible feature of k-mean partitioning and decision tree to generate pure local regions can simplify the decision trees and make them more precise through creating smaller sub-trees with fewer misclassified cases. The extracted rules from these trees can play the role of a knowledge base for a decision support system in further studies.

DECLARATION

that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

To the God, the unique creator who made all things possible

To the God`s great gift: love

To my father's hands and my mother's eyes

ACKNOWLEDGEMENTS

I would like to acknowledge and extend my gratitude to my supervisor Dr. Xiao-Jun Zeng for his help, guidance and encouragement.

Thanks to Iranian Ministry of Health and Medical Education for financial support during my PhD study.

Many thanks to TB & Leprosy Division, Disease Control Department of Iranian Ministry of Health and Medical Education for data access.

Especially heartfelt appreciation to my parents, for my father’s supports and advices during my study life and my mother`s wonderful encouragements and prayers.

Also, I would like to thank my brothers and sisters, my friends for their precious attention and endorsement.

Last but not least, special thanks to Dr. Bahman Abedi Kiasari for steadfast inspiration, generous supports, and valuable comments in medical point of view.

LIST OF PUBLICATION RELATED TO THE THESIS

- Predicting the Outcome of Tuberculosis Treatment Course in Frame of DOTS, From Demographic Data to Logistic Regression Model, Second International joint conference on Biomedical Engineering Systems and Technologies, Biostec 2009, January 14-17 2009, Porto, Portugal.

- Fuzzy Logic Approach to Predict the Outcome of Tuberculosis Treatment Course Destination, Proceedings of the World Congress on Engineering and Computer Science 2009 Vol. II, WCECS 2009, October 20-22, 2009, San Francisco, USA.

- A Logistic Regression Model to Predict High Risk Patients to Fail in Tuberculosis Treatment Course Completion, IAENG International Journal of Applied Mathematics, 40:2, IJAM_40_2_08.

Chapter 1

Introduction

1.1Purpose and Scope

Creating predictive (classification) models is one of the machine learning applications in order to uncover novel, interesting, and useful knowledge from large volumes of data in many medical domains such as diagnosis, prognosis and treatment. They are successfully developed through applying several machine learning techniques.

In the area of tuberculosis control, no model to predict the outcome of treatment courses has been developed. A predictive model should be able to define patient treatment destination and confirm whether or not each patient finishes a complete course of treatment entirely.

Tuberculosis (TB) is a global public health concern known as a major contributor to the global burden of disease, with around 9 million new cases worldwide in 2005 and 2 million deaths estimated to occur annually [Harries & Dye, 2006]. Over recent years, this infectious disease has been considered intensely, particularly in low- and middle-income countries where it is being fuelled by the HIV/AIDS epidemic [Thiam *et al*., 2007]. The main goals for TB control programme are case detection and treatment success; countries that detect 70% of all estimated TB cases and successfully treat 85% of detected TB cases should expect declines in incidence of 8-12% per year [Dye *et al*., 1998]. The current internationally recommended control strategy for TB is named DOTS (directly observed treatment, short course) and involves delivery of a standard short course of drugs, lasting 6 months for new patients and 8 months for patients diagnosed with TB. The delivery includes the direct observation of therapy (DOT), either by health staff or by a DOT supporter known by the patients for this purpose. Since 1997, the World Health Organization (WHO) has made an effort to fulfil tuberculosis control programme targets through DOTS, introduced as a widely promoted and globally implemented strategy. According to the DOTS approach, a major determinant of the outcome for a tuberculosis treatment is patient compliance; nevertheless, up to 50 percent of all patients with TB do not complete treatment and fail to adhere to their therapy [Cuneo& Snider, 1989]. It has been estimated that in industrialised countries non-completion of treatment is around 20% [Tangüis *et al*., 2000] and according to the Centres for Disease Control and Prevention in the United States, 25% of patients fail to complete their chemotherapy [Yew, 1999].

Noncompliance is a significant factor leading to the persistence of tuberculosis in many countries and the consequences of this well recognized fact are prolonged infectiousness, relapse, prolonged and more expensive therapy, development of drug resistance, and death [Thiam *et al*., 2007]. It has been revealed that noncompliance is associated with a 10-fold increase in the incidents of poor results from treatment and accounted for most treatment failures and although patients who fail to respond to therapy or suffer a relapse due to noncompliance are a minority of those with active tuberculosis, they may have an inconsistent effect on tuberculosis epidemiology [Burman *et al*., 1997]. Multidrug-resistant tuberculosis (MDRTB) is another noncompliance consequence well known as a formidable clinical and public health emergency.

1.1.1 Multidrug-Resistance Tuberculosis

Tuberculosis (TB) can usually be treated with a course of four standard or first-line anti-TB drugs. However, lack of drug therapy compliance, misuse or mismanagement of therapy which means taking the drugs in the wrong combination, taking fewer than prescribed, or taking insufficient doses or not at the proper time can lead to multidrug resistant TB (MDR-TB) [Yew, 1999]. In this condition, patient are resistance to the most important anti-TB drugs, *i.e.* isoniazid, and rifampicin happens and takes longer to treat with second-line drugs which are more expensive with more side-effects. Lack of adherence to the course, misuse or mismanagement of these drugs can also lead to extensively drug-resistant TB (XDR-TB) which is highly resistant to first- and second-line drugs where treatment options and the chances of cure are seriously restricted [Sampathkumar, 2008]. Compared with drug-sensitive TB cases, the treatment outcomes for this condition have been much less positive with the overall expected cure rate decreasing to 60%. The mortality rate is high, particularly for those subjects who have been co-infected by HIV. In addition, second-line drugs are much more expensive as poor patient affordability is also encountered [Yew, 1999]. For people with the MDR-TB condition, because of diagnostic delays, overcrowding and inadequate infection control, undesirable side of this difficulty might be accelerated; thus, prevention of MDR-TB should be of high priority, particularly in countries with limited resources.

1.1.2 DOTS & ‘’STOP TB Plan’’

DOTS is an internationally recommended approach currently implemented to control tuberculosis. It is aimed at preventing the transmission of *M. tuberculosis* and associated disease and death, through using combinations of anti-TB drugs to treat patients with active TB carefully. Case detection and completed treatment are known as two major targets of DOTS. The World Health Organization (WHO) has defined five distinct elements for DOTS including political commitment, microscopy services, drug supplies, surveillance and monitoring systems and use of highly efficacious regimens, and direct observation of treatment. In fact, ensuring that the patient completes therapy to cure the disease and prevent drug resistance from developing are major purposes of DOTS [Davies, 2003].

DOTS has played an impressive role in TB control since 1997 when it became the mostly widely-implemented and longest running global health intervention in health history. Known as a foundation strategy for TB control, it has encouraged patients to complete their therapy. In fact, the adoption of DOTS has been associated with reduced rates of treatment failure, relapse, and drug resistance [Burman *et al*., 1997]. Research demonstrates that the treatment compliance rate has increased from 25-50% with unsupervised treatment to 80-90% with DOTS application [Davies, 2003]. Likewise, according to Burman *et al*. (1997) DOTS application to TB control has descending effect on noncompliance rates as this rate decreased significantly from 13.3% by applying self administration of treatment to 5.9% through directly observed therapy, short-course ($p<0.05$) [Vieira & Ribeiro, 2008].

Hence, noncompliance which is widely acknowledged as the main cause of the failure of initial therapy and relapse is common in self-administered multidrug tuberculosis treatment regimens. Anuwatnonthakate *et al*. (2008), in the study about applying DOTS for TB control and its outcome improvement in Thailand, has proved the positive effect of directly observed therapy (DOT) compared with self administration with 93% rather that 69% treatment completion. Furthermore, there is a significant success rate for patient treatment compliance where they receive DOT by health care worker (93%) compared with a family member (69%). It has been expressed that increasing the provision of DOT by healthcare workers should be considered to prevent prolonged and expensive therapy that is less likely to be successful than the treatment of drug-susceptible tuberculosis.

In response to these findings, there is increasing emphasis on both use of DOT, in which a health-care worker observes the ingestion of each dose of anti tuberculosis therapy, and improving the DOTS performance by using a new scheme introduced by the World Health Organisation in 2006.

Global Plan to Stop TB (2006–2015) focuses on intensified TB-case finding, treatment of latent TB infection with isoniazid, prevention of HIV infection, cotrimoxazole preventative therapy, and antiretroviral therapy [Harries & Dye, 2006]. The Stop TB Strategy consists of six main components, which are as follows:

- Pursue high-quality DOTS expansion and enhancement
- Address TB/HIV and MDR-TB and other special challenges
- Contribute to health system strengthening
- Engage all care providers
- Empower people with TB, and communities
- Enable and promote research

DOTS expansion and enhancement is the foundation of both DOTS strategy and the other five components. It is vital to understand that the sequence and scale of implementation and the speed of activities building on DOTS will vary according to the setting and accuracy of basic DOTS implementation.

In order to put further focus on the first component of the ‘’Stop TB Plan’’ which is to pursue high-quality DOTS expansion and enhancement, addressing known limitations and meeting new challenges, additional reinforcement of the basic components of the DOTS strategy is required. The following are those lines that can improve the first component.

- Political commitment with increased and sustained financing
- Case detection through quality-assured bacteriology
- Standardized treatment, with supervision and patient support

- An effective drug supply and management system
- A system for monitoring and evaluation system, and impact measurement

For the element of supervision and patient support, it has been highlighted that health care services should identify and concentrate on interruption factors that halt treatment. Supervision, which plays a strong role in patient treatment adherence and preventing the development of drug resistance, must be carried out in a context-specific and patient-sensitive manner, and is expected to ensure commitment of both providers to give proper care and support and patients to receive regular treatment. It has been brought to light that selected patient groups, for example prisoners, drug users, and some people with mental health disorders may need intensive support including DOT [WHO, 2006]. Although WHO has highlighted the necessity of improving the quality of DOTS in terms of supervision and patient support in the ‘’Stop TB’’ plan, there is no way to measure how intensive health workers` support and supervision should be for patients. To make this supervision more context-specific and patient–sensitive, we may require a tool to predict the patient destination regarding TB treatment course completion. In conclusion, here is a summary of the reasons why we need a predictive system capable of forecasting the outcome of providing DOTS therapy for each patient specifically based on their own particular features:

- Non-adherence to the tuberculosis treatment course is a complex phenomenon and task-specific behaviour
- DOTS has produced better results than the method of patients’ self supervision
- DOTS is an expensive services package that result in active supervision and support for all patients
- Although WHO has understood that some patients need more support but it hasn’t pointed out how and to what extent it should be provided

Having considered these parameters including the complex entity of treatment course non-compliance, the relative success of DOTS, WHO` emphasis on supervision and support in different degrees for specific groups of patients, and the impossibility of serving all TB patients with active supervision and support, there is requirement for a tool to define the level of supervision and support each patient needs based on the predicted outcome defined by an accurate predictive model.

1.2 Research Objectives

This study pursues the overall objective of developing and then improving the most accurate and understandable predictive model to forecast the outcome of tuberculosis treatment courses through combining supervised and unsupervised learning methods. To meet this end, the detailed tasks can be listed as follows:

- Compare all potential methods
- Select the most effective one
- Develop the new combined algorithm which integrates supervised and unsupervised learning to improve the accuracy
- Enhance the interpretability of novel combined algorithm of the supervised and unsupervised learning method

1.3 Main Contributions

This thesis is the first systematic and quantitative analysis and prediction of TB based on the machine learning approach, which includes, firstly, a comparison of the main existing methods and identification of the most effective one using current techniques; secondly, recognition of the weaknesses of the existing methods and proposing/developing an innovative approach by combining the supervised and unsupervised learning which will improve the accuracy and enhance the interpretability for knowledge discovered. The more detailed contributions are listed as follows:

- Analysing a set of tuberculosis patient features to discover influential factors which effect the outcome of the tuberculosis treatment course
- Comparing the applied classifiers to predict the outcome of the tuberculosis treatment course and find the most accurate and valid classification algorithm
- Proposing the combination of clustering and classification methods to improve the classifier`s performance in terms of accuracy
- Utilising the k-mean clustering method to enhance the most accurate classifier`s interpretability

1.4 Overview of the Thesis

The contributions of this thesis have been presented in the related chapter as follows:

Chapter Two presents background study relating to the importance of the completion of treatment course, critical analysis to choose classification techniques and the algorithms of classifiers, influential factors effecting tuberculosis treatment adherence and also a brief literature review on applying machine learning tools to classify various tasks in the medical domain.

Chapter Three is an overview of the methodology employed to meet the objectives of this research. This chapter includes an explanation of the choice of supervised and unsupervised methods, and the decision to combine them to improve the accuracy and comprehensibility of predictive models, as well as a detailed look at the research method.

Chapter Four defines significant factors affecting the outcome of the tuberculosis treatment course through patient data analysis. Applied classification models developed by selected classifiers are also presented in this chapter. Various criteria for comparing them are utilised and the results which are yielded are represented and discussed.

Chapter Five focuses on the combination of cluster analysis to every single one of the developed classified models to improve their accuracy and the comparison of results based on different numbers of clusters and the applied algorithm of classifiers.

Chapter Six investigates the application of hierarchical clustering and classification method to reduce the tree size and misclassification rate through proposing and developing a cluster-based simplified decision tree (CSDT).

Chapter Seven concludes the major contributions and work fulfilled by this research. The significance of the result is summarised and major conclusions are drawn from the present work. Directions for future work are also proposed in view of how this research can be extended in the future.

Chapter 2

Research Background &

Literature Review

2.1 Introduction

In the first part of this chapter, the background of the prediction system for the course of tuberculosis treatment is discussed. The importance of the tuberculosis treatment course prediction, data analysis and feature selection methods, supervised and unsupervised learning tools, critical analysis to choose the proper techniques and evaluation frameworks are reviewed. Then, the literature review related to the predictive factors of the destination of tuberculosis treatment and a number of applied predictive systems in various medical areas are investigated.

2.2 Predicting the Outcome of Tuberculosis Treatment Course

In the former chapter, the importance of patients' adherence to tuberculosis therapy as a major determinant of tuberculosis control was illustrated. In international tuberculosis control approach, prediction of treatment course destination has not been the center of attention either in research or in practice.

Tuberculosis prevalence rate is reported as 11.1 million cases per year with 9.4 million annual deaths. It is fueled by HIV and needs to be controlled particularly in African country [WHO, 2006]. The erratic adherence to chemotherapy and irregular medication intake is the most common cause of relapse and development of drug resistance disease known considerably as more difficult cases because of no response to the standard treatments by the first or even the second line drugs [Harries & Dye, 2006].

Directly observing TB patients was piloted in the 1950s to ensure that patients adhered to and completed their treatment. Patients were observed taking their anti-TB treatment either daily or several times a week to insure adherence and treatment completion. Consequently, over the last four decades the tuberculosis cure rate has reached 82% and this significant progress in TB treatment delivery was made when both DOTS strategy was applied and the use of rifampicin (RMP) was initiated [Panjabi *et al.*, 2007]. However, prolonged infectiousness, drug resistance, relapse and death are still difficulties experienced by up to half of TB patients who do not complete their treatment course [Munro *et al*., 2007]. Efforts to improve treatment outcomes require a better understanding of the particular facilitators and barriers to TB treatment adherence since it is known as a complex behavioral issue influenced by the interaction of a number of factors. The "Stop TB" strategies designed by the WHO in 2006 focused on making the best use of currently available tools for the diagnosis, prevention and treatment of TB. This plan needs the improved tools that are likely to become available through research. Also, influential factors, including patient-centered interventions to address structural barriers to treatment adherence, have been classified. In the first component of the "STOP TB strategy", supervision and patient support are emphasized with the focus on identifying and addressing factors that may interrupt or stop treatment as well as supervising the treatment which helps patients to take their drugs regularly and completely, through direct observation of therapy (DOT). Although the WHO has highlighted that patient supervision and support should be carried out in a context-specific and patient-sensitive manner, no tool has been introduced which would enable the level of support a patient requires to be determined from their situation. Likewise, in this component of the "Stop TB" strategy, it is noted that certain patient groups, such as prisoners, drug users, and some people suffering from mental health disorders, may need intensive support including DOT, but the word 'intensive' is not defined specifically.

Currently, at clinic every patient interview and educational session is carried out by nurses when the therapy is initiated. At the onset of the course, nurses estimate the patients' understanding of the treatment process and pursuing the therapy up to completing the course based on their condition. Nurses record their own judgments in the nursing book in addition to telephone numbers and contact details. Furthermore, during follow-up, the nurses estimate the patient's compliance based on punctuality, attendance, telephone interviews and, in some instances, pill counts and home visits. These estimations and records are just for nursing purposes and they are not included in the hospital dossier, nor routinely reviewed by the treating physicians. Hence, there is a manual system to record the prediction of tuberculosis treatment course based on the nurses' estimation and no particular system is available to accurately define the treatment course destination according to patient features and conditions. A systematic method using all of the known influential features of TB patients to predict the outcome of tuberculosis treatment course, instead of nurses' estimation, would help DOTS to transform from a passive to an active system through finding patients who are at high risk of noncompliance of chemotherapy. In other words, based on various possible outcomes for the completion of a course of tuberculosis treatment which would be predicted by the designed system, the needed follow-up care and supervision can be defined. This may assist health providers and nurses to supply services at different levels.

To develop accurate predictive models, correct techniques and a suitable database related to tuberculosis cases are essential. The database should contain patients` records with details of the features of their cases of TB as well as the related outcome of their course of treatment; thus, the difficulties associated with limited data access with enough records and corresponding features might be the reason for no predictive system being available. Potential predictors and patients` variables should be analysed to find the influential factors as this knowledge would lead to the development of a predictive model. Besides, to develop a predictive model, there are technical requirements such as machine learning tools which are discussed in the next sections followed by a feature analysis to detect significant predictors.

2.3 Classification and Regression

The development of a predictive model can be categorised either by classification or regression tasks. In the case that the output is a continuous number, regression tasks should be considered whereas in the situation that the outcome is a discrete number (such as a predefined set of classes or categories) a classification task should be considered.

For the prediction of tuberculosis treatment course completion, the defined outcome related to each record of a TB patient contains five potential classes: cure and competed treatment (as desirable outcomes), failure, quit, and death (as undesirable outcomes). Cure has happened where the final sputum result is smear or culture negative, the case of treatment completion but no available proof of negative specimen is called completed treatment. The aim of the WHO is to achieve 85% or more TB cases in this category. Another category is death which is inevitable, even in developed counties 4-8% of cases might have this outcome for tuberculosis treatment course application. The category of failure is another outcome in which the patient sputum has not converted. In cases of relapse, sputum becomes positive after reverting to negative. Finally, there are those cases of patients who quit the therapy and forgo the follow-up which are regarded as undesirable outcome [Davies, 2003]. Thus, the problem of predicting the outcome of a course of tuberculosis treatment is a classification task which maps each item of the training dataset including patients' records with their corresponded set of attributes into one of a predefined set of classes.

2.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach for data analysis that applies a number of techniques to maximize insight into a dataset, uncover underlying structure, and test underlying assumptions [Field, 2005]. Data analysis investigates the common assumption about what type of model the available data follow with the more direct approach; it allows the dataset itself to reveal its underlying structure and producing model. To carry out EDA, two main techniques are available: quantitative test and graphical methods. There are collections of techniques that cover data analysis objectives; however, statistical analysis is one of those methods which perform very well in both quantitative tests and graphical presentations. This method provides numerous measurements and tests such as measures of location (mean and median), confidence limits for the mean, one sample t-Test, chi-square test, and skewness & kurtosis measures.

2.5 Feature Selection

For every domain in medicine, many candidate features are introduced; however, many of them are either to some extent or totally irrelevant or redundant to the considered concept. In the case of a large dataset, learning the dataset is not useful unless the unwanted features are removed since an irrelevant and redundant feature does not affect or add anything new to the target concept [Dash & Liu, 1997].

Assisting data visualization and data understanding, reducing the measurement and storage requirements, decreasing training and utilization times and challenging the curse of dimensionality to improve performance are some of possible benefits of feature selection.

One of the main aims of feature analysis is improving the prediction performance of the predictors and another is providing quicker and more cost-effective predictors. Feature selection attempts to choose a subset of features and reduce the size of structure without significantly decreasing the accuracy of the classifier developed through using the selected features. There are two major steps of feature analysis which are generation procedure and evaluation function.

Complete, heuristic, and random are three main categories of generation procedure. In the complete approach, a complete search is conducted for the optimal subset based on the evaluation function used. In the heuristic category, the generation process of subsets is basically incremental either by increasing or decreasing; in this approach all of the remaining features yet to be picked (or rejected) are considered for selection (or rejected); these steps are very simple and quick because of the quadratic number of features in the search space. The third generation method, random, requires a smaller number of parameters and iteration procedure than the other two methods. Assigning appropriate values to these selected parameters is a significant task for obtaining good results.

The evaluation function tries to measure the discriminating ability of a feature or a subset to distinguish the different class labels. There are five categories of evaluation function: distance, information (or uncertainty), dependence, consistency and classifier error rate [Dash & Liu, 1997]. Based on the types of generation procedure and evaluation function and their combination, there are thirty-two available methods [Dash & Liu, 1997]. Each can be applied in a suitable way and there is no single technique appropriate for all applications. The choice of a feature selection method depends on the dataset characteristics including data type, data size and noise. According to the method`s capability of handling different feature values (continuous, discrete, and nominal), dealing with binary or multiple classes, performing well for a small training set or large data size and finally producing an optimal subset from noisy data, the feature analysis technique is selected. Table 2.1 lists the ability regarding these above mentioned characteristics of sixteen feature selection methods which assist in the choice of the appropriate technique [Dash & Liu, 1997]. The combination of heuristic for generation and dependence for the evaluation function is the best option where there is a large dataset and the feature values are a mixture of continuous, discrete and nominal data with multiple classes. It is possible to use dependence or correlation measures to qualify the ability to predict the value of one variable based on the value of another. The correlation between a feature and a class is the main result of applying a coefficient. Pearson's coefficient and Spearman's rho are two measurements of correlations reflecting the degree of relationship between two variables [Field, 2005]. Also, in Table 2.1, important method in each category of feature selection is addressed. In heuristic generation function, there are several methods such as Relief, Rough sets, Koller & Sahami, decision trees and Average Correlation Coefficient. The feature selection process either starts from the empty set and in each iteration generates new subsets by adding a feature selected using some evaluation function or begins from the complete feature set and in every iteration new subsets is chosen by discarding a feature. Among these methods, average Correlation Coefficient is an easy to use and common method which qualifies the ability of each feature to predict the value of one variable from the value of another. This classical method produces the coefficient which indicates the degree of redundancy of the feature. In the case of too many features and assumption of linear relationship between feature and a class, correlation coefficient simply produces the list of correlated features [Guyon & Elissee, 2003]. However, other methods using rough set in this category of feature selection first finds a reduct and then remove all features not appearing in the reduct; then, it tries to rank the features based on their significance. This measure is based on dependence of attributes; however, the method which developed based on this theory, such as PRESET, is week to handle noise in the dataset and is not able of finding the optimal subset. It seems PRESET also suffer the imperfection of Pearson and Spearman correlation coefficient [Dash&Liu, 1997; Guyon & Elissee, 2003].

Table 2.1 list of feature selection methods and their ability to handle data characteristics

Category Generation/evaluation function	Important Method in each category	Data Type			Multiple Classes	Large Dataset	Noise
		C	D	Nd			
Heuristic/distance	Relief	N/Y	Y	Y	N/Y	Y	Y
Complete/distance	branch and bound (B & B)	Y	Y	N	Y	---	---
Heuristic/information	Decision tree Koller &Sahami	N/Y	Y	N/Y	Y	Y	---
Heuristic/dependence	Average Correlation Coefficient/rough set	Y	Y	Y	Y	Y	---
Random/ classifier error rate	Focus	N	N/Y	N/Y	Y	N	N
Random/consistency	LVF	N	Y	Y	Y	Y	Y

C=Continuous data, D=Discreet data,
Nd=Nominal data,
N=No, Y=Yes

2.6 Supervised Machine Learning Techniques

Supervised learning is applied to make predictions about future cases where current available instances are given with known labels (the corresponding correct outputs). Supervised machine learning involves trying to find the algorithms that learn from externally supplied instances in order to produce general hypotheses. The main goal of supervised learning is model development derived from the distribution of class labels in terms of predictor features selected by feature analysis. Then, the resulting classifier is applied to allocate class labels to the testing instances where the values of the predictor features are identified, but the value of the class label is unknown.

Many supervised classifiers are currently available; they have been categorized in main groups like logic-based methods, perceptron-based techniques, statistical learning algorithm and support vector machines [Alpaydin, 2004]. Pros and cons of some of the most applied techniques are discussed in the next section. The process of learning a set of rules from examples in the training set and consequently creating a classifier which is capable of generalizing from new-real life instances is described in Figure 2.1.

The first step is collecting a set of data associated with the problem that needs to be solved. Then, the most informative attributes should be identified by experts or using feature selection techniques. Data preprocessing and preparation is conducted mainly through handling missing data, noise and the detection of outliers; it reduces the amount of data which results in the data mining algorithm functioning more effectively, particularly in the case of very large sets of data.

After the pre-processing stages, the whole dataset is randomly divided into two parts named training and testing sets. The training set is two-thirds of the whole dataset which is used to construct the models by any classification algorithm. The testing set is the other third which is used to validate the developed model and to choose the optimal parameter configuration. After training classification algorithms by training set, models are developed; they are evaluated with the testing set and in the case of satisfactory accuracy and other criteria, the predictive model is chosen [Han & Kamber, 2006].

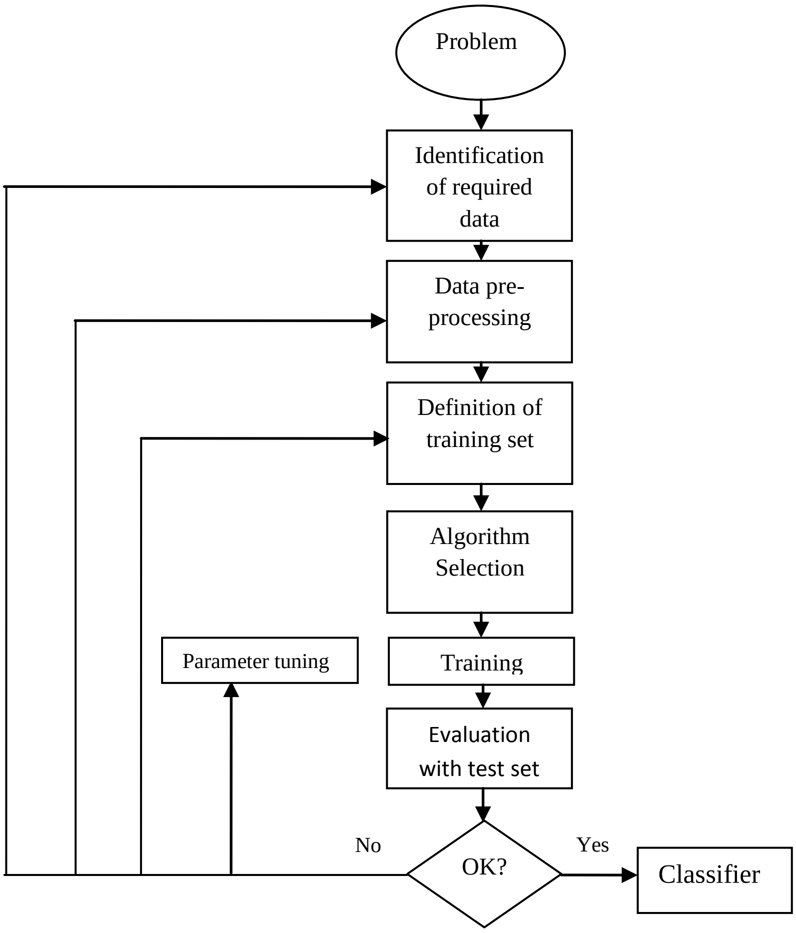


Figure 2.1 The process of supervised machine learning for classification task, adapted from Kotsiantis (2007).

2.6.1 Critical Analysis for Classification Algorithm Selection

The increasing number of electronic databases containing medical data has led to an increasing interest in building classification models by using a variety of statistical and machine learning approaches. However, it is recognized that critical analysis is required to demonstrate what features of an algorithm make it successful on specific dataset to support a particular algorithm. The major criteria including accuracy, the cost of misclassification, the time taken to produce results, the comprehensibility of the results and the ease of applying the algorithm in real life are defined [King *et al.*, 1995]. Each classification algorithm presents some characteristics which may be interesting in the context of clinical prediction tasks.

Decision trees (DT), neural networks (NN), support vector machines (SVM), Bayesian networks (BN), K-nearest Neighbor classifiers (K-NN), Logistic Regression (LR), and Radial Basis functions (RBF) are some of those applied classification algorithms for medical datasets and examples in several studies [Wetter, 2000]. Here, we discuss a number of their pros and cons to find out the most suitable algorithms for current research and the available set of data.

In the case of multi-dimensions and continuous features, the SVM and NN tend to perform well. For handling discrete/categorical features, the DT as a logic-based system performs better than other algorithms.

ANNs and SVM are two methods which need a large sample size to attain their maximum prediction accuracy whereas the BN may only need a relatively small dataset.

K-NN is very sensitive to irrelevant features because of the way that this algorithm works. In addition, the presence of irrelevant and redundant features decreases the level of the NN performance. In DT development, division of the instance space is orthogonal to the axis of one variable and parallel to all other axes; it makes the resulting regions into hyper-rectangles after partitioning. Thus, this classifier is unable to perform well where problems requiring diagonal partitioning are concerned.

BN train very quickly because they need only a single pass on the data either to count frequency in the case of discrete variables or to compute the normal probability density function when there are continuous variables under normality assumptions. They need little storage space over the training and classification phases; the minimum space required is the room needed to store the prior and conditional probabilities. However, the k-NN algorithm utilizes a large amount of storage space for the training phase, and its implementation space is at least as big as its training space. On the contrary, apart from the lazy learners, for other learning algorithms the execution space is usually much smaller than the training space, since the resulting classifier is usually a condensed abstract of the data. Lazy learning methods require zero training time since the training instance is simply stored. Decision trees are reputed to be a bit quicker than NNs and SVMs [Kotsiantis, 2007].

In terms of the ability to tolerate noise, kNN is known as a fine method [Kotsiantis, 2007]; measures similar to it can be easily distorted by errors in attribute values leading them to misclassify a new instance according to the wrong nearest neighbors. In contrast to kNN, rule learners and most DTs are judged as a tool resistant to noise because their pruning strategies avoid overfitting the data in general and for noisy data in particular.

Being very easy to interpret is another classifier characteristic that make logic-based algorithms like DTs preferable to other methods whereas NNs, SVMs, and k-NN are well known for poor understandability [Kotsiantis, 2007]. For medical interpretation, decision trees can help with the understanding of predictions. Although there is a wide interest in the application of NN, several limitations cause a number of difficulties in day-to–day practice. The most important critical weakness of NNs are their black box nature with not readily providing an explanation of their prediction; not being able to explicitly identify possible relationships among variables is the main reason that NNs are known for their poor interpretability [King *et al.*, 1995; Lee & Abbott, 2003]. Furthermore, the model that results from using a neural network is not fixed since the iterative learning process can continue on data. These methods probably have the potential to complement available statistical models and to contribute to the interpretation and presentation of computerized decision support systems. To predict the outcome of a course of tuberculosis treatment, the predictive model needs to be precise since, according to the destination of therapy, the level of therapy supervision and support is defined. The more accurate the model is, the higher the quality of health care provided to TB patients will be. Moreover, the supervision and support of TB patients is provided by health workers who are not expert enough in medicine and, also, play a core role in health promotion and maintenance. The resulting model needs to be as comprehensible as possible.

Table 2.2 Comparison of the characteristics of supervised learning techniques, adapted from Kotsiantis (2007)

	DT	NN	BN	SVM	RBF	LR
Accuracy in general	**	***	*	****	***	***
Speed of learning with respect the number of attributes and the number of instances	***	*	****	*	**	***
Speed of classification	****	****	****	****	****	****
Tolerance to missing values attributes	**	*	****	**	*	**
Tolerance to irrelevant attributes	***	*	**	****	*	**
Tolerance to redundant attributes	**	**	*	***	**	**
Tolerance to highly interdependent attributes	**	***	*	***	**	*
Dealing with discrete/binary/continuous attributes	****	*** (Not discrete)	*** (Not discrete)	** (Not discrete)	** (Not discrete)	*** (Not discrete)
Tolerance to noise	**	**	***	**	**	***
Dealing with danger of overfitting	**	*	***	**	**	***
Attempts for incremental learning	**	***	****	**	**	****
Explanation ability/transparency of knowledge	****	*	****	*	*	***
Model parameter handling	***	*	****	*	*	***

Poor interpretability is well-known feature of neural networks as it is difficult to interpret the symbolic meaning behind the learned weights and hidden units in the network. However, their high tolerance of noisy data and their capability of classifying patterns on which they have not been trained are two advantages of neural networks. Besides this, they could even be used in cases where limited information of the relationship between attributes and classes is available. Their successful application to a wide range of real life data, of different types, has been well documented [Han & Kamber, 2006].

In contrast of NNs with week presentation of produced output, the Bayesian network provide graphical diagram which represent relationships and influences among predictors. These qualities of BNs help experts to specify dependence and independence of variables through the network structure. However, they are week at handling discrete variables, redundant and irrelevant attributes.

Decision tree is capable of expressing the degree of relationships between output and input variables; however, they are week to consider relationships among input variables. They are also sensitive to outliers and inflexible with respect to missing data; so that their performance may be dependent on quality of available dataset. DTs are good at coping with discrete variables with high tolerance to irrelevant attributes.

A predictive system is required to be precise and understandable. After applying feature analysis to select the most relevant factors, there are no redundant features and k-NN has been deleted from the classification algorithm list. The large amount of spaces required for the training phase is another reason for not choosing K-NN for this study. The other six classifiers which can produce accurate and understandable classification algorithm are examined in this research. In the following section, the characteristics and algorithm of selected classifiers are introduced.

2.6.2 Logic Based Algorithms

2.6.2.1 Decision Trees

Decision tree induction is the process of learning a tree from class-labelled training dataset. It is a flowchart-like tree structure where the internal node, branch and leaf node means concepts associated with our training tuples. In this hierarchical data structure, the local region is identified in a sequence of recursive splits in a smaller number of steps by implementing a divide-and-conquer strategy. The decision tree is a nonparametric estimation that the input space is divided into local regions defined by a distance measure like the euclidean norm; using the training dataset in the region, the related local model is computed. In local data defining the local model should be identified which needs calculating of the distances from the given input to all of the training tuples through(N). Every $f_m(A)$ defines a discriminant in the d-dimensional input space driving it into smaller regions; as we take a path from the root to down, these regions get increasingly subdivided. $f_m(.)$ is a simple function to create a tree. A complex function is divided into a series of simple decisions. Based on the type of $f_m(.)$, various decision tree techniques are expected to be developed defining the shape of the discriminant and regions. Figure 2.2 represents a set of data and the corresponding decision tree. Oval nodes are decision nodes and rectangles are leaf nodes. The univariate decision node splits along one axis and successive splits are orthogonal to each other. After the first split, $\{x|x_1 < w_{10}\}$ is pure and is not split further. Each leaf node has an output label, which in classification cases, should be the class code; whereas, in a regression tree, it is a number value. A leaf node is the region of input area that instances in this region have the same output. These regions have boundaries that are defined by the discriminants that are coded in the internal nodes on the path from the root to the leaf node. Due to the hierarchical placement of decisions, a fast localization of the region to cover an input is possible. For instance, if the decision is binary then every decision removes half of the cases; however, if there are b regions, then in the best case, the proper region can be localized in $\log_2 b$ decisions. Decision trees can be easily converted to a set of IF-THEN rules which are easy to assimilate [Alpaydin, 2004]. They are, indeed, the most widely used method of supervised learning. Building a decision tree doesn't need any specific domain knowledge or parameter setting which makes this tool suitable for exploratory knowledge discovery as it can easily cope with high dimensional data. Their representation of generated knowledge in a tree structure is intuitive and normally simple to understand by humans. Also, the learning and classification process of decision tree induction is uncomplicated and quick [Han& Kamber, 2006]; however, this method is not able to consider relationship among input variables. It has also the disadvantages of being sensitive to outliers and inflexible regarding to missing data. This is a threat for decision tree's performance for predicting new case [Lee&Abbott, 2003].

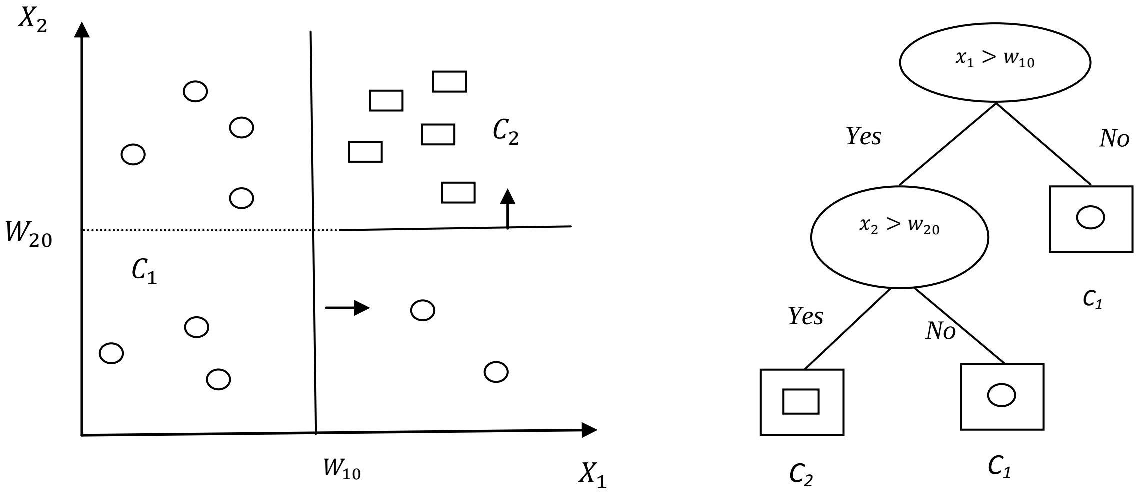


Figure 2.2 Example of data points and related decision tree, adapted from [Alpaydin, 2004].

2.6.2.2 Decision Tree Algorithm

ID3 (Iterative Dichotomiser) was the first decision tree algorithm developed by J.Ross Quinlan in the late 1970s and early 1980s. Afterwards, Quinlan introduced C4.5, a successor of ID3, and used as a scale to which newer supervised learning algorithms are often compared. The constructed decision trees have a top-down recursive structure; using divide and-conquer strategy, training set is applied to build trees where they recursively partitioned into small subsets over the process of tree growing [Alpaydin, 2004]. In more detail, the basic algorithm of decision trees contains three major parameters: a training data partition, an attribute-list (list of attributes describing the training data) and an attribute-selection-method. This method signifies a heuristic procedure for choosing the attribute that discriminates the given tuples in the best way on a class basis. In the first step, the tree starts as a single node, N, indicating the training tuples in A. Next, if all training set in A belongs to the same class, node N becomes a leaf and is labelled with that class; here is the finish point of condition.

The attribute-selection-method defines which attribute to test at node N, by determining the best way to separate or partition the training set in A into each class. Afterwards, the node N is labelled with the splitting criterion, which serves as a test at the node. Based on the next steps, a tree is grown from node N for every outcome obtained from the splitting criterion, and tuples in A are partitioned finally. Three possible settings could occur based on training tuples:

1. A is discrete-valued when the outcome of the test at node N matches the known values of A.

2. A is continuous-valued and the test at node N has two possible outcomes corresponding to the conditions $A \leq split - point$ and $A > split - point$ respectively.

3. *A* is discrete valued and a binary tree must grow.

2.6.2.3 Tree Pruning

When a decision tree is grown, noise or outliers in the training dataset can be addressed as various anomalies such as data overfiting. Pruning techniques resolve this by using statistical measures to delete the least reliable branches. Being less complex and smaller as well as easier to understand are the crucial features of a pruned tree. They are generally quicker and more accurate than un-pruned trees.

Pruning can be done by two approaches: pre-pruning and post-pruning. In the pre-pruning method, trees are pruned by halting their building early on through preventing further splits or partitions of learning tuples at a given node.

Whereas in post-pruning, which is known as a more regular method, sub-trees are removed from a fully grown tree and then a leaf is replaced at a given node. The most frequent class among the sub-tree would be applied there as the leaf. In the experiments of developing C4.5 decision tree, a pessimistic pruning approach is utilized. This method uses error rate estimation to make decisions about sub-tree pruning [Han & Kamber, 2006].

2.6.3 Perceptron-based Technique

2.6.3.1 Multilayer Perceptron Neural Network

The multilayer perceptron is known as an artificial neural network (ANN) structure and a nonparametric estimator that can be used for classification and regression. A neural network is a compound of linked input/output units in which every link has an associated weight. Adjusting the weights is the core phase of learning for predicting the correct class label of available input tuples. A feed-forward multilayer perceptron

(MLP) is a topology of the standard ANN, that a back propagation algorithm performs learning on [Han & Kamber, 2006]. In order to predict the class label of tuples, the back propagation algorithm performs learning on a multilayer feed-forward neural network iteratively. This network encompasses an input layer, one or more hidden layer and an output layer as shown in figure 2.3. Each layer consists of a number of units. The input layer is associated with the three variables (x_1, x_2, x_3) in the training dataset which concurrently feed into this layer. They pass through the input layer and, after weighting, feed to a second layer which is called neurolink units or the hidden layer. Afterwards, the output of the hidden layer units can be input to another hidden layer and so on. The number of hidden layers might vary and the weighted outputs of the last hidden layers are input to units composing the output layer to produce the network prediction for considered data (o_1, o_2, o_3).

The network is named feed-forward because none of the weights cycles back to input units or other units of previous layers. Classification by back propagation is the most applicable neural network learning algorithm in various fields. Back propagation learns by iteratively analysing the training dataset and comparing the network`s predictions for every tuple with the actual known target value this is either a class label of the training tuple from classification tasks or a continuous value for prediction problems.

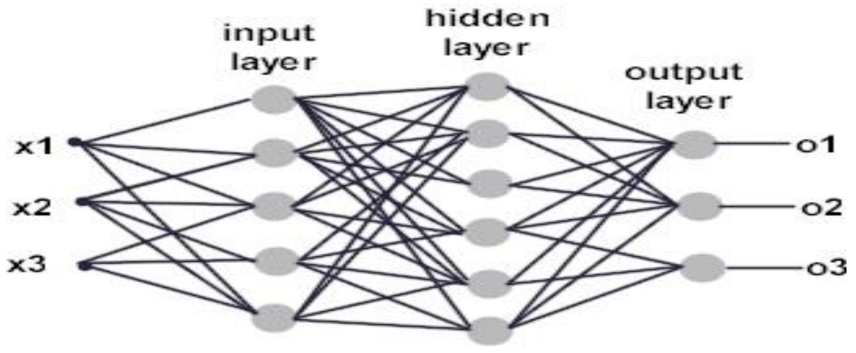


Figure 2.3 The architecture of a Multilayer perceptron neural network with one input, hidden and output layer, adapted from [Han & Kamber, 2006].

For every input variable, the weights are changing to minimize the mean squared error which is the difference between the neural network prediction and the actual target value. More precisely, the back propagation algorithm has three major stages as follows:

- Weight initialisation in the network with small random numbers varies from, for example, -1.0 to 1.0, or -0.5 to 0.5 when each unit has its own related bias.
- Propagation: the training data which feeds to the network`s input layer. For every input unit like j, the output O_j is equal and the same as its input I_j. That is, there is no change in the input layer. Afterwards, the net input and output for each unit in the hidden and output layers are computed; for unit j in a hidden or output layer, the net input I_j to unit for unit j in a hidden or output layer, the net input I_j to unit for unit j is:

$$I_j = \sum_i w_{ij}\, o_i + \theta_j \qquad (2.1)$$

Where, w_{ij} is the weight of the connection from unit i in the former layer to unit j, o_i is the output of unit i from the former layer, and θ_j is the bias of the unit. Every unit in the hidden and output layers utilizes its net input and then applies an activation function to it; this function can be either sigmoid or logistic and o_j, the output of unit j, is computed as:

$$O_j = \frac{1}{1+e^{-Ij}} \qquad (2.2)$$

Where I_j is the net input I_j to unit j. After calculating O_j for each hidden layer, the output layer produces a networks prediction.

- Backpropagate the error where the error is propagated backward by updating the weights and biases to reflect the error of the network’s prediction.

Neural networks have the critical weakness of being ‘‘black box’’ syndrome in which their models have no coefficients that can be interpreted clearly. These models therefore have a limited power of identifying possible relationships among variables. Although several works has been done to improve this weakness of ANNs by using sensitivity analysis or rule extraction, still this method suffers from black box’’ nature [Lee&Abbott, 2003].

2.6.3.2 Radial Basis Function

A radial basis function network (RBFN) is an artificial intelligence network in which its activation function is simply radial basis in a linear combination [Marsland, 2009]. This type of network was designed to view a problem in curve-fitting (approximation) and high dimensional space. The real inspiration behind the RBF technique is finding a multidimensional function that offers the best fit to the training tuples and then applies this multidimensional surface to interpolate the test data through regularization. In other words, the radial basis function breaks traditional interpolations into multidimensional space. In this category of neural networks, a set of functions that create a random basis for the input attributes compose the hidden units.

Basically, the radial basis function network comprises of three distinct layers. The input layer is the set of input pattern (vectors). The second layer is a hidden layer and the third layer is the output which is network’s response to the activation function applied to the input layer. The transformation process from the input layer to the hidden layer is nonlinear whereas it is linear from the hidden space to the output layer. Gaussian activation is a function which is used in the neurone structure of RBFN, meaning that normalising the input vectors is crucial here. In this case, every input that enters the neuron is assessed regarding whether it should be fired based on the distance between the weights and specific input in weight space. Indeed, these nodes are the hidden layer connecting up some output nodes in a second layer. If weights were added from every hidden (RBF) neuron to a set of output nodes, the RBFN would be built.

There are similarities between MLP and RBFN since these two supervised learning algorithms are good at providing a universal approximation; they can even be turned from one to another due to their neuron firing rules style which is based on distances and inner product for RBFN and MLP respectively. There are, however, some differences between them as the MLP applies the hidden nodes to divide the space by hyperplanes which are global, whereas the RBFN uses them to match functions locally. In contrast to the MLP, RBFN never has more than one layer of non-linear neurons and the given input to the perceptron is the nonlinear functions of the tuples. Interestingly, in an RBFN, for an input a number of the nodes will activate based on how close they are to the input and the combination of these activations will drive the network to decide how to respond properly. RBFN is significantly quicker than MLP because RBFN does not compute gradients for the hidden nodes. In the hidden layer of RBF, a nonlinear representation of the inputs can be found whereas the output layer intends to find a linear combination of those hidden nodes which are responsible to classification results. Hence, in the training process of RBFN, the RBF nodes should be positioned initially and afterwards the activation of these nodes to train the linear outputs should be applied [Marsland, 2009].

In the process of radial basis function development, setting the centers randomly to the training inputs is the simplest method but this approach is prone to overfitting. That is why clustering is applied to learn the training patterns leading into categories based on some similarity measurement and then assigning nodes to each cluster. Radial basis functions are slow to train and this is in contrast to preference of RBF over MLP because of its fast learning achieved by combining an unsupervised method with a supervised one [Mehrabi *et al*., 2009].

2.6.4 Statistical Learning Algorithm

2.6.4.1Logistic Regression

Logistic regression is an algorithm that constructs a separating hyperplane between two sets of data by using the logistic function to express distance from the hyperplane as a probability of dichotomous class membership:

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \varepsilon\ )}} \qquad (2.3)$$

In this equation, X_i symbolizes discrete or continuous predictor variables with numeric values; in the case of being dichotomous they are, for example, zero for a boy and one for a girl. The constants $\beta_0,\ \beta_1,\ \beta_2, \ldots, \beta_n$ are the regression coefficients estimated from training data which are typically computed by using an iterative maximum likelihood technique. Normally, this formula's justification is that the log of the odds, a number that goes from $-\infty$ to $+\infty$, is a linear function. Particularly by using this model, stepwise selection of the variables can be made and the related coefficients calculated. In producing the LR equation, the statistical significance of the variables used to be determined by the maximum-likelihood ratio. It has been highlighted that logistic regression provides an effective way of estimating probabilities from dichotomous variables.

2.6.4.2 Bayesian Analysis

Generally, Bayesian classifiers are statistical approaches capable of predicting class membership likelihoods like the probability of the training set belonging to a specific class. Bayesian classification is based on Bayes theorem and this classifier is known for its high accuracy and speed when applied to large data collection.

2.6.4.2.1Bayes Theorem

The name of Bayes theorem originates from Thomas Bayes who studied probability and decision theory during the 18th century [Han & Kamber, 2006]. Let X be a training dataset which is described by measurements made on a set of n attributions; H is the hypothesis, like the data tuple X belong to a specified class C. Bayes theorem produces a way to calculate the probability $P(H|\ X)$ from $P(H)$, $P(X|H)$, and $P(X)$ as shown in the following equation:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (2.4)$$

When $P\ (H|X)$ is the probability of H conditioned on X.

$P\ (X|H)$ is the probability of X conditioned on H.

$P\ (H)$ is the probability of H.

$P(X)$ is the probability of X.

2.6.4.2.2 Simple Bayesian Classifier

Let A be a training dataset including tuples and their associated class labels. Every tuple includes an n dimensional attribute vector $X = (\ x_1, x_2, \dots, x_n)$; thus n measurements made on the tuple from n attributes respectively D_1 , D_2 ,..., D_n . If we have m classes C_1,C_2, ..., C_m for tuple X, the classifier will predict that X belongs to the class having the most probability conditional on X. In other words, the Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$$P(c_i\,|x) > P\left(c_j\,|x\right) \quad \text{for} \quad 1 \leq j\ \leq m, j \neq i \qquad (2.5)$$

Hence, $P(C_i| \, x \,)$ should be maximised and the class C_i would be the maximum hypothesis according to Bayes theorem as in the following equation:

$$P\,(c_i|x\,) = \frac{P(x|c_i)P(C_i)}{P(x)} \qquad (2.6)$$

When $P(x)$ is a constant for every class, we have $P(x|c_i)P(C_i)$ to be maximised. The $P(x|c_i)$ calculation might be particularly complicated in computing and hence expensive. In order to reduce computation burden in the $P(x|c_i)$ evaluation, we presume that the values of the attributes are conditionally independent of each other based on the class label of the tuple. Thus,

$$P(x|c_i)= \prod_{k=1}^{n} P(x_k \; |c_i\,) = P(x_1|c_i)* P(x_2|c_i)*...* P(x_n|c_i) \qquad (2.7)$$

when x_k refers to the value of attribute A_k for tuple X. It is easy to estimate the probabilities $P(x_1|c_i), P(x_2|c_i), \ldots, P(x_n|c_i)$ by some simple calculation from the training tuples. Finally, to predict the class label of X, $P(x|c_i)$* $P(C_i)$ is calculated for every class C_i and the class label of tuple X is C_i if only if we have

$$P(x|C_i) * P(C_i) > P\left(x\middle|C_j\right) * P(C_j) \quad \mathit{for} \;\; 1 \le j \le m, j \neq i \qquad (2.8)$$

That is, the predicted class label is the class C_i when $P(X|C_i)$*$P(C_i)$ is maximum.

2.6.4.2.3Bayesian Networks

The Bayesian Network, known also as belief networks and probabilistic networks, provide a graphical model of casual and unconditional relationship from which learning can be performed through specifying joint conditional probability distributions and allowing class conditional independencies to be defined between subsets of attributes [Alpaydin, 2004]. There are two essential features that define a Bayesian network: a directed cyclic graph in which each node denotes a random variable, either discrete or continuous values, as well as a set of conditional probability tables.

Probabilistic dependence is shown by arcs in the graphs (figure 2.4); if an arc is drawn from a node y to a node z, for example, then y is a parent to the immediate predecessor of z and z is a descendent of y; every attribute is conditionally independent of its non-descendent in the graph given its parents. Hence, each variable has one conditional distribution $P\big(y\big|parents\ (y)\big)$ where parents (y) are parents of y. For instance, from the CPT represented in figure 2.4, we see that:

$$P(lung\ cancer = yes\ |family\ history = yes, smoker = yes\ ) = \ 0.8$$

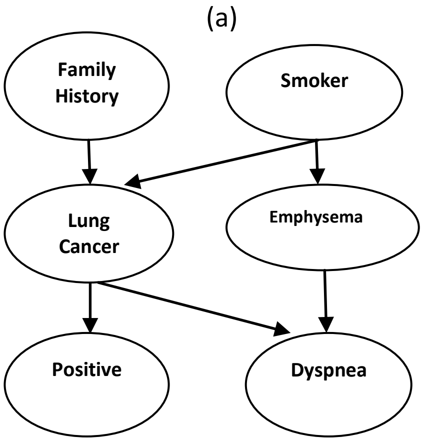
$$P(lung\ cancer = no\ |family\ history = no, smoker = no\ ) = \ 0.9$$

Let $x = (x_1, \dots, x_n)$ be a data tuple related to the correspondent attributes $y_1, y_2, \dots, y_n$. Given that every attribute is conditionally independent on its nondescendent and parents in the network graph, the following equation is a representation of joint probability distribution.

$$P(x_1, x_2, \dots, x_n) = \ \prod_{i=1}^{n} P(x_i\ \ |parents\ (y_i)) \qquad (2.9)$$

when the values for $P(x_i\ |parents\ (y_i))$ are related to the entries in CPT for y ; and $P(x_1, x_2, \dots, x_n\ )$ is the probability of a specific combination of values of X. Probability distribution, with the probability of each class, may be the output of this classification process.

The main weakness of BNs is their limitation to handle discrete variables, redundant and irrelevant attributes. They are not good at accuracy in general and suffer from the limitations of applying statistical methods in model building [Kotsiantis, 2007].



(b)

	FH, S	$FH, \sim S$	$\sim FH, S$	$\sim FH, \sim S$
LC	0.8	0.5	0.7	0.1
$\sim LC$	0.2	0.5	0.3	0.9

Figure 2.4 A simple Bayesian network: (a) is a casual model demonstrated by a directed acyclic graph. (b) the conditional probability table for the values of the variables Lung Cancer(LC) showing each possible combination of the values of its parent nodes, Family History (FH) and smoker (S), adapted from [Han& Kamber, 2006].

2.6.5 Support Vector Machines

Support vector machines (SVM) are a new classification method for both linear and nonlinear data. SVM applies nonlinear mapping to transform the original training tuple into a higher dimension. Inside this new dimension, it seeks the optimal linear separating hyperplane which is in fact the decision boundaries separating the tuples based on their class labels.

The SVM learning process takes a very long time and is known as an extremely slow method. However, it is highly accurate due to its capability of modelling complex nonlinear decision boundaries. Furthermore, it is much less prone to overfitting than other techniques [Han& Kamber, 2006].

2.6.5.1 Support Vector Machine Algorithm

Let the dataset *A* be considered as (x_1, y_1),(x_2, y_2) , …, $(x_{|A|}, y_{|D|})$, where (x_i) is the set of learning data with correspondent class label y_i. For a two-class related training dataset, for instance, every y_i can take either +1 or -1. This could also be generalized to n dimensions and the SVM duty is to find the best dividing lines that can be drawn and divide all of the tuples of every class from the others. For multidimensional classes the hyperplanes should be found as decision boundaries. This can be arranged by defining the maximum marginal hyperplane (MMH) since the hyperplanes with the larger hyperplane are more accurate at classification. Figure 2.5 depicts the role of MMH and hyperplanes to determine class and decision boundaries. This separating hyperplane can be defined as:

$$w.x + b = 0 \qquad (2.10)$$

Where w is a weight vector like $w = \{w_1, w_2, \dots, w_n\}$; n is the attribute number, b is a scalar and termed bias. For two input attributes like x_1 and x_2 the above separating hyperplane can be rewritten when we replace w_0 for b:

$$w_0 + w_1x_1 + w_2x_2 = 0 \qquad (2.11)$$

Hence, any point that is positioned above the dividing hyperplane satisfies

$$w_0 + w_1x_1 + w_2x_2 > 0 \qquad (2.12)$$

And any point that would be under the dividing hyperplane should satisfy

$$w_0 + w_1x_1 + w_2x_2 < 0 \qquad (2.13)$$

Thus, the weights can be adjusted in order to define the margin's side by hyperplane and for a two-class data tuple:

$$H_1 : w_0 + w_1x_1 + w_2x_2 \geq 1 \qquad for\ y_i = +1 \qquad (2.14)$$

$$H_2 : w_0 + w_1x_1 + w_2x_2 \leq \ -1 \quad for\ y_i = -1 \qquad (2.15)$$

It means that any tuple that is placed on or above H_1 falls into class +1 and any tuple that is on or below H_2 belongs to class -1. Combining the two above equations we have:

$$y_i(w_0 + w_1x_1 + w_2x_2 \geq 1 \qquad ,\forall_i \qquad (2.16)$$

Any training tuples that fall into H_1 or H_2 satisfy equation 2.16 and are named support vectors which are equally close to the MMH. After training the support vector machine, we need to use this tool to classify test tuples using the following equation:

$$d(X^T) = \ \sum_{i=1}^{l} y_i \ \alpha_i X_i X^T + b_0 \qquad (2.17)$$

Where, y_i is the class label of support vector x_i, x^T is a test tuple. α_i and b_0 are numeric parameters defined by the optimization or SVM algorithm automatically, and l is the number of support vectors.

In SVM, the number of support vectors defines the complexity of the learned classifiers compared with the dimensionality of the data. Thus, using an SVM gives less probability of overfitting than some other techniques. An SVM with a small number of support vectors can have good generalization even in the case of multidimensionality of the available data.

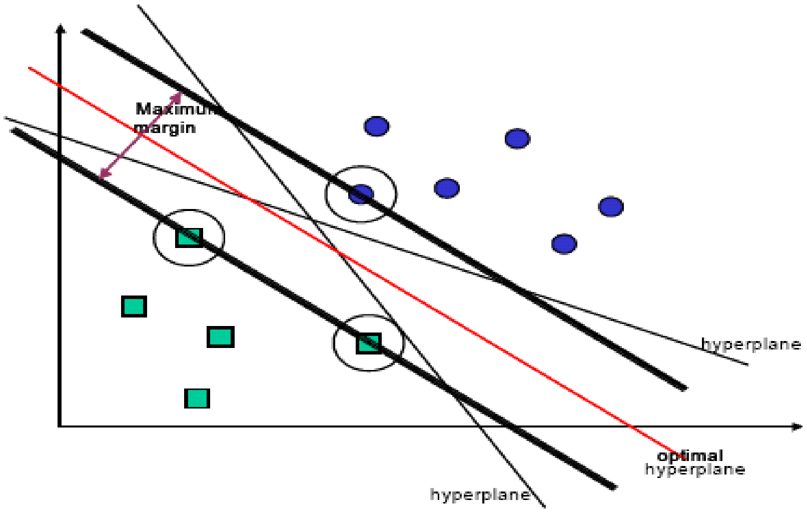


Figure 2.5 Maximum Margin in a support vector machine structure, adapted from Kotsiantis (2007).

2.7 Unsupervised Learning

In unsupervised learning, there is a set of training data tuples with no collection of labelled target data available. The aim of unsupervised learning is discovering clusters of close inputs in the data where the algorithm has to find similar data and the aim isn't confirming that certain data points belong to one class and others to a different class. In unsupervised learning all variables are treated the same way without the difference between dependent and independent attributions.

2.7.1 Cluster Analysis

The most important technique of unsupervised learning is cluster analysis which takes ungrouped data and applies automatic techniques in order to put them into groups. Cluster analysis, also called segmentation analysis or taxonomy analysis, creates groups or clusters of data. Clusters are formed in such a way that objects in the same clusters are very similar and objects in different clusters are very distinct.

Essentially, clustering is a type of learning by observation rather than examples. There are two main approaches for clustering analysis: hierarchical and nonhierarchical clustering. Hierarchical clustering, which is the most commonly used technique, organizes data in a nested sequence of groups that can be displayed in a tree-like structure. Another technique is K-mean clustering; it is a partitioning method, its function partitions data into k mutually exclusive clusters and returns the index of the cluster to which it has assigned each observation. Unlike hierarchical clustering, k-mean clustering operates on real observations rather than the average set of dissimilarity measures. It generates a single level of clusters. However, hierarchical clustering groups data over a variety of scales by creating a cluster tree. The tree is not a single set of clusters but rather a multilevel hierarchy, where clusters at one level are joined to clusters at the next level. This allows the user to decide the level or scale of clustering that is most appropriate for the given application. More details about clustering approaches are described in the next chapters.

2.8 Model Evaluation

A comparison framework is a critical step in deciding which specific learning algorithm should be chosen for the given prediction and classification task. Once the initial testing is judged to be satisfactory, the classifier with the duty of mapping from unlabeled instances to predict classes would be accessible for regular use.

Generally, there are three estimation methodologies for classification models. These include: *k*-Fold Cross Validation, Bootstrapping & Jackknifing and Simple Split (Holdout) [Olson & Delen, 2008].

In the cross validation method, subsets of the data collection is put aside for validating purposes and the remaining data are used as a training set to develop the predictive model. Then, the model is applied for prediction using the validation set. Through using the developed model the validation set is predicted. This could be regarded as a measurement of prediction accuracy. In the *k*-Fold cross validation approach, the dataset is divided into *K* subsets where each is held out in turns as the validation set.

In bootstrapping & jackknifing properties of an estimator (such as its variance) might be measured through sampling from an approximating distribution; the approximating distribution can be either empirical or an independent and identically distributed population. In the case of independent distribution, a number of resamples of the observed dataset with the same size are created. They are constructed by random sampling with replacement from the original dataset.

The simple Split (Holdout) method divides the dataset into three: two-thirds for training and the other third for testing randomly. This split estimation methodology partitions the data into two mutually exclusive subsets where the training set is used by the classifier and the built model is then tested on the holdout set.

Performance metrics for predictive modelling evaluate the inducer accuracy through a number of approaches such as prediction accuracy by coincidence matrix (classification matrix or contingency table).

2.8.1 Prediction Accuracy, Recall, Precision & F-measure

Sensitivity, specificity and F-measure are also utilized for assessing as well as calculating other aggregated performance measures like area under the ROC curve. The coincidence matrix structures the basis of these common matrixes so that for any classifier, four prediction outcomes are normally possible. If the instance is positive and it is also classified as positive, it is calculated as a true positive; however, if it is classified as negative, it is termed as a false negative. On the other hand, if the instance is negative and it is classified as negative, it is labelled as a true negative; if it is classified as positive, it is counted as a false positive. Given a classifier and a set of instances (the test set), a two- by- two coincidence matrix can be generated indicating the dispositions of the set of instances [Olson & Delen, 2008].

The prediction accuracy is the most precise measurement of classifier evaluation. Basically, it is the percentage of correct prediction (true positive + true negative) divided by the total number of predictions. Other parameters like sensitivity $P_r(+ \mid P)$, specificity $P_r(- \mid N)$, positive predicted value $P_r(P \mid +)$, and negative predicted value $P_r(N \mid -)$ can simply be calculated through the presented numbers in a contingency table and following formula. However, in machine learning, sensitivity is simply termed recall (r) and the positive predicted value is called precision. F-measure, another algorithm evaluator, is the harmonic means of precision and recall and the higher its value reveals better predictive performance. Figure 2.6 is the easy way to understand true positive (TP), true negative (TN), false positive (FP), false negative (FN) and consequently sensitivity and specificity. Furthermore, equation 2.18, 2.19, 2.20, and 2.21 present how these scales are calculated.

$$Prediction\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.18)$$

$$Precsion = \frac{TP}{TP+FP} \qquad (2.19)$$

$$Recall = \frac{TP}{TP+FN} \qquad (2.20)$$

$$F - measure = \frac{2}{\frac{1}{precision}+\frac{1}{Recall}} \qquad (2.21)$$

		True Class	
		Positive (P)	Negative (N)
Predicted Class	Positive (+)	True Positive Count (TP)	False Positive Count (FP)
	Negative (-)	False Negative Count (FN)	True Negative Count (TN)

Figure 2.6 A simple coincidence matrix demonstrating the TP, FP, FN, and TN which are applicable to calculate sensitivity, specificity, prediction accuracy, recall, and F-measure and drawing area under ROC curve, adapted from [Olson & Delen, 2008].

2.8.2 ROC Curve, Area under Curve (AUC)

Moreover, another performance evaluation technique for classification called the ROC curve (Receiver Operating Curve) has been applied for visualizing, organizing and selecting classifiers based on their performance. They are two-dimensional graphs in which the true positive (*TP*) rate is plotted on the *Y* axis and the false positive (*FP*) rate is plotted on the *X* axis. Basically, a ROC curve, which is a depiction of a classifier's performance, is capable of both comparing classifiers and judging the fitness of a single classifier simply by reducing the ROC measures to a single scalar value representing the expected performance.

The area under the ROC curve, abbreviated as AUC, is a portion of the area of the unit square. This value varies between 0 and 1.0. Perfect accuracy has a value of 1.0 and a classifier with an AUC of 0.5 would be a poor classifier. To calculate AUC two common methods are applied; one is on the basis of the trapezoid's construction under the curve as an approximation of the area, and the other method utilizes a maximum likelihood estimator to fit a smooth curve to the data points. Both methods can estimate area and standard error to compare different classifiers.

2.9 Influential Factors for Tuberculosis Treatment Course Non-compliance

Several factors play significant roles in adherence to anti-TB chemotherapy including the patient receiving the drugs, the doctor prescribing the drugs, the nurses and other healthcare staff `s supervision, the patient's follow-up progress and, finally, the selection of appropriate chemotherapy regimens and the organizing drug supplies programme. In spite of various interventions aimed at improving treatment completion, patients' independent adherence to drug treatment is still a vital determinant in terms of undesired consequences of noncompliance to treatment courses, like relapse and drug resistance occurrence. According to the study [Sbarbaro& Sbarbaro, 1994] methods like traditional health education, the consideration of cultural influences, behaviour analysis and emphasis on preventing physicians from misleading patients enhance the behaviour of TB patients, encouraging them to complete their treatment courses and get cured. However, it has been documented that the patient should not be solely charged with the task of adherence; adherence is the outcome of a process which has a long chain of responsibilities. Some studies put the duty of making sure that the patient has received the full course of treatment on the physician`s shoulders because mutual cooperation is required between the physician and the public health department [Sbarbaro & Sbarbaro, 1994]. This necessity of close mutual aids among various agents including patients and physicians has raised awareness of adherence as a complex behavioural issue influenced by many factors. Because of being multifaceted and complicated, it is still difficult to identify non-adhering patients. Recently, researchers have focused on investigating patients` features to find effective solutions for the lack of a comprehensive and holistic understanding of barriers to and facilitators of treatment adherence [Thiam *et al*., 2007]. The main aim of this research is to spot cases which are at a high risk of non-compliance. In this section more investigations are reviewed to highlight the significant risk factors for patient non-adherence to the treatment course. Also, the correlation between applied patient features in this study and the outcome of the tuberculosis treatment course will be investigated in the following chapters. Those significant risk factors create a golden opportunity to develop a predictive model which can forecast the prospect of applying chemotherapeutic regimens for each patient and define the level of supervision and support required, as the WHO has emphasised in the "Stop TB" plan.

Munro *et al.* (2007) in a systematic review study of qualitative research categorized the influential risk factors which effect adherence to tuberculosis treatment. They listed those factors as follows:

- Structural Factors such as poverty, gender, and discrimination
- Patient Factors like motivation, knowledge, beliefs, attitudes and interpretations of illness and wellness
- Social Context
- Health Care Service Factors

Inside these main categories, there are several factors which have been addressed by a number of studies. A retrospective review [Menzies *et al*., 1993] investigated the factors associated with compliance in treatment of tuberculosis by using medical and nursing records of TB patients treated between 1987 and 1988. The general therapy completion rate was 59% among identified TB patients. Compliance with preventive therapy was highest whereas the rate was lowest for those cases identified through a workforce screening survey (*P*<0.01). The length of the treatment course was shorter for those patients who were initially hospitalised, or had a better understanding of the treatment course. Their therapy lasted 6-9 months rather than 12 months. This was also the case for patients who returned for follow-up sessions during 4 weeks of onset of therapy (*P*<0.01). In summary, compliance can be enhanced by improving patient knowledge, providing closer follow-up and reducing the length of therapy, especially for those at lower risk of re-activation. Additional compliance enhancing interventions can be targeted to those patients with suboptimal compliance who can be accurately identified early in the course of therapy.

Another study [Burman *et al*., 1997] showed that patients who had not completed their treatment course were more likely to have a poor outcome (32% vs. 3.3%). Furthermore two risk factors, alcohol abuse (*OR= 3.0, CI=1.2 to 7.5, P = 0.02)* and homelessness (*OR= 3.2, CI=2.5 to 5.8, P = 0.004*), were introduced.

Tangüis *et al*. (2000) conducted an epidemiological study to determine the predictive factors of the non-completion of tuberculosis (TB) treatment among 2,201 HIV-infected patients where 76.2% of them were intravenous drug users (IDU). This study was conducted from 1987 to 1996 in Barcelona by calculating the x^2 test for bivariate analysis and developing a logistic regression model which revealed living in neighbourhoods of a low socio-economic level (*OR=1.61, CI=1.22 to 2.13*), homelessness (*OR= 3.56, CI= 2.01 to 6.31*), and history of TB (*OR =1.61, CI= 1.12 to 2.33*) as risk factors for quitting TB treatment. This study concluded that social and health factors influence non-completion of TB treatment in HIV-infected patients and patient support and supervision can improve treatment completion.

In 2006, a prospective survey [Shieh *et al*., 2006] aimed to predict Non-Completion of treatment for latent tuberculosis infection. The influence of factors like certain health attitudes, lifestyle, clinical and regimen association barriers were assessed by using x^2 and logistic regression analysis. Two independent predictors of non-completion were found including low risk awareness of developing active TB in the absence of LTBI treatment (*OR=0.31, CI=0.10 to 0.50, P ≤ 0.007*), and not wanting venipuncture (*OR=0.43, CI=0.22 to 0.69, P ≤ 0.015*). These two predictors accounted for 75% of Cases of non-completion in total. In conclusion, this study suggests that the predictors of latent tuberculosis treatment non-completion are identifiable from the first visit. Individuals, who are at a high risk of contracting TB, can be targeted to minimize difficulties, they could be educated about TB and diagnostic tests with improved specificity could be used on them.

A longitudinal non-concurrent cohort study [Vieira & Ribeiro, 2008] to determine treatment noncompliance rates among patients participating in a national tuberculosis control programme in the city of Carapicuíba in Brazil and to recognize the variables related to noncompliance was carried out in 2008. There were two patient groups; the first cohort contained 173 patients with tuberculosis who were self administrating treatment and the second consisted of 187 patients with tuberculosis who were being treated through DOTS. Besides the results verifying a significant reduction in the noncompliance rate from 13.3% to 5.9% because of applying DOTS ($P < 0.05$), four following variables were found to be associated with treatment noncompliance in the case of self administration therapy: being an unregistered worker (relative risk (*RR*) = 3.06); retreatment (*RR* = 2.73), alcoholism (*RR*= 3.10), and no investigation of contacts (*RR* = 8.94). This research emphasised the ability of DOTS to decrease the rate of non-compliance and produce better treatment outcomes.

Another piece of research [Machado *et al*., 2009] from Brazil, investigated the risk factors involved with the failure to complete a course of latent tuberculosis infection treatment. To meet this end, household contacts of patients hospitalized with pulmonary TB for 6 months after they initiated treatment with isoniazid (INH) were followed. 53.5% of household contacts completed the 6-month regimen and the risk of treatment non-compliance was significantly higher in household contacts with side effects to isoniazid treatment (*RR=2.69, CI=1.80 to 3.2, P= 0.01*). Furthermore, patients who had difficulty travelling to hospital due to transportation issues were at a higher risk of not completing their treatment (*RR =1.8, CI=1.5 to 1.9, P =0.04*). To sum up, completion of latent tuberculosis infection treatment was most affected by medication intolerance and commuting difficulties for follow-up visits.

In 2009, a retrospective epidemiological study [Rakotonirina *et al*., 2009] conducted analysing data relating to 442 tuberculosis cases in Antananarivo to recognize the risk factors of treatment default. The result of this study revealed that gender is a risk factor since males give up treatment more than women (*OR=1.81 , CI=1.13 to 3.03*); also, patients younger than 30 years old were more likely to quit the follow-up of their treatment (*OR=3.43 CI=1.16 to 10.15*). Hence, health workers and physicians should be more aware and alert regarding young male TB patients presenting these characteristics and should adjust the methods and means for follow-up according to these risk factors.

Finally, Yew (1999) has focused on patient characteristics which are more commonly associated with non-adherence like homelessness, alcohol or substance abuse, behavioural problems, mental retardation, and lack of social or family support. This study also highlighted the fact that identifying non-adherent patients is still a dynamic phenomenon and complex issue affected by a range of factors from patients' demographic features to qualities of the social and economic environment.

2.10 Predictive System in the Medical Domain

2.10.1 The Application of Classification Algorithms in Medical Areas

The increasing number of electronic data collections containing medical data has resulted in growing concentration on their utilization for generating valuable knowledge discoveries through machine learning techniques. Building classification models via learning from examples is one of the most common ways to find new knowledge. This is done by taking each instance and assigning it to a particular class label and then predicting categorical class labels including discrete and nominal.

In this section, a brief summary of applying six machine learning classifiers in medical areas – neural networks, decision trees, Bayesian networks, radial basis function, logistic regression, and support vector machines – has been given.

2.10.1.1Classification Decision Tree to Predict Medical Outcomes

In medical decision making where decisions must be made effectively and reliably in many situations such as the classification and diagnosis of disease, the decision tree (DT) has been introduced as a common tool. It provides high classification accuracy with a simple representation of produced medical knowledge. Some examples of the many DT applications in medicine are detailed below.

In the area of emergency medicine, a study [Tsien *et al*., 1998] was conducted to diagnose at an early stage myocardial infarction (MI) in the patients complaining of chest pain in the emergency room (ER). Because of the potential ability of decision trees to create simple but accurate decision aids, this technique was examined. The classification decision tree was developed to predict the probability that a patient with chest pain has an MI based solely upon the data available at the time of presentation to the ER. In this study, training and testing sets which came from a set of data collected in Edinburgh, Scotland were applied for model building and validity checking respectively. The results obtained showed that the decision tree performed well with ROC curve areas of 94.04%. This study highlights the classification trees` advantages and it shows accuracy as logistic regression does in this case.

In the field of maternal health and obstetrics, decision trees were used to predict cesarean delivery in a historical cohort study [Sims *et al*., 2000] of 24,661 records of women who had delivered live-born singleton neonates from 1995 to 1997 in Pittsburgh Hospital with 78 variables. Using 16 variables significantly related to the outcome, six different methods of decision trees were developed. The decision tree rule-based method was applied to the 50% of the sample to develop the predictive training model`s on the basis of the risk factors found for cesarean delivery and the remaining 50% were used to test the model`s accuracy. Using the C4.5 decision tree, the area under the curve for nulliparous and parous women was 0.82 and 0.93 respectively. It has been concluded that decision tree models can be used to predict cesarean delivery based on promising model accuracy and being small enough to be intelligible to physicians; it is because of their ability to disclose causal dependencies among variables, handle missing values easily, and predict the given outcome (cesarean deliveries) despite the absence of categorized risk factor variables.

In 2002, Lazarescu and colleagues [Lazarescu *et al*., 2002] investigated the machine learning application to classify glaucomatous progression into one of two possible classes including stable and progressive glaucoma. In this work in addition to introducing new influential factors for the considered outcome, stable glaucoma patients and progressive glaucoma patients were distinguished at the earliest possible stage of the disease. This allows ophthalmologists to decide whether or not the alternate treatments should be followed in order to preserve as much of the patient's sight as possible. In this work, 12 predictors of glaucomatous progression were found and related instances were split into training and testing sets. In order to classify the available data, the C4.5 software was used and after 50 times, each time with applying different random training and testing sets, a C4.5 decision tree with an average of 15 nodes using 7 features consistently was generated. To ensure the accuracy of the created decision tree, precision and recall were calculated and were 95% and 82.5% respectively.

Then decision tree application was reported [Pavlopoulos *et al*., 2004] to differentiate the diagnosis of Aortic Stenosis (AS) from Mitral Regurgitation (MR) via heart sounds. As for the background of this study, it has been explained that since the new technologies are costly, large in size and complex in operation; they are not appropriate for use in rural areas, homecare and primary healthcare set-ups. As well, the majority of internal medicine and cardiology training programmes underestimate the value of cardiac auscultation, so junior clinicians are not effectively trained in this field. Therefore efficient decision support systems would be very useful for supporting clinicians to make better heart sound diagnosis. Thus, this study pursues a rule-based method development based on decision trees. To meet this end, a collection of 84 heart sound signals was analysed, containing 42 heart sound signals with "clear" AS systolic murmur and 43 with "clear" MR systolic murmur. After pre-processing the first and second heart sounds, a total of 100 features were defined for each heart sound. Afterwards, a decision tree classifier with a training set of 34 records and a test set of 50 records was developed. The results of validity tests showed 90% classification accuracy when 45 correct/50 total records. The Classification accuracy of the decision trees, both in terms of partial classification and overall classification did not significantly change after pruning the decision tree. This work demonstrated that the decision tree algorithm can be effectively utilized as a basis for a decision support system to help young and inexperienced clinicians to make better heart sound diagnosis in health centres which are not-highly equipped.

2.10.1.2 Applying the Bayesian Network to Predict Medical Outcomes

Bayesian Networks (BN) have been used to acquire expert knowledge in medical areas; they handle uncertainty and tackle missing data via inference techniques. Disease diagnosis, treatment selection, and prognosis prediction can be carried out using BNs. As an example, BN has been used [Reiz & Csato, 2009] to develop a system to predict the bypass surgical survival probability. This involved undertaking 66 medical examinations of 313 patients. A tree-like Bayesian Network is the optimal tool for classifying logical data where the most relevant cause corresponding to the survival chance should be selected. Results showed that BNs model with 75.71% prediction accuracy, performing even better than logistic regression with 63.50%.

In terms of supporting medical decisions, online detection of Premature Ventricular Contraction beats (PVC) in electrocardiogram (ECG) records is another instance of BN application [De Oliveira *et al*., 2008]. PVC is a well known cardiac arrhythmia which can be analysed in standard ECG databases. BN with their ability to tackle uncertainty were applied to this task with its random character and random variables. After examining some topologies of static Bayesian networks, the one which was most suitable for the task was selected. The results produced verify that the combination of different ECG channels improves the performance of BN as classifiers and expresses the feasibility of using Bayesian networks as a tool to classify this type of signal as well.

2.10.1.3 Applying the Logistic Regression to Predict Medical Outcomes

Multivariable logistic regression (LR) as a method of statistical analyses is a widely utilized multivariable technique for modelling dichotomous outcomes. In fact, this technique serves two purposes: firstly, predicting the outcome variable for new values of the predictor variables, and secondly, answering questions about the area under study as the coefficient of each independent variable clearly illustrates the relative contribution of that predictor to the dependent variable [Bagley *et al*., 2001]. The following highlights some of those medical applications.

Through a retrospective cohort study [Dodek & Wiggs, 1998] a logistic regression model was developed to predict the outcome of post-hospital cardiac arrest along with assessing the validation, accuracy, sensitivity and specificity in numerous cut-off points. In this, available data were all from in-hospital cardiac arrests. The logistic regression model was created to estimate the probability of death before hospital discharge as a function of predictor variables including: patient and arrest descriptors, major underlying diagnosis, initial cardiac rhythm and time of year. Separate data collection was applied in order to model validation gathered from the same hospital in 1989-90. optimal sensitivity and specificity for testing set have been 0.75 at a cut-off probability of 0.75. By using validation dataset, optimal sensitivity and specificity are 0.6 at a cut-off probability of 0.85.

2.10.1.4 Applying Neural Networks to Predict Medical Outcomes

Acute coronary syndrome (ACS) is a heterogeneous condition that varies from severe, for which immediate medical treatment is necessary, to minor for which patients are advised to rest at home. Research [Green *et al*., 2006] has been carried out to diagnose ACS patients using artificial neural networks (ANN). These were trained on data from

634 patients with symptoms of chest pain. Only data immediately available at patient presentation were examined, such as electrocardiogram (ECG) data. Overall 18 variables consisting of 4 continuous and 14 discrete were available. Feed-forward multilayer perceptrons (MLP) with one hidden layer and no direct input-output connections were considered. The ANN ensemble approach together with ECG data after pre-processing by principal component analysis led to the result of an area under the ROC curve of 80%, sensitivity of 95% and specificity of 41%.

Another study [Garcla-Perez *et al.*, 1998] investigated the application of neural networks for differential diagnosis of Alzheimer`s Disease (AD) and Vascular Dementia (VD). There are great difficulties in diagnosing different types of dementia and Alzheimer’s and vascular dementia are often mixed up as their symptoms are very similar. This study applied neural network technology to assist neurologists for differential diagnoses of AD and VD. A three layer feed-forward neural net which was trained over 65 hours in order to reach the minimum average error of 0.0000002 with the back propagation learning algorithm. The number of neurons in the input layer was 46 based on the number of characteristics for each subject while the output layer was defined by one neuron, as the differential diagnosis might only be AD or VD. To check the model validity through using testing set, 82.6% correctly classified rate was calculated.

In 1999, an investigation [El-Solh *et al*., 1999] used NN technology on tuberculosis patients’ information to predict active pulmonary TB. It has been pointed out that identifying those with contagious active TB, isolating them over the contagious period and treating them effectively are the crucial aspects of the TB infection control programme. These current approaches to TB control have faced several difficulties. Thus, a prediction model to identify patients with active TB can play a substantial role.

It can be developed through using the clinical and radiographic information of TB patients besides modelling techniques to enhance the physician’s prediction ability. For this non-concurring prospective study, 21 distinct clinical and radiographic parameters of 563 TB patients were used as a training set to develop the predictive model and 119 cases were employed to check the model`s validity. The three–layer neural network structure consisted of an input layer, a hidden layer and an output layer where the input patterns comprised three groups: demographic variables, constitutional symptoms and radiographic attribution. The model significantly outperformed the physicians’ prediction, with calculated c-indices (6 SEM) of 0.947 ∓ 0.028 and 0.61 ∓ 0.045, respectively ($p < 0.001$). When it was applied to the validation group, the corresponding c-indices were 0.923 ∓ 0.056 and 0.716 ∓ 0.095 respectively. The obtained findings are interesting as the artificial neural network can recognize patients with active pulmonary TB more accurately than physicians’ clinical assessment.

In research [Silva *et al*., 2008] aimed at linking organ failure to adverse events, a neural network was able to develop a promising model which recognized early identification of organ impairment as a key issue. The importance of the final result of this research relates to the fact that sequential organ failure assessment (SOFA) is an expert-driven score that is widely used in European ICUs to quantify organ disorder and constitutes a complementary data-driven approach on the basis of adverse events defined from commonly monitored biometrics. This study investigated the impact of these events when predicting the risk of ICU organ failure. A large collection of data comprising 25,215 daily records taken from 4,425 patients and 42 European ICUs was used to form input variables including mixed cases (*i.e*. age, diagnosis, admission type and admission from) and adverse events were defined from four bedside physiologic variables. The output was the organ status (normal, dysfunction or failure) of six organ systems measured by the SOFA score. To develop the multilayer perceptron model, processing neurons were grouped into layers and connected by weighted links. The network was activated by feeding the input layer with the input variables and then propagating the activations in a feed-forward method through the weighted connections, over the entire network. With one hidden layer, the network output is the probability estimate of the rate of organ failure. To check the validity by a 5-fold cross-validation scheme, the area under the ROC curve (over all organs) was 64%, 69% and 74% for the dysfunction, normal and failure organ conditions, respectively. The ROC curve area for predicting renal failure was 76%. To sum up, adverse events obtained from bedside monitored data are important intermediate outcomes, contributing to a timely identification of organ dysfunction and failure during ICU length of stay. The results obtained demonstrated that it is possible to exploit neural networks to gain knowledge from easily obtainable data.

The following table (2.3) presents a small number of common neural network applications in various medical areas. All the mentioned studies have been carried out for risk factor analysis, diagnostic and predictive problems.

Table 2.3 Details of multi-layer perceptron application for different classification tasks in various medical areas

Application	Learning Algorithm	Network Structure	Accuracy %	References
Skin disease diagnosis	Back-propagation	Number of attributes:34 Input layer: 6 neurons Hidden layer:1, 3 neurons Output layer: 6 neurons	90	Chang & Chen, 2009
Predict the presence of coronary artery disease	Standard back-propagation	Subjects 1,245, variables: 8 Input layer: 8 neurons Hidden layer:1, Output layer: 2 neurons	79.3	Kurt *et al.*, 2008
Differentiate congestive heart failure and chronic obstructive pulmonary disease	Back-propagation	Subjects : 266 Input layer: 7 neurons Hidden layer:5, Output layer: 2 neurons	83.9	Mehrabi *et al.*, 2009
Risk factor analysis for Salmonella Typhimurium infections	Back-propagation	Variables: 18 Hidden neuron:56 neurons	89.67	Yang *et al.*, 2006

2.10.1.5 Applying the Radial Basis Function to Predict Medical Outcomes

There is not much research about the application of radial basis functions in medical areas. One of the rare reports [Mehrabi *et al*., 2009] is about the application of radial basis function neural networks alongside multilayer perceptron in differentiating between chronic obstructive pulmonary and congestive heart failure diseases since these two conditions show similar symptoms. Data from 266 patients with 42 clinical variables was used in this investigation.

The results showed that the radial basis function network with 6 neurons in the hidden layer and a threshold of 0.5174 performed well with sensitivity of 81.8%, specificity of 88.4% and AUC of 0.924 ± 0.017. However, the neural network, with 5 neurons in the hidden layer and a threshold of 0.5739 was not so different with sensitivity of 83.9%, specificity of 86% and an area under the receiver operating characteristic curve (AUC) of 0.889 ± 0.02. Figure 2.7 illustrates that the radial basis function performed as well, or even better, than neural networks at differentiating congestive heart failure and chronic obstructive pulmonary disease.

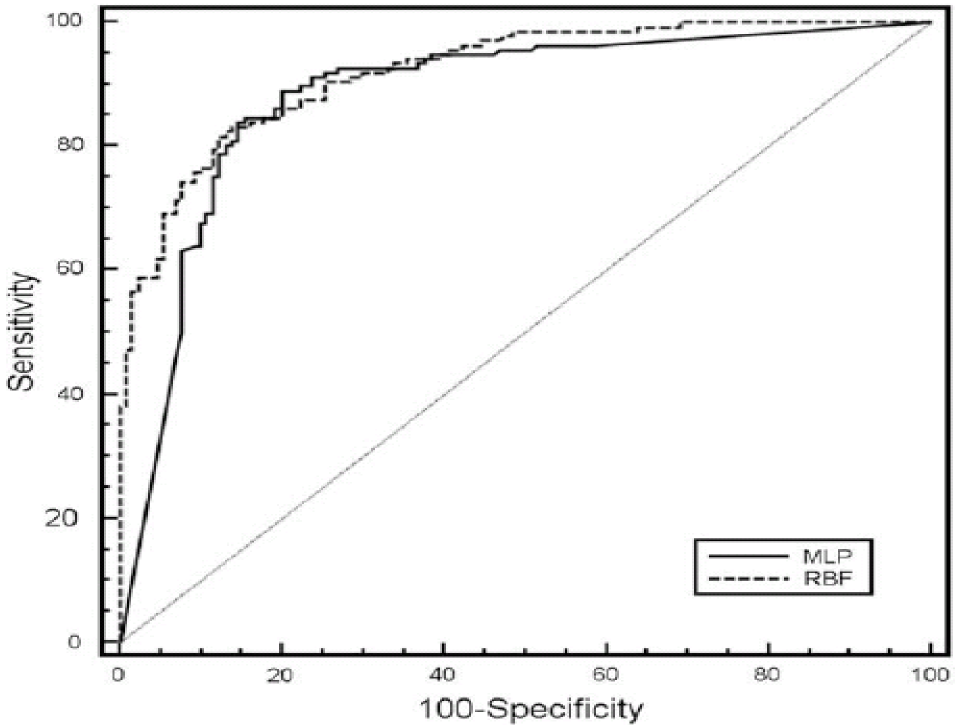


Figure 2.7 ROC curve comparison of MLP & RBF, adapted from Mehrabi *et al*. (2009).

2.10.1.6 Applying the Support Vector Machine to Predict Medical Outcomes

The support vector machine (SVM) is established as a high performance algorithm for solving classification tasks in many fields such as in biomedical and medical areas. This has been verified by several studies, including the two discussed below.

In the first example [Yu *et al.,* 2010], the potential power of SVM as an approach for classifying individuals into groups defined by disease status has been tested for detecting persons with diabetes and pre-diabetes.

Data from individuals related to six years of cross-sectional representative samples which were taken from 1999 to 2004 and were analysed to develop and validate SVM models in two classification plans. Firstly, diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes and, the secondly, undiagnosed diabetes or pre-diabetes vs. no diabetes. The SVM models choose sets of attributes that would produce the best classification of individuals into the aforementioned diabetes groups. Different patients` features were applied to two different schemes; these included family history, age, race and ethnicity, weight, height, body mass index (BMI), and hypertension for the first one and two further variables, sex and physical activity, were incorporated for the second classification scheme. The area under the curve, as an accuracy measurement, showed 83.5% and 73.2% for the first and second classification schemes respectively. This study has emphasised the SVM`s capability of yielding promising classification results using patients` characteristics in the medical domain, particularly for the detection of diabetic or pre-diabetic patients in certain populations.

In another investigation [Jiang *et al*., 2007], the SVM was proven to be one of the most accurate classifiers with a good capability of fault-tolerance and generalization. This approach was applied to classify digital mammography, analysing 322 images of patients in three big categories: normal, benign and malign. The experimental results revealed that The SVM classifier performed by 92.94% classification accuracy.

In summary, machine learning tools including DT, BN, LR, MLP, RBF, and SVM have played a crucial role in solving prediction tasks in the medical domain. Their application has yielded comprehensible, accurate and quick systems since every single one of them has specific features and outstanding capability. The similarity between tasks also may direct us to choose the appropriate technique to solve the problem of predicting the outcome of a course of tuberculosis treatment.

2.11 Summary

To sum up, not obtaining the desirable outcome from a course of tuberculosis treatment, which is either treatment course completion or getting cured, may lead to serious difficulties like multidrug-resistance TB which is increasingly common in several highly populated countries around the globe. DOTS needs to be actively applied in practice by intensive supervision and support which has not always been available due to the cost involved. A decision support system able to predict the course of tuberculosis treatment for each TB case at the onset of the therapy might be capable of defining the possible outcome. This can then be exploited to define the level of active support and supervision required for every TB case based on specific personal features. Some of those features have been identified from literature and some others will be revealed by feature analysis methods discussed in the next chapter as part of this study. There are a number of classification algorithms that have performed very well in terms of prediction accuracy in numerous medical topics; in this study the performance of six of them has been reviewed and the details of their application results will be presented in the following chapters.

Chapter 3

Research Methodology & Design

3.1 Introduction

This chapter discusses each stage of the research methodology, which includes data analysis, feature selection, supervised and unsupervised learning and their combination. It goes on to consider the methodology of decision tree simplification and accuracy improvement through proposing the cluster-based simplified decision tree (CSDT). Finally, an overview of the methodology is given and the chapter is summarized.

3.2 Statistical Analysis

Generally, at the start of developing predictive model, data analysis is carried out. One of the best techniques to obtain comprehensive understanding about a dataset in both quantitative and qualitative ways is statistical analysis. A fundamental task in statistical analysis is to estimate location measurements and normality parameters for a dataset. Statistical analysis by calculating descriptive measurements such as mean, median, standard deviation and normal distribution values like skewness and kurtosis provides insight into the underlying structure of a dataset and the resulting model.

The usual estimate of location is mean and median, whilst the common measure of variation is standard deviation. The mean is the sum of the data points divided by the number of data points. The median is the value of the point which has half the data smaller than that point and half the data larger than that point. Standard deviation is calculated through following formula:

$$Sd = \frac{1}{\sqrt{(N-1)}} \sqrt{\sum_{i=1}^{N} (Y_{i} - \bar{Y})^2} \qquad (3.1)$$

from N measurements $Y_1, Y_2, \dots, Y_N$.

Mode is another descriptive statistical measurement; it is the value that occurs most commonly in a data collection. More than one mode may be available in the case of having more than one value which appears the most.

In order to find out about how normal is the dataset, Skewness and Kurtosis and Kolmogorov-Smirnov are computed.

Skewness and Kurtosis are two measurements that should be zero in a normal distribution. Positive values of skewness demonstrate a pile-up of scores on the left of the distribution; however, negative values signify a pile-up on the right side. For Kurtosis, positive values mean a pointy distribution whereas negative values specify a flat distribution. The further the value is from zero, the higher the probability that the data is not normally distributed.

A z-score is a score from a distribution that has a mean of 0 and a standard deviation of 1 which are new comparative values. To transform any score to a z-score easily subtract the mean of the distribution and then divide by the standard deviation of the distribution as shown in the following formula:

$$^{Z}Skewness = S - 0 / Std.error \ \ Skewness \qquad (3.2)$$

$$^{Z}Kurtosis = K - 0 / std.error \ \ Kurtosis \qquad (3.3)$$

Where *S* and *K* are the values of skewness and Kurtosis and their related standard errors. Moreover, there are two tests which determine whether the distribution of the given dataset is normal: the Shapiro-Wilk and the Kolmogorov-Smirnov tests are suitable for a sample size under 2000 and greater than 2000 respectively. Here we have utilized the Kolmogorov-Smirnov test since our sample size is much more than 2000 cases. These tests compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. It leads to the conclusion that if a test is nonsignificant *(P>0.05)* then the distribution of the sample is not significantly different from a normal distribution and it is probably normal. Alternatively, in the case of a significant test *(P<0.05)*, the distribution in question is significantly different from a normal distribution which means there is a non-normal distribution available.

3.2.1 Kolmogorov-Smirnov Test

The Kolmogorov-smirnov test is a nonparametric statistical test which can be used specially to test the normality of a distribution. The test process begins with samples being standardized and then compared with a standard normal distribution. Contrary to the t-test process, this particular test does not rely only on the location of the sample mean and works even for non-normal data. This test is known to be more powerful than χ^2 and less sensitive than a t-test if the data is very normal.

For a random variable *X* and a sample {x_1, x_2, ..., x_n} the empirical distribution function of *X* is defined as

$$F_X(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

where *I* (*condition*) is the indicator function,

for two cumulative probability functions F_X and F_Y, the test statistics are:

$$D_+ = \max_x (F_X(x) - F_Y(x))$$
$$D_- = \max_x (F_Y(x) - F_X(x))$$

Usually the value *D* = max {D_+, D_-} is used.

3.3 Feature Selection

In real-world supervised learning where the underlying class probabilities are unknown and each record is associated with a class label, significant features are often unidentified. To determine the significant features, several methods were introduced in section 2.5. Factors such as data type, number of classes, size of dataset and noise in dataset are imperative in selecting the proper method.

The considered dataset in this research is a mix of continuous, discrete and nominal data types with multiple classes; the number of corresponding class is five. It is a large dataset encompassing seventeen independent variables which are patients' features and one dependent variable known as the outcome of tuberculosis treatment course for 6,450 cases. That is, the available dataset can be written as [6,450*18]. Hence, the suitable feature selection approach capable of handling dataset with these characteristics is heuristic/dependence through correlation coefficient measures. These methods tend to find the relationship between every independent variable in the dataset and the given dependent variable as well as the associations among them. A bivariate correlation is a correlation between two variables including independent and dependent parameters by calculating the correlation coefficient either through Pearson`s product-moment or Spearman`s *rho*. In the case of availability of directional hypothesis, the one-tailed test is utilized. In contrary, if there is no specific prediction and hypothesis two-tailed test is carried out.

3.3.1 Pearson Correlation Coefficient

The Pearson correlation coefficient *r* is a scale-free measure of linear association between two variables *x* and *y*, and is defined as follows: An optimal subset is always relative to a certain evaluation function. If the variable is normally distributed, then the

Person correlation value is defined to find the association between a feature and a class. To discover the degree to which the variables are related, correlation criteria are applied. The Pearson coefficient is the most popular measurement of correlation which is designated by the letter "r" when computed in a sample [Field, 2005]. It reflects the degree of linear relationship between two variables ranging from +1 to -1. A perfect positive linear relationship between variables is shown by +1; however, -1 implies an entire negative linear association.

The following formula is used to calculate the value of *r*

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_Y} \qquad (3.5)$$

where there are two variables X and Y and their means $\bar{X},\ \bar{Y}$ and standard deviations including $S_{X,}$ and S_Y respectively. The value of r would be positive if the values of *X* and the associated *Y* are both above the average; this causes the value of $(X_i - \bar{X})(Y_i - \bar{Y})$ greater than 0. If the *X* value and the *Y* value were both below average, then the product would be two negative numbers which would also be positive. In the case of the *X* value was below average and the *Y* value was above average, then the product would be negative.

3.3.2 Spearman`s *rho*

Spearman's rank correlation coefficient or Spearman's *rho* denoted by the Greek letter ρ (rho) is a non-parametric measure of statistical dependence assessing the relationship between two variables using a monotonic function. Generally, if the variable is not normally distributed, the Spearman correlation is applied to detect the association between a feature and a class. In the case of no repeated data available, a perfect Spearman correlation of +1 or -1 occurs if each of the variables is a perfect monotone function of the other. Suppose that there are n raw scores X_i (the independent variable), Y_i (the dependent variable) converted to ranks x_i, y_i ; the differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are computed. In the situation of tied ranks, then the value of rho is calculated through following formula:

$$rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (3.6)$$

The Spearman correlation coefficient is positive if both X and Y tends to increase. However, if Y tends to decrease when X increases, the Spearman correlation coefficient value would be negative. A Spearman correlation of zero signifies that there is no tendency for Y to either increase or decrease if X increases.

3.3.3 SPSS

SPSS (Statistical Package for Social Science) is one of the most widely applied computer programmes for statistical analysis; it is used by market researchers, health researchers, survey companies, governments and education organizations and others. SPSS is a comprehensive statistical package that includes, not only descriptive statistics like cross tabulation, frequencies, data analysis and exploration, but also bivariate statistics such as means, t-test, ANOVA, and correlation coefficient methods like Pearson correlation and Spearman-*rho*. This package has both user interface and command syntax programming facility. Furthermore, SPSS is capable of providing graphical plots for descriptive scales and drawing the degree of normality. These capabilities make SPSS a good choice to fulfil the requirements for this research with regards to exploratory data analysis and feature selection which are two underlining steps in the methodology.

3.4 Supervised Model Development and Selection

In supervised learning there is a set of data known as 'training data' that is composed of a set of input data in which each record has its own related target data, which is the answer that the algorithm should work out. This also has been shown as a set of data (x_i,t_i), where the inputs are x_i , the corresponded targets are t_i and the i index suggests that we have lots of pieces of data indexed by i starting from 1 to some upper limit N [Hardin & Chhieng, 2007]. The training process requires enough examples in the training data to minimise discrepancy and produce the learning answers. Each piece of data consists of values for a number of attributions and supervised learning simply attempts to reduce the difference between the expected and observed values(t_i).

Here, the available dataset of TB patients in order to predict the outcome of tuberculosis treatment course meets the conditions associated with applying the supervised learning.

For each patient, there is an ordered pair (x , t) and the training set containing 6,450 cases can be written as:

$$[X = \{x^i, t^i\}_{i=1}^{N=6,450}]$$

Where x^i is the symbol for the seventeen features and t^i is the associated target outcome for the *ith* TB patient. Thus, suitable and enough data is on hand since there are 6,450 records (N) containing seventeen features and the related real outcome. Supervised learning encompasses several tools and techniques and based on the criteria analysis discussed in section 2.6.1.1, six algorithms are selected. However, we need to focus on applying the algorithm for classification purpose.

Typically, data classification is divided into a two-step process: learning step (training phase) and accuracy test.

In the first step, to describe a predetermined set of data classes or concepts, a classifier is developed. Actually, in the learning step, a classification algorithm constructs the classifier by analysing or learning from a training set which consists of database tuples and their associated class label. In other words, let a tuple X, (e.g. patient`s dataset) be represented by an i-dimensional attribute vector, $X = \{x_1, x_2, \dots, x_i\}$, showing i measurement made on the tuple from i database attributes called A_1 , A_2 , ..., A_i respectively. Each tuple, X, is supposed to belong to a predefined class as determined by another database attribute named the class label attribute. The class label attribute is discrete-valued and unordered; they are categorical which means each value serves as a category or class. The training tuples is chosen from the database under analysis. In the process of this step, actually, by learning a mapping or function, $y = f(X)$, predict the associated class label y of a given tuple X can be made. Thus, we intend to learn a mapping or function that classify the data classes, mainly expressed in the form of classification rules, decision trees or other mathematical formula.

In the second step, the model is applied for classification purposes by estimating the classifier accuracy. If we apply only the training set to check the classifier accuracy, the result would most likely be optimistic due to the classifier tendency to overfit the data. That is, during learning it may incorporate some particular anomalies of the training data that are not present in the general dataset. Hence, a test set is usually composed of test tuples and their associated class label. Obviously, this set should be chosen from the general dataset randomly, independent of the training samples. However, as with classification, an independent test should be used to estimate the model accuracy. There are methods to assess the predictive model accuracy by calculating an error, based on the difference between predictive values and real values of y estimated for each of the test tuples.

3.5 Introduction to the WEKA Environment

According to the critical analysis discussed in section 2.6.1, six classifiers including decision trees, Bayesian networks, Logistic regression, Neural networks, Radial basis functions and Support vector machines are selected. To develop these algorithms, a suitable environment is required. A number of machine learning algorithms including the above mentioned classifiers are well known implementations in the freely available code package WEKA (Waikato Environment for Knowledge Analysis). This package is developed at the University of Waikato in New Zealand; it is available at http://www.cs.waikato.ac.nz/ml/weka/. The system provides a uniform interface to a number of different learning tasks such as classification, regression, clustering, associated rules and visualization. The algorithms can either be employed directly to a dataset or called from other JAVA code. Moreover, the environment is capable of pre-processing and post-processing to evaluate the result of a learning scheme on any offered dataset.

3.6 Un-supervised Model Development

In unsupervised learning, there are no class labels to train the system. In order to conduct clustering and unite observed examples into clusters, two major criteria should be satisfied:

- Every cluster should be homogeneous which means instances that belong to the same cluster are similar to each other.
- Every group should differ from other clusters; that is, instances that belong to one cluster should be different from the instances of other groups.

In the case of the large dataset on hand, partitioning methods can produce a more promising result than a hierarchical approach. K-means is the most well-known and frequently applied partitioning approach with a large dataset. Thus, in the current research with a large dataset available, the k-means method is recommended.

3.6.1 K-means Cluster Analysis

K-means is a centroid-based algorithm which takes the input parameter, normally named *k*, and then partitions a set of *n* objects into *k* clusters leading to high intra-cluster similarity and low inter-cluster similarity. The mean value of the objects in a cluster is the way to determine cluster similarity which can be viewed as the cluster`s centroid or centre of gravity. The k-means algorithm initially selects *k* objects, each of which primarily shows a cluster mean or centre. Then, for each of the remaining objects, one object is assigned to the cluster with most similarity according to the distance between the object and the cluster mean. Next, it computes the new mean for each cluster iterating until the centriod function converges. Generally, the square-error criterion is used which is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |P - m_i|^2 \qquad (3.7)$$

Where E is the sum of the square error for all in the dataset; P is the mean of cluster c_i instances when both P and m_i are multidimensional. That is, for every object in every cluster, the distance from the object to its cluster centre is squared and the distances are summed up. This method creates divided and compacted *k* clusters as much as possible. More detail about k-means clustering steps is illustrated in chapter 5.

3.6.2 Silhouette Analysis

After creating clusters indices by the k-means partitioning algorithm, the silhouette may reflect how well-separated the resulting clusters are. The silhouette is a plot where rows correspond to the objects of the *n*-by-*p* data matrix *X* and columns are associated with each cluster which can be a categorical variable, numeric vector, character matrix or cell array. A number of approaches are available to calculate distance between points; squared Euclidean distance is the most used way to compute distance between objects.

The produced silhouette plot in actual fact displays a measure of how close each point in one cluster is to points in the neighbouring clusters ranging from +1, indicating points that are very distant from neighbouring clusters, through 0, denoting points that are not distinctly in one cluster or another, to -1, signifying points that are probably assigned to the wrong cluster. Silhouette analysis is able to draw different number of clusters in plots which can help us to visually spot the most distinct and well separated clusters. In fact, the quality of clustering method performance is actually evaluated through silhouette analysis.

3.7 Combination of Supervised and Un-supervised Learning Methods

Supervised Learning methods application sometimes results in poor outcome when it deals with general estimation in high-dimensional space [Han& Kamber, 2006]. To overcome this limitation, local mapping with using more partitioned samples distributions is suggested. It is, in fact, the integrated supervised and unsupervised learning which intends to take the advantages of both methods in order to produce more accurate and promising results.

Combining approaches may lead to the advantages of both supervised and unsupervised learning methods to build up the integrated models that could best reflect the predicted class. In this way, comparable cases are collected in clusters according to their similarities discovered among their input features. Typically, this process is conducted before supervised learning and feeds the supervised learning algorithms by the more grouped and similar records. At the next stage, the learning process proceeds with the supervised learning paradigm in order to estimate the considered classes which in this case is the outcome of tuberculosis treatment course destination. This may affect the classification algorithm accuracy positively to amplify predictability; also the iteration times might decrease as the classification algorithm is trained from clustered data. This might be as a result of combined supervised and unsupervised learning algorithm ability to handle large bodies of data and, moreover, unsupervised learning performance in partitioning of the training dataset. That is, after creating partitions by clustering approaches, a supervised learning algorithm is applied to each partitioned dataset. Thus, instead of learning from the whole training dataset, combining the two learned results may lead to increased pace, accuracy and even comprehensibility of the produced predictive model. Prediction and generalization capabilities of this combined method may provide a basis for the strong and flexible mapping of input attributes into the single valued space of the tuberculosis treatment course.

Combined supervised and unsupervised learning methods have already been used to find significant features and cause-and-effect relationship between factors and the target variable.

In 1992, a piece of research [Pao & Sobajic, 1992] combined use of unsupervised and supervised learning for dynamic security assessment to estimate critical clearing time (CCT). CCT is highly advantageous in assessing the security and stability of electric power systems after exposure to large disturbances through offering significant information about the quality of the post-fault system behaviour. Feed-forward neural networks have been applied to learn the given mapping; under defined variable system, operating conditions and topologies it performed well. The combination of supervised and unsupervised learning found what arrangement of 'raw' features is significant to define CCT. Input patterns were clustered based on their similarities revealed among the input features. The range of label attributes effective for the given outcome (CCT) in every cluster denotes which associative actions should be taken in each instance; according to the level of clearing time, low or high, installed systems components or company policy advice for preventing control schemes may vary. In the next stage, based on unsupervised learning results, accurate estimation of the CCT parameter was carried out using supervised learning. To map clustered patterns into the single valued space of CCT values, Functional Link Net (FLN) which is a type of network architecture was used. Then, the covariance analysis of clustered patterns was done and only highly correlated features appeared in the enhanced representation. This helped to find the combination of features playing an influential role in specific situations which should be retained for the supervised learning mode. This report also concludes that applying a supervised and unsupervised learning combination method helps greatly to cope with large datasets due to the unsupervised learning ability to find similarities among data. The supervised algorithm afterwards used the clustered data to synthesize accurately the value of CCT.

In 2001, a twofold study [Šmuc *et al.*, 2001] was conducted to firstly address new or previously unknown cause-and-effect relationships between coronary heart disease (CHD) factors and then validate findings against cardiologist’s classification analysis. To meet this end, unsupervised learning partitioned data prior to learning models by classification. This created different and more informative, and possibly even more accurate descriptions, of separate patient subgroups independently. The dataset was related to 239 coronary heart diseases, patients with 40 defined parameters collected over 1.5 years; this database was composed of patients` data who were already suffering from CHD or having characteristic symptoms. Cardiovascular analysis for every case also was available based on their exercise ECG and long term ECG tests, they classified patients into five categories from 1-healthy to 5-patient in a critical stage of the disease. Using the WEKA (Waikato Environment for Knowledge Analysis) package, three main stages were conducted including first using clustering algorithm (EM-expectation maximization) to detect major regularities existed in the data base, second exploring the developed clusters and modelling their data by the decision tree algorithm C4.5, and finally splitting descriptors and their values using the most important basic set descriptor from the DT model which were closest to a root node as splitting criterion. Based on classification model accuracy for every cluster, final assessment of clusters and splitting was carried out.

Although a combined supervised and unsupervised learning method has been already used to fulfil aims such as feature analysis or cause-and-effect relationship detection, however, it is the first time that this approach is used for prediction accuracy improvement and interpretability enhancement. The novel methodology steps are described in chapter 5 in detail.

3.8 Using Cluster Analysis for Decision Tree Simplification and Accuracy Improvement

Here, the best desirable predictive model is the most accurate and also interpretable option. The desired model not only can predict the outcome of tuberculosis treatment course precisely, but also the predicted outcomes can be easily interpreted by health care provider even with low level of medical knowledge. Decision trees are the most well known comprehensible classification algorithm. The advantage of the model developed by decision trees over other methods is that they can be interpreted by users via produced decision rules. In fact, they extract decision rules from a database which makes them well suited approach for medical applications [Dreiseitl *et al.*, 2001].

Numerous systems have been developed to construct decision trees from collection of examples. In spite of the fact that those generated decision trees may perform well and are efficient, they usually suffer the disadvantage of excessive complexity and are therefore unintelligible to users. The number of nodes and tree size are two criteria to measure the decision tree complexity. No matter how accurate and efficient they are, it is under question whether their opaque structures can be described as a knowledge source. That is, if the two trees employ the same kind of data and have the same prediction accuracy, the one with fewer leaves is usually preferred.

Overfitting is another difficulty which occurs during learning the examples by a decision tree. In other words, a decision tree, or any learned hypothesis h, is said to overfit training data if another hypothesis h' exists with a bigger error than h when tested on the training data, but a smaller error than h when tested on the entire dataset [Kotsiantis, 2007]. There are two general methods that decision tree induction algorithms can apply to avoid overfitting training data. Firstly, stop the training algorithm before it reaches a point at which it completely fits the training data, and secondly, prune the induced decision tree. Cost-complexity pruning, reduced error pruning, pessimistic pruning are pruning techniques. Pruning including pre-pruning or post-pruning are the most straightforward ways to improve decision tree comprehensibility and tackle overfitting [Quinlan, 1999].

In pre-pruning of decision tree, we don’t allow a decision tree to grow to full size. Typically, a decision tree algorithm is able to employ post-pruning methods that evaluate the performance of decision trees using validation set. Any node of a decision tree can be removed and assigned to the most frequent class obtained from the corresponding training instances. Although the four above mentioned pruning techniques are capable of achieving significant simplification, there are still weak points like accuracy improvement and testing set requirement [Quinlan, 1999]. Methods have always attempted to pursue the twofold goal of improving decision tree comprehensibility along with upholding or improving accuracy. However, how successful these approaches are to fulfil this twin goal is the subject of debate since simplification procedures that significantly increase classification accuracy are unlikely to be functional. Pruning has been the most common method to simplify decision trees either by pre-pruning (imposing a non-trivial stopping criterion on tree expansion) or post-pruning (deleting sub-trees after induction the tree) while their use should be adjusted based on the data characteristics and distribution. Their focus is mainly on tree simplification and not accuracy enhancement.

The defect of available simplifying methods in accuracy improvement has led to a new algorithm requirement to fulfill the dual objective of comprehensibility and accuracy improvement. There is no report about applying any learning method like clustering to simplify decision trees with accuracy improvement. Developing a cluster-based simplified decision tree (CSDT) through applying an unsupervised learning method to simplify a decision tree along with accuracy improvement is a novel method that needs to be investigated. The partitioning entity of both the k-means clustering method and a classification decision tree may boost the learning algorithm performance. Normally, in the process of decision tree construction, by finding the feature that best divides the training data, the root node is selected. By creating sub-trees and dividing the training data into subsets of the same class, the process of partitioning the divided data continues. In the novel approach of CSDT development, applying the K-means clustering method which partitions the value of features without considering the target value for each record may enhance the partitioning stage at large branches of decision tree. Further investigation about the proposed CSDT is conducted in the chapter 6.

3.9 Methodology Design in Overview

The applied methodology in this research is drawn in a schematic process in figure 3.1. Available dataset is addressed as $X = \{x^t, r^t\}_{t=1}^N$, when X is the multidimensional dataset with *x* variables (the predictors) and *r* as a vector of outcome of tuberculosis treatment course (target class); *t* is the index of records which are TB patients; thus, for each case of *t* instances, there are several variables (*x)* and corresponding target class(r). In the first phase, the dataset needs to be pre-processed by exploratory data analysis and feature selection through statistical analysis and bivariate correlation respectively. Bivariate selection is recommended since there is a large body of dataset with both discreet and numerous independent variables and multiclass dependent variables. To find the best machine learning approach, decision trees (DT), Bayesian networks (BN), logistic regressions (LR), multilayer perceptrons (MLP), radial basis functions (RBF) and support vector machines (SVM) are examined. The best model on basis accuracy measurements is chosen by comparison frameworks like prediction accuracy, precision, recall and F-measure. In fact, this step of the methodology is aimed to examine the given algorithms to find out the best. In the second phase, unsupervised learning is used to feed supervised learning by more refined instances; in other words, supervised and unsupervised learning approaches are combined for each algorithm distinctly. Then, it is examined whether or not the learning accuracy is improved and what is the most accurate integrated algorithm to predict the outcome of tuberculosis treatment course. Here, K-means clustering method is proposed because of the large dataset on hand. Regarding the fact that the best possible predictive model is the most accurate, quick in learning and comprehensible one, phase 3 is aimed to simplify a decision tree. It intends to increase the trees` interpretability and precision through reducing the size of the developed trees and increasing their prediction accuracy respectively. Hierarchical clustering and classification method is proposed to simplify the huge branches of trees and develop CSDT. In phase 4, produced knowledge needs to be refined and managed leading to developing a decision support system in future work and research. Every part of this methodology is illustrated in the corresponding chapters in more detail.

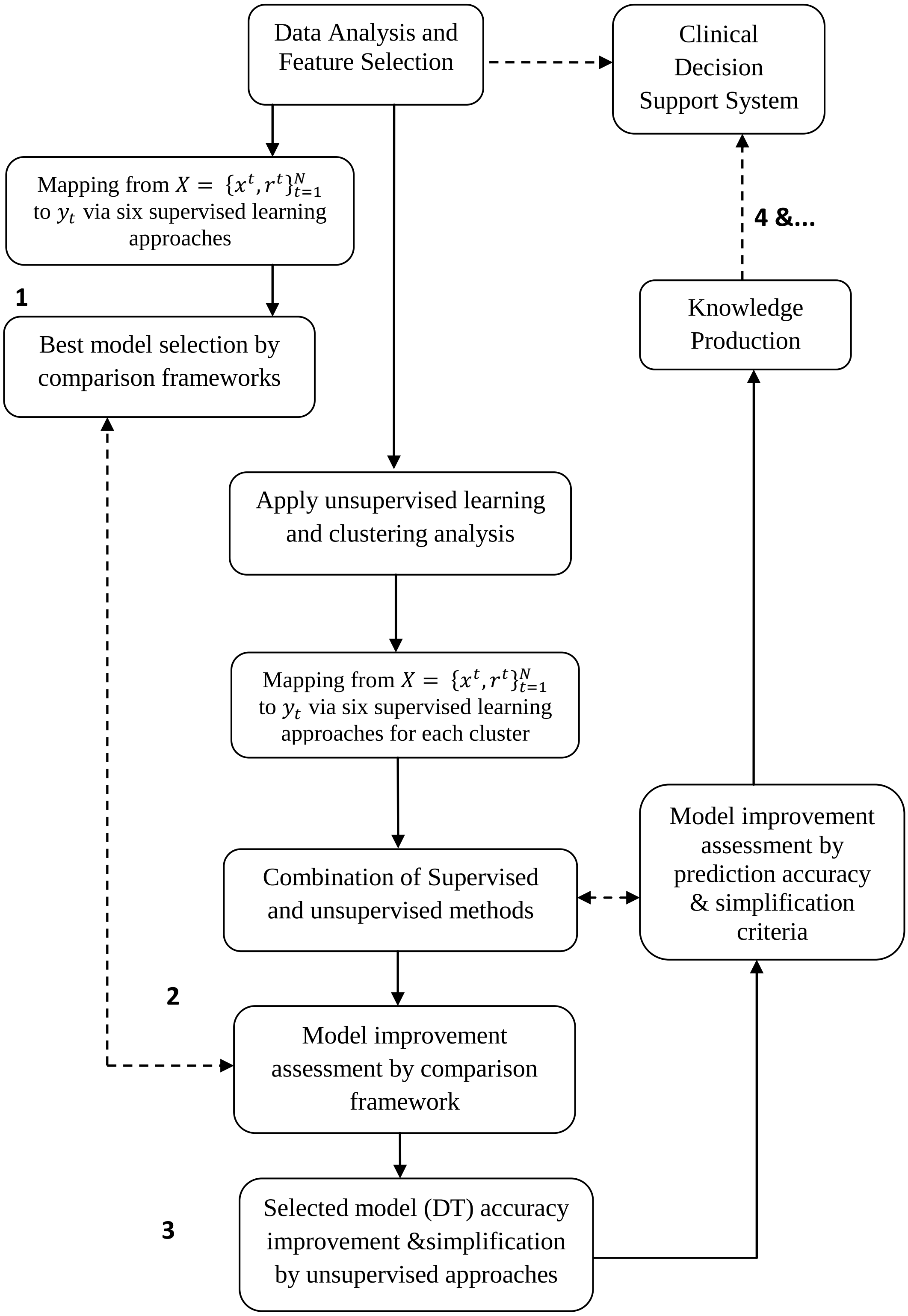

Data Analysis and Feature Selection
Clinical Decision Support System
Mapping from $X = \{x^t, r^t\}_{t=1}^N$ to y_t via six supervised learning approaches
4 &...
1
Best model selection by comparison frameworks
Knowledge Production
Apply unsupervised learning and clustering analysis
Mapping from $X = \{x^t, r^t\}_{t=1}^N$ to y_t via six supervised learning approaches for each cluster
Model improvement assessment by prediction accuracy & simplification criteria
Combination of Supervised and unsupervised methods
2
Model improvement assessment by comparison framework
3
Selected model (DT) accuracy improvement &simplification by unsupervised approaches

Figure 3.1 The conceptual diagram of applied methodology in this research.

3.10 Summary

In conclusion, this chapter has introduced the methodology design of the study. Statistical analysis and feature selection methods designed to be used in order to prepare necessary data related to the influential factors of tuberculosis treatment course. Next, they feed supervised and unsupervised learning algorithms in a combined way. Then, it is considered to find the most accurate classifiers among six applied algorithms and reveal the effect of a combination approach on classification accuracy. Furthermore, to simplify and improve the decision tree accuracy and comprehensibility, at the same time, k-means clustering approach is considered. The designed methodology is summarized in figure 3.1.

Chapter 4

Comparison of Supervised Machine Learning Techniques to Predict the Outcome of Tuberculosis Treatment Course

4.1 Introduction

The importance of predicting tuberculosis treatment course for TB control has already been highlighted. In the previous chapter, the overview of this study`s methodology was described and six considered classifiers were introduced in four main categories including logical based algorithms, perceptron-based techniques, statistical learning algorithms and support vector machines. This chapter is aimed at finding the most accurate classifier among six examined algorithms to predict tuberculosis treatment course destination.

To predict various outcomes in medical domains a number of studies have been conducted by applying various machines learning tools. It has led to the development of several imperative clinical decision support systems which are applicable to assist physicians and health assistances in practice. HELP which is a Knowledge-based hospital information system and MYCIN which is to identify micro-organisms that cause bacteremia and meningitis are two examples of applied clinical decision support system used in routine health care system [Hardin& Chhieng, 2007].

As reviewed in the second chapter, supervised machine learning including decision tree, logistic regression, bayesian networks, support vector machine, radial basis functions and neural networks are applicable in several medical domains. However, they have not been tested to predict the outcome of treatment course yet.

Classifiers have performed with varying results in different situations. The choice of algorithm always depends on the task at hand and the most promising results may emerge based on numerous factors and situations which are rarely perfect, as in the case of multidimensionality, outliers or missing data. In other words, no single algorithm can uniformly outperform other methods over all data collections. Thus, the simplest solution is to estimate the accuracy of the candidate techniques on the given task and choose the one that appears to be most accurate [Kotsiantis, 2007].

This chapter has been organized as follows. In section 4.2, the experimental methodology and setup is explained by detailing the available dataset, and applied classifiers. In next two sections, 4.3 and 4.4, the obtained results are reported and discussed. Finally the last section there is a summary.

4.2 Material and Experimental Method

4.2.1 Subjects and Data

The dataset has been gathered by health practitioners, nurses, and physicians at local TB control centres throughout Iran in 2005. In tuberculosis control centres, health deputies of each province in a network system collected data related to tuberculosis patients from every appointment in their associated regions using ‘Stop TB’ software to register TB patients. They also entered their data into a database and transferred the gathered data to the Iranian Ministry of Health. At the Centre of Disease Prevention and Surveillance, the data for 30 Iranian provinces was gathered and completed. By using ‘Stop TB’ software, more than 35 parameters for TB patients were collected. In this study we chose seventeen variables as well as the outcome for every TB patient in the frame of DOTS application.

After applying bivariate correlation, those independent variables which are significantly correlated with the target outcome ($P<=0.05$) are selected as predictors. The refined dataset consists of 6,450 cases with seventeen attributes categorized in three main categories such as demographical, clinical factors and social factors. The attributes are presented in Table 4.1 based on their related category.

Demographic factors encompass a number of patients` features such as age, sex, nationality, area of residency, as well as indicating whether or not a patient is living in prison during treatment.

The second group of variables, clinical factors, is composed of five variables related to tuberculosis or other disease history. LBW standing for low body weight is the ratio of weight based on patient age by using the standard chart specified for gender, (male/female). For each case, LBW has been defined through checking the case's weight plotted against age on the chart which is special for each gender. The categorical variable LBW is defined as whether the patient has low body weight (1) or not (0).

Diabetes and HIV are two diseases that the health practitioners and doctors quizzed the patients on at the onset and during treatment course.

Length of tuberculosis is calculated for every patient at the end of the treatment course in terms of month. Typically, the treatment course lasts six months consisting of four months for antibiotic therapy and two months for complementary treatment course.

There are two kinds of treatment scheme for TB patients applied based on their status. In treatment category *A*, the typical treatment for new patient should be used normally. However, a patient who has been unsuccessful in the type *A* plan, needs to go through the *B* scheme which is more expensive with stronger antibiotics.

Case Type denotes the patient status in terms of being a new case, imported from other countries, returned and Cure after absence. Returned implies the patient is smear positive even after completed antibiotheic therapy. Cure after absence means that patient restart the treatment course again even after quitting the treatment course.

Another attribute in clinical features is TB type which can be either pulmonary which affects the lung and respiratory system or other areas such as bones, genitalia, urinary, plover, eye, digestion, skin, ear and CNS systems.

Recent TB infection indicates whether or not the patient has been affected by TB as recently as the last six months.

The last category of attributes is social factors which could include the following: Imprisonment, referring to the history of being in prison in the patient's life. IV drug usage means whether the patient consumed any type of drugs intravenously. Finally, unprotected sex indicates whether the patient has a history of having unsafe sex. The data related to social factors has been classed as yes (1), No (2), or suspected (3). Since there is a probability that many patients who are infected with HIV, had unprotected sex and imprisonment history, or consume drugs are unlikely to reply honestly to these questions or declare these behaviours in spite of some symptoms, they have been coded as suspected here.

Initially, to get the best accuracy the dataset was divided into training and testing datasets from two-third and another one-third respectively. Table 4.2 represents the number of instances in original dataset and corresponding training and testing sets based on multiple related classes. Completed, cured, quit, failed and dead are five target outcomes which place every patient in a level of treatment from the best outcome (cure) to the worst one (dead). Cure implies that the result of sputum has been smear negative. Dead implies that the patient has died during the process of treatment because of tuberculosis or a related reason.

Table 4.1 Patients` attributes used for experiments and their range of values

variable	Categories of values
Demographic Characteristics	
Gender	Male(1) /Female(2)
Age	(Continuous var.) 0.05-99
Weight	(Continuous var.) 4-110
Nationality	Iranian(1), Central Asians(2), Iraqi(3), Pakistani(4), Afghani(5)
Area of residence	Abroad(1), Mobile(2), Rural(3), Urban(4)
current stay in prison	No(1) /Yes(2)
Clinical Features	
Case type	new(1), Imported(2), Cure after absence(3), returned(4)
Treatment categories	A(1)/ B(2)
TB type	Pulmonary(1)/extra-pulmonary(2)
Recent Tb infection	No(1)/yes(2)
Diabetes	No(1)/yes(2)
HIV	No(1)/Suspected(2)/yes(3)
Length (Month)	(Continuous var.) 0.03-90.77
Low Body Weight(LBW)	No(1)/yes(2)
Social Risk Factors	
Imprisonment	No(1)/Suspected(2)/yes(3)
IV drug using	No(1)/Suspected(2)/yes(3)
Risky sex	No(1)/Suspected(2)/yes(3)

Table 4.2 The outcomes for Tuberculosis patients for Training and Testing sets with their related codes in the dataset

	Cured (1)	Complete (2)	Quit (3)	Failed (4)	Dead (5)	Total
Training Set	1510	1274	790	462	479	4515
Testing Set	572	841	179	207	136	1935
Initial dataset	2082	2115	969	669	615	6450

4.2.2 Applied Classifiers

Decision trees, Logistic regression, Bayesian Networks, Multilayer Perceptrons (MLP), Radial Basis Functions (RBF) and support vector machines are the classifiers examined on the available dataset. Using the WEKA package, the dataset was split into training (two-third) and testing (the other one-third) datasets each containing seventeen significantly correlated attributes and the outcome variables for every record without any missing data. The Six above named classifiers were applied to the training dataset to estimate the relationship among the attributes and to build predictive models. Afterwards, the testing dataset which was not used to model development was utilized to calculate the predicted classes and compare the predicted values with the real ones available in the testing dataset. This was also done to check model fitness by comparing the predicted classes by applying the training set to the built model with the real value of the corresponding outcomes variable. However, due to the fact that testing data was not applied to model building, their application to check the accuracy was much more critical to judge the model quality. That is, the training accuracy percentage and prediction accuracy percentage are measurements to check the model fitness and model accuracy respectively. Recall, Precision, F-measure and ROC area are other criteria used to assess the model`s validity. This process has been carried out for all six classifiers in an identical way. The process of developing the predictive models and their validation is drawn in figure 4.1.

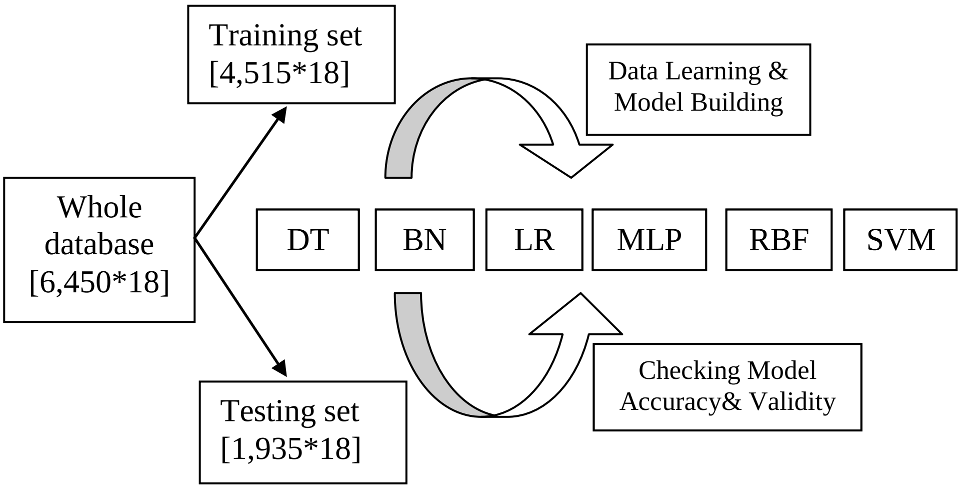


Figure 4.1 Schematic presentation of applied methodology of this chapter including six model development and validation process.

4.3 Experimental Results

Results of statistical analysis for exploratory data analysis, feature analysis as well as accuracy of six considered classifiers are addressed in the following sections, Tables (4.3 to 4.16) and Figures (4.2 to 4.11).

4.3.1 Statistical Analysis

To reveal possible errors in the data or outliers, a better understanding of the features of the dataset was obtained. This was done through applying parametric or non-parametric statistical tests and observing the skewness, kurtosis and the normality of distribution. Dataset exploration was conducted by using statistical software, SPSS 14. Table 4.3, 4.4, and 4.5 show statistical criteria such as mean, median and standard deviation values for three continuous variables including age, weight, and length of disease. In addition, skewness and Kurtosis, their related standard error as well as their normal values of ^{Z} Skewness and ^{Z} Kurtosis have been listed in above mentioned tables. ^{Z} Skewness and ^{Z} Kurtosis are two parameters for checking the normality which are defined as follows:

Z_skew=skew/Standard_Error(skew)

Z_kurtosis = kurtosis / Standard_Error(kurtosis)

Table 4.3 Descriptive statistical analysis for demographical attributions

		Sex	age	weight	nationality	Area	prison
N	Valid	6450	6450	6450	6450	6450	6450
	Missing	0	0	0	0	0	0
Mean		------	46.65	52.07	------	------	------
Median		------	46.00	51.00	------	------	------
Std. Deviation		------	20.92	12.73	------	------	------
Minimum		1	0.05	4	1	1	1
Maximum		2	99	110	5	4	2
Skewness		.093	.042	.181	1.674	-1.414	3.475
Std. Error of Skewness		.030	.030	.030	.030	.030	.030
^{z} Skewness		3.1	1.4	6.03	55.8	-4.71	11.58
Kurtosis		-1.992	-1.136	1.597	.813	2.723	10.080
Std. Error of Kurtosis		.061	.061	.061	.061	.061	.061
^{z} Kurtosis		-32.65	-18.62	-26.18	13.55	44.63	165.24

Table 4.4 Descriptive statistical analysis for clinical attributions

	LBW	Tcat	TBtype	Case Type	Length	RTB infection	Diabetes	HIV
Valid	6450	6450	6450	6450	6450	6450	6450	6450
Missing	0	0	0	0	0	0	0	0
Mean	------	------	------	------	11.33	------	------	------
Median	------	------	------	------	8.9700	------	------	------
Std. Deviation	------	------	------	------	8.29	------	------	------
Minimum	1	1	1	1	0.03	1	1	1
Maximum	2	2	2	4	90.77	2	2	3
Skewness	3.701	3.274	.644	3.185	2.943	2.355	3.666	2.118
Skewness Std. Error	.030	.030	.030	.030	.030	.030	.030	.030
^{z} Skewness	123.36	109	21.46	106.16	98.1	7.85	188.86	70
Kurtosis	11.698	8.723	-1.586	8.782	12.333	3.548	30.108	2.618
Std. Error Kurtosis	.061	.061	.061	.061	.061	.061	.061	.061
^{z} Kurtosis	197.77	143	26.43	143.96	202.13	58.16	501.8	42.91

Table 4.5 Descriptive statistical analysis for social attributions

		Imprisonment	IV drug Using	Risky Sex
N	Valid	6450	6450	6450
	Missing	0	0	0
Minimum		1	1	1
Maximum		3	3	3
Skewness		2.245	3.728	4.589
Skewness Std. Error		.030	.030	.030
^{z} Skewness		74.83	124.26	152.96
Kurtosis		3.972	12.899	21.799
Std. Error Kurtosis		.061	.061	.061
^{z} Kurtosis		65.11	211.45	357.36

It has been documented that an absolute value of ^{z} Skewness and ^{z} Kurtosis greater than 1.96 is significant at $P<0.05$; values above 2.58 are significant at $P<0.01$ and an absolute value bigger than 3.29 is significant as $P<0.001$ [Field, 2005]. Thus, it seems that for many of the variables (whit ^{z} Skewness and ^{z} Kurtosis represented in Tables 4.3, 4.4, and 4.5), the skewness and kurtosis are highly significant as $P<0.01$. However, due to the large sample size, [6,450*17], we have a small standard error for both skewness and kurtosis and consequently the values of ^{z} skewness and ^{z} kurtosis may occur through even small deviations from normal distribution. As Table 4.3, 4.4, and 4.5 demonstrate, apart from skewness for age, almost the values of other variables` skewness and kurtosis are significant ($P<0.05$).

Having presented the result of the Kolmogorov-Smirnov test for this study`s dataset in table 4.6, all considered attributions and the outcome of tuberculosis treatment course are not normally distributed $D\ (6450) = 0.235,\ P<0.001$.

Table 4.6 Kolmogorov-Smirnov test of normality for dependent and independent variables used in this study

List of Variables	Kolmogorov-Smirnov(a)		
	Statistic	df	Sig.
Outcome	.235	6450	.000
LBW	.540	6450	.000
Sex	.353	6450	.000
Age	.087	6450	.000
Weight	.059	6450	.000
Nationality	.499	6450	.000
Area	.384	6450	.000
Prison	.539	6450	.000
Treatment category	.537	6450	.000
TB type	.420	6450	.000
Case Type	.519	6450	.000
Length	.193	6450	.000
Recent TB infection	.524	6450	.000
Diabetes	.539	6450	.000
HIV	.509	6450	.000
Imprisonment	.490	6450	.000
IV drug Using	.530	6450	.000
Risky Sex	.534	6450	.000

However, in some circumstances, this test can produce misleading results particularly in the case of testing a large sample size which can easily lead to a significant result; hence, many statisticians asses graphical plots as well as this test to get the best results. Here, as we have a large sample size which normally causes a small standard error for skewness and kurtosis assessment and a significant Kolmogorov-Smirnov test, we need to have a look at the shape of attribute`s distribution visually rather than just trust the skewness and Kurtosis calculation or Kolmogorov-Smirnov test. Having looked at Figures 4.2, 4.3, and 4.4 we can find that many of the attribute distribution are negatively or positively skewed apart from variables like sex, age , TB Type and weight which seem more normally distributed. This helps to make a decision regarding which bivariate correlation should be chosen for each variable specifically, either the Pearson correlation coefficient for normally distributed variables or Spearman`s correlation coefficient for non-normally distributed data known as non-parametric statistics. Having analysed the shape of considered variables in Figure 4.3, 4.4, and 4.5, it seems that the four first histograms related to sex, age, weight, and TB type have roughly normal distributions. However, as confirmed by the significant skewness and kurtosis as well as the result obtained from Kolmogorov-Smirnov test, other left variables are non-normally distributed verified by non-normal histograms in Figure 4.3, 4.4, and 4.5.

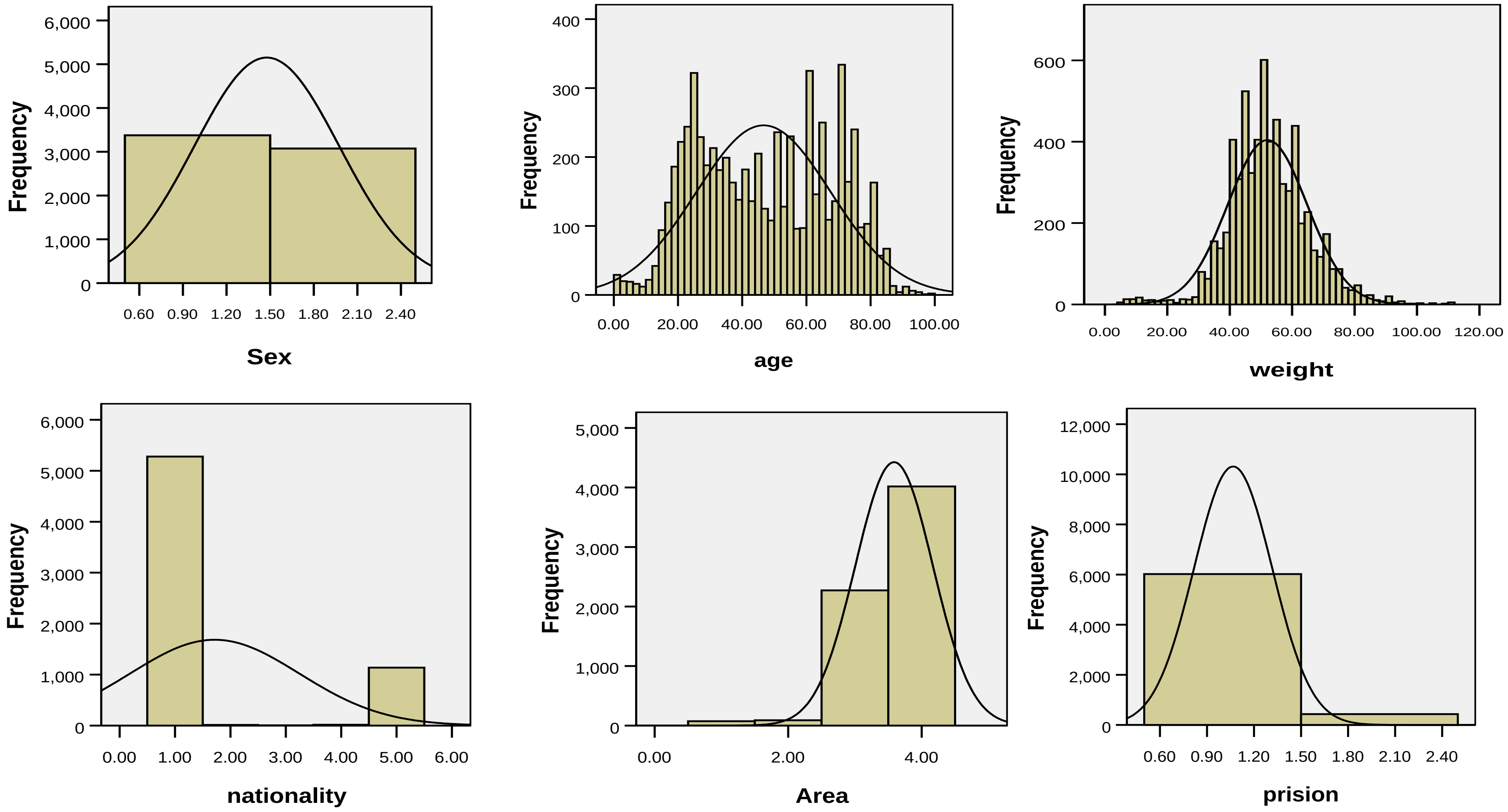


Figure 4.2 Histograms of the six variables in demographic category including six, age, weight, nationality, area, and prison.

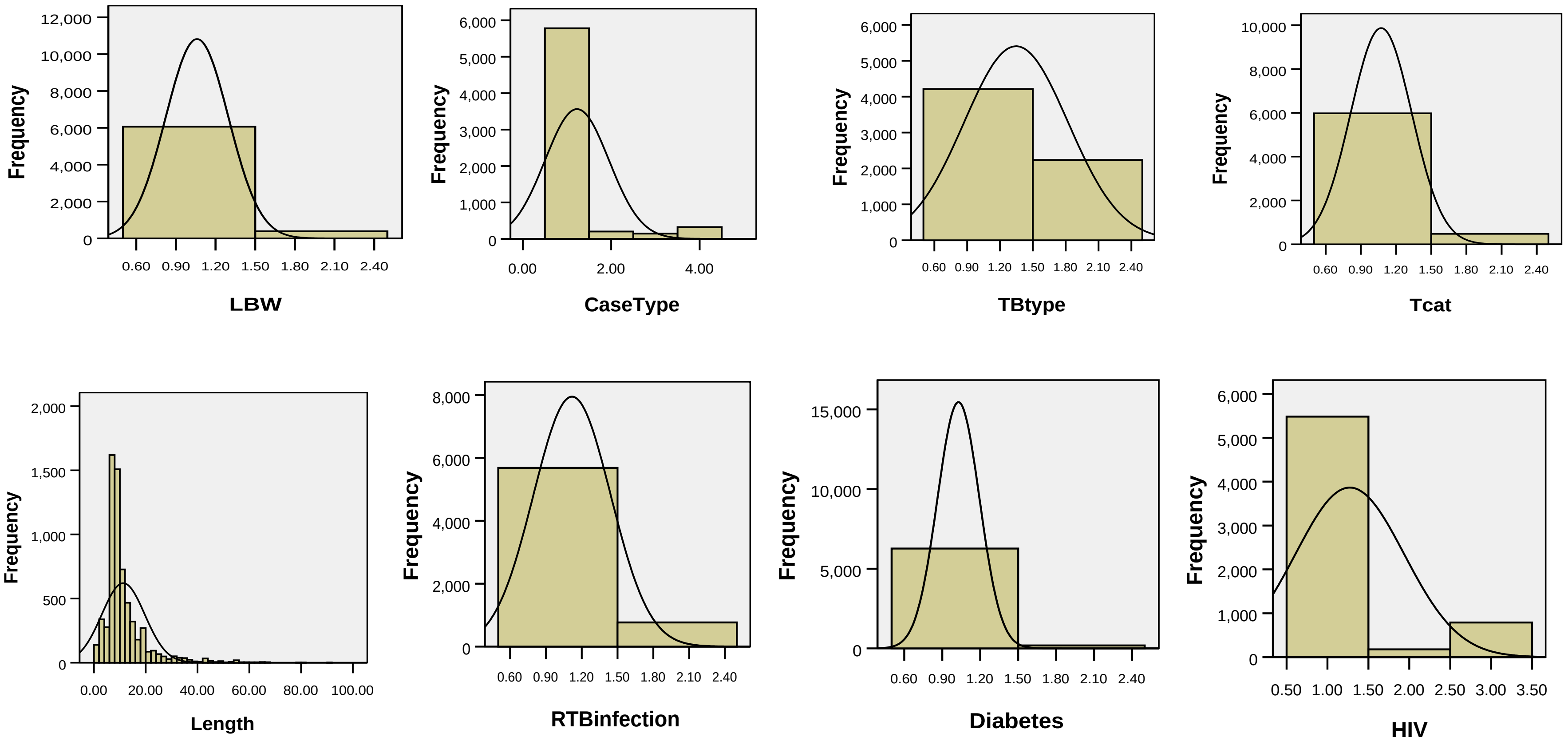


Figure 4.3 Histograms of the eight variables in clinical category including LBW, CaseType, TBtype, Tcat, Length, RTBinfection, Diabetes, and HIV.

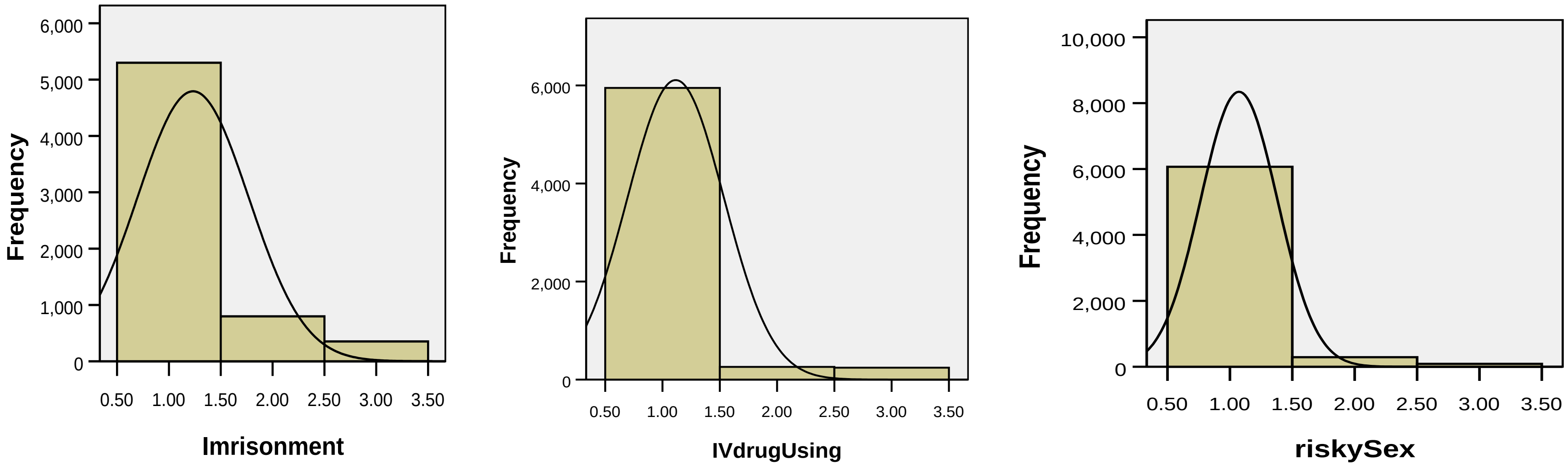


Figure 4.4 Histograms of the three variables in social category including imprisonment, IV drug using, and risky sex.

4.3.2 Results of Features Analysis

A bivariate correlation which is a heuristic/dependence method of feature analysis defines correlation between two variables including independent variable in the dataset and given dependent variable. The Pearson product-moment or Spearman`s *rho* are two approaches to calculate a correlation coefficient in normal and non-normal distributed variables respectively. Because of the availability of directional hypothesis, here the one-tailed test has been applied.

For four of the independent variables, sex, age, TB Type and weight their histograms, shown in Figures 4.2 and 4.3 seem relatively normal. Pearson's correlation coefficient has been used whereas for the other thirteen independent variables which have both significant skewness and kurtosis and non-normal distributed histograms, shown in Figures 4.2, 4.3, and 4.4, Spearman`s correlation coefficients for nonparametric correlation have been employed. The results of these applications have been shown in Tables 4.7, 4.8, 4.9 and 4.10.

Table 4.7 represents Pearson`s correlation coefficients for four of the given variables along with both the significance value of the correlation and the sample size (*N*) on which it is based.

Results show that variables including sex, age, weight, and TB type have a significant relationship with the outcome of tuberculosis course according to the significant values of the Pearson correlation coefficient; this conclusion also is drawn because of degree of freedom's values which are listed in the Tables (4.7-4.10); the low values of *p*-value which are still significantly related to the given outcome are justified with the higher values of degree of freedom (*P<.0001).*

There is a negative association between sex and outcome of the tuberculosis treatment course implying that males are more likely to not be cured and complete the treatment course rather than females (*rho = -.082, P<.0001, CI = -.055 to -.099).*

Outcome of tuberculosis treatment course appears to be positively related to the patient age (*rho =.158, P<.0001, CI=.954 to .342)* indicating that as they get old, there is more probability to get a worse result from the treatment course like failing or even a dead outcome.

Patient weight has emerged negatively with the study showing that the more under-weight the patients are, the more they are at the risk of a non-desirable outcome. This could be as minor as quitting the treatment or failing at the course or as serious as death (*rho =-.056, P<.0001, CI =-.012 to -.087).*

The last parameter in Table 4.7 illustrates that the type of tuberculosis is significantly associated with the outcome of a tuberculosis course and cases with extra-pulmonary TB are more likely at the risk of non-desirable outcomes such as quitting, failing or death (*rho =-.066, P<.0001, CI = -.033 to -.095*).

In Table 4.8, the relationship between the outcome of tuberculosis and those demographic features which are non-normally distributed has been analysed using Spearman`s correlation coefficient. The table displays that there is a significant association between nationality and outcome of tuberculosis showing that immigrants from Iraq, Pakistan and Afghanistan are 0.127 more likely to quit, fail or die on the treatment course (*rho = .127,P<.0001,CI = .087 to .221*). It implies nationality is a crucial risk factor for tuberculosis treatment non-compliance. Furthermore, TB patients who are living abroad as well as mobile cases are -0.027% more likely to have an undesirable outcome compared with those who are living in urban and rural area; it is because area of residency and the given outcome are negatively related (*rho = -.027, P<.0001, CI=-.013 to -0.341*).

The last variable in Table 4.8 is related to being in prison at the time of treatment; it reveals that prison residency and desirable treatment outcome are significantly correlated (*rho =-.026, P<.0001, CI =-.0121 to -.0321).* This might be due to the high level of supervision and support provided to those TB patients who are living in prison.

In Table 4.9, non-normally distributed clinical features have been assessed in terms of their association with the outcome of the tuberculosis treatment course using Spearman`s correlation coefficient.

There is a positive relationship between case type and the given outcome *(rho =0.048*) denoting that as the case moves from being a new TB patient to an imported, returned and cure after absence, the outcome of tuberculosis treatment course changes from cure and completion to quitting, returning or even death. It seems that new cases are safer from undesirable outcomes than returned patients.

The normal length of treatment course varies from 6-12 months. The result presented in Table 4.9 shows that there is a positive correlation between length of disease and the considered outcomes; the more time TB patients have this infectious disease, there is a 0.073 times more likely chance of non-desirable effects like quitting, failing or death. This verifies the importance of patient supervision, support and encouragement to not quit treatment which protects them from developing multi-drug TB which is the consequence of long lasting disease. Furthermore, the hypothesis of having been recently infected by TB has a positive effect on inappropriate outcome like quitting, failing or death (*rho = .251, P<0.001, CI = .195 to .342*).

Diabetes and HIV are two clinical condition which have been asked about by patients and those cases who are declared that they have these two disease are positively prone to have a worse result of treatment course completion, to either quit the therapy or fail in getting cured properly (*rho diabetes=.029, P<0.001* & *rho HIV=.045, P<0.05).*

Obtained results for treatment category 'Tcat' demonstrate that if the patient treatment category changes from *A* to *B* which lasts longer with stronger antibiotics, there is a 0.022 higher chance that they will quit or fail the treatment course. In fact, if the length of disease lasts longer, the category of treatment might be changed to *B* with both leading to worse outcome clearly.

The result for Low Body Weight 'LBW' also might point out the weight attribute finding. LBW is based on the weight of patients according to their age and gender. The hypothesis is that 'more probability that patient has LBW, more likely an improper outcome (quit, fail, or dead) might occur (*rho = .130, P<0.001, CI=.987 to .230*).

Table 4.10 also relates correlation of outcome of tuberculosis treatment course and social factors such as imprisonment, Intravenous drug usage (IV drug using), and unprotected sex which have a non-normal distribution (shown in figure 4.4). Using Spearman`s correlation coefficient, it is revealed that all three parameters are positively associated with the outcome of a tuberculosis treatment course; that is, the higher the probability of an imprisonment history in his/her life, consuming drugs intravenously, or having unprotected sex increases the likelihood of undesirable outcome during a TB treatment course such as quitting, failing or dying with *rho = 0.157, 0.0172*, and *0.16* respectively (*P<0.000).*

These presented results show the significant effect of seventeen inputs which were selected 35 assessed factors collected in 2005 in Iran from TB registered patients. We will use these significant correlated attributions in the next stage experimental analysis. Consequently, we will develop and validate predictive models for outcome of tuberculosis treatment course and attempt to improve their accuracy and comprehensibility in the next chapters.

Table 4.7 pearson`s correlation coefficients for four relatively normally-distributed variables for the outcome of tuberculosis treatment course

Attribution and type of Bivariate correlation by Pearson Correlation		Outcome (tuberculosis treatment course)
Sex	Pearson Correlation	-.082(**)
	Sig. (1-tailed)	.000
	Degree of freedom	35
	N	6,450
Age	Pearson Correlation	.158(**)
	Sig. (1-tailed)	.000
	Degree of freedom	28
	N	6,450
Weight	Pearson Correlation	-.056(**)
	Sig. (1-tailed)	.000
	Degree of freedom	36
	N	6,450
TB type	Pearson Correlation	.066(**)
	Sig. (1-tailed)	.000
	Degree of freedom	26
	N	6450

**Highly correlated (p<0.001)

Table 4.8 Spearman's correlation coefficients for three non-normal distributed variables for the outcome of tuberculosis treatment course

Attribution and type of Bivariate correlation by Spearman's rho		Outcome (tuberculosis treatment course)
Nationality	Correlation Coefficient	.127(**)
	Sig. (1-tailed)	.000
	Degree of freedom	19
	N	6450
Area	Correlation Coefficient	-.027(*)
	Sig. (1-tailed)	.017
	Degree of freedom	38
	N	6450
Prison	Correlation Coefficient	-.026(*)
	Sig. (1-tailed)	.020
	Degree of freedom	36
	N	6450

**Highly correlated (p<0.001)

Table 4.9 Spearman's correlation coefficients for seven non-normal distributed variables for the outcome of tuberculosis treatment course

Attribution and type of Bivariate correlation by Spearman's rho		Outcome (tuberculosis treatment course)
Case Type	Correlation Coefficient	.048
	Sig. (1-tailed)	.000
	Degree of freedom	41
	N	6450
Length	Correlation Coefficient	.073(**)
	Sig. (1-tailed)	.000
	Degree of freedom	38
	N	6450
Recent TB infection	Correlation Coefficient	.251(**)
	Sig. (1-tailed)	.000
	Degree of freedom	29
	N	6450
Diabetes	Correlation Coefficient	.029(*)
	Sig. (1-tailed)	.010
	Degree of freedom	40
	N	6450
HIV	Correlation Coefficient	.014(*)
	Sig. (1-tailed)	.045
	Degree of freedom	41
	N	6450
Treatment Category	Correlation Coefficient	.022(*)
	Sig. (1-tailed)	.041
	Degree of freedom	40
	N	6450
LBW	Correlation Coefficient	.130(**)
	Sig. (1-tailed)	.000
	Degree of freedom	31
	N	6450

**Highly correlated (p<0.001)

Table 4.10 Spearman's correlation coefficients for three non-normal distributed variables for the outcome of tuberculosis treatment course

Attribution and type of Bivariate correlation by Spearman's rho		Outcome (tuberculosis treatment course)
Imprisonment	Correlation Coefficient	.157(**)
	Sig. (1-tailed)	.000
	Degree of freedom	32
	N	6450
IV drug Using	Correlation Coefficient	.172(**)
	Sig. (1-tailed)	.000
	Degree of freedom	34
	N	6450
Risky Sex	Correlation Coefficient	.160(**)
	Sig. (1-tailed)	.000
	Degree of freedom	33
	N	6450

**Highly correlated (p<0.001)

4.3.3 Results of Classifiers Application

Table 4.11 shows the result of applying three types of decision tree including C4.5, Rep Tree and FT tree. C4.5 has been able to build a model with greatest accuracy since the prediction accuracy obtained by applying the model to the testing set is 74.21%. This accuracy is slightly better than Rep Tree with 71.62% and much more accurate than Rep Tree with 67.59%. This is confirmed by comparing other criteria such as recall, Precision, F-measure and ROC Area. C4.5 has performed better than the other tested algorithms using all measurements. The values of model fitness also show the C4.5

superiority where the training accuracy is 84.45% for C4.5 which is greater than 73.48 and 79.02 for Rep Tree and FTree respectively.

Results from various Bayesian networks application have been represented in Table 4.12 denoting that the best performance is obtained from Bayesian Network with 62.06% prediction accuracy. This algorithm predicts the outcome of treatment course completion much better than Navie Bayes which is known as an algorithm typically with promising results. However in this case it performs worse than a Bayesian net with 54.52% model accuracy.

Model accuracy obtained from other classifiers such as Logistic Regression, Multi layer perceptron neural networks, Radial Basis Functions, and support vector machines (shown in Tables 4.13 to 4.16) are 57.88, 57.31, 53.74 and 51.36 respectively. This, besides, is happened for ROC area that the percentage of ROC area do not address a great deal of differences among those applied classifiers when they are 0.82% , 0.81%, 0.79%, and 0.76% for LR, MLP, RBF, SVM respectively.

Comparing training accuracy percentage, likewise, it is understandable that there is no big gap between these four classifiers` performance in terms of model fitness. The values of training accuracy for logistic regression, multilayer perceptron neural networks, radial basis functions, and support vector machines are 56.5%, 64.93%, 50.65%, and 53.04% respectively.

Hence, comparing all six classifiers` performance can be conducted simply as shown in Figure 4.5 and 4.6; decision trees (C4.5) has the best performance with 74.21% prediction accuracy compared with other methods like Bayesian nets (62.06) or logistic regression (57.88), MLP (57.31), RBF (53.74) or SVM (51.36). The developed SVM is polykernel support vector machine. Six optimal values for the kernel parameters are selected which include C =1.0, L= 0.0010, P= 1.0E-12, N=0, V= -1, W=1.

The goodness of fit for the logistic regression model was 12.132, p-value <= 0.001.

It is the same story for the model fitness assessment by evaluating training accuracy which gives 84.45 for a C4.5 decision tree; this is considerably more for Bayesian nets (58.56), logistic regression (56.5), MLP (64.93), RBF (50.65) and SVM (53.04). Comparing the Roc curve area reveals that the area under curve for C4.5 has the most value both for model fitness and model accuracy with 0.960, 0.963 respectively. This measurement is less for Bayesian nets (0.85), logistic regression (0.82), MLP (0.81), RBF (0.79) and SVM (0.76) in terms of model accuracy.

Results of comparing values for recall, precision and F-measure which are, in fact, sensitivity and specificity evaluation of the developed models for both model fitness and model accuracy are presented in Tables 4.11 to 4.16 verifying the superior performance of a C4.5 decision tree in this classification task as well.

Figure 4.7 to 4.11 give the comparative Roc curves based on the given outcome of tuberculosis treatment including cure, complete, quit, failed or dead. Each figure shows the Roc curves for the six considered models based on considered outcome. The *X* axis is the number of false positive cases and the *Y* axis is the true positive rate plotting to draw the Roc curve for each classifier`s accuracy. For the outcome ‘’cure’’, C4.5 has outperformed the other classifiers with an area under curve of 0.958. Similarly, the most accurate result for outcomes ‘completed’ and ‘quit’ are from C4.5 with 0.966 and 0.956 respectively. C4.5 has also performed best for the outcome ‘failed’ and ‘dead’ by classifying 0.986 and 0.963 of cases correctly. Overall, the results for area under curve reveal the excellent performance of the C4.5 decision tree classification algorithm.

Table 4.11Comparison on model fitness and model accuracy of three various decision trees

Classifier		Model Fitness				Model Accuracy				
Decision Tree	TA* %	Recall	Precision	F-measure	ROC area	PA** %	Recall	Precision	F-measure	ROC Area
C4.5	84.45	0.845	0.845	0.843	0.96	74.21	0.742	0.753	0.746	0.963
Rep Tree	73.48	0.735	0.734	0.731	0.91	71.62	0.716	0.726	0.719	0.884
FT tree	79.02	0.79	0.79	0.789	0.91	67.59	0.676	0.696	0.684	0.83

*Training Accuracy

* *prediction Accuracy Percentage

Table 4.12 Comparison on model fitness and model accuracy of five various Bayesian networks

Classifier		Model Fitness				Model Accuracy				
Bayesian Network	TA* %	Recall	Precision	F-measure	ROC Area	PA** %	Recall	Precision	F-measure	ROC Area
Bayesian Net	58.56	0.586	0.591	0.579	0.83	61.70	0.621	0.659	0.621	0.85
Navie Bayes	49.78	0.498	0.494	0.476	0.78	54.52	0.545	0.548	0.539	0.80
Navie Bayes simple	49.72	0.49	0.494	0.47	0.78	54.72	0.547	0.55	0.541	0.80
Naïve Bayes Updated	49.78	0.498	0.494	0.476	0.78	54.52	0.545	0.548	0.539	0.80

*Training Accuracy

* *prediction Accuracy Percentage

Table 4.13 Model fitness and model accuracy of logistic regression

Classifier	Model Fitness					Model Accuracy				
LR	TA*	Recall	precision	F-measure	ROC Area	PA**	Recall	Precision	F-measure	ROC Area
	56.5	0.566	0.574	0.553	0.81	57.82	0.579	0.628	0.578	0.82

*Training Accuracy

* *prediction Accuracy Percentage

Table 4.14 Model fitness and model accuracy of multilayer perceptron (MLP) neural network

Classifier	Model Fitness					Model Accuracy				
MLP	TA*	Recall	Precision	F-measure	ROC area	PA**	Recall	Precision	F-measure	ROC Area
	64.93	0.649	0.68	0.644	0.86	57.82	0.573	0.677	0.57	0.81

*Training Accuracy

* *prediction Accuracy Percentage

Table 4.15 Model fitness and model accuracy of radial basis function

Classifier	Model Fitness					Model Accuracy				
RBF	TA*	Recall	Precision	F-measure	ROC Area	PA**	Recall	Precision	F-measure	ROC Area
	50.65	0.507	0.503	0.491	0.77	53.74	0.537	0.554	0.536	0.79

*Training Accuracy

* *prediction Accuracy Percentage

Table 4.16 Model fitness and model accuracy of support vector machine

Classifier	Model Fitness					Model Accuracy				
SVM	TA*	Recall	Precision	F-measure	ROC Area	PA**	Recall	Precision	F-measure	ROC Area
	53.04	0.53	0.555	0.503	0.76	57.47	0.514	0.621	0.50	0.76

*Training Accuracy

* *prediction Accuracy Percentage

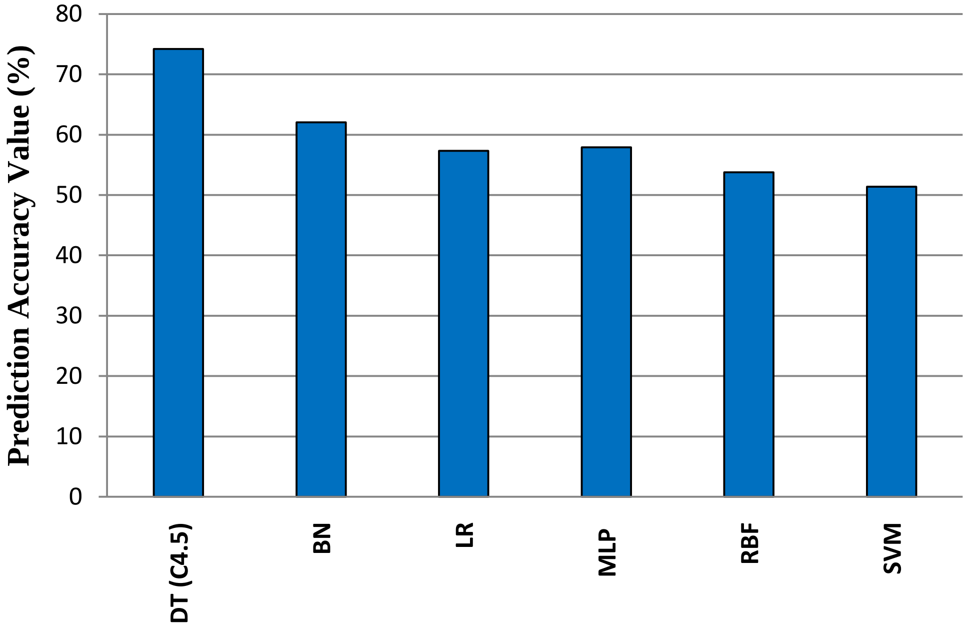


Figure 4.5 Comparisons of prediction accuracy percentage for six machine Learning tools.

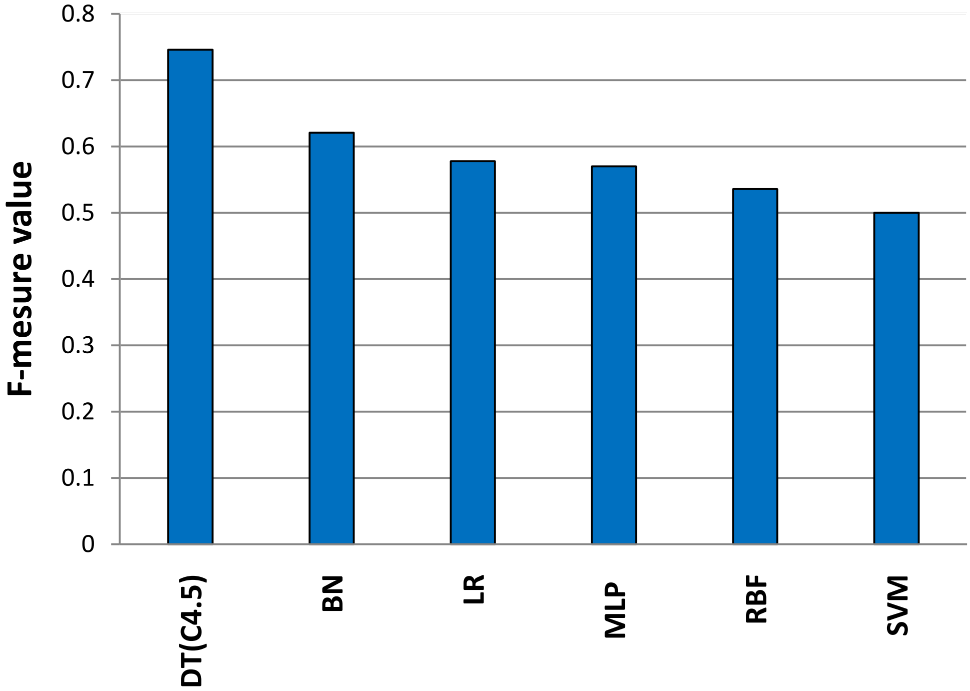


Figure 4.6 Comparisons of F-measure for six machine Learning tools.

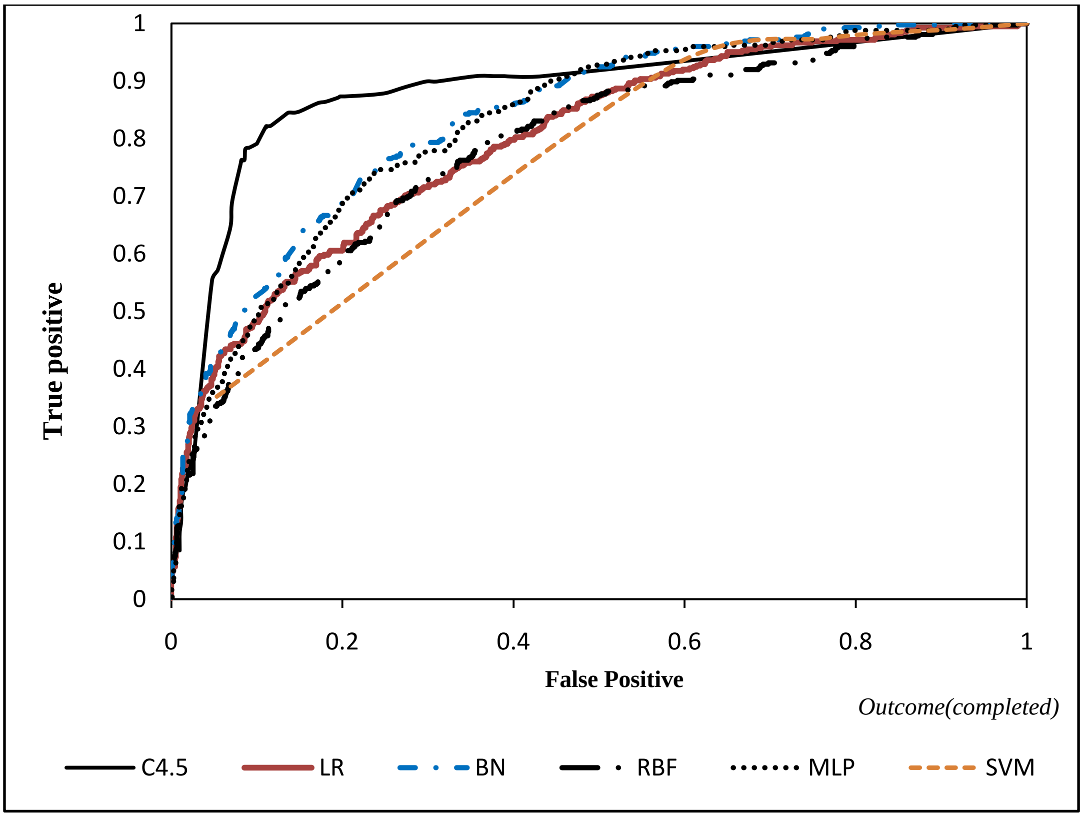


Figure 4.7 Comparison Roc Curve Area for outcome (completed) for six classifiers.

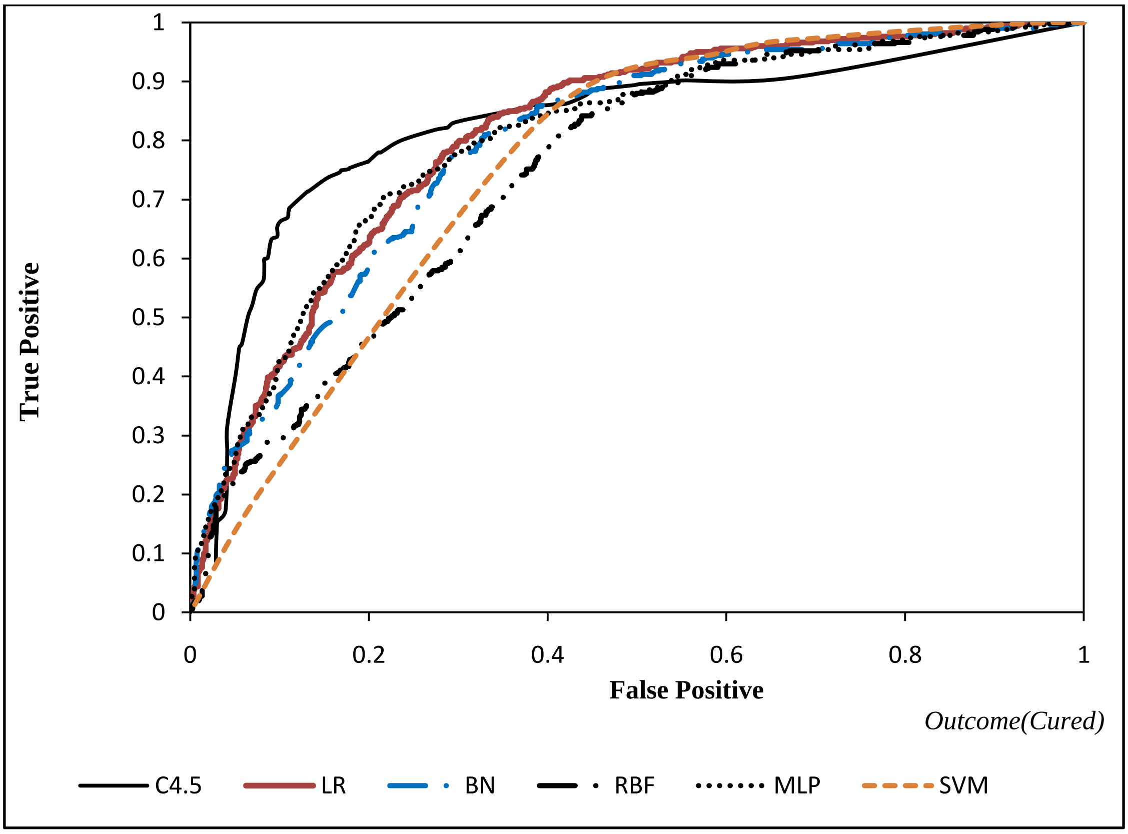


Figure 4.8 Comparison Roc Curve Area for outcome (cured) for six classifiers.

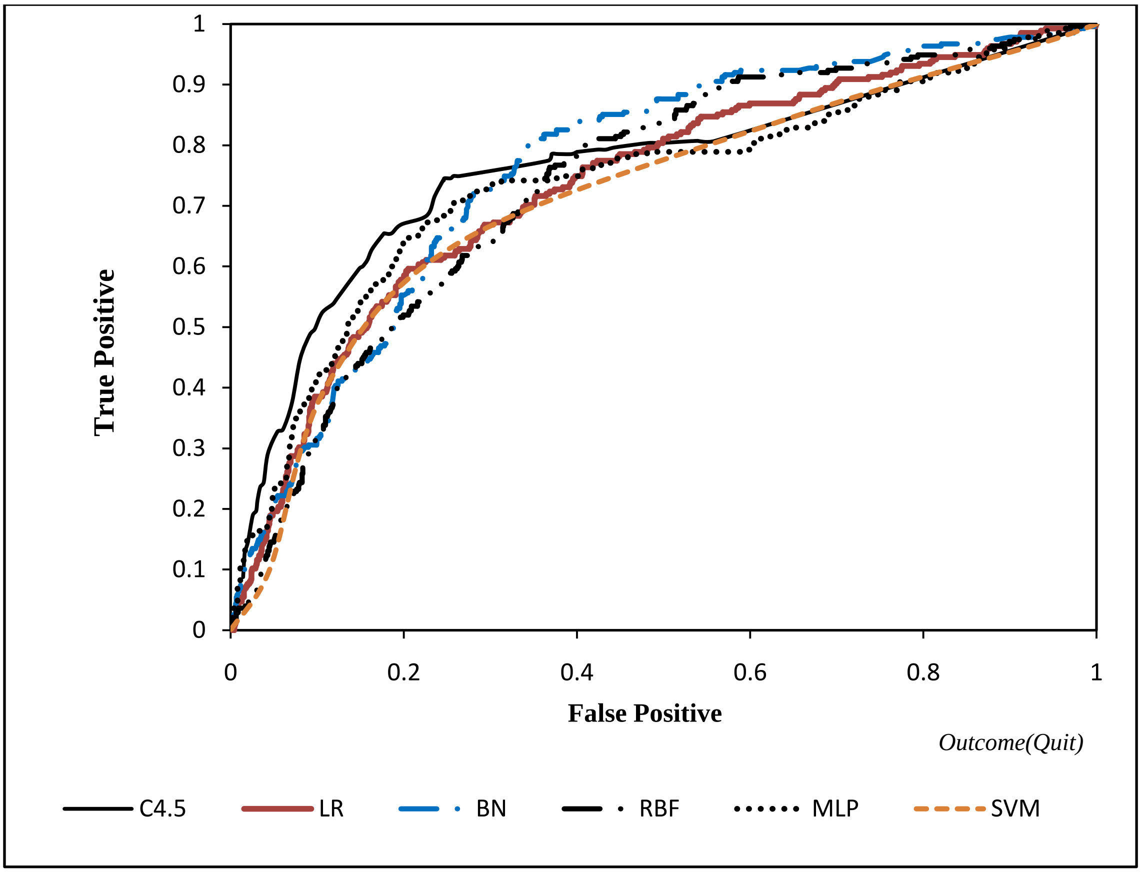


Figure 4.9 Comparison Roc Curve Area for outcome (quit) for six classifiers.

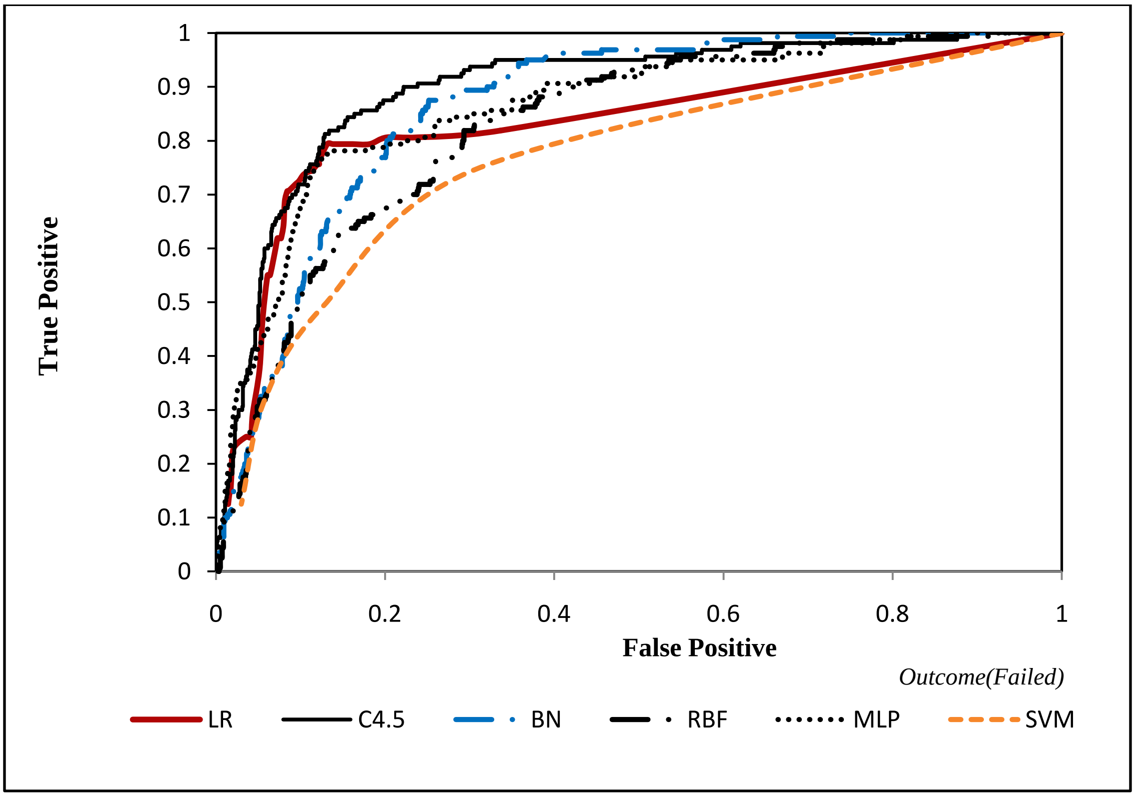


Figure 4.10 Comparison Roc Curve Area for outcome (failed) for six classifiers.

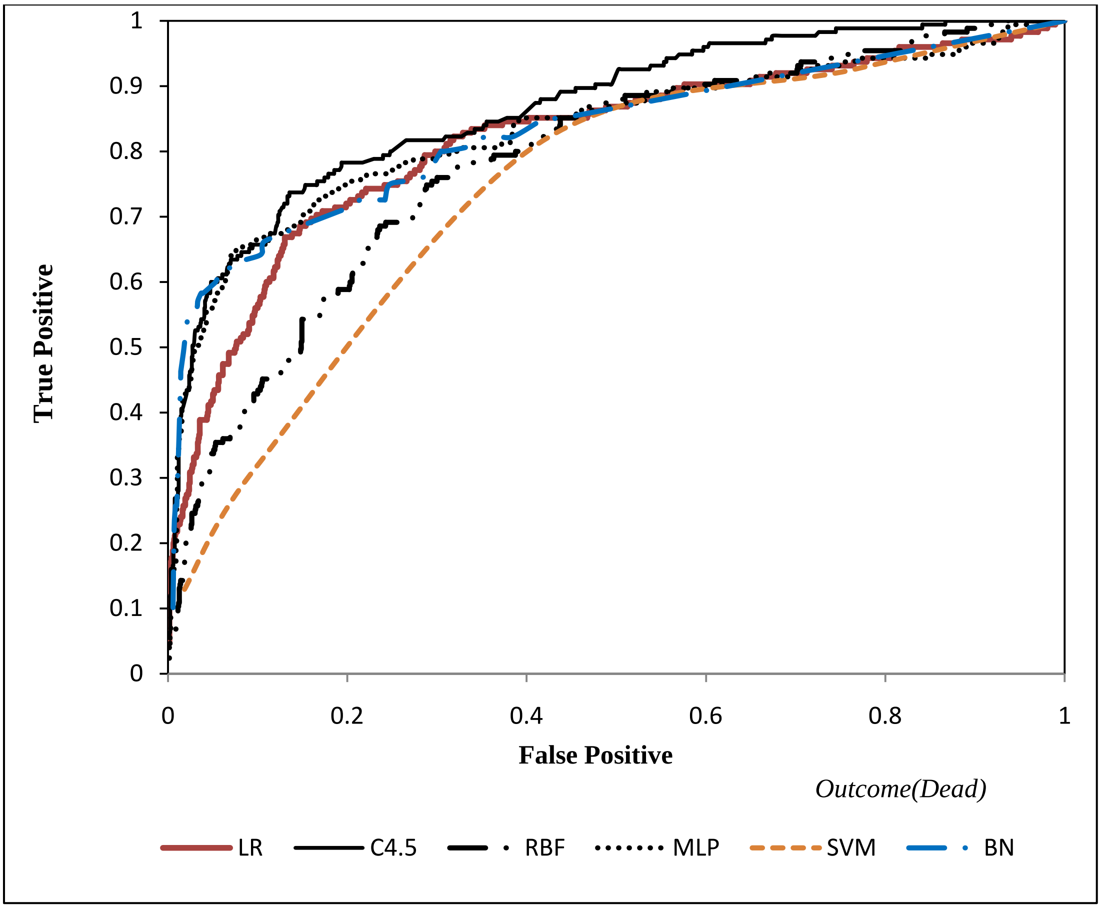


Figure 4.11 Comparison Roc Curve Area for outcome (dead) for six classifiers

4.4 Discussion

Findings presented in tables 4.7 to 4.10 reveal that this study identified seventeen influential factors effecting tuberculosis treatment course destination. Having reviewed several investigations, the influential factors on the target outcome are introduced in section 2.9. Although other work has already verified nationality, age, imprisonment, and TB case as influential factors for TB treatment course non-compliance, patient's weight is a new effective attribute *(OR = -0.056,* $P \leq 0.0001$*)*. Opposite of the reviewed studies, males are known as the high risk gender. This study strongly confirmed the role of nationality and imprisonment since, like the previous studies, immigrant people who are mainly Afghani and Pakistani in Iran are prone to failure; as the WHO indicates imprisonment as risk factor, prisoners has been indicated as high risk cases. This study introduce several new factors like diabetes, low body weight, HIV, recent TB infection, unprotected sex, TB type as well as treatment category *A* or *B*; however, it hasn't applied variables like homelessness and alcohol abuse which have been already known to be influential factors on treatment course non-compliance. It may be because of the Iranian laws related to alcohol consumption or the weakness of data collection of not paying attention to some recognized influential factors.

As mentioned before, choosing a 'best model' for a considered classification problem depends on factors such as model's discriminatory and interpretability. In this study we test the same dataset on all selected classifiers and determining the classification performance to discover the most accurate algorithm has been the main focus of this chapter. Furthermore, the model interpretability does matter as the variables applied in the dataset are human interpretable in the real world.

Of the six investigated methods, decision trees have achieved the best performance in terms of both accuracy and interpretability while other classifiers have given relatively close results in the lower rank.

According to previous studies, the technique with the best classification performance might behave differently from another one and there is no single best method for every circumstance. Decision trees that classify instances by sorting them based on feature values have performed variously in different investigations. Here, we review briefly the decision trees performance in terms of accuracy in several other studies and associated emerging condition, rather then, just this study`s parameters which may have caused decision trees outperformance rather than other classifiers.

In a piece of research [Bradley, 1997], C4.5 was compared with other classifiers by using six various real world data collections. Table 4.17 presents the main features of the applied data and the obtained results for three of the utilized classifiers. Results show that C4.5 performs variously in different conditions. It has been reported that there is an association between the performance of applied tools and following issues including the type of problem we are analysing, the type of input data either discrete or continuous, and finally emerging overlapping in outcome classes. Having looked at the result presented in Table 4.17, it is concluded that BN and MLP perform better than a decision tree (C4.5) in terms of overall accuracy in the case of continuous input with overlapping classes. Likewise, the models used by BN and MLP are specifically well suited to the type of problems in which their related data are mainly images and continuous values with noise and residuals.

Table 4.17 Comparison of produced results from 6 major dataset tested by three classifiers, adopted from [Bradley, 1997]

Data Collection	[examples *features]	Data type	outcome class number	Classifier accuracy			Area under curve		
				C4.5	MLP	BN	C4.5	MLP	BN
Cervical cell	[117*6]	images	2 Normal/ malignant	89.2	91.7	89.2	92.1	98.6	96.7
Post operating bleeding	[134*4]	continuous values	2 Normal/ excessive	71.7	78.3	79.1	48.7	66.7	73.3
Breast cancer diagnosis	[683*9] +noise& residual	continuous values	2 Normal/ malignant	90.7	93.5	94.2	93.7	96.5	98.2
Pima Indian diabetes	[768*8] +noise& residual	continuous values	2 yes/no	71.7	78.4	75.	80.2	85.3	76.3
Cleveland data (heart disease diagnosis)	[297*14] (removed missing values)	continuous values	2 yes/no	77.5	81.3	86.	84.2	85.9	90.8
Hungarian data (heart disease diagnosis)	[261*58] (removed missing values)	continuous values	2 yes/no	73	75.5	79.	79.2	84.7	83.8

In another report [Dreiseitl *et al*., 2001] C4.5 performed at a lower level of accuracy than other employed techniques including k-nearest neighbour (KKN), logistic regression, artificial neural networks (ANN) and support vector machine (SVM). The task was classification of pigmented skin lesions with three possible outcome classes. The input data was a 1619 PLS images and six other clinical data items recorded for every diagnosed case. Results showed that the top three methods were logistic regression, ANN and SVM producing identical results and, in contrary, along with

KNN, decision tree performance was considerably weak in comparison. It has been concluded that a decision tree is not ideally suited for the classification of PSL images since almost all the variables in the dataset are continuous related to images. Although in this task a decision tree`s superiority about human interpretability has been emphasized, it has been noted that C4.5 is not properly applicable due to machine generated input variables (from the vision segmentation system) with no direct correspondence to visible features of the lesion.

However, in the case of discrete data, C4.5 not only performs as accurately as other classifiers such as MLP and logistic regression, but also outperforms others such as radial basis functions. In work [Kurt *et al*., 2008] to predict the occurrence of coronary artery disease (CAD), five classification methods, regression (LR), classification and regression tree (CART), multi-layer perceptrons (MLP), radial basis functions (RBF) and self-organizing feature maps (SOFM) were applied and compared. The dataset was composed of 1,245 subjects and applied independent variables were mainly discrete. The area under curves (AUC), as shown in Figure 4.12, compared the performance of classifiers including MLP, LR, CART, RBF and SOFM. The author categorized applied techniques based on their performance in two groups; the first group, including MLP, CART, LR and RBF, performed very similarly and the second group included only SOFM which performed poorly.

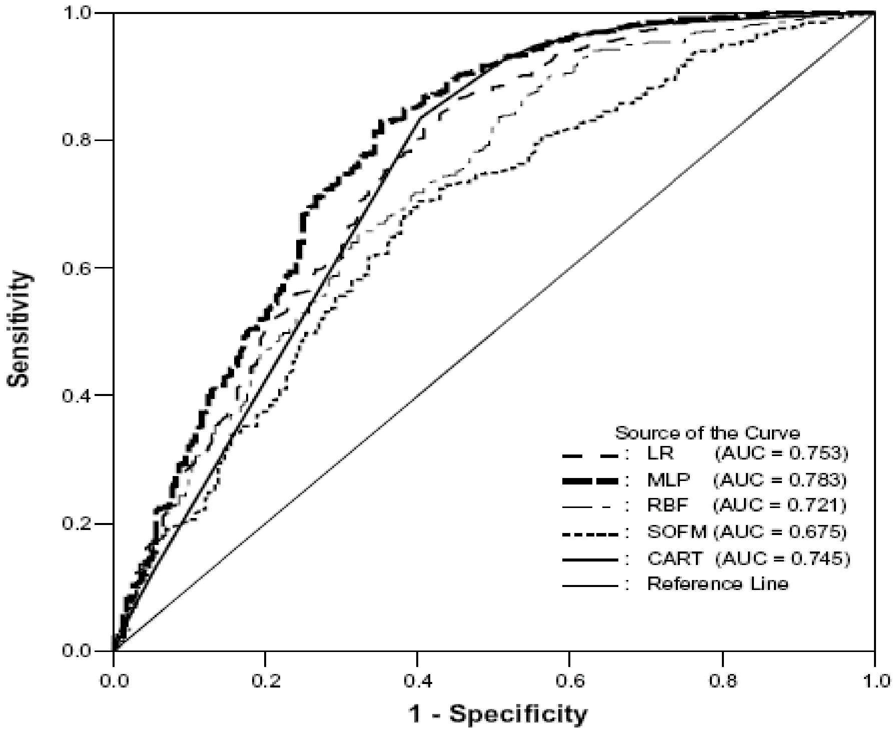


Figure 4.12 the ROC curves for five classification techniques performed similarly to predict coronary artery disease, adopted from [Kurt *et al*., 2008].

Likewise, classification and regression tree (CART) performed as well as other techniques like multilayer perceptrons, and logistic regression to predict cardiovascular risk where there were n-categorical variables transformed for the model into n-1 binary variables. Results (figure 4.13) show that the applied algorithms performed almost identically since the areas under the curve were 0.78, 0.78, and 0.76 with 95% confidence interval for logistic regression, MLP, and CART respectively. This similar performance from these classifiers may be because binary variables which are compatible with decision trees algorithms [Colombet *et al*., 2000].

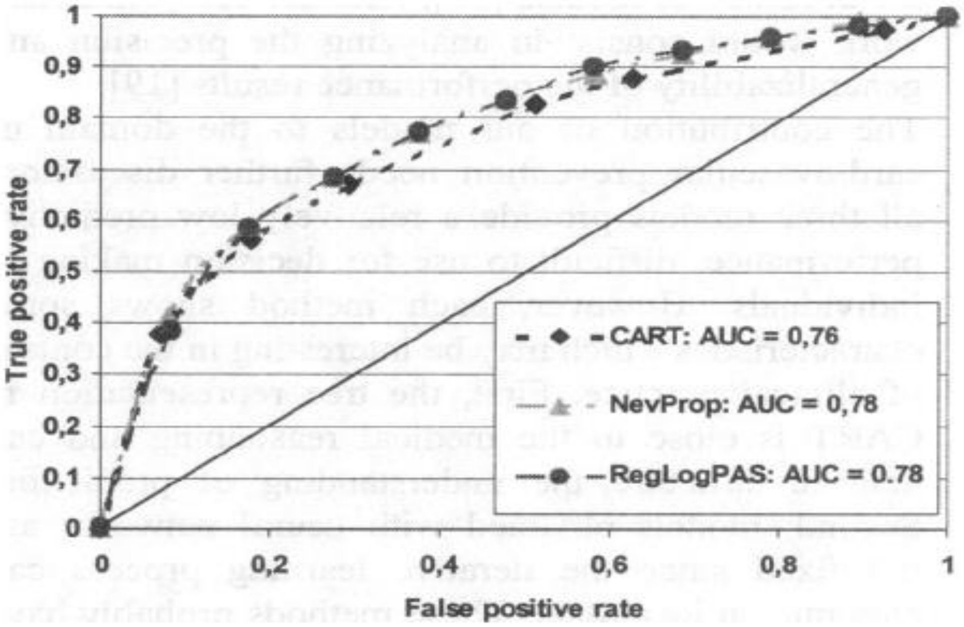


Figure 4.13 Roc Curve comparisons for CART decision tree, MLP neural network and logistic regression to predict cardiovascular risk, adopted from [Colombet *et al.*, 2000].

We have mentioned some examples related to decision tree performance in which they have behaved with different level of success. They were either worse or similar to other algorithms because of reasons such as data and available variables characteristics and number of outcome classes. However, there are several studies that report decision trees outperformance; having investigated the condition of applied parameters in this research may shed light on the reasons behind decision trees superiority.

In 1995, King and colleagues carried out a project called Statlog which was aimed at finding the answer to the question why certain algorithms perform better on particular datasets. Using twelve datasets related to image analysis, medicine, engineering and finance, several tentative conclusions were drawn which pointed out that finding the best algorithm for a specific dataset depends essentially on features of the dataset. In addition to the effect of data characteristics, on the other hand, algorithms also vary in terms of the ways that they deal with different types of data and the task in hand. Also, King and colleagues compared three categories of classifiers including symbolic, statistics and neural network learning techniques. In fact, StatLog aimed to find out why certain algorithms perform better on particular datasets. Symbolic learning mainly consists of decision tree algorithms like C4.5 and CART whereas there were Naïve Bayes, K-nearest neighbour, Bayesian networks and logistic regression in the statistical package. The last category was made up of back propagation and radial basis function algorithms. All aforementioned algorithms were tested on twelve large real-world datasets including five from image analysis, three from medication and two each from engineering and finance. Table 4.18 shows the obtained result from all available datasets using five classifiers as well as a very short explanation about the applied data features. Comparing the produced results for testing accuracy of the chosen algorithms, C4.5 outperformed the other four algorithms for six datasets with a high level of accuracy (more than 94%) in three cases. As stated before, in this study seventeen features have been employed and compared the result obtained by those algorithms, C4.5 or other symbolic learning algorithm also did very well for three of datasets including shuttle, segment and credit in terms of accuracy with 99.96, 96.0, and 94.3 percentage of accuracy respectively [King *et al*., 2005].

Table 4.18 Brief definition of 12 real world data collection applied in Statlog project and the obtained testing accuracy for various algorithms, gathered from [King *et al*., 1995]

Data collection	Train No.	Test No.	Attribute No.	Classes	Skew	Kurtosis	Back propagation	RBF	Logistic Regression	C4.5	Bayesian network
Handwritten digit	9000	9000	16	10	0.856	5.1256	92.0	91.7	91.4	85.1	76.7
Karhunen Loere digits	9000	9000	40	10	0.18	2.92	95.1	94.5	94.9	82.0	77.7
Vehicle Silhoutte	752	94	18	4	0.828	5.180	79.3	69.3	80.9	73.4	44.2
Satellite image	4435	2000	36	6	0.731	4.1737	86.1	87.9	83.1	84.9	69.3
Segment data	2079	231	11	7	2.95	24.48	----	----	89.1	96.0	73.5
Credit risk data	6230	2670	16	2	1.208	4.404	88.2	87.5	87.3	94.3	85.0
Belgian data	1250	1250	21	2	0.433	2.6581	----	----	99.3	95.5	93.8
Shuttle control data	43500	14500	97	7	4.43	160.310	95.10	----	96.17	99.96	95.45
Diabetes data	704	64	8	2	1.058	5.827	----	----	77.73	73.05	93.8
Heart disease	240	30	13	2	0.956	3.649	0.57	0.78	0.39	0.78	0.37
Head injury	800	100	6	3	1.007	5.0408	21.5	63.1	18.0	82.0	25.0
German	900	100	20	2	1.698	7.794	----	----	0.49	0.64	0.60

For segment data, the values of skewness (2.9580) and kurtosis (24.4813) show that the attributes are much further from normality compared with a dataset like the Vehicle dataset; here, logistic regression outperformed the other algorithms. This high level of skewness and kurtosis emerged to outfit symbolic learning well and C4.5 was one of the top seven algorithms.

Using the credit risk dataset with many irrelevant attributes and eight categorical variables, most symbolic algorithms including C4.5 with 94.3% accuracy outperformed through producing the most simple and efficient rules. Neural networks achieved close level of accuracy with less comprehensibility results. Statistical discriminants failed to discover the simple rules even with low level accuracy.

Furthermore, the Shuttle control dataset with 58,000 examples and numerical attributes was again far from normal distribution since the values of skewness and kurtosis were 4.43 and 160.31 respectively. Symbolic algorithms generally achieved accuracy close to 100% on the test set. Having analysed the developed decision tree of this dataset, it was revealed that only a few attributes are required to classify the data; for instance, a tree with five leaves applying two attributes would perfectly classify two classes indicating that attributes selected in this dataset are very well suited to the symbolic algorithms. It seems that decision trees are capable of finding the perfect rule with some tuning of the pruning parameter, but it is not the same in every situation since it is occasionally necessary to override the splitting criterion according to [Gordon & Olshen, 1978]. This study has emphasized that the best algorithm performance depends significantly on the investigated dataset. Therefore, it suggests a set of dataset descriptors to help to decide which algorithm is well suited to specific datasets. As an example, datasets with extreme distributions (skew> 1 and kurtosis> 7) and with many binary/categorical attributes (>38%) tend to favour symbolic learning algorithms like CART and C4.5 decision trees [King *et al*., 1995].

Another study [Delen *et al*., 2005] has suggested that data cleansing and preparation strategies may have an effect on the decision tree`s accuracy. The SEER Breast cancer dataset related to the breast cancer cases of 1973-2000 was applied and after a long process of cleaning it consisted of 202,932 records and 16 variables associated to socio-demographic and cancer specific information concerning cancer occurrence. The generated results showed the C5 decision tree as the best predictor with 93.6% accuracy on the holdout sample, a best ever result of research conducted in this area. Artificial neural networks and logistic regression placed in the second and third positions with 91.2% and 89.2% respectively. It has been reported that medical databases may consist of a large amount of heterogeneous data, which complicates the use of classifiers tools and techniques. Additionally, in the case of large databases, missing values in the database must be tackled, prior to the use of the decision tree. Further, medical databases may contain data that is redundant, incomplete, imprecise or inconsistent with noise related to data collection affecting decision trees application. All of the above may create problems for learning by decision trees or even other classification algorithms. That is, the results of decision trees application are directly affected by the quantity and quality of the data and through improving the collection of the data, decision trees can yield even greater results and benefits. Here we discuss our results concerning decision tree performance based on above mentioned critical analysis.

Due to the fact that our available dataset is a large volume of data with multidimensional structure [6,450*18], normally it is expected that SVMs, neural networks and decision tree outperform others [Kotsiantis, 2007]. However, the dataset is mainly composed of fourteen discrete variables and three continuous attributions (age, weight, and length of disease); in this case, the decision tree has produced the most promising result due to its double ability to tackle both continuous and discrete/categorical features which is superior to other aforementioned techniques that are good at handling only continuous variables. Thus, not having any images in the available dataset has caused DT outperformance.

BNs perform well when the input variables are conditionally independent of the class, as was true for two medical datasets used by King *et al*. (1995). In the case of emerging relationship among attributions, they don't perform well to manage learning properly. Tables 4.7 to 4.10 present the significant relationship between seventeen predictors and outcome class which may cause the weak performance of Bayesian networks. Discriminant algorithms like logistic regression also fail on this type of data with high correlation between the attributes [King *et al*., 1995]. In this study, there are many correlations among variables, like weight and nationality (r = -0.052, P<0.001), LBW and Sex (r=-0.047, P<0.001), Imprisonment and Sex (r = -0.156, P<0.001), prison and weight (r=0.065, P<0.001), length and nationality(r =0.099, P<0.001). Those correlations in addition to applying fourteen discrete inputs might cause weaker results from BN and LR rather than DT.

As mentioned earlier, Delen *et al*. (2005) emphasized that removing missing values, outliers and irrelevant and redundant features has a positive effect on decision trees accuracy and a decision tree is more capable of handling redundant features than BN. Also, DT is good at coping with irrelevant data. This might be the case in this particular study since here there are some variables with very low correlation coefficient that decision tree has not used them very much to build the model and not at all as the main root nodes. For example, as shown in Table 4.8 and 4.9, area (-0.027), prison (-0.026), diabetes (0.029) have low correlation coefficient, where these values for recent TB infection (rtbinf), imprisonment, IV drug using, sex are 0.25, 0.151, 0.172, 0.16 respectively. Some of the defined parameters by decision tree have been presented in Figure 4.14; it is the first page of constructed decision tree which is developed in this part of study. As can be seen the variables with high correlation coefficient such as recent TB infection (rtbinf), length, imprisonment and treatment category (Tcat) have played a major role as root and main nodes whereas the variables with small correlation coefficient have been recognised as less important factors and placed as very sub-nodes close to leaves which can even be pruned. Decision tree`s ability to utilize significant input factors on the basis of their degree of contribution to estimate outcome of tuberculosis treatment course (presented in Tables 4.7 to 4.10) create a better predictive model than classifiers such as MLP, RBF and SVM which use every input uniformly by weightening which affects the results transparency [Kukar *et al.*, 1992].

In all built decision trees in this study, either in this chapter or in chapter 6, length has been chosen as a root node which is, in fact, a very determining attribute if we consider it from medical point of view. Interestingly decision trees initiate partitioning the dataset at root node as either length <=6.07 or length > 6.07, almost shown in figure 4.14. In the process of DOTS therapy, firstly, patient must take four antibiotics (isoniazid, rifampicin, pyrazinmide, and ethambutol) for two months and then carry on with only isoniazid and rifampicin for further four-months. This is called six month DOTS therapy which is currently the core treatment plan carried out throughout the world. This schedule is for new TB cases; however, TB patients who have been already diagnosed should take the 8-months programme. In other words, if patient’s length of disease is 6 month or less s/he is a new case whereas therapy duration of more than 6 months implies a relapse TB case who has possibly failed or quit. Hence, dividing patients initially as either under or over six months by a decision tree is a meaningful partition from a medical point of view. Clearly, this highlights DT ability to divide cases in the meaningful partition and find the underlining relationships among input-output space intelligently.

As explained before, King *et al.* (1995) reported that the highest values for Kurtosis (>7) and skew (>1) denote that they are the furthest from normality and decision tree like other symbolic methods are basically nonparametric; that is, this methods do not make any assumption about the underlying distributions. Therefore, they tackle robustly distributions with large kurtosis and skew. In this research`s dataset, the average values of skew and kurtosis are 2.169 and 7.469 respectively; this is shown in Tables 4.3 to 4.5 in more detail for each variable distinctly leading to significant skewness and kurtosis ($P<0.05$) and non-normal distribution. Hence, the only available nonparametric symbolic learning algorithm in the current study is the decision tree which performed well at partitioning the input space. In actuality, High skew (>1) or kurtosis (>7) along with the presence of binary/categorical variables, using relevant and correlated predictors without any missed instances or noised data have decision trees to predict more accurately than other algorithms.

```
Len <= 6.07
|  Imprisonment = No
|  |  Rtbinf = No
|  |  |  AGE <= 52
|  |  |  |  LBW = No
|  |  |  |  |  AGE <= 20
|  |  |  |  |  |  Weig <= 53.5
|  |  |  |  |  |  |  Nat = Iran
|  |  |  |  |  |  |  |  Weig <= 42.5
|  |  |  |  |  |  |  |  |  Len <= 1.9: cured (2.0)
|  |  |  |  |  |  |  |  |  Len > 1.9: dead (9.0)
|  |  |  |  |  |  |  |  Weig > 42.5: cured (8.0/1.0)
|  |  |  |  |  |  |  Nat = Afghani
|  |  |  |  |  |  |  |  Len <= 4.73: cured (8.0/2.0)
|  |  |  |  |  |  |  |  Len > 4.73: quit (5.0/1.0)
|  |  |  |  |  |  |  Nat = CnAs: cured (1.0)
|  |  |  |  |  |  |  Nat = iraq: cured (0.0)
|  |  |  |  |  |  |  Nat = pakist: cured (0.0)
|  |  |  |  |  |  Weig > 53.5
|  |  |  |  |  |  |  AGE <= 19: cured (3.0)
|  |  |  |  |  |  |  AGE > 19: complet (5.0/1.0)
|  |  |  |  |  AGE > 20
|  |  |  |  |  |  HIV = No
|  |  |  |  |  |  |  Ivdrg = No
|  |  |  |  |  |  |  |  Castp = new
|  |  |  |  |  |  |  |  |  Area = rural
|  |  |  |  |  |  |  |  |  |  Nat = Iran: dead (11.0/1.0)
|  |  |  |  |  |  |  |  |  |  Nat = afghani: quit (11.0)
|  |  |  |  |  |  |  |  |  |  Nat = CnAs: quit (0.0)
|  |  |  |  |  |  |  |  |  |  Nat = iraq: quit (0.0)
|  |  |  |  |  |  |  |  |  |  Nat = pakist: dead (1.0)
|  |  |  |  |  |  |  |  |  Area = urban
|  |  |  |  |  |  |  |  |  |  Pris = No
|  |  |  |  |  |  |  |  |  |  |  Nat = Iran
|  |  |  |  |  |  |  |  |  |  |  |  SEX = female
|  |  |  |  |  |  |  |  |  |  |  |  |  Weig <= 36: dead (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Weig > 36
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Len <= 2.2
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  AGE <= 27: cured (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  AGE > 27: quit (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Len > 2.2: quit (11.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  SEX = male
|  |  |  |  |  |  |  |  |  |  |  |  |  AGE <= 37
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Weig <= 52
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Len <= 4.23
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Weig <= 45.5: cured (2.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Weig > 45.5: dead (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Len > 4.23: quit (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Weig > 52: quit (9.0)
```

Figure 4.14 First page of developed decision tree with root node length <= 6.07 as right side, *len* in the first line means length of disease.

Apart from DT, regarding to the rank of other employed classifiers which are represented in figures 4.5 and 4.6, BN has outperformed other four classifiers. LR, MLP, RBF AND SVM have performed relatively similar with prediction accuracy percentage ranging from 53.74% to 57.82%. Tu (1996) reviewed a number of studies comparing neural networks and logistic regression performance concluding that regression models usually have close predictive performance in testing datasets; it is the case here that LR and MLP performed very similarly with identical prediction accuracy (57.82%). RBF is actually a type of neural network and it might be a postulation that based on their algorithm similarities and data type entity the results are comparable.

Logic-based algorithms are considered very easy to understand and their comprehensibility of structure has received much attention, whereas neural networks and SVMs have very extremely poor interpretability and remained a ‘black box’ model [Lim *et al.*, 2000; Tsien *et al.*, 1998; Kotsiantis, 2007]. In fact, the main objective of this study is creating a system for which the user may vary from a physician to a health practitioner with a lower level of medical knowledge playing the role of TB patient’s supervision and support. Hence, a decision tree with a flowchart-type structure is superior to other methods such as logistic regression, neural networks, and support vector machines which are less likely to be acceptable in general use based on their less understandable results. Produced results of decision trees can be simply interpretable and applicable; their rules can be understood either by doctors or health practitioners who implement DOTS in rural areas and make decisions alone; this suggests preference of decision trees even if performance is not as high as other methods. Even Bayesian network application in practice, needs some knowledge like probabilistic explanations in diagnosing disease and their related factors which may not be available through the health staff at low organizational level. For example in Iran, Behvarzes are health workers who have limited medical knowledge but have a core role in health promotion and maintenance; based on my experience and knowledge, they can easily handle decision tree-based system application.

4.5 Summary

An available large set of data related to tuberculosis patients has created an excellent opportunity to generate predictive models which may lead to a system capable of defining which specific patient needs what level of supervision and support since it is not possible to give every single one of them full supervision. Covariance analysis for seventeen applied variables reveals that they are influential factors impacting outcome of tuberculosis treatment course. Based on the fact that there is no single learning algorithm which can uniformly outperform other algorithms over all datasets, six classifiers from different algorithm categories have been chosen. Using those factors and 6,450 records of TB patients to be tested by six classifiers, decision trees have outperformed the other methods in estimating the outcome of tuberculosis treatment course shown by five prediction comparison criteria. This outperformance is mainly attributed to the data characteristics. For example, the DT has tackled its non-normal distribution with high value of skew or kurtosis, possessing fourteen discrete/categorical variables with some relationship among their variables and different level of correlation coefficient. However, the degree of accuracy of developed models may need improvement since the most accurate tool has performed with only 74.21%. It seems that this level of accuracy is still not satisfactory and needs to be improved by a clustering-based combination algorithm which is discussed in the next chapter.

Chapter 5

Combined Use of Supervised & Unsupervised Learning for Tuberculosis Treatment Course Prediction

5.1 Introduction

In this chapter in order to generate a more precise and accurate system, it is suggested that we combine classification techniques as supervised and clustering methods as unsupervised learning algorithms. Supervised and unsupervised learning combination has already been used in a few studies that have been conducted to estimate a given outcome [Pao & Sobajic, 1992; Boudour & Hellal, 2005], to address cause-and-effect relationship [Šmuc *et al*. 2001] other than accuracy improvement. As illustrated in chapter 4, fitting the whole dataset to develop one model has led to complication and inaccuracy. In the best case scenario, a decision tree has been found as the most accurate algorithm to deal with the available dataset in respect of prediction accuracy of 74.21%. It might be because of the fact that mapping a big volume of patient data with various conditions at once to fit a model may create a problem. Here, applying the "divide and concur" concept might be useful. Based on this idea through recursively breaking down, a problem is divided into two or more sub-problems of the same or associated categories until the problem becomes straightforward enough to be solved directly. Afterwards, the obtained solutions to sub-problems are combined to give the main solution to the original problem. Thus, this concept may generate the idea of combination approach of supervised and unsupervised learning. This allows segmentation of the different patients/conditions and should then find the right (*i.e.*, simpler and more accurate) classification model for each segment of the patients/conditions.

The K-means clustering algorithm is able to divide the whole dataset into segments based on the number of clusters, *K*, defined initially. Here we have chosen *K* = 2, 3 and 4; since every developed cluster should be used to fit a classification model through a classification algorithm, the high number of *K* creates further complexity which is out of this research feasibility. For example, for *K*=4, there would be 4 models mapped by 4 related clusters (training sets) in which they should be tested by 4 testing sets for each of six classifiers. In this research, by defining *K*=2, 3, and 4 there are 9 models fitted by 9 developed clusters for every classification algorithm. In total, for six classifiers we have got 36 models which need to be validated by their associated testing set. Hence, we have started from the simplest number which is two. To find out the best performing cluster number an optimization methodology is suggested.

Through combining unsupervised learning such as the k-means clustering approach with six applied classifiers including decision trees (DT), Bayesian Network (BN), logistic regression (LR), multilayer perceptrons (MLP), radial basis functions (RBF), and support vector machines (SVM), this chapter aims to improve the prediction accuracy percentage. Details of classification estimation from each of the above mentioned classification tools and their comparison by various measurements was illustrated in chapter 4. This chapter is aimed at improving these measurements through k-means clustering application prior to the classification process in order to decrease the number of misclassified cases. Furthermore, this part of the study is aimed at evaluating the effect of supervised and unsupervised learning combination on the accuracy of the six developed models to predict the outcome of a tuberculosis treatment course. This aim can be considered in more detail as follows:

- Determine which of the examined cluster number is most optimized for the given classification task. Here we have examined 2, 3, and 4 partitions
- Which classification algorithm performs best with cluster-based input-output mapping
- How effectively has combined unsupervised and supervised learning algorithm performed to improve the prediction accuracy

In order to respond to these questions, firstly the experimental methodology for this part of the research is explained. Secondly, the obtained findings are presented and discussed, followed a summary conclusion of the chapter.

5.2 Experimental Methodology

The available dataset which was initially applied to estimate the outcome of a tuberculosis treatment course by six classification algorithms was used to assess their prediction and training accuracy via clustering method; the steps of approach are illustrated in the next three sections in more detail.

5.2.1 Unsupervised Learning by Clustering Method

It is denoted in tables 4.1 and 4.2 that there are 17 variables as independent inputs and the outcome of a tuberculosis treatment course as a dependent variable with five classes. Let us represent these seventeen input variables as $\{x = x_1, x_2, \dots, x_n\}$, $n = 17$, and the correspondent target outputs of the outcome of tuberculosis treatment course as $\{r = r_1, r_2, \dots, r_n\}, n = 5$ when

$$X = \{x^t, r^t\}_{t=1}^{N}$$

Where *t* indexes different examples in the dataset where here the whole dataset is 6,450; however, based on the fact that the dataset was divided with two-thirds for training and the other third for estimating performance, we will have two datasets including *R* and *T* denoting training and testing datasets respectively as follows:

$$R = \{x^t, r^t\}_{t=1}^{N}, N = 4{,}515$$

$$T = \{x^t, r^t\}_{t=1}^{N}, N = 1{,}935$$

Where *t* represent the pair of numbers of an input x^t and the corresponding target output r^t; *R* and *T* consist of 4,515 and 1,935 pairs of examples for the training and testing set respectively. In order to apply a clustering learning algorithm for both training and testing set, r^t is removed from the dataset at the beginning of clustering learning.

Because of the partitioning method capacity to handle a large volume of data compared with hierarchical clustering method and the large body of dataset available in this study the k-means clustering method has been examined here. The K-means clustering method is a centroid-based technique employed to group a dataset into K partitions ($K = \{2, 3, 4\}$) through the following stages:

First, it randomly chooses K of the object; every one of which initially represents a cluster mean or centre for every remaining objects, an object is assigned to the cluster based on its similarity with one of the embedded clusters. Actually, the distance between the object and the cluster mean defines whether this object should be added to this cluster or not. That is, every object is distributed to a cluster on the basis of the cluster centres whichever is nearest. This distribution forms silhouettes. Then, the cluster centres are updated *i.e* the mean value of each cluster is calculated according to the current objects in the cluster. Again, by considering the new cluster centres, the objects are redistributed to the cluster according to which cluster is the nearest leading to new silhouettes. This iteration process is carried on until the time that no iterative relocation would occur to any further extent and consequently no redistribution of the objects in any cluster will occur and so the process terminates. Here, $K = \{2, 3, 4\}$, the iteration process is carried out 10 times and when the same index for a given object was yielded repeatedly, those indexes determined to which cluster the object belongs. Related silhouettes for each $K = \{2, 3, 4\}$, have been represented in figure 5.2, 5.3, and 5.4 respectively.

Now, the training and testing datasets was divided into K clusters separately in the MATLAB environment. After adding the target output r^t for each cluster we have 9 cluster-based training sets and 9 cluster-based testing sets where $K = \{2, 3, 4\}$. We denote training sets as R_i^k and testing sets as T_i^k where *i* is the i^{th} cluster-based training or testing set and *K* is the number of partitions produced by *K*-means clustering varying from 2 to 4 in this study. Table 5.1 shows the number of cases in each cluster based on the partition number *K* and training or testing datasets separately.

5.2.2 Supervised Learning by Classification Methods

Applying each R_i^k to train each considered classifier including decision trees (DT), bayesian networks (BN), logistic regression (LR), neural networks (NN), radial basis functions (RBF), and support vector machines (SVM), the correspondent models are built distinctly.

To learn these datasets, the WEKA package, illustrated in Figure 3.5 is utilized. For every partition number *K*, we have a corresponding number of constructed models named M_i^k ; indicating the i^{th} constructed model trained by the i^{th} cluster-based training set and *K* is the number of partitions constructed by the K-means clustering approach $K = \{2, 3, 4\}$.

5.2.3 Combination & Comparison Stage

To check the validity and generalization ability of this mapping from R_i^k to M_i^k , every one of the developed models are checked by the corresponding testing data T_i^k to which they haven't been applied in model development. Now, by this application for every T_i^k, we are going to calculate y_i^k where y is the class label of outcome of a tuberculosis treatment course. It is defined by the corresponding model parameters, *i* is the index of patient records x^t in the testing set in every cluster *K*, partitioned by the *K*-means clustering method. Then, for every *K*, including 2, 3, and 4 partitions, the correspondents y_i^k are put together to make up the whole yielded y as the classification label for the whole testing set together. For example, for *K* =2, we have two series of y_i^k on which *i* in the first series comprises the calculated classification label of a tuberculosis treatment course for the 1^{th} to the 966^{th} patient and for the second series includes the second cluster from the 967^{th} to the $1{,}935^{th}$ cases. These series of y_i^k converged to compose y_i which are obtained based on both clustering and classification methods.

Having compared these produced y_i and the corresponding r^t for each x^t by using accuracy comparison measurement such as prediction accuracy, the impact of clustering and classification methods combination has been revealed. To calculate the prediction accuracy, confusion matrices are developed for y_i yielded from each partition number $K = \{2, 3, 4\}$.

The process of y_i calculation based on $K = \{2, 3, 4\}$ has also been conducted by using training set R_i^K leading to training accuracy calculation which shows the degree of model fitness. However, for judgment of a model, the importance of model accuracy addressed by a measurement like prediction accuracy is the subject of high interest.

At the final stage, yielded prediction and training accuracy for two, three, and four-clustered based models are compared; these results compared six classification algorithms to find out which outperforms others. The combination stage including confusion matrix construction and comparison process are carried out in WEKA and SPSS (statistical software) environment.

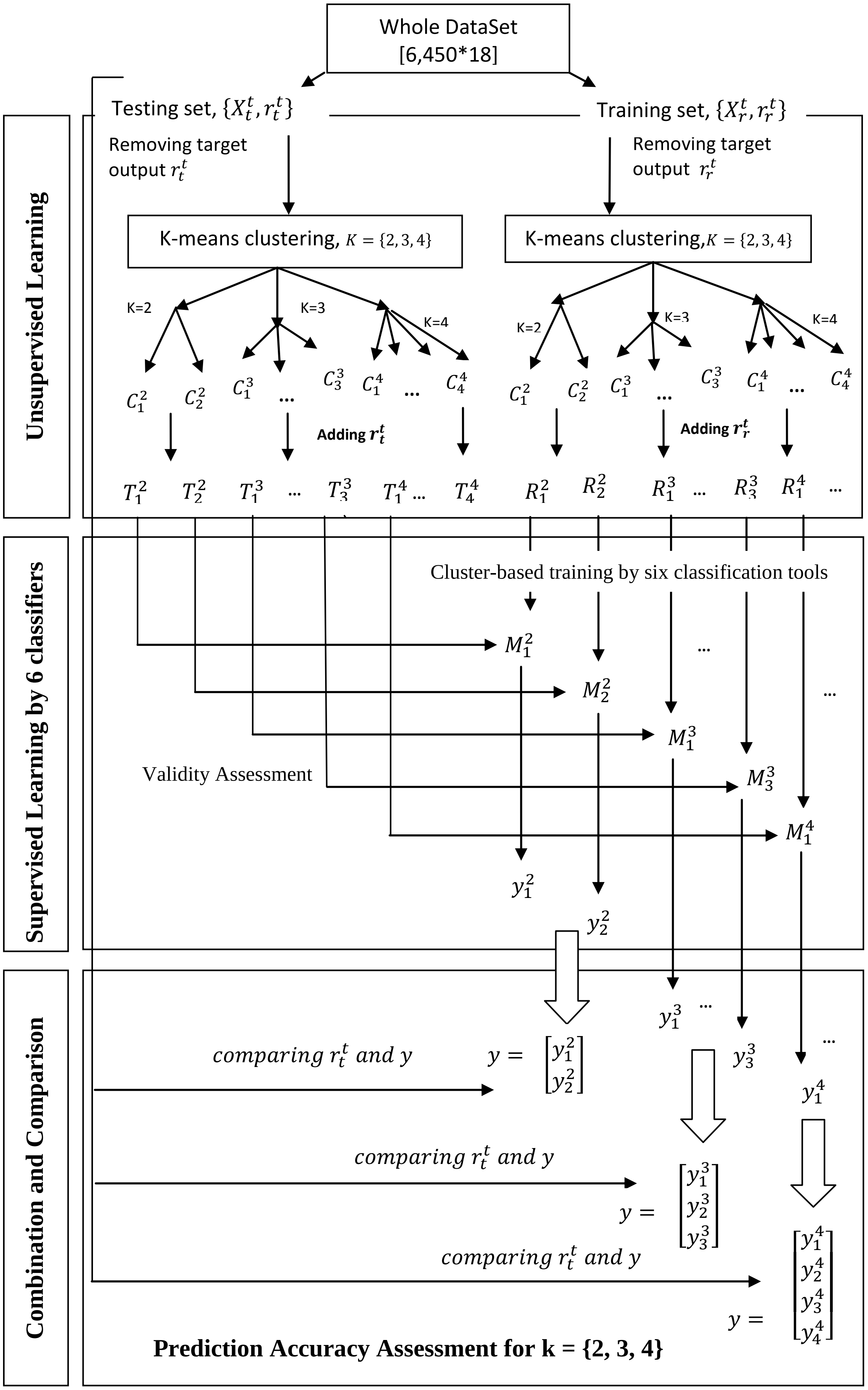

Whole DataSet
[6,450*18]
Testing set, $\{X_t^t, r_t^t\}$
Training set, $\{X_r^t, r_r^t\}$
Removing target output r_t^t
Removing target output r_r^t
K-means clustering, $K = \{2,3,4\}$
K-means clustering, $K = \{2,3,4\}$
K=2
K=3
K=4
K=2
K=3
K=4
C_1^2 C_2^2 C_1^3 ... C_3^3 C_1^4 ... C_4^4
C_1^2 C_2^2 C_1^3 ... C_3^3 C_1^4 ... C_4^4
Adding r_t^t
Adding r_r^t
T_1^2 T_2^2 T_1^3 ... T_3^3 T_1^4 ... T_4^4
R_1^2 R_2^2 R_1^3 ... R_3^3 R_1^4 ...
Unsupervised Learning
Supervised Learning by 6 classifiers
Combination and Comparison
Cluster-based training by six classification tools
M_1^2
M_2^2
M_1^3
M_3^3
M_1^4
...
...
Validity Assessment
y_1^2
y_2^2
y_1^3
...
y_3^3
...
y_1^4
comparing r_t^t and y
$y = \begin{bmatrix} y_1^2 \\ y_2^2 \end{bmatrix}$
comparing r_t^t and y
$y = \begin{bmatrix} y_1^3 \\ y_2^3 \\ y_3^3 \end{bmatrix}$
comparing r_t^t and y
$y = \begin{bmatrix} y_1^4 \\ y_2^4 \\ y_3^4 \\ y_4^4 \end{bmatrix}$
Prediction Accuracy Assessment for k = {2, 3, 4}

Figure 5.1 Schematic processes of supervised & unsupervised learning combination and evaluation.

5.3 Experimental Results

The results of this part of study can be categorized in two main sections including the first group of findings which are obtained from different numbers of cluster K, $K = \{2, 3, 4\}$ and the second category of results which are related to different classification algorithms comparison.

5.3.1 Silhouette Analysis

The returned silhouette for $K = \{2, 3, 4\}$ are displayed in Figure 5.2, 5.3, and 5.4 respectively. Having compared the drawn silhouette plots, the silhouette values related to three numbers of clusters (*K=3)* is slightly more well-separated than others; furthermore, in figure 5.4, clusters contain negative silhouette values indicating that those four clusters are not well separated. This is not the case for Figure 5.2 and 5.3 which don’t have any negative points.

5.3.2 Predictive Accuracy Comparison Based on Cluster Number

The number of objects in each cluster has been represented in Table 5.1 based on training and testing sets. Using the R_i^k as training and T_i^K as testing sets, nine models are built with the training and prediction accuracy shown in table 5.2, 5.3, 5.4, 5.5, 5.6, 5.7. Obviously, the models yielded from 3-cluster based datasets outperform others since the training accuracy values are 91.66, 84.88, and 90.67 for training accuracy of two, three, and four-cluster based models respectively. The percentages of prediction accuracy also verifies the 3-cluster performance with 80.26, 82.92, and 87.28 which are higher than other models, particularly those constructed by 4-cluster sets.

Table 5.1 Applying k-means clustering method to cluster the training and testing set after removing outcome parameter

Data	2-cluster		3-cluster			4-cluster			
	C1	C2	C1	C2	C3	C1	C2	C3	C4
Training Set	2255	2260	1560	1707	1248	1309	1227	940	1039
Total	4515		4515			4515			
	2-cluster		3-cluster			4-cluster			
Testing Set	C1	C2	C1	C2	C3	C1	C2	C3	C4
	966	969	669	732	534	561	526	403	445
Total	1935		1935			1935			

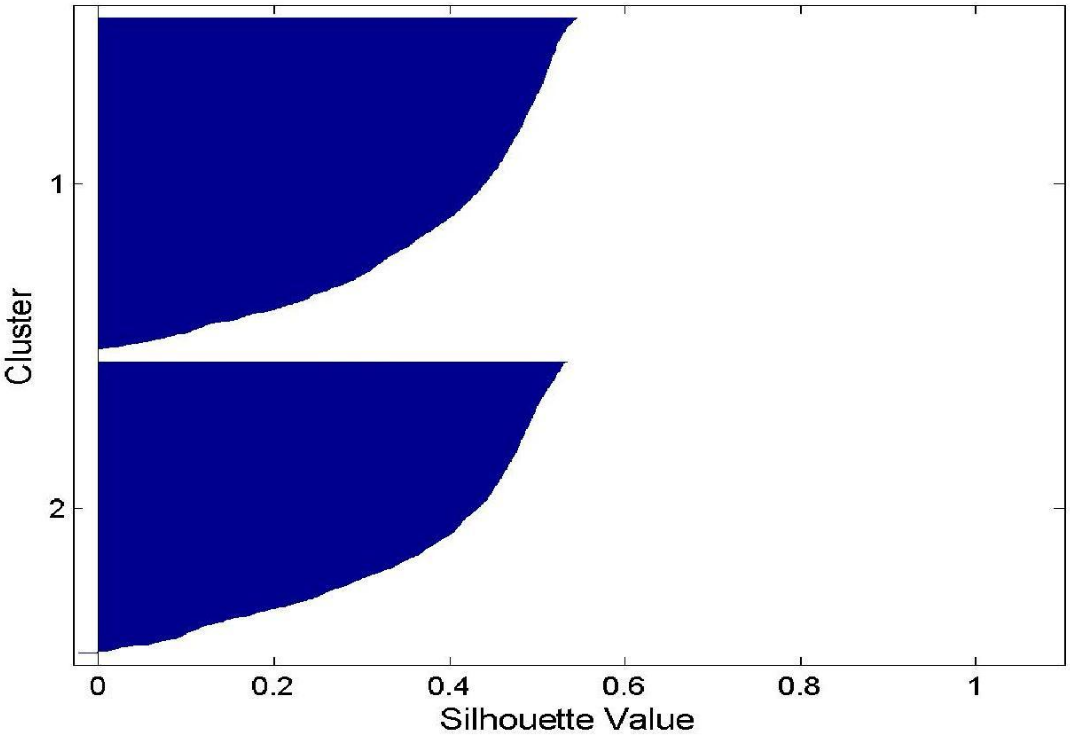


Figure 5.2 The silhouette plot for two partition number clustered by k-means.

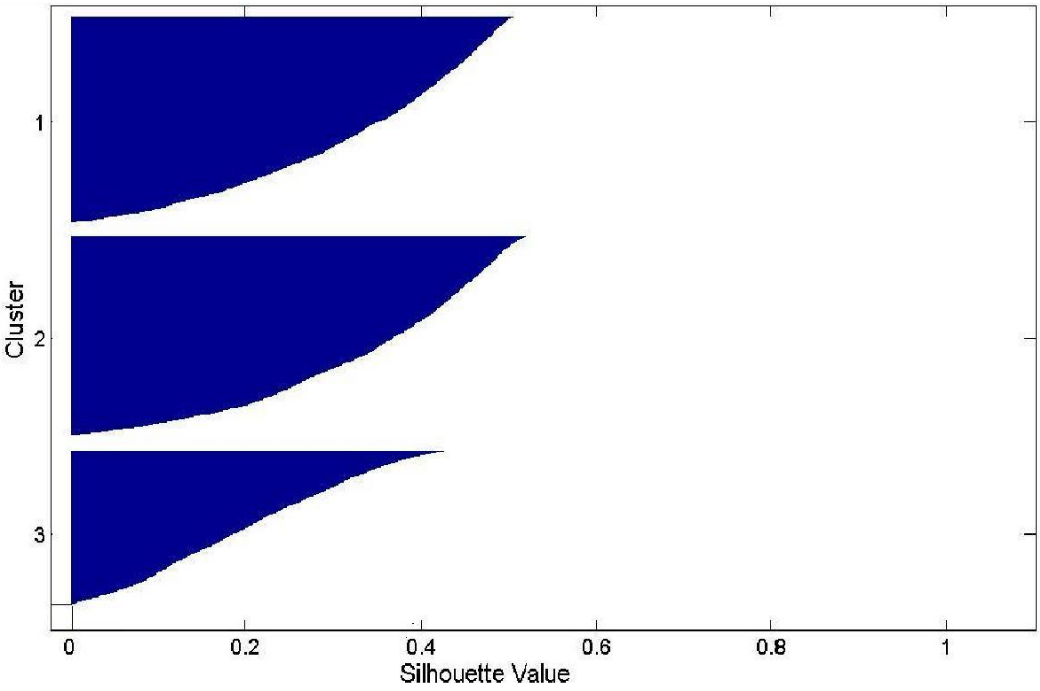


Figure 5.3 The silhouette plot for three partition number clustered by k-means.

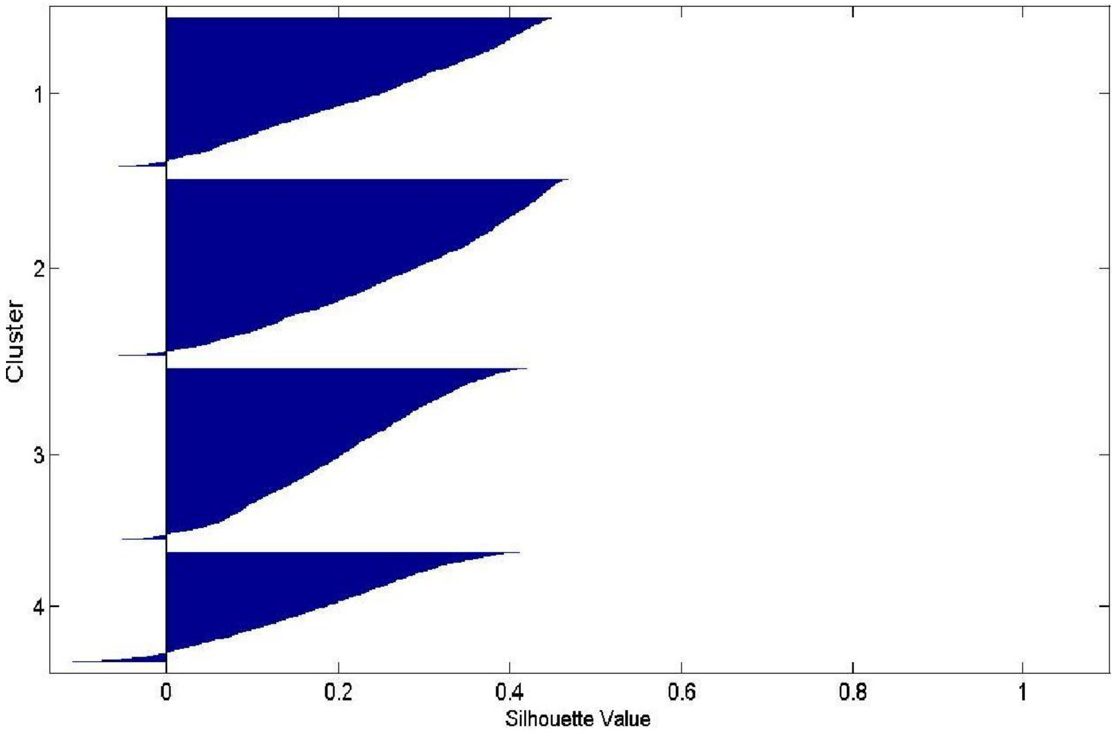


Figure 5.4 The silhouette plot for four partition number clustered by k-means.

Table 5.2 Comparison of prediction accuracy percentage for 2, 3, and 4 cluster-based Decision Trees

		Model Fitness	Model Accuracy
Number of Cluster		Training Accuracy	prediction Accuracy
2-cluster	Cluster1	84.70	80.74
	Cluster2	85.53	73.78
3-cluster	Cluster1	91.66	80.26
	Cluster2	84.88	82.92
	Cluster3	90.67	87.28
4-cluster	Cluster1	87.31	61.31
	Cluster2	85.33	53.61
	Cluster3	83.72	43.67
	Cluster4	80.4	49.88

Table 5.3 Comparison of prediction accuracy percentage for 2, 3, and 4 cluster-based Bayesian Network

		Model Fitness	Model Accuracy
Number of Cluster		Training Accuracy	prediction Accuracy
2-cluster	Cluster1	54.01	52.48
	Cluster2	56.37	56.86
3-cluster	Cluster1	63.71	55.00
	Cluster2	62.09	70.21
	Cluster3	66.47	64.67
4-cluster	Cluster1	59.28	52.76
	Cluster2	53.38	52.85
	Cluster3	56.91	49.87
	Cluster4	54.28	53.25

Table 5.4 Comparison of prediction accuracy percentage for 2, 3, and 4 cluster-based Logistic Regression

		Model Fitness	Model Accuracy
Number of Cluster		Training Accuracy	prediction Accuracy
2-cluster	Cluster1	52.5	55.79
	Cluster2	61.28	62.02
3-cluster	Cluster1	68.65	74.14
	Cluster2	56.18	61.33
	Cluster3	71.21	70.65
4-cluster	Cluster1	57.83	52.40
	Cluster2	60.14	56.32
	Cluster3	64.14	56.65
	Cluster4	54.76	52.58

Table 5.5 Comparison of prediction accuracy percentage for 2, 3, and 4 cluster-based Multilayer Perceptron Neural Network

		Model Fitness	Model Accuracy
Number of Cluster		Training Accuracy	prediction Accuracy
2-cluster	Cluster1	65.72	59.10
	Cluster2	72.78	60.78
3-cluster	Cluster1	76.73	63.22
	Cluster2	70.47	60.00
	Cluster3	78.26	85.23
4-cluster	Cluster1	72.65	52.58
	Cluster2	68.94	48.47
	Cluster3	77.34	43.42
	Cluster4	69.10	48.98

Table 5.6 Comparison of prediction accuracy percentage for 2, 3, and 4 cluster-based Radial Basis Function

		Model Fitness	Model Accuracy
Number of Cluster		Training Accuracy	prediction Accuracy
2-cluster	Cluster1	48.55	46.16
	Cluster2	55.88	55.72
3-cluster	Cluster1	63.52	54.55
	Cluster2	51.08	53.68
	Cluster3	56.93	53.27
4-cluster	Cluster1	56.91	43.13
	Cluster2	51.50	43.53
	Cluster3	58.61	40.94
	Cluster4	46.19	46.96

Table 5.7 Comparison of prediction accuracy percentage for 2, 3, and 4 cluster-based Support Vector Machine

		Model Fitness	Model Accuracy
Number of Cluster		Training Accuracy	prediction Accuracy
2-cluster	Cluster1	49.40	56.21
	Cluster2	60.97	62.64
3-cluster	Cluster1	69.93	67.86
	Cluster2	51.84	63.52
	Cluster3	68.00	74.20
4-cluster	Cluster1	56.22	56.22
	Cluster2	56.15	56.46
	Cluster3	61.48	53.10
	Cluster4	53.60	51.46

5.3.3 Predictive Accuracy Comparison Based on Classifiers

To assess how the six considered classifiers have worked in the combination of supervised and unsupervised learning, a confusion matrix is developed for each classifier and for every partition number separately. This step leads us to calculate model fitness and accuracy. Thus, there are 36 confusion matrices produced for six tools and three *K*; we have presented three matrices as an example for all three partition numbers: two, three, and four, executed for the decision tree. As shown in tables 5.8, 5.9, and 5.10 here the confusion matrix represents the number of cases which have been predicted correctly by the given classifier application (decision trees) and the outcome of tuberculosis treatment course in reality. The model fitness and accuracy for each classification algorithm application is shown in Figures 5.5, 5.6, 5.7, 5.8, 5.9, and 5.10 for two, three, and four partitioning number comparatively. The 3-cluster based models have been the best in all cases where, the model accuracy has been 80% for 3-cluster based model partitioning decision tree whereas this value has been 75% and 48% for two and four clusters respectively. This is the same for Bayesian networks, signified in Figure 5.6, where the model accuracy is 65.43 for $K=3$ which is greater than 60% and 57.3% for two and four clusters.

Likewise, as shown in figure 5.7 for logistic regression the prediction accuracy is calculated as 67.60% which is 15% and 18% more than the results for two and four clusters respectively.

Produced results by MLP confirm the 3-cluster outperformance when the prediction accuracy obtained from 3-cluster based model is 64.80 which is %4 and 6.5% for two and four cluster-based learning results. This is shown in figure 5.8.

For radial basis functions, 3-cluster based learning has given the best result for prediction accuracy (presented in Figure 5.9) with 55.80% compared with 49% and 43% for two and four cluster number respectively.

The last example of three-cluster base learning superiority with 63.11% rather than the partition number two with 56% and four with 50% has been obtained by support vector machine performance.

Comparisons among two, three, and four cluster-based learning results by six classification algorithms have confirmed that three-cluster is the best partition number; after 3 clusters, two and then four clusters have produced the best results respectively.

5.3.4 Prediction Accuracy Comparison Before and After Clustering

After applying combined clustering and classification method for six considered classification methods, there is the opportunity to compare prediction accuracy before and after this method application. Figures 5.11 and 5.12 demonstrate the prediction accuracy percentage and F-measure values for decision trees (DT), Bayesian net (BN), logistic regression (LR), multilayer perceptron (MLP), radial basis functions (RBF) and support vector machines (SVM) comparatively. The improvement in these two measurements is clear; where for DT, BN, LR, MLP, RBF, and SVM, the prediction accuracy improvement are reported as 7%, 5%, 10%, 7%, 3.5%, and 4.8% respectively. This improvement for all employed classifiers through combination method by F-measure values improvement is verified with improvement of 0.11, 0.08, 0.10, 0.11, 0.09, and 0.20 for DT, BN, LR, MLP, RBF and SVM respectively. Figure 5.12 also shows findings.

Table 5.8 Confusion matrix for Decision Tree, C4.5, model accuracy for whole 2 clusters, cluster1+cluster2

Predicted Outcome	Outcome in Reality					Total	Prediction Accuracy (%)
	Cured	Complete	Quit	Failed	Dead		
Cured	622	85	33	15	17	772	0.80
Complete	41	354	15	7	7	424	0.83
Quit	55	27	231	18	13	344	0.67
Failed	16	15	21	168	6	226	0.74
Dead	24	7	14	4	120	169	0.71
Total	758	488	314	212	163	1935	0.75

Table 5.9 Confusion matrix for Decision Tree, C4.5, model accuracy for whole 3 clusters, cluster1+cluster2+cluster3

Predicted Outcome	Outcome in Reality					Total	Prediction Accuracy (%)
	Cured	Complete	Quit	Failed	Dead		
Cured	695	59	19	10	16	799	0.869
Complete	37	368	7	8	1	421	0.874
Quit	37	9	275	20	7	348	0.79
Failed	11	6	12	149	5	183	0.81
Dead	18	6	28	8	124	184	0.67
Total	798	448	341	195	153	1935	0.80

Table 5.10 Confusion matrix for Decision Tree, C4.5, model accuracy for whole 4 clusters, cluster1+cluster2+cluster3+cluster4

Predicted Outcome	Outcome in Reality					Total	Prediction Accuracy (%)
	Cured	Complete	Quit	Failed	Dead		
Cured	394	82	42	26	26	570	0.69
Complete	211	352	29	24	9	625	0.563
Quit	93	46	107	42	31	319	0.33
Failed	43	34	44	88	10	219	0.40
Dead	50	12	35	21	84	202	0.41
Total	791	526	257	201	160	1935	0.48

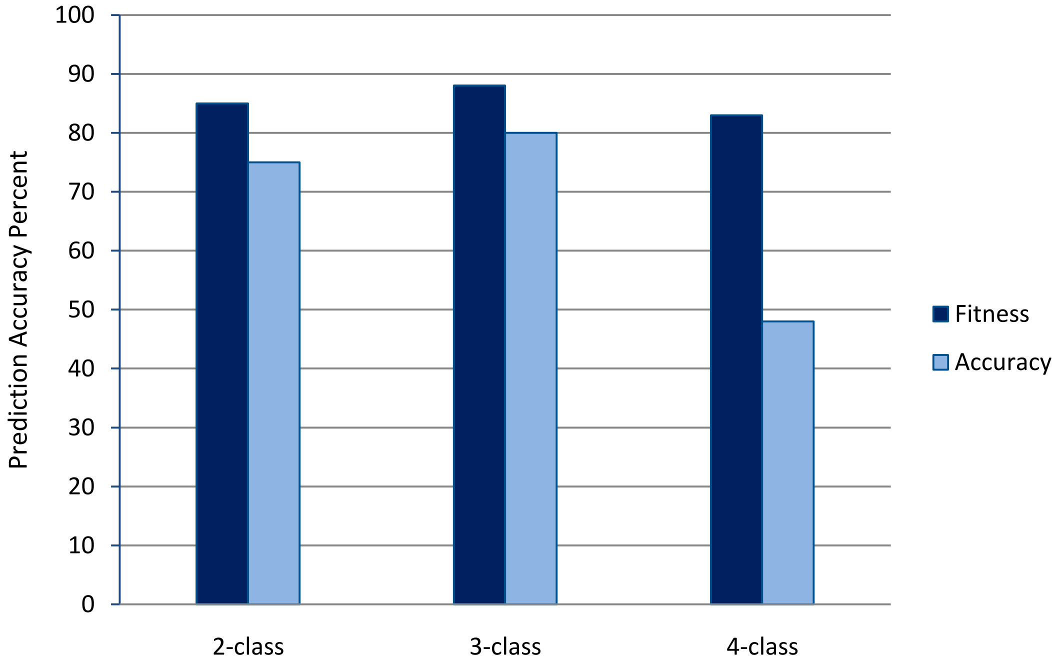


Figure 5.5 Decision Tree prediction accuracy percent (%).

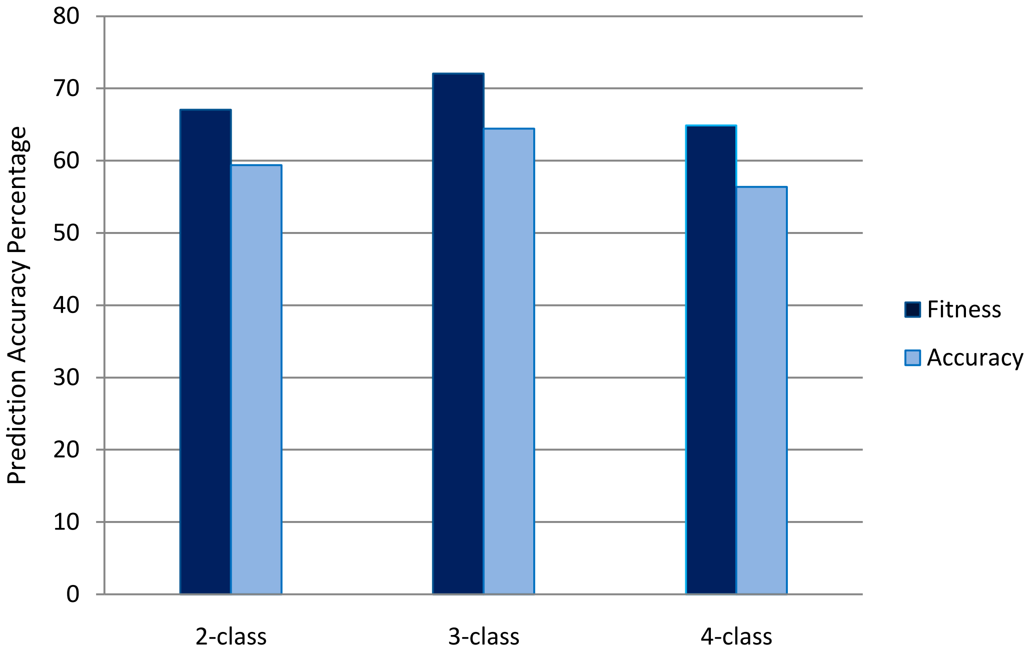


Figure 5.6 Bayesian Network prediction accuracy percent (%).

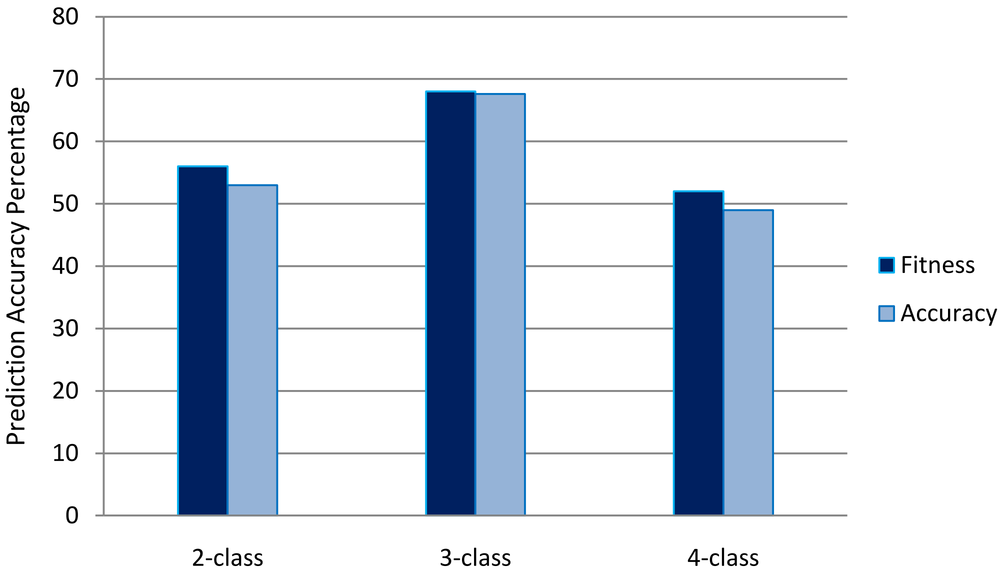


Figure 5.7 Logistic Regression prediction accuracy percent (%).

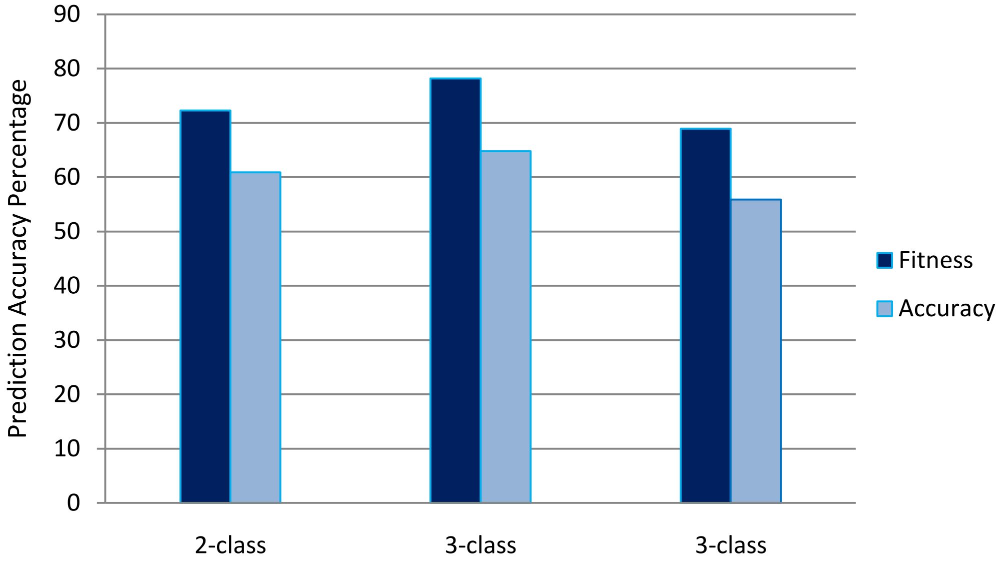


Figure 5.8 Multi Layer Perceptron (Neural Network) prediction accuracy percent (%).

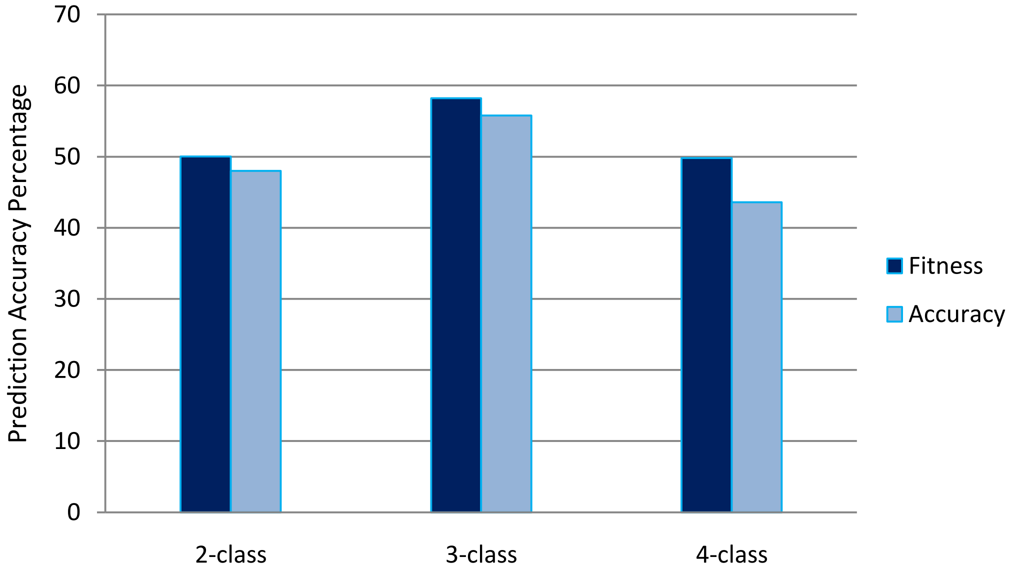


Figure 5.9 Radial Basis Network prediction accuracy percent (%).

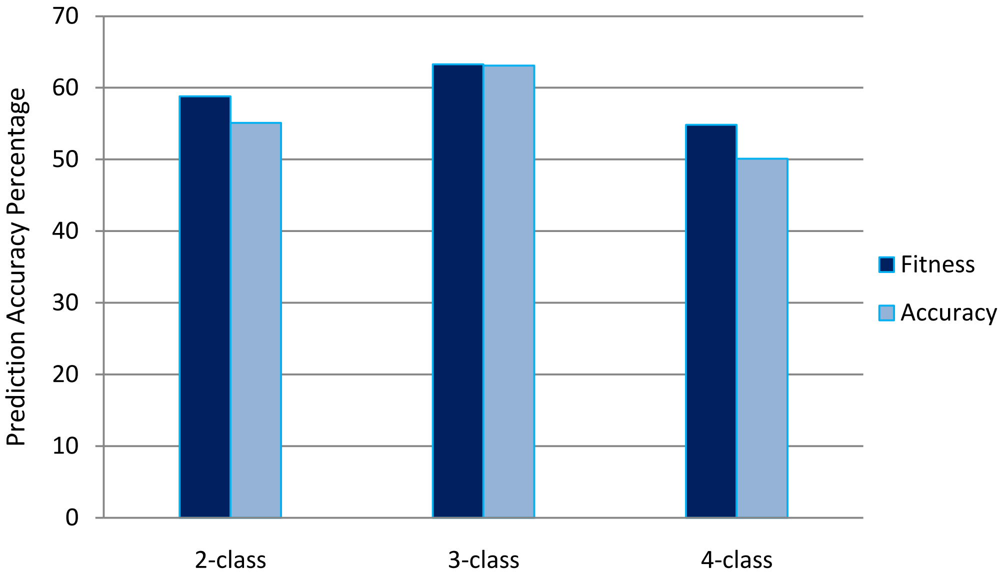


Figure 5.10 Support Vector Machine prediction accuracy percent (%).

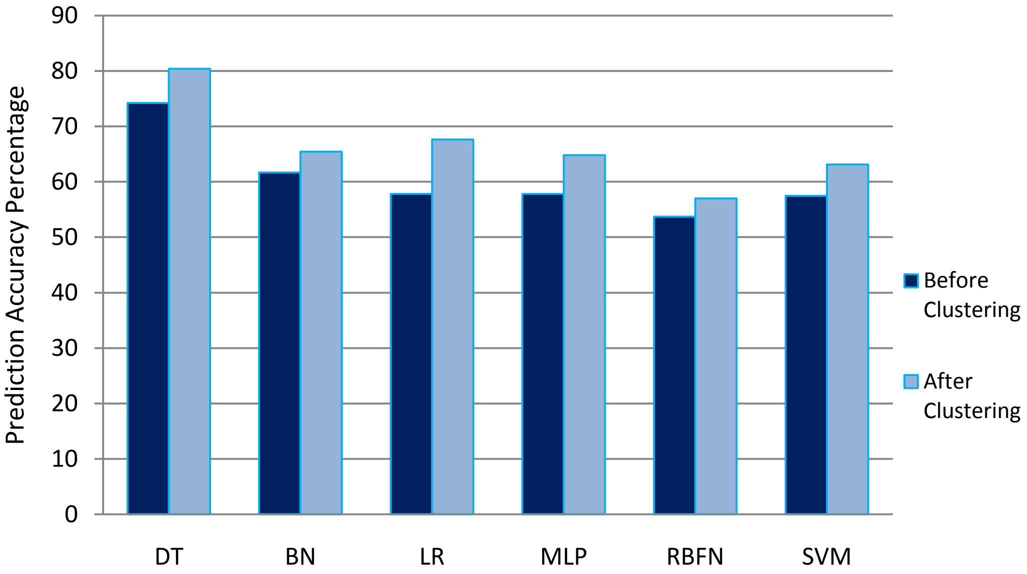


Figure 5.11 Comparison of six machine learning tools Prediction accuracy percentage for model accuracy before and after clustering.

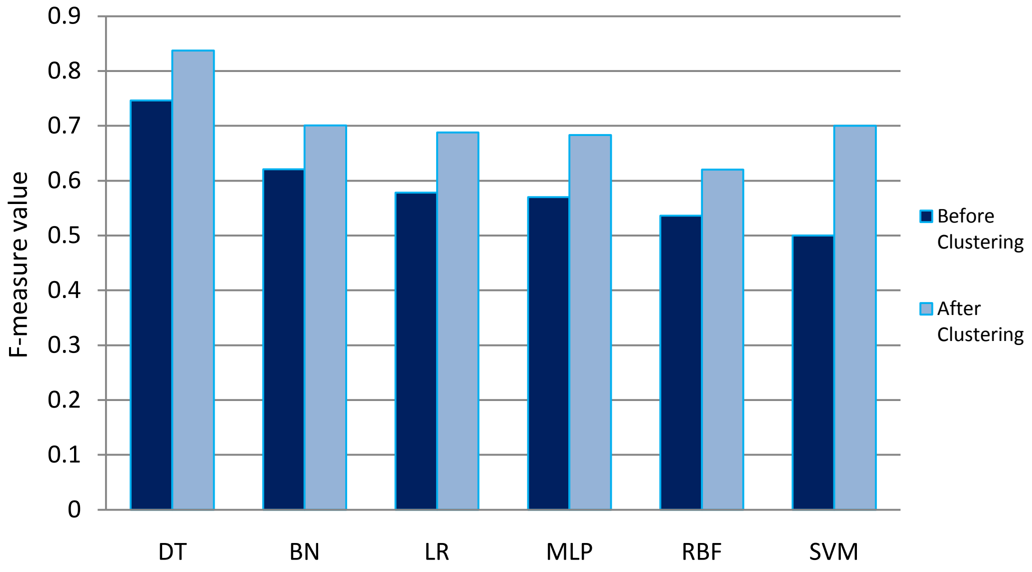


Figure 5.12 Comparison of six machine learning tools F-measure for model accuracy before and after clustering.

5.4 Discussion

Applying the integrated supervised and unsupervised learning method increased the perdition accuracy for six classifiers. Table 5.11 presents the prediction model accuracy and F-measure as two measurements of accuracy improvement for six classifiers. There are accuracy improvements (4-10%) for six classifiers.

Table 5.11 Summary of accuracy improvement for six classifiers presented by changes in Prediction Model Accuracy and F-measure

Classifiers	Prediction Model Accuracy		F-measure	
	Before Clustering	After Clustering	Before Clustering	After Clustering
DT	74.21	80.4	0.746	0.837
BN	61.7	65.43	0.621	0.701
LR	57.82	67.6	0.578	0.688
MLP	57.82	64.8	0.570	0.683
RBFN	53.74	57	0.536	0.620
SVM	57.47	63.116	0.500	0.700

The supervised and unsupervised combination approach has already been employed for other purposes; however, here we have examined this approach in a new application to improve prediction accuracy. The common point about the combined approach is that in all of them the unsupervised learning phase is conducted prior to supervised learning. It makes the supervised learning algorithms use clustered data to produce more accurate results. However, in this study which uses this method for the purpose of prediction improvement, the combination stage is carried out through merging the predicted outcome of each cluster from the testing set and putting it together to make up the whole predicted outcomes. Classification is basically a type of clustering which identifies similarities among inputs that belong to the same class. Furthermore, when the clustering algorithm exploits similarities between inputs in order to segment those similar inputs together, this might perform classification automatically [Marsland, 2009]. Using the k-means clustering algorithm to produce a well-segmented input space, based on the number of partitions, two, three, and four clusters are generated. Due to using an iterative technique in the k-means algorithm and moving the objects` location, there is new partitioning and topology. This changes the location of objects from one group to another. We intend to locate similar objects in the same cluster as ‘close’ as possible whereas objects of different clusters should be as ‘far apart’ as possible.

Having compared the silhouette values and their corresponding plots for different number of partitions (*K*=2, 3, and 4*)* in Section 5.3.2 and Figure 5.2, 5.3 and 5.4, it is obvious that *K*=3 has returned the most well-separated clusters with greater mean silhouette values and no negative silhouette values. Thus, we focus on the clusters where *K*=3. To describe each cluster, we calculate the mode for each variable in the boundary of every cluster. Here, mode is the most occurring values for each variable in each cluster`s boundary. We have investigated the mode of each attributes values before and after clustering. Table 5.11 presents the mode measurement which is the most frequent values of variables in the training set. Having compared the mode before and after clustering in the training set and R_1^3, R_2^3, R_3^3, it is revealed that by using clustering, proper segmentation has been conducted. Apparently, the majority of variables` modes have been altered before and after clustering due to the change in object locations which have been updated in the process of partitioning; this may result in developing groups of patients with new members since clustering aims to put similar cases in a cluster. According to the k-means clustering requirement that each object must belong to exactly one group, similar cases are placed in one cluster. Developed groups may increase the model`s accuracy since similar patients/condition might be placed in the same sector and mapping these consistence segments might lead to more accuracy and precision.

From Table 5.11, the values of mode in the training set (before clustering) are different from the mode values of each attribute in clusters. For example, the most frequent value for TB type is pulmonary before clustering; however, applying clustering makes change where the mode of TB type in R_2^3 is different as extra-pulmonary. That is, it has been changed after clustering into three splits where TB type is pulmonary for two of them and extra-pulmonary for another. It seems clustering has been strong enough to divide cases and put similar conditions together.

Table 5.12 The value of mode measurement for the variable of training sets before and after partitioning, *K=3*

Input factors	Before partitioning	After partitioning		
	The most frequent value (mode) of input factors in training set	The most frequent value (mode) of input factors in cluster R_1^3	The most frequent value (mode) of input factors in cluster R_2^3	The most frequent value (mode) of input factors in cluster R_3^3
Gender	Male	Male	Female	Male
Age	70	25	70	50
Weight	50	50	50	60
Nationality	Iranian	Iranian	Afghani	Iranian
Area of residence	Urban	Rural	Urban	Urban
current stay in prison	No	No	No	No
Case type	new	new	returned	new
Treatment categories	A	A	B	A
TB type	Pulmonary	Pulmonary	Extra-Pulmonary	Pulmonary
Recent TB infection	No	No	yes	No
Diabetes	No	No	No	No
HIV	No	No	No	suspected
Length (Month)	7.07	6.03	19	28.5
Low Body Weight(LBW)	No	No	No	yes
Imprisonment	No	No	No	suspected
IV drug using	No	No	No	suspected
Risky sex	No	No	No	suspected

Furthermore, there are connections among values of variables from a medical point of view. To be precise, by clustering and changing the object`s partition, the most common values of variables in each cluster have been arranged in a meaningful way. For instance, in the first cluster, the most repetitive cases are young new cases with a short length of TB who are under good supervision in rural areas. In cluster 2, there are those cases who are old females from Afghanistan living in urban regions under treatment type 2 who are returned cases having had the disease for about 19 months. Here, being immigrants, long term infection, returned, extra-pulmonary cases and treatment category *B* are really associated in medical knowledge terms. In the third cluster, the most repetitive conditions are related to middle-aged Iranian men, who have pulmonary TB and live in urban regions and are suspected to have had unprotected sex, consume drugs or be HIV positive. Typically those people who have these features are involved with TB for longer durations resulting in the outcome of them quitting treatment which is the case here as well. Due to the high association of HIV, IV drugs, unprotected sex as social related risk factors, it is fairly obvious that partitioning these cases together is a success for the k-means algorithm leading to improve classification accuracy.

Obviously, this integration stage of clustering and classification may improve the prediction accuracy of classifiers performance through strengthening local mappings instead of a general approximation approach which was addressed in the previous chapter.

Three clusters have achieved the best results. However, we have applied a number of clusters: two, three and four. Furthermore, two clusters has performed better than two clusters; thus, it looks as if no strong conclusion can be drawn with regards to an optimized cluster number and further optimizing investigation might be required.

Combined supervised and unsupervised learning method has affected all classifiers` accuracy positively. Each classification algorithm has been fed clustered sets; In other words, in the process of input-output mapping, here, similar objects in a clusters have been applied to produce the given output and model development. Apparently, more consistent objects in separated segments might result in fewer misclassified prediction than any of the applied classification algorithms.

5.5 Summary

According to the result of chapter 4, decision trees outperformed other classifiers; however, they still needed accuracy improvement which was fulfilled through applying the approach of k-means clustering and classification techniques combination. Having compared the results of prediction accuracy before and after clustering, there is an improvement in model accuracy. Creating partitions by the k-means clustering algorithm and carrying out the local learning returns promising results, mainly for the 3-cluster-based model with meaningful results from a medical point of view. Further investigation for more clustering number might be required. Although the produced decision tree is improved in accuracy, it is too large to be easily understandable. In fact, a large number of branches in a decision tree damages comprehensibility and accuracy. In the next chapter, we apply a cluster-based method to develop smaller and more accurate decision trees. Combination of smaller trees might lead to development of a novel algorithm producing more understandable and accurate decision trees. This could be used to predict the outcome of a tuberculosis treatment course which is usable even by staff at a low level of the health system.

Chapter 6

Decision Tree Accuracy Improvement & Simplification using Unsupervised & Supervised Learning Approaches

6.1 Introduction

In the previous chapter we attempted to improve the accuracy of the decision tree through developing the clustering-based decision tree. This accuracy improvement is expected to enhance the tree`s simplicity and interpretability. Simplicity is a prominent feature of the decision tree and this algorithm has been always at the centre of attention due to its scalability, interpretability and comprehensibility [Lim *et al.*, 2000]. Interpretability refers to the level of understanding and insight that is provided by the classifier or predictor for which decision tree are famous [Han & Kamber, 2006].

The level of comprehensibility typically diminishes with increase in tree size and complexity. In fact, in the case of having two trees employing the same kind of test and yielding the same prediction accuracy, the one with fewer leaves is generally preferred [Lim *et al.*, 2000]. Typically, developed decision trees suffer from the weakness of excessive complexity and therefore are incomprehensible to experts [Quinlan, 1999]. That is, the induced decision trees may often not be very clear which prevents them from concisely illuminating classification behaviour. Consequently, they are not able of satisfying the needs of domain expert or even novice users [Breslow & Aha, 1997].

The developed cluster-based decision trees are still fairly complicated with 509 numbers of tree size and 888 numbers of nodes; it needs to be more simplified and understandable. In this chapter, we intend to simplify the cluster-based decision tree as well as improve its accuracy. First the novel methodology of developing cluster-based simplified decision tree (CSDT) is illustrated. This is done through applying hierarchical clustering and classification learning. Then, the new method is applied to the available dataset in the experimental methodology and the results are recorded. The following discussion investigates the results obtained in this chapter by comparing them with previous research. A summary concludes the findings and the chapter.

6.2 Cluster-based Simplified Decision Tree to Predict the Outcome of Tuberculosis Treatment Course

In order to construct a cluster-based simplified decision tree (CSDT) three main stages are required. These are explained in the following sections.

6.2.1 Branches Selection

Suppose there is a dataset where j clusters (C_i^j) are developed by k-means clustering producing K_2^j where *i* is *ith* number of *j* partitions; in fact, *j* is the cluster number which can be set at 2 as a minimum, to create at least 2 clusters. Using the corresponding cases in each cluster, they are learned by a decision tree algorithm to produce (dt_n^i) where *n* is the *nth* decision tree induced from the *ith* number of clusters. Every dt_n^i has several branches which are called br_m^n where *m* and *n* imply the *mth* branch from the *nth* corresponding tree. As the aim of this chapter is both the tree size reduction and accuracy improvement, it is required to select large branches of the initial decision trees with more misclassification cases which can display, at the end, the effect of learning method sensibly. That is, the main criteria for selecting a set of br_m^n from dt_n^i are for them to be both large in size and contain more misclassification cases. However, branches with accurately classified cases have been ignored, even if they are large in size.

6.2.2 Branches Training by Hierarchical Clustering & Classification Approach

In this step, the hierarchical clustering and classification methods are applied to the selected branches. Thus, firstly, the selected branches are segmented by using the K-means clustering algorithm and corresponding cases of every br_m^n are partitioned into *K* clusters. Thus, there is C_p^m where *p* and *m* signify the number of the *pth* partition from the *mth* branch. Using the related cases of each $C_l^j(n)$, a C4.5 decision tree was induced and there would be $subdt_l^p$ where *l* is the *lth* number of learned sub-decision trees and *p* is the related cluster in which its data are used to map and develop the *lth* sub-trees. By composing these sub-trees ($subdt_l^p$) instead of the original large branches (br_m^n), there would be, in fact, branches ($B\acute{r}n$) with new structure in terms of size and misclassified cases. The new branches would make up the new trees which would be more precise as well as smaller in size. These trees can be referred to as $CSDT_n^i$ where *n* is the tree number related to the corresponding cluster. Each one of the $CSDT_n^i$, is compared with the corresponding dt_n^i to find the effect of partitioning and sub-tree development on decision tree misclassification and size reduction.

6.2.3 Evaluating the Accuracy and Simplicity of Cluster-based Decision Trees and CSDT

In order to find if higher accuracy and simplification has been achieved by CSDT, the change of tree size, node number and misclassification rate before and after applying hierarchical clustering and classification algorithms are evaluated; in fact, tree size and node number assessment are considered to estimate the alteration on decision tree interpretability; the change in misclassification rate is measured as decision trees precision icon. Tree size signifies the total number of nodes, including decision nodes and leaf nodes, whereas a "node number" only indicates the number of decision nodes. Misclassification rate is another measurement which is the number of misclassified predictions per selected cases in corresponding branches.

Moreover, to calculate the effect size, a T-test is applied to check if the generated change is significant and has not occurred by chance. This effect size (r) which is an objective and standardized measure of the magnitude of observed effect (Field, 2005) is carried out by the following formula converting a t-value into an r-value:

$$r = \sqrt{\frac{t^2}{t^2+df}} \qquad (6.1)$$

Where t is the value of T-test and df stands for degree of freedom. The value of t-test can be calculated by equation 6.2:

$$t = \frac{\bar{D}-\mu D}{SD/\sqrt{N}} \qquad (6.2)$$

Where ($\bar{D}$) is the mean of the given sample, μD is the population mean, *SD* is standard error and N is the sample number. The value of effect size (r) is calculated for node number, tree size and misclassification number separately.

The systematic process of hieratical clustering and classification which improves the accuracy of decision trees and simplified them is presented in Figure 6.1.

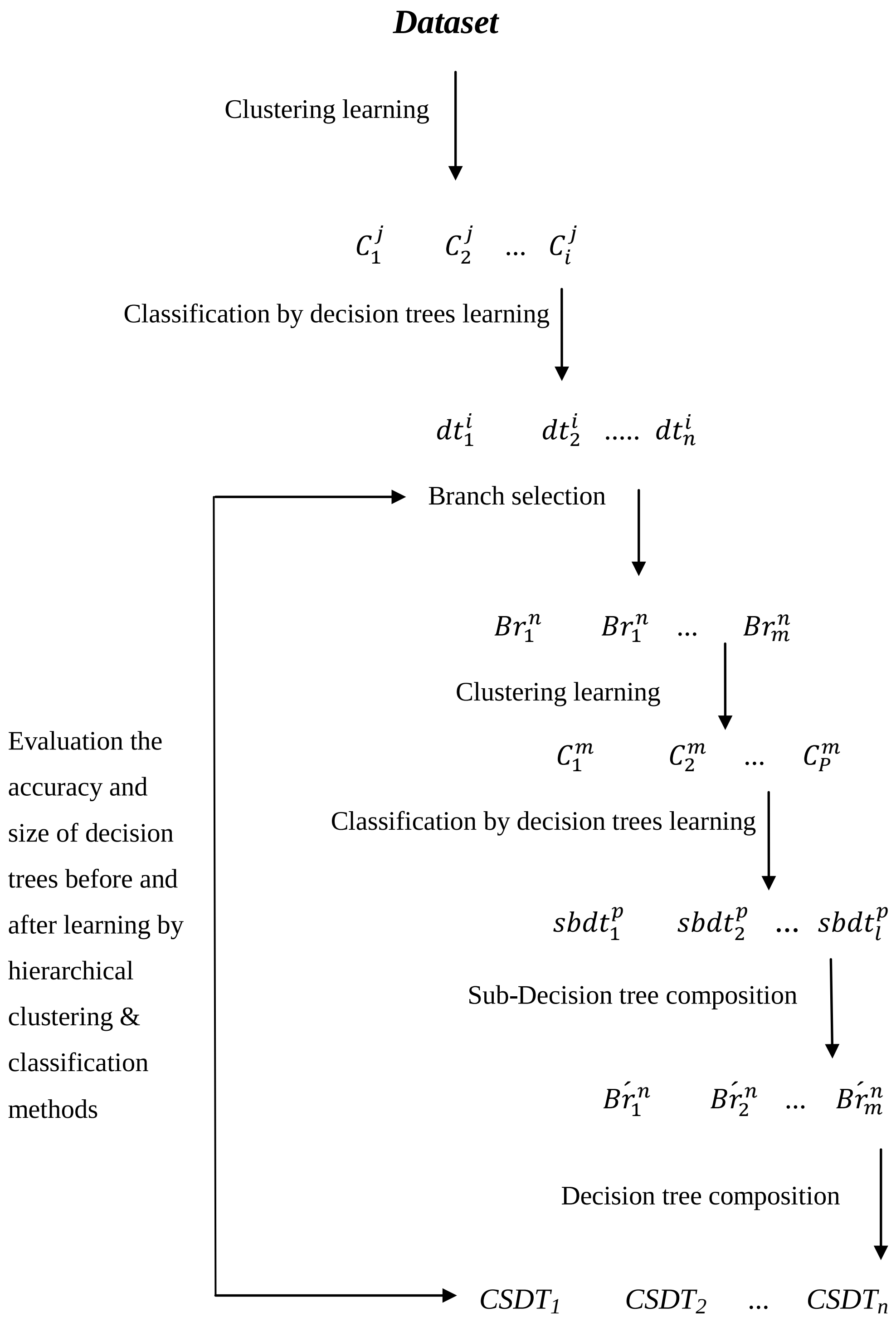

Dataset
Clustering learning
C_1^j C_2^j ... C_i^j
Classification by decision trees learning
dt_1^i dt_2^i ..... dt_n^i
Branch selection
Br_1^n Br_1^n ... Br_m^n
Clustering learning
C_1^m C_2^m ... C_P^m
Classification by decision trees learning
$sbdt_1^p$ $sbdt_2^p$ ... $sbdt_l^p$
Sub-Decision tree composition
$B\acute{r}_1^n$ $B\acute{r}_2^n$ ... $B\acute{r}_m^n$
Decision tree composition
$CSDT_1$ $CSDT_2$ ... $CSDT_n$
Evaluation the accuracy and size of decision trees before and after learning by hierarchical clustering & classification methods

Figure 6.1 the process of hierarchical clustering & classification learning, from cluster-based decision trees to cluster-based simplified decision trees *(CSDT)*.

6.3 Experiment Design

As discussed earlier, available data used to construct a pruned C4.5 decision tree; gave 74.21% prediction accuracy. Accuracy of the decision tree and the other employed classifiers were improved through integrated clustering and classification methods. Although generated cluster-based decision trees are more accurate and automatically pruned, they are still large with 312, 470, and 332 tree sizes for the first, second and third tree respectively; basically, there are 3 pruned trees (dt_n^i) induced from three clusters partitioned by the k-means algorithm. The reason for using three clusters, *k*=3, is that the earlier results were more precise than for 2 or 4 cluster-based decision trees. Looking at the big branches containing many leaves and high misclassification rates, 3, 5, and 4 branches were chosen in the three mentioned cluster-based trees. The criteria is to find the branches (Br_m^n) which have the most misclassification cases and are large from the top of the bottom to the down where it reaches the tree`s leaves. Three branches were chosen for the first tree labelled: Br_1^1, Br_2^1, Br_3^1, five for the second tree: Br_1^2, $Br_2^2, \dots,$ Br_5^2 and four branches for the third tree: Br_1^3, $Br_2^3, \dots, Br_4^3$. The trees are shown in Figures 6.2, 6.3, and 6.4 respectively. Table 6.1 compares the tree size, misclassification number for each of the cluster-base trees and their entire branches, which are the target for this experiment. Through selected branches cross the first, second and third trees, 46%, 38%, and 50% of tree sizes are considered respectively. These branches contain 68%, 47%, and 60% of the misclassified cases of the entire tree. With regards to the number of cases which are selected through these branches in each cluster-based tree, the details are as follows: for the first tree there are 979/1560 which is 63%. That is, 979 cases are selected from the 1560 objects in cluster one, which is already developed in 5.2.1. This portion is 885/1707=52% and 622/1248=50% for the chosen branches from the second and third trees respectively. These trees are learned from the corresponding clusters.

Thus, it seems these long 3, 5, and 4 branches shown in Figures 6.2, 6.3, and 6.4 for the three cluster-based decision trees are well selected to contain the maximum object number, misclassification rate and tree size. Further details about each $\mathrm{dt}_{\mathrm{i}}^{\mathrm{j}}$ size, node number, misclassification rate as well as scattered objects throughout the branches are demonstrated in table 6.2.

The cases related to these selected 12 branches (Br_m^n) are applied to develop 36 clusters (C_P^m) where *K=3*. Then, these clusters are applied to use a decision tree algorithm to develop 36 sub-decision trees ($sbdt_l^p$). If we use these sub-trees instead of the selected branches Br_m^n , we are going to have new branches $B\acute{r}_m^n$ which compose the decision tree (*CSDT*$_n$) with a revised structure in terms of size and accuracy. The process of clustering is implemented in MATLAB *R2007a* environment; the WEKA package was used to generate decision trees as a classification tool.

Table 6.1 The number of tree size and misclassification number for cluster-based trees dt_n^i and $\sum Br_m^n$ for each dt_n^i

Criteria	Cluster-Based trees			Sum of all selected branches		
	dt_1^3	dt_2^3	dt_3^3	dt_1^3	dt_2^3	dt_3^3
Tree Size	312	470	332	144	179	167
Misclassification No.	130	258	116	89	121	69

Table 6.2 The characteristics of the selected branches from three clustered decision tree before integrated learning approach

Tree No.	Branch No.	Number of objects (cases) in selected branches	Tree Size	Node No.	Misclassification rate
dt_1^3	Br_1^1	160	30	12	21/160= 0.13125
	Br_2^1	700	80	35	52/700= 0.0742
	Br_3^1	119	34	13	16/119= 0.1344
	Total1	979	144	80	89/979= 0.0909
dt_2^3	Br_1^2	54	29	10	12/54= 0.222
	Br_2^2	94	46	20	18/94= 0.191
	Br_3^2	62	24	10	14/62= 0.225
	Br_4^2	357	58	26	41/357= 0.114
	Br_5^2	318	51	25	36/318= 0.113
	Total2	885	179	91	121/885 = 0.136
dt_3^3	Br_1^3	241	32	14	34/241 = 0.141
	Br_2^3	240	64	23	15/240 = 0.062
	Br_3^3	87	41	20	9/87= 0.103
	Br_4^3	54	30	11	11/54 = 0.203
	Total3	622	167	68	69/622= 0.110

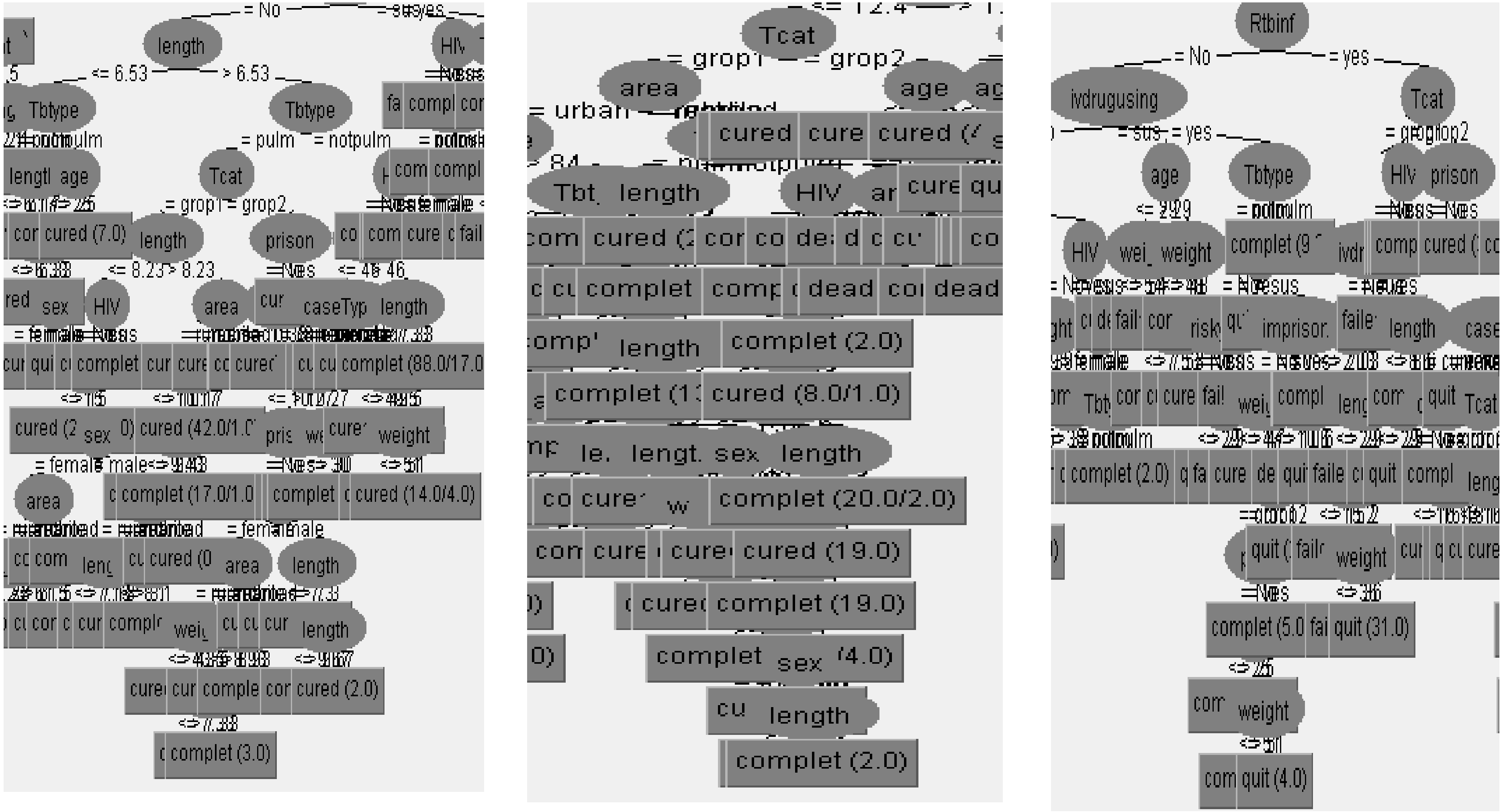


Figure 6.2 The sample of selected branches for the first tree to improve the accuracy and simplicity

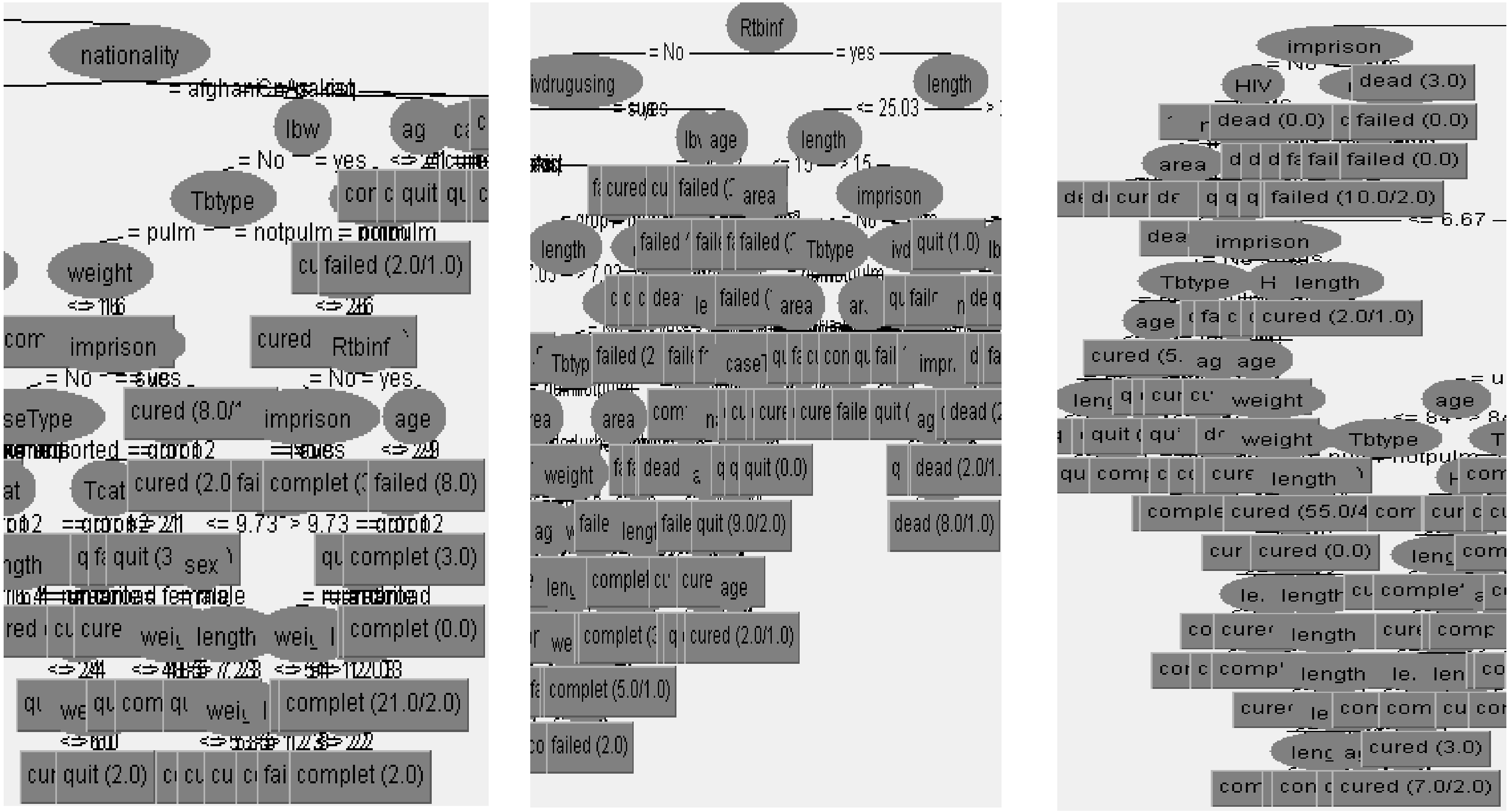


Figure 6.3 The sample of selected branches for the second tree to improve the accuracy and simplicity

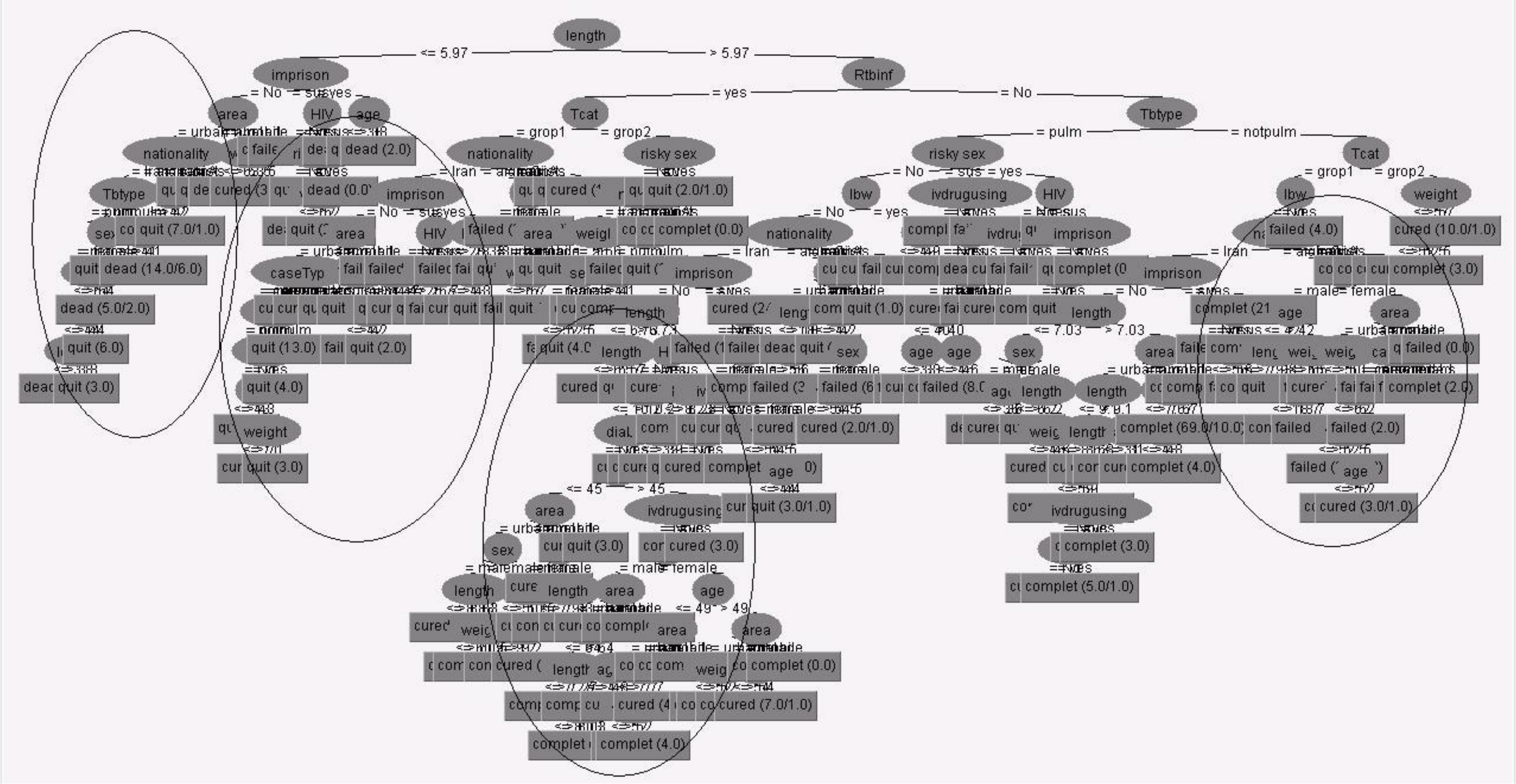


Figure 6.4 The third cluster-based Decision Tree and four spotted branches.

6.4 Experimental Results

After developing $C_i^j(n)$ by k-means clustering and then, three sub-trees construction for each branch, there are new values of tree size and misclassification rate, and object numbers for each Br_i^j demonstrated in Table 6.3, 6.4, and 6.5. The value of misclassification rate is calculated through using training sample.

Sum of the considered criteria for the selected branches of every tree are presented in Table 6.6 where there are reductions in both size and misclassification rate. For $\sum Br_1^1, Br_2^1, Br_3^1$, the tree size is reduced from 144 to 110. It is also verified by reduction in node number from 80 to 50. There is 1% drop in misclassification rate yielded from 10 less misclassification number for the corresponding cases of the three chosen branches.

In $\sum Br_1^2, Br_2^2, Br_3^2, Br_4^2, Br_5^2$ tree size and subsequently node number have 40 and 37 numbers fewer than the original five related branches in the original tree after hierarchical clustering and classification process. The misclassification rate becomes 0.108 after learning which contains 2.8% fewer misclassified cases than before the learning process.

Results for $\sum Br_1^3, Br_2^3, Br_3^3, Br_4^3$ verify the effect of learning on the tree size and misclassification rate where there are 55, 18, and 3% decline in tree size, node number and misclassification rate respectively. Figure 6.5, 6.6, and 6.7 show the reduction in size and misclassification number before and after learning for $\sum Br_i^j$ of dt_1^3, dt_2^3, dt_3^3 comparatively.

The results demonstrated in Table 6.7 disclose the overall decrease obtained for each tree by applying this novel methodology. This comparison reveals the positive effect of hierarchical clustering and classification learning on a given tree`s overall accuracy and simplicity. By comparison of cluster-based decision trees and CSDT, tree size reduced 34, 40, and 55 numbers for the first, second and third tree respectively. The rate of misclassification for trees before and after applying hierarchical learning decreased by 0.7% (first tree), 1.5% (second tree), and 1.6% (third tree).

Figure 6.8 and 6.9 focus on the reduction of tree size and misclassification rate for the three given trees before and after the methodology application. This is done by comparing the criteria for dt_i^j and CSDT. The most decreased values for tree size (40) and misclassification rate (3%) occurred in the second and third trees identically.

Results of assessing the effect size (r) for trees simplification and their accuracy improvement verify above mentioned findings. It is revealed that the values of effect size, r, for tree size is 0.864 (t (11) = 6, ($P \leq 0.001$)), for node number is 0.876 (t (11) =5, ($P \leq 0.001$)) and for misclassification rate is 0.727 (t (11) = 4, ($P \leq 0.007,$)) (presented in Table 6.10); where t is the value of t-test and 11 shows the degree of freedom. That is, the size of tree, number of nodes and misclassification rates are significantly less than their corresponding values before learning. This conclusion is due to the effect sizes values of r which are bigger than 0.50. This indicates that there is a large effect which accounts for 25% of variance [Field, 2005].Tables 6.8, 6.9 and 6.10 show the value of t-test, r and the related parameters in calculation process in more detail.

Table 6.3 Tree sizes, node number and misclassification rate produced after learning process by clustering in branches of first decision tree and classification decision tree for every partition $C_i^j(1)$

Branches and related clusters		Number of objects (cases) in partitioned branches after clustering	Tree Size	Node Number	Misclassification
Br_1^1	$C_1^1(1)$	57	3	1	6/57=0.105
	$C_2^1(1)$	41	5	2	4/41=0.097
	$C_3^1(1)$	62	10	4	7/62=0.112
	sum_1^1	160	18	7	17/160=0.106
Br_2^1	$C_1^2(1)$	224	11	5	18/224=0.080
	$C_2^2(1)$	213	19	9	11/213=0.051
	$C_3^2(1)$	263	39	19	19/263=0.072
	sum_2^1	700	69	33	48/700=0.068
Br_3^1	$C_1^3(1)$	39	9	4	5/39=0.128
	$C_2^3(1)$	32	9	4	4/32=0.125
	$C_3^3(1)$	48	5	2	5/48=0.104
	sum_3^1	119	23	10	14/119=0.117
SUM1		979	110	50	79/979=0.080

Table 6.4 Tree sizes, node number and misclassification rate produced after learning process by clustering in branches of second decision tree and classification decision tree for every partition $C_i^j\,(2)$

Branches and related clusters		Number of objects (cases) in partitioned branches after clustering	Tree Size	Node Number	Misclassification
Br_1^2	$C_1^1\,(2)$	24	11	5	2/24=0.083
	$C_2^2\,(2)$	7	3	1	2/7=0.258
	$C_3^1\,(2)$	23	3	1	4/23=0.173
	sum_1^2	54	17	7	8/54=0.148
Br_2^2	$C_1^2\,(2)$	34	19	9	3/34=0.088
	$C_2^2\,(2)$	31	3	1	5/31=0.161
	$C_3^2\,(2)$	29	9	1	2/29=0.068
	sum_2^2	94	31	11	10/94=0.106
Br_3^2	$C_1^3\,(2)$	21	13	6	2/26=0.076
	$C_2^3\,(2)$	26	1	0	6/21=0.258
	$C_3^3\,(2)$	15	5	2	3/15=0.2
	sum_3^2	62	19	8	11/62=0.177
Br_4^2	$C_1^4\,(2)$	104	19	9	19/104=0.182
	$C_2^4\,(2)$	107	5	2	12/107=0.112
	$C_3^4\,(2)$	145	15	7	5/145=0.034
	sum_4^2	357	39	18	36/357=0.100
Br_5^2	$C_1^5\,(2)$	94	11	5	9/94=0.095
	$C_2^5\,(2)$	113	17	8	10/113=0.088
	$C_3^5\,(2)$	111	5	2	12/111=0.108
	sum_5^2	318	33	10	31/318=0.097
SUM2		885	139	54	96/885=0.108

Table 6.5 Tree sizes, node number and misclassification rate produced after learning process by clustering in branches of third decision tree and classification decision tree for every partition $C_i^j\,(3)$

Branches and related clusters		Number of objects (cases) in partitioned branches after clustering	Tree Size	Node Number	Misclassification
Br_1^3	$C_1^1\,(3)$	85	3	1	15/85= 0.176
	$C_2^1\,(3)$	69	17	8	3/69 = 0.043
	$C_3^1\,(3)$	87	1	0	11/87 = 0.126
	sum_1^3	241	21	9	29/241 = 0.120
Br_2^3	$C_1^2\,(3)$	59	21	10	3/59 = 0.050
	$C_2^2\,(3)$	95	1	0	4/95 = 0.042
	$C_3^2\,(3)$	86	13	6	4/86 = 0.046
	sum_2^3	240	35	16	11/240 = 0.045
Br_3^3	$C_1^3\,(3)$	24	13	6	0/24 = 0.000
	$C_2^3\,(3)$	35	13	6	3/35 = 0.085
	$C_3^3\,(3)$	28	9	4	3/28 = 0.107
	sum_3^3	87	35	16	6/87 = 0.068
Br_4^3	$C_1^4\,(3)$	14	5	2	2/14 = 0.142
	$C_2^4\,(3)$	19	7	3	1/19 = 0.052
	$C_3^4\,(3)$	21	9	4	1/21 = 0.047
	sum_4^3	54	21	9	4/54 = 0.074
SUM3		622	112	50	50/622 = 0.080

Table 6.6 Comparing three sum of branches size and misclassification rate in three clustered–based trees before and after learning by k-mean clustering and decision tree algorithm, the change of overall accuracy before and after learning presented in misclassification rate measurement

Sum of Branches	Measurements	Before Learning	After Learning
$\sum Br_i^1 \text{in } dt_1^3$	Tree Size	144	110
	Node Number	80	50
	Misclassification	89/979= 0.090	79/979= 0.080
$\sum Br_i^2 \text{in } dt_2^3$	Tree Size	179	139
	Node Number	91	54
	Misclassification	121/885= 0.136	96/885= 0.108
$\sum Br_i^3 \text{in } dt_3^3$	Tree Size	167	112
	Node Number	68	50
	Misclassification	69/622= 0.110	50/622= 0.080

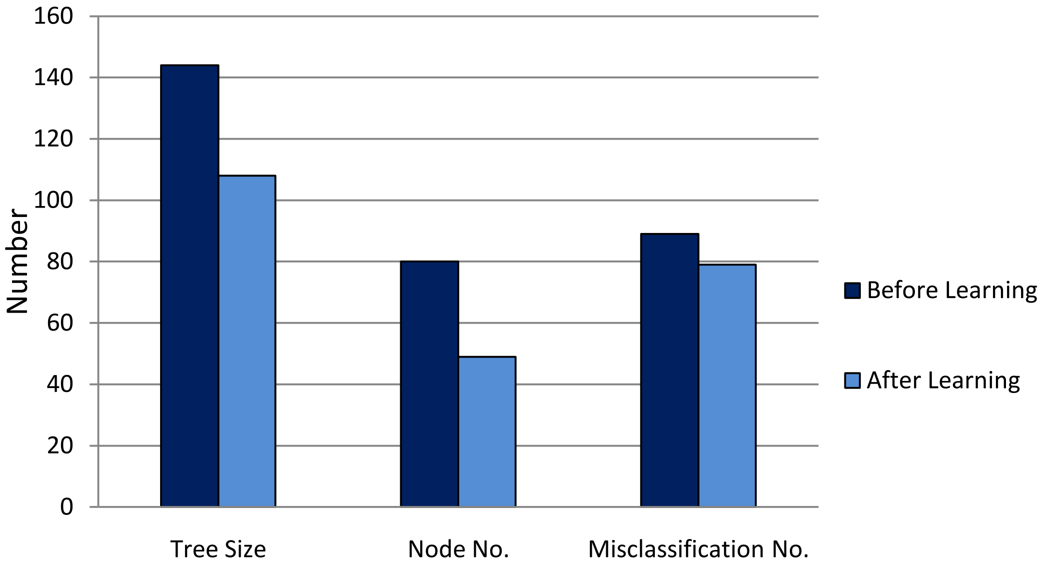


Figure 6.5 Comparing the sum of three selected branches of first tree characteristics before and after learning method application; the change of overall accuracy before and after learning presented in the measurement of misclassification number.

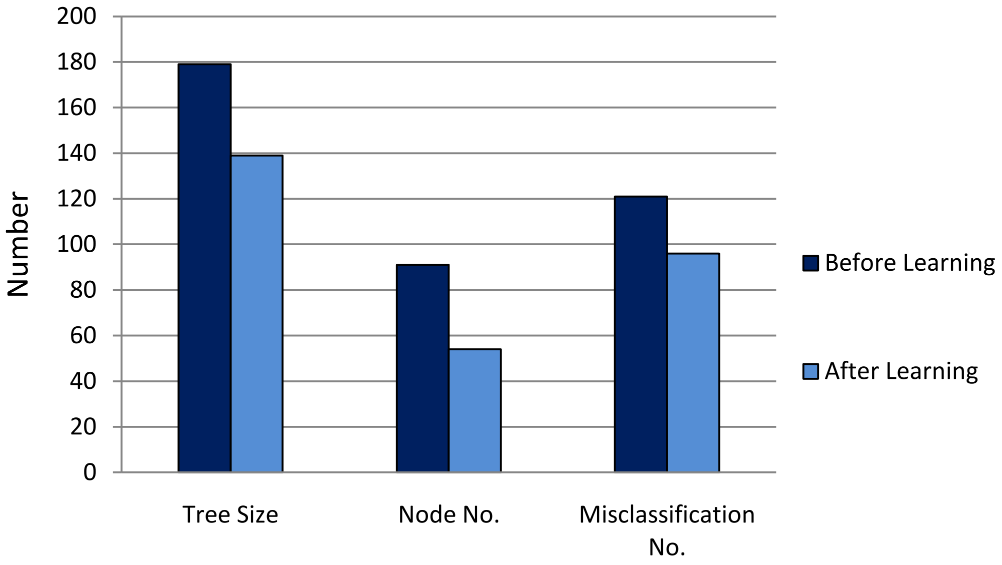


Figure 6.6 Comparing the sum of five selected branches of second tree characteristics before and after learning method application; the change of overall accuracy before and after learning presented in the measurement of misclassification rate.

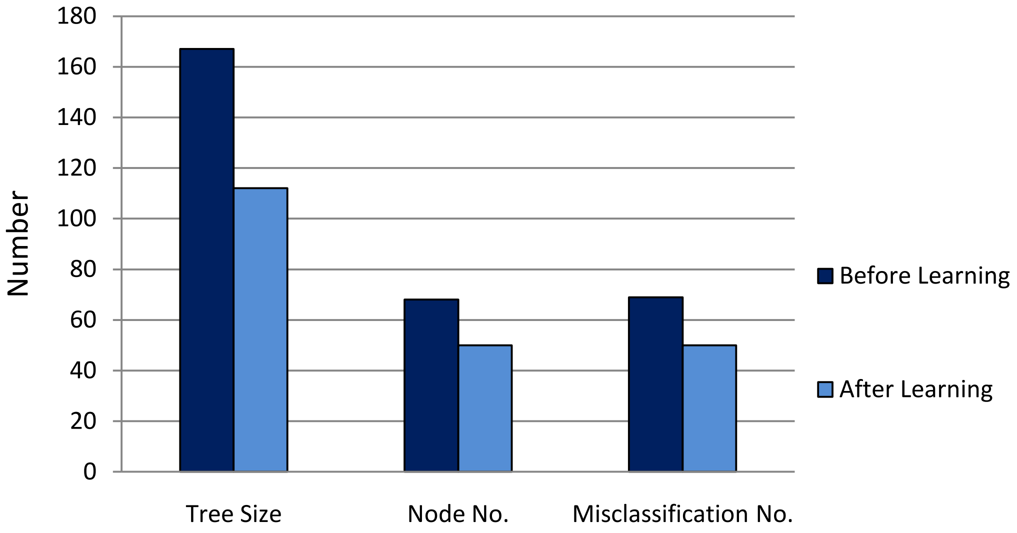


Figure 6.7 Comparing the sum of four selected branches of third tree characteristics before and after learning method application; the change of overall accuracy before and after learning presented in the measurement of misclassification rate.

Table 6.7 Comparing the size and misclassification rate of three cluster-based decision tree $CSDTn$ before and after integrated learning; the change of overall accuracy before and after learning presented in misclassification rate measurement

Trees	Measurements	Before Learning	After Learning
CSDT1	Tree Size	312	278
	Misclassification rate	0.083	0.076
CSDT2	Tree Size	470	430
	Misclassification rate	0.151	0.136
CSDT3	Tree Size	332	277
	Misclassification rate	0.093	0.077

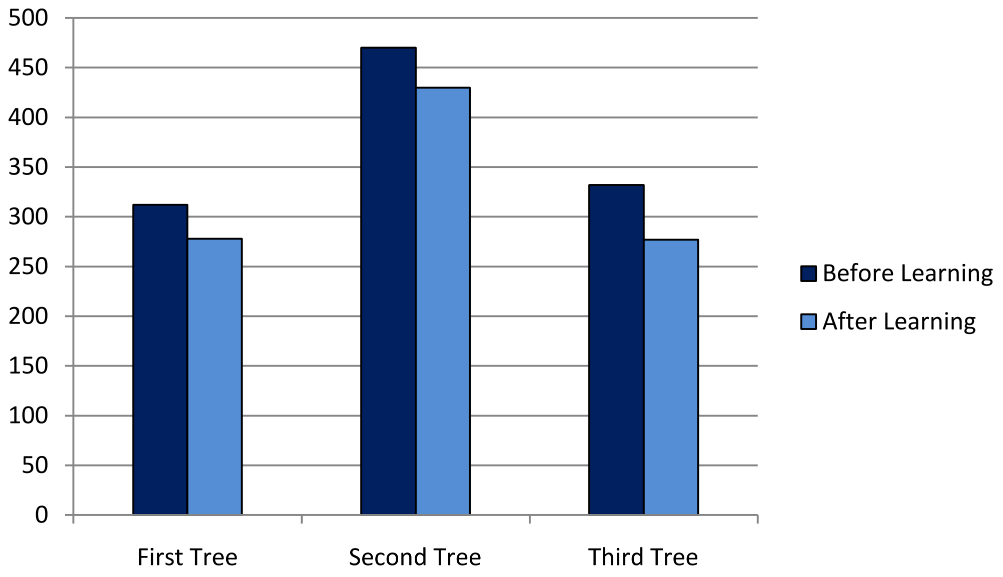


Figure 6.8 Comparing the tree size for *CSDT$_1$*, *CSDT$_2$*, and *CSDT$_3$* before and after applying the hierarchical clustering &classification approach.

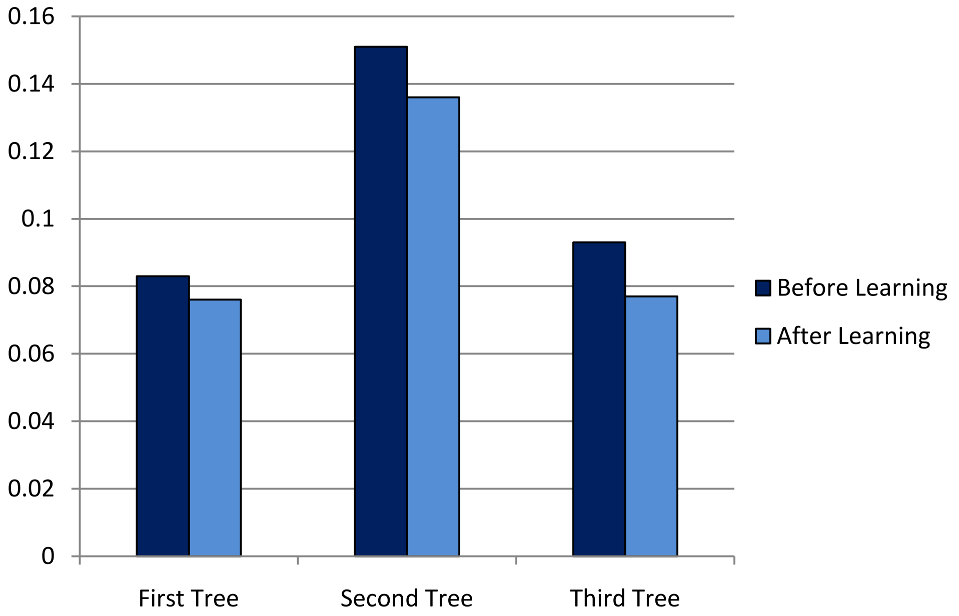


Figure 6.9 Comparing the misclassification rate for *CSDT$_1$*, *CSDT$_2$*, and *CSDT$_3$* before and after learning method application; the change of overall accuracy before and after learning presented in the misclassification rate measurement.

Table 6.8 The paired samples statistics for tree size, node number and misclassification rate before and after hierarchical clustering &classification approach

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Tree size Before	43.2500	12	17.07803	4.93000
	Tree size After	28.4167	12	15.42406	4.45254
Pair 2	Node no. Before	18.2500	12	7.91001	2.28342
	Node no. after	12.7500	12	7.42386	2.14308
Pair 3	Misclassification Rate Before	.1407	12	.05508	.01590
	Misclassification Rate After	.0975	12	.03696	.01067

Table 6.9 The paired samples correlations for tree size, node number and misclassification rate before and after hierarchical clustering &classification approach

		N	Correlation	Sig.
Pair 1	Tree size Before & Tree size After	12	0.864	0.000
Pair 2	Node no. Before & Node no. after	12	0.876	0.000
Pair 3	Misclassification Rate Before &Misclassification Rate After	12	0.727	0.007

Table 6.10 Paired Samples Test for three pairs of tree size, node number and misclassification rate before and after hierarchical clustering &classification approach

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Tree size Before – Tree size After	14.83	8.63	2.49	9.34	20.31	6.00	11	.000
Pair 2	Node no. Before – Node no. after	5.50	3.84	1.11	3.05	7.94	5.00	11	.000
Pair 3	Misclassification-Rate Before Misclassification-Rate After	.043	.037	.010	.019	.067	4.00	11	.002

6.4.1 Sample of Extracted Rules

Each leaf of a decision tree produces a rule in the following structure:

$$\text{IF } X_1 \wedge X_2 \wedge \ldots \wedge X_n \textbf{ THEN } \textit{class a}$$

Where the X_{is} are conditions of the rule and *a* is the corresponding class of the leaf. The produced decision trees can be merely rewrite as the collection of rules. Here, some of the produced rules are listed as follows:

IF Length >6.13and nationality= Iranian and recent TB infection= yes and treatment category group =1 and HIV= positive THEN class= *failed*

IF length <6.13 and age<=21 and weight >37 and case Type=new class= *quit*

IF length >6.13 and nationality =Iranian and recent TB infection =No and IV drug using=no and LBW=no and TB type= non-pulmonary THEN class=*complete*

IF length <=5.97 and imprisonment=no and HIV=no and age<24 and LBW=no and IV drug using=no THEN class=*cured*

IF length >5.97 and age>50 and lbw=yes and area=urban and nationality =Afghani and TB type = extra-pulmonary THEN class=*dead*

IF length <5.97 and age <22 and LBW =no and nationality =Iranian and Diabetes =no and TB type=pulmonary and recent TB infection=no THEN class = *cured*

IF length >6.13 and gender=male and risky sex=suspected and HIV=positive and IV drug using =yes and treatment category =A and case type=returned THEN class *failed*

IF Length>5.97 and area of residency=rural and gender=female and nationality=Afghani and case type = imported THEN class *quit*

IF length <5.97 and LBW=no and area of residency = mobile and gender = female and current stay in prison = no and age <35 THEN class *cured*

6.5 Discussion

Decision trees, according to a few studies [Kurt *et al*., 2008; Colombet *et al.*, 2000; Tsien *et al*., 1998] has been able of performing as accurately as other algorithms or even more accurately than others [King *et al*., 1995; Delen *et al*., 2005]. However, in some reports, the accurate trees might be along with more leaves and complexity [Lim *et al*. 2000]. Tree comprehensibility improvement by simplification is typically carried out by pruning illustrated in [Breslow & Aha, 1997; Quinlan, 1999].

Simplification is an approximately accurate concept which might be more beneficial than an absolutely precise description defined with a lot of detail. This postulation is generally noted in reports of decision tree simplification since smaller and simpler trees trade accuracy mainly by pruning approaches [Bohance & Bratko, 1994]. This is shown in figure 6.10 presenting two typical stages of pruning. In these steps, a pruned tree is produced where the smallest tree (T_*) has the accuracy a(T_*) which is not less than the original tree accuracy a(T_0). This is the ultimate optimal pruned tree where the process of pruning has increased the trees accuracy; however, in real world it doesn't happened very often and the accuracy is either decreased or ideally remains unchanged [Bohance & Bratko, 1994].

Apart from the pruning limitation to improve the tree accuracy, they also modify the tree structure either by stopping tree expansion and removing sub-trees after induction or incrementally resizing the tree to control its size; that is, they focus mainly on tree structure simply by evaluating the significance of the role of the given part on tree performance. However, integrated clustering and classification learning are essentially focused on both tree structure and objects which are derived from tree induction. In other words, to resize and simplify the tree structure this approach exploits the capacity of k-means clustering and benefit from a decision tree`s characteristic to generate sub-trees with improved mutually comprehensibility and precision.

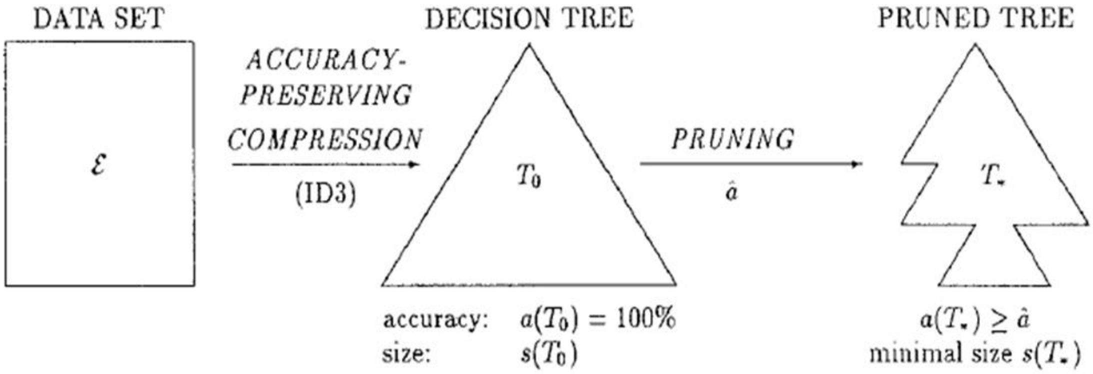


Figure 6.10 A two-stage approach to represent concepts by a pruned tree in an ideal form in which the accuracy has been improved along with tree simplification; obtained from [Bohance& Bratko, 1994].

In the process of integrated unsupervised and supervised learning, the cases corresponding to each selected branches are relocated through iteration process; these repositions are induced through applying k-means partitioning algorithm. This is repeated until similar cases are placed in the identical partition. Thus, using the k-means algorithm *(k=3)* for each of the 12 chosen branches from three trees, 36 partitions were created with similar objects within every cluster and dissimilar to the objects in other clusters. This makes the duty of the applied classification algorithm (decision tree) in the next step more straightforward to distinguish similar groups or classes of objects within a partition.

The applied classification algorithm is C4.5; the tree is constructed in a top-down recursive divide-and-conquer manner. This top-down approach, here, starts running on the partitioned objects obtained by k-means clustering (C_i^j) and their associated class labels. That is, using the C4.5 algorithm, each C_i^j is recursively partitioned into smaller subsets as the sub-trees are being built. Generally, the algorithm of C4.5 imports three inputs to generate a decision tree including:

- The Data partition (*D*) which is a set of training cases.
- The attribute–list and the associated class labels
- The attribute-selection-method specifies a heuristic procedure for selecting the attribute that ‘best’ discriminate the given tuples according to the given class

This process employs an attribute selection measures such as information gain or Gini index. In the process of tree growing, the splitting criterion is determined by attribute-selection-method clearing which attribute to test at node *N* by defining the ‘best’ way to partition the tuple in *D* into individual classes. Moreover, the splitting criterion describes which branches to grow from node *N* regarding the outcome of the chosen test. Ideally, the main aim of the splitting criterion is creating a pure partition through a correct splitting attribute, point and subsets; the pure partition is generated where all of the classes in it belong to the same class. The splitting criterion is the point that clustering effects the decision tree learning procedure when it has already partitioned the given cases based on their attribute and placed similar objects in the same group; afterwards, the splitting criterion produces much purer local regions with lower distance measures like euclidean norm. In fact, clustered tuples with as many similar cases in a partition as possible influences the splitting criterion to stop the tree growing. To be exact, when a clustered tuple is learned instead of *D*, the process of local region identification in a sequence of recursive splits is accomplished in a smaller number of steps with lower nodes and a smaller tree size production. Improving to the splitting criterion to detect the best splitting attribute and their corresponding splitting points is carried out easily since these selections are from cluster objects with more consistency and similarity.

Having looked at the produced sub-trees, there are reductions in the number of applied attributes in the root nodes with many repetitive application of attributes which have been found to be highly correlated with the outcome of the tuberculosis treatment course. For instance, the root node in 33% of 36 developed sub-trees is length and in just under 20% is TB type. Likewise, in the structure of developed sub-trees, 10 attributes have been used whereas very large trees before this integrated learning application have been composed of many sub-nodes from all 17 attributions. Six out of ten variables in sub-trees are weight, age, sex, TB type, length, and recent Tb infection with 27, 21, 20, 14, 41, and 5 times playing the role of node in the 36 developed sub-trees respectively; actually, these exact attributes contribute as a node of sub-trees in

83% of cases. They are highly correlated variables to the outcome of the Tb treatment course signified with two asterisks in tables 4.7 to 4.10. In addition to the promising results shown in figure 6.10 and 6.11 in decreasing tree size and misclassification for three considered trees, the performance of the new algorithm, made up of unsupervised and supervised learning to select the best splitting attributes in sub-trees is a considerable achievement.

6.6 Summary

This chapter introduced an innovative approach to improve a decision tree`s comprehensibility and precision through the new integrated unsupervised and supervised algorithm. The applied hierarchical clustering & classification approach firstly divided the data related to the selected branches of cluster-base trees and then learns them by the C4.5 decision tree algorithm; this has yielded smaller and more accurate trees with reducing tree size and fewer misclassified cases. Produced sub-trees use more number of significantly correlated attributes with the outcome of Tb treatment course in their small and precise structures. This can be a significant step to improve decision trees` application since they usually produce more nodes and leaves when they are at the top accuracy; in other words, the high level of accuracy damages the best feature of decision trees which is their comprehensibility. The novel method of developing cluster based simplified decision tree (CSDT) removes this imperfection and improves both accuracy and simplicity of a decision tree. In conclusion, this chapter has introduced a new method to improve both DTs` comprehensibility and precision concurrently through the integration of the supervised and unsupervised learning methods.

Chapter 7

Conclusions & Future Work

The thesis is concluded by firstly reviewing and discussing the significance of our contributions and then suggesting the directions toward further research in the area of predicting the course of tuberculosis treatment. The contributions of this research are highlighted as follows.

7.1 Conclusions

The conclusion of this study mainly consists of two aspects. Firstly, predicting the outcomes for TB patients under DOTS therapy and secondly proposing the integration algorithms to improve the accuracy and comprehensibility of a tuberculosis treatment course.

A number of systems have been widely used in a variety of medical situation like disease diagnosis or treatment prognosis, presence of disease prediction, differentiating two conditions, and risk factor analysis. The analysis of existing work, pointed out that tuberculosis treatment course prediction is still at a very early stage as only a limited number of reports reveal the influential risk factors on the outcome of a tuberculosis treatment course. However, to the best of our knowledge, no systematic method even in prototype form, for the given outcome is available; the system would be applied to support the level of patient's supervision and support dynamically and effectively. Aspects like the high burden of tuberculosis (nine million new cases and two million new death per annum), five different possible outcomes for TB patients under DOTS therapy, as well as the requirement of defining the level of supervision and support for each specific patient based on possible outcome of treatment course to improve DOTS and shift it from passive to active services are those major reasons behind developing a predictive decision support system for the tuberculosis treatment course postulation. The World Health Organization has designed a global plan named “Stop TB”, a key element of which is treatment with patient supervision and support. This plan requires prediction of patient treatment course destination, to determine how intensive the level of supplying services and support in DOTS therapy should be. As there is no predictive tool, which can be used to predict the outcome of DOTS therapy nor applied, to decide the level of supervision and support to TB patients. There is a rather urgent demand for a solution to this problem. For this purpose, this study has developed a more precise and understandable model to predict the given outcome.

Using feature analysis methods, data of 6,450 Iranian TB patients under DOTS therapy were analysed to initially diagnose the significant predictors. Then, they were applied to find the best classification tool from six examined algorithms including decision trees (DT), Bayesian networks (BN), logistic regression (LR), multilayer perceptron (MLP), radial basis function (RBF) and support vector machine (SVM).

The first results of this research were finding seventeen significantly correlated features, which were: age, sex, weight, nationality, area of residency, current stay in prison, low body weight, Tb type, treatment category, length of disease, TB case type, recent TB infection, being affected by diabetes or HIV, and social risk factors like history of imprisonment, IV drug using, and risky sex $(P \leq 0.048)$. Although former research has already verified nationality, age, imprisonment, and TB case type as influential factors for the non-compliance of TB treatment course, patient’s weight is a new effective attribute *(OR=-0.056, P ≤ 0.0001)*. Males are known as a high risk gender. This study strongly confirmed the role of nationality and imprisonment; Afghani and Pakistani nationals who are living in Iran were more prone to failure in treatment course compliance. Furthermore, diabetes, low body weight, HIV, recent TB infection, risky sex, TB type as well as treatment category A or B are newly diagnosed factors affecting outcome of DOTS therapy.

Using these influential factors and their related patient examples to do classification task using supervised machine learning tools revealed that decision trees outperformed others with the best prediction accuracy (74.21%) whereas the other methods such as BN, LR, MLP, RBF, or even SVM produced 62.06, 57.88, 57.31, 53.74, and 51.36 percentages of prediction accuracy respectively.

Due to the large volume of data with multidimensional structure, it was expected that SVMs, neural networks, and decision trees would outperform others. However, the available dataset is composed of fourteen discrete variables and three continuous attributes. In this investigation, decision trees produced the most promising result. This was due to its dual ability to tackle both continuous and discrete/categorical features, in comparing with other applied techniques which are only good at handling continuous variables. Data type and how normal is the distribution effect prediction accuracy of the classifiers; fourteen available discrete predictors with average value of skew (2.169) and kurtosis (7.469) have enabled decision trees to outperform other algorithms.

Next, the prediction accuracy of the applied classification technique was improved by using the integrated method of *k*-mean clustering approach and each of the above mentioned classification tools. Using two, three and four partitions, the combination method was examined. Results of the next part of the study which was pre-learning by a k-mean partitioning algorithm, combined with the classification approach improved prediction accuracy of all applied classifiers between 4 to 10% with partition number of three (*K*=3). The most and least improvement for prediction accuracy were for logistic regression and support vector machines respectively. Pre-learning by k-mean clustering to relocate the objects and placing similar cases in the same group may improve the classification accuracy.

Although the proposed novel methodology of the combination of k-mean clustering and classification algorithm has improved the accuracy of decision trees with a higher level of accuracy than other tools, the constructed trees are huge with poor comprehensibility. To simplify the identified trees and further improve their prediction accuracy, a novel method of hierarchical clustering and classification algorithm is proposed, which leads to the new CSDT methods. In this process, twelve branches from three cluster-based trees were built by learning from the first, second and third clusters. After partitioning objected in the selected branches by k-means and re-learning them by a decision tree algorithm, more accurate and much simpler sub-trees were generated. Simplicity of three cluster-based decision trees were improved where there are reduction for trees size by 34, 40, and 55 number of nodes for the first, second and third tree respectively. Also, misclassification rate percentage of three –cluster- based trees are increased by 0.7%, 1.5%, and 1.6% for the first, second and third tree respectively. Replacing these sub-trees instead of those selected branches created more accurate and simplified trees. This process may lead to achieving the main aim of this study. That is, to develop the most accurate and understandable model to predict the outcome of tuberculosis treatment course at the onset of anti-biotic therapy. The compatible feature of k-mean partitioning and decision trees to generate pure local regions can simplify and improve accuracy of the decision trees through creating smaller sub-trees with fewer misclassified cases. The extracted rules from these trees can play the role of a knowledge-base for a decision support system in further studies.

7.2 Future Work

Although the project fulfils the aims of study to develop highly accurate predictive model for the outcome of tuberculosis treatment course and the system prototypes now has basic functionality, a number of further developments and improvements could be made. This is to create a clinical decision support system in order to define the level of supervision and support based on the patient status specifically. It enhances the efficiency of DOTS as the international TB control service around the globe and converts it from a passive to active health care service. The suggestions for further enhancements and possible directions for future research are specified as follows:

Firstly, other influential circumstances could be included in future studies. These include structural factors (poverty and discrimination), environmental factors, health care system, management factors as well as other patient related factors which were not considered in this study but still are related to the outcome of tuberculosis treatment course. Thus, considering more comprehensive features can be the subject of further investigation.

Furthermore, in health care system, based on the condition which is defined by predictive system, there is requirement of listing the services that nurses and physician should provide to TB patients. This may add an agenda to the predictive system which assists the staff to carry out the patient supervision in more defined framework. That is, the agenda instructs health staff to know how to deal with different TB patients with various destination of taking DOTS therapy. As an example, the number of patient visits based on the predicted status is one of those issues needs to be defined; obviously, TB patients who are going to be failed in completing DOTS therapy need more visits rather than those cases who have the cure class for predicting in treatment course completion.

Moreover, in this study two, three and four clusters are examined to improve the classifier`s accuracy by k-mean clustering method. Applying more clusters or ideally finding the optimized number can be the subject for further study. Also, only three clusters have been examined to develop the cluster-based simplified decision tree (CSDT) and improve the accuracy of the selected branches and the entire trees. It may be examined with more clusters and produce more promising results.

Due to time and resource limitation, there is only one database related to TB patient and we used one-thirds of this dataset for testing the validity of built models. In the case of further dataset availability, the level of generalisation and model accuracy can be defined more precisely.

Finally, there is a need to develop software based on the proposed method, then users can use this software to achieve potential benefits from the proposed methods. In fact, the prototype which is designed to predict the outcome of tuberculosis treatment course based on the real world database of TB patient needs to be implemented to develop a proper clinical decision support system (CDSS). These computer systems are designed to impact clinician`s decision making about individual patients at the point that these decisions are made. CDSS have the potential to change the way health care is provided such as supervising TB patient actively based on their status which is defined by predictive models. This needs guidelines and plans to define the set of cares that the health organizations such as the World Health Organization work on in depth. The system works based on the produced rules which are interpretable and understandable in medical point of view; further study about their interpretability might be carried out by physicians and medical staff who are expert in the TB patient’s condition and their corresponding outcome of treatment in real world. They can consider each rule to recognize how meaningful they are in medical point of view.

Looking at the misclassified cases which are presented in trees` leaf, many of them are not meaningful in medicine and should not be considered as a true condition. This confirms the models performance even when it has misclassified a patient’s condition.

The innovated technique can be further evaluated through application in other field with different databases. These new methodology might work in other areas like business which has classification task with a big database available. However, not availability of big body of data with different characteristics of database might cause difficulty with this new innovative system.

Chapter 8

References

Alpaydin, E. (2004). *Introduction to machine learning* (1^{th} ed.). Cambridge: The MIT Press.

Anuwatnonthakate, A., Limsomboon, P., Nateniyom, S., Wattanaamornkiat, W., Komsakorn, S., Moolphate, S., Chiengsorn, N., Kaewsa-ard, S., Sombat, P., Siangphoe, U., Mock, P. A. & Varma, J. K. (2008). Directly Observed Therapy and Improved Tuberculosis Treatment Outcomes in Thailand. PLoS ONE, 3(8), e3089.

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54, 979–985.

Bohanec, M., & Bratko, I. (1994). Trading accuracy for simplicity in decision trees. *Machine Learning*, 15, 223-250.

Boudour, M., & Hellal, A. (2005). Combined use of supervised and unsupervised learning for power system dynamic security mapping. *Engineering Applications of Artificial Intelligence*, 18, 673–683.

Bradley, A. E. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145-1159.

Breslow, L. A., & Aha, D. W. (1997). Simplifying decision trees, A survey. *The knowledge Engineering Review*, 12(1), 1-40.

Burman, W. J., Cohn, D. L., Rietmeijer, C. A., Judson, F. N., Sbarbaro, J. A., & Reves, R. R. (1997). Noncompliance with directly observed therapy for tuberculosis, Epidemiology and effect on the outcome of treatment. *Chest*, 111(5), 1168-73.

Chang, C., & Chen, C. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, 36, 4035-4041.

Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., & Jaulent, M. (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression, *AMIA*, 156-160.

Cuneo, W. D., & Snider, DE. Jr. (1989). Enhancing patient compliance with tuberculosis therapy. *Clinics In Chest Medicine,* 10(3), 375-80.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1, 131–156.

Davies, P. D. O. (2003). The role of DOTS in tuberculosis treatment and control. *American Journal of Respiratory Medicine*, 2(3), 203-209.

De Oliveira, L. S., Andreão, R. V., & Sarcinelli-Filho, M. (2008). The use of Bayesian networks for heart beat classification. *Advances in Experimental Medicine and Biology,* 657, 217-31.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34, 113-127.

Dodek, P. M., & Wiggs, B. R. (1998). Logistic regression model to predict outcome after in-hospital cardiac arrest: validation, accuracy, sensitivity and specificity. *Resuscitation*, 36, 201–208.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhart, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of Biomedical Informatics,* 34, 28–36.

Dye, C., Garnett, G. P., Sleeman, K., & Williams, B. G. (1998). Prospects for worldwide tuberculosis control under the WHO DOTS strategy. *Lancet,* 352, 1886-1891.

El-Solh, A. A., Hsiao, C., Goodnough, S., Serghani, J., & Grant, B. J. B. (1999). Using an Artificial Neural Network predicting active pulmonary tuberculosis. *Chest,* 116, 968-973.

Field, A. (2005). *Discovering Statistics Using SPSS*. (2nd ed.). London: SAGE Publication LTD.

Garcla-Perez, E., Violante, A., & Cervantes-Perez, F. (1998). Using neural networks for differential diagnosis of alzheimer disease and vascular dementia. *Expert Systems with Applications,* 14, 219-225.

Gordon, L., & Olshen, R. A. (1978). A symptotically efficient solutions to the classification problem. *Annals of Statistics*, 6, 515-33.

Green, M., Bjo¨rk, J., Forberg, J., Ekelund, U., Edenbrandt, L., & Ohlsson, M. (2006). Comparison between neural networks and multiple logistic regressions to predict acute coronary syndrome in the emergency room. *Artificial Intelligence in Medicine,* 38, 305-318.

Guyon, I., & Elissee., A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3 ,1157-1182.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). USA: Morgan Kaufmann Publishers.

Hardin, J. M., & Chhieng, D. C. (2007). Data Mining and Clinical Decision Support Systems. In Hannah, K. j., & Ball, M. J. (Eds.), *Health Informatics, formerly computers in health care* (2nd ed.), PP. 44-64. New York, NY: Springer.

Harries, A. D., & Dye, C. (2006). Tuberculosis. *Annals of Tropical Medicine & Parasitology*, 100 (5 and 6), 415–43.

Jiang, Y., Li, Z., Zhang, L., & Sun, P. (2007). An improved SVM classifier for medical image classification. *Springer-Verlag Berlin Heidelberg.*

King, R. D., Feng, C., & Sutherland, A. (1995). Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, 9(3), 289- 333.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica,* 31, 249-268.

Kukar, M., Kononenko, I., & Silvesterb, T. (1996). Machine learning in prognosis of the femoral neck fracture recovery. *Artificial Intelligence in Medicine,* 8, 431-451.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34, 366–374.

Lazarescu, M., Turpin, A., & Venkatesh, S. (2002). An application of machine learning techniques for the classification of glaucomatous progression. *Springer-Verlag*, 2396, 243–251.

Lee, S. & Abbott, P.A. (2003). Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of Biomedical Informatics*, 36, 389-399.

Lim, T., Loh, W., & Shih, Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine Learning*, 40, 203–228.

Machado, Jr. A., Finkmoore, B., Emodi, K., Takenami, I., Barbosa, T., Tavares, M., Reis, M. G., Arruda, S., & Riley, L. W. (2009). Risk factors for failure to complete a course of latent tuberculosis infection treatment in Salvador, Brazil. *The International Journal of Tuberculosis and Lung Disease*, 13(6), 719–725.

Marsland, S. (2009). *Machine Learning: An Algorithmic Perspective* (1th ed.). Chapman and Hall.

Mehrabi, S., Maghsoudloo, M., Arabalibeik, H., Noormand, R., & Nozari, Y. (2009). Application of multilayer perceptron and radial basis function neural networks in differentiating between chronic obstructive pulmonary and congestive heart failure diseases. *Expert Systems with Applications*, 36, 6956–6959.

Menzies, R., Rochert, I., & Vissandjee, B. (1993). Factors associated with compliance in treatment of tuberculosis. *Tubercle & and Lung Disease,* 74, 32-37.

Munro S. A., Lewin S. A., Smith, H. J., Engel, M. E., Fretheim, A., & Volmink, J. (2007). Patient adherence to tuberculosis treatment: A systematic review of qualitative research. *PLoS MEDICINE,* 4(7), e238, 1230-1245.

Olson, L., & Delen, D. (2008). *Advanced Data Mining Techniques*. Berlin Heidelberg: Springer.

Panjabi, R., Comstock, G. W., Golub, J. E. (2007). Recurrent tuberculosis and its risk factors: adequately, treated patients are still at high risk. *The International Journal of Tuberculosis and Lung Disease*, 11(8), 828–837.

Pao, Y., & Sobajic, D. J. (1992). Combined use of unsupervised and supervised learning for dynamic security assessment. *Transactions on Power Systems*, 7(2), 878-884.

Pavlopoulos, S. A., Stasis, A. C., & Loukis, E. N. (2004). A decision tree–based method for the differential diagnosis of Aortic Stenosis from Mitral Regurgitation using heart sounds. *BioMedical Engineering OnLine*, 3(21).

Quinlan, J. R. (1999). Simplifying decision trees. *International Journal of Human-Computer Studies*, 51, 497-510.

Rakotonirina, el-CJ., Ravaoarisoa, L., Randriatsarafara, F. M., Rakotomanga, Jde. D., & Robert, A. (2009). Factors associated with tuberculosis treatment non-compliance in Antananarivo city, Madagascar. *Sante Publique,* 21(2), 139-46.

Reiz, B., & Csató, L. (2009). Bayesian network classifier for medical data analysis. *Introduction Journal of Computers, Communications & Control*, 4(1), 65-72.

Sampathkumar, P. (2008). Drug resistant tuberculosis: a global public health issue. *International Journal of Dermatology*, 47(10), 985-988.

Sbarbaro, J. A., & Sbarbaro, J. B. (1994). Compliance and supervision of chemotherapy of tuberculosis. *Seminars in Respiratory Infections*, 9(2), 120-127.

Shieh, F. K., Snyder, G., Horsburgh, C. R., Bernardo, j., Murphy, C., & Saukkonen, J. J. (2006). Predicting non-completion of treatment for latent tuberculosis infection, A prospective survey. *American Journal of Respiratory and Critical Care Medicine*, 174, 717–721.

Silva, A., Cortez, P., Santos, M. F., Gomes, L., & Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit. *Artificial Intelligence in Medicine*, 43(3), 179-193.

Sims, C. J., Meyn, L., Caruana, R., Bharat Rao, R., Mitchell, T., & Krohn, M. (2000). Predicting cesarean delivery with decision tree models. *American Journal of Obstetrics & Gynecology*, 183(5), 1196-1206.

Šmuc, T., Gamberger, D., & Krstacic, G. (2001). Combining unsupervised and supervised machine learning in analysis of the CHD patient database. *The American Invitational Mathematics Examination*, 109–112.

Tangüis, H. G., Caylà, J. A., García de Olalla, P., Jansà, J. M., & Brugal, M. T. (2000). Factors predicting non-completion of tuberculosis treatment among HIV-infected patients in Barcelona (1987–1996). *The International Journal of Tuberculosis and Lung Disease*, 4(1), 55–60.

Thiam, S., Le Fevre, A. M., & Hane, F. (2007). Effectiveness of a strategy to improve adherence to tuberculosis treatment in a resource-poor setting: A cluster randomized controlled trial. *The Journal of the American Medical Association,* 297(4), 380-386.

Tsien, C. L., Fraser, H. S. F., Long, W. J., & Kennedy, R. L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infarction. *MEDINFO*, 493-497.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11), 1225-1231.

Vieira, A. A., & Ribeiro, S. A. (2008). Noncompliance with tuberculosis treatment involving self administration of treatment or the directly observed therapy, short-course strategy in a tuberculosis control program in the city of Carapicuíba, Brazil. *Brazilian Journal of Pulmonology,* 34(3), 159-166.

Wetter, T. (2000). Medical Decision Support Systems. Springer-Verlag Berlin Heidelberg.

World Health Organization (2006). *The Stop TB strategy, Building on and enhancing DOTS to meet the TB-related Millennium Development Goals*. Geneva, WHO/HTM/STB/2006.37

Yang, S. X., Qin, L., Pollari, F., Dore, K., Fazil, A., Ahmed, R., Buxton, J., Grimsrud, K., & Middleton, D. (2006). Modeling and analysis of risk factors for salmonella typhimurium DT104 and non-DT104 infections: A comparison of logistic regression and neural networks. *Proceedings of the Fifth Mexican International Conference on Artificial Intelligence IEEE, (MICAI'06).*

Yew, W. W. (1999). Directly observed therapy, Short-Course: The best way to prevent multidrug-resistant tuberculosis. *Chemotherapy,* 45, 26-33.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modelling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(16).