CHAPTER VI

CONCLUSIONS AND SCOPE FOR FUTURE WORK

6.1 INTRODUCTION

This chapter concludes the thesis with a summary of research findings. It begins with a discussion of findings in relation to major research questions and gives the detailed description of how the outcomes of work emerged. This research also remains as the representation of the proposition of grounded hypothesis that can be experimented in further research, about the extent to which the prediction model can assist the physicians in diagnosing at clinical level. Future scope for the potential research is also discussed in a detailed way that certainly would pave a way for researchers in the future.

6.2 MAJOR RESEARCH QUESTIONS AND PROPOSITION OF GROUNDED HYPOTHESIS

A major research question is:

6.2.1 How various data mining techniques can be used in health care industry and to identify their performance in prediction?

In proportion to the investigate character of the research, the grounded theory has been developed through a comprehensive literature, a process of system development and experimentation. Exploration on the major research question and assertion of the grounded hypothesis are stated as follows:
Data mining for healthcare is useful in evaluating the effectiveness of medical treatments and ensures detection of fraud and abuse. The data mining techniques give the necessary standard in prediction. However, use of data mining techniques can reduce number of tests and physicians identify effective treatments and best practices. This reduced test set plays an important role in time and performance. The performance in prediction depends on the various attributes which are helpful in predicting disease efficiently and patients receive better and more affordable healthcare services.

6.2.2 How does a classification techniques help in developing the prediction model so as to predict accurately the risk of heart disease among diabetic patients?

This thesis has contributed some important aspects:

Firstly, it has windowed a new perspective on how classification techniques help in developing the prediction model. Classification in data mining is used to predict group membership for data instances and it involves the use of sophisticated data analysis tools to discover the relationship in large datasets. Thus, there is a great potential for the use of data mining techniques for medical data classification.

Secondly, an innovative approach in predicting accurately the risk of heart disease among diabetic patients is needed. In this research, prediction of the chances of getting a heart disease for the diabetic using various attributes has been attempted. This helps physicians to help patients by categorizing them based on the risk factors. A number of experiments have been conducted for comparison of the performance of predictive data mining techniques on the same diabetic dataset with 1000 records using different attributes.

In the first experiment, we have presented naïve Bayes data mining classifier technique has been applied which produces an optimal prediction model using minimum training set to predict the chances of diabetic patient getting heart disease.
Using diabetic’s diagnosis, the proposed system uses attributes such as age, sex, blood pressure and blood sugar and predicts the chances of a diabetic patient getting a heart disease.

In the next method, in this work support vector machine data mining classifier technique has been used with radial basis function kernel to diagnose vulnerability of diabetic patients to heart disease. Support vector machine is generally considered being a successful tool for classification purpose and evidence is shown in the present work. The outcome of the proposed system is quite good, since it has proved and shown good accuracy on the prediction of vulnerability of diabetic to heart diseases.

On comparing Support vector machine and Decision tree, this work aims to determine the most accurate technique to predict the risk in diabetic patients for heart disease. Rapid Miner has been used as a tool due to its learning operators and operator framework. Apart from accuracy are trying to determine the ROC, AUC and Lift chart for measuring the performance.

In the final experiment, a comparative analysis of the classifiers which can classify the risk of diabetic patients getting heart disease from a machine learning perspective has been provided. It aims to evaluate and compare using three different data mining classification techniques such as naïve Bayes, support vector machine and decision tree to determine the possible ways to predict the risk of heart disease for diabetic patients based on their predictive accuracy.

Out of the three chosen methods, the decision tree provides the highest classification accuracy in all of the models. Not only in overall accuracy, even in terms of precision and recall of the three classes (High, Medium and Low), decision tree exhibited a very consistent performance in the final results.
6.3 FINDINGS RELATED TO RESEARCH AIMS AND OBJECTIVES

In data mining, intelligent methods are applied in order to extract data patterns. There are tremendous opportunities to assist physicians deal with this large amount of data. The medical data mining has great potential for exploring the hidden patterns in the data sets. These patterns can be utilized for clinical diagnosis. Recognition and classification of patterns in multivariate patient attributes enable prediction of future outcomes based on previous experiences. Our study predicts and classifies the data with a reasonable accuracy. It helps in quality healthcare services based on the patient’s needs, symptoms and preferences. It minimizes the waiting time for medical treatment.

The naïve Bayes model was able to classify 74% of the input instances correctly. It exhibited a precision of 71% on an average, recall of 74% on an average, and F-measure of 71.2% on an average.

SVM classifier has been used effectively and the results show a very high classification accuracy i.e 94.60% overall, and a very high precision for the positive class (97.52%) also the recall of the positive class is quite good (83.10%). In the case of negative classes, the classifier exhibits high precision (93.67%) as well as high recall (99.10%).

The use of the decision tree using various split methods such as gain ratio, information gain and gini index has been attempted in the current study. The information gain is used as split parameter in decision trees. The results are exhibited by average precision, and recall and we found that accuracy of this technique is 90.79 % followed by naïve Bayes (81.58%) and support vector machine (61.26%) for predicting the heart disease for diabetic patients using diagnostic features. The performances are compared through accuracy, sensitivity, specificity and F-score.
Decision tree model was consistent in its performance and outperformed naïve Bayes and SVM model. So we finally fine tuned the decision tree model for optimal performance for predicting the chances of heart disease for diabetic patients.

6.4 FUTURE ENHANCEMENT

It has been shown that, by using a decision tree, it is possible to predict heart disease vulnerability in diabetic patients with reasonable accuracy. Classifiers of this kind can help in early detection of the vulnerability of a diabetic patient to heart disease. Thus, the patients can be forewarned to change their lifestyles. This will result in preventing diabetic patients from being affected by heart diseases, thereby resulting in low mortality rates as well as reduced cost on health care for the state. This can be extended in future to predict other types of ailments which arise from diabetes, such as visual impairment. The proposed work can be further enhanced and expanded, to use stacking techniques to increase the accuracy of decision trees and reduce the number of leaf nodes.