Prediction of Heart Diseases Using Data Mining Techniques: Application on Framingham Heart Study

Nancy Masih, Chitkara University, Khanna, India
Sachin Ahuja, Chitkara University, Punjab, Chandigarh, India

ABSTRACT
Health care organizations accumulate large amount of healthcare data, but it is not ‘extracted’ to draw hidden patterns which can prove efficient for the decision making process. Data mining techniques can be used to gain insights by discovering hidden patterns which remain undetected manually. Data analytics proves to be useful in detection and identification of the diseases. A complete analysis has been conducted on the FHS (Framingham Heart Study) using various data analytic techniques viz. Decision tree, Naïve Bayes, Support vector machine (SVM) and Artificial neural network (ANN) and the results were ranked according to the accuracy. ANN produce better results than other classification algorithms. The output helps to find out the prominent features that cause heart disease and also identifies the most common features that must be analyzed for prediction of deaths due to heart disease. Despite various studies carried out on heart diseases, the main focus of this study is prediction of heart disease on the dataset of FHS by using various classification algorithms to achieve high accuracy.

KEYWORDS
Artificial Neural Network, Decision Tree, Framingham Heart Study, Heart Disease, Naïve Bayes, Prediction, Support Vector Machine

1. INTRODUCTION
Coronary heart disease (CHD) is convicted as the leading reason for mortality rate worldwide according to WHO (World Health Organization). It is observed that CHD is the cause considered for 17.7 million deaths every year and more than twenty-four million ratio of people anticipated passing from cardiovascular sickness by the year of 2030 (Kinge & Gaikwad, 2018) CHD dominates other diseases with its severe effects on a person’s wellbeing worldwide (Wilson et al., 1998). In earlier times, the risk of the respective disease used to be analyzed by personal experience of doctors and patients and which was highly vulnerable to various errors and lack of hidden patterns were also observed (Palaniappan & Awang, 2008). Manual decisions are based upon the knowledge and experience of doctors’ which cannot be always accurate. Therefore, it is vulnerable to various errors and hence diminishes quality of treatment given to patients. The errors of traditional methods give rise to the various new techniques

DOI: 10.4018/IJBDAH.2018070101
such as machine learning, data mining and artificial intelligence which can provide more accurate prediction results (Huang, Chen, & Wang, 2007). A prominent terminology of co-relations in features was also introduced which is used for better training of neural network outputs by connecting them to hidden layers of neural network. The feature co-relation analysis includes a feature selection process which is performed prior to analysis to assign ranking to the features selected. After that, the features associated with lower ranks will be eliminated because it will reduce the complexity of network and it has been proved that more the number of relevant co-relations of features are, less the risk factors would be for the particular disease (Kim & Kang, 2017). For instance, Systolic Blood Pressure and Diastolic Blood Pressure are co-related with each other for diagnosis of Blood Pressure and they are connected to the input layer. A variety of data sets are available on internet for various types of diseases but we are have used Framingham Heart Disease dataset for prediction as heart disease is spreading widely worldwide and it is one of the datasets which is very popular and old but still in use. All the considered attributes in dataset such as Last exam, Cause of death, Exam first CHD, Death from CVD, Death from CA (Coronary Artery) disease, Sex, Age, Height, Weight, BMI (Body Mass Index), Serum cholesterol Exam 1, Serum cholesterol Exam 2, Diastolic Blood Pressure, Systolic Blood Pressure, Metropolitan relative weight along with Amount smoked were analyzed and then the combined information of these factors is used for classifying various risk probabilities for prediction (Anderson, Odell, Wilson, & Kannel, 1991). The paper introduces extreme studies conducted on Framingham Heart Disease dataset. The comparison of different classification algorithms conducted on this dataset will organize a well-structured knowledge format of CHD disease along with their risk measurements and treatment strategies for medical practitioners (Kim & Kang, 2017).

Moreover, these techniques reduce expenses and can represent the results in a far better way which is easy to understand by the ordinary people also (Palaniappan & Awang, 2008). Various machine learning classification algorithms were used to analyze the useful patterns of dataset for evaluating risks of diseases, but neural network tends to provide better results than the other algorithms of classification (Dangare & Cse, 2012). ANN dominates other data mining techniques with its more accurate results. It discovers new intelligible form and data associated with the heart disease. Four data mining techniques i.e. Decision tree, Naïve Bayes, SVM and ANNs are implemented on the Framingham dataset. Efficiency of all the four algorithms is based on the accuracy they achieve. Accuracy can be measured as the proportion of decisions that are correctly estimated.

The choice of the algorithm used is based on the task associated for prediction of the heart disease. The data available for prediction was found suitable for applying the classification techniques. Minor changes related to data cleaning and pre-processing were done before applying following techniques:

- **Decision Tree:** Decision tree is branching graph which demonstrates every outcome for final decision. Every attribute is considered as a branching node in which each node is supposed to split the data until a decision is made. Tree selects the best attribute as its first node which is known as root node and terminates with the final nodes which are its leaf nodes. Major advantage of using the decision trees is its simple and easy approach (Design et al., 2018). This approach is user friendly where each node consists of some variable which denotes a decision and it does not hide any kind of information and remains transparent throughout (Harrison et al., 2018). Decision Trees are known to be vulnerable against overtraining/over fitting. Over fitting occurs when a model learns its current training data in a very well manner which will asserts a negative impact on training of new data. However, if a tree expands completely and all last nodes score minimal impurity then it is known as over fitting (Li, Laima, & Li, 2018). Under fitting is reverse: here, the model is too simple that it can neither memorize its current training set nor perform well with new training data to analyze the patterns of the data. To prevent your tree becoming over trained they have the option to (pre)prune themselves. Pruning is a way to cut away leaves which are not statistically meaningful, after the tree was built. Pre-pruning prevents that such leaves being built at all. The building of new nodes is repeated until the stopping criteria are
Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection

A Survey on Prediction Using Big Data Analytics
[www.igi-global.com/article/a-survey-on-prediction-using-big-data-analytics/197438?camid=4v1a](www.igi-global.com/article/a-survey-on-prediction-using-big-data-analytics/197438?camid=4v1a)