CONCLUSION AND RECOMMENDATIONS

The main intent of this study is to predict heart disease through machine learning data analysis techniques and platforms. Further, this study proposes a comparative study on the accuracy of the technique for multiple platforms. Logistic Regression has been used to analyze the Cleveland heart dataset with the help of data analysis tools or platforms python and R. The dataset contains 303 records with 14 necessary attributes with some missing values. At first, the raw dataset has been prepared to visualize and analyze the dataset through raw coding with the help of different libraries and packages imported on both python and R. As it is a classification problem, the outcome has been classified into two classes 1(one) and 0(zero). "1" indicates having heart diseases. Further work involves the development of the system using the mentioned technique and hence training and test the system. It produces a satisfactory accuracy score in both python and R platforms. As we know if the ROC curve is over 0.50 then it can be considered a good result. In this study logistic regression produces an accuracy score over 0.50 for both python and R. However, when the logistic regression is applied with python, it produces a more accuracy score of 0.9230 than R accuracy score of 0.7444. Moreover, if more data has been used then the model could have improved more. The result of the logistic regression shows that men have more chance to have heart disease than women. Age, number of cigarettes each day, and systolic blood pressure are the main risk of heart disease.

Though the Logistic Regression technique has produced a significant-good result, this heart disease prediction can be done using other machine learning techniques like Artificial Neural Network, Random Forest. The Artificial Neural networks perform well with the ensemble method to this dataset. And the Random forest can perform better with decision trees. However, an Artificial neural network can produce more precise results than any other technique. Moreover, instead of using confusion matrix validation, this study could have used k-fold cross-validation to find better accuracy in the trained and testing dataset.