Effective heart disease prediction system using data mining techniques

Abhinav Saurabh
MT20127
abhinav20127@iiitd.ac.in

Apoorv Lokhande
MT20022
apoorv20022@iiitd.ac.in

Shivam Mittal
MT20041
shivam20041@iiitd.ac.in

1. BACKGROUND INFORMATION

Heart disease is one of the most notable causes of death in the world today. Heart disease encompasses a wide range of cardiovascular problems and has become an important health issue and estimated that approximately one person dies per minute due to heart disease. Several diseases and conditions fall under the umbrella of heart disease. Heart disease accounts for a set of range of conditions that affect your heart, such as stress, lack of exercise, high blood pressure, smoking, alcohol, drug abuse, high cholesterol, fast blood sugar. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease. There are several heart diseases, including Arrhythmia, Atherosclerosis, Cardiomyopathy, Congenital heart defects, etc. So, there is a need for a reliable, accurate, and feasible system to diagnose such disease in time for proper treatment.

2. OBJECTIVE

To develop an effective heart disease prediction system using data mining and machine learning techniques.

Many Hospitals today have information systems that store data about patient health records. These sometimes amount to a massive amount of data. But hardly these data are used in clinical decision making. These data could be hidden gold of knowledge that is mainly ignored. Here our main objective is to develop a system to tap into the massive potential of hidden information. We will try to implement possible machine learning techniques. And optimization techniques to get an accurate prediction of heart disease prediction. It could potentially save lives if doctors could know early the probability of disease.

3. METHODOLOGY

A) Preprocessing

The dataset contains a total of 303 patient records, where six records are with some missing values. Those six records have been removed from the dataset, and the remaining 297 patient records are used in pre-processing. The Multiclass target variable changed to a binary class variable. In the instance of the patient having heart disease, the value is set to 1, and else the value is set to 0, indicating the absence of heart disease in the patient. The results of data preprocessing for 297 patient records suggest that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value of 0, indicating the absence of heart disease.

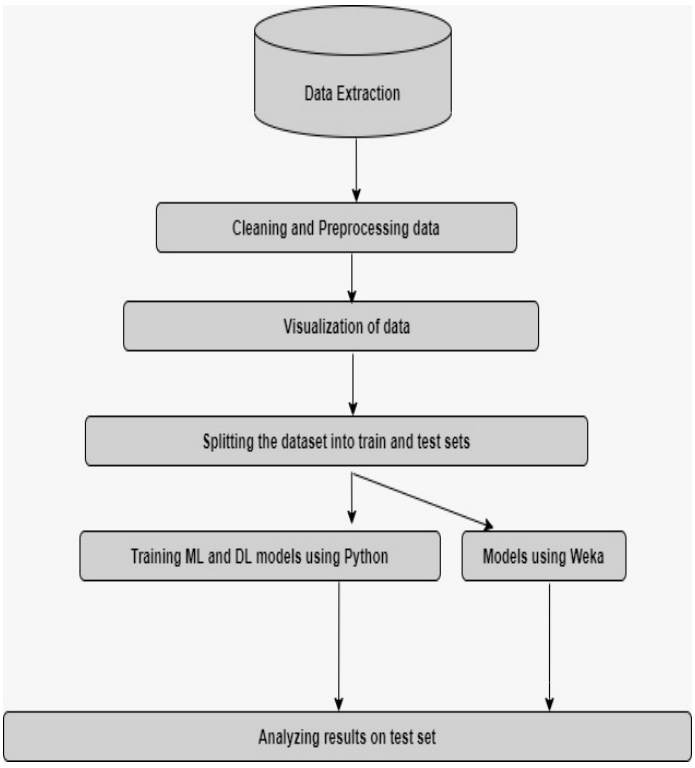

Figure 1: Flow of Work

B) Visualization of Data
We visualize some properties to find insights about the dataset. Some of our understanding is that people of age 58 are mostly suffering from heart disease. The heart rate of a person may fall between 80 to 200. As you can see in Figure-2 and Figure-3, respectively.

C) Splitting
The whole dataset has been split into a training and testing set. The 60% data is taken for training, while the remaining 40% data is used for testing.

D) Classification Models
The training data is trained by using five different machine learning algorithms, i.e. Decision Tree, KNN, SVM, Random Forest, Multilayer Perceptron. For each algorithm, we first fit on the training data, and then we predict the result using testing data and calculate each model's accuracy.

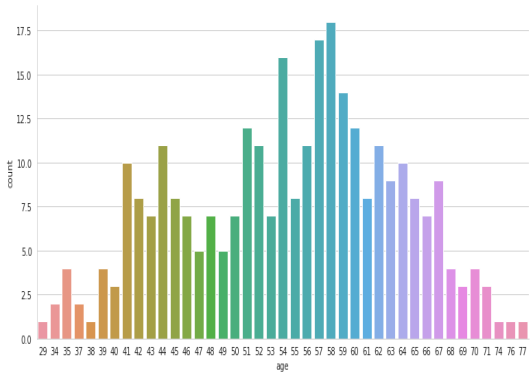

Figure 2: Age wise data classification

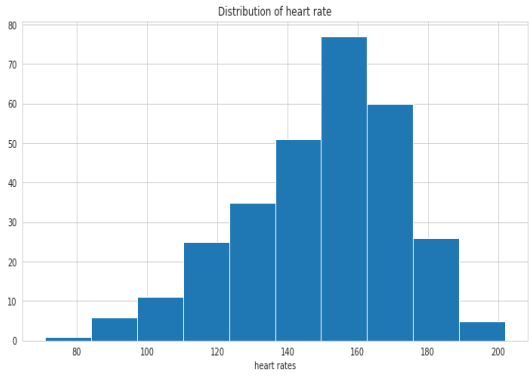

Figure 3: Distribution of Heart rate

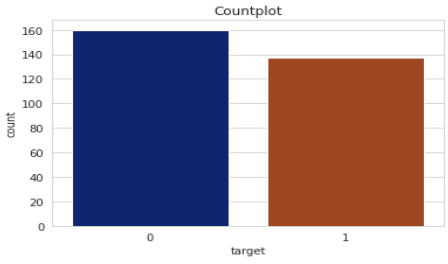

Figure 4: Count Of Target class

4. RESULTS

We have implemented various models such as Decision Tree, Random Forest, SVM, KNN, Multilayer Perceptron. Following are the results obtained.

4.1 Decision Tree

In the Decision tree, we obtained an accuracy of around 75.63% which is not good enough. As we can see in the confusion matrix, False Positives and False negatives are pretty significant in this case, i.e. 16 FP and 13 FN. These are primarily responsible for less accuracy.

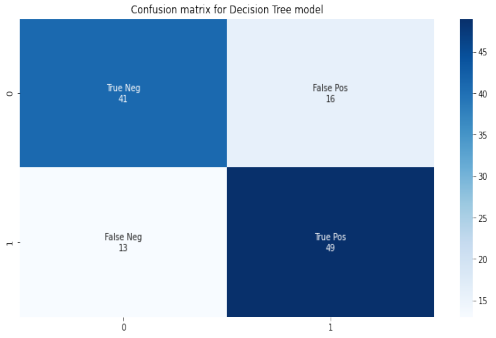

Figure 5: Decision Tree Confusion matrix

4.2 Random Forest

In Random Forest, we obtained an accuracy of around 74.78% which is not good enough. As we can see in the confusion matrix, False Positives and False negatives are pretty large, even in this case. The False Positives were 7, and False negatives were 23. Mostly false negatives are responsible for low accuracy.

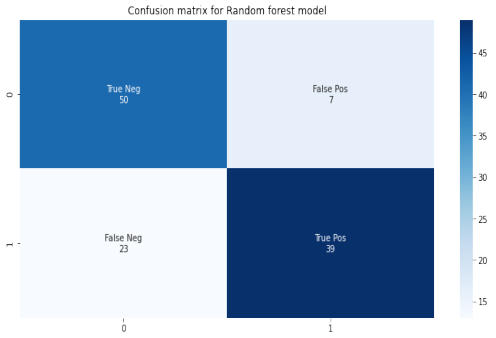

Figure 6: Random Forest Confusion matrix

4.3 SVM

In SVM, we obtained an accuracy of around 82.35%. We can see in the confusion matrix 4 False Positives and 17 False Negatives. Here false-negative accounts for most of the accuracy losses.

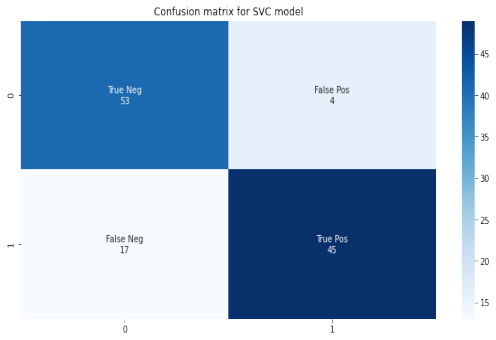

Figure 7: SVM Confusion matrix

4.4 KNN

In KNN, we obtained an accuracy of around 80.67% which is quite good. As we can see in the confusion matrix less number of False Positives and False negatives.Here we had 7 False Positives and 16 False negatives.

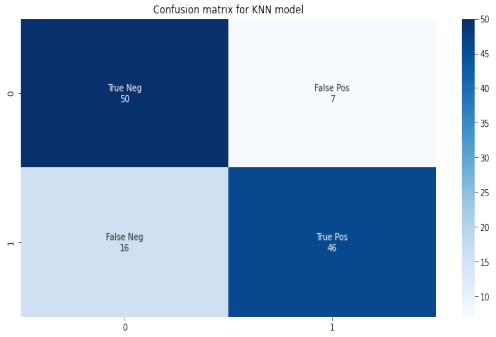

Figure 8: KNN Confusion matrix

4.5 MLP

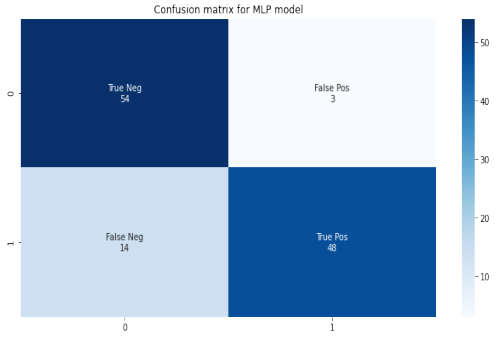

Figure 9: Multilayer Perceptron Confusion matrix

In Multilayer Perceptron, we obtained an accuracy of around 85.71% which is very good. As we can see in the confusion matrix significantly fewer False Positives and False negatives. Here we had 3 False Positives and 14 False Negatives.

The results show that Multilayer Perceptron outperformed most of the algorithms used here. It gave an accuracy of 85.71% which is the highest among all.

The following pipeline was applied to different Machine Learning algorithms and the results are tabulated :

Models	Accuracy
Decision Tree	75.63 %
Random Forest	74.78 %
SVM	82.35 %
KNN	80.67 %
Multilayer Perceptron	85.71 %

Table 1: Accuracy Table

4.6 REPLICATION OF THE WORK

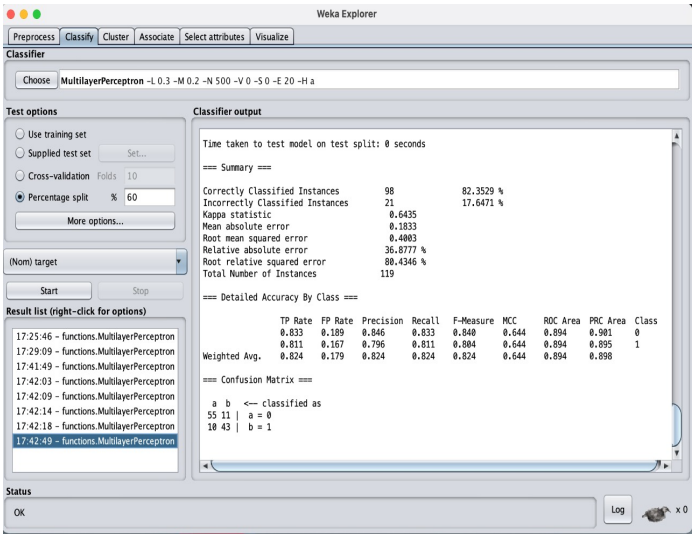

Figure 10: Multilayer Perceptron in Weka

We also used weka to implement multilayer perceptron as done by the author. We divided the training and testing data into a 60:40 ratio. Then converted the target variable to a nominal attribute. Further applied Multilayer Perceptron, and we obtained an accuracy of 82.35%.

5. CONCLUSION

Here MultiLayer Perceptron gives 85.71% accuracy on results which is quite good. This provided quite an accurate prediction of heart disease. Pre-processing the medical data of heart-related and early detection of heart problems will help save lives in the long term. The techniques used here, such as Decision Tree, Random Forest, SVM, KNN, Multilayer Perceptron, are to process the data and provide sound judgement towards heart disease. Predicting heart disease is challenging, but if it is detected at early stages, prevention measures could be taken, and the person could be saved. This could help in controlling the mortality rate very effectively.