Prediction of Human Heart Disease using Data Mining Techniques

*

Swati Verma
Computer Science and Engineering Department
Indraprastha Institue of Information Technology
Delhi,India
swati190732@iiitd.ac.in

***Abstract*—One of the major challenges faced by data mining and machine learning algorithms is in the field of medicals. Due to the change in environmental conditions, several diseases have become common nowadays. Also as the lifestyle is changing, it is leading to several dangerous diseases. Especially, heart disease has become very common nowadays. According to research, it is found that around 17 million people die every year due to heart disease. So, diagnosing heart disease is very important. Although many machine learning techniques effectively predict such disease on the basis of different parameters still they cannot guarantee whether a person is suffering from heart disease or not. This paper presents different data mining methods such as support vector machine, logistic regression and ensemble classifiers to predict whether a person will have heart disease or not.**

I. Introduction

One of the biggest reasons for death nowadays is heart disease and predicting such disease so that the patient can be provided accurate treatment is an even more difficult task. There are various factors using which we can predict whether a person will have heart disease or not. Like if a person is having high cholesterol values and suffers from frequent chest pain then we can say that there is a higher chance of having heart diseases. Also, smoking, age, bp, etc are some of the factors for heart disease. Various techniques like SVM classifies whether a person is having a heart problem or not by using hyper plane which divides the data into two halves. Each half contains one class either positive or negative. Similarly, the random forest is one of the ensemble method which uses a decision tree as a classifier. Also, logistic regression also performs classification by using the logistic function to model binary dependent variables.

II. Data Description

The dataset used for analysis and prediction is taken from Cleveland dataset which consists of 1025 records and 14 attributes on which classification is done. Attributes are age, sex, chol(which represents cholesterol level), cp(represents a type of chest pain), fbs (which is fasting blood pressure that is when a person is having fast then his/her blood pressure is measured and used for analysis), trestbps(represents resting blood pressure of a person), restecg(represents values from ECG machine),thalach(represents maximum heart rate achieved), exang(represents the discomfort that is noted when heart does not get enough oxygen due to exercise), oldpeak(represents ST depression induced by exercise relative to rest), slope(represents slope of the ECG machine), ca(which represents the number of blood vessels),thal(a blood disorder passed down through family) and the target values which is 0 if a person is not having heart disease and 1 if the person is having heart disease.

III. Data Preprocessing

A. Data Cleaning

As the process of data cleaning all the null and zero values should be eliminated.In this data set, all 14 attributes have numerical data. Only, oldpeak contains float datatype all others are integers. None of the attributes have null or zero values.

B. Class Imbalance

Out of 1025 records, 526 records belong to a positive class and 499 records belong to negative class. So there is no class imbalance problem in this data set.

C. Data distribution and outlier detection

Data distribution for various attributes are shown using histograms. Then correlation matrix is used to find correlation among different attributes.Chest pain and thalach which is maximum heart rate achieved is the most important of all the attributes according to the correlation matrix.For outlier detection boxplot is used which shows how many values are suspected outliers

IV. Feature Selection

[3] For feature selection, SelectKBest feature selection technique which selects best n features from all the features. But again we cannot comment on which features are more important than others because we cannot know the actual cause of disease. After feature selection, normalization is performed on

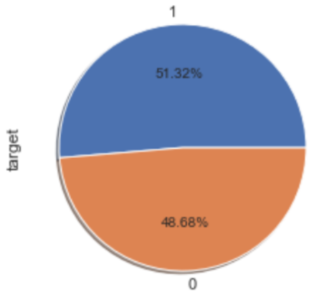

Fig. 1. Distribution of data into two classes

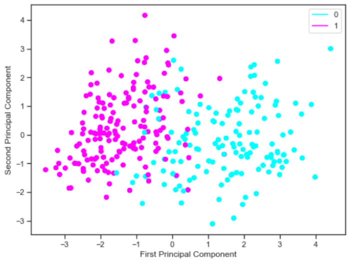

Fig. 2. Principle Component Analysis

the data using standard scalar. For data visualization, Principle Component Analysis is used which selects two components from all the features which give the most variation in our data set. After analysis data is coming linearly separable.In this dataset there are two weak attributes exang and oldpeak. Exang is weak feature because we cannot predict whether a person is having heart disease or not on the basis of discomfort a person is feeling at the time of exercise.

V. Data Mining Technique used

A. Support Vector Machine

[1] SVM is a supervised machine learning algorithm that classifies data using a plane called hyperplane. As it uses supervised learning so class labels are provided into it and SVM gives optimal hyperplane which classifies new unknown data. It uses different types of kernels for different classification problems. As the data is linearly separable so here kernel used is linear.As all the outliers are dropped and then SVM is applied so accuracy is coming about .91.Also, cross validation is used and accuracies after every validation is plotted in graph.

B. Logistic Regression

Logistic Regression is another type of supervised learning model which uses binary classification. It is one of the simplest and most efficient machine learning algorithm which is used in a wide variety of applications. It uses variables that are binary such as true or false, 1 or 0, yes or no, etc. It uses a sigmoid function for classification. Logistic regression here is used for classification but can also be used for regression models.As logistic regression works best for binary classification so it is giving accuracy about .90. After applying cross validation graph is plotted against different accuracies.

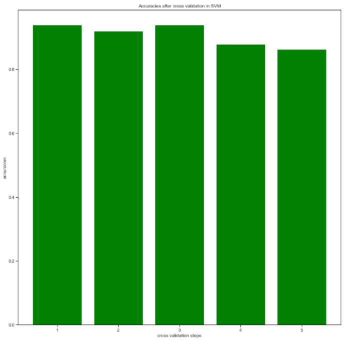

Fig. 3. Accuracy after applying cross validation in SVM

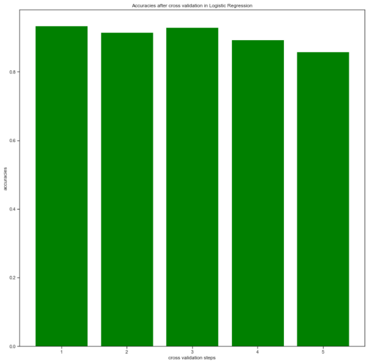

Fig. 4. Accuracy after applying cross validation in Logistic Regression

C. Random Forest Classifier

[2] Random forest is one type of ensemble classifier which uses decision tree classifier as the underlying classifier. It splits the data into multiple sets and then these sets are fed into the decision tree which provides output. This classifier uses a weak classifier to build a string classification model. Also, it solves the problem of overfitting and also handles the missing values.Accuracy is coming better by using this classifier.Accuracy is coming about .92.

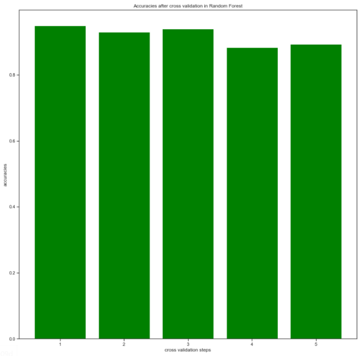

Fig. 5. Accuracy after applying cross validation in Random Forest

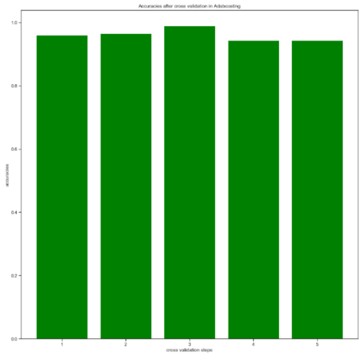

Fig. 6. Accuracy after applying cross validation in Adaboost

D. Adaboost Classifier

Adaboost classifier is another type of ensemble classifier which uses different classifiers and merges the results from all the classifiers and gives output. As by using individual classifiers, we might not get the accurate classification but by applying all the classifiers and using votes to decide what will the output will give more accurate results. This is why Adaboost perform well in the classification and accuracy is coming best for this classification. After cross validation graphs are plotted for different accuracies.

VI. Equation Used

For finding accuracy true positive rates(TPR), true negative rate(TNR), false positive rate(FPR) and false negative rates(FNR). Accuracy= TN+TP/TP+TN+FP+FN TPR=TP/TP+FN FPR=FP/TN+FP

VII. Implementation and Results

A. ROC

After applying different classifiers, the most important thing is performance measurement. In this implementation, Receiver Operating Characteristics or ROC curve is used for performance analysis of classification models.Roc curve implies how much our classification model is capable of differentiating between the classes.Graph is plotted against true positive rate and false positive rate. If the area under curve is more then the classifier performs best.In my result,the area under curve is best is coming for Adaboost classifier because it is ensemble classifier which uses many weak classifiers to create strong classifier.

B.

In this paper, different classifiers are used to classify whether a person is having heart disease or not.Among SVM, Random forest , Logistic Regression and Adaboosting, Adaboost classifier is performing best with accuracy about .97.after that Random Forest is giving best accuracy. Because both the Adaboost and Random Forest are ensemble classifiers they are performing better than other classifiers.

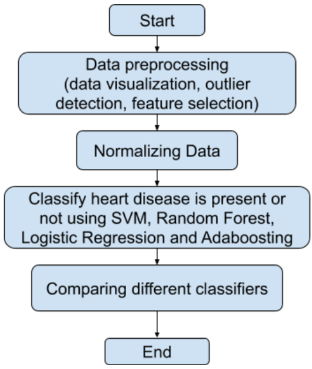
Start
Data preprocessing (data visualization, outlier detection, feature selection)
Normalizing Data
Classify heart disease is present or not using SVM, Random Forest, Logistic Regression and Adaboosting
Comparing different classifiers
End

Fig. 7. Implementation

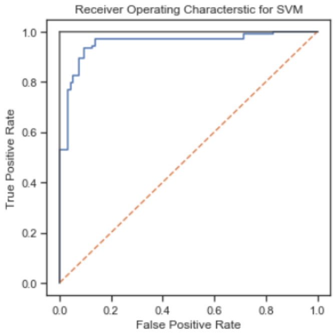

Fig. 8. ROC curve for SVM

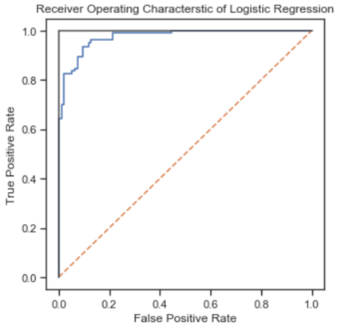

Fig. 9. ROC curve for Logistic Regression

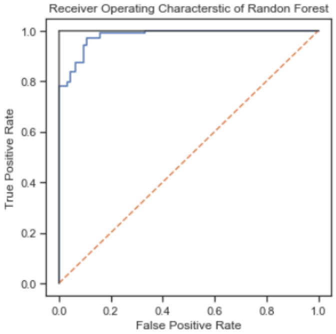

Fig. 10. ROCcurve for Random Forest

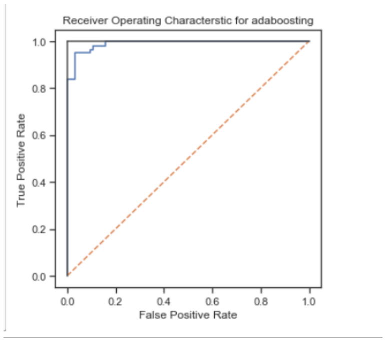

Fig. 11. ROC curve for Adaboosting

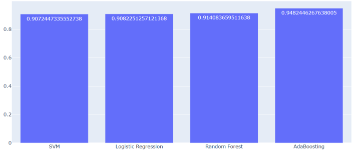

Fig. 12. Comparing accuracies of all the four classifier

VIII. Conclusion and Future Work

[4] The main objective of this paper was to build a classifier which can best classify whether a person has heart disease or not.Different data mining techniques are used so that we can get the best classifier which will help for the better and early treatment of the patients suffering from heart disease.Using PCA and some feature selection techniques, best attributes which affects the most is selected and different models and trained and tested among which Adaboosting is giving the best performance.Still there can be many reasons for heart disease we cannot actually say that a particular feature is more important than others because it is natural phenomena and can occur any time.So, in future what we can do is that we can merge our data mining techniques with IOT. We can use sensors for sensing any unusual activity of the body and can immediately report to our classifier so that the person can be informed about unusual activities as soon as possible.In future, number of attributes can be increased by taking some other attributes like whether a person smokes or not a=or how many cigarettes does a person takes per day.

References

[1] H A Esfahani and M Ghazanfari. “Cardiovascular disease detection using a new ensemble classifier”. In: *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*. Dec. 2017, pp. 1011–1014. DOI: 10.1109/KBEI.2017.8324946.

[2] B S S Rathnayakc and G U Ganegoda. “Heart Diseases Prediction with Data Mining and Neural Network Techniques”. In: *2018 3rd International Conference for Convergence in Technology (I2CT)*. Apr. 2018, pp. 1–6. DOI: 10.1109/I2CT.2018.8529532.

[3] G Shanmugasundaram et al. “An Investigation of Heart Disease Prediction Techniques”. In: *2018 ieee international conference on system, computation, automation and networking (icscan)*. July 2018, pp. 1–6. DOI: 10.1109/ICSCAN.2018.8541165.

[4] J Thomas and R T Princy. “Human heart disease prediction system using data mining techniques”. In: *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. Mar. 2016, pp. 1–5. DOI: 10.1109/ICCPCT.2016.7530265.