Heart Failure Prediction Using Predictive Models

Animesh Guchhait

January 05, 2022

Abstract

This paper describes various methods of exploratory data analysis along with predictive modeling for predicting the heart disease. Data mining and machine learning modeling plays an important role in building an important model for medical system to predict heart disease or cardiovascular disease. Medical experts can help the patients by detecting the cardiovascular disease before occurring. Now-a-days heart disease is one of the most significant causes of fatality. The prediction of heart disease is a critical challenge in the clinical area. But time to time, several techniques are discovered to predict the heart disease using various predictive models.This paper aims to describe some predictive modeling techniques and train those model on a suitable heart disease data and compare their results.

1 Introduction

Over the last decade, heart or cardiovascular disease remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths,80% are because of coronary artery disease and cerebral stroke[1].The vast number of deaths is common amongst low and middle-income countries[2]. Many predisposing factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various habitual risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are predisposing factors for heart disease. The efficient and accurate and early medical diagnosis of heart disease plays a crucial role in taking preventive measures to prevent death.

Data mining refers to the extraction of required information from huge data sets in various fields such as the medical field, business field, and educational field.Machine learning is one of most rapidly evolving domain in artificial intelligence. Using various machine learning algorithms, we can extract hidden complex pattern from huge data set in various fields.In medical field, there are lots of huge data sets to analyze their complex pattern.By using suitable machine learning algorithm, a computer can easily extract hidden crucial decision making information from a collection of a past repository for future analysis.Healthcare professionals do analysis of these data to achieve effective diagnostic decision.

In the rest of the paper, I discuss about some predictive machine learning model like decision tree, random forest, naive bayes classifiers and logistic regression and how those models performs on a heart disease data.

1.1 Problem Statement and Data set

In machine learning techniques, models learn from the past data and try to extract hidden crucial decision making information for future analysis.The Heart Disease data which is used for predictive models, has 918 observations with 12 attributes and is collected from Kaggle. The attributes are the following:

Table 1: Attributes and details of dataset of heart disease

Sr. no.	Attribute	Details
1	Age	age of the patient [years]
2	Sex	sex of the patient [M: Male, F: Female]
3	ChestPainType	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4	RestingBP	resting blood pressure [mm Hg]
5	Cholesterol	serum cholesterol [mm/dl]
6	FastingBS	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7	RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8	MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]
9	ExerciseAngina	exercise-induced angina [Y: Yes, N: No]
10	Oldpeak	oldpeak = ST [Numeric value measured in depression]
11	ST Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12	HeartDisease	output class [1: heart disease, 0: Normal]

Objective : Objective of this paper is to predict the above mentioned 'Heart Disease' attribute from rest of the attributes.So, different predictive models are to be trained using the above mentioned data to predict 'Heart disease' and their results will be compared.

2 Methodology

Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technology that helps computers to build and classify various attributes. This research paper uses classification techniques to predict heart disease. This section gives a portrayal of the related subjects like machine learning and its methods with brief descriptions.

2.1 Machine Learning

Machine learning is an emerging subdivision of artificial intelligence. Its primary focus is to design systems, allow them to learn and make predictions based on the experience.

It trains machine learning algorithms using a training dataset to create a model. The model uses the new input data to predict heart disease. Using machine learning, it detects hidden patterns in the input dataset to build models. It makes accurate predictions for new datasets. The dataset is cleaned and missing values are filled. The model uses the new input data to predict heart disease and then tested for accuracy. Machine learning techniques can be classified in three categories. Those are 1. Supervised learning , 2. Unsupervised learning and 3. Reinforcement learning. In this paper, I use some supervised learning models for prediction.

Supervised Learning:

The model is trained on a dataset that is labelled. It has input data and its outcomes. Data are classified and split into training and test dataset. Training dataset trains our model while testing dataset functions as new data to get accuracy of the model. The classification and regression are its example.In this heart disease data we can apply classification models for prediction.

2.2 Classification Machine Learning Techniques

Here I describe some supervise learning classification method that I use for modelling

2.2.1 Decision Tree

Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

- Root node: main node, based on this all other nodes functions.
- Interior node: handles various attributes.
- Leaf node: represent the result of each test.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having minimum entropy or gini index:

$$Entropy(S) = \sum_{i=1}^{c} -P_i \log(P_i)$$

$$Gini(S) = 1 - \sum_{i=1}^{c} {P_i}^2$$

where c is number of type of categories, S is the state in the decision tree where next branching will be performed and P_i are the weightage average of probability of finding ith category in the next branches.
The results obtained are easier to read and interpret and most of time this model performs well in labelled datasets.

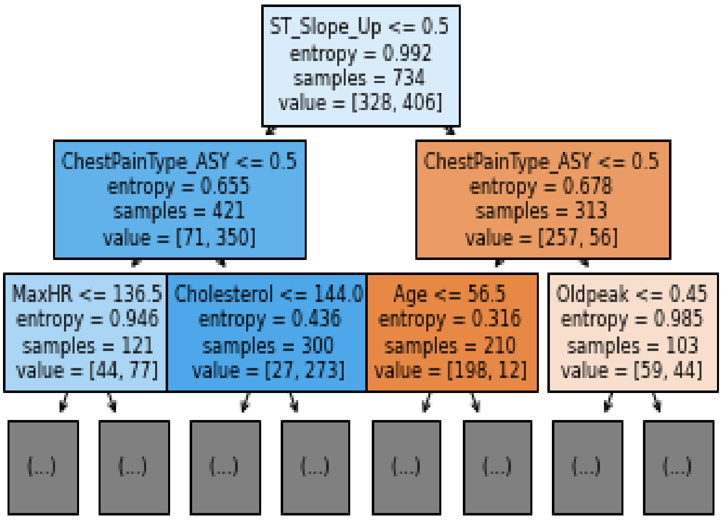

Figure 1 Visual Representation of Decision Tree

2.2.2 Random Forest

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several decision trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more number of trees give higher accuracy.It is used for classification as well as regression task, but can do well with classification task. It is slow to obtain predictions as it requires large data sets and more trees.

Feature Importance: Feature Importance is a important tool in Random Forest to measure the relative importance of each feature by looking at how much the tree nodes that use that feature reduce impurity on average(across all trees in the forest). More precisely, it is a weighted average, where each node's weight is equal to the number of training samples that are associated with it.

2.2.3 Naïve Bayes' Classifier

Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes strong (Naive) independence among attributes. Bayes theorem is a mathematical concept to get the probability. The predictors are neither related to each other nor have correlation to one another. All the attributes independently contribute to the probability to maximize it. It is able to work with Naïve Bayes model and does not use Bayesian methods.

$$P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}$$

P(X/Y) is the posterior probability, P(X) is the class prior probability, P(Y) is the predictor prior probability, P(Y/X) is the likelihood, probability of predictor. Naïve Bayes is a simple, easy to implement, and efficient classification algorithm that handles non-linear, complicated data. However, there is a loss of accuracy as it is based on assumption and class conditional independence.But on many real life datasets, this algorithm performs very well.

2.2.4 Logistic Regression

Logistic regression is one of the most popular Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function(sigmoid function), which predicts two maximum values (0 or 1).The curve from the logistic function indicates the likelihood of something such as whether heart disease output class is 0 or 1.It is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

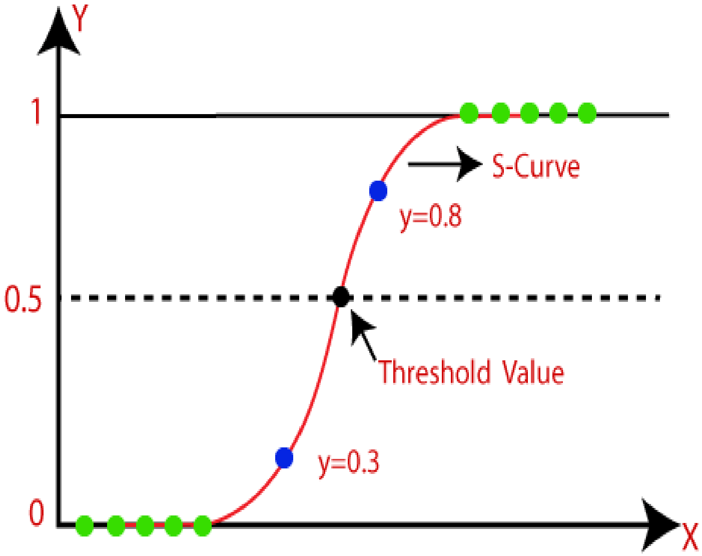

Figure 2 Sigmoid curve of logistic regression

3 Results and Analysis

Aim of this research is to predict whether or not a patient will develop heart disease and to find relationship among these features and heart disease using EDA. This research was done on supervised machine learning classification techniques using Naïve Bayes, decision tree, random forest, and Logistic Regression on UCI heart disease data(collected from Kaggle). This research was performed on the device of 'Intel(R) Core(TM) i3-1005G1 CPU @ 1.20GHz 1.19 GHz' processor configuration with 4 GB RAM.

3.1 EDA and results:

I get some useful results from EDA of this data. It is clear from the Figure 3 that man are suffering from heart disease more than female.

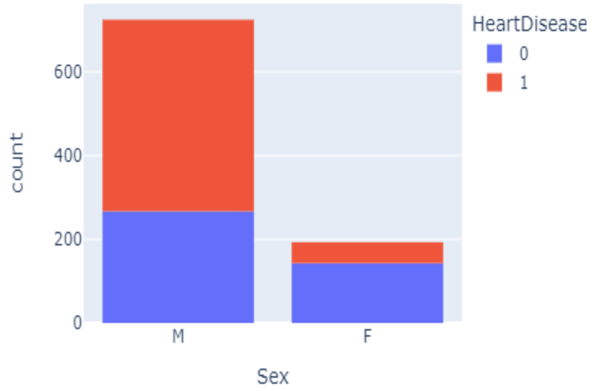

Figure 3 Bar plot of Sex

According to the plot, the result of cases:

Male: 725 cases(79%), 458 of them have heart disease(63%)

Female: 193 cases(21%), 50 of them have heart disease(26%)

So it is clear that male are suffering more than female.

Figure 4 shows the bar plot for Chest Pain type. According to the plot:

ATA(Atypical angina): 173 cases (20%), 24 of them have heart disease (14%)

NAP(Non-anginal pain): 203 cases (22%), 72 of them have heart disease (35%)

TA(Typical angina) : 46 cases (5%), 20 of them have heart disease (43%)

ASY(Asymptomatic): 496 cases (53%), 392 of them have heart disease (80%)

About 80% of the asymptomatic people has heart disease problem and that means most of them are healthy apparently!

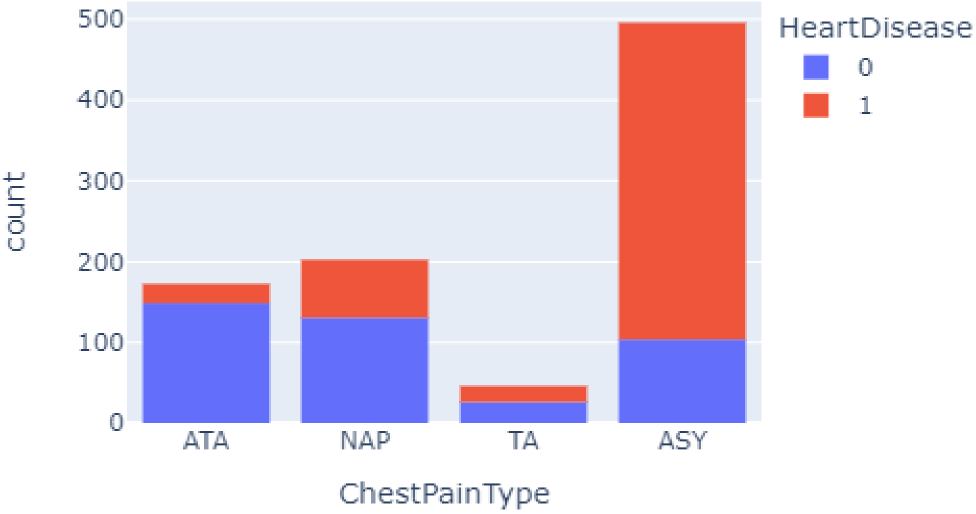

Figure 4 Bar plot of Chest Pain Type

According to the bar plot of ST slope in Figure 5:

UP: 395 cases (43%), 78 of them have heart disease(20%)
Flat: 460 cases (50%), 381 of them have heart disease(82%)
Down: 63 cases (7%), 49 of them have heart disease(77%)

We can get cases with 'Flat' and 'Down' ST slop are in high risk of heart disease.

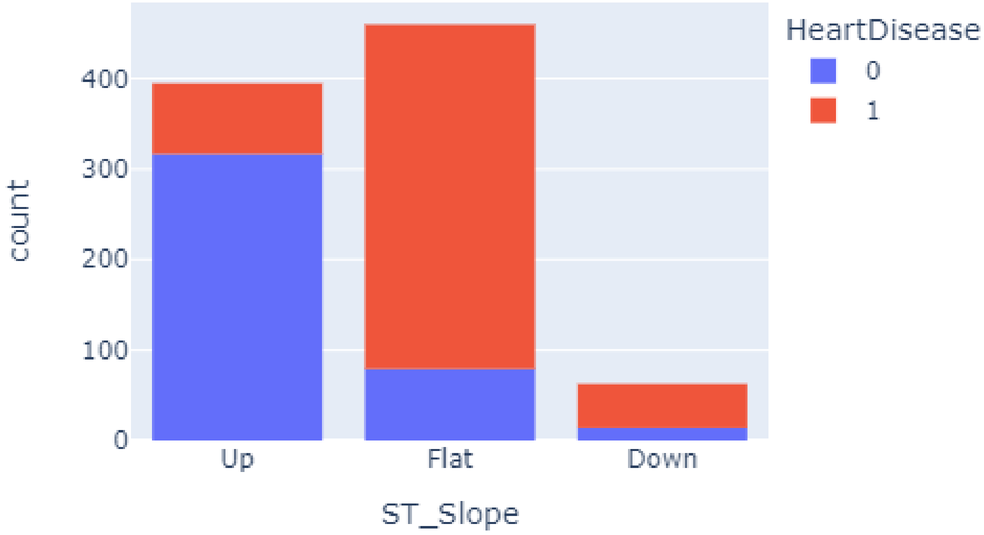

Figure 5 Bar plot of ST Slope

Now, in Figure 6, we can see bar plot of Exercise angina and from where we can see that:

'Exercise Angina =Yes' has 371(40%) cases where 316 (85%) of them have heart disease.

'Exercise Angina =No' 547(60%) cases where 192 (35%) of them have heart disease.

So We can say that if a person has exercise angina, there is a high chance that he has heart disease.

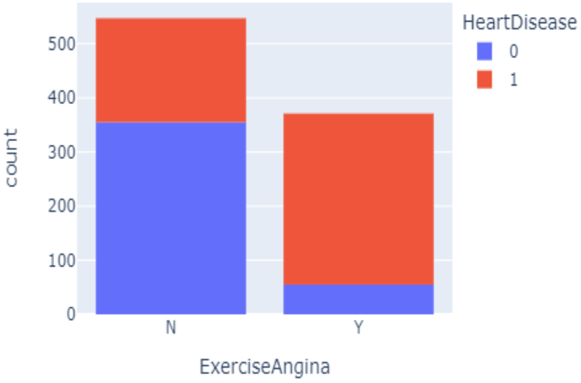

Figure 6 Bar plot of Exercise Angina

In Figure 7, We can see strip plot of ST-Slope and Exercise Angina.From where we can say that

If the result of exercise angina for somebody is 'yes' and her/his ST slope is 'flat', she/he is more in danger of heart disease. It's true for ST slope = 'down' too.

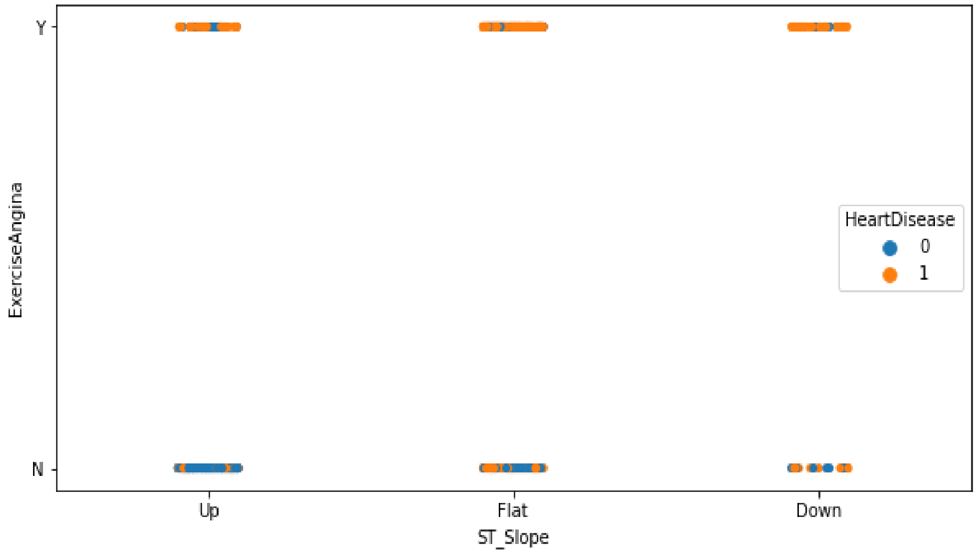

Figure 7 Strip plot of ST-Slope and Exercise angina

Now, in Figure 8, We can see box plot of Maximum Heart rate (Max HR) and Exercise Angina.From where we can say:

Persons who have their exercise angina test as 'Yes', have lower maximum heart rate than person having 'No' as exercise angina test.We also can say that heart patients have lower heart rate compare to healthy persons.

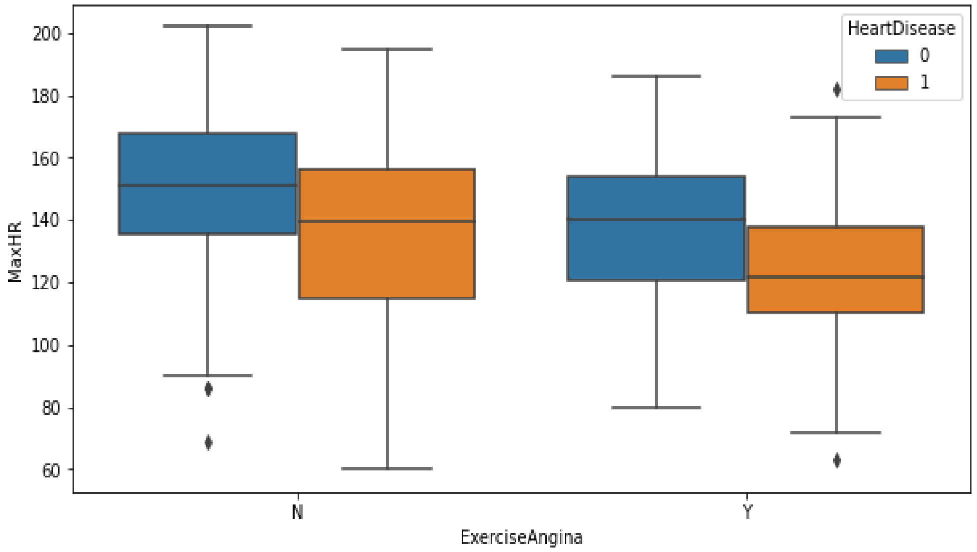

Figure 8 Box plot of Max HR and Exercise Angina

3.2 Classification Techniques results:

Data set is split into train and test data set and data pre-processing is performed as necessary.Supervised classification techniques such as Naïve Bayes, decision tree, random forest and Logistic Regression are applied to get accuracy score. The accuracy score results of different classification techniques were noted using Python Programming for test data sets.Percentage accuracy scores are depicted in Figure 9 for different algorithms.

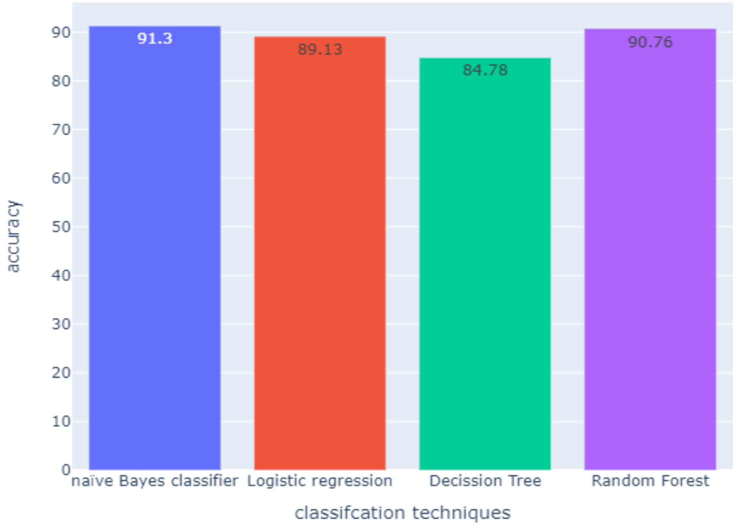

Figure 9 Comparative result of classification techniques

So,from Figure 9, we can see that Naïve Bayes classifier gives the best accuracy and then Random Forest classifier.

We already know that, from Random Forest classifier, we can also get to know about the important features for the classification.Figure 10 shows bar plot of mean decrease in impurity of each features.Which feature has greater mean decrease in impurity, is more important.

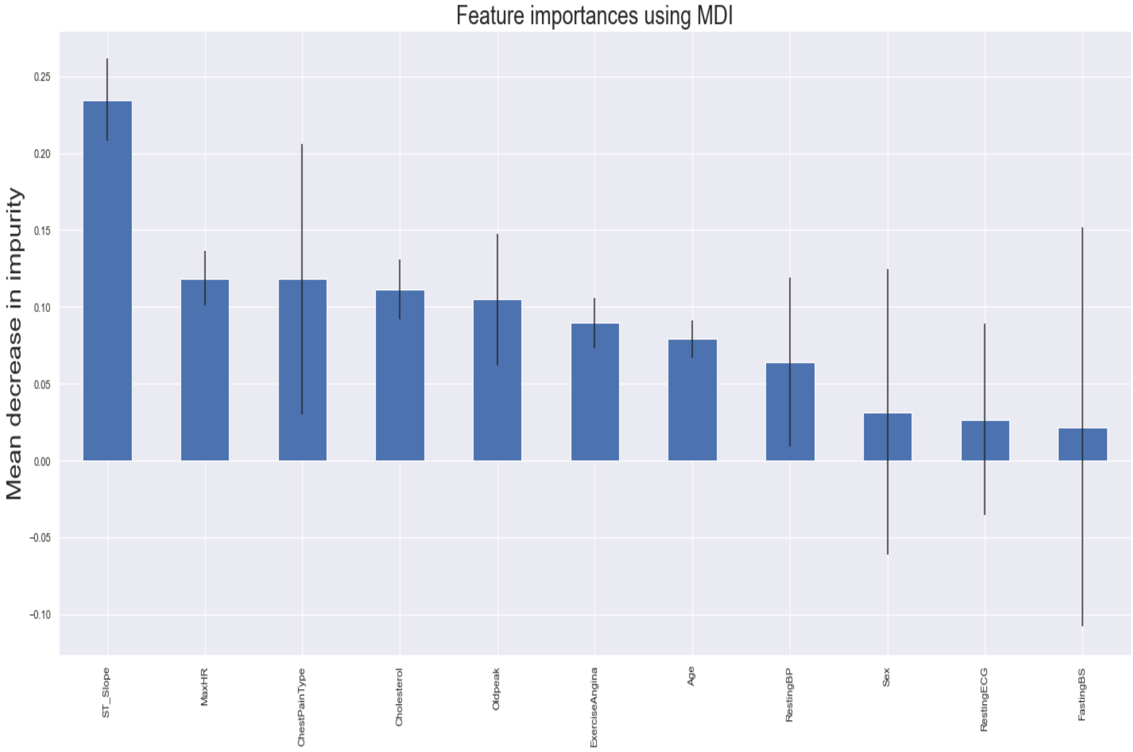

Figure 10 Bar plot for features importance

From plot, we can see that ST-Slope is the most important features for predicting heart disease and then MaxHR, chest pain type, cholesterol and old peak have more importance than other features.

4 Conclusion

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes is our goal. In this study, I consider 12 essential attributes.By EDA, I find some beautiful observations helpful for treatments and detection of heart disease. I applied four data mining classification techniques,Naïve Bayes, decision tree, Random forest and Logistic Regression and get best accuracy for Naïve Bayes classifier (91.30%). We can further expand this research incorporating other data mining techniques such as time series, clustering and neural network in hope of better accuracy.Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.

References

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
[2] Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low-and middle income countries. Curr Probl Cardiol. 201
[3] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron
[4] C4.5: PROGRAMS FOR MACHINE LEARNING J. ROSS QUINLAN
[5] Link of my notebook code of this project