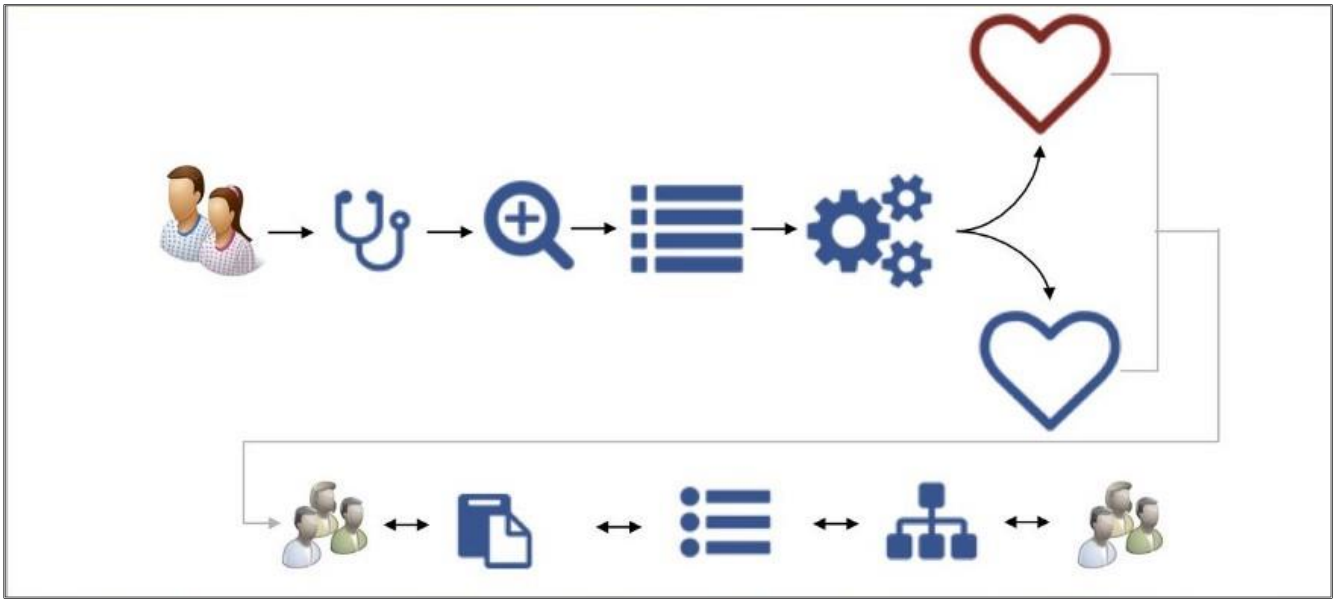
Symbiosis Skills and Professional University

Kiwale, Pune

PROJECT REPORT

On

"Heart Disease Prediction using Machine learning algorithms"



Submitted by

Vaibhav Arjun Jumde

ML-Batch-III

Under The Guidance of

Trainers' Name: 1) Mr. Sanjay Bhorekar 2) Dr. Ruby Jain

STUDENT DECLARATION AND ATTESTATION BY TRAINER

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources. All information included from other sources has been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

Signature of student

Name of student: Vaibhav Arjun Jumde

Registration Number:

Signature of trainer

Name of trainer:

CERTIFICATE

This is to certify that the report entitled, "**Heart Disease Prediction using Machine Learning Algorithms**" submitted by "**Vaibhav Arjun Jumde** to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a record of bonafide Project work carried out by him under my supervision and guidance and is worthy of consideration for the completion of certificate course in 'Machine Learning Engineer".

Signature of Trainer
Name of Trainer

Date: / / 2021

_____ _____

Supervisor Supervisor

Date:

Abstract

Heart related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need of reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. In this project we perform various algorithms that is based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Random Forest (RF)

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. In this project we are doing classification techniques that is use to predict heart disease. This section gives a portrayal of the related subjects like machine learning and its methods with brief descriptions, data pre-processing, evaluation measurements and description of the dataset used in this research. The diagnosis of heart disease in most cases depends on a complex combination and huge volume of clinical and pathological data. Machine learning has been shown to be effective assisting in making decisions and predictions from the large quantity of data produced by the health care industry.

Sr. No	Index		Page No.
1	Acknowledgment		6
2	Introduction		7
3	Objectives		8
4	Description Of Data And Method		9
5	Method and Algorithm Used		10-13
6	Statistical Terminologies		14-15
7	Data Pre-processing		16
8	Exploratory Data Analysis		
	8.1	Count Plot	17
	8.2	Heatmap Plot	18
9	Code		20
10	Conclusion		21
11	References		22

ACKNOWLEDGEMENT

It is really a matter of pleasure for me to get an opportunity to thank all those who contributed directly or indirectly for the successful completion of the project report, "**Heart Disease Prediction using Machine Learning Algorithms**".

My sincere thanks to Mr. Sanjay Bhorekar sir and Dr. Ruby Jain madam who took keen interest to give important information & valuable guidance.

I would like to thank all those who offered this course for their support and co-operation. Without their support we would not have been able to complete our project under this course.

Introduction

As indicated by the World Health Organization, consistently 12 million passings happen worldwide because of Heart Disease. The heap of cardiovascular illness is quickly expanding everywhere on the world from the previous few years. Numerous explores have been directed in endeavor to pinpoint the most powerful factors of coronary illness just as precisely anticipate the general danger. Coronary illness is even featured as a quiet executioner which prompts the passing of the individual without clear side effects. The early conclusion of coronary illness assumes a crucial part in settling on choices on way of life changes in high-hazard patients and thusly decrease the complexities.

Heart is a significant organ of the human body. It siphons blood to all aspects of our life structures. Assuming it neglects to work effectively, the mind and different organs will quit working, and inside couple of moments, the individual will pass on. Change in way of life, business related pressure and awful food propensities add to the increment in pace of a few heart related sicknesses. The European Public Health Alliance detailed that cardiovascular failures, strokes and other circulatory sicknesses represent 41% of all passings (European Public Health Alliance 2010). A few unique manifestations are related with coronary illness, which makes it hard to analyze it speedier and better. Chipping away at coronary illness patients information bases can measure up to genuine application. Specialists information to allot the load to each ascribe. More weight is relegated to the quality exceptionally affecting sickness expectation. In this manner it seems sensible to have a go at using the information and experience of a few experts gathered in data sets towards helping the Diagnosis cycle. It likewise gives medical care experts an additional wellspring of information for deciding.

Objectives

The main objective of developing this project are:

1. The main purpose of this project is how many people suffering from heart disease, by implementing various machine learning Algorithm.
2. To determine accuracy by applied algorithms on that particular medical dataset which may lead to heart disease.
3. To determined accuracy through analyzing the data.

Description of Data and Method

The dataset is publicly available on the github Website, Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. Heart disease are the diverse conditions by which the heart is affected. This dataset contains 303 rows and 14 columns. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol etc. Where each row corresponds to a single record, All attributes are listed in 'Table 1'.

Table 1. Various Attributes used are listed

Sr. No	Observation	Discription	Values
1	Age	Age in years	Continous
2	Sex	Sex of subject	Male/Female
3	Cp	Chest Pain	Four Type
4	Trestbps	Resting Blood Pressure	Continous
5	Chol	Serum cholesterol	Continous
6	Fbs	Fasting Blood Sugar	< , or > 120 mg/dl
7	Restecg	Resting electrocardiograph	Fivbe Values
8	Thalach	Maximum Heart Rate Achieved	Continous
9	Exang	Exercise Induced Angina	Yes/No
10	Oldpeak	ST Depression when Workout compared to the amount of Rest Taken	Continous
11	Slope	Slope of Peak Exercise ST segment	Up/Flat/Down
12	Ca	Gives the number of major vessels Fluroscopy	0-3
13	Thal	Defect type	Fixed/Reversible/Normal

Methods and Algorithms Used

In this project I have applied machine learning algorithm to classify whether a person is suffering from heart disease or not. For that prediction purpose I have used Random Forest and Knn algorithm that is described below in details.

- **KNN Algorithm:**

KNN is stands for K- Nearest Neighbors Algorithm works by finding the distances between a query and all the examples in the data, selecting the specified number.

For examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

The K-NN classifier calculates the distances between the point and points in the training data set. Usually, the Euclidean distance is used as the distance metric.

The following two properties would define KNN well:

- **Lazy learning algorithm**:

  KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

- **Non-parametric learning algorithm**:

  KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

- **Euclidean Distance:**

To find the Euclidean distance from this formula shown in below figure:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

when p = 2,

when p is set to 2 we get Euclidean distance

K-NN work based on euclidean distance, so the data needs tro be scaled we select K=5 and find the nearest neighbors. The new point belong to class which has the Highest neighbors, In this example for two point for class A and 3 points for class B, so the new point would be assigned to class B

- **SVM Algorithm:**

SVM (Support Vector Machine) is a supervised machine learning algorithm that is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyper plane, which acts like a decision boundary between the various classes. SVM can be used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data.

Support vectors are the data points nearest to the hyperplane, points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the as possible, while still being on the correct side of it. So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

- **Linear SVM:**

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:**

Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non linear data and classifier used is called as Non-linear SVM classifier.

➢ Random Forest Algorithm:

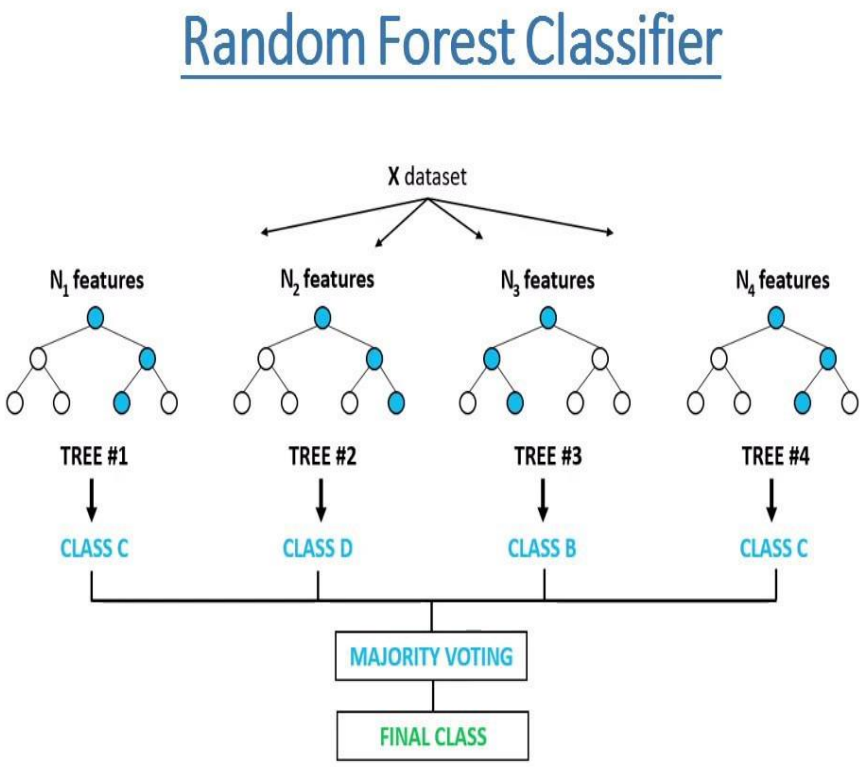
Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more number of trees give higher accuracy. Random forest is a supervised learning algorithm which is used for both classification as well as regression.

But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on

data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

The "forest" it builds, is an ensemble of decision trees,usally trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

The below figue shows the graphical representation of Random Forest Algorithm:



Statistical Terminologies

- **Heatmap Plot** :

  A Heatmap is a two-dimensional representation of data in which values are represented by colours.

- **KNN Classifier** :

  K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

- **Random Forest Classifier :**

  Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

- **Confusion Matrix** :

  A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

	TRUE POSITIVE	FALSE POSITIVE
Positive :	33	8
	FALSE NEGATIVE	**TRUE NEGATIVE**
Negative:	7	43

- **Standardization** :

  In statistics, standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables. Typically, to standardize variables, you calculate the mean and standard deviation for a variable.

  The standardization formula is shown as below :

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean

$\sigma =$ Standard Deviation

Data Pre-Processing

Data pre-processing is important part of any statistical analysis. Because without data cleaning we might get result with less accuracy. Data pre-processing includes missing values handling,

So in this project we have performed following data pre-processing task:

- **Missing Value:**

  We have checked whether missing values are present or not. No missing values are found in the data.

  Data Pre-processing follow the given flow chart:

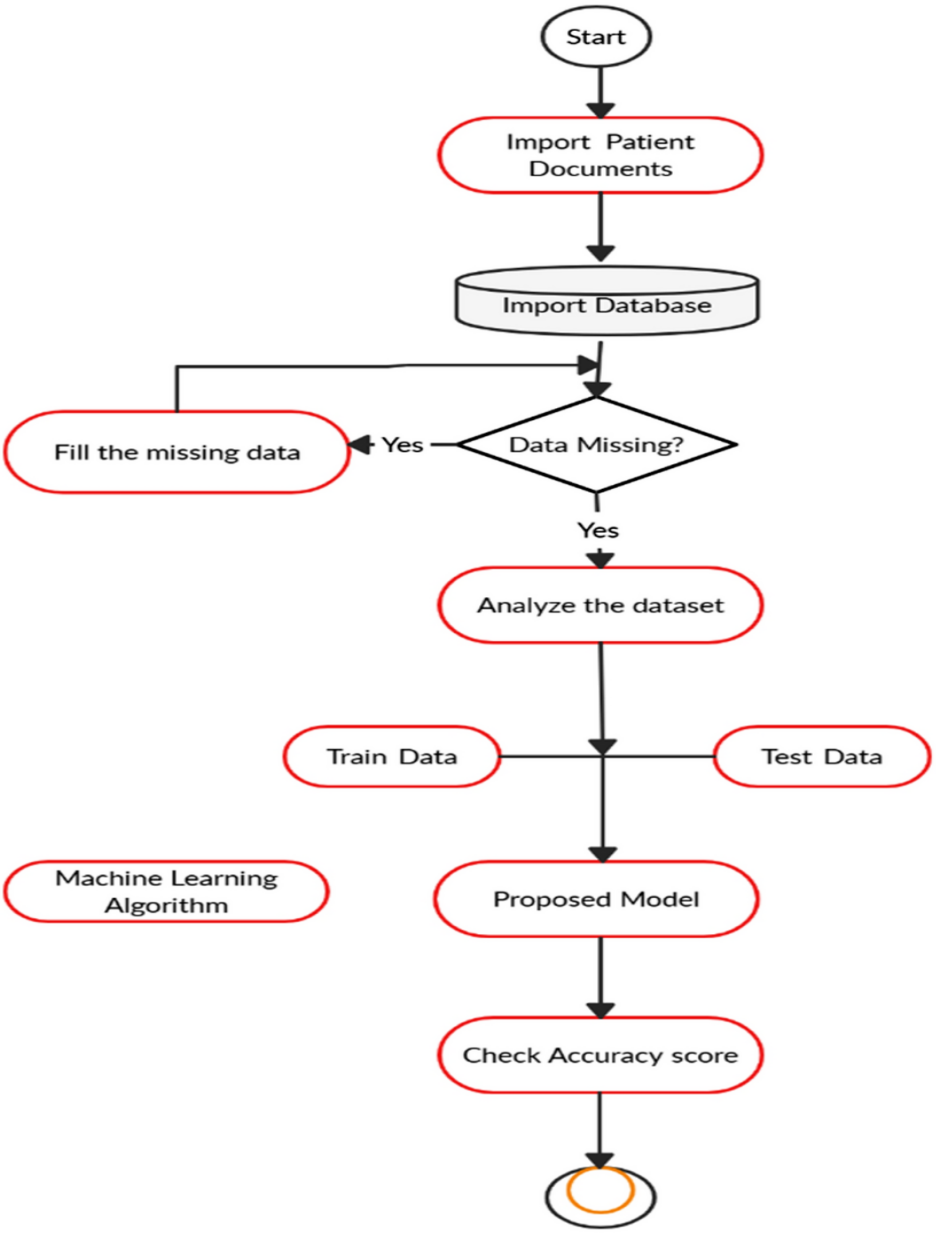
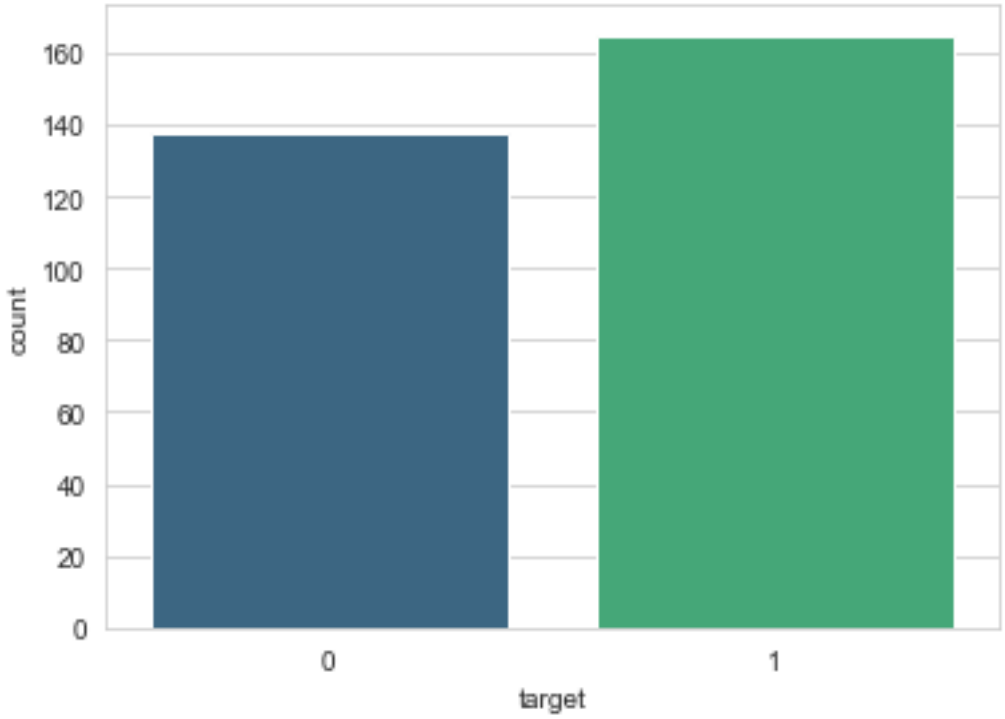


Fig. Data preprocessing

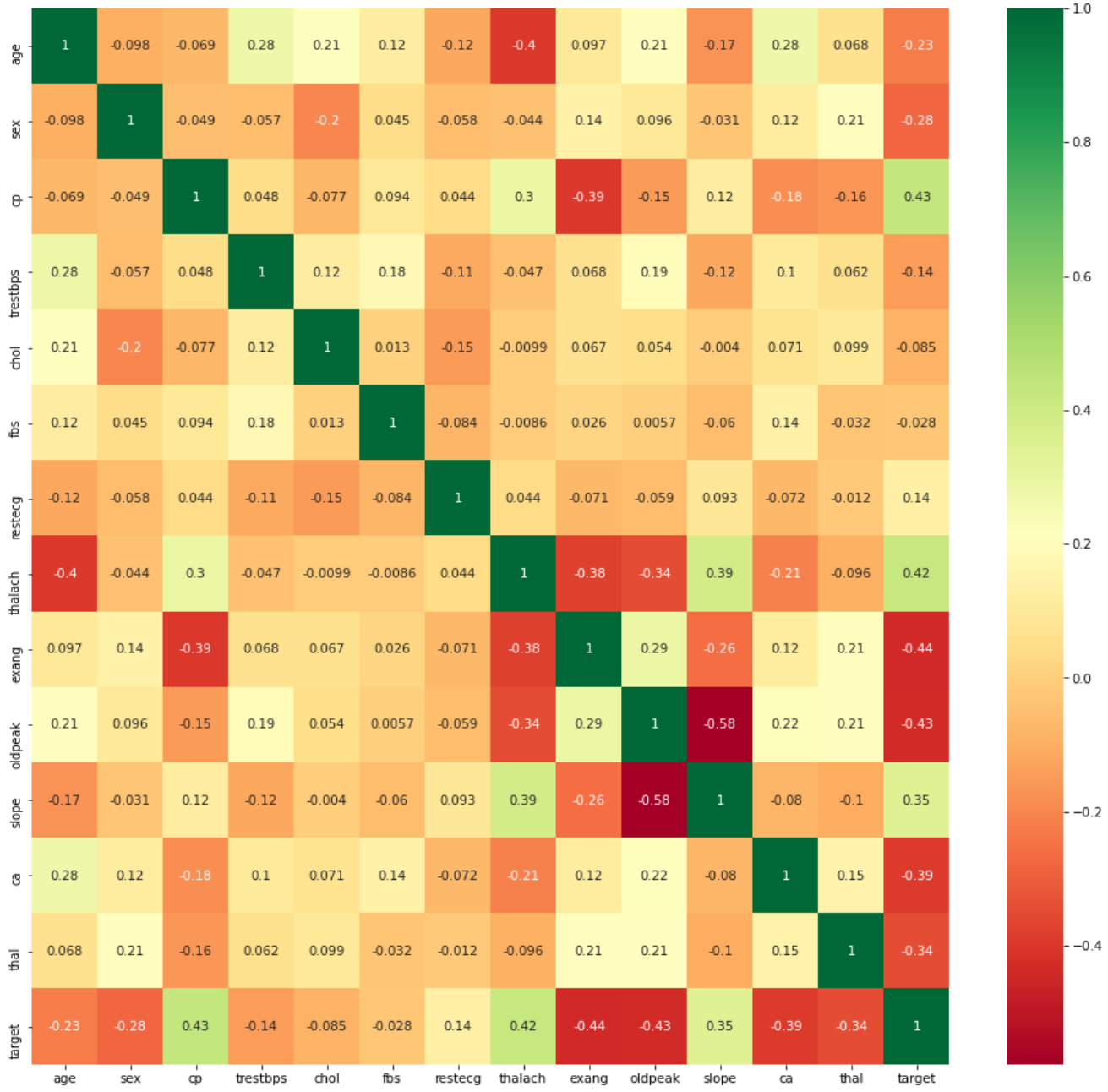
Exploratory Data Analysis

➢ Count Plot:

In Python, we can normally plot barcharts for numerical variables. when it comes to the case of categorical variables then we cannot normally plot the count for each category. It allows you to plot the count of each category for non-numerical/categorical variables. Here we plot the simple bar plot of Target variables from this plot we see the shape of dataset we We can also said the data is imbalance or balance.



- **To find the relationship between various parameters and target variables using Heatmap plot**



From the above heatmap, where the target classes are of approximately equal size. it's easy to see that a few features have negative correlation with the target value while some have positive. But there is higher relationship between cp, slope, thalach with the target variables.

Code

The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it. The code has been written in Python programming language using Jupyter Notebook as IDE. The experiments and all the models building are done based on python libraries.

The code is available in the Github repository given in following link:

https://github.com/vaibhav209

Libraries used:

- NumPy
- Pandas
- Matplotlib
- Sklearn

Conclusion

We have done project on Heart Disease Prediction using Machine Learning Algorithm to predicts a how many people suffering from heart disease. So by doing Statistical analysis we have achieved following conclusions.

I. By using KNN algorithm we have achieved 84% accuracy
II. By using SVM algorithm we have achieved 83% accuracy
III. By using Random forest algorithm we have achieved 81% accuracy

Hence, we can use KNN Algorithm for prediction

References:

[1] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.

[2] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine- learning-36f00f3edb2c. [Accessed 2 March 2020].

[3] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[4] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.