**BANA 6620**

**Group Project**

**A STUDY TO PREDICT HEART DISEASES AND THE FACTORS CAUSING THEM**

**Anurag Lahon**

**Ruthvik Reddy**

**Namratha Reddy K**

**Vaishnavi Yeluri**

**Group Project Documentation Page**

**Title: “***A Study to predict Heart Diseases and the factors causing them”*

## 1.**Python packages**: While completing the project we have used sklearn (scikit-learn), numpy, matplotlib, seaborn, plotnine and statsmodels packages.

## 2.**Data Preparation**: The dataset consists of 4238 rows and 16 columns.

A screenshot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generatedSummary of the dataset include those that summarize the central tendency, dispersion and shape of a dataset’s distribution, excluding NaN (Not a number) values.

## 3. **Handing Missing Values:**

Missing values are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Checking Missing Values: There are missing values in education, cigsPerDay, BPMeds, totChol, BMI, heartrate, glucose columns.

A screenshot of a cell phone

Description automatically generated

Filling Missing Values: Filling the missing value with foreward and backward fill because there are categorical values also.  Forward filling means fill missing values with previous data. Backward filling means fill missing values with next data point. We are not using the mean for missing value as there are categorical features which can’t take the mean. Forward and Backward fill is used to fill the null values, so the distribution is not affected.



## 4. **Data Understanding and EDA:**

Categorical column: male,education,currentSmoker,BPMeds,prevalentsStroke,prevalentHyp,diabetes and TenYearCHD.

Continuous column: age, cigsPerDay, totChol, sysBP, diaBP, BMI, heartrate and glucose.

TenYearCHD is the response variable while other features are predicator variable.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedIdeal BMI: A BMI of between 18.5 and 24.9 is ideal, which can be an important factor affecting heart disease.

A screenshot of a cell phone

Description automatically generated 44.43 % people have ideal BMI and 14.19% people have heart disease among ideal BMI.

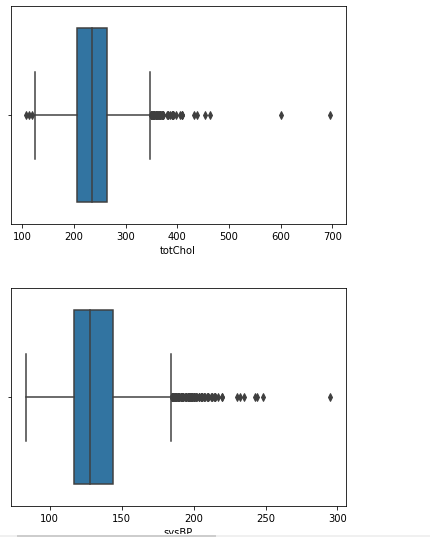
Correlation between different features:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

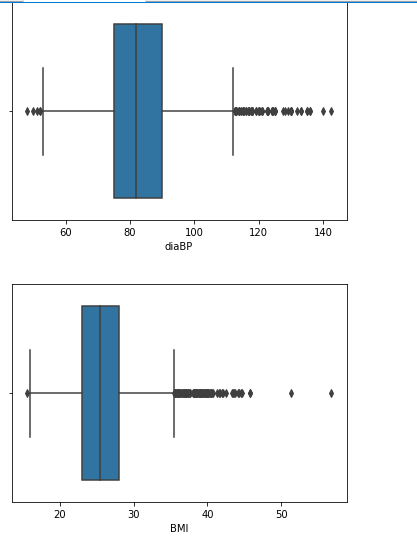
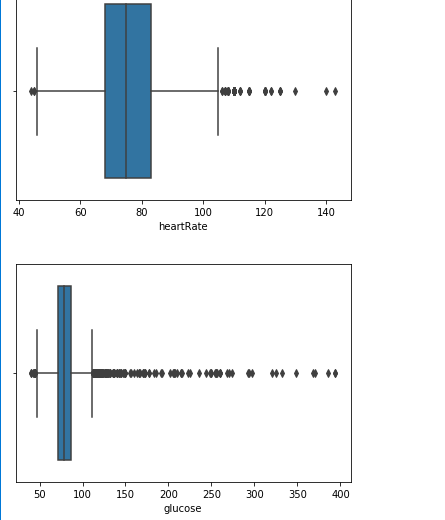
Description automatically generatedWhen correlation is 1 ,it is strong positive correlation(it is dark red as given in the scale above) and when it is -1 ,it is strong negative correlation(it is dark blue as given in the scale above).There is no correlation when correlation is 0.

Checking for outliers in the quantitative column of data set:

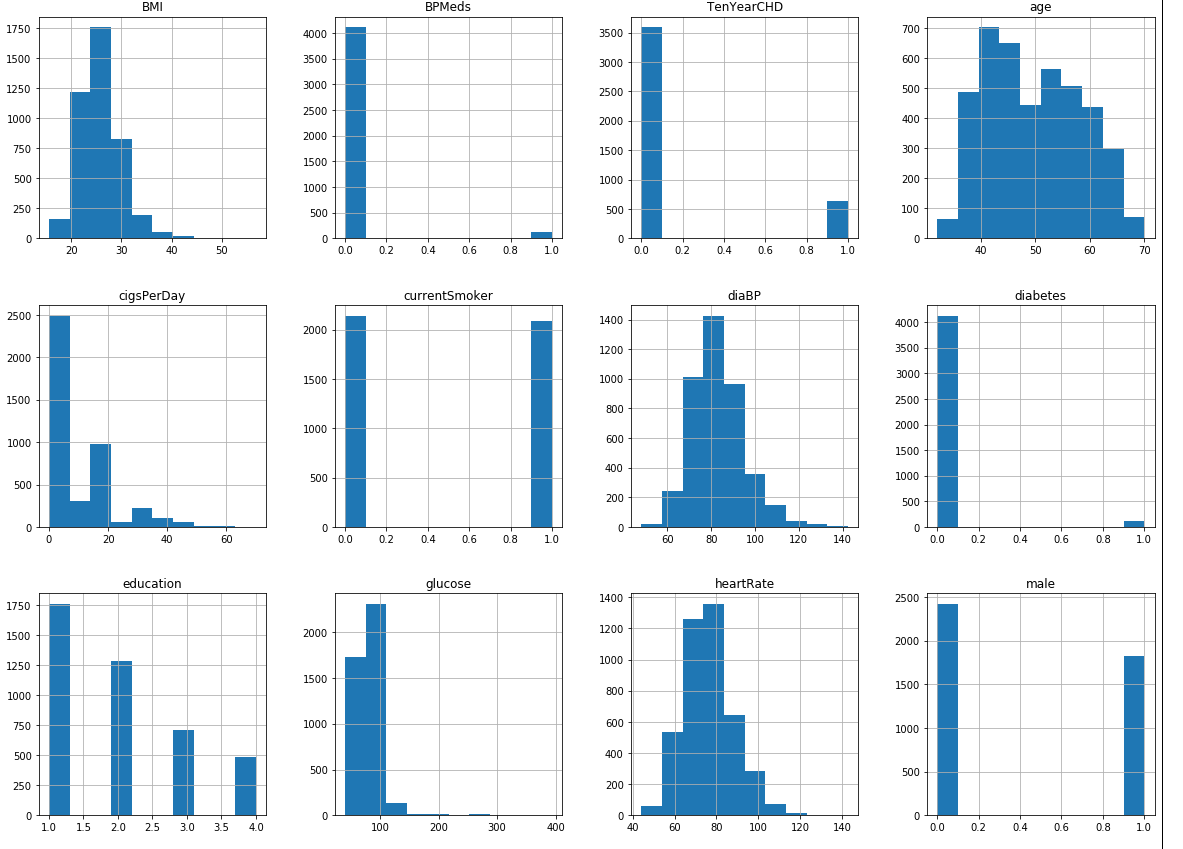
A picture containing clock

Description automatically generatedA screenshot of a cell phone

Description automatically generated



Histogram plots of the data:



A screenshot of a social media post

Description automatically generatedScatter plot between Age and heart Rate: Maximum number of heart diseases are found in upper age group (above 50 years).

A screenshot of a cell phone

Description automatically generatedScatter plot between BMI and glucose:

A screenshot of a cell phone

Description automatically generatedCategorical Plotting with TenYearCHD, age and male:

Scatter Plot between sysBP and glucose:

A close up of a map

Description automatically generated

Scatter plot between age and sysBP with hue as male: Men tend to have high systolic blood pressure (above 190). thus, it is possible that men are more likely to get heart disease.

A screenshot of a social media post

Description automatically generated

## Education and smoking habit: The trend in male and female are slightly different. In male, low-level of education have higher number of smokers as compared to female. The number of smokers decreases as the level of education increases.

## A screenshot of a cell phone Description automatically generated

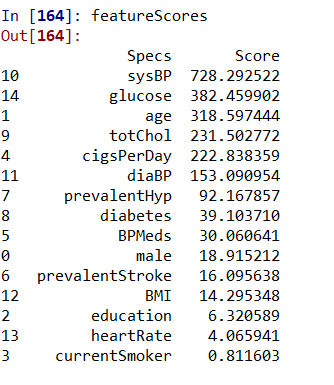
Smoking habits with Education level and male, female:

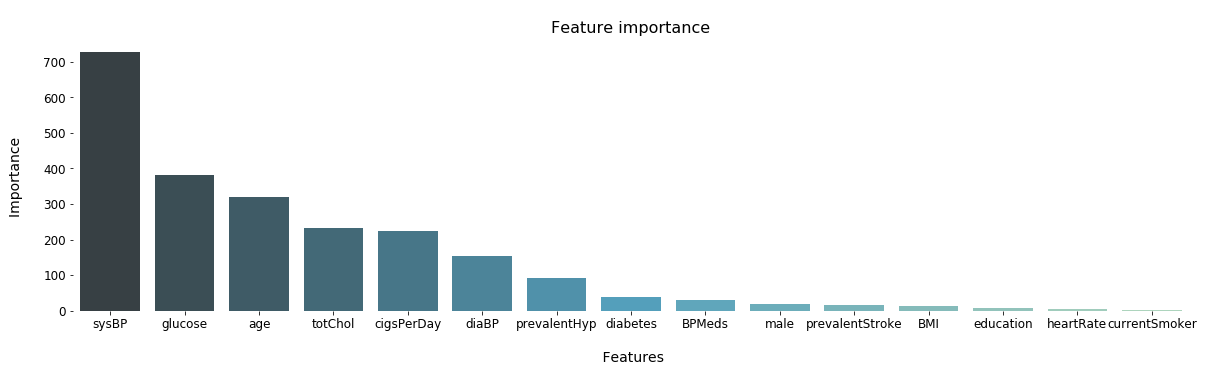
A screenshot of a cell phone

Description automatically generatedAlmost all the categories of education level have same level of both the genders except level 3.

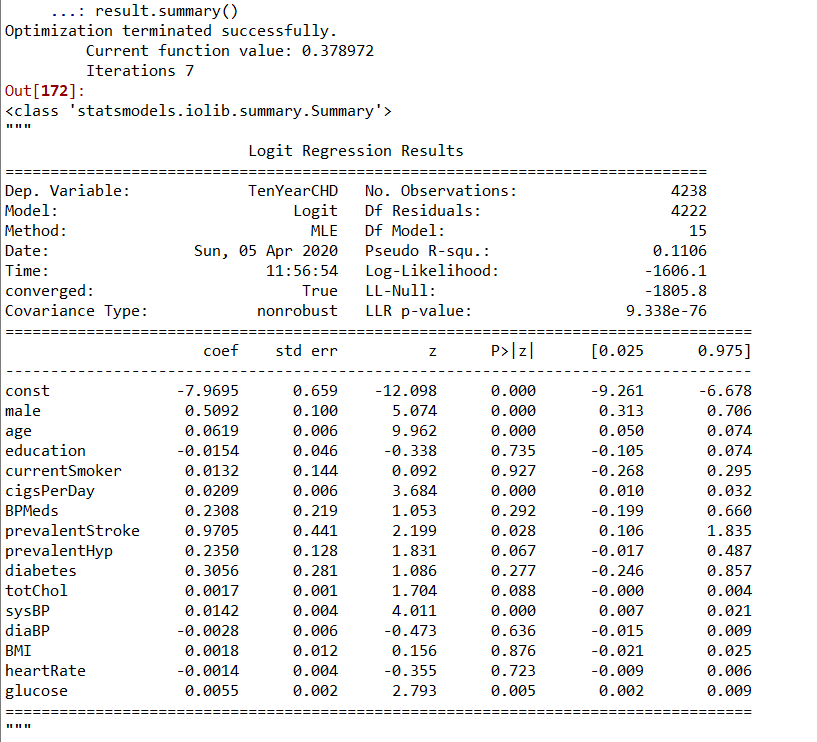
5. **Feature Selection:** Feature Selection is the process where we automatically or manually select those features which contribute most to our prediction variable. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features

Feature scores with sort in descending order:

 Plot to show the features that are important:

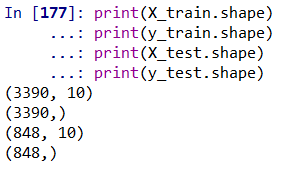


We are dropping the education, currentSmoker, heartRate and BMI, prevalentStroke column The above features are not as important features compared to others while determining TenYearCHD. Current Smoker may not be Smoker in the past years. Education is a very subjective and it is not handy to put in practice.

6. **Results of the Statistical Logit Model on all the features**:

## 7. **Splitting into training and testing data set:**

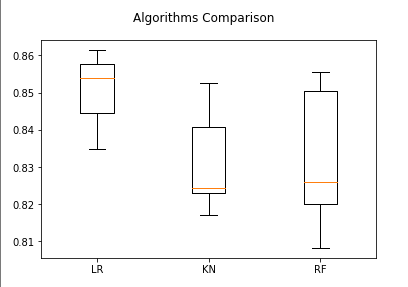
We have divided the data into testing data and training data in the ratio of 20:80. Our training dataset consists of 3390 rows and 10 most important and relevant features while our testing dataset consists of 848 rows with the most important features.



# 8. **Modeling and Evaluation:**

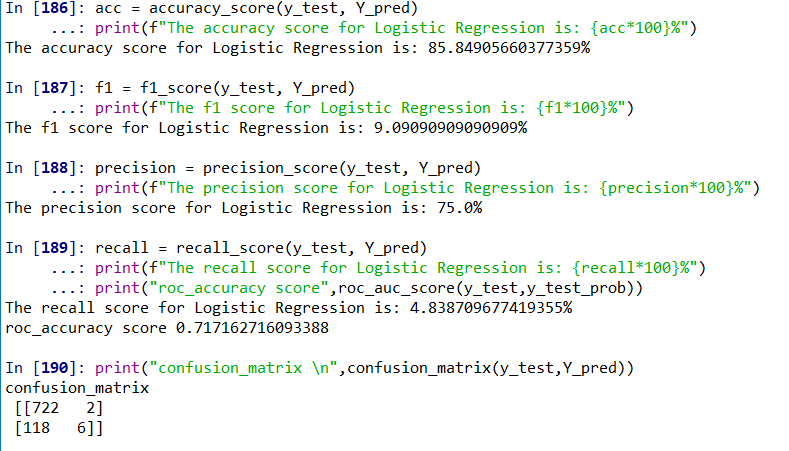
Before making prediction on the testing data set and to decide which model to use it is better to do cross validation on training dataset. Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement.

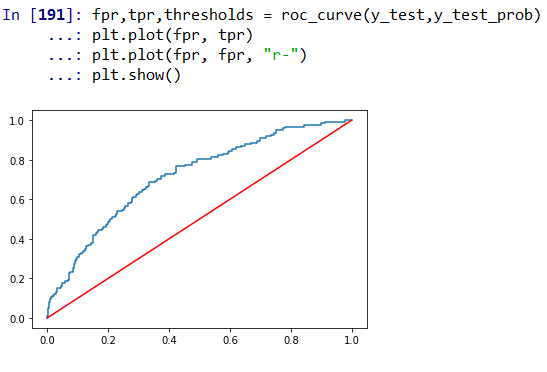
The performance of our model on the training dataset with KFolds (K=10) before evaluating our data to testing dataset (new data set). The Logistic Regression performs better than other models.



**Comparing different classifiers on testing dataset:**

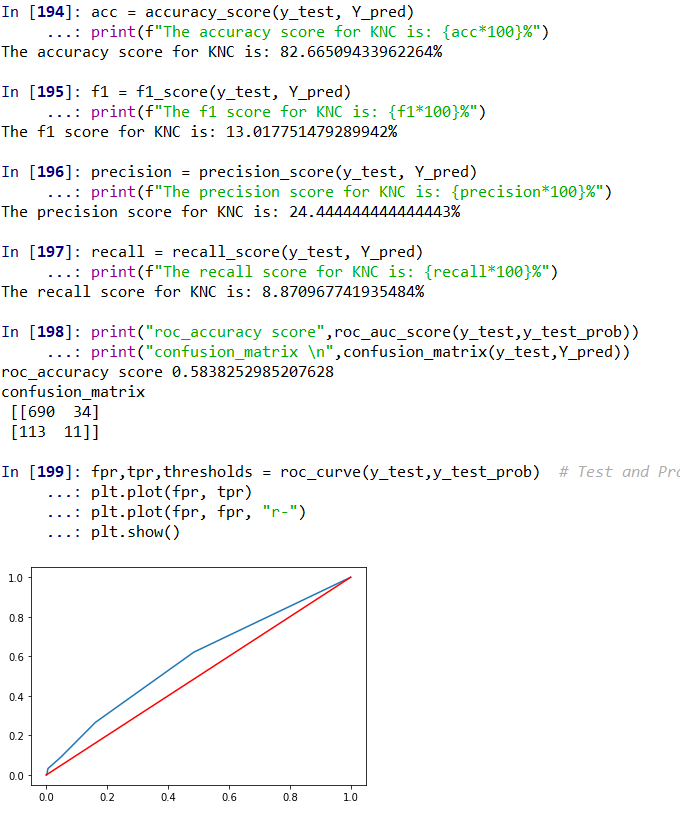
**Logistic Regression:**





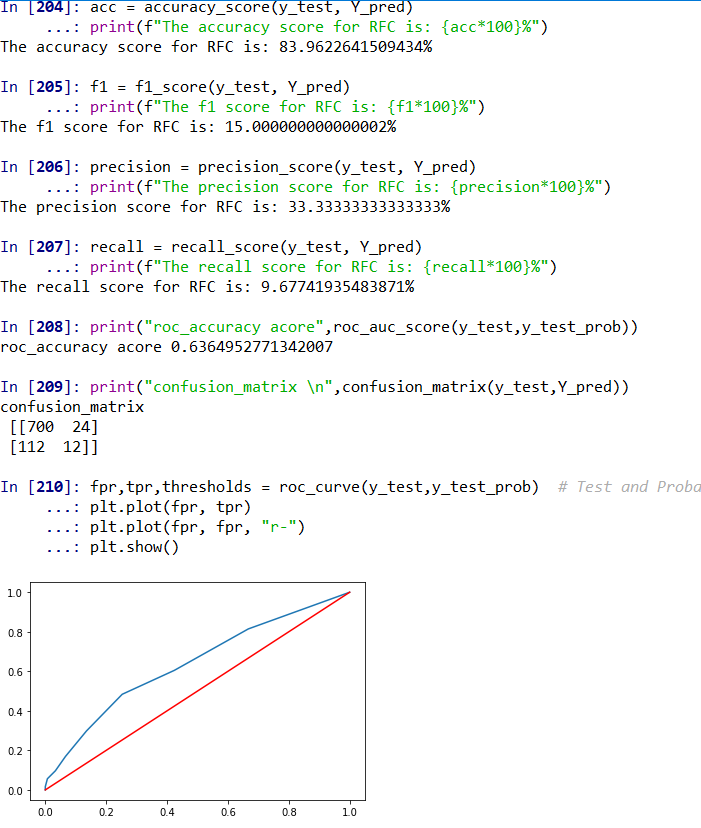
**The accuracy of Logistic Regression is 85.84% on the testing data set.**

**K Neighbors Model:**



**The accuracy of KNN classifier is 82.66% on the testing data set.**

**Random Forest Model:**

 **The accuracy of Random Forest Classifier is 83.96% on the testing data set.**

1. For the given dataset, Logistic Regression performs better on testing dataset than KNN classifier and Random Forest Classifier.
2. The Logistic Regression model accuracy is 85.84 %. So, the model does a good job in predicting whether the patient has 10-year risk of future coronary heart disease (CHD).
3. The accuracy of Logistic Regression is 85.84% and the model is not overfitting or underfitting.