UNIVERSITY OF TEXAS ARLINGTON

DATA MINING (IE 6318)

Heart Disease and Risk prediction using Machine learning Algorithms

Submitted by

ANUJ ANANDAKUMAR

1001746608

(12/15/2020)

Project Advisor: Shyoui wang
Submitted on: 12/15/2020

List of contents:

List of Figures:

SECTION I

Abstract:

Over the last few decades, heart attacks or cardiovascular disease is the primary basis of death worldwide. An estimate made by the WHO, that over 17.9 million deaths occur every year all over the world because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral strokes. Various habitual risk factors are in process 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, optimum and accurate and early medical diagnosis of heart health plays a crucial role in taking preventive measures to prevent death.

Heart disease, also known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the past few decades. Data analyzing/ Data mining is a frequently used technique for processing enormous data in the healthcare industry.

This research presents various factors related to cardiovascular disease, and the model on basis of supervised learning algorithms as Logistic regression, decision tree, K-nearest neighbor, and random forest algorithm. The accuracy proposed by different techniques varies with list of attributes. This research provides diagnostic accuracy score for improvement of better health results. We are using Python to carry all these activities.

The dataset is publicly found on the Kaggle website, and it is from an current cardiovascular disease research on residents of the city of Framingham, Massachusetts. The classification objective is to predict whether the patient has risk of future modifications of cardiovascular disease. The dataset contains the all the information. It contains almost 4,000 records and 15 attributes.

Keywords: Heart disease prediction, Data mining, Decision tree, Naïve Bayes, K-NN, Random forest, Machine learning, WHO-World Health Organization

Project Plan:

The dataset is publicly found on the Kaggle website, and it is from an current cardiovascular disease research on residents of the city of Framingham, Massachusetts. The classification objective is to predict whether the patient has risk of future modifications of cardiovascular disease. The dataset contains the all the information. It contains almost 4,000 records and 15 attributes.

The real-life information will have large numbers with missing and noisy data. These data will be pre-processed to overcome such issues and make predictions vigorously.

We are going to carry this research with Python as a programming language with various machine learning libraries. For implementation of Python programming Anaconda jupyter notebook is best tool, which have many libraries, header file, that make the work more accurate and precise.

As already stated in proposal, this research contains various factors related to cardiovascular disease, and the model on basis of supervised learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm etc. In this research, we calculate accuracy of machine learning algorithms for predicting heart disease.

We can say this project as an application-based project where the research is carried with real time datasets from online platform.

Attributes

- **sex**: male or female (Nominal)
- **age**: age of the patient (Continuous)
- **currentSmoker**: patient is a current smoker in that zone (Nominal)
- **cigsPerDay**: the number of cigarettes that the person smokes on average in one day.(continious)
- **BPMeds**: whether patient was on blood pressure medication (Nominal)
- **prevalentStroke**: whether patient had previously had a stroke (Nominal)
- **prevalentHyp**: whether patient was hypertensive (Nominal)
- **diabetes**: whether patient had diabetes (Nominal)
- **totChol**: total cholesterol level (Continuous)
- **sysBP**: systolic blood pressure (Continuous)
- **diaBP**: diastolic blood pressure (Continuous)
- **BMI**: Body Mass Index (Continuous)
- **heartRate**: heart rate (Continuous)
- **glucose**: glucose level (Continuous)
- **10-year risk of coronary heart diseases CHD** (binary: “1” means “Yes”, “0” means “No”) - Target Variable

Objective:

For The dataset provides the risk factors associated with heart disease for the patients and whether they have a risk of coronary heart disease in the next coming years.

Based on the dataset provided:

1. Predict the probability of a patient suffering a coronary heart disease in the next 10 years
2. Identify the most important factors that influence for this cardiovascular disease
3. Come up with recommendations for

   a. Preventing / reducing chances of getting the disease

   b. Extrapolated applications of the model we build and its results.

SECTION II

Cleaning of Data:

Data cleaning is the task that identifies the one which is not correct, not complete, inaccurate, or irrelevant data, fixes the errors, and makes sure that all such issues will be fixed automatically in the future aspects. We find the data, Missing values, we replace missing values with zeros and mean of that column wherever necessary.

Data Visualization

Box Plot:

We create box plot to find the percentage of outliers

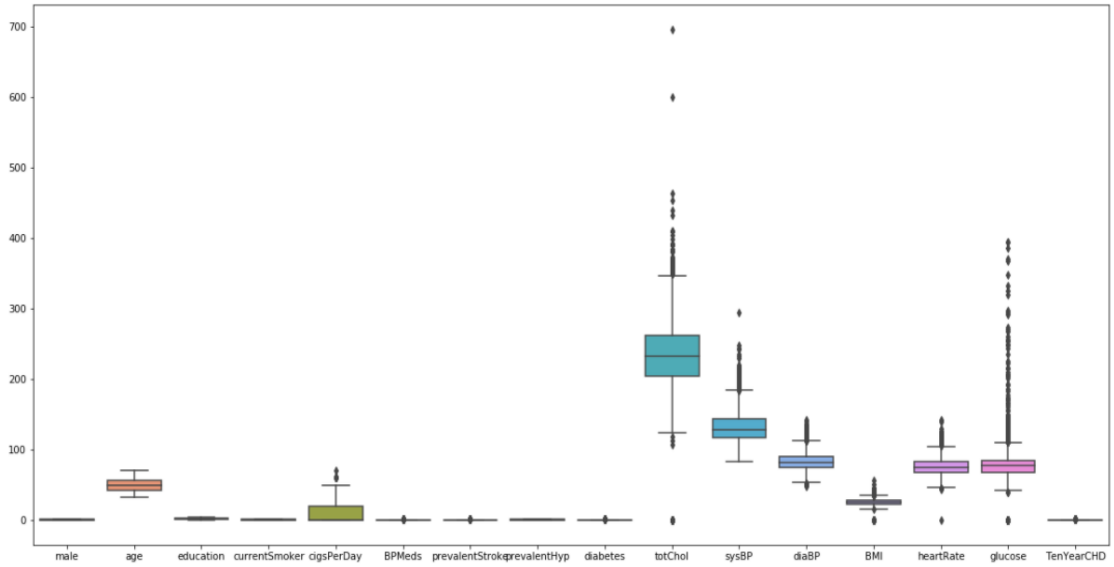


Fig1: Boxplot

There are outliers in the columns of totChol, sysBP, diaBP, BMI, heartRate and Glucose.We will remove the outliers as anything over 75% of the maximum value.

Histogram:

We here understand the frequency distribution for various attributes.

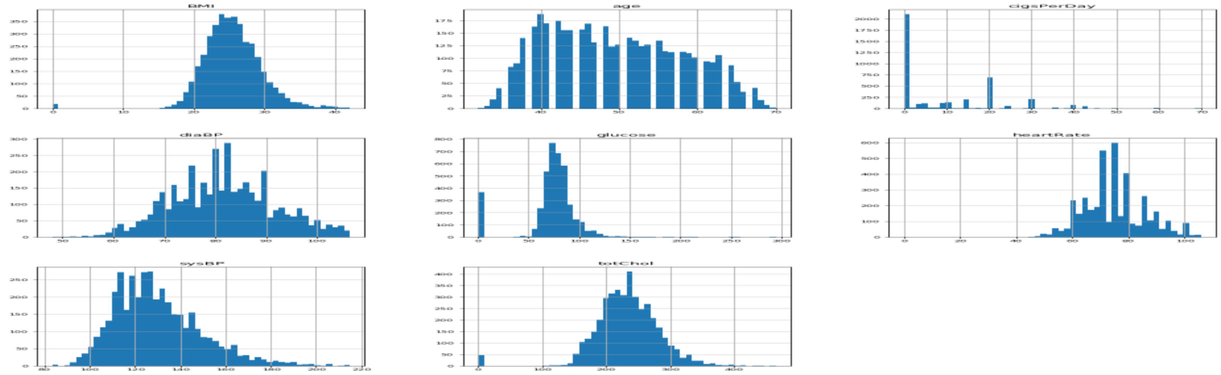


Fig 2: Histogram

The average BMI of the category is around 25 and the most number of people have a BMI between 23 to 28. This means that most of the people in the group are bordering near overweight.

Age of people in the study varies between 32 to 70 and the average age is around 50. So the people in this research are middle aged to old.

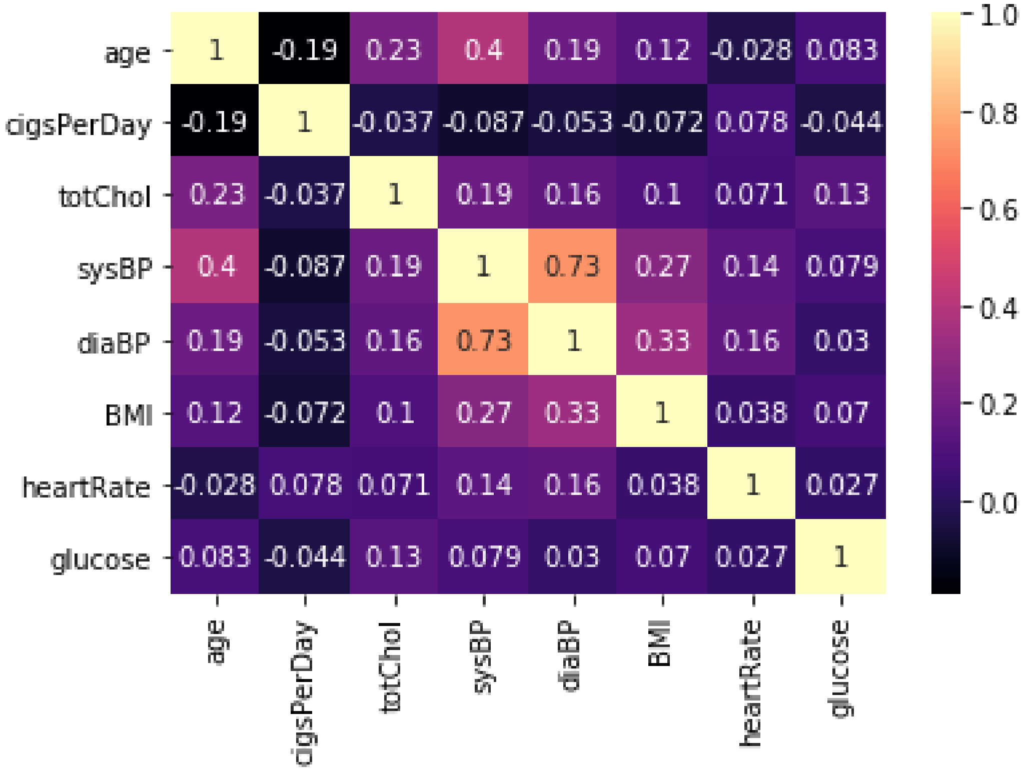
Even though the number of cigarettes smoked is 9 on aveage, Half of the people did not smoke at all. One outlier has smoked 70 cigarettes a day.

The Glucose level on average is 82 and it varies between 70 to 88 which means none of them are diabetic presently.

Heart rate on average is 74 with the maximum number of people between 65 to 89.

Many people have a cholesterol levels of 206 to 263, with average being 237. This indicates Borderline of high cholesterol level.

Corelation Plot:



	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
age	1	-0.19	0.23	0.4	0.19	0.12	-0.028	0.083
cigsPerDay	-0.19	1	-0.037	-0.087	-0.053	-0.072	0.078	-0.044
totChol	0.23	-0.037	1	0.19	0.16	0.1	0.071	0.13
sysBP	0.4	-0.087	0.19	1	0.73	0.27	0.14	0.079
diaBP	0.19	-0.053	0.16	0.73	1	0.33	0.16	0.03
BMI	0.12	-0.072	0.1	0.27	0.33	1	0.038	0.07
heartRate	-0.028	0.078	0.071	0.14	0.16	0.038	1	0.027
glucose	0.083	-0.044	0.13	0.079	0.03	0.07	0.027	1

Fig 3 Correlation Plot

Systolic BP & Diastolic BP are highly correlated to each other.

Systolic BP & Diastolic BP each are partly correlated to BMI (Body Mass Index).

Systolic BP & Age are partly correlated to each other.

Exploratory Data Analysis Findings

The research consists of 4,240 borderline of overweight Middle-Aged to Older patients, with the sex ratio slightly skewed towards Women, consisting of an almost equal split between smokers and non-smokers. There are only 25 people or 0.6% of the study group have had a history of strokes in past. Most of them in the study has high cholesterol levels.

Data Insights:

1. Data is unbalanced overall.
2. Education level does not seem to be related to the heart disease risk.
3. Systolic BP & Diastolic BP are highly correlated to each other.

Feature Selection

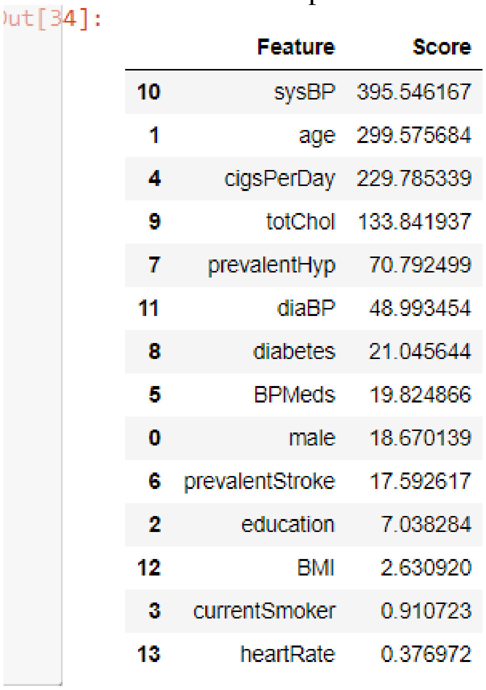
Feature selection is a process where you automatically select the essential features in our data that contribute most to the prediction variable or output in which you are interested.

Having irrelevant features in our data can decrease the accuracy of many models we predict, especially linear algorithms like linear and logistic regression.

There are three benefits of performing feature selection before modeling your data are:

1. It reduces Overfitting.
2. It improves the Accuracy.
3. It reduces the training time.

There are lots of ways to do feature selection - Univariate Feature Selection, Recursive Feature Selection, Model Feature Selection. We are using Univariate Feature Selection as it checks how each feature affects the predicted variable.



)ut[34]:

	Feature	Score
10	sysBP	395.546167
1	age	299.575684
4	cigsPerDay	229.785339
9	totChol	133.841937
7	prevalentHyp	70.792499
11	diaBP	48.993454
8	diabetes	21.045644
5	BPMeds	19.824866
0	male	18.670139
6	prevalentStroke	17.592617
2	education	7.038284
12	BMI	2.630920
3	currentSmoker	0.910723
13	heartRate	0.376972

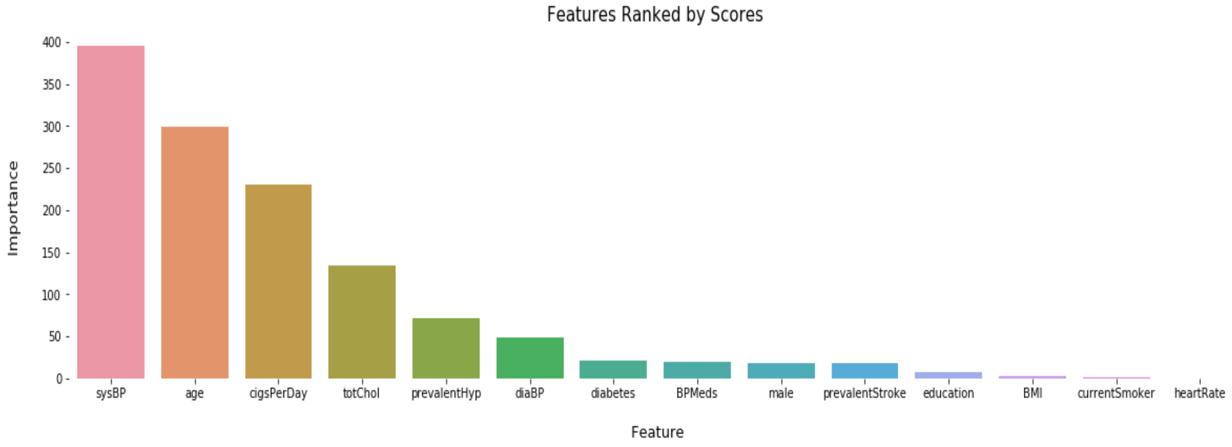


Fig 4: graph and value based on feature score

The above images and outputs shows the top 10 features based on scores.

The most important features that the datasets are

- Systolic BP
- Age
- Cigerattes Per Day
- Total Cholestoral
- Prevalent Hypertension
- Diastolic BP
- Diabetics
- BP Medicines
- Sex
- Prevalent Stroke

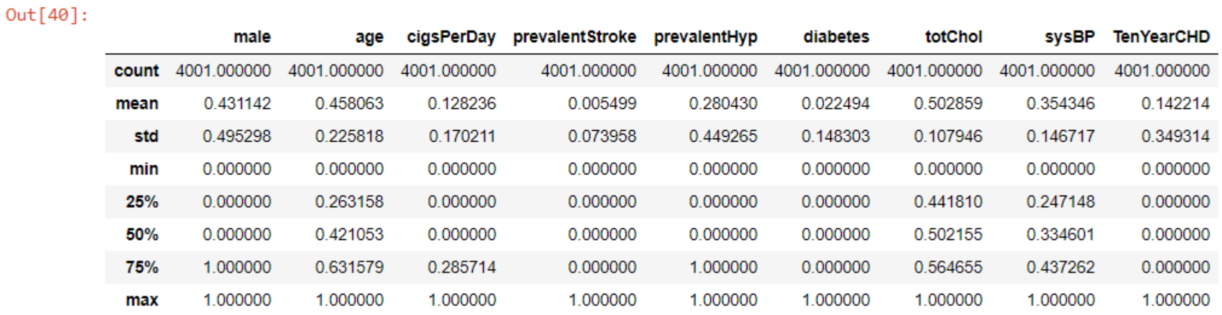
The features that can be ignored are education. We will only be using these factors for all further analysis.

Scaling Data

Since each of the features can have impact on different scales, which leads to data models giving certain features without any value. Scaling the data brings all the columns onto the same levels of process.

There are many ways to scale the data, like Standard Scaler, MaxAbs Scaler, MinMax Scaler and Robust Scaler.

We will be using MinMax Scaler as the datasets is too sensitive of outliers.



Out[40]:

	male	age	cigsPerDay	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	TenYearCHD
count	4001.000000	4001.000000	4001.000000	4001.000000	4001.000000	4001.000000	4001.000000	4001.000000	4001.000000
mean	0.431142	0.458063	0.128236	0.005499	0.280430	0.022494	0.502859	0.354346	0.142214
std	0.495298	0.225818	0.170211	0.073958	0.449265	0.148303	0.107946	0.146717	0.349314
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.263158	0.000000	0.000000	0.000000	0.000000	0.441810	0.247148	0.000000
50%	0.000000	0.421053	0.000000	0.000000	0.000000	0.000000	0.502155	0.334601	0.000000
75%	1.000000	0.631579	0.285714	0.000000	1.000000	0.000000	0.564655	0.437262	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig 5: Datahead after scaling

The images shows the dataset after scaling.

SECTION III

Model Building

Divide Train and Test Datasets

We divide data into train and test data with a 70-30 ratio.

Modeling and Balencing the Datasets:

The dataset is very unbalanced as only 15% of the sampled population are at risk of developing heart risk at the end of 10 years. The problem with unbalanced dataset The most common methods are - oversampling and under sampling

Oversampling — Duplicating samples from the minority class.

Undersampling — Deleting samples from the majority class.

These change to the class distribution is only applied to the training dataset. This is done to influence the fit of the models.

Smote

Smote is called Synthetic Minority Oversampling Technique. Oversampling duplicates the existing minority class value unlike smote draws lines between existing minority class members and creates brand new minority class which levels between the existing ones. Generally, smote is used to prepare the Train dataset

Modelling and Evaluation

These metrics can be used to determine the best models:

Confusion Matrix - it gives the information between actual and predicted value.

a. True Positive = TP i.e. actual value (y_actual) is 1 and the predicted value (y_pred) is 1.

b. True Negative = TN i.e., actual value (y_actual) is 0 and the predicted value (y_pred) is 0.

c. False Positive = FP i.e. the actual value (y_actual) is 0 but the predicted value (y_pred) is 1.

d. False Negative = FN i.e. the actual value (y_actual) is 1 but the predicted value (y_pred) is 0.

Accuracy - Accuracy predicts how accurately the y predict matches the y actual value in the system. The closer the accuracy of the model is to 100%, the more better the model.

Accuracy = (True Positive + True Negative)/ Total

Precision Score - Precision is the ratio of True positives and the predicted positives. The lesser the false positive, better the value it is.

Precision = True Positive / (True Positive + False Positive)

Recall Score / Sensitivity – Recall/Sensitivity is the ratio of true positives vs actual positives. The lesser the number of False Negatives, the better the recall/sensitivity of score.

Recall/ Sensitivity = True Positive / (True Positive + False Negative)

F1 Score - It combines Recall/sensitivity and Precision scores. The closer the F1 score is to 100%, better the model.

F1 Score = 2 / ((1/Recall) + (1/Precision))

ROC Curve - ROC (Receiver Operating Characteristics) curve is a visualization of the mod el output in the terms of False Positive Rate (FPR = FP / (FP + TN)) on the x-axis and True Positive Rate (TPR = TP / (TP + FN)) on the y-axis.

AUC (Area Under Curve)- Area under the curve represents degrees or measurements of se parability of the model. The more higher the AUC, the better the model is at predicting 0s co rrectly as 0s and 1s as 1s.

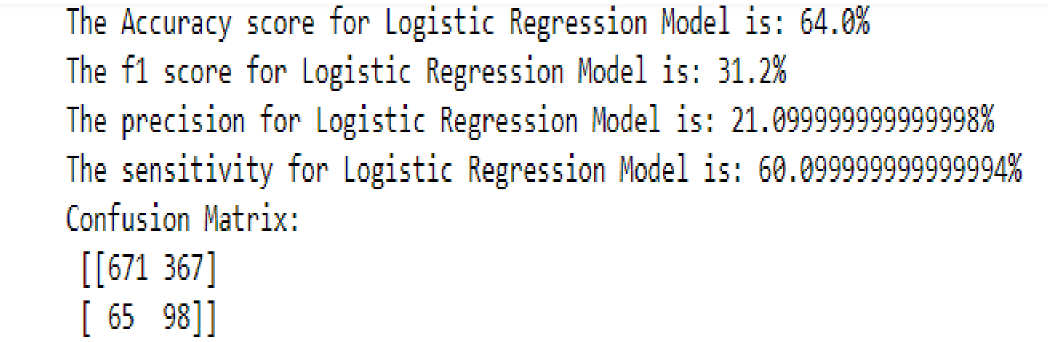
AUC thumb rule:

0.9 – 1→excellent, 0.8 - 0.9→ good, 0.7 - 0.8→ fair, 0.6 - 0.7→ poor, 0.5 - 0.6→ fail

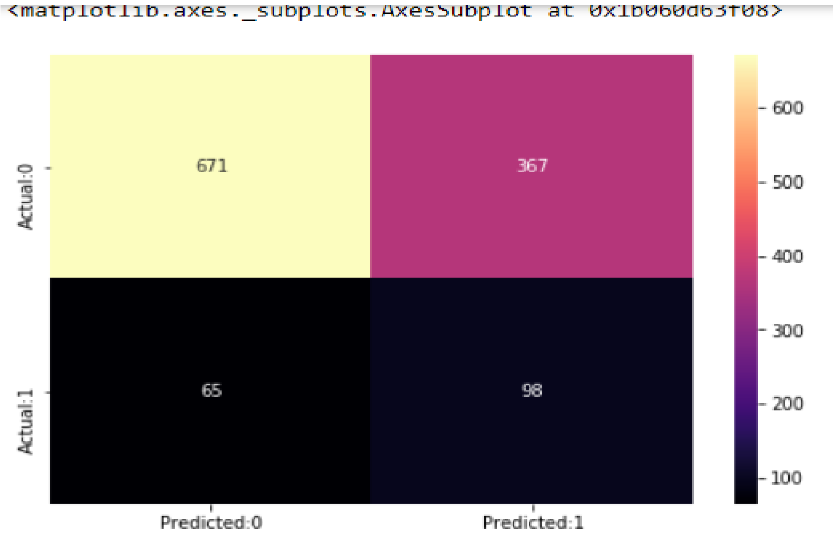
SECTION IV

Logistic Regression:

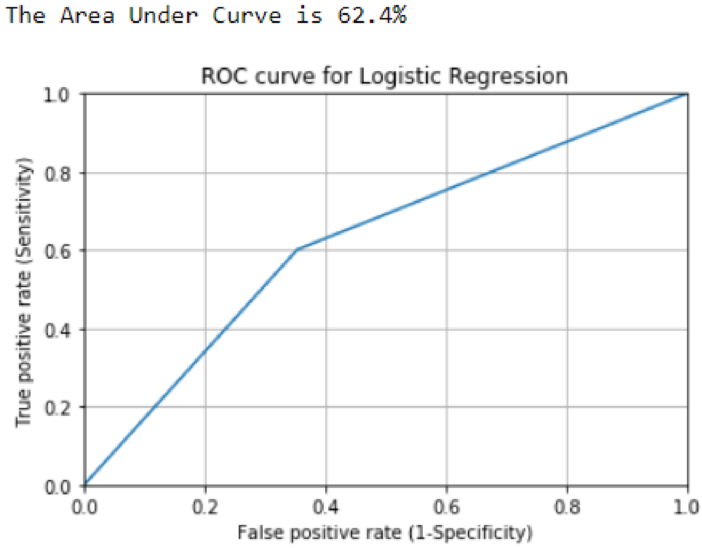
A logistic regression model predicts the dependent data variable by analyzing the relationship between one or more existing independent variables in the datasets. It is a very common statistical model used in the machine learning process to determine if an independent variable has an effect on a binary dependent variable. The sigmoid fuction/logistic function consists the values to lie between 0 to 1.



```
The Accuracy score for Logistic Regression Model is: 64.0%
The f1 score for Logistic Regression Model is: 31.2%
The precision for Logistic Regression Model is: 21.099999999999998%
The sensitivity for Logistic Regression Model is: 60.099999999999994%
Confusion Matrix:
 [[671 367]
 [ 65  98]]
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x1b060d63f08>
```



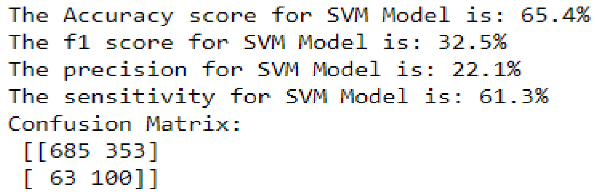
```
The Area Under Curve is 62.4%
```

Fig 6:Plots for logistic regression

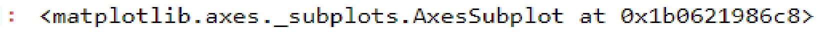
A 64% accuracy, 31% F1 Score and AUC of 62% are observed which are not up to the mark

Support Vector Machines:

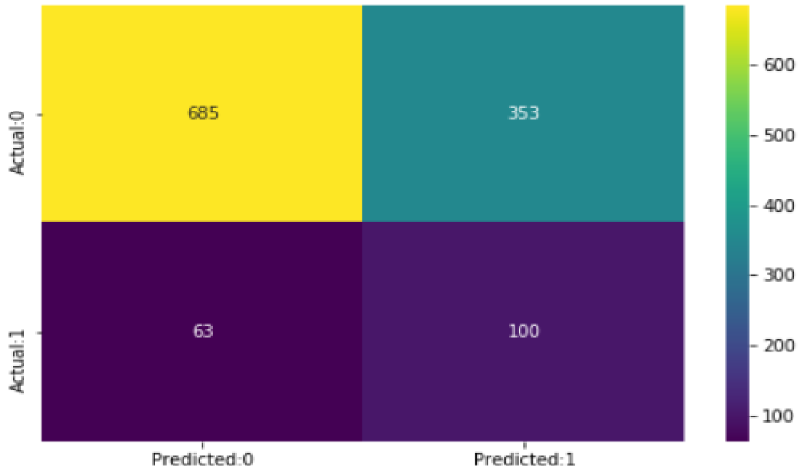
The important objective of the support vector machine(svm) algorithm is to find a hyperplane in an N-dimensional space(N→the number of features) that distinctly classifies the data points in the datasets. Support vectors(sv)are data points in the sets that are very closer to the hyperplane and influence the position and orientation of the hyperplane. Using the help of support vectors Machines algorithm, we maximize the margin of the classifiers where the values generally lie between -1 to 1.

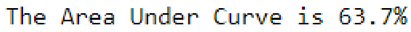


```
The Accuracy score for SVM Model is: 65.4%
The f1 score for SVM Model is: 32.5%
The precision for SVM Model is: 22.1%
The sensitivity for SVM Model is: 61.3%
Confusion Matrix:
 [[685 353]
 [ 63 100]]
```



```
: <matplotlib.axes._subplots.AxesSubplot at 0x1b0621986c8>
```





```
The Area Under Curve is 63.7%
```

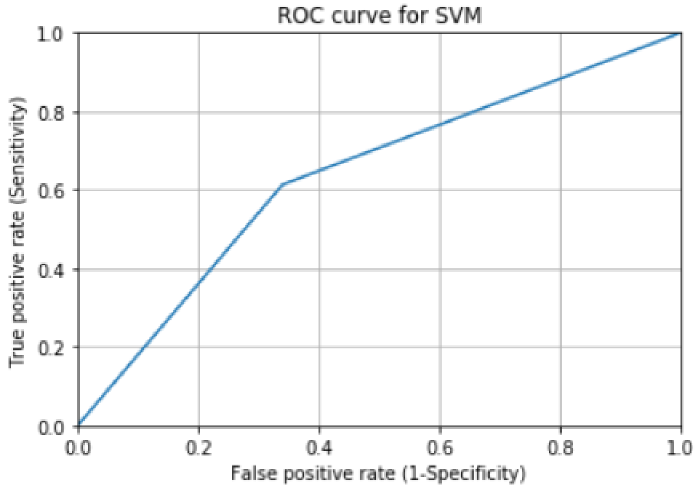
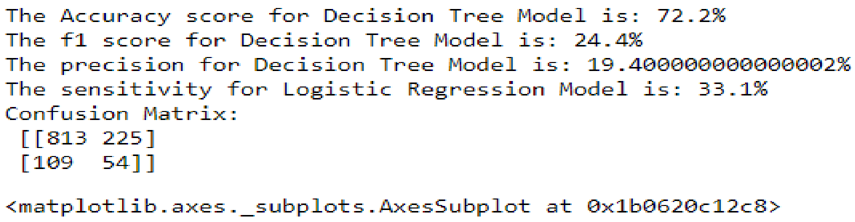


Fig 7: Plots for SVM

From the values we can consider that SVM also does not showing any positive result.

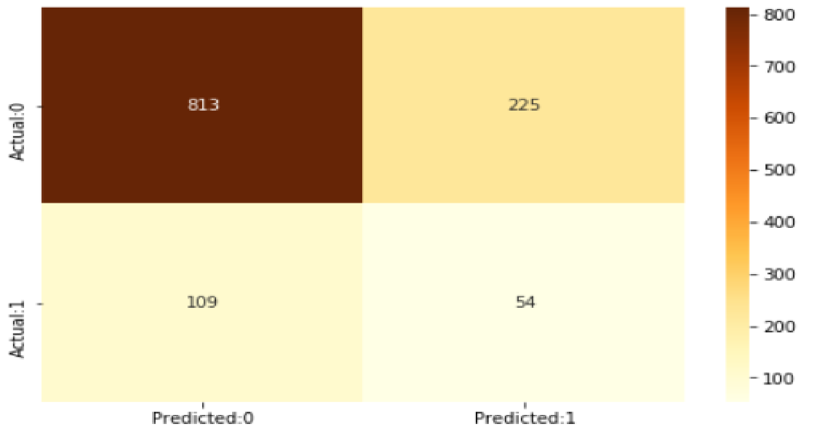
Decision tree:

A Decision tree is a form of flowchart which has many branches ie tree like structure, where each internal node denotes a test on the attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label in its position. A decision tree is always drawn upside down with the root pointing on the top. One of the main use of decision trees is that they provide a clear indication of which fields i.e. variables that are most important to prediction or classification.



```
The Accuracy score for Decision Tree Model is: 72.2%
The f1 score for Decision Tree Model is: 24.4%
The precision for Decision Tree Model is: 19.400000000000002%
The sensitivity for Logistic Regression Model is: 33.1%
Confusion Matrix:
 [[813 225]
 [109  54]]

<matplotlib.axes._subplots.AxesSubplot at 0x1b0620c12c8>
```



```
The Area Under Curve is 55.7%
```

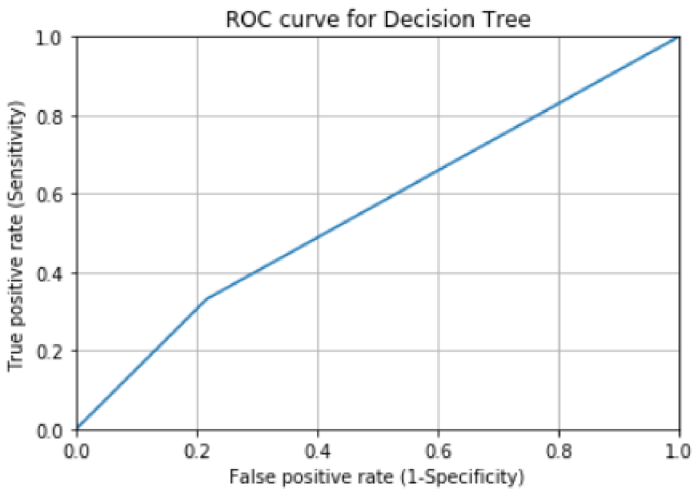
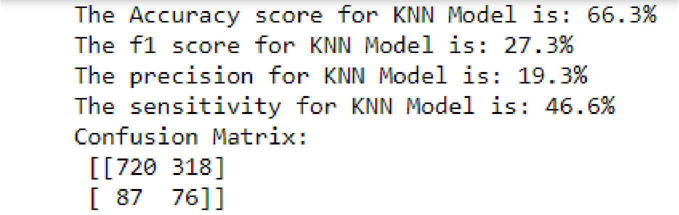


Fig 8: Plots for Decision Tree

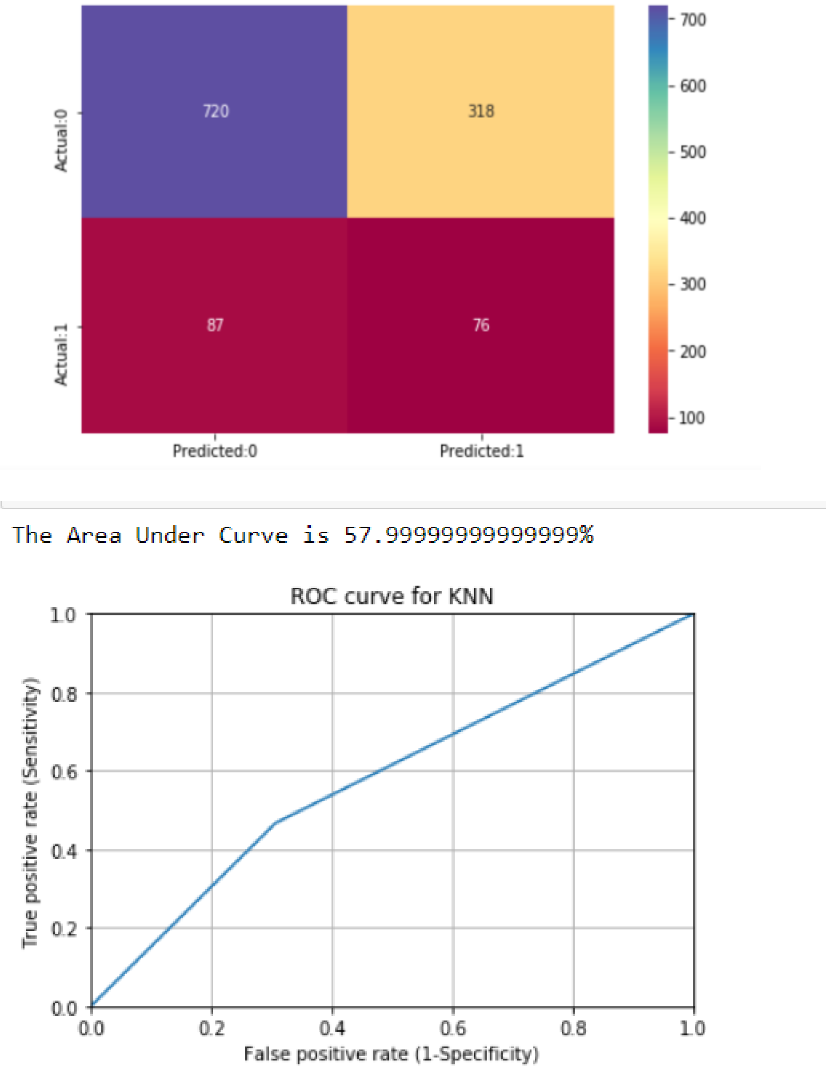
The values are much better than SVM and Logistic Regression but still not upto the mark.

Knn -K-nearest Neighbors:

KNN - K Nearest Neighbor assumes that the similar points are its neighbors- that is points with similar characteristics/structures which lie close to each other. One of the main use of KNN is that it is a non-parametric learning algorithm I.e. the algorithm starts out with no assumptions.



```
The Accuracy score for KNN Model is: 66.3%
The f1 score for KNN Model is: 27.3%
The precision for KNN Model is: 19.3%
The sensitivity for KNN Model is: 46.6%
Confusion Matrix:
 [[720 318]
 [ 87  76]]
```



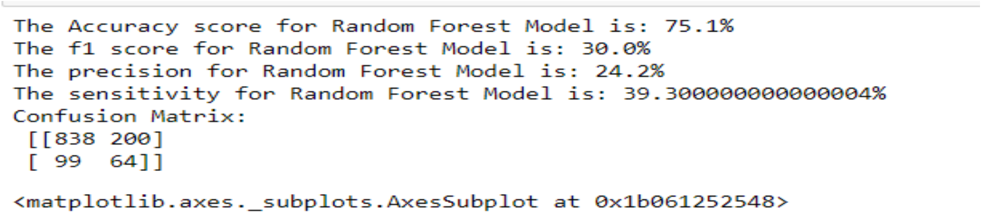
```
The Area Under Curve is 57.99999999999999%
```

Fig 9:Plots for Knn

The results of Knn is represented above.

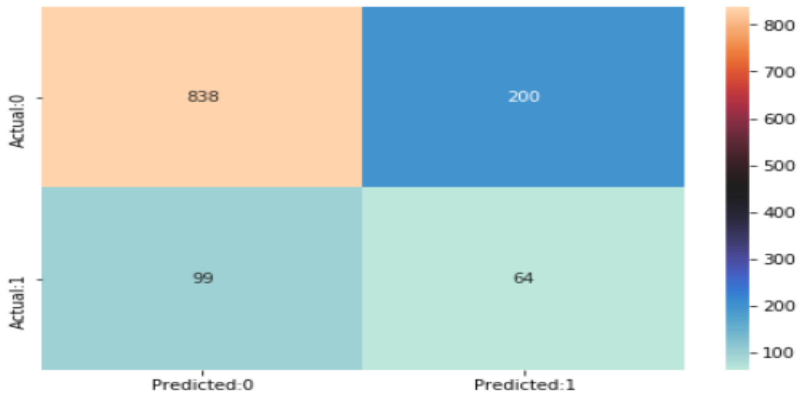
Random Forests:

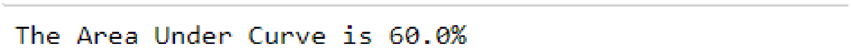
Random forest is an assemble of many relatively uncorrelated trees, each of which make their own decisions. In other words, it is called collections of decision trees. The main use of this is, each tree is unbiased by the output or workings of the other, moving the decision in an overall right way of path.



```
The Accuracy score for Random Forest Model is: 75.1%
The f1 score for Random Forest Model is: 30.0%
The precision for Random Forest Model is: 24.2%
The sensitivity for Random Forest Model is: 39.300000000000004%
Confusion Matrix:
 [[838 200]
 [ 99  64]]
```

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x1b061252548>
```





```
The Area Under Curve is 60.0%
```

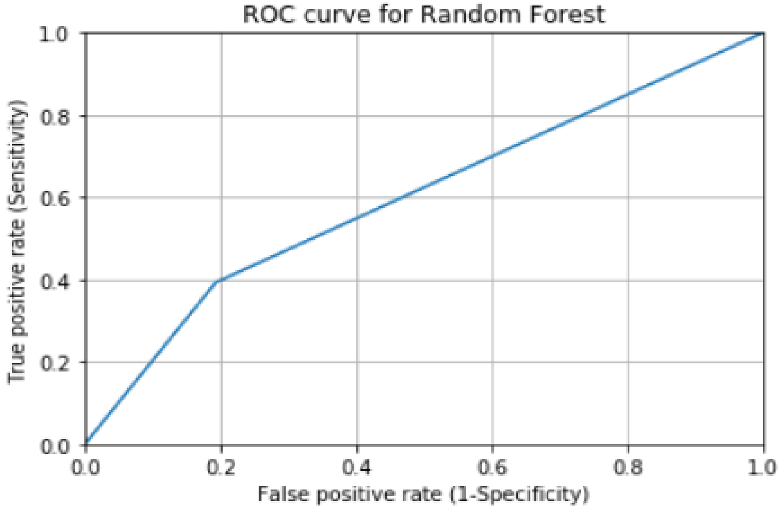
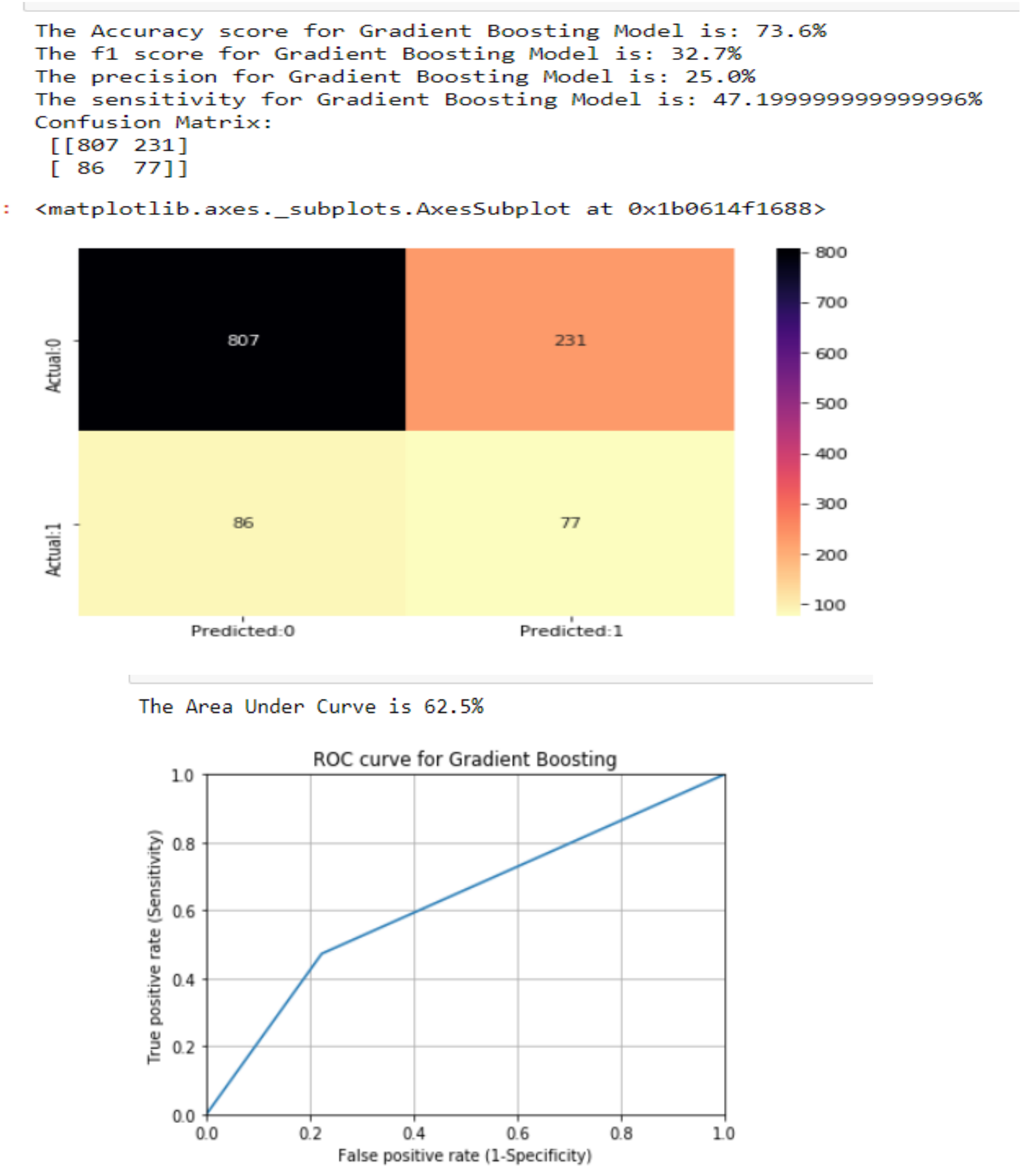


Fig 10: Plots for Random forests

The results of Random forests are stated above.

Gradient Boosting:

Boosting is a method which converts weak learners into strong learners ie Weak learners here means decision trees, Gradient Boosting trains many models in a gradual, additive and sequential manner in this process. XGB involves 3 elements - a loss function to be optimized, a weak learner to make predictions and an additive model to add the weak learners to minimize the loss function which is stated.



```
The Accuracy score for Gradient Boosting Model is: 73.6%
The f1 score for Gradient Boosting Model is: 32.7%
The precision for Gradient Boosting Model is: 25.0%
The sensitivity for Gradient Boosting Model is: 47.199999999999996%
Confusion Matrix:
 [[807 231]
 [ 86  77]]

: <matplotlib.axes._subplots.AxesSubplot at 0x1b0614f1688>
```

Actual:0
Actual:1
807
231
86
77
Predicted:0
Predicted:1
800
700
600
500
400
300
200
100

```
The Area Under Curve is 62.5%
```

ROC curve for Gradient Boosting
True positive rate (Sensitivity)
False positive rate (1-Specificity)
0.0
0.2
0.4
0.6
0.8
1.0

Fig 11: Plots for Gradient Boosting

The above output represents the result of gradient boosting.

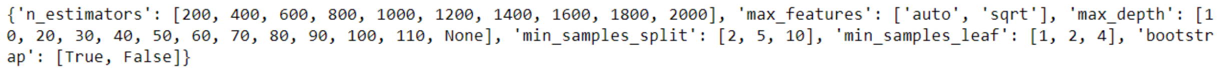
Hyperparameter tuning for best classifier:

A Machine Learning model is also defined as a mathematical model with a n number of parameters that need to be learned from the datasets. By training a model with existing data, we are able to fit the model parameters for the given datasets. However, there is another kind of parameters which is known as Hyperparameters, that cannot be directly learned from the regular training methods. They are usually fixed before the actual training process take place. These parameters express important properties of the model such as its complexity and how fast it should learn the model.

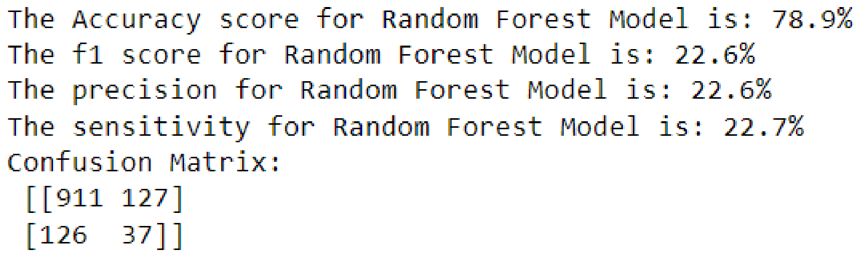
The 2 methods used are:

1. GridSearchCV - This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values present in the datasets. These are very expensive since it goes to all the grids.
2. RandomizedSearchCV - RandomizedSearchCV solves the negatives of GridSearchCV, as it goes through only a limited/fixed number of hyperparameter settings.

Random Forests:



```
{'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000], 'max_features': ['auto', 'sqrt'], 'max_depth': [1
0, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'bootstr
ap': [True, False]}
```



```
The Accuracy score for Random Forest Model is: 78.9%
The f1 score for Random Forest Model is: 22.6%
The precision for Random Forest Model is: 22.6%
The sensitivity for Random Forest Model is: 22.7%
Confusion Matrix:
 [[911 127]
 [126  37]]
```

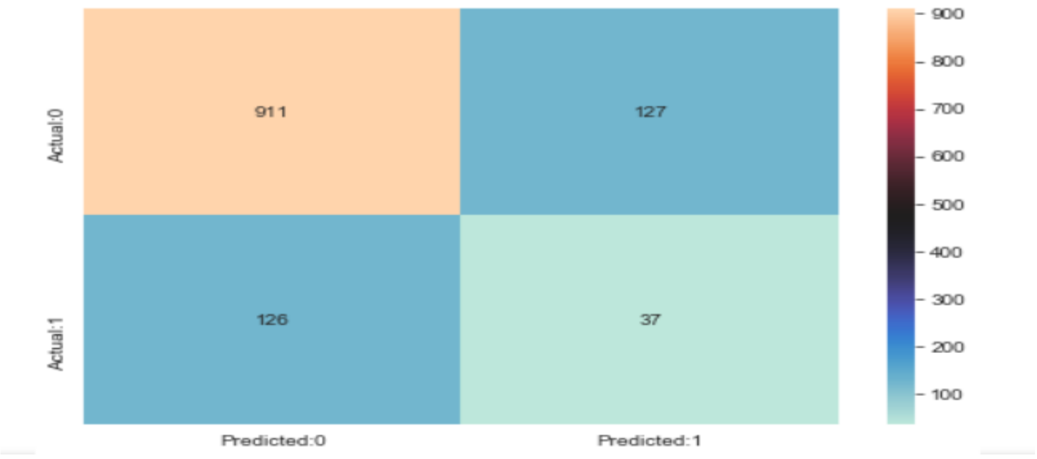
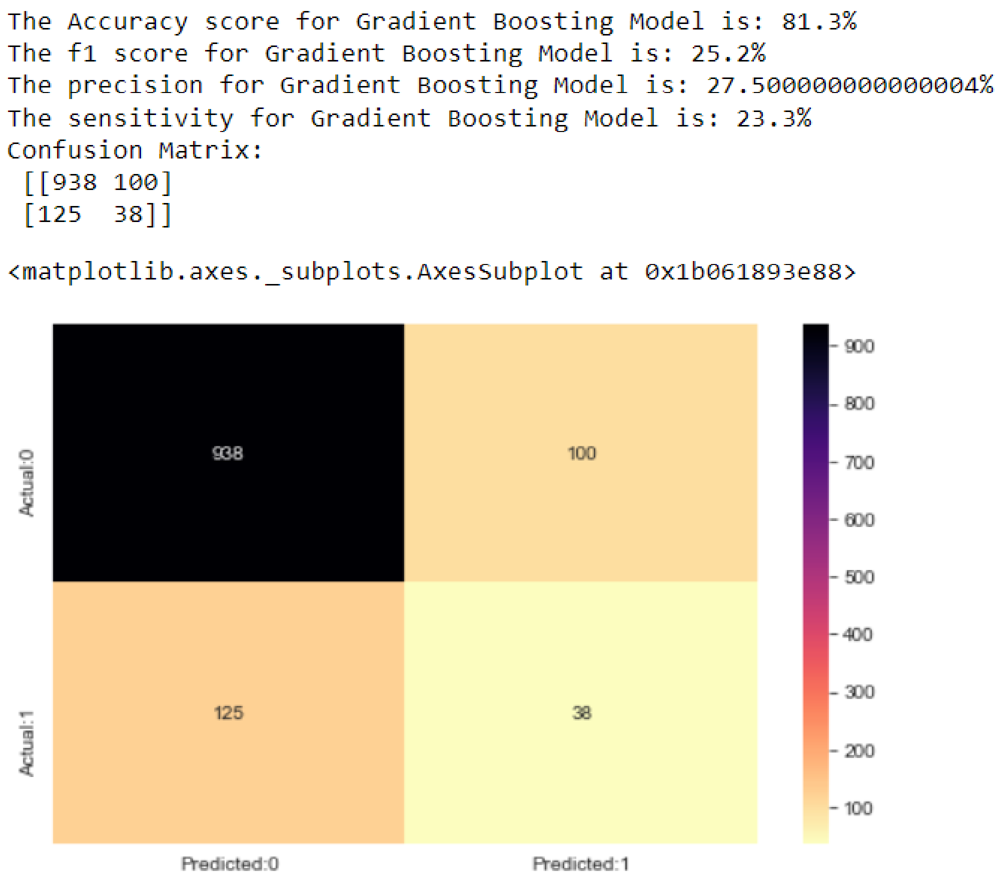


Fig 12: Plots for Random forests after HTC

Gradient Boosting:



```
The Accuracy score for Gradient Boosting Model is: 81.3%
The f1 score for Gradient Boosting Model is: 25.2%
The precision for Gradient Boosting Model is: 27.500000000000004%
The sensitivity for Gradient Boosting Model is: 23.3%
Confusion Matrix:
 [[938 100]
 [125  38]]

<matplotlib.axes._subplots.AxesSubplot at 0x1b061893e88>
```

Fig 13: Plots for Gradient Boosting after HTC

ROC FOR EACH MODEL:

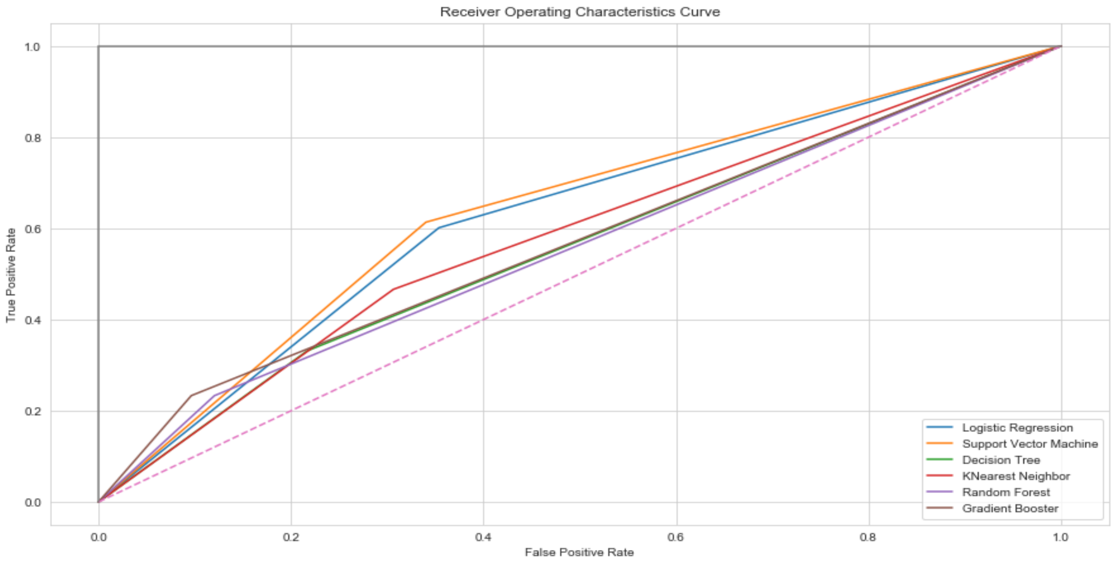


Fig 14: Plots for ROC for each model

SECTION V

Conclusion:

The recommendations from the models predicts or shows that there are 2 factors under our control to reduce chance of heart attack and 1 factor that is not.

The 2 factors that we can control in our lifestyle are:

- Systolic Blood Pressure is one of the most important factors to determine risk of cardiovascular disease. So keeping BP in control is essential.
- The number of cigarettes we smoke also have a massive Impact with the Heart disease which ultimately proves smoking kills.

The factor which is not in control with human hands is Age, so exercise regularly, do Yoga and Practice a healthy lifestyle.

Findings from the Data:

- The data is highly unbalanced
- The gradient boosting showed the highest accuracy of almost 82%. We can use this model many health organizations such as hospitals.
- The support vector machine had a high accuracy in terms of F1 score, precision and Recall scores. So, if any information needed to be predicted based on population, age, demographic SVM can be used.

SECTION VI

Acknowledgement:

I would like to express my deepest appreciation to all those who provided me the possibility to complete this project/research. A special gratitude I give to(IE 6318) Professor Dr Shoyui Wang, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in doing my analysis work. I would also like to thank my Teaching Assistant Bahareh Nasirian for wonderful guidance throughout my course work.

References:

- Alberto, Túlio C, Johannes V Lochter, and Tiago A Almeida. "Tubespam: comment spam filtering on YouTube." In Machine Learning and Applications (Icmla), Ieee 14th International Conference on, 138–43. IEEE. (2015).
- "Definition of Algorithm." https://www.merriam-webster.com/dictionary/algorithm. (2017).
- https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a
- https://www.geeksforgeeks.org/machine-learning/
- https://www.kaggle.com/datasets?utm_medium=paid&utm_source=google.com+search&utm_campaign=datasets&gclid=CjwKCAiAq8f-BRBtEiwAGr3DgQpfwSuALLK4C6bRPCtbzrdridTANu5AtgdmVqf4FBijHLSyj11NeRoC2_8QAvD_BwE