Capstone Project-3
Cardiovascular Risk Prediction

By

Neeraja C

Data Science Trainee|AlmaBetter

INTRODUCTION

- World Health Organization has estimated 12 million deaths occur worldwide,every year due to heart diseases.Half the deaths in the United States and other developed countries are due to cardiovascular diseases.

- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

PROBLEM STATEMENT

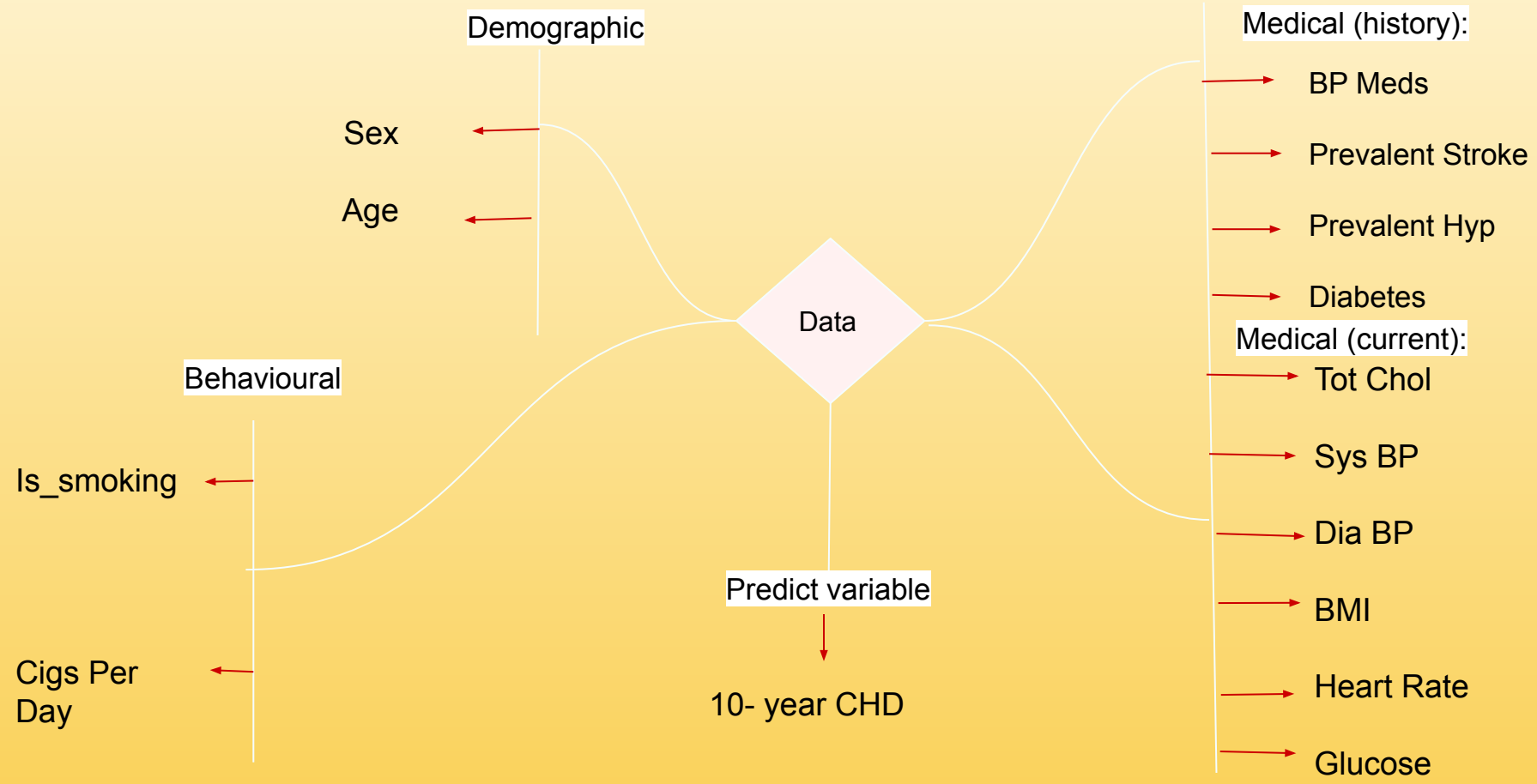
The dataset is from an ongoing cardiovascular on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

The dataset provides the patient's information. It includes over 4,000 records and 15 attributes.

Variables

Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors.

DATA SUMMARY



COLUMN DETAILS

Demographic

- Sex: Male or Female (“M” or ”F”)
- Age: Age of the patient (Continuous- Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioural

- is_smoking: Whether or not the patient is a current smoker (“YES” or “NO”)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day (can be considered continuous as one can have any number of cigarettes, even half a cigarette)

Medical (history):

- BP Meds: Whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: Whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: Whether or not the patient was hypertensive (Nominal)
- Diabetes: Whether or not the patient had diabetes (Nominal)

COLUMN DETAILS

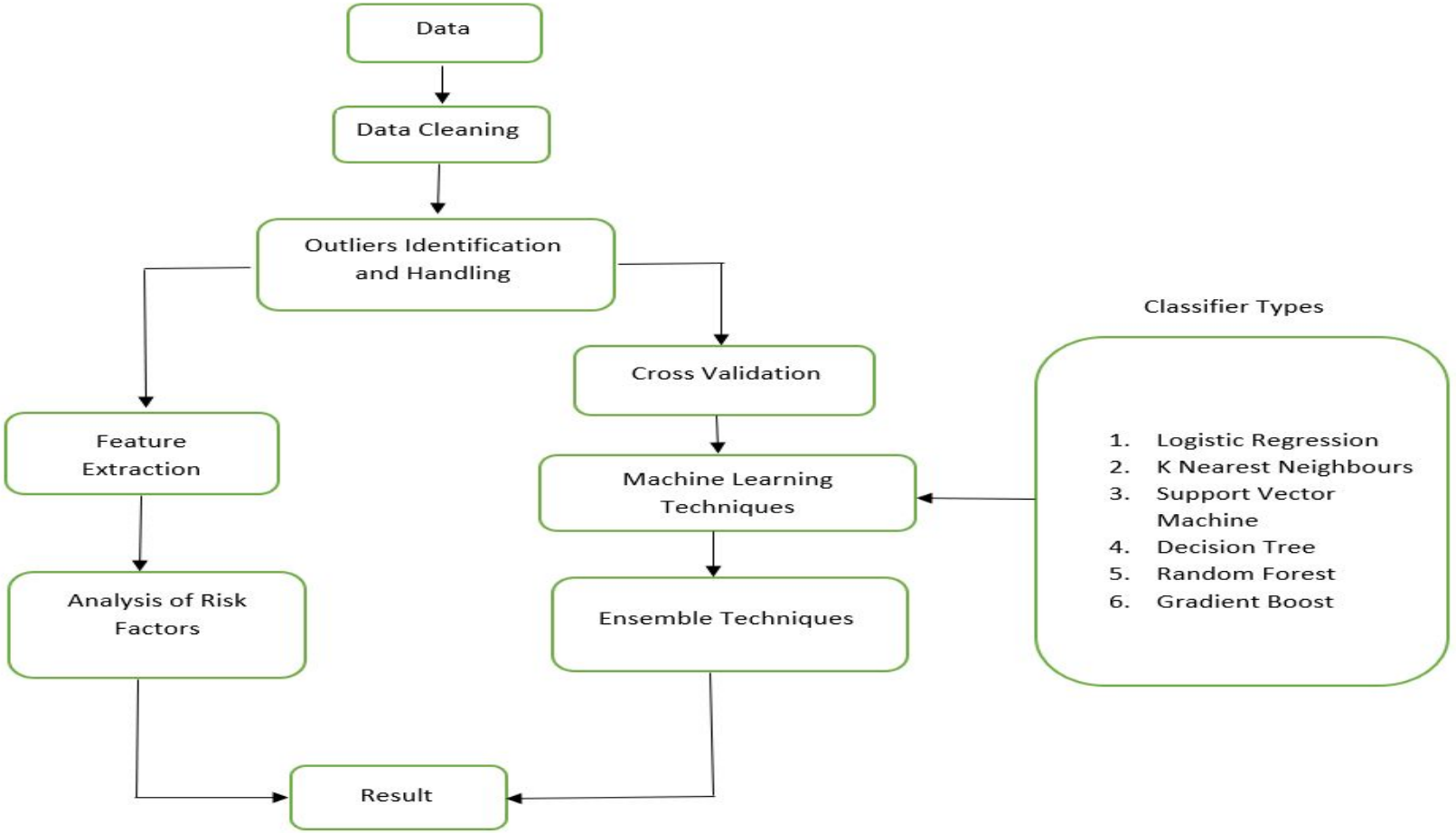
Medical (current):

- Tot Chol: total cholesterol level (Continuous)
- Sys BP: systolic blood pressure (Continuous)
- Dia BP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- Heart Rate: heart rate (Continuous- In medical research, variables such as heart rate through in fact discrete, yet are considered continuous because of large number of possible values)
- Glucose: glucose level (Continuous)

Predict variable (desired target)

- 10- year risk of coronary heart disease CHD (binary:”1”, means ”Yes”, “0” means “No”)-Dependent Variable

OVERVIEW


Data
Data Cleaning
Outliers Identification and Handling
Feature Extraction
Analysis of Risk Factors
Cross Validation
Machine Learning Techniques
Ensemble Techniques
Result
Classifier Types
1. Logistic Regression
2. K Nearest Neighbours
3. Support Vector Machine
4. Decision Tree
5. Random Forest
6. Gradient Boost

Data Preprocessing

Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model.

Handling null values

It is not the best option to remove the rows and columns from our dataset as it can result in significant information loss. So we can handle null values with the help of Imputation.

Imputation:

It is not the best option to remove the rows and columns from our dataset as it can result in significant information loss. So we can handle null values with the help of Imputation.
Imputation is simply the process of substituting the missing values of our dataset. We can do this by defining our own customised function or we can simply perform imputation by using the Simple Imputer class provided by sklearn.

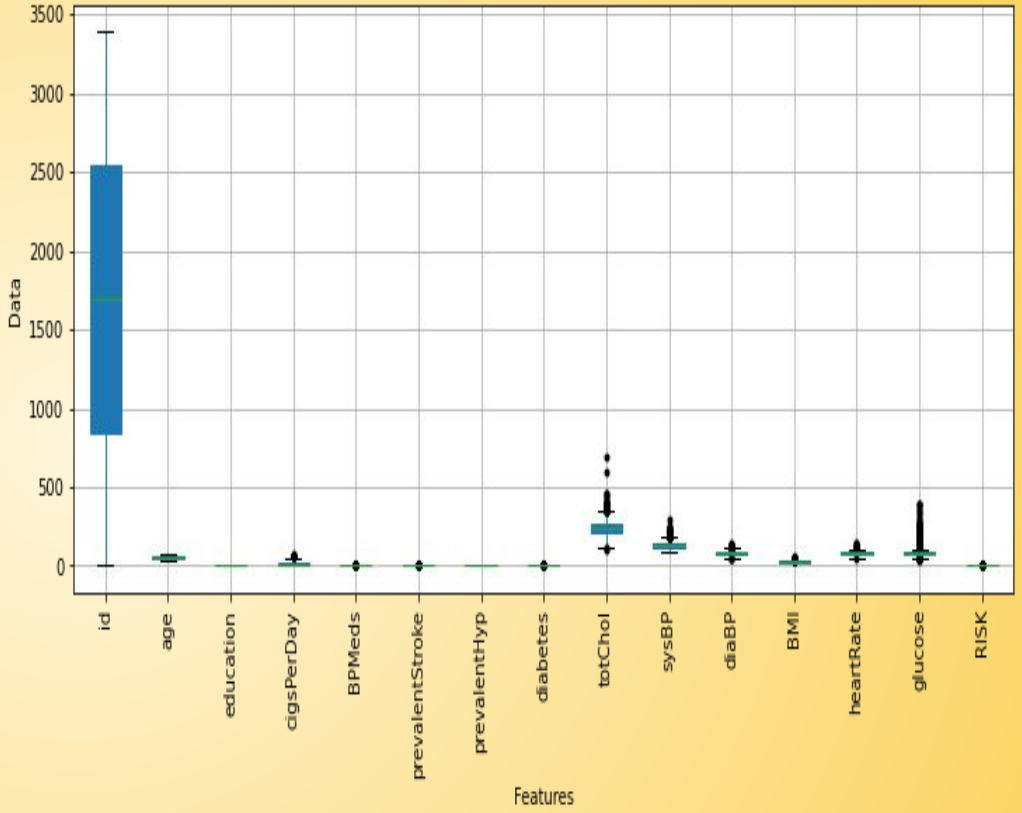
Label Encoding

We have to know that computers do not understand text data and thus, we need to convert these categories to numbers. A simple way of doing that can be to use label encoding

Outliers

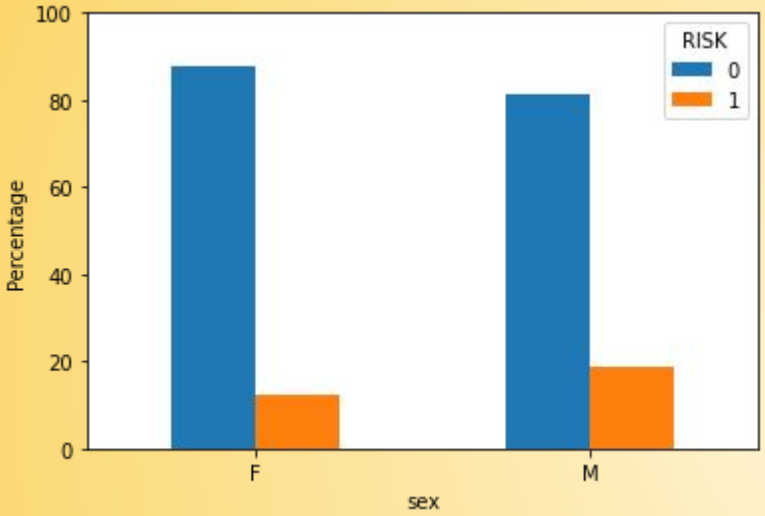
A box plot shows the distribution of quantitative data in a way that facilitates comparisons between variables.The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution.The box plot is a standardized way of displaying the distribution of data based on the five number summary:

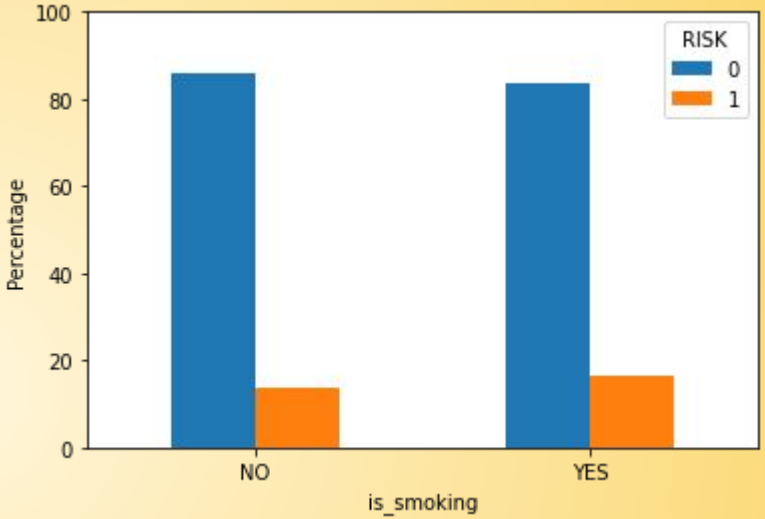
- Minimum
- First quartile
- Median
- Third quartile
- Maximum



In the simplest box plot the central rectangle spans the first quartile to the third quartile (the interquartile range or IQR).A segment inside the rectangle shows the median and “whiskers” above and below the box show the locations of the minimum and maximum.

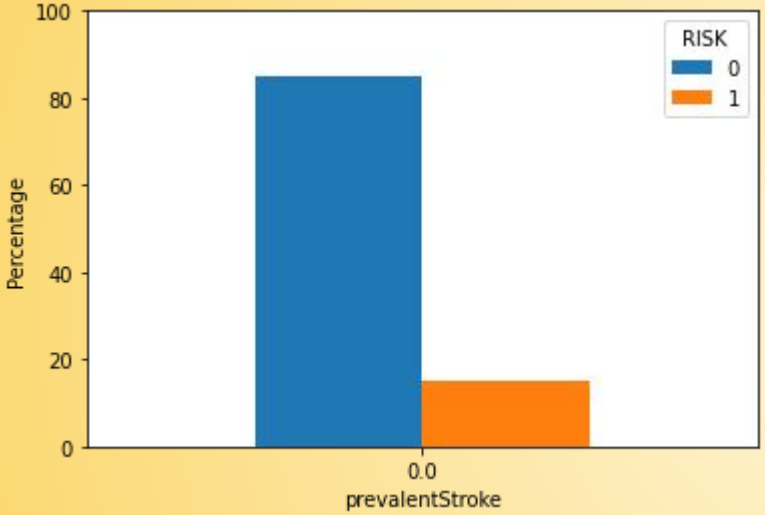
EDA

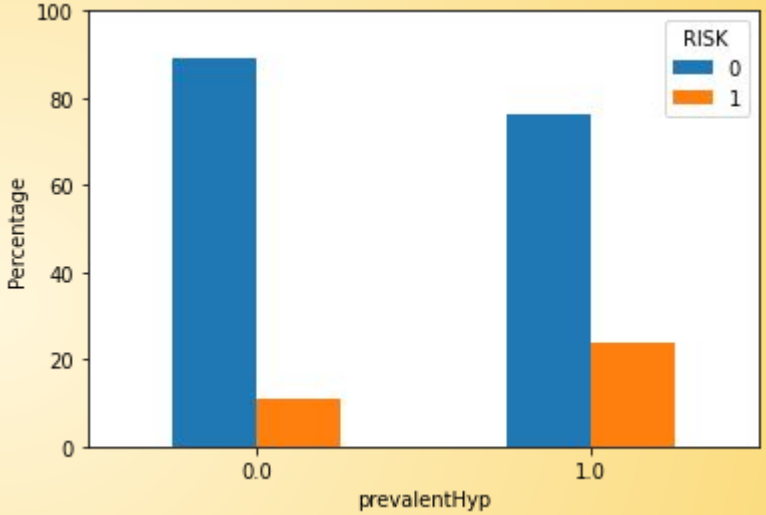




- **Male people have higher risk of coronary heart disease.**
- **Also smoking people are at higher risk.**

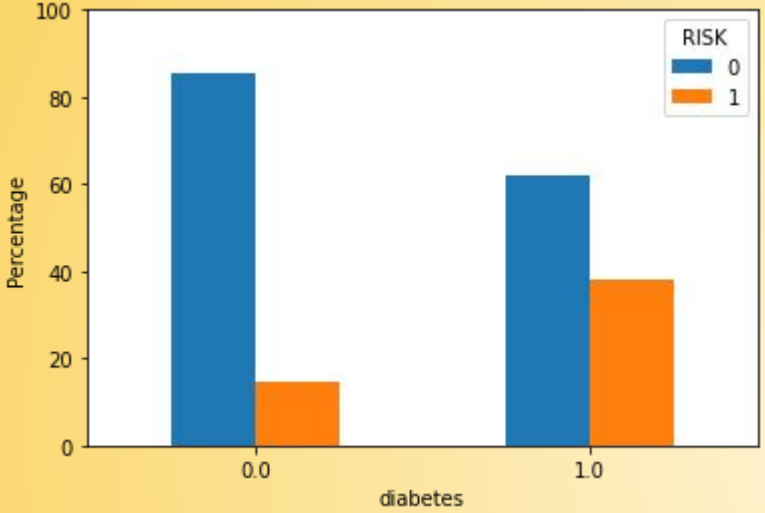
EDA

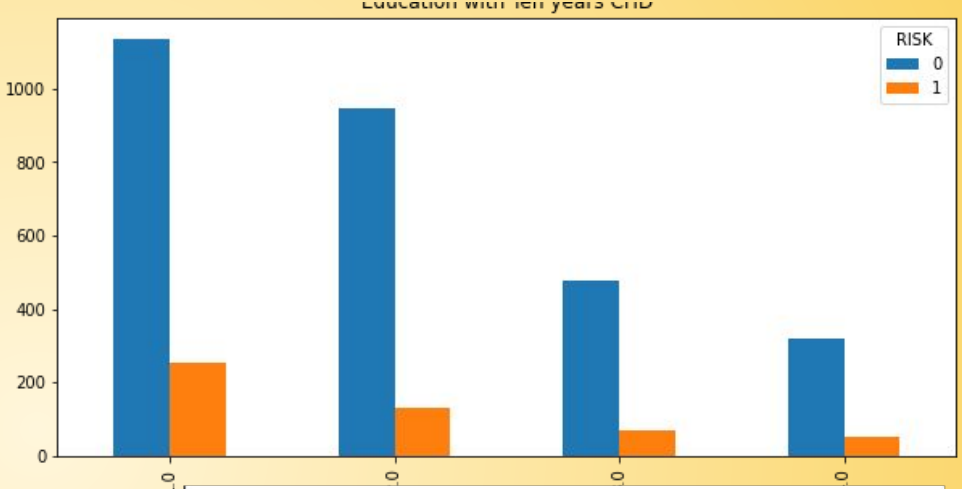


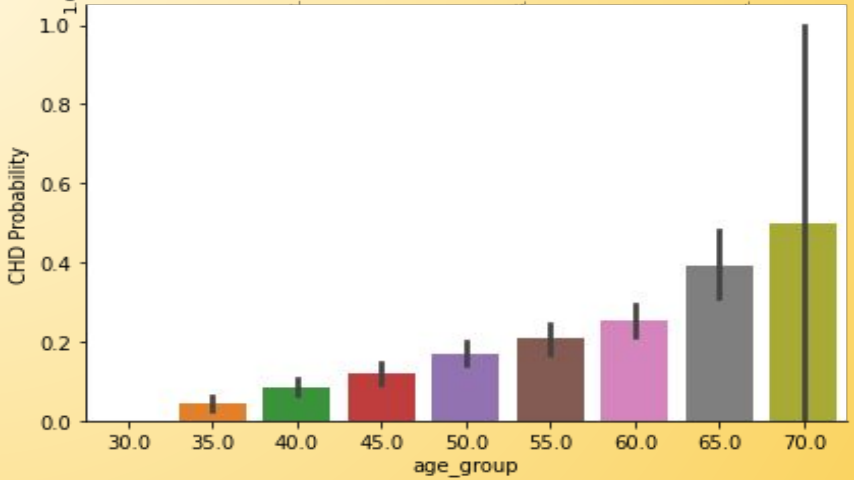


- **Patients, who have no prevalent stroke have higher chances of getting heart disease. Also people who have no history of hyper -tension have lesser chances of heart disease.**

EDA







- Diabetes patients have higher chances of risk.
- As education increases, the risk of heart disease decreases.
- Aged people have higher chances of heart diseases.

Feature Selection:

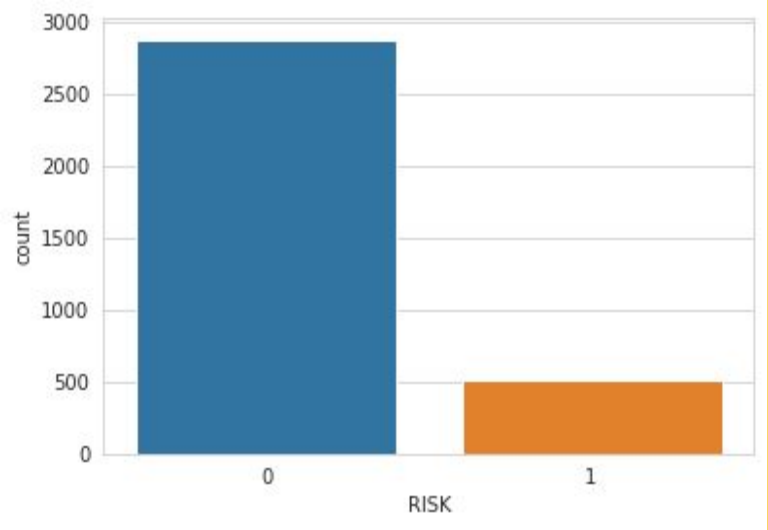
Using one-way ANOVA test to determine dependency between continuous variables and the target categorical variable.The test shows us that Ten year risk of CHD deos not dependent on the continuous variables like heartRate,id and is_smoking.

Imbalanced Data

The data is not properly balanced as the number of people without the disease greatly exceeds the number of people with the disease.

SMOTE:

It is a statistical technique for increasing the number of cases in the dataset in a balanced way.The component works by generating new instances from existing minority cases that you supply as input.



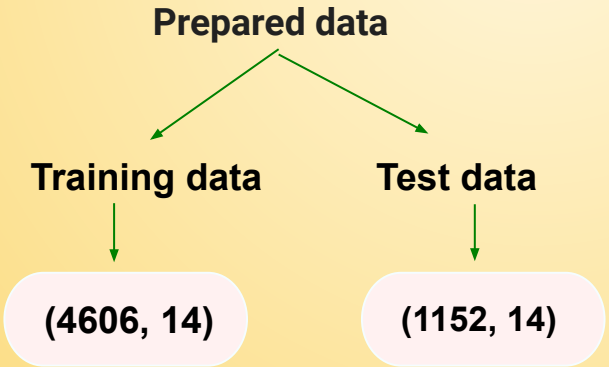
PREPARING DATASET FOR MODEL

Feature Scaling

Some machine learning algorithms are sensitive to feature scaling while others are virtually invariant to it.

Therefore, we scale our data before employing a distance based algorithm so that all the features contribute equally to the result.

StandardScaler will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. Given the distribution of the data, each value in the dataset will have the sample mean value subtracted, and then divided by the standard deviation of the whole dataset.



Splitting the 80% of the dataset into train_data and 20% of the dataset into test_data.

LOGISTIC REGRESSION

AI

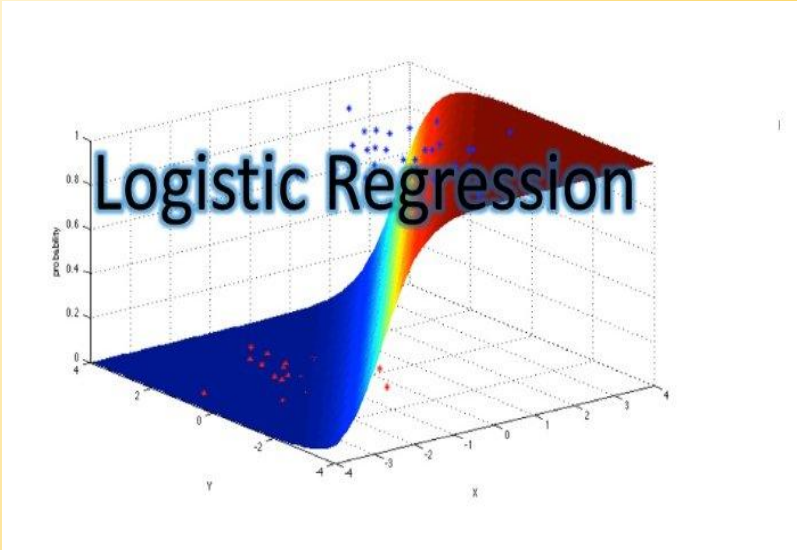
Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.

Choosing the right Cross-Validation

Choosing the right cross-validation depends on the dataset you are dealing with, and one's choice of cross-validation on one dataset may or may not apply to other datasets. However, there are a few types of cross-validation techniques which are the most popular and widely used. Here we are using:

k-fold cross-validation

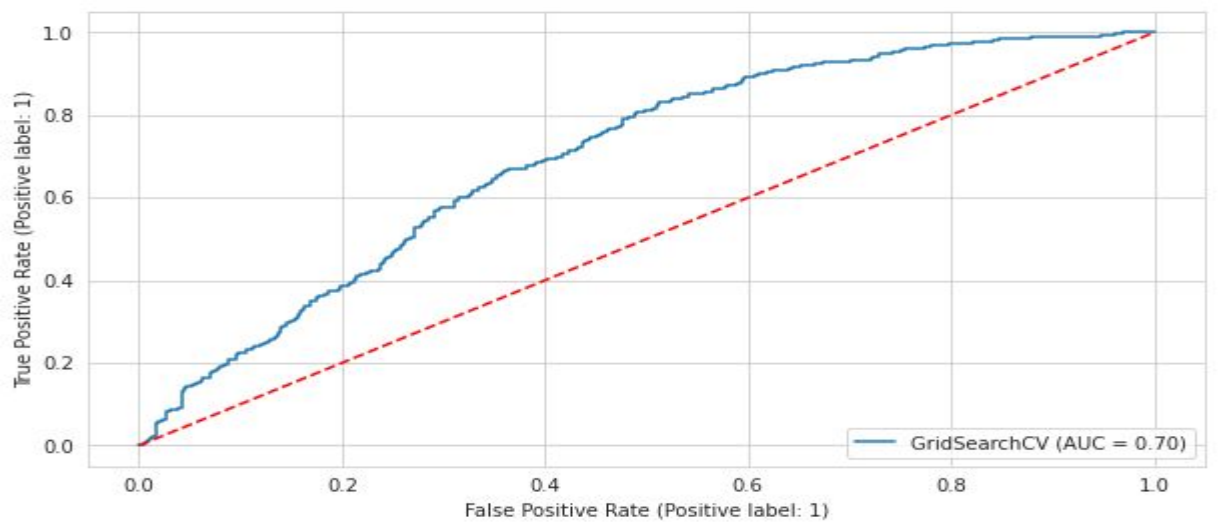
As you can see, we divide the samples and the targets associated with them. We can divide the data into k different sets which are exclusive of each other. This is known as k-fold cross-validation, We can split any data into k-equal parts using KFold from scikit-learn. Each sample is assigned a value from 0 to k-1 when using k-fold cross validation.

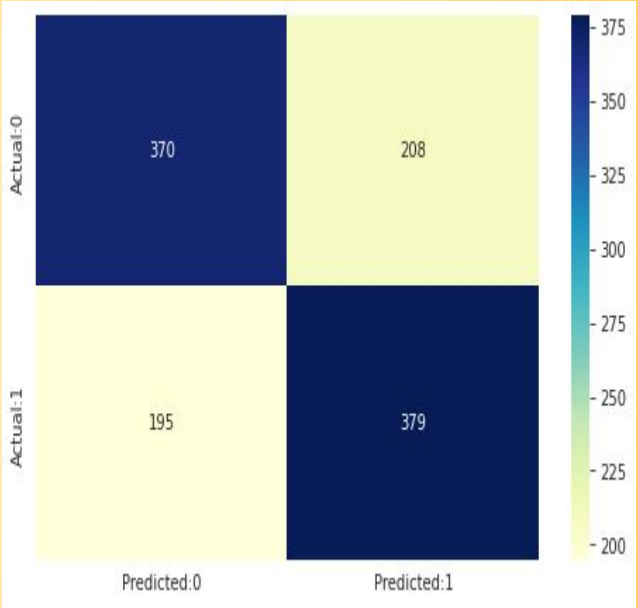


PERFORMANCE METRICS

accuracy=0.6501736111111112

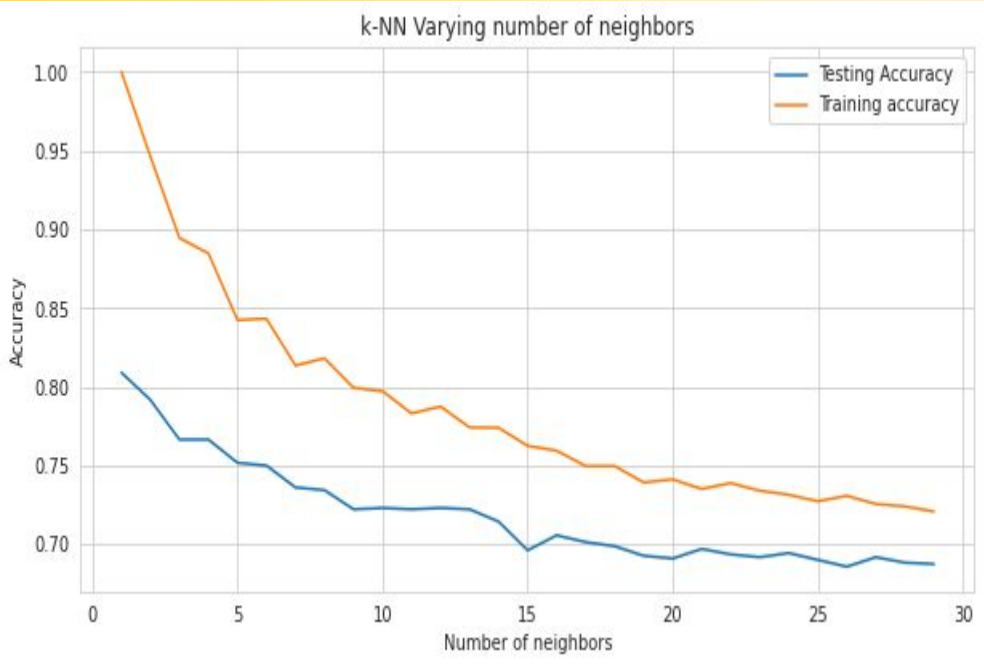
	precision	recall	f1-score	support
0	0.65	0.64	0.65	578
1	0.65	0.66	0.65	574
accuracy			0.65	1152
macro avg	0.65	0.65	0.65	1152
weighted avg	0.65	0.65	0.65	1152

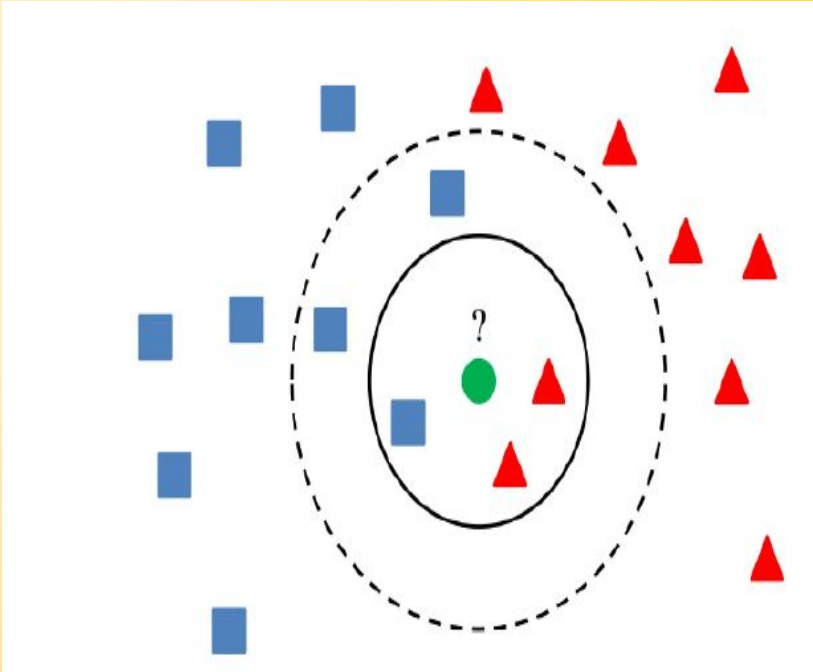




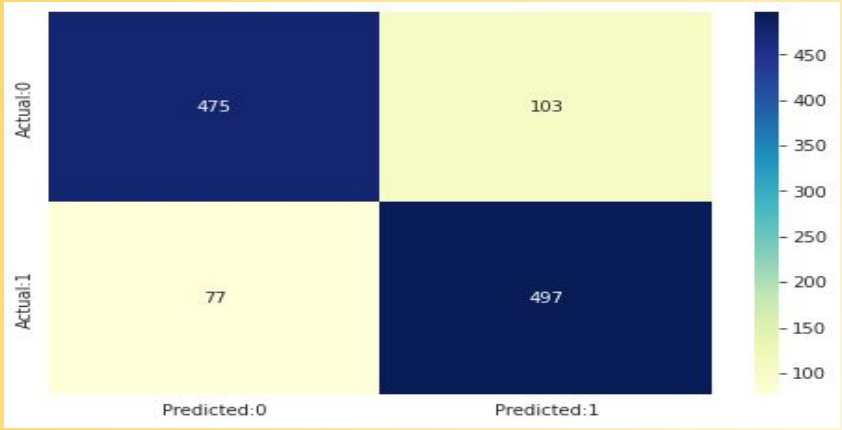
K Nearest Neighbor

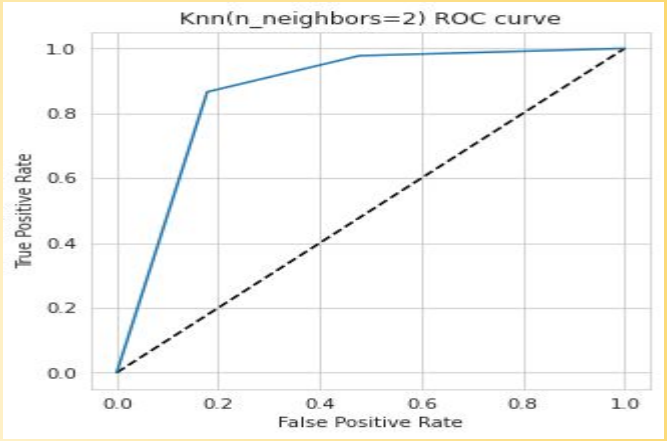
K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition.





PERFORMANCE METRICS



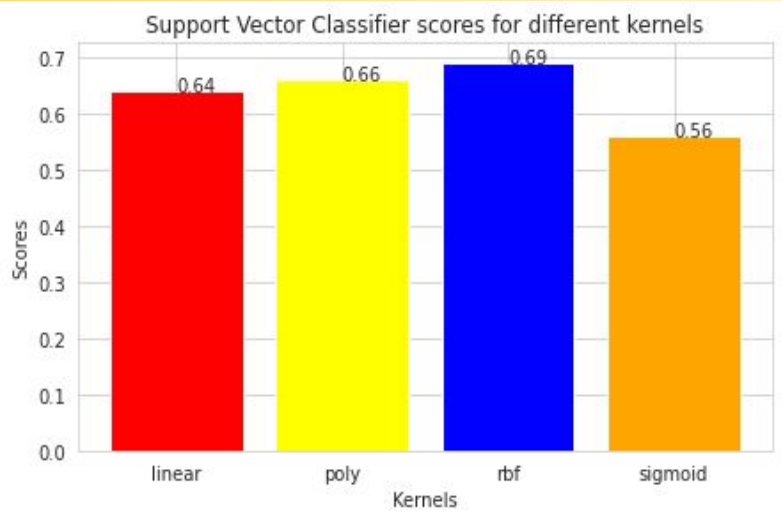


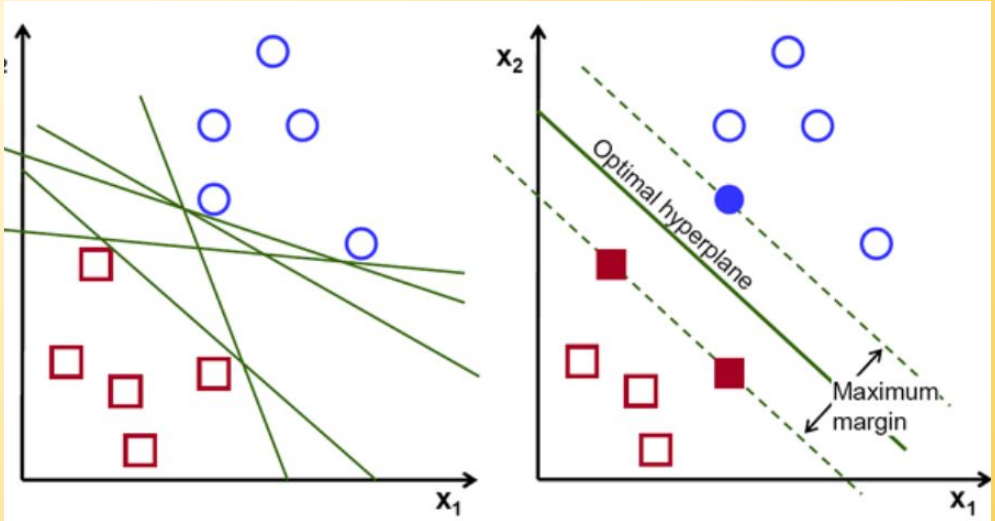
accuracy=0.84375

	precision	recall	f1-score	support
0	0.86	0.82	0.84	578
1	0.83	0.87	0.85	574
accuracy			0.84	1152
macro avg	0.84	0.84	0.84	1152
weighted avg	0.84	0.84	0.84	1152

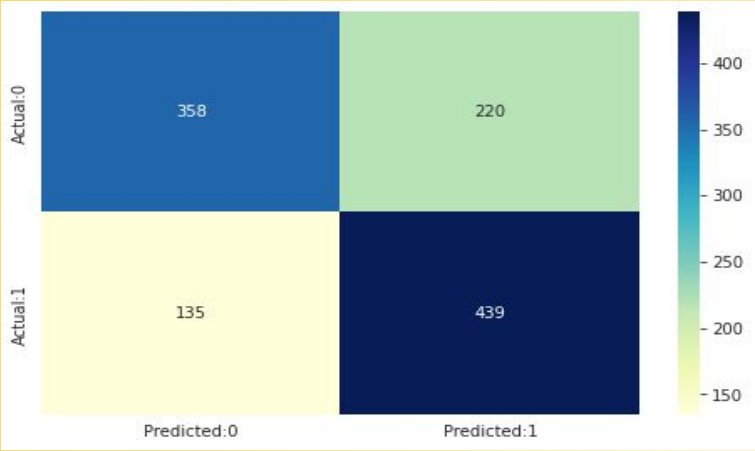
SVM

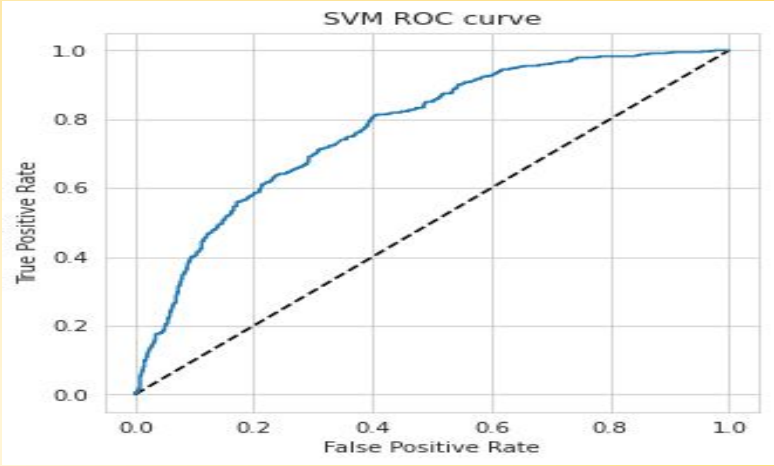
A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.From the classification approach, the goal of SVM is to find a hyperplane in an N-dimensional space that clearly classifies the data points. Thus hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes.





PERFORMANCE METRICS



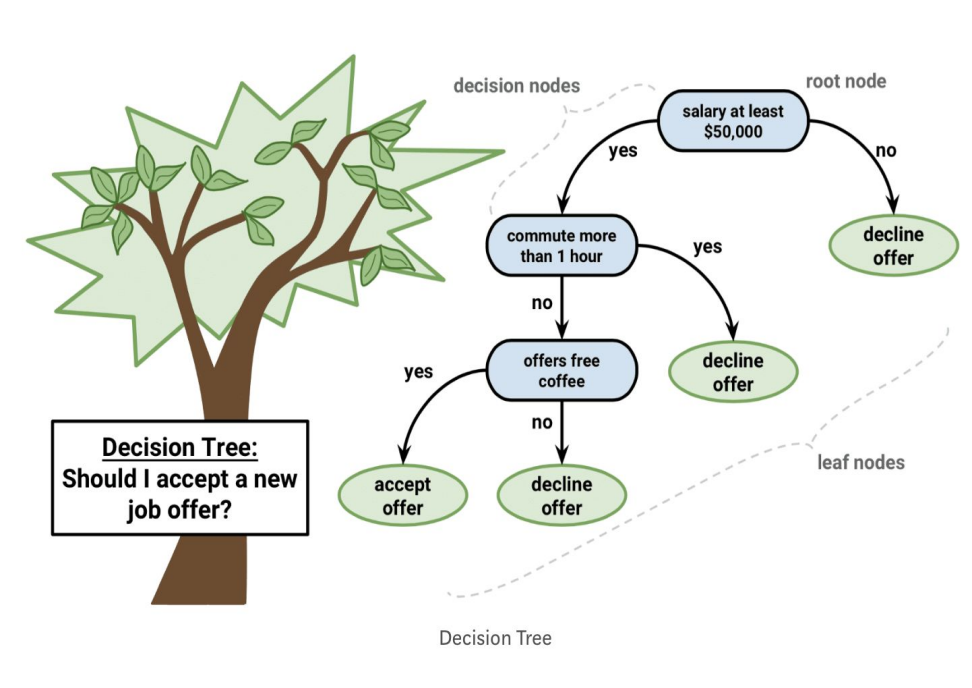


accuracy=0.6918402777777778

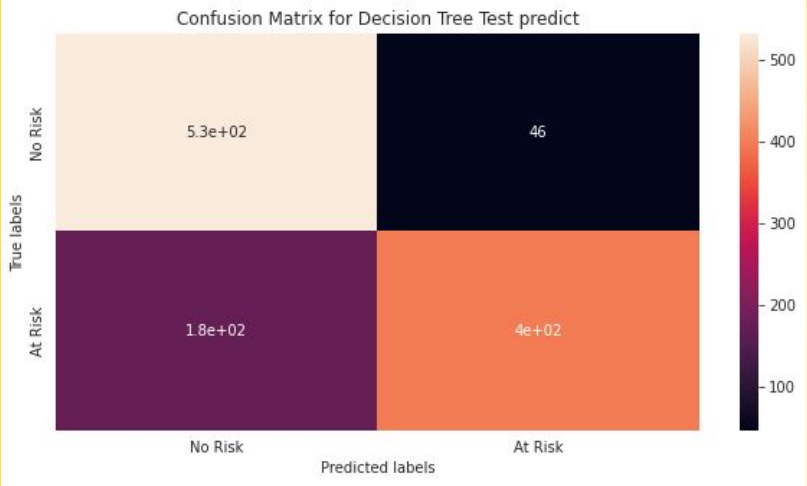
	precision	recall	f1-score	support
0	0.73	0.62	0.67	578
1	0.67	0.76	0.71	574
accuracy			0.69	1152
macro avg	0.70	0.69	0.69	1152
weighted avg	0.70	0.69	0.69	1152

Decision Tree

Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables.

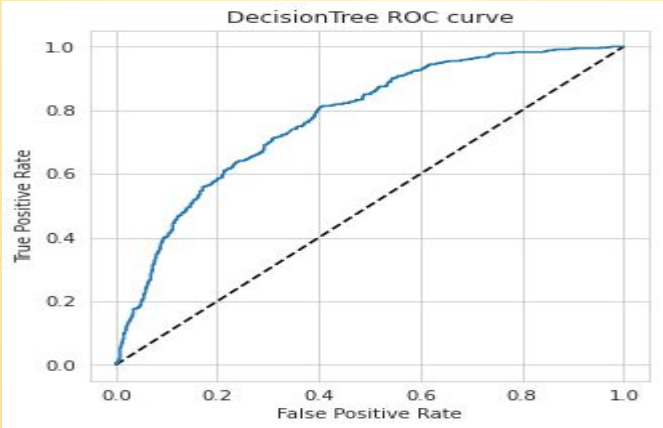


PERFORMANCE METRICS



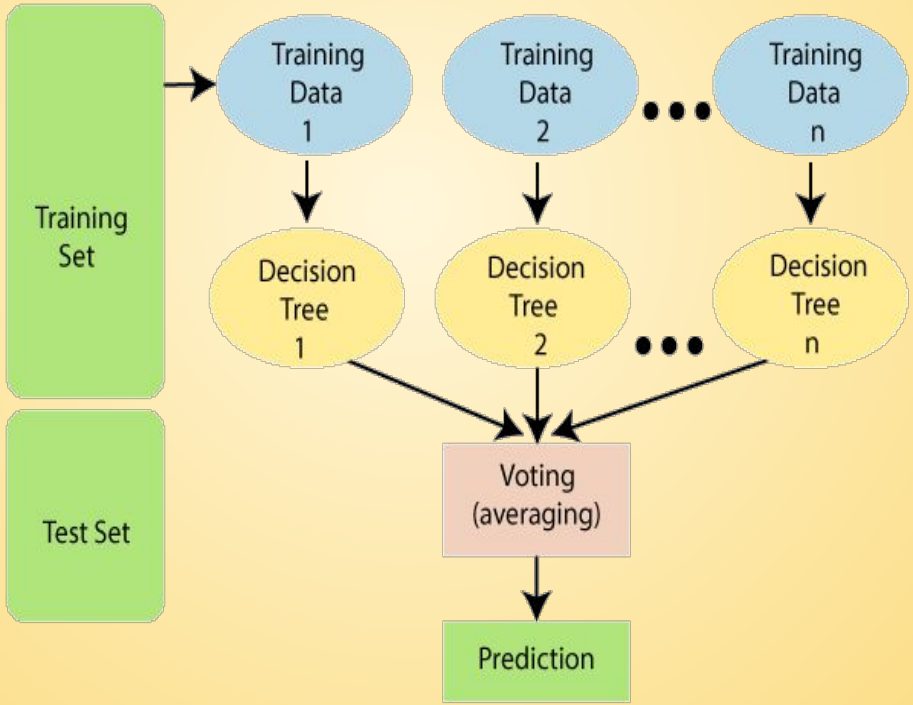
accuracy=0.8081597222222222

	precision	recall	f1-score	support
0	0.75	0.92	0.83	578
1	0.90	0.70	0.78	574
accuracy			0.81	1152
macro avg	0.82	0.81	0.81	1152
weighted avg	0.82	0.81	0.81	1152



Random Forest

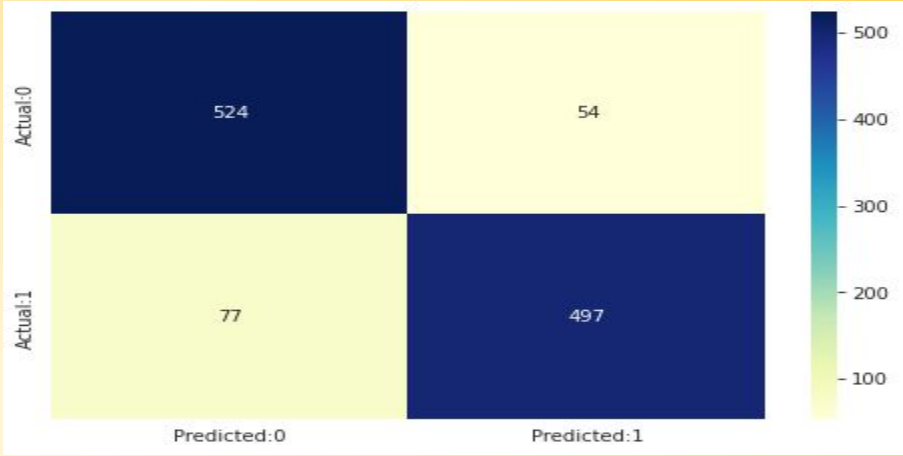
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

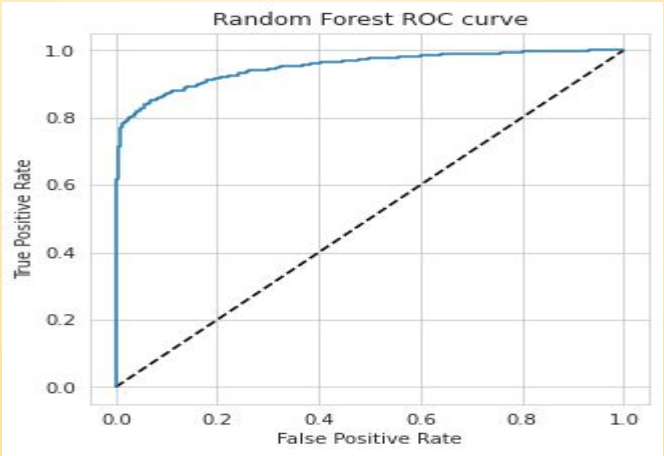


PERFORMANCE METRICS

accuracy=0.8862847222222222

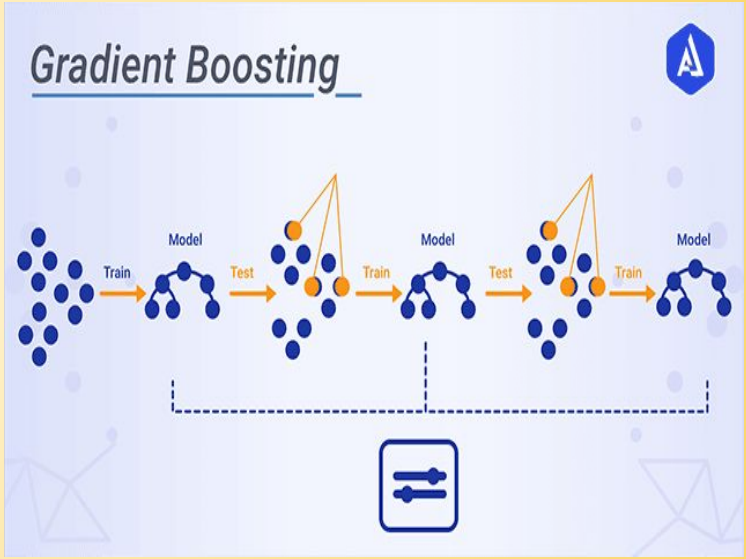
	precision	recall	f1-score	support
0	0.87	0.91	0.89	578
1	0.90	0.87	0.88	574
accuracy			0.89	1152
macro avg	0.89	0.89	0.89	1152
weighted avg	0.89	0.89	0.89	1152





Gradient Boosting

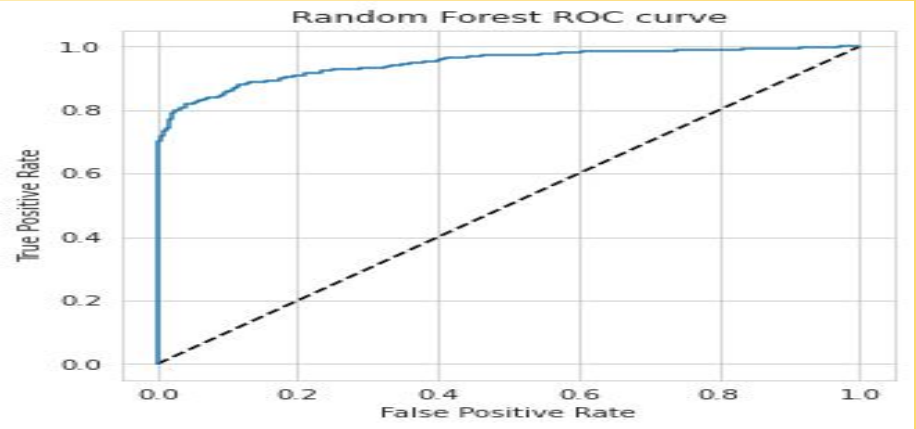
Gradient boosting is a type of machine learning boosting.
It relies on the intuition that the best possible next
Model, when combined with previous models,minimizes
the overall prediction error.
The key idea is to set the target outcomes for
This next model in order to minimize the error.

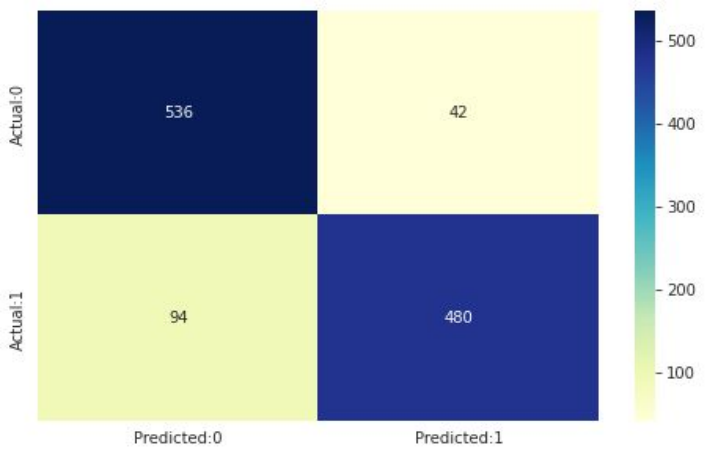


PERFORMANCE METRICS

accuracy=0.8819444444444444

	precision	recall	f1-score	support
0	0.85	0.93	0.89	578
1	0.92	0.84	0.88	574
accuracy			0.88	1152
macro avg	0.89	0.88	0.88	1152
weighted avg	0.89	0.88	0.88	1152





Plotting the Accuracy of the models

Here we plot the performance or the accuracy of the different machine learning model, in this plot we observe that the different models have different performance.

Pretty Table

SL NO	MODEL_NAME	Accuracy	F1-score_class0	F1-score_class1	Precision_class0	Precision_class1	Recall_class0	Recall_class1
1	Logistic Regression	0.65	0.65	0.65	0.65	0.65	0.64	0.66
2	KNearest Neighbors	0.84	0.84	0.85	0.86	0.83	0.82	0.87
3	SVM	0.69	0.67	0.71	0.73	0.67	0.62	0.76
4	Decision Tree	0.80	0.83	0.78	0.75	0.90	0.92	0.70
5	Random Forest	0.88	0.89	0.88	0.87	0.90	0.91	0.87
6	Gradient Boosting	0.88	0.89	0.88	0.85	0.92	0.93	0.84

- Logistic Regression is the poor model and Random forest is the best model.
- Gradient Boosting has almost the the same accuracy as Random Forest.
- Both Random Forest and Gradient Boosting are Ensemble Techniques.They are performing Equally good.
- Logistic Regression and SVM are not good models.

CONCLUSION

- We started with the data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important. During this process we used Plotly, seaborn and matplotlib to do the visualizations.
- During the data preprocessing part, we converted features into numeric ones, grouped values into categories and created a few new features. Afterwards we started training machine learning models, and applied cross validation on it.
- Of course there is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features.
- Lastly, we looked at it's confusion matrix and computed the models precision.
- The highest accuracy score was achieved with the random forest classification method and Gradient Boosting method.Both these techniques are ensemble techniques.