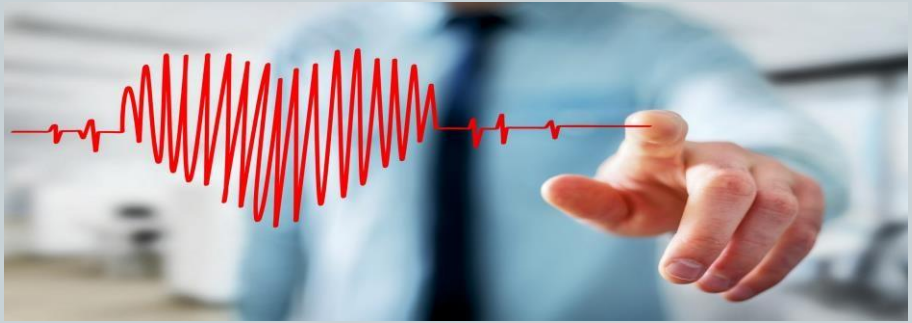
Capstone Project-3



CARDIOVASCULAR RISK PREDICTION

Kumar Vijay Mhaske

About Project:

The dataset is from an ongoing cardiovascular study 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 patients' information. It includes over 4,000 records and 15 attributes.

POINTS FOR DISCUSSION :

- Problem Statement
- DATA DESCRIPTION
- Data Cleaning
- Extra Data Analysis(EDA)
- FEATURE ENGINEERING/SELECTION
- Splitting Data
- MODELLING AND PREDICTING WITH MACHINE LEARNING
- MODELS
- Conclusion

Problem Statement

- Perform classification analysis using multiple models to predict risk of future coronary heart disease (CHD) and compare the evaluation metrics for all of them to find the best model.
- Predict the overall risk of heart disease using Classification regression
- Data balancing for Train Model
- Getting accuracy score of several machine learning model.

DATA DESCRIPTION

- The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.
- VARIABLES: - EACH ATTRIBUTE IS A POTENTIAL RISK FACTOR. THERE ARE BOTH DEMOGRAPHIC, BEHAVIORAL, AND MEDICAL RISK FACTORS

DEMOGRAPHIC

1. Sex: male or female ("M" or "F")

2.Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of
age is continuous)

BEHAVIORAL

3. is_smoking: whether or not the patient is a current smoker ("YES" or "NO")

4.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)

5. BP Meds: whether or not the patient was on blood pressure medication (Nominal)
6. Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
7. Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
8. Diabetes: whether or not the patient had diabetes (Nominal)

MEDICAL(CURRENT)

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

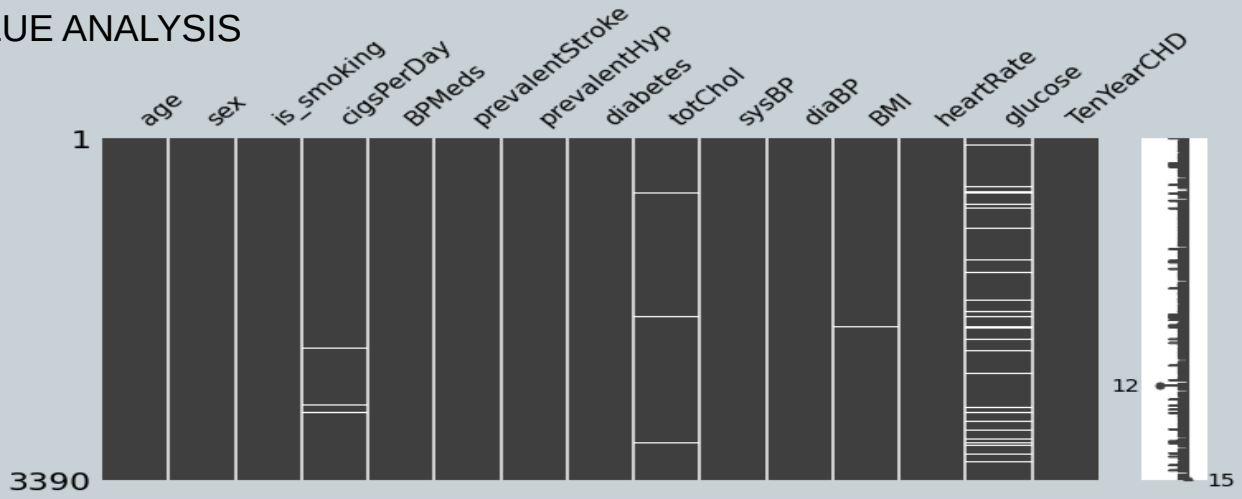
PREDICT VARIABLE (DESIRED TARGET)

15. TenYearCHD: 10-year risk of coronary heart disease CHD (binary: 1 means “Yes”, 0 means “No”) - DV

DATA EXPLORATION

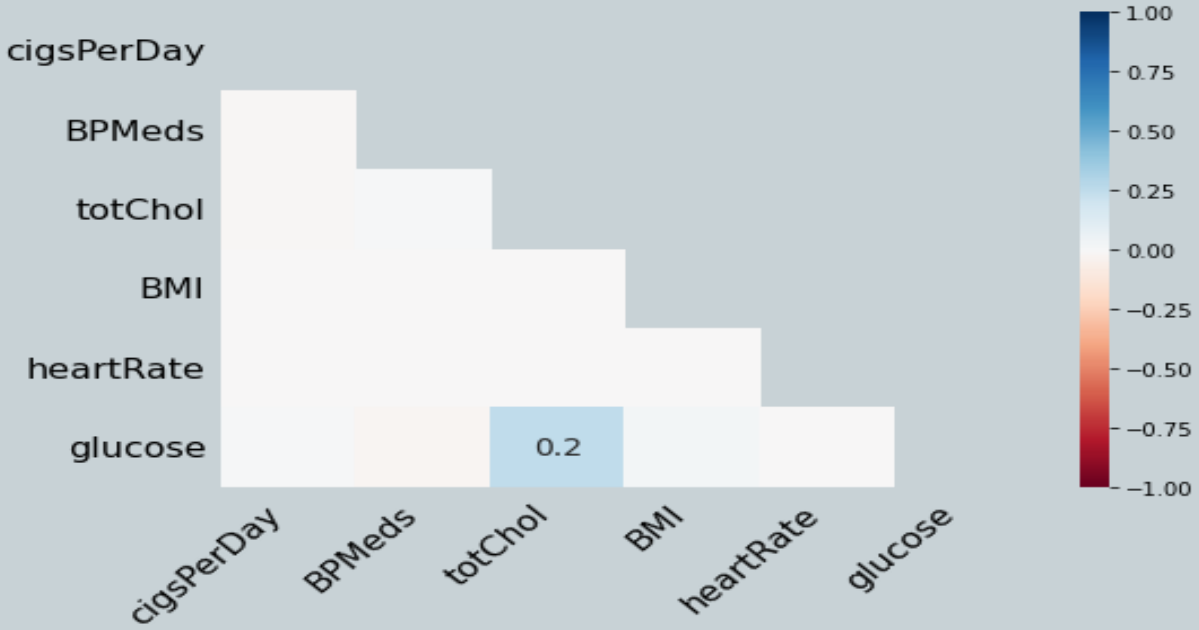
- This Dataset has 3390 observations in it with 17 columns(features)
- Before any analysis, we just wanted to take a look at the data. So, we used the info () method.
- There are a total of 16 features and 1 target variable. Also, there are some missing values so we need to take care of null values. Next, we used describe() method.

MISSING VALUE ANALYSIS



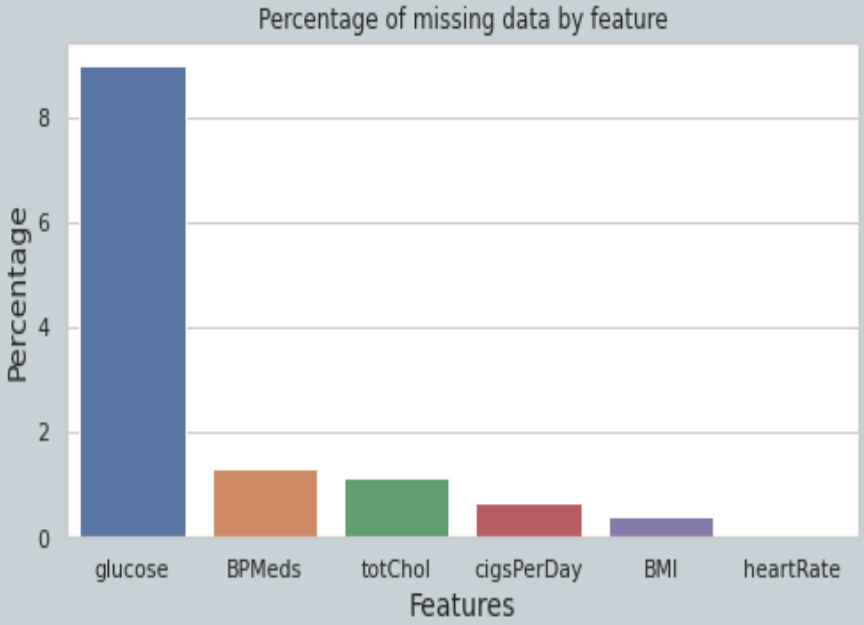
- These trends give an idea about how the features are correlated with one another

But to get a better idea about correlations we need to use heatmaps.



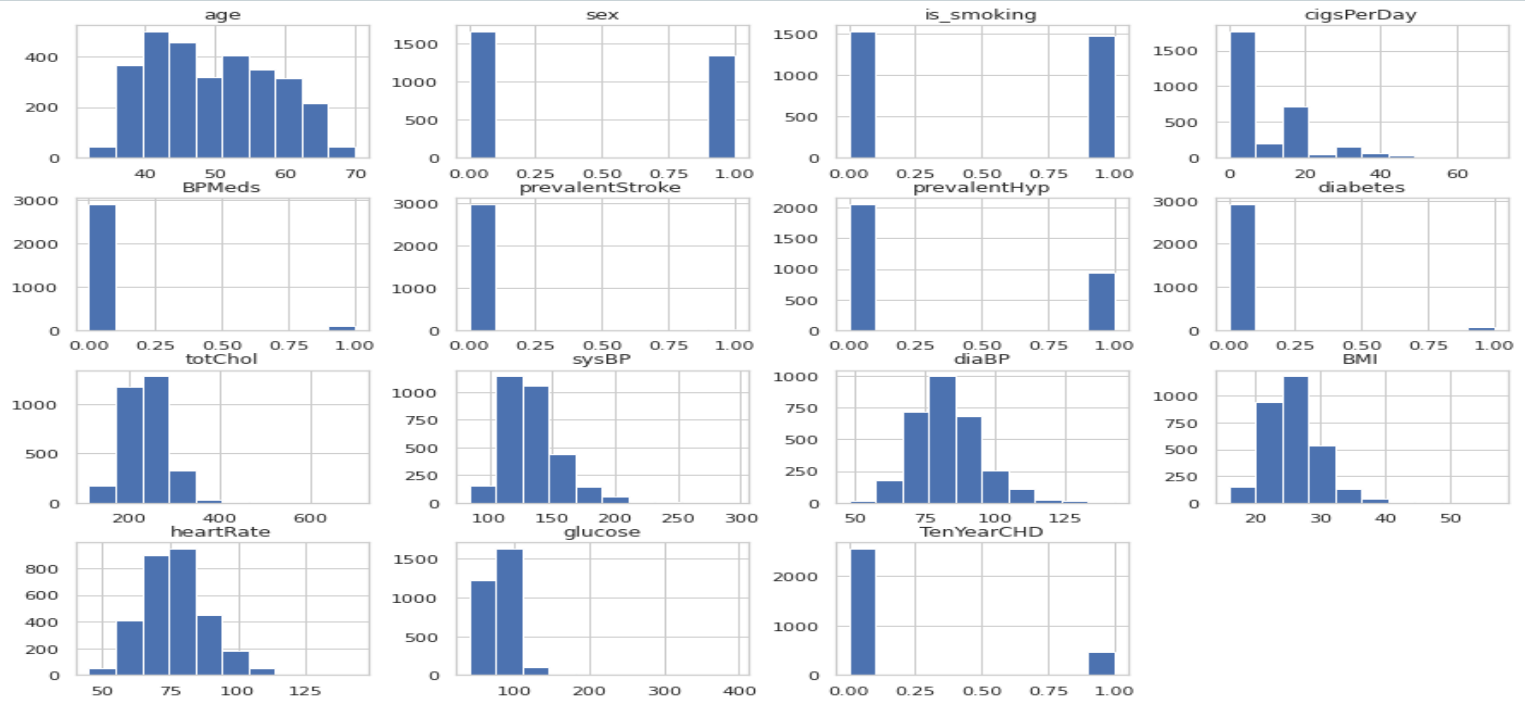
- There is no strong Correlation between any features.

LET'S CHECK PERCENTAGE OF MISSING DATA OF EACH FEATURES



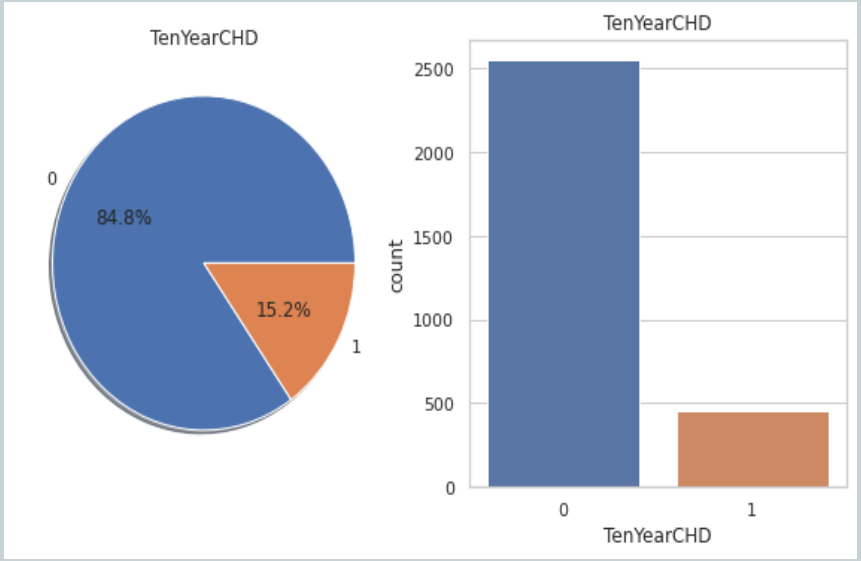
- At 8.97%, the blood glucose entry has the highest percentage of missing data. The other features have very few missing entries.
- BPMeds have near to 1.29% of missing data
- totChol has near to 1.12% missing data.
- cigsPerDay has near to 0.64% missing data.
- BMI has near to 0.41% missing data.
- heartRate has near to 0.02% missing data.
- Since the missing entries account for only 11% of the total data so, we can drop these entries without losing a lot of data.

NOW, LET'S VISUALIZE DATA DISTRIBUTION



- From above distribution plot we can say that the data on the prevalent stroke, diabetes, and blood pressure meds(BPMeds) are poorly balanced.

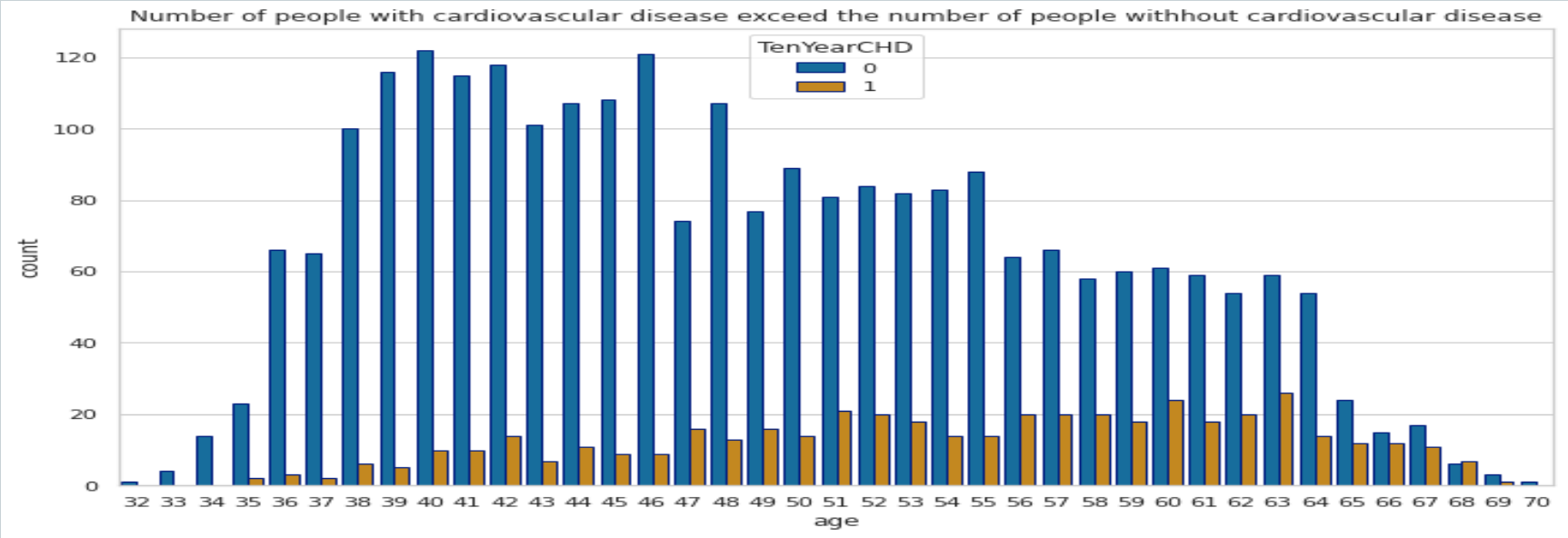
TARGET VARIABLE ANALYSIS



1 --> Person do not have risk of coronary heart disease
2 --> Person has risk of coronary heart disease

- There are 2547 patients without heart disease and 457 patients with the disease.
- We can see above that we have the imbalanced data set as the number of people without the disease greatly exceeds the number of people with the disease.

Let's look at the number of people with cardiovascular disease exceed the number of people without cardiovascular disease respect to age.



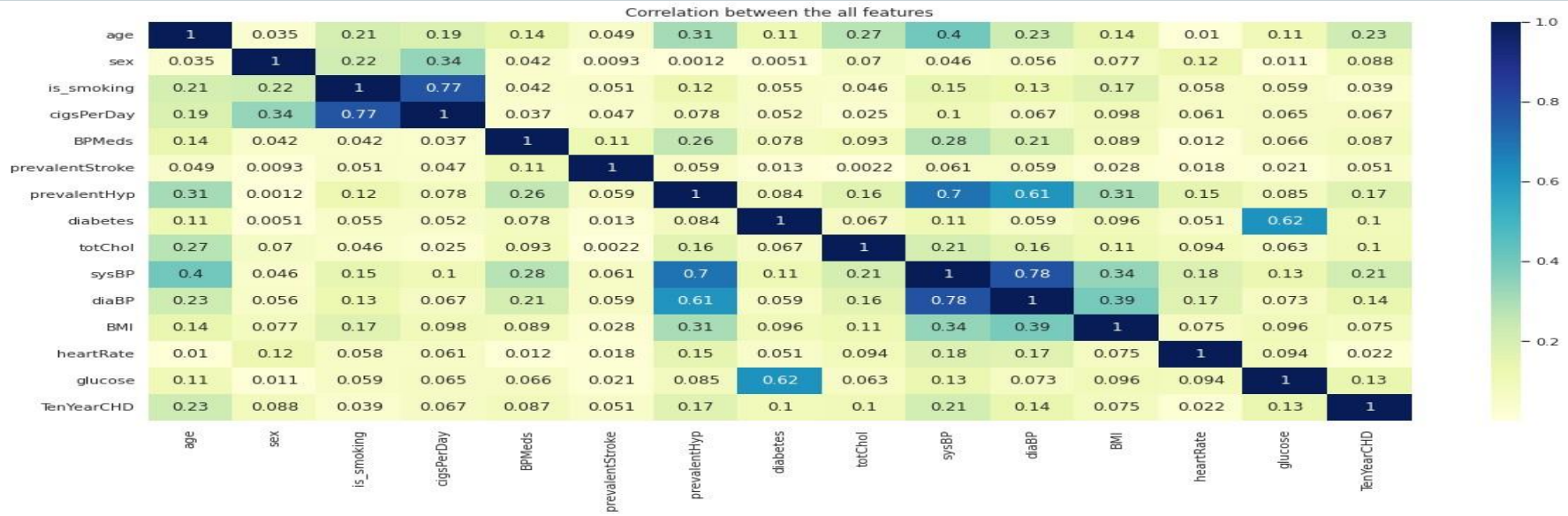
- As we can see in above plot The people with the highest risk of developing heart disease are between the ages of 51 and 63.
- Because the number of sick people generally increases with age.

CATEGORICAL VARIABLE COMPARISONS WITH TARGET VARIABLE(TENYEARCHD)

From the above categorical variable comparison plot we can conclude that,

- Slightly more males are suffering from Cardiovascular heart disease than females.
- The people who have Cardiovascular heart disease is almost equal between smokers and non-smokers.
- The percentage of people who have Cardiovascular heart disease is higher among the diabetic patients and also those patients with prevalent hypertension have more risk of Cardiovascular heart disease compare to those who don't have hypertensive problem.
- The percentage of people who are on medication of blood pressure have more risk of Cardiovascular heart disease compare to those who are not on medication

NOW, LET'S SEE THE CORRELATION BETWEEN THE ALL FEATURES



From the above correlation plot we can conclude that,

- There are no features with more than 0.2 correlation with the Ten-year risk of developing CHD and this shows that the features a poor predictor. However, the features with the highest correlations are age, prevalent hypertension(prevalentHyp) and systolic blood pressure(sysBP).
- Also, there are a couple of features that are highly correlated with each other and it makes no sense to use both of them in building a machine learning model.

These includes:

- Blood glucose and diabetes;
- systolic and diastolic blood pressures;
- cigarette smoking and the number of cigarettes smoked per day.

Therefore, we need to carry out feature selection to pick the best features.

FEATUREENGINEERING/SELECTION

Tree-based: SelectFromModel

SelectFromModel is an Embedded method. Embedded methods use algorithms that have built-in feature selection methods.

Here,

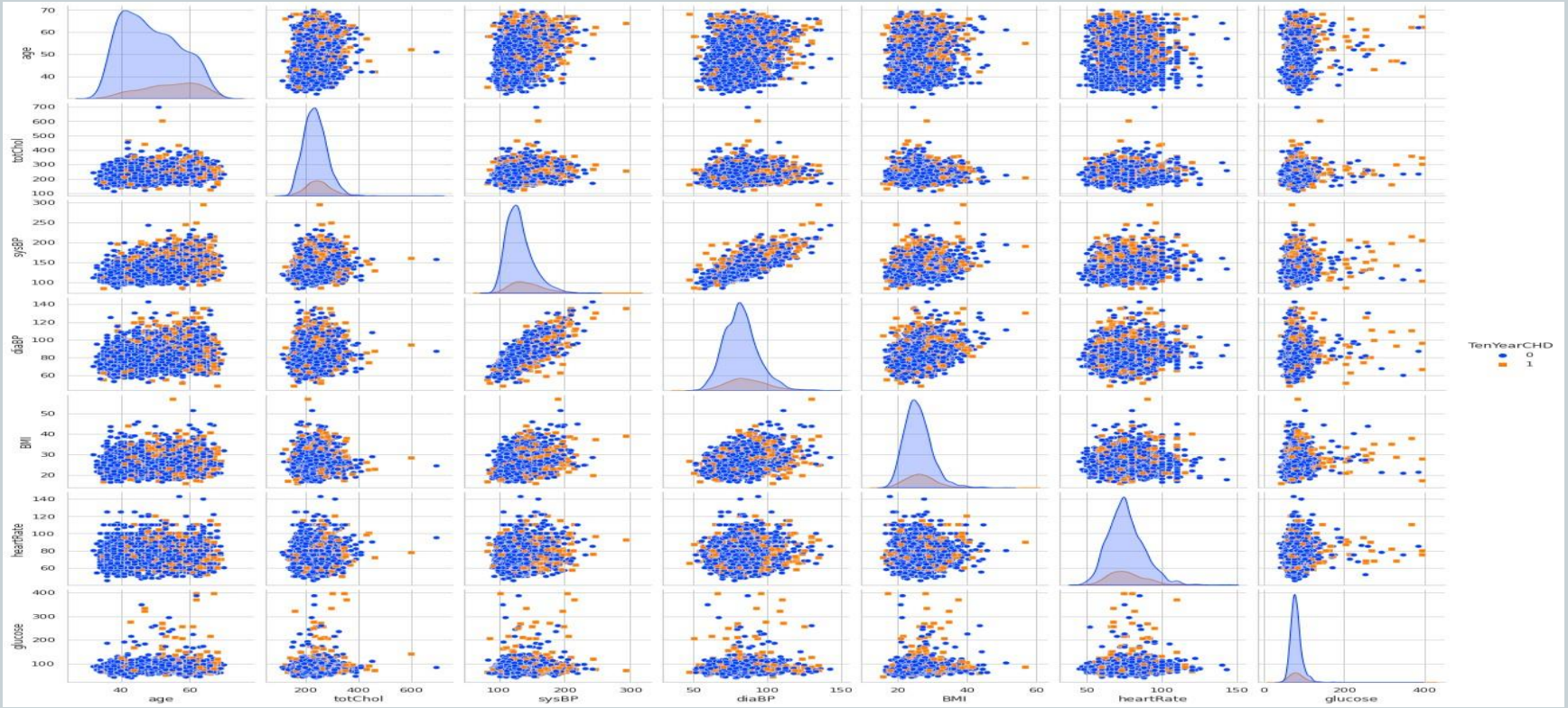
We have used Random Forest() to select features based on feature importance. We calculate feature importance using node impurities in each decision tree.

In Random Forest, the final feature importance is the average of all decision tree feature importance.

The top features are:

1. Age
2. Total cholesterol
3. Systolic blood pressure
4. Diastolic blood pressure
5. BMI
6. Heart rate
7. Blood glucose

LET’S VISUALIZE THROUGH PLOTTING PAIR PLOT OF TOP FEATURES VS TARGET VARIABLE



- So, we can easily find relation between all features with target variable

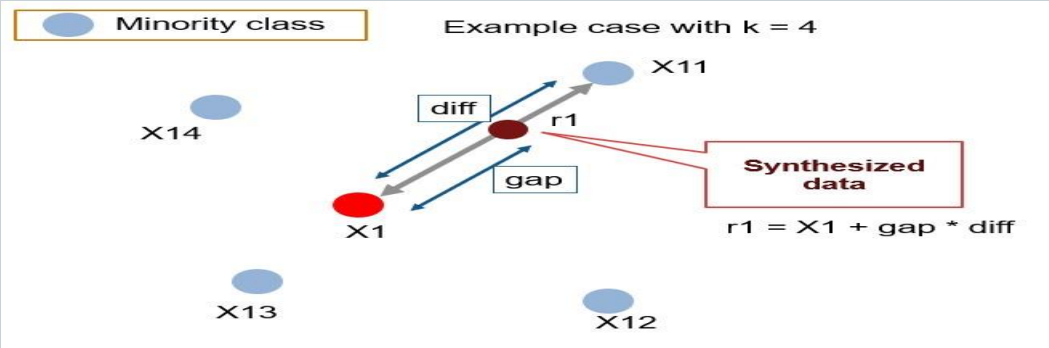
MODELLING AND PREDICTING WITH MACHINE LEARNING

Since our dataset is imbalanced i.e. for every positive case there are about 5-6 negative cases. We may end up with a classifier that is biased to the negative cases. The classifier may have a high accuracy but poor a precision and recall.

- To handle this problem, we will balance the dataset using the Synthetic Minority Oversampling Technique(SMOTE).

SMOTE: Synthetic Minority Oversampling Technique

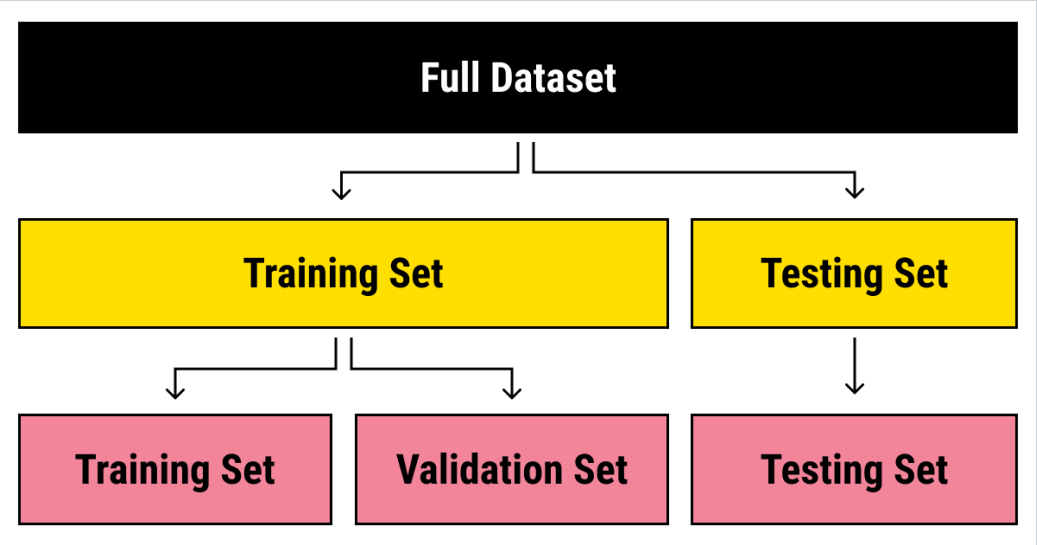
SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.



- As seen after applying SMOTE, the new dataset is much more balanced.

Splitting Data

- Data splits into training dataset and testing dataset.
- Training dataset is for making algorithm learn and train model.
- Test dataset is for testing the performance of train model.
- Here 80% of data taken as training dataset & remaining 20% of dataset used for testing purpose.
- Training features have 4075 records and Testing features have 1019 records.



MODELS

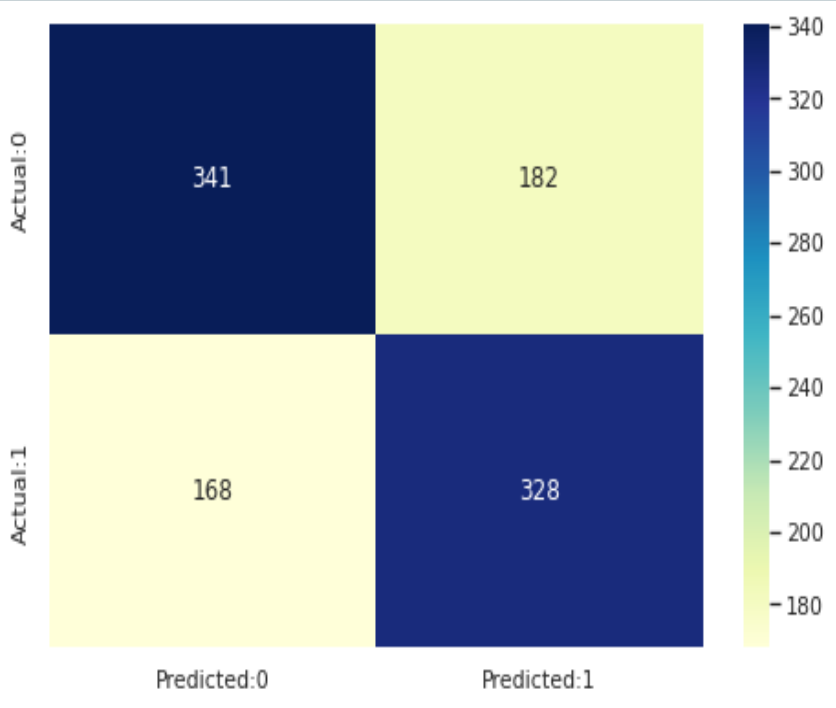
The four algorithms that we will be using are:
1. Logistic Regression
2. Random Forrest
3. XGBoost
4. Support Vector Machine

Here, we will be using GridsearchCV search algorithm for above algorithms

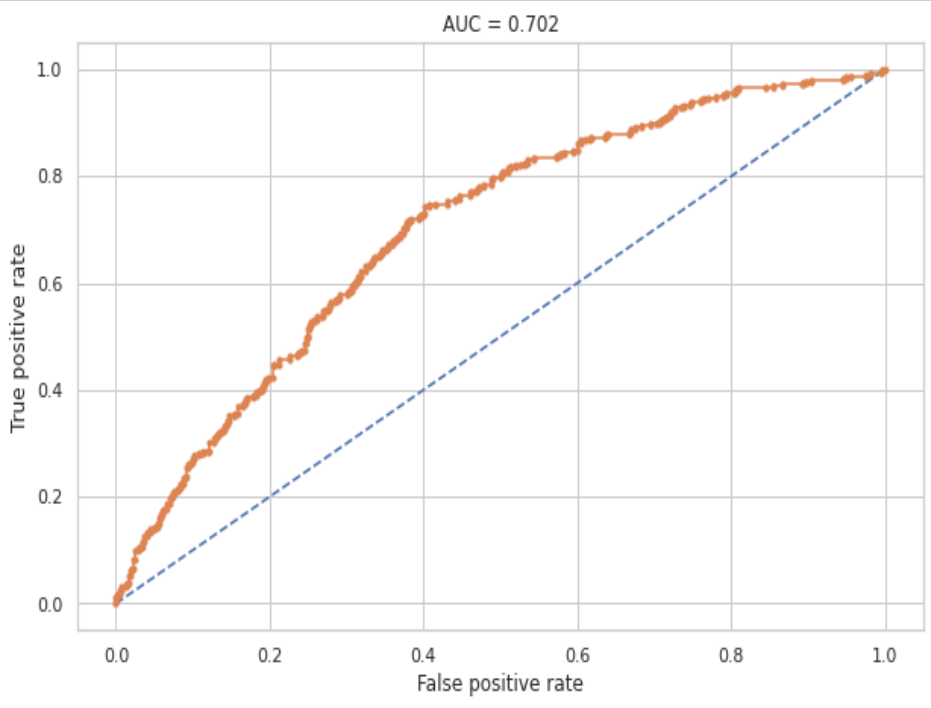
LOGISTIC REGRESSION

Logistic regression aims to measure the relationship between a categorical dependent variable and one or more independent variables (usually continuous) by plotting the dependent variables' probability scores.

Confusion matrix of Logistic Model



Auc Score of Logistic Model

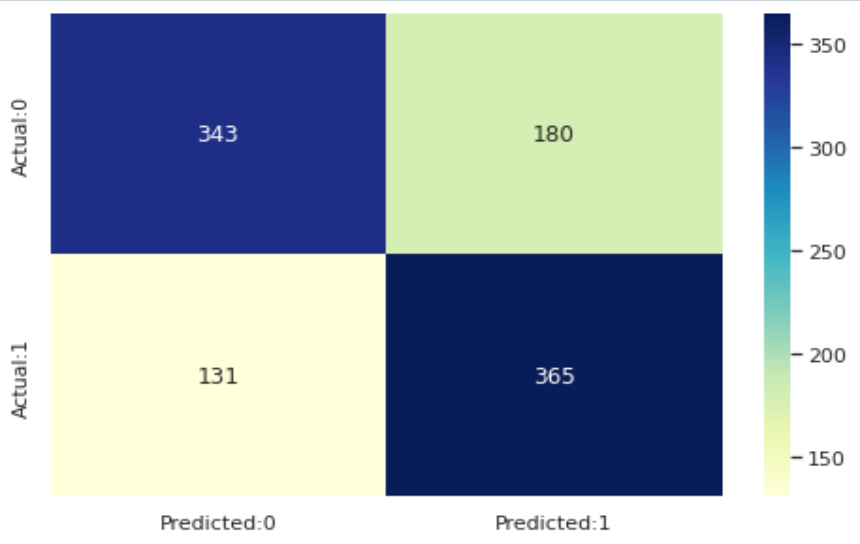


- ✓ Using logistic regression, we get an accuracy of 65.95%

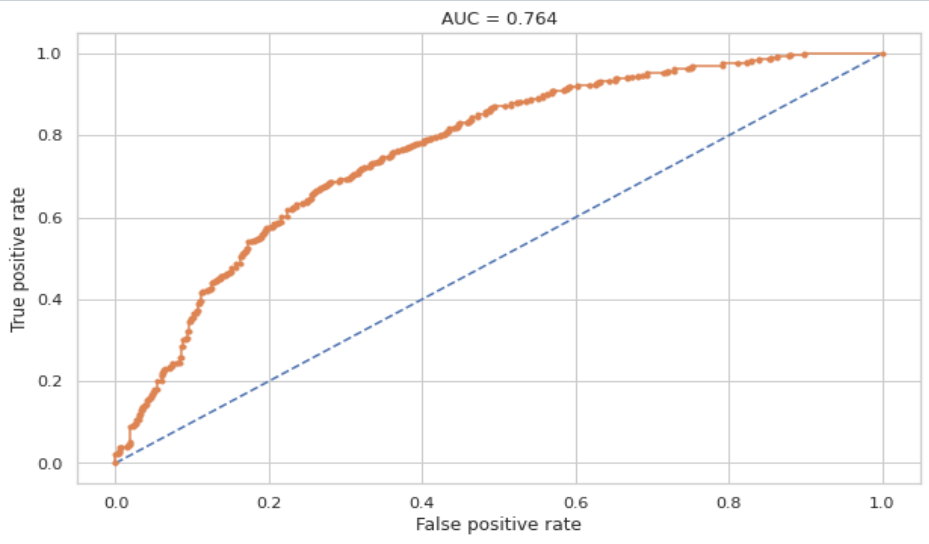
RANDOM FOREST

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model

Confusion matrix of Random Forest



Auc Score of Logistic Model

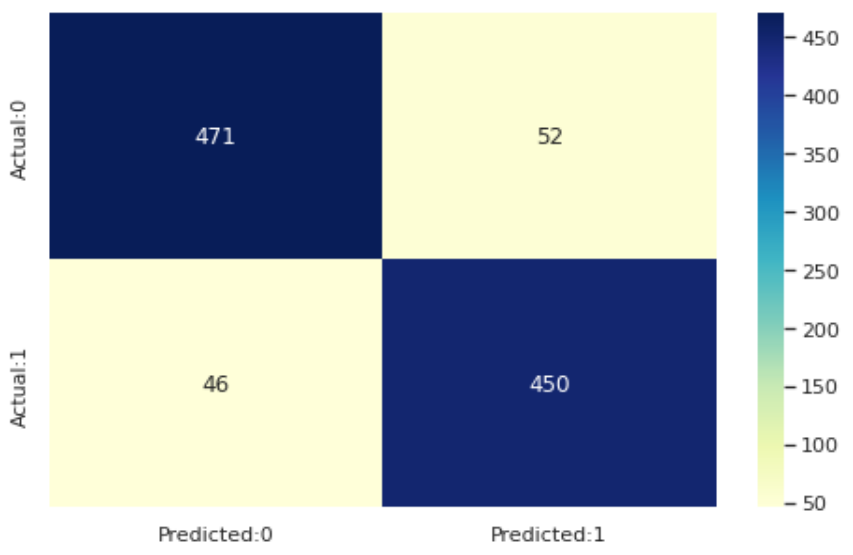


✓ Using Random Forest, we get an accuracy of 68.4%

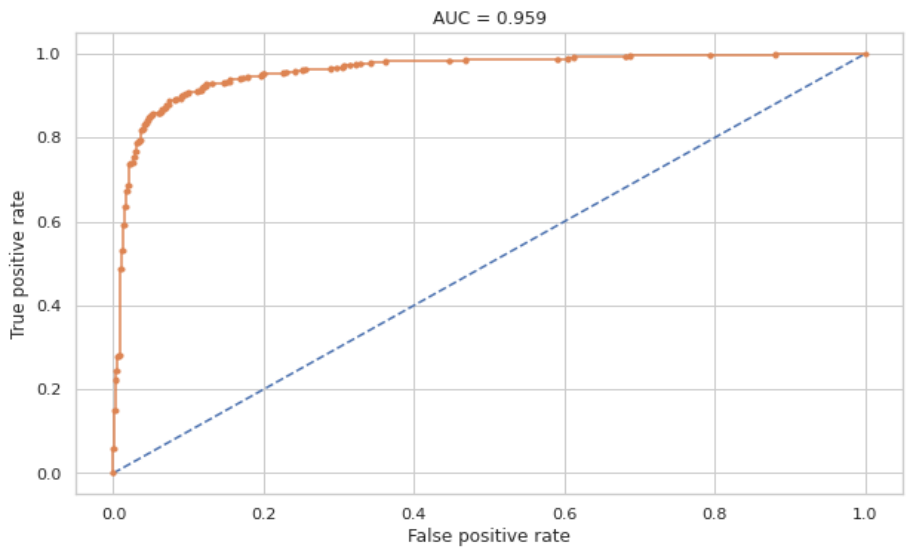
XGBOOST

XGBoost stands for extreme Gradient Boosting. The name xgboost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms

Confusion matrix of XG boost Classifier



Auc score of XG boost Classifier

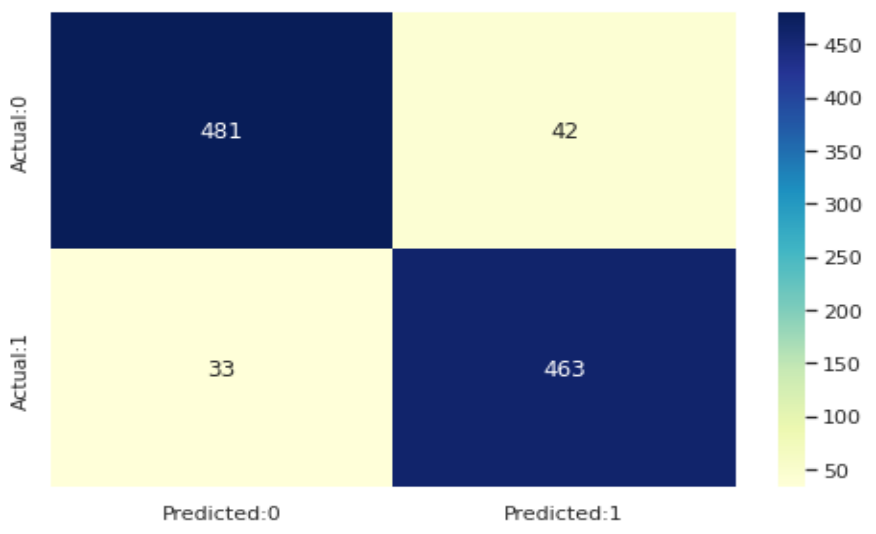


✓ Using XG boost we get an accuracy of 89.7%

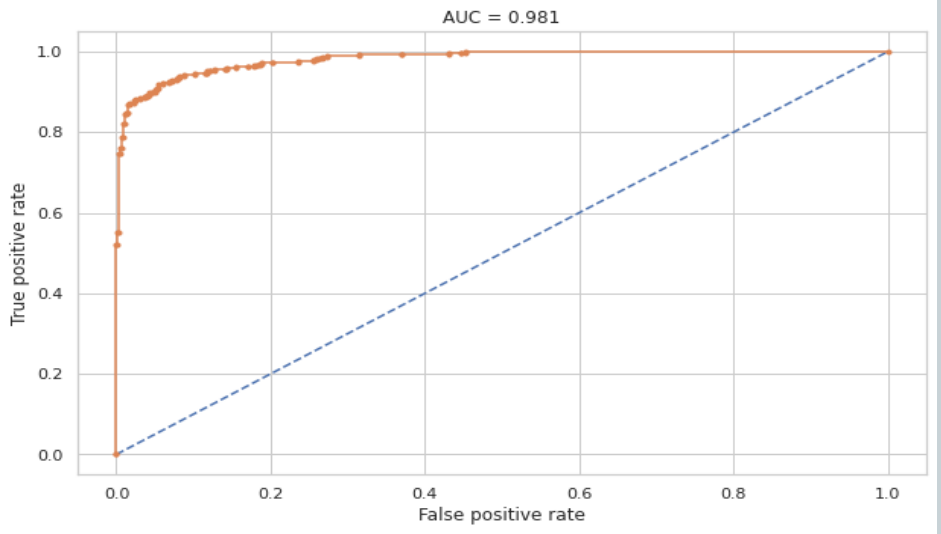
SUPPORT VECTOR MACHINE

- Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems.
- An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

Confusion matrix of SVM



Auc score of SVM



✓ Using Support Vector Machine, we get an accuracy of 92.64%.

LET'S COLLECT ALL OUR BEST MODELS

Creating data frame which shows the performance metrics of each model

	Test Accuracy	Precision	Recall	F1 Score	AUC
Logistic regression	0.66	0.64	0.67	0.66	0.71
Random Forest	0.68	0.66	0.74	0.70	0.76
XG Boost	0.90	0.88	0.91	0.90	0.96
Support vector machine	0.93	0.92	0.93	0.93	0.98

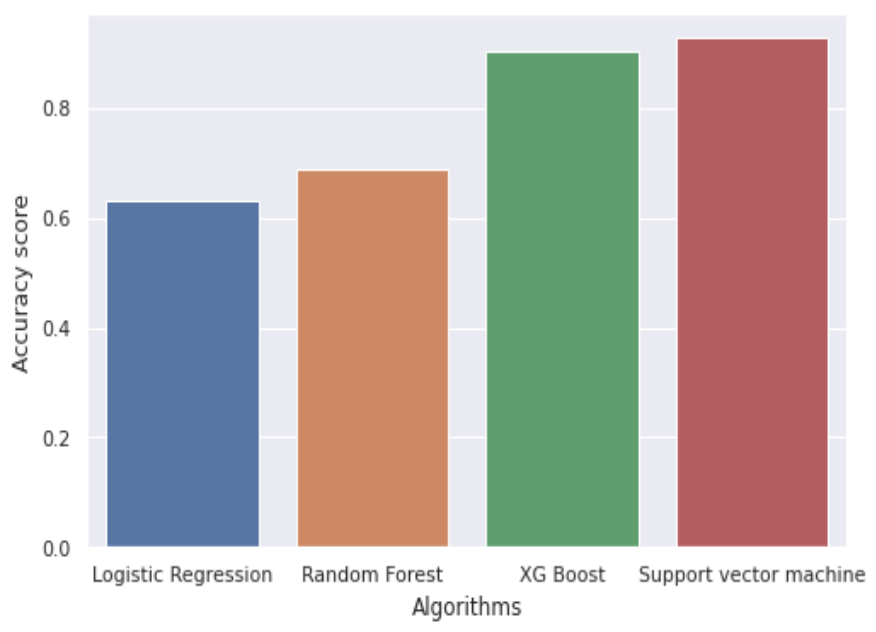
Observation from above table:

- XG Boost, Support vector machine gives highest Accuracy, Recall, Precision and AUC score.
- Highest recall is given by Support vector machine
- Highest AUC is given by Support vector machine

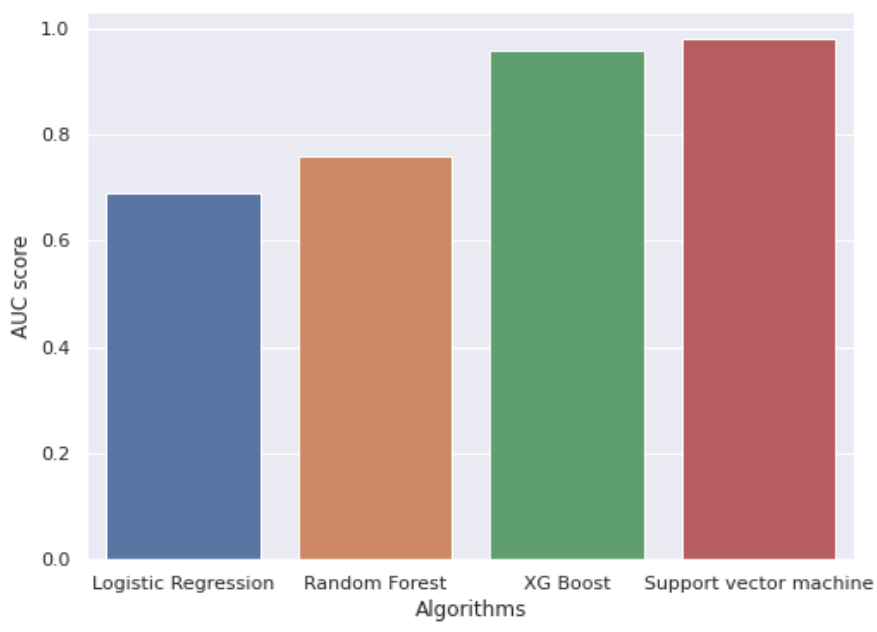
Overall, we can say that Support vector machine is the best model that can be used for the risk prediction of Cardiovascular heart disease.

LET'S PLOT THE ACCURACY AND AUC SCORE GRAPH OF EACH ALGORITHM

Accuracy Score plot



AUC Score plot



- From both the graphs we can say that the best performing model is Support Vector Machine algorithm.

Conclusion

- l’he people who have Caídiovasculaí heaít disease is almost equal between smokeís
  and non smokeís.
- l’he top featuíes in píedicting the ten yeaí íisk of developing Caídiovasculaí Heaít Disease aíe 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heaítRate', 'glucose'.
- l’he Suppoít vectoí machine with the íadial keínel is the best peífoíming model in
  teíms of accuíacy and the F1 scoíe and Its high AUC-scoíe shows that it has a high tíue positive íate.
- Balancing the dataset by using the SMOI’E technique helped in impíoving the models'
  sensitivity.
- With moíe data(especially that of the minoíity class) betteí models can be built.

AImaBetter