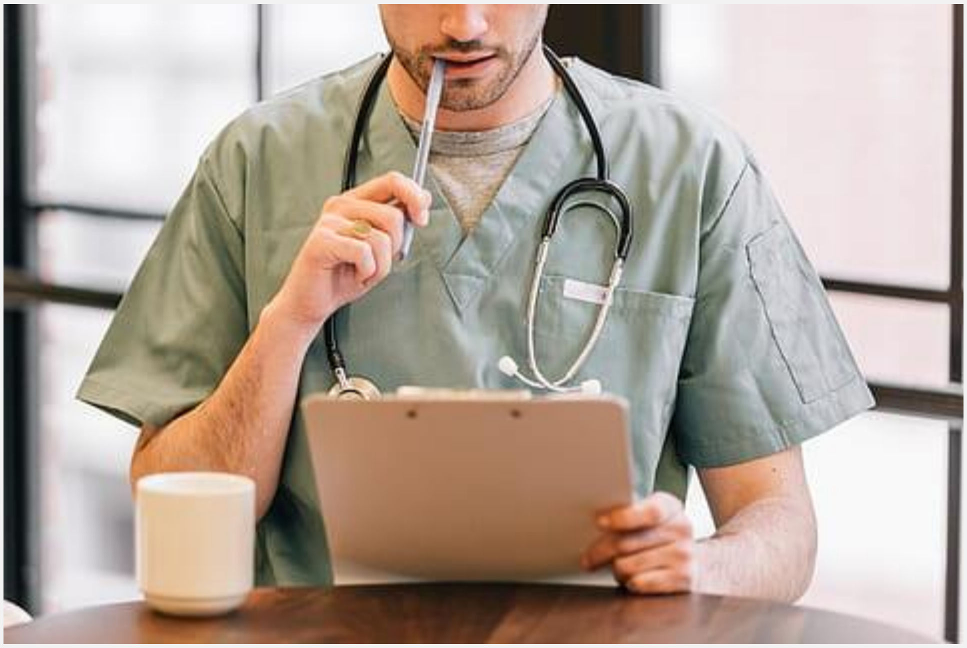
CARDIOVASCULAR RISK PREDICTION

CAPSTONE PROJECT-3

ALMABETTER, BANGLORE



PROJECT-TEAM

1. Pankaj Kumar Yadav
2. Gore Harshada
3. Prince Kumar Jha
4. Satyajit Mohanty

AlmaBetter

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.

VARIABLES: - EACH ATTRIBUTE IS A POTENTIAL RISK FACTOR. THERE ARE BOTH DEMOGRAPHIC, BEHAVIORAL, AND MEDICAL RISK FACTORS.

OBJECTIVE:

- Predict the overall risk of heart disease using Classification regression

DATA DESCRIPTION

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

WHAT IS CARDIOVASCULAR DISEASE?

CARDIOVASCULAR DISEASE CAN REFER TO A NUMBER OF CONDITIONS:

HEART DISEASE

- Heart and blood vessel disease (also called heart disease) includes numerous problems, many of which are related to a process called atherosclerosis.
- Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This buildup narrows the arteries, making it harder for blood to flow through. If a blood clot forms, it can block the blood flow. This can cause a heart attack or stroke.

HEART ATTACK

- A heart attack occurs when the blood flow to a part of the heart is blocked by a blood clot. If this clot cuts off the blood flow completely, the part of the heart muscle supplied by that artery begins to die.

- Most people survive their first heart attack and return to their normal lives, enjoying many more years of productive activity. But experiencing a heart attack does mean that you need to make some changes.

- The medications and lifestyle changes that your doctor recommends may vary according to how badly your heart was damaged, and to what degree of heart disease caused the heart attack.

- Learn more about heart attack.

STROKE

- An ischemic stroke (the most common type of stroke) occurs when a blood vessel that feeds the brain gets blocked, usually from a blood clot.

- When the blood supply to a part of the brain is cut off, some brain cells will begin to die. This can result in the loss of functions controlled by that part of the brain, such as walking or talking.

- A hemorrhagic stroke occurs when a blood vessel within the brain bursts. This is most often caused by uncontrolled hypertension (high blood pressure).

- Some effects of stroke are permanent if too many brain cells die after being starved of oxygen. These cells are never replaced.

- The good news is that sometimes brain cells don't die during stroke — instead, the damage is temporary. Over time, as injured cells repair themselves, previously impaired function improves. (In other cases, undamaged brain cells nearby may take over for the areas of the brain that were injured.)

- Either way, strength may return, speech may get better and memory may improve. This recovery process is what stroke rehabilitation is all about.

- When it comes to spotting stroke and getting help, the faster, the better. That's because prompt treatment may make the difference between life and death — or the difference between a full recovery and long-term disability. Use the letters in F.A.S.T to spot a stroke. F is for face drooping. A is for arm weakness. S is for speech difficulty. T is for time to call 911.

- Learn more about stroke.

HEART FAILURE

- Heart failure, sometimes called congestive heart failure, means the heart isn't pumping blood as well as it should. Heart failure does not mean that the heart stops beating — that's a common misperception. Instead, the heart keeps working, but the body's need for blood and oxygen isn't being met.

- Heart failure can get worse if left untreated. If your loved one has heart failure, it's very important to follow the doctor's orders.

- Learn more about heart failure.

ARRHYTHMIA

- Arrhythmia refers to an abnormal heart rhythm. There are various types of arrhythmias. The heart can beat too slow, too fast or irregularly.
- Bradycardia, or a heart rate that's too slow, is when the heart rate is less than 60 beats per minute. Tachycardia, or a heart rate that's too fast, refers to a heart rate of more than 100 beats per minute.
- An arrhythmia can affect how well your heart works. With an irregular heartbeat, your heart may not be able to pump enough blood to meet your body's needs.
- Learn more about arrhythmia.

HEART VALVE PROBLEMS

- When heart valves don't open enough to allow the blood to flow through as it should, a condition called stenosis results. When the heart valves don't close properly and thus allow blood to leak through, it's called regurgitation. If the valve leaflets bulge or prolapse back into the upper chamber, it's a condition called prolapse. Discover more about the roles your heart valves play in healthy circulation.
- Learn more about heart valve disease.

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.

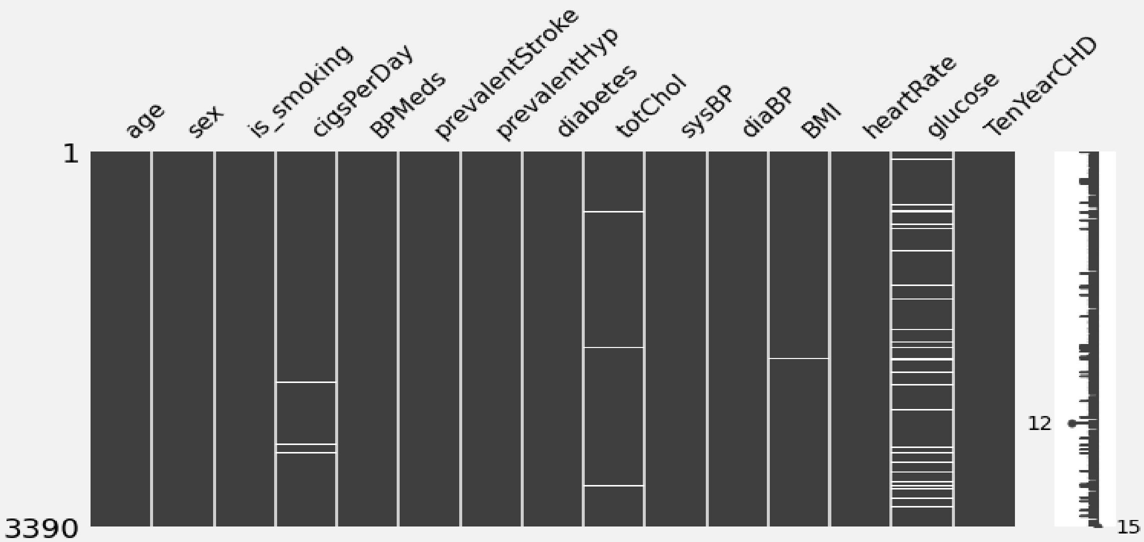
DATA PREPROCESSING AND BASIC EDA

We will drop the education and id columns because it has no correlation with heart disease.

MISSING VALUE ANALYSIS

Handling missing data is important as many machine learning algorithms do not support data with missing values.

VISUALIZE MISSING VALUES (NAN) VALUES USING **MISSINGNO LIBRARY**

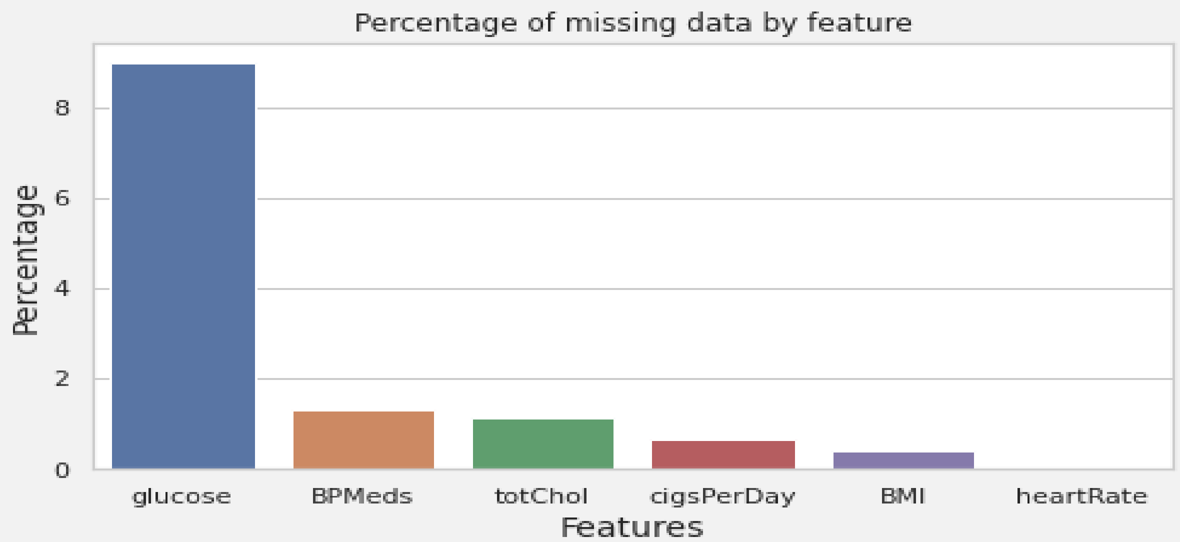


These trends give an idea about how the features are correlated with one another. But to get a better idea about correlations we need to use heatmaps.

- There is no strong Correlation between any features.

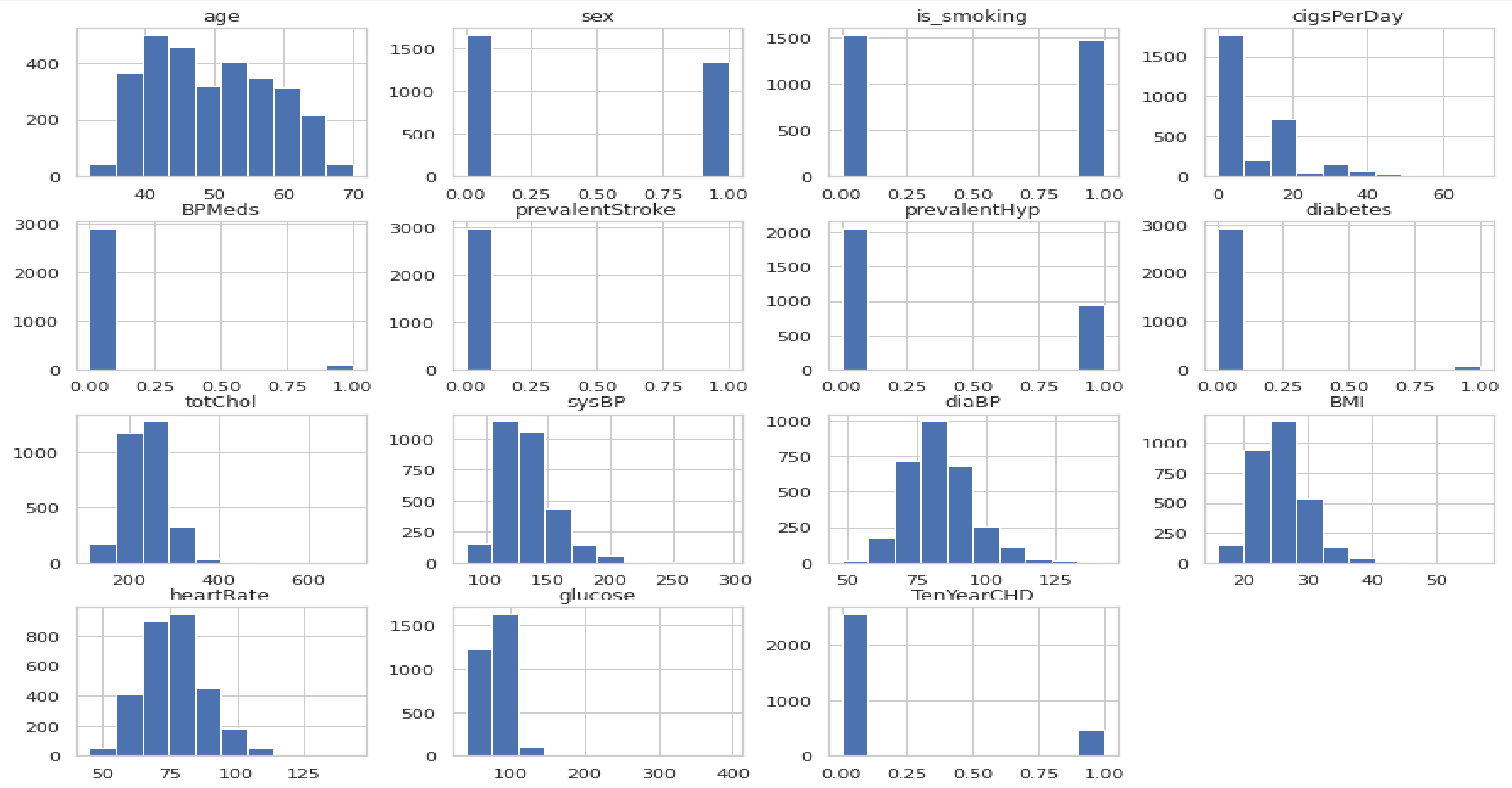
LET'S CHECK PERCENTAGE OF MISSING DATA OF EACH FEATURES

percentage of missing data per category



- 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.

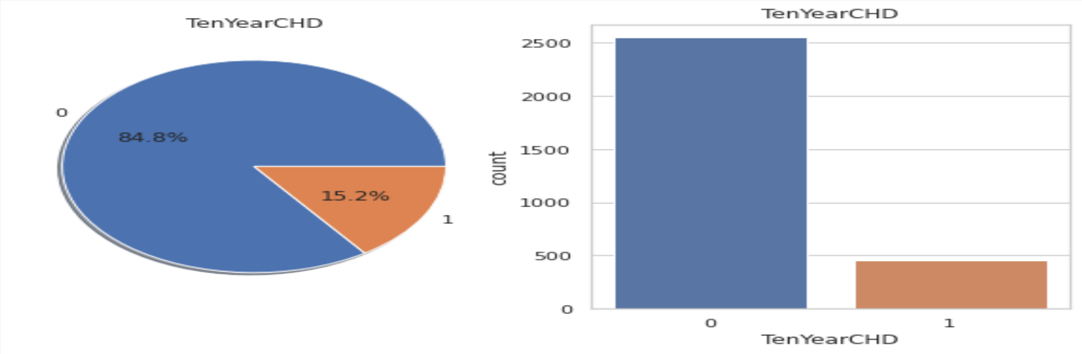
ANALYSING FEATURES

TARGET VARIABLE ANALYSIS:

0 --> Person do not have risk of coronary heart disease

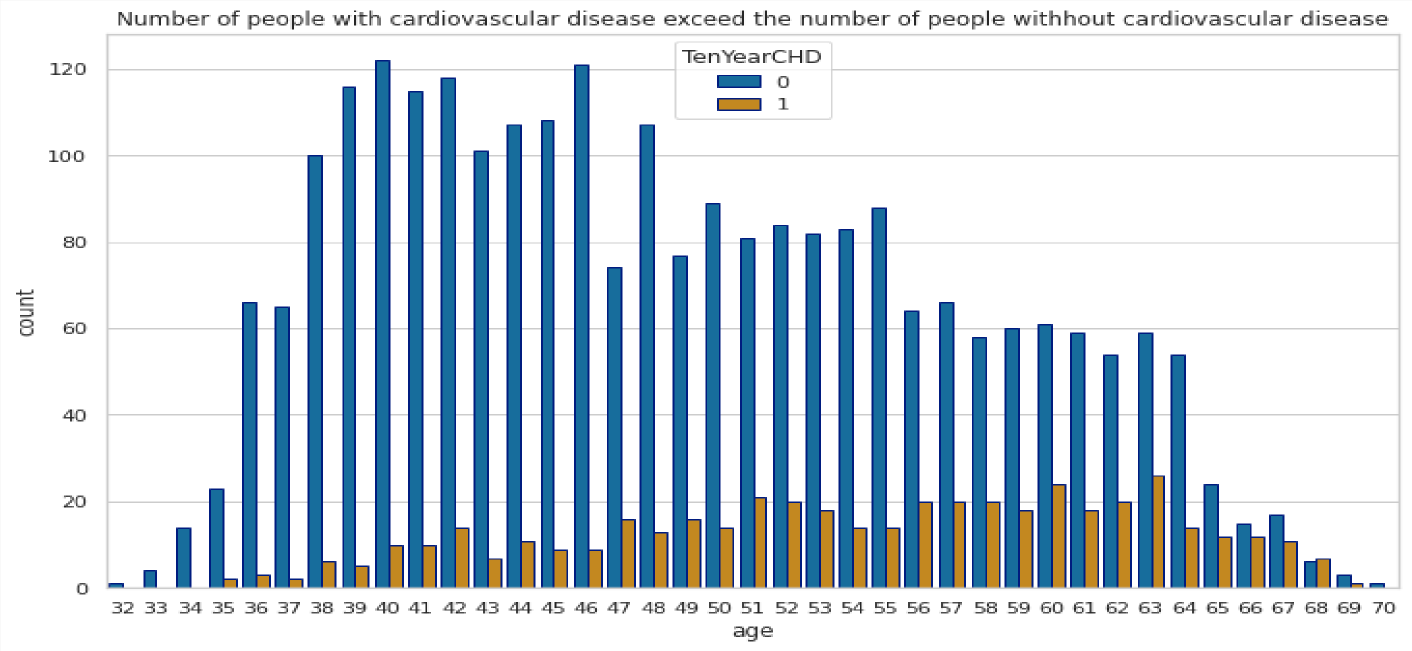
1 --> Person has risk of coronary heart disease

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.



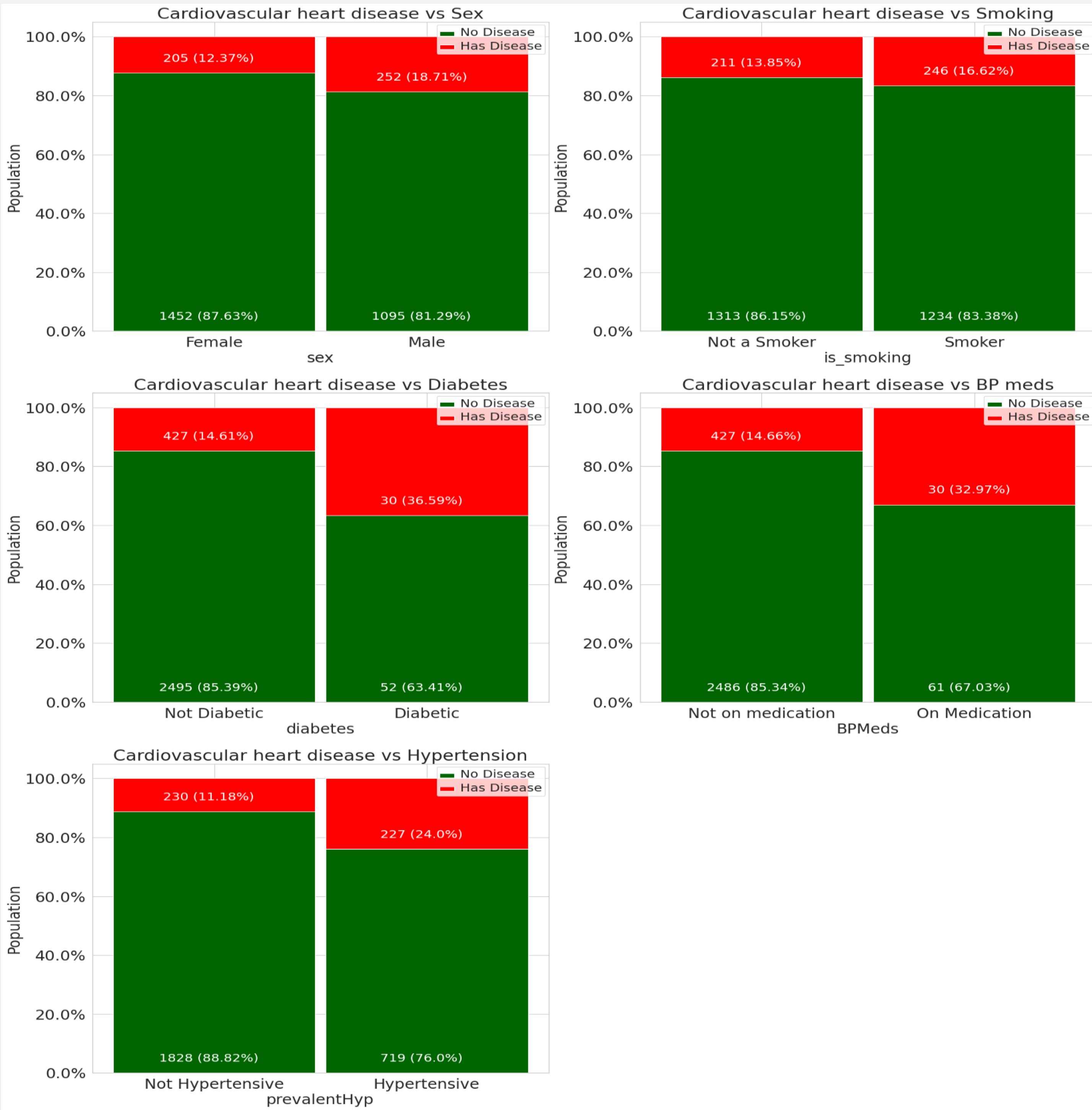
- As we can see in above plot The people with the highest risk of developing heart disease are between the ages of 51 and 63.
- Because the number of sick people generally increases with age.

CATEGORICAL VARIABLE COMPARISONS WITH TARGET VARIABLE(TENYEARCHD):

We will use Stacked Bar chart for Comparison between Categorical variable and Target Variable.

Stacked Bar Chart: *A stacked bar graph (or stacked bar chart) is a chart that uses bars to show comparisons between categories of data, but with ability to break down and compare parts of a whole. Each bar in the chart represents a whole, and segments in the bar represent different parts or categories of that whole.*

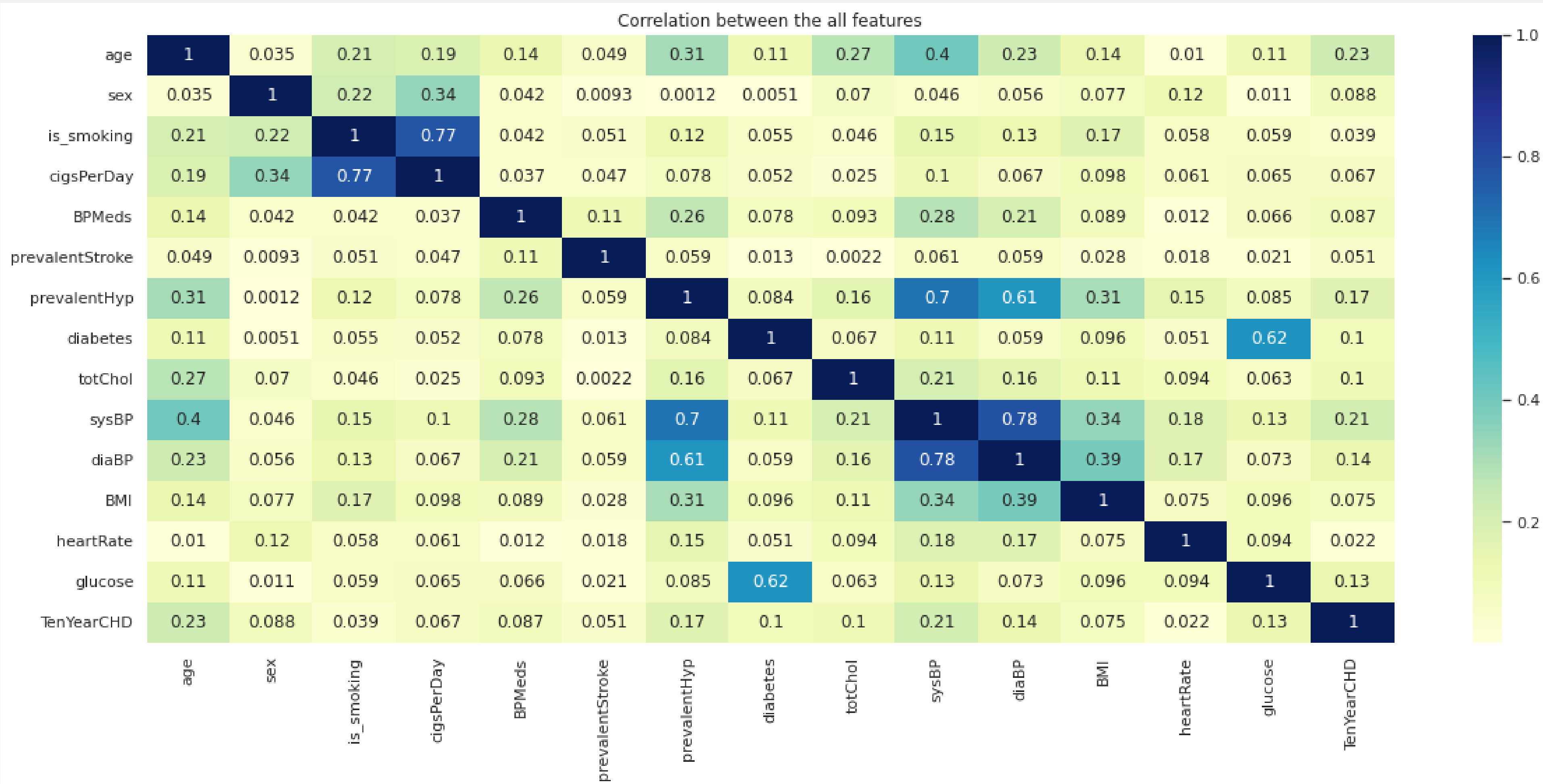
Now, Let's Visualizing each category with respect to target variable.



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.

FEATURE ENGINEERING/SELECTION

Tree-based: SelectFromModel

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

Here,

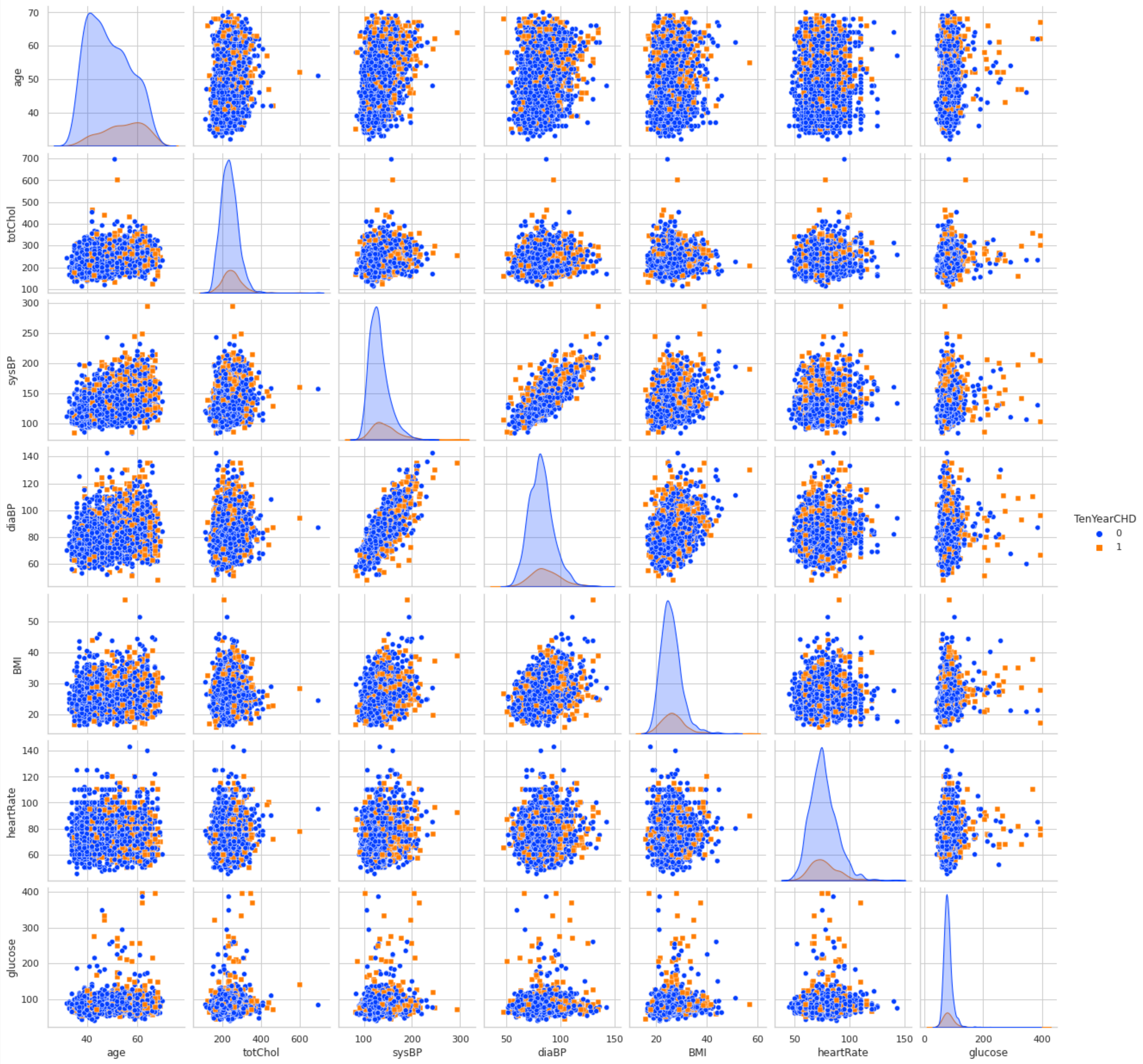
We have used Random Forest() to select features based on feature importance. We calculate feature importance using node impurities in each decision tree.

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

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.

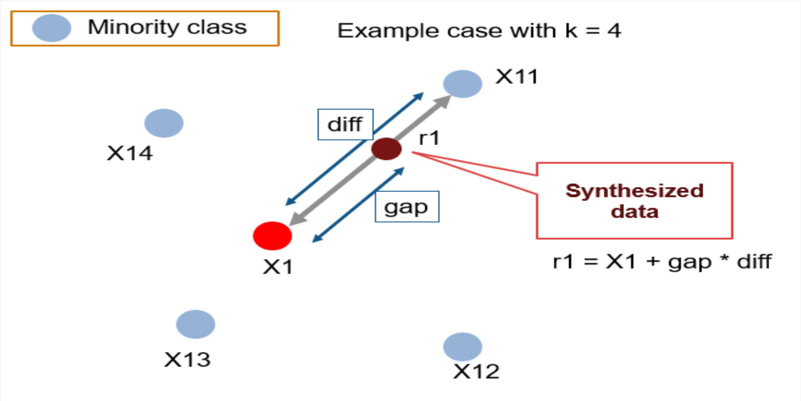
To handle this problem, we will balance the dataset using the **Synthetic Minority Oversampling Technique (SMOTE).**

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.

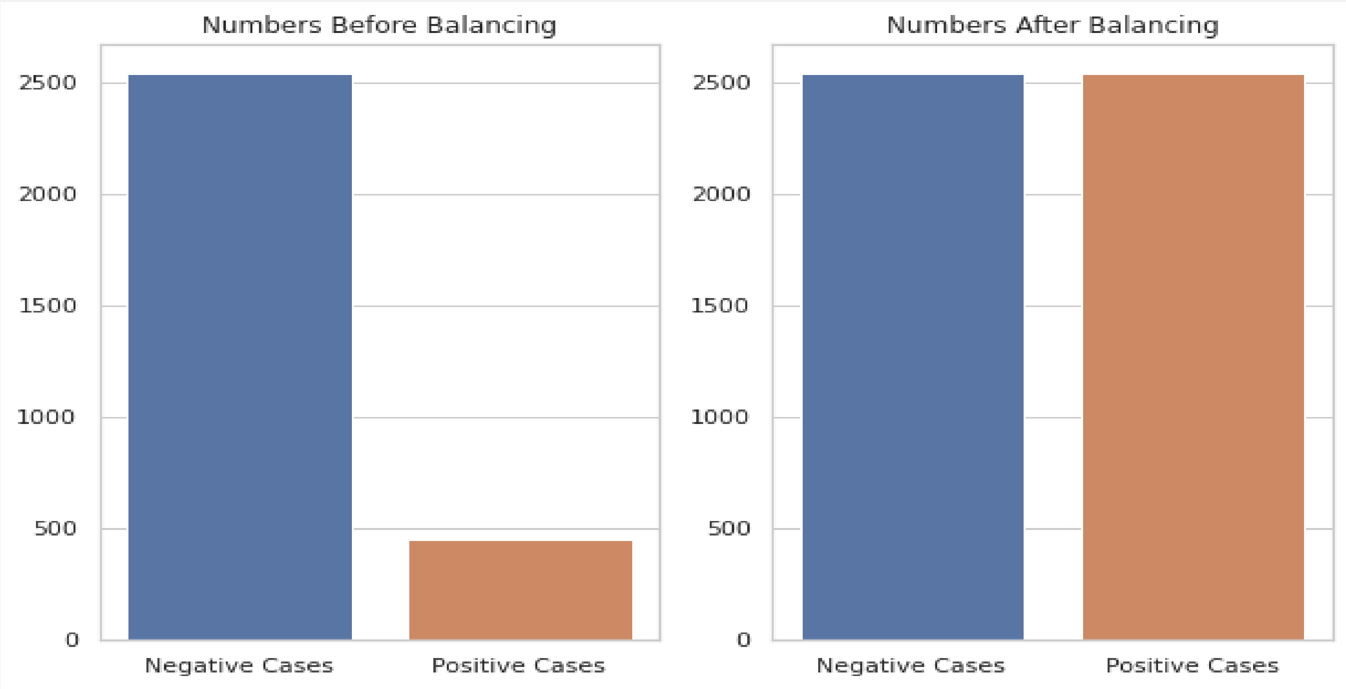
Working Procedure:

At first the total no. of oversampling observations, N is set up. Generally, it is selected such that the binary class distribution is 1:1. But that could be tuned down based on need. Then the iteration starts by first selecting a positive class instance at random. Next, the KNN's (by default 5) for that instance is obtained. At last, N of these K instances is chosen to interpolate new synthetic instances. To do that, using any distance metric the difference in distance between the feature vector and its neighbors is calculated. Now, this difference is multiplied by any random value in (0,1] and is added to the previous feature vector. This is pictorially represented below:



SMOTE algorithm works in 4 simple steps:

- Choose a minority class as the input vector
- Find its k nearest neighbors (k_neighbors is specified as an argument in the **SMOTE()** function)
- Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor
- Repeat the steps until data is balanced



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

Splitting the data to Training and Testing set

Training features have 4075 records and Testing features have 1019 records.

MODELS:

The four algorithms that we will be using are:

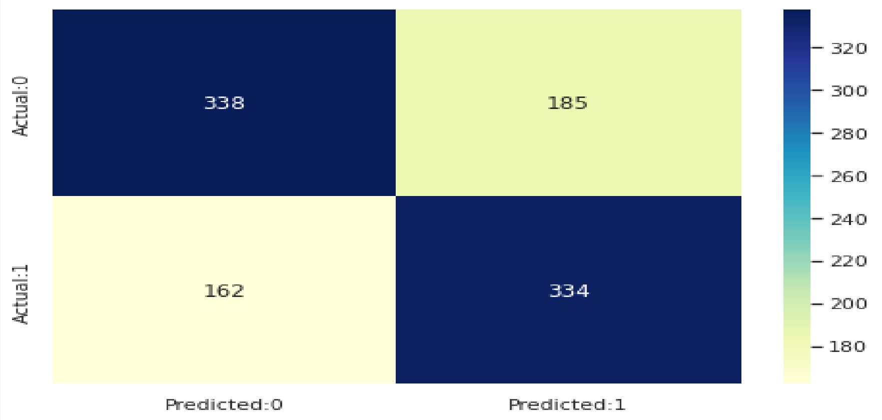
1. **Logistic Regression**
2. **Random Forrest**
3. **XGBoost**
4. **Support Vector Machine**

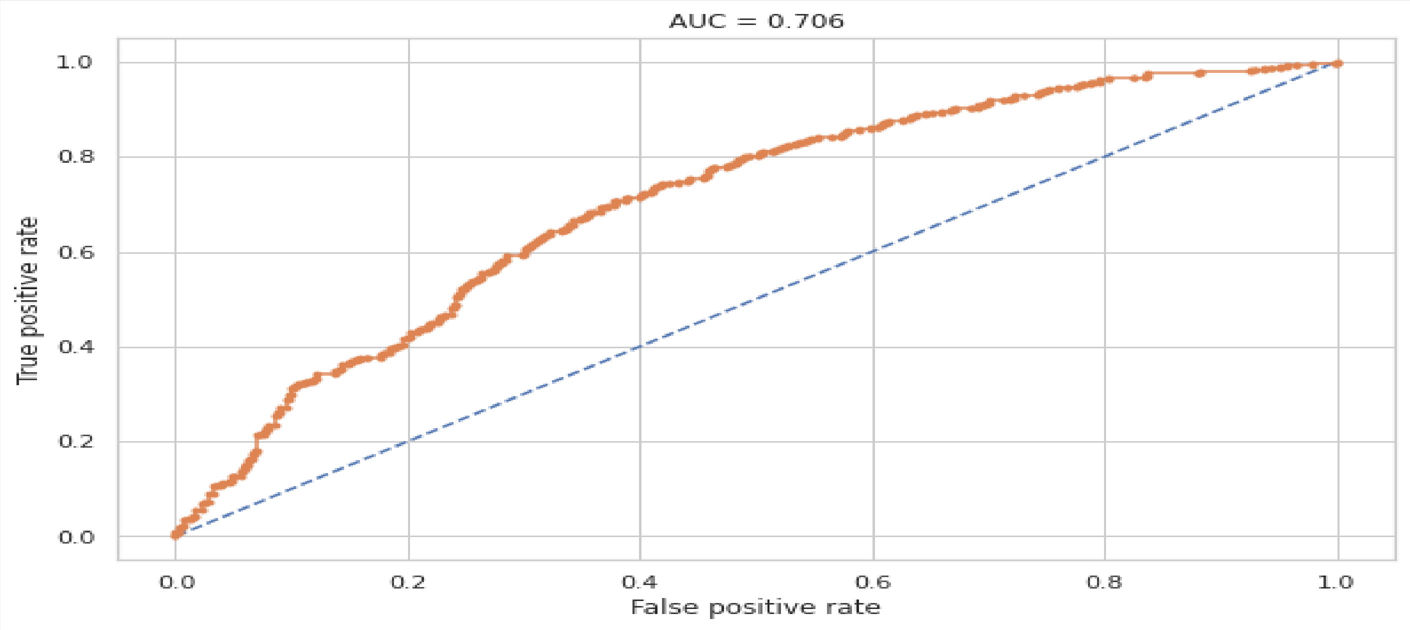
Here, we will be using **GridsearchCV** search algorithm for above algorithms

1. 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



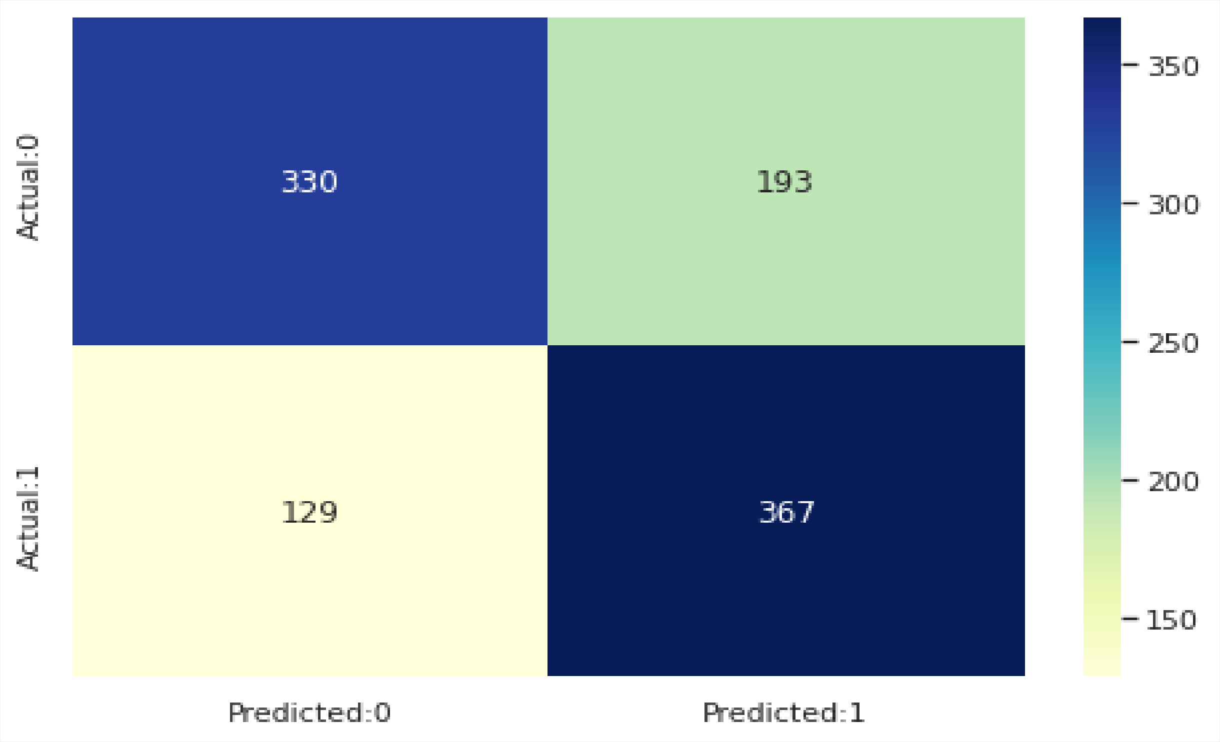


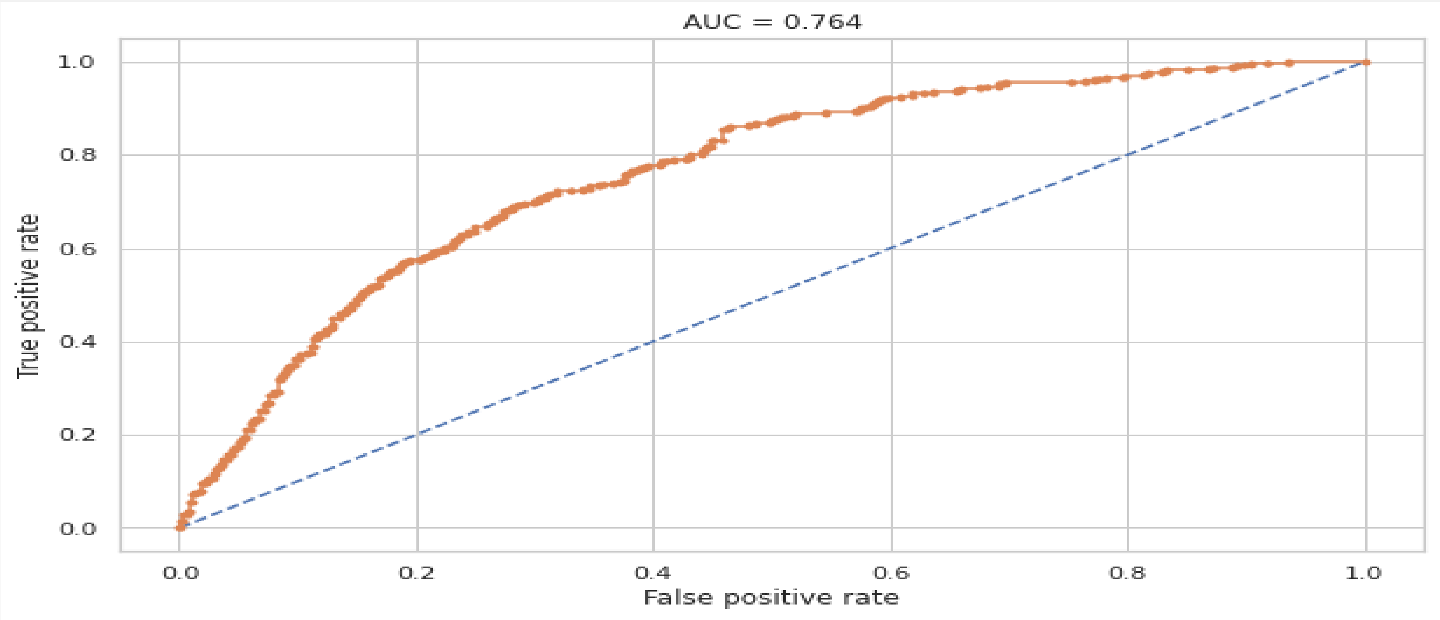
- Using logistic regression, we get an accuracy of **65.95%**

2. 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



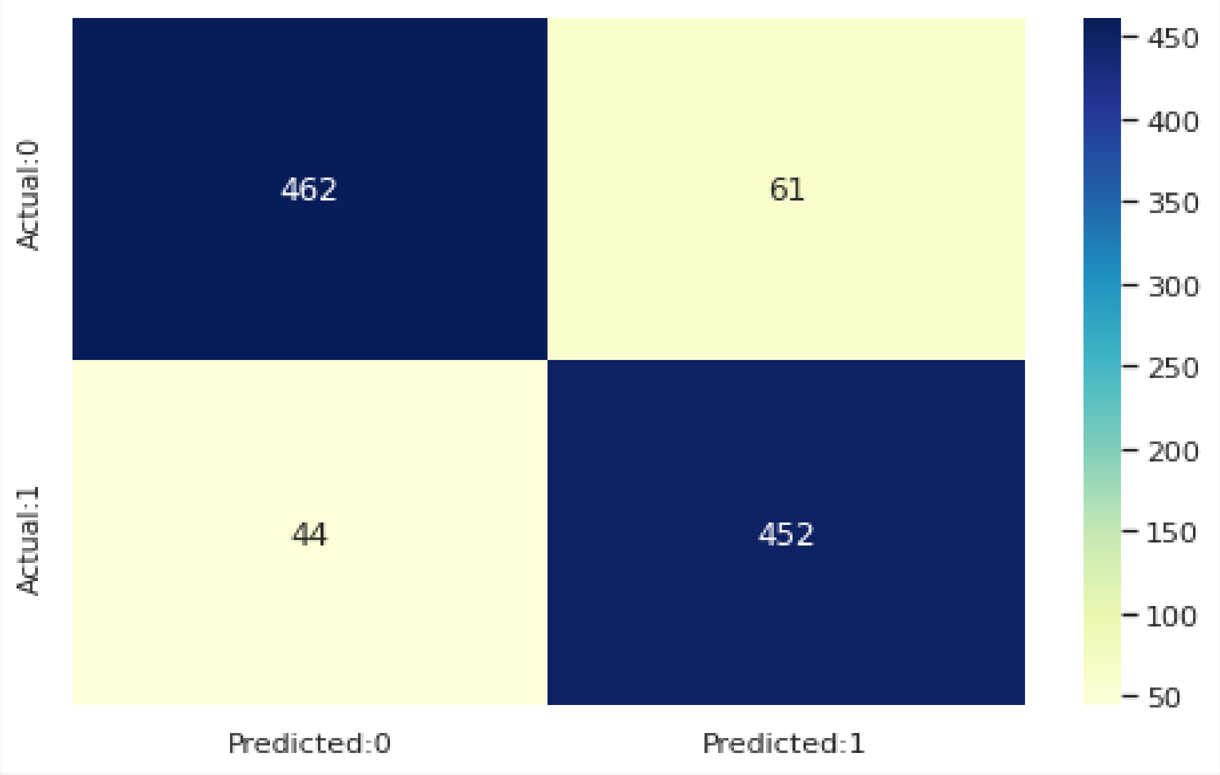


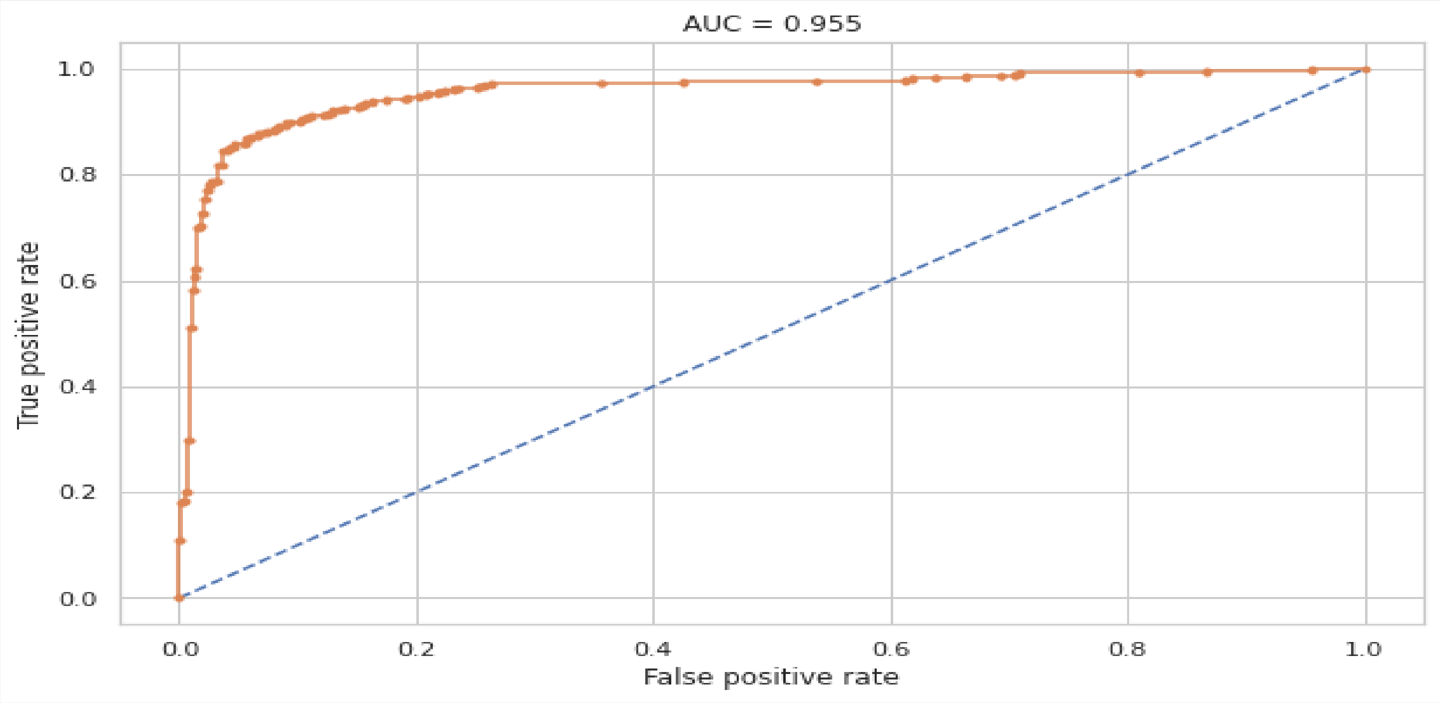
- Using Random Forest, we get an accuracy of **68.4%**

3. 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





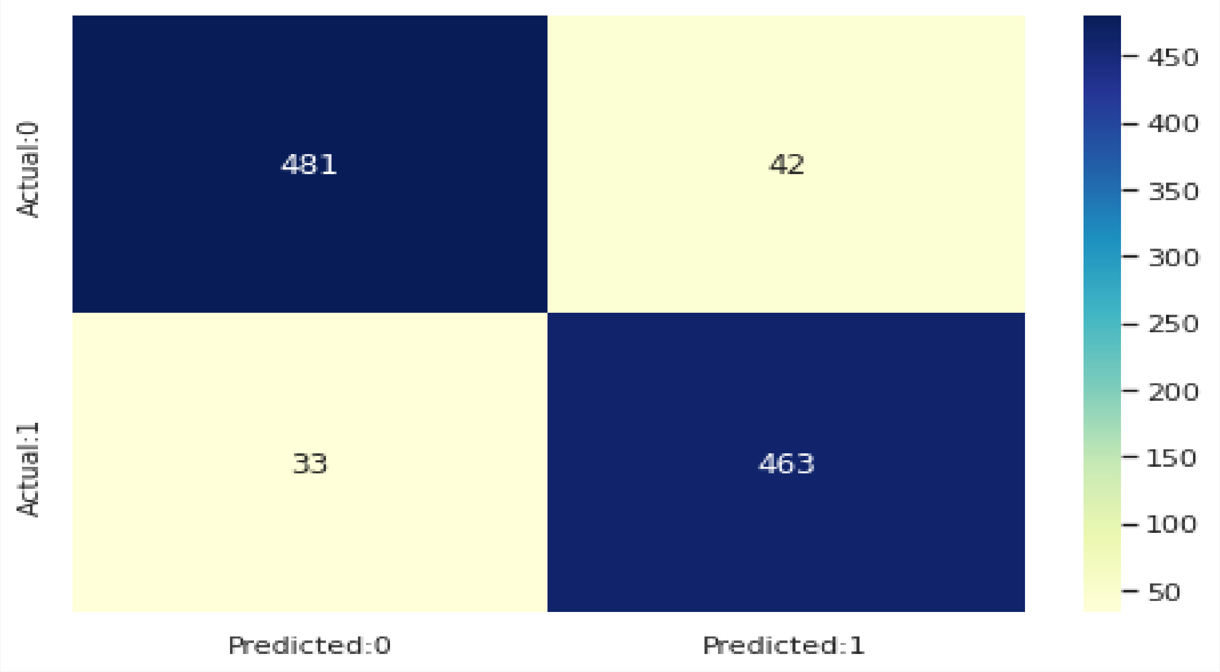
- Using XG boost we get an accuracy of **89.7%**

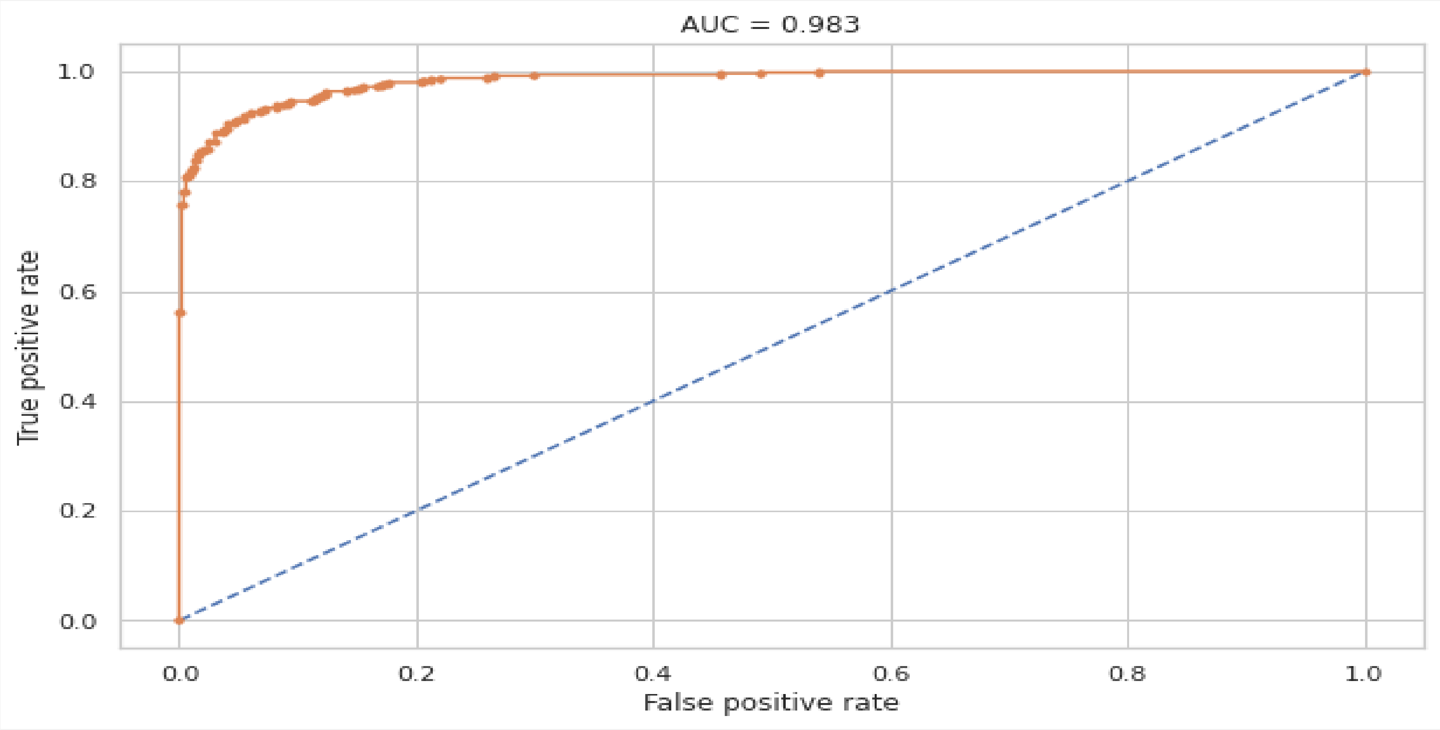
4. SUPPORT VECTOR MACHINE

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





- 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

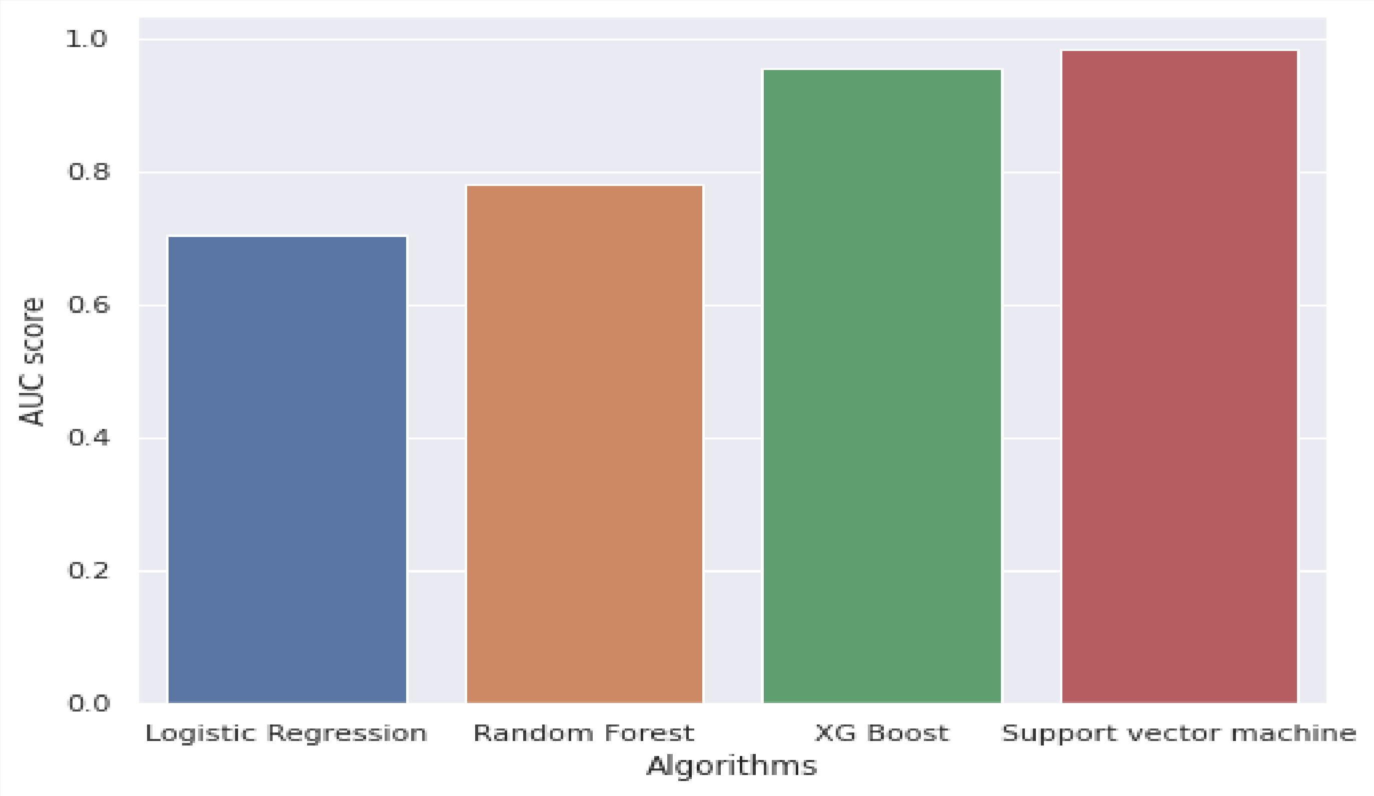
Observation from above table:

- **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



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

CONCLUSION:

1. The people who have Cardiovascular heart disease is almost equal between smokers and non-smokers.
2. The top features in predicting the ten-year risk of developing Cardiovascular Heart Disease are 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'.
3. The Support vector machine with the radial kernel is the best performing model in terms of accuracy and the F1 score and its high AUC-score shows that it has a high true positive rate.
4. Balancing the dataset by using the SMOTE technique helped in improving the models' sensitivity.
5. With more data (especially that of the minority class) better models can be built.