Capstone Project

Cardiovascular Risk Prediction

Shubhankar Rathod
Abhishek Tiwari

CONTENTS

Problem Statement:

- **Coronary heart disease is caused due to accumulation of plaque in major heart blood vessels leading to blockage of oxygen- rich blood to heart.**
- **It is the most common type of heart disease, killing around 300 K people in US alone every year.**
- **The goal of my project is to come up with a ML model that correctly predicts 10-year risk of a patient having Coronary Heart Disease(CHD).**

Data Pipeline

1. **Data Preprocessing:** At this stage, we check for missing values, duplicate values and treat them accordingly. Also review the descriptive statistics of the numerical features. Furthermore, we check for the features present in our dataset and transform the columns if necessary.

2. **Exploratory Data Analysis:** In EDA we conducted Univariate and Multivariate Analysis of independent and dependent feature to understand their spread, pattern and relationship with each other. It helps us to better understand our data and make inferences out of them.

3. **Feature Engineering:** At this stage we created new columns from the existing features which helps simplifying and speeding up data transformations while also enhancing model accuracy.

4. **Model Building:** In model building we have applied 5 classification algorithms on our training dataset and after training the models, we received the prediction for the risk of 10 Year CHD.

5. **Model Evaluation:** In this stage we have used different evaluation metrics used for classification problem to test the performance of our models in validation dataset. The metrics which we have used are Accuracy score, AUC_ROC score, Precision, Recall and F1 Score.

Data Summary

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 factor variables.

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)

Behavioral:

- **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.)

Data Summary

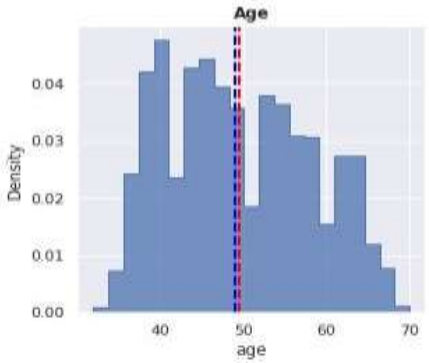
Medical(history)

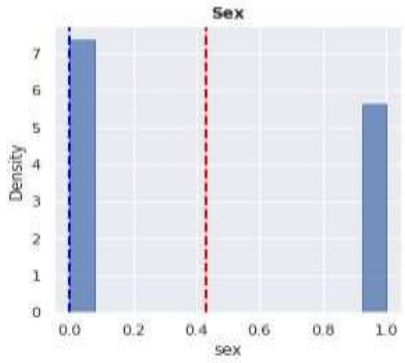
- **BP Meds:** Whether or not the patient was on blood pressure medication (Nominal)
- **Prevalent Stroke:** Whether or not the patient had previously had a stroke (Nominal)
- **Prevalent Hyp:** Whether or not the patient was hypertensive (Nominal)
- **Diabetes:** Whether or not the patient had diabetes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

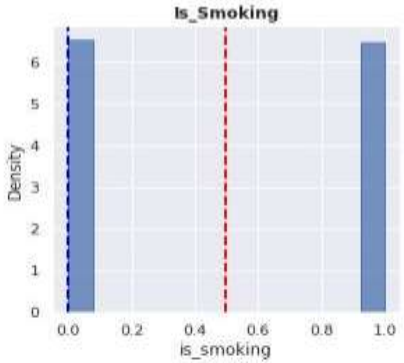
Medical(current)

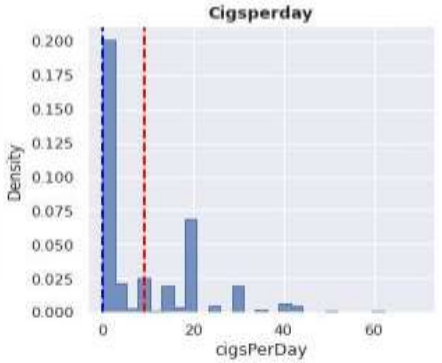
- **Tot Chol:** total cholesterol level (Continuous)
- **Sys BP:** systolic blood pressure (Continuous)
- **Dia BP:** diastolic blood pressure (Continuous)
- **BMI:** Body Mass Index (Continuous)
- **Heart Rate:** heart rate (Continuous -In medical research, variables such as heart rate though in fact 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**

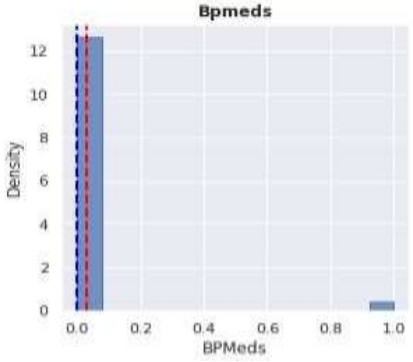
EDA (Univariate Analysis)

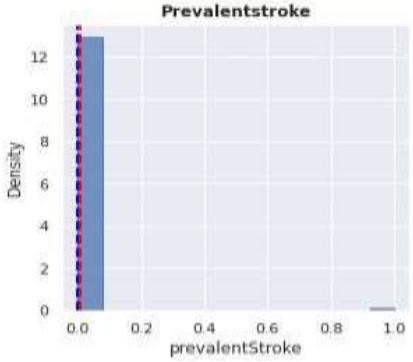


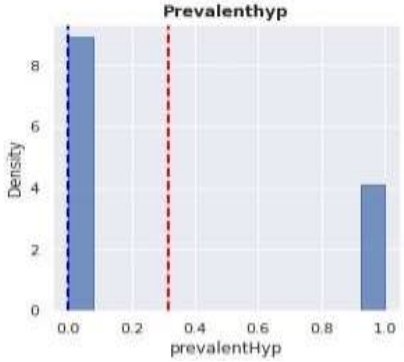


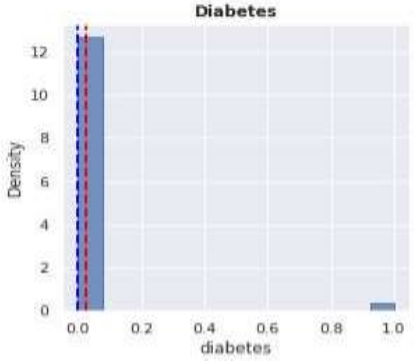




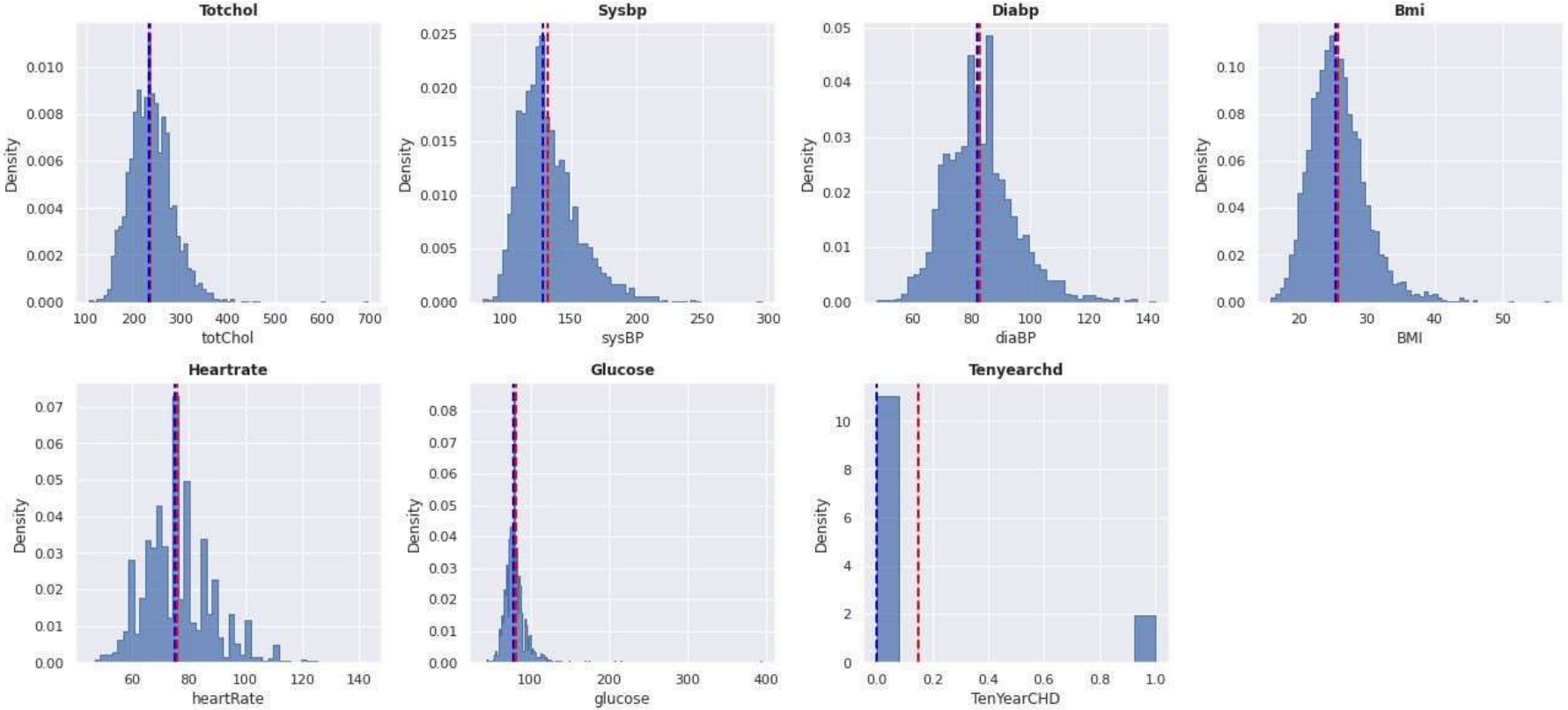








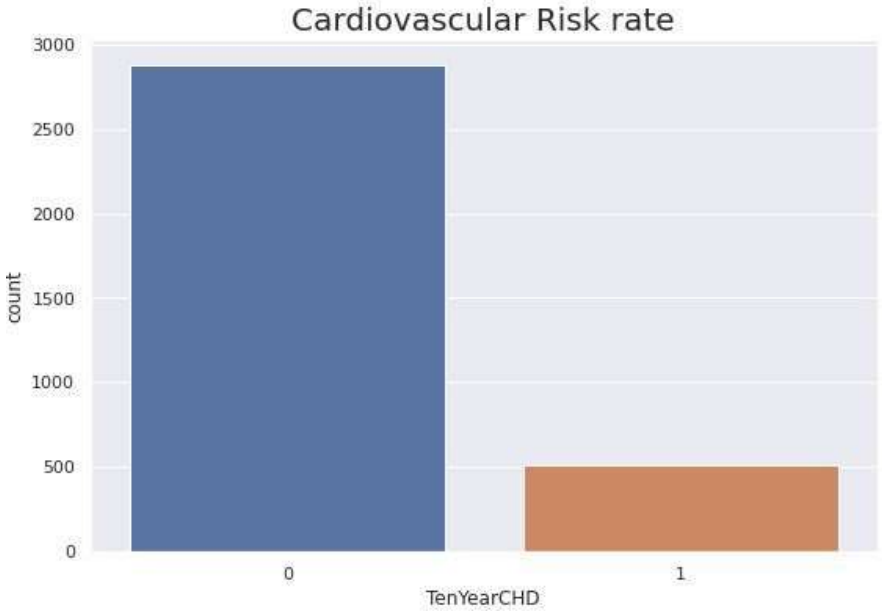
Univariate Analysis

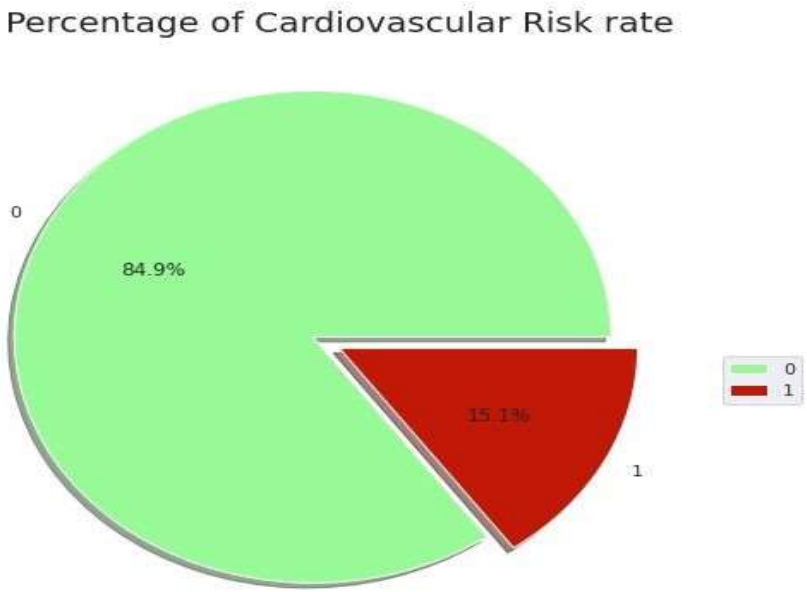


Observations:

- ➔ Most of the people in our dataset are around 40 - 60 years old.
- ➔ Most people smokes less than 10 cigarettes a day.
- ➔ From above distribution plot we can say that the data on the prevalent stroke, diabetes, and blood pressure meds are poorly balanced. Also continuous features like totChol , sysBP, BMI etc are right skewed.
- ➔ Rest all the features appears to be normally distributed.

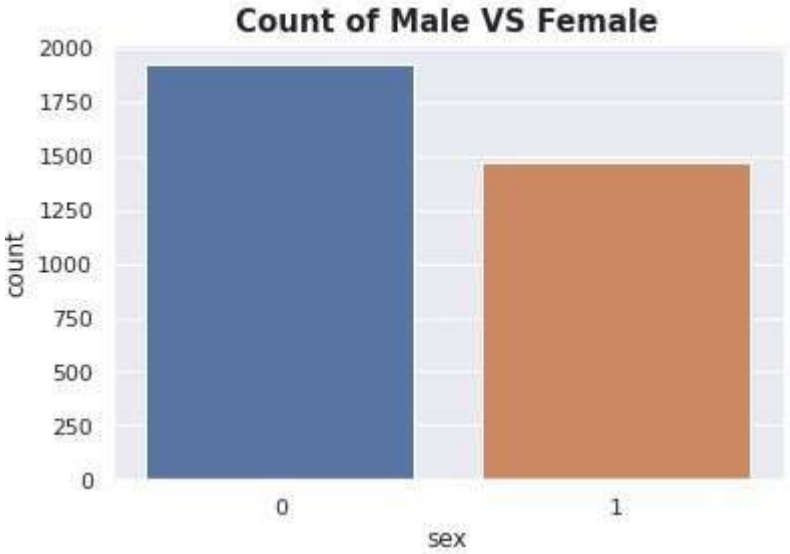
Univariate Analysis

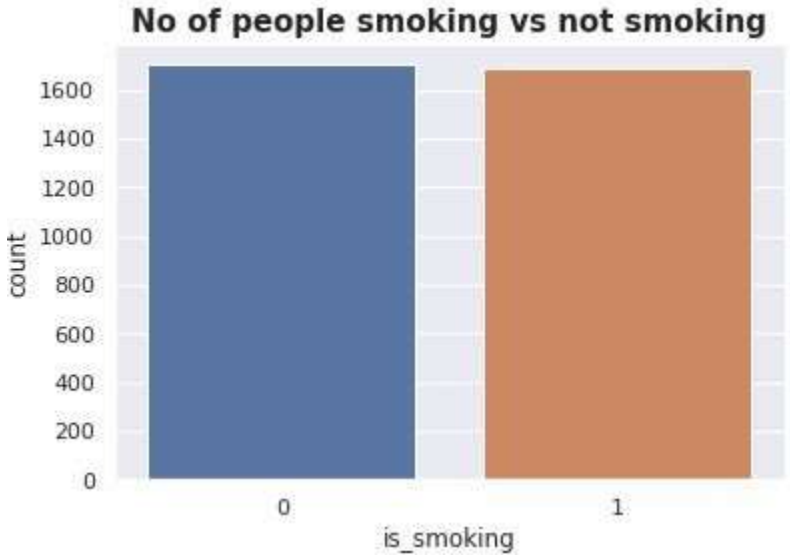




- From the above charts it's clearly visible that there is a class imbalance problem with our dataset!
- Number of people without the disease greatly exceeds the number of people with the disease.
- An imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes. So we would need to deal with this problem for better predictions.

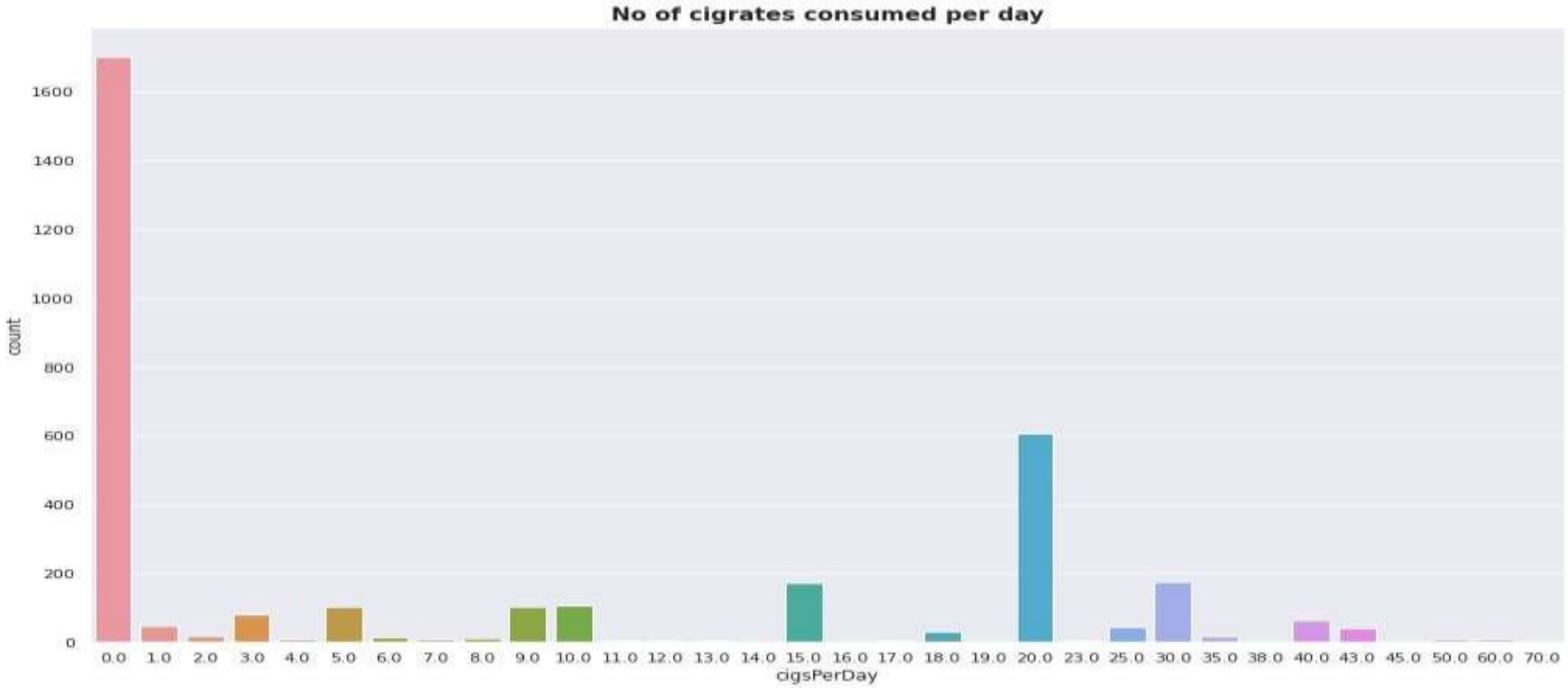
Univariate Analysis





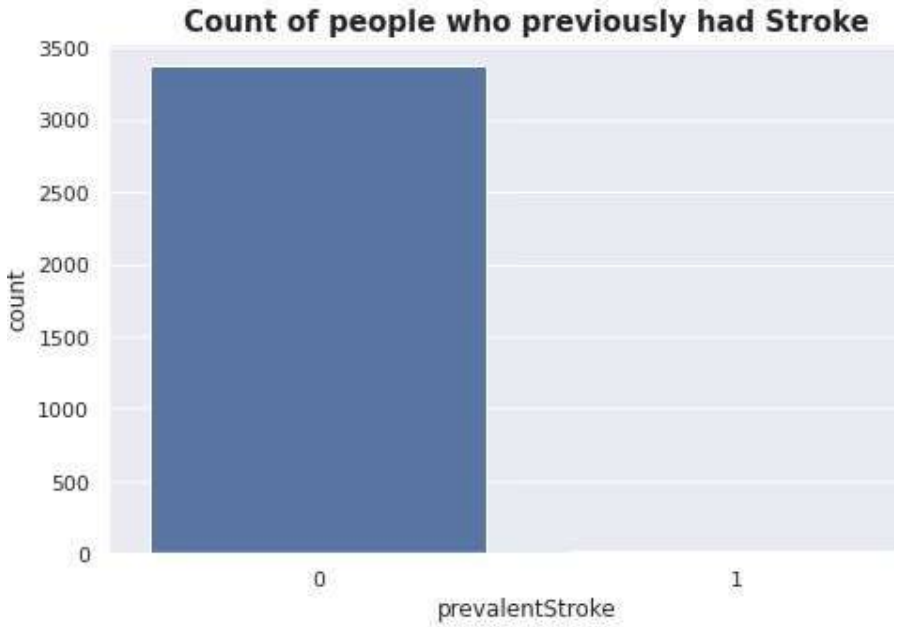
- From the above Barplots it can be seen that the data of Female population is more than that of Men.
- Also it can be seen that there are almost equal number of smokers and non smokers in the dataset.

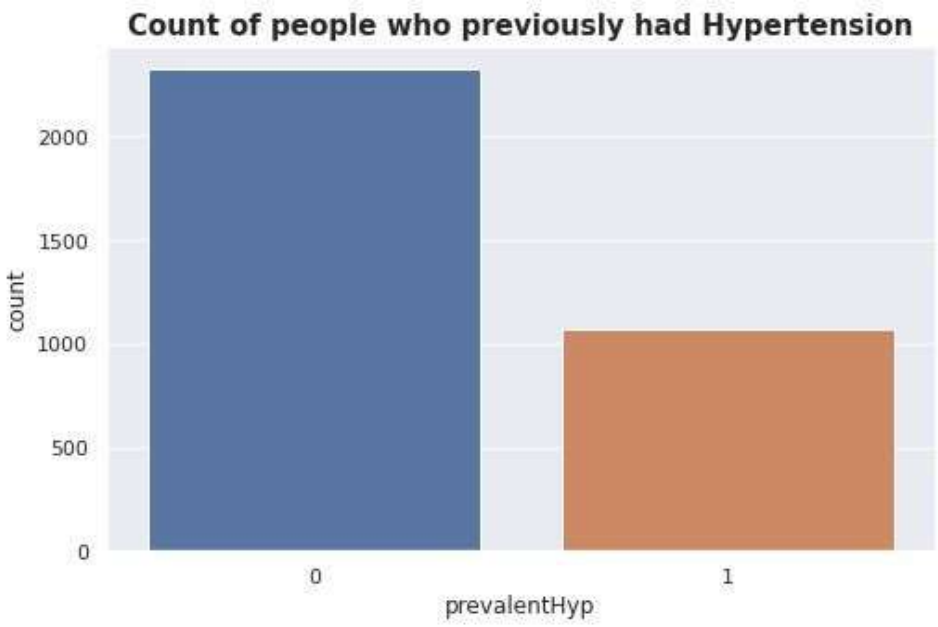
Univariate Analysis



- From the above countplot it can be said that most of the people don't smoke cigarettes. Among people who smoke, most of them smoke 20 cigarettes per day, followed by 30 cigarettes , 15 cigarettes etc.

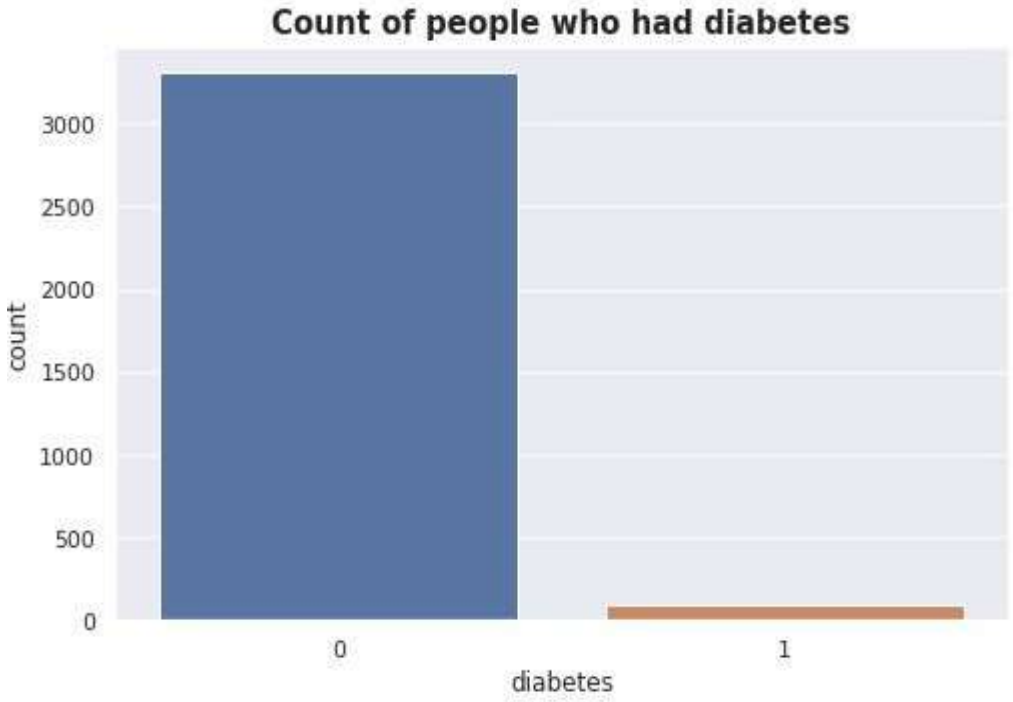
Univariate Analysis(cont.)





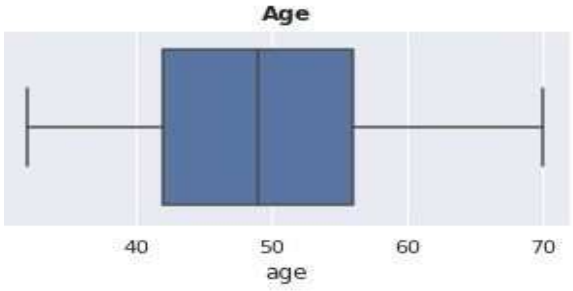
- It can be said that number of people having Stroke in the past is very less, which is 22 vs the number of people who didn't have a stroke.
- Also out of 3390 patients from the records 1069 people had Hypertension before.

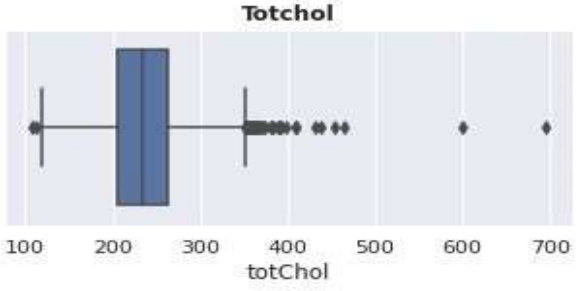
Univariate Analysis(cont.)

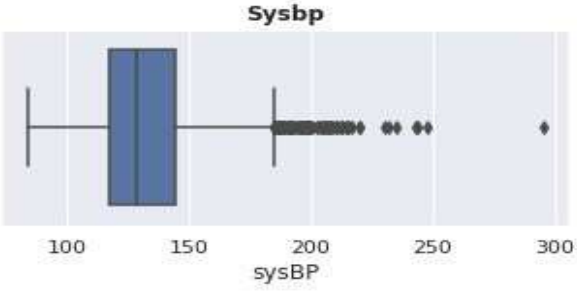


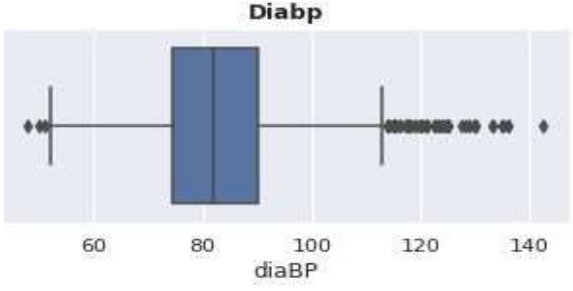
- Out of 3390 patients from our records 87 people had diabetes.
- **Interesting fact about diabetes is that if a person have diabetes, then they are twice as likely to have heart disease or a stroke than someone who doesn't have diabetes—and at a younger age. The longer anyone have diabetes, the more likely they are to have heart disease.**

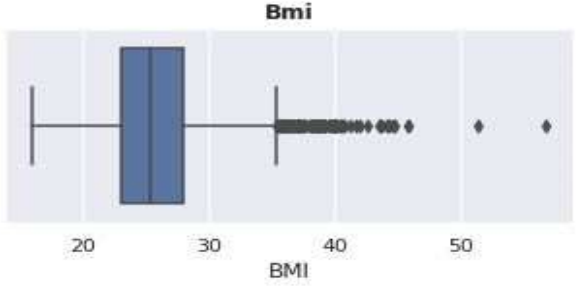
Univariate Analysis(cont.)

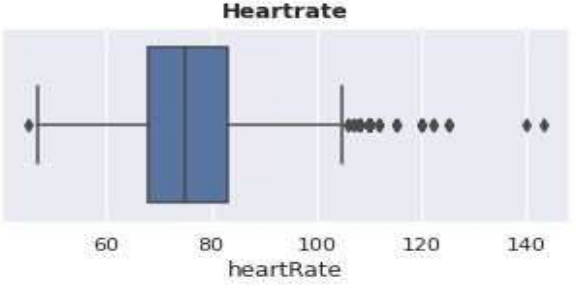


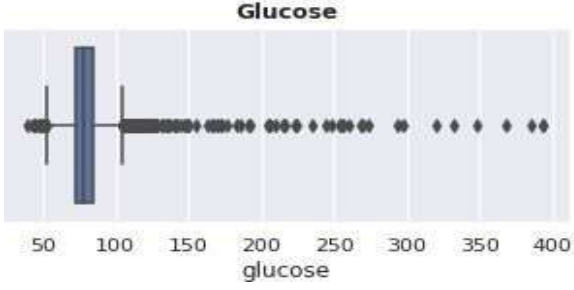






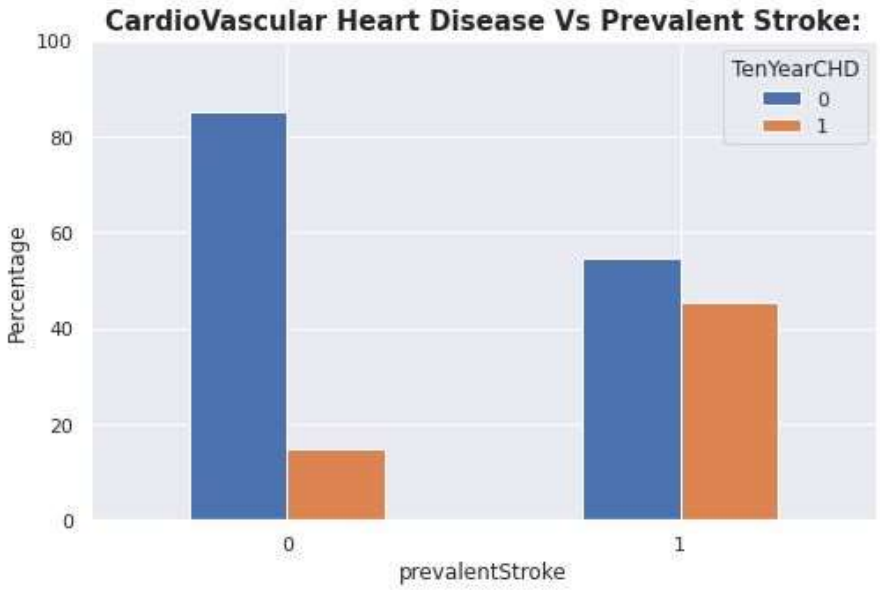


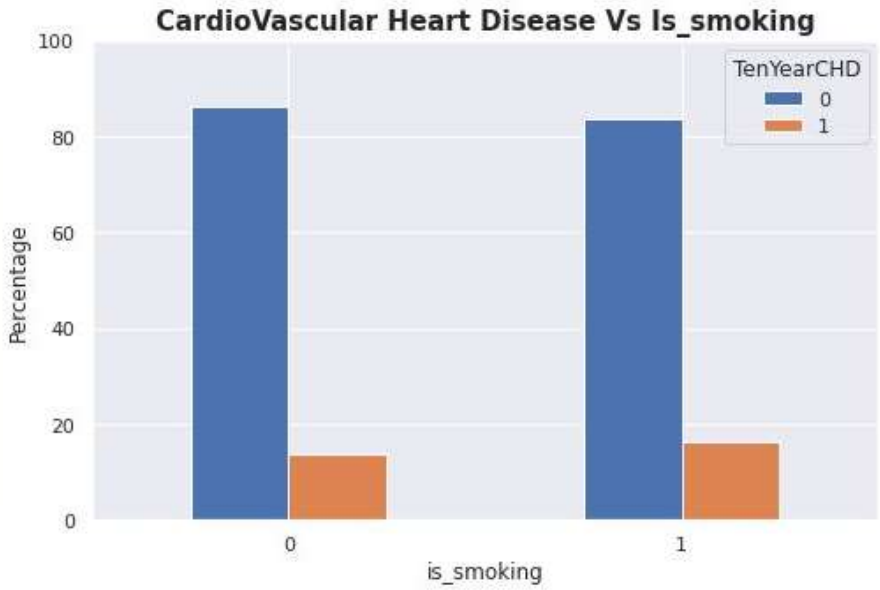




- The above Boxplots is a visualization of continuous features present in our dataset.
- From these Boxplots the median values and interquartile range of each feature can be inferred.
- Also we can see that there is a presence of outliers in all the continuous features except ‘age’.

Bivariate Analysis

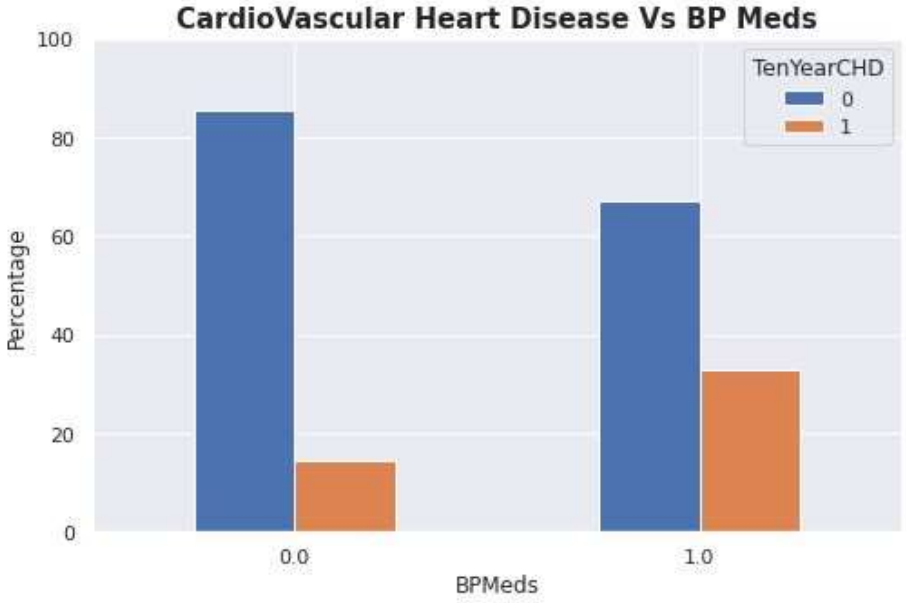


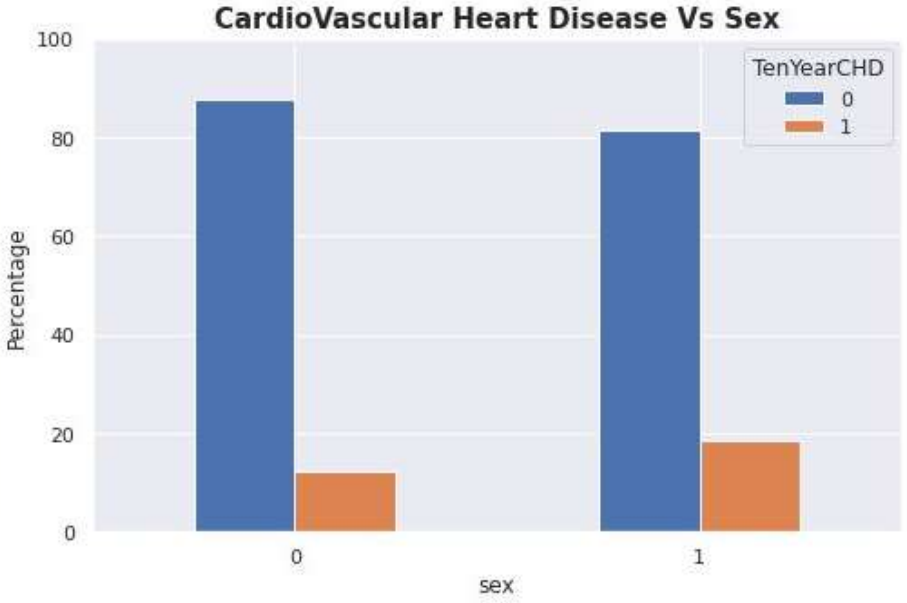


From the above two Bar chart's following inferences can be made:

- Patients with prevalent stroke symptoms have a high 10-year risk of CHD vs patients who don't have any symptoms.
- Patients irrespective of smokers vs non-smokers shows similar risks of 10 Year CHD.

Bivariate Analysis(cont.)

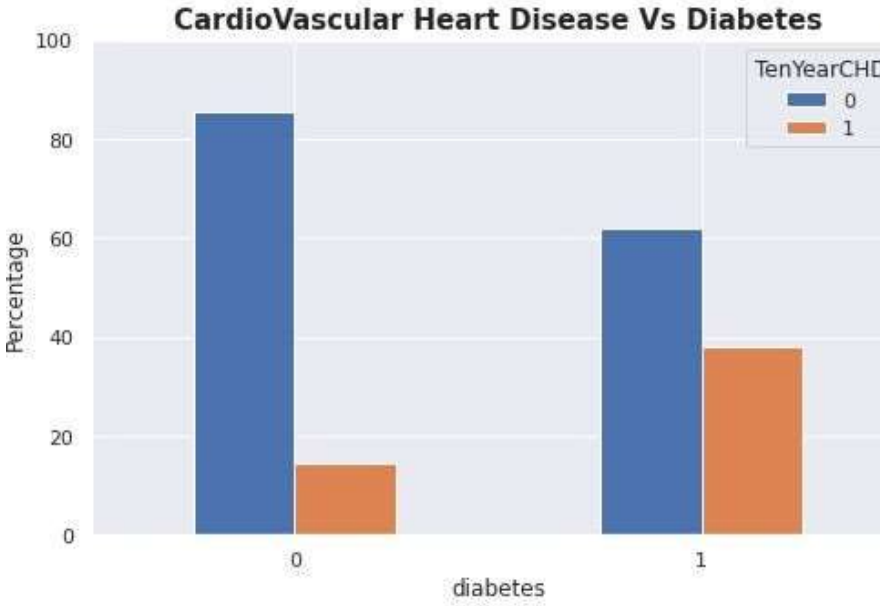


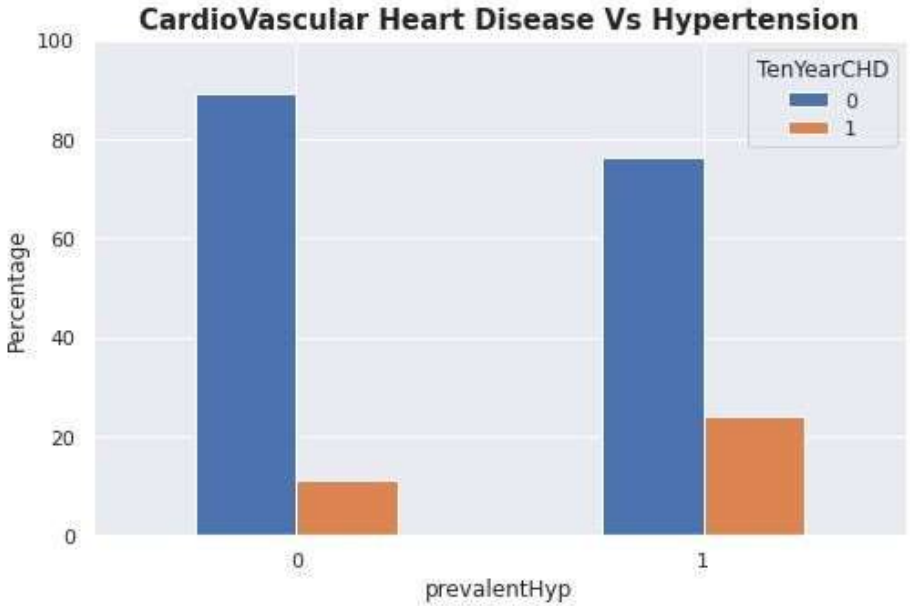


From the above two Bar Plot comparison of CHD with BP Meds & Sex following inferences can be made:

- Patients on Blood Pressure medication have higher risks of getting 10 year CHD as compared to patients who are not on medications.
- Males has slight higher risk of having 10 Year CHD.

Bivariate Analysis(cont.)





From the above two Bar Plot comparison of CHD with Diabetes and Hypertension following inferences can be made:

- Patients with Diabetes tend to have higher 10 year Risk of CHD vs patients who don't have diabetes.
- Hypertensive patients tend to have higher risk of 10 Year CHD.

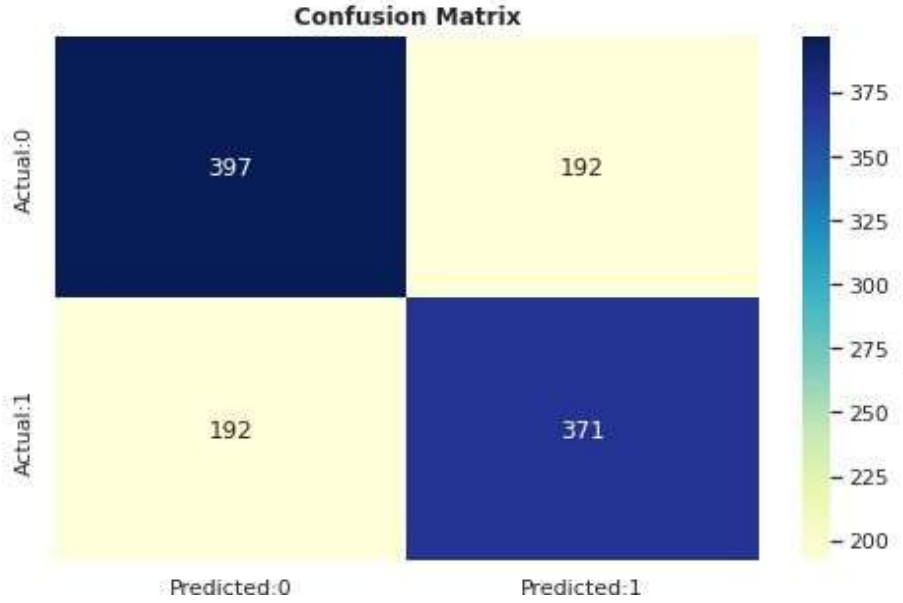
Correlation Matrix

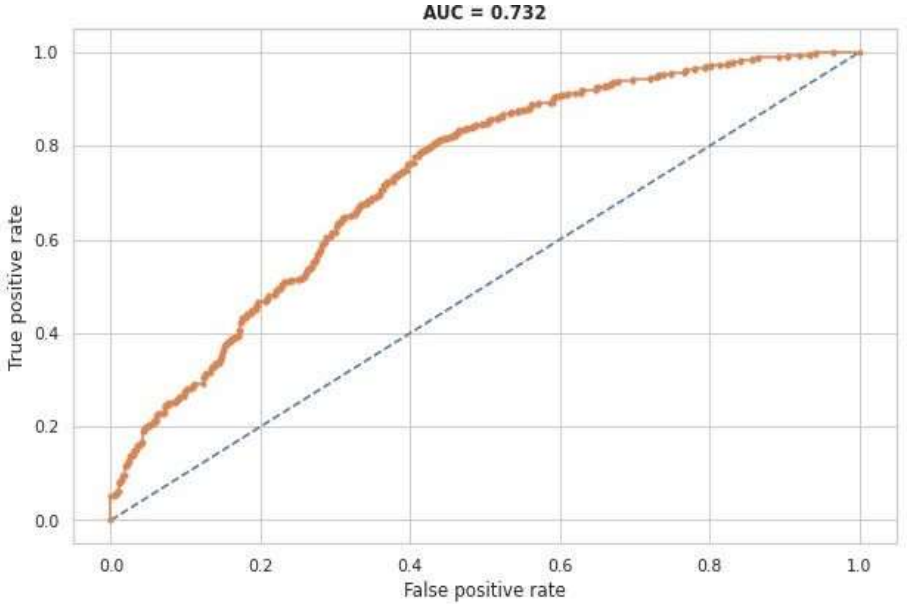
From this correlation plot we can conclude that,

There are no features with more than 0.5 correlation with the Ten year risk of developing CHD and this shows that the features are poor predictors. However the features with the highest correlations are age, prevalent hypertension and systolic blood pressure(sysBP).

Also there are a couple of features that are highly correlated with each other and for model building it's best to pick features which are independent of each other.

Logistic Regression

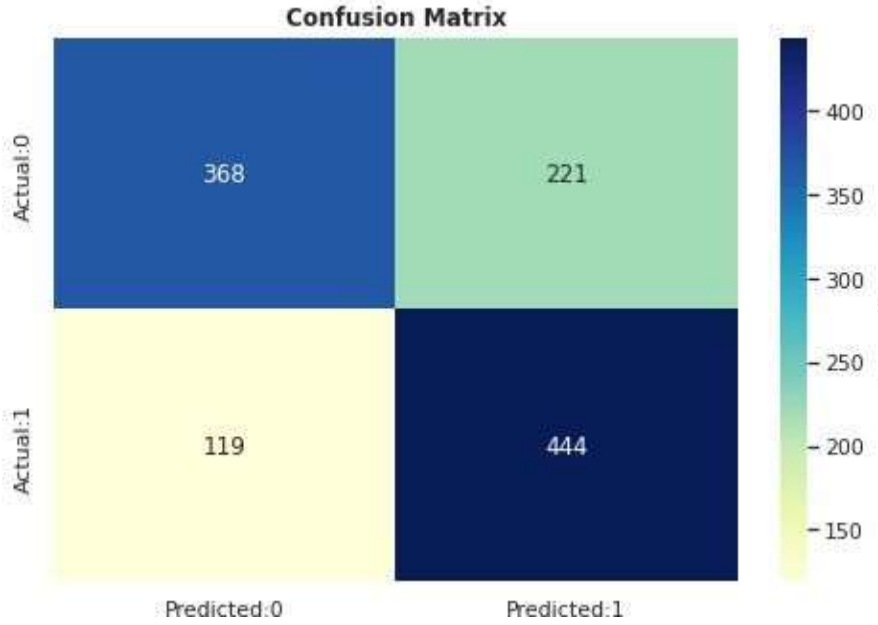


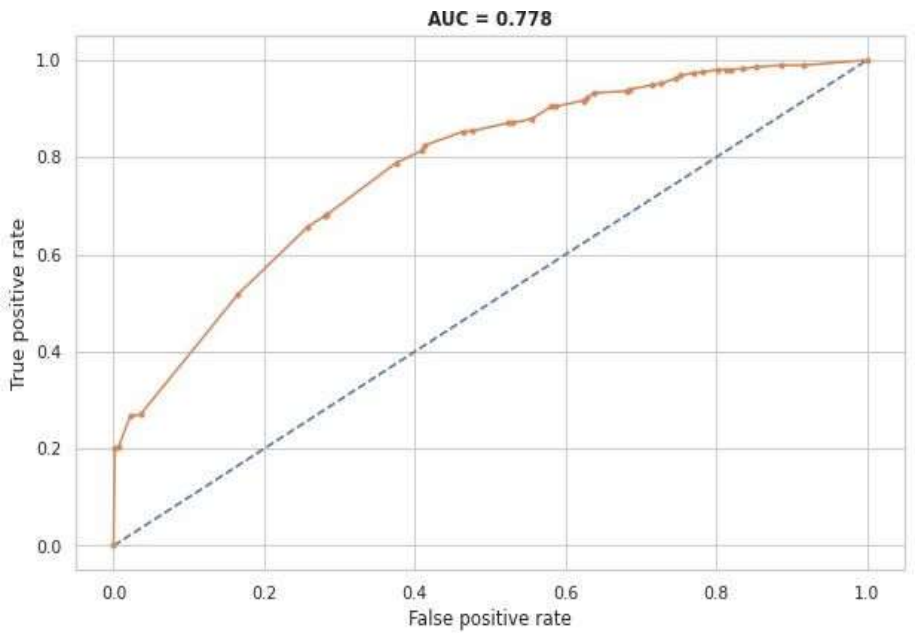


Classification Report

	precision	recall	f1-score	support
0	0.67	0.67	0.67	589
1	0.66	0.66	0.66	563
accuracy			0.67	1152
macro avg	0.67	0.67	0.67	1152
weighted avg	0.67	0.67	0.67	1152

Decision Trees

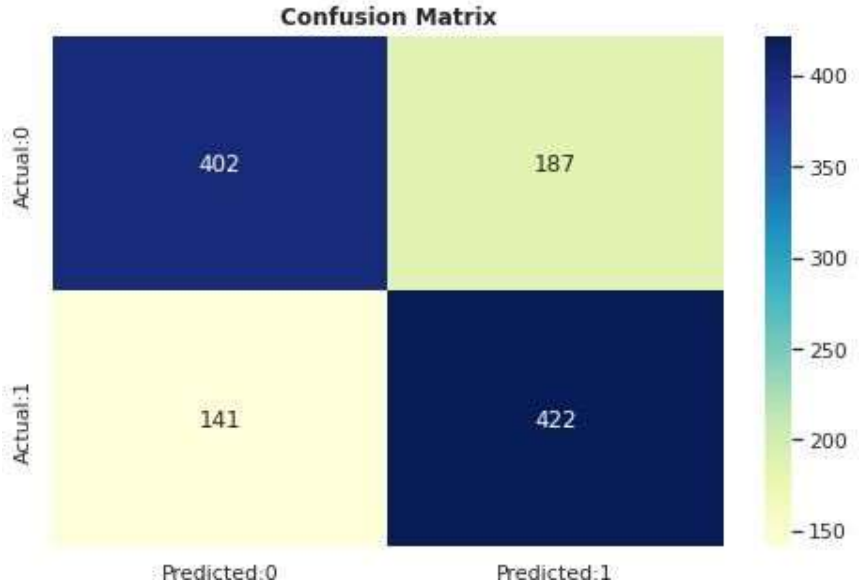


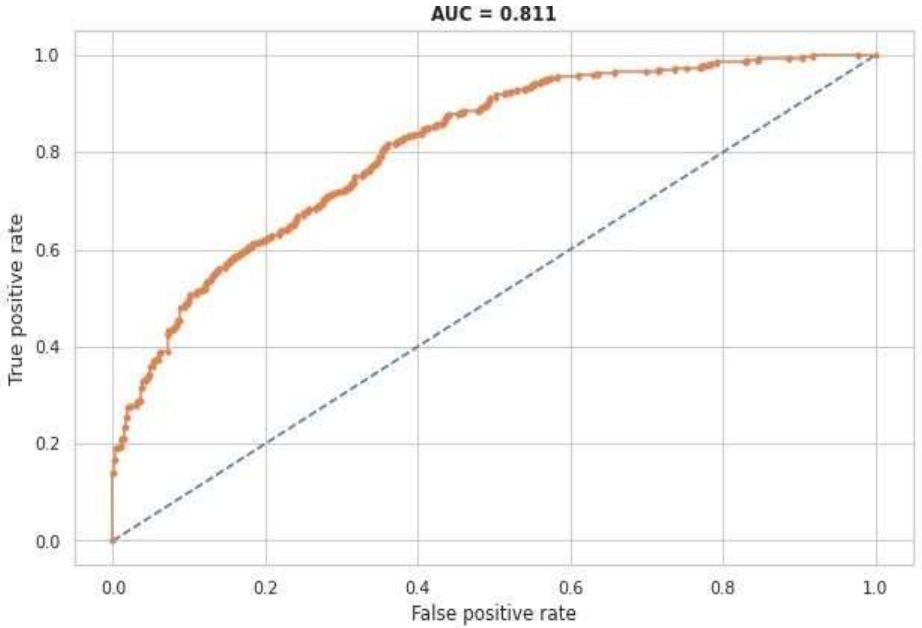


Classification Report

	precision	recall	f1-score	support
0	0.76	0.62	0.68	589
1	0.67	0.79	0.72	563
accuracy			0.70	1152
macro avg	0.71	0.71	0.70	1152
weighted avg	0.71	0.70	0.70	1152

Random Forest

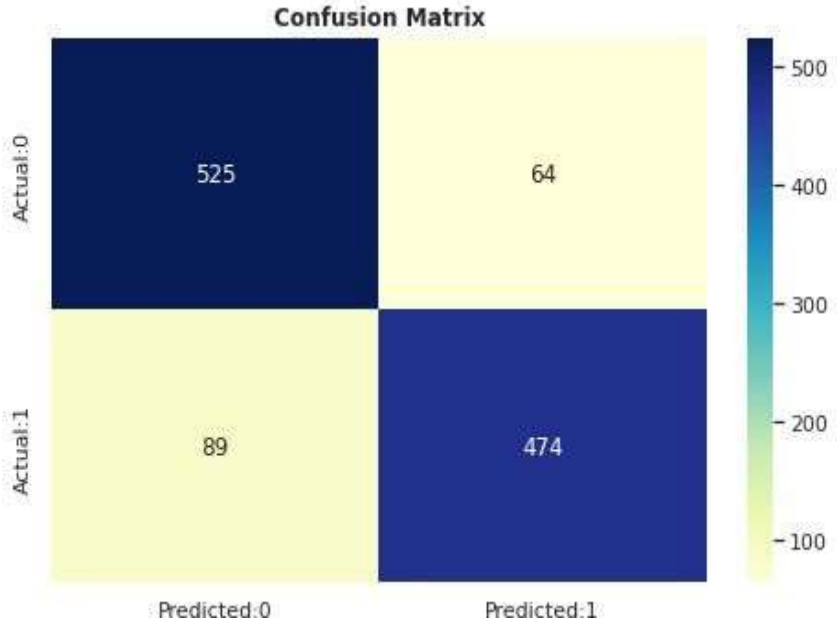


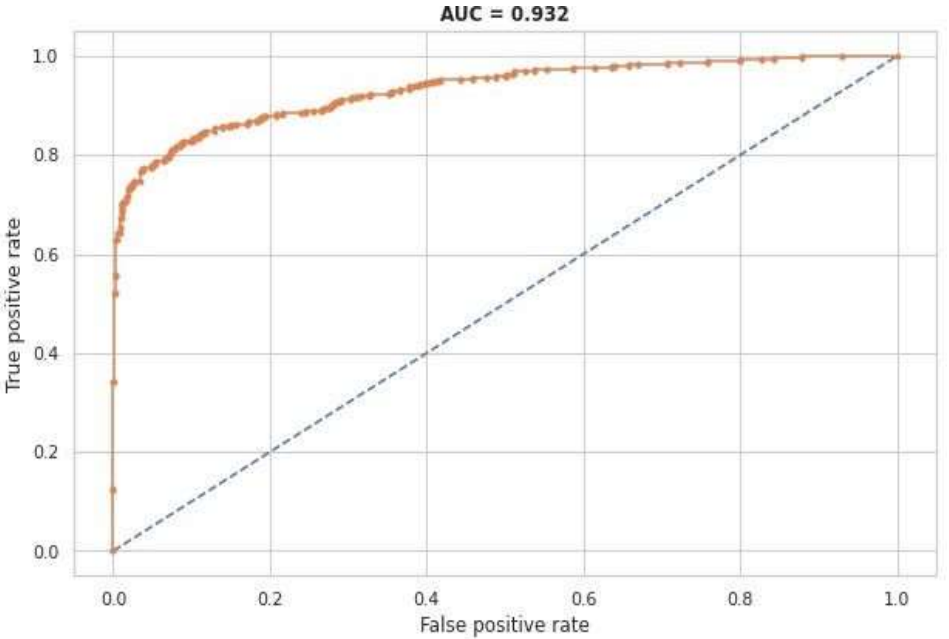


Classification Report

	precision	recall	f1-score	support
0	0.74	0.68	0.71	589
1	0.69	0.75	0.72	563
accuracy			0.72	1152
macro avg	0.72	0.72	0.72	1152
weighted avg	0.72	0.72	0.72	1152

XGBoost

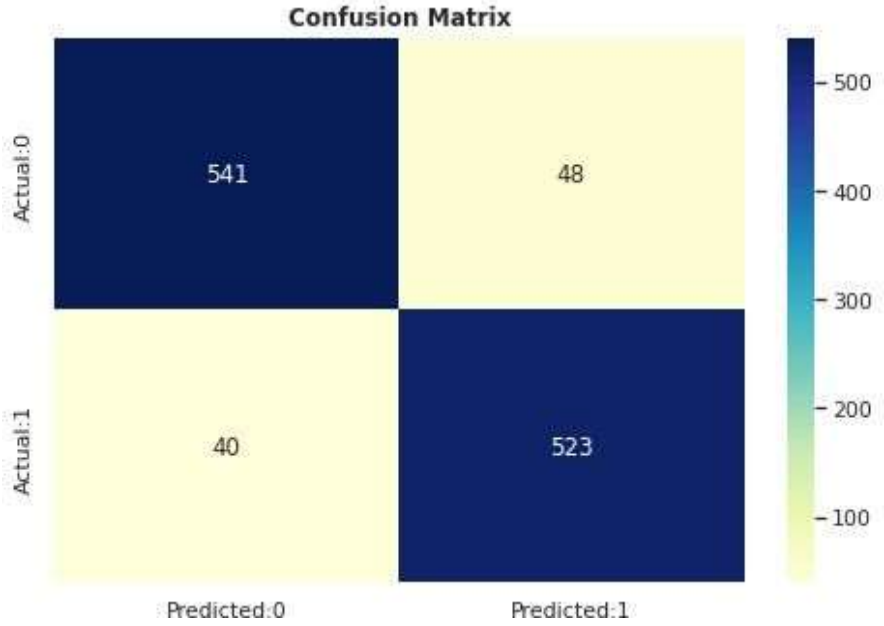


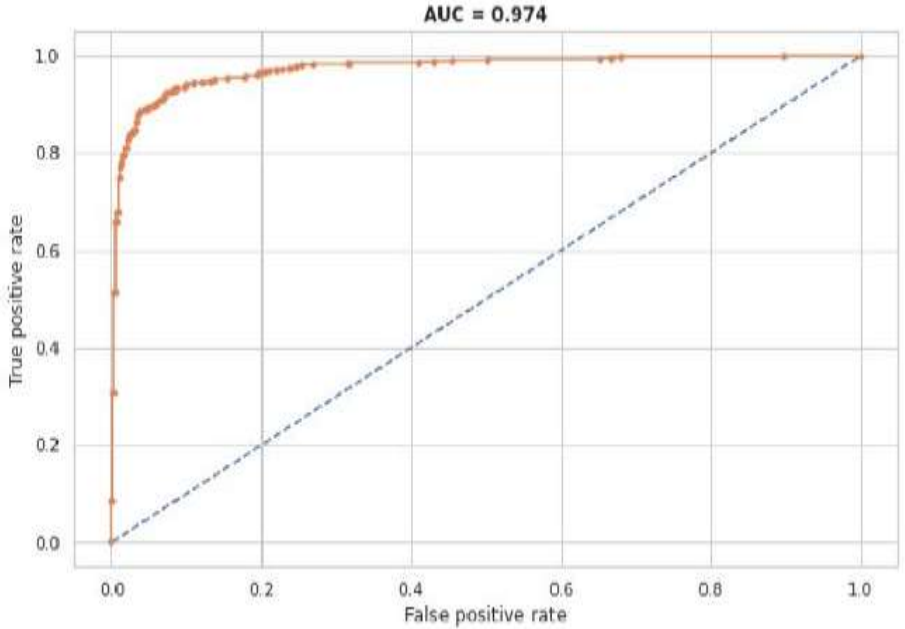


Classification Report

	precision	recall	f1-score	support
0	0.83	0.88	0.86	589
1	0.87	0.82	0.84	563
accuracy			0.85	1152
macro avg	0.85	0.85	0.85	1152
weighted avg	0.85	0.85	0.85	1152

Support Vector Machine





Classification Report

	precision	recall	f1-score	support
0	0.94	0.92	0.93	589
1	0.91	0.94	0.93	563
accuracy			0.93	1152
macro avg	0.93	0.93	0.93	1152
weighted avg	0.93	0.93	0.93	1152

Model Comparison Matrix

	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	Auc
Logistic Regression	68.71	66.67	65.90	65.90	65.90	73.18
Decision Trees	73.40	70.49	66.77	78.86	72.31	77.78
Random Forest	75.68	71.53	69.29	74.96	72.01	81.06
XG Boost	93.38	86.72	88.10	84.19	86.10	93.25
SVC	99.91	92.36	91.59	92.90	92.24	97.37

Summary

- We have used Logistic Regression, Decision Trees, Random Forest, XGBoost and SVC for modelling. Based on our observations, the Support vector machine with the radial kernel is the best performing model in terms of accuracy and the F1 score and it's high AUC-score shows that it has a high true positive rate.

- Balancing the dataset by using the SMOTE technique helped in improving the models' sensitivity and improving the scores of all the models.

- With scaling the dataset the performance of SVM models is worst. Without scaling, the performance of SVM is best among all other models.

- With more data(especially that of the minority class) better models can be built.

Thank you