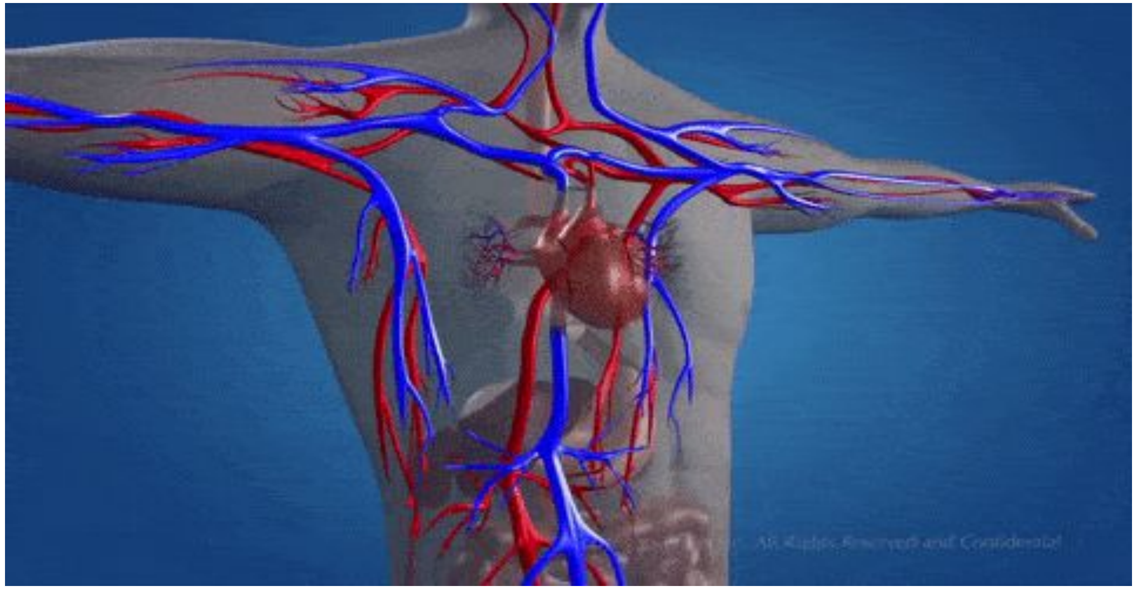
# Cardiovascular Risk Prediction

**TEAM MEMBERS**

**Saurabh Arawad**

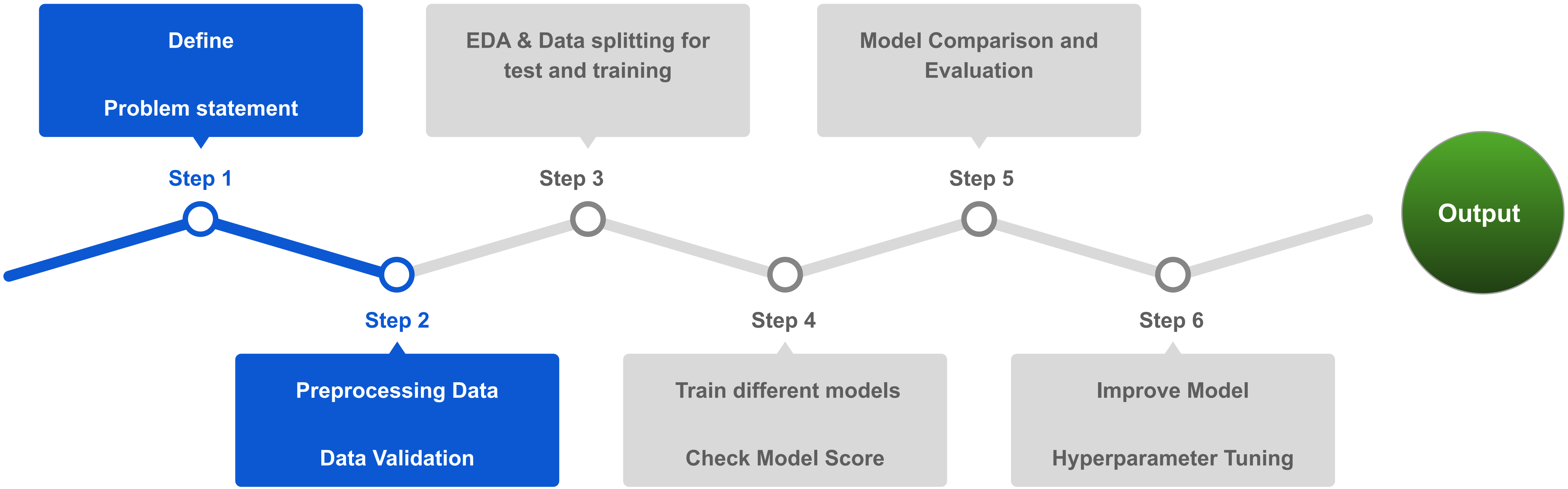
**Aman Guleria**

**Rishika Rai**



# Content

# Steps In Supervised ML Classification



Step 1

Define

Problem statement

Step 2

Preprocessing Data

Data Validation

Step 3

EDA & Data splitting for test and training

Step 4

Train different models

Check Model Score

Step 5

Model Comparison and Evaluation

Step 6

Improve Model

Hyperparameter Tuning

Output

# Problem statement

The objective of the project is to come up with the machine learning model to predict whether a patient has 10-year risk of developing coronary heart disease (CHD) using the residents of the town of Framingham, Massachusetts dataset.

# Data Summary:

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.

**Attributes:**

**Demographic:**

- Sex: male or female("M" or "F")(Nominal)
- Age: Age of the patient(Continuous)

**Behavioral:**

- is_smoking: Whether or not the patient is a current smoker ("YES" or "NO") (Nominal)
- cigs Per Day : the number of cigarettes that the person smoked on average in one day. (Continuous)

# Contd..

**Medical(history):**

- BPmeds : whether or not the patient was on blood pressure medication (Nominal)
- prevalentStroke : whether or not the patient had previously had a stroke (Nominal)
- prevalentHyp : whether or not the patient was hypertensive (Nominal)
- diabetes: whether or not the patient had diabetes (Nominal)

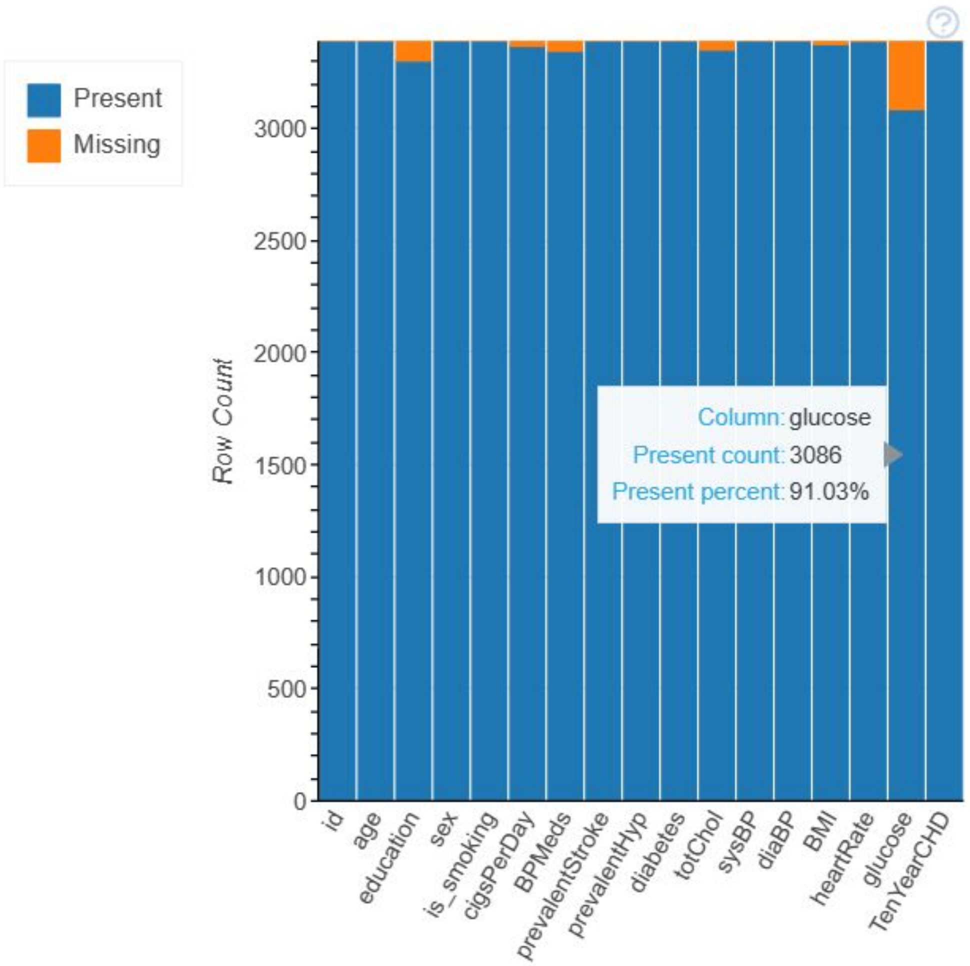
**Medical(current)**

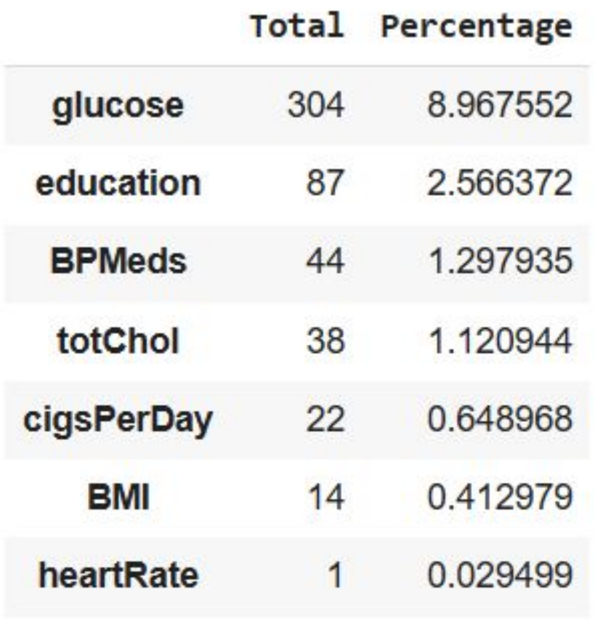
- totChol: total cholesterol level (Continuous)
- sysBP: systolic blood pressure (Continuous)
- diaBP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- heartRate: heart rate (Continuous)
- glucose: glucose level (Continuous)

**Predict variable (desired target):**

- **TenYearCHD:**10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0”  means “No”) –Discrete variable

# Missing Value Treatment :





|  | Total | Percentage |
|---|---|---|
| **glucose** | 304 | 8.967552 |
| **education** | 87 | 2.566372 |
| **BPMeds** | 44 | 1.297935 |
| **totChol** | 38 | 1.120944 |
| **cigsPerDay** | 22 | 0.648968 |
| **BMI** | 14 | 0.412979 |
| **heartRate** | 1 | 0.029499 |

- At **8.97%**, the blood glucose entry has the highest percentage of missing data. The other features have very few missing entries.
- Since the missing entries account for only **11%** of the total data, we can exclude these entries without losing most of the data.

# Exploratory Data Analysis

The EDA for this dataset is done with the help of DataPrep Library.

The goal of create_report is to generate profile reports from a pandas DataFrame. create_report utilises the functionalities and formats the plots from dataprep. It provides the following information:

- ➔ **Overview**: detect the types of columns in a dataframe
- ➔ **Variables**: variable type, unique values, distinct count, missing values
- ➔ **Quantile statistics** like minimum value, Q1, median, Q3, maximum, range, interquartile range
- ➔ **Descriptive statistics** like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- ➔ **Text analysis** for length, sample and letter
- ➔ **Correlations**: highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
- ➔ **Missing Values**: bar chart, heatmap and spectrum of missing values
- ➔ In the following, we break down the report into different sections to demonstrate each part of the report.

# Integer Treatment

Before we go ahead, an important step to do is to convert our string feature  into an integer.

➔ In sex feature  M will be converted to 1 and F will be converted to 0 .

➔ In is_smoking feature YES will be converted to 1 and NO will be converted to 0 .

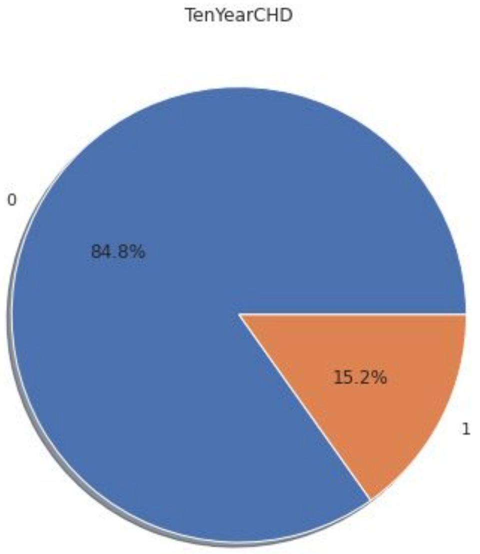
# Feature Analysis

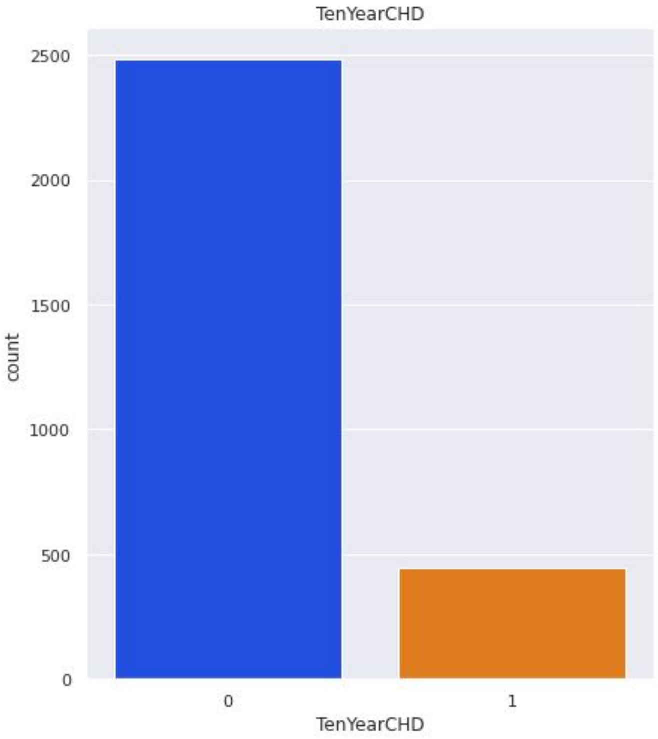
Our target variable is **TenYearCHD**

There are 2483 patients without heart disease and 444 patients with the disease.

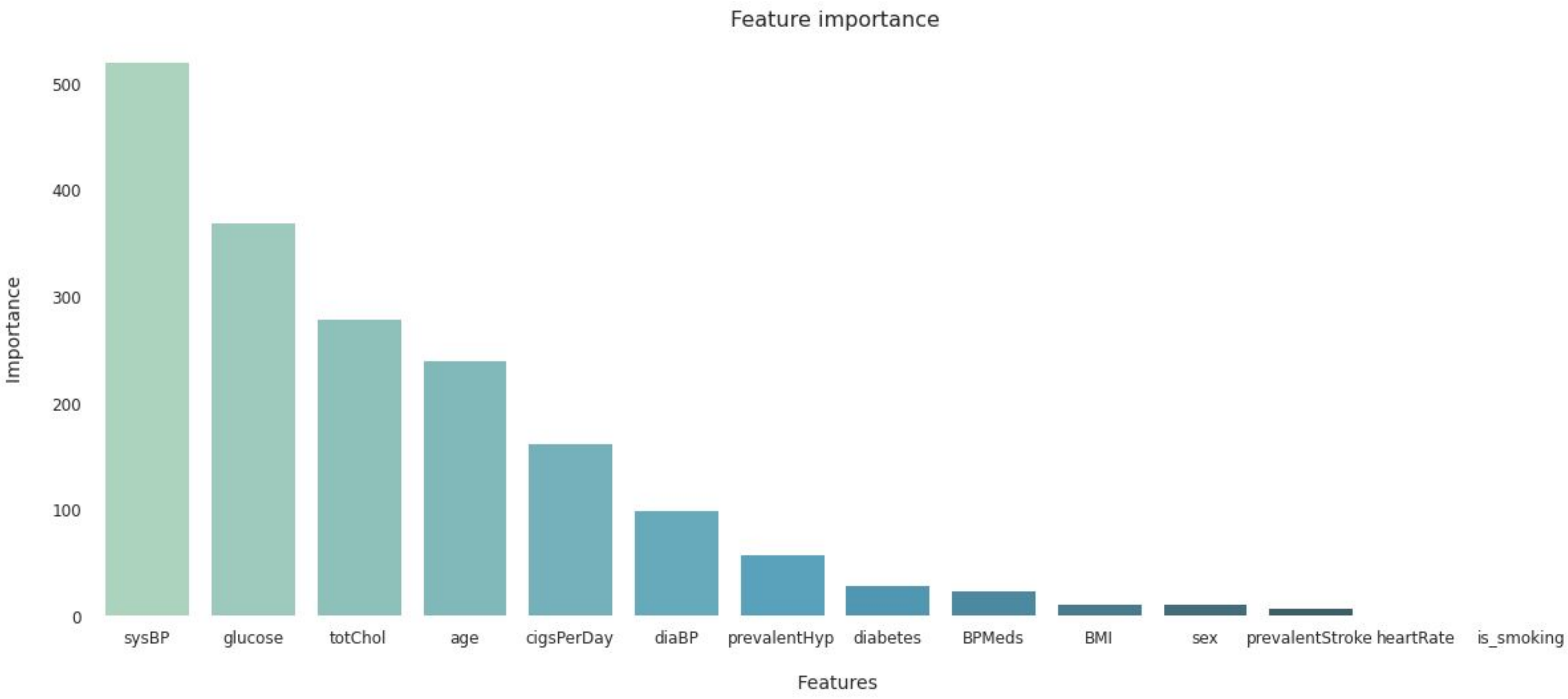
➔ **1** indicates  person **have risk** of **coronary heart disease**

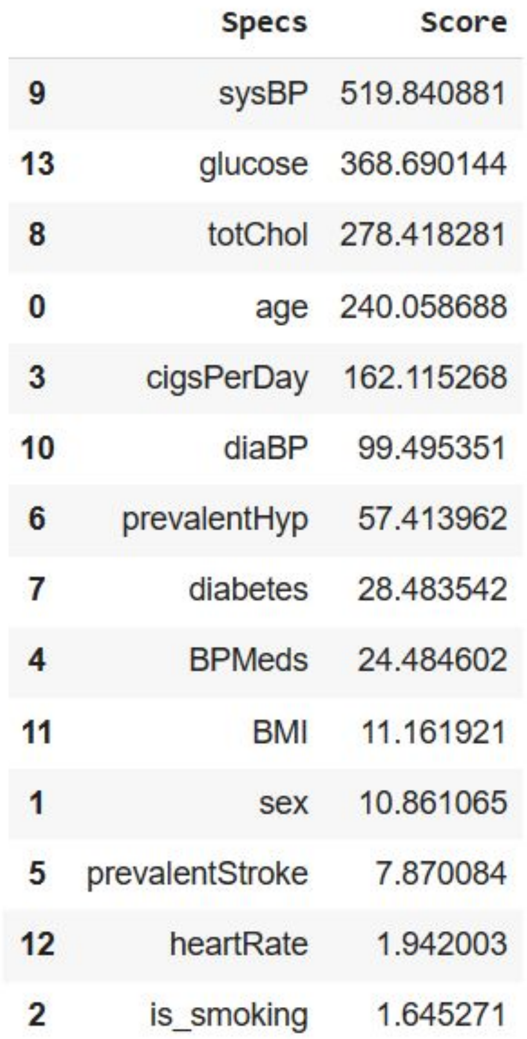
➔ **0** indicates person **do not have risk** of **coronary heart disease**





# Feature Selection

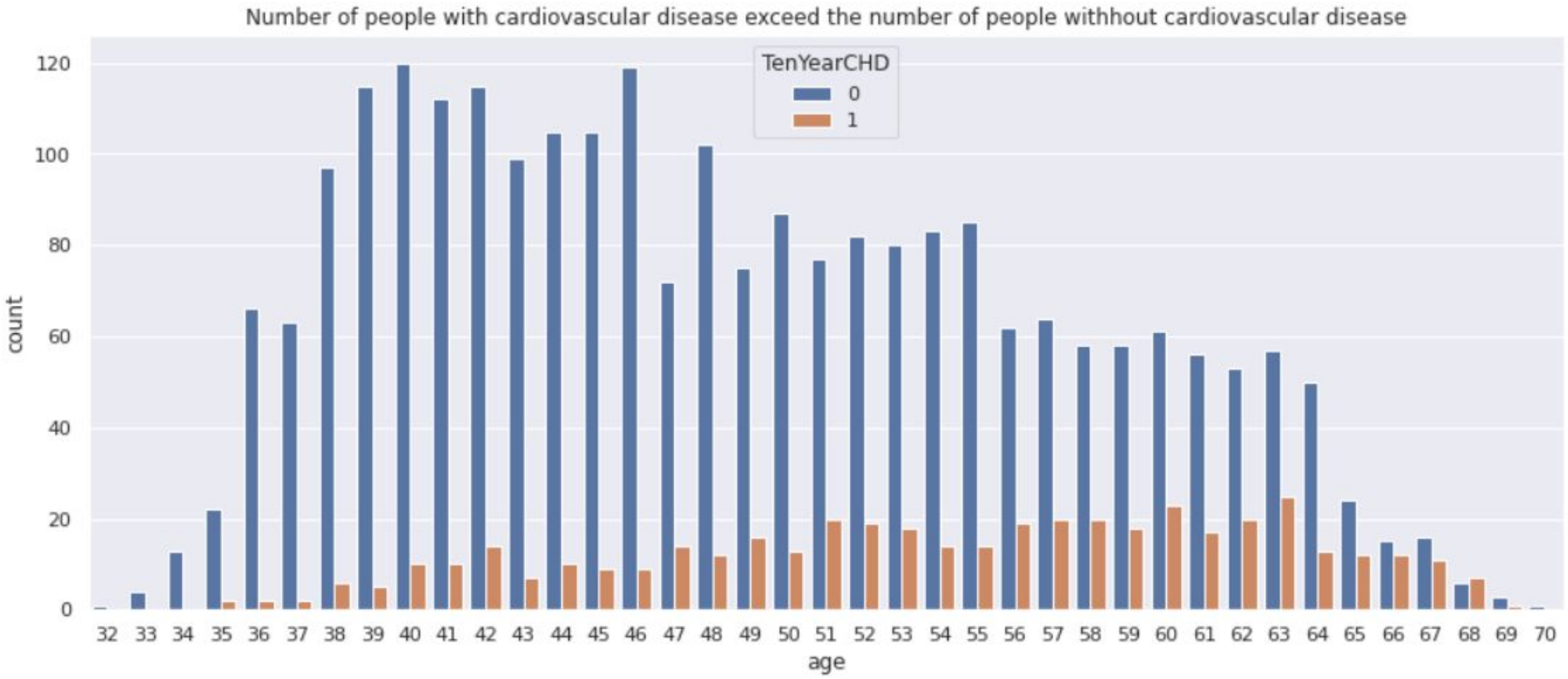




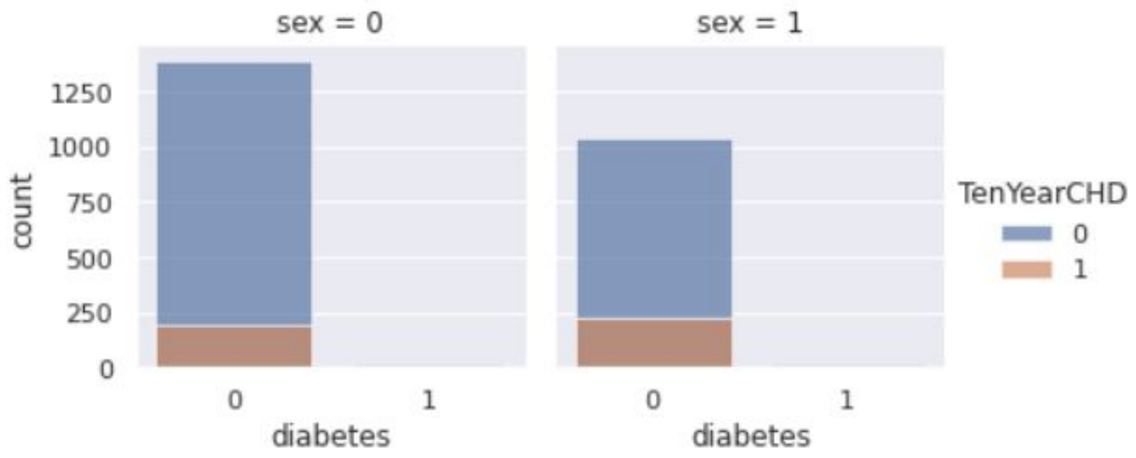
| | Specs | Score |
|---|---|---|
| 9 | sysBP | 519.840881 |
| 13 | glucose | 368.690144 |
| 8 | totChol | 278.418281 |
| 0 | age | 240.058688 |
| 3 | cigsPerDay | 162.115268 |
| 10 | diaBP | 99.495351 |
| 6 | prevalentHyp | 57.413962 |
| 7 | diabetes | 28.483542 |
| 4 | BPMeds | 24.484602 |
| 11 | BMI | 11.161921 |
| 1 | sex | 10.861065 |
| 5 | prevalentStroke | 7.870084 |
| 12 | heartRate | 1.942003 |
| 2 | is_smoking | 1.645271 |

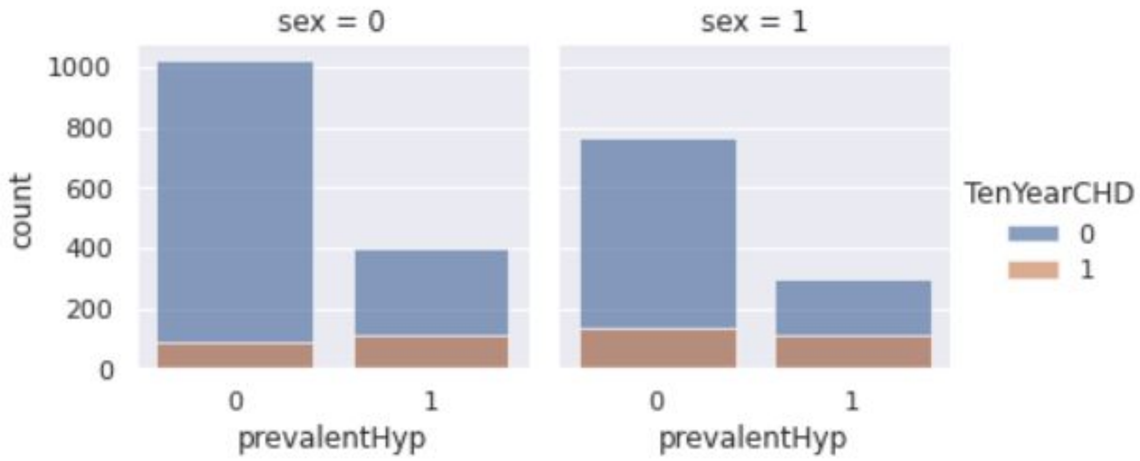
From above visualisation we can consider features such as sysBP, glucose, totChol, age, cigsPerDay, diaBP who have very scores in model creation.

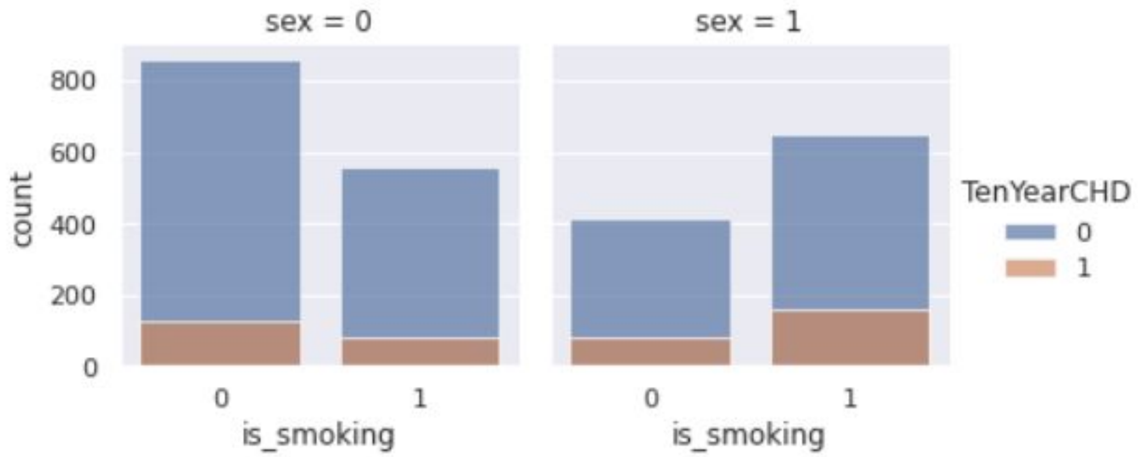
# Target Variable Analysis

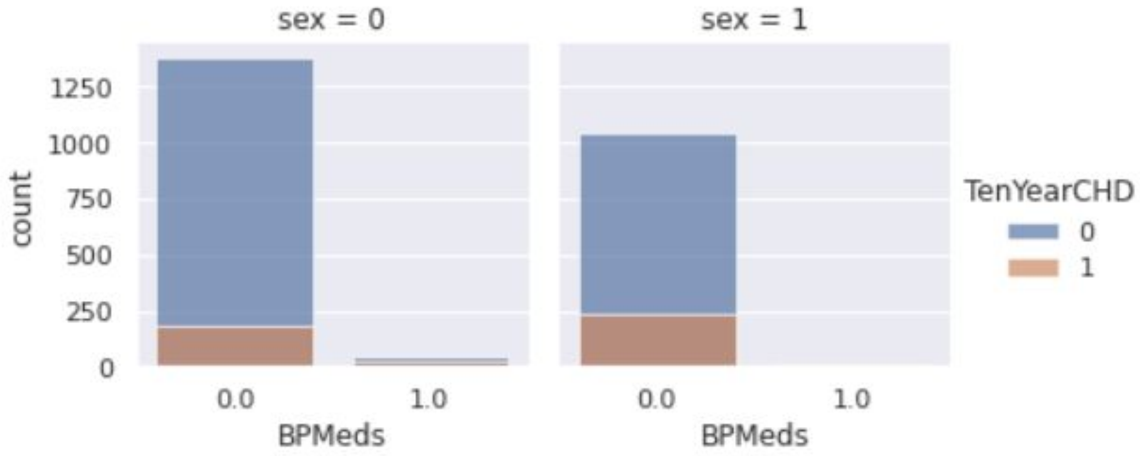


# Categorical comparison

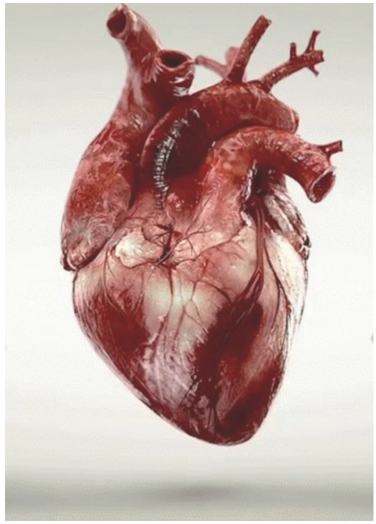




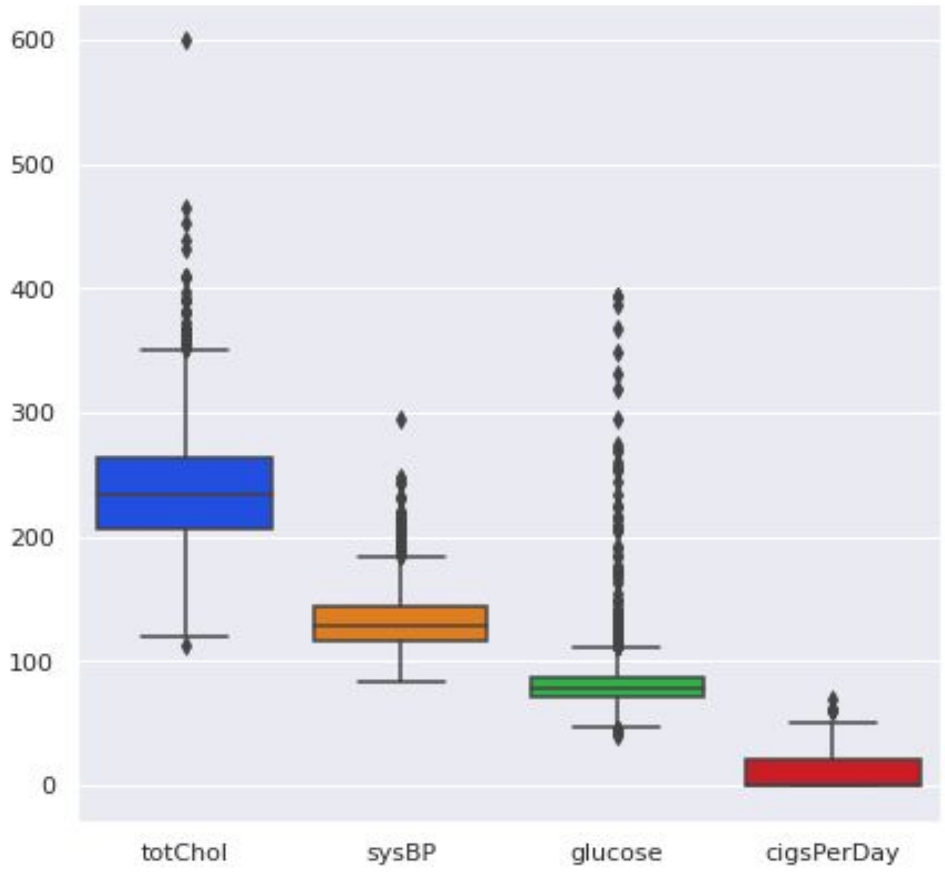




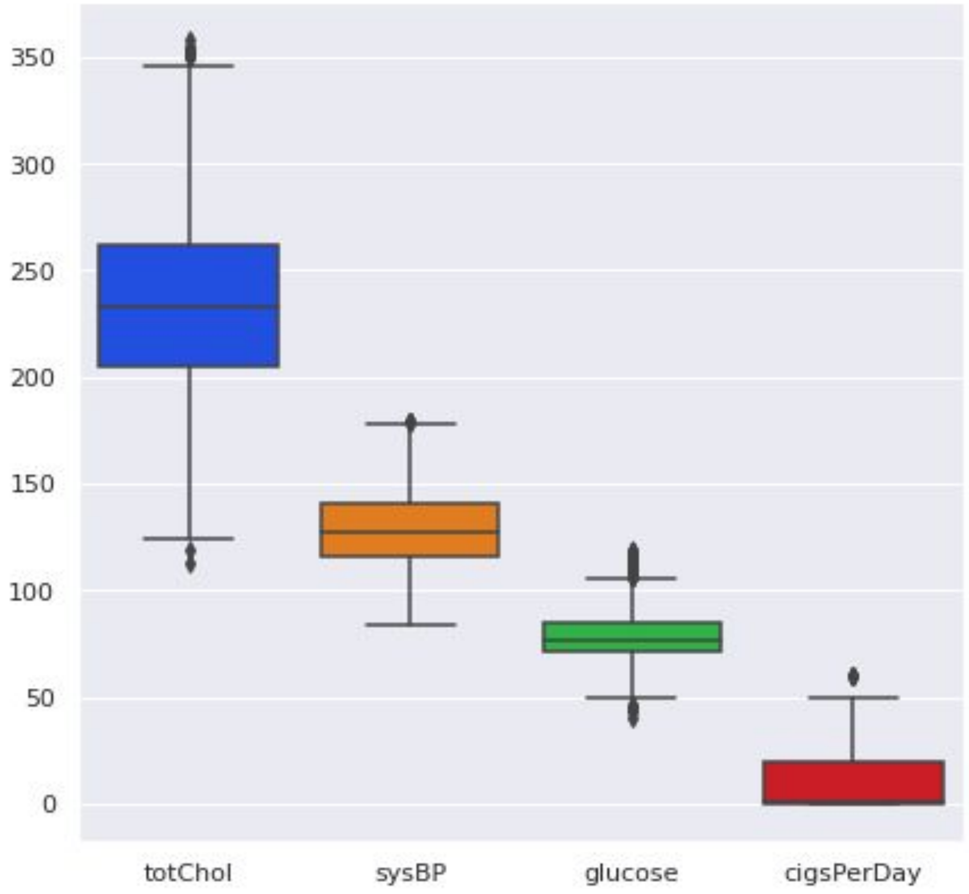
- Slightly more males are suffering from Cardiovascular heart disease than females.
- The people who have Cardiovascular heart disease is almost equal between smokers and non smokers
- The percentage of people who have Cardiovascular heart disease is higher among the diabetic patients and also those 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.



# Outlier Treatment:

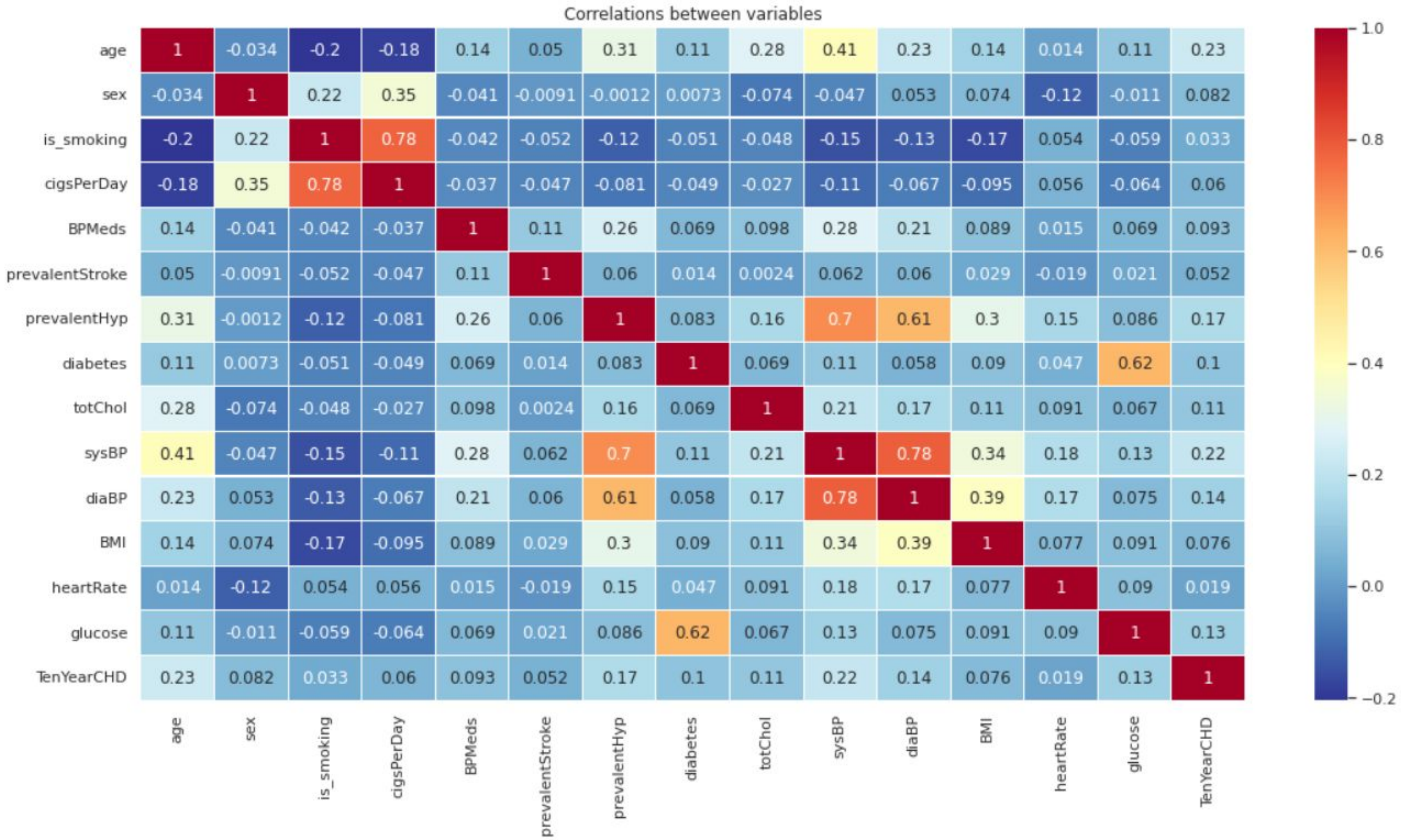


**Before Outlier Treatment**



**After Outlier Treatment**

# Correlation Analysis



Correlations between variables

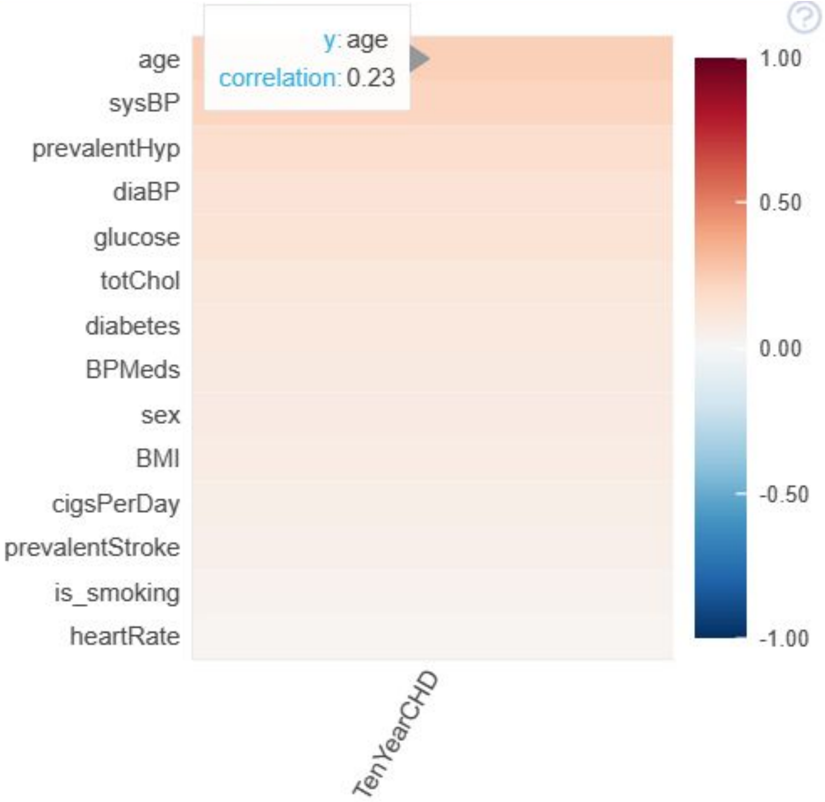
| | age | sex | is_smoking | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP | BMI | heartRate | glucose | TenYearCHD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1 | -0.034 | -0.2 | -0.18 | 0.14 | 0.05 | 0.31 | 0.11 | 0.28 | 0.41 | 0.23 | 0.14 | 0.014 | 0.11 | 0.23 |
| sex | -0.034 | 1 | 0.22 | 0.35 | -0.041 | -0.0091 | -0.0012 | 0.0073 | -0.074 | -0.047 | 0.053 | 0.074 | -0.12 | -0.011 | 0.082 |
| is_smoking | -0.2 | 0.22 | 1 | 0.78 | -0.042 | -0.052 | -0.12 | -0.051 | -0.048 | -0.15 | -0.13 | -0.17 | 0.054 | -0.059 | 0.033 |
| cigsPerDay | -0.18 | 0.35 | 0.78 | 1 | -0.037 | -0.047 | -0.081 | -0.049 | -0.027 | -0.11 | -0.067 | -0.095 | 0.056 | -0.064 | 0.06 |
| BPMeds | 0.14 | -0.041 | -0.042 | -0.037 | 1 | 0.11 | 0.26 | 0.069 | 0.098 | 0.28 | 0.21 | 0.089 | 0.015 | 0.069 | 0.093 |
| prevalentStroke | 0.05 | -0.0091 | -0.052 | -0.047 | 0.11 | 1 | 0.06 | 0.014 | 0.0024 | 0.062 | 0.06 | 0.029 | -0.019 | 0.021 | 0.052 |
| prevalentHyp | 0.31 | -0.0012 | -0.12 | -0.081 | 0.26 | 0.06 | 1 | 0.083 | 0.16 | 0.7 | 0.61 | 0.3 | 0.15 | 0.086 | 0.17 |
| diabetes | 0.11 | 0.0073 | -0.051 | -0.049 | 0.069 | 0.014 | 0.083 | 1 | 0.069 | 0.11 | 0.058 | 0.09 | 0.047 | 0.62 | 0.1 |
| totChol | 0.28 | -0.074 | -0.048 | -0.027 | 0.098 | 0.0024 | 0.16 | 0.069 | 1 | 0.21 | 0.17 | 0.11 | 0.091 | 0.067 | 0.11 |
| sysBP | 0.41 | -0.047 | -0.15 | -0.11 | 0.28 | 0.062 | 0.7 | 0.11 | 0.21 | 1 | 0.78 | 0.34 | 0.18 | 0.13 | 0.22 |
| diaBP | 0.23 | 0.053 | -0.13 | -0.067 | 0.21 | 0.06 | 0.61 | 0.058 | 0.17 | 0.78 | 1 | 0.39 | 0.17 | 0.075 | 0.14 |
| BMI | 0.14 | 0.074 | -0.17 | -0.095 | 0.089 | 0.029 | 0.3 | 0.09 | 0.11 | 0.34 | 0.39 | 1 | 0.077 | 0.091 | 0.076 |
| heartRate | 0.014 | -0.12 | 0.054 | 0.056 | 0.015 | -0.019 | 0.15 | 0.047 | 0.091 | 0.18 | 0.17 | 0.077 | 1 | 0.09 | 0.019 |
| glucose | 0.11 | -0.011 | -0.059 | -0.064 | 0.069 | 0.021 | 0.086 | 0.62 | 0.067 | 0.13 | 0.075 | 0.091 | 0.09 | 1 | 0.13 |
| TenYearCHD | 0.23 | 0.082 | 0.033 | 0.06 | 0.093 | 0.052 | 0.17 | 0.1 | 0.11 | 0.22 | 0.14 | 0.076 | 0.019 | 0.13 | 1 |

## Contd..

- There are no features with more than 0.5 correlation with the Ten year risk of developing CHD and this shows that the features a poor predictors.
- 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



# Feature Engineering/ Selection:

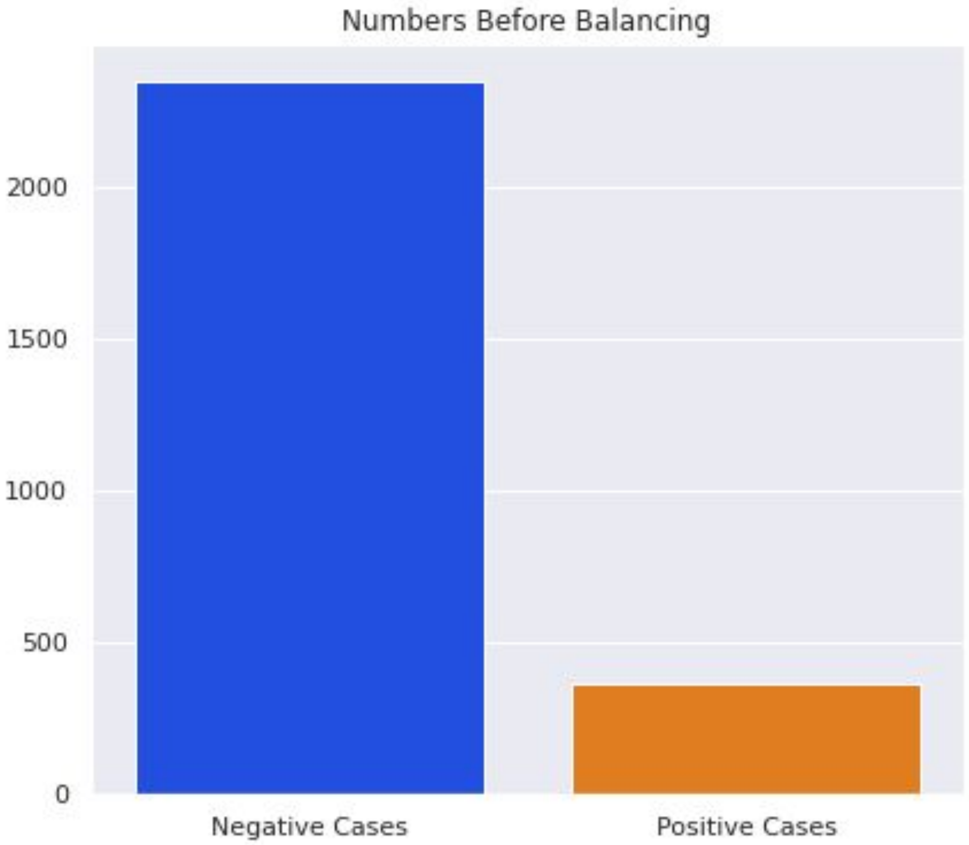
- For feature selection we’ve used **Tree-based: SelectFromModel** which is an embedded methods use algorithms that have built-in feature selection methods
- For feature selection we’ve used **Tree-based: SelectFromModel** which is an embedded methods use algorithms that have built-in feature selection methods
- In Random forest, the final feature importance is the average of all decision tree feature importance.
- After performing the feature selection the important features are:

**['age', 'totChol, 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose']**

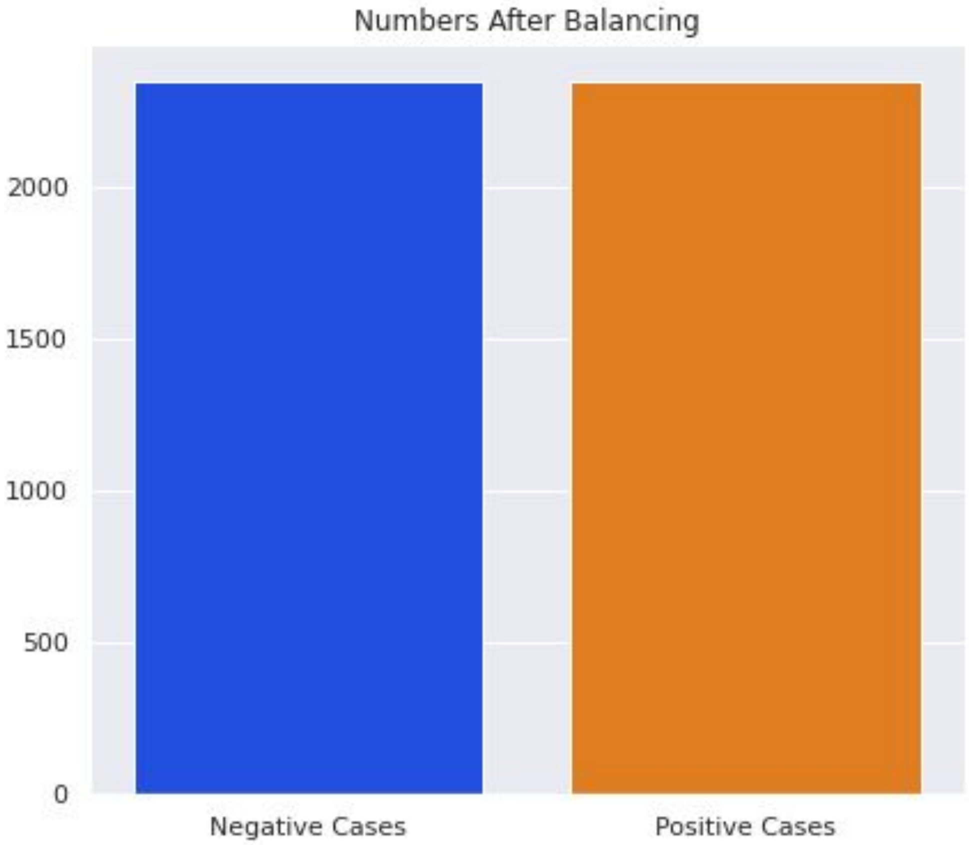
# Data Preprocessing:

**SMOT -** Let’s balance our dataset

- First we balance our dataset because for every positive case there are about **5-6** negative cases.
- To handle this problem we will balance the dataset using the **Synthetic Minority Oversampling**



Numbers Before Balancing



Numbers After Balancing

# Data Preprocessing:

### Split

- Data is splitted for test and training purpose.
- We have splitted our new balanced data into 80% and 20% split for training and testing.

### Metrics

We use evaluation metrics for comparing our model. For this classification problem we have use some metrics and they are as follows:

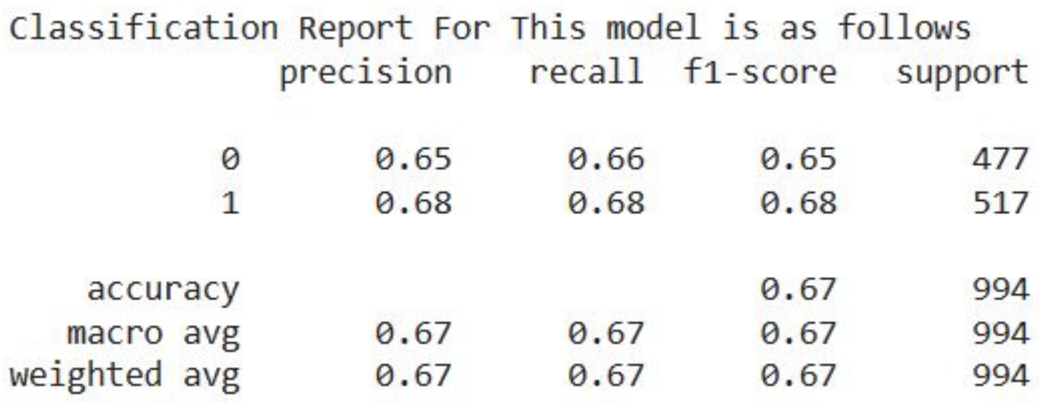
- GridSearchCV: for the selection of best parameters.
- confusion_matrix: for evaluation of classification model accuracy.
- classification_report: Text report creation for main classification matrix.
- Scores: accuracy_score, f1_score, recall_score, prescision_score, roc_auc_score, roc_curve.

# Classification models Used:

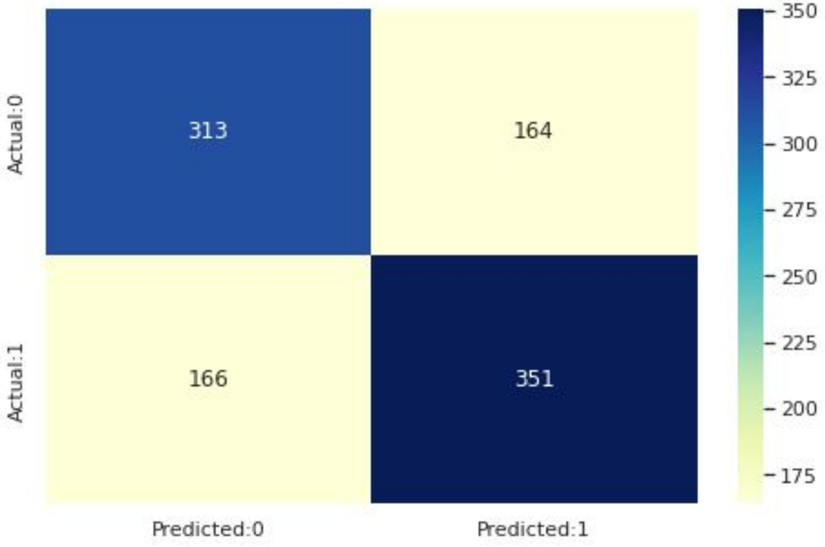
Here, we'll be using this 4 algorithms along with **GridsearchCV** for finding optimum parameters:

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

# 1. Logistic regression :



```
Classification Report For This model is as follows
              precision    recall  f1-score   support

           0       0.65      0.66      0.65       477
           1       0.68      0.68      0.68       517

    accuracy                           0.67       994
   macro avg       0.67      0.67      0.67       994
weighted avg       0.67      0.67      0.67       994
```

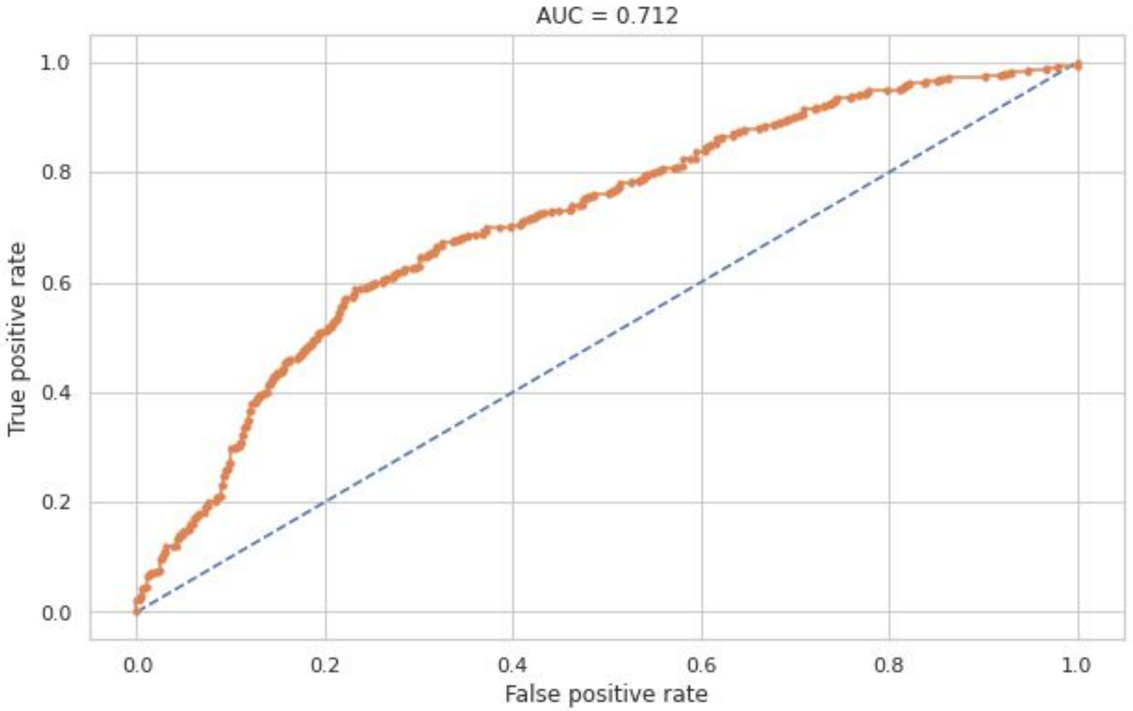


## Best parameters:

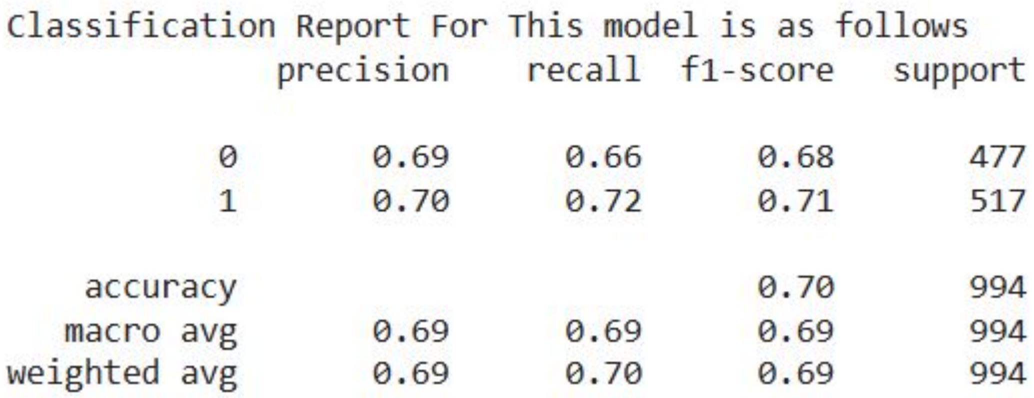
{'C': 10,

'class_weight': None,

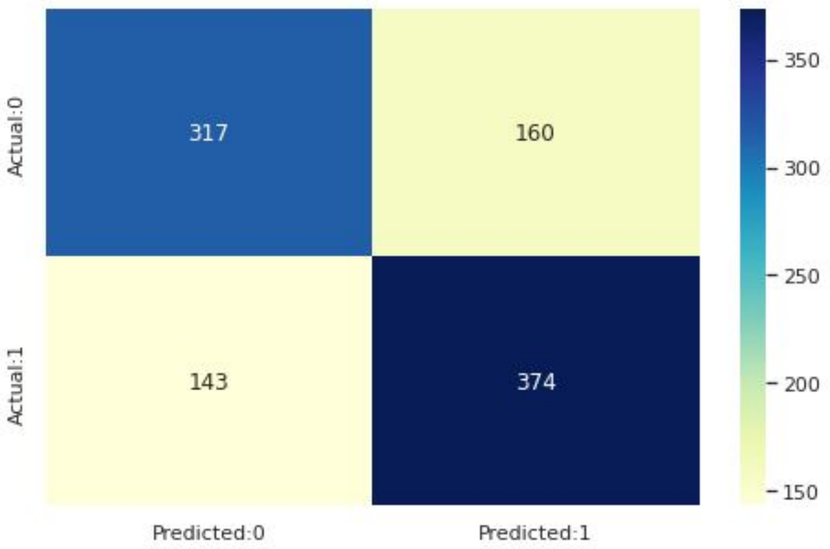
'penalty': 'l2'}



# 2. Random Forest :



```
Classification Report For This model is as follows
              precision    recall  f1-score   support

           0       0.69      0.66      0.68       477
           1       0.70      0.72      0.71       517

    accuracy                           0.70       994
   macro avg       0.69      0.69      0.69       994
weighted avg       0.69      0.70      0.69       994
```



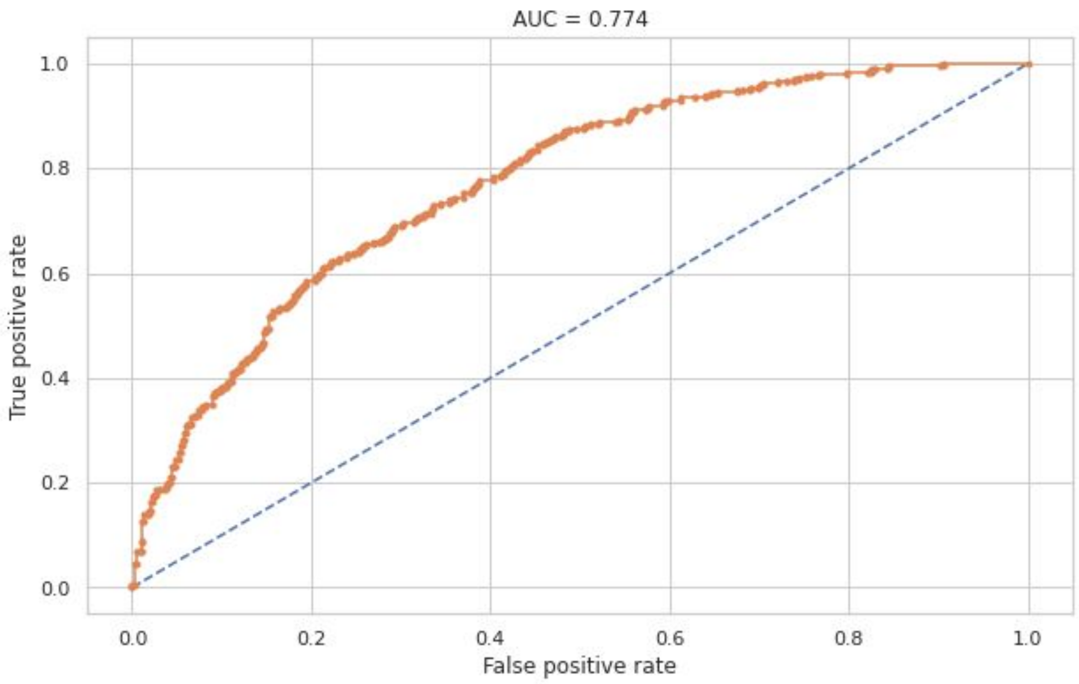
**Best parameters:**

{'max_depth': 8,

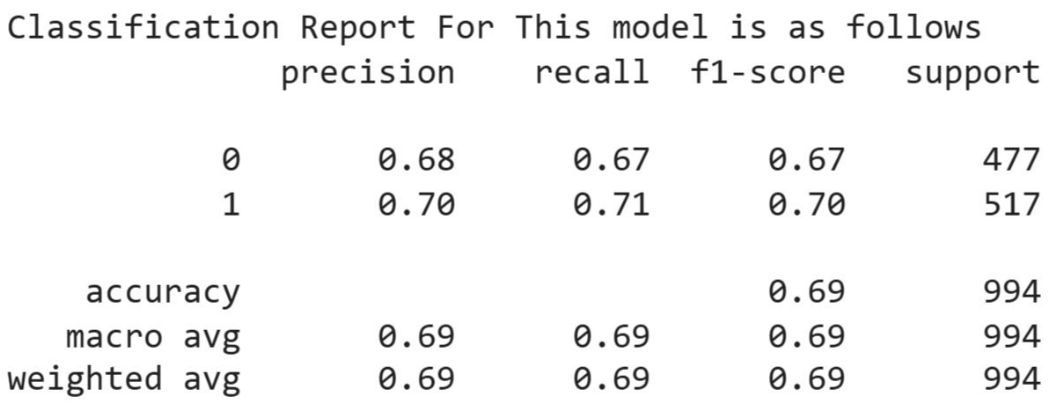
'min_samples_leaf': 40,

‘min_samples_split': 50,

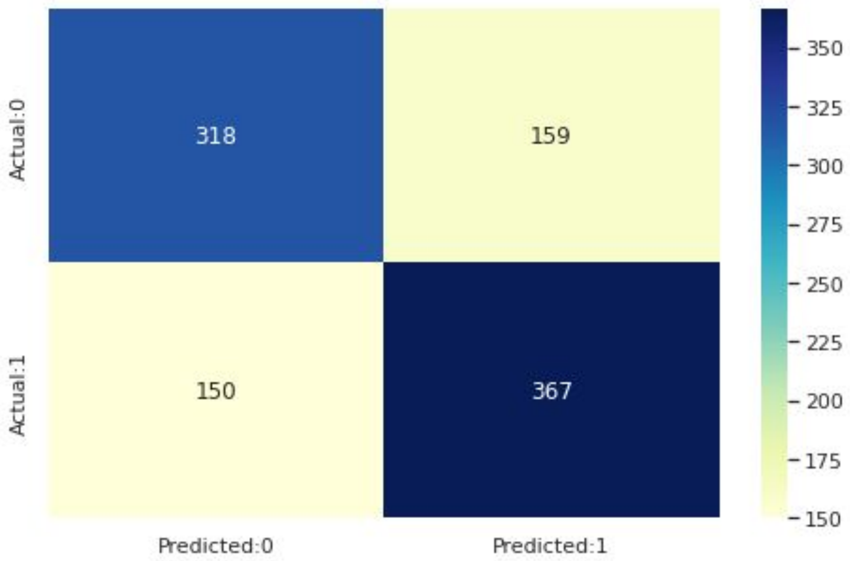
'n_estimators': 80}



# 3. XGBoost :

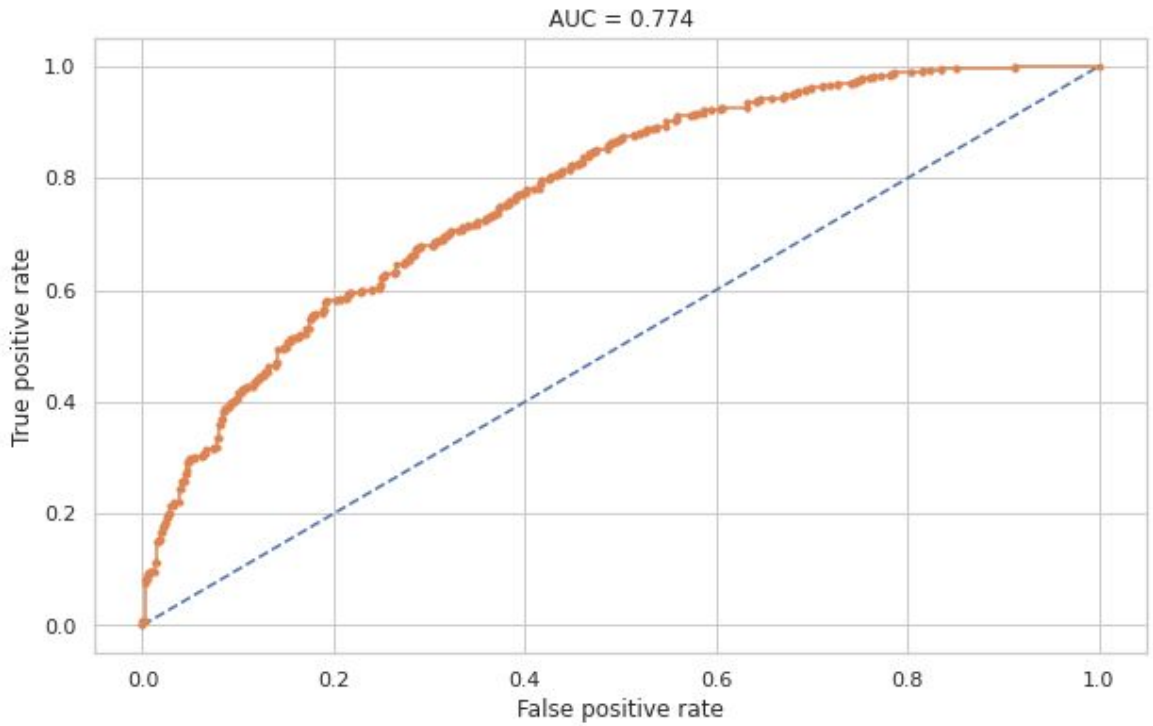


```
Classification Report For This model is as follows
              precision    recall  f1-score   support

           0       0.68      0.67      0.67       477
           1       0.70      0.71      0.70       517

    accuracy                           0.69       994
   macro avg       0.69      0.69      0.69       994
weighted avg       0.69      0.69      0.69       994
```

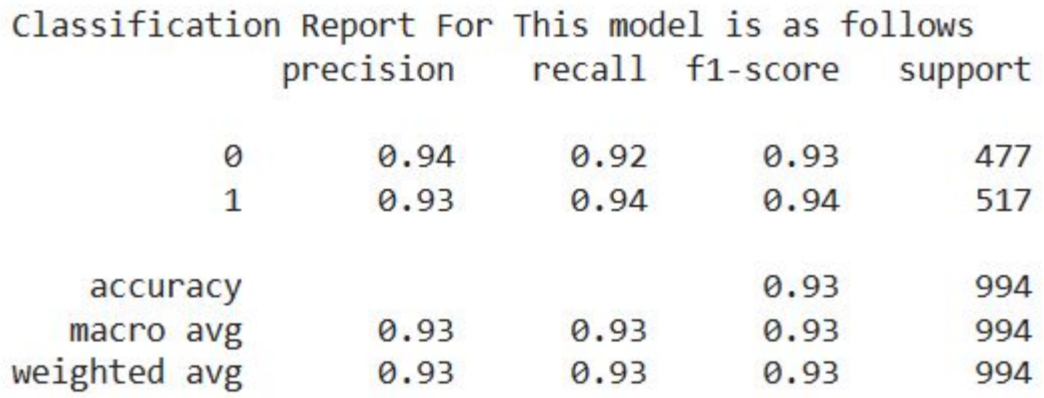


**Best parameters:**

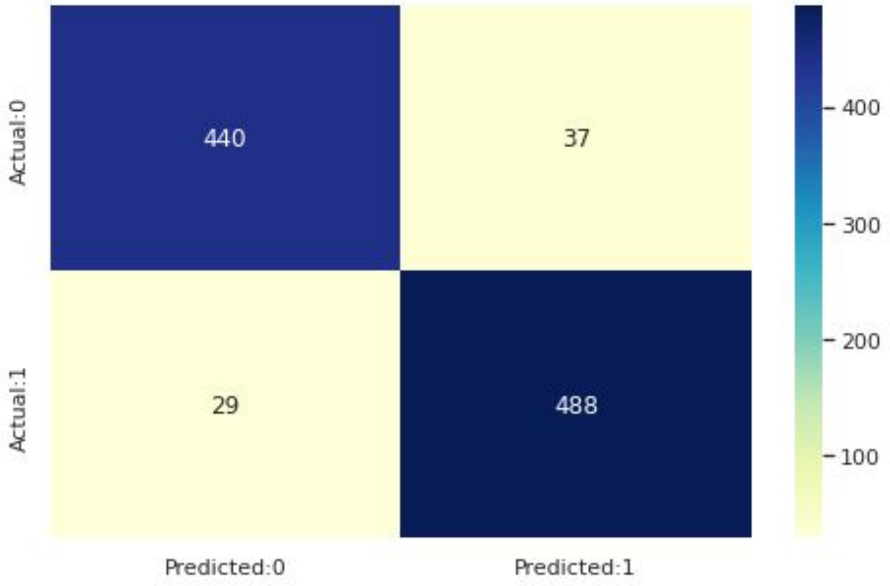
{'learning_rate': 0.1,
'max_depth': 11,
'n_estimators': 200}



# 4. Support Vector Machine :

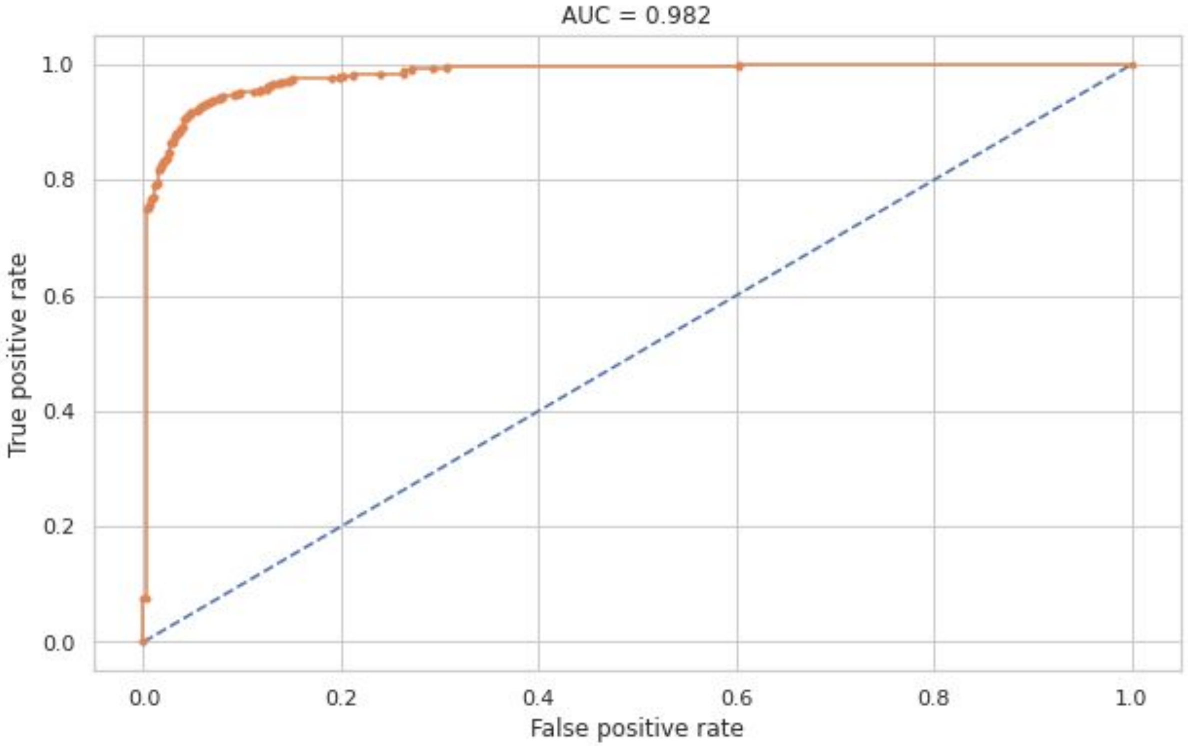


```
Classification Report For This model is as follows
              precision    recall  f1-score   support

           0       0.94      0.92      0.93       477
           1       0.93      0.94      0.94       517

    accuracy                           0.93       994
   macro avg       0.93      0.93      0.93       994
weighted avg       0.93      0.93      0.93       994
```

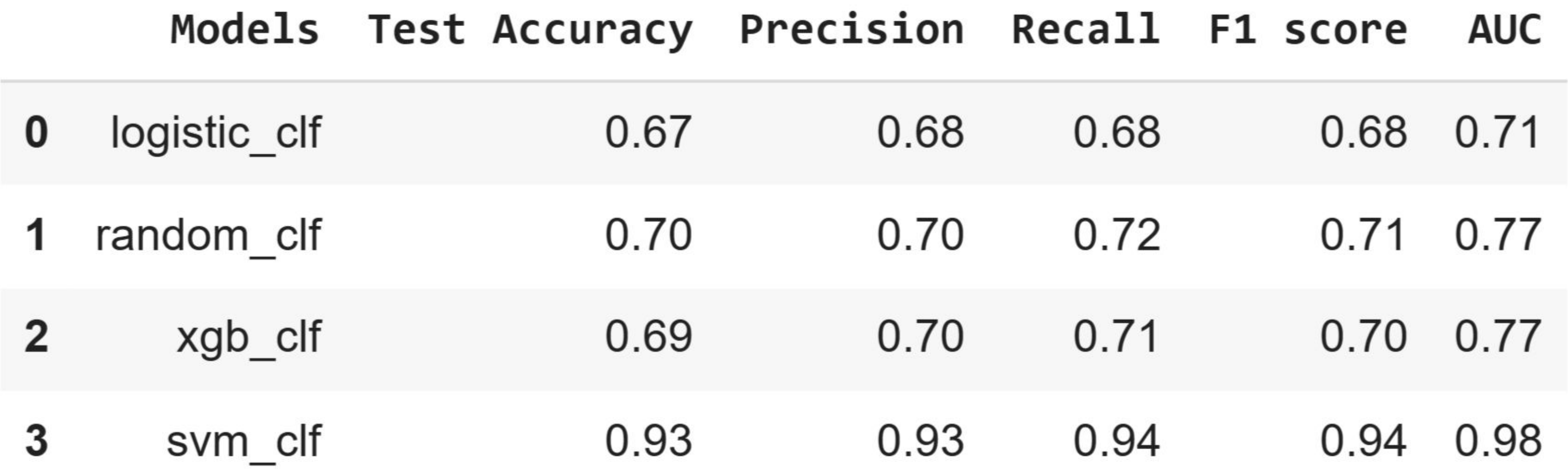


## Best parameters:

{'learning_rate': 0.1,
'max_depth': 11,
'n_estimators': 200}



# Comparison of Models Performance Metrics:



|   | Models | Test Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|---|
| **0** | logistic_clf | 0.67 | 0.68 | 0.68 | 0.68 | 0.71 |
| **1** | random_clf | 0.70 | 0.70 | 0.72 | 0.71 | 0.77 |
| **2** | xgb_clf | 0.69 | 0.70 | 0.71 | 0.70 | 0.77 |
| **3** | svm_clf | 0.93 | 0.93 | 0.94 | 0.94 | 0.98 |

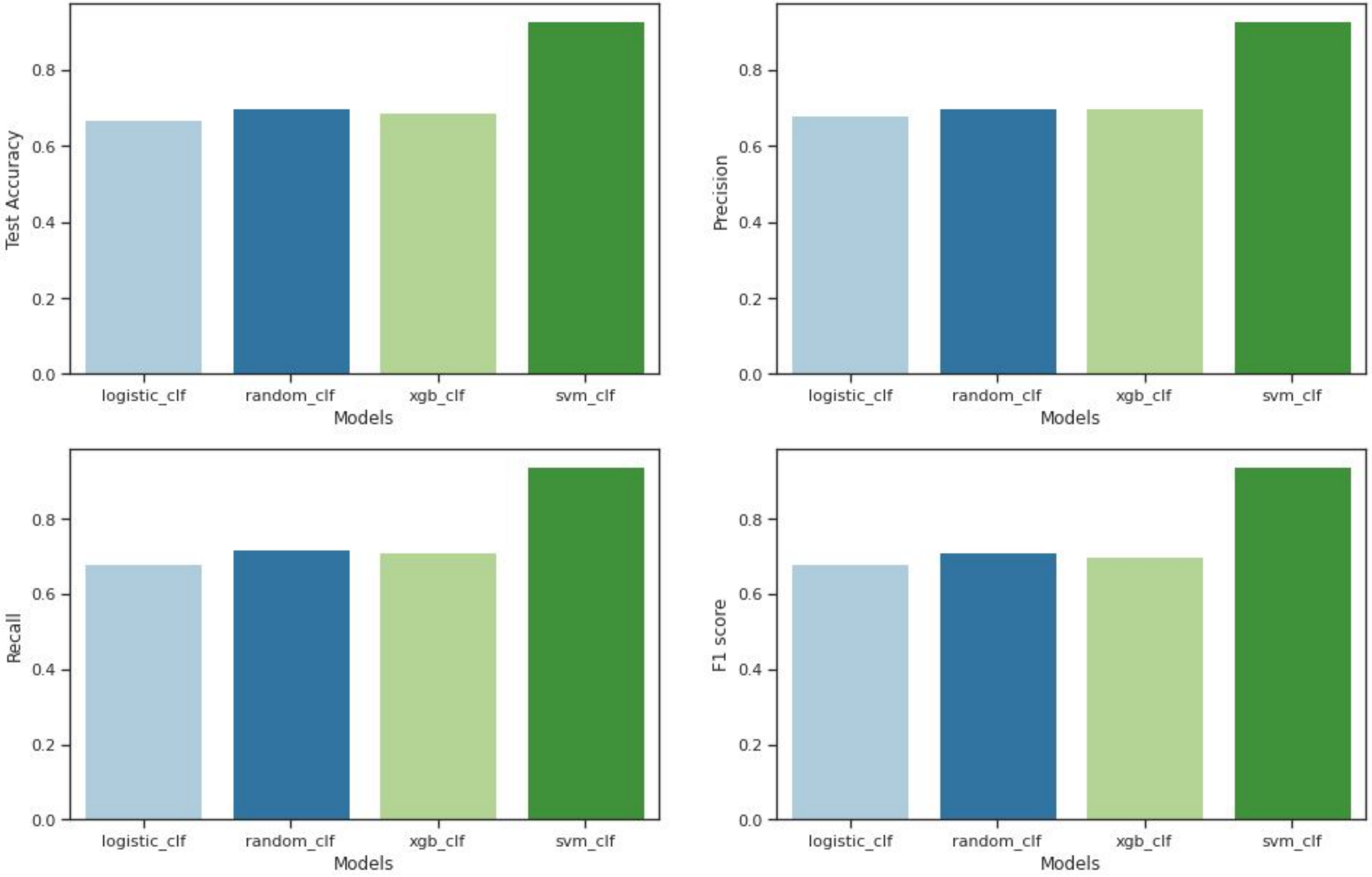
## **Observation from above table**:

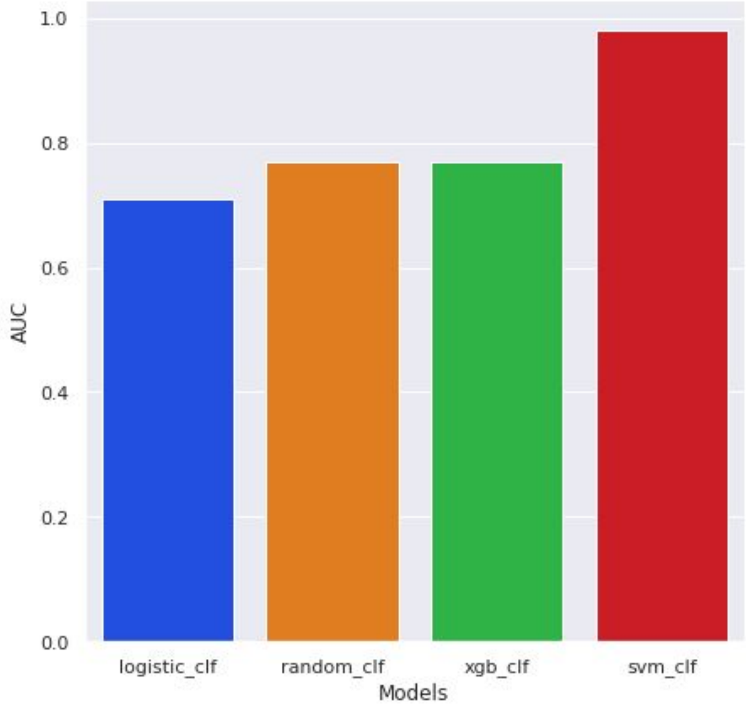
**XGBoost, Support vector machine** gives highest Accuracy, Recall, Precision and AUC score.

Highest recall and AUC score 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.

# Plotting Accuracy score with respect to each models :

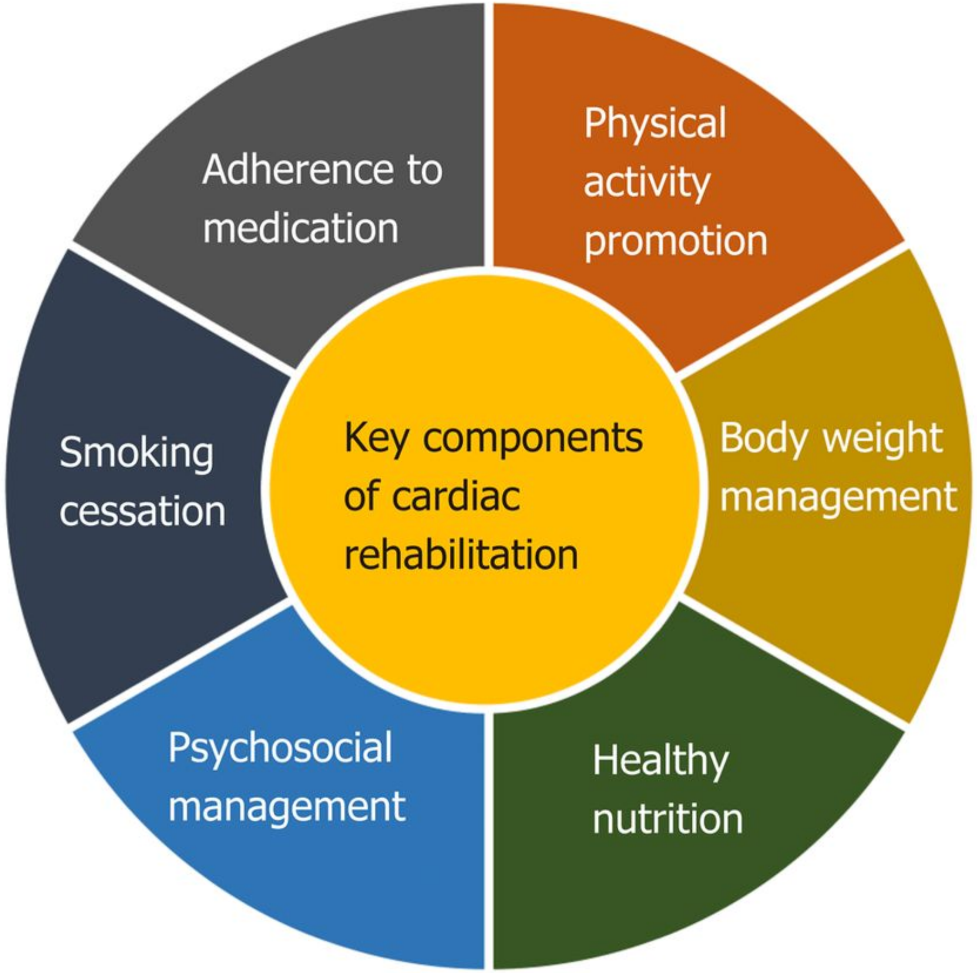




- From graphs we can say that the best performing model is **Support Vector Machine** algorithm.

## Challenges :

- Handling the missing values.
- Making data more accurate.
- Selection of important features.



# Conclusion :

- **Risk** of **Cardiovascular heart disease** is almost **equal** between the **smokers and non-smokers** and same goes with **gender** it is pretty much **same** for both ***male*** and ***females***.
- **Correlation** obtained is **very poor** for this dataset but still due to **tuned parameters** and **strong** classification **algorithms** model **efficiency** obtained is about **93%**.
- The top **contributing features** in predicting the **ten year risk** of developing Cardiovascular Heart Disease are **'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'**.
- 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**.
- **Balancing** the dataset by using the **SMOTE technique** helped in **improving** the **models' sensitivity**.
- With **more data** & with more correlated features **better models** can be built.