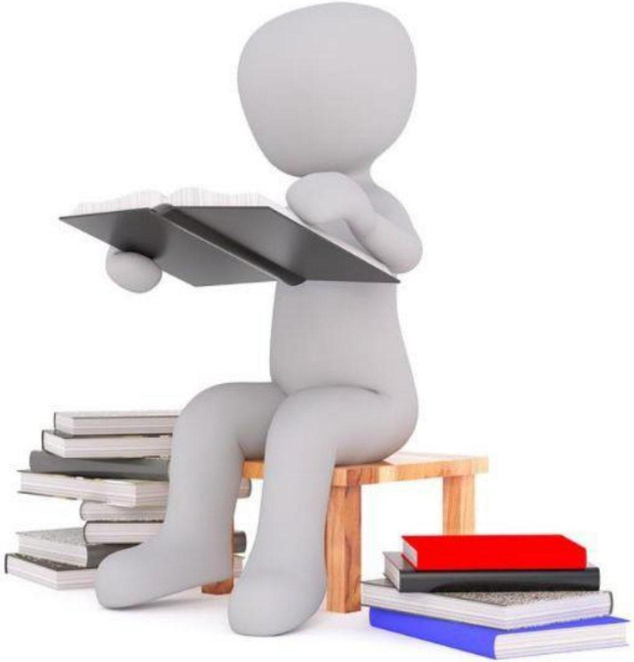
Capstone Project-3
Cardiovascular Risk Prediction

Member

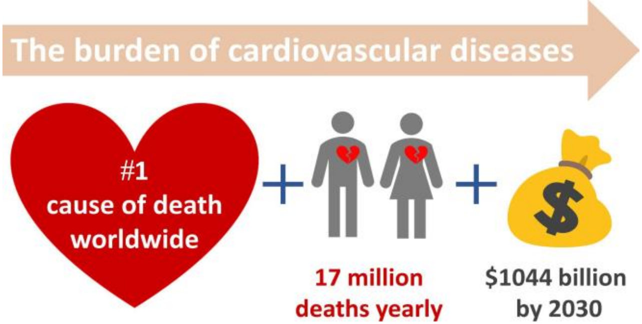
Sumit Kumar Dhibar

Contents



Abstract

- Heart disease is the major cause of morbidity and mortality globally, it accounts for more deaths annually than any other cause.



- According to WHO, an estimated 17.9 million people died from CVDs in 2019, accounting for 32% of all global fatalities.

- The World Heart Federation has estimated that by 2030, the total global cost of CVD treatment will increase from approximately USD 863 billion in 2010 to a staggering USD 1,044 billion.

- Though CVDs cannot be treated, predicting the risk of the disease and taking the necessary precautions and medications can help to avoid severe symptoms and in some cases, even death.

Problem Statement

- A heart attack happens when the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen. If blood flow isn't restored quickly, the section of heart muscle begins to die.

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

- Our goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD) based on their present health conditions using different Machine Learning Techniques.

Data Summary

Independent Variables

Categorical Column

- Education
- Sex
- Is_Smoking
- BP_Meds
- Prevalent Hypertension
- Prevalent Stroke
- Diabetes

Continuous Column

- Age
- Cigs_Per_Day
- Total Cholesterol
- Sys BP
- Dia BP
- BMI
- Heart Rate
- Glucose

Target Variable

10 Year CHD

Data Summary

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

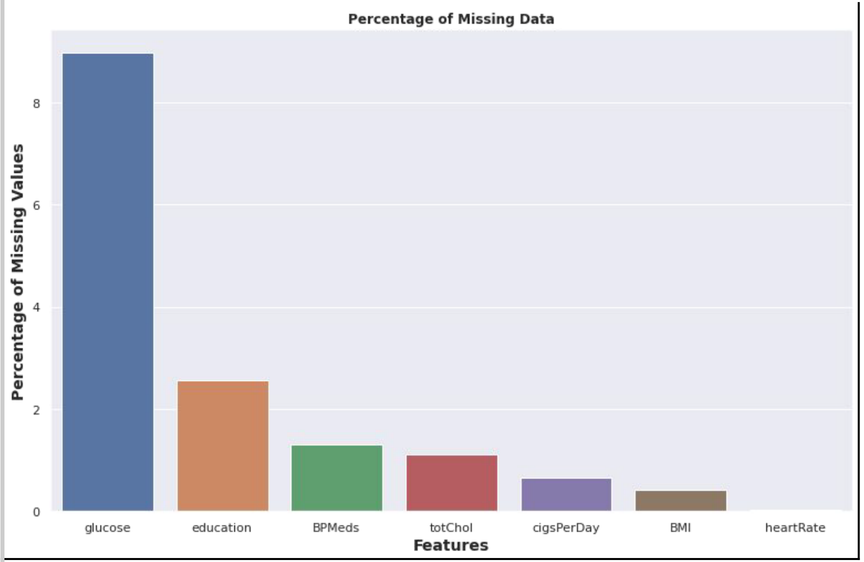
- **This Dataset contains 3390 rows and 17 columns.**

- **The Target variable namely ‘Ten Year CHD’ refers to whether the patient suffers from coronary heart disease depending upon the values of current medical parameters.**

- **The dependent variable consists of the binary values where, 1 - Risk of Coronary Heart disease and 0 - No risk of Coronary Heart Disease.**

Handling Missing Values

Null Values



	Total No of Missing Values	Percentage of Missing Values
glucose	304	8.97
education	87	2.57
BPMeds	44	1.30
totChol	38	1.12
cigsPerDay	22	0.65
BMI	14	0.41
heartRate	1	0.03

Handling Missing Values

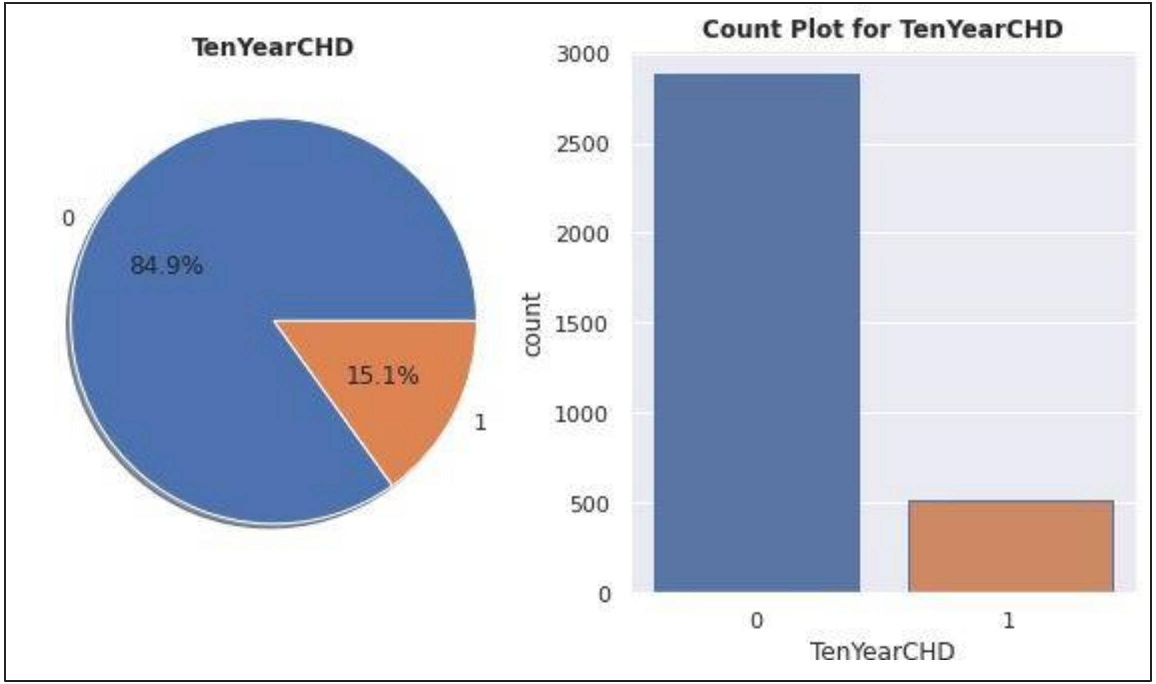
After Handling Missing Values

```
id                 0
age                0
education          0
sex                0
is_smoking         0
cigsPerDay         0
BPMeds             0
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol            0
sysBP              0
diaBP              0
BMI                0
heartRate          0
glucose            0
TenYearCHD         0
dtype: int64
```

Exploratory Data Analysis

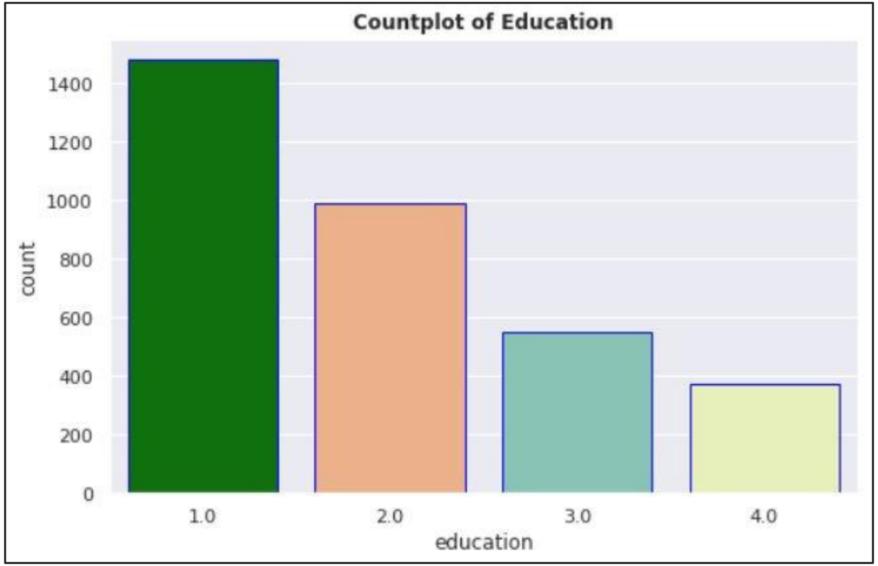
10 Year CHD - Dependent Variable

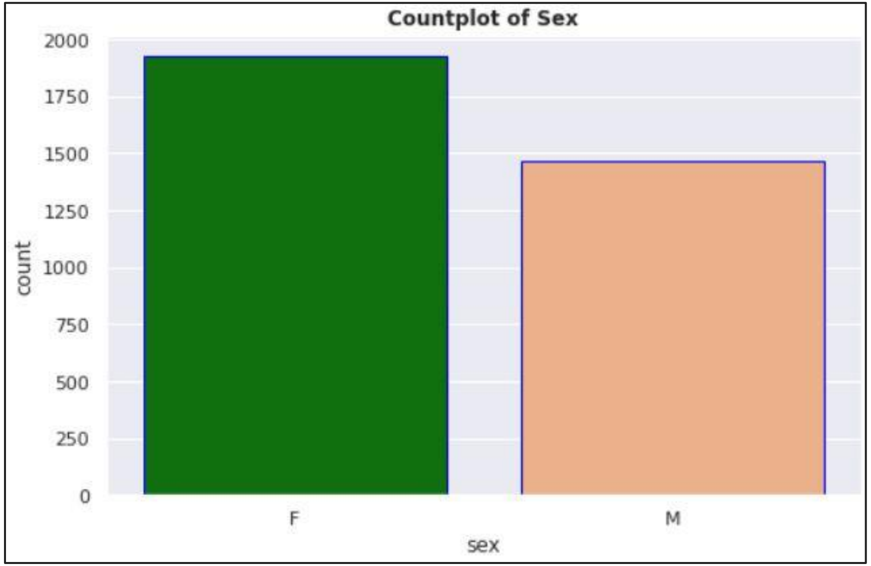
The dependent variable is imbalanced with just ~15% of patients testing positive for CHD.



EDA (Univariate Analysis) - Categorical Variables

Education & Sex

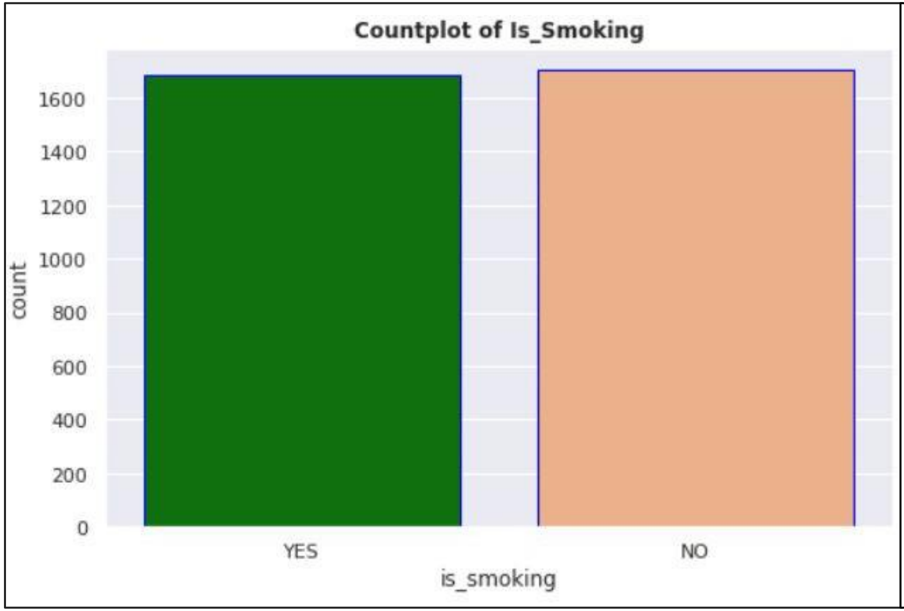


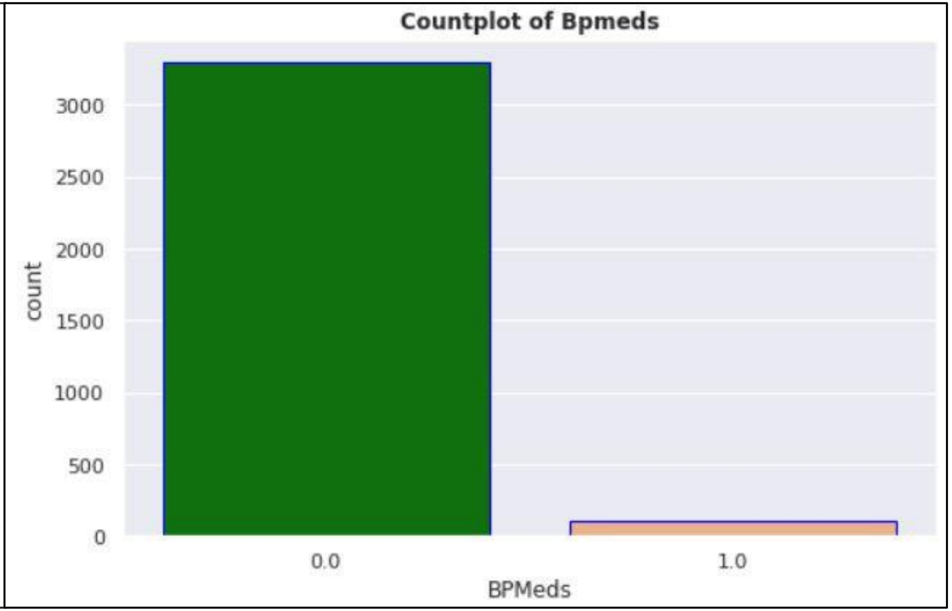


- **Most patients have education level 1.**
- **There are more Female patients than Male.**

EDA (Univariate Analysis) - Categorical Variables

Is Smoking & BP Meds

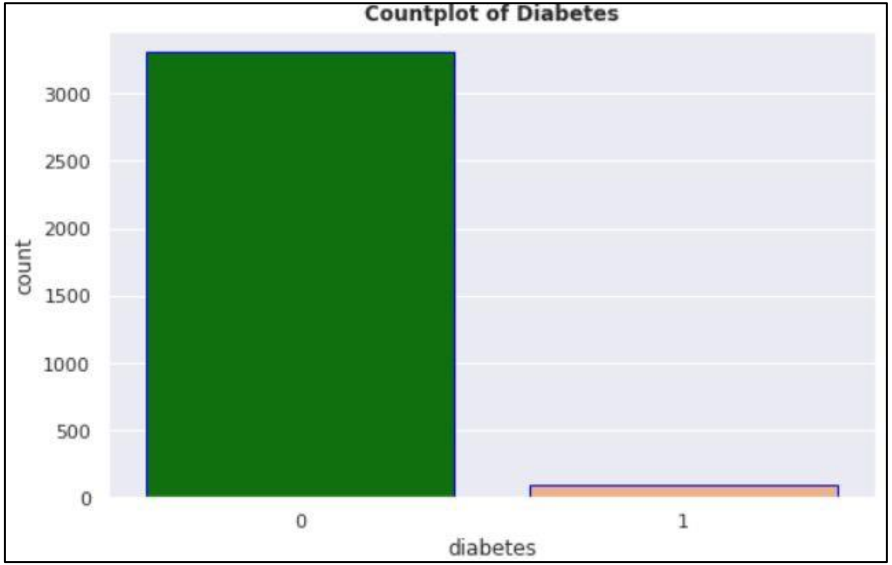


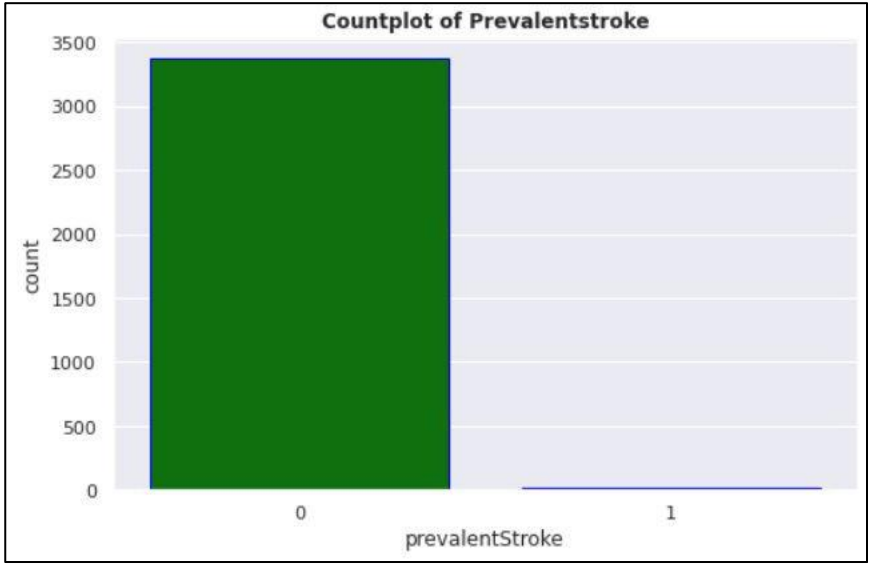


- Half of the patients are smokers
- There are very few individual who are using blood pressure medicine.

EDA (Univariate Analysis) - Categorical Variables

Diabetes & Prevalent Stroke



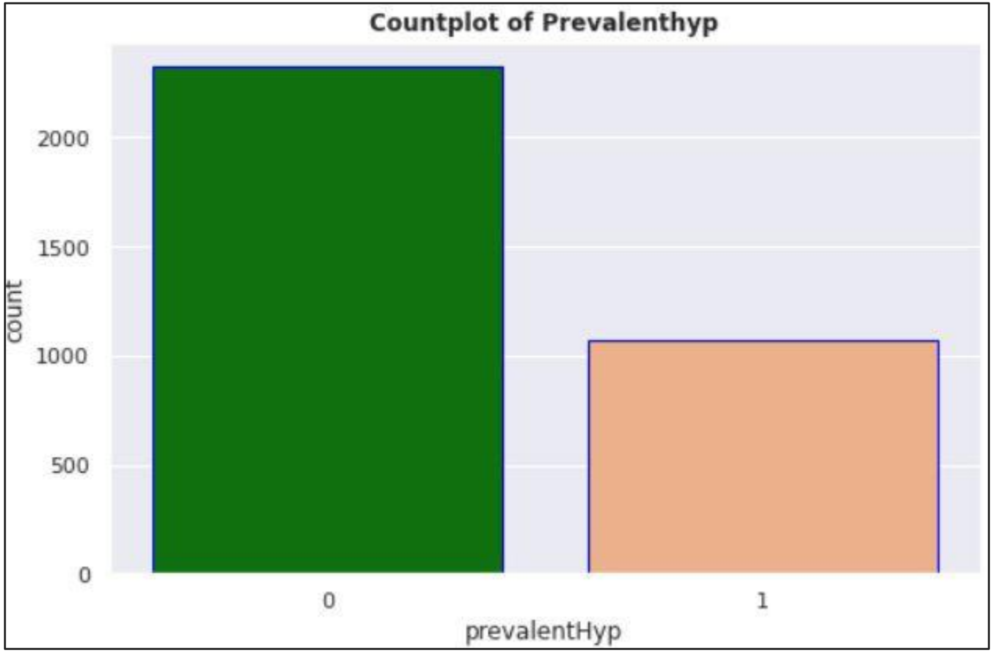


- There are very few individual who have had previously stroke or Diabetes.

EDA (Univariate Analysis) - Categorical Variables

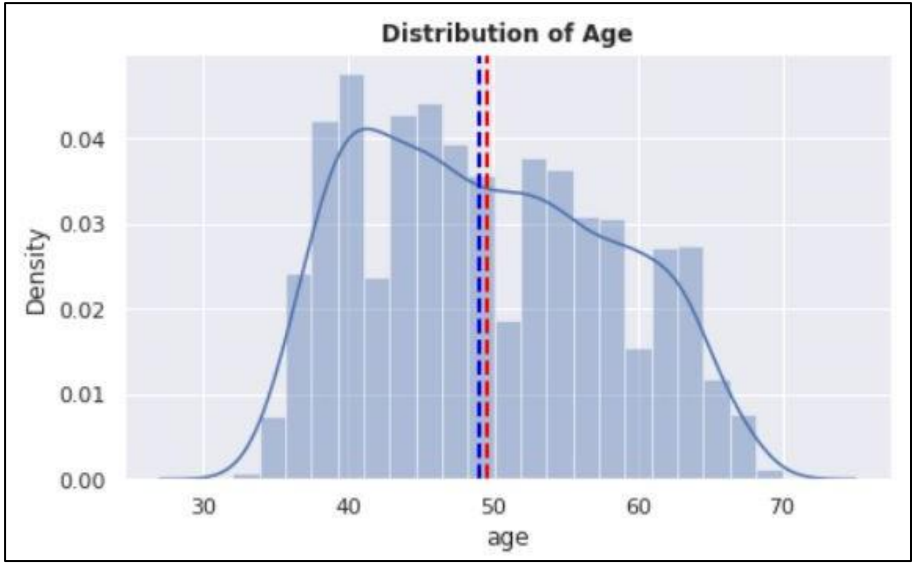
Prevalent Hypertension

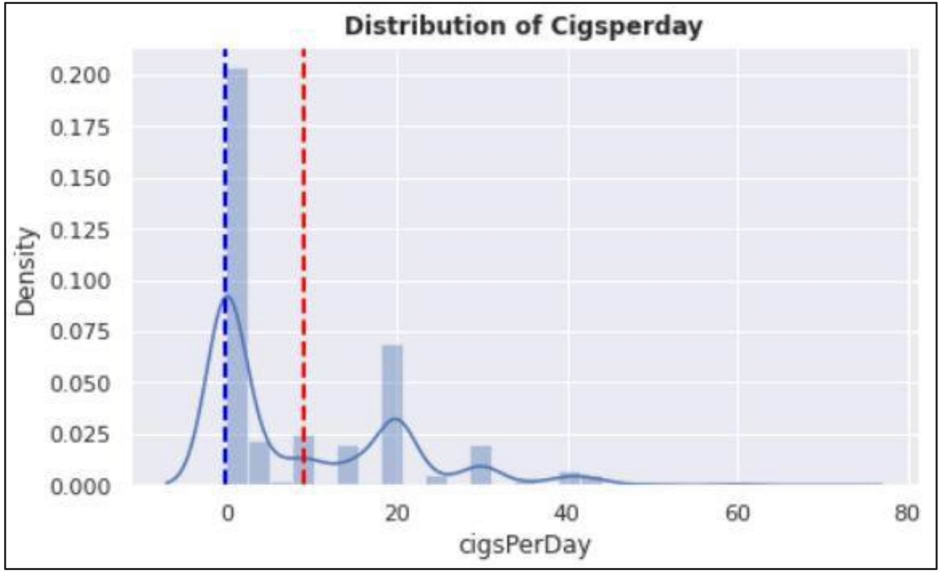
- Almost 30% individual has prevalent hypertension.



EDA (Univariate Analysis) - Continuous Variables

Age & Cigs_per_day

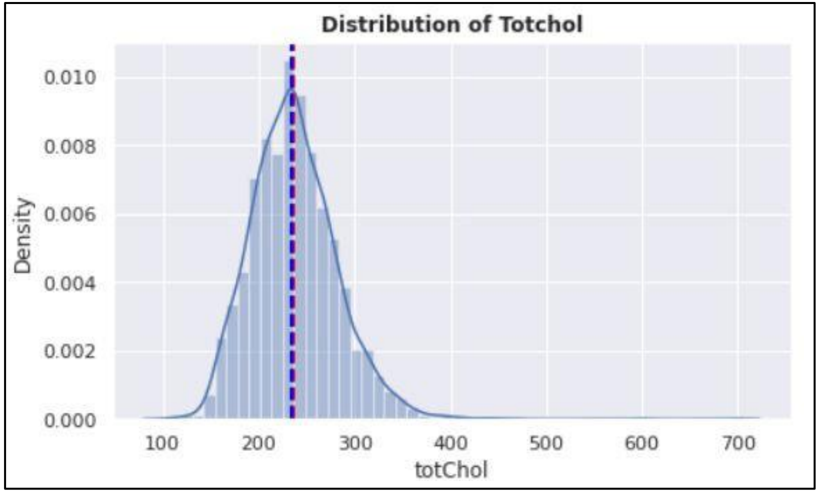


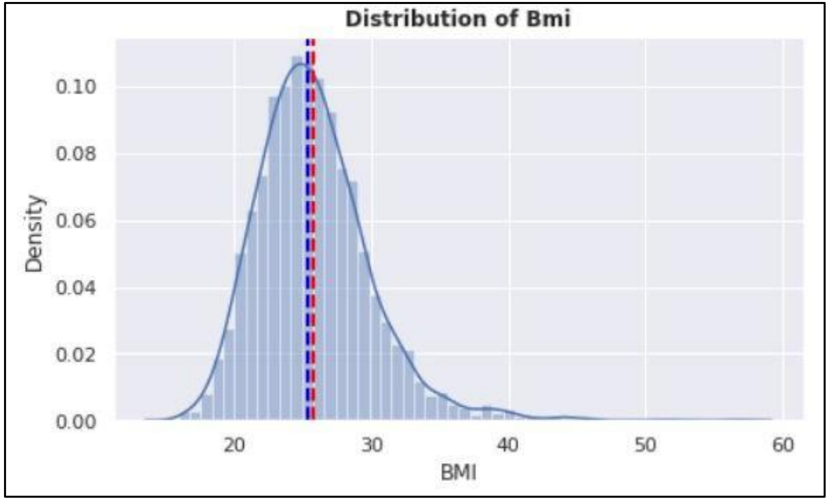


- ❑ Most of the people are around 40 to 50 years old.
- ❑ Most of the patient smoke less than 10 Cigarette a day

EDA (Univariate Analysis) - Continuous Variables

Total Cholesterol & BMI



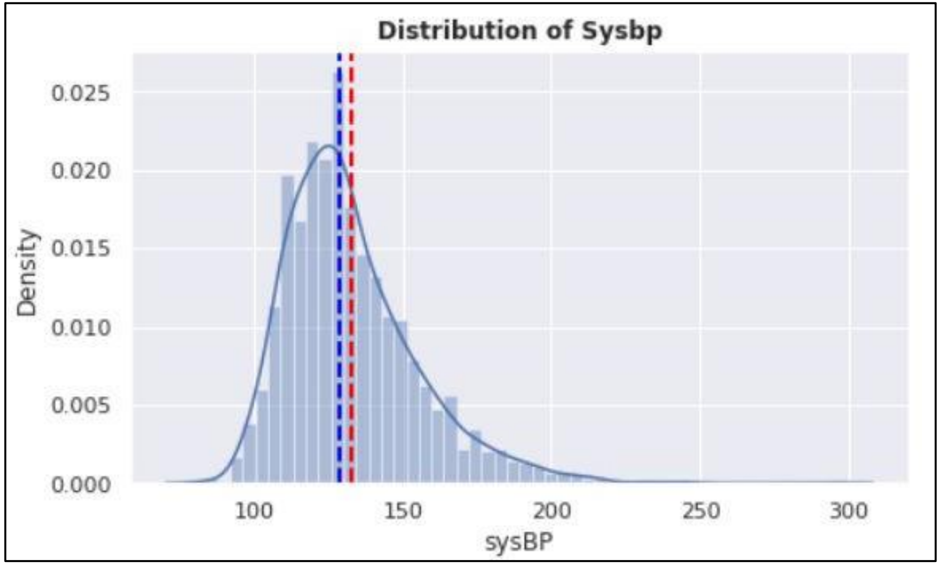


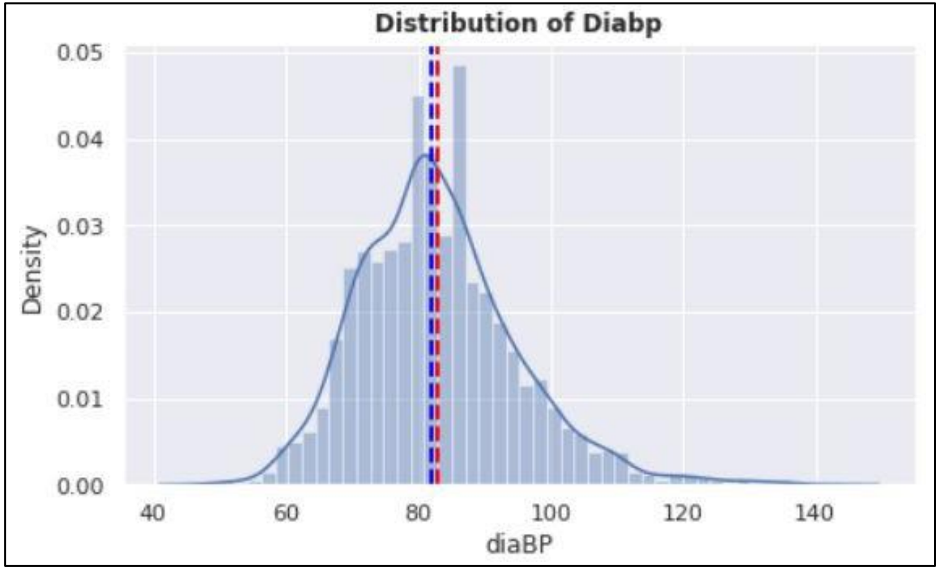
- ❑ Cholesterol range is 200 to 250 , which is borderline high level.

- ❑ Most of the patient are in healthy weight and overweight range.

EDA (Univariate Analysis) - Continuous Variables

Sys BP & Dia BP

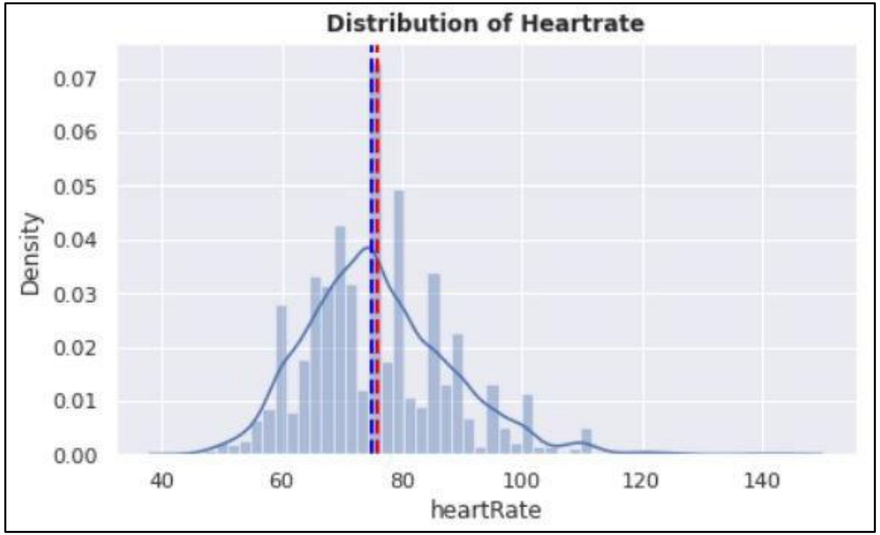


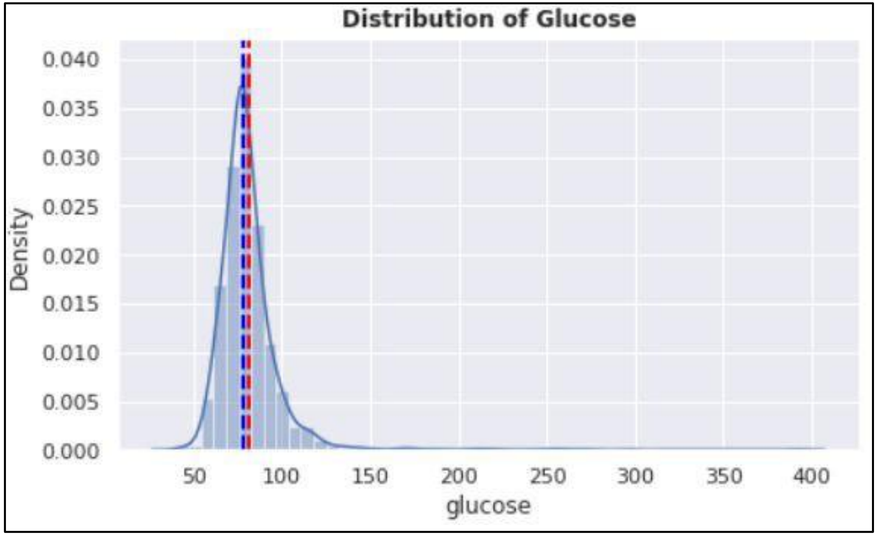


- Systolic blood pressure and Diastolic blood pressure are in normal range

EDA (Univariate Analysis) - Continuous Variables

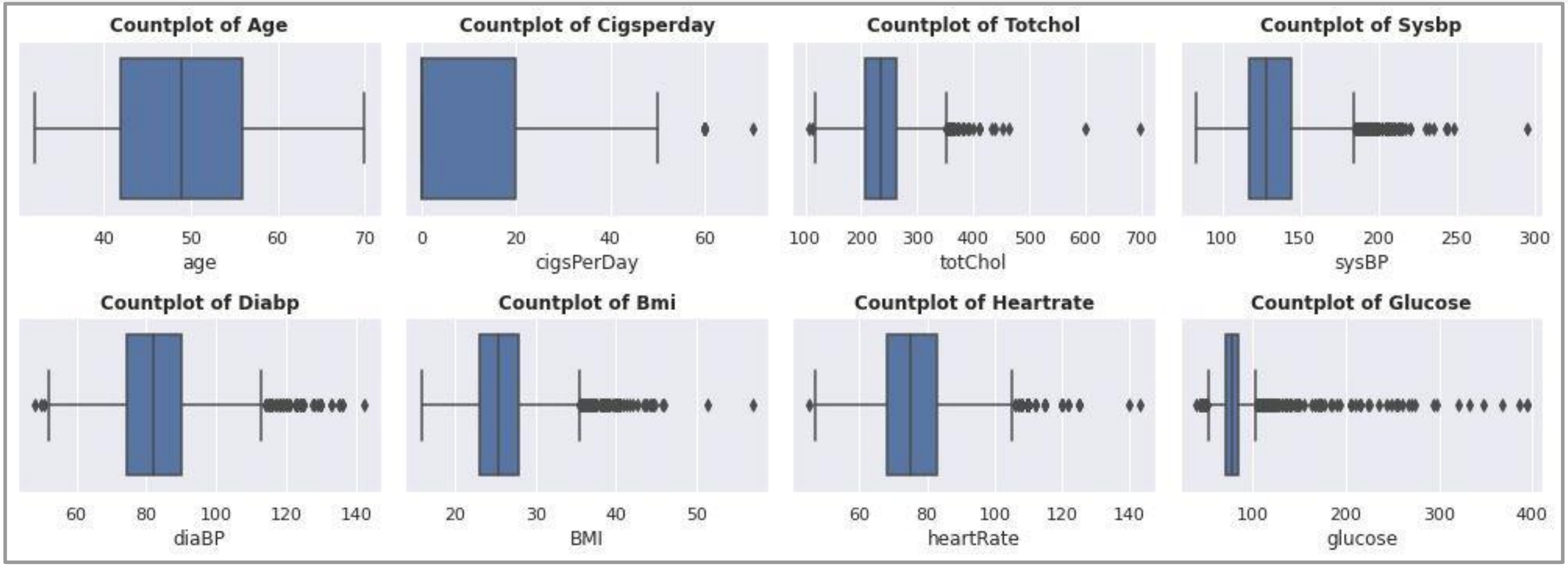
Heart Rate & Glucose





- ❑ **Heart rate of patients are in normal range.**
- ❑ **Glucose level is also in normal range.**

Outliers Analysis

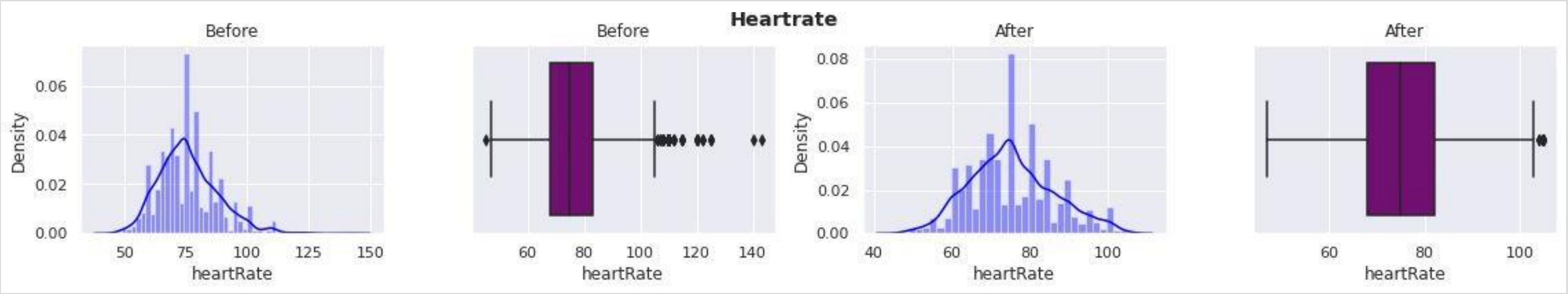


- It can be clearly seen from above plot that outliers present in some columns and we treated them by replacing with median value.

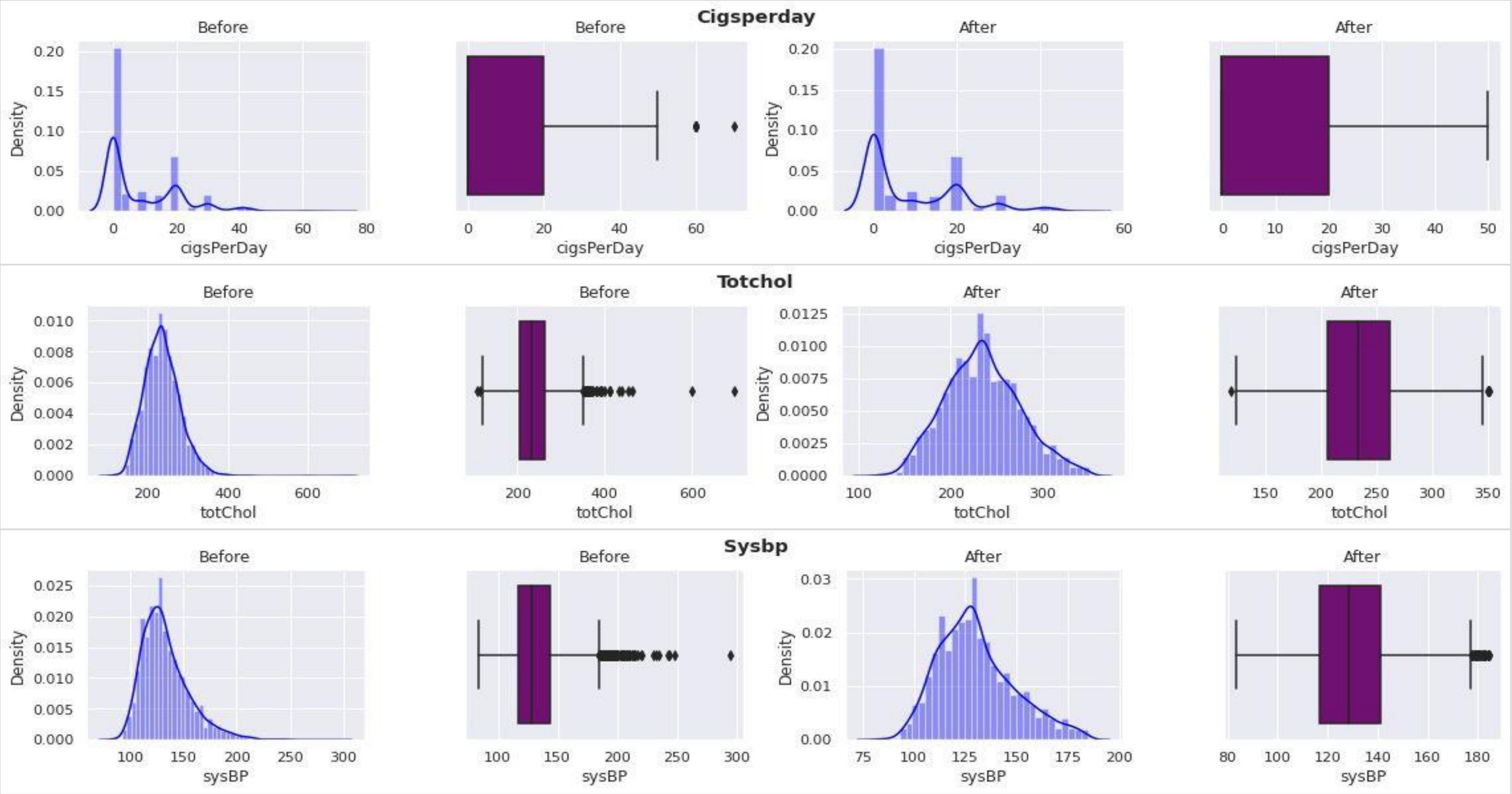
Handling Outliers

- **IQR or Interquartile range is the difference between third quartile (Q3) and first quartile(Q1) . To detect outliers we create a boundary by considering 1.5 times the Interquartile range (IQR) and then subtract this value from Q1 and also we add this value to the Q3. This gives a minimum and maximum range, with this we compare every observation.**

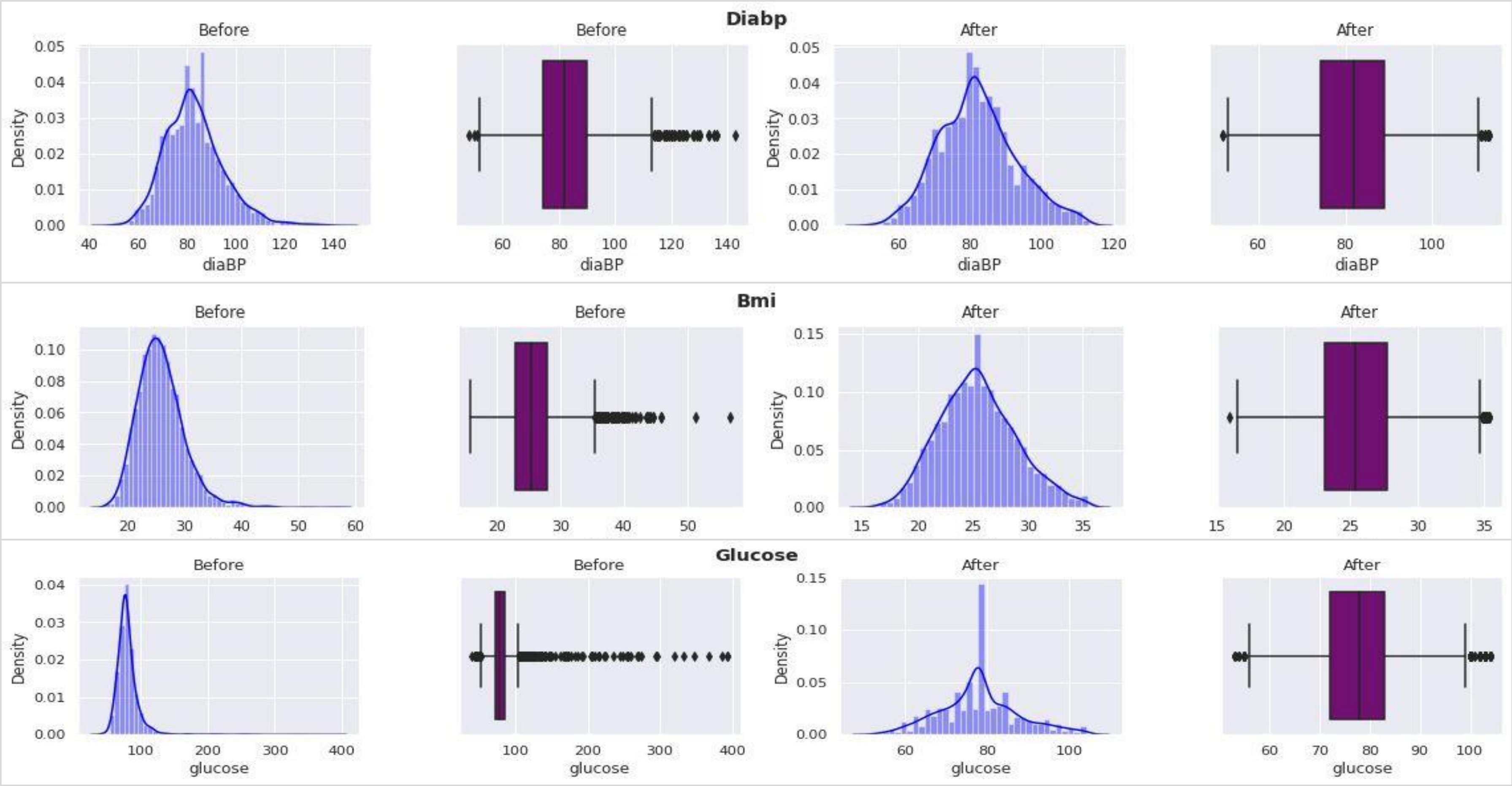
- **Any observation that are more than 1.5 times IQR below Q1 or more than 1.5 times IQR above Q3 are considered outliers. We replaced the outliers with median values (50th percentile) of that column.**



Handling Outliers

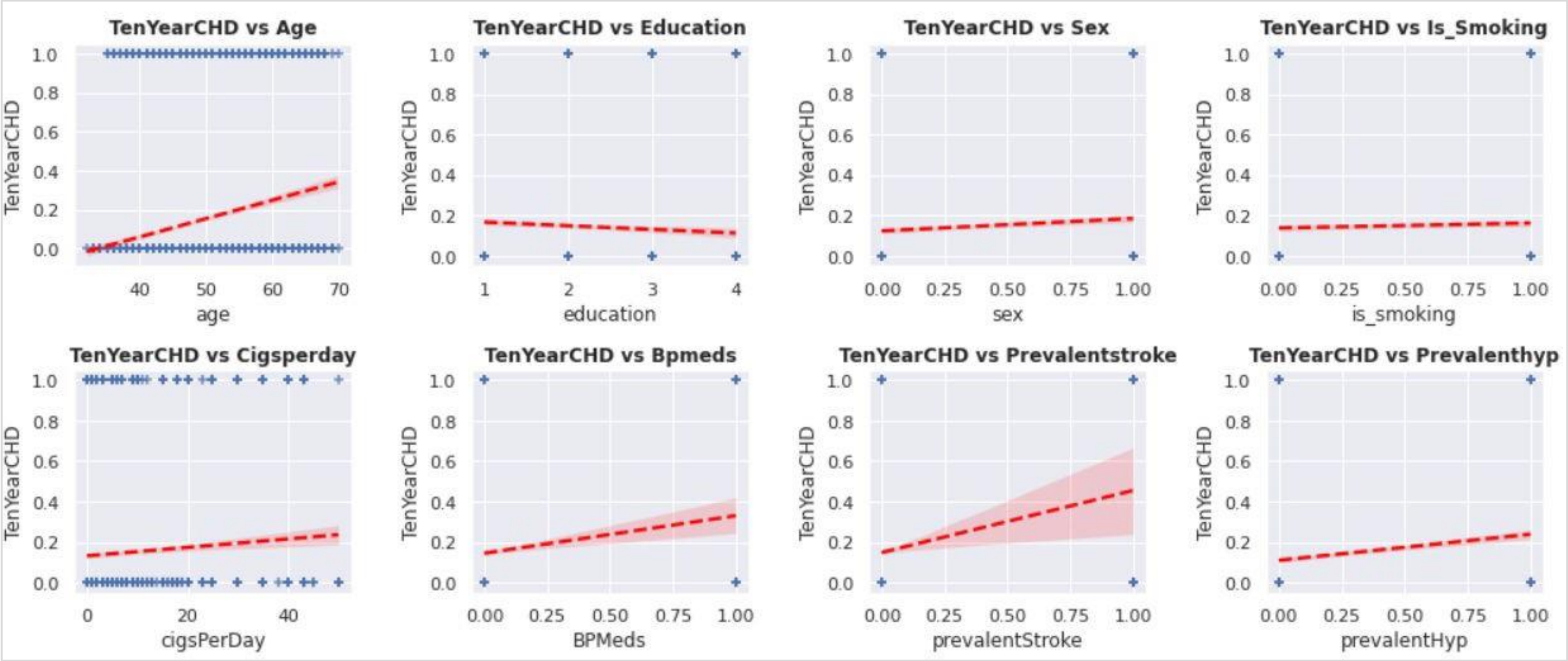


Handling Outliers

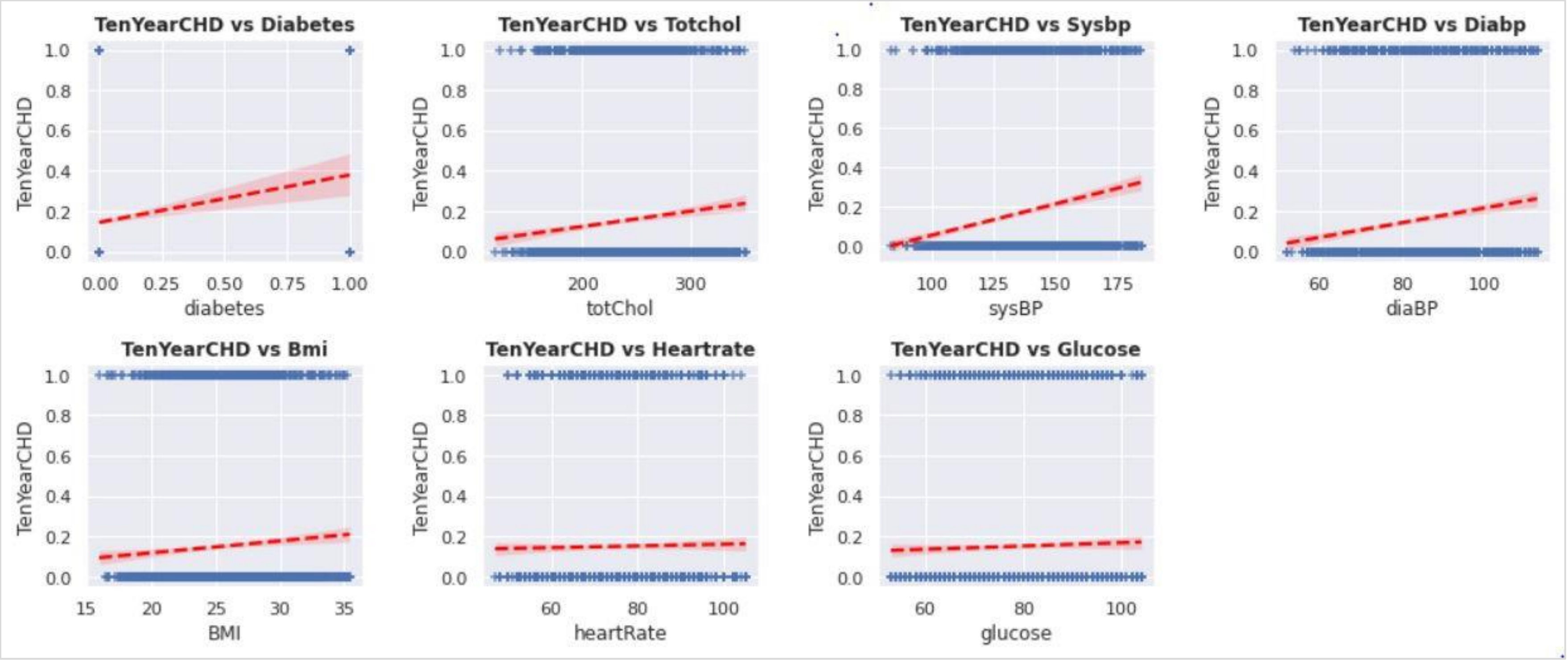


EDA (Bivariate Analysis)

- In Bivariate analysis we are visualizing the relation between dependent variable and independent variable.

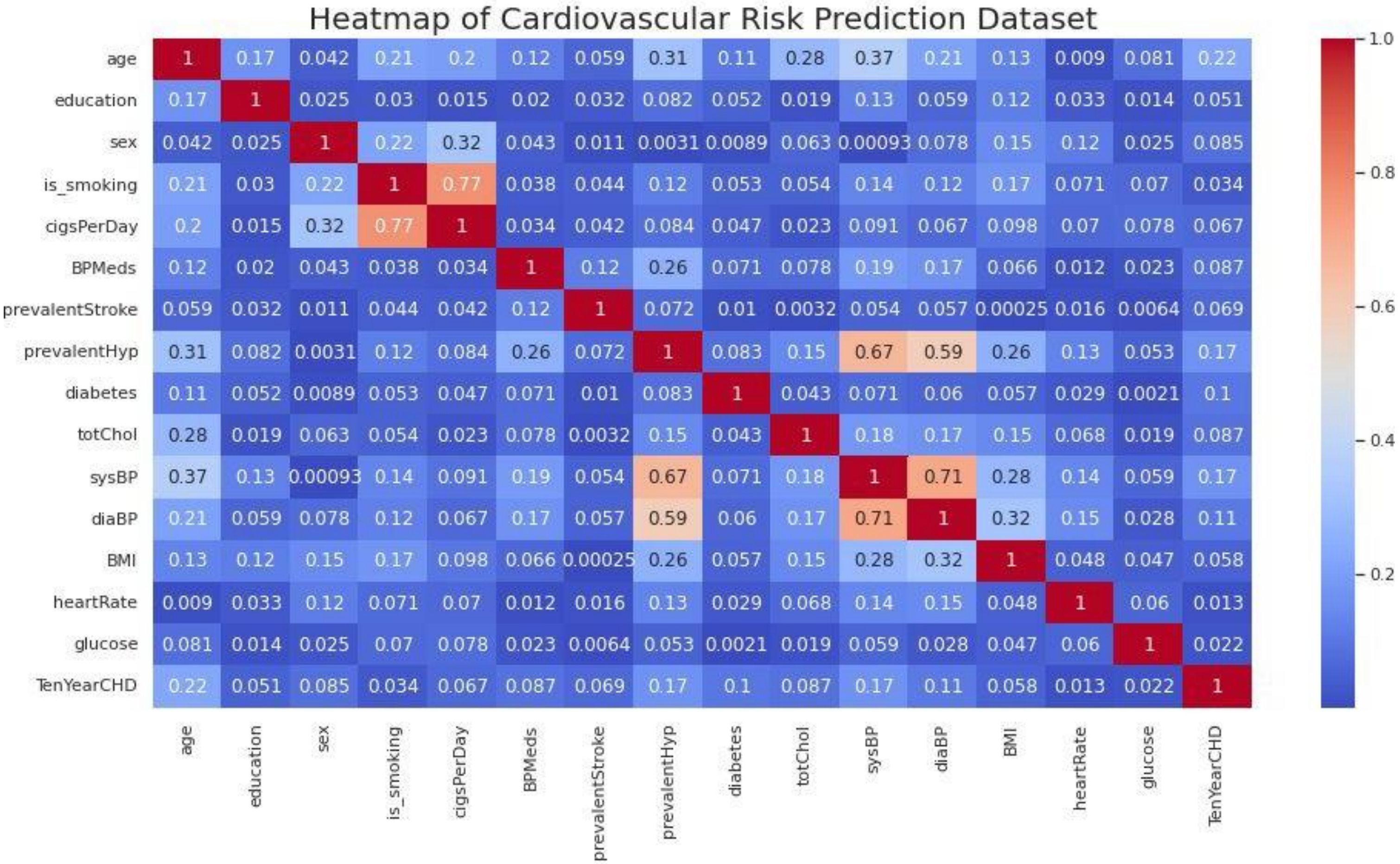


EDA (Bivariate Analysis)



- It is clear from above plots that ‘Age’, ‘Cigs_per_day’ , ‘Total Cholesterol’, ‘Sys BP’, ‘Dia BP’, ‘BMI’ These are having positive relation with the dependent variable.

Multicollinearity Analysis

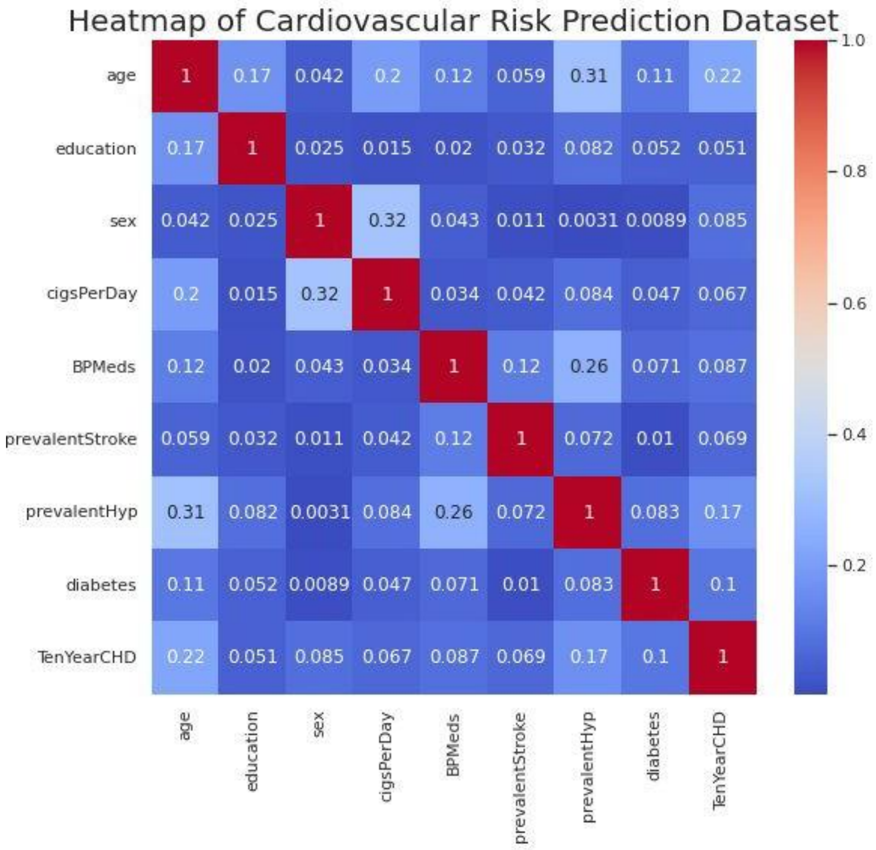


VIF Analysis

	variables	VIF
10	sysBP	132.655302
11	diaBP	127.212212
12	BMI	58.866609
14	glucose	55.671761
13	heartRate	47.789265
0	age	42.772276
9	totChol	37.653008
3	is_smoking	4.954371
1	education	4.649621
4	cigsPerDay	4.194075
7	prevalentHyp	2.357598
2	sex	2.147585
5	BPMeds	1.128250
8	diabetes	1.047244
6	prevalentStroke	1.026716

	variables	VIF
0	age	5.381373
1	education	3.965520
2	sex	1.966446
3	cigsPerDay	1.734502
6	prevalentHyp	1.685134
4	BPMeds	1.120404
7	diabetes	1.044744
5	prevalentStroke	1.024797

Updated Heatmap



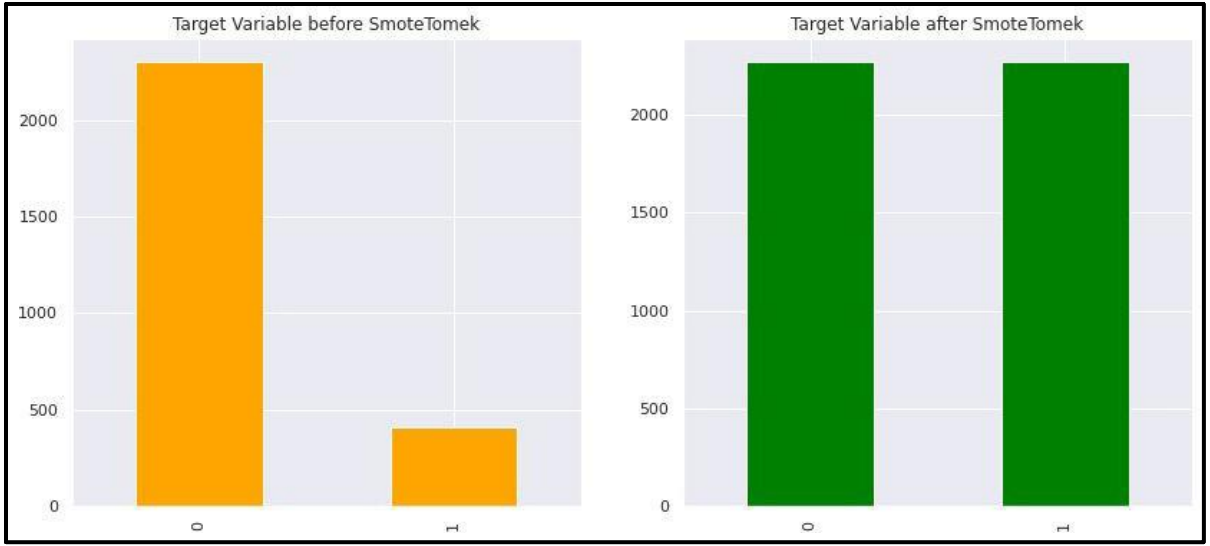
Machine Learning models

Pre - processing

- Defining X and Y variables, splitting the data in 80 - 20 ratio as train and test sets.

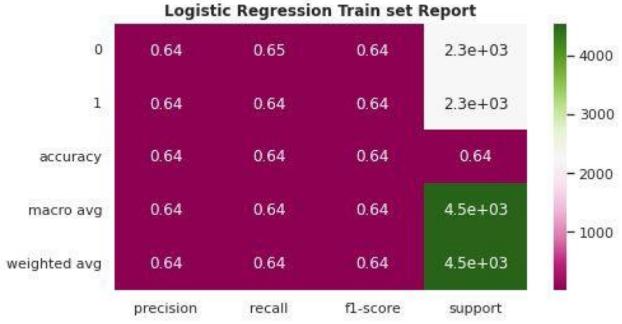
Handling Class Imbalance

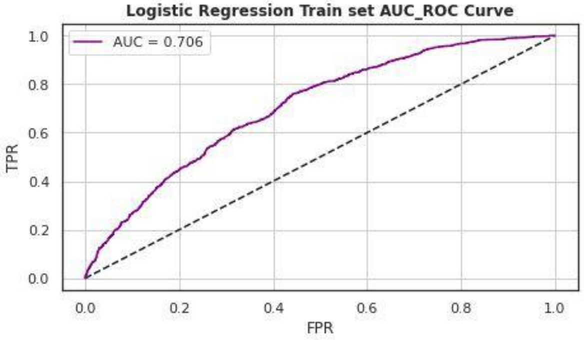
```
before handling class imbalance :
0    2303
1     409
Name: TenYearCHD, dtype: int64


after handling class imbalance :
0    2271
1    2271
Name: TenYearCHD, dtype: int64
```

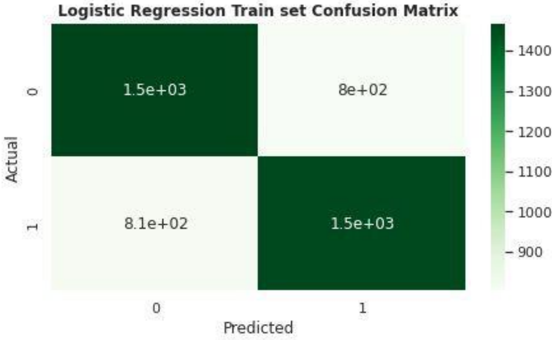


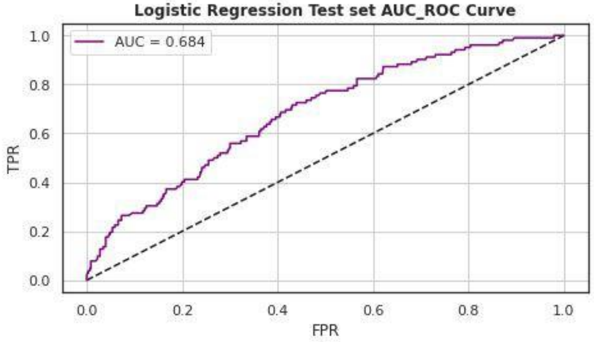
- **Min - Max Scaler** - for scaling the features

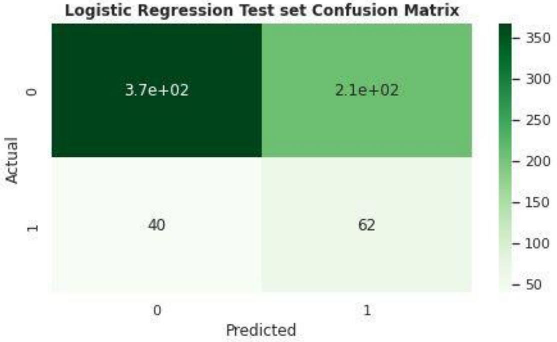
Logistic Regression





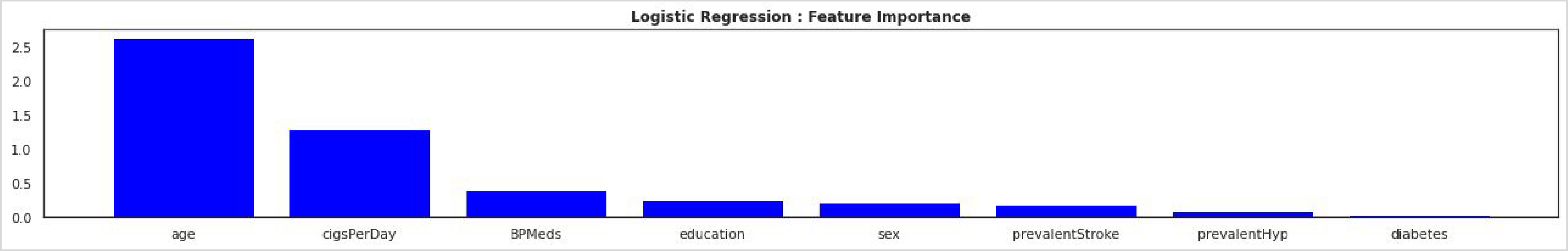






Logistic Regression

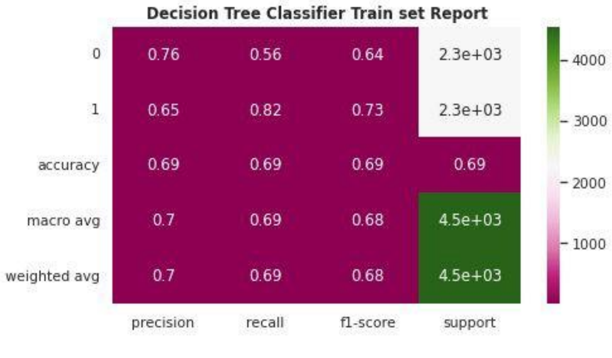
Feature Importance

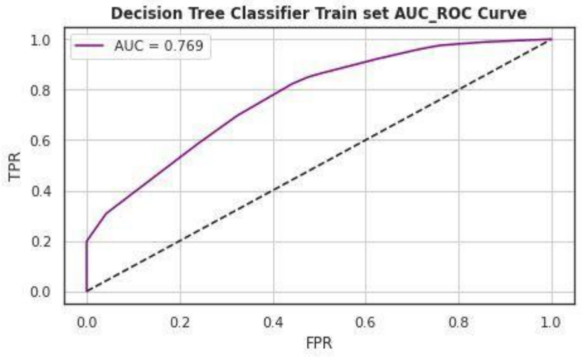


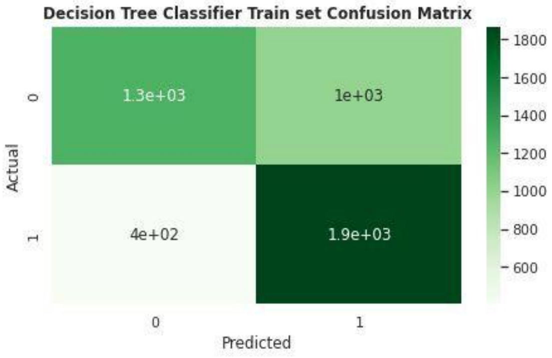
Logistic Regression results for class 1 on Test data :

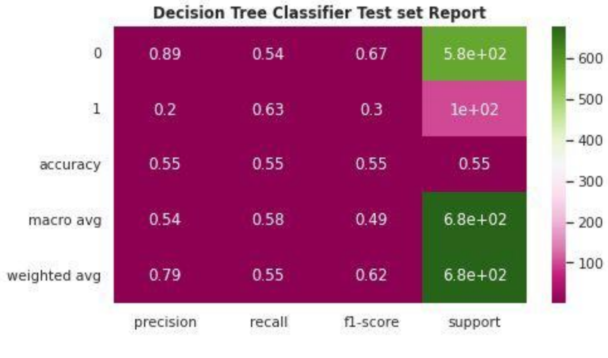
- Precision : 0.23
- Recall : 0.61
- F1 Score : 0.33
- AUC : 0.684

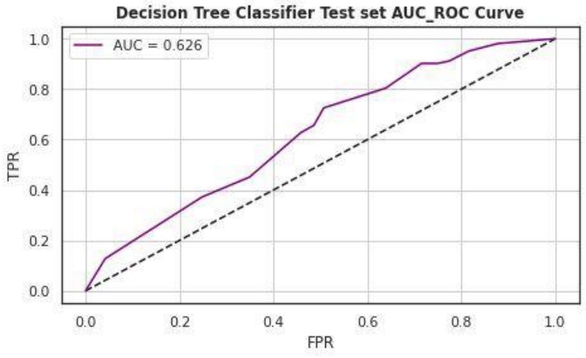
Decision Tree Classifier

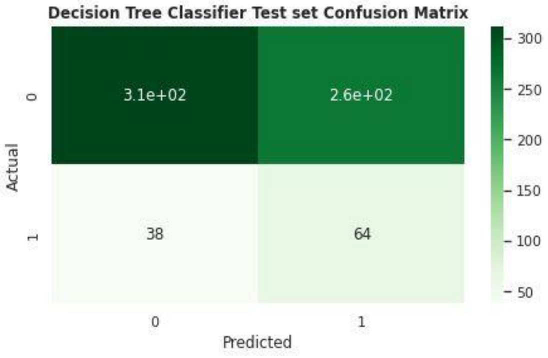






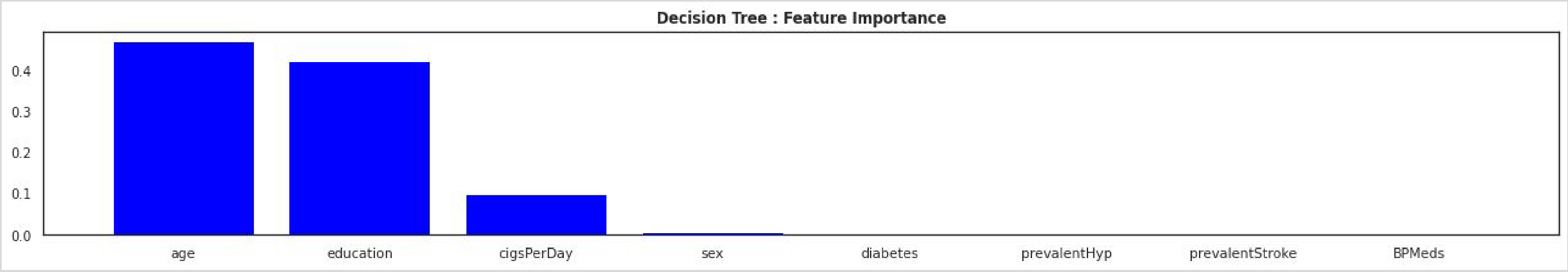






Decision Tree Classifier

Feature Importance

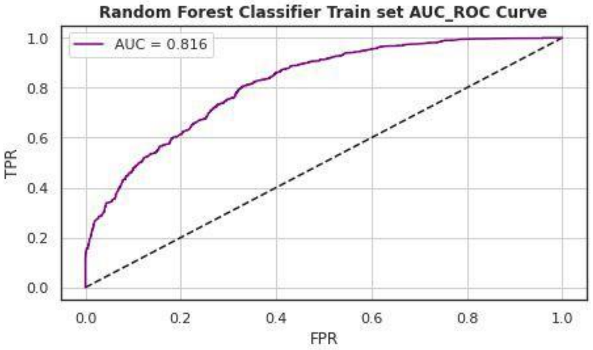


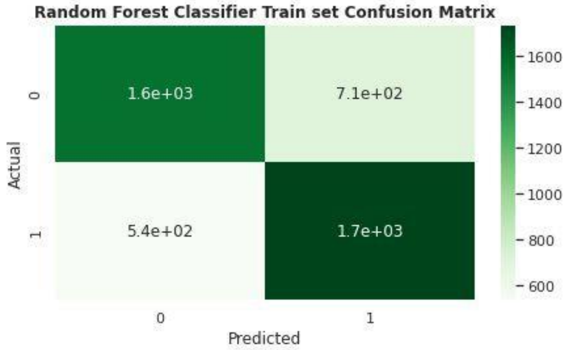
Decision Tree results for class 1 on Test data :

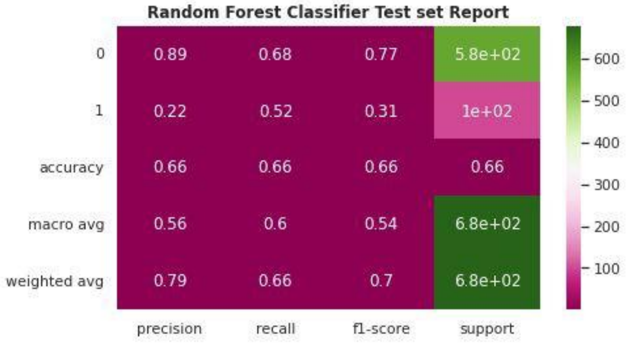
- ❑ Precision : 0.20
- ❑ Recall : 0.63
- ❑ F1 Score : 0.30
- ❑ AUC : 0.626

Random Forest Classifier

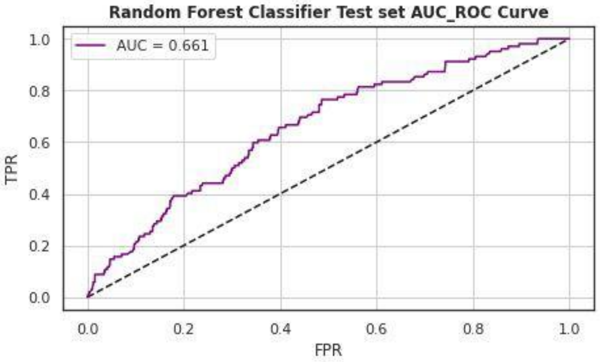
	precision	recall	f1-score	support
0	0.74	0.69	0.71	2.3e+03
1	0.71	0.76	0.74	2.3e+03
accuracy	0.73	0.73	0.73	0.73
macro avg	0.73	0.73	0.73	4.5e+03
weighted avg	0.73	0.73	0.73	4.5e+03

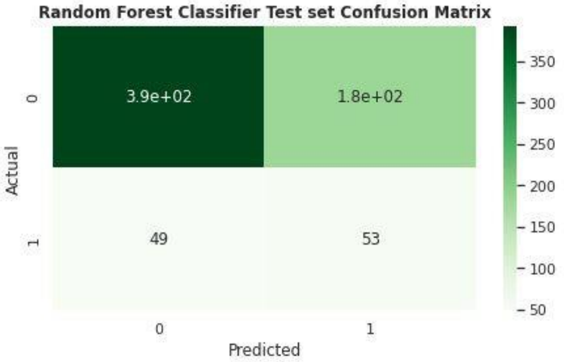






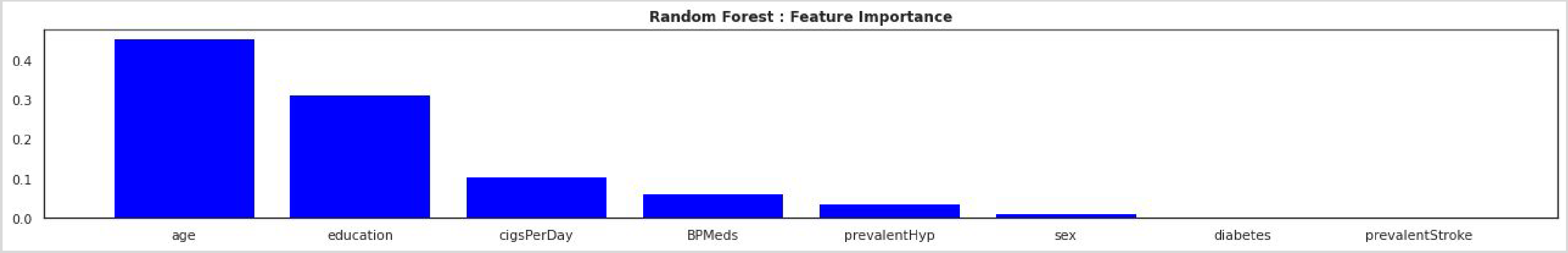
	precision	recall	f1-score	support
0	0.89	0.68	0.77	5.8e+02
1	0.22	0.52	0.31	1e+02
accuracy	0.66	0.66	0.66	0.66
macro avg	0.56	0.6	0.54	6.8e+02
weighted avg	0.79	0.66	0.7	6.8e+02





Random Forest Classifier

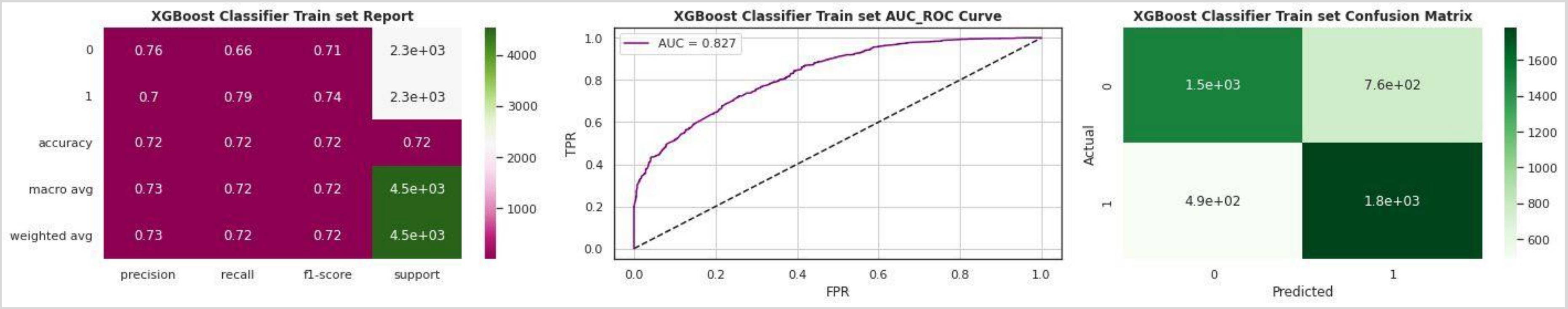
Feature Importance

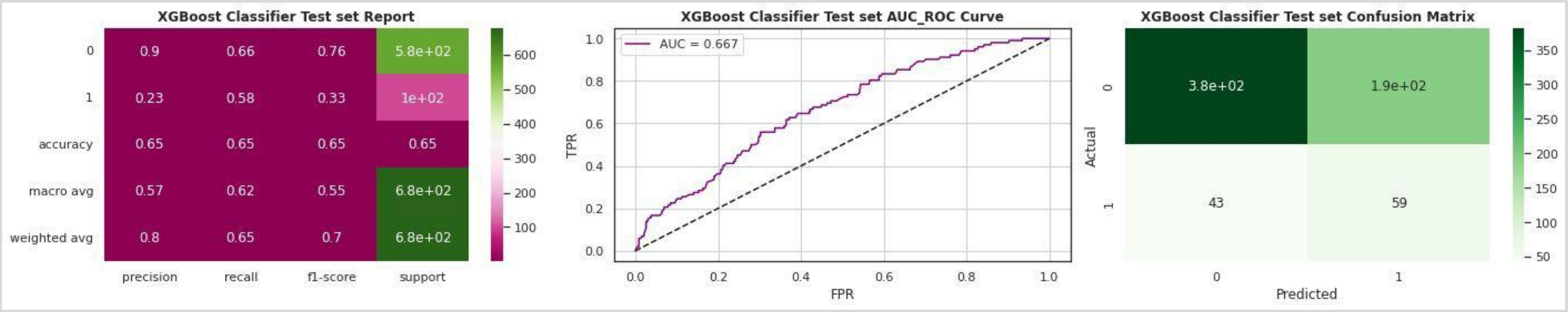


Random Forest results for class 1 on Test data :

- Precision : 0.22
- Recall : 0.52
- F1 Score : 0.31
- AUC : 0.661

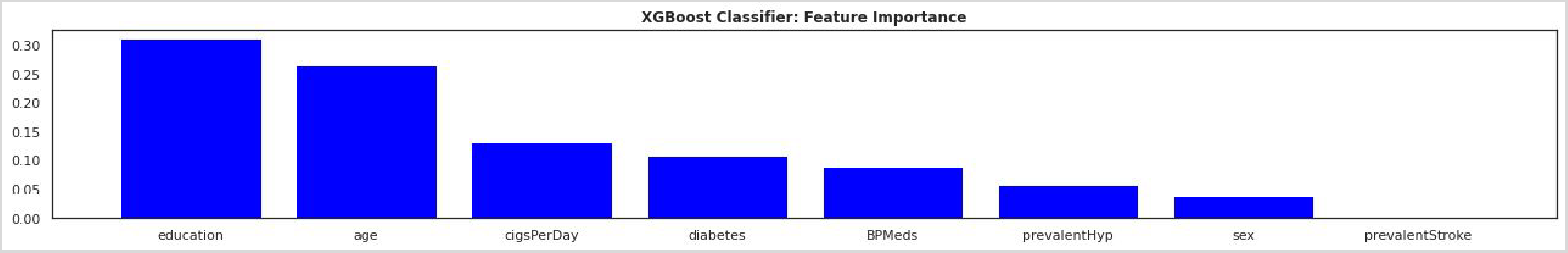
XGBoost Classifier





XGBoost Classifier

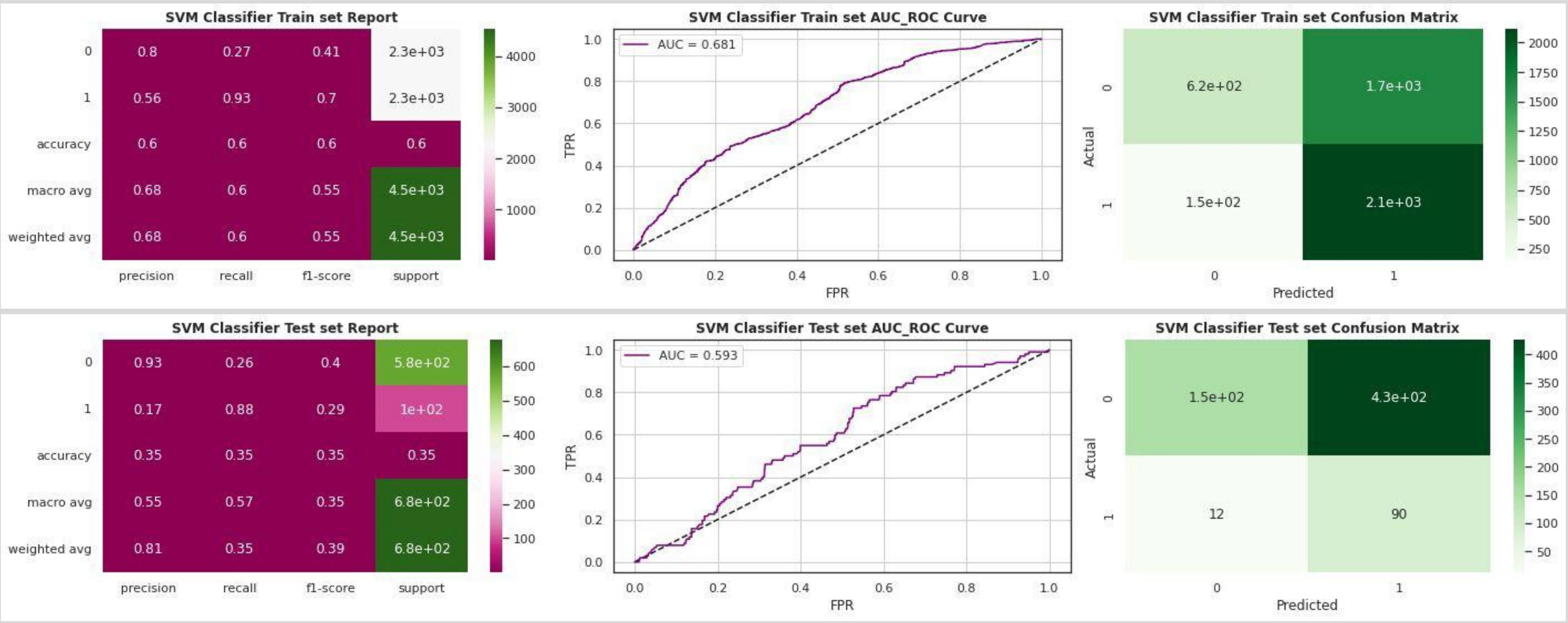
Feature Importance



XGBoost results for class 1 on Test data :

- ❑ Precision : 0.23
- ❑ Recall : 0.58
- ❑ F1 Score : 0.33
- ❑ AUC : 0.667

Support Vector Machine

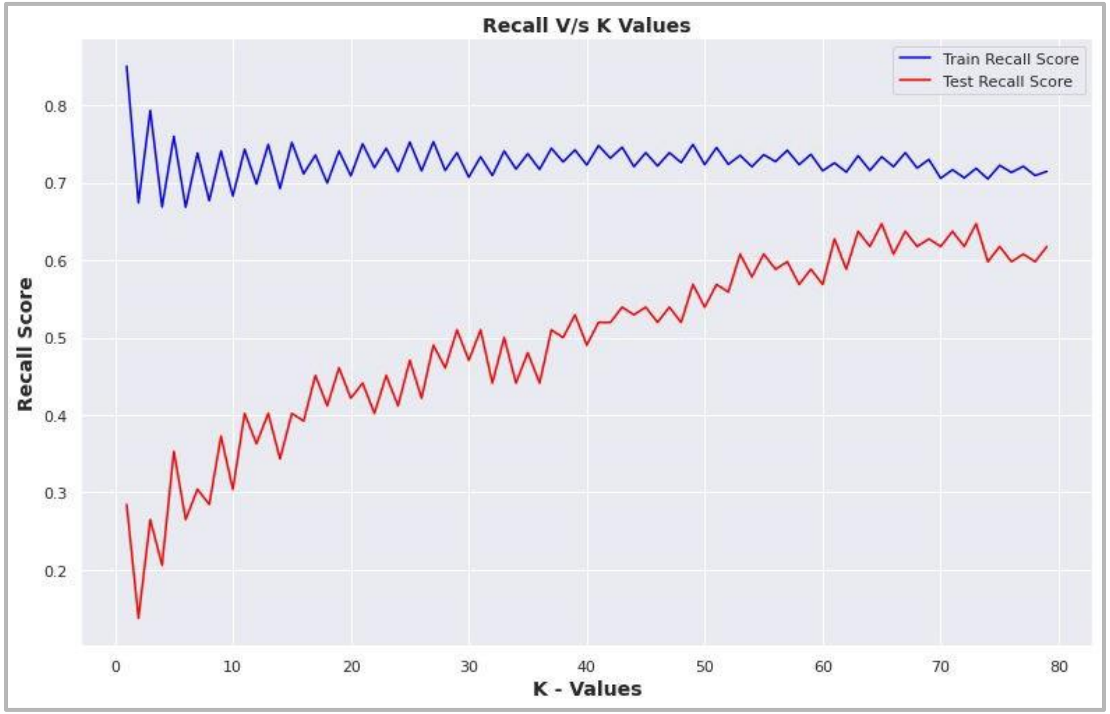


SVM results for class 1 on Test data :

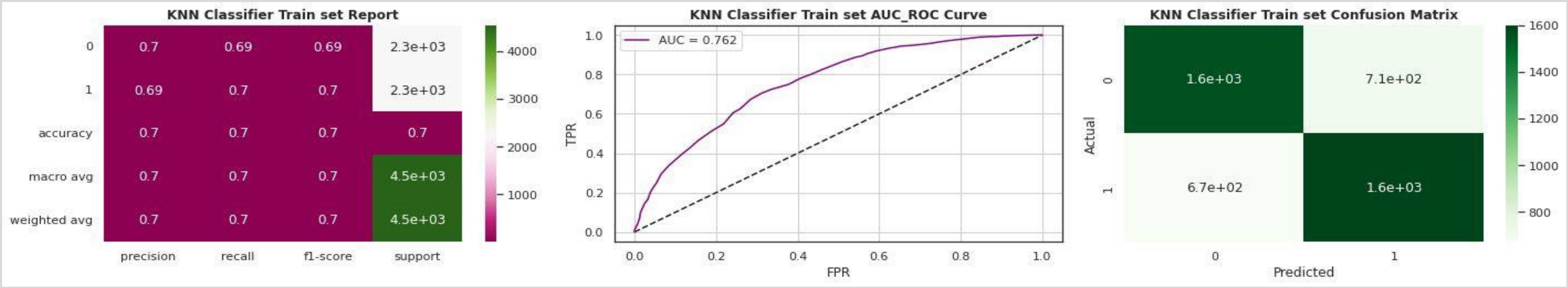
Precision : 0.17 || Recall : 0.88 || F1 Score : 0.29 || AUC : 0.593

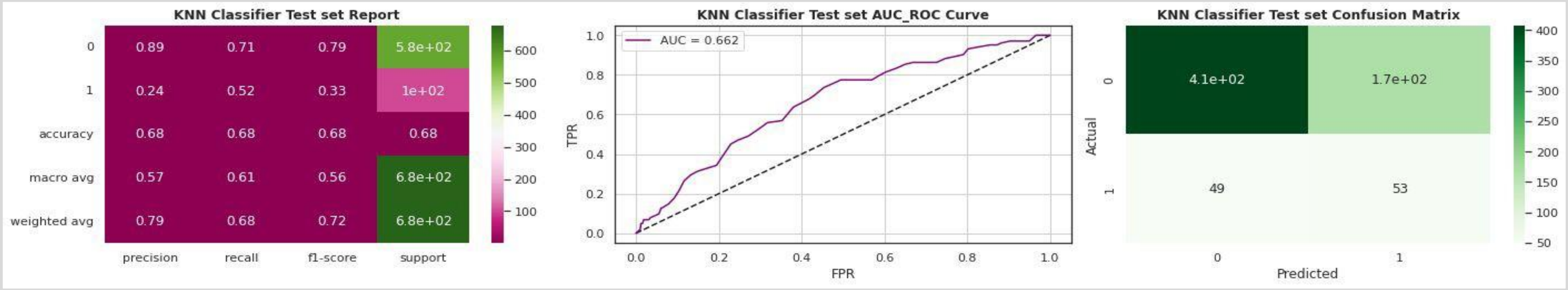
K- Nearest Neighbors

Estimating Optimum K Value



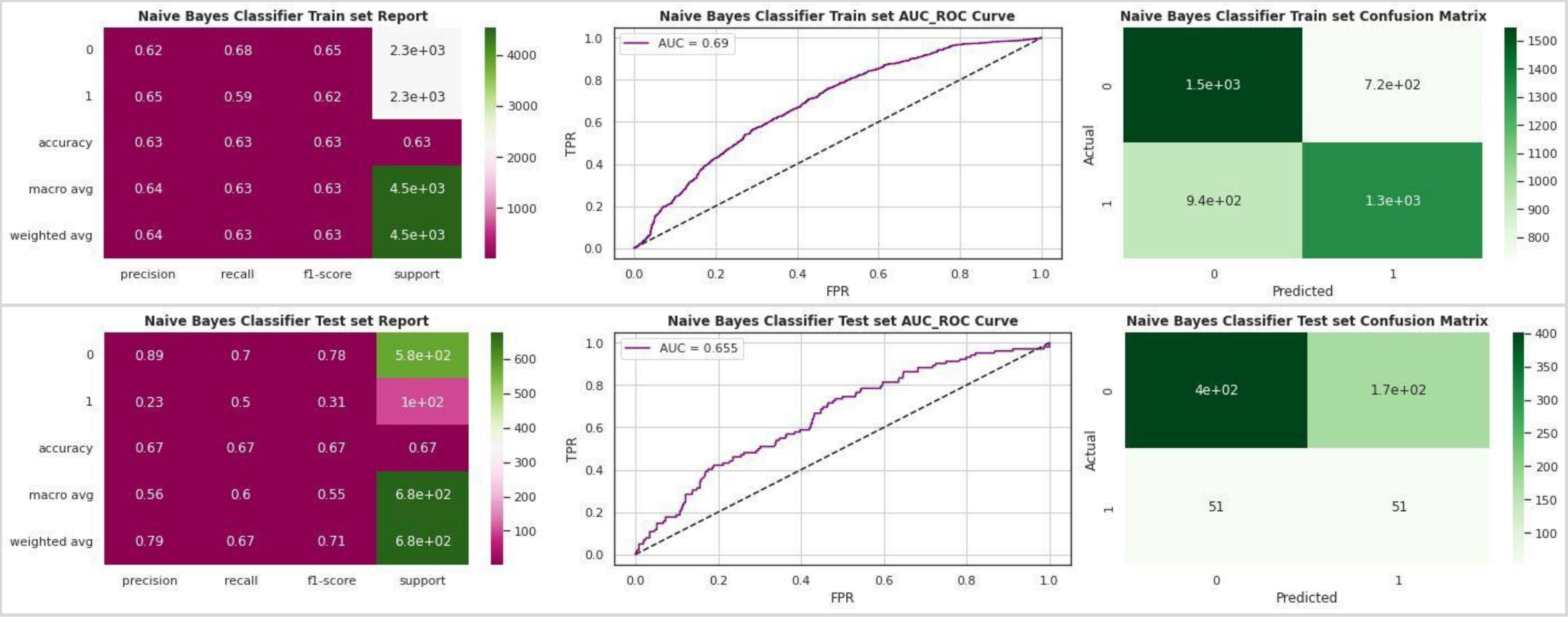
K- Nearest Neighbors





K-NN results for class 1 on Test data :
Precision : 0.24 || Recall : 0.52 || F1 Score : 0.33 || AUC : 0.662

Naive Bayes Classifier

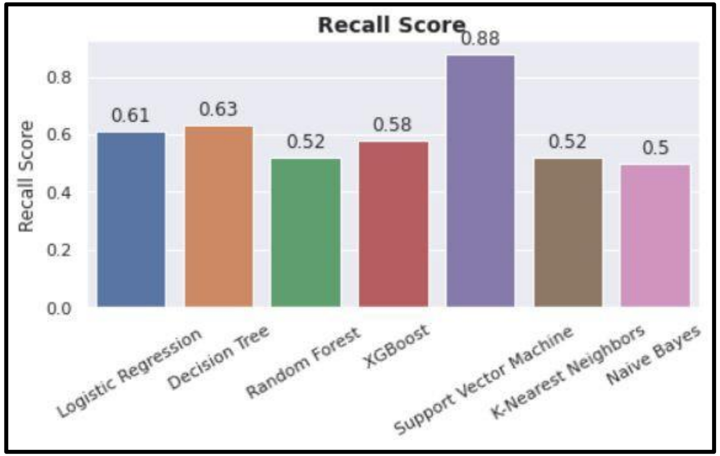


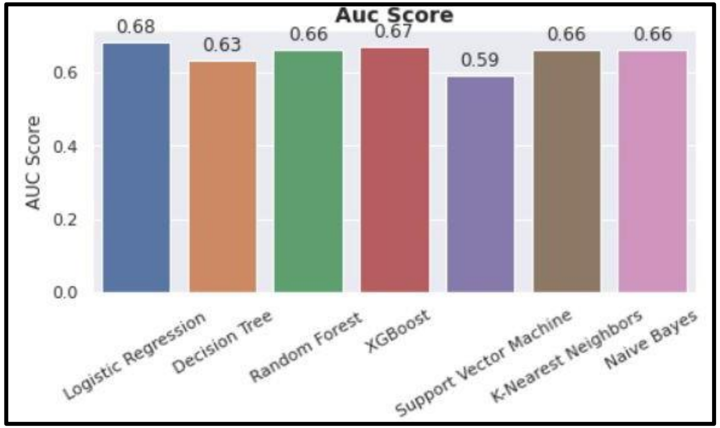
Naive Bayes results for class 1 on Test data :

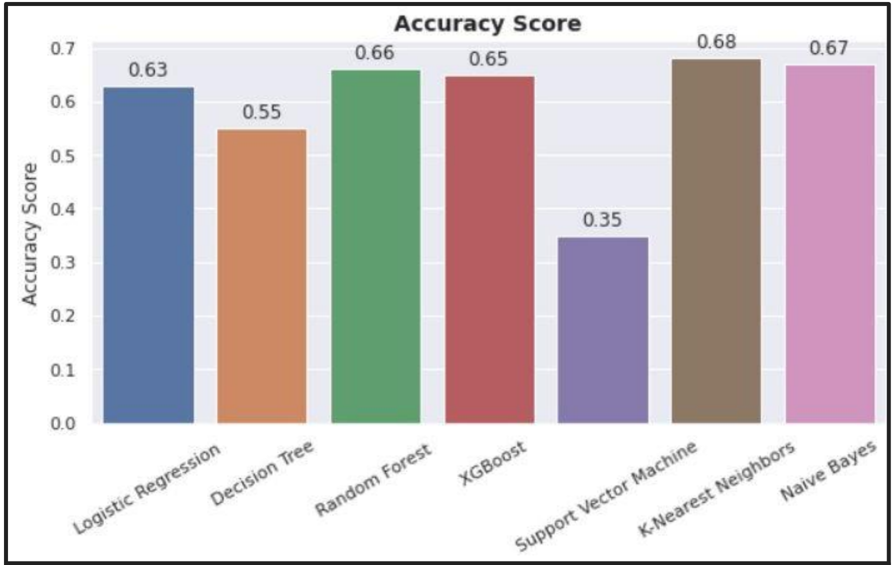
Precision : 0.23 || Recall : 0.50 || F1 Score : 0.31 || AUC : 0.655

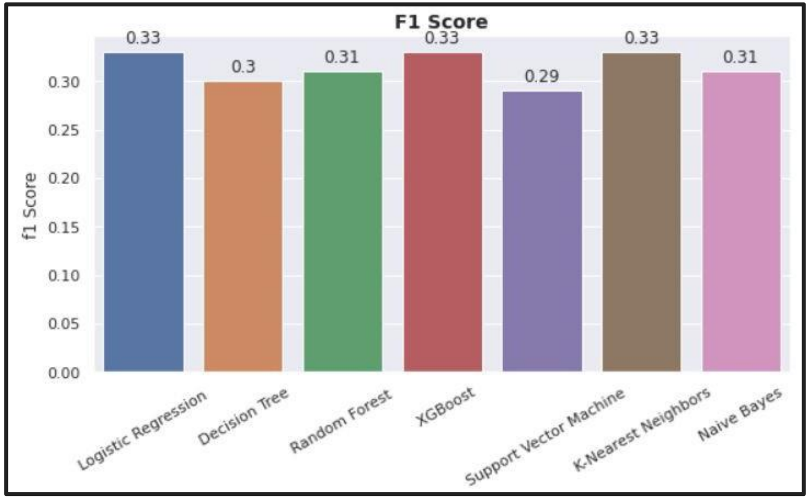
Model Comparison Matrix

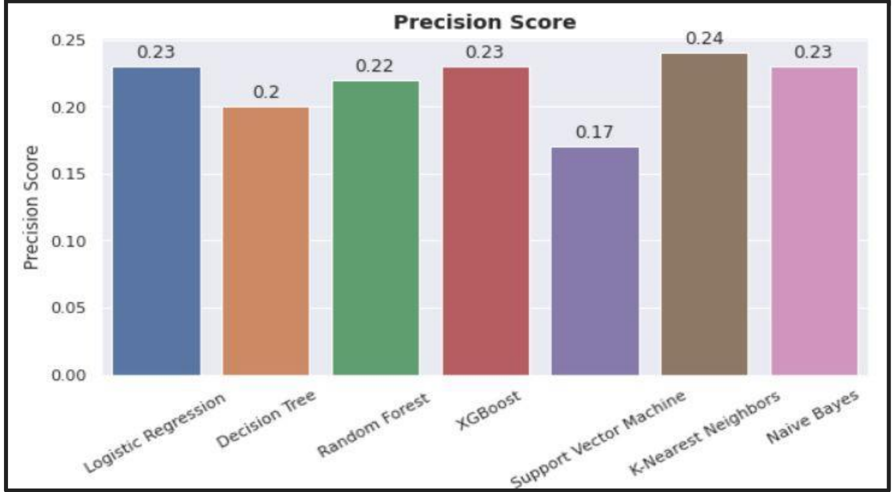
		Model	Accuracy Score	Precision Score	Recall Score	f1 Score	AUC Score
Training Dataset Results	0	Logistic Regression	0.64	0.64	0.64	0.64	0.71
	1	Decision Tree	0.69	0.65	0.82	0.73	0.77
	2	Random Forest	0.73	0.71	0.76	0.74	0.82
	3	XGBoost	0.72	0.70	0.79	0.74	0.83
	4	Support Vector Machine	0.60	0.56	0.93	0.70	0.68
	5	K-Nearest Neighbors	0.70	0.69	0.70	0.70	0.76
	6	Naive Bayes	0.63	0.65	0.59	0.62	0.69
Test Dataset Results	0	Logistic Regression	0.63	0.23	0.61	0.33	0.68
	1	Decision Tree	0.55	0.20	0.63	0.30	0.63
	2	Random Forest	0.66	0.22	0.52	0.31	0.66
	3	XGBoost	0.65	0.23	0.58	0.33	0.67
	4	Support Vector Machine	0.35	0.17	0.88	0.29	0.59
	5	K-Nearest Neighbors	0.68	0.24	0.52	0.33	0.66
	6	Naive Bayes	0.67	0.23	0.50	0.31	0.66





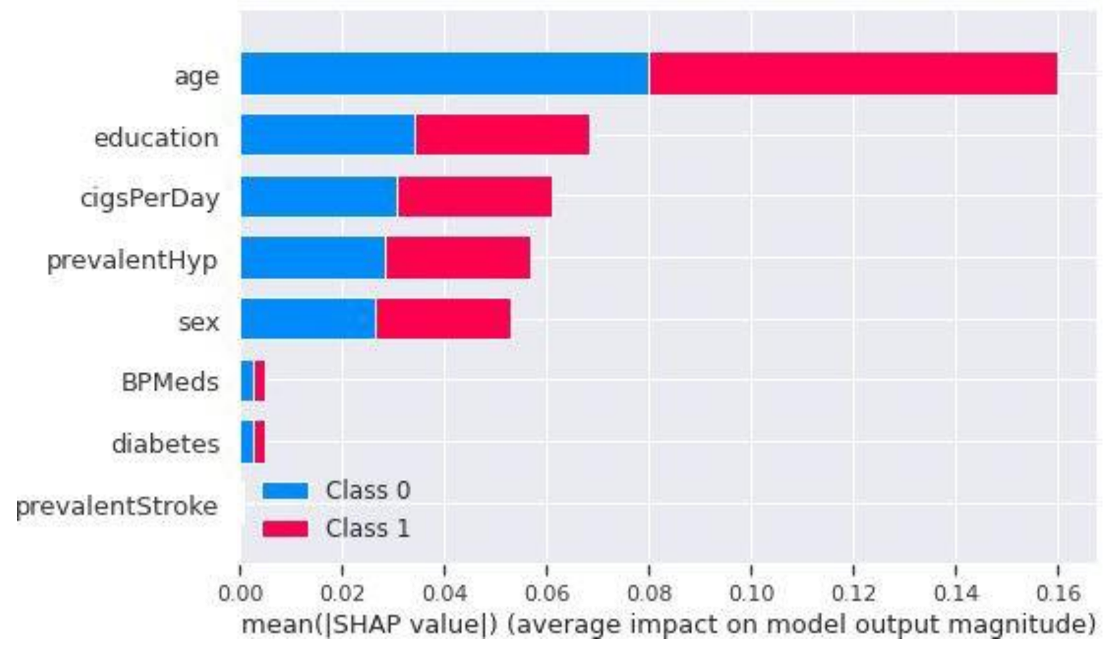

Accuracy Score
Accuracy Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.63
0.55
0.66
0.65
0.35
0.68
0.67
Logistic Regression
Decision Tree
Random Forest
XGBoost
Support Vector Machine
K-Nearest Neighbors
Naive Bayes


F1 Score
f1 Score
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.33
0.3
0.31
0.33
0.29
0.33
0.31
Logistic Regression
Decision Tree
Random Forest
XGBoost
Support Vector Machine
K-Nearest Neighbors
Naive Bayes


Precision Score
Precision Score
0.25
0.20
0.15
0.10
0.05
0.00
0.23
0.2
0.22
0.23
0.17
0.24
0.23
Logistic Regression
Decision Tree
Random Forest
XGBoost
Support Vector Machine
K-Nearest Neighbors
Naive Bayes

Model Explanation

Summary Plot - SVM Model



SVM - The most important features are ’Age’, ‘Education’ , 'cigarettes per day' and 'Prevalent Hypertension'.

Conclusion

- Predicting the risk of coronary heart disease is critical for reducing fatalities caused by this disease, we can avert deaths by taking required medications and precautions if we can foresee the danger of this illness ahead of time.

- It is important to have a high recall score in this scenario because It is okay if the model incorrectly identifies a healthy person as a high risk patient, because it will not result in death, but if a high risk patient incorrectly identified as healthy, it may result in fatality. Support Vector Machine with rbf kernel is the best model with recall score of 0.88.

- There may be a case where the patients who are incorrectly classified as suffering from heart disease is equally important as patients who are correctly classified as suffering from heart disease, because patients who are incorrectly classified they may have some other illness, so in that case high f1 score is desired. Logistic Regression, XGBoost, K-NN these are the model with most F1 score.

- From our analysis, it is found that the 'Age' of the patient is the most important feature in determining the risk of coronary heart disease, middle and older age people are more prone to coronary heart disease than younger people followed by 'cigarettes per day', 'BP Meds', 'Prevalent Hypertension' are also very important feature in determining risk of heart disease

- Future developments must include a strategy to improve models scores with the help of more data from people with different medical history.

Challenges Faced

- Handling missing values in the dataset and working with limited availability of data.

- Exploring all the columns and calculating VIF for multicollinearity , deciding on which features to be dropped/kept/transformed was challenging; it might decrease the models performance.

- Selecting the appropriate models and choosing the best hyperparameters to maximize the performance of our models and to prevent overfitting was one of the challenges faced.

References

- MachineLearningMastery
- GeeksforGeeks
- Analytics Vidhya Blogs
- Towards Data Science Blogs
- Stack Overflow

Thank You!