Capstone Project - 3

Cardiovascular Risk Prediction

Faraz Ahmad
(Individual)

Problem statement

- 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. Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors.

Data Description

- **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)

- **Behavioural**
  - 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.)

- **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)

Data Description

- **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)

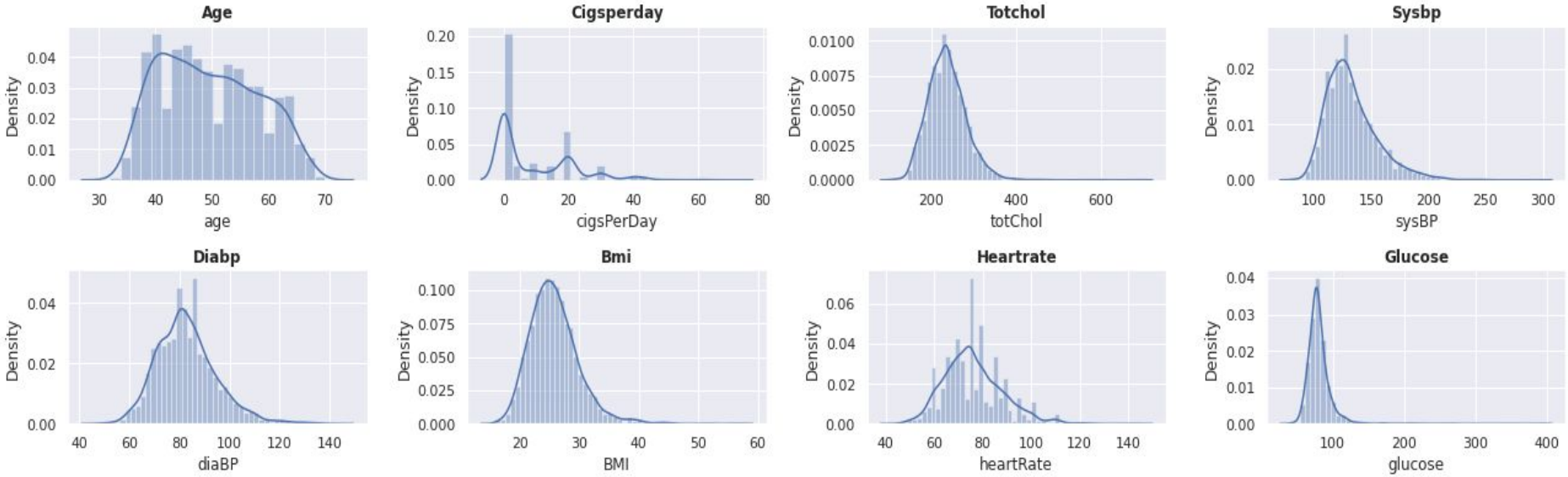
- **Predict variable (desired target)**
  - 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

Inspecting dataset

- The dataset contains 3390 rows and 17 columns.
- Dropping ‘id’ column because it contains unique id numbers of the patients and will not be helpful in prediction.
- Missing data count and percentage for each column are as follows:
  - 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%)

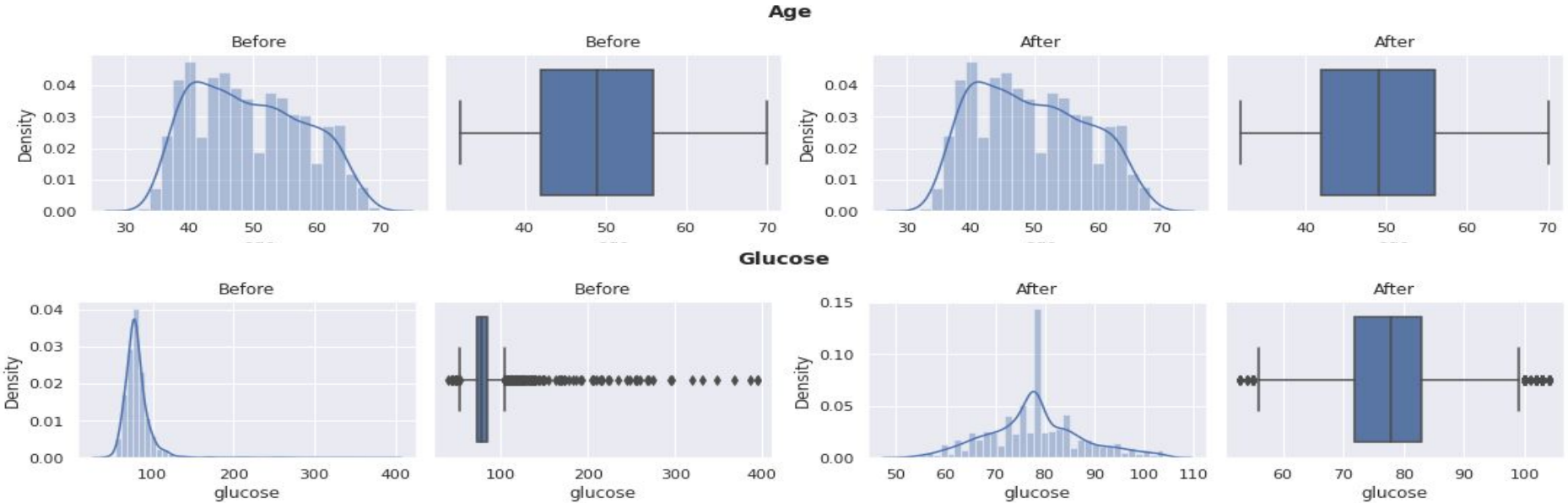
- Replacing null values with median in all columns.
- There are no duplicate values in the dataset.

Checking distribution of numerical features

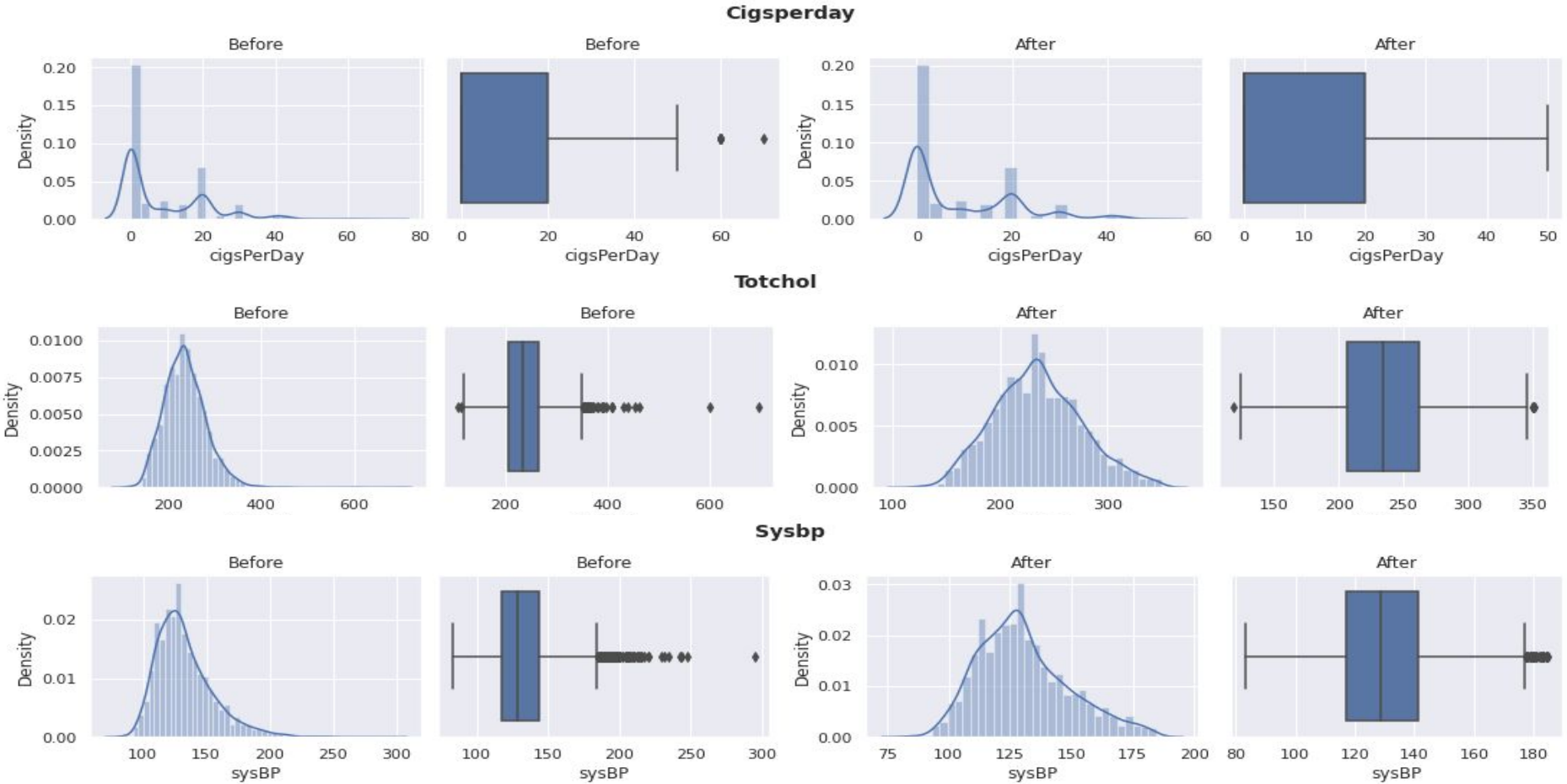


Handling Outliers

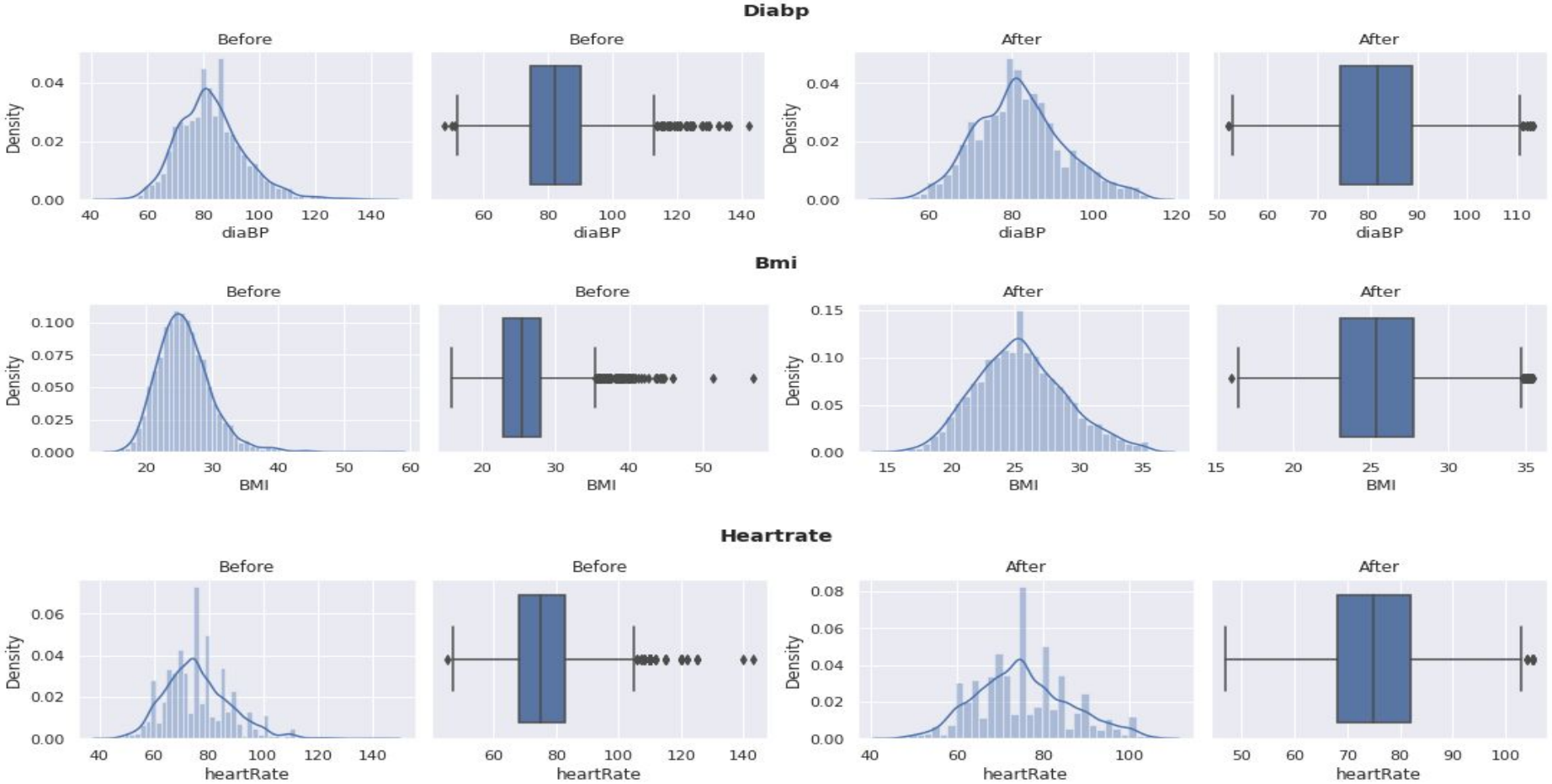
The IQR method for identifying outliers is used to set up a “fence” outside of Q1 (first quartile) and Q3 (third quartile). Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. We replaced the outliers with median.



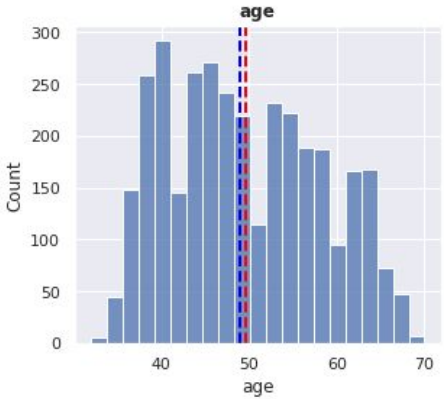
Handling Outliers

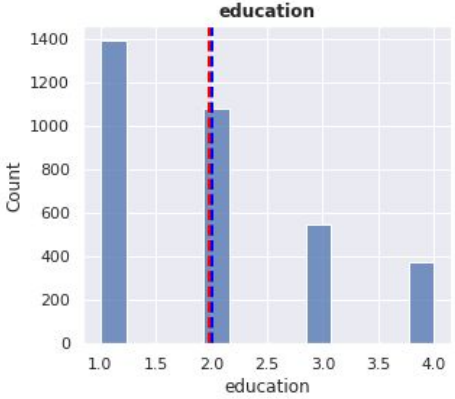


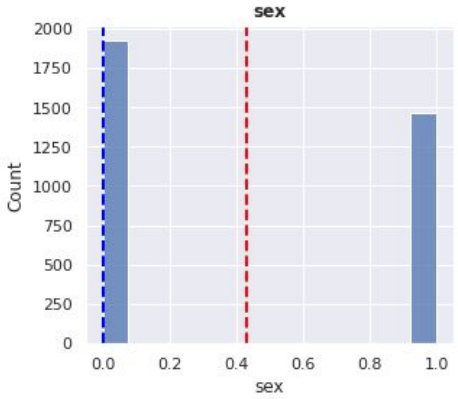
Handling Outliers

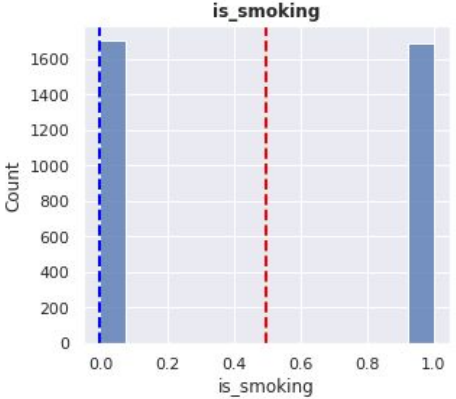


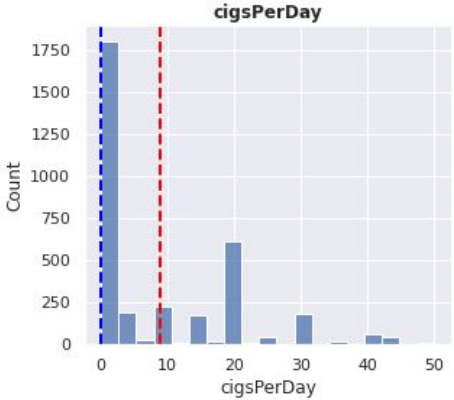
Univariate analysis

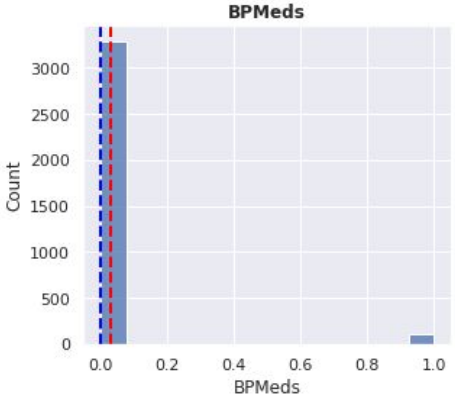


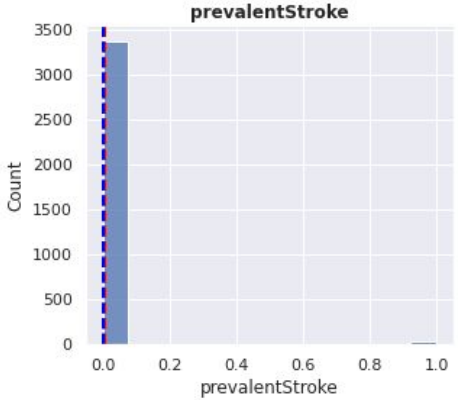


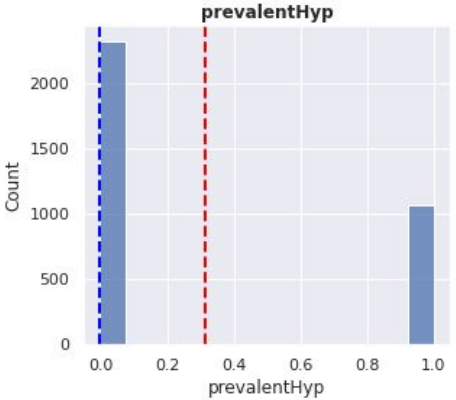




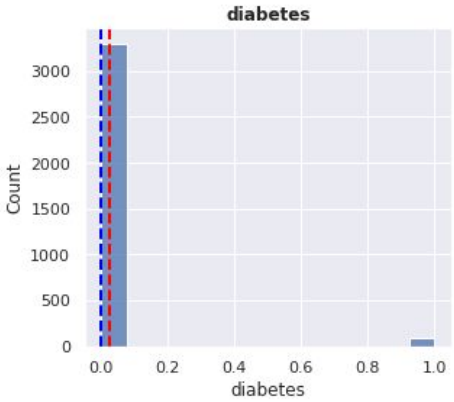


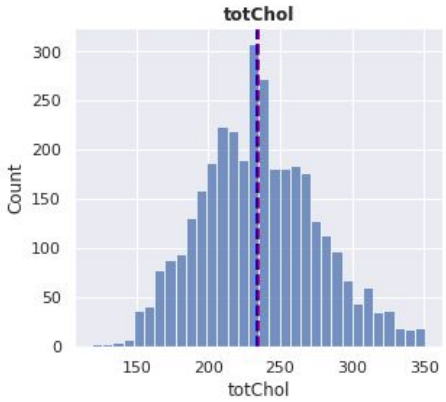


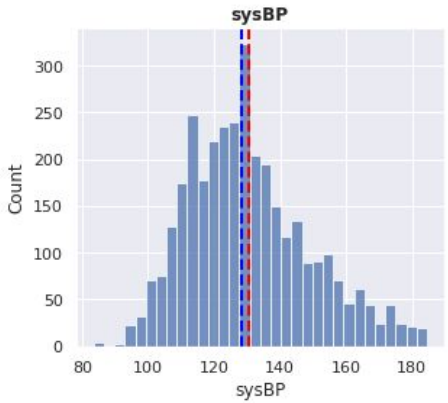


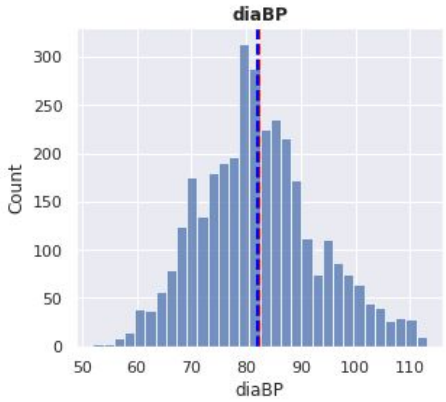


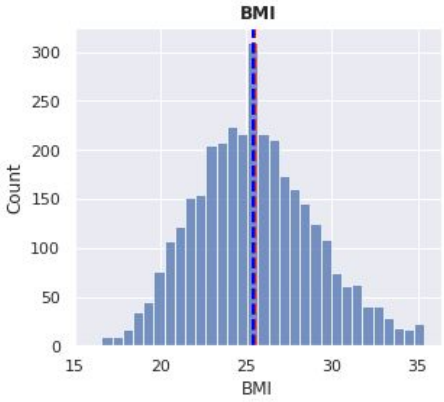
Univariate analysis

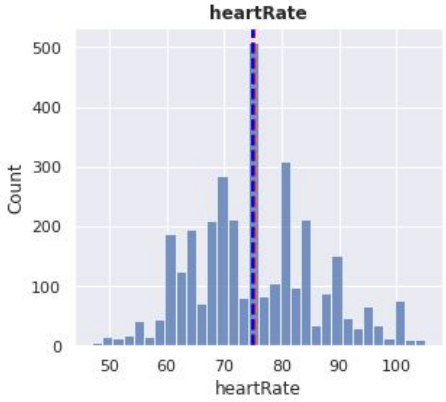


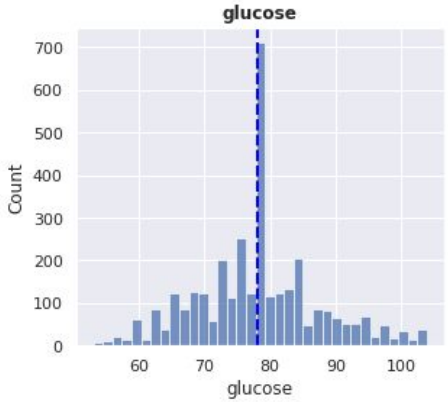


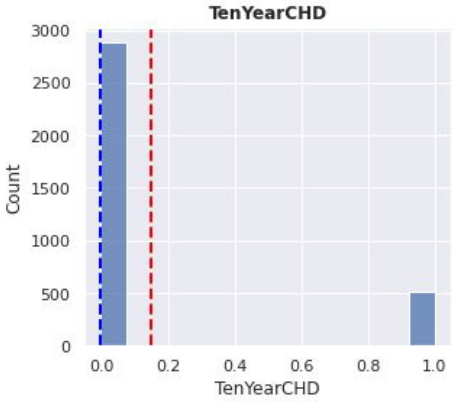








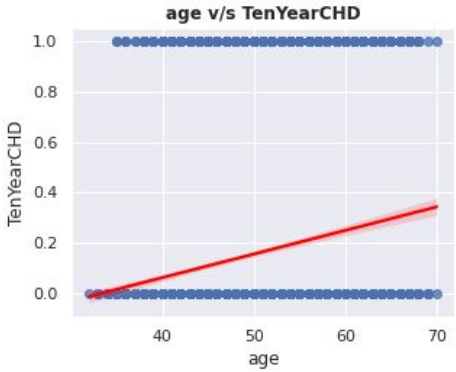


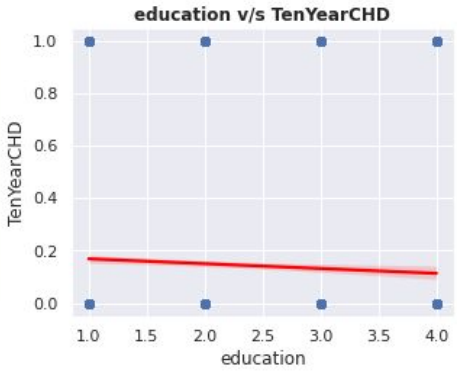


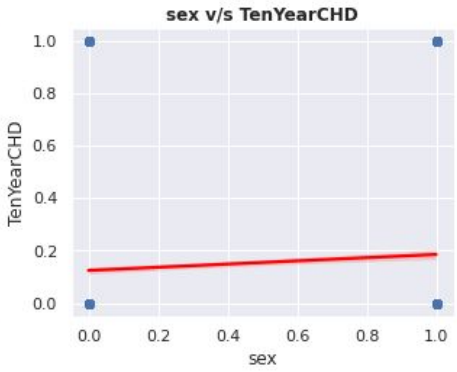
Univariate analysis

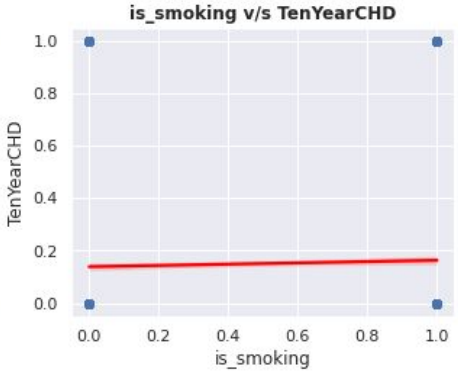
- In our dataset, most people are between 40 and 50 years old.
- The number of females is more than that of males in the dataset.
- The number of smokers and non-smokers is almost equal in the dataset.
- There are only few people which are on blood pressure medication and have diabetes and had previously a stroke.
- The average person smokes fewer than ten cigarettes a day.
- Also in the dataset provided, very few number of people have the risk of coronary heart disease. So we have to deal with class imbalance problem which we will discuss in the later slides.

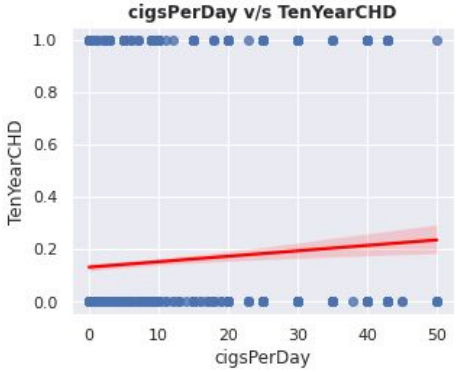
Bivariate analysis

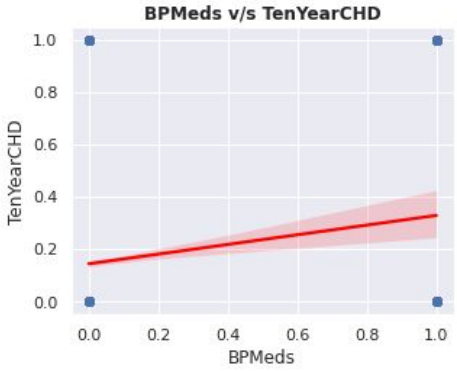


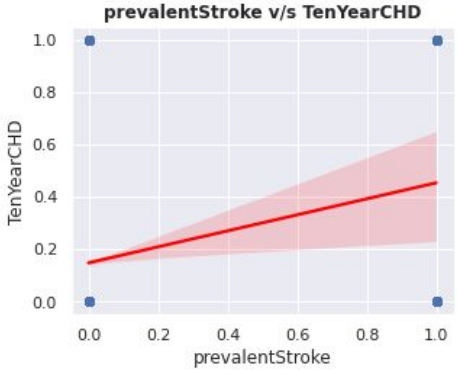


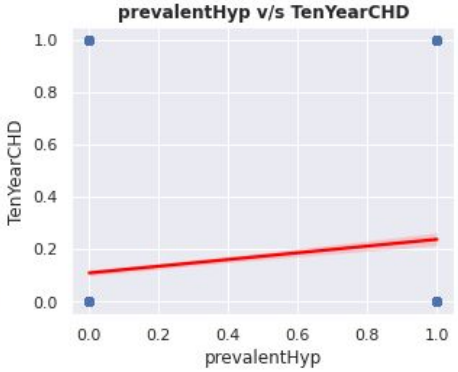




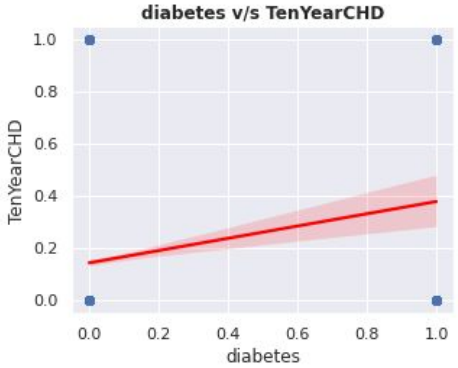


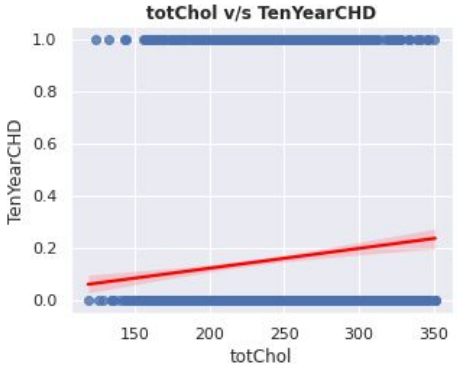


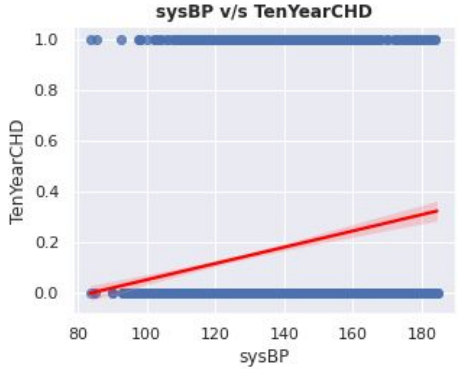


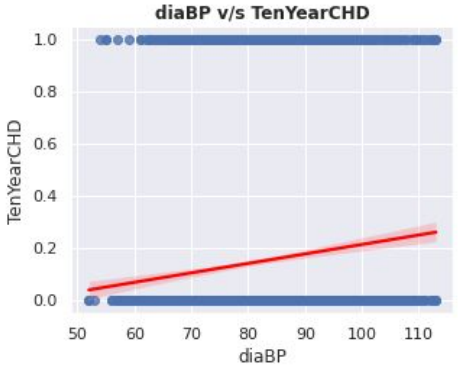


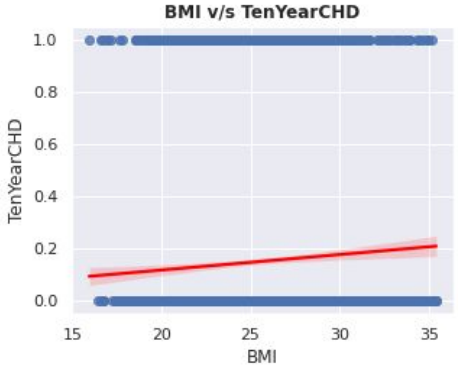
Bivariate analysis

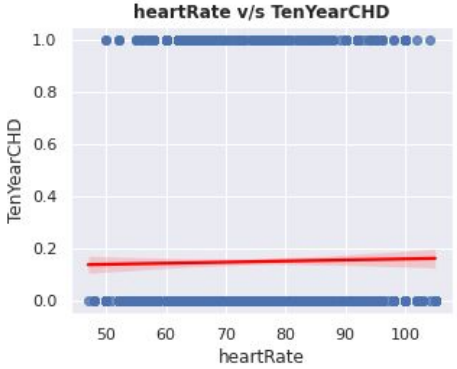


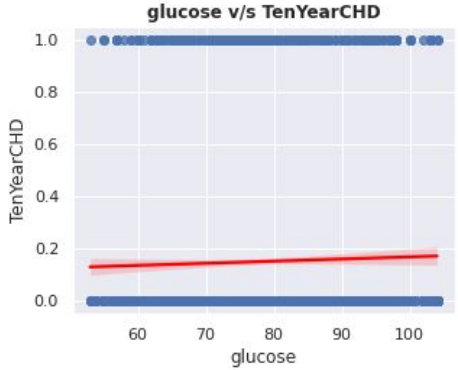






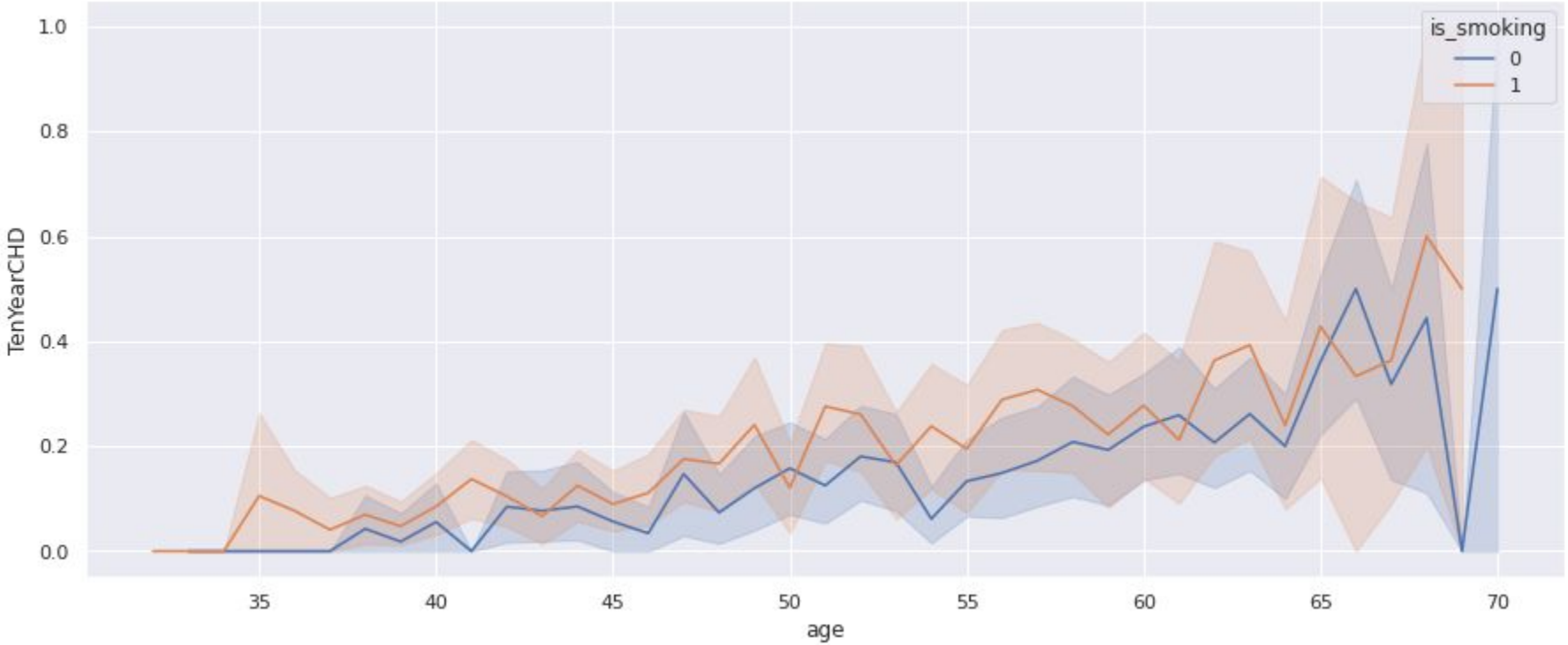






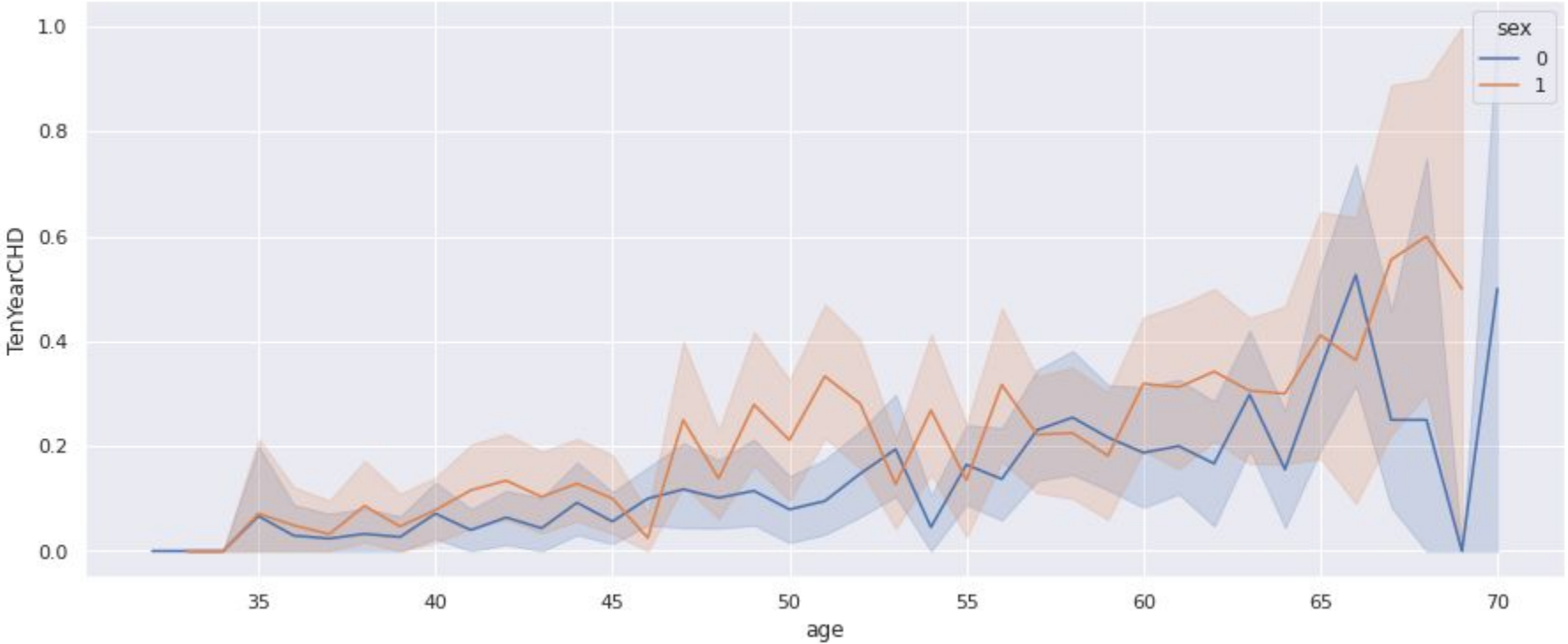
Multivariate analysis

- The chances of coronary heart disease increases with the increase in age.
- Furthermore, if the person is a smoker, he/she is at a higher risk as compared to non smoker.

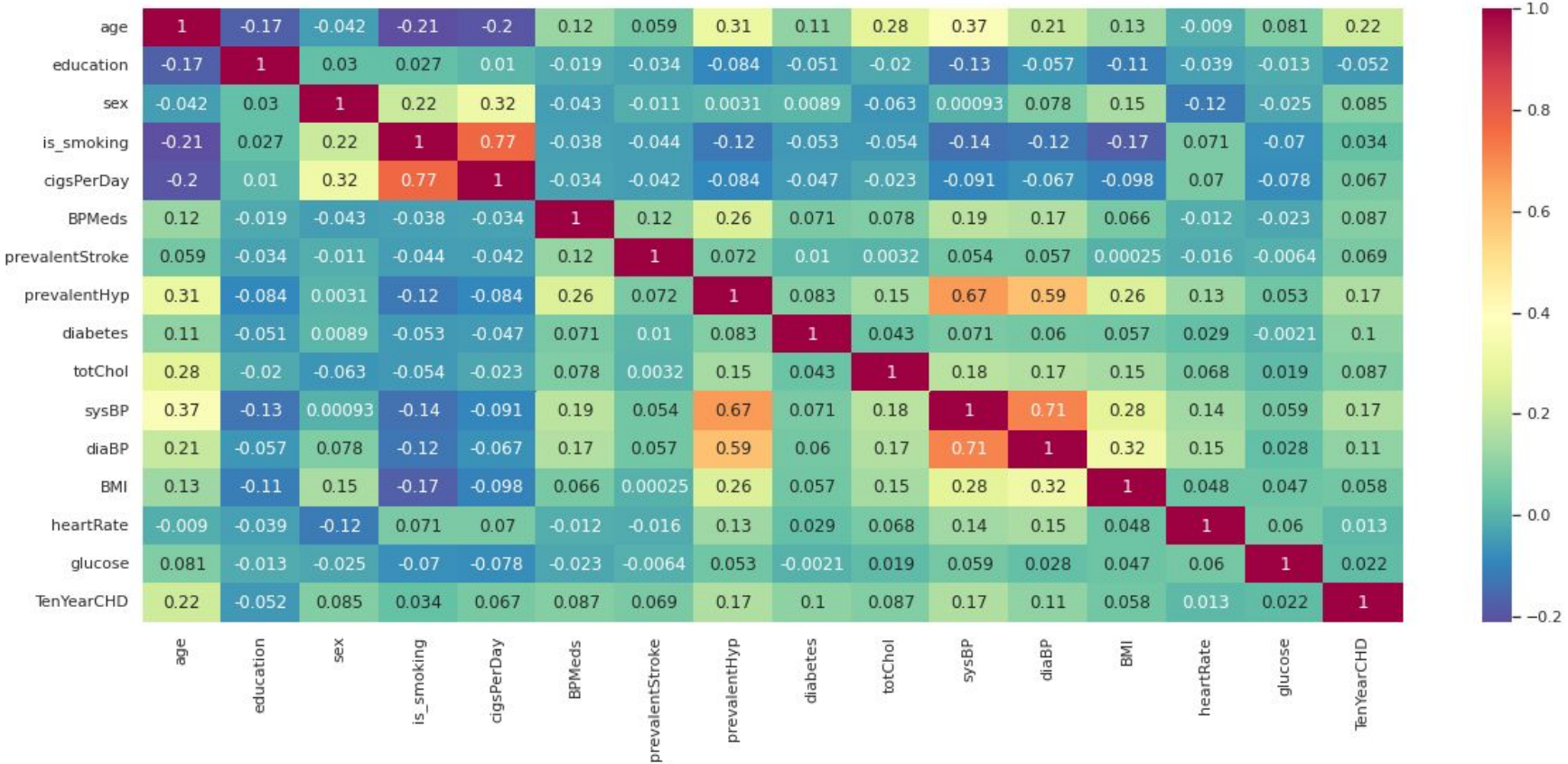


Multivariate analysis

The males have higher chances of getting a CHD as compared to females for the same age group.



Handling Multicollinearity

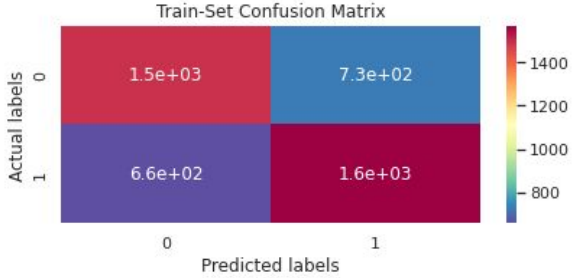


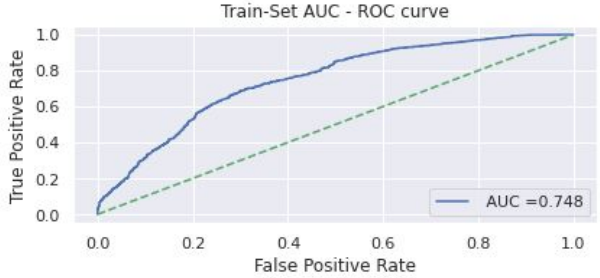
Model building prerequisites

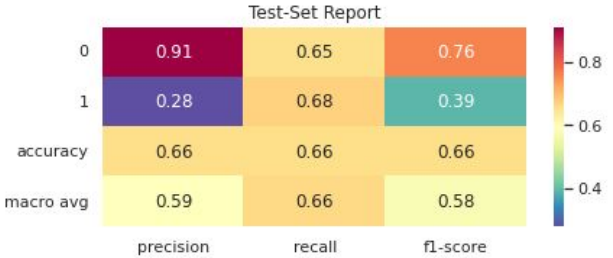
- Using Minmax scaler for feature scaling.
- Splitting the X and y variables into 80-20 ratio as train and test sets.
- Handling class imbalance by oversampling using SMOTE followed by removing Tomek links. Finally, checking value counts for both classes before and after handling class imbalance.

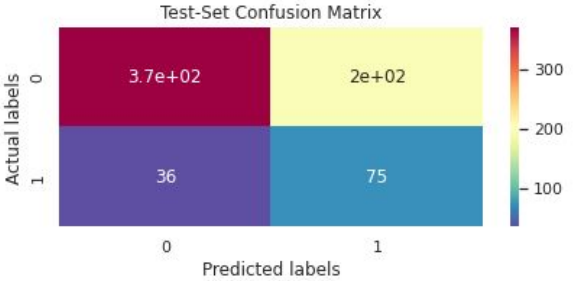
```
Before Handling Class Imbalace:
0    2300
1     412
Name: TenYearCHD, dtype: int64

After Handling Class Imbalace:
0    2195
1    2195
Name: TenYearCHD, dtype: int64
```

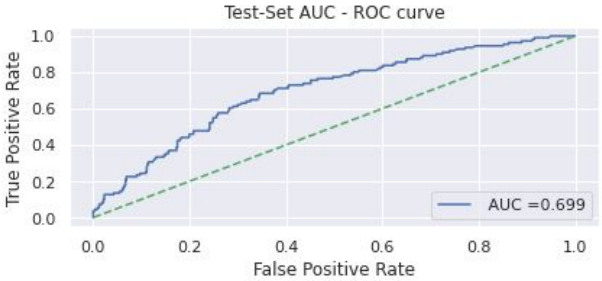
Logistic Regression

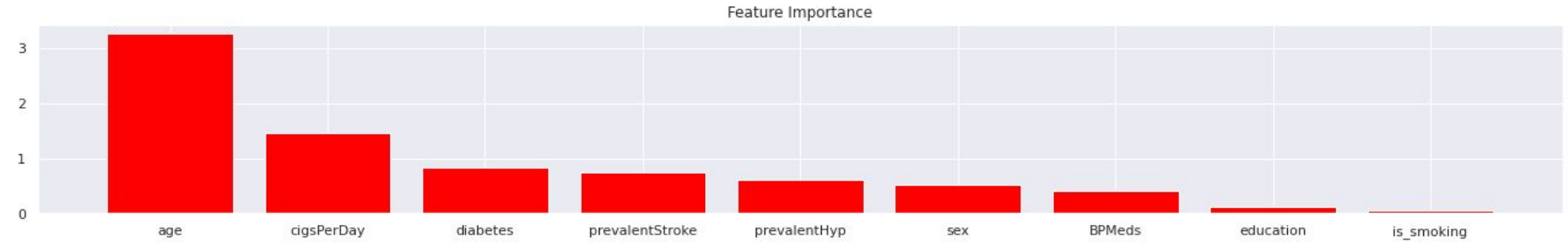




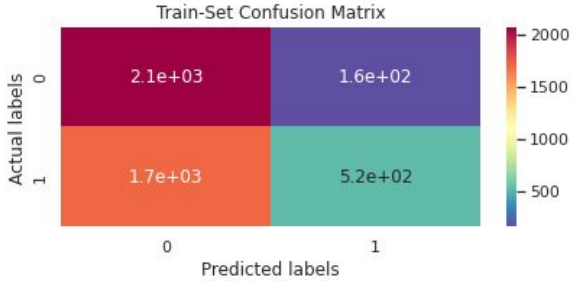


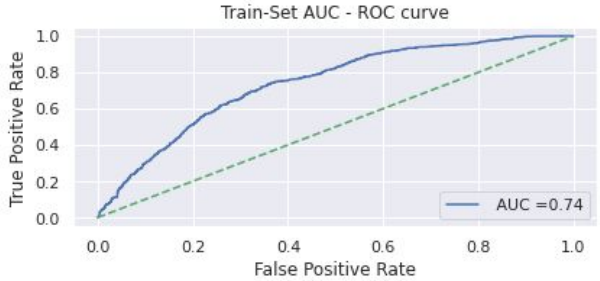


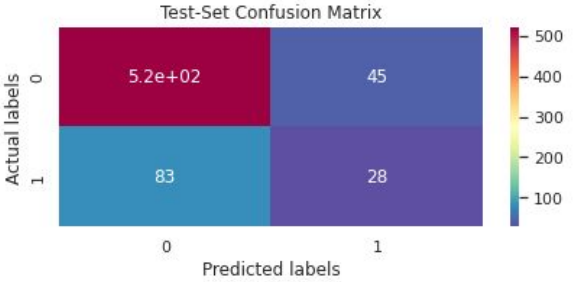


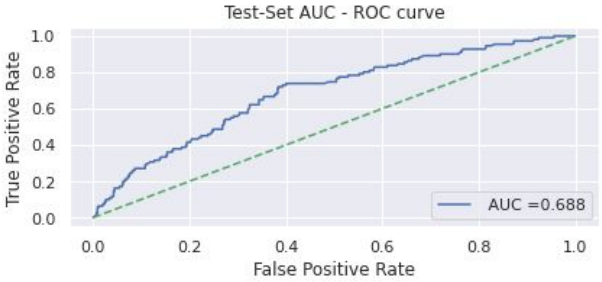


Naive Bayes Classifier

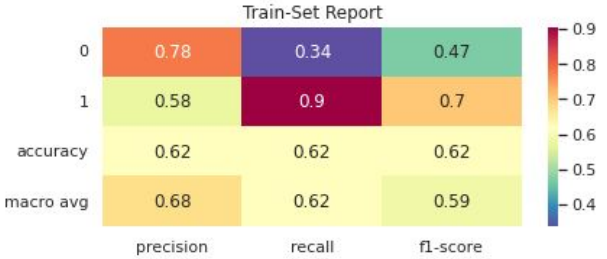


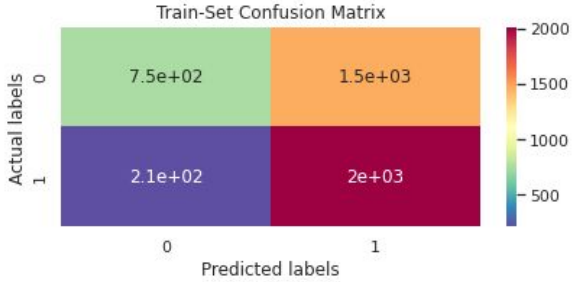


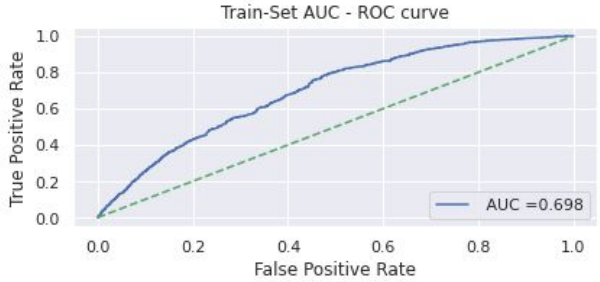


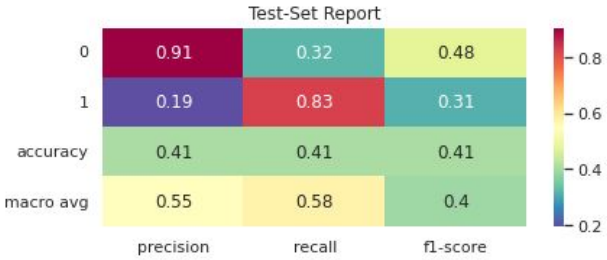


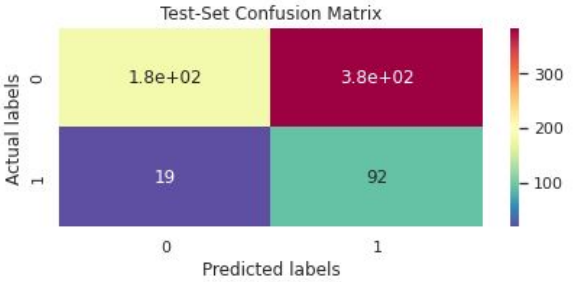
Support Vector Classifier

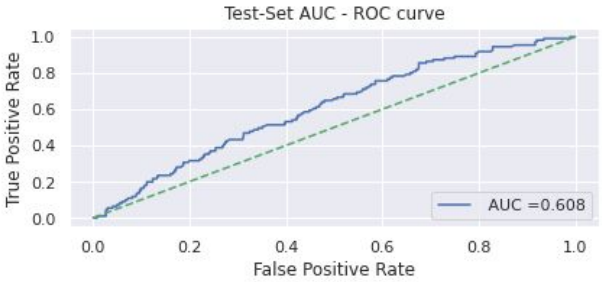




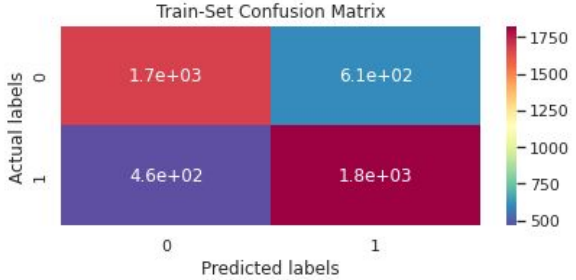


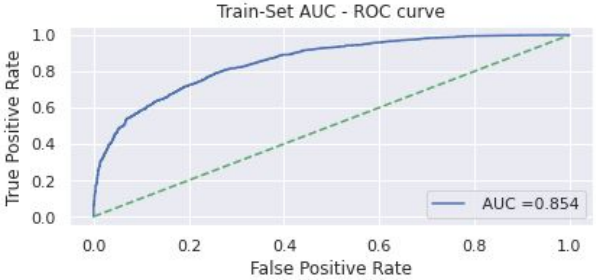


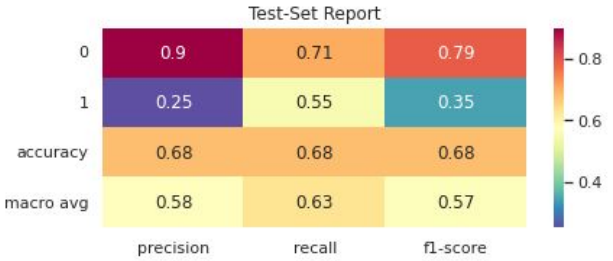


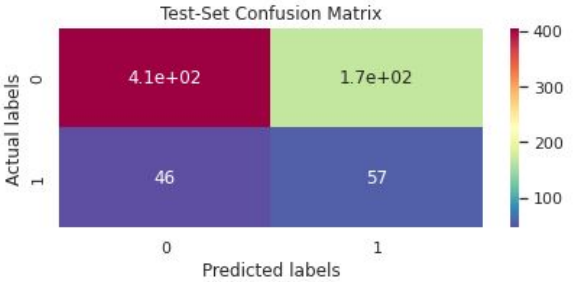


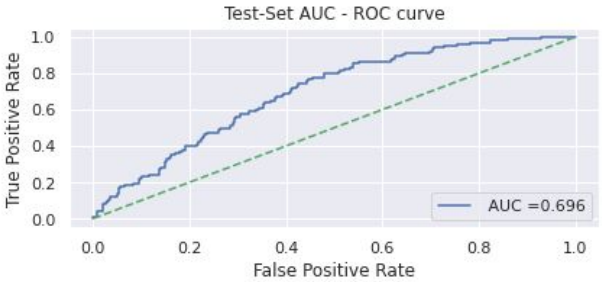
Random Forest Classifier

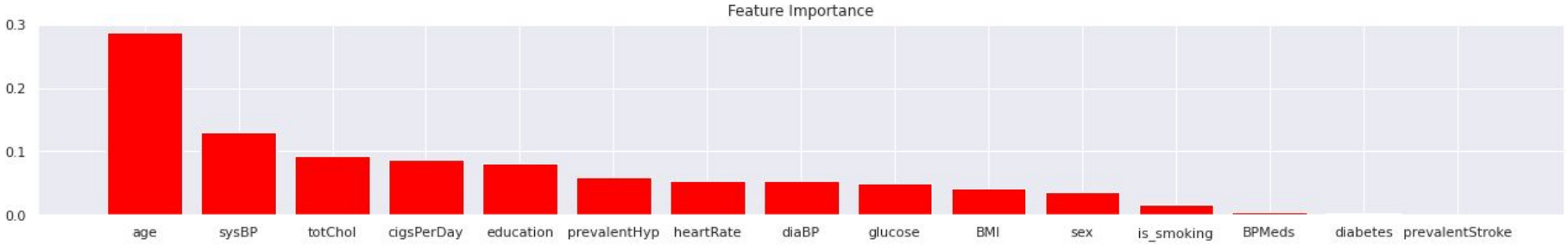




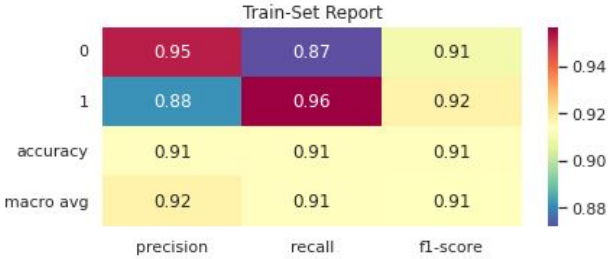


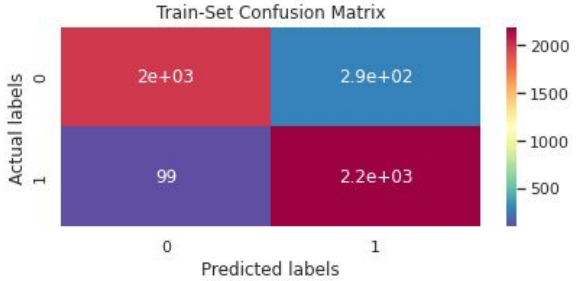


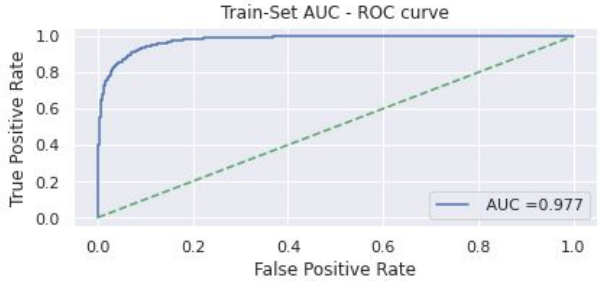


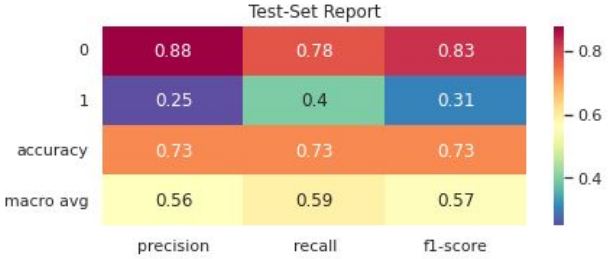


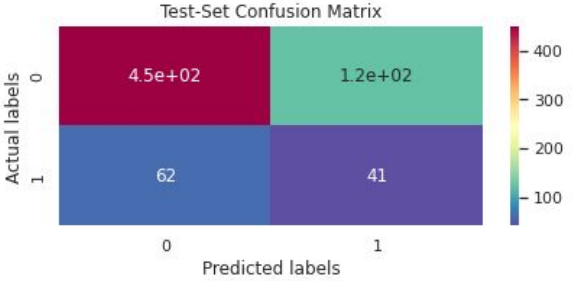
XGBoost Classifier

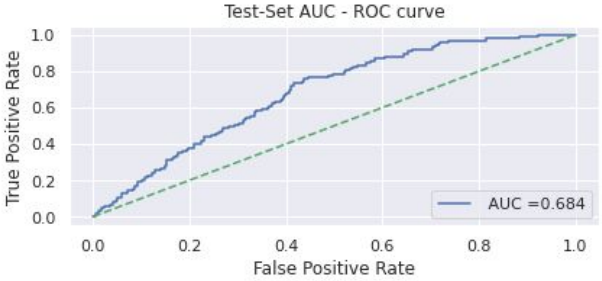


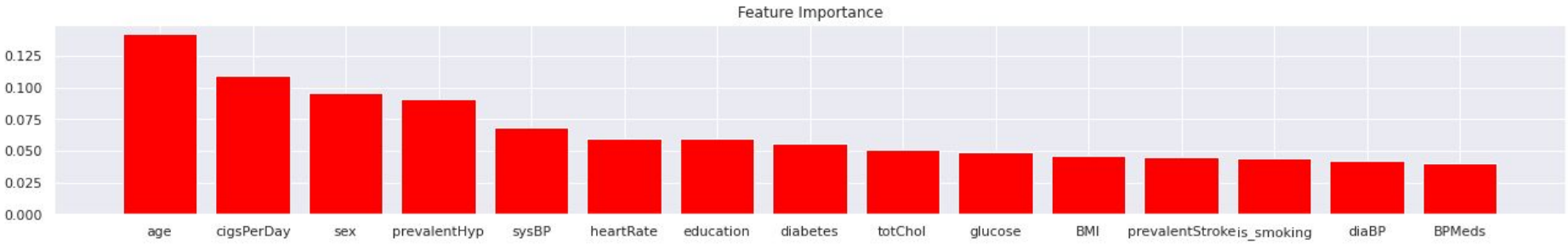




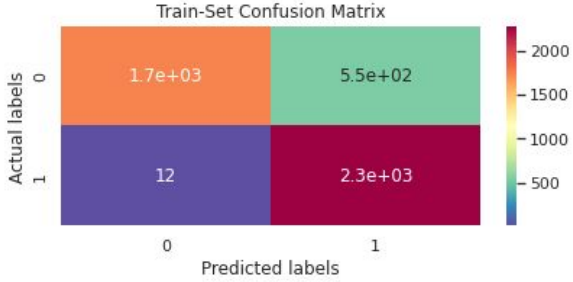


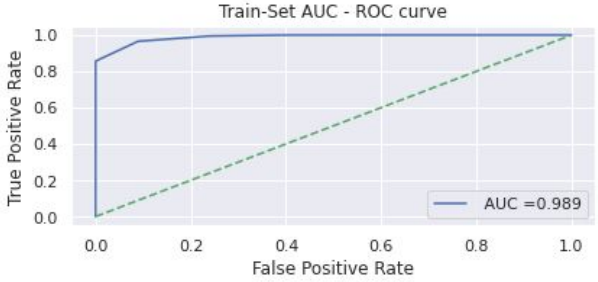


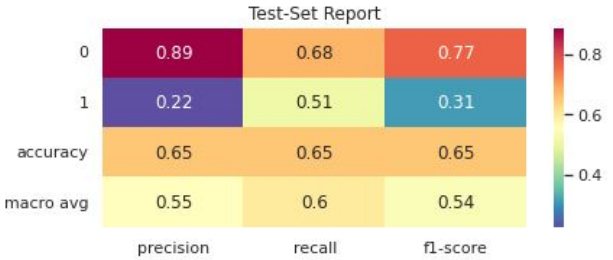


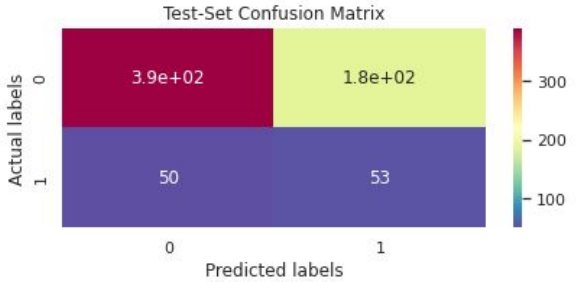


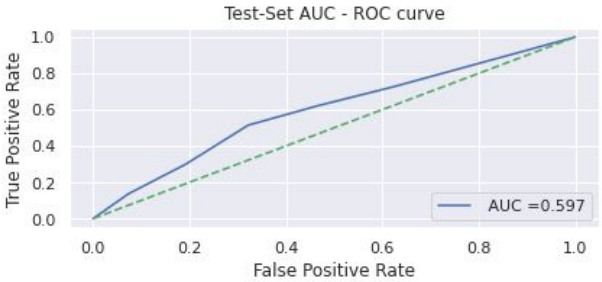
KNN Classifier











Conclusion

- If we want to completely avoid any situations where the patient has heart disease, a high recall is desired. Whereas if we want to avoid treating a patient with no heart diseases a high precision is desired.
- Assuming that in our case the patients who were incorrectly classified as suffering from heart disease are equally important since they could be indicative of some other ailment, so we want a balance between precision and recall and a high f1 score is desired.
- Since we have added synthetic data points to handle the huge class imbalance in training set, the data distribution in train and test are different so the high performance of models in the train set is due to the train-test data distribution mismatch and not due to overfitting.
- Best performance of Models on test data based on evaluation metrics for class 1:
    - Recall - SVC
    - Precision - Naive Bayes Classifier
    - F1 Score - Logistic Regression
    - Accuracy - Naive Bayes Classifier