Supervised ML Project (Classification)

CardioVascular Risk Prediction

By-Aditi Rajguru

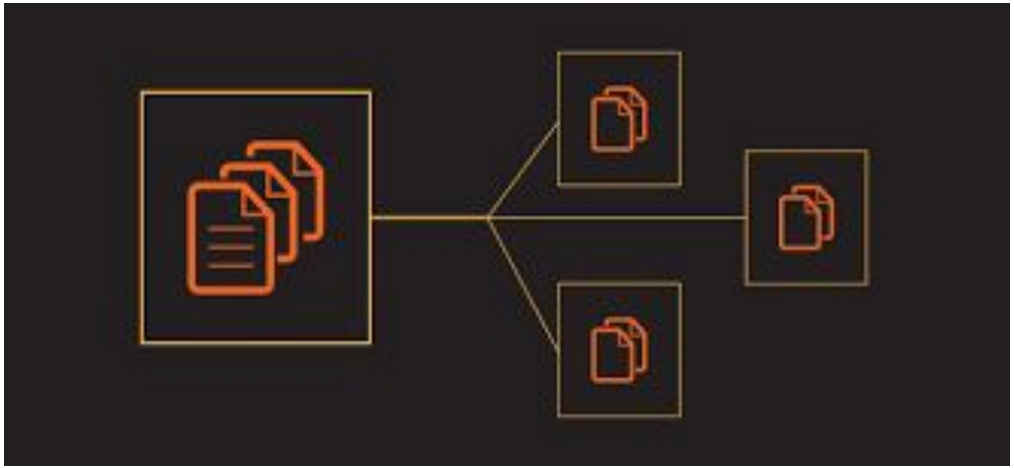
Problem Statement:

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 demographic, behavioral, and medical risk factors.Attributes includes various factors based on the mentioned three categories like sex,age,diabetes,TenYearCHD,is_smoking,etc.For this project,We have used various analysing techniques which includes Data acquisition,Data Description,Missing values imputation,Graphical Representation,Modelling,etc.

Contents

Datasets used and Feature Representation

Cardiovascular_risk_df:This dataframe is a created directly from the given file and since data was not very large all the attributes was included in this single dataframe and manipulated through out the process.



Feature Representation:

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.)

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”) - DV

Feature Engineering:

Duplicate Values:

There were no duplicate values in our dataset.

Dimension Reduction:

We imputed id feature as index after checking duplicates in it.

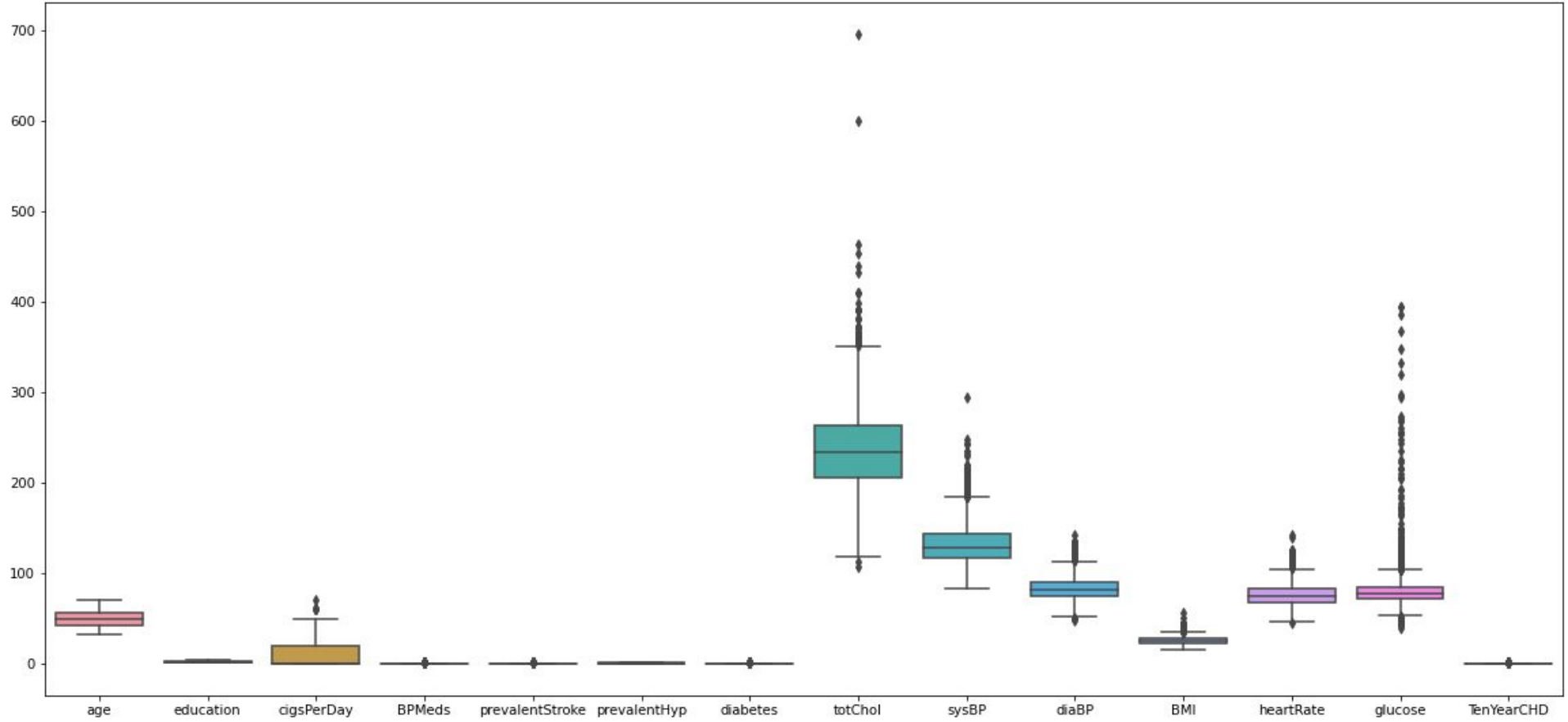
Missing Values:

For this classification project imputing of missing values was very important as it would been reflect on the results of the models.Missing values were simply filled by their respective mean,median and mode according to their data.

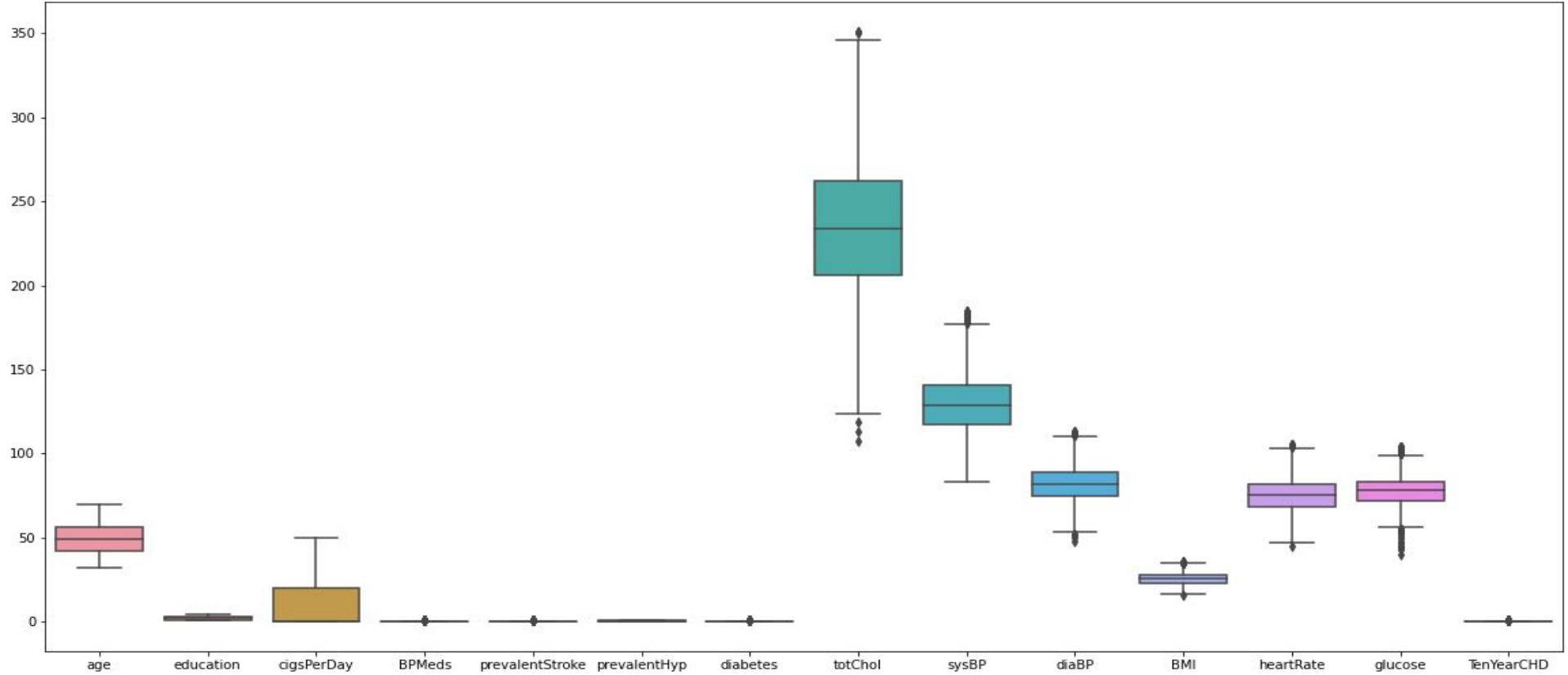
Outliers:

There were outliers in many attributes rather than removing them completed we replace them with median values as our data set is already having less data.

Before and After removing Outlier:



After Imputing Outlier:

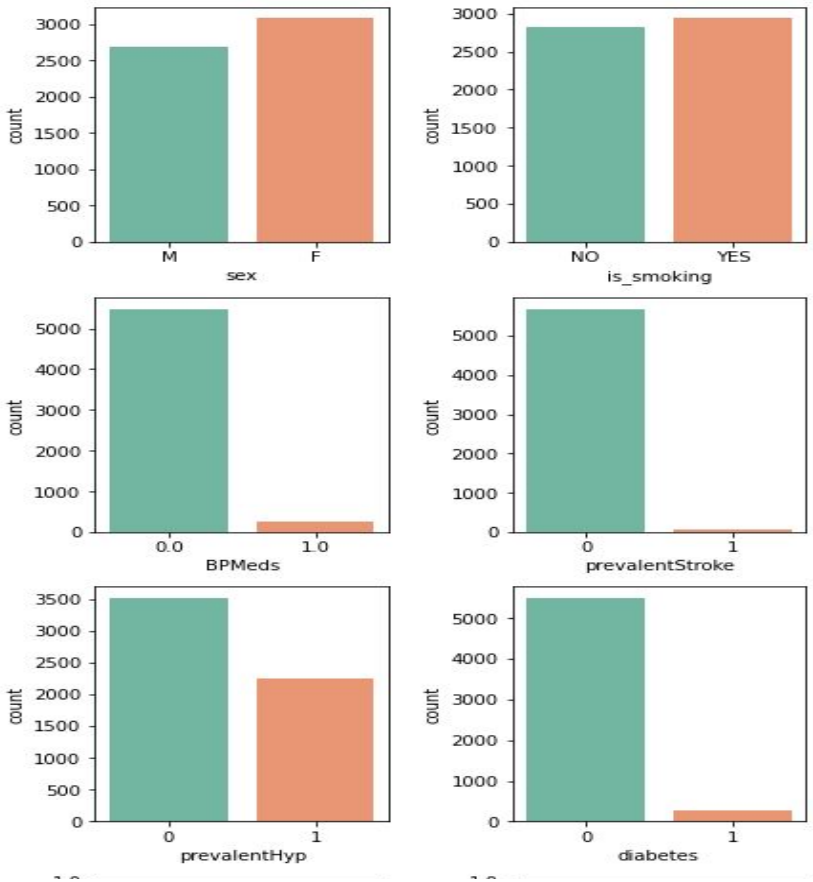


Exploratory Data Analysis:

Why EDA is important?

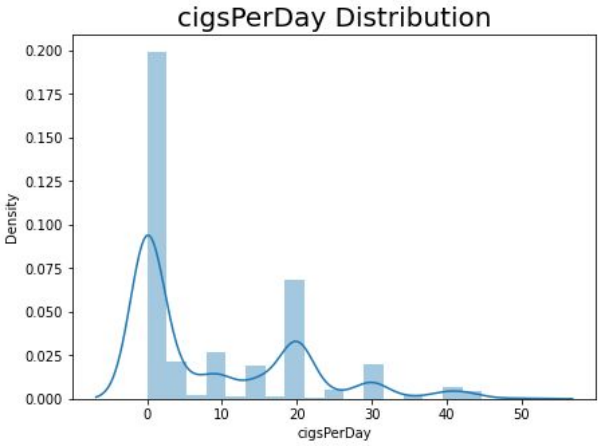
For an understanding of your data, it's characteristics, and it's distributions is vital to any successful data science task, whether it's inference or prediction. And contrary to what you might expect, the reason for EDA's importance is not technical, and has nothing to do with programming. It's the thing that separates a mediocre data scientist from a great one — decisions.

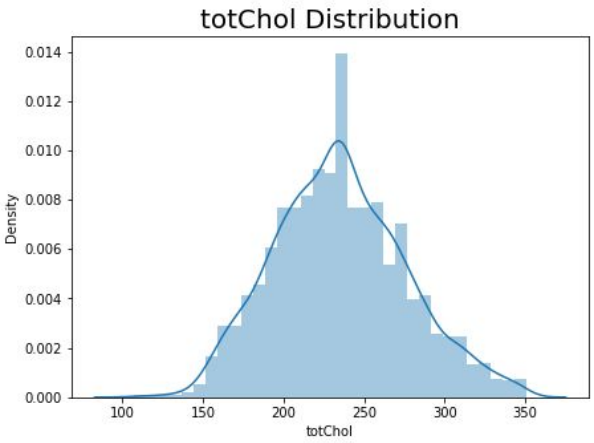
Categorical Features:

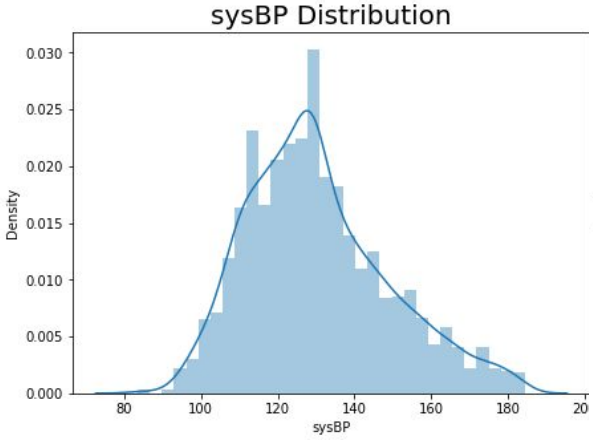


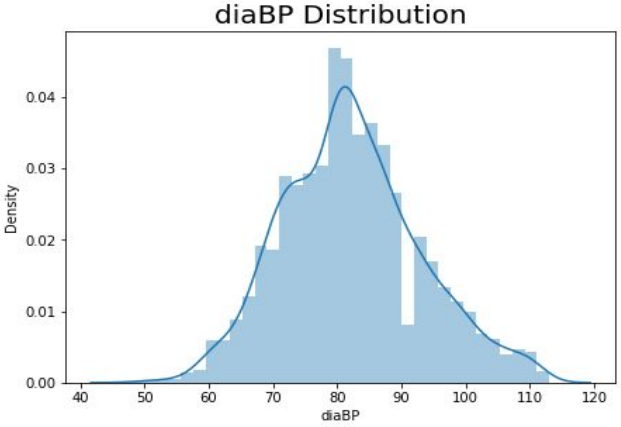
- BPmeds, prevalentStroke and diabetes are highly imbalanced.
- The number of Smokers and non-Smokers in is_smoking is almost the same.

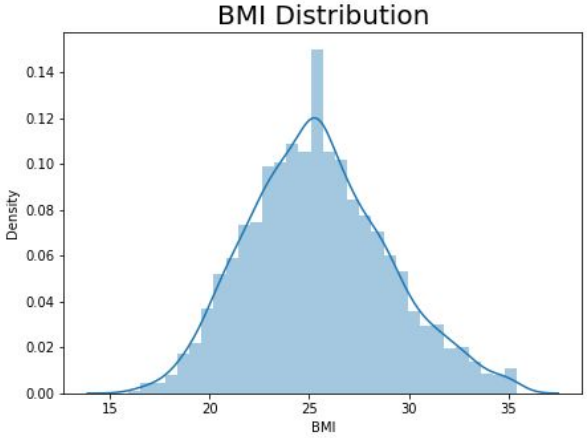
Numerical Features:

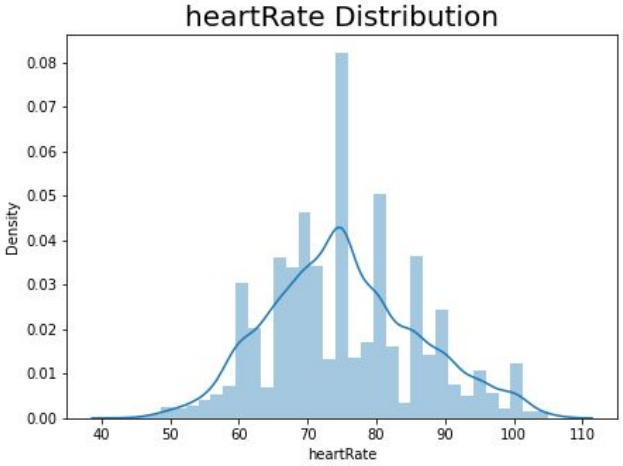


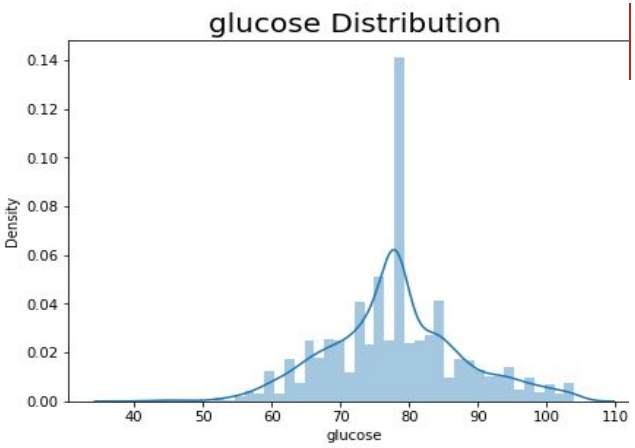






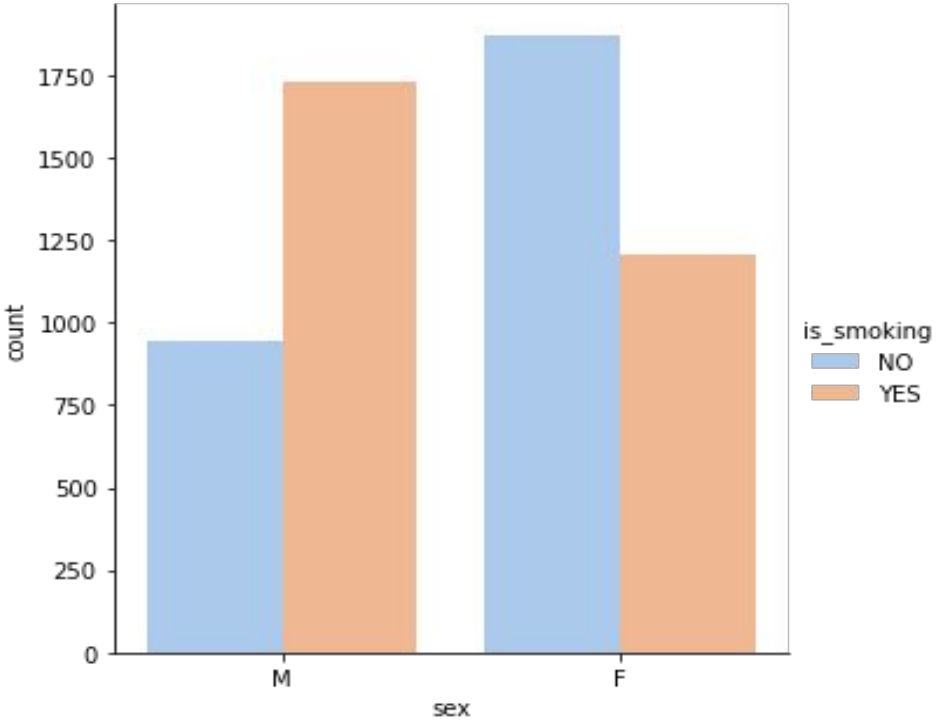




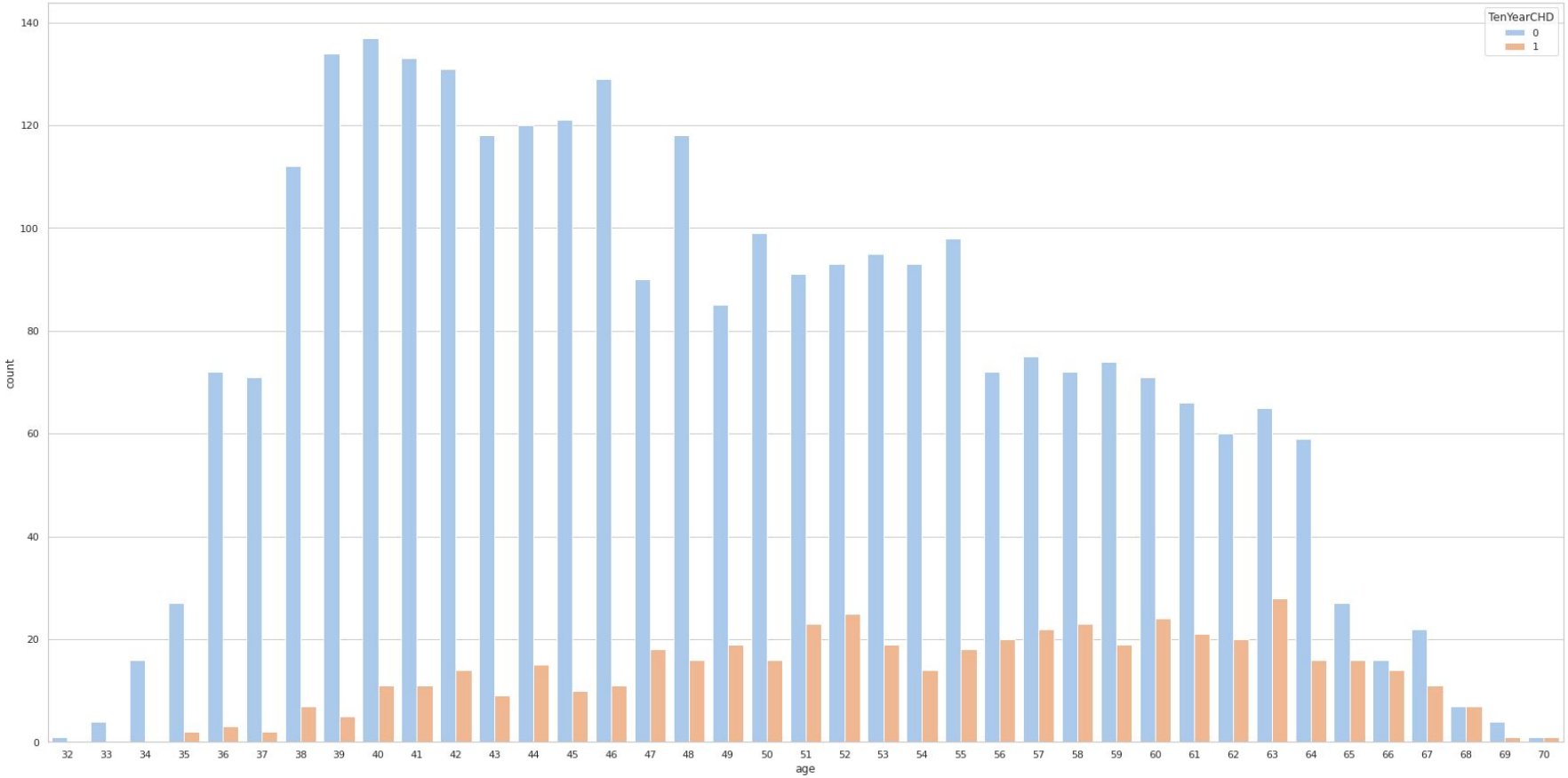


- totChol,sysBP,diaBP,BMI have uniform distribution while heartRate, glucose and cigsPerDay are unevenly distributed.
- cigsPerDay and sysBP are slightly right skewed.
- cigsPerDay has most data present in 0 as well as it is highly unevenly distributed.
- Heartrate also has highly uneven distribution most data is present around 80.

Smokers as per Gender

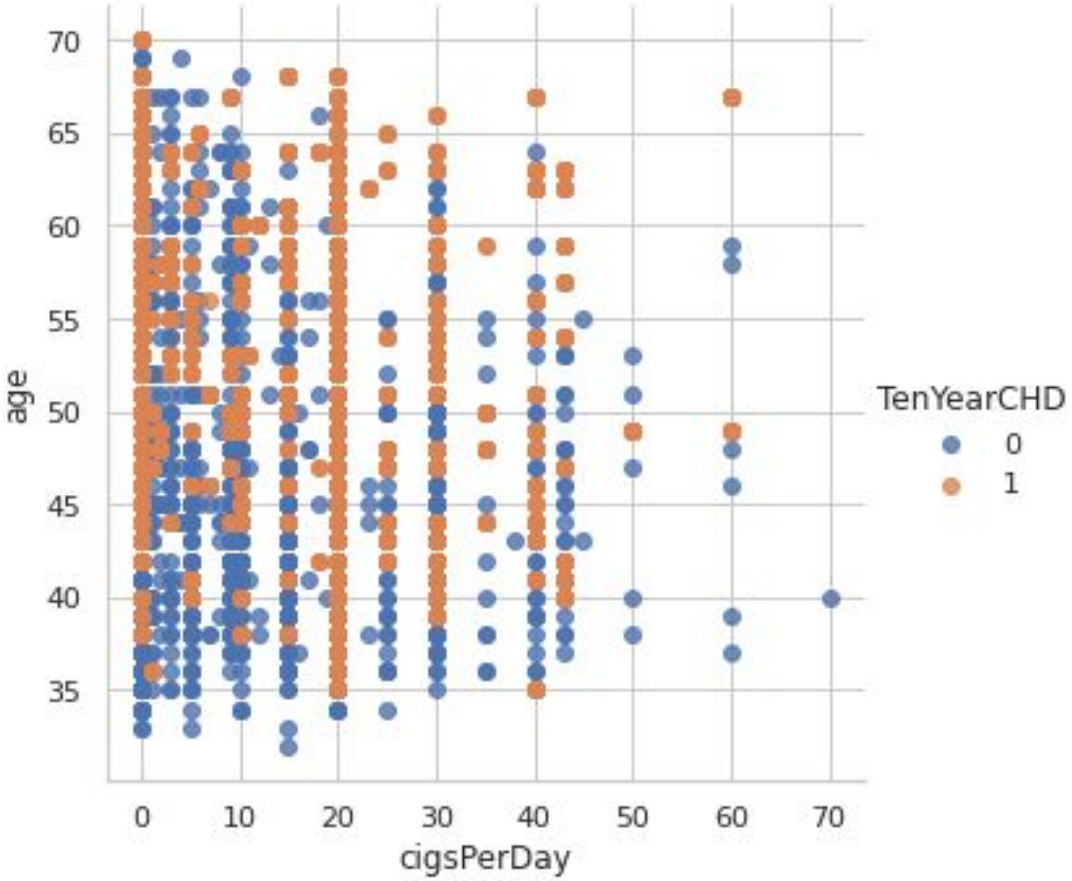


Chances of getting CHD according to their age

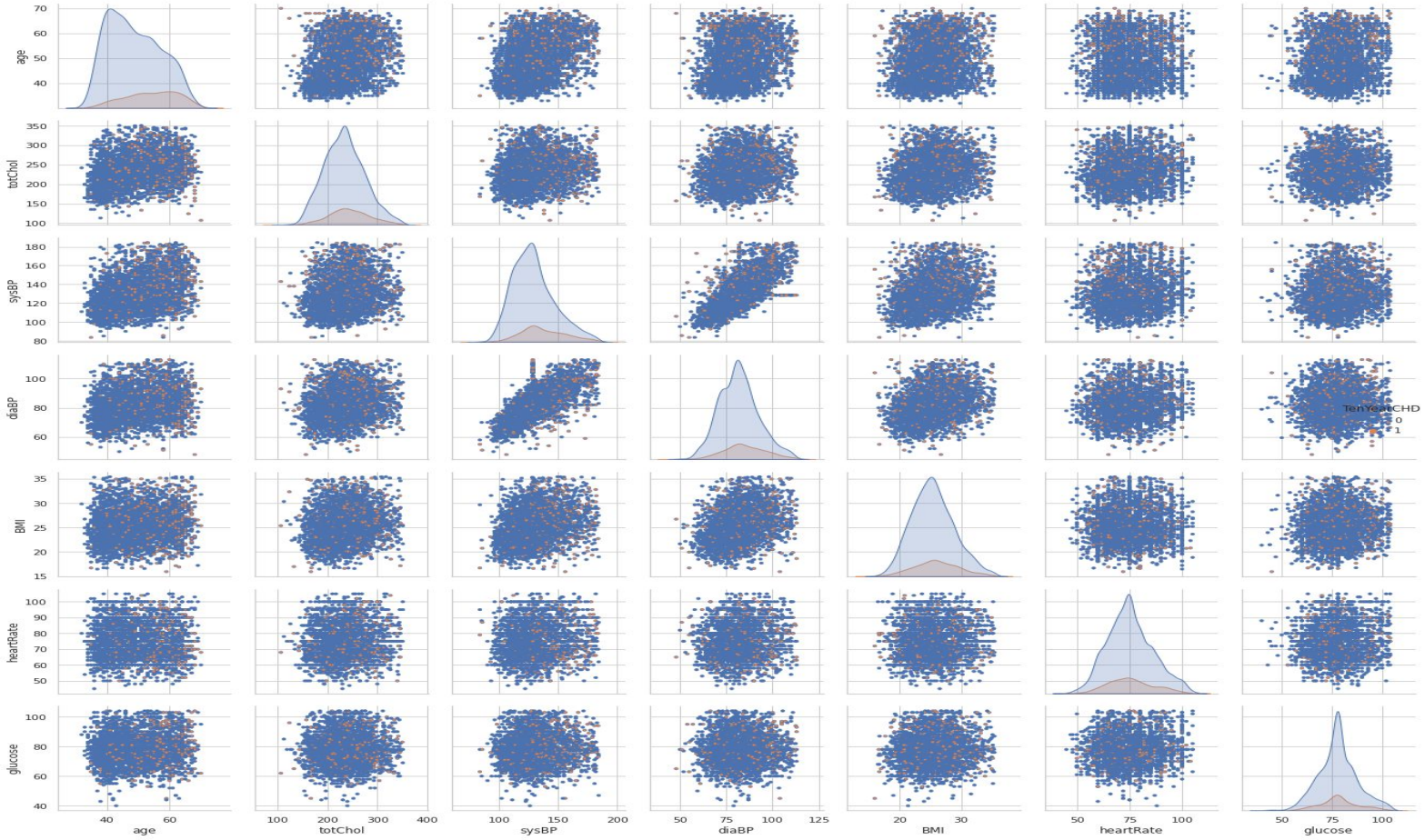


Relationship between age,cigsPerDay and CHD:

- Cigarette smokers belong to age group of 35 to 70.
- Those who smoke 20 cigarette a day have higher risk of detecting CHD.



Spread of Numerical features with target variable

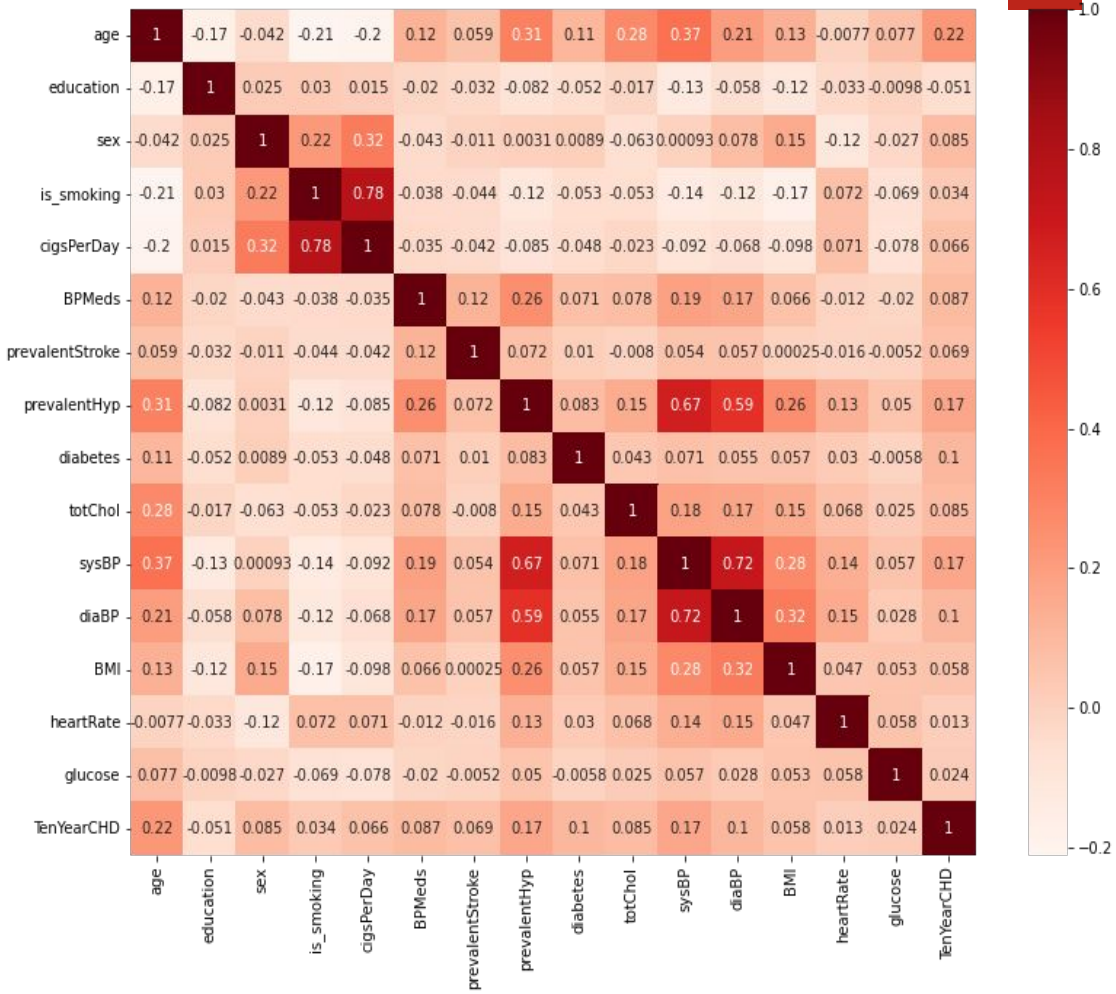


Feature Selection:

1.Multicollinearity.

2.Variance Threshold.

3.Chi-square.



Feature selection

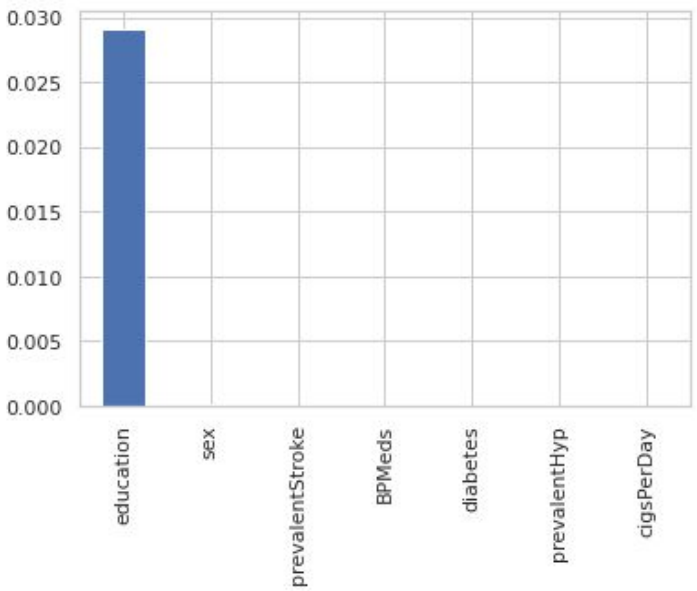
Using variance threshold:

- Variance threshold from sklearn is a simple baseline approach to feature selection.
- It removes all features which variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e., features that have the same value in all samples.

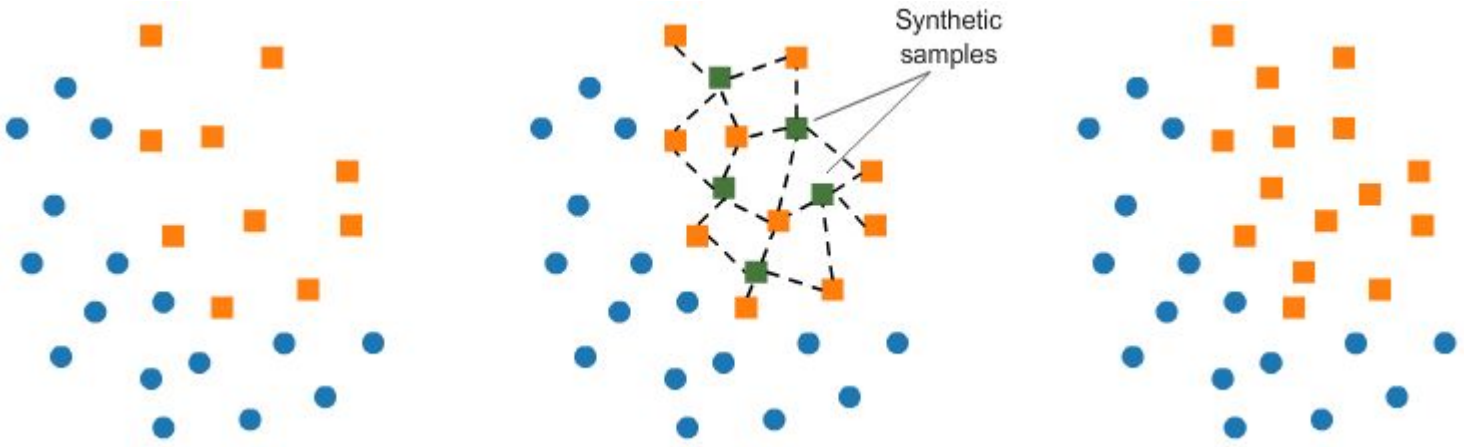
We saw that 99% of the Observations show one value 0.Therefore,this feature is almost constant.

Using Chi-square test:

Here,Since education has higher the p-value, it says that this variables is independent of the response and can not be considered for model training.



Target Variable: TenYearCHD



Since the target variable TenYearCHD was highly imbalanced and it would have been lead problem in future analysis it was handled using SOMTE

Models Used:

1.Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

2.K-nearest Neighbour:

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

3.Decision Tree:

Decision trees help you to evaluate your options. Decision Trees are excellent tools for helping you to choose between several courses of action. They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options.

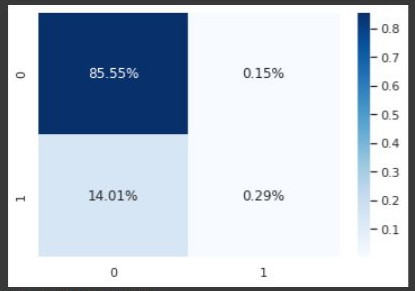
4.Random Forest Classifier:

A Random Forest is a reliable ensemble of multiple Decision Trees (or CARTs); though more popular for classification, than regression applications. Here, the individual trees are built via bagging (i.e. aggregation of bootstraps which are nothing but multiple train datasets created via sampling of records with replacement) and split using fewer features. The resulting diverse forest of uncorrelated trees exhibits reduced variance; therefore, is more robust towards change in data and carries its prediction accuracy to new data.

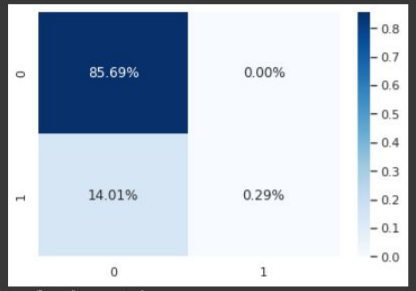
However, the algorithm does not work well for datasets having a lot of outliers, something which needs addressing prior to the model building.

Model Evaluation:

Logistic regression with and without hyperparameter tuning



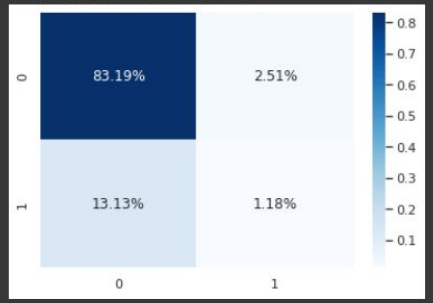

```
confussion matrix
[[580   1]
 [ 95   2]]


Accuracy of Logistic Regression: 85.84070796460178

Recall  2.0618556701030926

F1 Score 4.0

              precision    recall  f1-score   support

           0       0.86      1.00      0.92       581
           1       0.67      0.02      0.04        97

    accuracy                           0.86       678
   macro avg       0.76      0.51      0.48       678
weighted avg       0.83      0.86      0.80       678
```



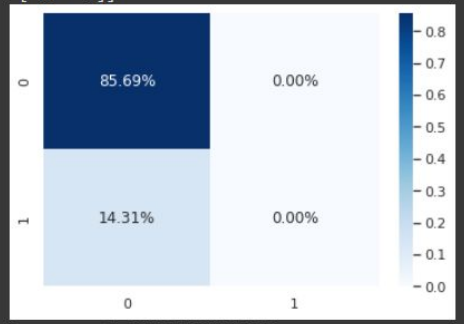

```
confussion matrix
[[581   0]
 [ 95   2]]


Accuracy of Logistic Regression: 85.9882005899705

Recall  2.0618556701030926

F1 Score 4.040404040404041

              precision    recall  f1-score   support

           0       0.86      1.00      0.92       581
           1       1.00      0.02      0.04        97

    accuracy                           0.86       678
   macro avg       0.93      0.51      0.48       678
weighted avg       0.88      0.86      0.80       678
```

KNN with and without hyperparameter tuning

```
Confusion Matrix:
[[564  17]
 [ 89   8]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.97      0.91       581
           1       0.32      0.08      0.13        97

    accuracy                           0.84       678
   macro avg       0.59      0.53      0.52       678
weighted avg       0.79      0.84      0.80       678

Accuracy: 0.8436578171091446
Recall 0.08247422680412371
F1 Score 0.13114754098360656
```



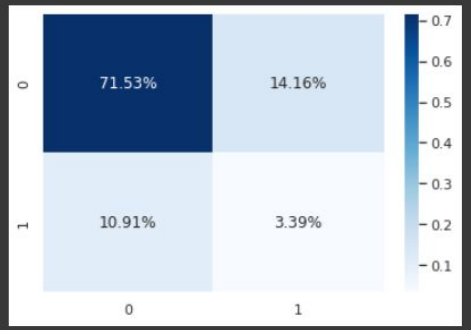

```
Confusion matrix
[[581   0]
 [ 97   0]]
```



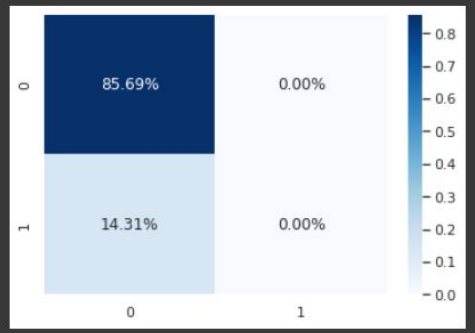

```
Accuracy: 0.8569321533923304
Recall 0.0
F1 Score 0.0
              precision    recall  f1-score   support

           0       0.86      1.00      0.92       581
           1       0.00      0.00      0.00        97

    accuracy                           0.86       678
   macro avg       0.43      0.50      0.46       678
weighted avg       0.73      0.86      0.79       678
```

Decision Tree with and without hyperparameter tuning

```
Accuracy 0.7492625368731564
Recall 0.23711340206185566
f1 score 0.21296296296296294
[[485  96]
 [ 74  23]]

              precision    recall  f1-score   support

           0       0.87      0.83      0.85       581
           1       0.19      0.24      0.21        97

    accuracy                           0.75       678
   macro avg       0.53      0.54      0.53       678
weighted avg       0.77      0.75      0.76       678
```




```
Accuracy 0.8569321533923304
Recall 0.0
f1 score 0.0
[[581   0]
 [ 97   0]]

              precision    recall  f1-score   support

           0       0.86      1.00      0.92       581
           1       0.00      0.00      0.00        97

    accuracy                           0.86       678
   macro avg       0.43      0.50      0.46       678
weighted avg       0.73      0.86      0.79       678
```



Ensemble Technique based on bagging:Random Forest Classifier

```
Accuracy on training set is : 0.9022861356932154
Accuracy on validation set is : 0.8480825958702065
Accuracy of Hyper-tuned Random Forest Classifier: 84.80825958702066

Recall 0.020618556701030927
f1 score 0.037383177570093455
              precision    recall  f1-score   support

           0       0.86      0.99      0.92       581
           1       0.20      0.02      0.04        97

    accuracy                           0.85       678
   macro avg       0.53      0.50      0.48       678
weighted avg       0.76      0.85      0.79       678
```

Model Observations:

AI

Observation 1:
For logistic regression the accuracy was 85.8407 which increased to 85.9882 after hyper parameter Tuning.

Observation 2:
For K- nearest neighbors the accuracy was 84.3657 which increased to 85.6932 after hyper parameter Tuning.

Observation 3:
For Decision Tree the accuracy was 74.9262 which increased to 85.6932 after hyper parameter Tuning.

Observation 4:
Ensemble Random Forest Classifier the accuracy was 84.8082.

Model Comparison:

	Model	Accuracy	Recall	F1_Score
0	Logistic Regression	85.840708	0.020619	0.040000
1	logisticWith tunning	85.988201	0.020619	0.040404
2	knn	84.365782	0.082474	0.131148
3	Knn with tunning	85.693215	0.000000	0.000000
4	Decision Tree	74.926254	0.237113	0.212963
5	Decision Tree with tunning	85.693215	0.000000	0.000000
6	Random Forest	84.808260	0.020619	0.037383

Conclusion:

- Those who were smoking 20 Cigarettes per day above were having more risk of detecting CHD.
- Males smokes more compared to female.
- Females and males have almost equal chances of getting CHD .
- Chances of getting CHD are mostly among the age group 35-50.While age group below 35 has lowest chances of getting CHD.
- Education feature is irrelevant for our target variable which we came know after applying Chi-square test.
- Features like Is_smoking and CigsPerDay were having multicollinearity so we chose we keep cigsPerDay out of these two

- Features like BPmeds, prevalentStroke and diabetes were not able to meet variance threshold of 99% so we discarded them.
- After Data processing , Feature Selection and Balancing our Target Variable we got approx same range of accuracy for different models.
- After Hyper parameter tuning we are able to improve the accuracy but by very small range.
- Random forest Bagging Ensemble technique is also giving the approx same result as our classification models.

Challenges faced:

- We have outliers in our dataset which accounts for 14% of data. So instead of discarding them we chose to impute them with median values of respective features.
- We have 2 categorical feature as string type which we labelled using label encoding.
- We have so many feature irrelevant for our model so we Adjust them like putting id feature as index and Those which have multicollinearity and values lower than variance threshold we discarded them from our model.
- Our target feature was highly Imbalanced. So we use SMOTE to balance it before training.
- Some computations were complex and had to be done correctly since wrong computations would have lead to loss of data.

- We Use several models with and without feature selection and tried different data processing steps. Out of all the process we selected the these as they were giving better result.
- We tried our model over simple classification algorithms and then tune them with hyperparameter the result was not so drastic change but it improves from simple models.
- We tried Different Ensemble techniques But all of them were giving approx same result so we chose to have Random Forest.