Capstone Project

Cardiovascular Risk Prediction

By- Jayesh Dahiwale

STEPS INVOLVED :

Steps that significantly contribute towards achieving the results:

1. Defining The Problem Statement.
2. Exploratory Data Analysis.
3. Applying The Data Pre-Processing Steps
4. Feature Selection And Transformation.
5. Classification Model Fitting.
6. Comparing The Metrics.
7. Selecting The Best Model.

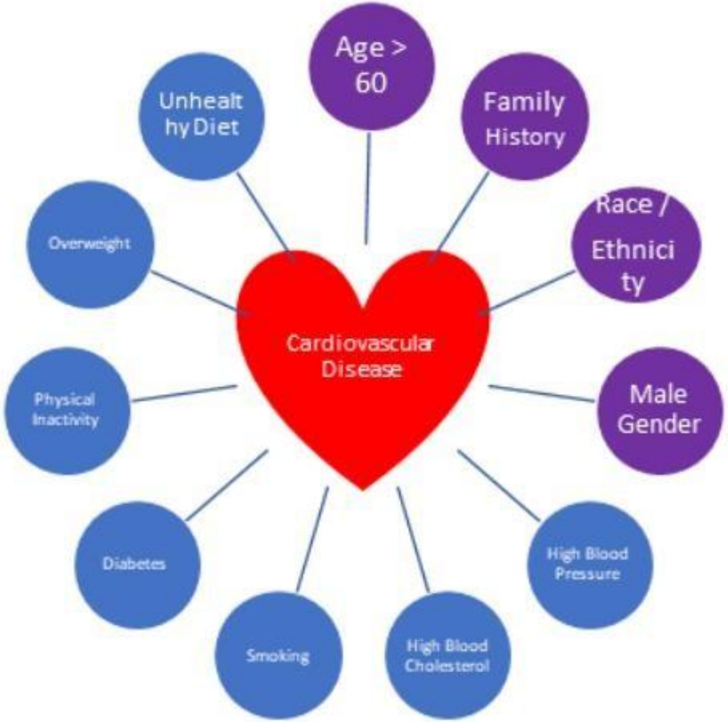
WHY DO WE NEED CARDIOVASCULAR RISK PREDICTION?

1. Predicting and diagnosing heart disease is the biggest challenge in the medical industry. There are many factors which influence heart diseases.

2. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms.

3. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

4. Machine learning can play a vital and accurate role in predicting chances of heart disease in coming potential years based upon the current way of living.



DATA PIPELINE:

1. **Data Processing-1**: In this initial step we went to look for different features available and tried to uncover their relevance with the target variable and we made id column as index column it is one of that most irrelevant features from analysis perspective.

2. **Data Processing-2:** During this stage, we looked for the data types of each feature and corrected them. After that comes the null value and outlier detection and treatment. For the null values imputation we used Mean, Median and Mode technique and for the outlier we used Capping method to handle the outliers without any loss to the data.

3. **EDA:** EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data. So, through this we have observed certain trends and dependencies and drawn certain conclusions from the dataset that will be useful for further processing.

4. **Feature Selection and Transformation:** During this stage, we went on to select the most relevant features using the chi-square test, information gain, extra trees classifier and next comes the feature scaling in order to bring down all the values in similar range. After that comes the treatment of class imbalance in the target variable that is done using random oversampling.

5. **Model Fitting and Performance Metric**: Since the data is transformed to an appropriate form therefore, we pass it to different classification models and calculate the metrics based on which we select a final model that could give us better prediction.

DATASET SUMMARY:

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
- The dataset provides the patients information.
- The data set consists of record of almost 3390 records and 17 features.
- Some of the features are Categorical in nature while other holds numeric data type.
- The target variable namely ‘TenYearCHD’ refers to whether the patient suffers from cardiovascular heart disease depending upon the values of current medical parameters.
- The dependent variable consists of the binary value where, 1-Risk of Cardiovascular Heart Disease and 0-No Risk of Cardiovascular Heart Disease.

INDEPENDENT VARIABLES:

These are divided in certain categories. All the dependent variables are listed below:

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).

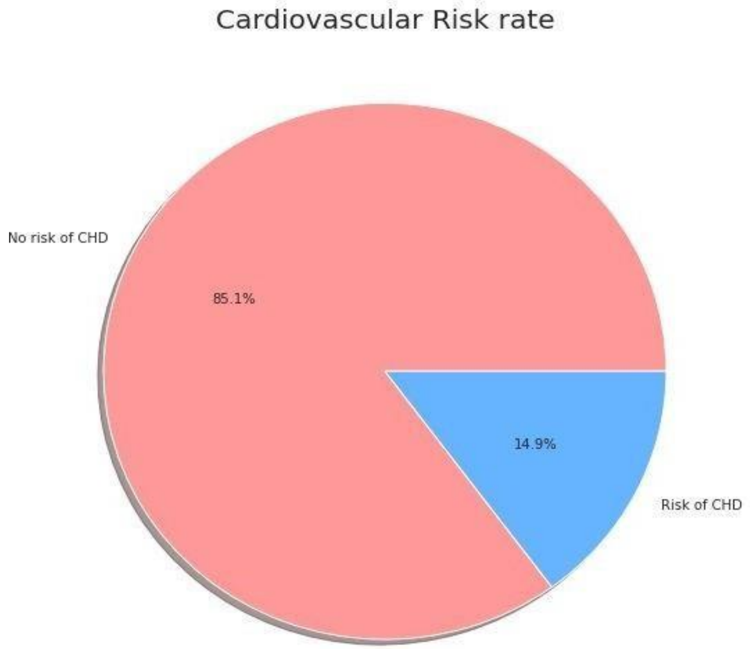
Medical(current):

- 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).

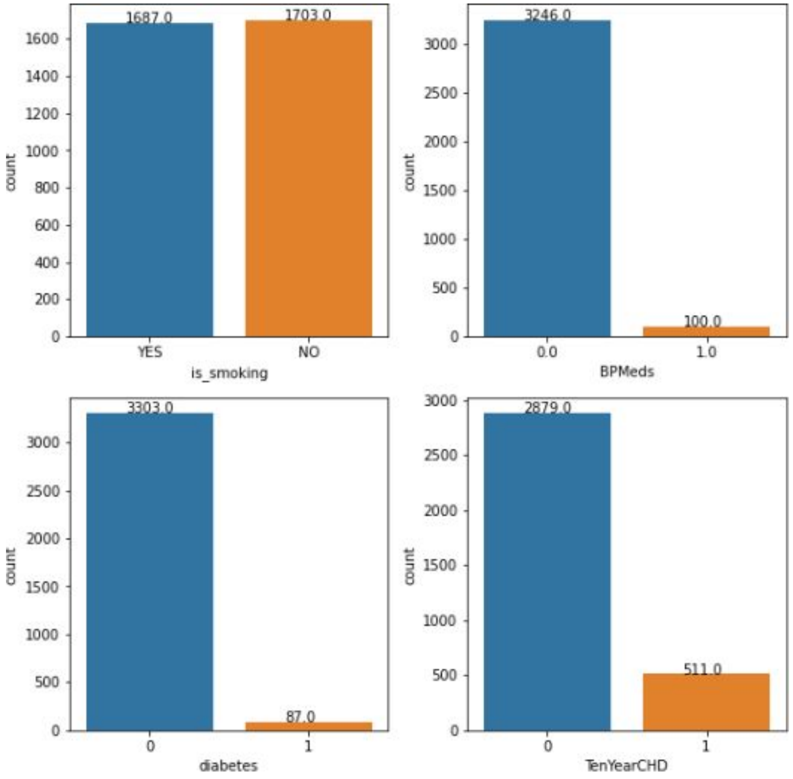
DEPENDENT VARIABLES:

- Our dependent variable i.e TenYearCHD that refers to the Risk of Cardiovascular Heart Disease in coming 10 years.
- 10-year risk of cardiovascular heart disease CHD is binary i.e. it only hold two discrete values:
- 1 - Yes: There is a risk of CHD

  0 - No: There is no risk of CHD

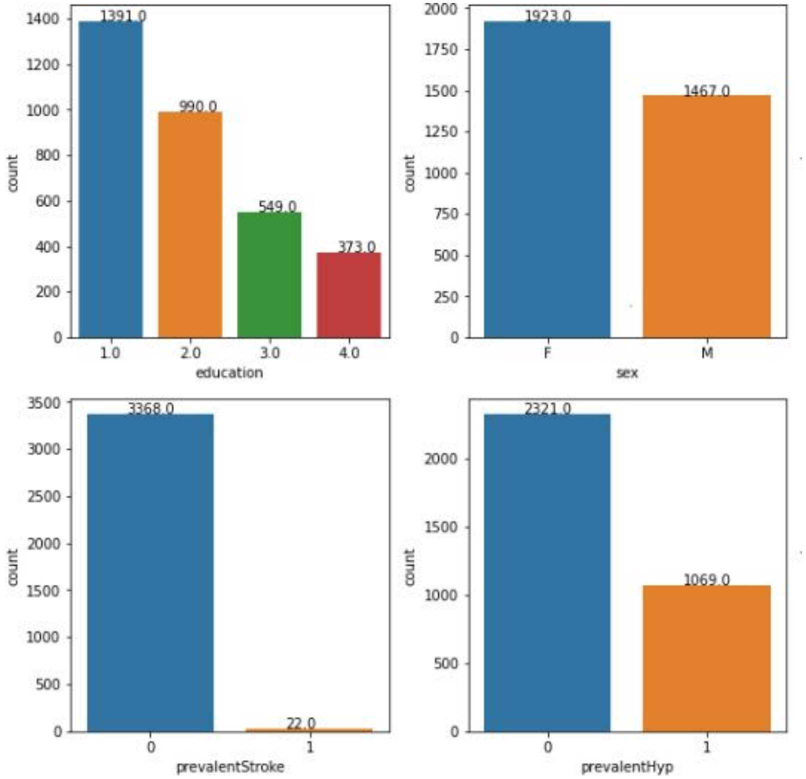


EXPLORATORY DATA ANALYSIS: (UNIVARIATE ANALYSIS)



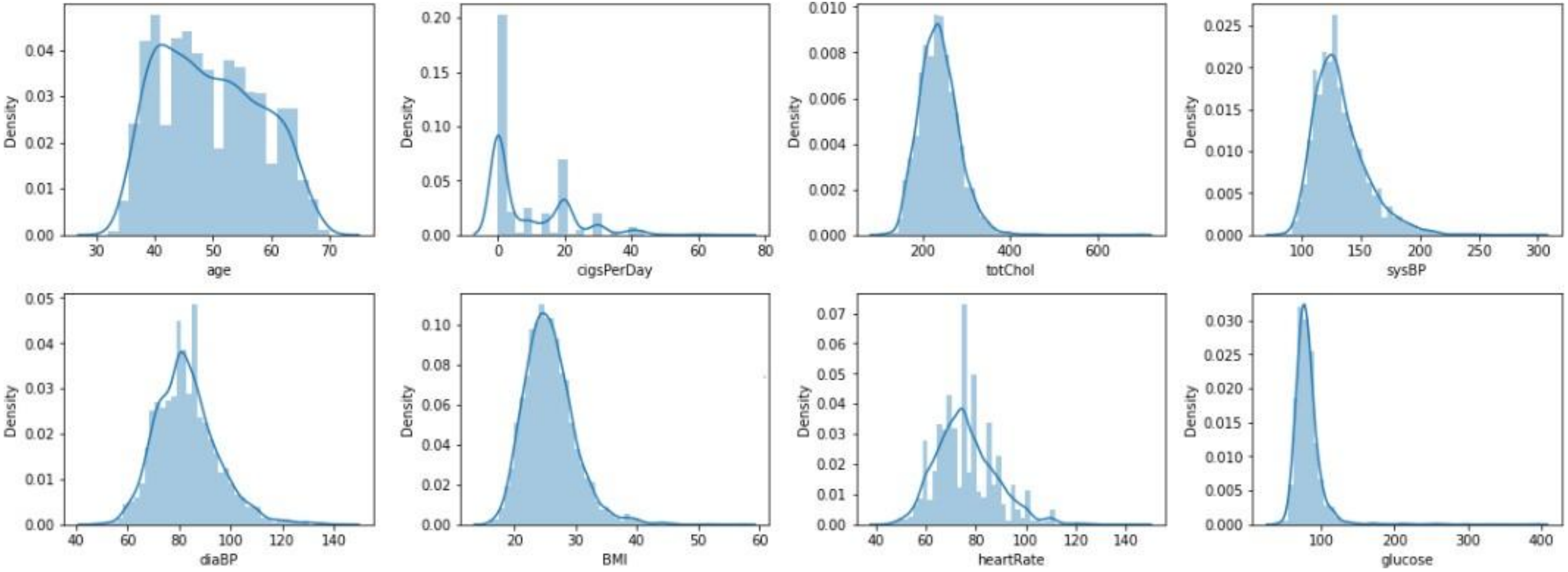
- There are a greater number of count is non-smokers than smokers, but the count for both the class is comparable.
- More than 3000 people are not on BP medication.
- A large number (> 3000) of the people do not have diabetes.
- Around 500 patients having 10 years risk of CHD.

EXPLORATORY DATA ANALYSIS: (UNIVARIATE ANALYSIS)



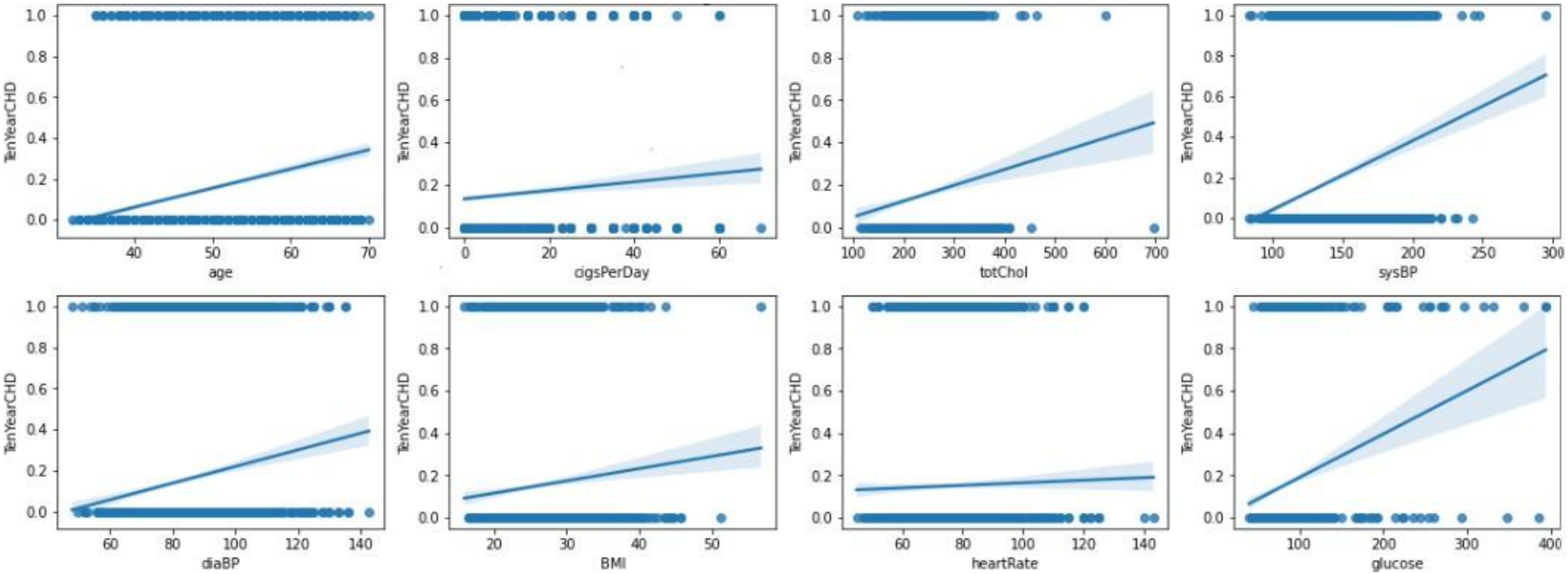
- Highest number of count belong from class 1 category.
- Females are more in proportion than male by a small margin.
- Only a small number of people have suffered a stroke previously.
- Around 1000 people were hypertensive.

EXPLORATORY DATA ANALYSIS: (UNIVARIATE ANALYSIS)



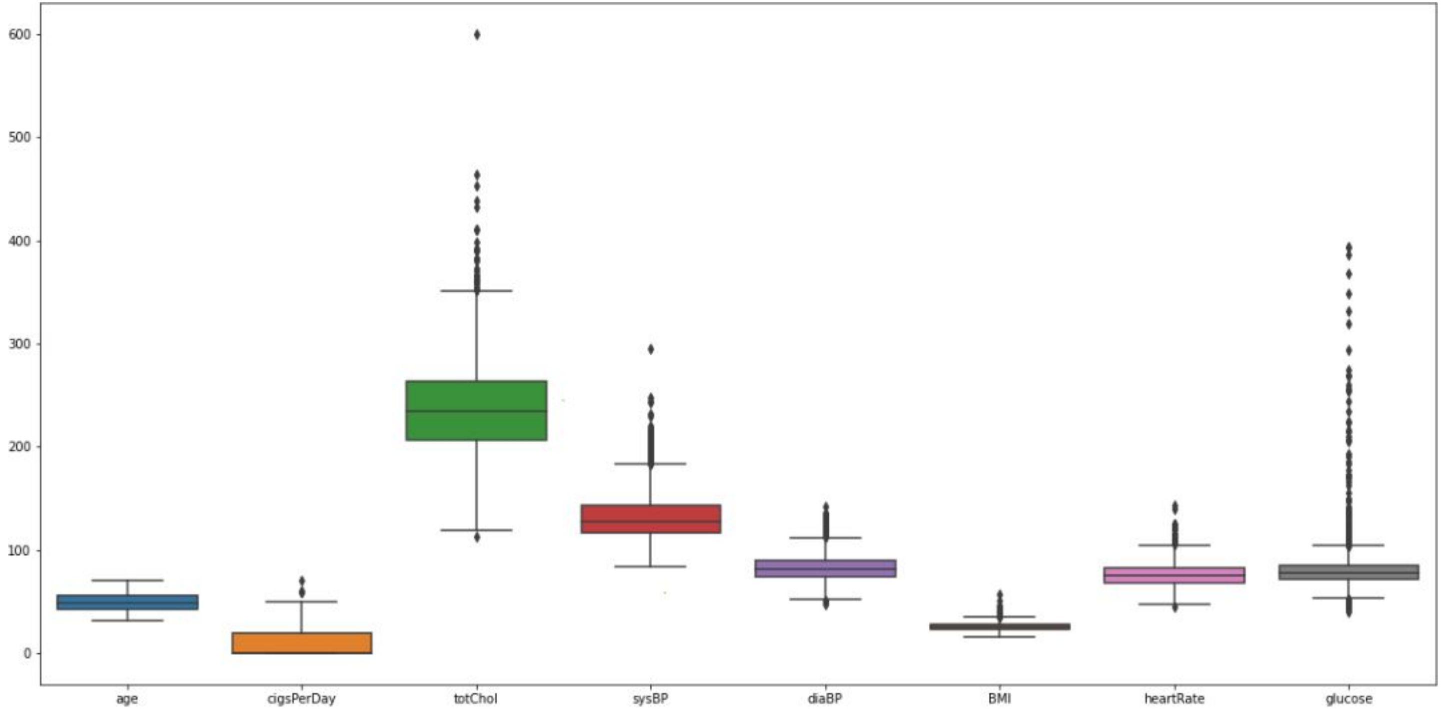
We can observe that most of the distributions are right skewed. totChol (total cholesterol) and BMI have roughly similar distributions. Glucose have a highly right skewed distribution. It shows Glucose has a lot of outliers.

EXPLORATORY DATA ANALYSIS: (BIVARIATE ANALYSIS)



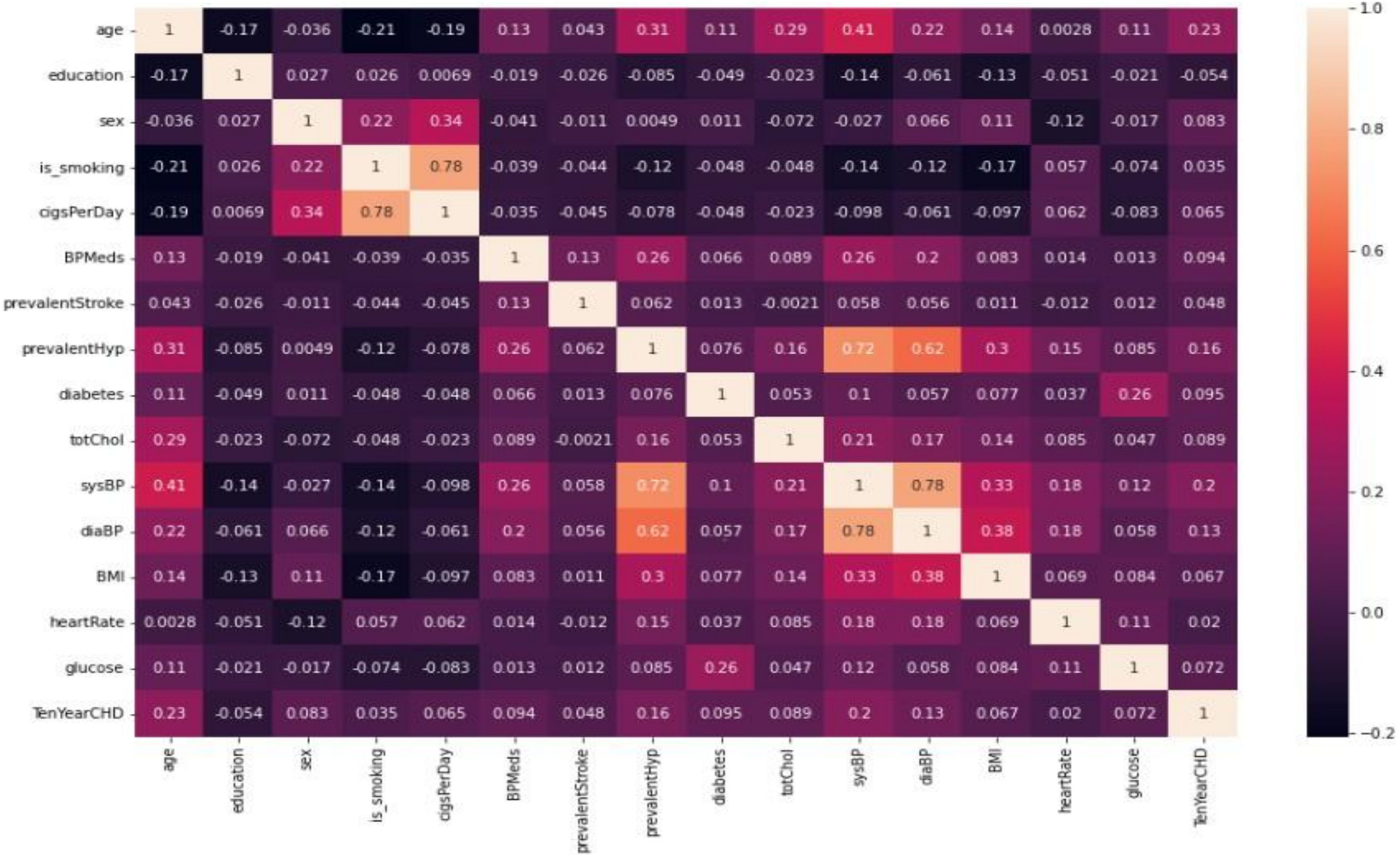
Almost all Independent continuous variables show relation with Target variable (TenYearCHD).

OUTLIERS DETECTION:



A **capping method** for outliers is a technique used to handle extreme values in a dataset that may be causing issues with the analysis or modeling of the data. One common method for capping outliers is to replace the outlier values with a maximum or minimum value that is within a certain range of the other values in the dataset.

CORRELATION MATRIX:



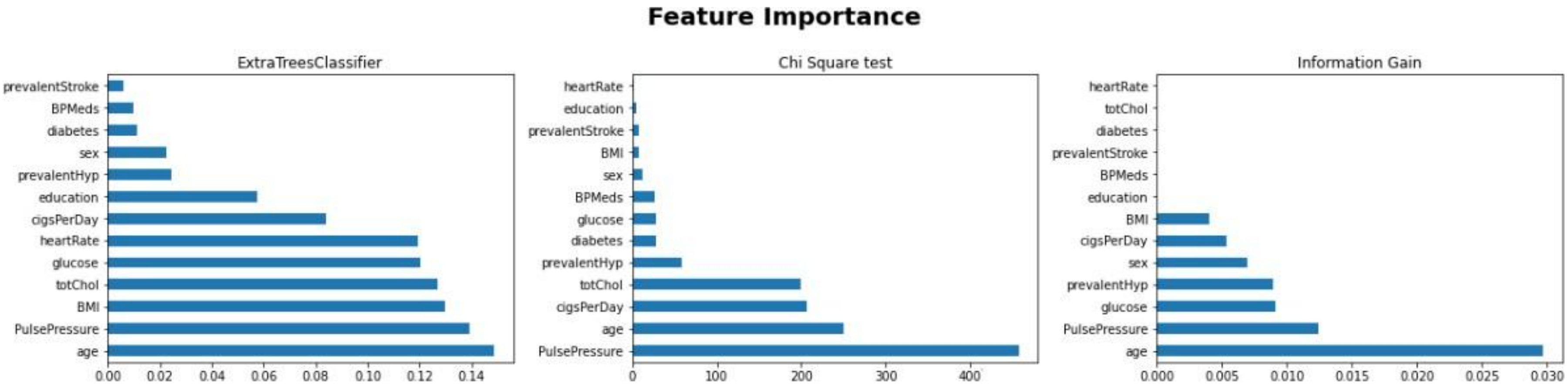
	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	-0.17	-0.036	-0.21	-0.19	0.13	0.043	0.31	0.11	0.29	0.41	0.22	0.14	0.0028	0.11	0.23
education	-0.17	1	0.027	0.026	0.0069	-0.019	-0.026	-0.085	-0.049	-0.023	-0.14	-0.061	-0.13	-0.051	-0.021	-0.054
sex	-0.036	0.027	1	0.22	0.34	-0.041	-0.011	0.0049	0.011	-0.072	-0.027	0.066	0.11	-0.12	-0.017	0.083
is_smoking	-0.21	0.026	0.22	1	0.78	-0.039	-0.044	-0.12	-0.048	-0.048	-0.14	-0.12	-0.17	0.057	-0.074	0.035
cigsPerDay	-0.19	0.0069	0.34	0.78	1	-0.035	-0.045	-0.078	-0.048	-0.023	-0.098	-0.061	-0.097	0.062	-0.083	0.065
BPMeds	0.13	-0.019	-0.041	-0.039	-0.035	1	0.13	0.26	0.066	0.089	0.26	0.2	0.083	0.014	0.013	0.094
prevalentStroke	0.043	-0.026	-0.011	-0.044	-0.045	0.13	1	0.062	0.013	-0.0021	0.058	0.056	0.011	-0.012	0.012	0.048
prevalentHyp	0.31	-0.085	0.0049	-0.12	-0.078	0.26	0.062	1	0.076	0.16	0.72	0.62	0.3	0.15	0.085	0.16
diabetes	0.11	-0.049	0.011	-0.048	-0.048	0.066	0.013	0.076	1	0.053	0.1	0.057	0.077	0.037	0.26	0.095
totChol	0.29	-0.023	-0.072	-0.048	-0.023	0.089	-0.0021	0.16	0.053	1	0.21	0.17	0.14	0.085	0.047	0.089
sysBP	0.41	-0.14	-0.027	-0.14	-0.098	0.26	0.058	0.72	0.1	0.21	1	0.78	0.33	0.18	0.12	0.2
diaBP	0.22	-0.061	0.066	-0.12	-0.061	0.2	0.056	0.62	0.057	0.17	0.78	1	0.38	0.18	0.058	0.13
BMI	0.14	-0.13	0.11	-0.17	-0.097	0.083	0.011	0.3	0.077	0.14	0.33	0.38	1	0.069	0.084	0.067
heartRate	0.0028	-0.051	-0.12	0.057	0.062	0.014	-0.012	0.15	0.037	0.085	0.18	0.18	0.069	1	0.11	0.02
glucose	0.11	-0.021	-0.017	-0.074	-0.083	0.013	0.012	0.085	0.26	0.047	0.12	0.058	0.084	0.11	1	0.072
TenYearCHD	0.23	-0.054	0.083	0.035	0.065	0.094	0.048	0.16	0.095	0.089	0.2	0.13	0.067	0.02	0.072	1

FEATURE SELECTION:

Most important features is age and Pulse Pressure to predict target variable.

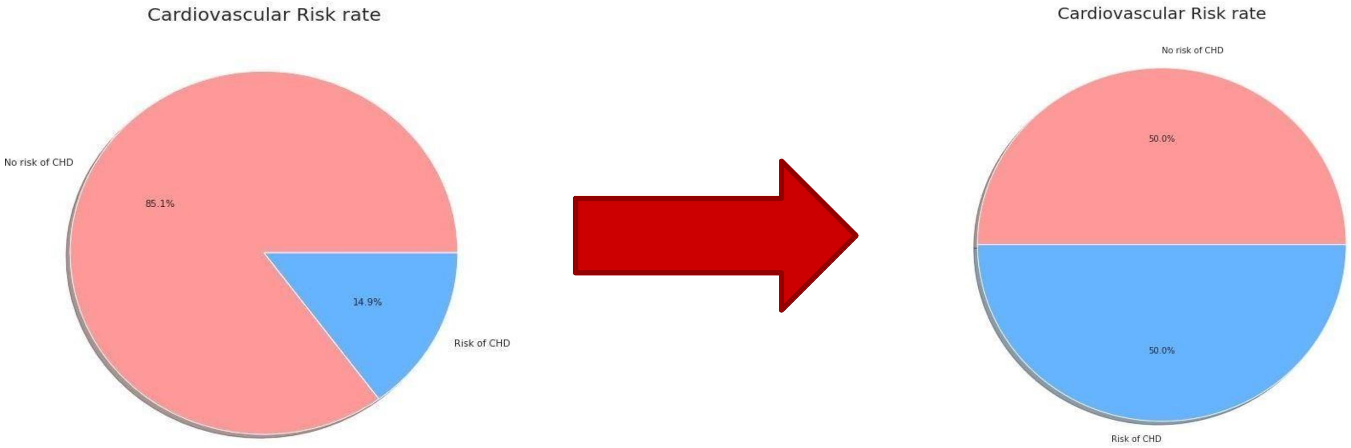
In Healthcare industry, every single data is important to analyze or to make prediction on target variable. In this case the dataset is related to medical domain, the entries in this dataset are person specific and the values vary among different individuals and all the features are very much important.

We are taking all features to train the model except multicollinearity one.



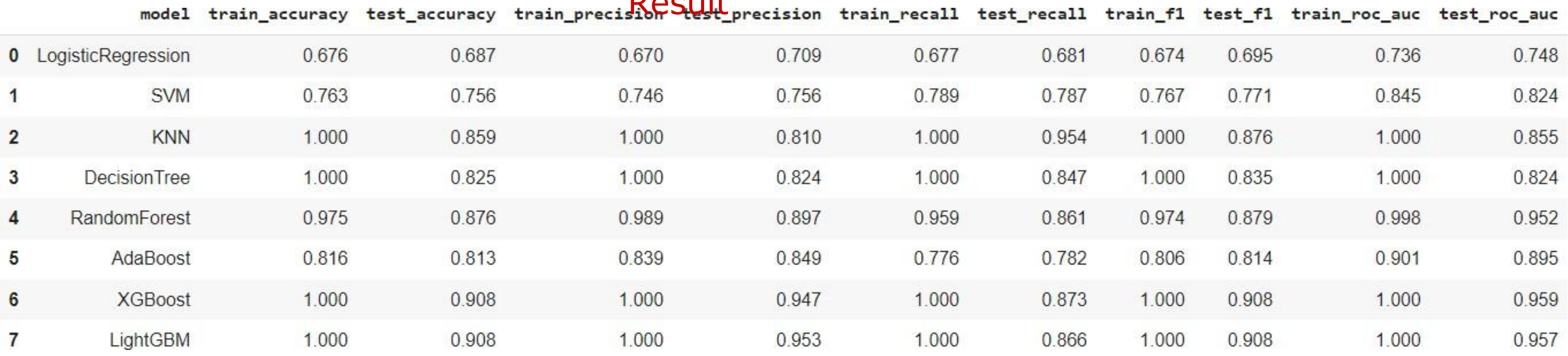
DEALING WITH CLASS IMBALANCE:

- The dataset that we have majority count belong from negative class and minority count belong from positive class.
- We have used the SMOTE for oversampling the minority class.
- SMOTE (Synthetic Minority Oversampling Technique) is another technique to oversample the minority class. Simply adding duplicate records of minority class often don’t add any new information to the model. In SMOTE new instances are synthesized from the existing data.
- SMOTE works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors. Repeat the steps until data is balanced.



CLASSIFICATION MODEL AND MODEL PERFORMANCE:

Model Performance Result

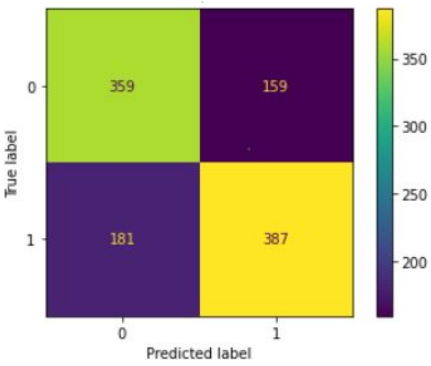


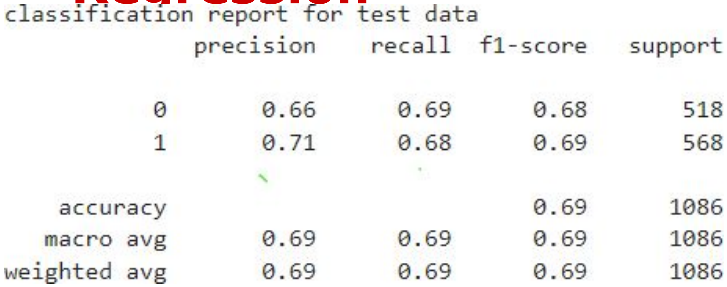
	model	train_accuracy	test_accuracy	train_precision	test_precision	train_recall	test_recall	train_f1	test_f1	train_roc_auc	test_roc_auc
0	LogisticRegression	0.676	0.687	0.670	0.709	0.677	0.681	0.674	0.695	0.736	0.748
1	SVM	0.763	0.756	0.746	0.756	0.789	0.787	0.767	0.771	0.845	0.824
2	KNN	1.000	0.859	1.000	0.810	1.000	0.954	1.000	0.876	1.000	0.855
3	DecisionTree	1.000	0.825	1.000	0.824	1.000	0.847	1.000	0.835	1.000	0.824
4	RandomForest	0.975	0.876	0.989	0.897	0.959	0.861	0.974	0.879	0.998	0.952
5	AdaBoost	0.816	0.813	0.839	0.849	0.776	0.782	0.806	0.814	0.901	0.895
6	XGBoost	1.000	0.908	1.000	0.947	1.000	0.873	1.000	0.908	1.000	0.959
7	LightGBM	1.000	0.908	1.000	0.953	1.000	0.866	1.000	0.908	1.000	0.957

- In Medical domain recall score is the most important evaluation metrics.
- While dealing with Imbalanced data Accuracy score doesn't help much in Imbalanced class situations.
- In cases where positives are as important as negatives, balanced accuracy is a better metric for this F1 score. F1 is a good scoring metric for imbalanced data.

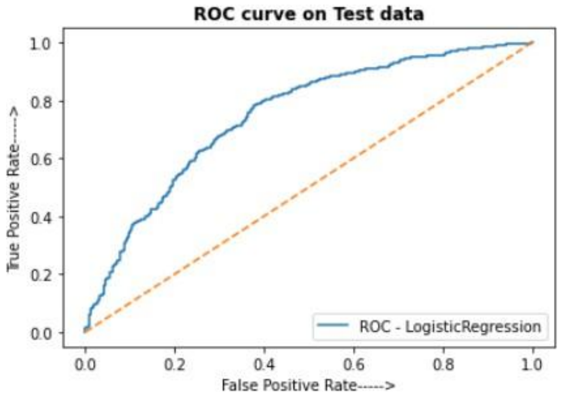
Selecting Final model as KNN because it has Highest Recall score, and we don't want to mispredict a person safe when he has the risk of 10 years of CHD)

1. Logistic Regression

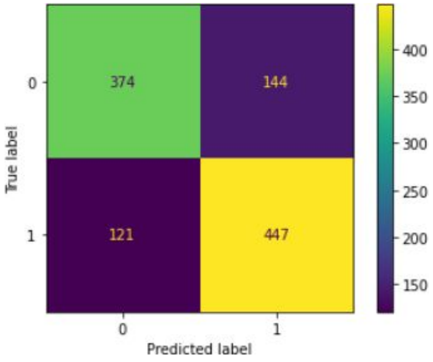


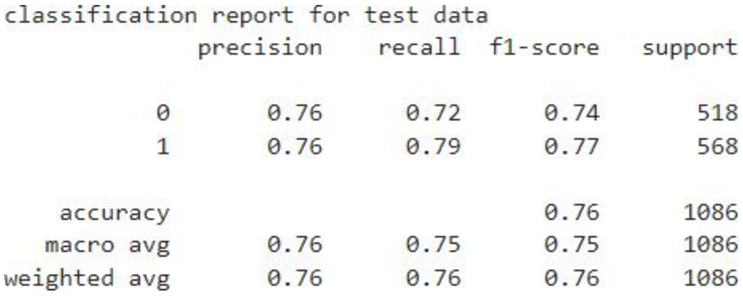


```
classification report for test data
              precision    recall  f1-score   support

           0       0.66      0.69      0.68       518
           1       0.71      0.68      0.69       568

    accuracy                           0.69      1086
   macro avg       0.69      0.69      0.69      1086
weighted avg       0.69      0.69      0.69      1086
```

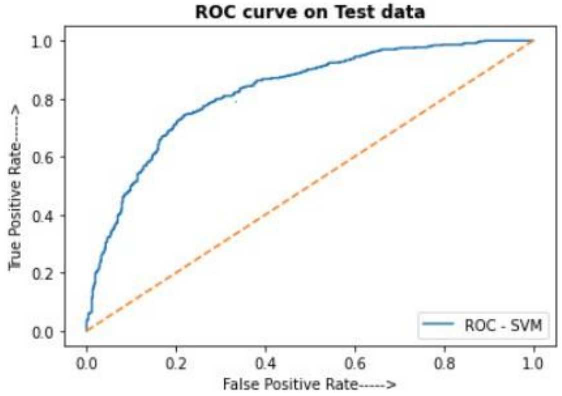


2. Support Vector Machine

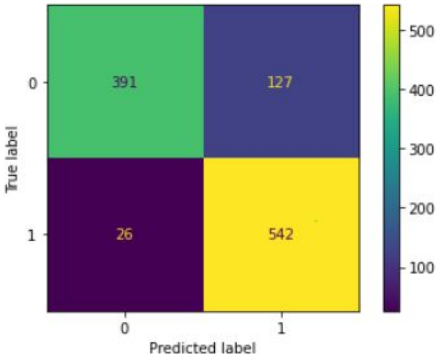


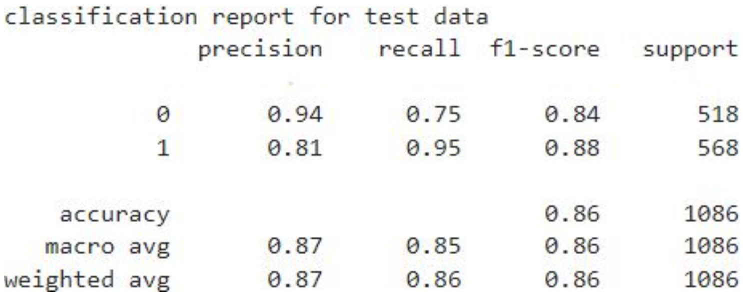


```
classification report for test data
              precision    recall  f1-score   support

           0       0.76      0.72      0.74       518
           1       0.76      0.79      0.77       568

    accuracy                           0.76      1086
   macro avg       0.76      0.75      0.75      1086
weighted avg       0.76      0.76      0.76      1086
```

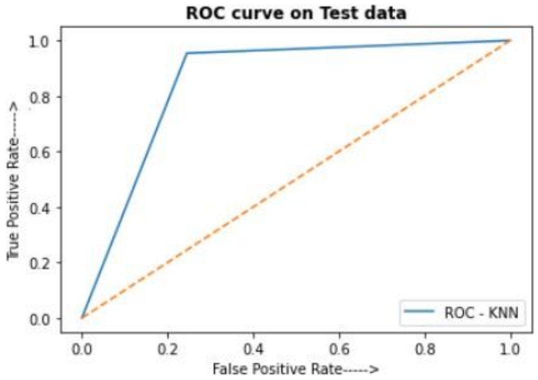


3. K-Nearest Neighbours

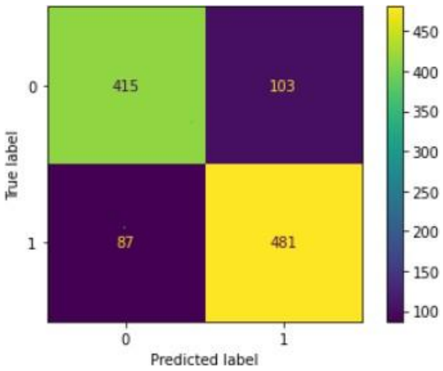


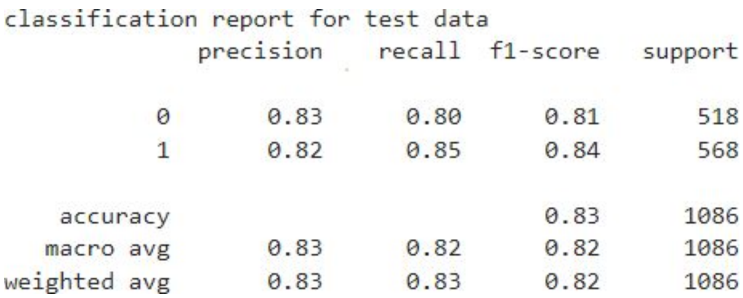


```
classification report for test data
              precision    recall  f1-score   support

           0       0.94      0.75      0.84       518
           1       0.81      0.95      0.88       568

    accuracy                           0.86      1086
   macro avg       0.87      0.85      0.86      1086
weighted avg       0.87      0.86      0.86      1086
```

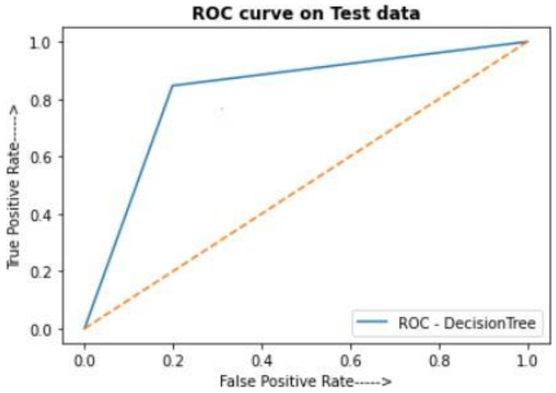


4. Decision Tree

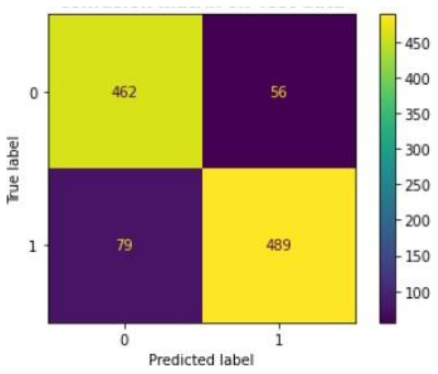


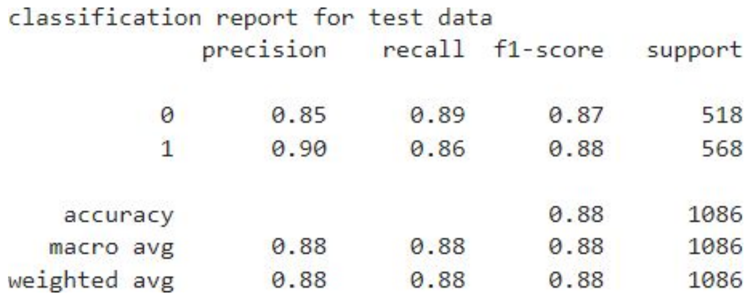


```
classification report for test data
              precision    recall  f1-score   support

           0       0.83      0.80      0.81       518
           1       0.82      0.85      0.84       568

    accuracy                           0.83      1086
   macro avg       0.83      0.82      0.82      1086
weighted avg       0.83      0.83      0.82      1086
```

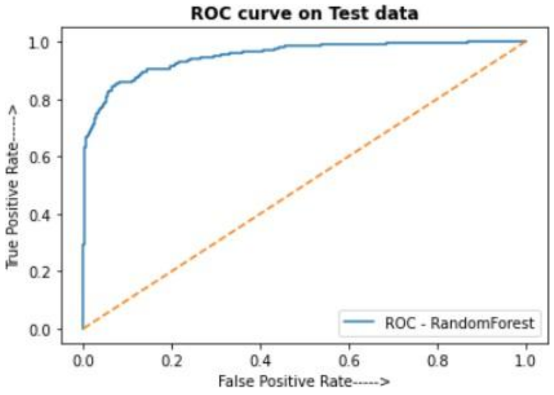


5. Random Forest

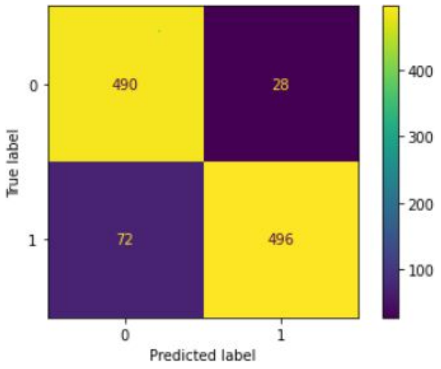


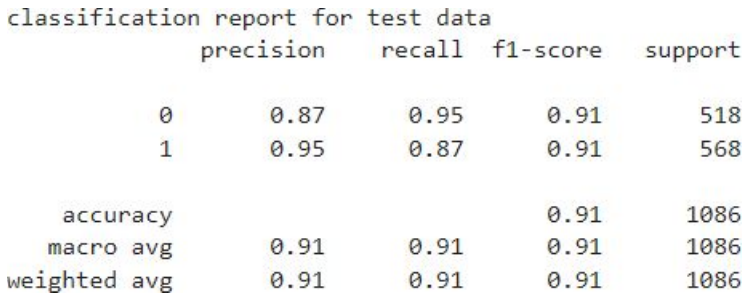


```
classification report for test data
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       518
           1       0.90      0.86      0.88       568

    accuracy                           0.88      1086
   macro avg       0.88      0.88      0.88      1086
weighted avg       0.88      0.88      0.88      1086
```

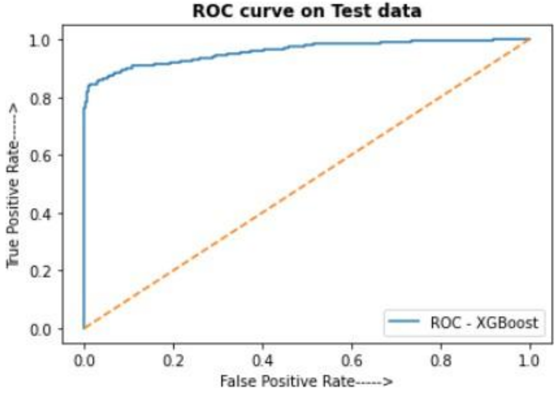


6. Xtreme Gradient Boosting

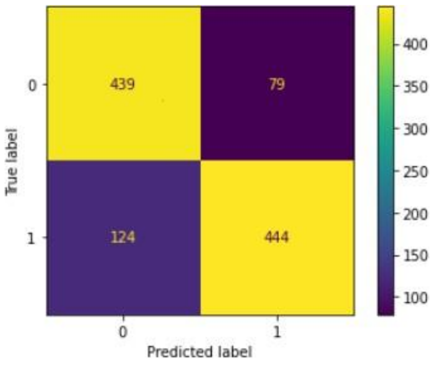


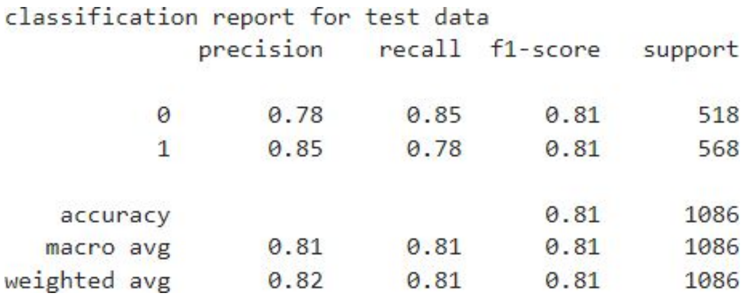


```
classification report for test data
              precision    recall  f1-score   support

           0       0.87      0.95      0.91       518
           1       0.95      0.87      0.91       568

    accuracy                           0.91      1086
   macro avg       0.91      0.91      0.91      1086
weighted avg       0.91      0.91      0.91      1086
```

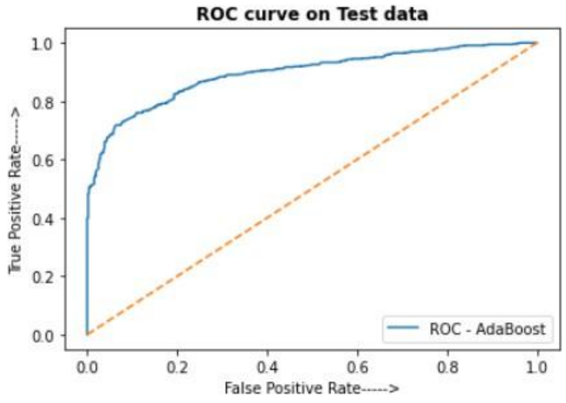


7. Adaptive Boosting

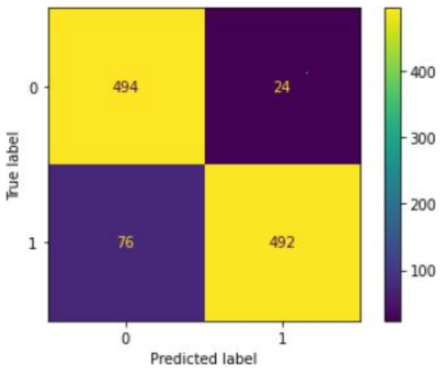


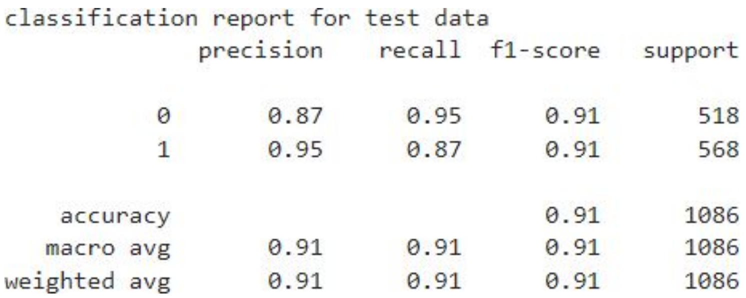


```
classification report for test data
              precision    recall  f1-score   support

           0       0.78      0.85      0.81       518
           1       0.85      0.78      0.81       568

    accuracy                           0.81      1086
   macro avg       0.81      0.81      0.81      1086
weighted avg       0.82      0.81      0.81      1086
```

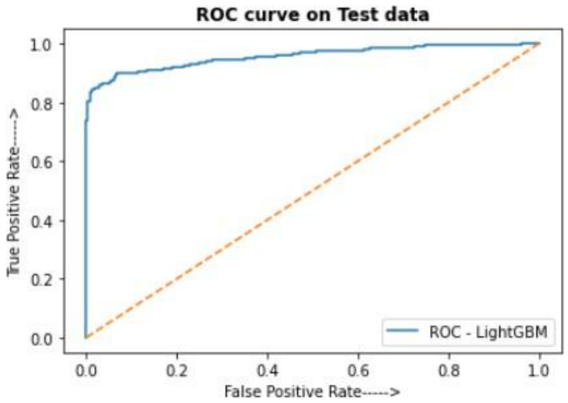


8. Light Gradient Boosting





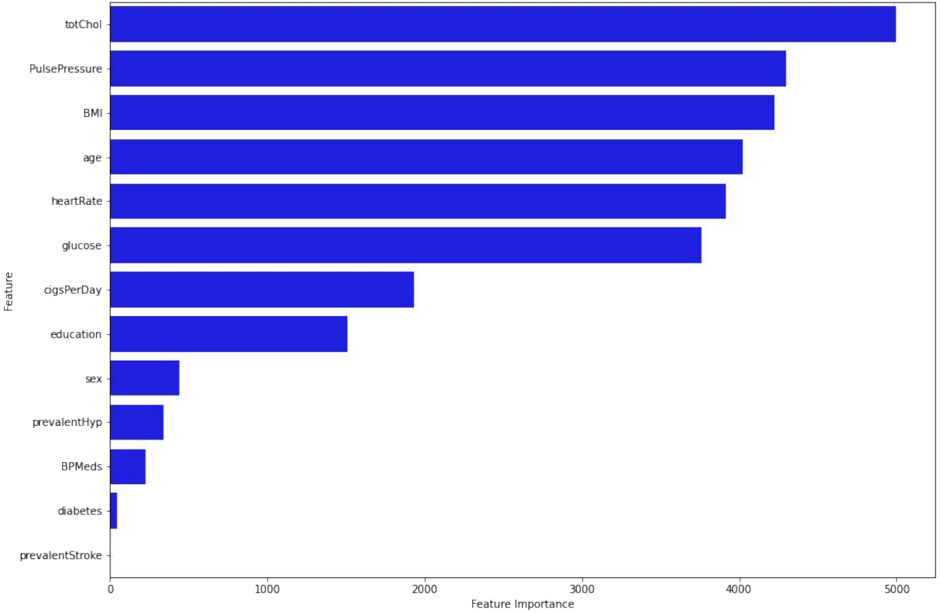
```
classification report for test data
              precision    recall  f1-score   support

           0       0.87      0.95      0.91       518
           1       0.95      0.87      0.91       568

    accuracy                           0.91      1086
   macro avg       0.91      0.91      0.91      1086
weighted avg       0.91      0.91      0.91      1086
```



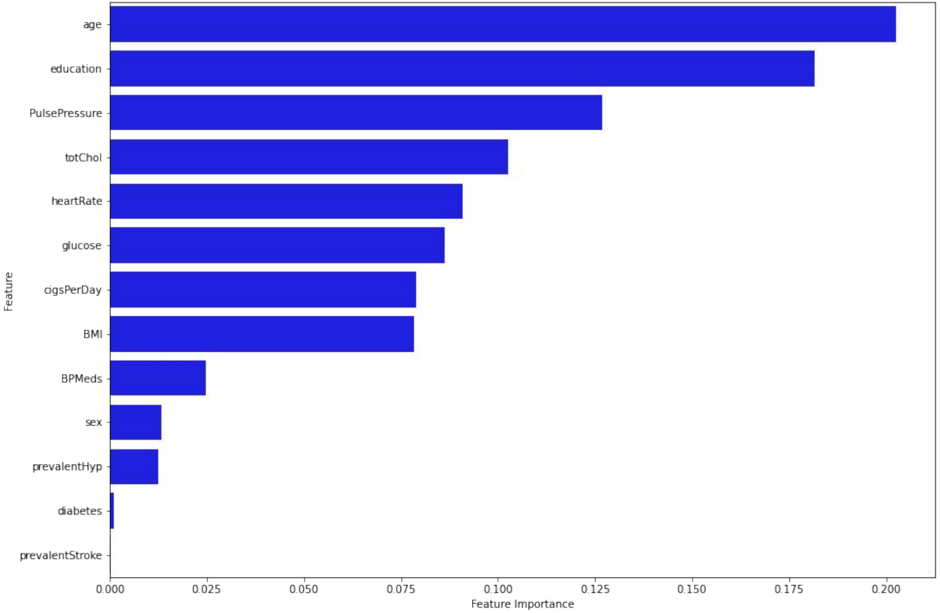
HYPER TUNED MODEL:

Feature Importance:

Light Gradient Boosting



Random Forest



OBSERVATIONS:

1. As we can observe from the model performance table, out of all model Logistic Regression gave us lowest performance.

2. Talking of Decision tree it gave us good results than Logistic regression. If we compare with recall score even that case also it performed well than SVM & AdaBoost model.

3. K-Nearest Neighbours showed the best results as the best accuracy if we don’t consider hyper tuned models but as far recall is considered, it gave us highest recall score out of all models i.e. 0.954.

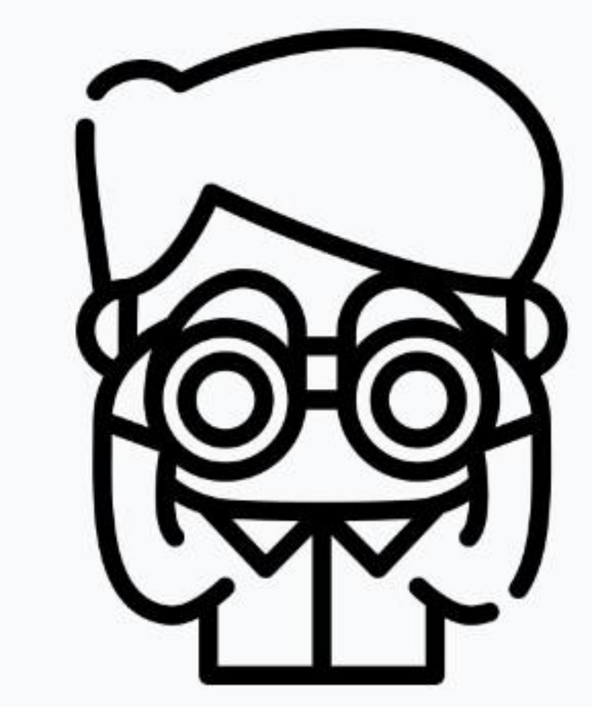
0.954

OBSERVATIONS:

4. We've noticed that XBG Classifier is the standout performer among all models with f1-score of 0.908 and recall score of 0.873 on test data. it's safe to say that XGB Classifier provides an optimal solution to our problem.

5. Out of the tree-based algorithms, LGBM Classifier and Random Forest Classifier was also providing an optimal solution towards achieving our objective. We were able to achieve an f1-score of 0.908 and 0.884, respectively.

6. Looking at the business problem recall is utmost important to us as there should be no case where a person having risk of CHD left unattended. Therefore, we choose KNN as final model it give the highest value of recall i.e. 0.954.

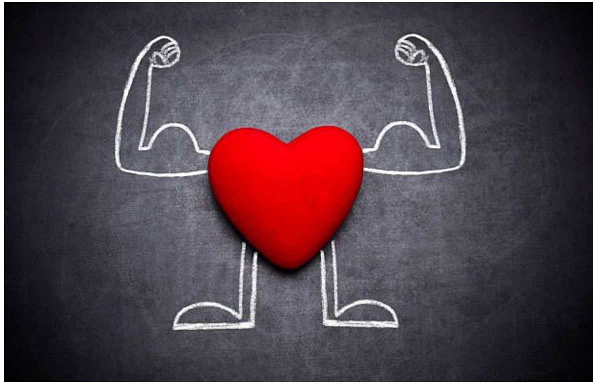


CONCLUSION:

1. Better model can be developed that can predict the risk of coronary heart disease with the help of the subject matter experts we can engineer an extensive amount of variable that could make our prediction more accurate.

2. In the Medical domain (**more focus towards the reducing False Negative values, as we don't want to mispredict a person safe when he has the risk**) here the recall score is the most importance. KNN, XGB, LGBM & Random Forest gave the best recall score 0.954 ,0.873 ,0.866, 0.863, respectively.

3. The **models that can be deployed according to our study is KNN because it has highest Recall score** and It's okay to classify a healthy person as having 10-year risk of coronary heart disease CHD (false positive) and following up with more medical tests, but it is not definitely okay to miss identifying a disease patient or classifying a disease patient as healthy (false negative).



ThankYou