Capstone Project 3

Supervised ML - classification

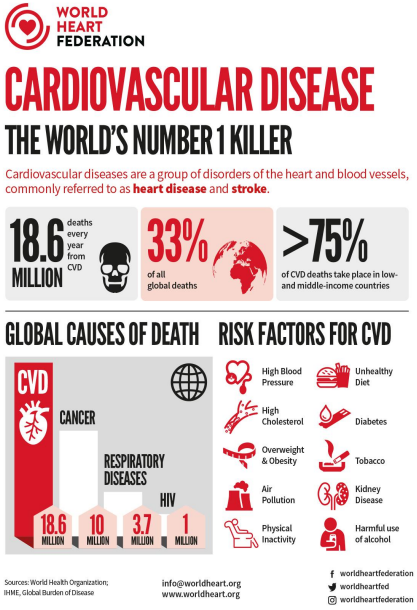
Cardiovascular risk prediction

by
Syed Adnan S

Problem statement:

- Coronary heart diseases (CHDs) are the leading cause of death globally, taking an estimated 18.6 million lives each year, which accounts for 33% of all the global deaths.

- Therefore It is important to detect cardiovascular diseases as early as possible so that management with counselling and medicines can begin.
- Our main aim here is to predict if a patient has a ten year risk of future coronary heart diseases based on a set of metrics.



Understanding the data:

- To increase the efficiency of our analysis we will first have to understand the data and also check if there are some corruptions in the data and if any found we will try to treat it.
- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.
- Each attribute(column) is a potential risk factor. These attributes include demographic, behavioral, and medical risk factors.
- Now we will look at what each column means.

The columns involved:

Demographic:

1. Sex.
2. Age.

Medical (Current):

1. Tot Chol.
2. Sys BP.
3. Dia BP.
4. BMI.
5. Heart rate.
6. Glucose.

Behavioral:

1. is_smoking.
2. Cigs Per Day.

Medical (history):

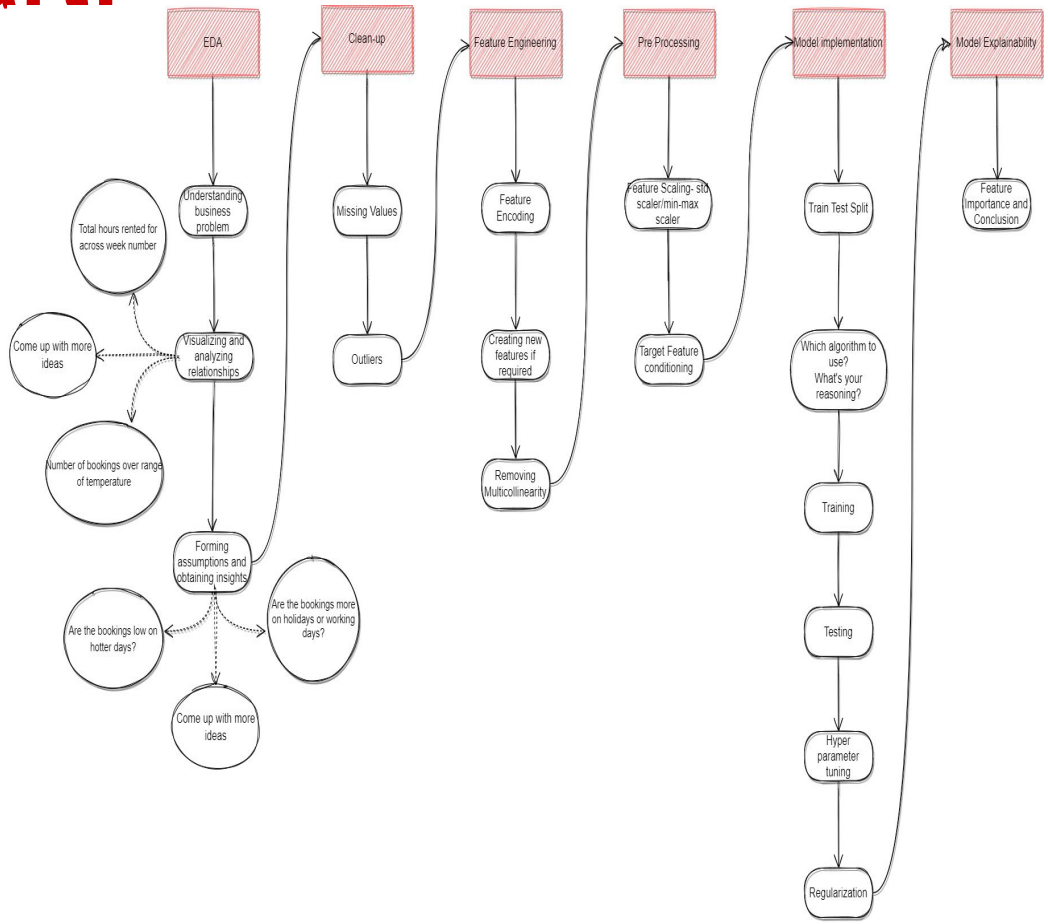
1. BP Meds.
2. Prevalent Stroke.
3. Prevalent Hyp.
4. Diabetes.

Target variable

Ten Year CHD

Project Flowchart:

1. Initial preparations.
2. EDA.
3. Clean up.
4. Feature Engineering.
5. Pre processing the data.
6. Model implementation.
7. Model explainability.



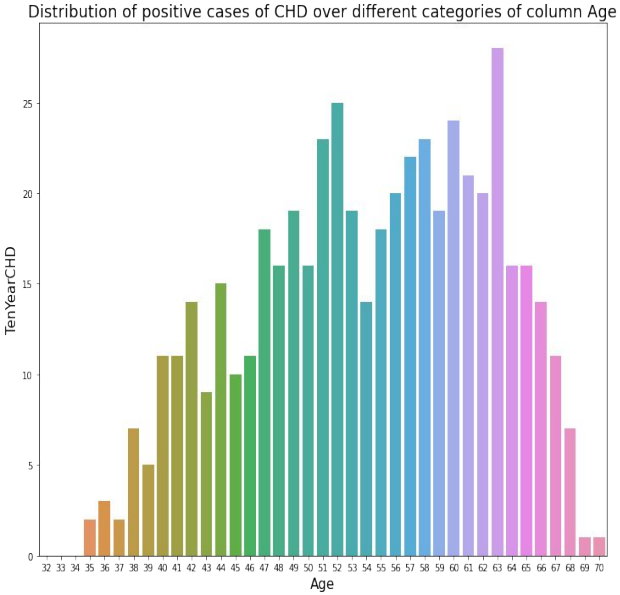
1. Initial preparation:

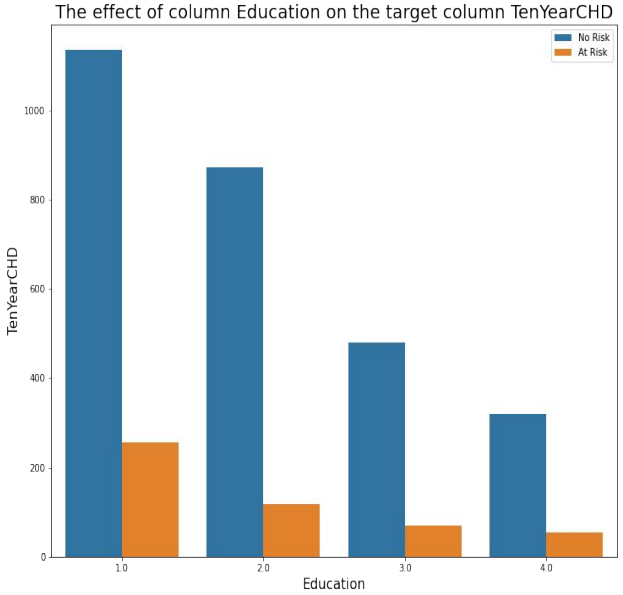
- In this section I've loaded in the dependencies, like pandas, seaborn, and many more from the scikit learn library.
- The next step was to mount the drive where the data was stored.
- After mounting the drive I used the pandas.read_csv() function to read the data given to us in csv format.

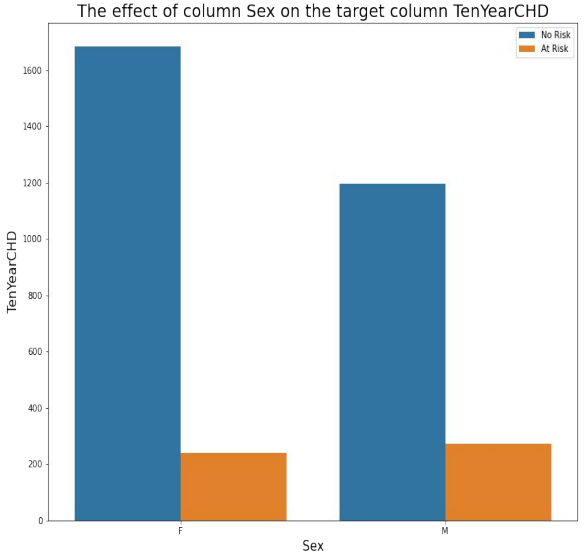
Note: The data for this project is given to us by the company, AlmaBetter.

2. EDA:

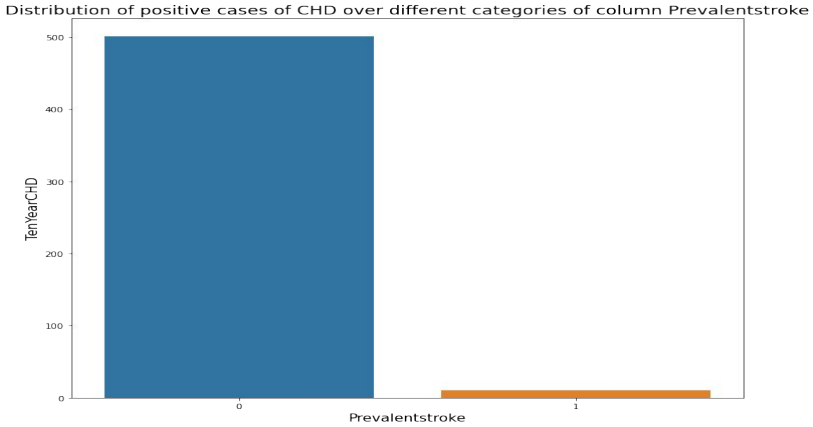
- In this step I have done exploratory data analysis on the data to see if I can find some valuable insights that can be directly applied to reduce the chances of having a positive CHD risk factor.
- I have plotted each attribute against the target variable to see how it affects the target variable.

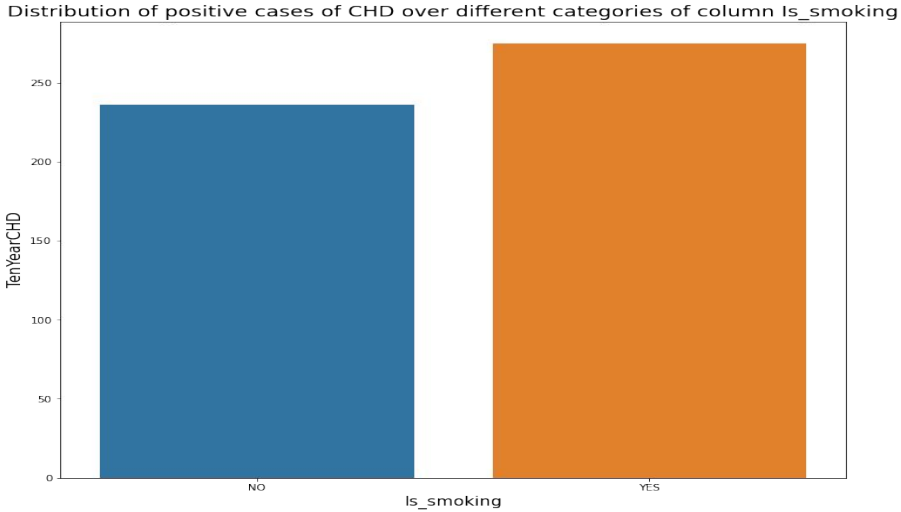


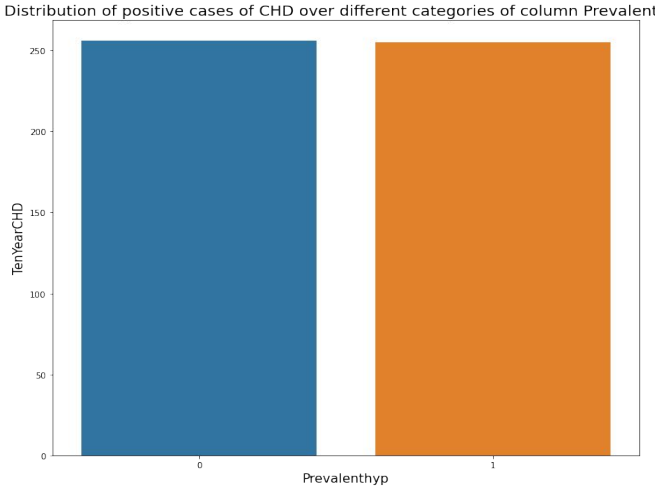


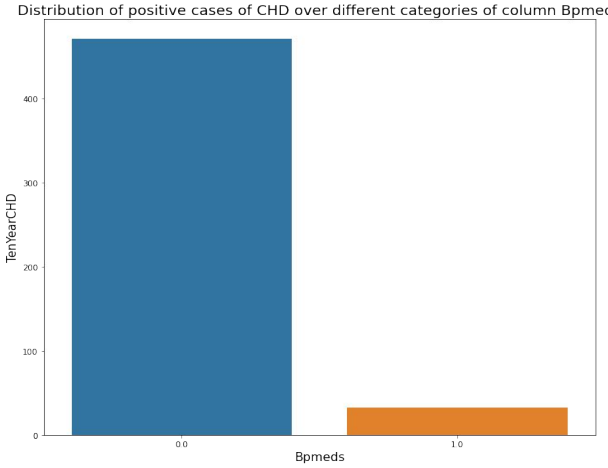


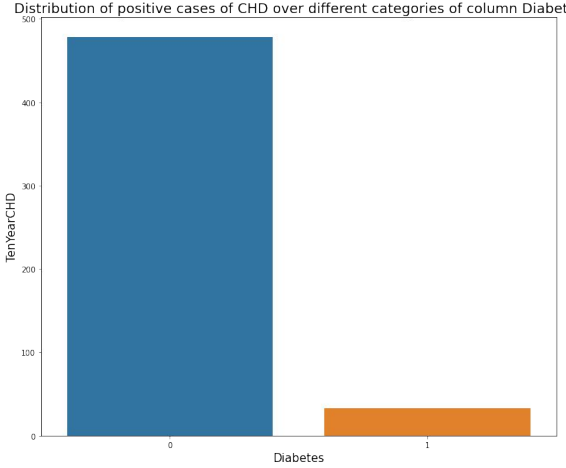
2. EDA(Contd):


Distribution of positive cases of CHD over different categories of column Prevalentstroke
TenYearCHD
Prevalentstroke
0
1
0
100
200
300
400
500

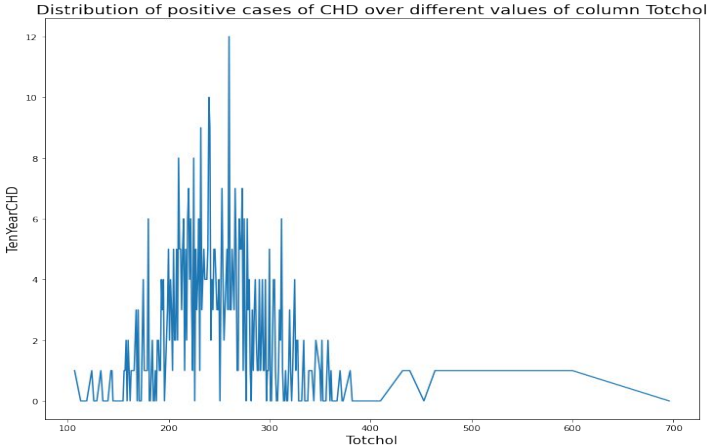

Distribution of positive cases of CHD over different categories of column Is_smoking
TenYearCHD
Is_smoking
NO
YES
0
50
100
150
200
250

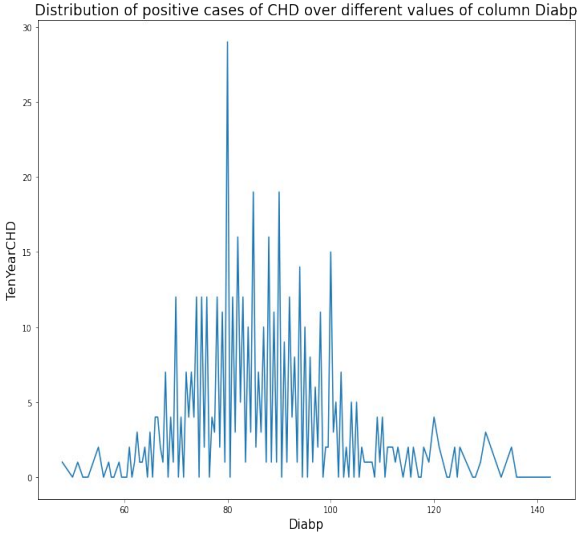

Distribution of positive cases of CHD over different categories of column Prevalenthyp
TenYearCHD
Prevalenthyp
0
1
0
50
100
150
200
250

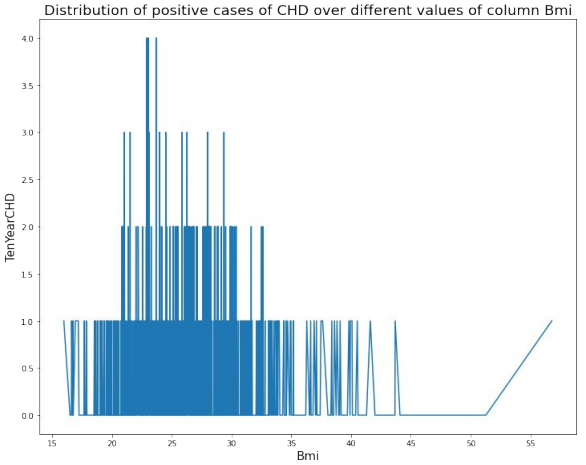

Distribution of positive cases of CHD over different categories of column Bpmeds
TenYearCHD
Bpmeds
0.0
1.0
0
100
200
300
400

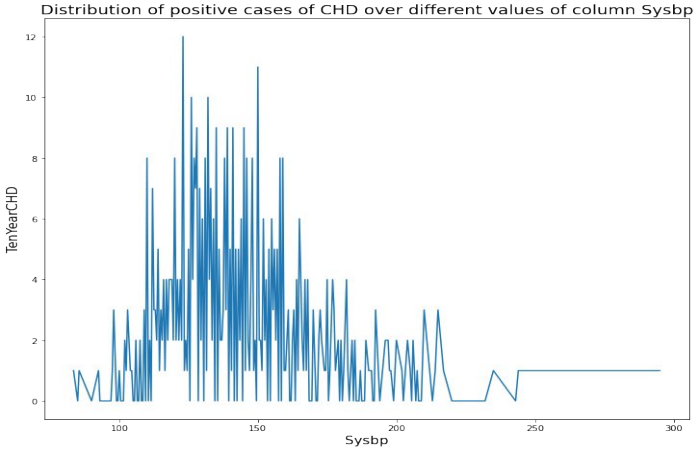

Distribution of positive cases of CHD over different categories of column Diabetes
TenYearCHD
Diabetes
0
1
0
100
200
300
400
500

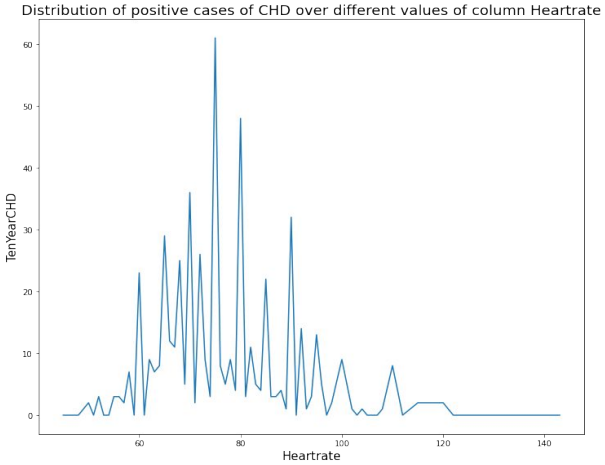
2. EDA(Contd):

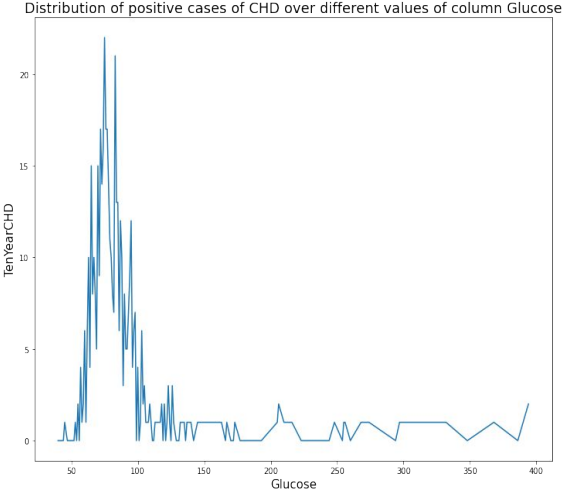

Distribution of positive cases of CHD over different values of column Totchol
TenYearCHD
Totchol


Distribution of positive cases of CHD over different values of column Diabp
TenYearCHD
Diabp

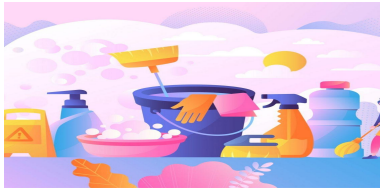

Distribution of positive cases of CHD over different values of column Bmi
TenYearCHD
Bmi


Distribution of positive cases of CHD over different values of column Sysbp
TenYearCHD
Sysbp


Distribution of positive cases of CHD over different values of column Heartrate
TenYearCHD
Heartrate


Distribution of positive cases of CHD over different values of column Glucose
TenYearCHD
Glucose

3. Clean up:

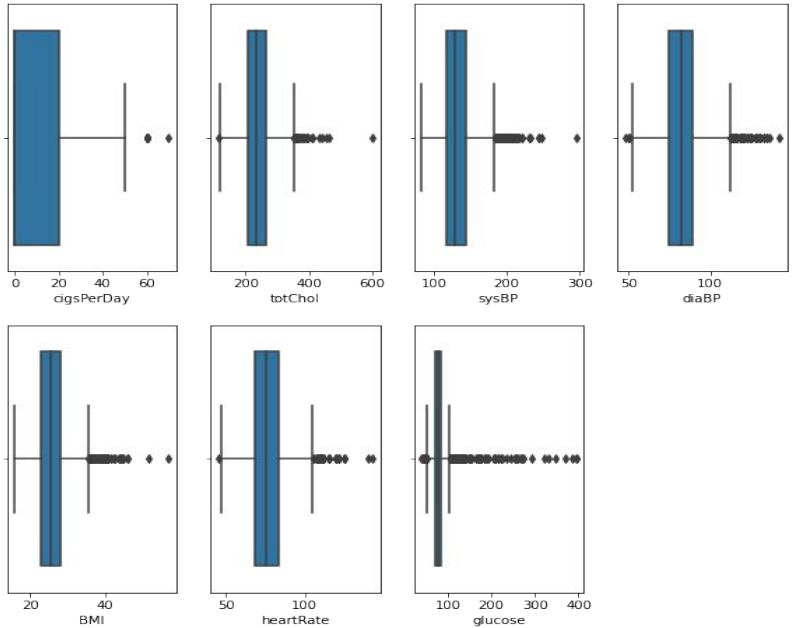


- Handling Null values: In this project the first step in Data cleaning is Handling the null values. Null values can affect the accuracy and quality of our ML models, therefore it is a good practice to handle null values. In this project I have used a combination of imputing and deleting the null values, where I have imputed columns with a great number of null values with the mean of that column and then deleted other observations that contain null values.
- Handling duplicate values: Duplicate values can have adverse effects on our ML models, therefore we have to try and remove it. Luckily we don’t have any duplicate values in our data so we can move on to the next step in data cleaning.

3. Clean up(Contd):

- Removing Outliers: First of all we find the variables that may contain outliers, to detect this I've used the box plot offered by seaborn library.

- Here we can see that many columns contain outliers, but they are all practically possible values.
- Keeping in mind this idea of practicality, I will allow these outliers to stay in our data because they will impact our predictions.

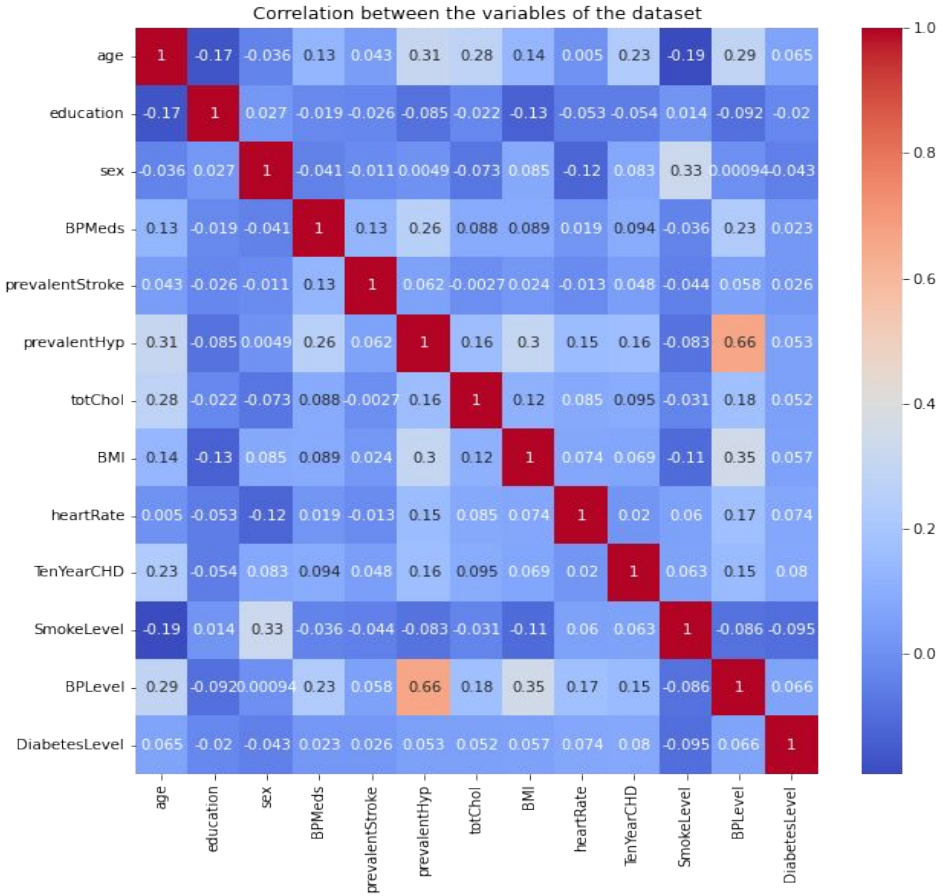


4. Feature Engineering:

- Feature Encoding: Machine learning models can only work with numerical values and therefore we have to turn the categorical columns to numeric columns, and this is achieved by feature encoding. In our dataset we have two such categorical columns, sex and is_smoking which we have to convert to numeric columns. I've used one hot encoding technique here.
- Grouping columns for better understanding: There are multiple columns which can be combined to form single columns that convey the same information in a better and understandable way. There are 3 such pairs of columns, is_smoking and cigsPerDay which is combined to form SmokeLevel, diabetes and glucose which is combined to DiabetesLevel, and sysBP and diaBP which is combined to BPLevel.

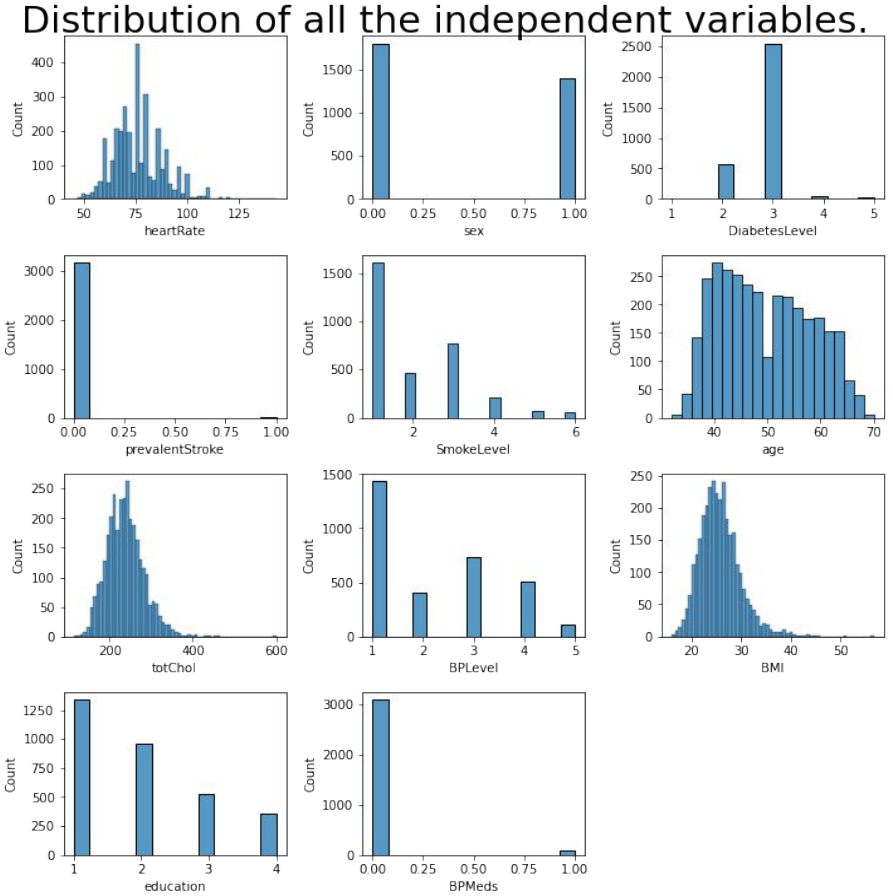
4. Feature Engineering(Contd):

- **Checking correlation for feature removal:**
  I've plotted the correlation matrix using the heatmap function offered by the seaborn library. As we can see, the prevalentHyp column is highly correlated to BPLevel, and also since it doesn't really convey much information, I'll just delete it.



4. Feature Engineering(Contd):

- **Checking the distribution of the data:** In this step I've just plotted the data in each individual column to check their distribution, using the histplot function offered by the seaborn library. The main reason I've done this is to check if the columns could really contribute to the prediction of the target variable. Using this step I've deleted the columns BPMeds, prevalentStroke and Education.



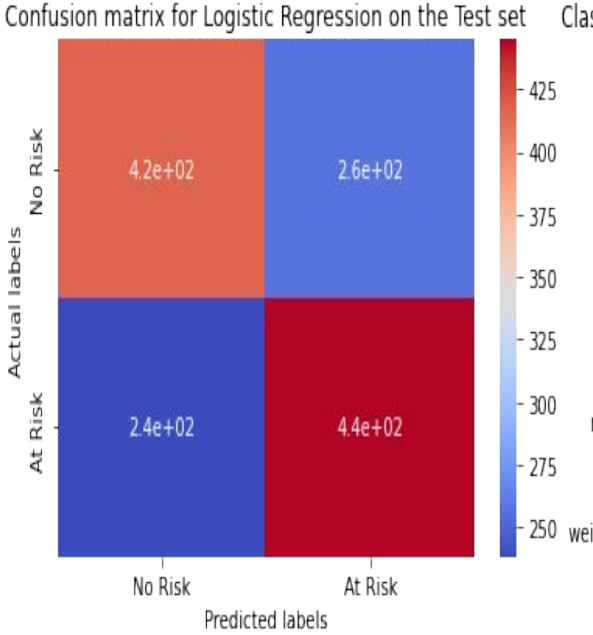
5. Pre Processing the data:

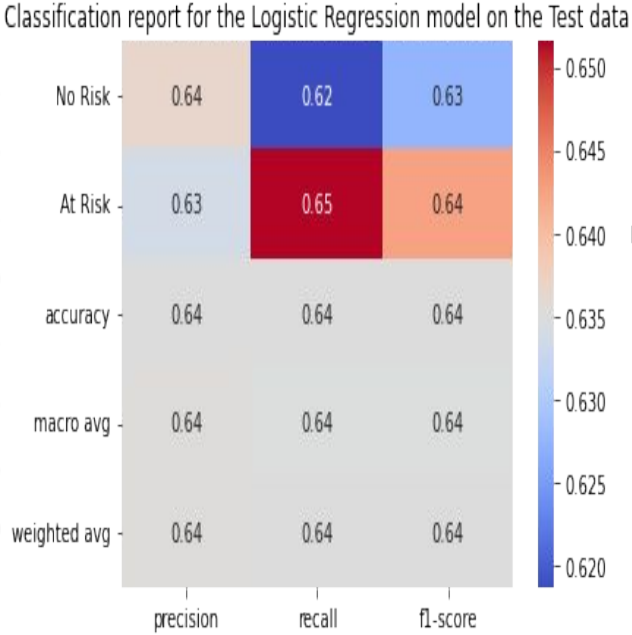
- **Dealing with class imbalance:** Our target variable has two classes, 0-No risk and 1-At risk(Risk of having a CHD in the next 10 years). There is a high class imbalance here, that is the number of observations with 0 is significantly greater than the number of observations with 1 as the target variable. To solve this problem I've used the SMOTE(synthetic minority oversampling technique) technique. After this step the class imbalance is completely removed.
- **Splitting and scaling the data:** In this step I've split the data into the independent columns and the target variable and further into train and test set and then scaled the independent columns to avoid giving more weightage to columns that have higher values and low weightage to columns that have low numeric value.

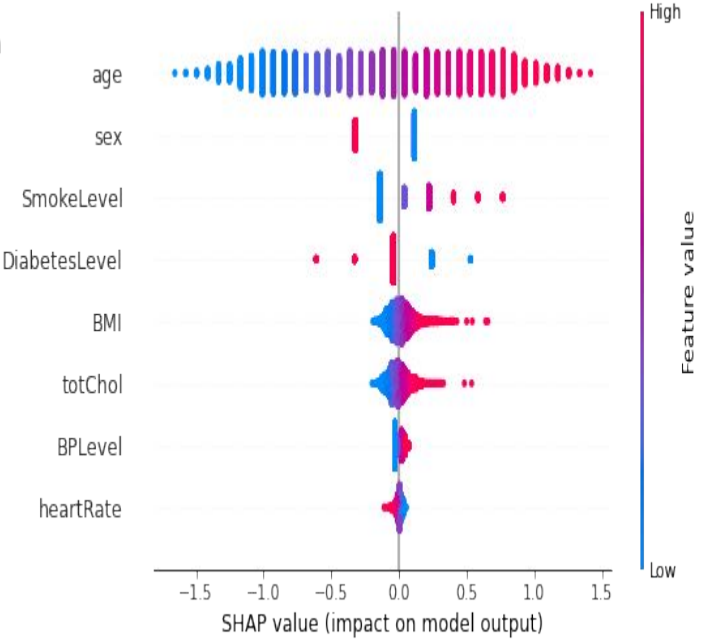
6&7. Model implementation and explainability:

These are two different processes in the flowchart, but for explaining purposes I'm clubbing the two into a single process.

- **Logistic regression: ROCAUC score on test set = 0.6351**

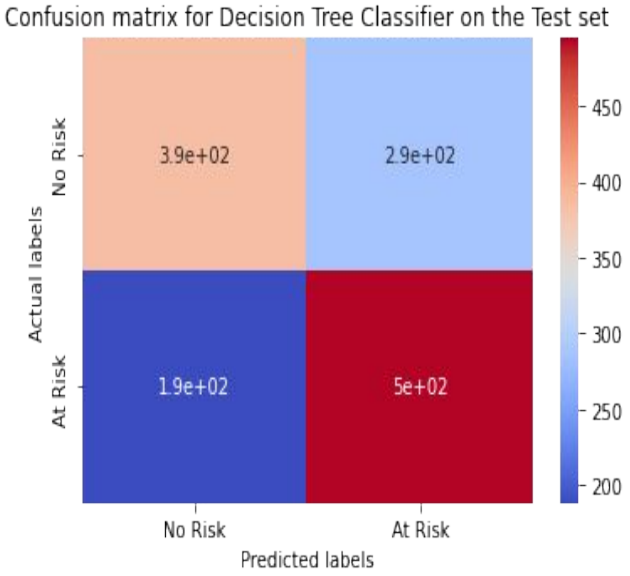


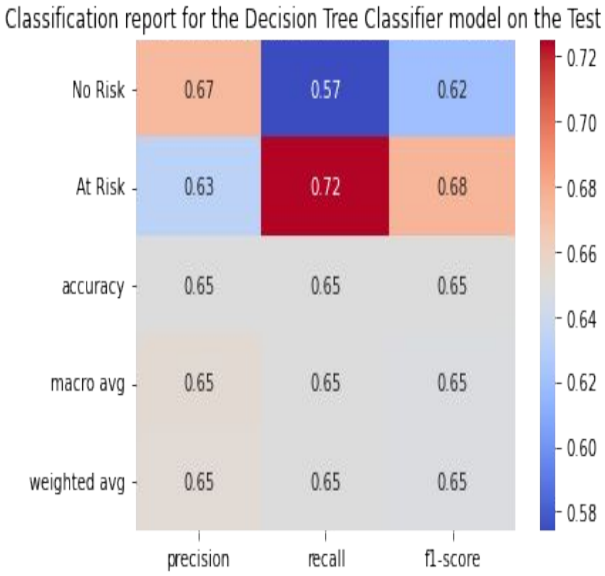


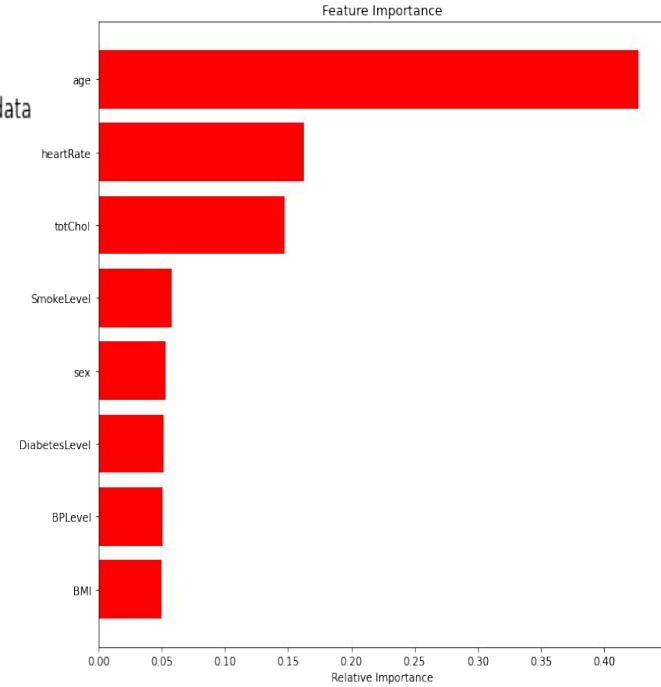


6&7. Model implementation and explainability(Contd):

- **Decision tree classifier :**
  ROCAUC score on test set = 0.6494

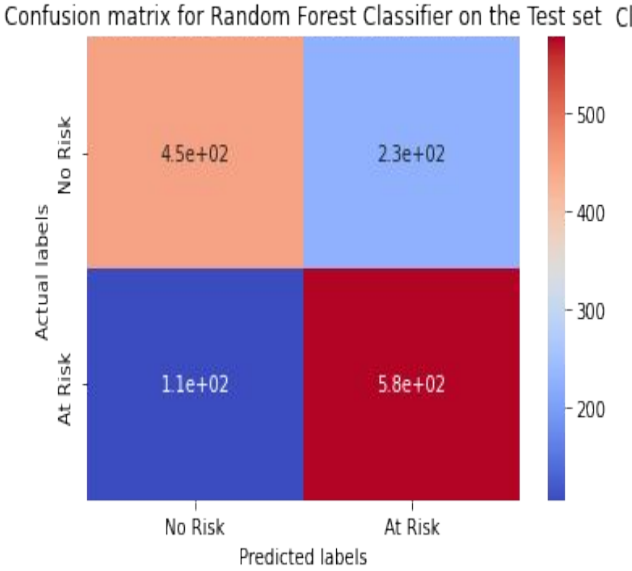


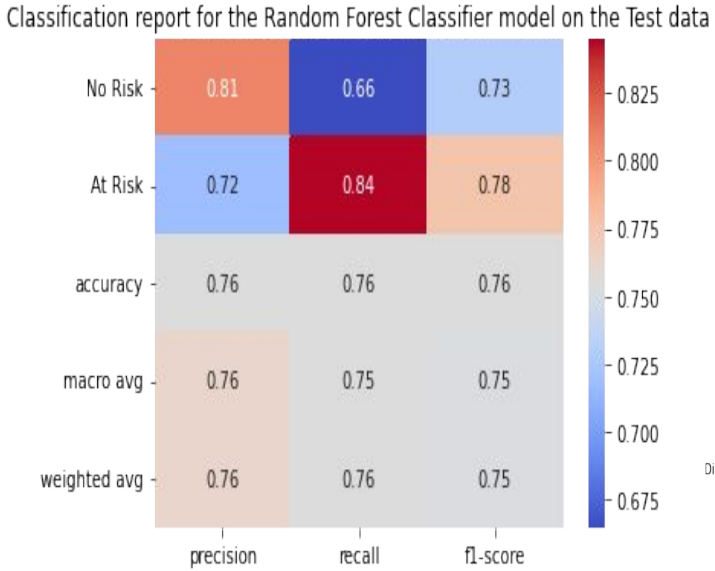


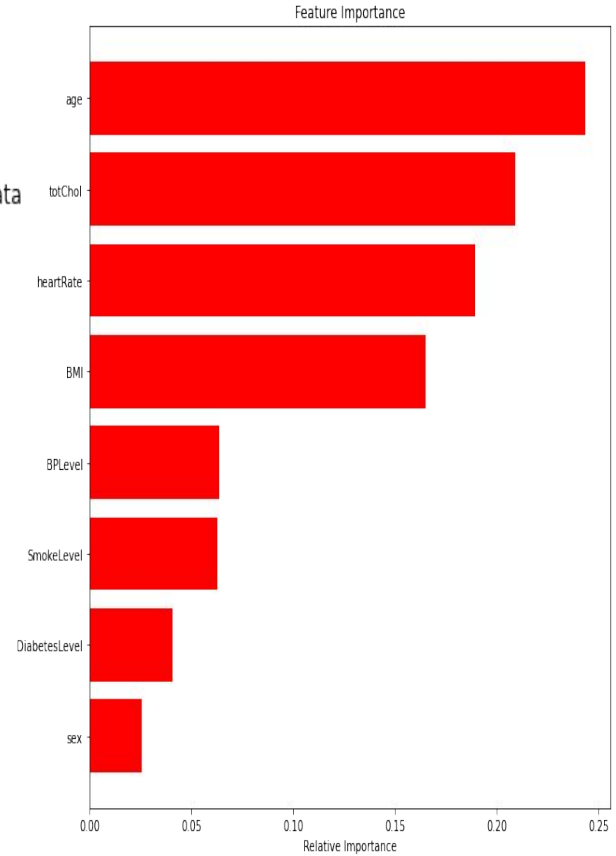


6&7. Model implementation and explainability(Contd):

- **Random Forest classifier :**
  ROCAUC score on test set = 0.7547

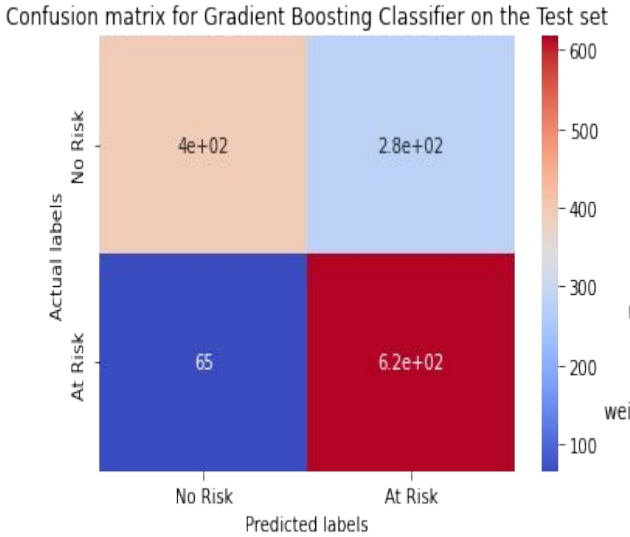


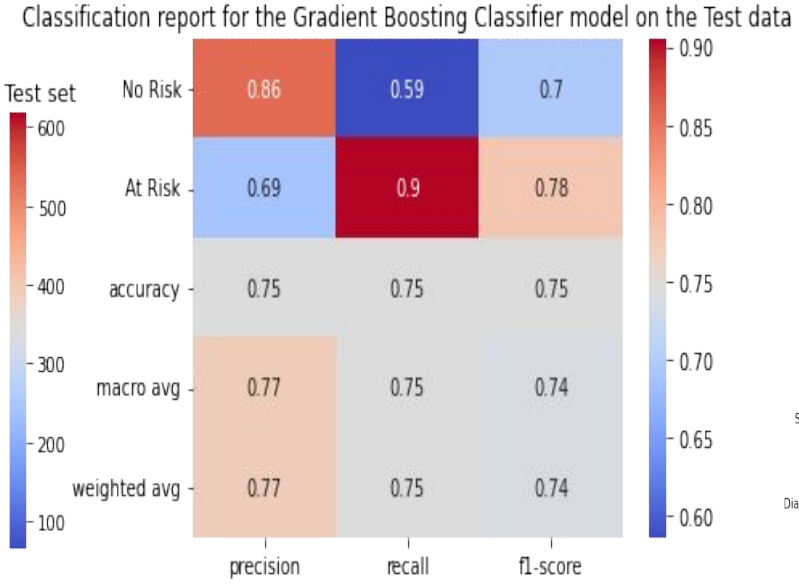


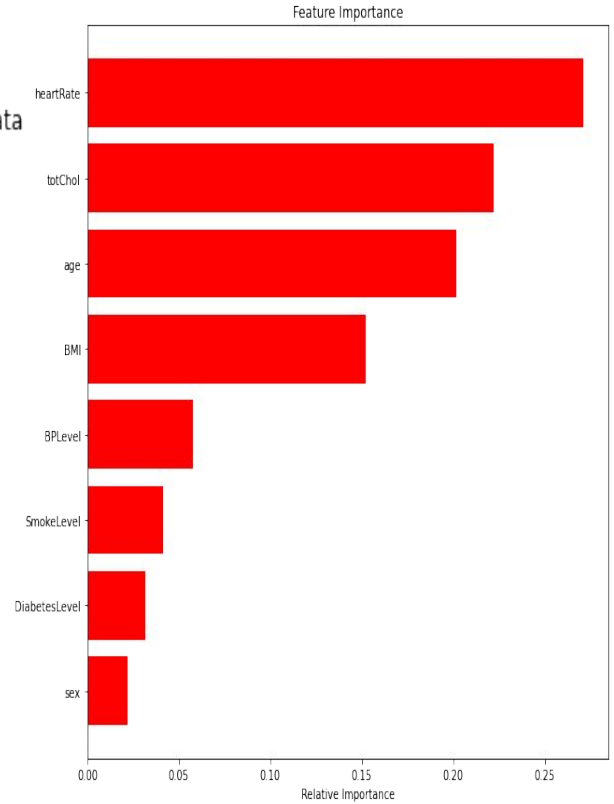


6&7. Model implementation and explainability(Contd):

- **Gradient Boosting classifier :**
  ROCAUC score on test set = 0.7454







❖ Conclusion:

EDA insights:

- The age group that is most likely to have a positive CHD risk factor is 47-65.
- Education level is not a great factor to determine the CHD risk factor.
- Males have a 7% greater chance of having a positive CHD risk factor.
- Smoking increases the chances of a positive CHD risk factor by around 3%.
- Having BP medications increases the chances of a positive CHD risk factor by around 19%.
- Having a prevalent stroke increases the chances of a positive CHD risk factor by around 31%.
- Prevalent Hypertension increases the chances of a positive CHD risk factor by around 13%.
- Diabetes increases the chances of a positive CHD risk factor by around 24%.

Conclusion(Contd):

Results from ML models:

- Logistic regression gives a ROCAUC score of 0.6365 on the testing set. This is worst performing model.
- Decision tree model gives a ROCAUC score of 0.6617 on the testing set.
- Random Forest Classifier model gives a ROCAUC score of 0.7584 on the testing set. This is the best performing model.
- Gradient Boosting Classifier model gives a ROCAUC score of 0.7416 on the testing set.
- Classification report and confusion matrix has been plotted for all the models.
- Model explainability has been achieved by SHAP library's summary plot and an attribute called feature_importances_ of the tree based algorithms.
- Total cholesterol and age are the two most important factors to predict the CHD risk factor.

Challenges faced:

- Feature engineering.
- Handling class imbalance.
- Choosing model explainability techniques.
- Running the slow Gradient Boosting Classifier.

THANKYOU!!

I’m very glad that you have tagged along until the end. I hope you enjoyed it and if you have any suggestions about my work, please let me know :)