# Cardiovascular Risk Prediction

Efforts By:

TEAM MIND

Mukund Jha

Nikhila NS

Shubhankit Sirvaiya

Yogesh Patil

# Table of Contents



# Machine Learning in Health care !

- Why it is possible?
  - Digitization
  - Advancement in computation power
  - Curated Algorithm
- How is it reliable?
  - Algorithms are transparent
  - Based on real world cases
  - Domain expertise included
- Why is it better ?
  - Cheaper
  - Easily available to all
  - Takes less time training , compared to a medical professional

# Introduction

Heart disease is the major cause of mortality globally. It accounts for more deaths annually than any other cause.

According to an estimate in 2016 there were nearly 18 million deaths globally due to heart disease, representing  31% of all global deaths.

Most heart diseases are highly preventable by simple lifestyle modification.

So we wish to know if Machine Learning is superior in pattern recognition and classification of those who are in risk of cardiovascular disease.

# Problem Statement

We have been given the cardiovascular data of nearly 3,300 people, where we have health characteristics. we want to make a machine learning model , which can learn from the pattern of these data and can predict if a person is going to have a cardiovascular risk in next 10 years.

# Understanding our data

- 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.
- 10-year risk of coronary heart disease CHD(binary: "1", means "Yes", "0" means "No") -DV

# Understanding the Data – (contd)

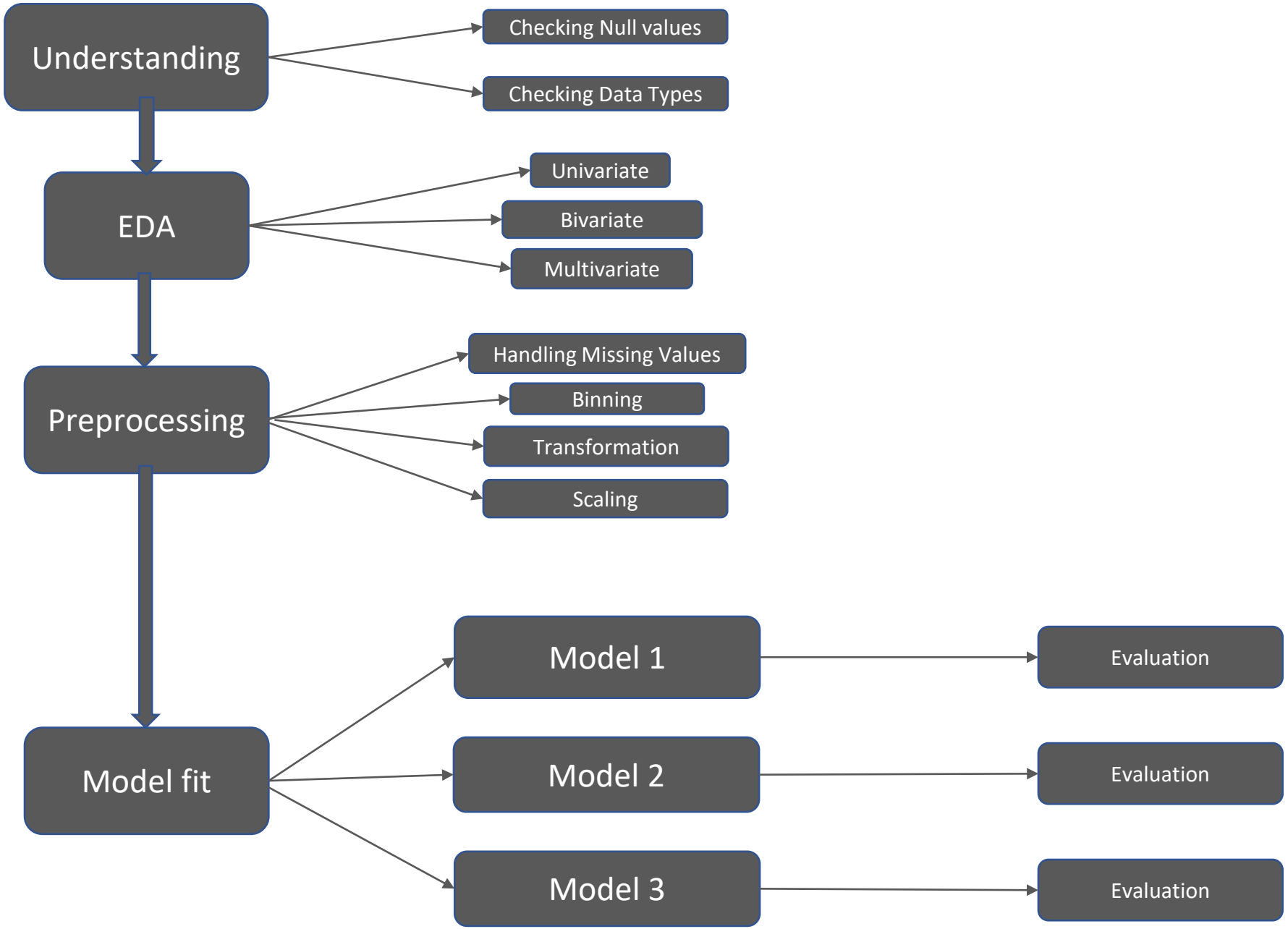
- **Demographic:**
  - **Sex**: male or female("M" or "F")
  - **Age**: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- **Medical( history):**
  - **BP Meds**: whether or not the patient was on blood pressure medication (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)

# Understanding the Data – (contd)

- **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)

**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.)


Understanding
Checking Null values
Checking Data Types
EDA
Univariate
Bivariate
Multivariate
Preprocessing
Handling Missing Values
Binning
Transformation
Scaling
Model fit
Model 1
Model 2
Model 3
Evaluation
Evaluation
Evaluation

# EDA

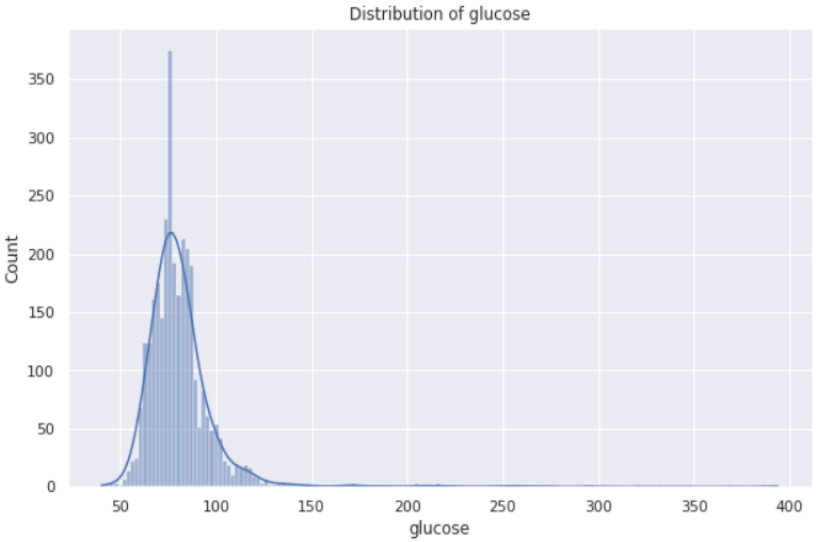
# Missing Values

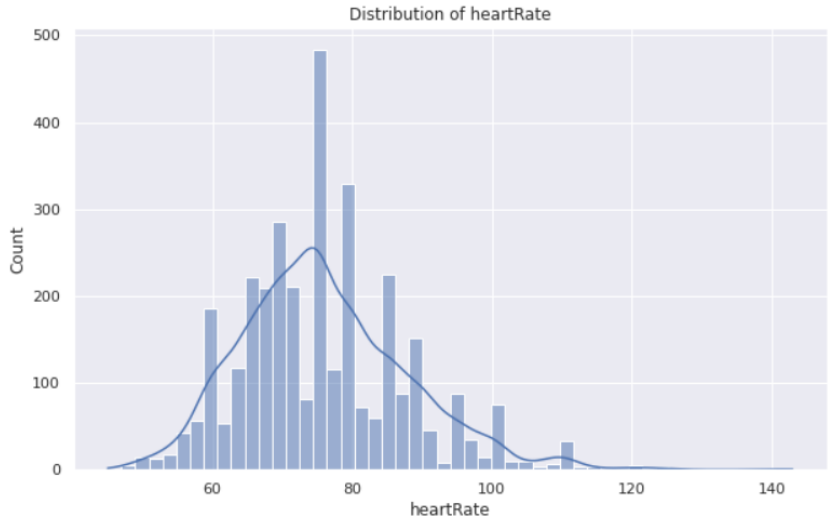
| columns | missing values |
|---|---|
| id | 0 |
| age | 0 |
| education | 87 |
| sex | 0 |
| is_smoking | 0 |
| cigsPerDay | 22 |
| BPMeds | 44 |
| prevalentStroke | 0 |
| prevalentHyp | 0 |
| diabetes | 0 |
| totChol | 38 |
| sysBP | 0 |
| diaBP | 0 |
| BMI | 14 |
| heartRate | 1 |
| glucose | 304 |
| TenYearCHD | 0 |

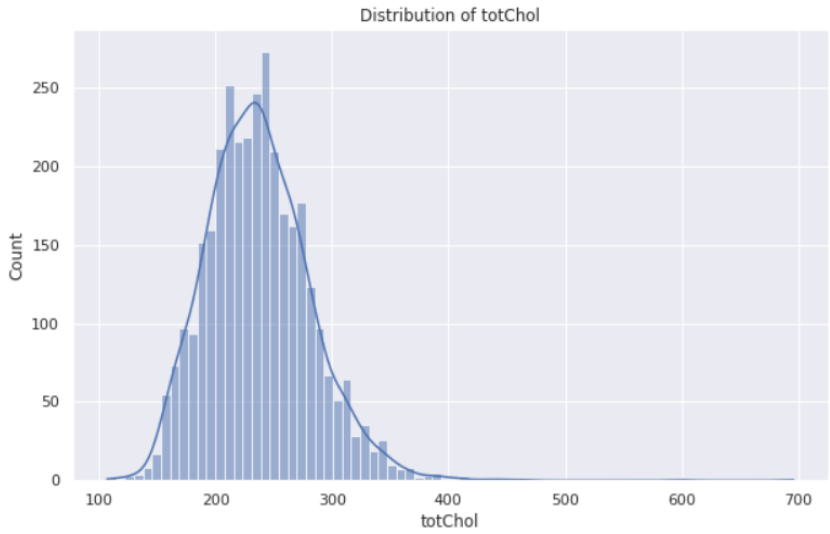
- There are some null values in our Data that we will have to deal with

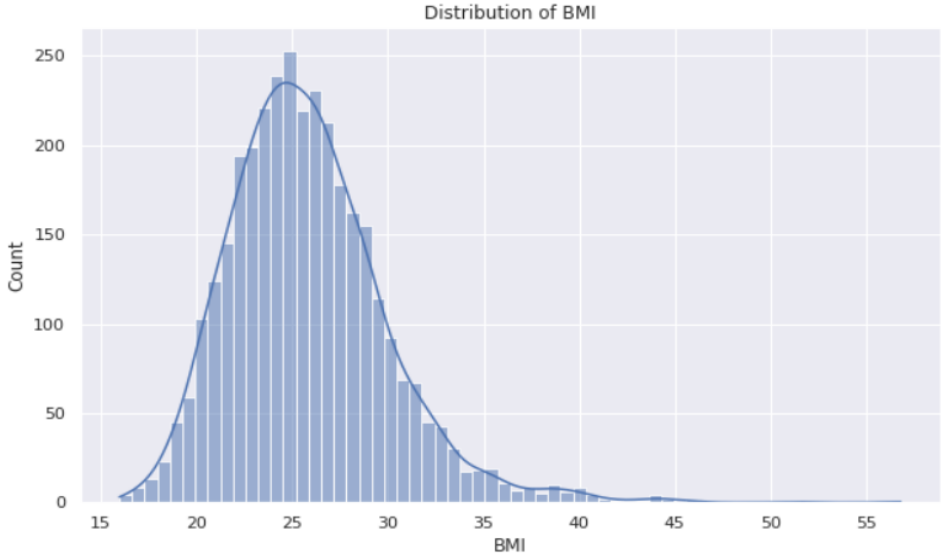
- We have imputed the missing values with median for continuous features - as there was less skewness in our data; and mode for categorical features

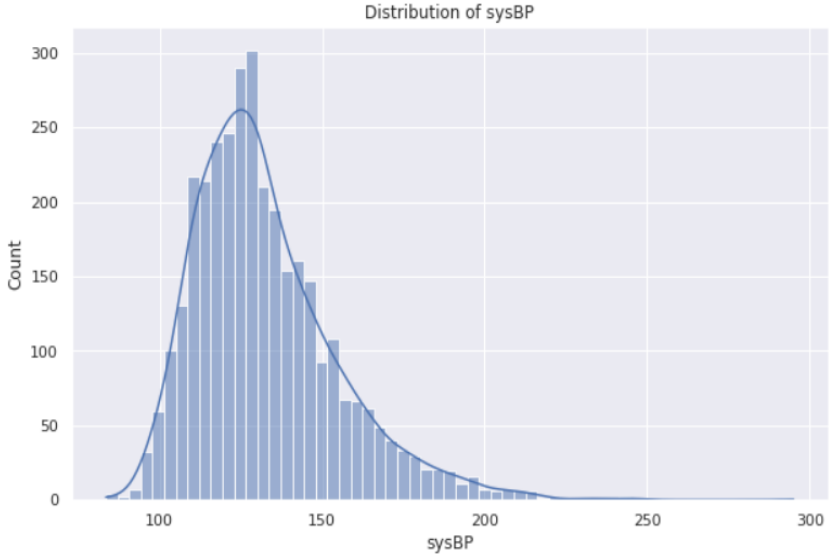
# Checking for Distribution and skewness in the continuous independent variable



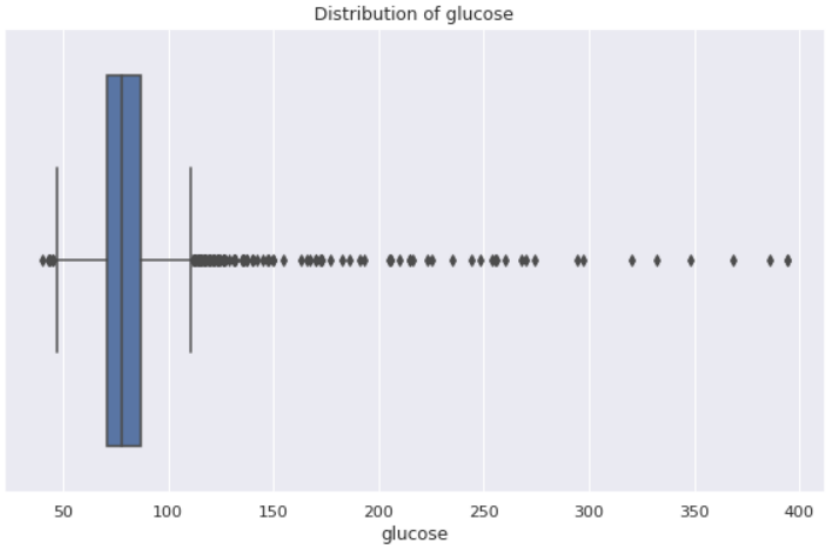



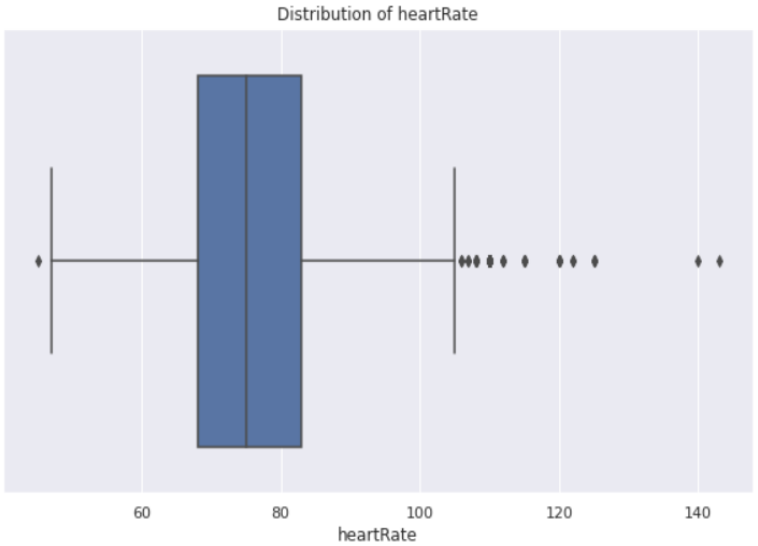



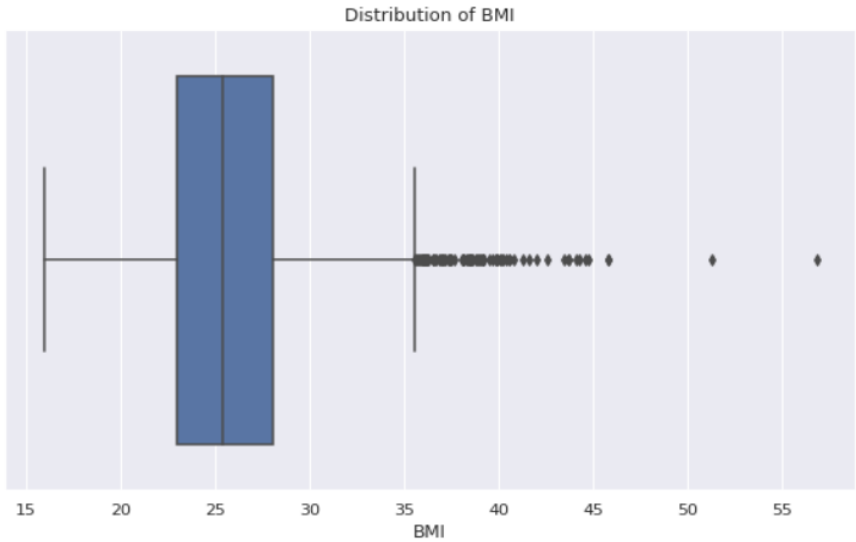



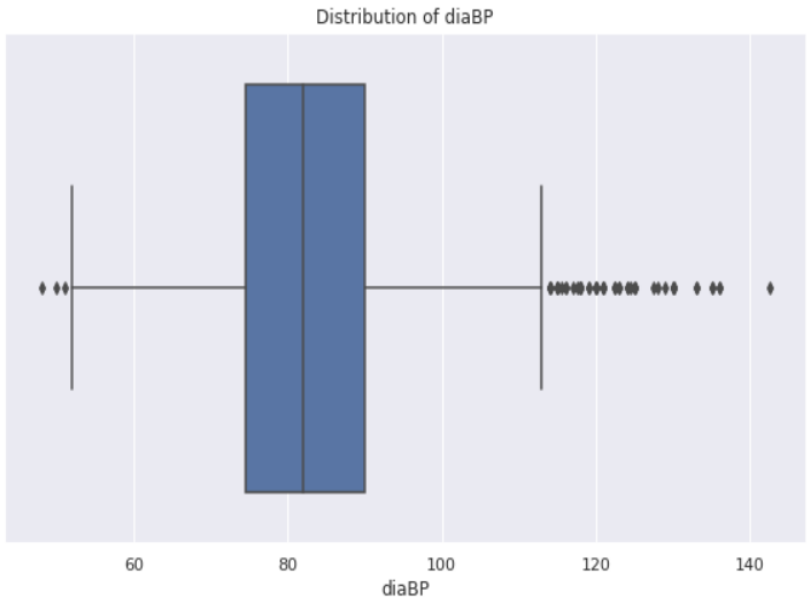



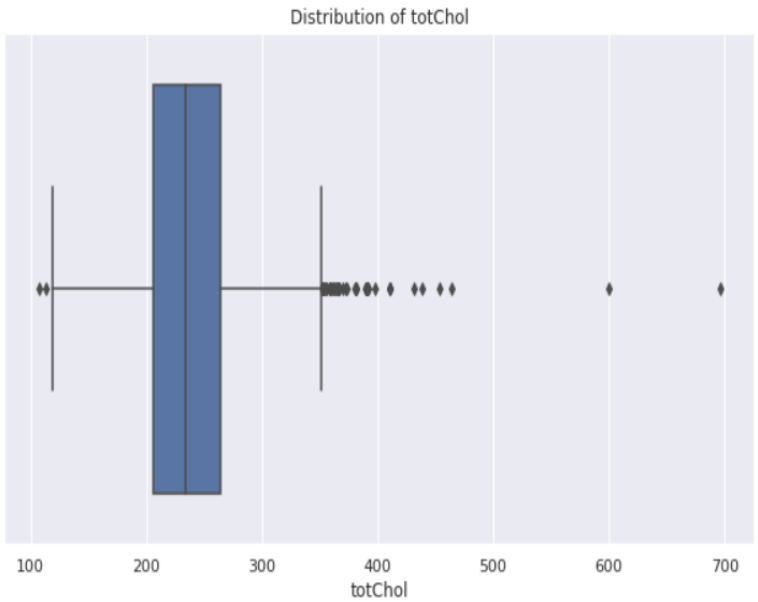
# Checking for outliers in the Data

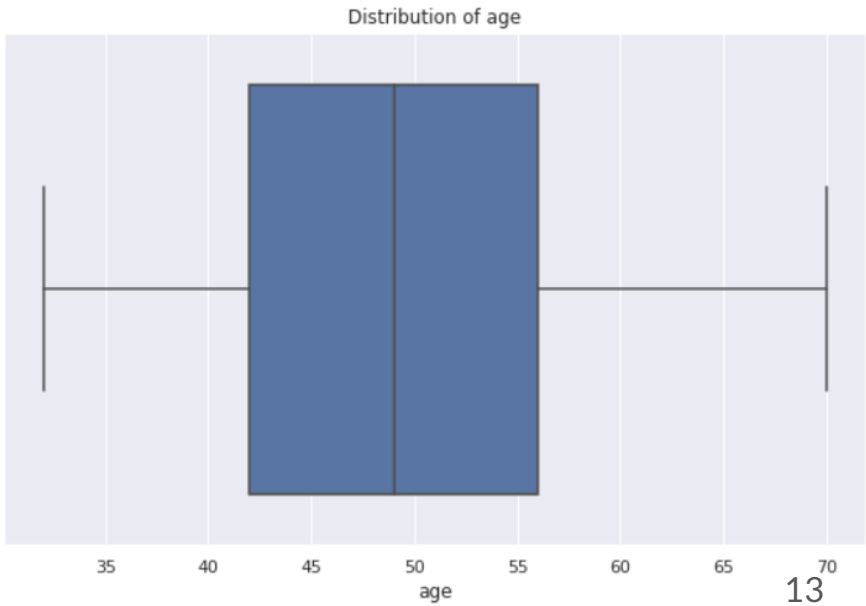


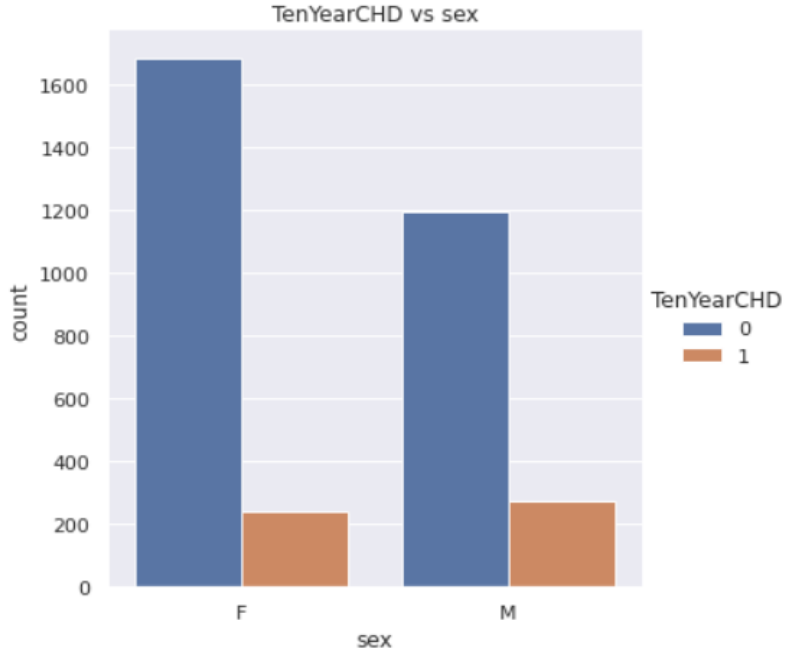



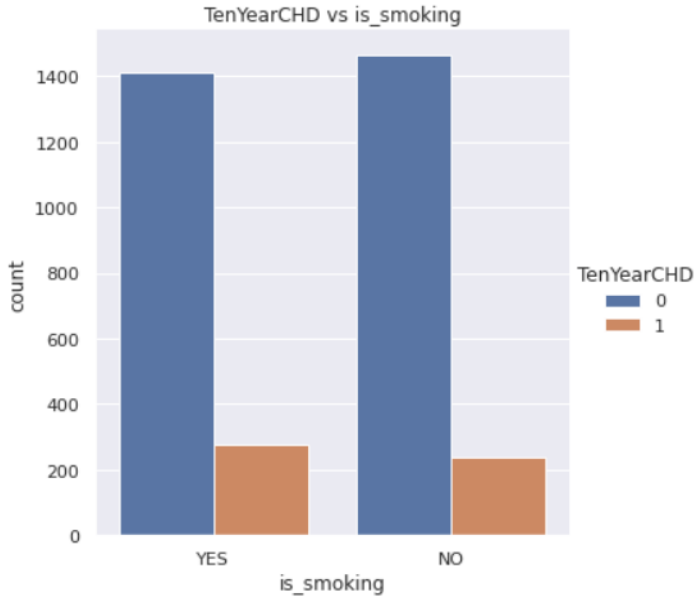



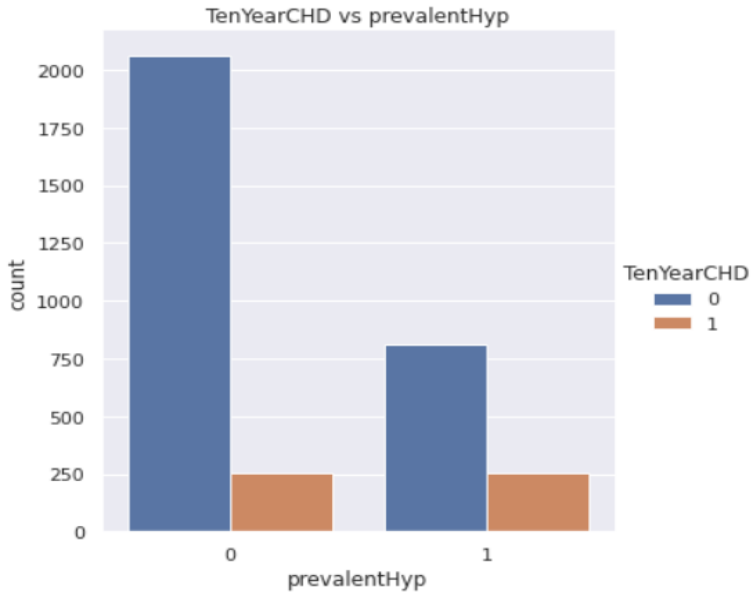



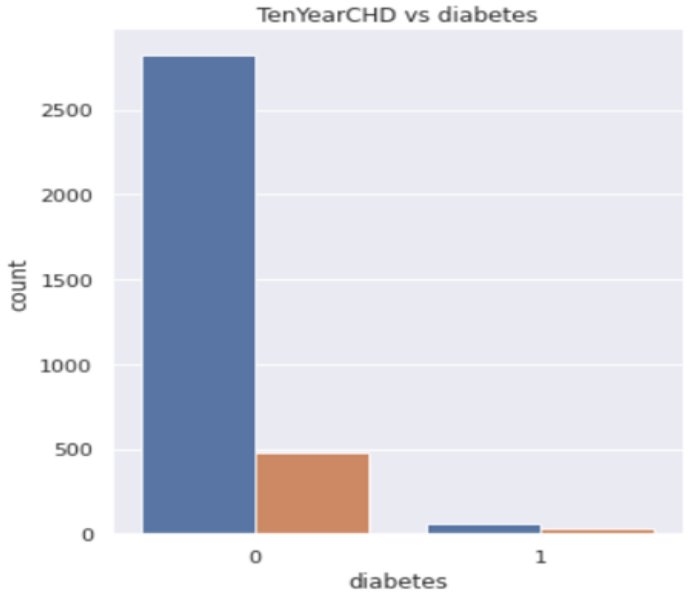



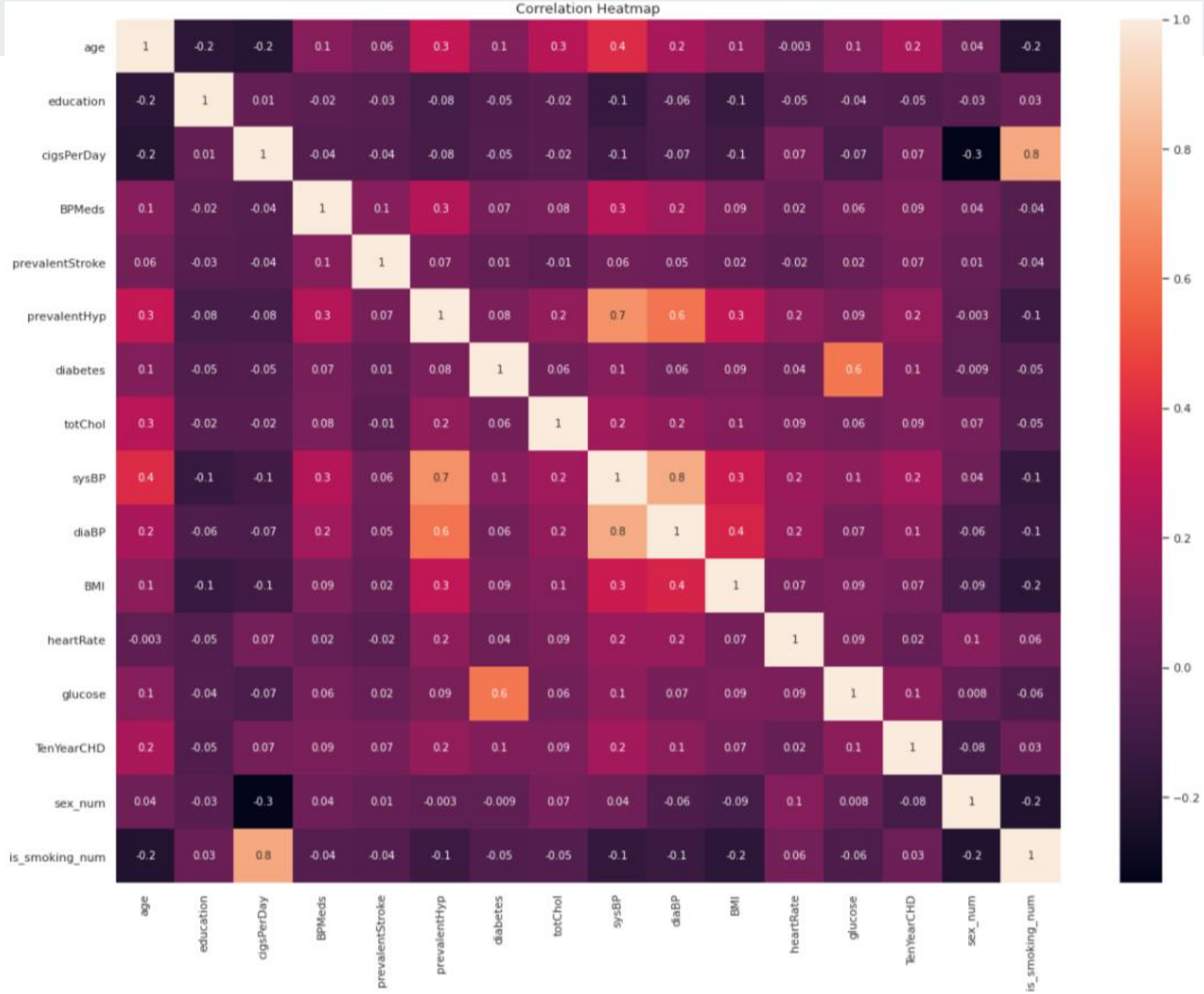
# Bivariate Analysis



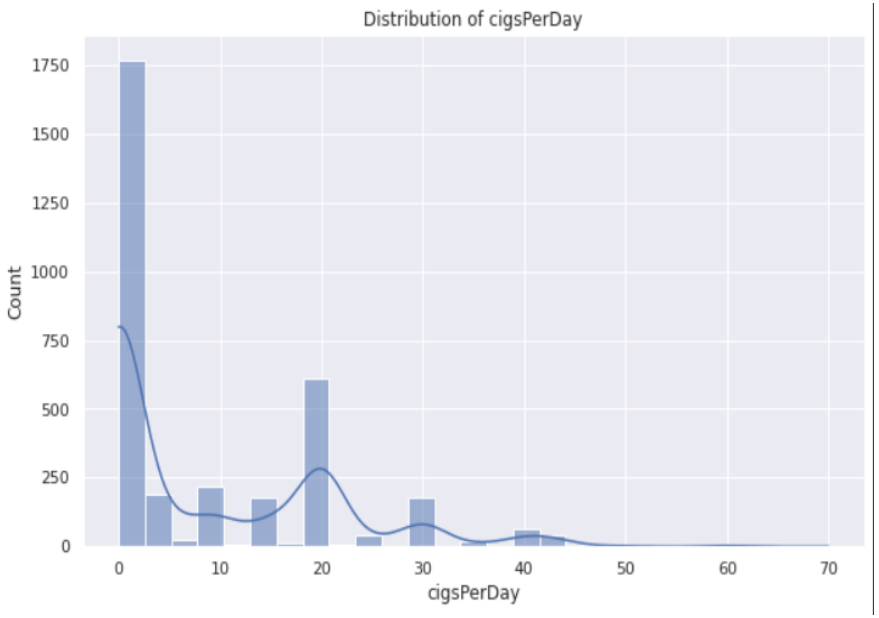



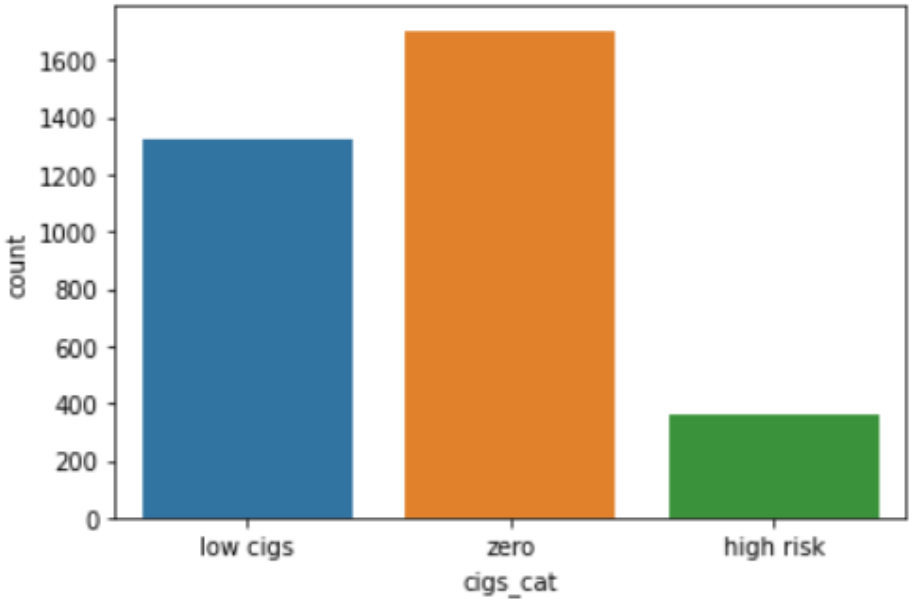



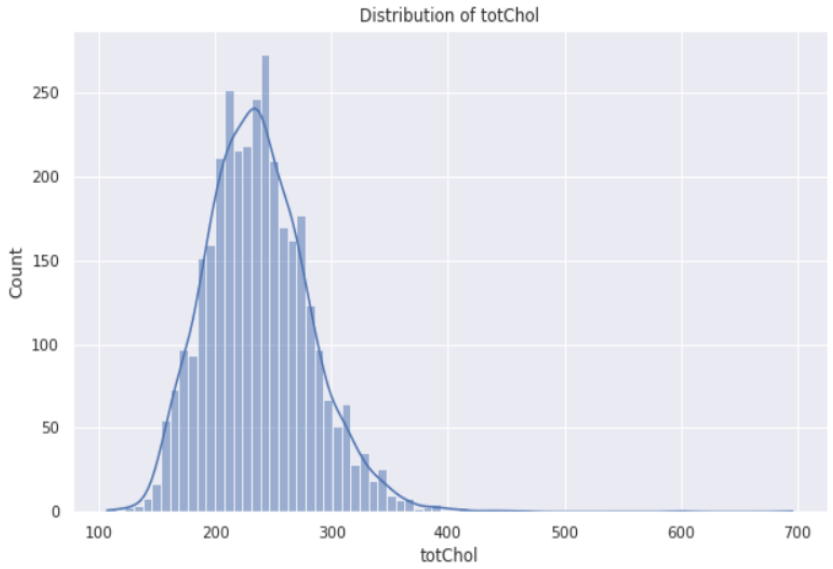



# Multivariate Analysis - Correlation

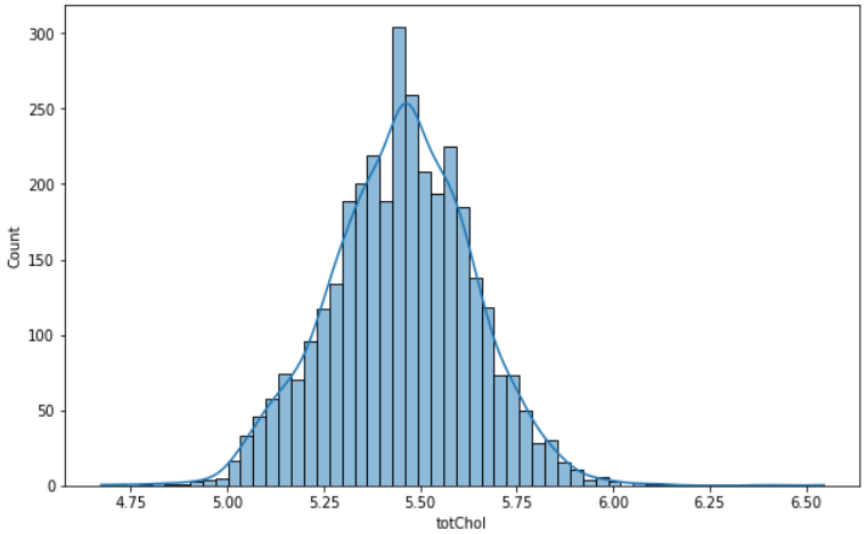



Binning to convert continuous to categorical variable




Using Log Transformation to deal with Skewness of the Data

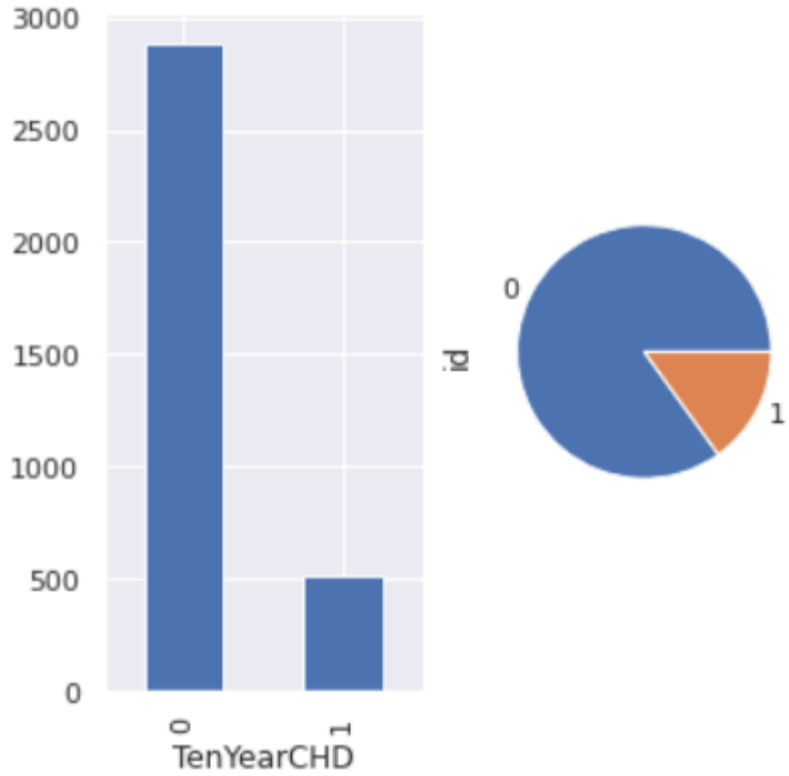


# Handling Class imbalance

A High Class imbalance is present in our Data

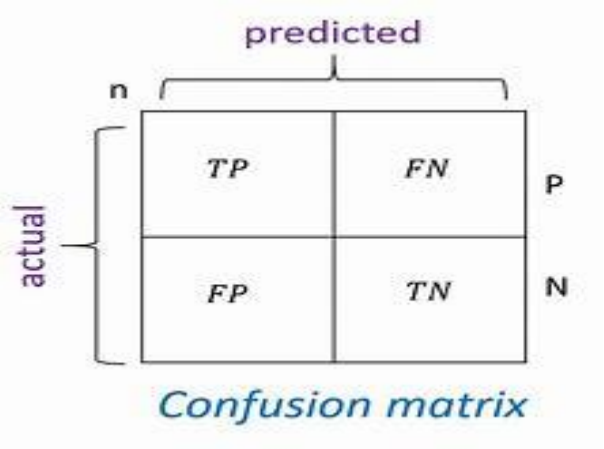
Following technique can be used to deal with class imbalance

- Random Under Sampling
- Random Over Sampling
- SMOTE
- Class weights



# Choosing Right Evaluation Metrics

- Right Metrics for evaluation depends on the Domain
- Different evaluation metric can be taken into account , based on the Problem at hand.
- Here we have taken Recall as our main evaluation metrics
- Recall measures the sensitivity of the model
- We will aim , that our model should not Misclassify any minority Point




$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

# Modelling

# Modelling using SMOTE to handle imbalance

| | Model_Name | Train accuracy | Test accuracy | Train recall | Test recall | Train ROC-AUC | Test ROC-AUC |
|---|---|---|---|---|---|---|---|
| **0** | LogisticRegression | 0.6803 | 0.7052 | 0.5924 | 0.5859 | 0.6715 | 0.6562 |
| **1** | DecisionTreeClassifier | 0.8824 | 0.7524 | 0.8454 | 0.7422 | 0.8787 | 0.7482 |
| **2** | RandomForestClassifier | 0.7529 | 0.7311 | 0.6792 | 0.6094 | 0.7455 | 0.6811 |
| **3** | XGBClassifier | 0.6762 | 0.6309 | 0.7151 | 0.7031 | 0.6801 | 0.6606 |

- Decision Tree Overfits the model
- We get the best recall score for XGBoost

# Modelling using Class Weights to handle Imbalance

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train ROC | Test ROC |
|---|---|---|---|---|---|---|
| Logistic | 0.660897 | 0.630896 | 0.704961 | 0.679688 | 0.737491 | 0.704112 |
| SVM | 0.670732 | 0.610849 | 0.754569 | 0.679688 | 0.765444 | 0.682823 |
| Xg Boost | 0.447679 | 0.419811 | 0.874674 | 0.859375 | 0.708819 | 0.665647 |

- We have tried to get better recall score
- XGBoost gives the best recall score

# Suggested Model

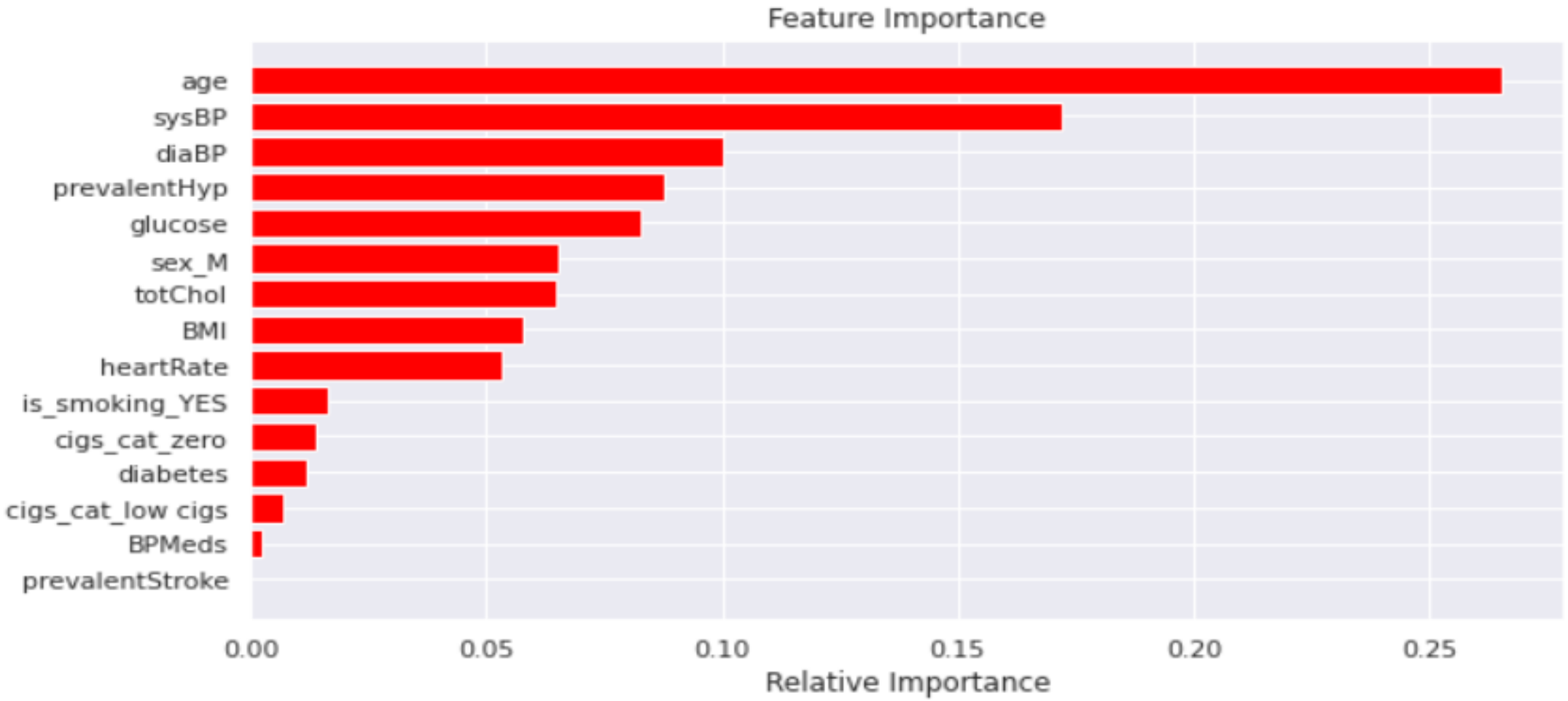
- Parameters

```
base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1,
colsample_bytree=0.7725876038663544, gamma=4.128689439174278, learning_rate=0.1,
max_delta_step=0, max_depth=9, min_child_weight=3.0, missing=None, n_estimators=100, n_jobs=1,
nthread=None, objective='binary:logistic', random_state=1, reg_alpha=176.0,
reg_lambda=0.44453811221465295, scale_pos_weight=6, seed=None, silent=None, subsample=1,
verbosity=1
```

- Evaluation

| Model | Train Accuracy | Test Accuracy | Train Recall | Test Recall | Train ROC | Test ROC |
|---|---|---|---|---|---|---|
| Xg Boost | 0.447679 | 0.419811 | 0.874674 | 0.859375 | 0.708819 | 0.665647 |

# Feature importance



# Summary

- As the first step, we understand the data &  perform some cleaning on the null values and checking the data types and EDA on data.
- After EDA We divided our project into different model building and preprocessing before model.
- As so we have a high imbalance data set, we used different techniques to balance the data like SMOTE and assign class weights.
- We tried different models and evaluated their performance scores. Models Built: Logistic Regression, SVM, Decision Tree, RandomForest, XGBoost
- Based on our targeted evaluation metric - recall, we chose XGBoost as the suggested model

# Challenges Faced

- High Class Imbalance in the data
- Unfamiliarity with the domain
- Number of records were less
- We have deal with the trade-off between accuracy and recall.

# Conclusion

- This model explored the potential of applying ML approaches to predict which patients will have Cardiovascular Risk in next 10 years.

- The results showed that the Machine Learning performs well with established risk tools in identifying a potential candidate for CHD.

- Multiple models were built in order to achieve high recall so that we do not misclassify a patient who is at potential risk. We got the best results on XGBoost Model.

- Thus ML methods should be considered in the development of future cardiovascular risk prediction along with a domain expert.

# Q / A