Capstone Project
Cardiovascular Risk Prediction

Sonali Kaushal
Somya Hingorani
Nitesh Verma
Harsh Vardhan
Prateek Gupta

Problem statement

- 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. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

Agenda

To discuss the analysis of given cardiovascular risk data set.

Topics covered for the project :

- Data Pipeline
- Data Summary
- Data Description
- Feature engineering
  - Multicollinearity
- Exploratory Data Analysis
- Model Overview
- Model Analysis
  - Model's Evaluation Matrices
  - ROC-AUC curve
  - Model Features
  - Accuracy of Models Performed
- Conclusion

Data Pipeline

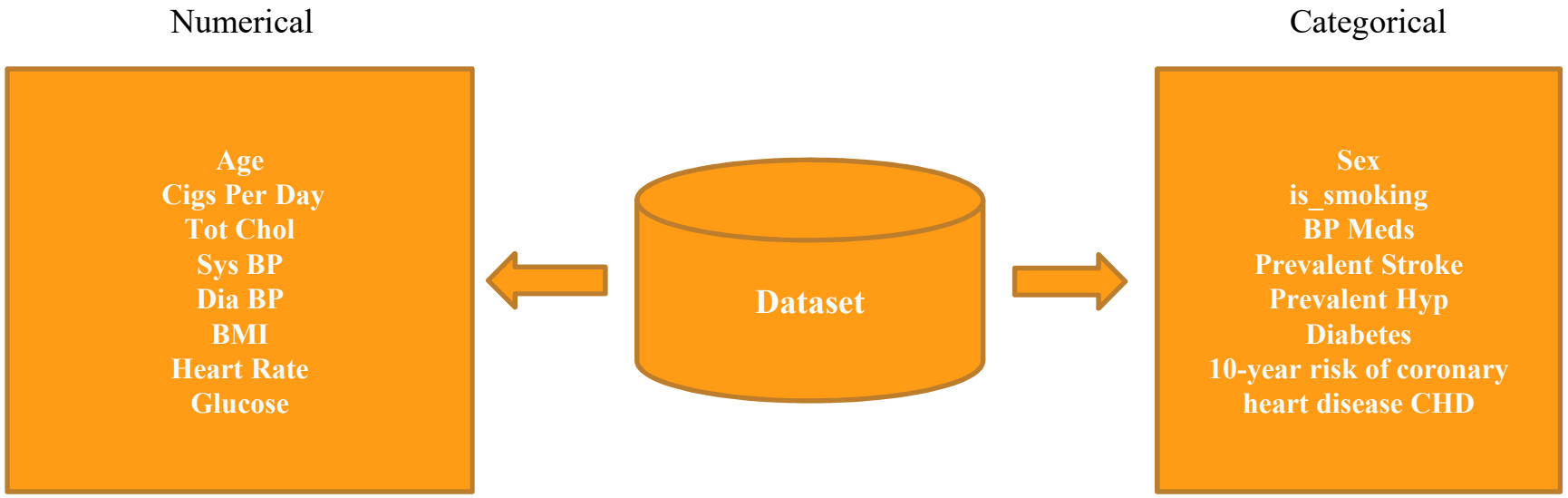
❖ Data pre-processing: We pre processed the data by dealing with the outliers, null values, and duplicate data.

❖ Feature engineering: In this part we went through each attributes and encoded the categorical features.

❖ Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to get insights to know the effects of various given parameters on heart disease

❖ Model Creation: Finally in this part we created the various models. These various models are being analysed and we tried to study various models so as to get the best performing model for our project.

Data Summary



Data Description

Dependent variable:

• 10-year risk of coronary heart disease CHD

Independent variables:

❖ **Demographic:**
- Sex: male or female("M" or "F")
- Age: Age of the patient

❖ **Behavioural:**
- is_smoking: whether or not the patient is a current smoker
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day

❖ **Medical(history):**
- BP Meds: whether or not the patient was on blood pressure medication
- Prevalent Stroke: whether or not the patient previously had a stroke
- Prevalent Hyp: whether or not the patient was hypertensive
- Diabetes: whether or not the patient had diabetes

Data Description

- **Medical(current):**
  - Tot Chol: total cholesterol level
  - Sys BP: systolic blood pressure
  - Dia BP: diastolic blood pressure
  - BMI: Body Mass Index
  - Heart Rate: heart rate level
  - Glucose: glucose level

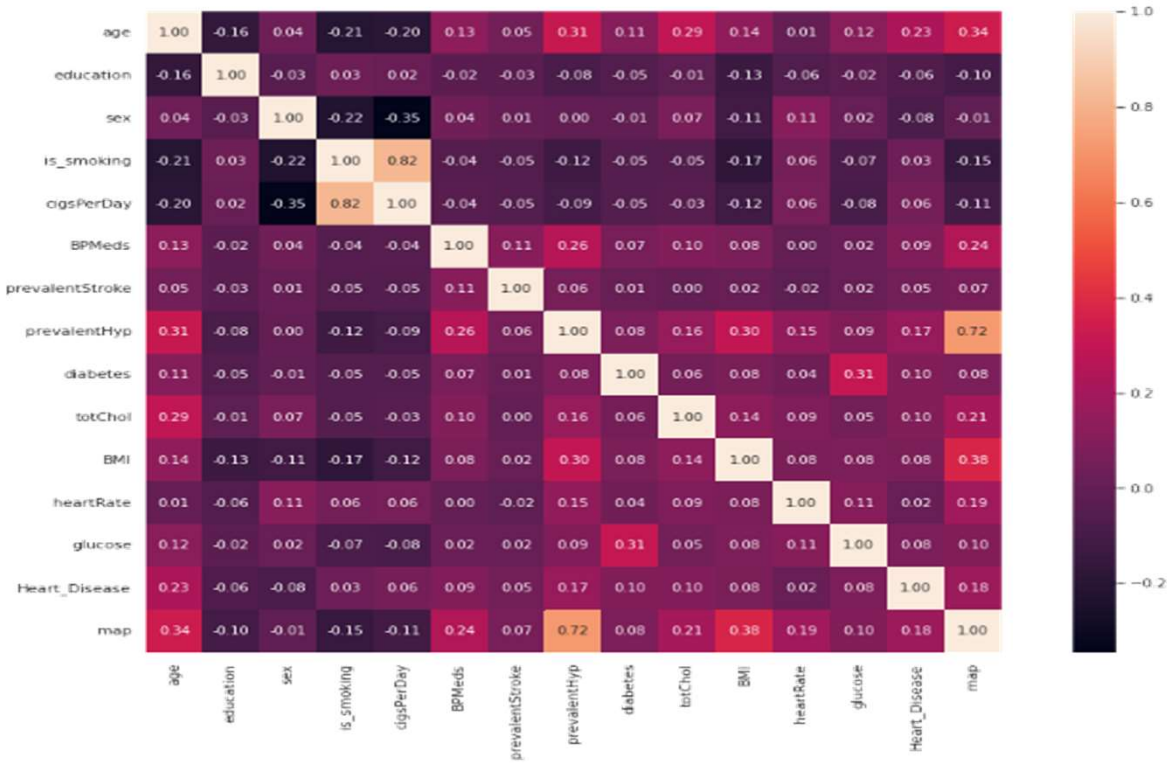
Feature engineering

Steps followed

- There are 7 features in the dataset containing null values. They are education, cigsPerday, BPMeds, totChol, BMI, heartRate and glucose. We dropped the rows with null values from the dataset as medical data of different individuals can't be compared and replaced with mean, median or mode.
- We have the features like 'id' and 'education' which does not provide any relevant information so we removed that columns.
- We've the columns 'sex' and 'is_smoking' which are of string type so we convert them into integer by applying the function which converts the following:
  - In sex feature M(Male) will be converted to 0 and F(Female) will be converted to 1.
  - In is_smoking feature YES will be converted to 1 and NO will be converted to 0.
- As sysBP, diaBP are highly correlated with one another, we combined them and created a new feature MAP(Mean Arterial Pressure). The MAP is calculated as below.

$$MAP = \frac{SBP + 2(DBP)}{3}$$

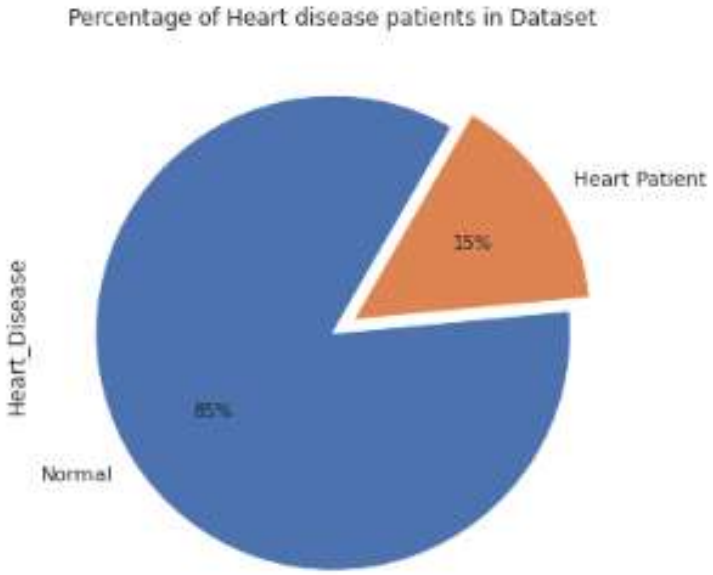
Multicollinearity



- As per the correlation matrix, cigsperday and is_smoking are highly correlated (0.82), also map and prevalentHyp are highly correlated (0.72) .
- We finally dropped is_smoking and prevalentHyp as they have comparatively less impact on target variable.

Exploratory Data Analysis

Percentage of heart disease patients



- According to the pie chart, dataset contains 85% normal persons and 15% heart patients. The class of the dataset is highly imbalanced, we have used SMOTE technique to handle class imbalance.

SMOTE
(synthetic minority oversampling technique)

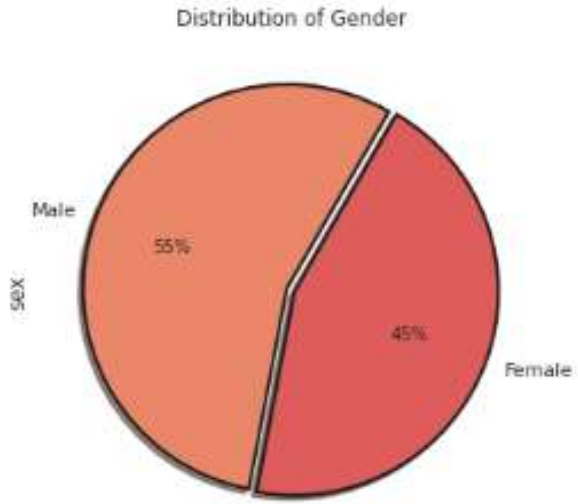
- This was a class imbalanced dataset so we used SMOTE(Synthetic minority oversampling technique) which is a class imbalance handling technique before running our algorithms.

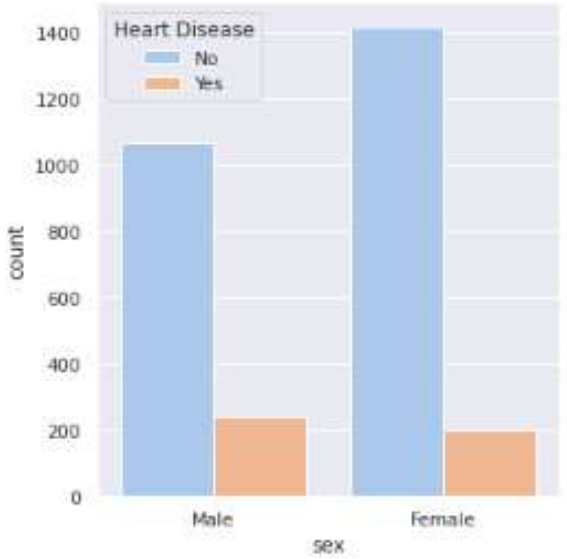
WHAT IS SMOTE ?

- This is a statistical technique for increasing the number of cases in the given classes in our dataset in a balanced way. The module works by generating new instances from existing minority class that we supply as input. This implementation of SMOTE does not change the number of majority class.

- SMOTE takes the entire dataset as an input, but it increases the percentage of only the minority class.

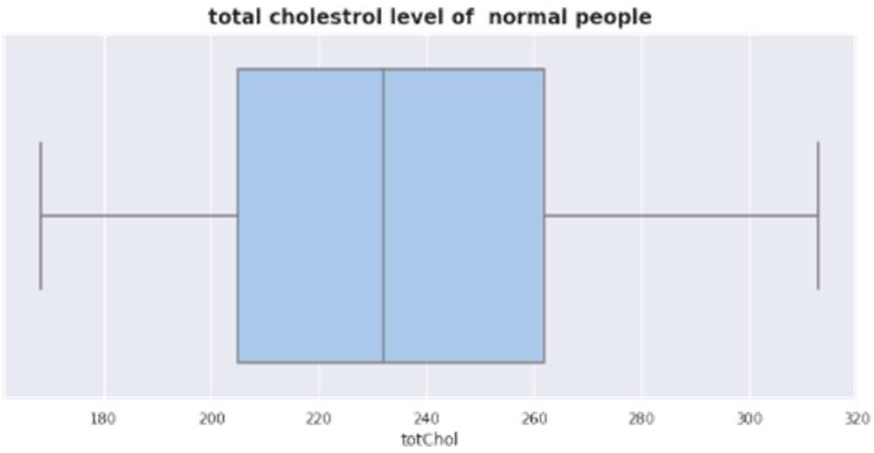
Analysis on the basis of gender

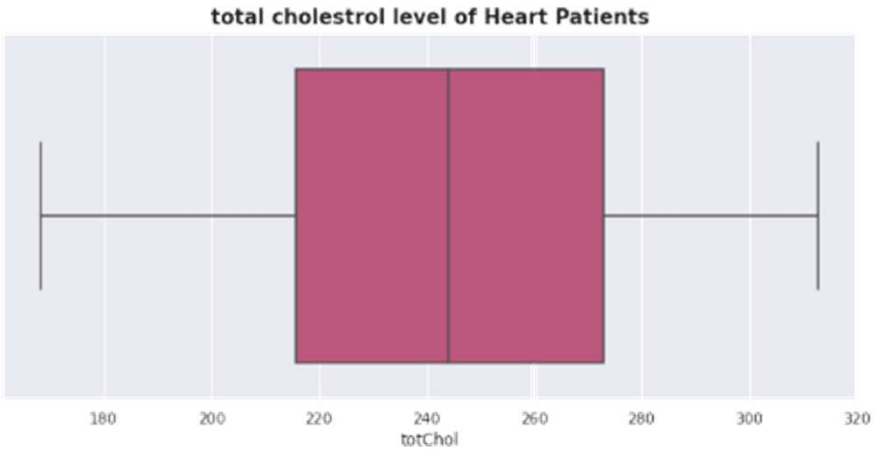




- According to the pie chart, given dataset contains 55% male and 45% female.
- According to the bar chart, males are more prone to heart disease as compared to females.

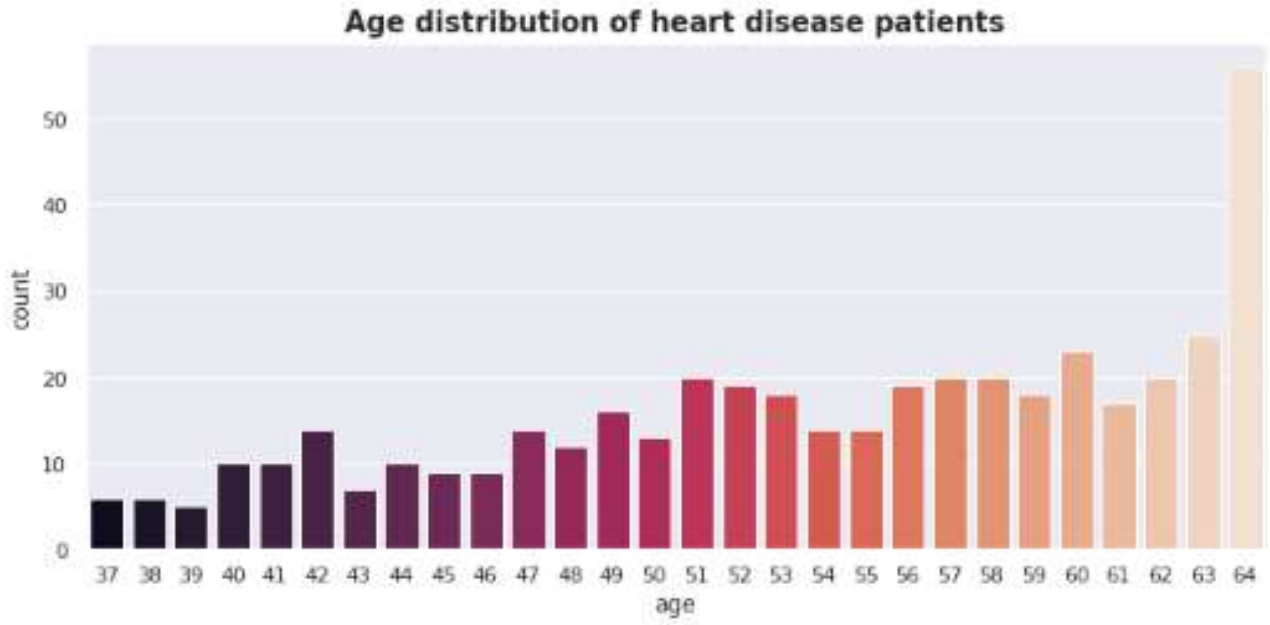
Analysis on the basis of cholesterol level





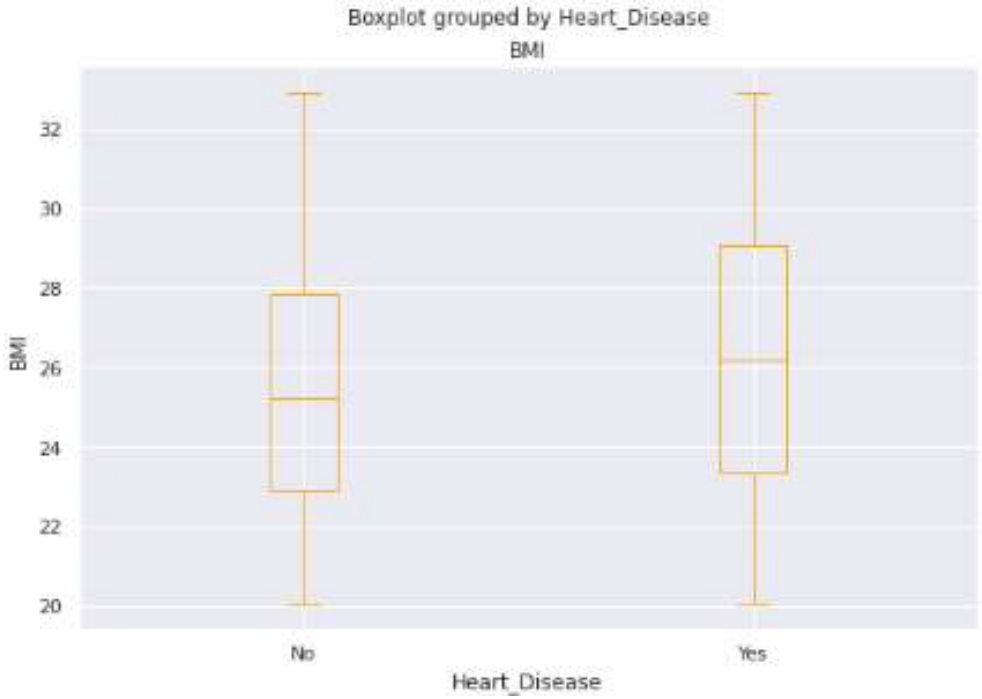
- Total Cholesterol level of heart patient seems to be slightly higher than normal patient
- People who have cholesterol level more than 240 are prone to heart problems.

Analysis on the basis of age



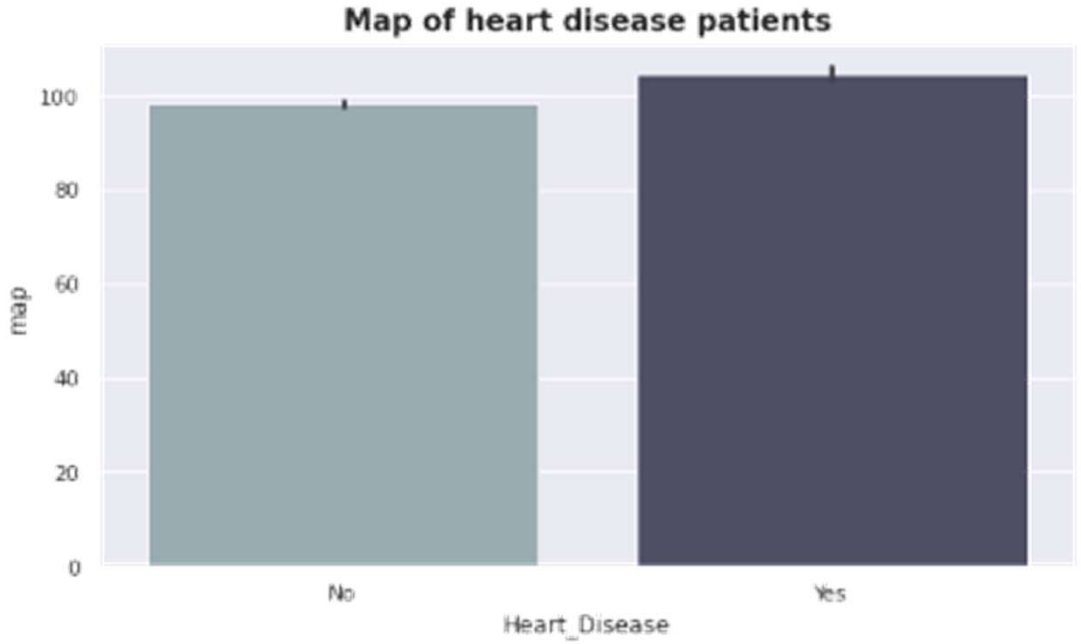
- According to the chart, as age increases, the chances of suffering from heart problems are more likely.

Analysis on the basis of BMI



- According to the boxplot, Higher BMI leads to higher chances of Heart Disease

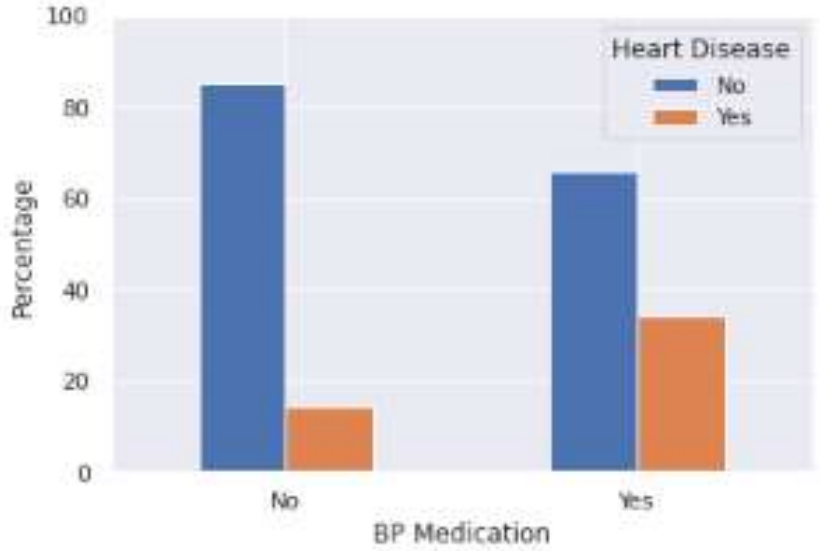
Analysis on the basis of MAP (Mean Arterial Pressure)



24-H MAP Categories	24-H MAP Thresholds, mm Hg
Normotension	<90
Elevated BP	90 to <92
Stage-1 HT	92 to <96
Stage-2 HT	≥96

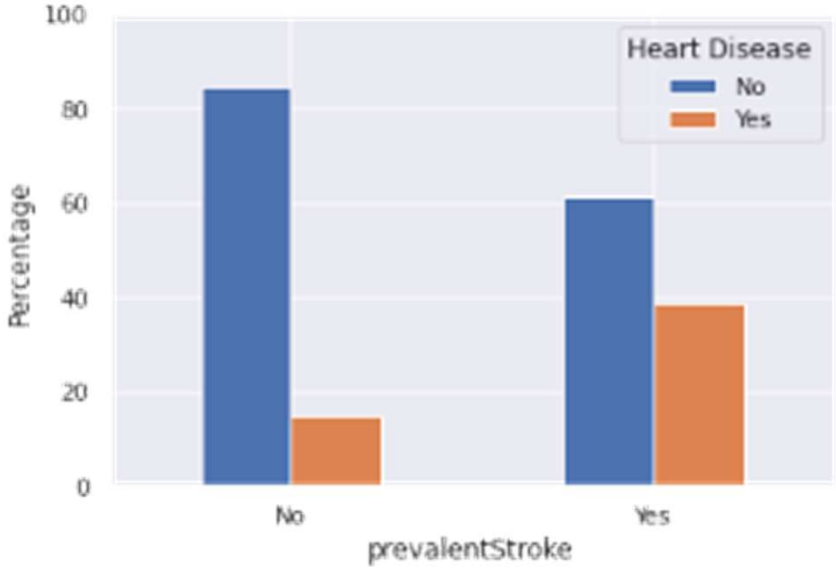
- In the above graph, the heart disease patients have higher MAP
- If the value of MAP is above 96, the patient is more prone to Heart Disease or suffer Hypertension

Analysis on the basis of Blood Pressure medication



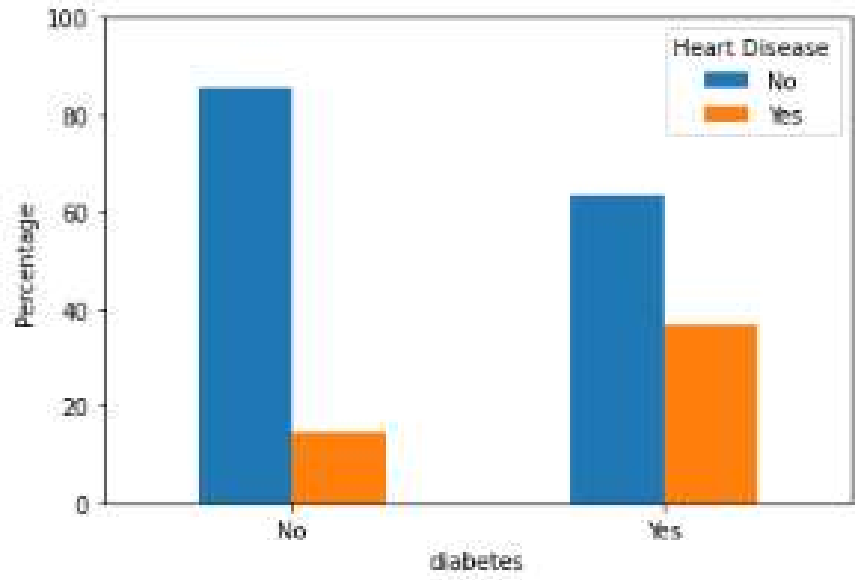
- According to the chart, People who take Blood pressure medication have a higher chance of suffering from heart disease.

Analysis on the basis of Prevalent Stroke



- According to the chart, people who previously had a stroke are more likely to suffer from Heart Disease.

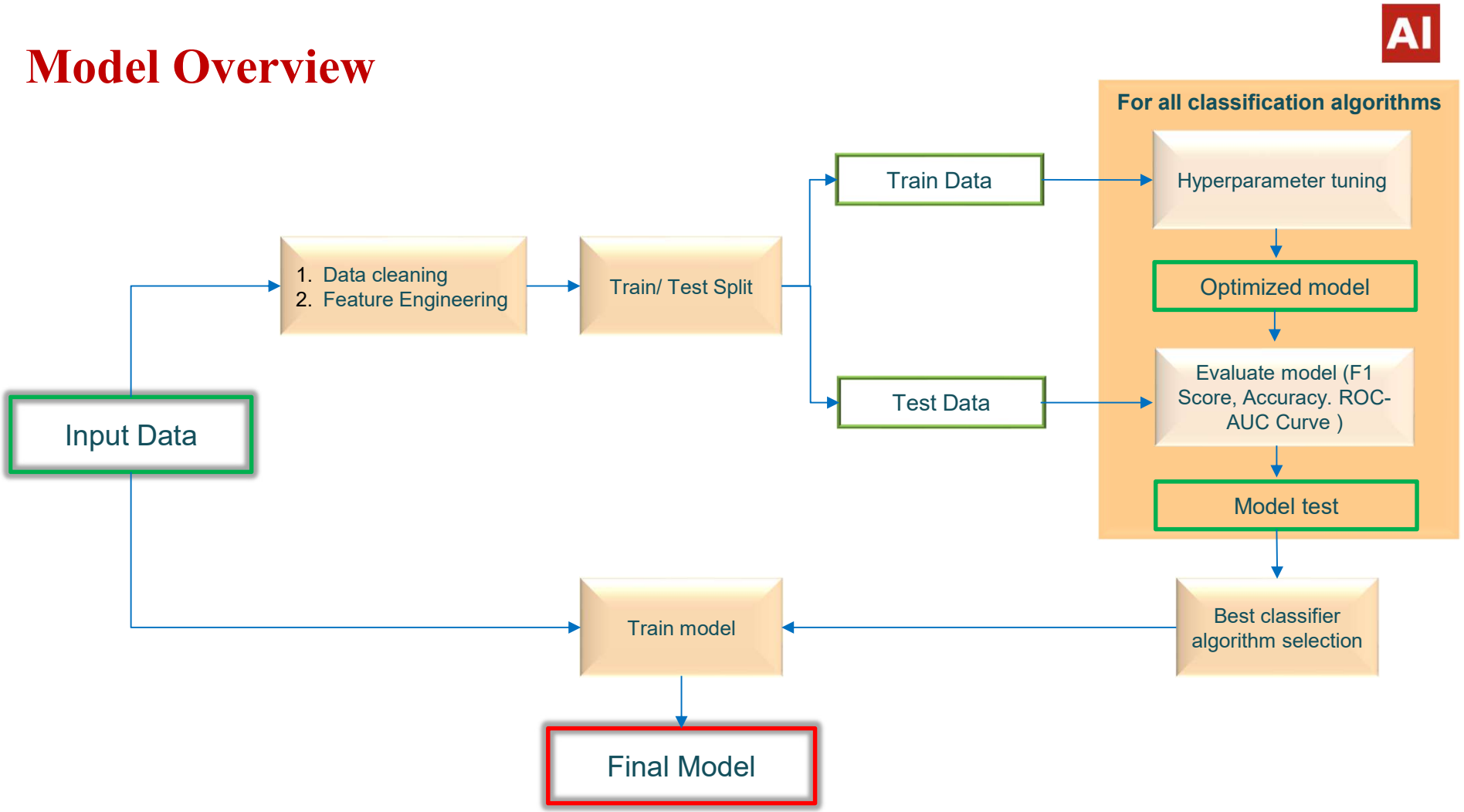
Analysis on the basis of Diabetes



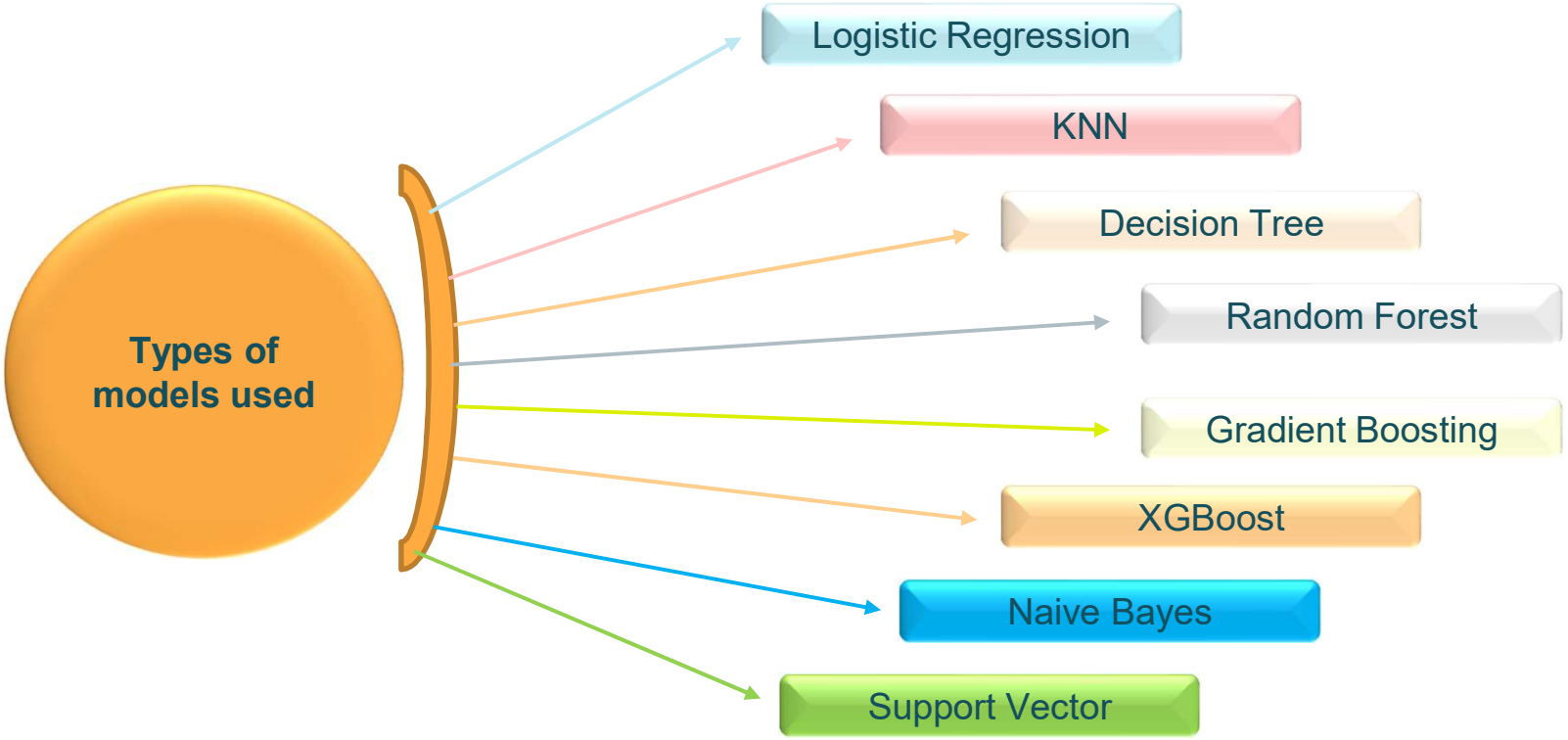
- According to the bar chart, Diabetic person is more likely to suffer from a heart disease.

Model Overview

Model Overview



Types of models used



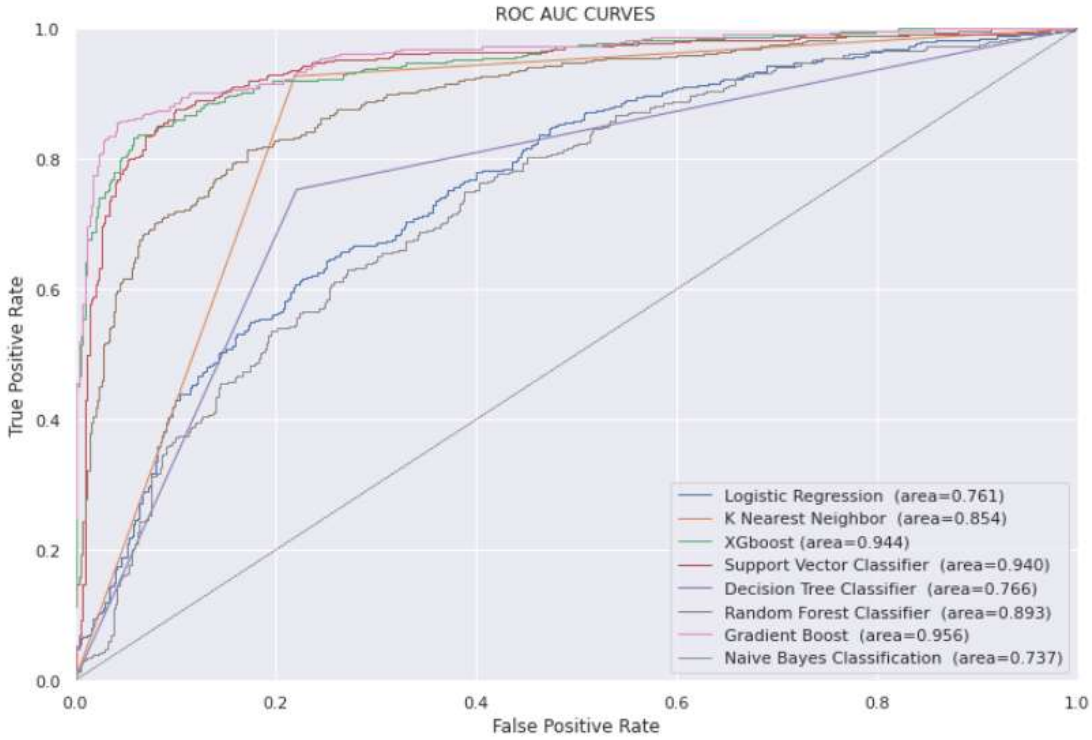
Model Analysis

Model's Evaluation Matrices

Model	Accuracy	Precision	Recall	Specificity	F1 Score	ROC
Logistic Regression	0.70	0.66	0.66	0.72	0.66	0.76
KNN	0.85	0.77	0.93	0.78	0.84	0.85
DecisionTree	0.77	0.73	0.75	0.78	0.74	0.77
Random Forest	0.82	0.80	0.79	0.84	0.79	0.89
GradientBoosting	0.90	0.89	0.87	0.91	0.88	0.96
XGBoost	0.88	0.87	0.86	0.89	0.86	0.94
Naive Bayes	0.68	0.68	0.54	0.79	0.60	0.74
SupportVector	0.88	0.86	0.88	0.88	0.87	0.94

- According to the table, Gradient Boosting has performed best among all the models in terms of evaluation parameters such as Accuracy, Precision, F1 score, and ROC value.

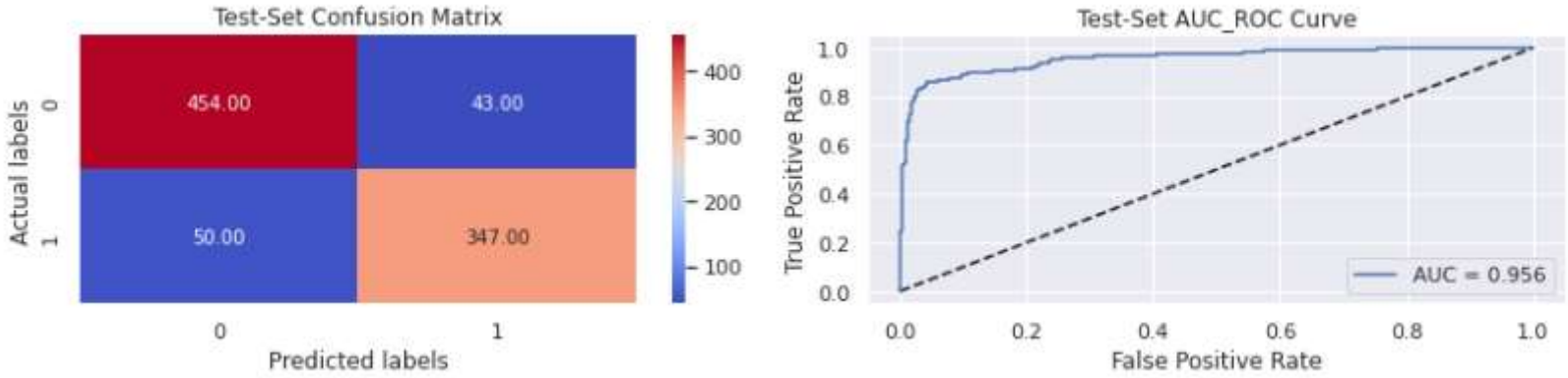
ROC-AUC curve



- Here, we can see the highest average area under the curve (AUC) of 0.96 is attained by Gradient Boost Classifier and second highest is of 0.94 attained by Support Vector Classifier and XG Boost

Model Features

Gradient boosting

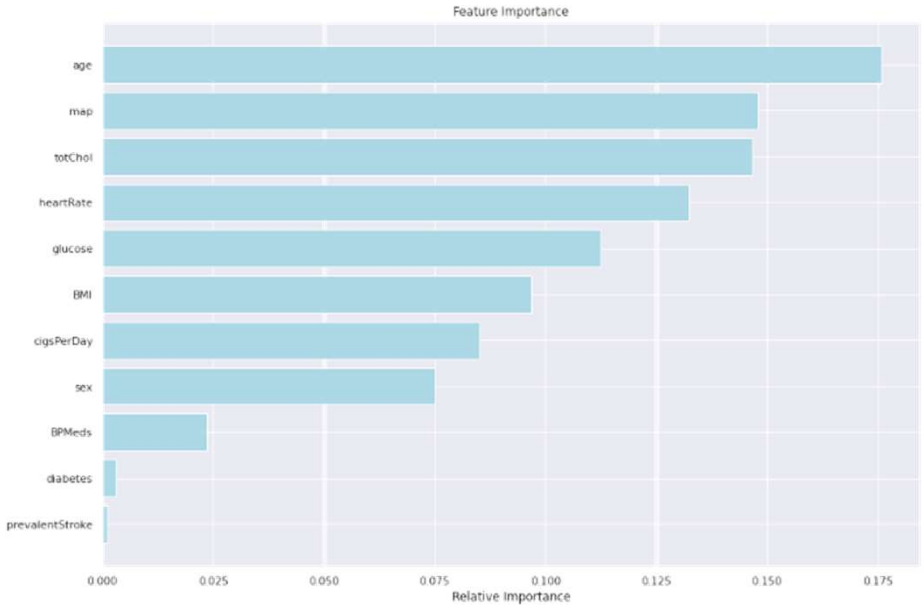


	Model	Accuracy	Precision	Recall	Specificity	F1 Score	ROC
0	GradientBoosting	0.895973	0.889744	0.874055	0.913481	0.88183	0.955739

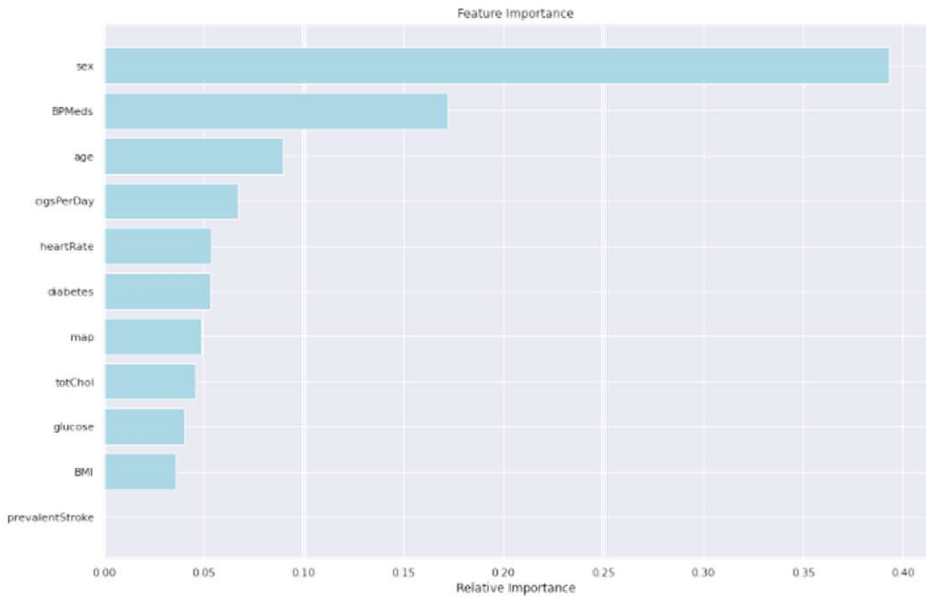
- The average area under the curve (AUC) attained by Gradient Boost Classifier is 0.95
- Gradient Boost Classifier has model accuracy of 0.89

Important Features

Gradient Boosting

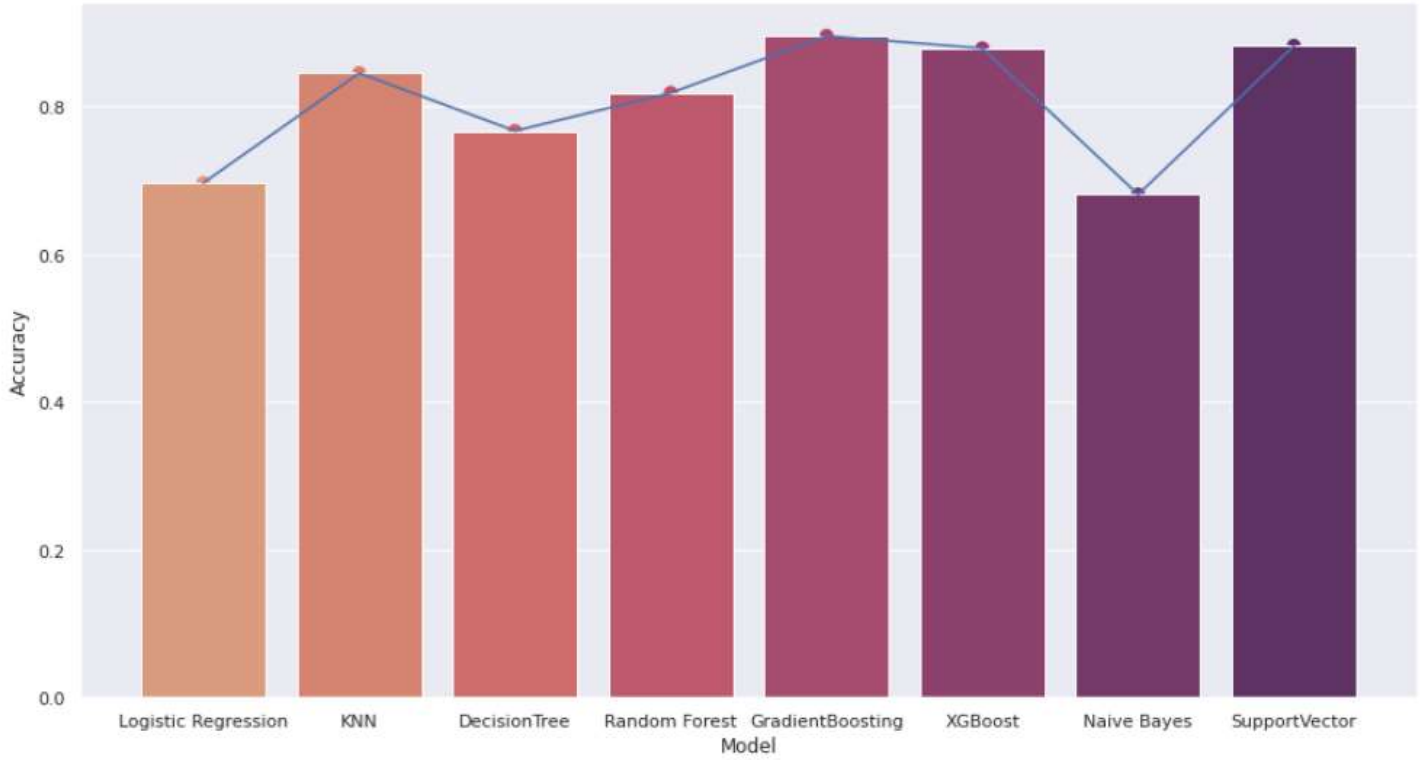


XGBoost



- According to Gradient Boosting, age is the most important feature and has the highest impact on 10-year risk of coronary heart disease CHD
- According to XGBoost, sex is the most important feature and has the highest impact on 10-year risk of coronary heart disease CHD

Accuracy of Models Performed



- We can see the Highest Accuracy among all the models is of Gradient Boosting followed by Support Vector Classifier and XG Boost models.

Conclusion - EDA

- The dataset contains 85% normal persons and 15% heart patients
- Given dataset consists of 55% male and 45% female.
- Males are more prone to heart disease as compared to females.
- As age increases, the chances of suffering from heart problems are more likely.
- Higher BMI leads to higher chances of Heart Disease.
- Higher cholesterol indicates the higher chances of getting Heart Disease.
- If the value of MAP is above 96, the patient is more prone to Heart Disease or suffer Hypertension
- People who take Blood pressure medication have a higher chance of suffering from heart disease.
- People who previously had a stroke are more likely to suffer from Heart Disease.
- Diabetic person is more likely to suffer from a heart disease.

Conclusion – Classification Models

- **Gradient boost model is the most accurate model** among all the models, on the basis of evaluation parameters such as **Accuracy (90%), Precision (89%), Specificity (91%), F1 score (88%), and AUC-ROC score (96%).**

- **Age** is the most important feature according to **Gradient boost**

- **Logistic Regression model** has the **least Accuracy (70%).**

- Best performance of Models on test data based on evaluation metrics for class 1:
  1. Recall - KNN
  2. Precision – Gradient Boost
  3. F1 Score – Gradient Boost
  4. Accuracy – Gradient Boost

Thank You