

# Capstone Project - 3

## Cardiovascular Risk Prediction

Submitted by

Patil Prasad

Wagh Kiran

Deore Prasad

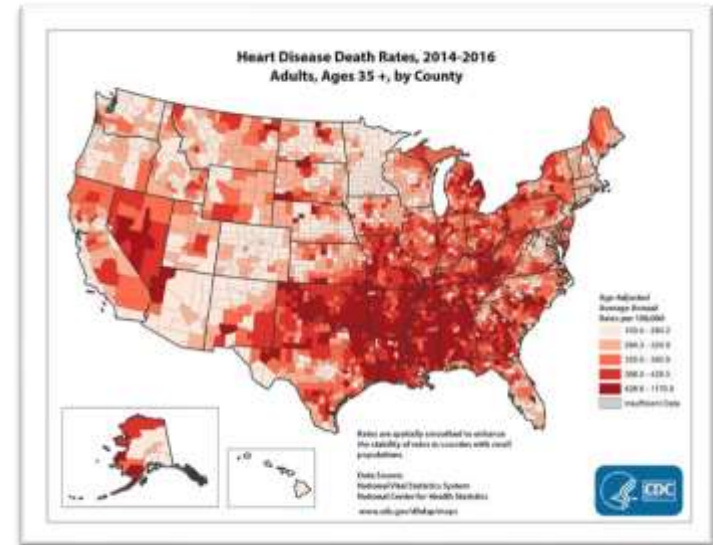
# Agenda

- Problem Statement
- Data Summary
- Handling Missing Values
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Handling Skew
- Modelling Approach
- Predictive Modelling
- Model Comparison
- Challenges Faced
- Conclusions



# Abstract

- Cardiovascular diseases (CVDs) are the major cause of mortality worldwide.
- According to WHO, **17.9 million** people died from CVDs in 2019, accounting for **32%** of all global fatalities.
- Though CVDs cannot be treated, predicting the risk of the disease and taking the necessary precautions and medications can help to avoid severe symptoms and, in some cases, even death.



# Problem Statement

- The goal of this project is to develop a classification model that can predict whether a patient is at risk of **coronary heart disease (CHD)** over the period of 10 years, based on demographic, lifestyle, and medical history.
- The data was gathered from **3390** adults participating in a cardiovascular study in Framingham, Massachusetts.



# Data Summary

## ➤ Demographic

- Sex
- Age
- Education

## ➤ Current Medical Status

- Total cholesterol
- Systolic BP
- Diastolic BP
- BMI
- Heart rate
- Glucose

## ➤ Behavioral

- Is smoking
- Cigarettes per day

## ➤ Medical History

- BP medication
- Prevalent stroke
- Prevalent hypertension
- Diabetes
- Ten year risk of CHD → DV

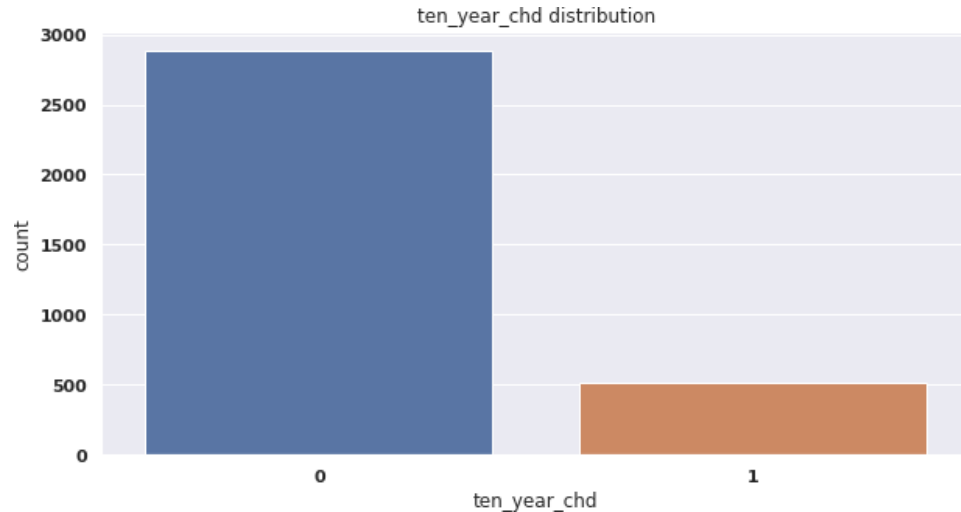
# Handling Missing Values

- There were a total of **510** missing values in the dataset.
- Total missing values and how they were handled are as follows:
  - Education(87) , BP Medication(44) – **mode** imputation
  - Cigarettes per day(22) – imputed with **median** cigarettes per day for **smokers**
  - Total cholesterol(38), BMI(14), Heart rate(1) – **median** imputation
  - Glucose(304) – **KNN** imputation with  $k = 10$



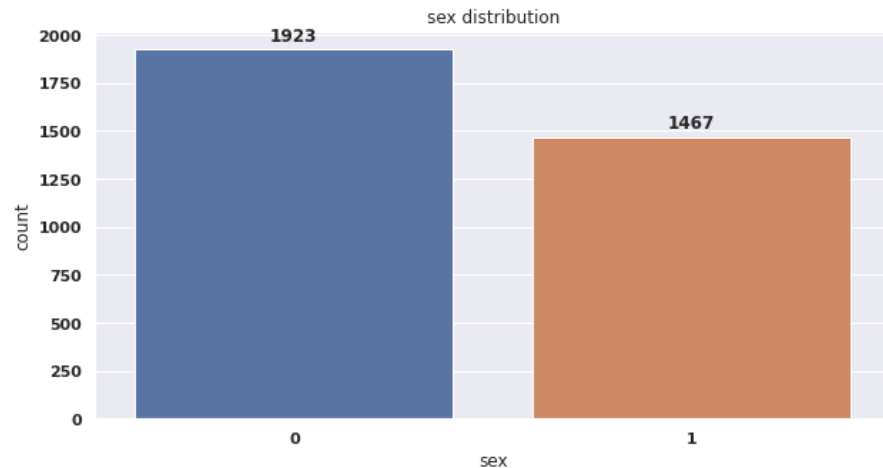
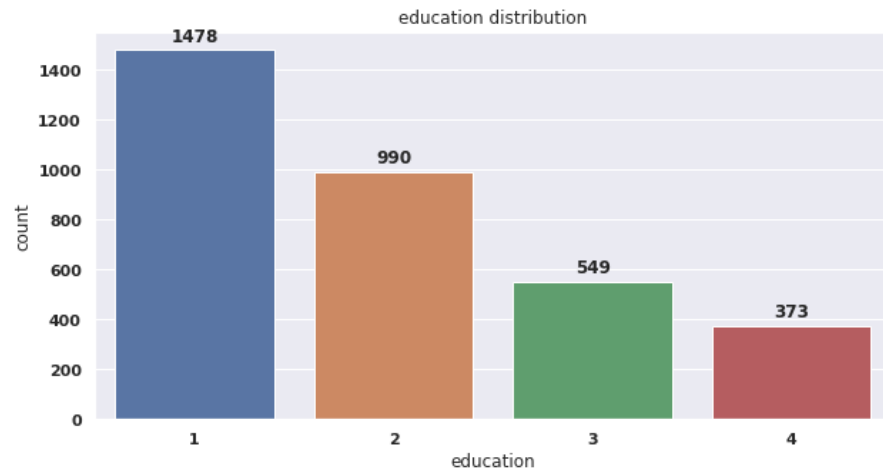
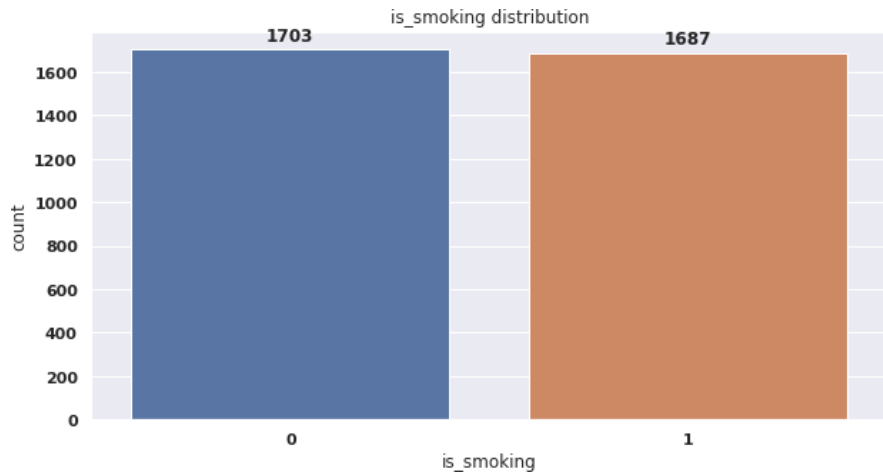
# Exploratory Data Analysis (EDA)

- The dependent variable is imbalanced, with just **~15%** of patients testing positive for CHD.
- All continuous independent variables are **positively skewed** except age, which is almost normally distributed.



# EDA (Contd.)

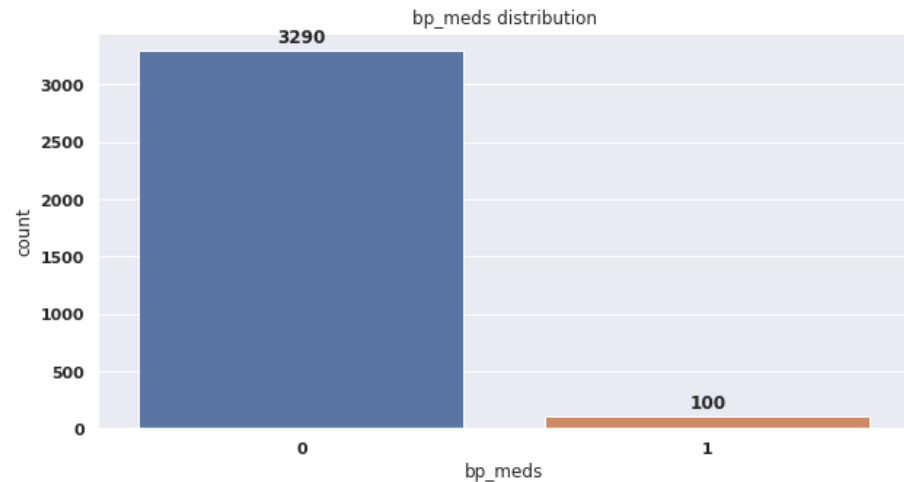
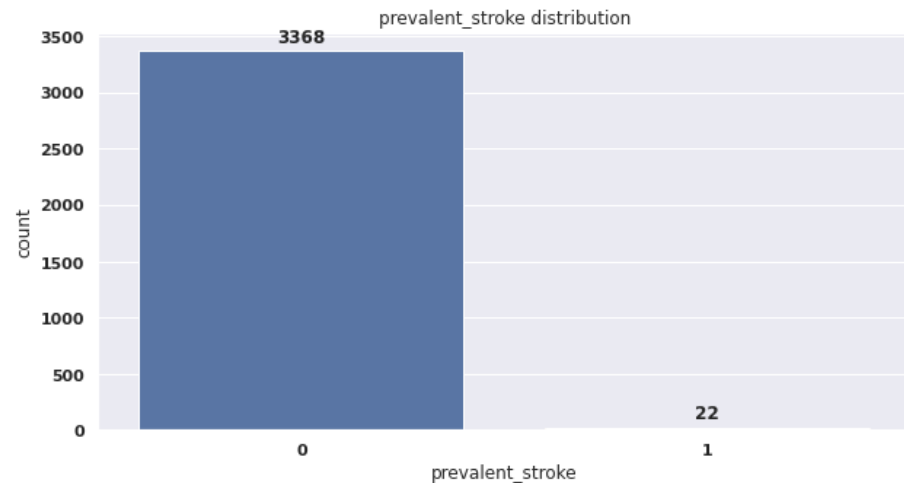
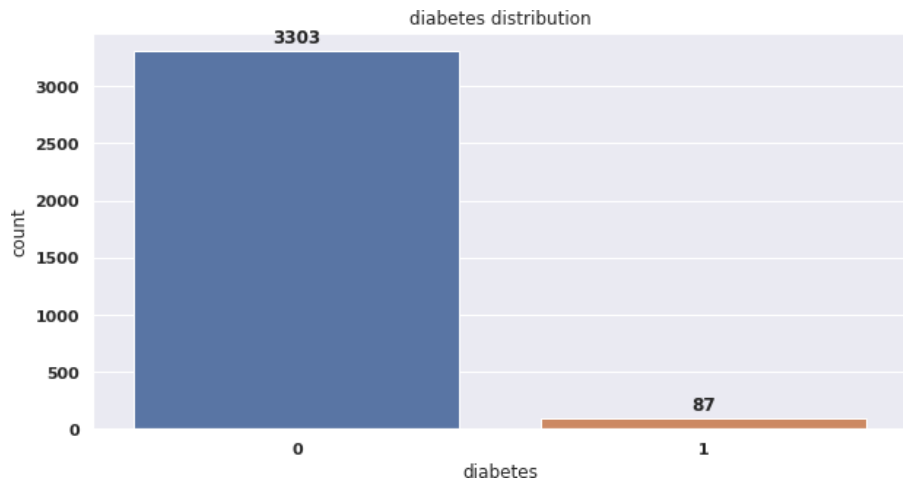
- **Half** the patients are smokers
- There are **more female** patients than male
- Most patients have education **level 1**





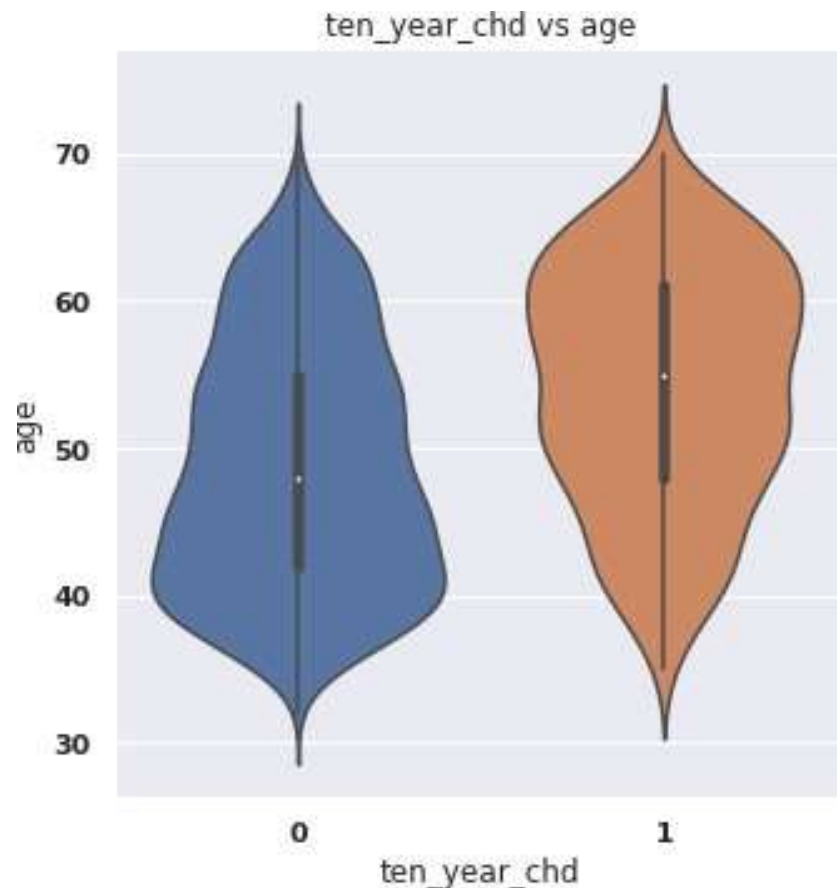
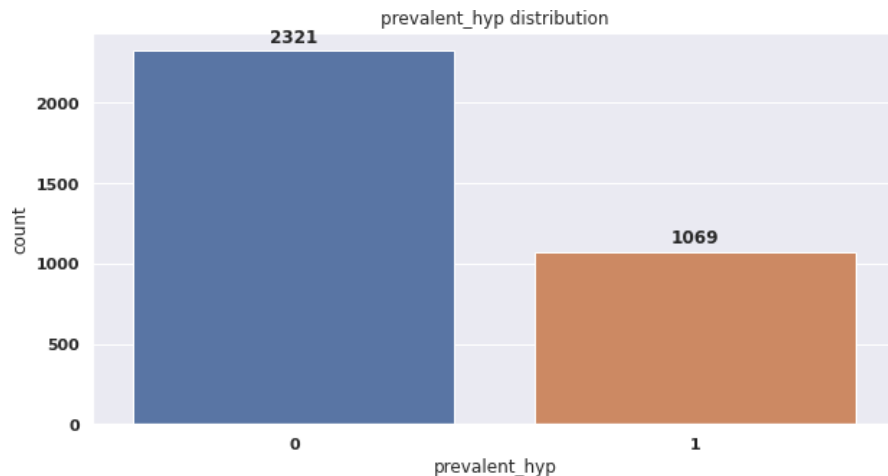
# EDA (Contd.)

- There are relatively **few** individuals who have had a stroke, have diabetes, or are using blood pressure medicine.



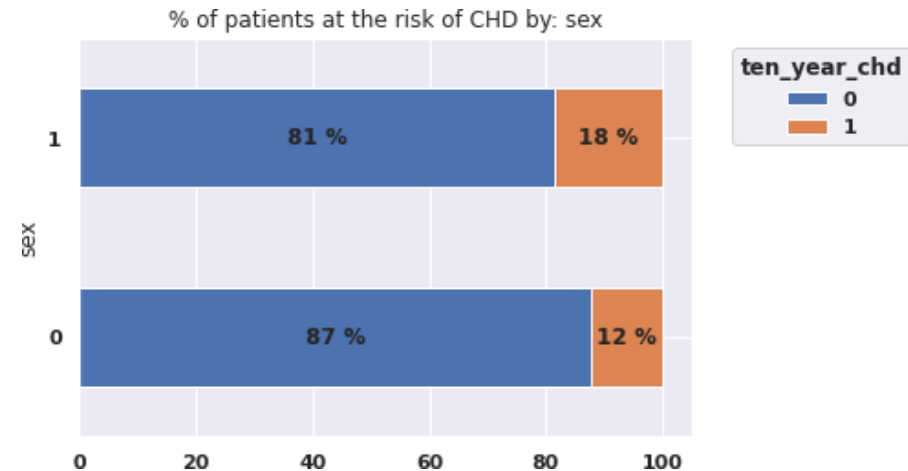
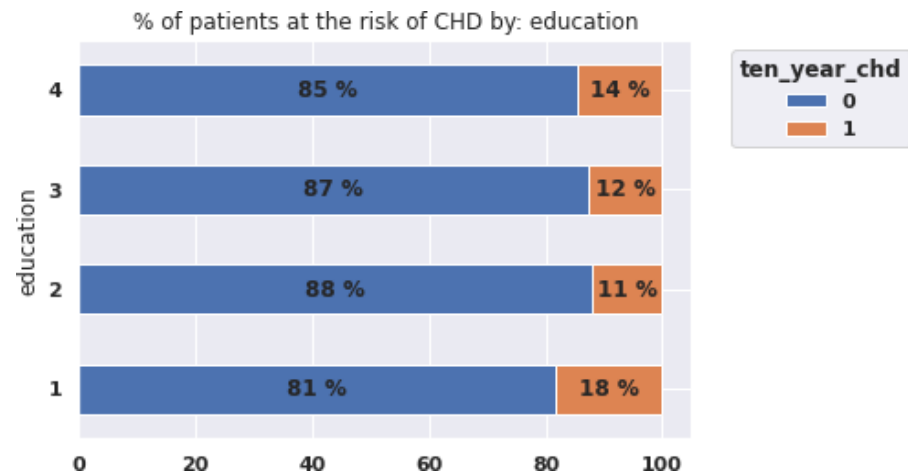
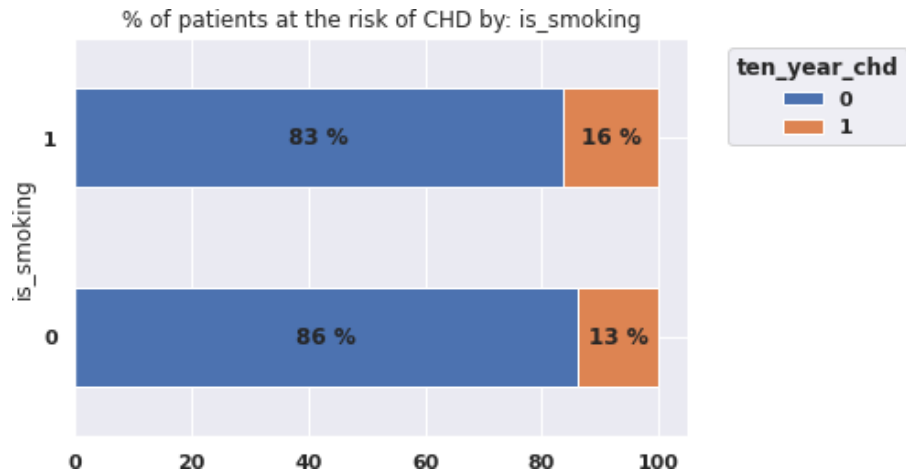
## EDA (Contd.)

- 1069 patients have **hypertension**
- The risk of CHD **increases with age**



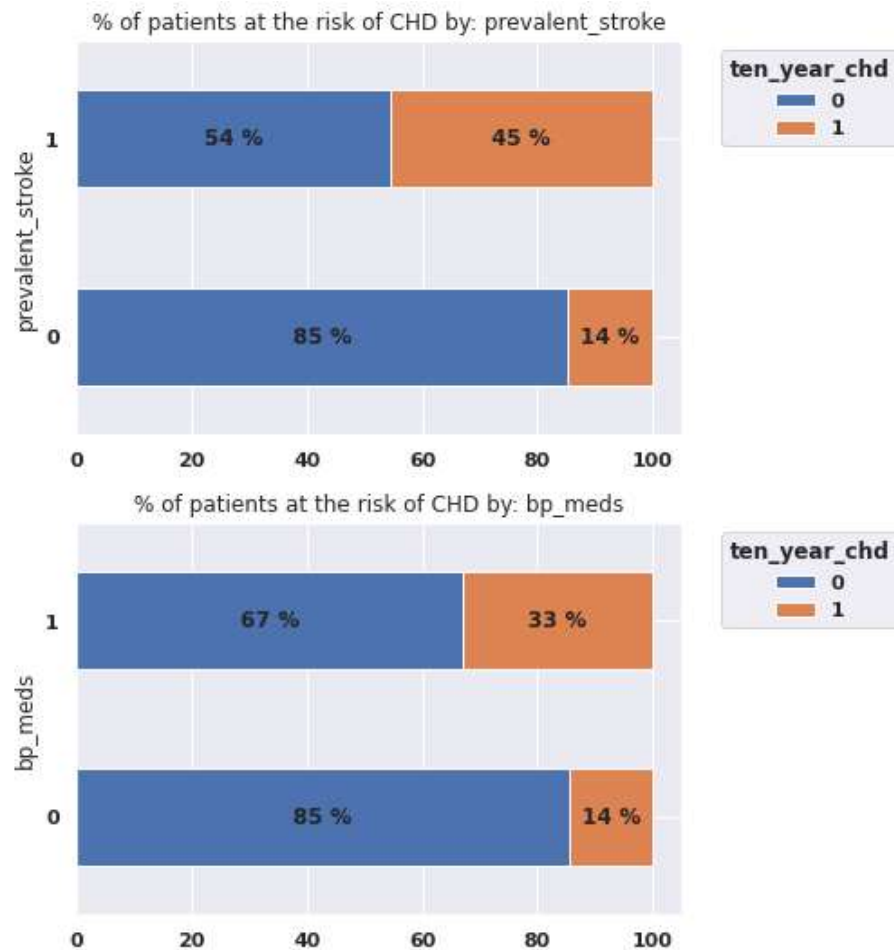
# EDA (Contd.)

- The risk of CHD **varies** by educational level, gender, and whether or not the patient smokes.



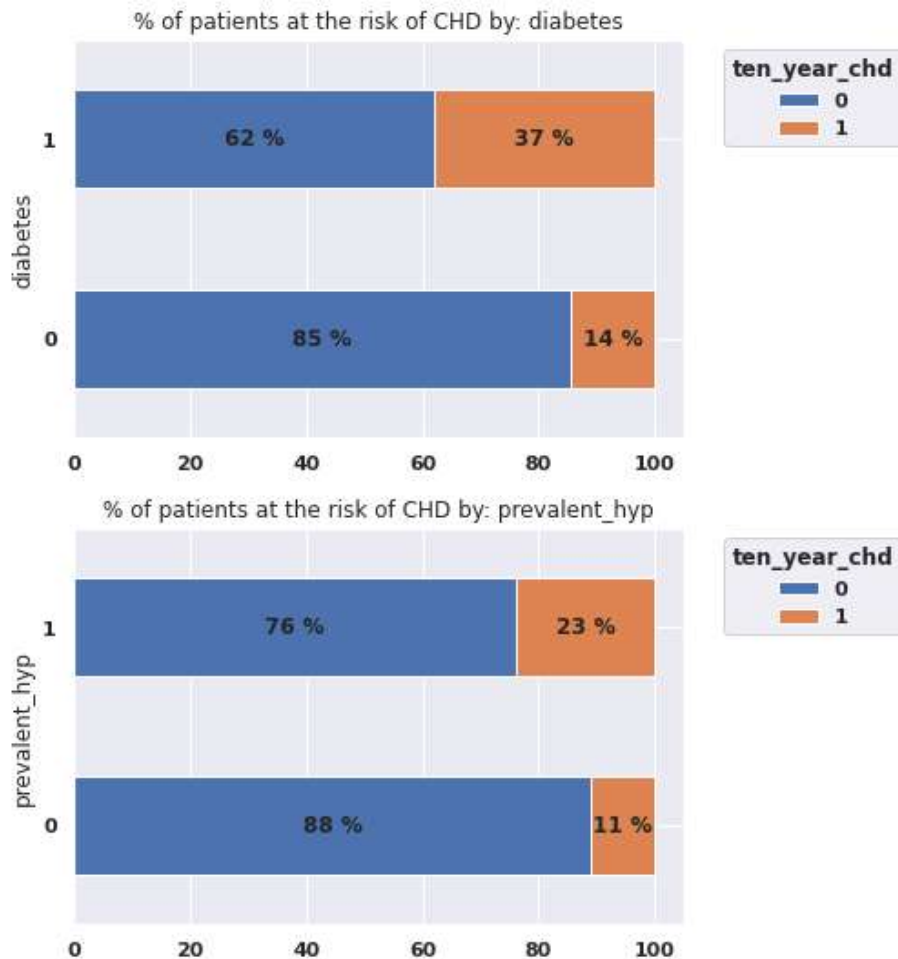
## EDA (Contd.)

- Patients who have had a **stroke** or are presently on **blood pressure medication** are more likely to test positive for CHD.



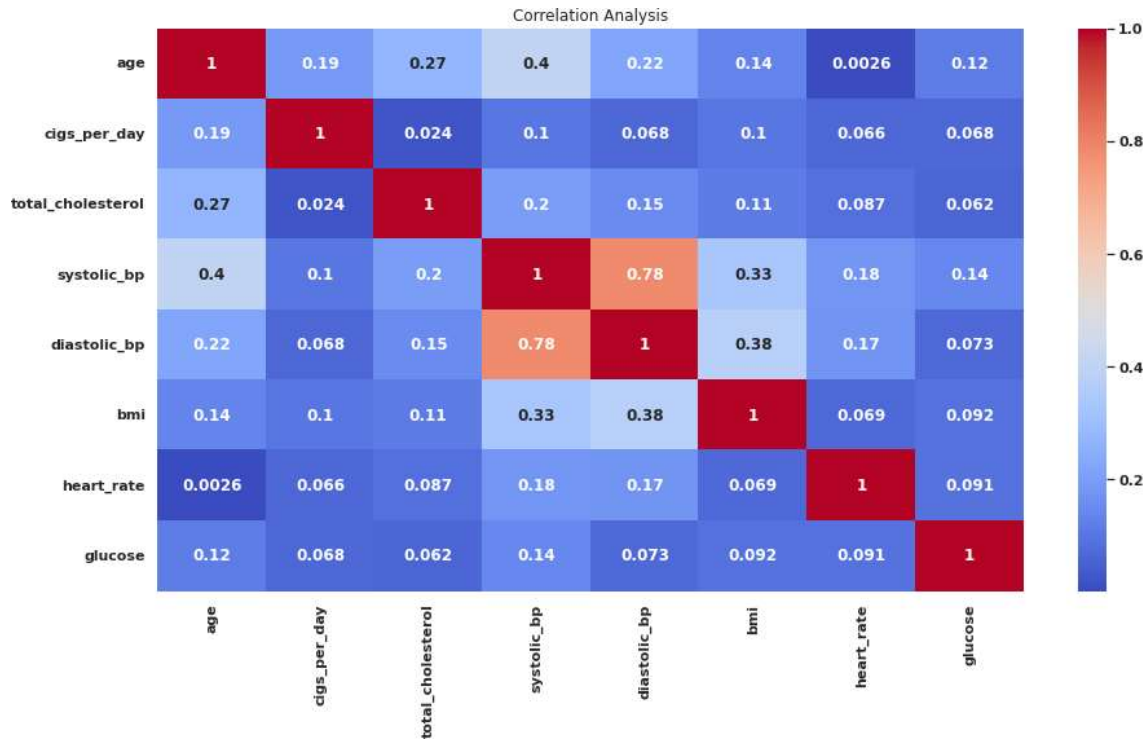
## EDA (Contd.)

- Patients with **hypertension** or **diabetes** are more likely to be diagnosed with CHD.



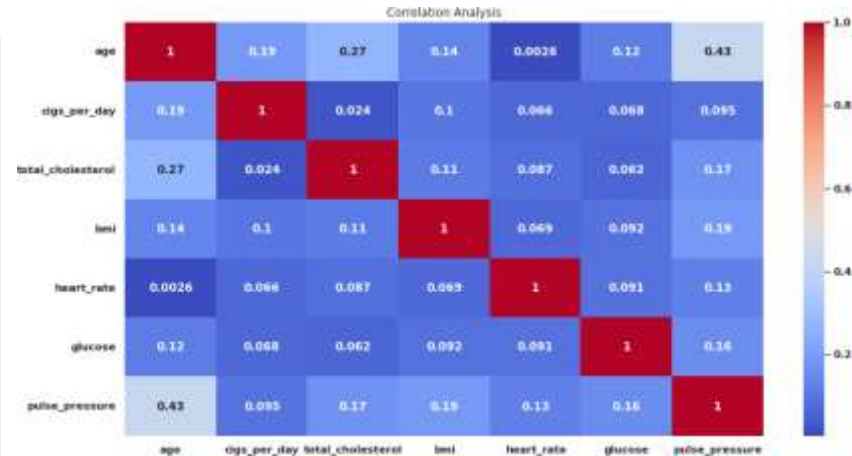
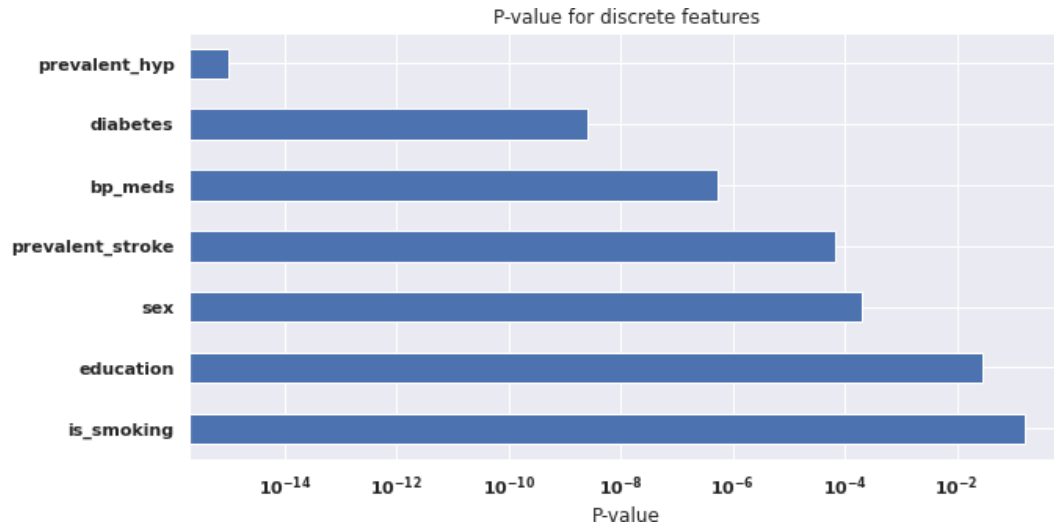
# EDA (Contd.)

- Systolic BP and diastolic BP are **highly correlated**



# Feature Engineering

- **Pulse Pressure = Systolic BP – Diastolic BP**
- The **Chi2 test** on discrete features indicates that the **'is\_smoking'** column has the highest p-value and so is the **least relevant feature**. As a result, we **drop** it.



# Handling Skew and Outliers

- The skew in numeric variables is reduced by performing **log transformation**.
- The outliers beyond **3 standard deviations** from the mean were **imputed** with the **median** value

Attribute	Original skew	Skew After Transformation
Age	0.225796	-0.015053
Cigarettes Per Day	1.204077	0.275072
Total Cholesterol	0.948170	0.011860
BMI	1.025551	0.370422
Heart Rate	0.676660	0.165898
Glucose	6.361911	2.310918
Pulse Pressure	1.412382	0.354174



# Summary so far...

- We defined the problem statement
- Handled the missing values
- Created data visualizations
- Performed feature engineering and feature selection
- Transformed numeric variables to reduce skew.
- Handled outliers.

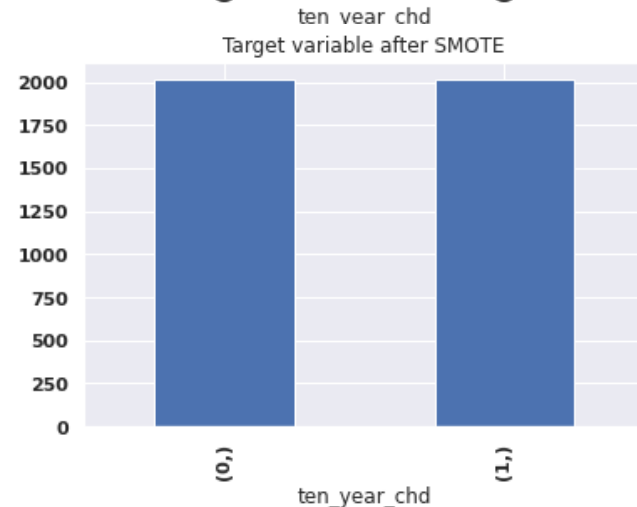
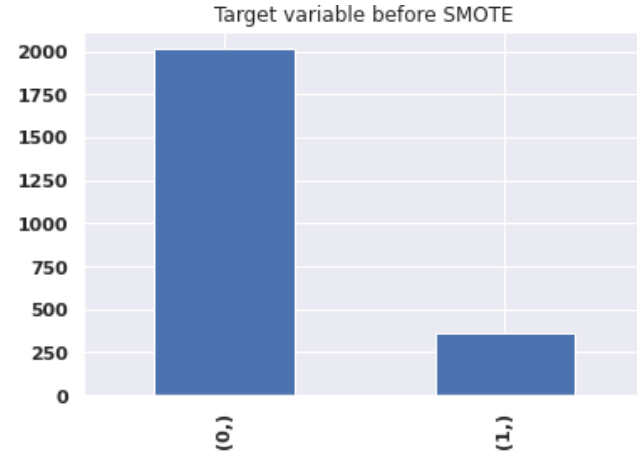


# Modelling Approach

- Data points in test data set = 30%
- Choice of split: **Repeated stratified K fold**,  $k = 4$
- Evaluation metric: **Recall**

$$\text{Recall} = \frac{\text{True Positive}}{\text{False Negative} + \text{True Positive}}$$

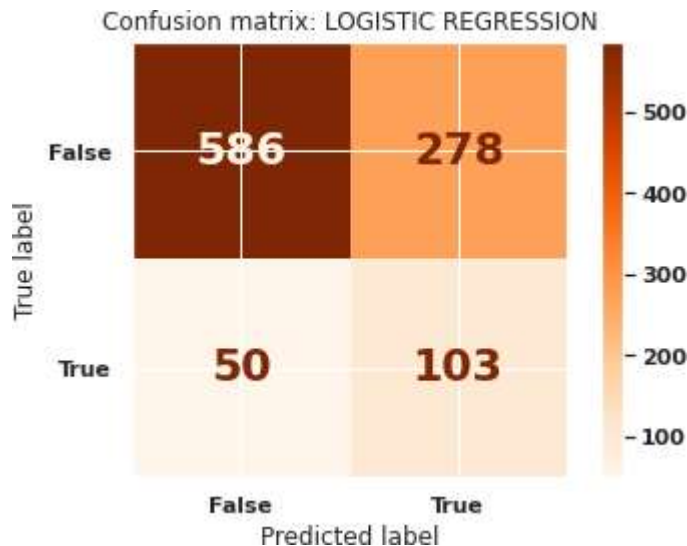
- Hyperparameter tuning: **Grid search**
- Oversampling strategy: **SMOTE**
- Data points before SMOTE = **2373**
- Data points after SMOTE = **4030**
- Scaler used: **Standard Scaler**



# Logistic Regression

## Evaluation metrics:

- Train Recall = 0.6987
- Test Recall = 0.6732
- Test Accuracy = 68%



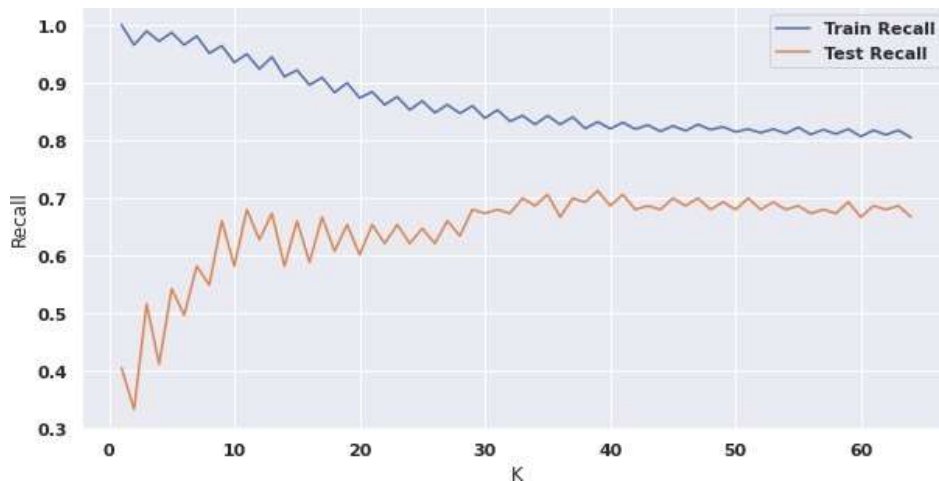
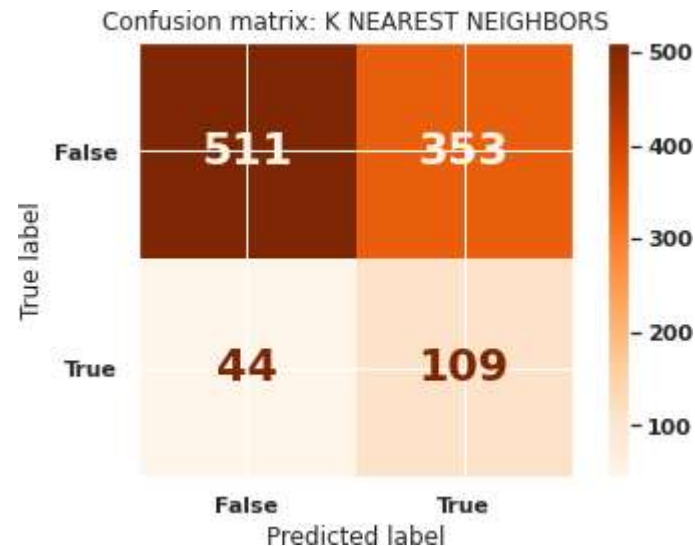
# K-Nearest Neighbors

## Parameters:

- $K = 39$

## Evaluation metrics:

- Train Recall = 0.8317
- Test Recall = 0.7124
- Test Accuracy = 61%



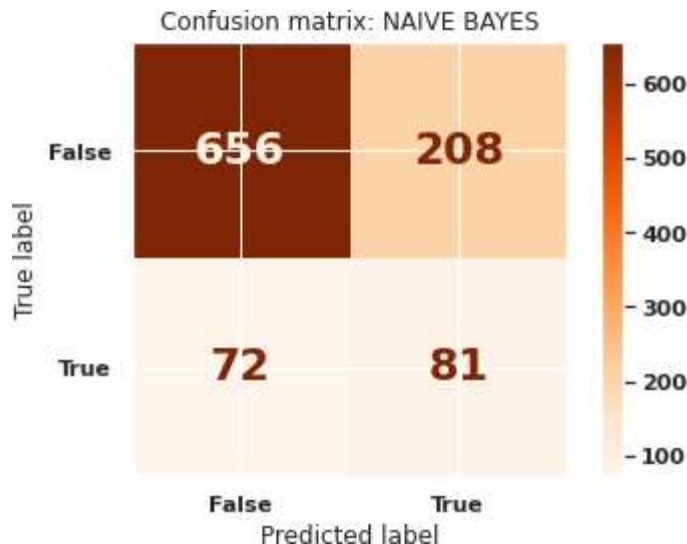
# Naïve Bayes

## Parameters:

- `var_smoothing=10`

## Evaluation metrics:

- Train Recall = 0.5811
- Test Recall = 0.5294
- Test Accuracy = 72%



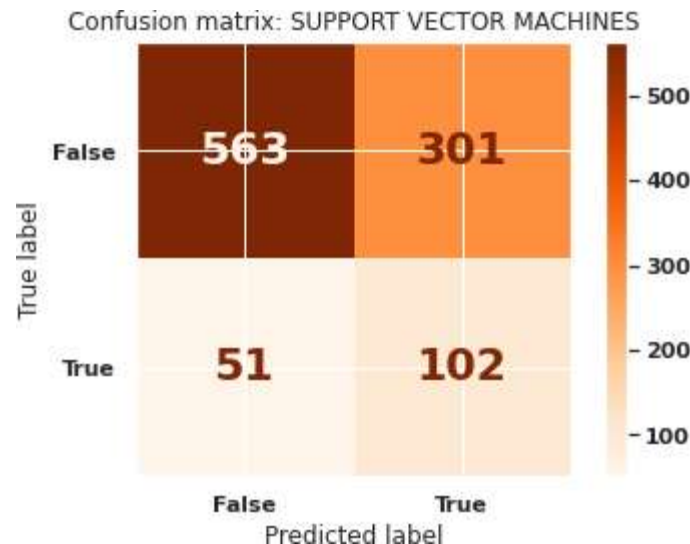
# Support Vector Machines

## Parameters:

- $C = 1$
- $\text{Gamma} = 0.01$
- $\text{Kernel} = \text{rbf}$

## Evaluation metrics:

- Train Recall = 0.7652
- Test Recall = 0.6666
- Test Accuracy = 65%



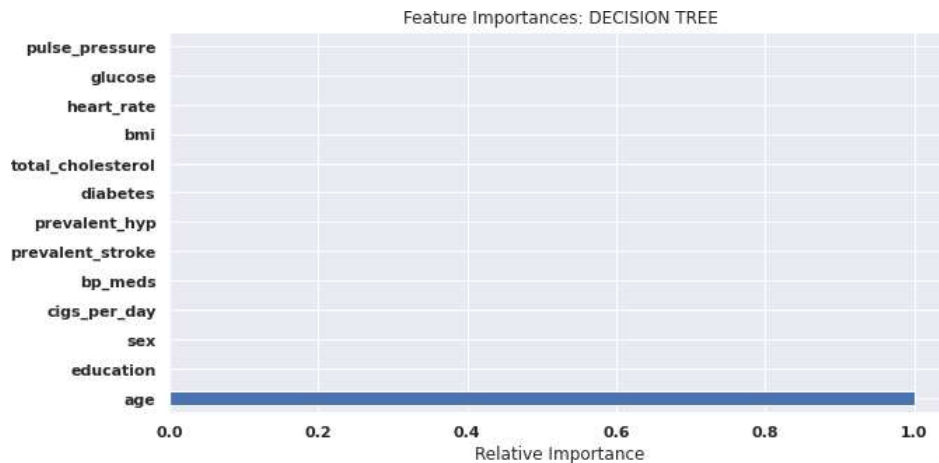
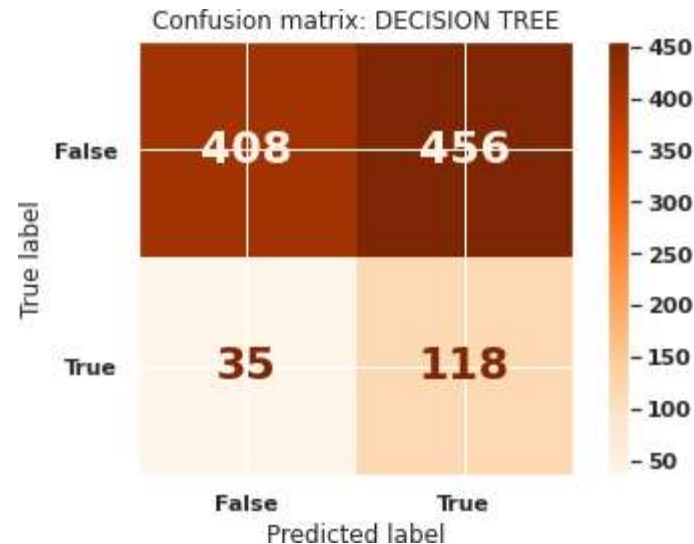
# Decision Tree

## Parameters:

- `max_depth = 1`
- `min_samples_leaf = 0.1`
- `min_samples_split = 0.1`

## Evaluation metrics:

- Train Recall = 0.8595
- Test Recall = 0.7712
- Test Accuracy = 52%



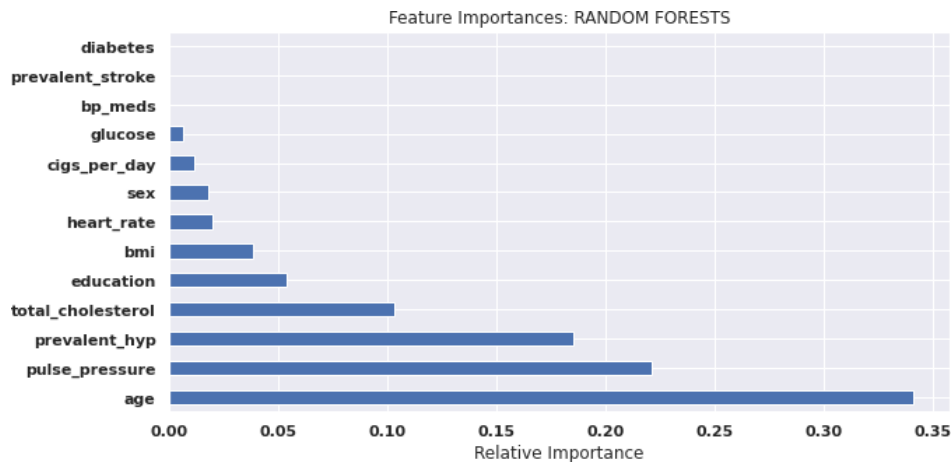
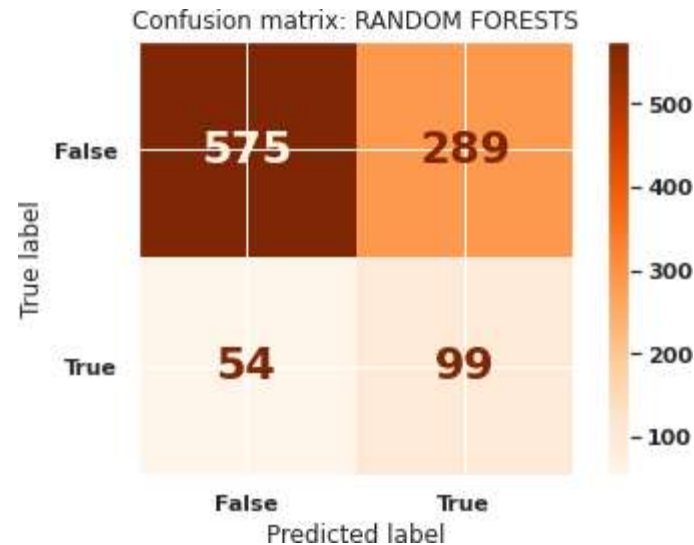
# Random Forests

## Parameters:

- `max_depth = 2`
- `min_samples_leaf = 0.1`
- `min_samples_split = 0.1`
- `n_estimators = 500`

## Evaluation metrics:

- Train Recall = 0.6997
- Test Recall = 0.647
- Test Accuracy = 66%





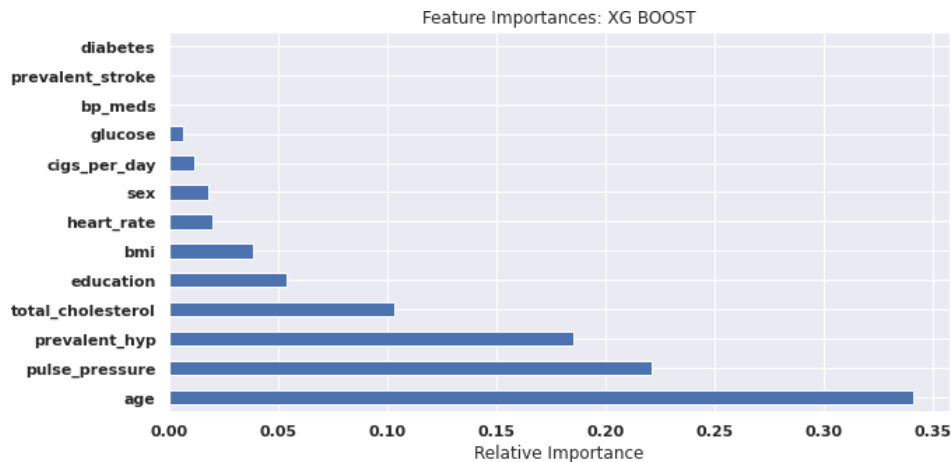
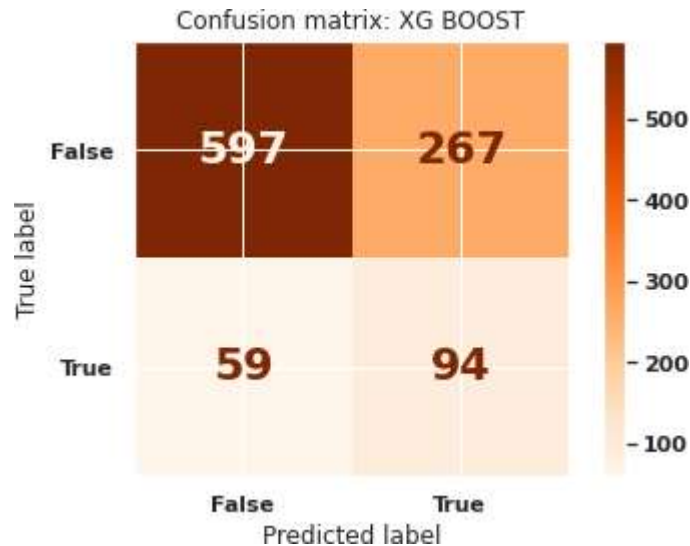
# XG Boost

## Parameters:

- `max_depth = 1`
- `min_samples_leaf = 0.1`
- `min_samples_split = 0.1`
- `n_estimators = 500`

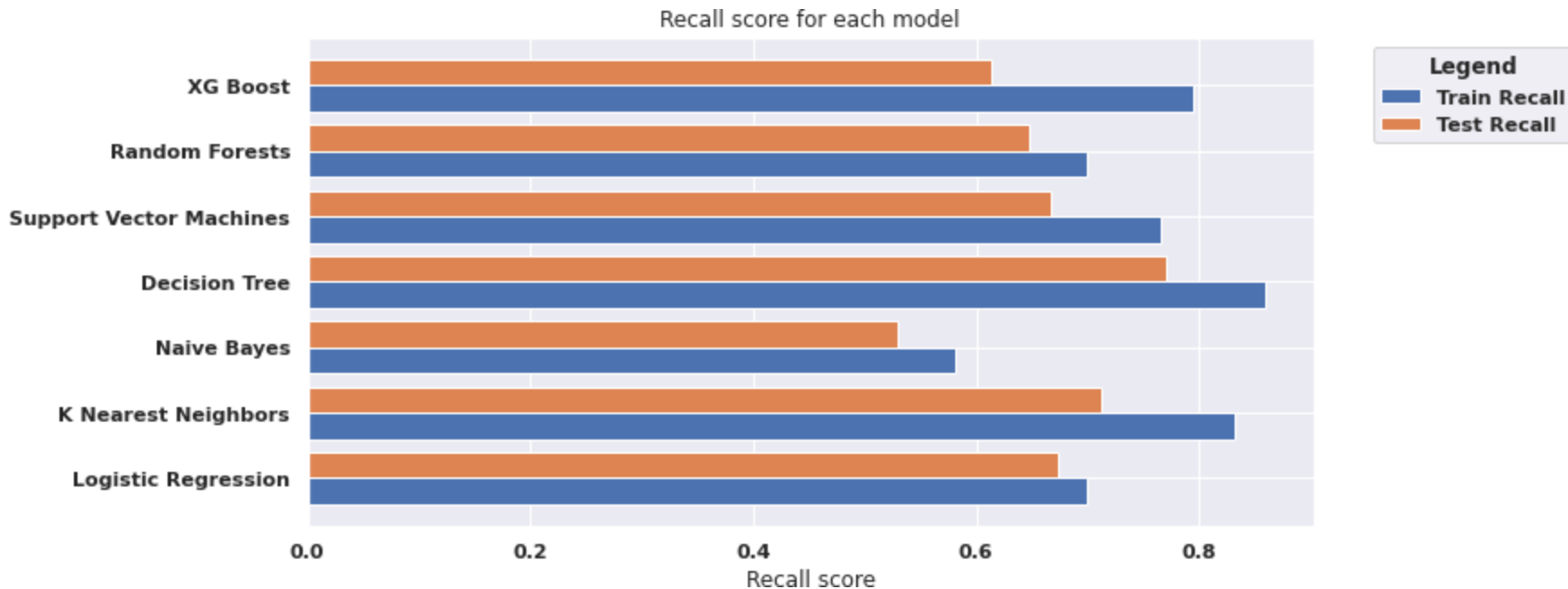
## Evaluation metrics:

- Train Recall = 0.7945
- Test Recall = 0.6143
- Test Accuracy = 68%



# Model Comparison

- The **Decision Tree** model has the **highest test Recall** compared to other models.



# Challenges Faced

- Comprehending the problem statement, and understanding the business implications – understanding the importance of predicting the risk of this disease
- Handling missing values in the dataset, and working with limited availability of data
- Feature engineering – deciding on which features to be dropped / kept / transformed
- Choosing the best visualization to show the trends among different features clearly in the EDA phase
- Deciding on ways to handle skew and outliers
- Choosing the best hyperparameters, which prevents overfitting

# Conclusion

- We have successfully built predictive models that can predict a patient's risk for CHD based on their demography, lifestyle, and medical history.
- The predictive models built were evaluated using Recall, and it was found that decision tree (0.77) has the highest test recall compared to other models.
- Efforts must be put into gathering more data, and also include people who have undergone different medical conditions.
- Future developments must include a strategy to improve the model recall score, enabling us to save even more lives from this disease.

**Thank You!**