Capstone Project -3

Cardiovascular Risk Prediction

(Supervised Learning-Classification)

BY

Snehal D. Ramteke

Content

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

Introduction

- Cardiovascular disease (CVD) is **a general term for conditions affecting the heart or blood vessels**. It's usually associated with a build-up of fatty deposits inside the arteries and an increased risk of blood clots.
- Risk factors are attributes, characteristics, or exposures of a person that play a role in the development of cardiovascular disease, for example, your smoking status or your blood pressure.
- 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.
- Our major objective is to analyze the dataset and determine whether someone is suffering from Cardiovascular disease using machine-learning concepts.

Data Summary

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
- 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(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)

❖ **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).

Data Summary(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, are considered continuous because of the large The number of possible values.)

• Glucose: glucose level (Continuous) Predict variable (desired target)

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

Handling Missing Values

```
df.isna().sum()
```

```
id                   0
age                  0
education           87
sex                  0
is_smoking           0
cigs_per_day        22
bp_meds             44
prevalent_stroke     0
prevalent_hyp        0
diabetes             0
total_cholesterol   38
systolic_bp          0
diastolic_bp         0
bmi                 14
heart_rate           1
glucose            304
ten_year_chd         0
dtype: int64
```

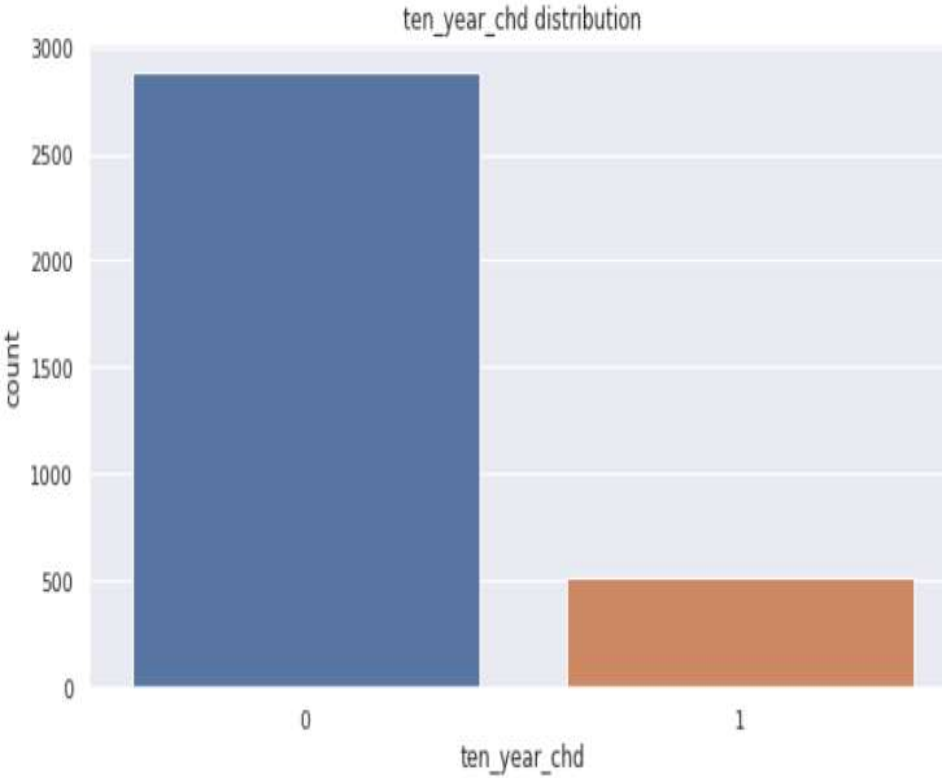
```
[22] # total null values
     df.isna().sum().sum()
```

```
510
```

- 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.

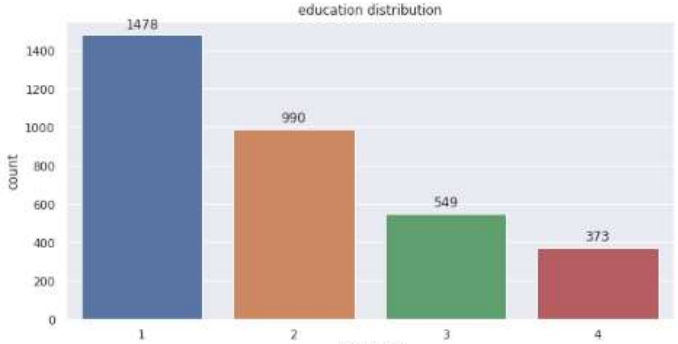
Exploratory Data Analysis (EDA)

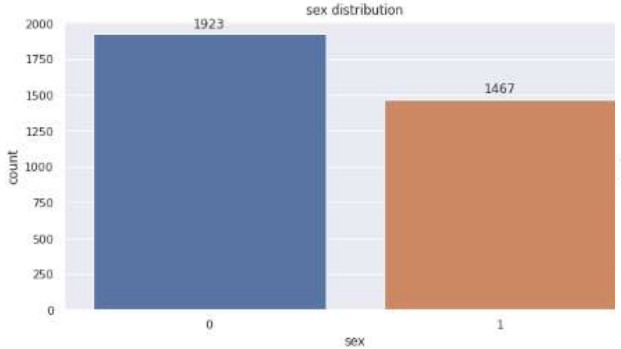
Distribution of the dependent variable

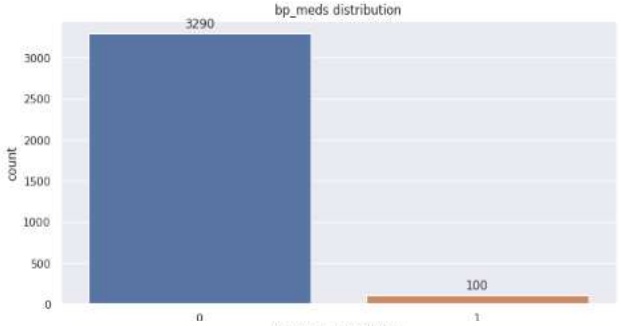


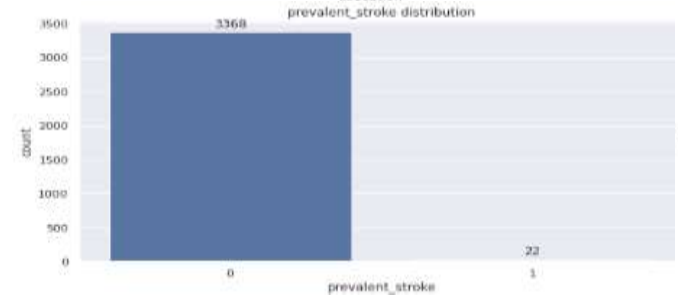
- The dependent variable – the 10-year risk of coronary heart disease is unbalanced. Only ~15% of the patients in the study were eventually exposed to the risk of this heart disease, the rest of the patients were not exposed to this disease after the end of the 10-year study.
- All continuous independent variables are positively skewed except age, which is almost normally distributed.

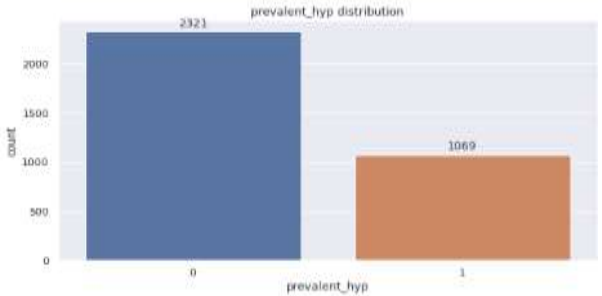
distribution of the discrete independent variables

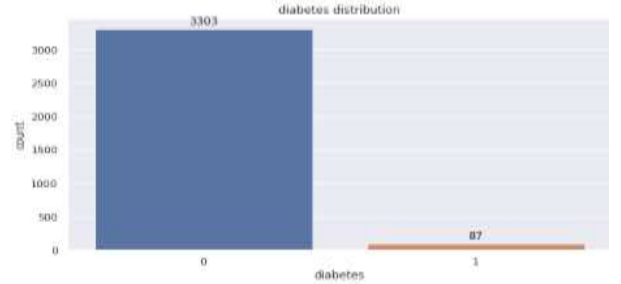






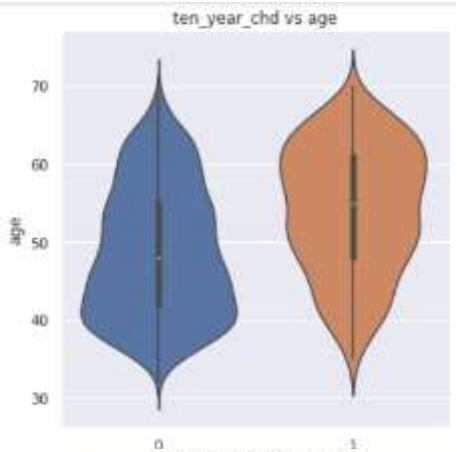


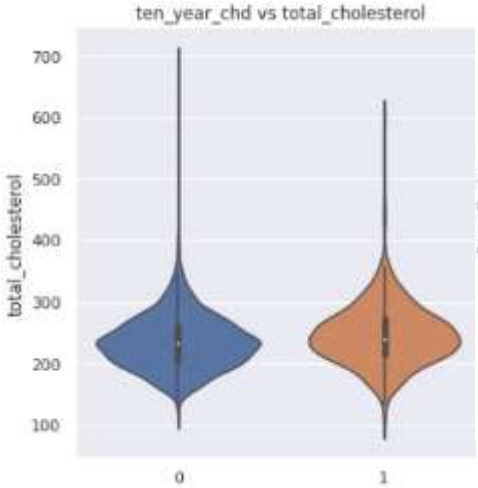


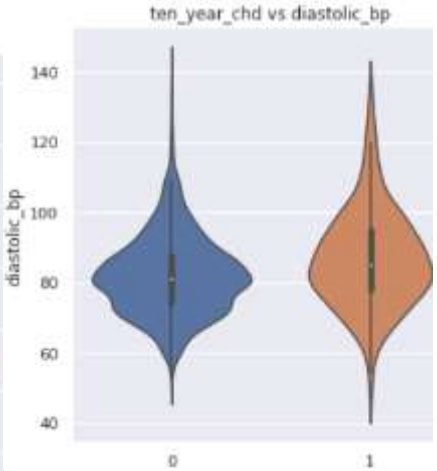


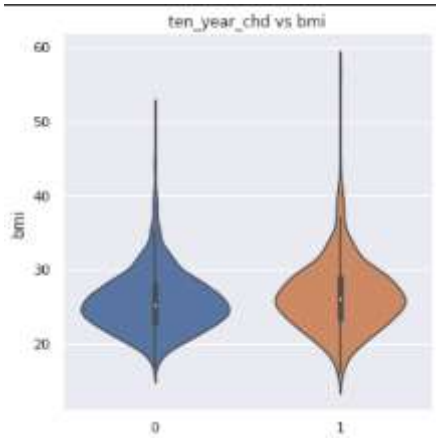
- **Majority of the patients belong to education level 1, followed by 2, 3, and 4 respectively.**
- **There are more female patients compared to male patients.**
- **Almost half the patients are smokers.**
- **100 patients under the study are undertaking blood pressure medication.**
- **22 patients under the study have experienced a stroke.**
- **1069 patients have hypertension.**
- **87 patients have diabetes.**

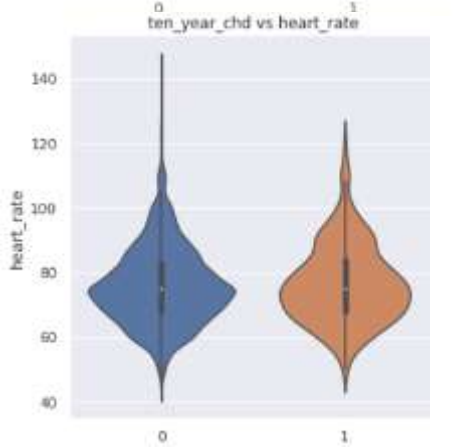
relationship between the dependent variable and the continuous variables

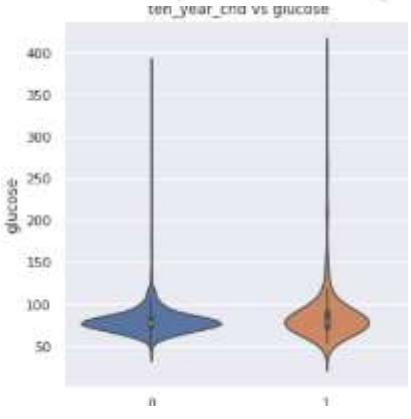






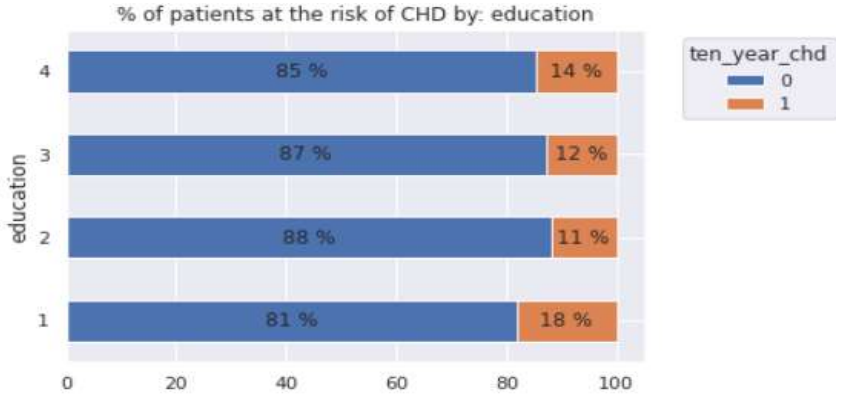


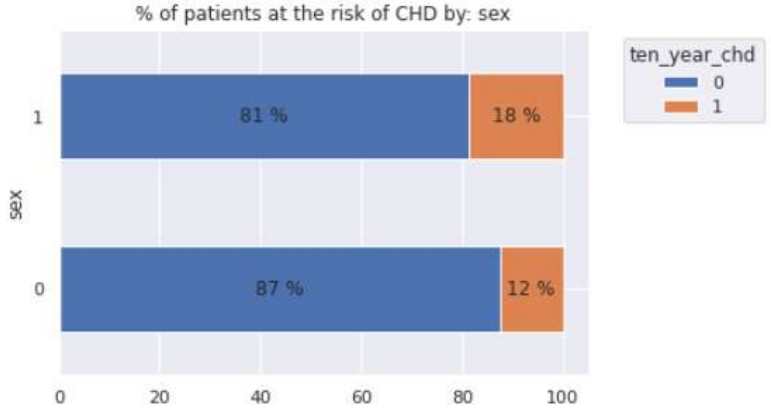




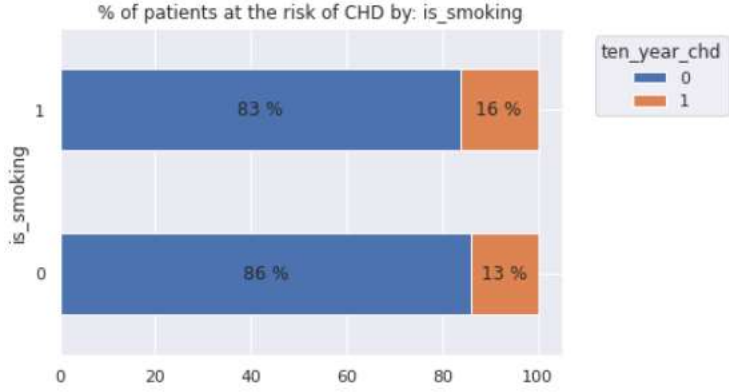
- The risk of CHD is higher for older patients than younger patients.
- No conclusive evidence on the risk of CHD with other continuous variables.
- The risk of CHD increases with age, cholesterol, BP, BMI, heart rate, and glucose.

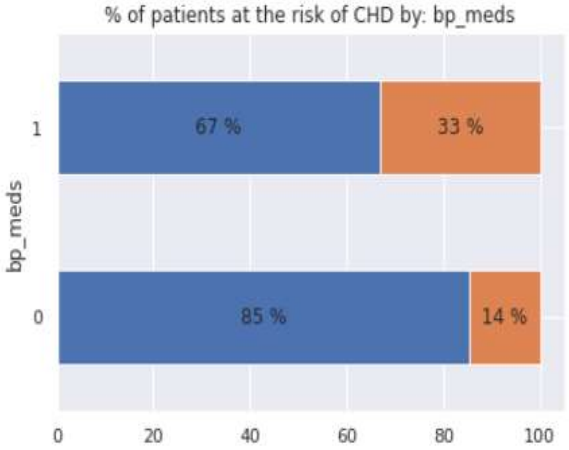
relationship between the dependent variable and the discrete variables

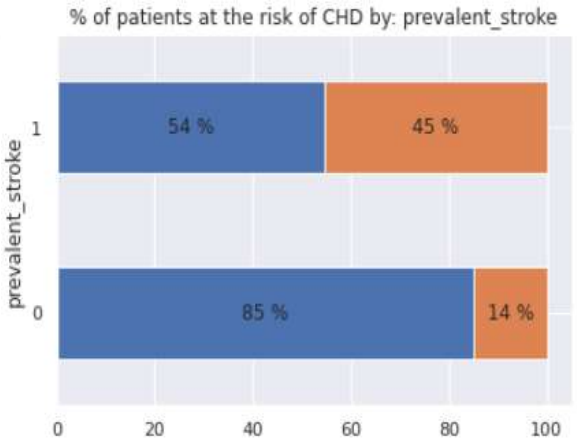


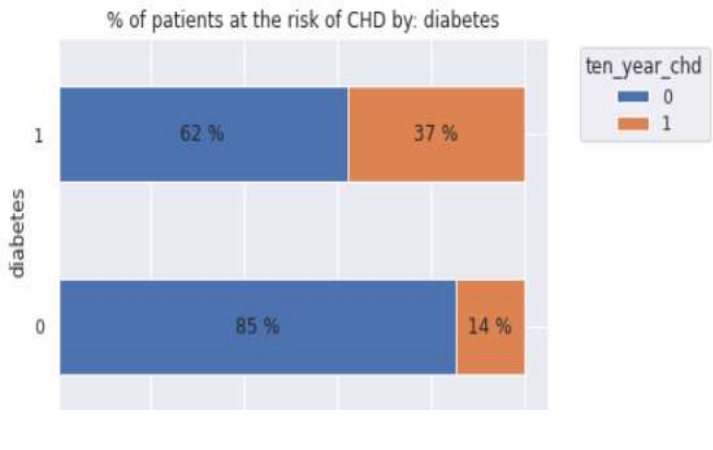


- The risk of CHD varies by **educational level, gender,** and whether or not the patient **smokes.**



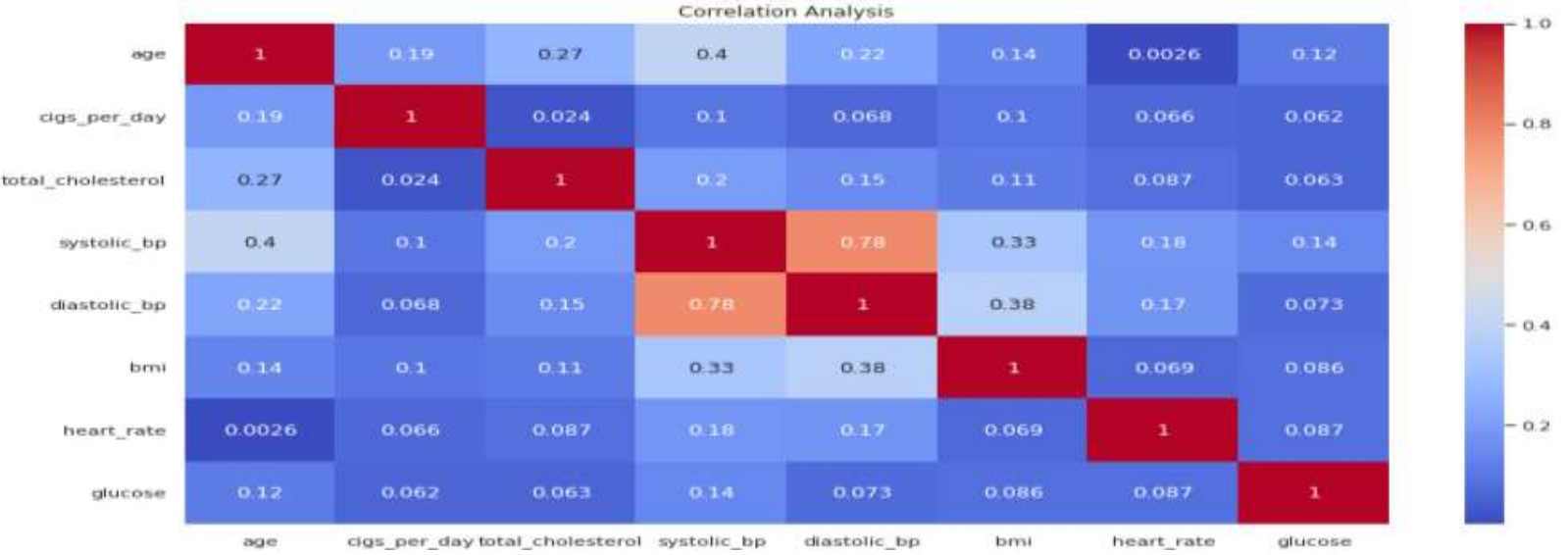






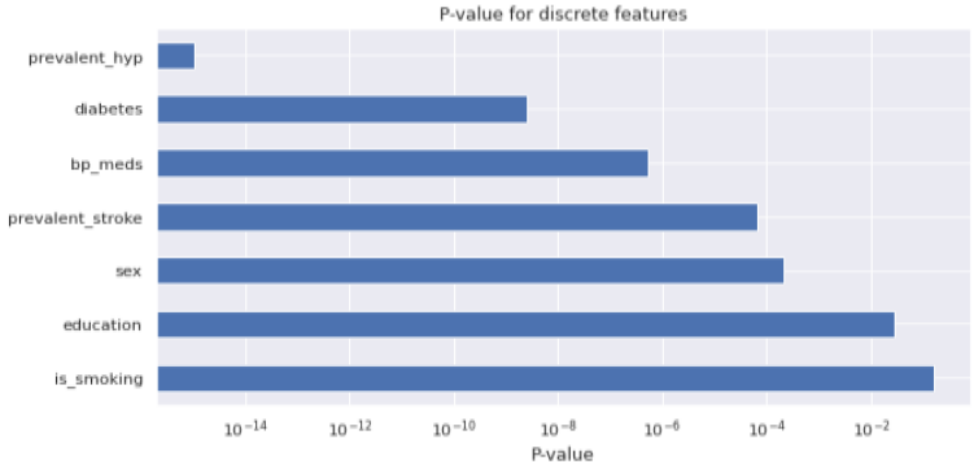
•18%, 11%, 12%, and 14% of the patients belonging to education levels 1, 2, 3, and 4 respectively were eventually diagnosed with CHD.
•Male patients have a significantly higher risk of CHD (18%) than female patients (12%).
•Patients who smoke have a significantly higher risk of CHD (16%) than patients who don't smoke (13%).
•Patients who take BP medicines have a significantly higher risk of CHD (33%) than other patients (14%).
•Patients who had experienced a stroke in their life have a significantly higher risk of CHD (45%) than other patients (14%).
•Hypertensive patients have a significantly higher risk of CHD (23%) than other patients (11%).
•Diabetic patients have a significantly higher risk of CHD (37%) than other patients (14%).

Correlation Matrix



- Above the correlation magnitude heatmap for all the continuous variables in the dataset.
- The variables systolic BP and diastolic BP are highly correlated.
- Cigs_per_day & smoking are highly correlated.
- combined systolic BP and diastolic BP to denote a new feature pulse rate.

Feature Engineering



- Since the prevalent hypertension column (prevalent_hyp) has the smallest p-value, we can say that it is the most important feature (among the categorical independent variables) which determines the outcome of the dependent variable.
- The is_smoking feature has the highest p-value, which indicates that it is the least important feature (among categorical independent variables).
- We can drop this column since we already have a column cigs_per_day, which gives the number of cigarettes smoked by the patient in a day. The patients who don't smoke have entered zero in this column.

Handling Skew & Outliers

```
# skewness along the index axis
(df[continuous_var]).skew(axis = 0)
```

```
age                0.225796
cigs_per_day       1.204077
total_cholesterol  0.948170
bmi                1.025551
heart_rate         0.676660
glucose            6.342892
pulse_pressure     1.412382
dtype: float64
```

```
# Skew for log10 transformation
np.log10(df[continuous_var]+1).skew(axis = 0)
```

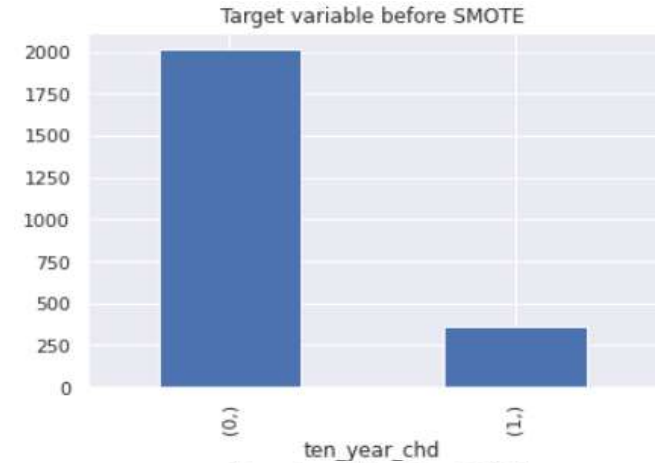
```
age               -0.015053
cigs_per_day       0.275072
total_cholesterol  0.011860
bmi                0.370422
heart_rate         0.165898
glucose            2.309072
pulse_pressure     0.354174
dtype: float64
```

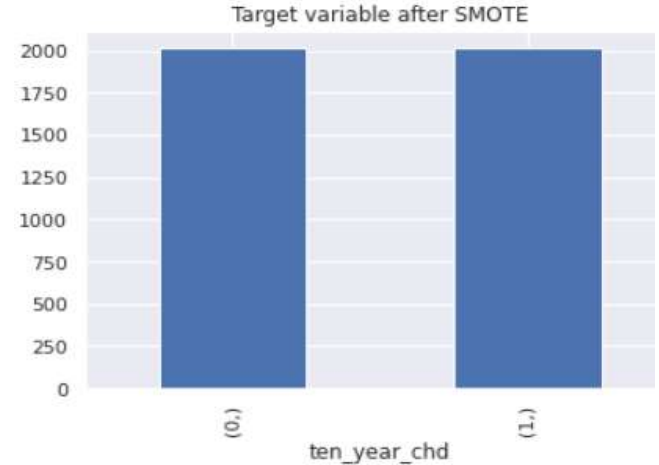
- The skew in numeric variables is reduced by performing log transformation.
- The outliers beyond 3 standard deviations from the mean were imputed with the median value.
- Except for cigs_per_day, we have successfully been able to reduce the skewness in the continuous variables. Now, these distributions are closer to symmetric distribution.

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





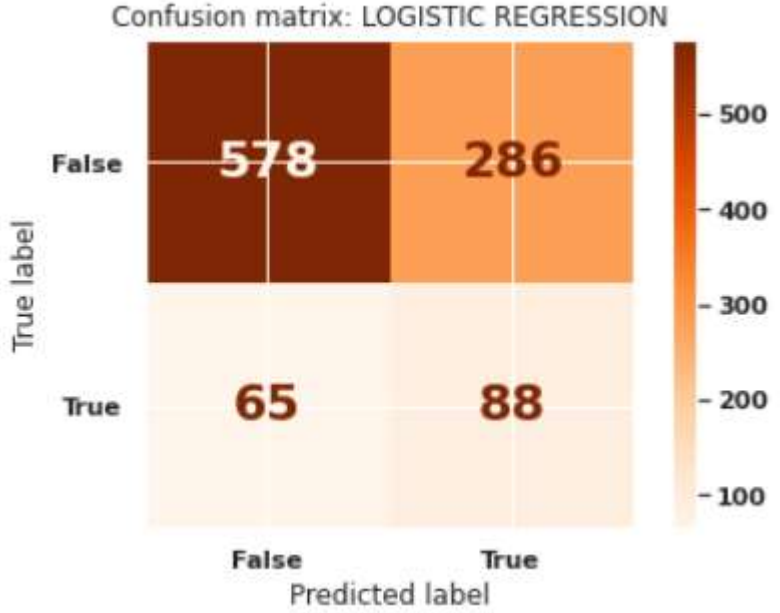
- Data points in the test data set = 30%
- Choice of split: Repeated stratified K fold, k = 4
- Evaluation metric: Recall
- **Recall = *True Positive False / Negative+True Positive***
- Hyperparameter tuning: Grid search
- Oversampling strategy: SMOTE
- Data points before SMOTE = 2373
- Data points after SMOTE = 4030
- Scaler used: Standard Scale

Model Evaluation

- We preprocess the data to train the dataset with different classification models:
- Logistics Regression
- K Nearest Neighbors
- Naive Bayes
- Decision Tree
- Support Vector Machines
- XG Boosts

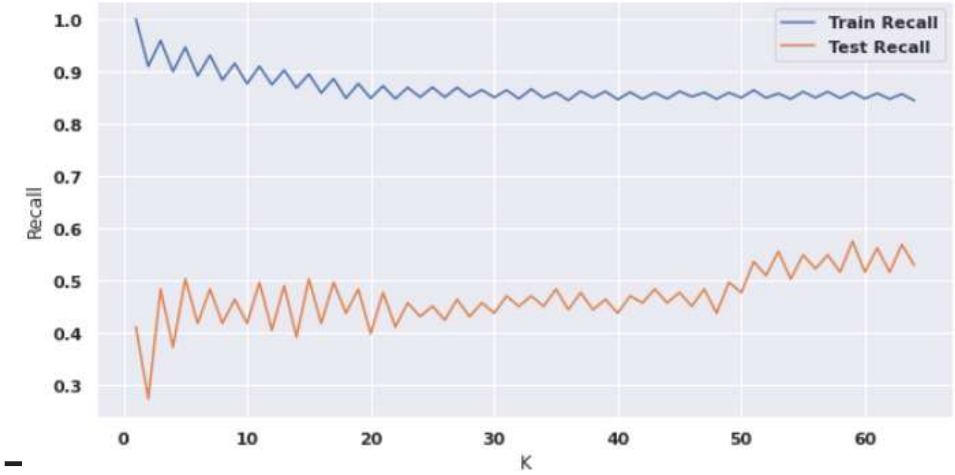
Logistics Regression

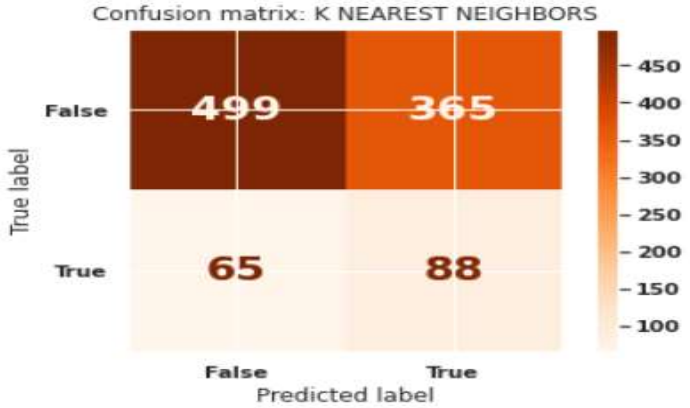
- **Evaluation metrics:**
  - Train Recall = 0.6987
  - Test Recall = 0.6732
  - Test Accuracy = 68%



K Nearest Neighbors

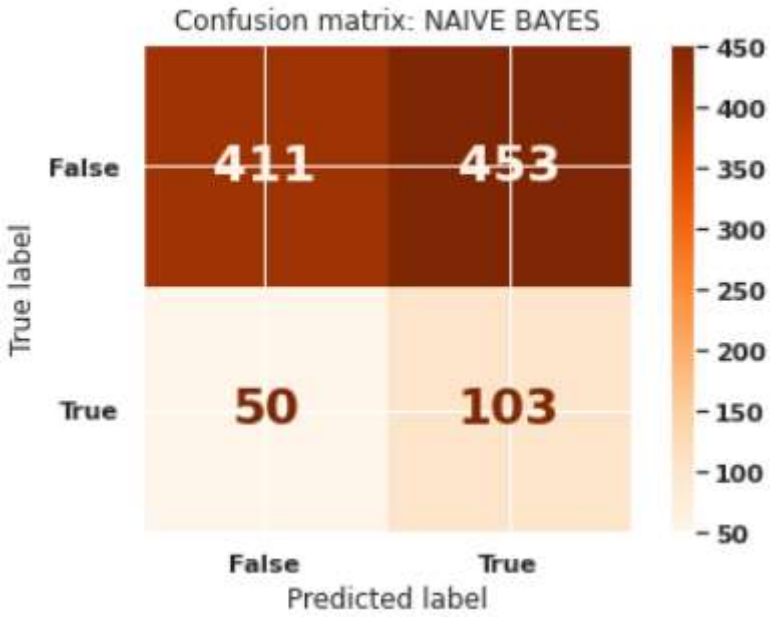
- **Parameters:**
  - K = 39
- **Evaluation metrics:**
  - Train Recall = 0.8317
  - Test Recall = 0.7124
  - Test Accuracy = 61%





Naive Bayes

- **Parameters:**
- ● var_smoothing= 1.0
- **Evaluation metrics:**
- ● Train Recall = 0.5811
- ● Test Recall = 0.5294
- ● Test Accuracy = 72%

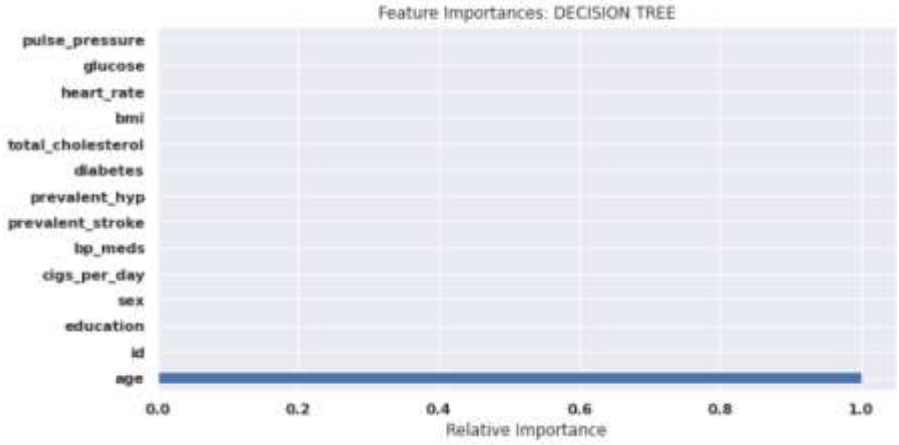


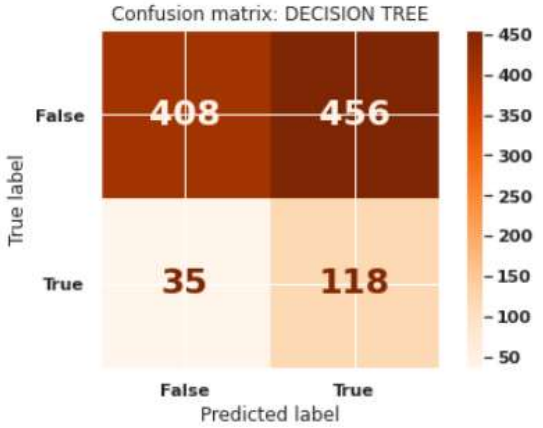
Decision Tree

- **Parameters:**
- ● max_depth = 1
- ● min_samples_leaf = 0.1
- ● min_samples_split = 0.1

Evaluation metrics:

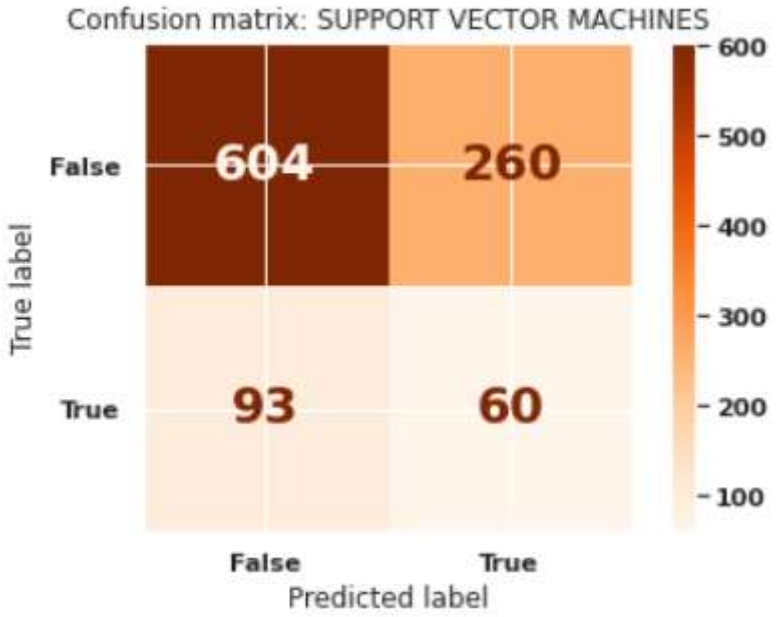
- ● Train Recall = 0.8595
- ● Test Recall = 0.7712
- ● Test Accuracy = 52%





Support Vector Machines

- **Parameters:**
- C = 1
- Gamma = 0.01
- Kernel = RBF
- **Evaluation metrics:**
- Train Recall = 0.7652
- Test Recall = 0.6666
- Test Accuracy = 65%

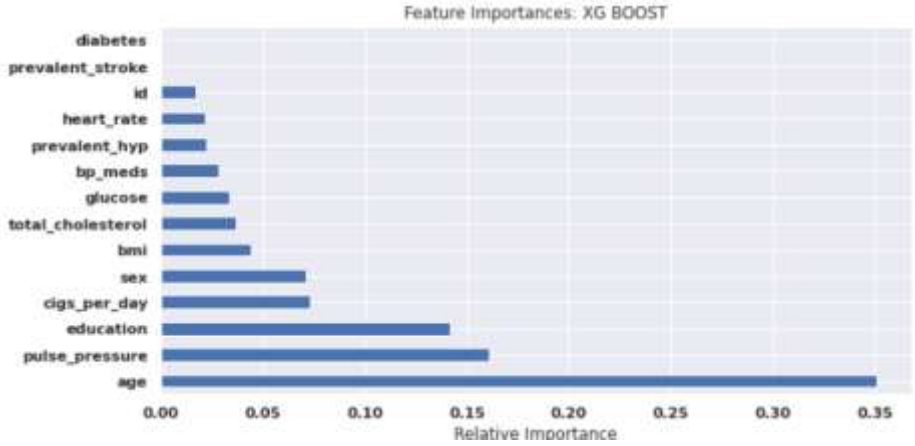


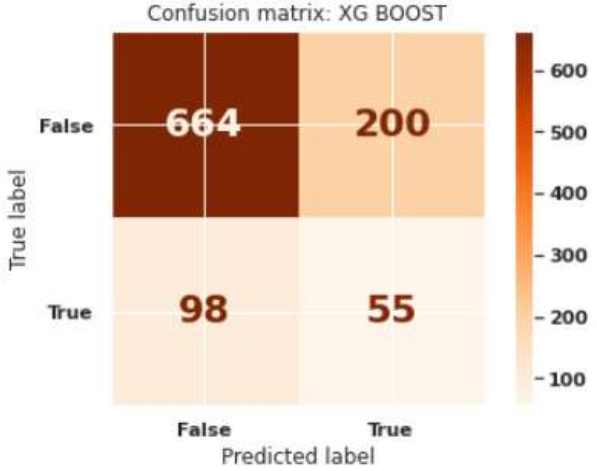
XG Boosts

- **Parameters:**
  - max_depth = 1
  - min_samples_leaf = 0.1
  - min_samples_split = 0.1
- n_estimators = 500

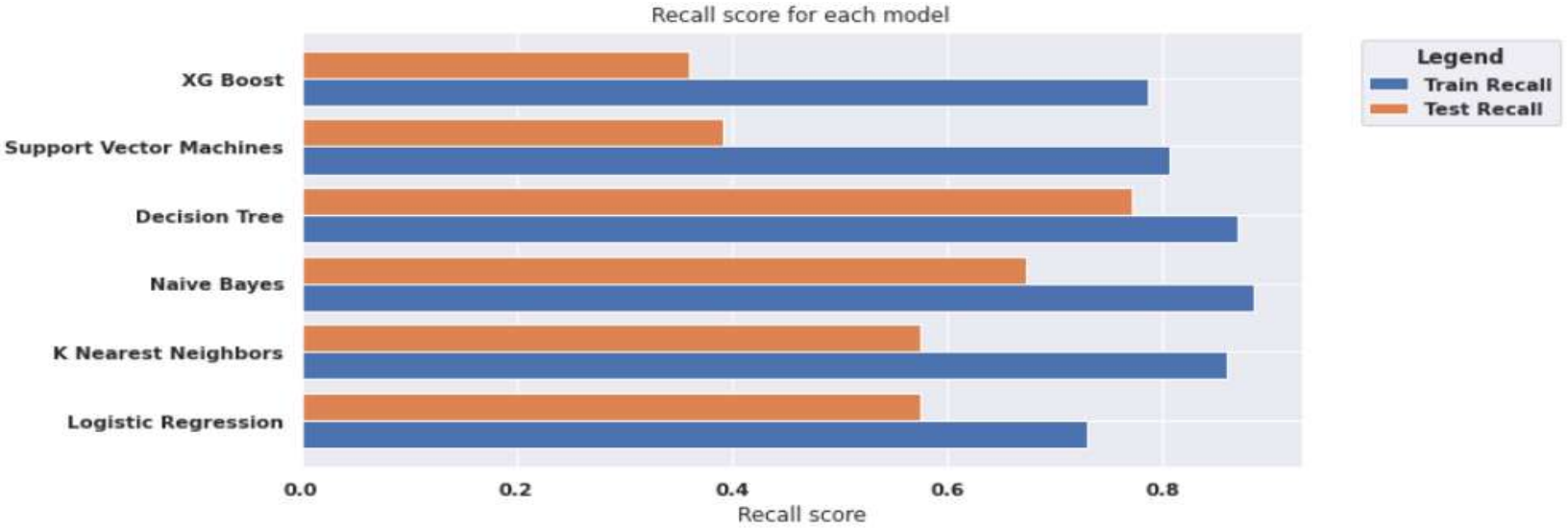
Evaluation metrics:

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





Model Comparison



- The **decision tree** has the highest train and test recall score compared to other models built.

Challenges

- 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 the 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