Capstone Project 3

Cardiovascular Risk Prediction

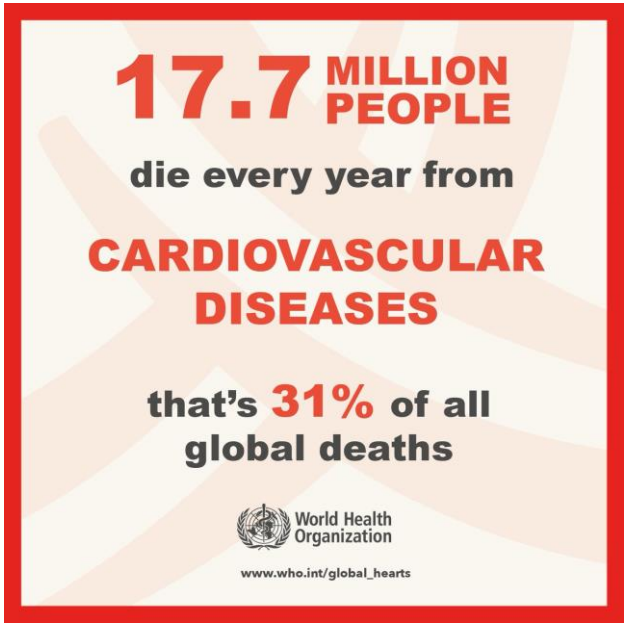
By:- Viral Bhatu Shewale

Content:

- Problem Statement
- Data Summary
- Handling Missing Values
- Exploratory Data Analysis (EDA)
- Multicollinearity
- Data Processing
- Modelling Approach
- Predictive Model
- Model Comparison
- Conclusion

Abstract:

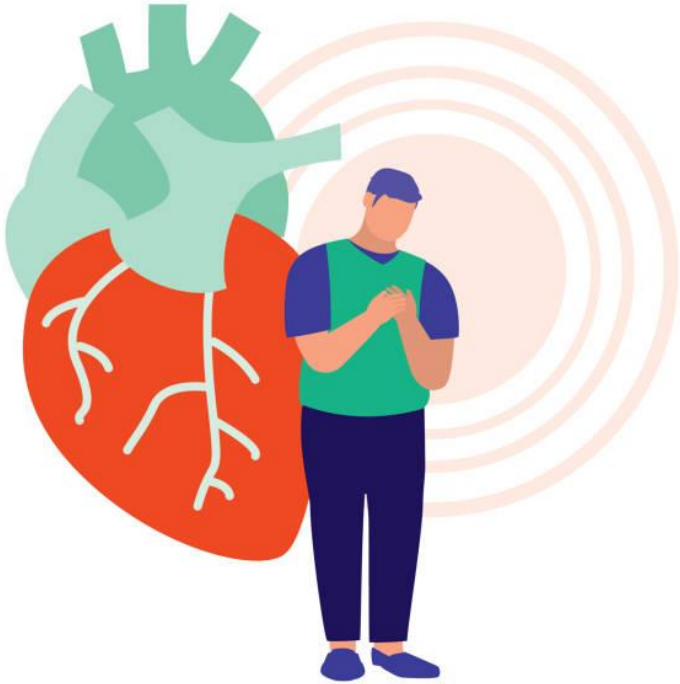
- **Cardiovascular disease** (**CVD**) is a group of disorders of the heart and blood vessels and includes coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions.
- Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated **17.9 million lives each year**.
- 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 the 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.



Data Summary:

- There are **3390 rows** and **16** different **attributes** (columns) in the dataset.
- **Dependent Variable: TenYearCHD**

- **Demographic Data:**
  - Sex
  - Age
  - Education

- **Medical History:**
  - BP medication
  - Prevalent hypertension
  - Prevalent stroke
  - Diabetes

- **Current Medical status:**
  - Total cholesterol
  - BMI
  - Heart Rate
  - Glucose
  - Systolic BP
  - Diastolic BP

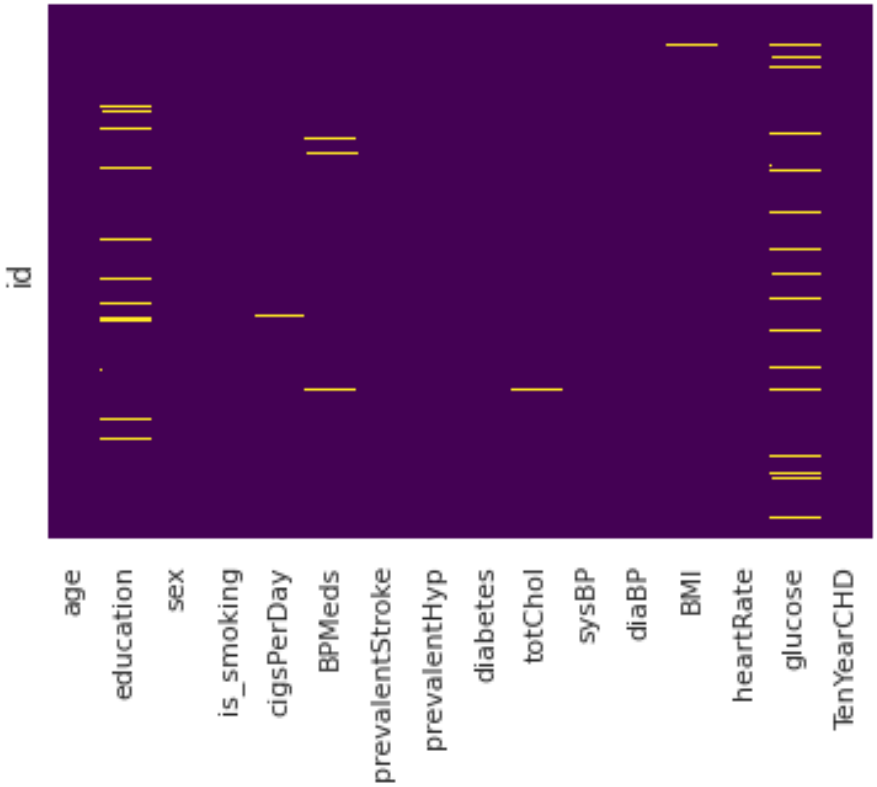
- **Behavioural:**
  - Is smoking
  - Cigarettes per day

Handling Missing Values:

- In the data, there were **some columns** which have **missing values** (Education, cigsPerDay, BPMeds, Totchol, BMI, heartrate, glucose).

- Missing Values imputed by the aggregate (**mean or median**).

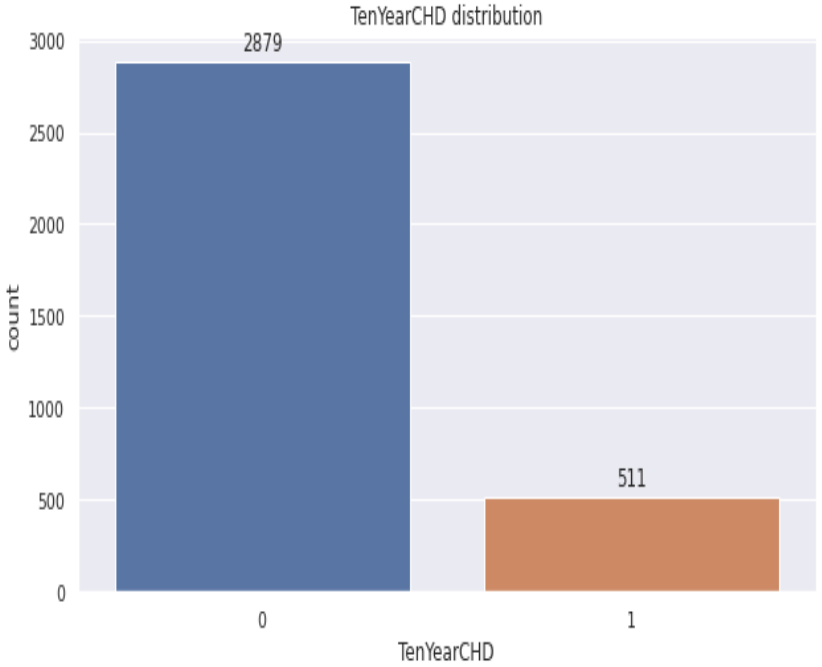
- **Mean imputation** Education(87), totChol(38)
- **Median imputation** CigPerDay(22), BMI(14), Glucose(304)
- **Mode imputation** BPMeds(44)
- heartRate column has only one missing value hence I dropped that missing value column.



Exploratory Data Analysis (EDA):

- The **dependent variable** - 10 year risk of CHD is **Imbalanced**.
- Only 15% of the patients in the study were eventually exposed to the risk of this heart disease, rest of the patients were not exposed to this disease after the end of 10 year study.

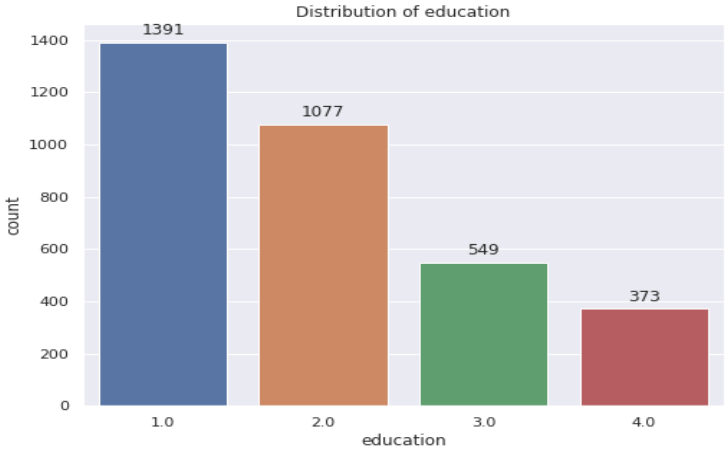
- All the continuous independent variables are **positively skewed** except age, which is almost normally distributed.

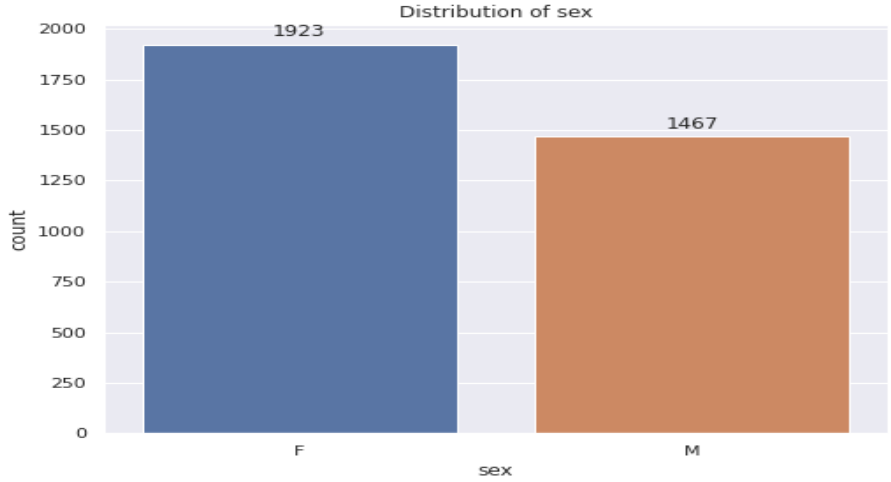


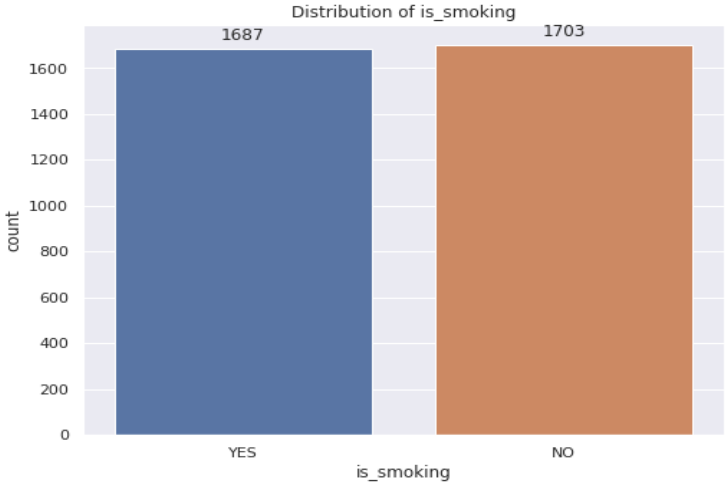
EDA:

Distribution of the categorical variables:

- There are **more female** patients compared to male patients.
- Majority of patent belong to the education **level 1**.
- Half the patient are smokers.



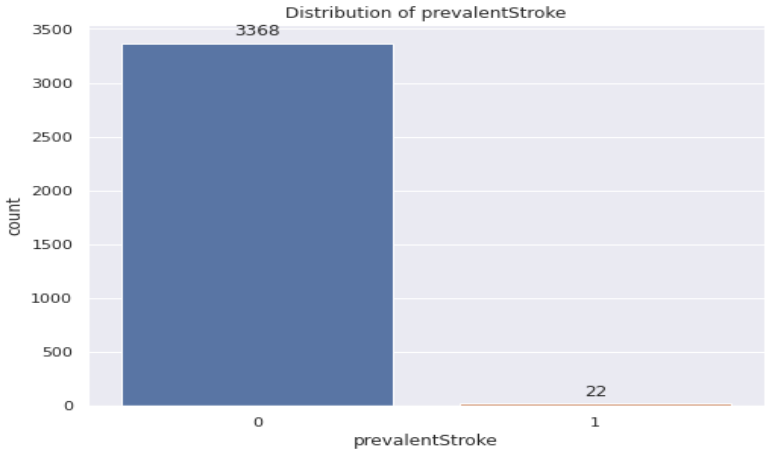


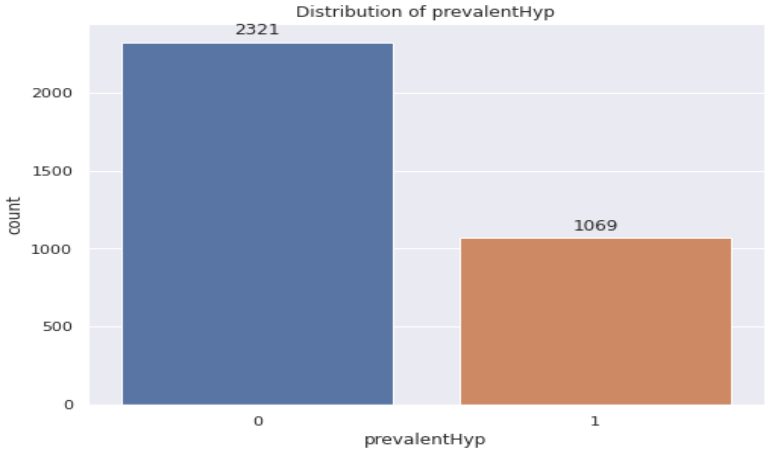


EDA:

Distribution of the categorical variables:

- There are relatively few individuals who have experienced a **stroke**
- Of the total participants, 31.53% had prevalent **hypertension**.

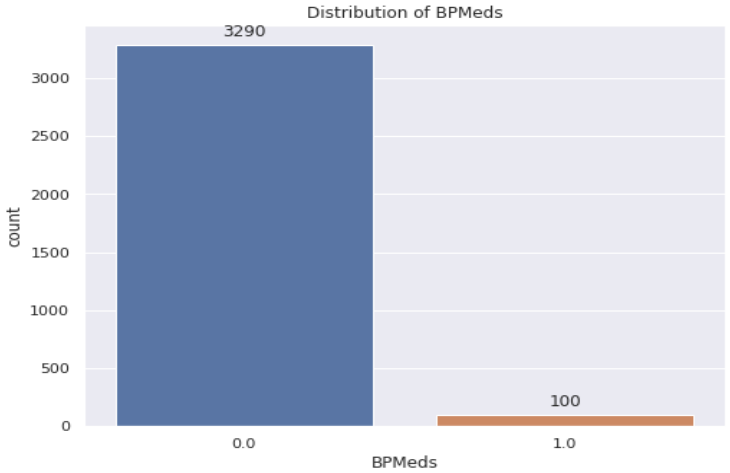


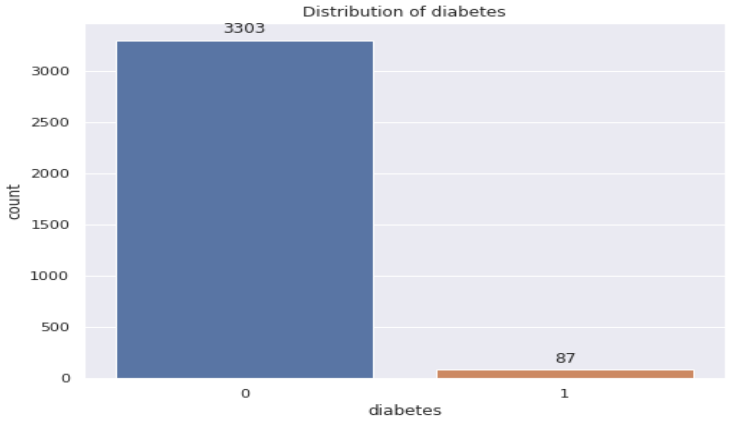


EDA:

Distribution of the categorical variables:

- Of the total participants, 100 participants were under **blood pressure medication**.
- There are relatively few participants who have **diabetes**.

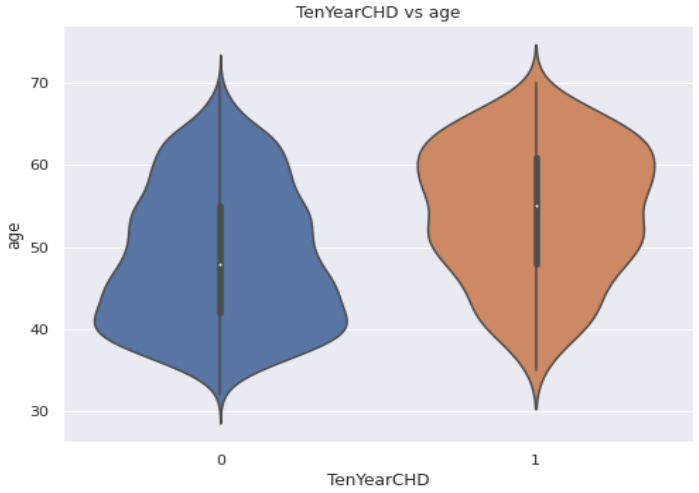


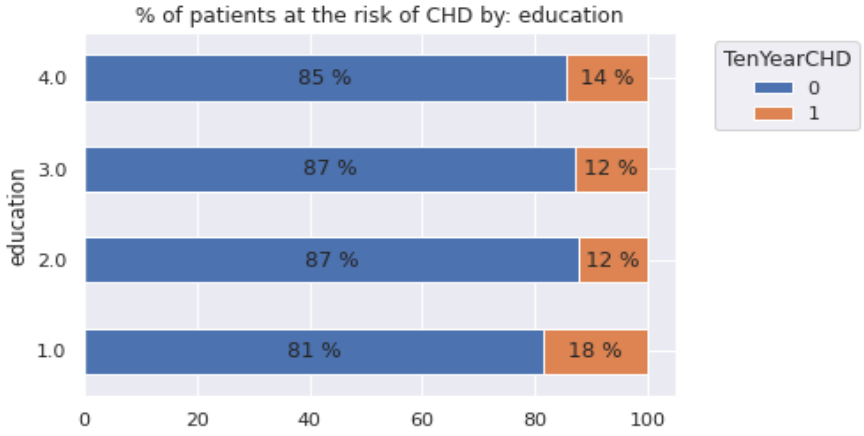


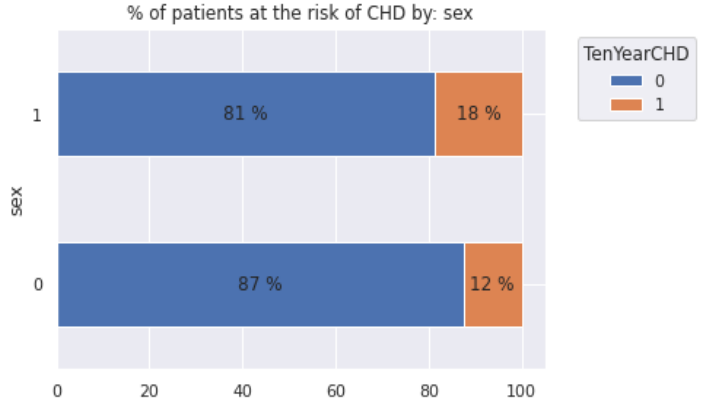
EDA:

The relation between dependent variable and independent variables:

- The risk of CHD increases with the **age**.
- The risk of CHD varies by **education** level and **gender**.



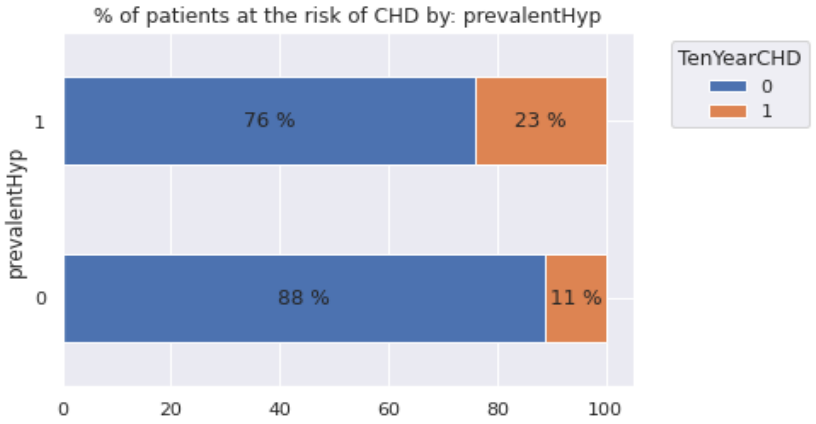


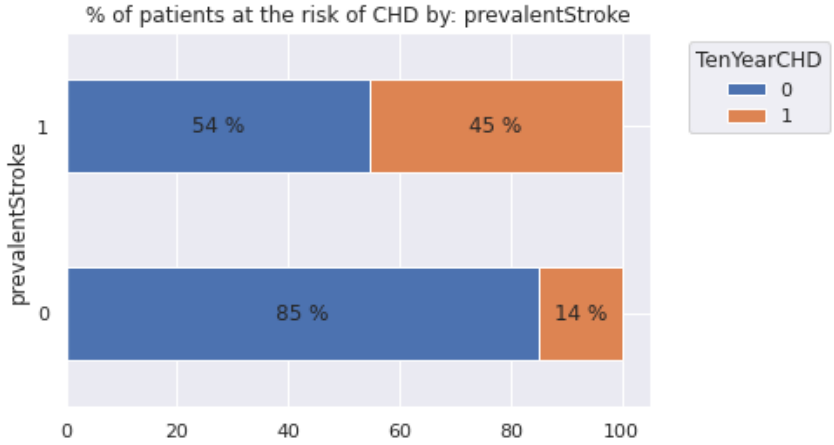


EDA:

The relation between dependent variable and independent variables:

- Patients who have had a **stroke** or **hypertension** are more likely to test **positive** for **CHD**.

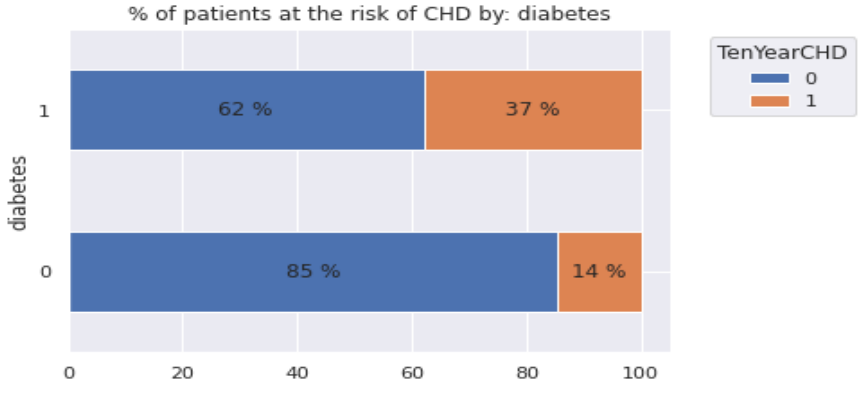


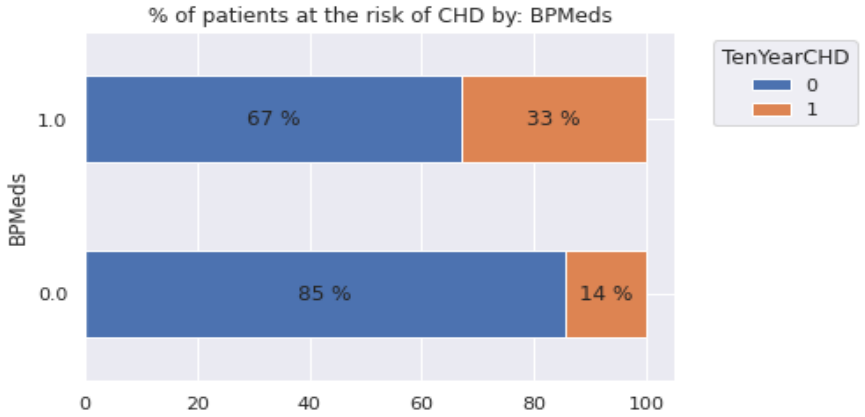


EDA:

The relation between dependent variable and independent variables:

- Patients with **diabetes** or are presently on **blood pressure medication** are more likely diagnosed with **CHD**.





Multicollinearity:

- **Variance Inflation Factor(VIF)** is a measure of collinearity among predictor variables.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- systolic pressure and diastolic pressure have higher VIF.
- Added new feature **Pulse Pressure.**
- **Pulse pressure** = (systolic pressure - diastolic pressure)

	variables	VIF
0	age	41.230057
1	education	4.762567
2	sex	2.129752
3	is_smoking	5.025583
4	cigsPerDay	4.321178
5	BPMeds	1.133739
6	prevalentStroke	1.027151
7	prevalentHyp	2.413136
8	diabetes	1.583230
9	totChol	30.810583
10	sysBP	135.108477
11	diaBP	131.565537
12	BMI	44.643386
13	heartRate	38.764361

Data Processing:

- **Transform data to reduce skewness:-** The values of a certain independent variable (feature) are **skewed**, depending on the model, skewness may violate model assumptions or may impair the interpretation of feature importance. Hence **log transformations** used to transform skewed data.

- **Handling outliers:-** **(Mean/Median imputation)** As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

- **Handling Unbalanced data:-** Since we are dealing with unbalanced data, SMOTE (Synthetic Minority Oversampling Technique) is used to oversample train data.

- **Scaling the data:-** StandardScaler removes the mean and scales the data to the unit variance.

Model Selection Approach:

- We are working on a binary classification problem.
- **Standard binary classification** models like **Naive Bayes, decision tree classifiers, ensemble of decision tree** and **support vector machines**.

- **Choice of split:-** k fold cross validation k=5.

- **Hyperparameter tunning:-** GridsearchCV

- **Evaluation Metrics:-** since we are dealing with data related to healthcare, **False Negatives are big concern** than False Positives. The **recall score** would be the best **evaluation metric**.

  Recall = TruePositives / (TruePositives + FalseNegatives)

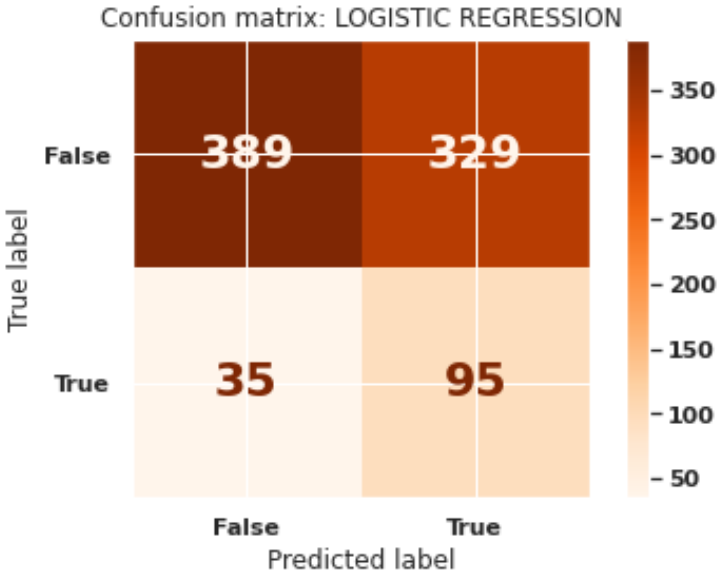
Logistic regression:

- **Evaluation Metrics:**
- **Recall:**

  Test Recall = **0.7307**

- **Confusion Metrics:**

  There are **35 False Negative**.



Decision Tree Classifier:

➢ **Parameters:**

- max_depth = 4
- min_samples_leaf = 0.1
- min_samples_split = 0.1

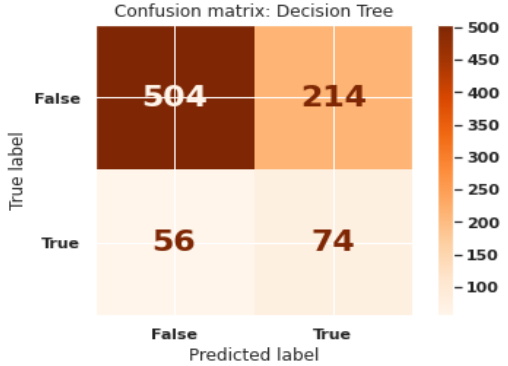
➢ **Evaluation Metrics:**

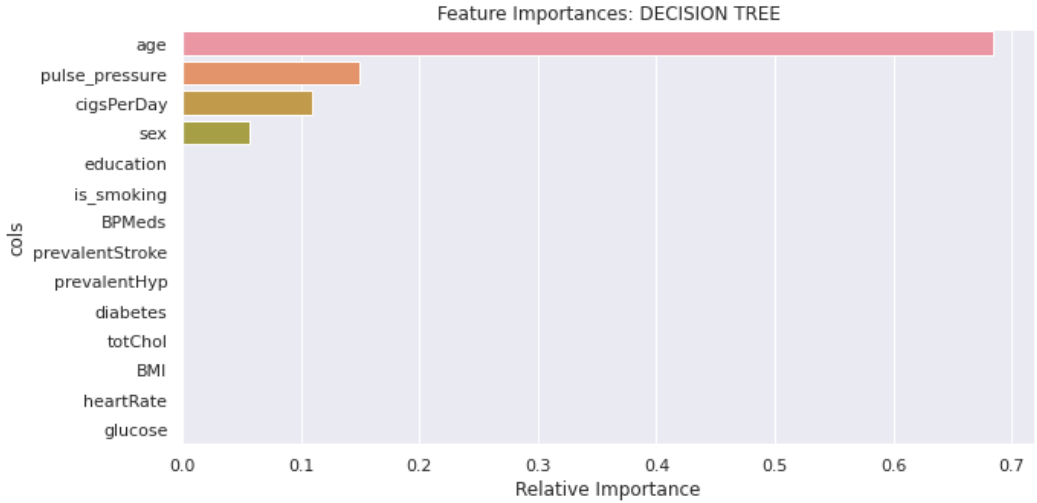
- Recall:

  Test Recall = **0.5692**

- Confusion Metrics:

  There are **56 False Negative**.





Naive Bayes Classifier:

➢ **Parameters:**

- var_smoothing = 1.0

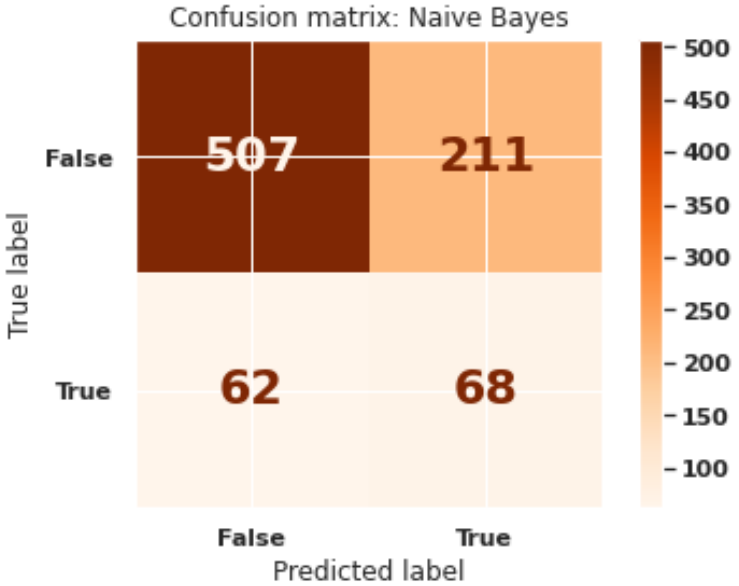
➢ **Evaluation Metrics:**

- Recall:

  Test Recall = **0.5230**

- Confusion Metrics:

  There are **62 False Negative**



Support Vector Machine (SVM):

➢ **Parameters:**

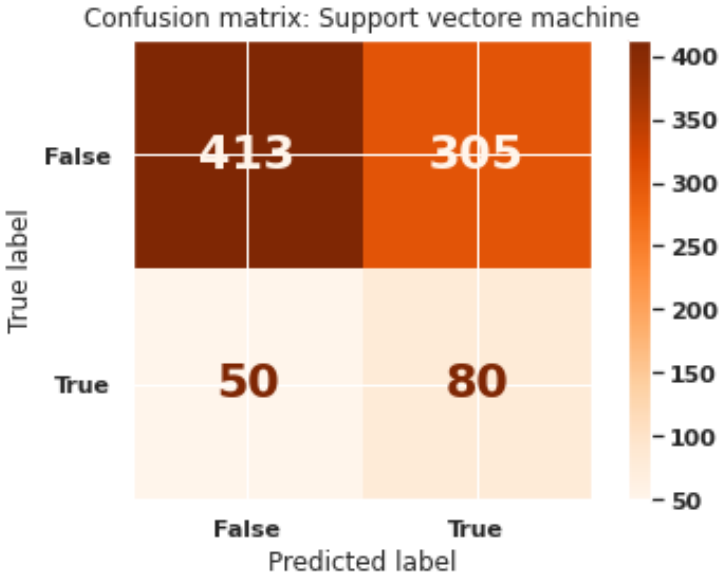
- kernel = rbf
- gamma = 0.01
- C = 10

➢ **Evaluation Metrics:**

- Recall:

  Test Recall = **0.6153**
- Confusion Metrics:

  There are **50 False Negative**.



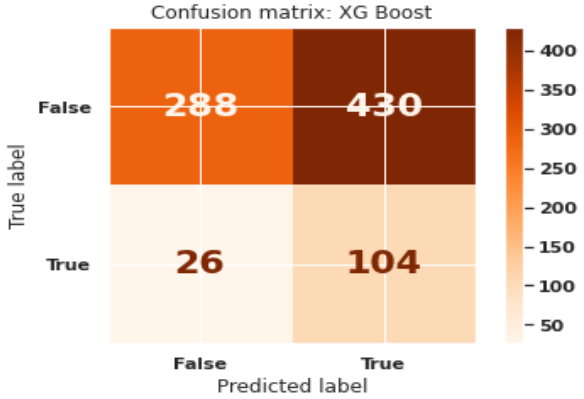
XG Boost:

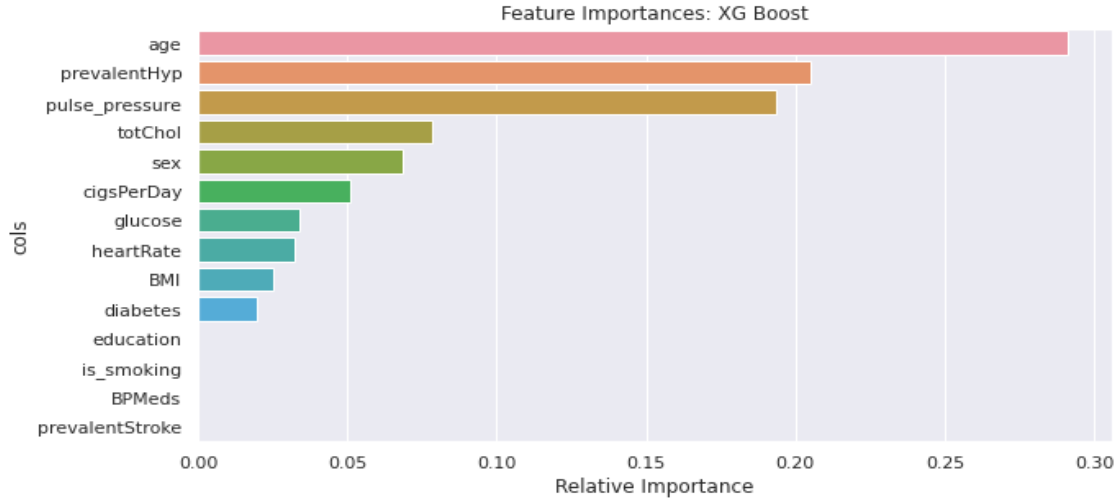
- **Parameters:**
- max_depth = 1
- min_samples_leaf = 0.1
- min_samples_split = 0.1
- n_estimators = 500

- **Evaluation Metrics:**
- Recall:

  Test Recall = **0.80**
- Confusion Metrics:

  There are **26 False Negative**.





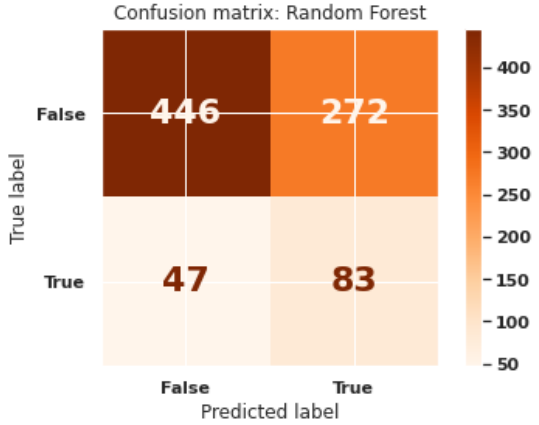
Random Forest:

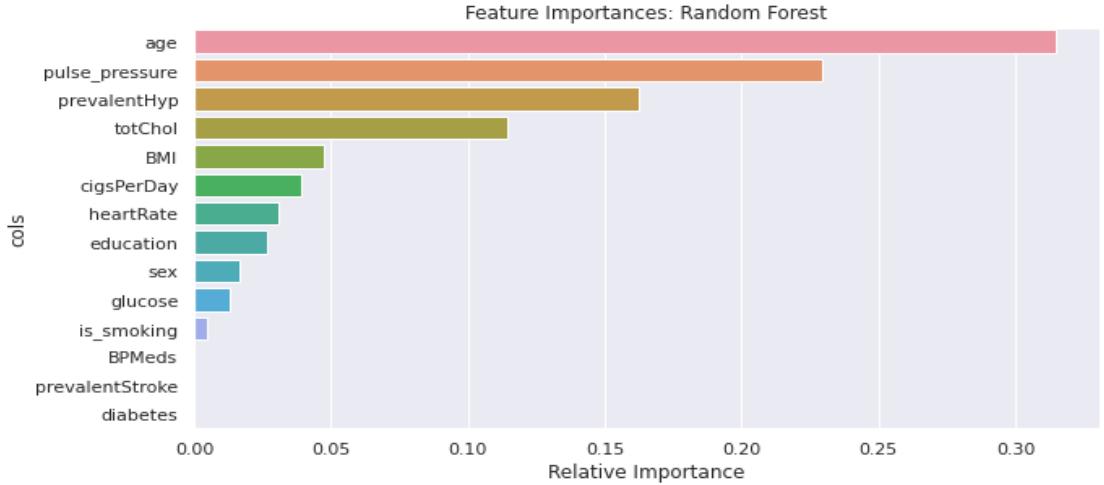
- **Parameters:**
- max_depth = 2
- min_samples_leaf = 0.1
- min_samples_split = 0.1
- n_estimators = 500

- **Evaluation Metrics:**
- Recall:

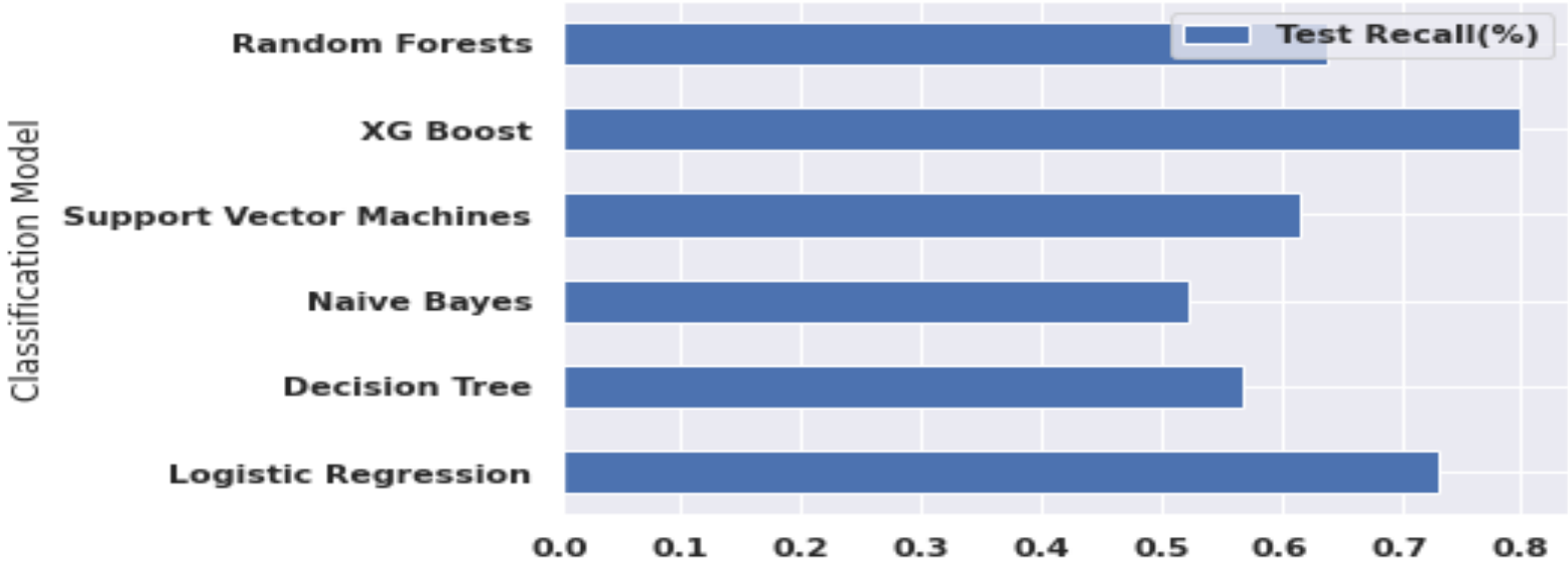
  Test Recall = **0.6384**
- Confusion Metrics:

  There are **47 False Negative**





Model Comparison:



- The **XG Boost** model has the highest test recall compare to other models.

Conclusion:

- Predictive models have been successfully built, that can predict a patient's risk for CHD based on their demography, lifestyle, and medical history.
- The Built models were evaluated using Recall, and the XG Boost (0.80) has the highest test recall compared to other models.
- A recall score of 0.80 indicates that out of 100 individuals with the illness, our model will be able to classify only 80 as high-risk patients, while the remaining 20 will be misclassified.
- From the analysis, I found that the age of a person was the most important feature in determining the risk of a patient getting infected with CHD, followed by pulse pressure, prevalent hypertension and total cholesterol.
- Diabetes, prevalent stroke and BP medication were the least important features in determining the risk of CHD.