Cardiovascular Health Assessment and Risk Prediction Model Project

Prepared by: Miral Naik

Index

- Problem Statement
- Dataset Overview
- Exploratory Data Analysis & Visualization
- Data Preprocessing & Feature Engineering
- Model Building
- Results & Insights

Problem Statement

- Visiting hospitals for regular check-ups, it is almost always seen that they encourage people to get special check-ups to identify if they are at risk of heart disease. Heart diseases have unfortunately become very common. It may be due to various reasons such as lifestyle, work pressure, lack of exercise, etc. In this project, we will be working on predicting the 10-year risk of Coronary Heart Disease (CHD). We are given a set of variables that impact heart diseases. These variables are related to demographic, past, and current medical history.
- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It include over 3390 records and 16 attributes Variables. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.
- This problem statement outlines the task of predicting the 10-year risk of Coronary Heart Disease (CHD) using a dataset from an ongoing cardiovascular study conducted in Framingham, Massachusetts. The dataset comprises demographic, behavioural, and medical risk factors for over 3390 individuals, with a total of 16 attribute variables. The goal is to develop a predictive model that can accurately classify whether a patient is at risk of future CHD based on these factors. Given the prevalence of heart diseases and their impact on public health, this project holds significant importance in the realm of preventive healthcare.

Dataset Overview

- There are 3390 records and 16 features in the dataset.
- The features include age, education, sex, is_smoking, cigsPerDay, BPMeds, prevalent stroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose, TenYearCHD
- The data types of the columns are integers (int64, float64) and objects.
- Integers are used for numerical information such as age, education, cigsPerDay, BPMeds, prevalent stroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose, TenYearCHD
- Objects are used for categorical information such as sex, and is_smoking.
- There is a column named 'id' that is used as a unique identifier for each record. it is not necessary for our prediction model. So, dropped it.

Overview of the features

➢ **Demographic:**

- Sex: Male or female (M/F)
- Age: Age of the patient (Continuous)
- Education:
  - 1: Higher Secondary
  - 2: Graduate
  - 3: Post Graduate
  - 4: Doctorate or PhD

➢ **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)

➢ **Behavioral:**

- is_smoking: Current smoker (Yes/No)
- Cigs Per Day: Average number of cigarettes smoked per day (Continuous)

➢ **Medical (current):**

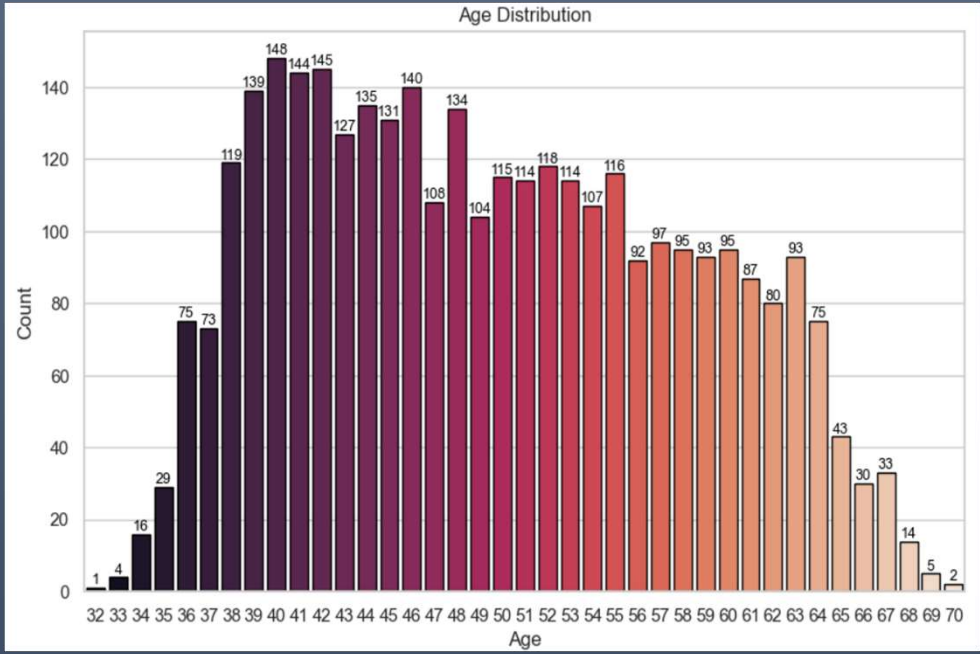
- Tot Chol: Total cholesterol level (Continuous)
- Sys BP: Systolic blood pressure (Continuous)
- Dia BP: Diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- Heart Rate: Heart rate (Continuous)
- Glucose: Glucose level (Continuous)

➢ **Predict variable (Target Variable):**

- TenYearCHD: 10-year risk of coronary heart disease CHD (Binary: 1 for "Yes", 0 for "No")

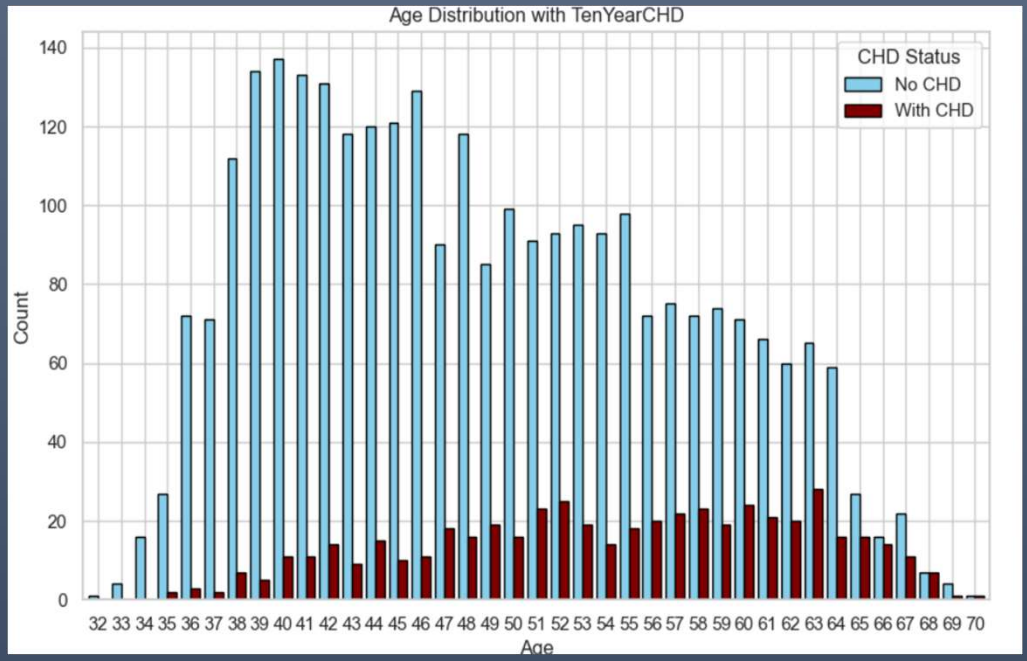
Exploratory Data Analysis & Visualization

Age



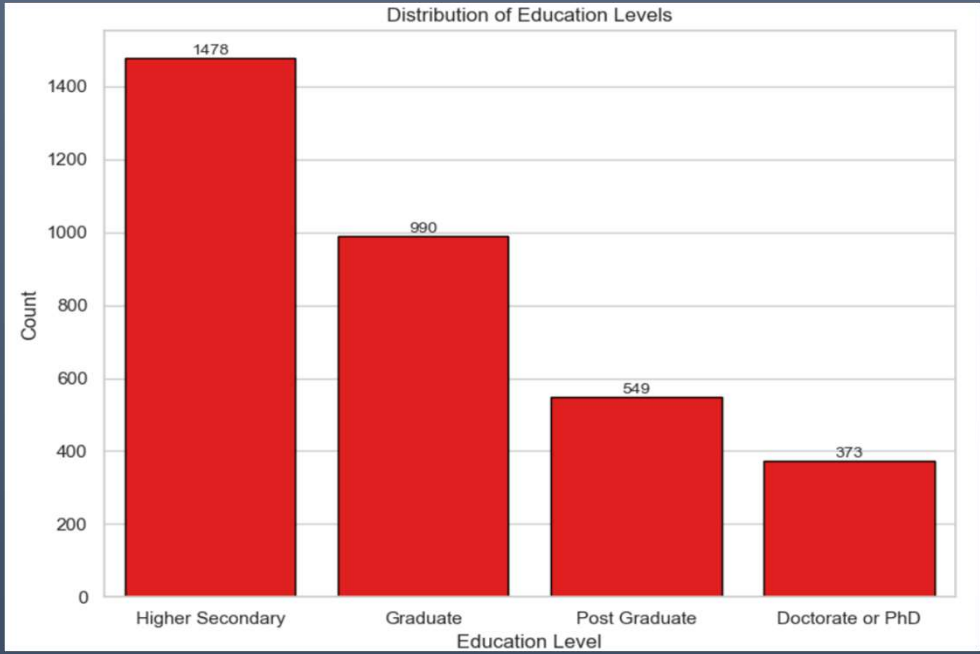
- The dataset shows a right-skewed distribution, with peaks in the age groups of 39-42 and 44-46, indicating more individuals in younger to middle-aged categories compared to older ones.

Age with TenYearCHD



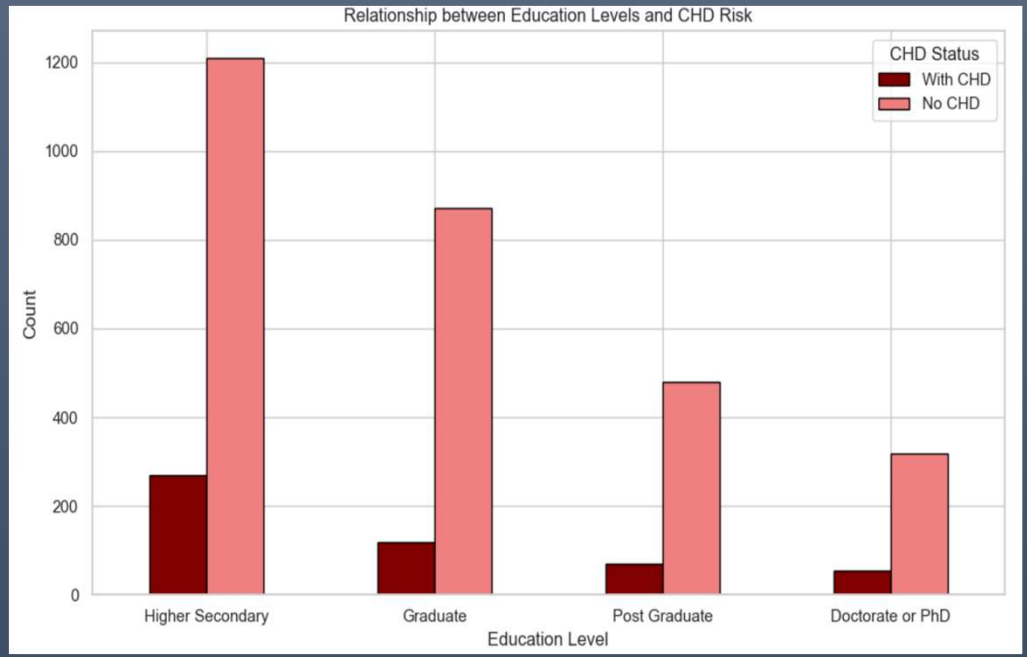
- CHD risk generally rises with age, peaking in middle age (around 38-46) and declining in older age groups (65+). This trend begins around age 35 and continues until age 70.

Education



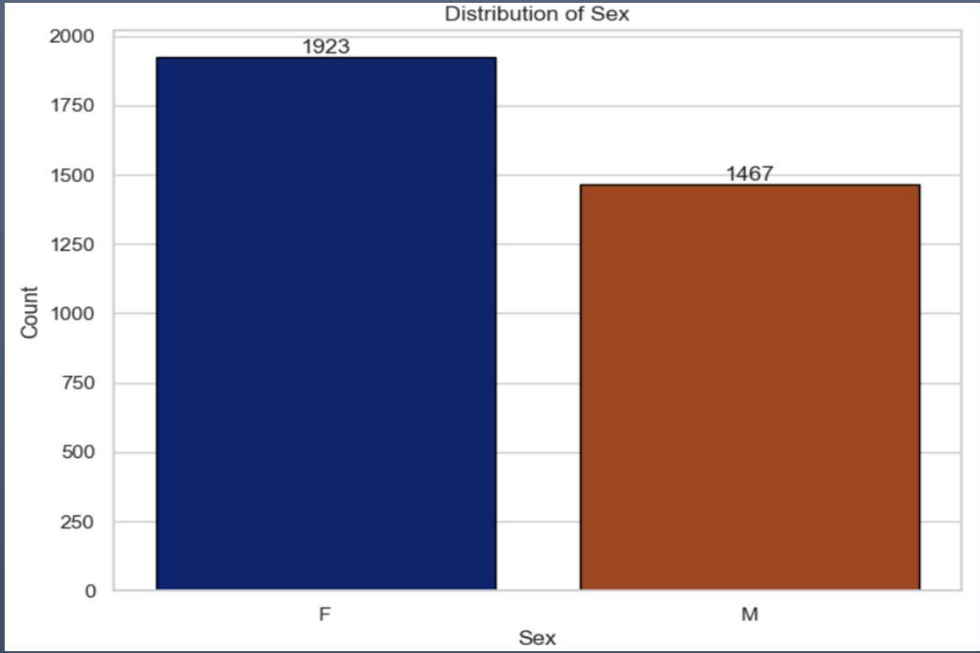
- The majority of individuals in the dataset have attained a "Higher Secondary" education, followed by "Graduate," "Post Graduate," and "Doctorate or PhD".

Education with TenYearCHD



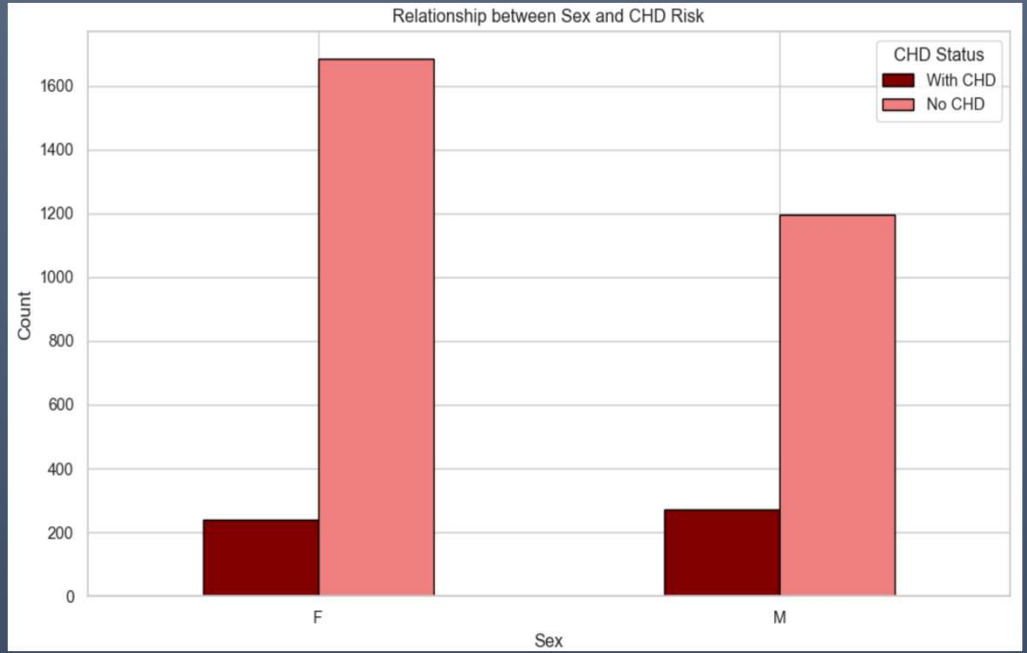
- People who finished "Higher Secondary" or "Graduate" studies, have more cases of CHD compared to those with higher education, such as "Post Graduate" or "Doctorate or PhD."

Sex



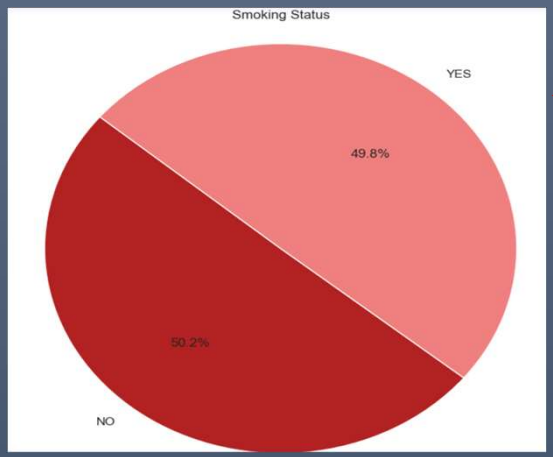
- The dataset has 1923 females and 1467 males, indicating significantly more females compared to males.

Sex with TenYearCHD

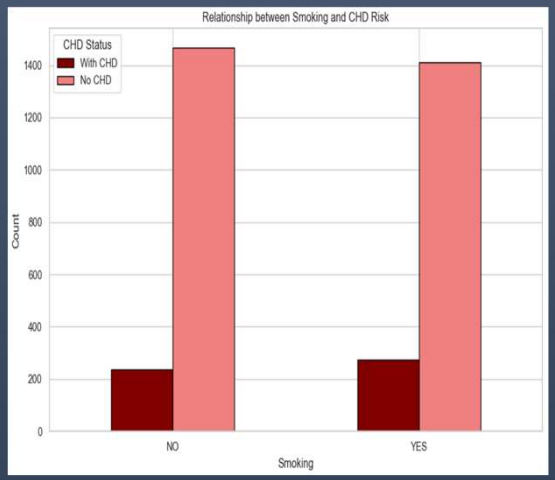


- Among females, 239 have CHD compared to 272 among males, suggesting a slightly higher risk of CHD in males compared to females.

Is_smoking & with TenYearCHD

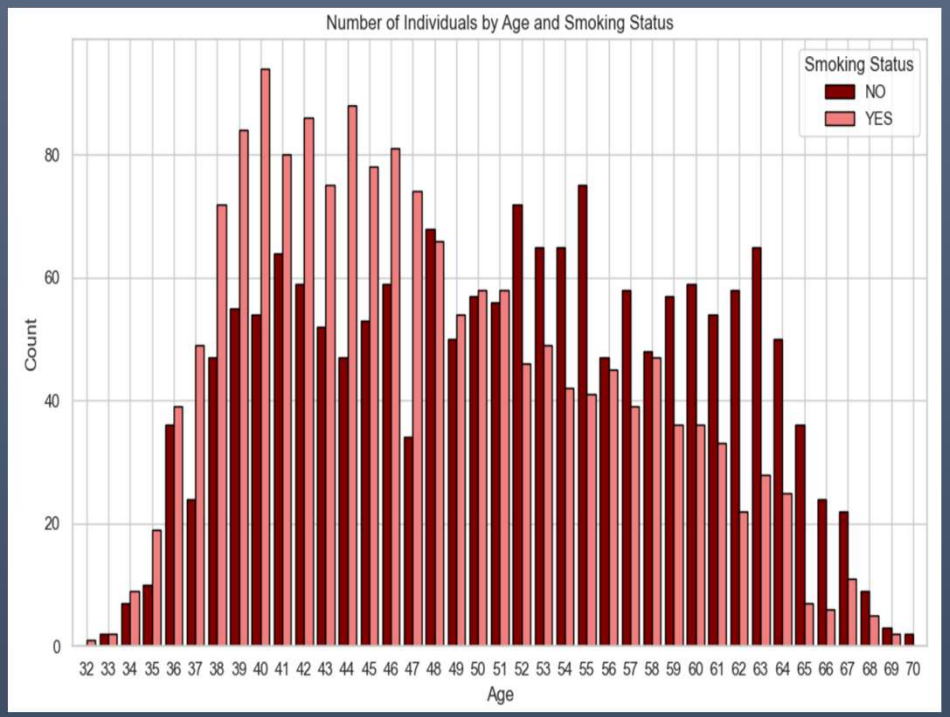


- Approximately 50.2% of individuals do not smoke (1703), while around 49.8% are smokers (1687).



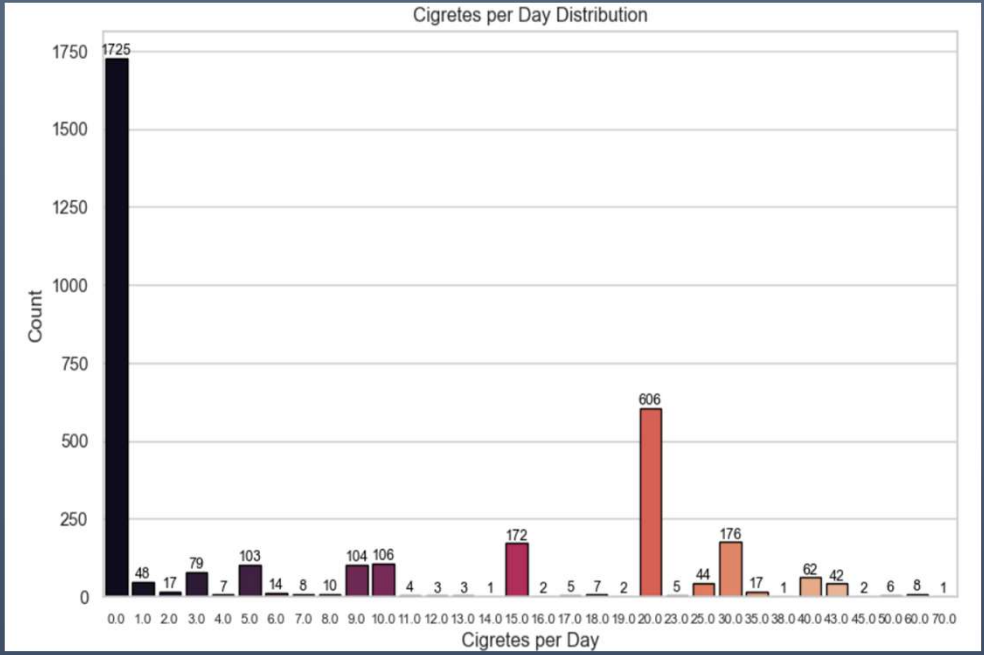
- There are 236 cases of CHD among non-smokers and 275 among smokers, indicating slightly more cases of CHD among smokers.

Is_smoking with Age



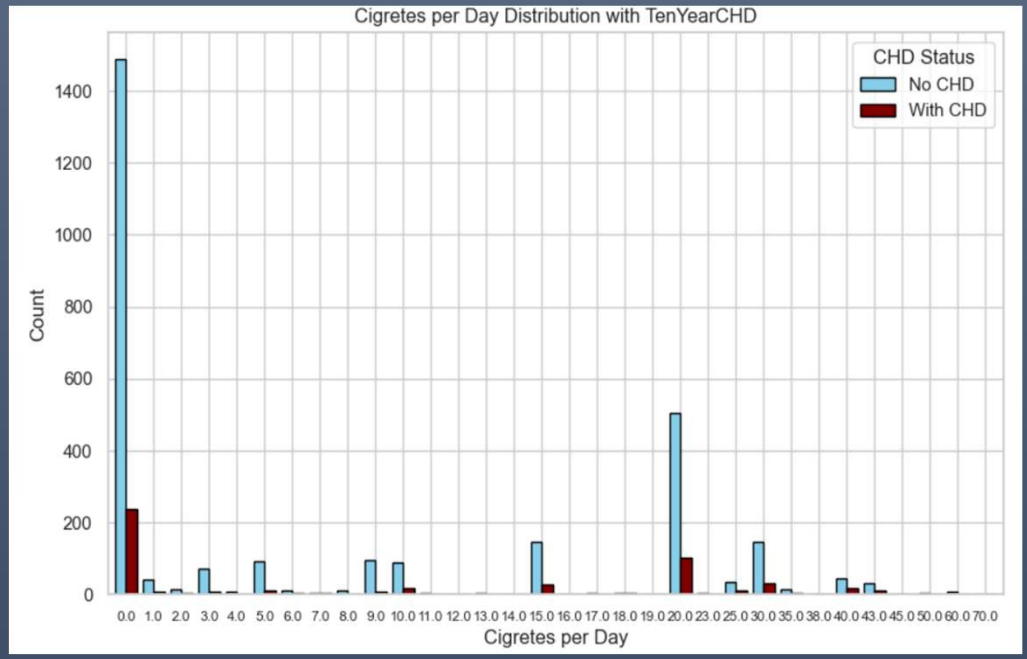
- In age groups like 34 to 47, the number of smokers is close to or slightly exceeds non-smokers. Smoking prevalence is higher in younger and middle-aged groups, declining with age.

Citrates Per Day



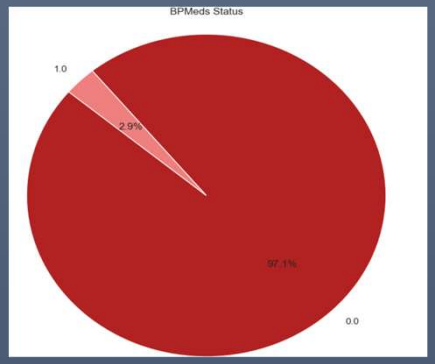
- "Non-smokers are most common (1725 individuals), followed by moderate smokers (1-15 cigarettes/day), with fewer heavy smokers (over 20 cigarettes/day).
- The distribution of cigarette consumption varies, with peaks at certain values.

CigsperDay with TenYearCHD

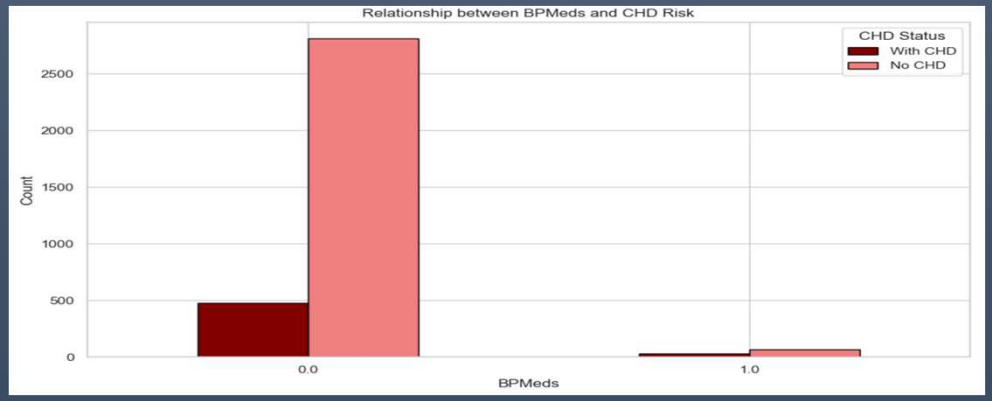


- Those who smoke 20 cigarettes/day have a higher CHD risk (103 cases) compared to those who smoke fewer or none (503 cases). Even smoking just 1 cigarette/day shows CHD cases (7 cases vs. 41 without). Fewer individuals smoke very high amounts, making CHD risk less clear at extreme levels.

BPMeds & BPMeds with TenYearCHD

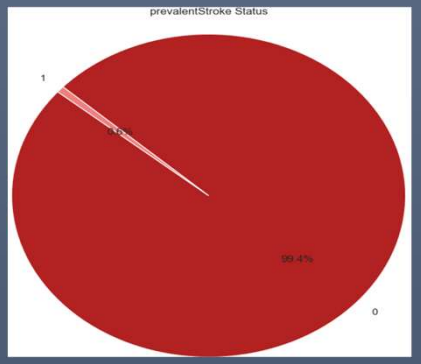


- The majority (97.1%) of individuals aren't taking blood pressure medication, with only 2.9% using it.

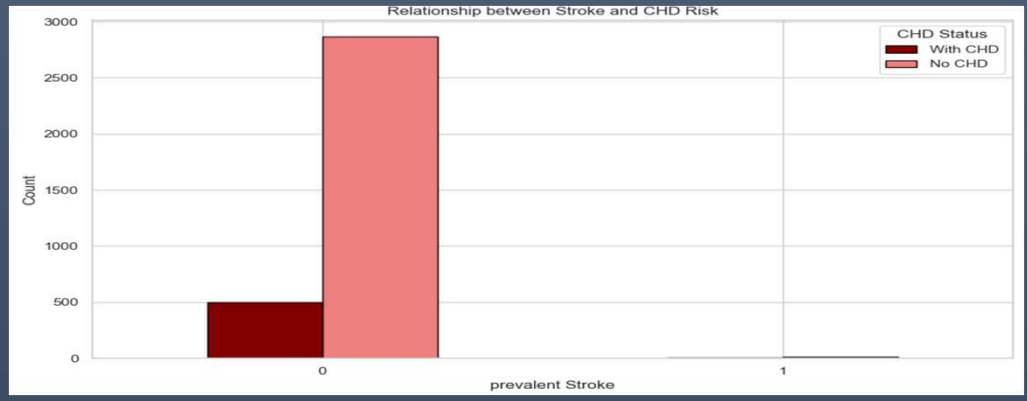


- Most individuals with and without CHD (478 out of 511 and 2812 out of 2879, respectively) aren't using BPMeds, indicating a significant proportion of individuals, whether they have CHD or not, are not using blood pressure medication.

Stroke & Stroke with TenYearCHD

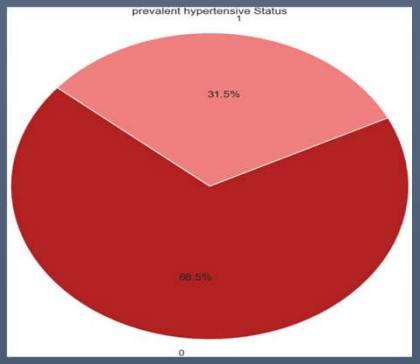


- The majority (3368 out of 3390) haven't had a stroke, with only a few (22 out of 3390, 0.6%) experiencing one. Prevalent stroke appears rare.

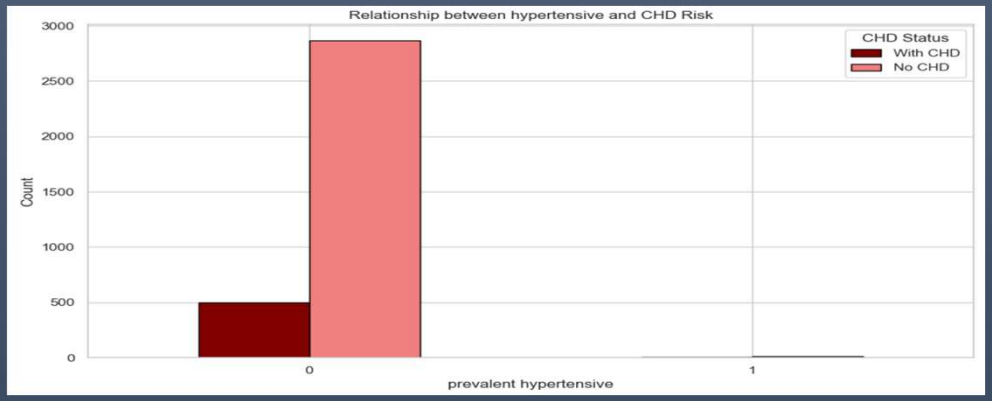


- Among those with CHD, 501 haven't had a prevalent stroke (0), compared to 10 who have (1). For those without CHD, 2867 haven't had a stroke (0), while 12 have (1), indicating that individuals with CHD are more likely to have had a stroke compared to those without CHD.

Hypertensive & Hypertensive with TenYearCHD

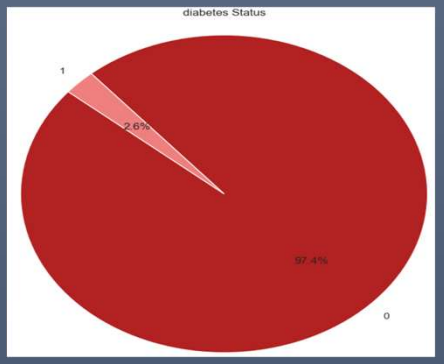


- About 68.5% (2321 out of 3390) of individuals don't have prevalent hypertension, while approximately 31.5% (1069 out of 3390) do.

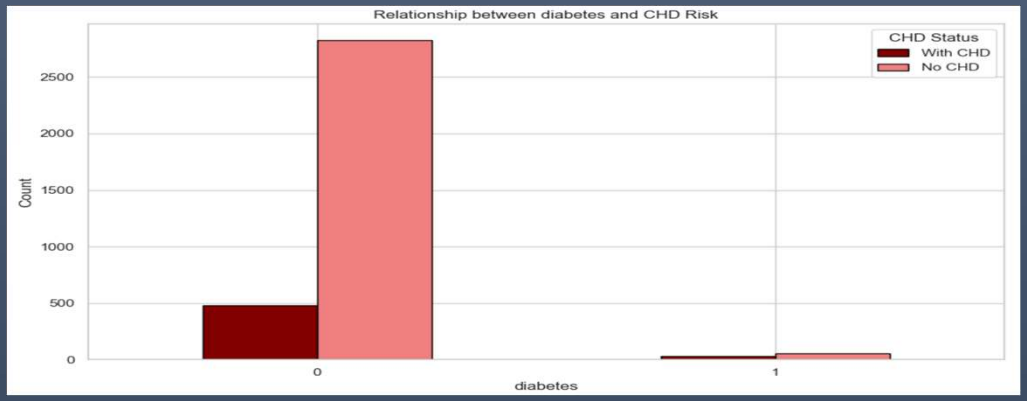


- Among individuals with CHD, slightly more have prevalent hypertension, while among those without CHD, slightly more don't, indicating a relatively equal distribution of CHD among hypertensive and non-hypertensive individuals.

Diabetes & Diabetes with TenYearCHD

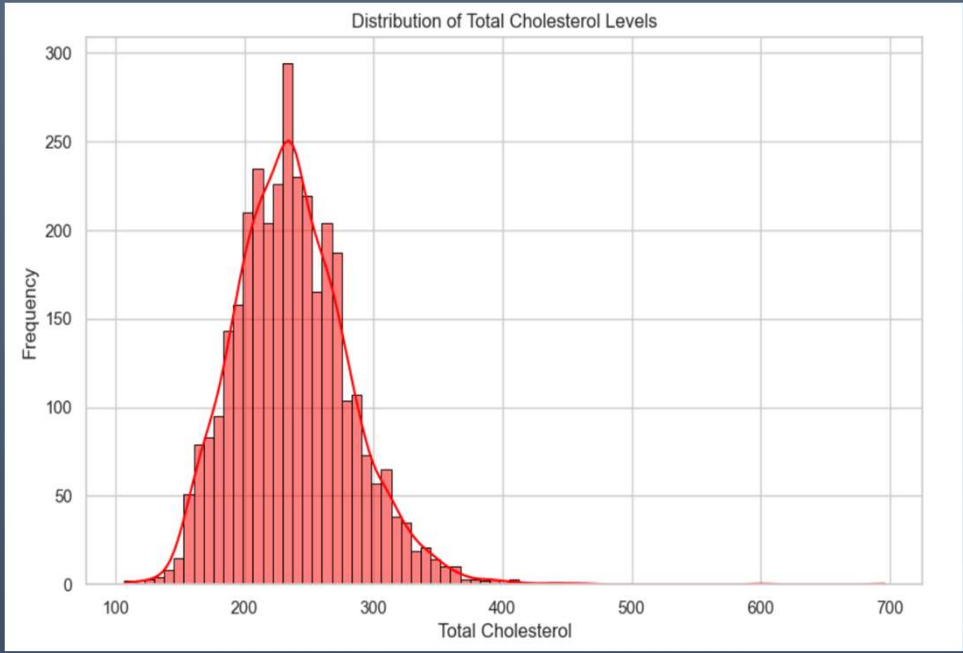


- Most individuals (97.4%) don't have diabetes (3303 out of 3390), while a small percentage (2.6%) do (87 out of 3390).



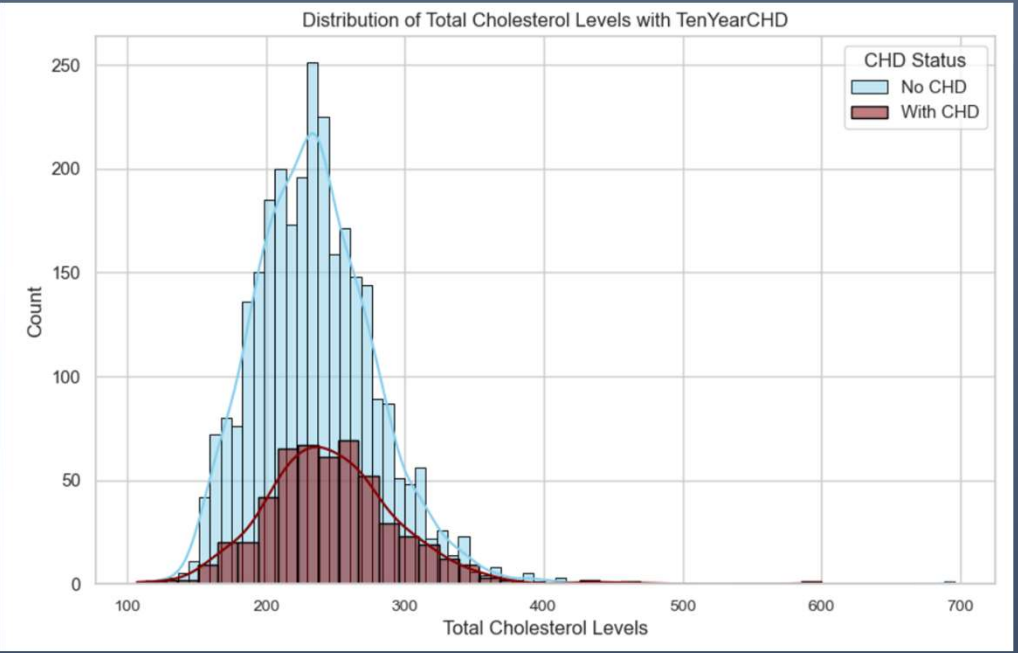
- Most people with CHD (478 out of 511) and without CHD (2825 out of 2879) don't have diabetes, indicating a higher occurrence of CHD among individuals who do not have diabetes compared to those who do have diabetes.

Total Cholesterol



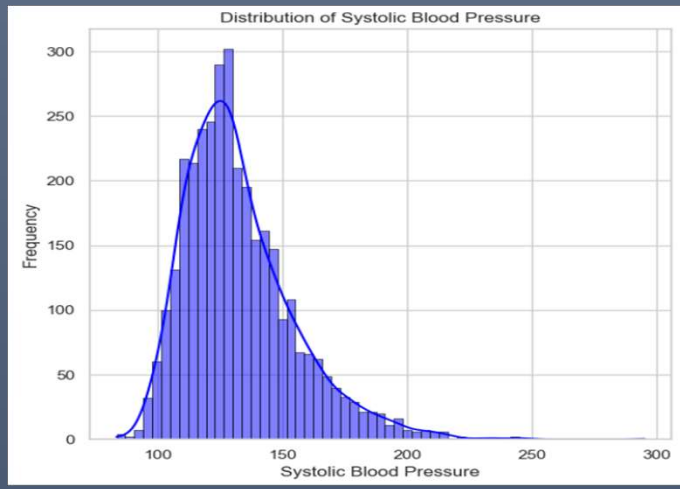
- The dataset's cholesterol levels range from 77 to 696, positively skewed towards lower values. Most individuals have lower cholesterol, with fewer having higher levels.

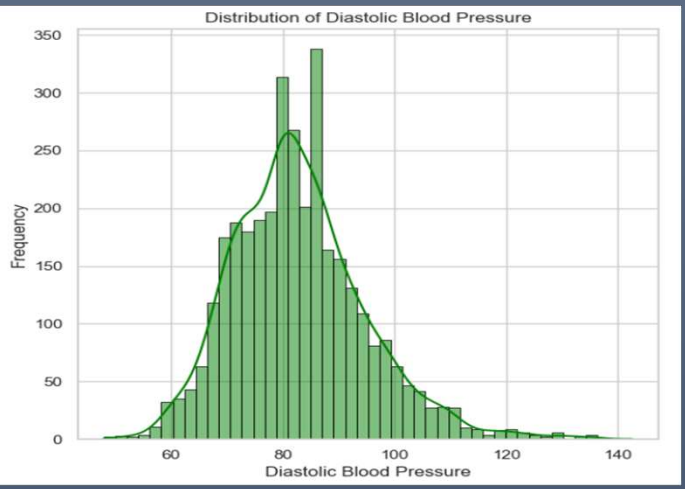
Total Chol with TenYearCHD



- For individuals without CHD, cholesterol levels peak around certain values, gradually decreasing as levels increase. However, for those with CHD, cholesterol levels vary widely, lacking a clear peak in distribution.

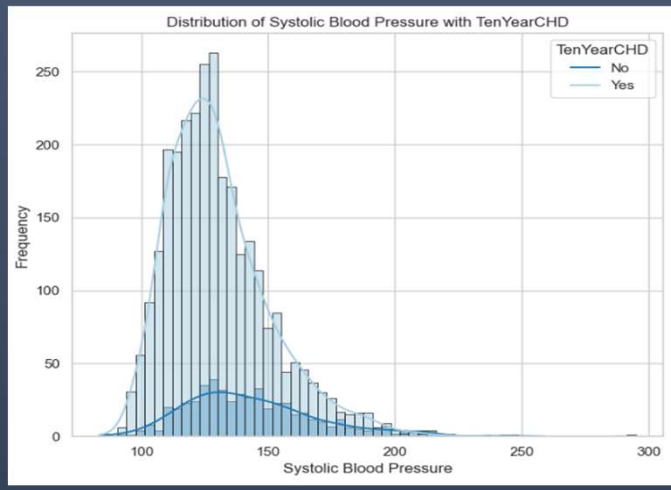
Systolic & Diastolic Pressure

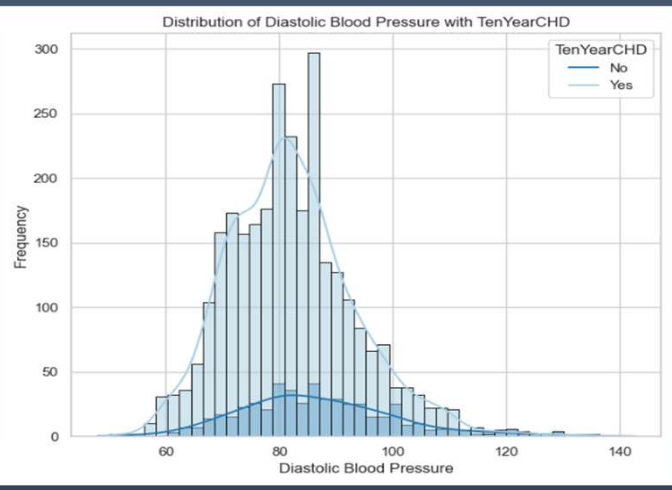




- Both systolic and diastolic blood pressure distributions display peaks around specific ranges, with KDE curves indicating smooth distributions and some variability around the peaks.

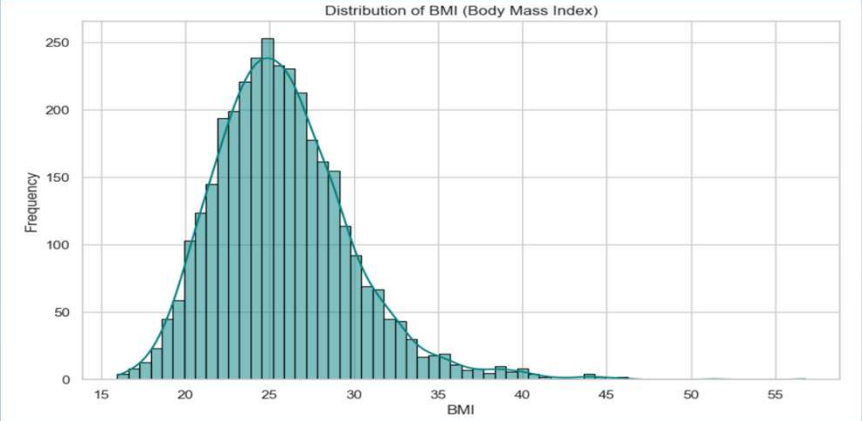
Systolic & Diastolic Pressure with TenYearCHD



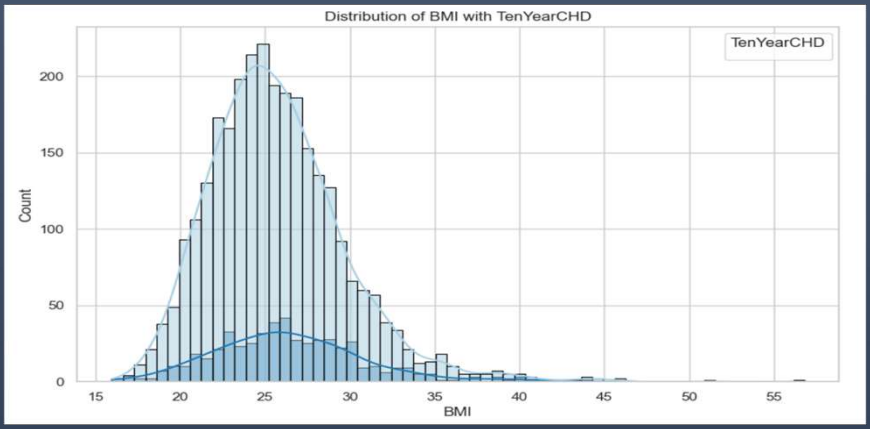


- Individuals who have CHD generally have higher systolic and diastolic blood pressure levels compared to those without CHD, suggesting the importance of both in predicting coronary heart disease over ten years.

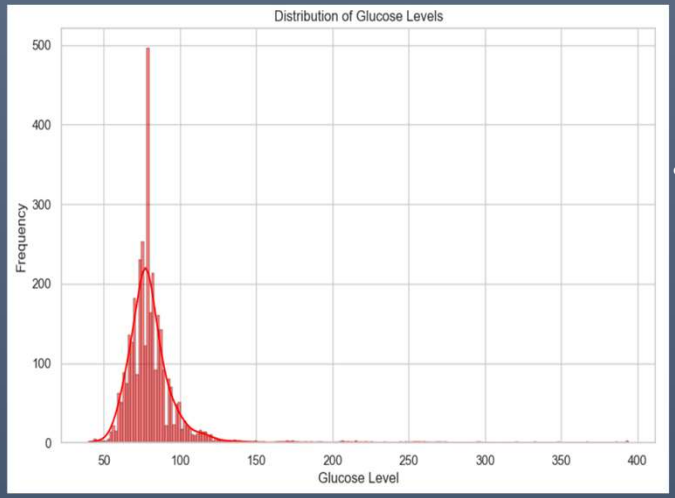
BMI & BMI with TenYearCHD



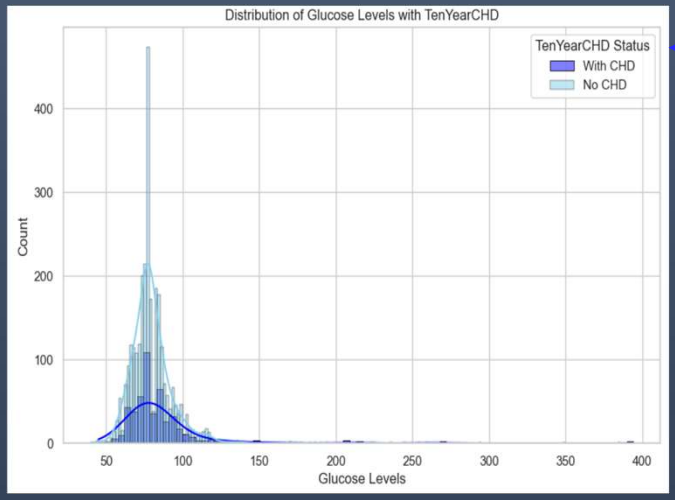
- The most common BMI value is 25.38, occurring 24 times. Other BMI values are also present, each with their frequency of occurrences.



Glucose & Glucose with TenYearCHD

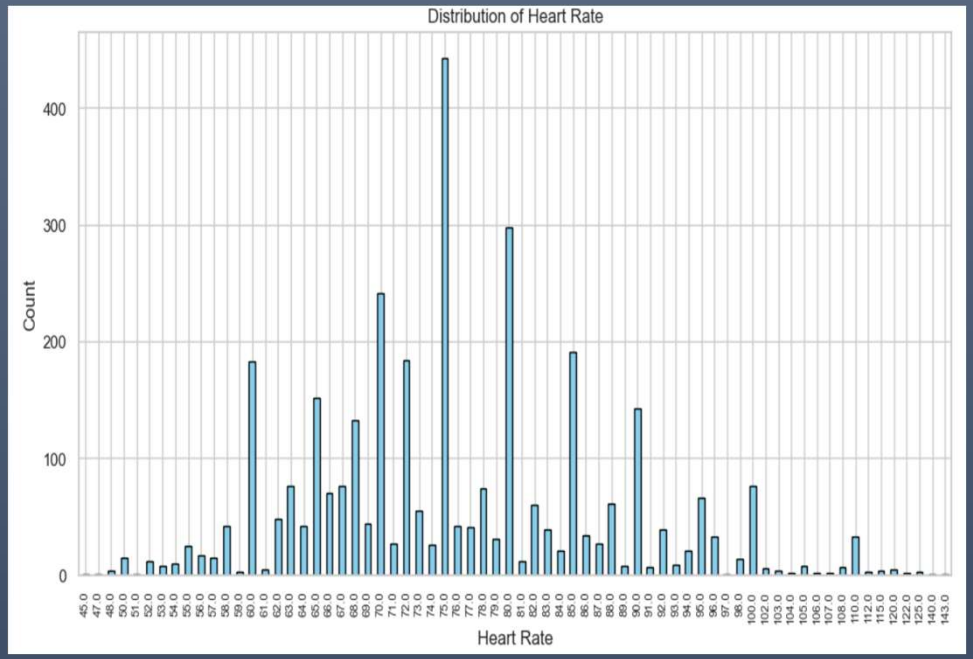


- Glucose levels range from 40 to 394, with the most frequent level being 78.0(count: 421). However, some levels, like 191.0 or 119.0, occur only once.



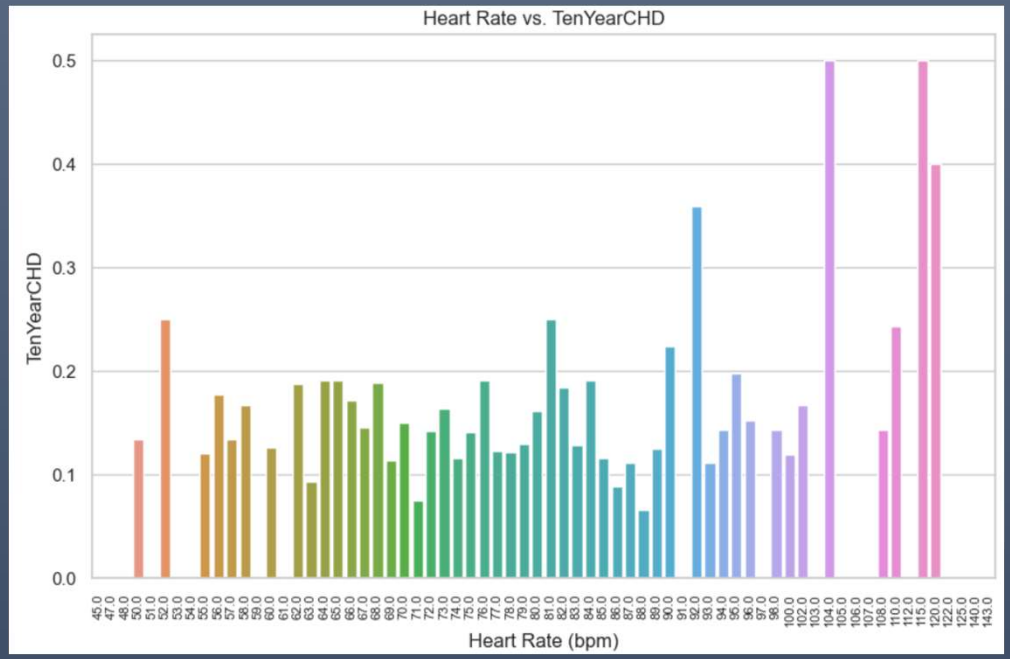
- Glucose level 78.0 is common in both individuals with and without CHD, but it's more frequent in individuals without CHD than those with it

Heart Rate



- The majority of people have heart rates clustered around specific values like 75, 80, and 70 beats per minute (bpm).

Heart Rate with TenYearCHD



- There doesn't appear to be a clear pattern or trend indicating a strong relationship between heart rate and the likelihood of TenYearCHD.

Data Preprocessing
&
Feature Engineering

➢ Encoding Part

- The categorical columns “sex” and “ is_smoking” are transformed into numerical columns separately using the Label Encoder.
- The Label Encoder assigns a unique numerical label to each category within each column, thereby converting categorical data into a format suitable for Machine Learning models requiring numerical input.

➢ Missing Value Treatment

- The highest percentage of missing values is found in the 'glucose' variable, with approximately 9% missing. Other variables with notable missing values include 'education' (2.57%), 'cigsPerDay' (0.65%), 'BPMeds' (1.30%), and 'totChol' (1.12%).
- So I used the imputation Method to handle the missing values.
- The "Education" feature represents discrete categories rather than a continuous scale. So we are using Mode to fill in the missing values.
- The “BPMeds” feature is a nominal variable indicating whether or not the patient was on blood pressure medication. So we are using Mode to fill in the missing values.
- “cigsPerDay”, “totChol”, “BMI”, “heartrate”, and “glucose” are continuous. So we are using the Median for filling in the missing values.
- This ensures the dataset is more comprehensive and ready for analysis.

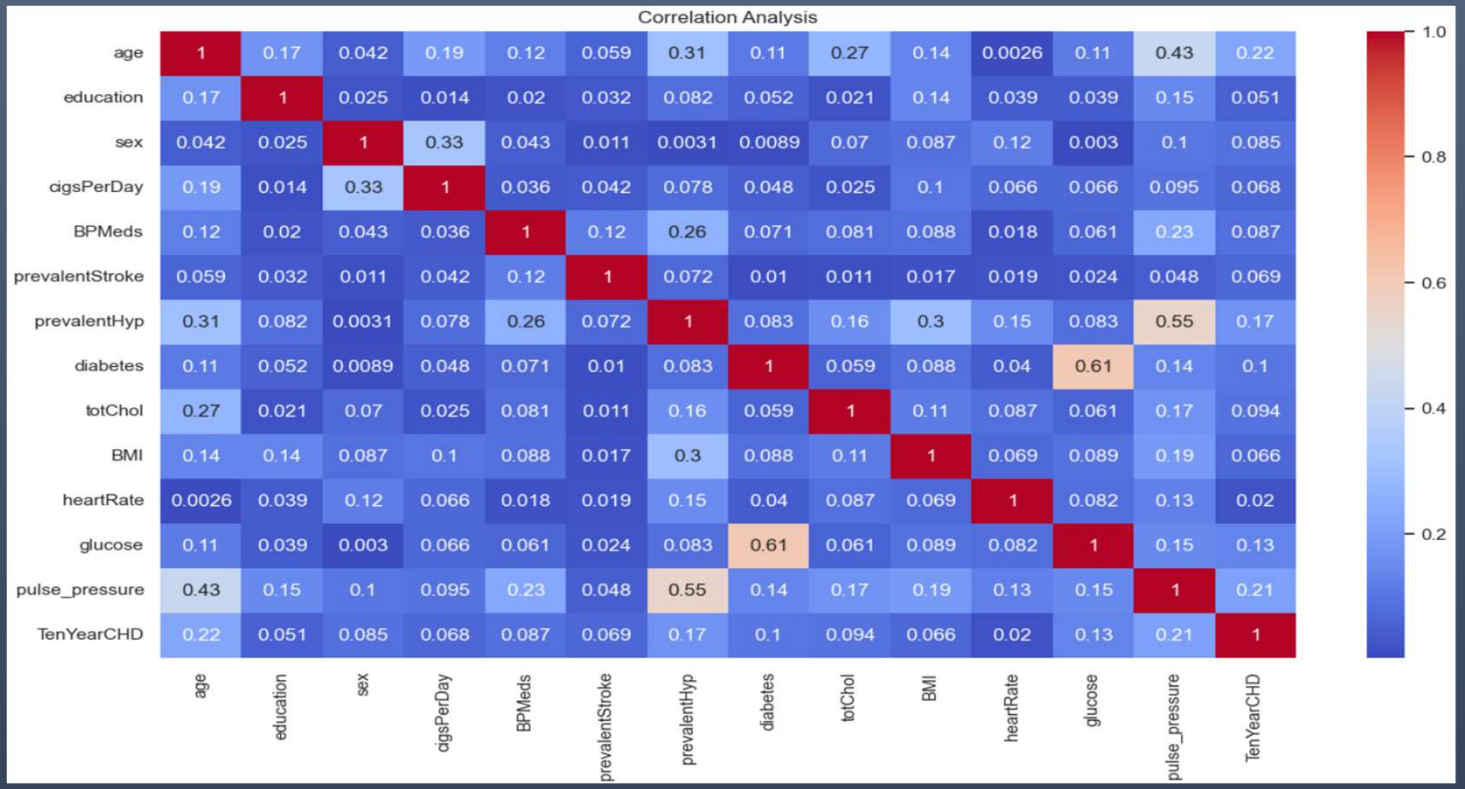
➢ Correlation using Heatmap

- There are moderate positive correlations between TenYearCHD and other cardiovascular risk factors such as age (0.22), prevalentHyp (0.17), sysBP (0.21), and diaBP (0.14), indicating that individuals with higher blood pressure levels are more likely to develop TenYearCHD.
- Glucose levels (glucose) also show a positive correlation with TenYearCHD (0.13), suggesting that higher glucose levels may be associated with a higher risk of developing coronary heart disease over ten years.
- Systolic blood pressure (sysBP) and diastolic blood pressure (diaBP) exhibit a strong positive correlation of 0.78. So we have to handle this multicollinearity.

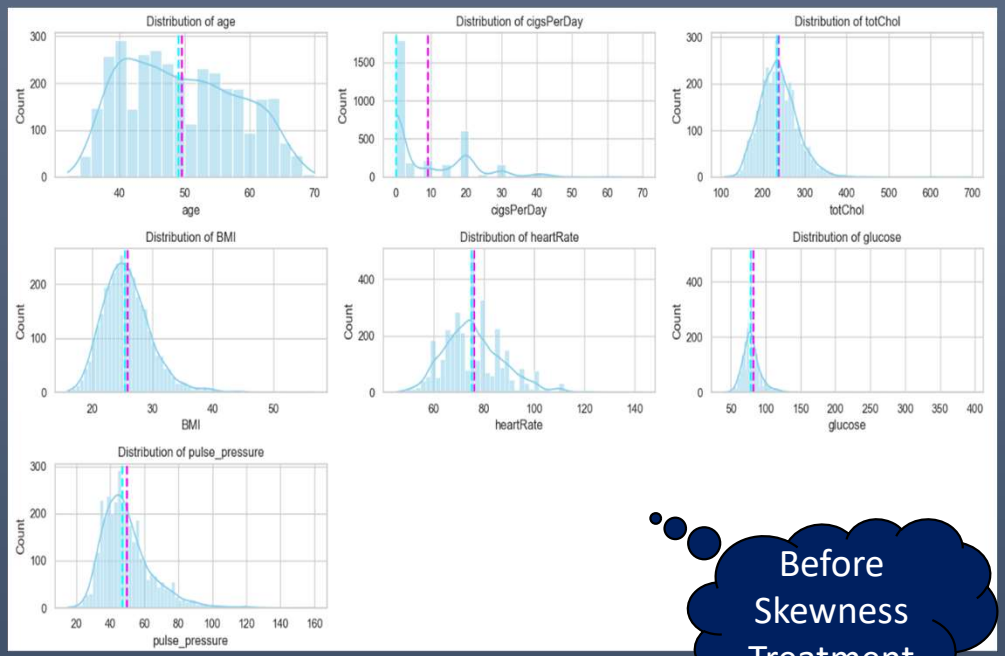
- Also, the correlation between "is_smoking" and "cigsPerDay" is 0.76, indicating a strong relationship.
- Since "cigsPerDay" already captures smoking behaviour by representing the number of cigarettes smoked daily, "is_smoking" doesn't offer additional insights.
- Therefore, we can drop the "is_smoking" column.

➢ Handling Multicollinearity

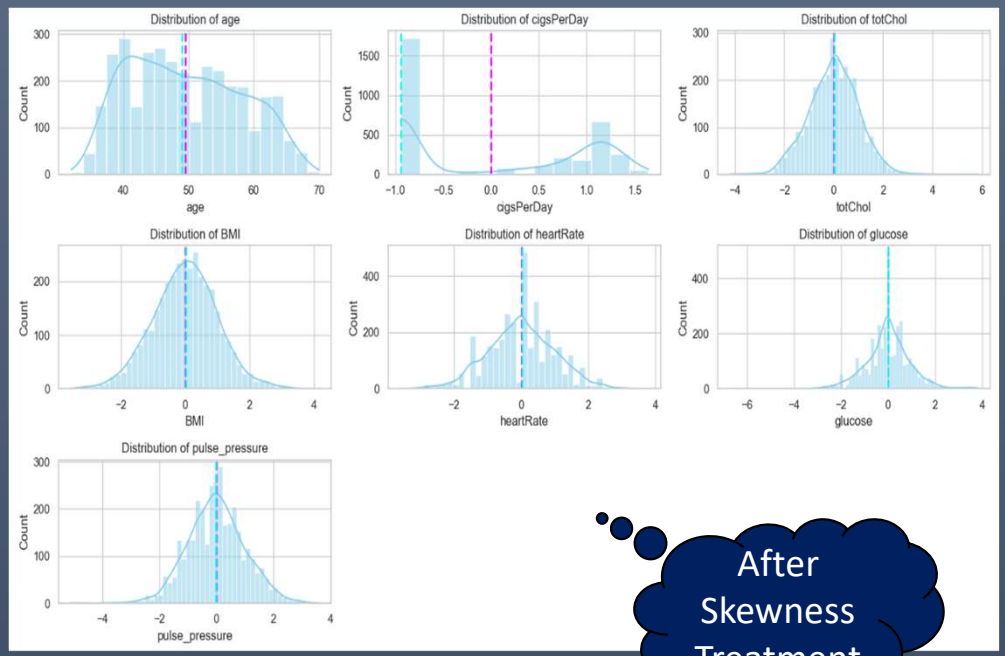
- To handle multicollinearity between two independent continuous variables Systolic BP & Diastolic BP, I have replaced these two columns with a new variable 'pulse pressure', which is given as follows: Pulse Pressure = Systolic BP - Diastolic BP
- I can make the classification model more accurate and reliable.



➢ Skewness treatment



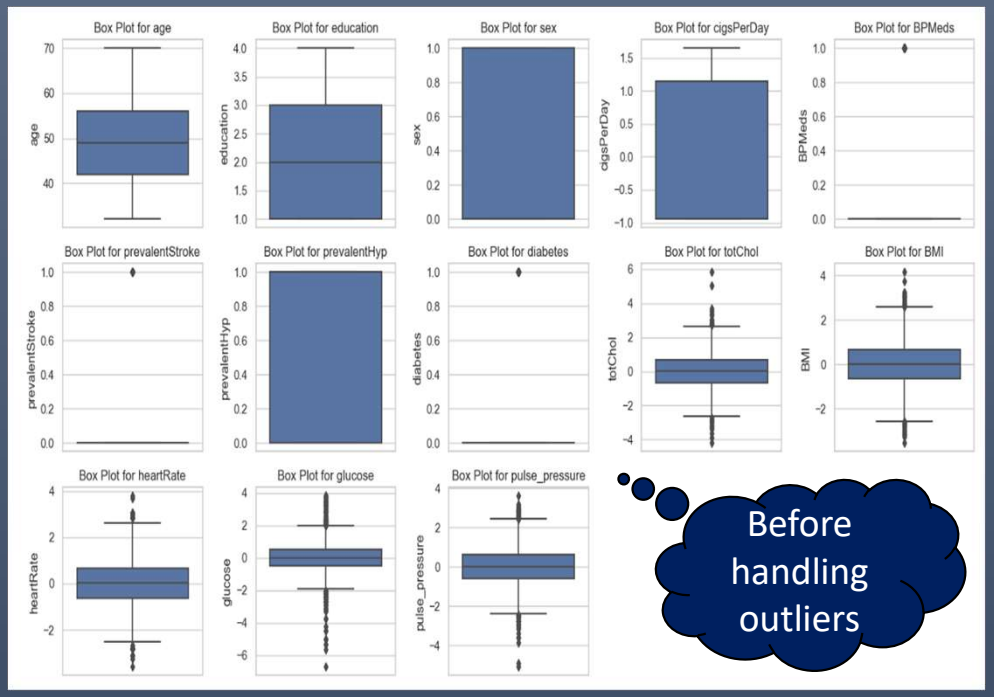
Before Skewness Treatment

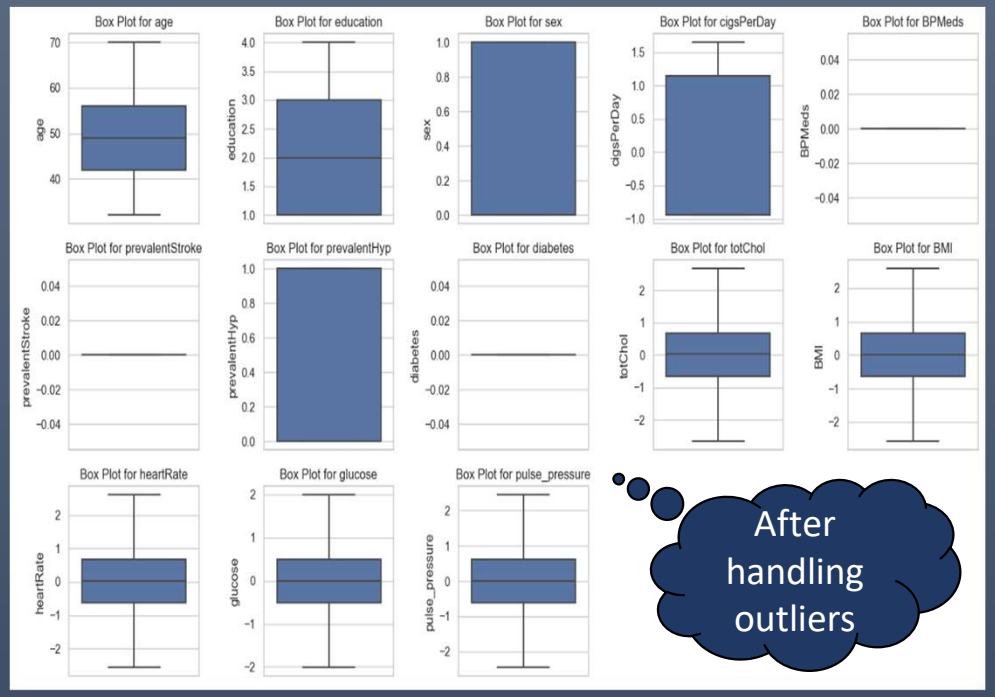


After Skewness Treatment

- To preprocess continuous numerical columns ("age", "cigsPerDay", "totChol", "BMI", "heartRate", "glucose", "pulse_pressure") by applying a power transformation to reduce skewness, and it provides visualizations to assess the effectiveness of the transformation. Also, it can handle slightly heavier tails and outliers.
- Except for cigsPerDay, I have successfully been able to reduce the skewness in the continuous variables.
- Now these distributions are closer to symmetric distribution.

➢ Handling Outliers





- There are outliers present in our data. It will make an impact on model prediction so we have to handle these outliers.
- I have used the IQR method to remove those outliers.
- If I removed the outliers, I lost about 12% of the data. So instead of removing outliers, I have replaced them with lower or upper bound.
- So the dataset remains the same as the original dataset with 3390 rows and 14 features.

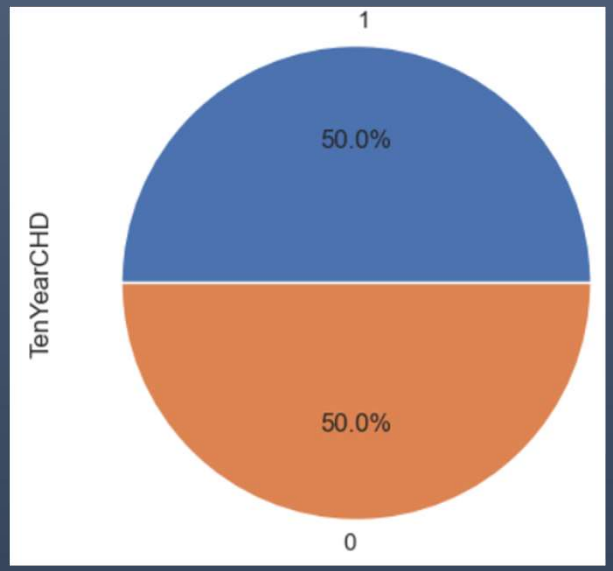
➢ Scaling the Data(StandardScaler)

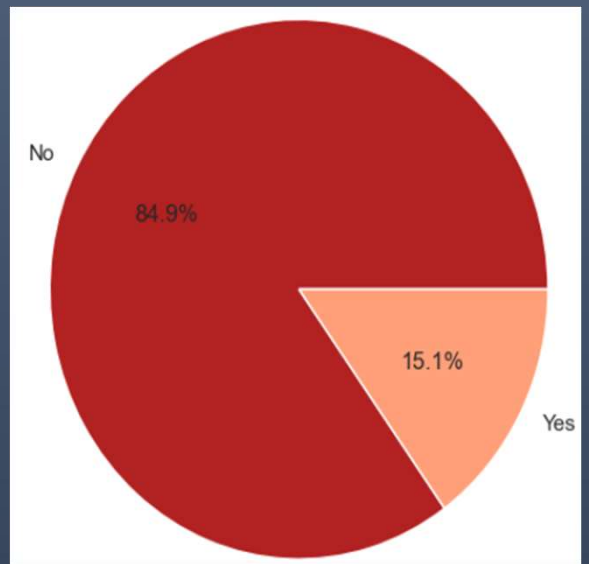
- Standardization is a common preprocessing step that ensures all features have a mean of 0 and a standard deviation of 1, which can be important for certain machine learning algorithms that are sensitive to the scale of the input features.

➢ Handling Imbalance Data

- Imbalanced datasets, where one class is significantly underrepresented compared to another, can pose challenges for machine learning models, especially in classification problems. Class imbalance can lead to biased models that perform poorly on the minority class (0:No – 84.9%, 1:Yes - 15.1%)

- I used SMOTE (synthetic minority oversampling technique) over the under-sampling technique (near miss) because I got better results in terms of model performance
- It will take random samples from 1:Yes values and make the duplicates till they both have the same proportion.
- Data is balanced now, both the category of TenYearCHD have 50% data each.





Model Building

- I have Split the data into test and train in a 20:80 ratio.
- For choosing the base model I have used all the classification models such as Logistic Regression, KNN, Naive Bayes, Decision Tree, Random Forest and XGboost.
- I Checked their performance by different metrics Accuracy Score, AUC ROC Score, Precision, Recall, and F1 score.
- Some of these models were fine-tuned with hyperparameter tuning to get the best parameters for these three models so that I could choose the final model.
- I have used the Grid search CV method to get the best parameters for the base models.

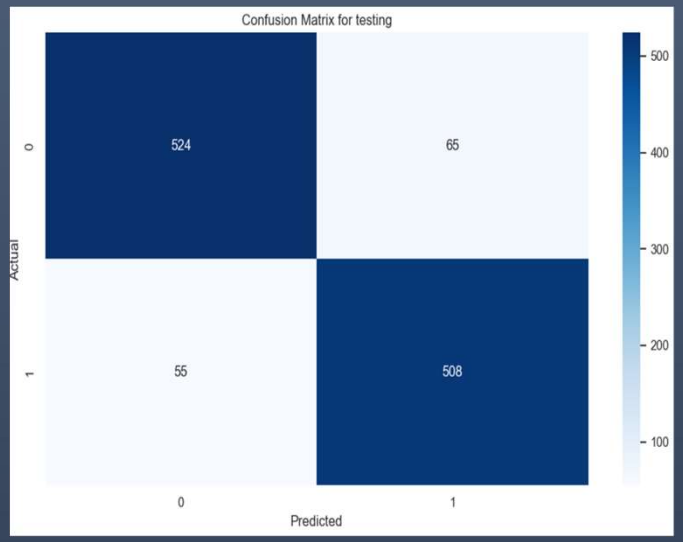
	Model Name	Accuracy Score	ROC-AUC Score	Precision Score (0)	Precision Score (1)	Recall Score (0)	Recall Score (1)	F1-Score (0)	F1-Score (1)
0	Logistic Regression	0.68	0.68	0.69	0.67	0.68	0.68	0.69	0.68
1	KNN	0.86	0.86	0.98	0.79	0.75	0.98	0.85	0.87
2	Naive Bayes	0.66	0.66	0.67	0.65	0.67	0.66	0.67	0.65
3	Decision Tree	0.73	0.73	0.80	0.69	0.63	0.83	0.71	0.75
4	Random Forest	0.90	0.90	0.91	0.89	0.89	0.90	0.90	0.89
5	XG Boost	0.89	0.89	0.88	0.92	0.93	0.86	0.90	0.89

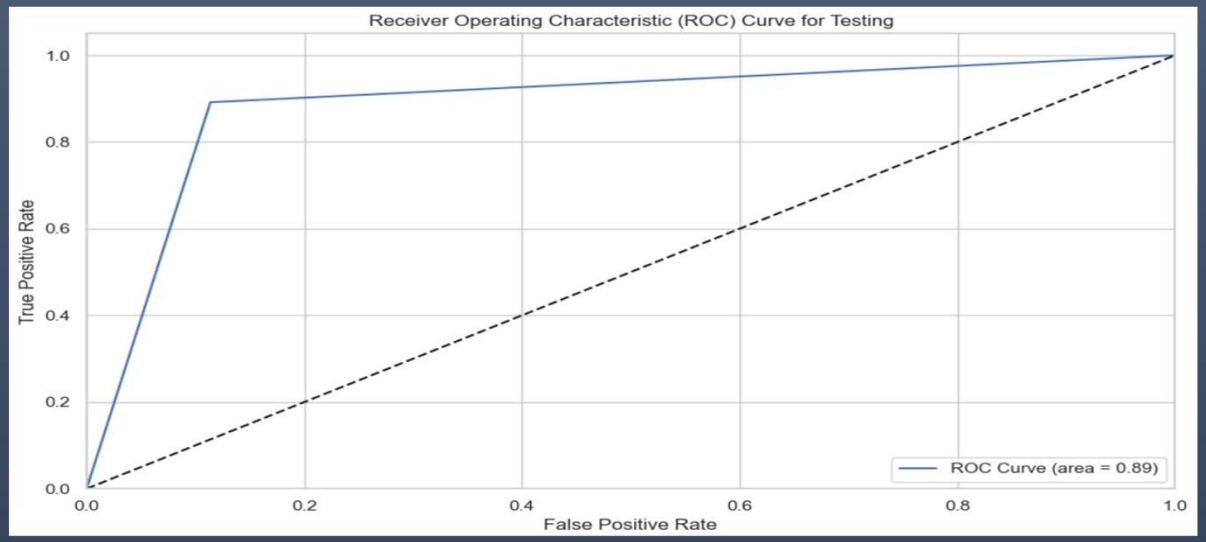
- Therefore, based on the provided information, **Random Forest** seems to be the most suitable choice for predicting the 10-year risk of Coronary Heart Disease in this scenario.

Results & Insights

✓ Result

- Random Forest has the highest accuracy scores, indicating they have the highest overall correctness in predicting the 10-year risk of coronary heart disease.
- Random Forest has the highest recall score for class 1, indicating it identifies true positive cases effectively.
- Random Forest also has the highest precision scores for class 0, indicating fewer true negative predictions.
- Random Forest has the highest F1 scores, indicating a good balance between precision and recall.





✓ Insights

- CHD risk tends to increase with age, with a peak in middle age and then decreases in older age groups.
- Pulse pressure might be associated with the risk of developing CHD over the next ten years.
- An individual's heart rate can give important clues about how their heart is doing and whether they might be at risk for heart problems over the next ten years. So heart rate is an important feature for prediction.
- Total cholesterol, Glucose, BMI, education, Cigsperday, Prevalent Hypertension and sex were the least important features in determining the risk of CHD.

• Usage:

- The project helps doctors find people who might get heart disease in the next ten years. By finding them early, talking to specialists about their health and getting help with specific things to lower their risk of getting heart disease.
- The project can contribute to raising awareness about CHD risk factors and promoting healthier lifestyles. By understanding the factors that contribute to CHD risk, individuals can make informed decisions about their health behaviours and take proactive steps to reduce their risk.
- The project findings can help researchers learn more about what causes heart disease, how it gets worse over time, and what works best to treat it. Also, leaders can use this information to come up with plans to help both individuals and whole groups of people lower their risk of heart disease.

Thank You!