Capstone Project

Cardiovascular Risk Prediction

Anuj Menaria

Points to Discuss:

- Defining the problem statement
- Data Cleaning
- Data visualization & EDA
- Data preprocessing
- Feature selection
- Preparing Dataset for model
- Applying model
- Model validation and selection

Problem Statement:

- 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 includes over 4,000 records and 15 attributes. Each quality has the potential to be risky. Risk factors might be medical, behavioral, or demographic.

DATA DESCRIPTION:

- ✓ **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 average number of cigarettes the person smoked daily.
- ✓ **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 DESCRIPTION(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)

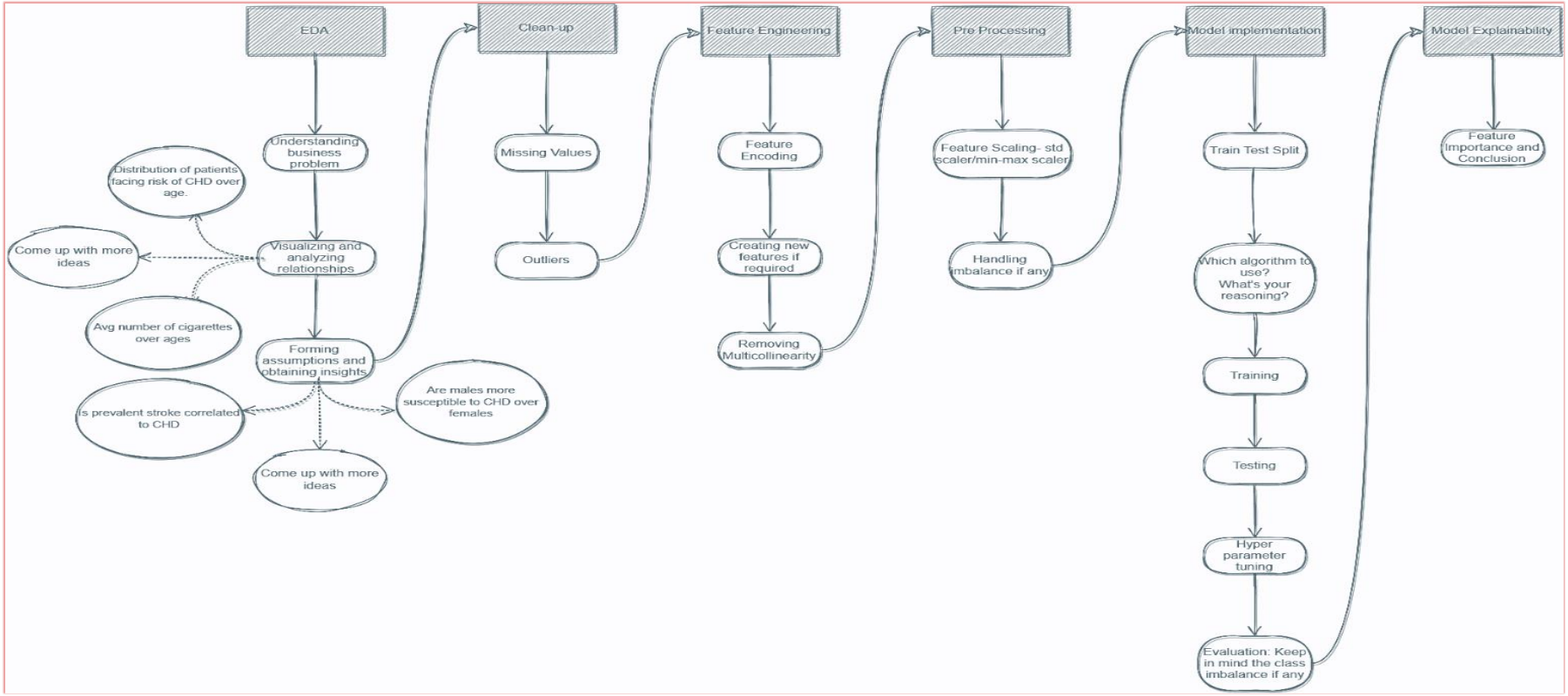
  Heart Rate: In medical research, variables such as heart rate though discrete, are considered continuous because of the large number of possible values.

  Glucose: Glucose level (Continuous).

- ✓ **Predict variable (desired target):**

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

Flow-chart:


EDA
Understanding business problem
Distribution of patients facing risk of CHD over age.
Visualizing and analyzing relationships
Come up with more ideas
Avg number of cigarettes over ages
Forming assumptions and obtaining insights
Is prevalent stroke correlated to CHD
Are males more susceptible to CHD over females
Come up with more ideas
Clean-up
Missing Values
Outliers
Feature Engineering
Feature Encoding
Creating new features if required
Removing Multicolinearity
Pre Processing
Feature Scaling- std scaler/min-max scaler
Handling imbalance if any
Model implementation
Train Test Split
Which algorithm to use? What's your reasoning?
Training
Testing
Hyper parameter tuning
Evaluation: Keep in mind the class imbalance if any
Model Explainability
Feature Importance and Conclusion

Dataset:

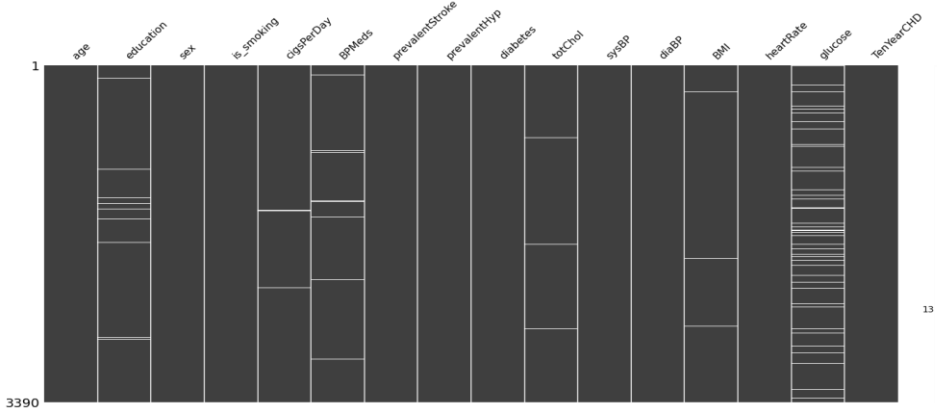
After Loading the dataset we can observe that it has :

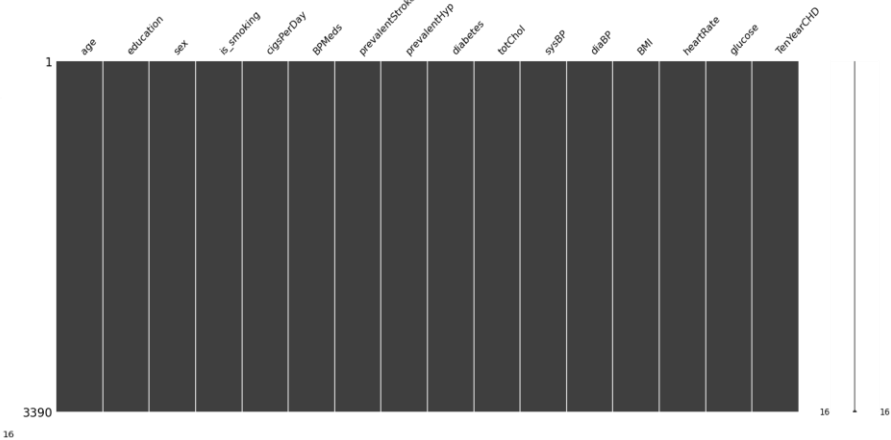
Rows: **3390**

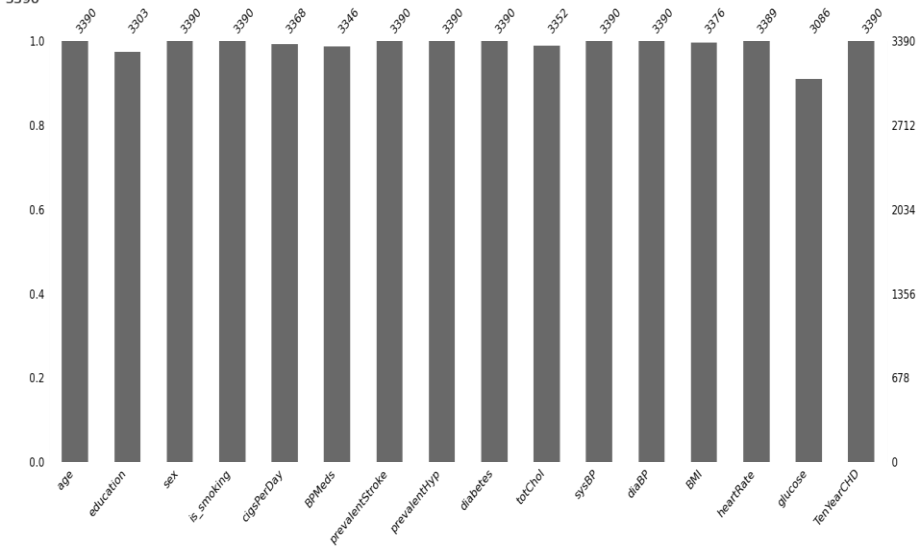
Columns: **17**

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

Data Cleaning:

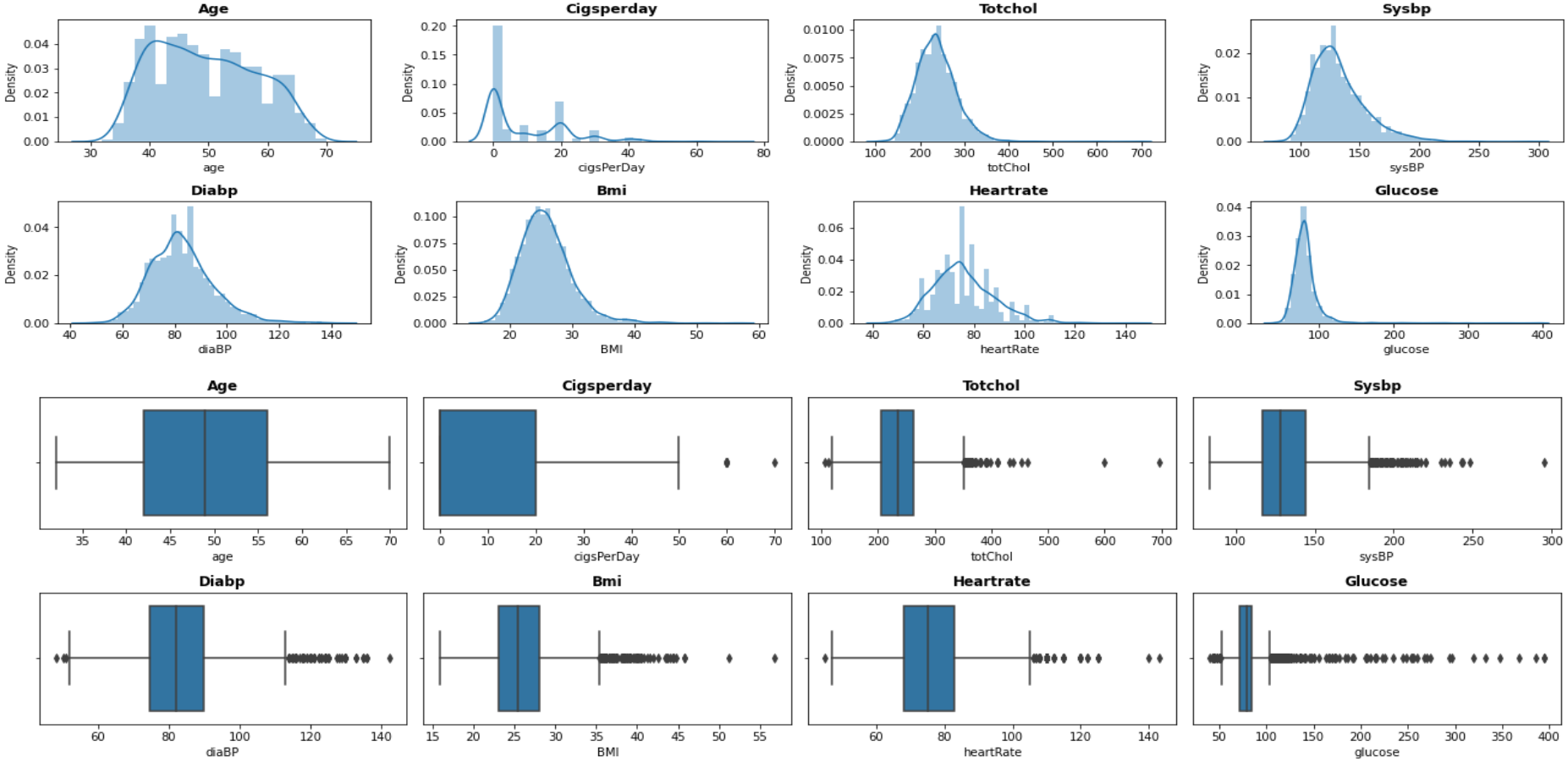




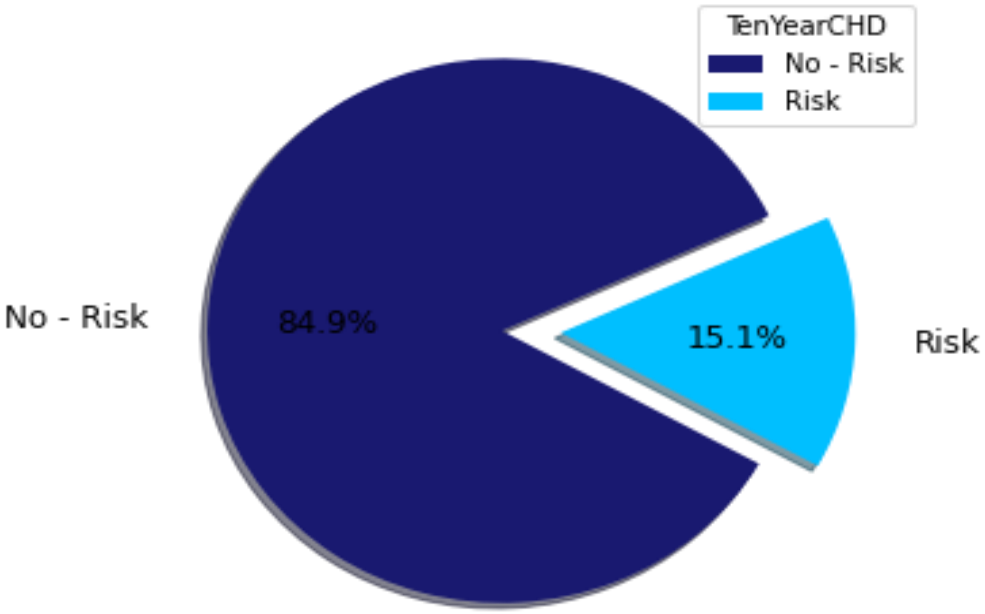


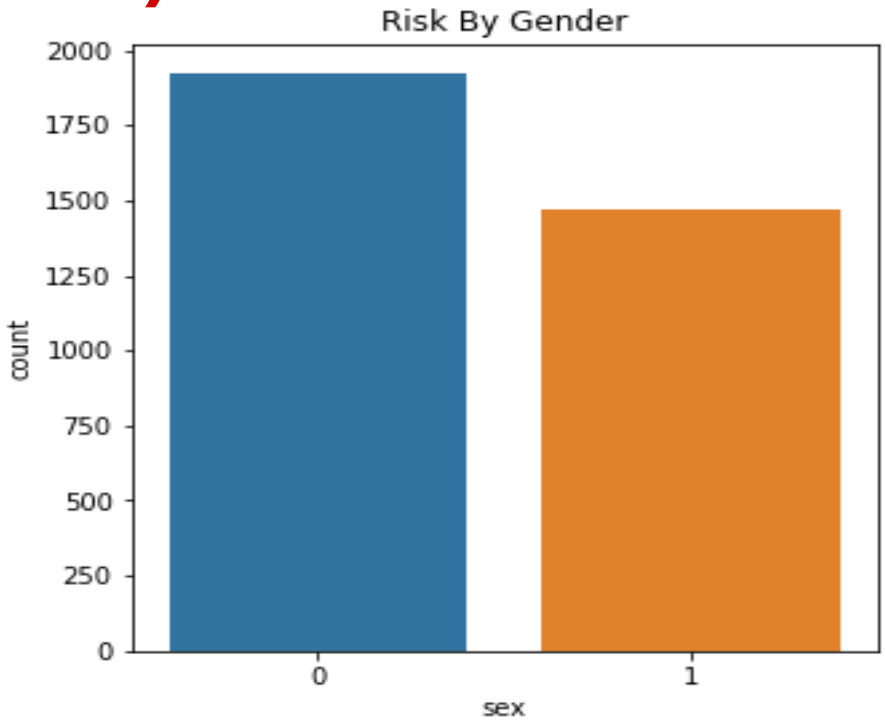
- We ran into missing values in our data when preparing the data.
- We were able to see the missing data both before and after addressing the missing values with the aid of the missngno library.
- Understanding the distribution of missing values in data and the relationship between features is made easier with the help of the missingno library.
- The graphs show that the two categories with the greatest amount of missing data are glucose and education.

Data Visualization & EDA:



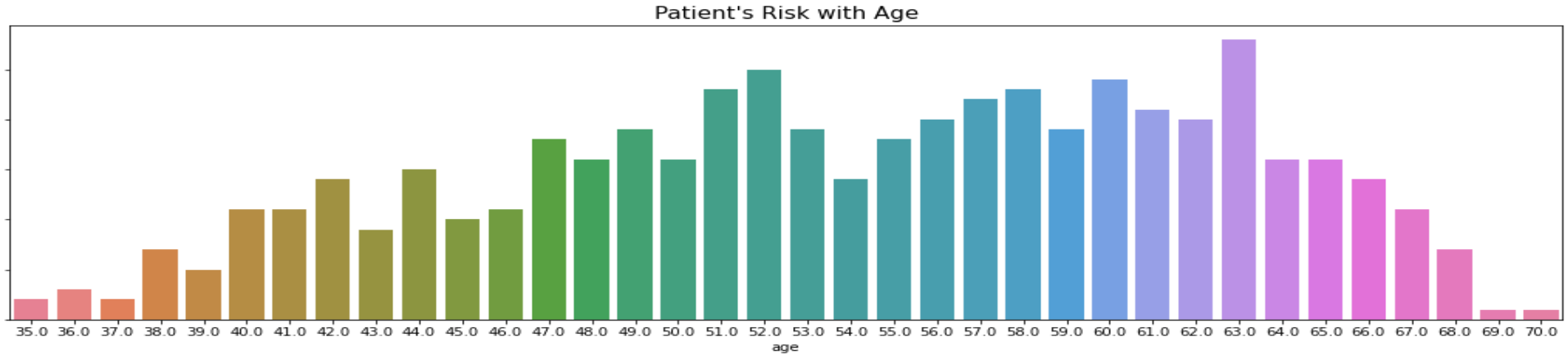
Data Visualization & EDA(Contd.):





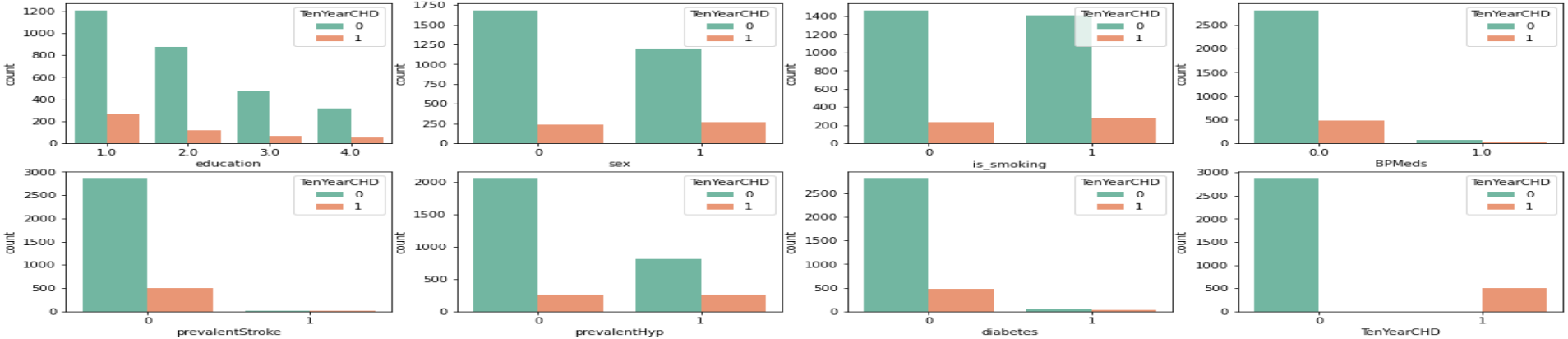
- We can see that while the vast majority of people do not have a cardiovascular risk, only a little portion 15.1% does.
- The ratio of males to females is greater than anticipated.

Data Visualization & EDA(Contd.):



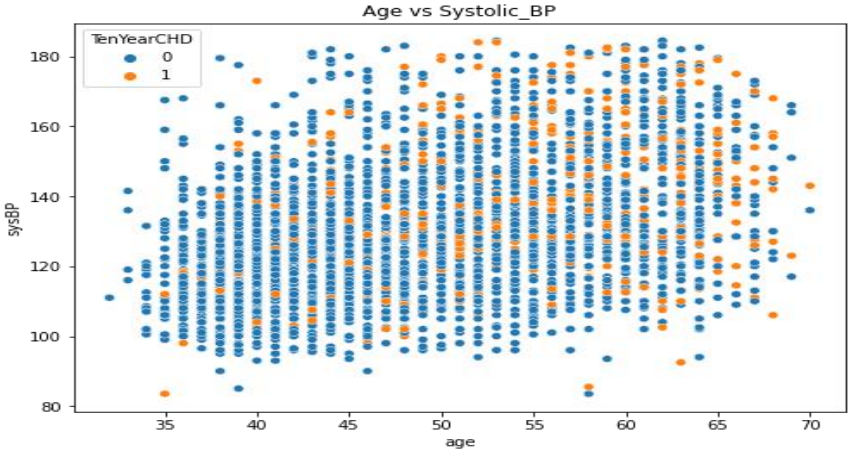
- In the provided dataset, the number of persons in the middle age (between 44-58) and old age (above 60) groups is the largest.
- There is the least danger among those who are young (between 35 and 43 years old).
- Whether a person smokes or not, age is a major factor in cardiovascular risk.

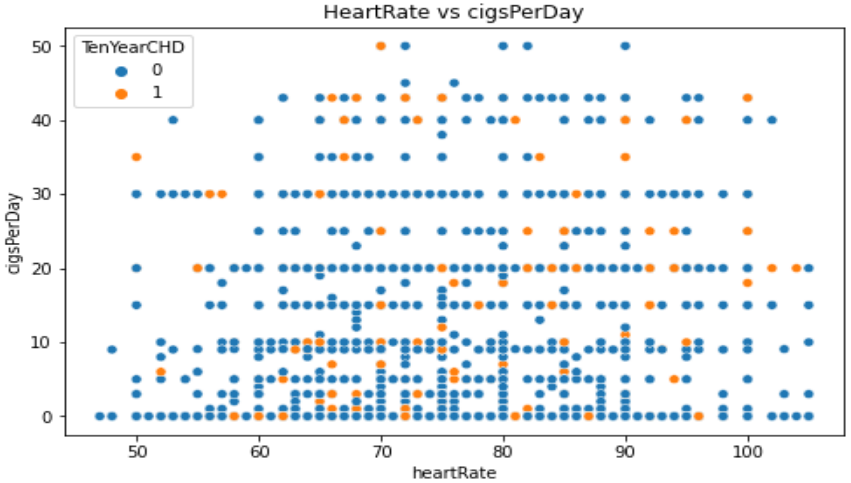
Data Visualization & EDA(Contd.):



- Compared to those with greater education levels, those with lower education levels are more likely to face risk.
- Since both men and women face the same danger, gender has no effect on that risk.
- Smokers are at higher risk regardless of their gender.
- While those who don't use blood pressure medications are in danger, having a common stroke has no impact on the count.
- One may be more at risk if they have prevalent hypertension, however, danger is not always present even in the absence of hypertension.
- People with no disabilities are also at a risk.

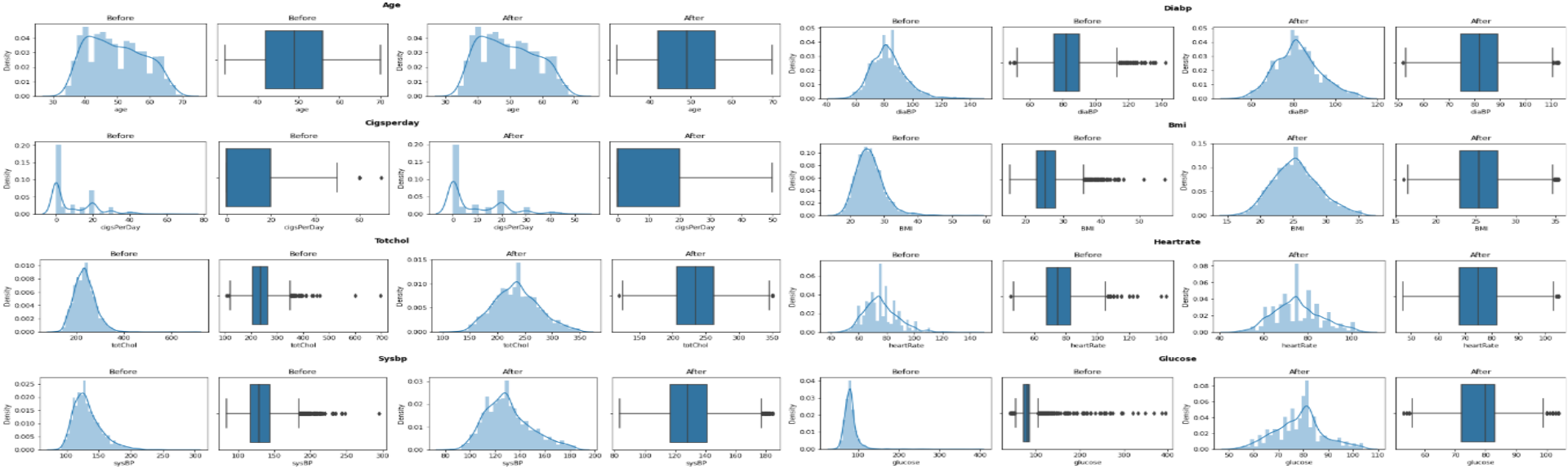
Data Visualization & EDA(Contd.):





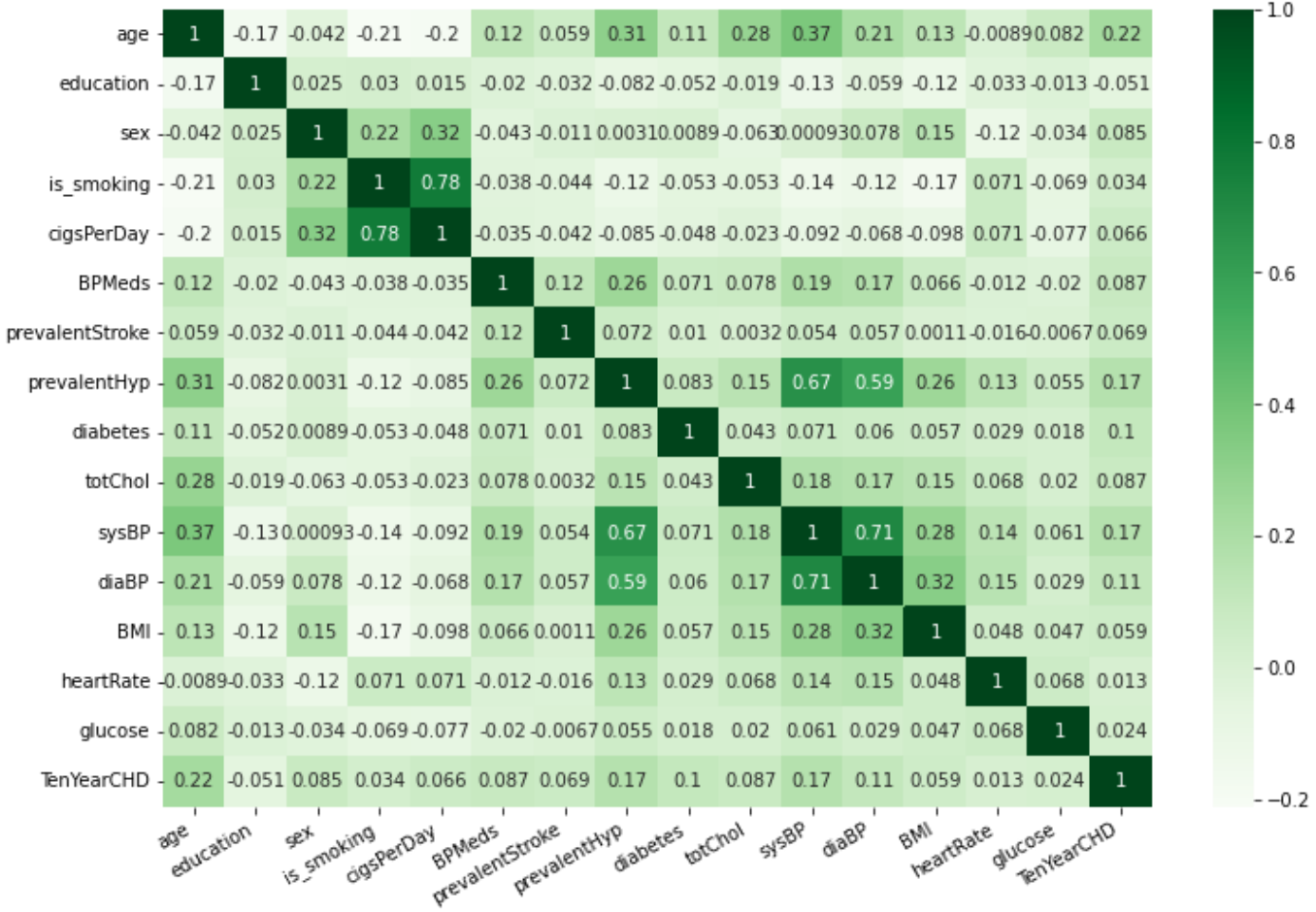
- With the use of a scatter plot, we can compare the relationship between blood pressure and age in the first graph. By looking at the graph, we can conclude that as blood pressure rises with age, the risk of cardiovascular illnesses also rises. The risk may rise because of the age factor.
- The largest number of cigarettes smoked in a day is about 50, and we have compared heart rate with that number in the following graph.
- However, as we can see from the graph, smoking only has a very slight negative impact on a person. This may be because of the age aspect; as a person's age grows, the likelihood that they may develop a problem may increase.

Handling Outliers:



- The "fence" is constructed outside of Q1 and Q3 using the IQR method of finding outliers. Outliers are any values that lie outside of this range.
- Outliers are any observations that fall more than 1.5 IQR outside Q1 or rise more than 1.5 IQR outside Q3. The median values, or the 50th percentile of that column, were used to replace the outliers.

Heatmap:



- Is_smoking and cigsPerDay have a positive correlation of 0.78.
- sysBP and diaBP show the strongest positive correlation of 0.71 with one another.

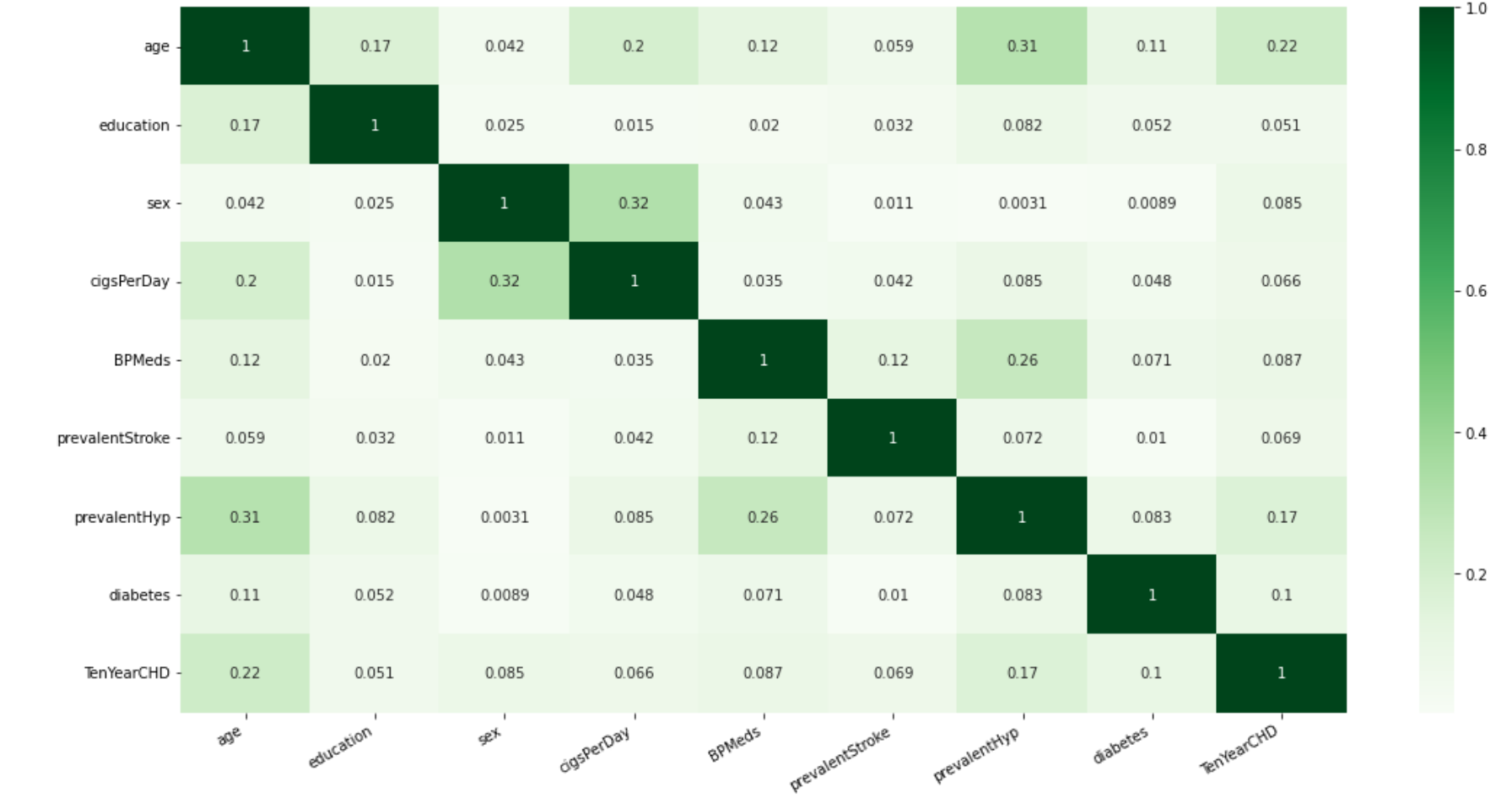
Multicollinearity removal:

	variables	VIF
0	sysBP	132.639335
1	diaBP	127.184097
2	BMI	58.895029
3	glucose	55.694635
4	heartRate	47.895845
5	age	42.769996
6	totChol	37.678636
7	is_smoking	5.080989
8	education	4.652023
9	cigsPerDay	4.337061
10	prevalentHyp	2.355436
11	sex	2.152121
12	BPMeds	1.128187
13	diabetes	1.046447
14	prevalentStroke	1.026684

- The features with VIF scores greater than 10 have been eliminated. The VIF scores of the variables before and after the multicollinearity treatment are shown in the images on the left and right, respectively.
- When all the characteristics are checked for multicollinearity, certain features, such as is_smoking and cigsperday, have a significant correlation with one another.
- The VIF score of all independent variables, which measures how well a variable is explained by other independent variables, has been used to address multicollinearity.

	variables	VIF
0	age	5.385514
1	education	3.965761
2	sex	1.969319
3	cigsPerDay	1.747477
4	prevalentHyp	1.685379
5	BPMeds	1.120397
6	diabetes	1.044828
7	prevalentStroke	1.024822

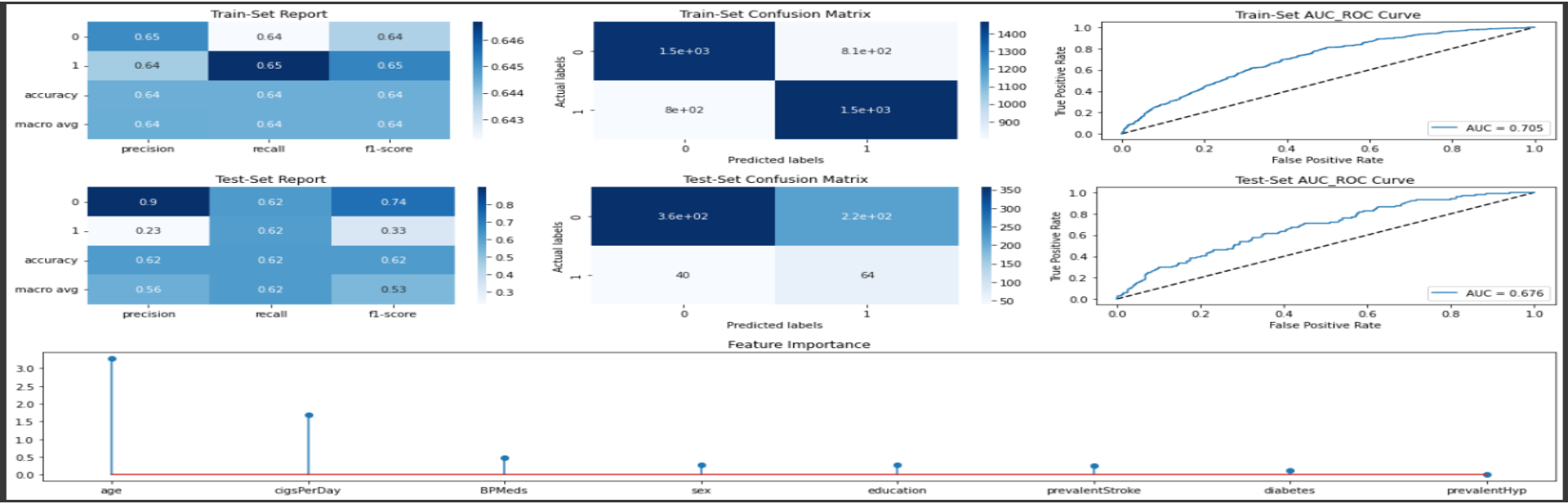
Updated Heatmap:



Model Building:

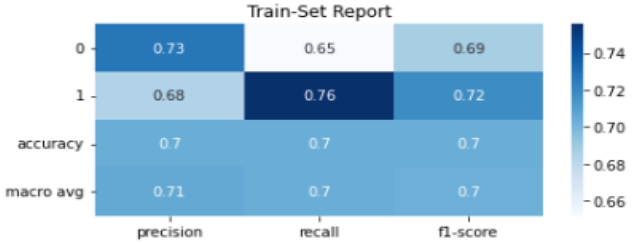
- We do the train-test split before creating the models. 20% of the data were used as test data, while the remaining 80% were used as train data.
- SMOTE (Synthetic Minority Over-sampling Technique) oversampling is used to address the class imbalance, and the Tomek links are then eliminated. After resolving class imbalance, check value counts for both classes at the end.
- It helps deal with significantly varying magnitudes, values, or units during the pre-processing of data. A machine learning algorithm would often weight larger values as higher and evaluate smaller values as lower, regardless of the unit of measurement, if feature scaling is not done.
- Defining a function that provides a classification report for a model's performance on train and test data using the classifier model and train/test splits as inputs. Plots the feature significance as well.

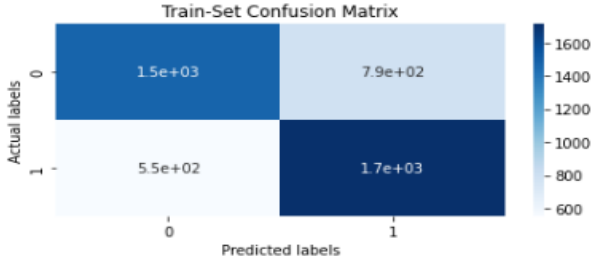
Logistic Regression:

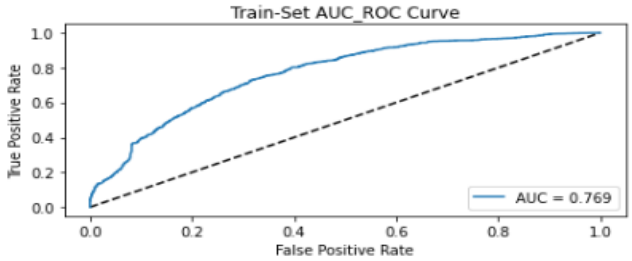


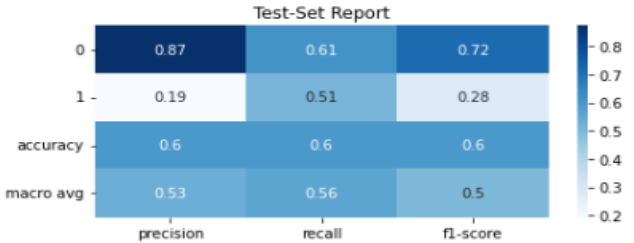
- The method of estimating the likelihood of a discrete result given an input variable is known as logistic regression.
- On test data, logistic regression produces the following outcome for class 1: Precision is 0.23, recall is 0.62, and F1 score is 0.33.
- The most important factor is age, then cigsperday, then BPMeds.

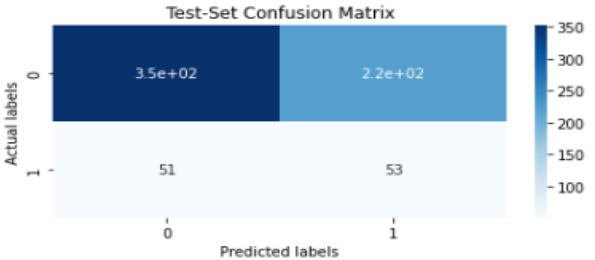
Support Vector Classifier:

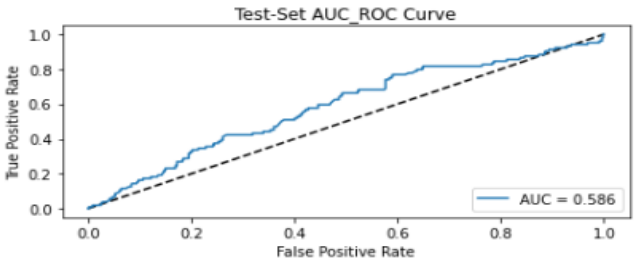






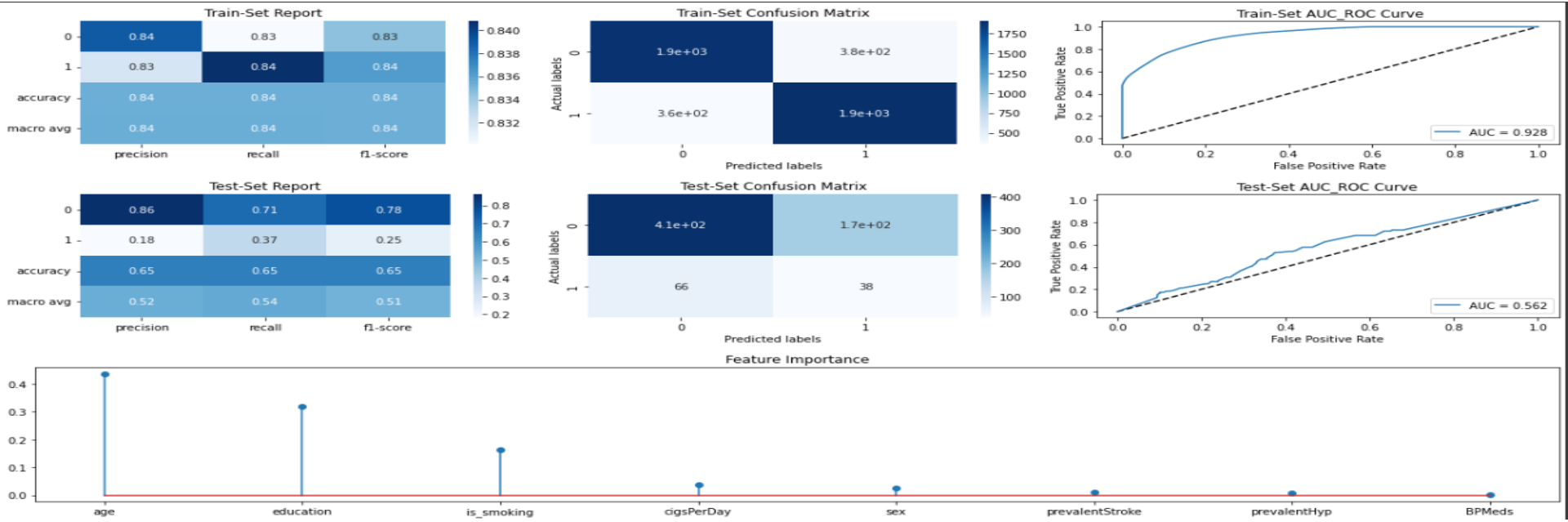






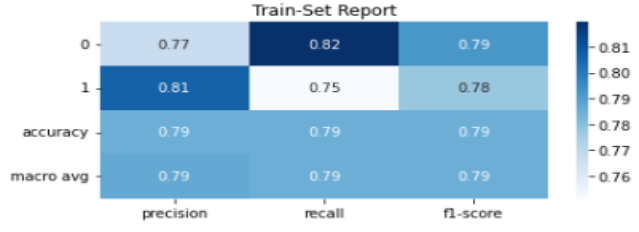
- Here, each piece of data is represented as a point in n-dimensional space, with each feature's value being the value of a certain coordinate. Then, we carry out classification by identifying the hyper-plane that effectively distinguishes the two classes.
- On test data, the Support Vector Classifier produces the following outcome for class 1: Precision is 0.19, recall is 0.51, and F1 score is 0.28.

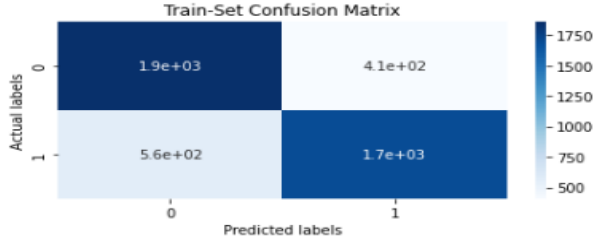
Decision Tree Classifier:

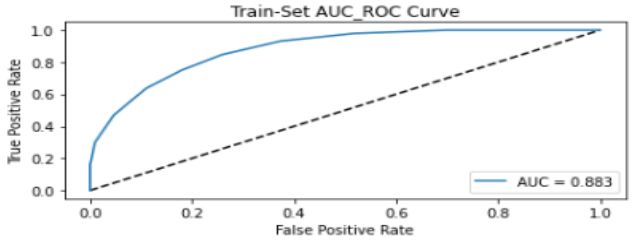


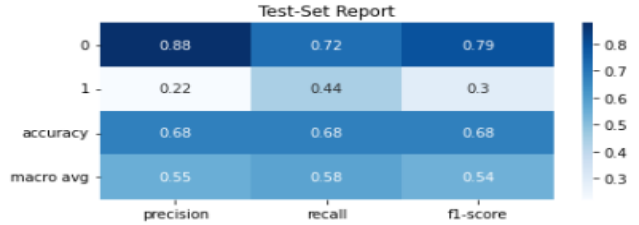
- A decision tree is a methodical strategy to make a choice by breaking the options down into smaller choices. Leaves and decision nodes make up the tree. Its foundation is the idea of entropy.
- On test data, the Decision Tree Classifier produces the following outcome for class 1: Precision is 0.18, recall is 0.37, and F1 score is 0.25.

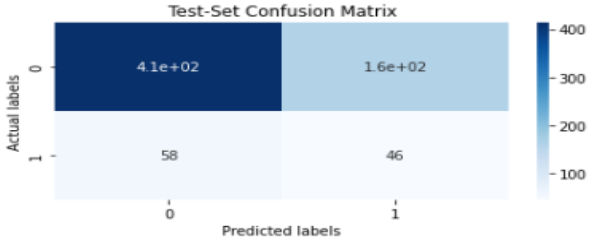
KNeighbor Classifier:

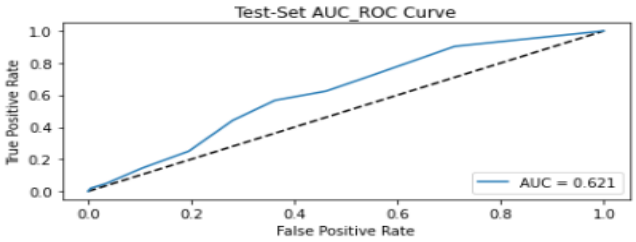






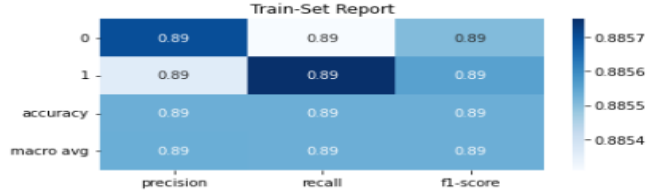


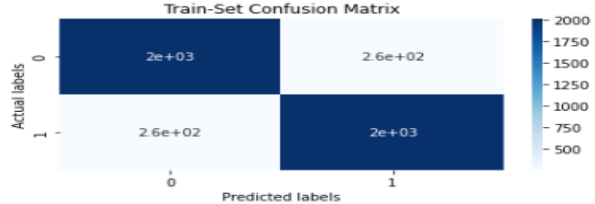


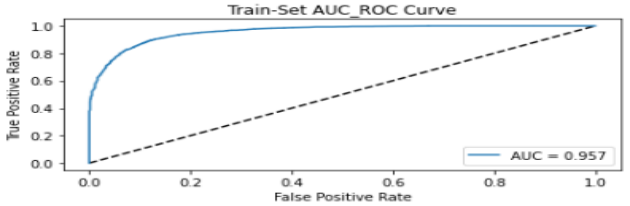


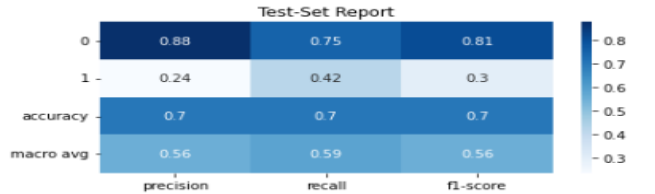
- A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that by utilizing the K-NN method, fresh data may be quickly and accurately sorted into a suitable category.
- On test data, the KNeighbor Classifier produces the following outcome for class 1: Precision is 0.22, recall is 0.44, and F1 score is 0.3.

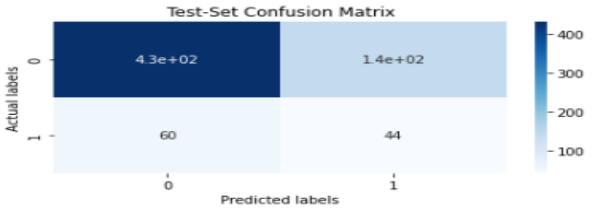
Random Forest Classifier:

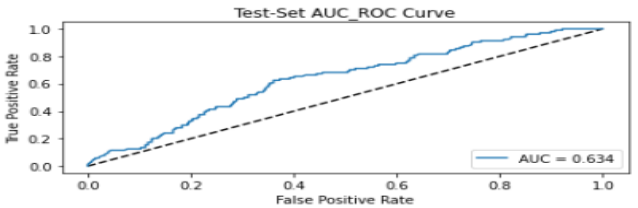


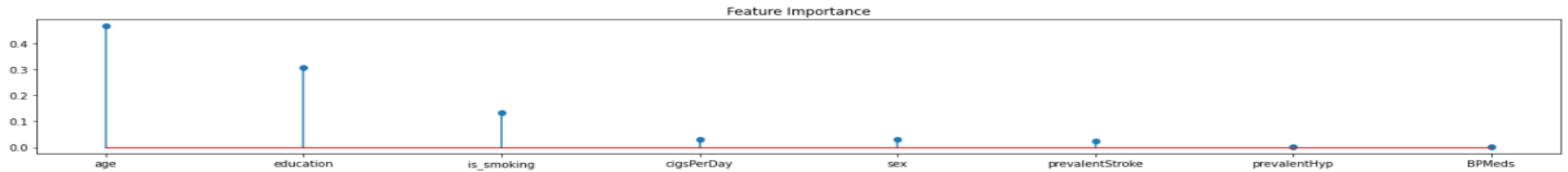












- An ensemble learning technique for classification, regression, and, other problems, random forests or random decision forests work by building a large number of decision trees during training.
- On test data, the Random Forest Classifier produces the following outcome for class 1: Precision is 0.24, recall is 0.42, and F1 score is 0.3.

Conclusion:

EDA insights:

- 15.1% of the population in our dataset is at risk for cardiovascular disease, while 84.9% of the population is unaffected (TenYearCHD).
- Cardiovascular disease is more likely to affect persons between the ages of 51 and 63.
- We can't prove smoking causes heart disease since, as we can see from the count plot, there isn't much of a difference between these two groups, and our severe smoker, who smokes 70 cigarettes a day, doesn't have a ten-year risk.
- Approximately 2800 patients are safe, while 500 patients who have not yet experienced a stroke are at risk.
- We can see that there are more people without diabetes here, and 500 or so people without diabetes are in danger. And only a small percentage of persons with diabetes are at danger.
- Most people with normal cholesterol levels range from 210 to 280, while those at risk have cholesterol levels between 215 and 285; this is a small but perfectly normal difference.
- Most healthy individuals have a heart rate that ranges from 68 to 83, while those who are at risk have a heart rate that ranges from 68 to 84. which holds true for both risky and non-risky individuals.
- The average person smokes between 1 and 10 cigarettes per day, with a heart rate between 60 and 100.

Conclusion(Contd.):

ML Model Results :

- The substantial class imbalance in the training set was addressed by the addition of synthetic data points, which changed the data distribution in the train and test sets. As a result, the large class imbalance in the train set and the mismatch in the data distribution between the train and test sets are to blame for the high performance of the train set models rather than overfitting.
- The testing set's ROCAUC score for SVC is 0.60.
- The testing set's ROCAUC score for logistic regression is 0.62.
- The testing set's ROCAUC score for DTC is 0.66.
- The testing set's ROCAUC score for KNN is 0.68.
- The testing set's ROCAUC score for Random Forest Classifier is 0.72.
- For each model, a classification report and a confusion matrix have been plotted.

Thank You.