Capstone Project

Cardiovascular Risk Prediction

(Classification Algorithm)

By- Navneet Keshri

AI

Steps Involved

- Problem Statement
- Understanding Data
- Exploratory Data Analysis
- Data Cleaning
- Feature Engineering
- Model Building
- Model Implementation
- Conclusion

Project Goal

A group of conditions affecting the heart and blood vessels is known as cardiovascular diseases. They consist of heart disease, which affects the blood vessels that supply the heart muscle.

The goal of the project is to predict the 10-year risk of future coronary heart disease (CHD) for patients.

AI

Problem Statement

- The greatest obstacle facing the medical industry is accurately predicting and diagnosing heart disease. Heart diseases are influenced by numerous factors.
- This project aims to predict the 10-year risk of future coronary heart disease (CHD) for patients in Framingham, Massachusetts.
- A dataset containing demographic, behavioral, and medical risk factors for over 4000 patients is used to build a predictive model.
- The outcome of the project will be a predictive model that can be used by healthcare providers to make informed decisions regarding patient care.

AI

Data Pipeline

1. **Analyze Data:** In this initial step, we attempted to comprehend the data and searched for various available features. We looked for things like the shape of the data, the data types of each feature, a statistical summary, etc. at this stage.

2. **EDA:** EDA stands for Exploratory Data Analysis. It is a process of analyzing and understanding the data. The goal of EDA is to gain insights into the data, identify patterns, and discover relationships and trends. It helps to identify outliers, missing values, and any other issues that may affect the analysis and modeling of the data.

3. **Data Cleaning:** Data cleaning is the process of identifying and correcting or removing inaccuracies, inconsistencies, and missing values in a dataset. We inspected the dataset for duplicate values. The null value and outlier detection and treatment followed. For the imputation of the null value we used the Mean, Median, and Mode techniques, and for the outliers, we used the Clipping method to handle the outliers without any loss to the data.

AI

Data Pipeline

4. **Feature Selection:** At this step, we did the encoding of categorical features. We used the correlation coefficient, chi-square test, information gain, and an extra trees classifier to select the most relevant features. SMOTE is used to address the class imbalance in the target variable.

5. **Model Training and Implementation:** We scaled the features to bring down all of the values to a similar range. We pass the features to 8 different classification models. We also did hyperparameter tuning using RandomSearchCV and GridSearchCV.

6. **Performance Evaluation:** After passing it to various classification models and calculating the metrics, we choose a final model that can make better predictions. We evaluated different performance metrics but choose our final model using the f1 score and recall score.

Dataset Summary

- The data comes from an ongoing cardiovascular study of Framingham, Massachusetts, residents. The dataset provides the patient's information.
- There are approximately 3390 rows and 16 columns available in the dataset.
- While some of the features have a numerical data type, others are categorical. Our target variable is "TenYearCHD"

ATTRIBUTES INFORMATION

Demographic

- age: Age of the patient (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- education: level of education from 1 to 4 (Ordinal Variable)
- Sex: male or female ("M" or "F")

Behavioral

- is_smoking: whether or not the patient is a current smoker ("YES" or "NO")
- cigsPerDay: 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.)

ATTRIBUTES INFORMATION

Medical(history)

- BPMeds: whether or not the patient was on blood pressure medication (Nominal)
- prevalentStroke: whether or not the patient had previously had a stroke (Nominal)
- prevalentHyp: whether or not the patient was hypertensive (Nominal)
- diabetes: whether or not the patient had diabetes (Nominal)

Medical(current)

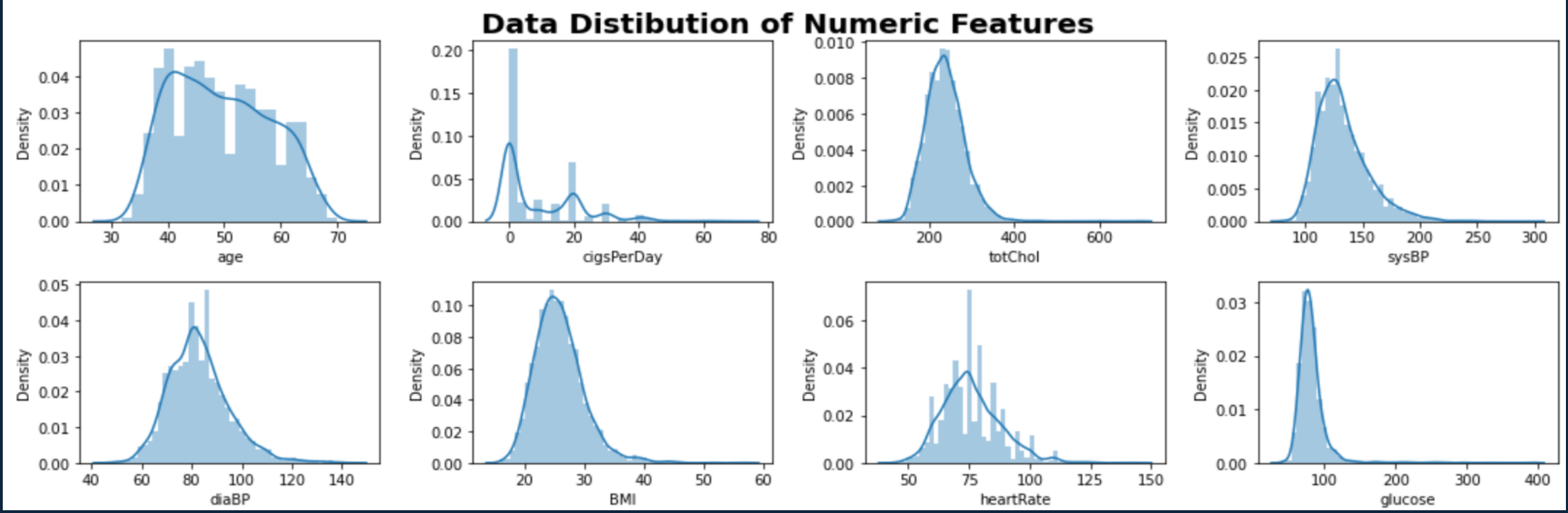
- totChol: total cholesterol level (Continuous)
- sysBP: systolic blood pressure (Continuous)
- diaBP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- heartRate: heart rate (Continuous)
- glucose: glucose level (Continuous)

Predict variable (desired target)

- TenYearCHD: (binary: “1”, means “Yes”, “0” means “No”)

Exploratory Data Analysis

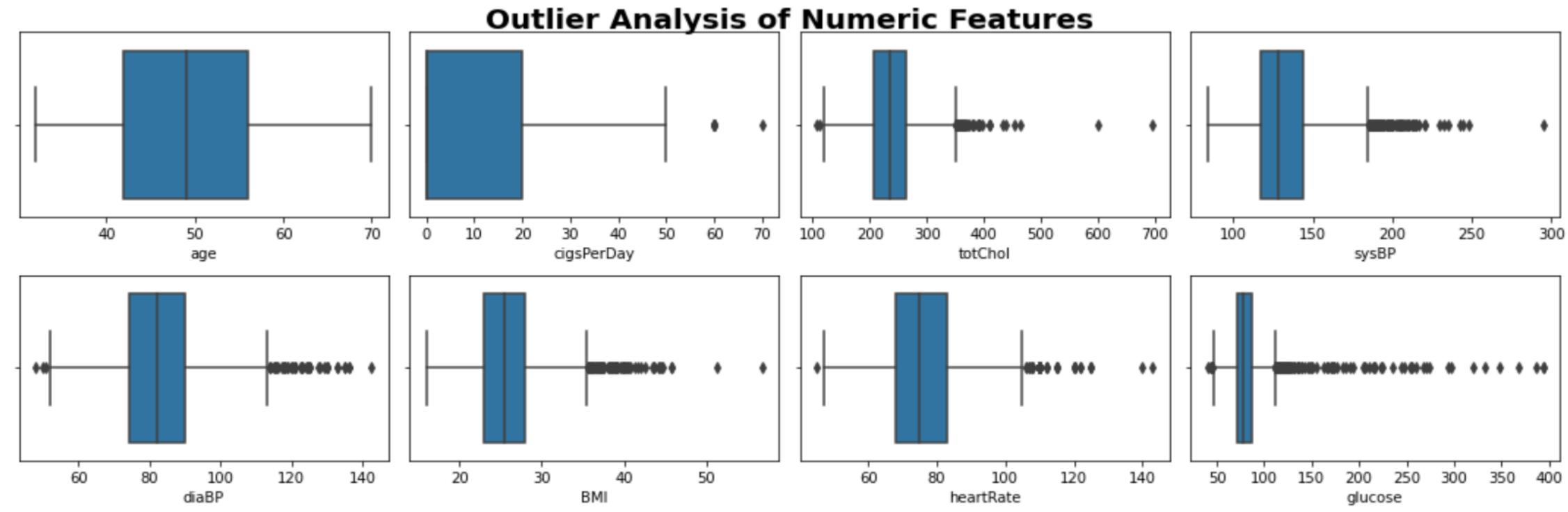
Data Distribution of Numeric Features



For numerical features, we can see that the majority of distributions are right-skewed. The distributions of totChol (total cholesterol) and BMI are roughly comparable. The distribution of glucose is highly skewed to the right. It demonstrates that glucose has many outliers.

Exploratory Data Analysis

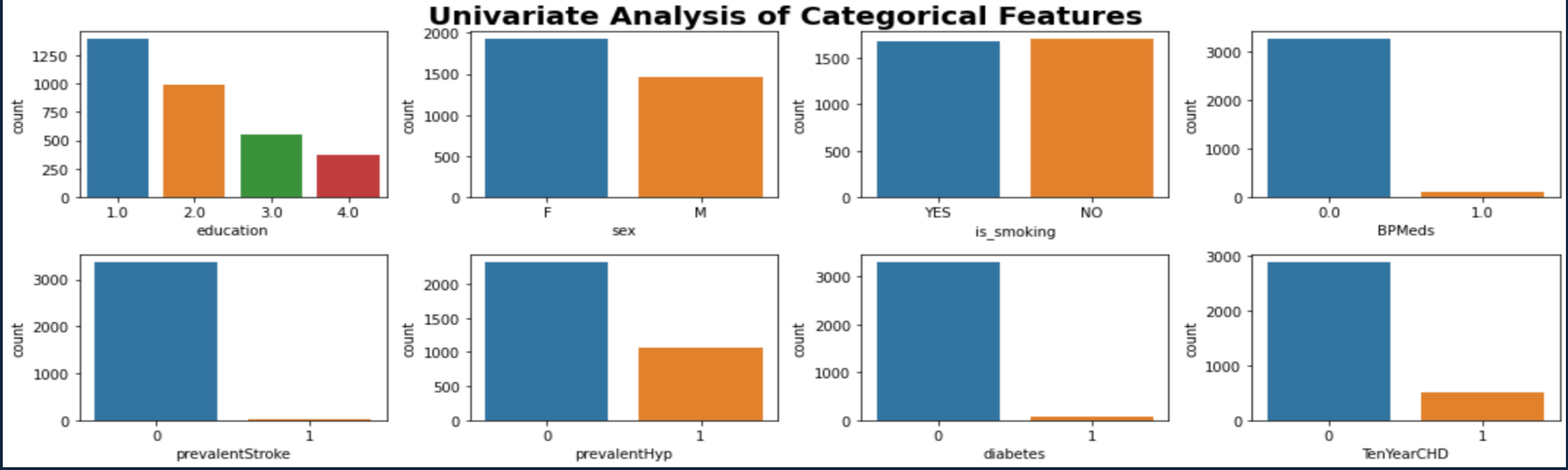
Outlier Analysis of Numeric features



Outliers are visible in the 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', and 'glucose' columns.

Exploratory Data Analysis

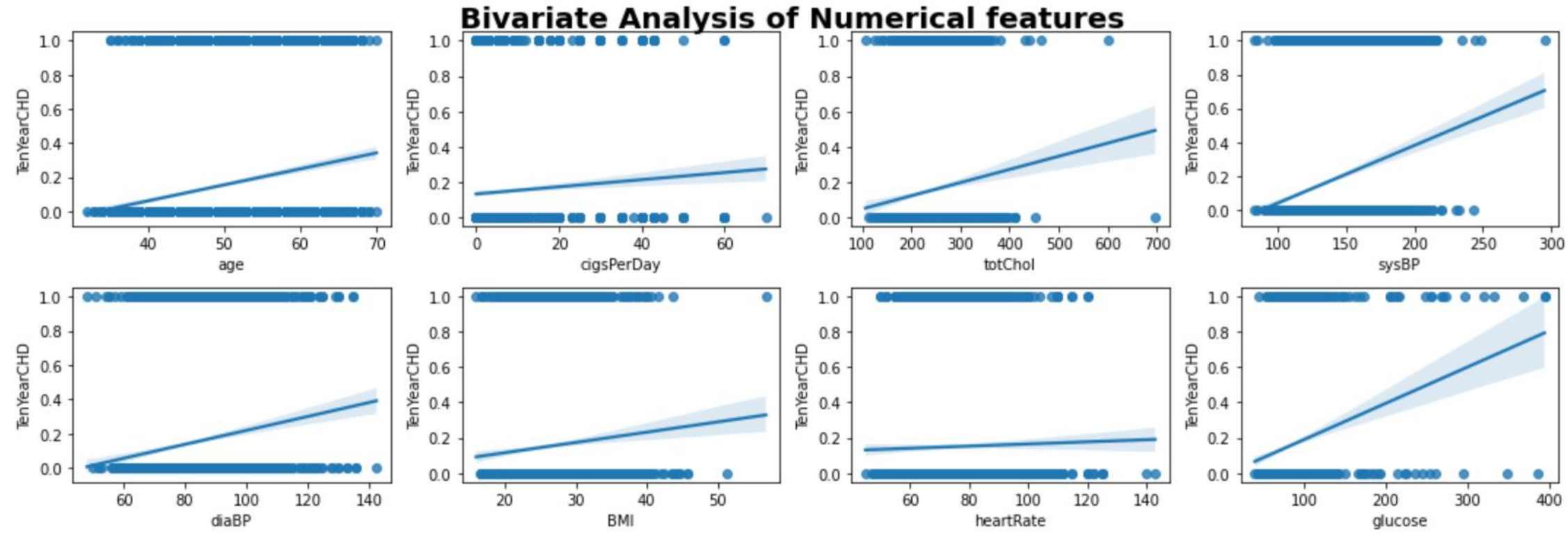
Univariate Analysis of Categorical Features



- In the 'education' column majority of the count falls under category 1, with fewer falling under category 4.
- When compared to male patients, female patients are greater in numbers.
- The proportion of non-smokers and smokers is nearly identical.
- Majority of patients do not take blood pressure medication, do not suffered a stroke, had hypertension, and had Diabetes

Exploratory Data Analysis

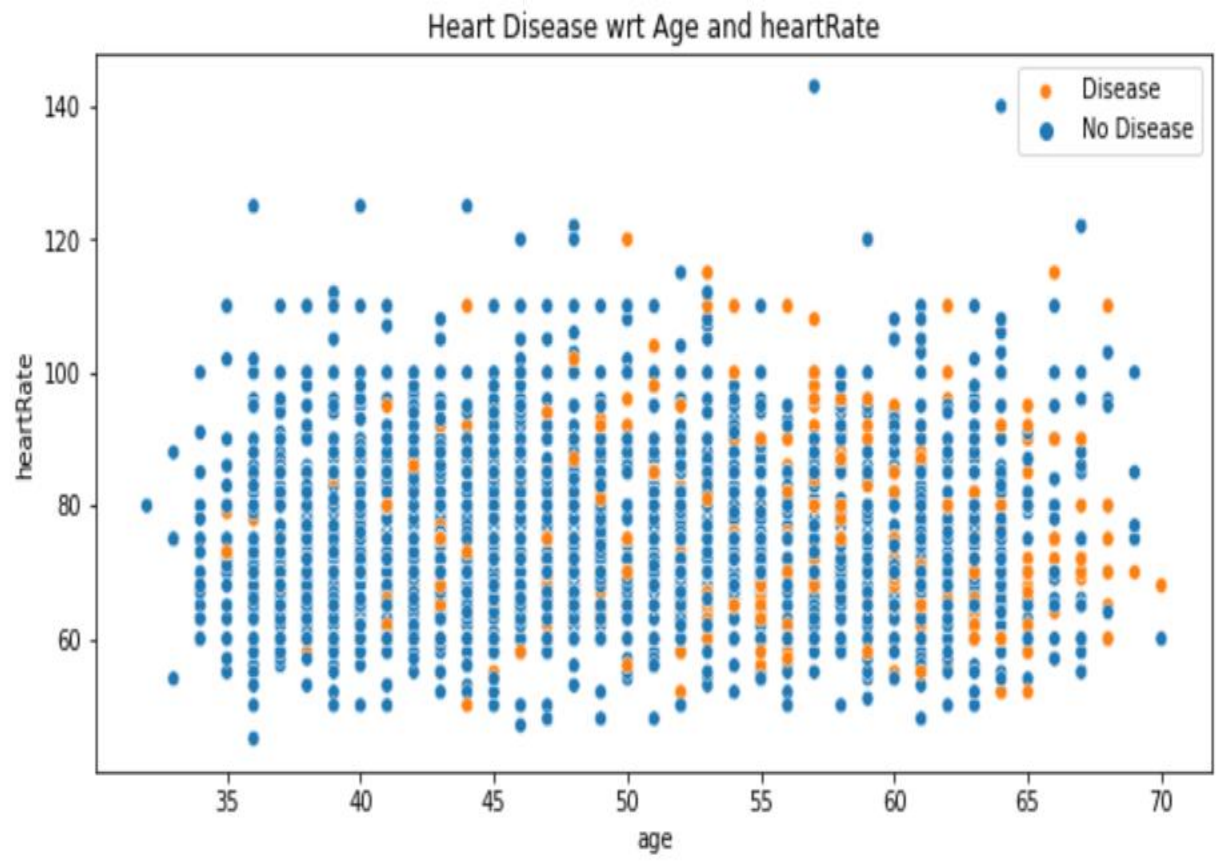
Regression plot between target variable and numerical features



Numerous Independent numerical variables are linked to our Target variable and have a positive relationship with our TenYearCHD.

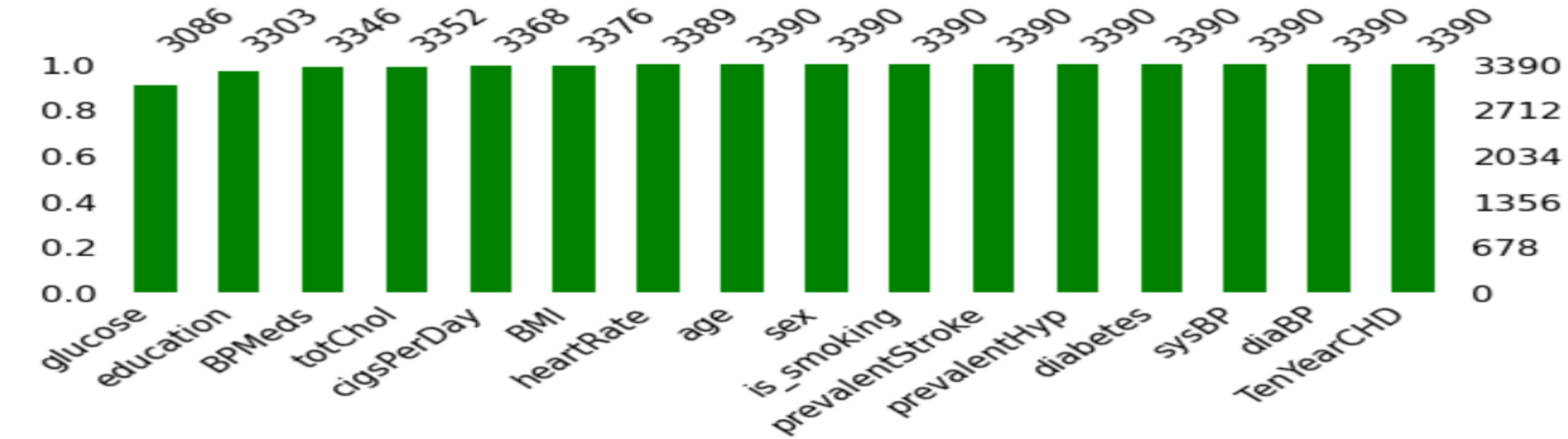
Exploratory Data Analysis

Scatter plot between target varible wrt to age and heartRate



- This is a multi variate analysis between age, TenYearCHD and heartRate.
- There is a clear relation between age and Heart Disease, with the increase in age the chance of heart disease increases. There is no significant relationship between heart disease and heart rate

Data Cleaning



- The goal of data cleaning is to improve the quality of the data and make it suitable for further analysis and modeling.
- Null values are present in the 'glucose', 'education', 'BPMeds', 'totChols', 'cigsPerDay, 'BMI', and 'heartRate' columns.
- Typically, we use other records to replace these null values. However, the entries in this dataset are person-specific.
- If we attempt to impute null values using advanced methods, it may affect the outcome because the values will be incorrect.

Data Cleaning

```
# Missing Values Percentage
round(risk_df.isna().sum()/len(risk_df)*100, 2)
```

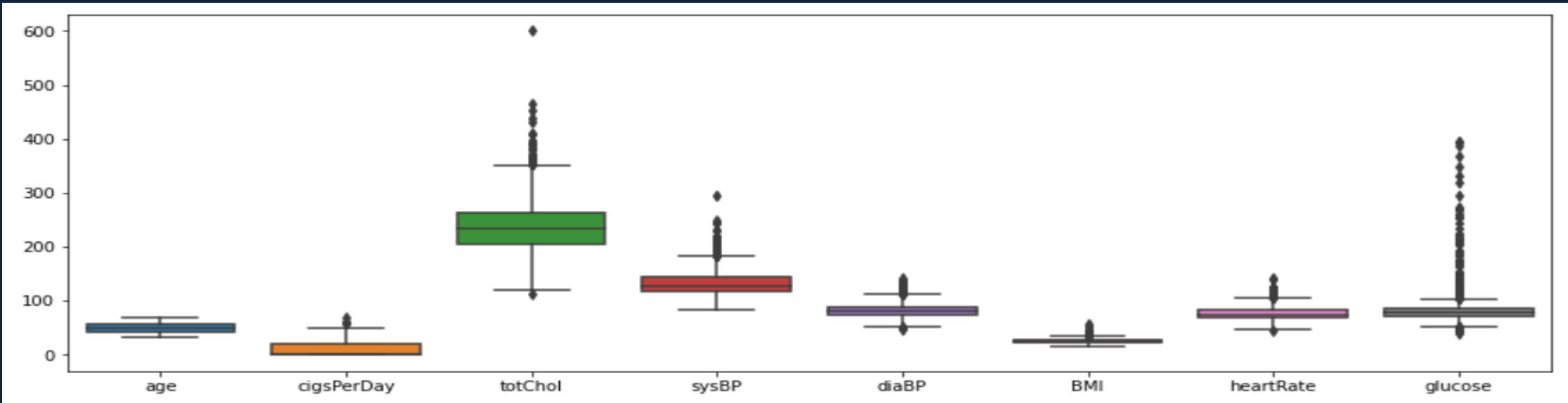
```
age                0.00
education          2.57
sex                0.00
is_smoking         0.00
cigsPerDay         0.65
BPMeds             1.30
prevalentStroke    0.00
prevalentHyp       0.00
diabetes           0.00
totChol            1.12
sysBP              0.00
diaBP              0.00
BMI                0.41
heartRate          0.03
glucose            8.97
TenYearCHD         0.00
dtype: float64
```

Null Values Treatment

In the healthcare industry, every piece of data is crucial. Because of this, we came up with a solution by setting a threshold value. If a feature has less than 5% null values, we decide to drop those rows, and the remaining rows are imputing, which will affect prediction but not significantly

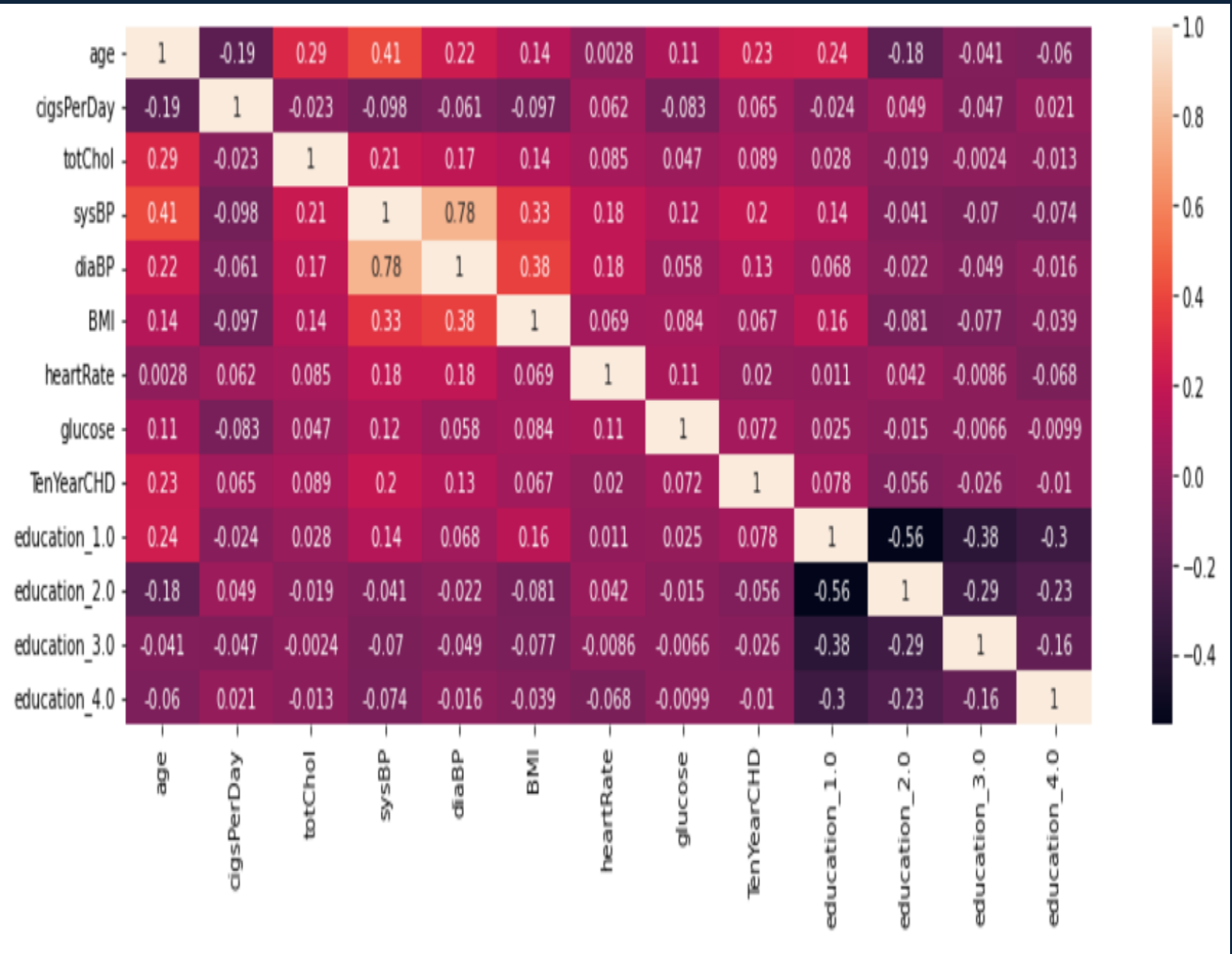
Data Cleaning

Outliers Detection



- Since we have limited datapoint hence we are not simply removing the outlier instead of that we are using the clipping method.
- **Clipping Method:** In this method, we set a cap on our outliers data, which means that if a value is higher than or lower than a certain threshold, all values will be considered outliers. This method replaces values that fall outside of a specified range with either the minimum or maximum value within that range.
- Some of the data were skewed before handling outliers, but after outlier treatment, the features almost follow the normal distribution. Therefore, we need not utilize the numerical feature transformation techniques.

Feature Engineering

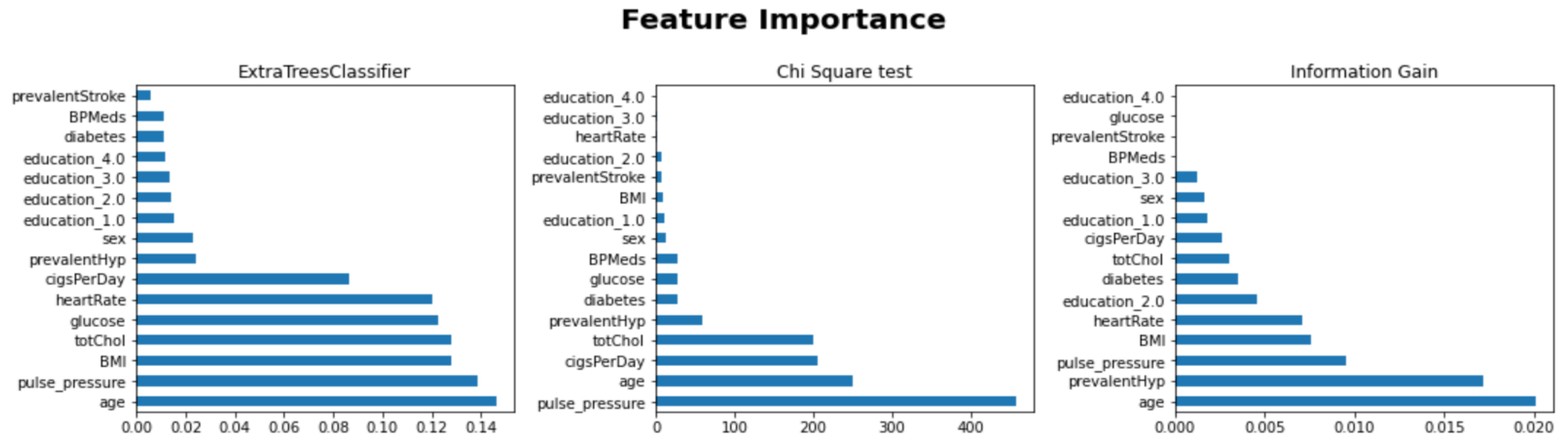


1. **Encoding:** Encoding is a technique in feature engineering that is used to convert categorical variables into numeric values that can be used by machine learning algorithms.
   - Except for the 'sex' and 'is_smoking' columns, almost all of the categories in the dataset are already represented numerically (ordinal). Therefore, we are encoding these two columns.
   - 'education' column is represented in 4 categories from 1 to 4 and We're using algorithms that treat the categorical variables as unordered, such as decision trees and random forests, hence one-hot encoding can be more effective in the 'education' feature.

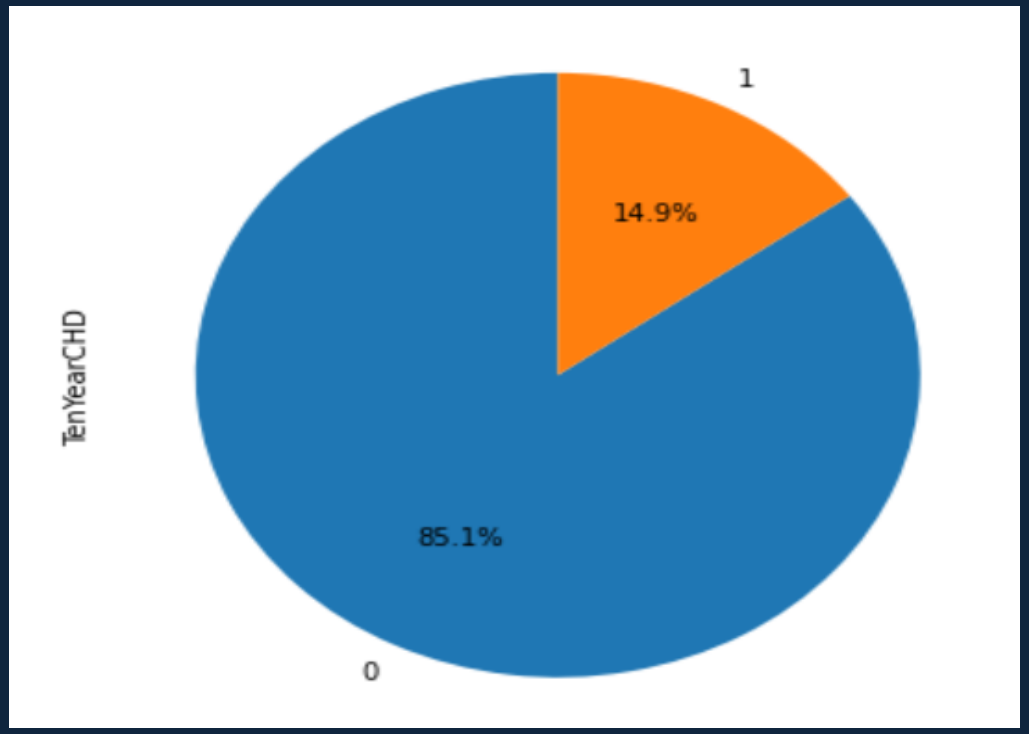
2. **Correlation Matrix:** The correlation coefficient is a numerical measure of the strength and direction of a linear relationship between two variables. age and Pulse Pressure are important features to predict target variable.

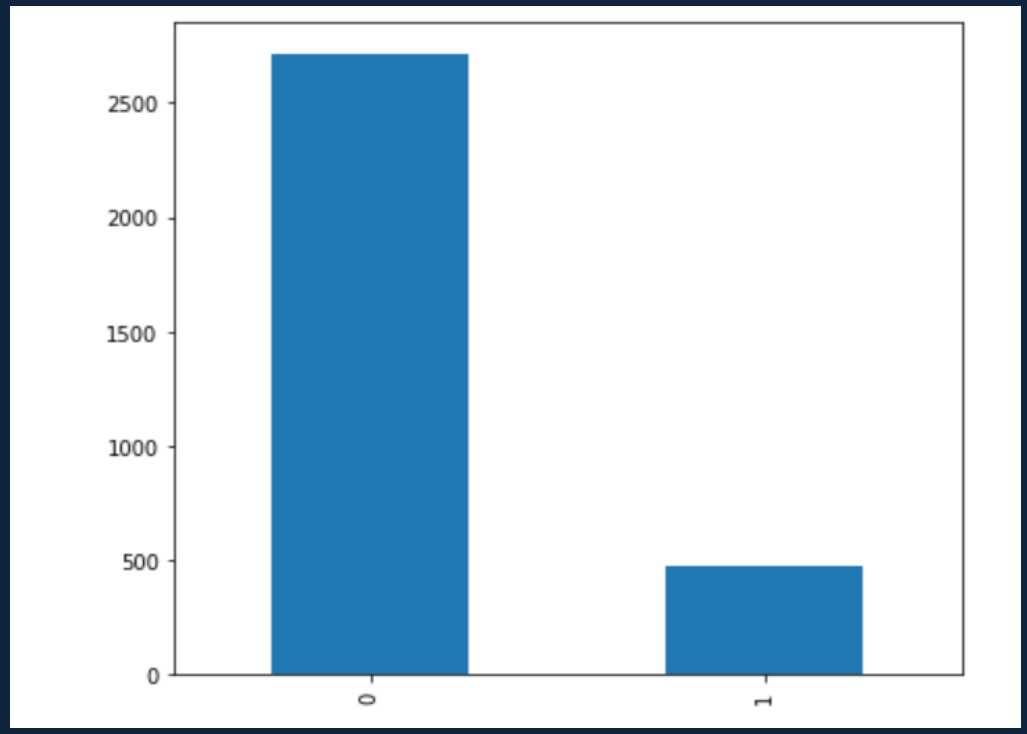
Feature Engineering

3. **Feature Manipulation:** From Correlation Matrix we have found out that sysBP (systolic BP and diaBP (diastolic BP) are highly correlated hence we converted those two columns into one using the subtraction operation.
4. **Feature Selection:** We have used Extra Trees Classifier, Chi-square Test, and Information Gain methods to understand important features yet The entries in this dataset are person-specific, and the values vary between individuals, hence we are using all features, except multi-collinear features, to train the model.



Feature Engineering





5. **Handling Class Imbalance:** When there are significantly more instances of certain classes than others, the issue of class imbalance typically arises. Class imbalance in the target class is a problem for machine learning models because it can result in biased predictions. We have used the Synthetic Minority Oversampling Technique (SMOTE) to resolve this issue.

Model Implementation

	model	train_accuracy	test_accuracy	train_precision	test_precision	train_recall	test_recall	train_f1	test_f1	train_roc_auc	test_roc_auc
0	LogisticRegression	0.801	0.808	0.861	0.883	0.712	0.731	0.779	0.800	0.883	0.892
1	SVM	0.857	0.833	0.938	0.909	0.762	0.757	0.841	0.826	0.939	0.910
2	KNN	1.000	0.895	1.000	0.871	1.000	0.938	1.000	0.903	1.000	0.893
3	DecisionTree	1.000	0.832	1.000	0.835	1.000	0.847	1.000	0.841	1.000	0.832
4	RandomForest	0.959	0.894	0.992	0.945	0.925	0.847	0.957	0.893	0.997	0.955
5	AdaBoost	0.865	0.855	0.920	0.924	0.796	0.787	0.854	0.850	0.929	0.919
6	XGBoost	1.000	0.903	1.000	0.941	1.000	0.870	1.000	0.904	1.000	0.951
7	LightGBM	1.000	0.907	1.000	0.955	1.000	0.863	1.000	0.907	1.000	0.949

- After all these steps, here we are to build and implement the model. We did split the features using **train-test split** and scaled the data using the **standard scaler**.
- Then finally cleaned and scaled data was sent to 8 different models. We used multiple metrics like accuracy score, recall score, f1-score, confusion matrix, and auc-roc to measure the performance of our model.
- Since recall score is the most important evaluation metric in the medical domain, we are selecting the final model as **KNN** because it has the **Highest Recall score**, and we don't want to mispredict a person.

AI

Conclusion:

In this project, we tackled a classification problem in which we had to classify and predict the 10-year risk of future coronary heart disease (CHD) for patients. The goal of the project was to develop a tool for the early detection and prevention of CHD, addressing a significant public health concern using machine learning techniques.

- There were approximately 3390 records and 16 attributes in the dataset.
- We started by importing the dataset, and necessary libraries and conducted exploratory data analysis (EDA) to get a clear insight into each feature by separating the dataset into numeric and categoric features.
- After that, the outliers and null values were removed from the raw data.
- In feature engineering we transformed raw data into a more useful and informative form, by creating new features, encoding, and understanding important features. We handled target class imbalance using SMOTE.
- Then finally cleaned and scaled data was sent to 8 various models, the metrics were made to evaluate the model, and we tuned the hyperparameters to make sure the right parameters were being passed to the model.
- We are, however, focusing more on the Recall score and F1 score because we are dealing with healthcare data and our data is unbalanced.

AI

Conclusion:

- With an f1-score of 0.907 and a recall score of 0.863 on test data, we have noticed that LightGBM Classifier outperforms most of the performance metrics.
- Our highest recall score, 0.938%, came from KNN.
- The XGBoost and RandomForestClassifier tree-based algorithms also provided the best approach to achieving our goal. We were successful in achieving a respective f1-score of 0.904 and 0.893.
- We can select the Final model as our KNN classifier due to its highest recall score. It is acceptable to classify a healthy individual as having a 10-year risk of coronary heart disease CHD (false positive) and to follow up with additional medical tests; however, it is categorically unacceptable to miss identifying a particular patient or to classify a particular patient as healthy (false negative).
- This is not the end with the assistance of subject matter experts, we can engineer a large number of variables that could increase the accuracy of our prediction, resulting in the development of a better model.

Thank you

Navneet Keshri

linkedin.com/in/navneet-keshri-28650918b