Capstone Project – 3
Classification
"Cardiovascular Risk Prediction”

Presented By :
Mr. Hasnain Mazhar Rizvi

Points of Discussion

1. Problem Statement
2. Data Description
3. Data Preparation and Cleaning
4. EDA (Exploratory Data Analysis)
5. Hypothesis Testing
6. Feature Engineering
7. Model Implementation
8. Model Interpretation
9. Conclusion

1. Problem Statement

- **Cardiovascular diseases (CVDs) are the major cause of mortality worldwide**.
- 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 16 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

2. Data Description

- There are a total of **16 feature columns where 'TenYearCHD'** is the dependent variable column. The total number of observations(rows) are **3390.**
- There are **no duplicate rows** in the dataset.
- Also there are missing values in the columns **education, cigs per day, BP meds, totChol, BMI, heart rate and glucose.**
- **Sum of all missing Values are 510.**

2. Data Description

Fields	Description
Sex	gender
Age	age
education	The level of education of the patient
is_smoking	Whether smoking currently or not
Cigs_Per_Day	Cigarettes smoked per day
BP_Meds	Whether taking BP meds or not
Prevalent Stroke	If the patient has a history of stroke
Prevalent hyp	If the patient has a history of hypertension
Diabetes	Patient has diabetes or not
Tot Chol	Cholesterol measure
Sys BP	BP measure
Dia BP	diastolic BP measure
BMI	Body Mass Index
Heart Rate	Heart Rate measure
glucose	glucose level
TenYearCHD	10-year risk of coronary heart disease CHD

2. 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)
- Education: The level of education of the patient (categorical values - 1,2,3,4)
- **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)

2. Data Description

- **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, yet are considered continuous because of large number of possible values.)
- Glucose: glucose level (Continuous)
- **Predict variable (desired target):**
- TenYearCHD: 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”)

3. Data Preparation and Cleaning

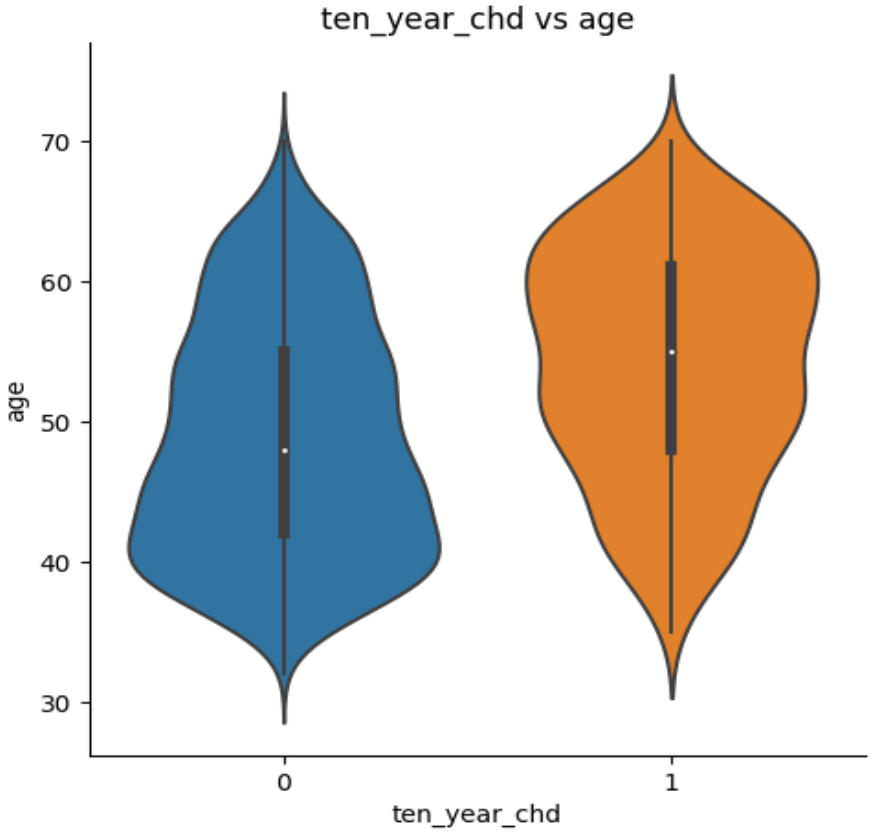
- There are **no duplicate rows** in the dataset.
- There are **missing values** in the columns **education**, **cigs per day**, **BP meds**, **totChol**, **BMI**, **heart rate** and **glucose**.
- **Changed the names** of all the **columns** for ease of use.
- I have also **defined** the **continuous variables**, **dependent variable** and **categorical variables** for ease of plotting graphs.

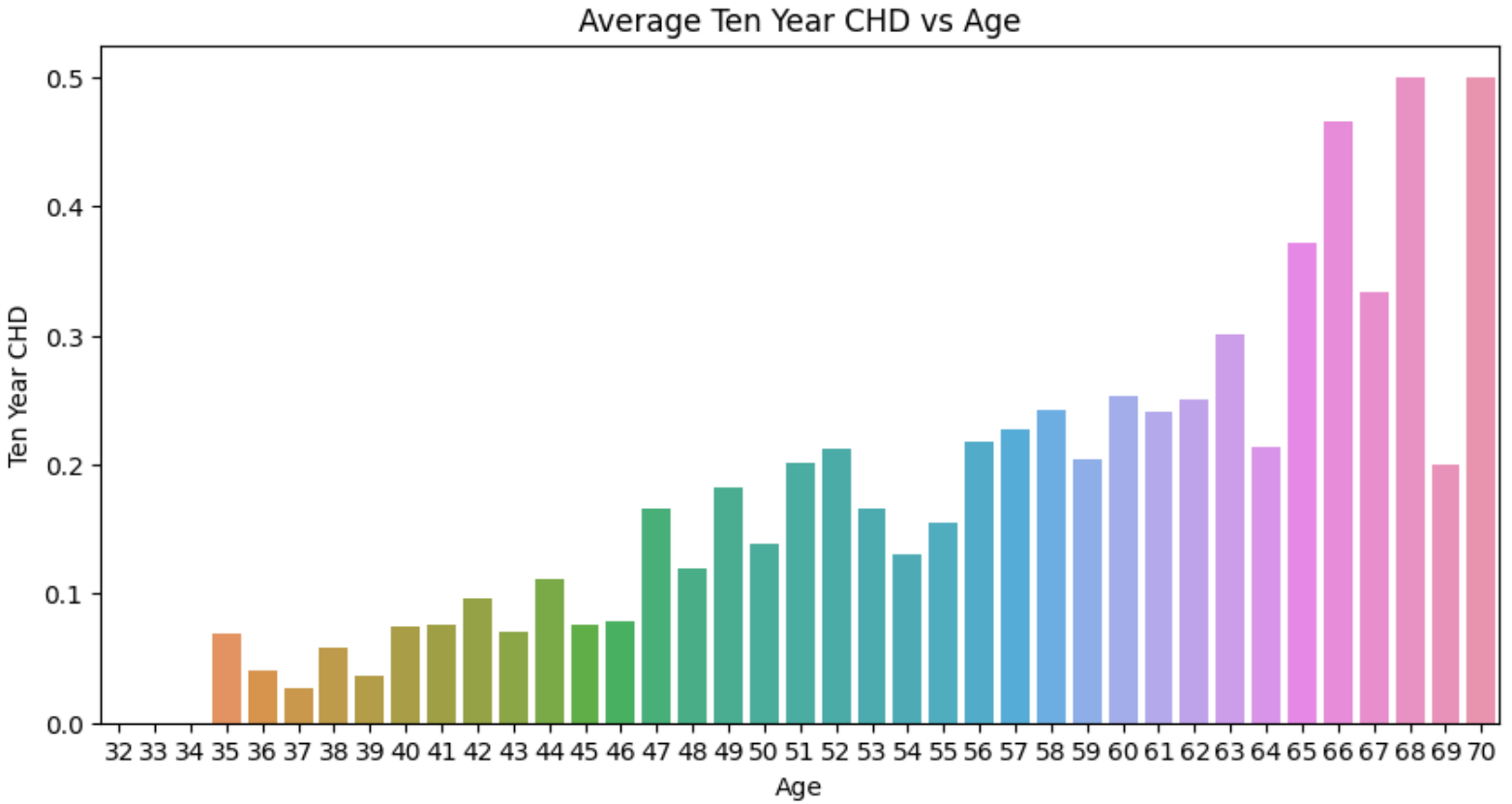
4. EDA (Exploratory Data Analysis)

Ten Year CHD by Age

CHD probability is **high** for **above 65+ aged** peoples.

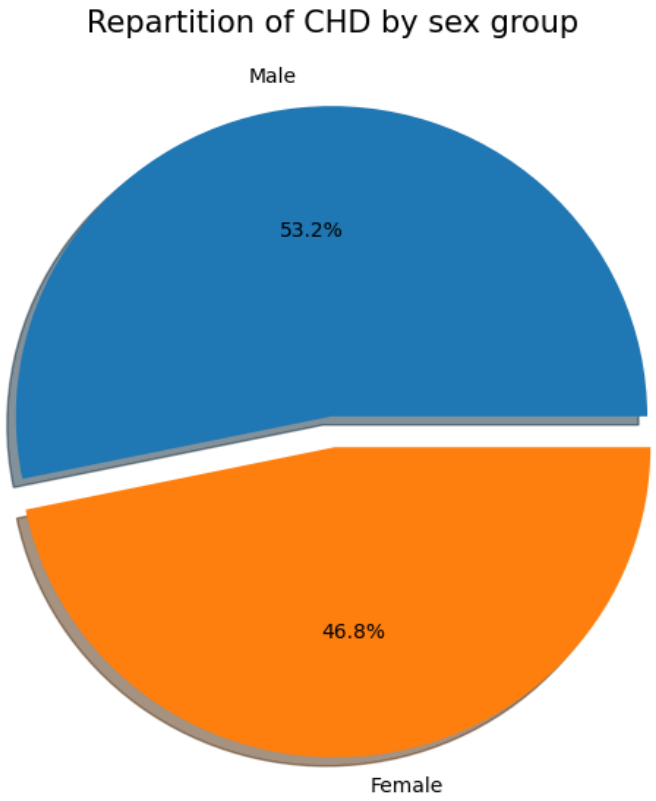
So, **older people** have a **higher risk** of having **coronary heart disease** in next **10 years**.

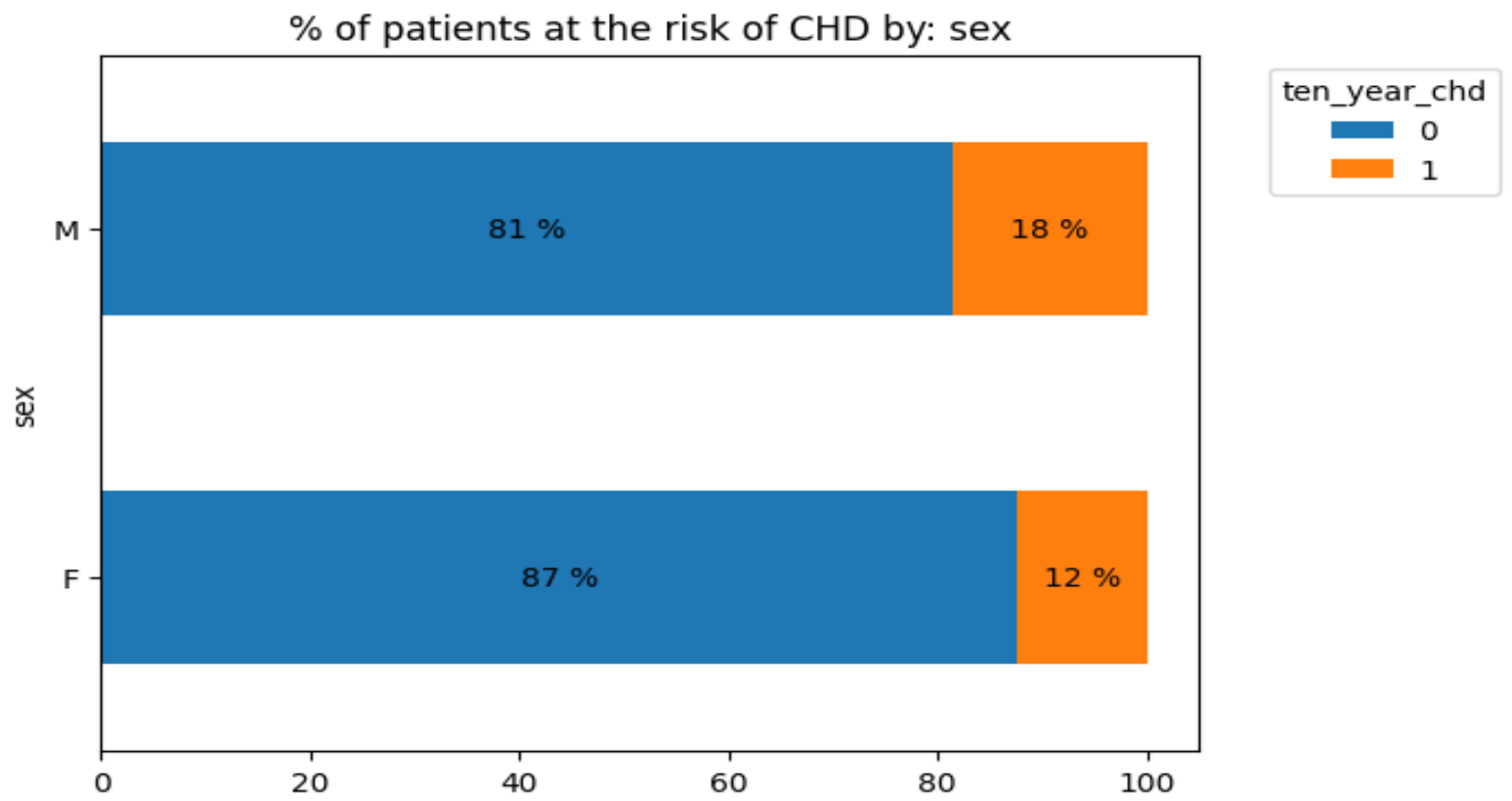




Ten Year CHD by Sex

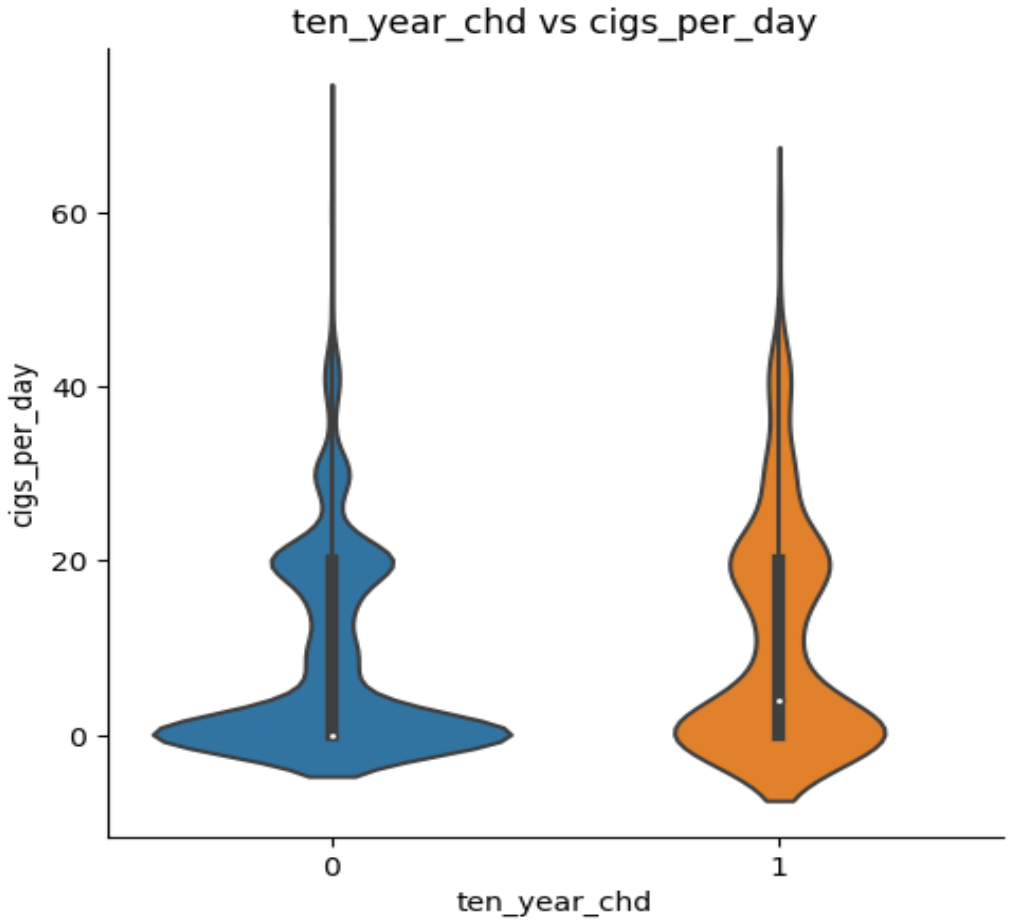
- The **gender distribution** is not even with high count for **females**. **53.2%** are there for **males** and **46.8%** for **females**.
- **Men** are generally at a **higher risk** of **having coronary heart disease**.

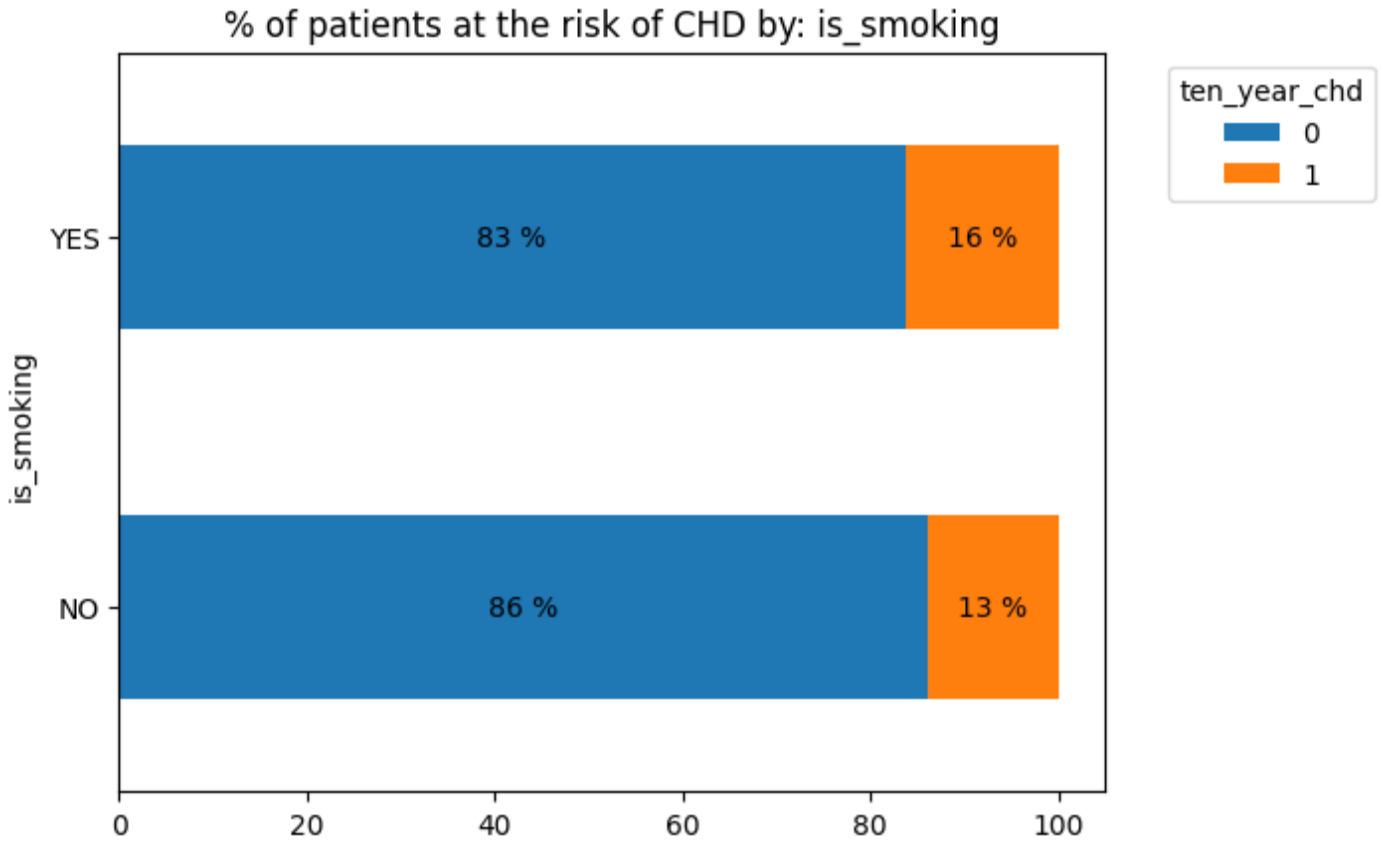




Ten Year CHD by Smoking

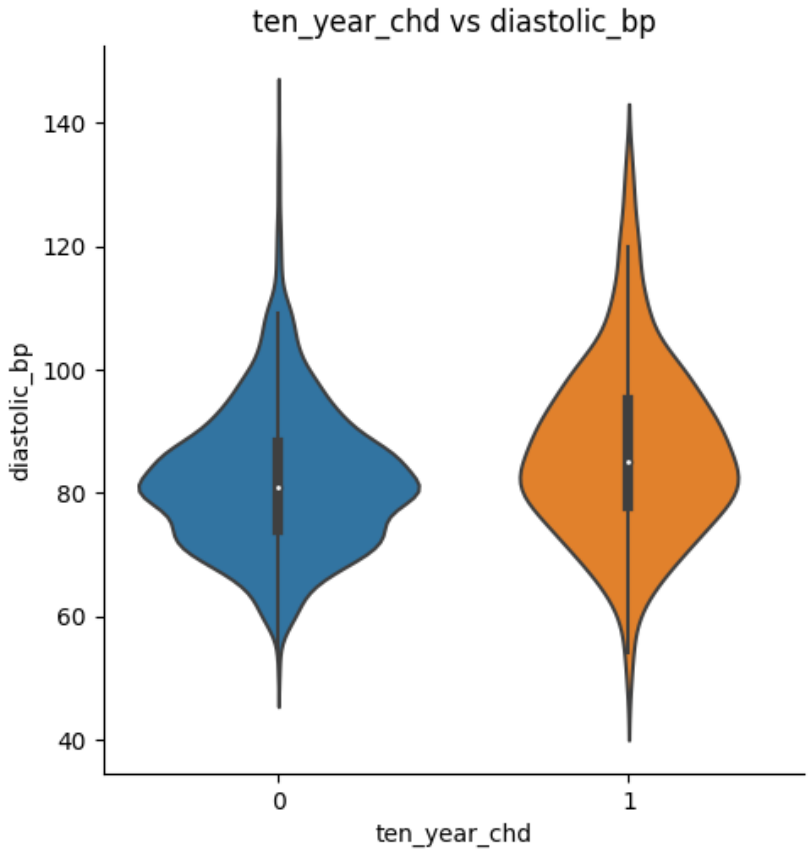
- The **negative cases** are **more** for the **non smokers** compared to the positive cases for non smokers.
- Statistically, **10 year risk of CHD** is **not dependent** on **smoking** with a 95% confidence.

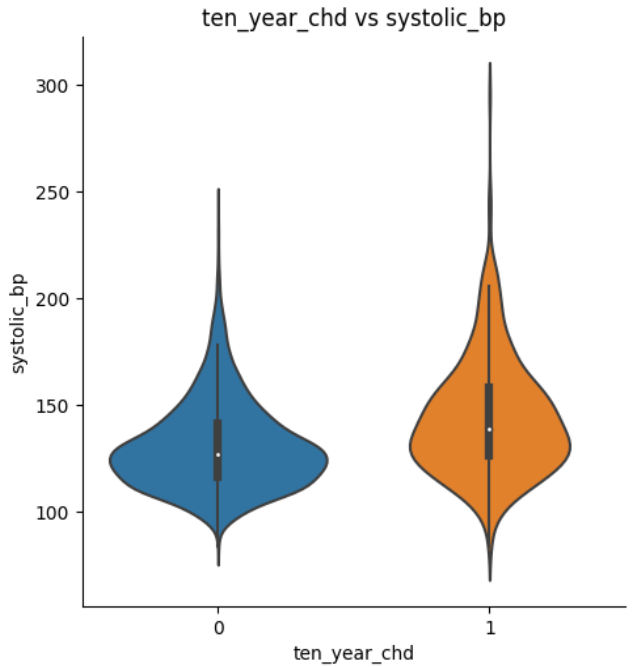


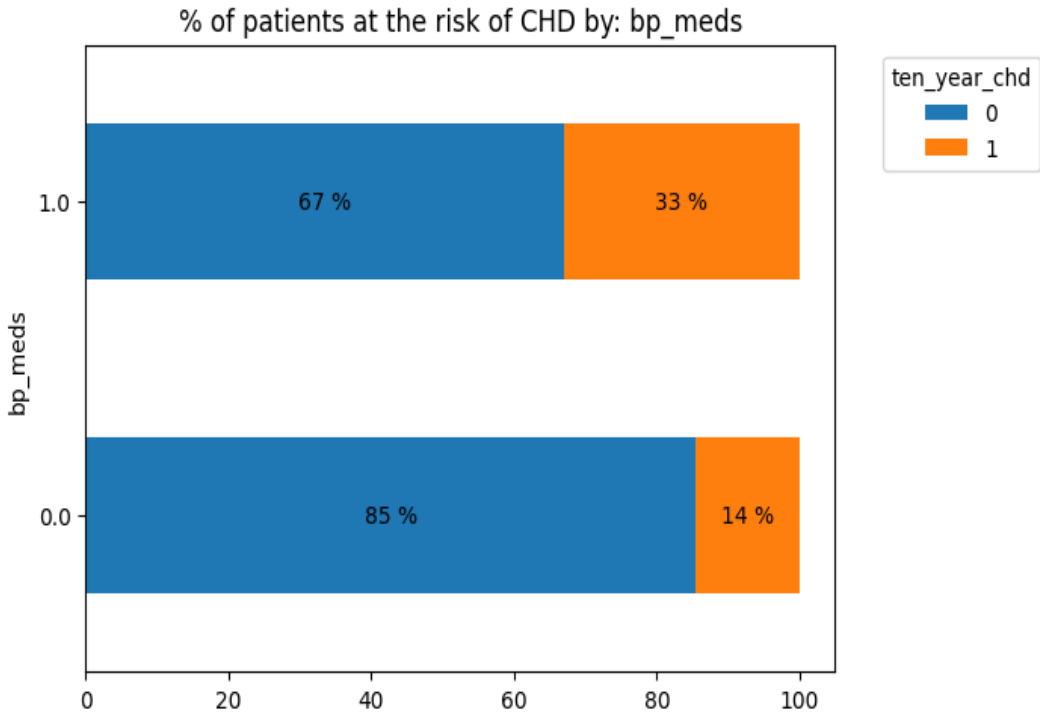


Other Notable Observations

- **Patients** who have **high blood pressure** and have been **taking BP medication** have comparatively **higher risk of CHD**.

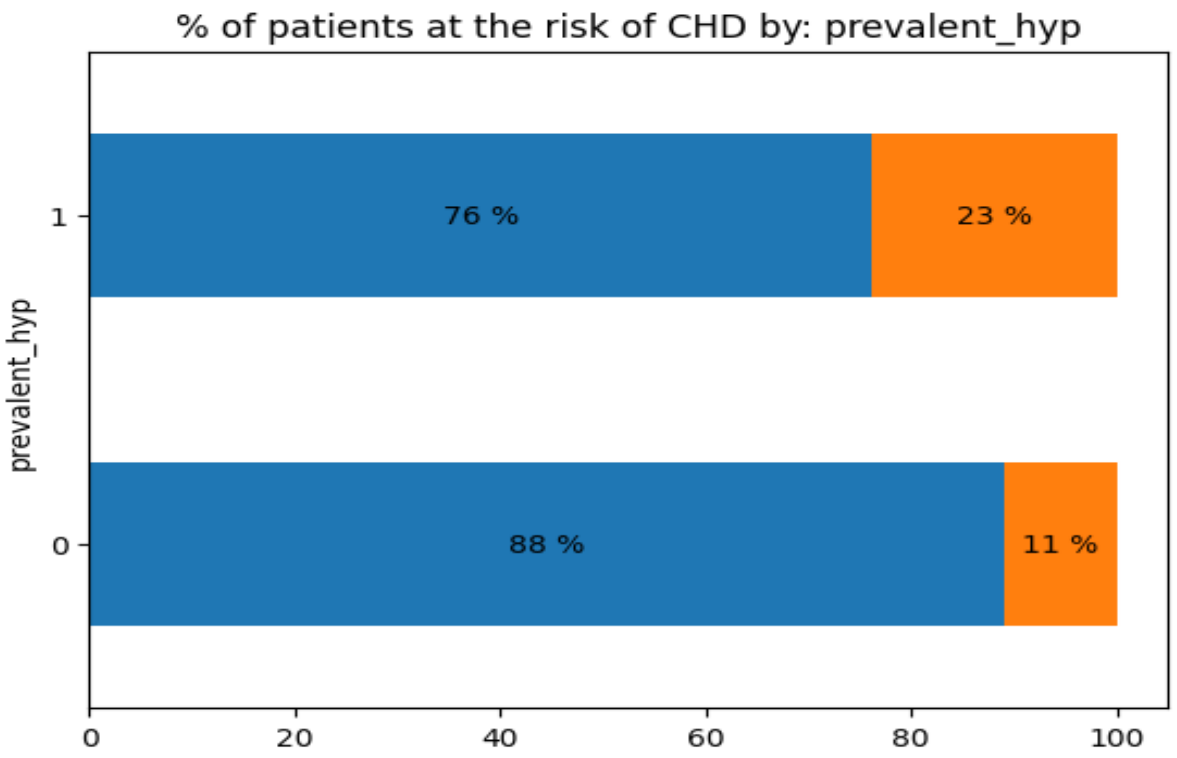


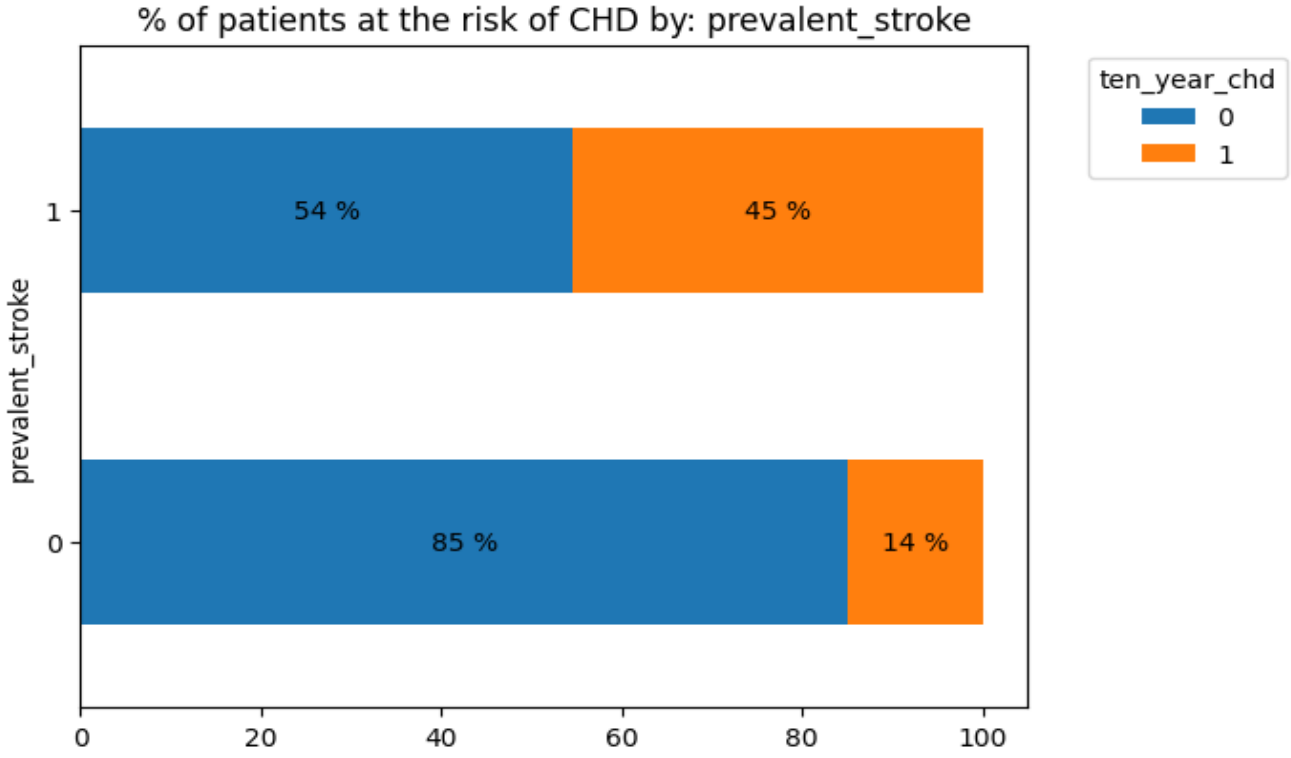




Other Notable Observations

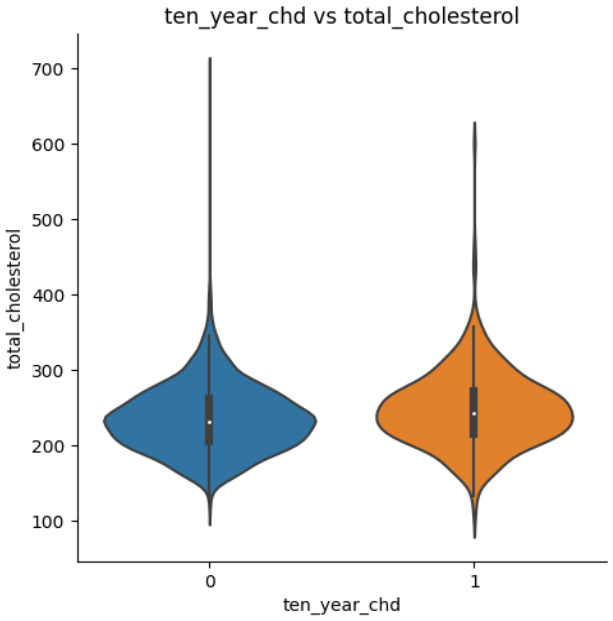
- **Patients** who have a **history of hypertension** and had a **stroke previously** have comparatively **higher risk of CHD**.

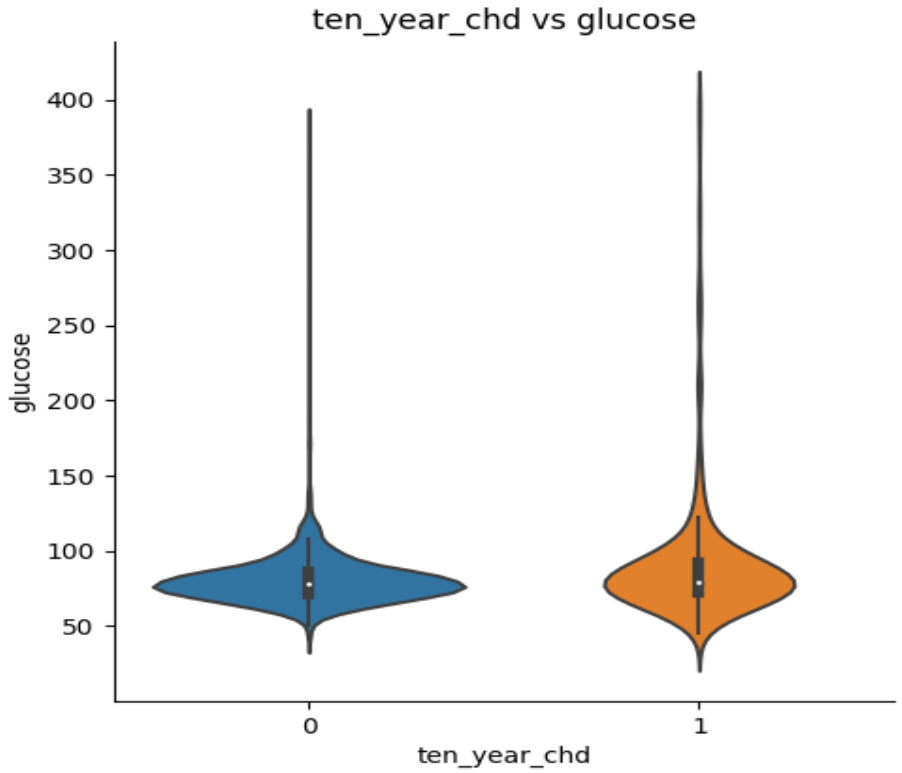


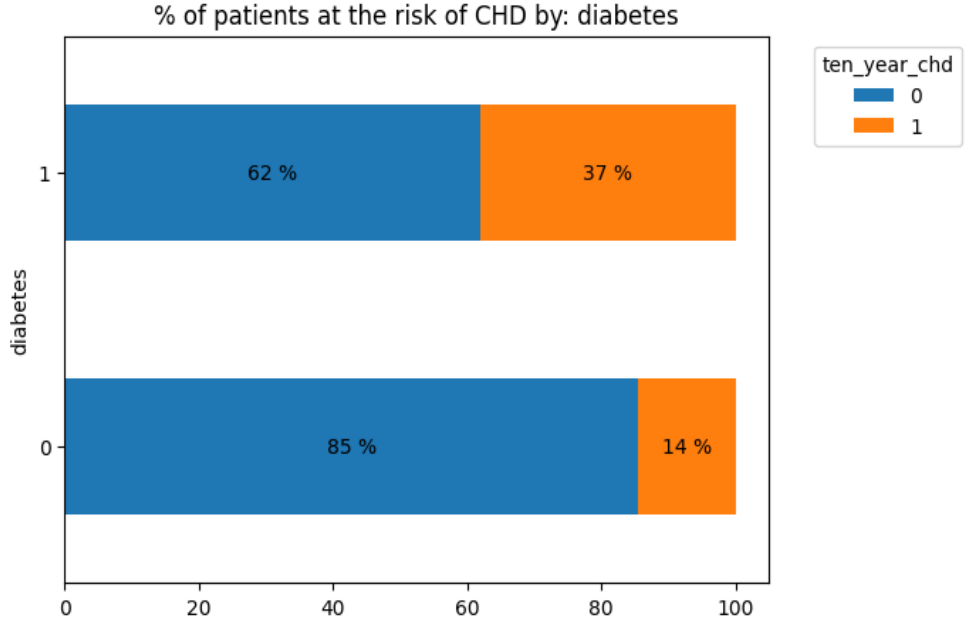


Other Notable Observations

- Similarly**, patients** with **high cholesterol** and **glucose level** (with **diabetes**) have **higher risk** of having **CHD**.

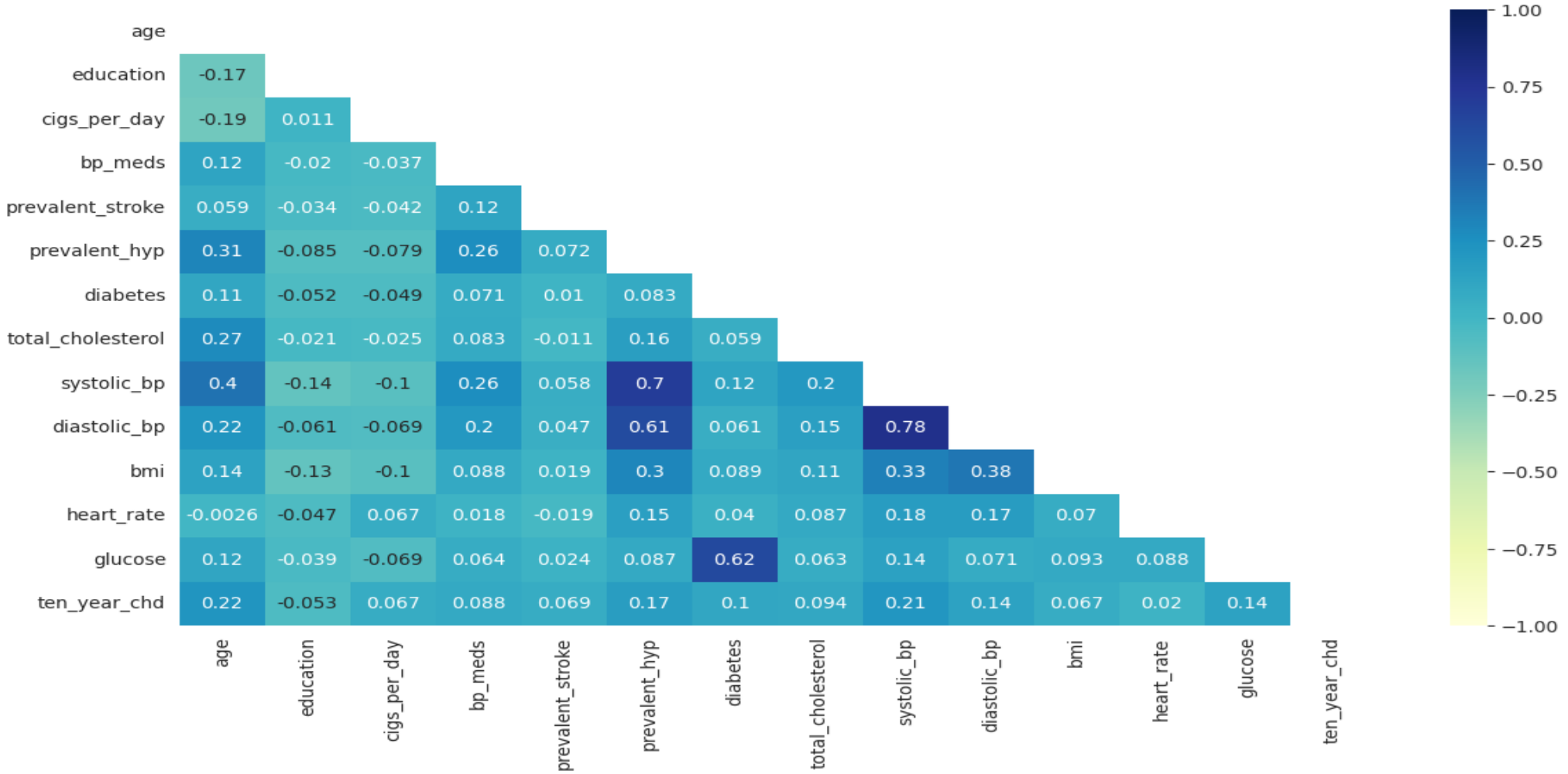






Correlation of features

- There is a **significant correlation** between **systolic BP** and **prevalent hypertension**.
- Similarly **diastolic BP** and **systolic BP** are **highly correlated**. Also **glucose level** and **diabetes** are **correlated**.



Hypothesis Testing

Null hypothesis: There is no association between education level and CHD outcome.

Alternate hypothesis: There is an association between education level and CHD outcome.

- I choose the **chi-squared test** of independence **to test** the **hypothesis** that the **‘education’** column **does not impact** the **outcome of chronic heart disease (CHD)**.
- In this case, both **education level** and **CHD outcome** are **categorical variables**.

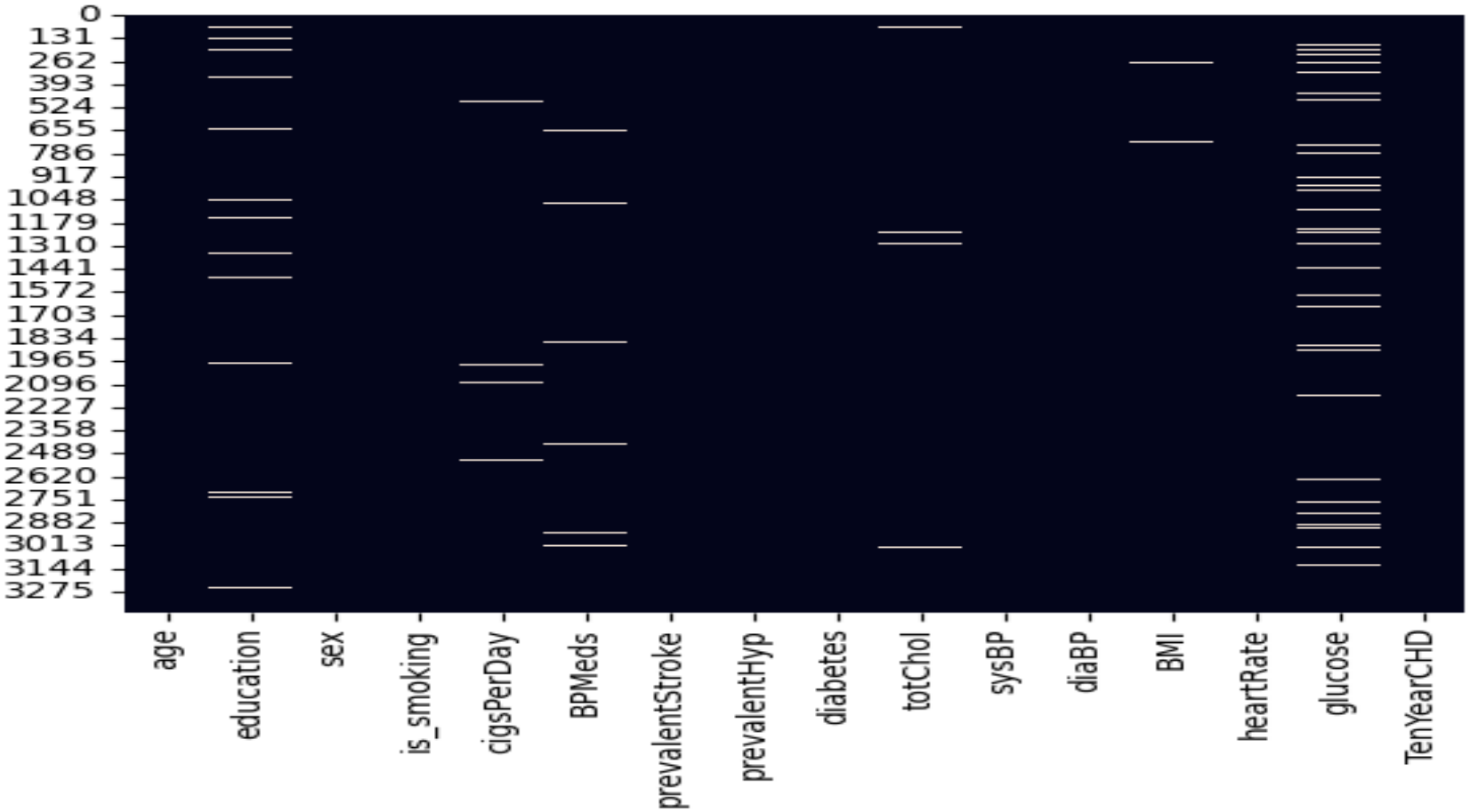
ten_year_chd	0	1
education		
1.0	1135	256
2.0	872	118
3.0	479	70
4.0	319	54

p-value: 6.038646749234552e-05

- The **p-value** is significantly **lower than 0.05** so we **reject** the **null hypothesis**.

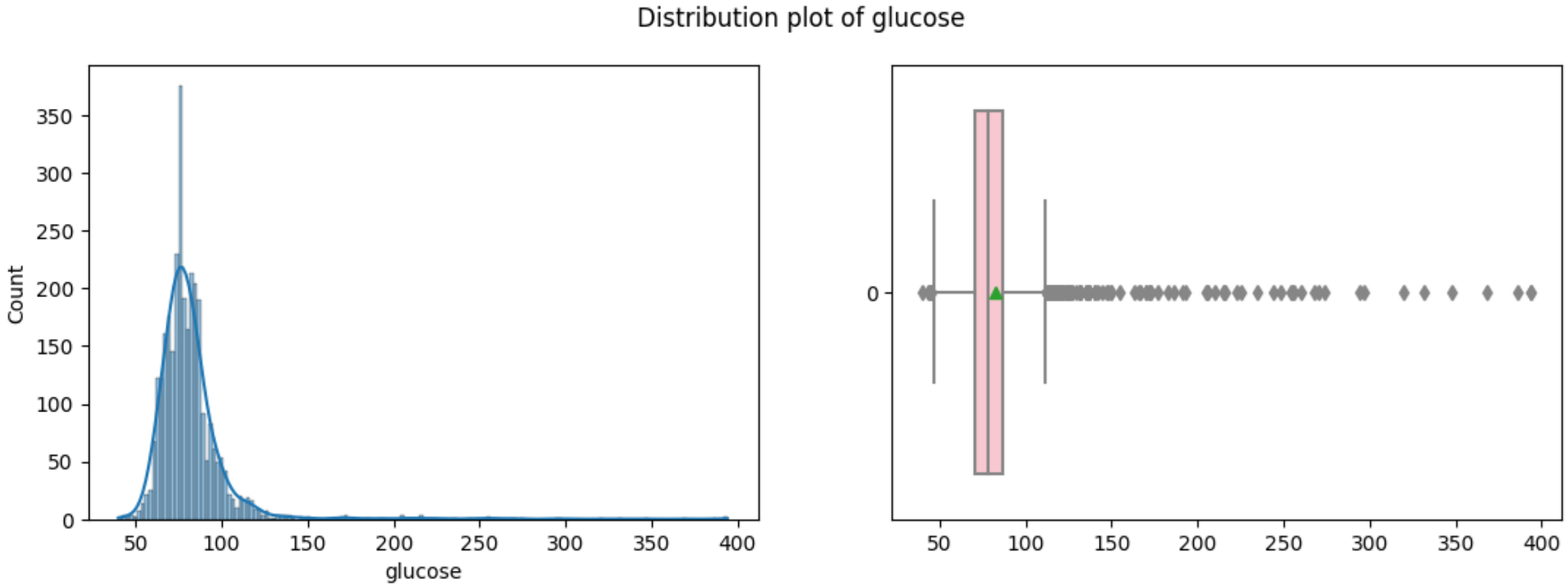
Feature Engineering

- **We Encounter some Missing Values so we have to handle it first.** To fill up the absence of data in our **categorical variables** i have used **simple imputer** that **imputes** the **null values** with feature label that is **most frequent** in the **feature column**.
- In **continuous variables**, i have used **KNN imputer** which uses a **unsupervised clustering algorithm** to come up with values of the features.



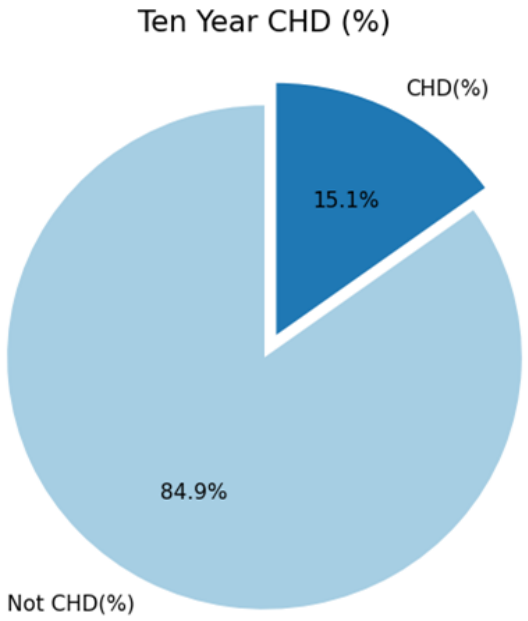
Handling Outliers

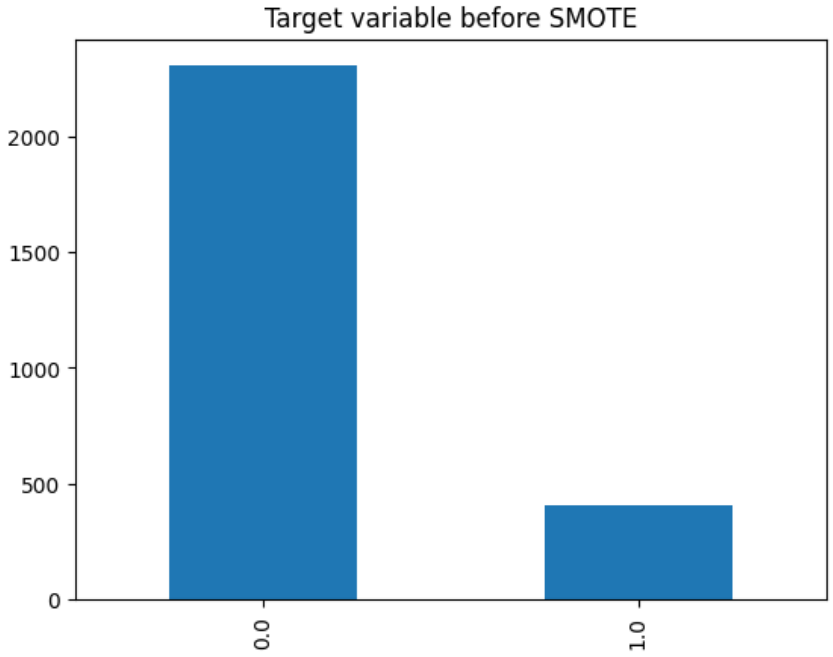
- Used the **Interquartile Range (IQR)** method to **identify** and **remove outliers** in the **continuous columns** (**systolic_bp**, **diastolic_bp**, **total cholesterol**, **glucose** etc.) of the dataset.

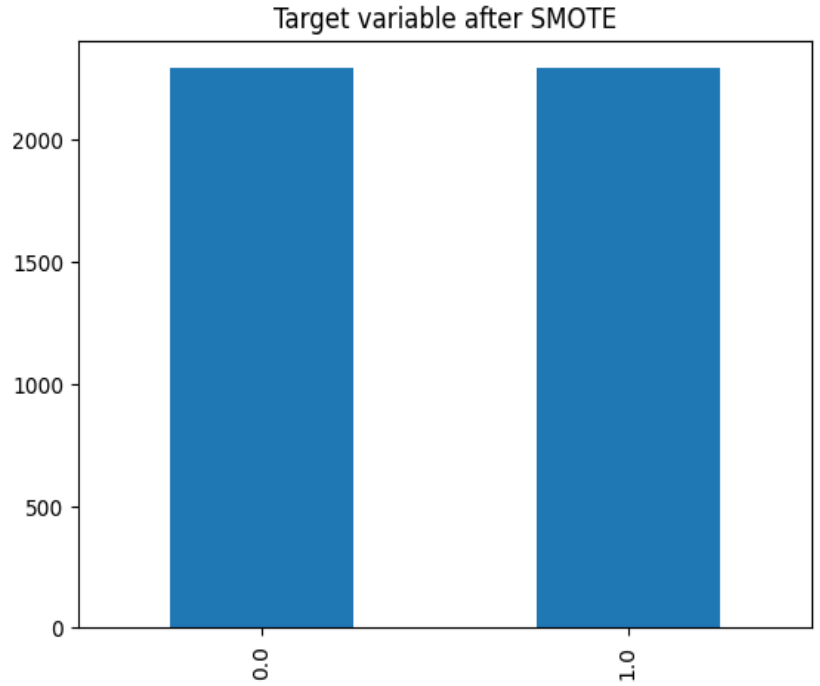


Feature Engineering

- **Handling Imbalanced Dataset**
- **After splitting data into train and test sets with ratio 80:20, i have used SMOTETomek links to handle the imbalanced dataset.**
- **By combining oversampling of the minority class with undersampling of the majority class, I was able to achieve a balanced dataset, where train set of size 4586 with 2712 samples of each of the class.**





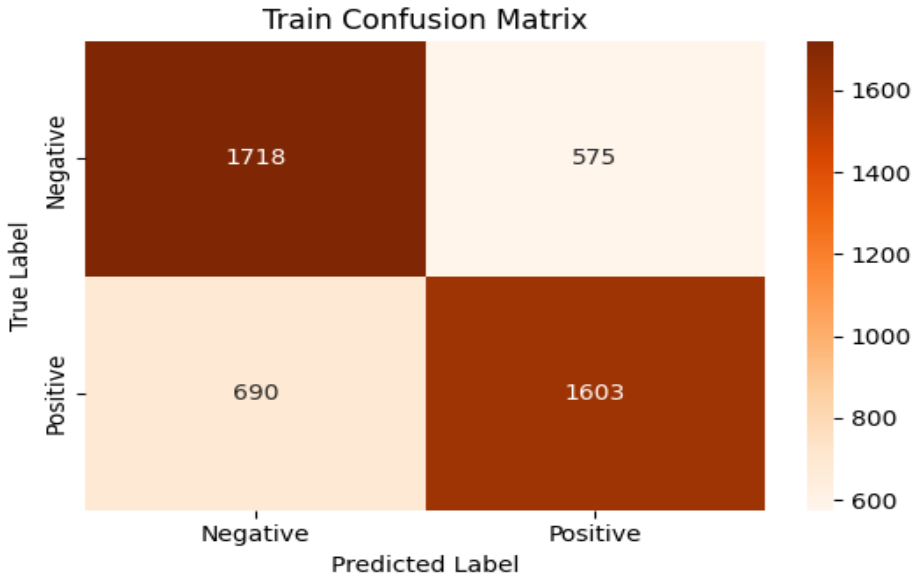


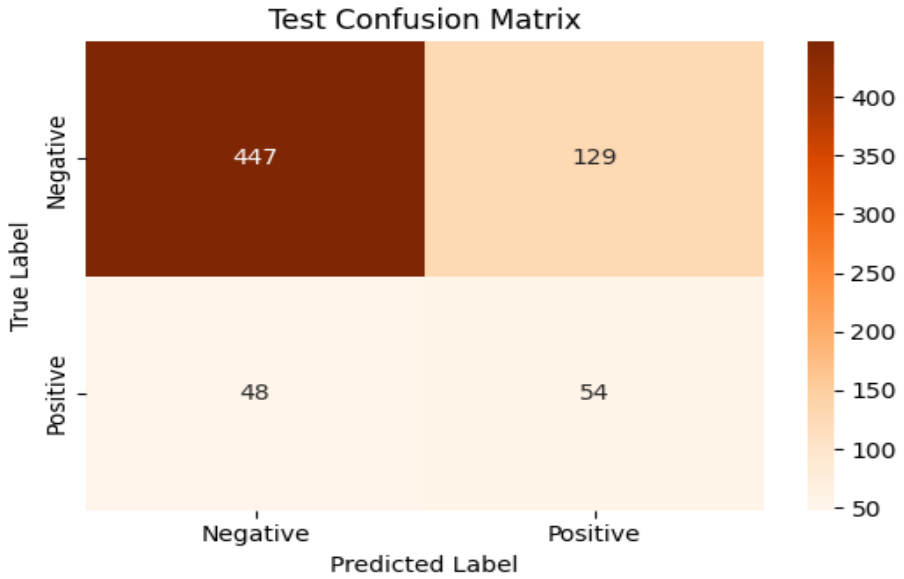
Model Implementation

- **Since we're trying to predict continuous variable, I trained various classification algorithms along with hyperparameter tuning and cross validation to get the best model.**

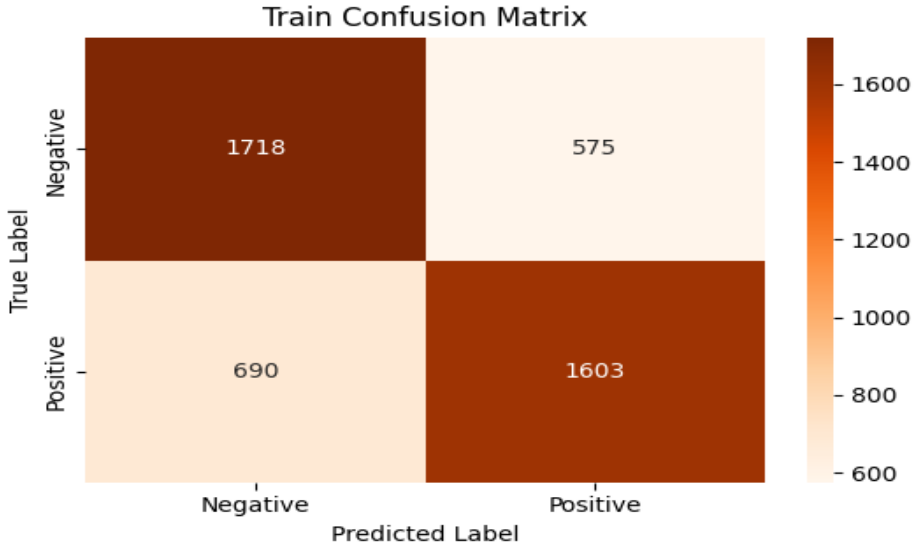
1) **Logistic Regression**
2) **Decision Tree**
3) **Random Forest**
4) **Support Vector Machine**
5) **Xtreme Gradient Boosting**
6) **Naive Bayes**
7) **Neural Network**

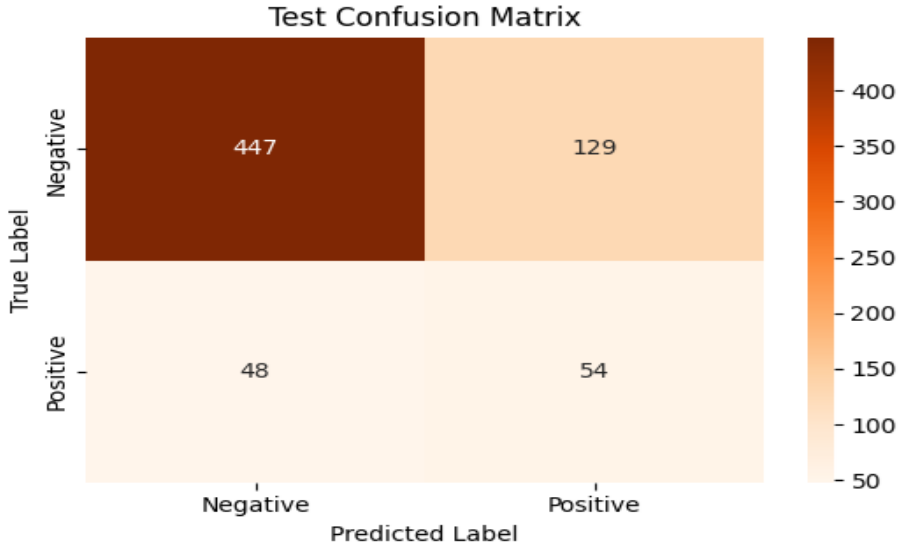
1. Logistic Regression



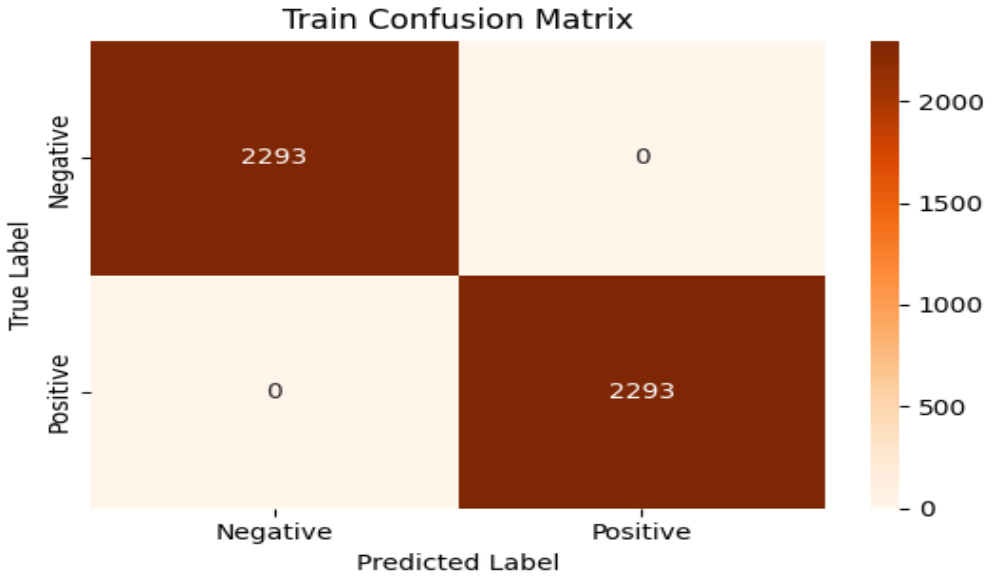


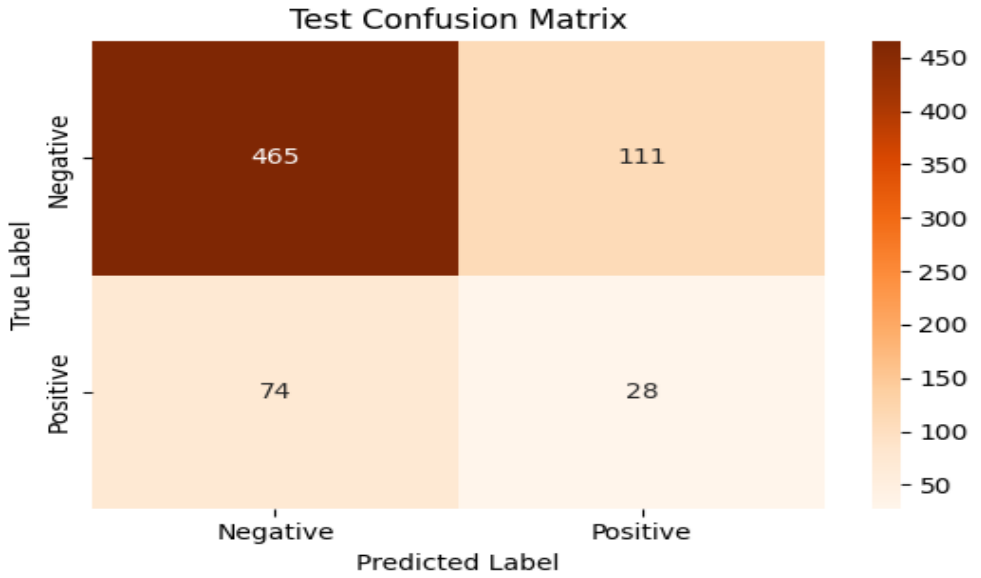
- **After Tuned**



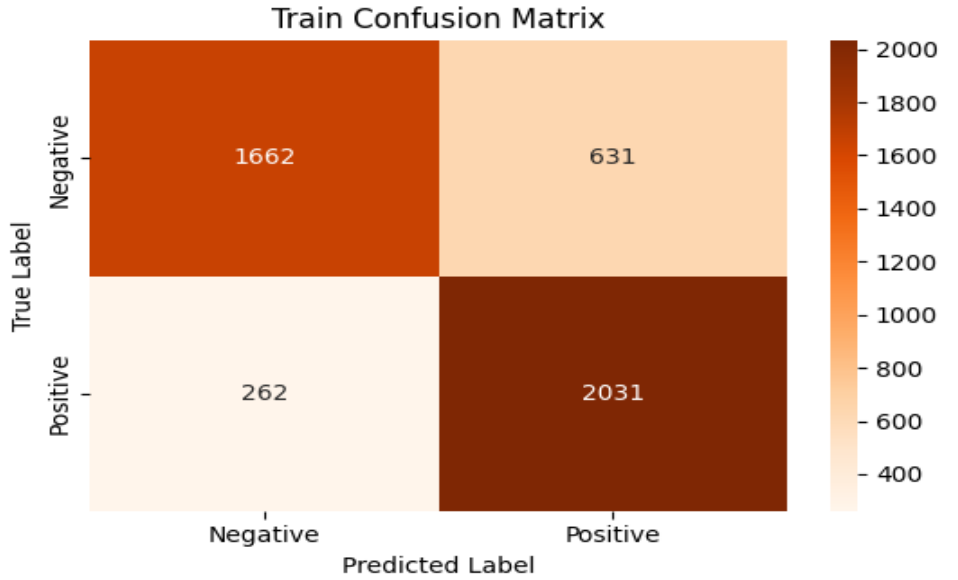


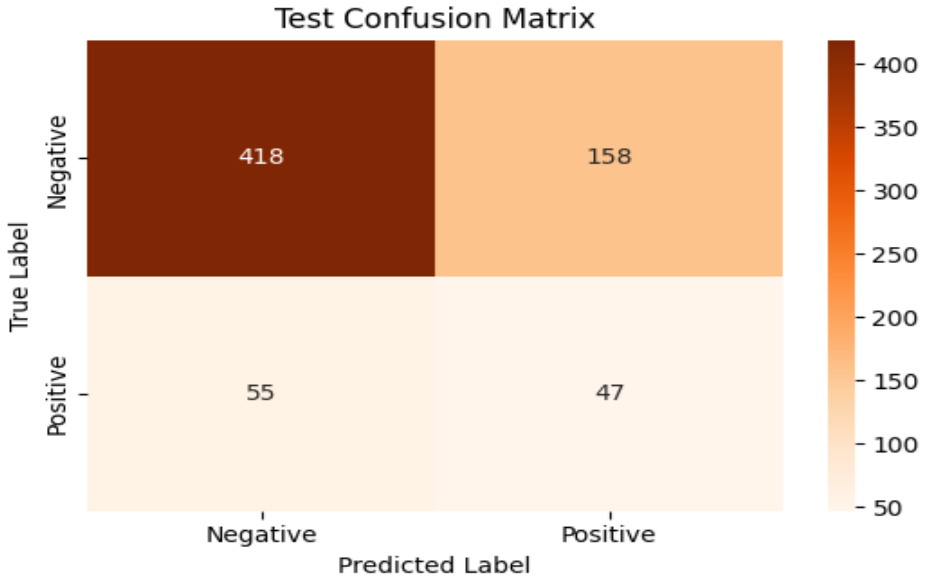
2. Decision Tree



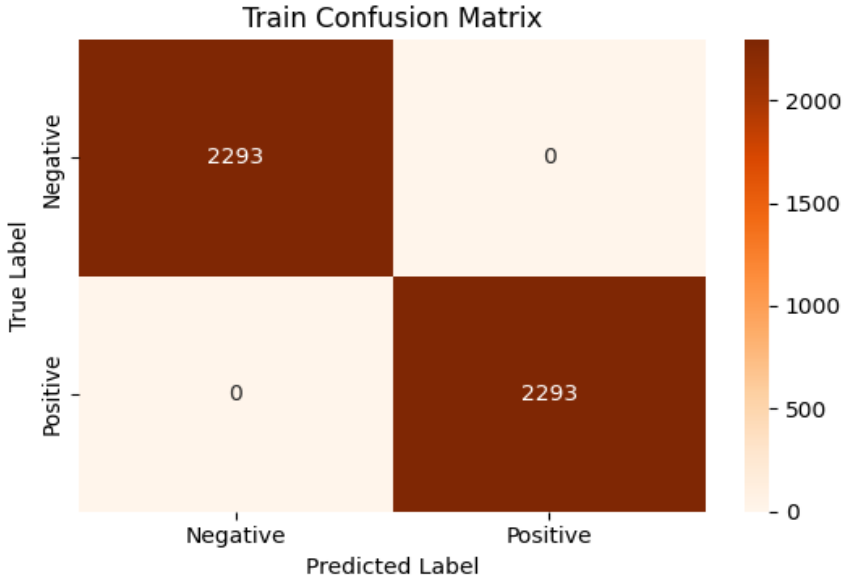


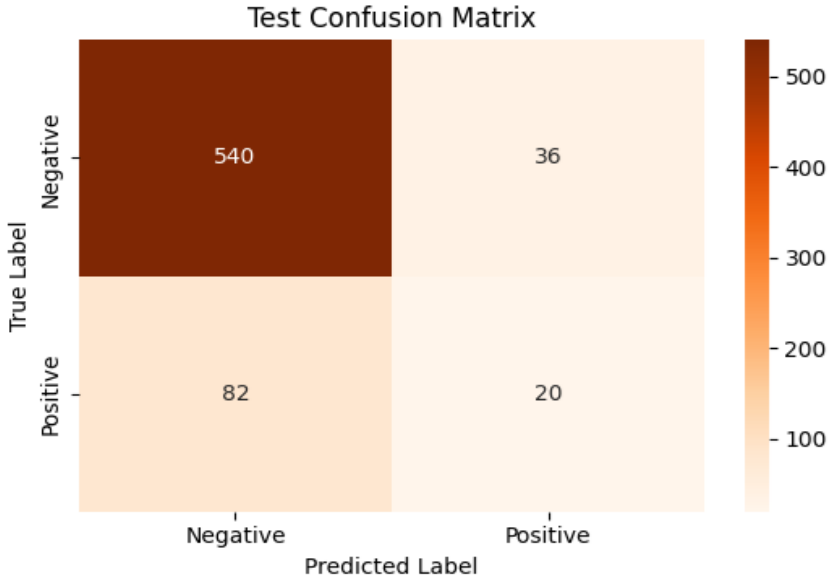
❖ After Tuned



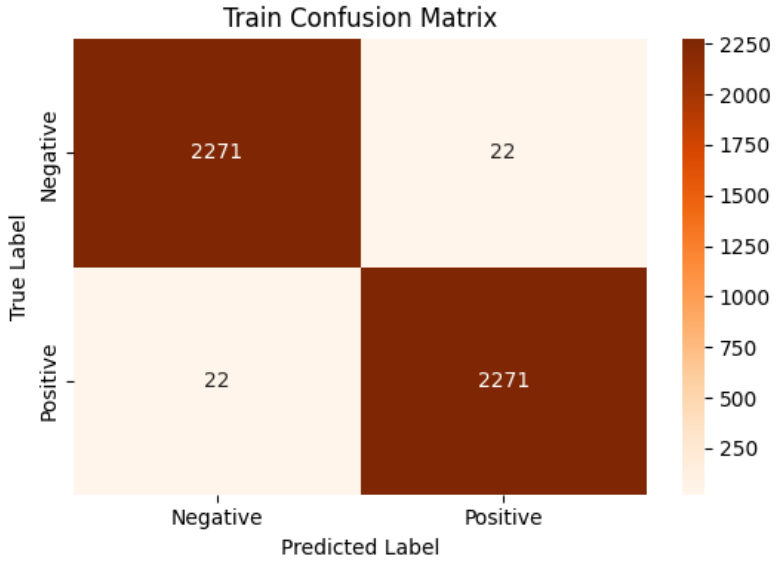


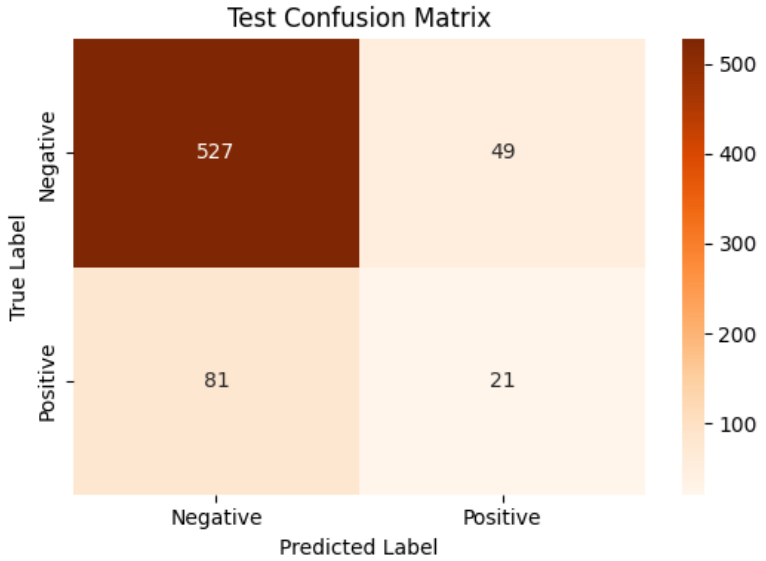
3. Random Forest



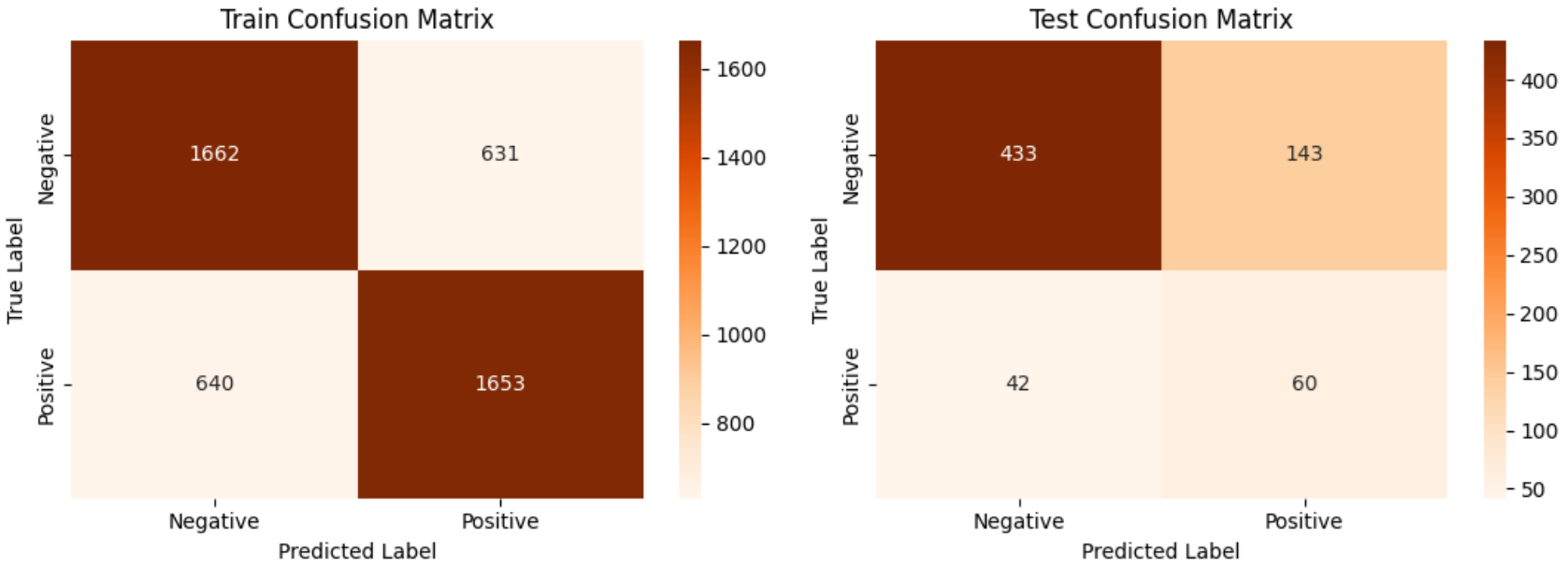


❖ After Tuned

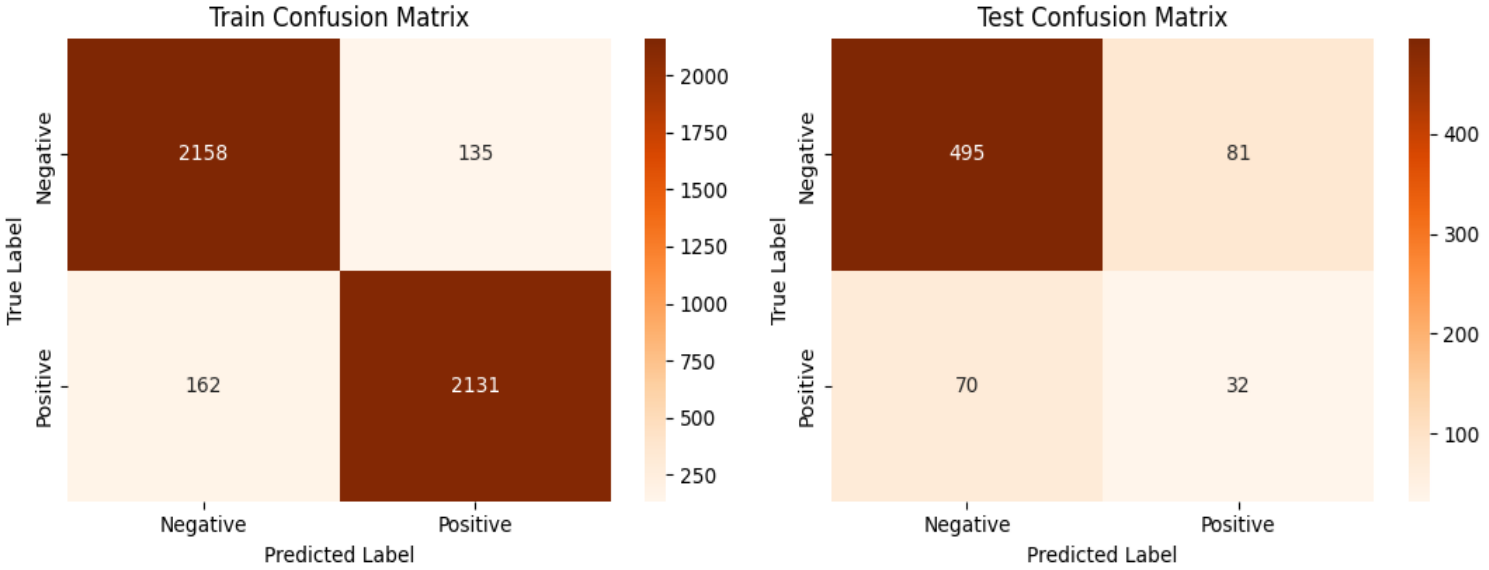




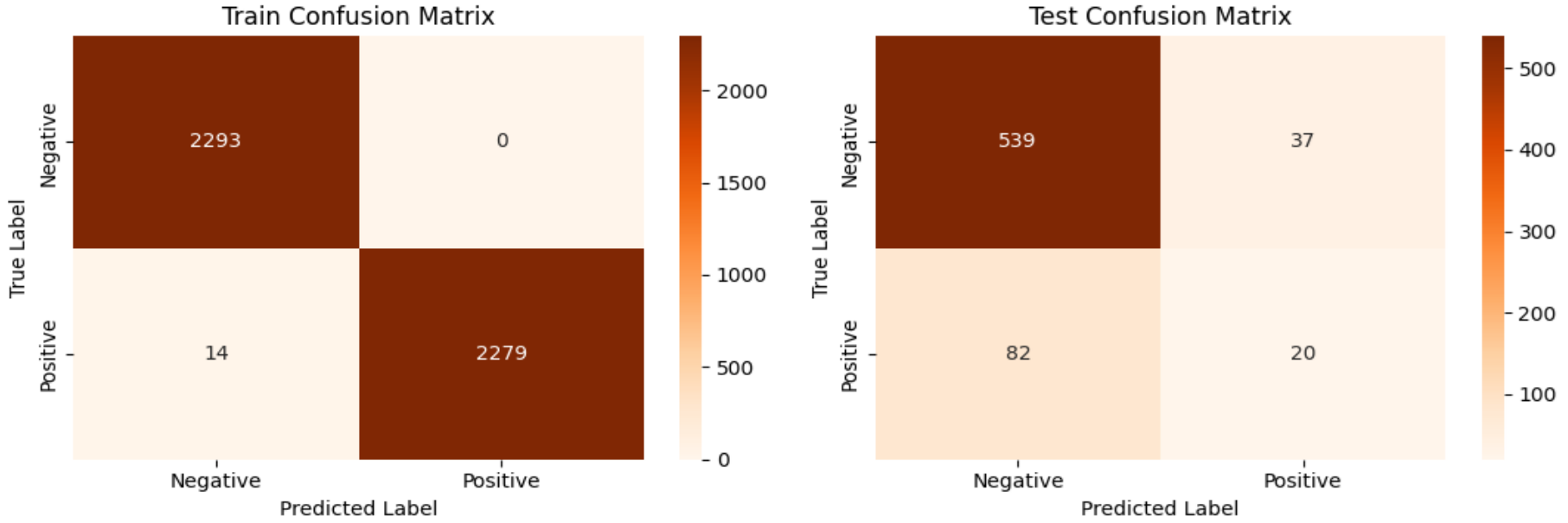
4. SVM (Support Vector Machine)



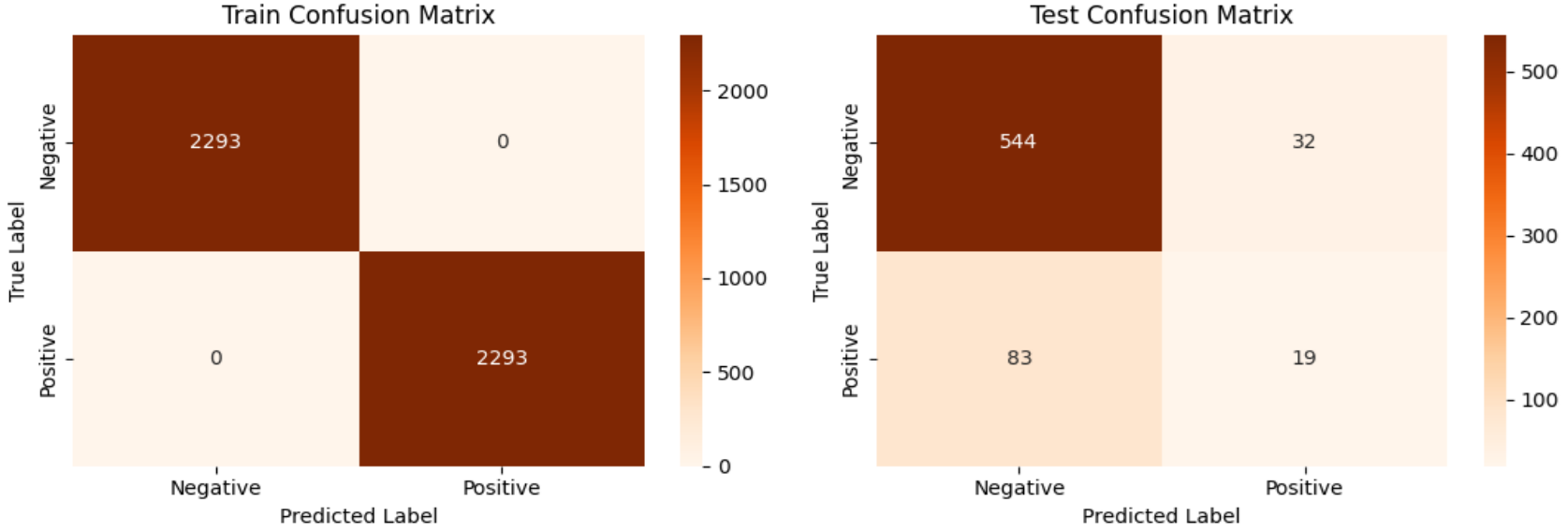
❖ After Tuned



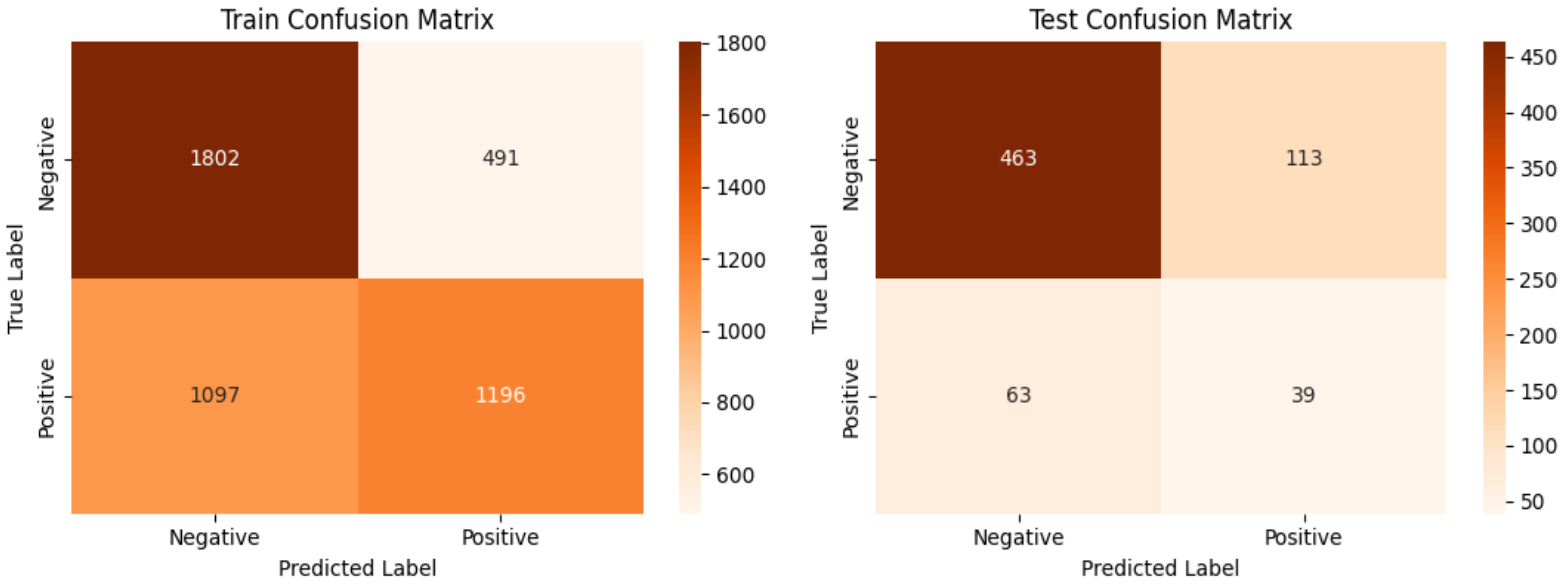
5. Xtreme Gradient Boosting



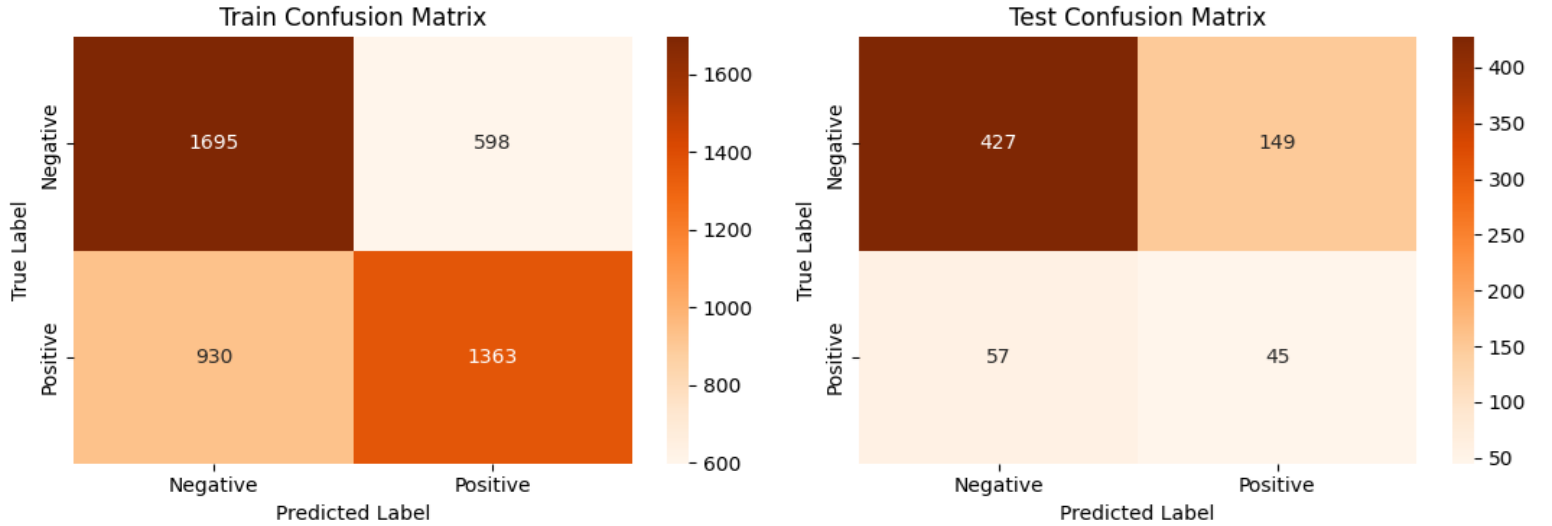
❖ After Tuned



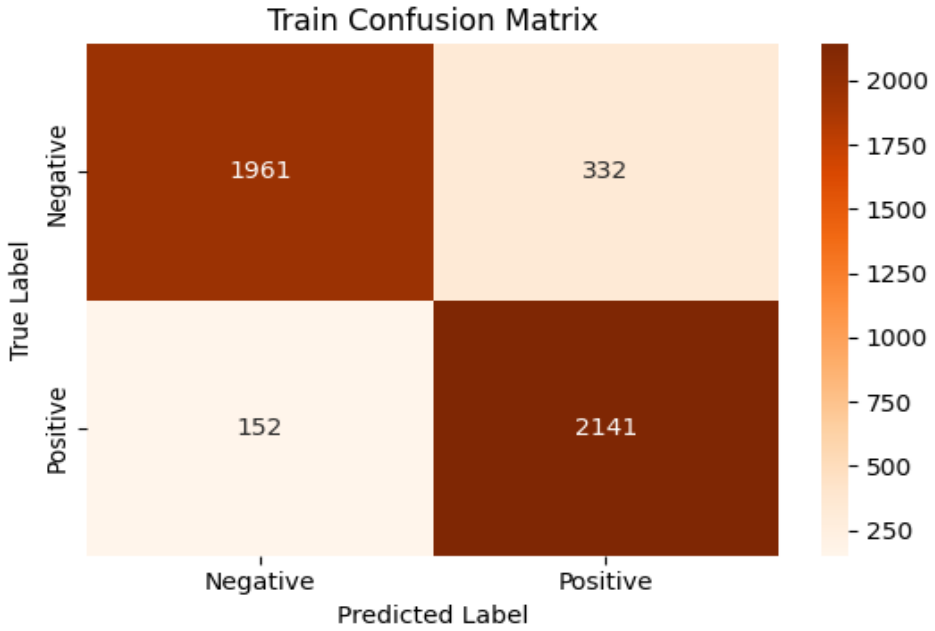
6. Naive Bayes

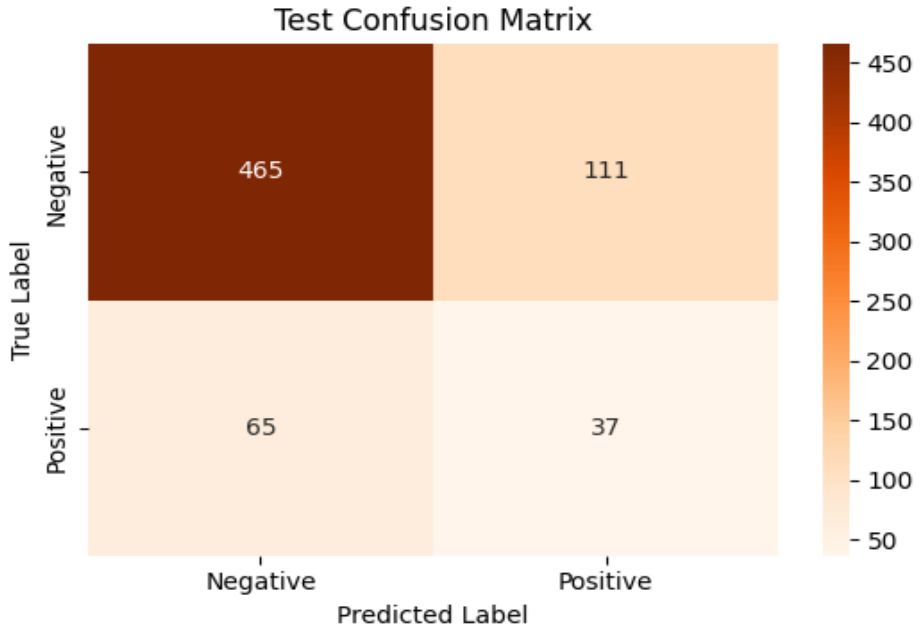


❖ After Tuned

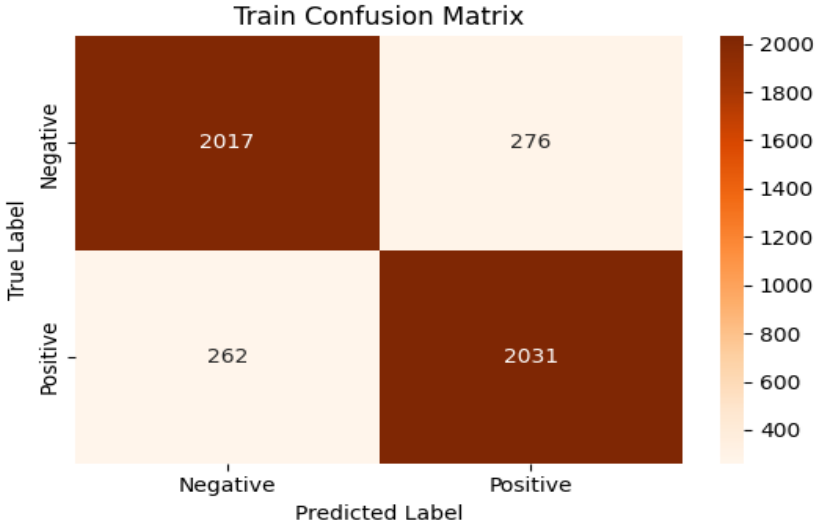


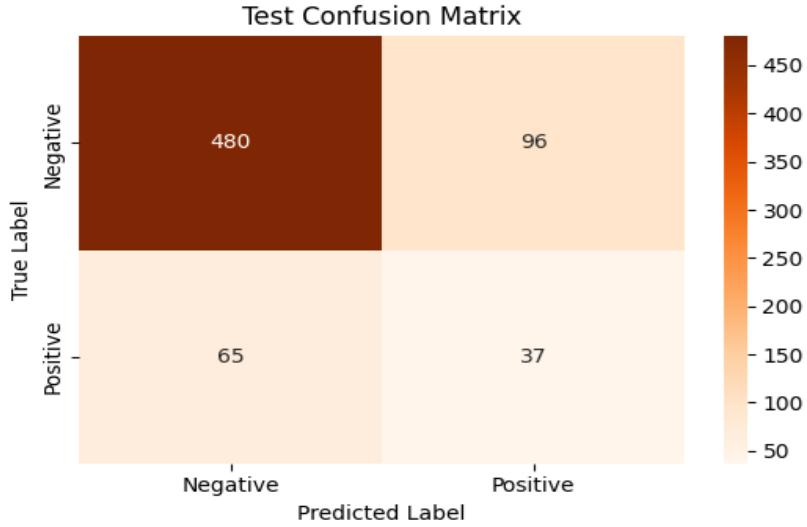
7. Neural Network





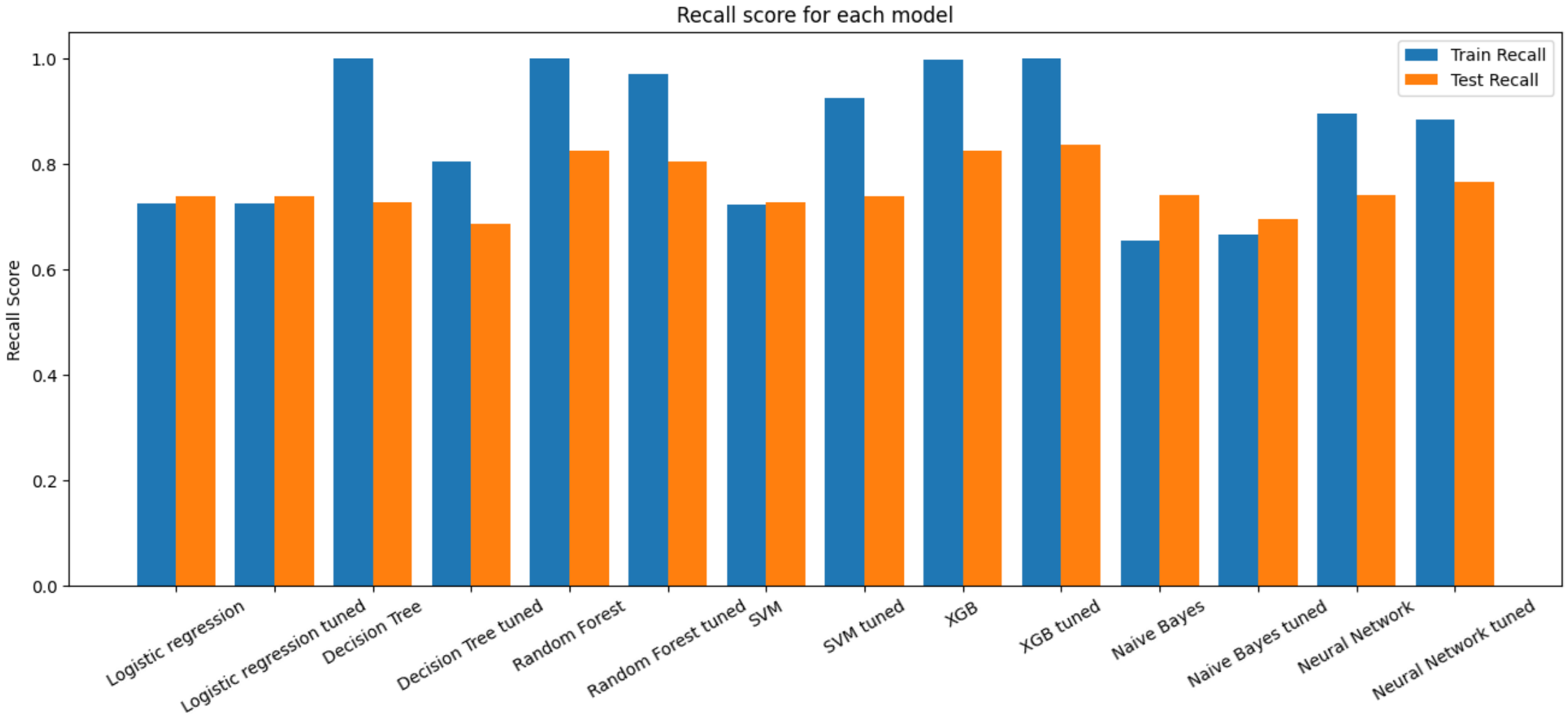
- **After Tuned**





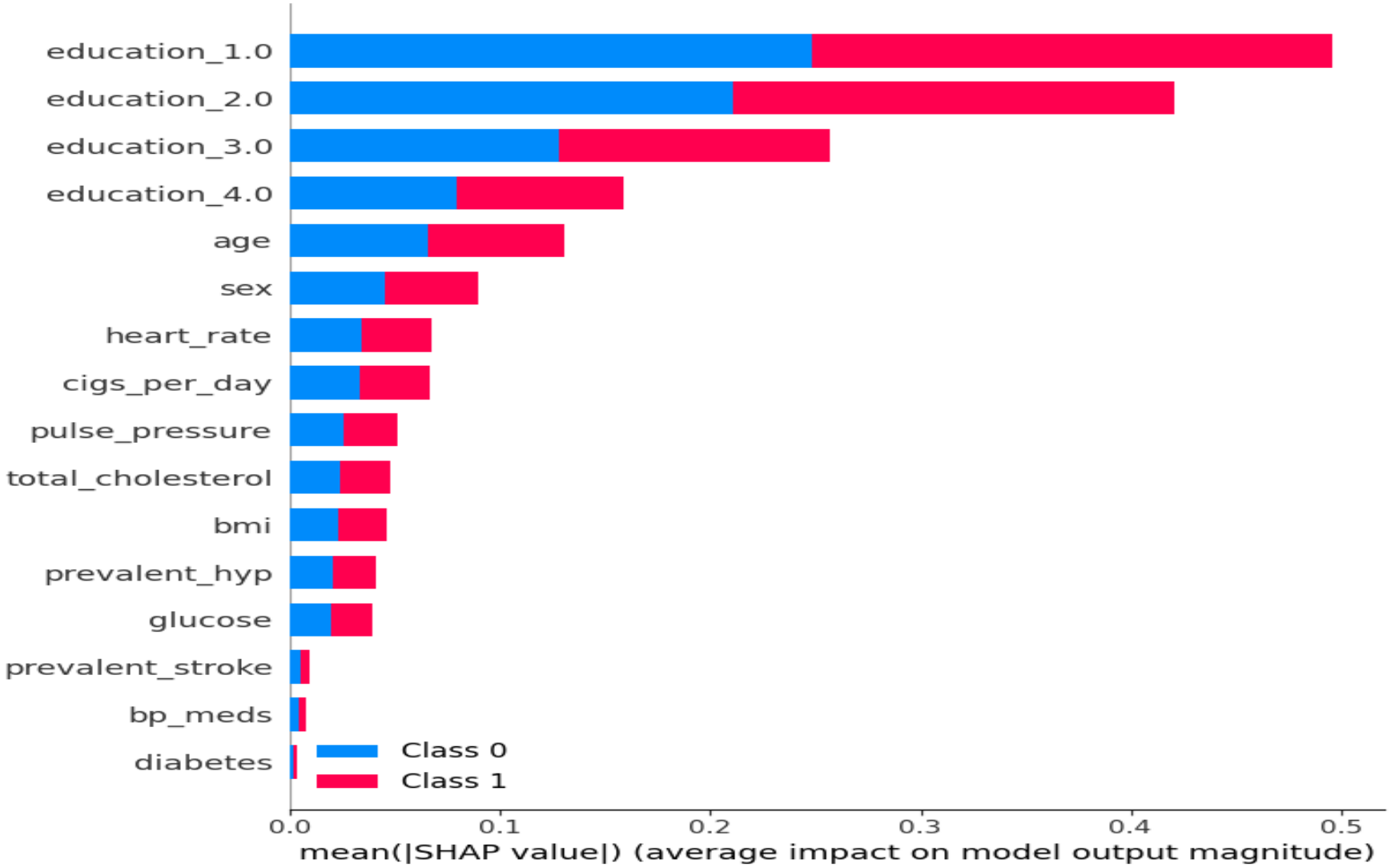
Selection of Best Model

- **Removing** the **overfitted models** which have **recall**, **ROC-AUC**, **f1 scores** for **train as 1**.
- Selected **recall** as the **primary evaluation metric**.



Model Interpretation

- ## SHAP (SHapley Additive exPlanations)



Conclusion

- The **Neural Network model (tuned)** was **chosen** as the **final prediction model** due to its **high recall score** compare to the other models.
- Due to the **presence** of much **missing/ null values** in dataset, the **accuracy** is **less**. But, its ok because it **not affects** in **life risk**.

Thank You AlmaBetter!!!