Capstone Project

Cardiovascular Risk Prediction

By - Arindam Paul
(aripaul05@gmail.com)

Points of Discussion

01

Problem Statement

02

Data Description

03

Data Preparation and Cleaning

04

EDA (Exploratory Data Analysis)

Points of Discussion

05

Hypothesis Testing

06

Feature Engineering

07

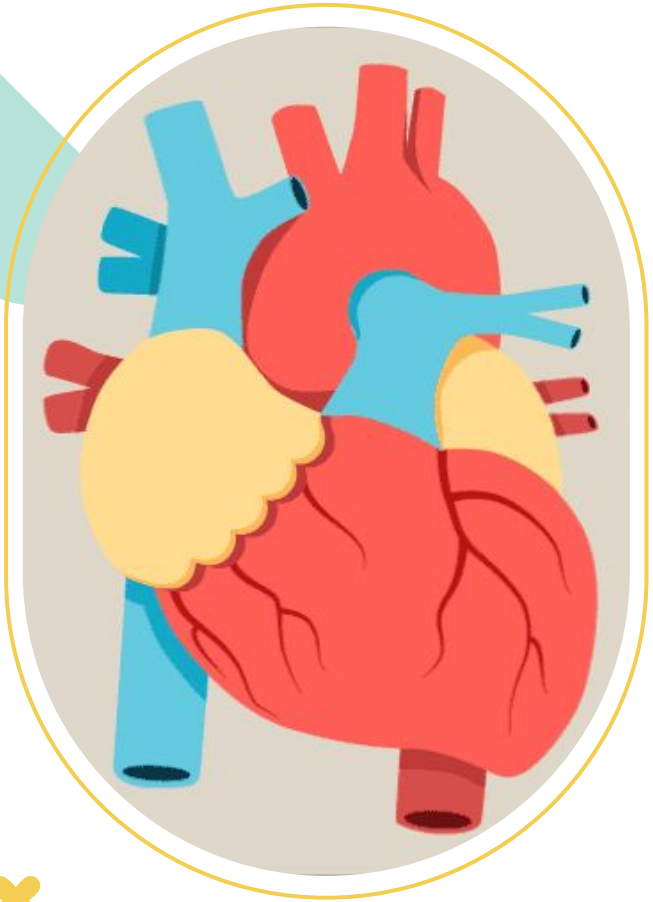
Model Implementation

08

Model Interpretation

09

Conclusion



01

Problem Statement

Classification Goal of the project

Problem Statement

Cardiovascular diseases (CVDs) are the **major cause** of **mortality worldwide**.

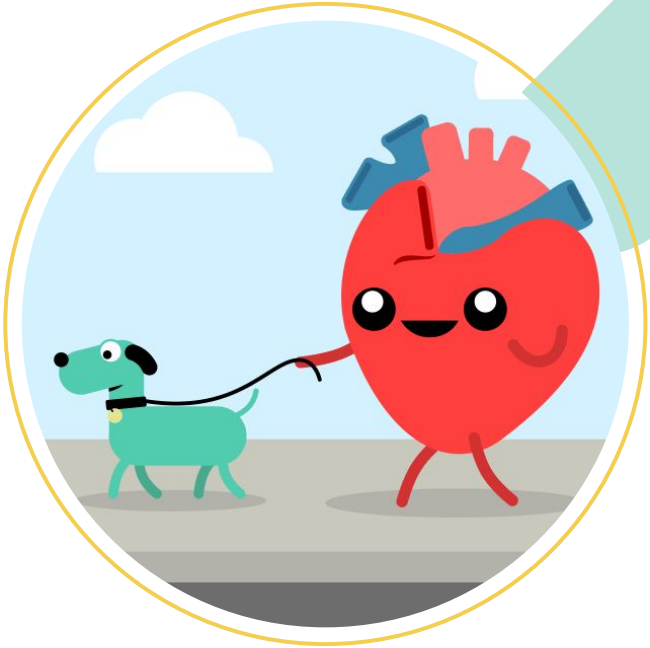
The **dataset** is from an **ongoing cardiovascular study** on **residents** of the town of **Framingham, Massachusetts**. The **classification goal** is to **predict** whether the **patient** has a **10-year risk** of **future coronary heart disease (CHD)**. The **dataset provides** the **patients information**. It includes over **4,000 records** and **15 attributes**. **Each attribute** is a **potential risk factor**. There are both **demographic**, **behavioral**, and **medical risk factors**.

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**.



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)

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”)

Data Preparation & Cleaning

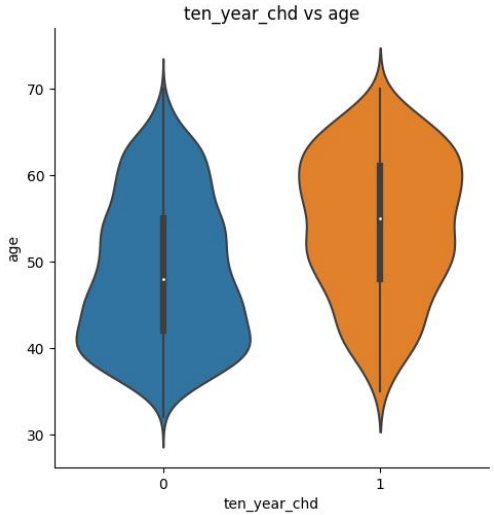
- There are **no duplicate rows** in the dataset.
- There are **missing values** in the columns **education**, **cigs per day**, **BP meds**, **totChol**, **BMI**, **heart rate** and **glucose**.
- **Changed the names** of all the **columns** for ease of use.
- I have also **defined** the **continuous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

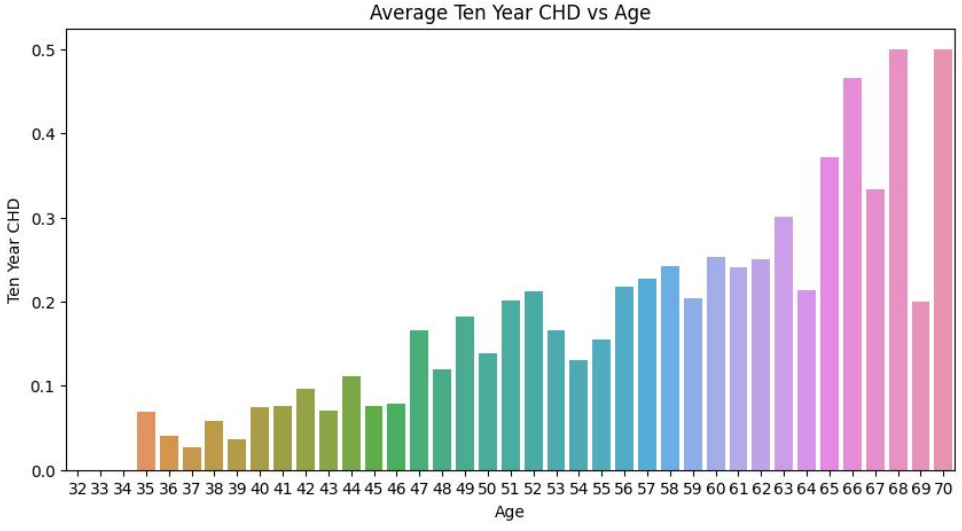
04
Exploratory Data Analysis

Analyzing data sets with statistical graphs

Ten Year CHD by Age

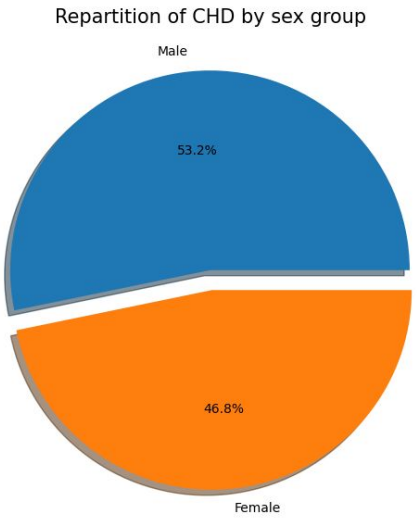
- **CHD probability** is **high** for **above 65+ aged** peoples.
- So, **older people** have a **higher risk** of having **coronary heart disease** in next **10 years**.

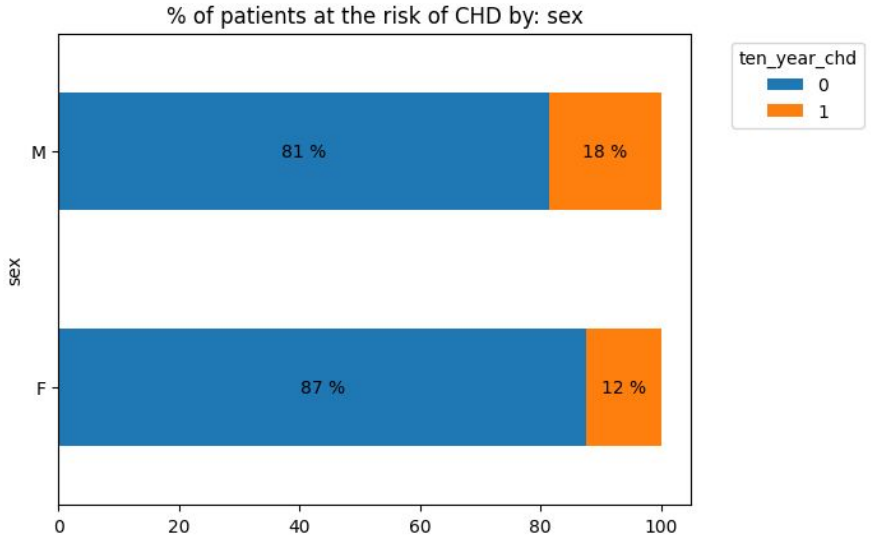




Ten Year CHD by Sex

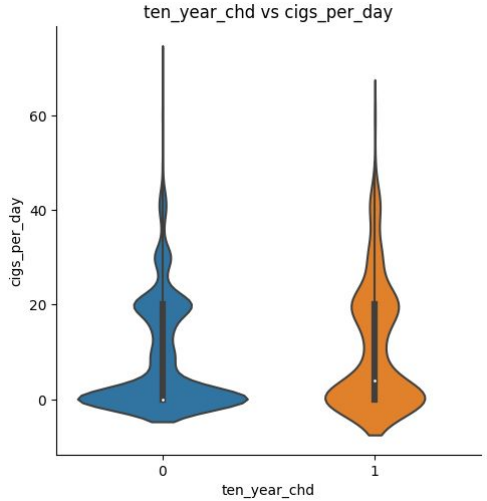
- The **gender distribution** is not even with high count for **females**. **53.2%** are there for **males** and **46.8%** for **females**.
- **Men** are generally at a **higher risk** of **having coronary heart disease**.

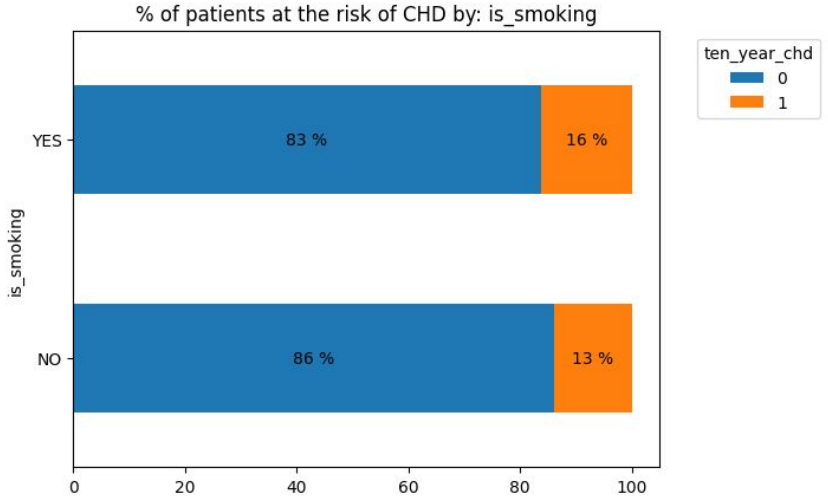




Ten Year CHD by Smoking

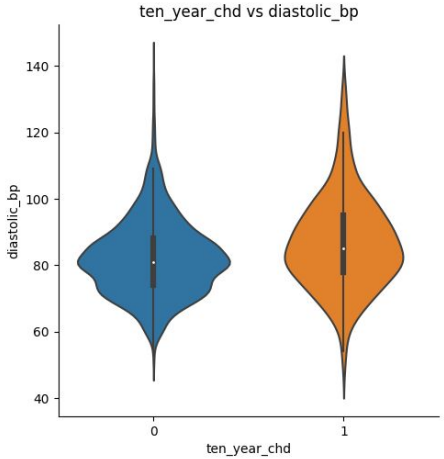
- The **negative cases** are **more** for the **non smokers** compared to the positive cases for non smokers.
- Statistically, **10 year risk of CHD** is **not dependent** on **smoking** with a 95% confidence.

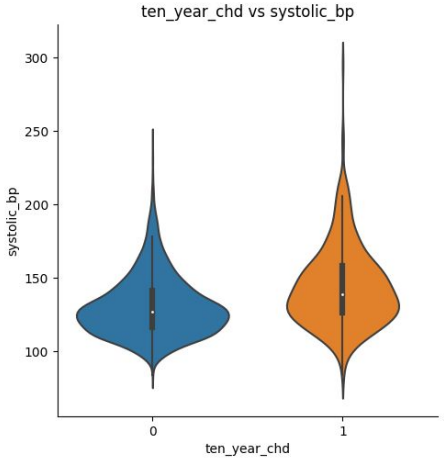


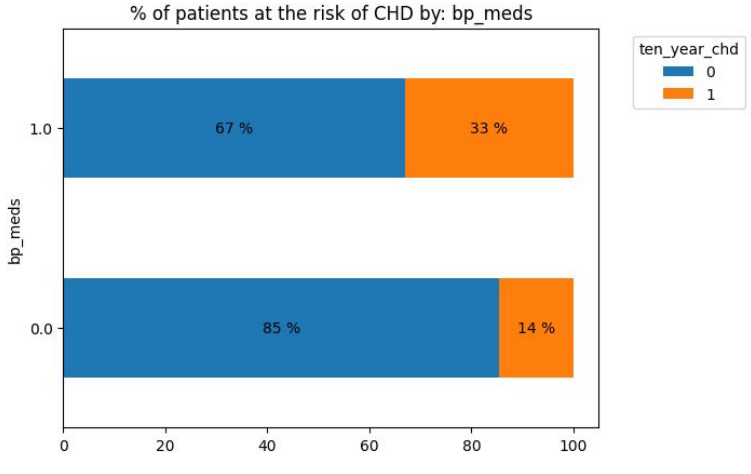


Other Notable Observations

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

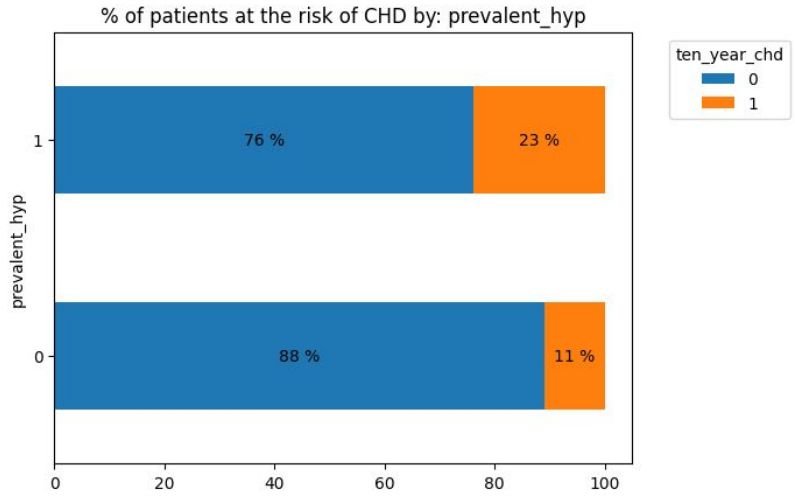


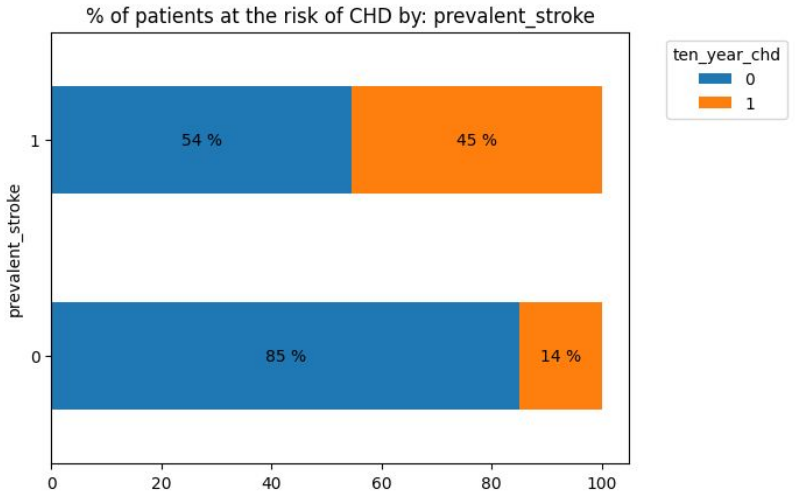




Other Notable Observations

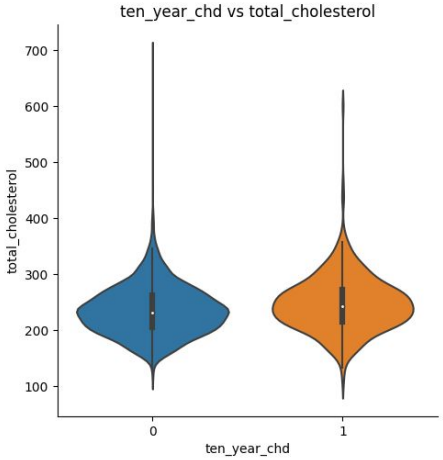
- **Patients** who have a **history of hypertension** and had a **stroke previously** have comparatively **higher risk of CHD**.

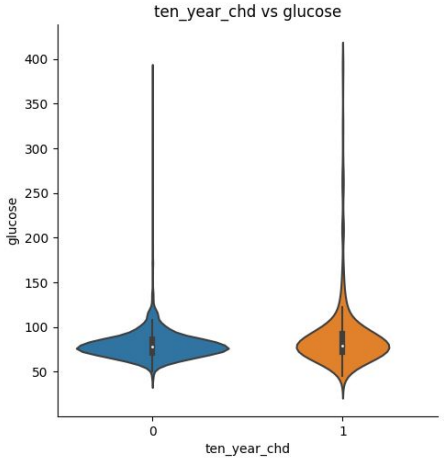


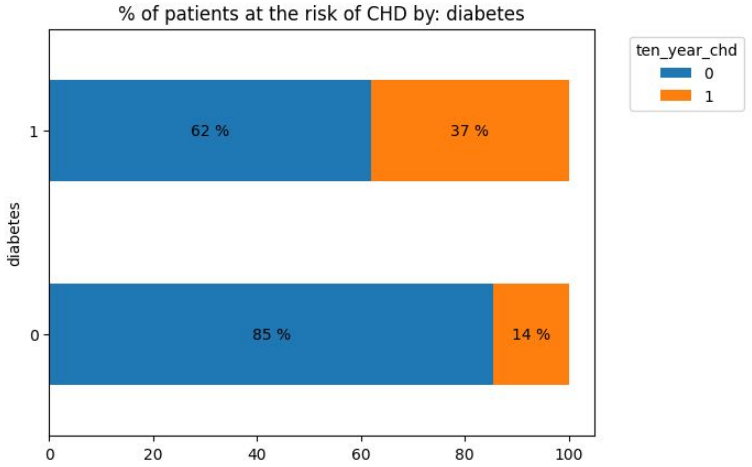


Other Notable Observations

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

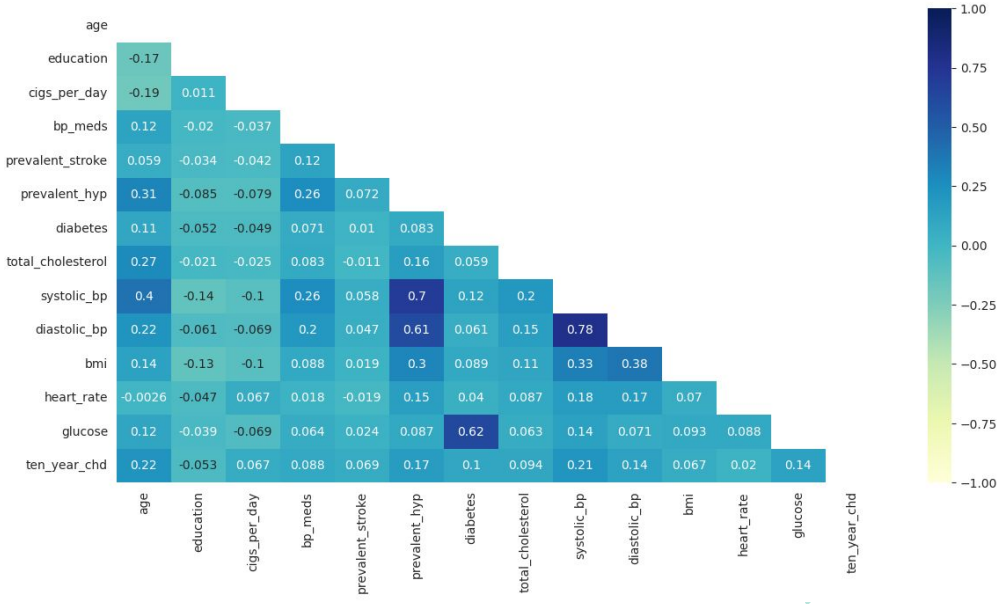






Correlation of features

- 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**.



05

Hypothesis Testing

Observation of an experiment under a given assumption

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

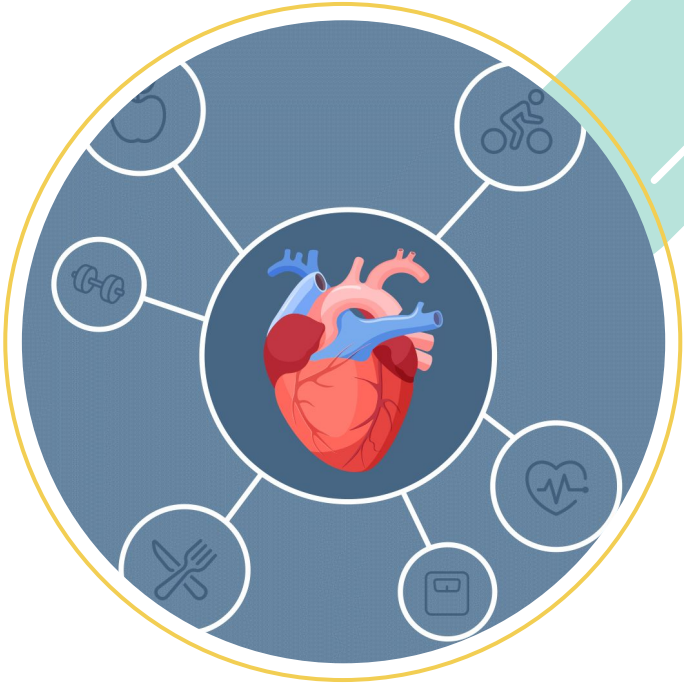
p-value: 6.038646749234552e-05

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

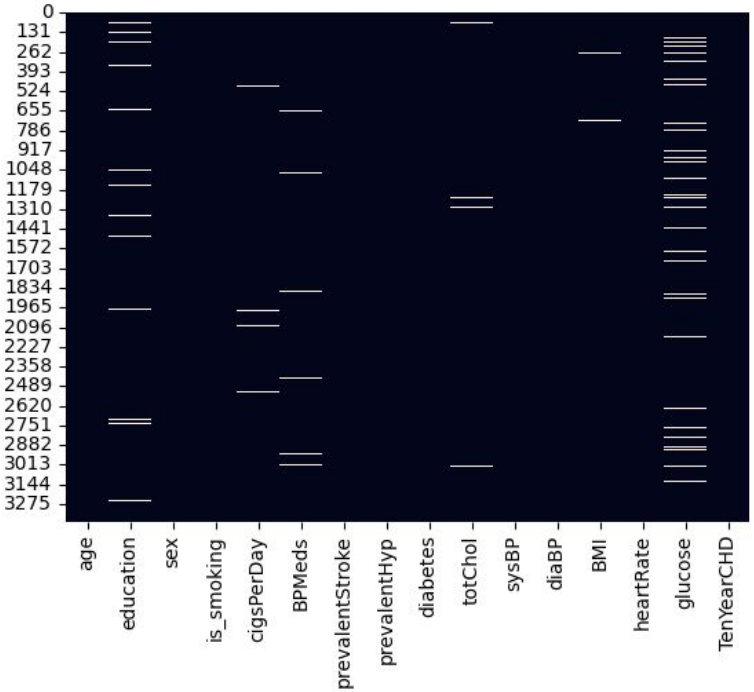
06

Feature Engineering

Extracts features from raw data



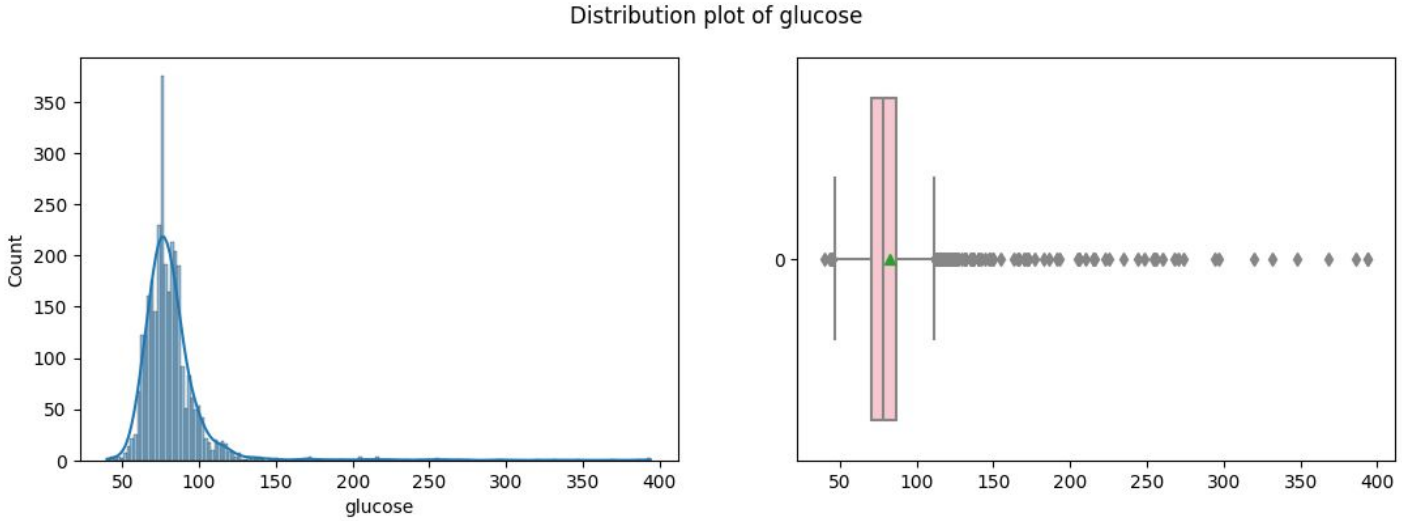
Handling Missing Values



- 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.

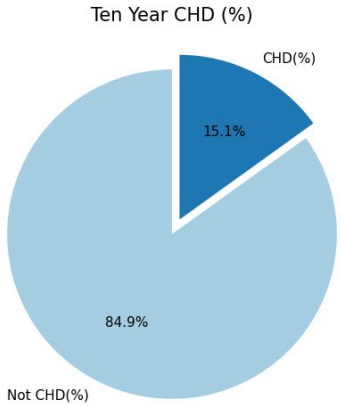
Handling Outliers

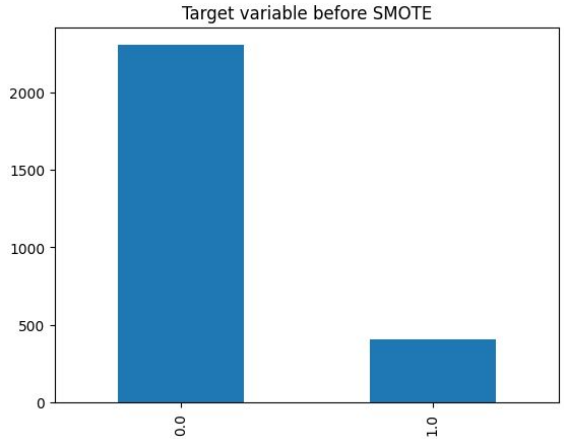
- Used the **Interquartile Range (IQR)** method to **identify** and **remove outliers** in the **continuous columns** (**systolic_bp**, **diastolic_bp**, **total cholesterol**, **glucose** etc.) of the dataset.

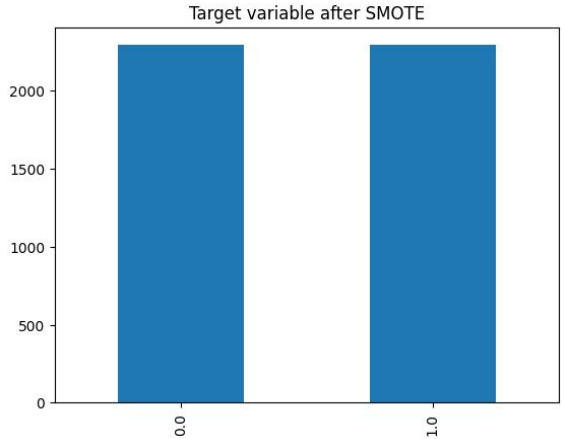


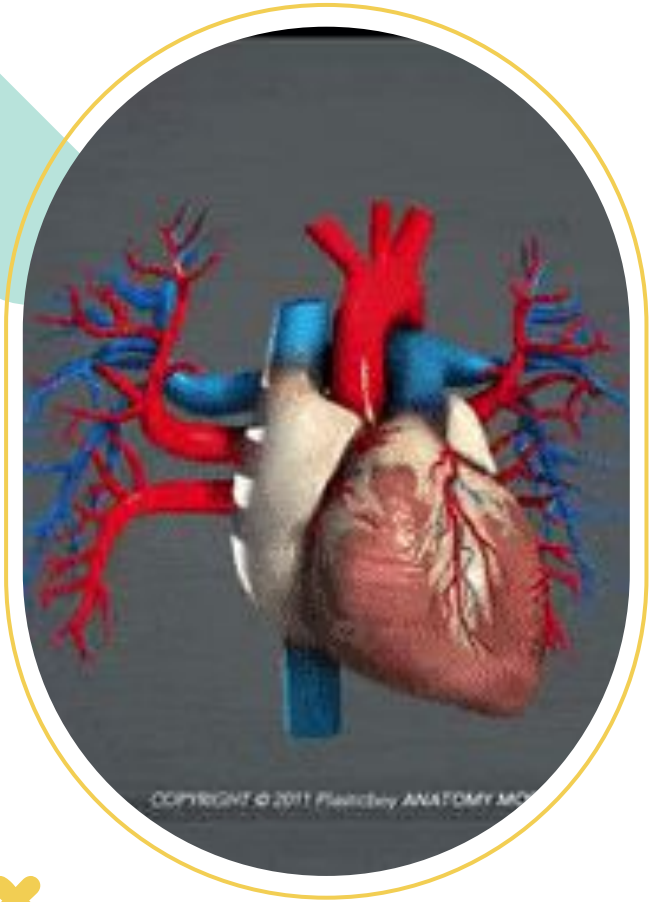
Handling Imbalanced Dataset

- 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.









07

Model Implementation

Train Ml Algorithms to get best model

ML Model Implementation

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

01

Logistic Regression

02

Decision Tree

03

Random Forest

04

Support Vector Machine

05

Xtreme Gradient Boosting

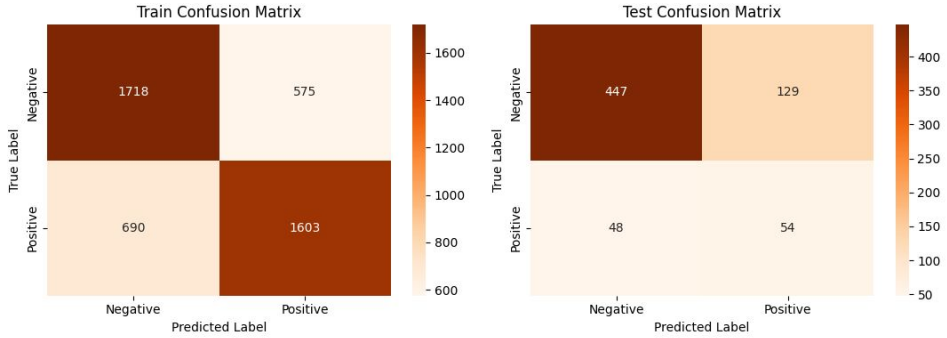
06

Naive Bayes

07

Neural Network

Logistic Regression



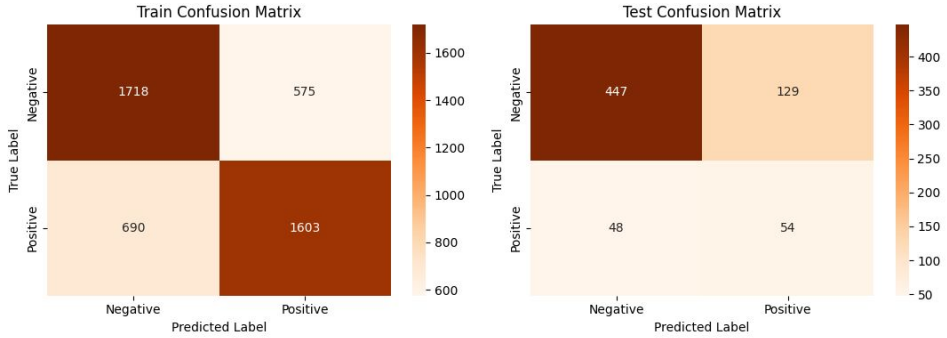
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.713455	0.749237	0.730908	2293
1.0	0.735996	0.699084	0.717066	2293
accuracy	0.72416	0.72416	0.72416	0.72416
macro avg	0.724726	0.72416	0.723987	4586
weighted avg	0.724726	0.72416	0.723987	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.90303	0.776042	0.834734	576
1.0	0.295082	0.529412	0.378947	102
accuracy	0.738938	0.738938	0.738938	0.738938
macro avg	0.599056	0.652727	0.606841	678
weighted avg	0.811569	0.738938	0.766164	678

After Tuned



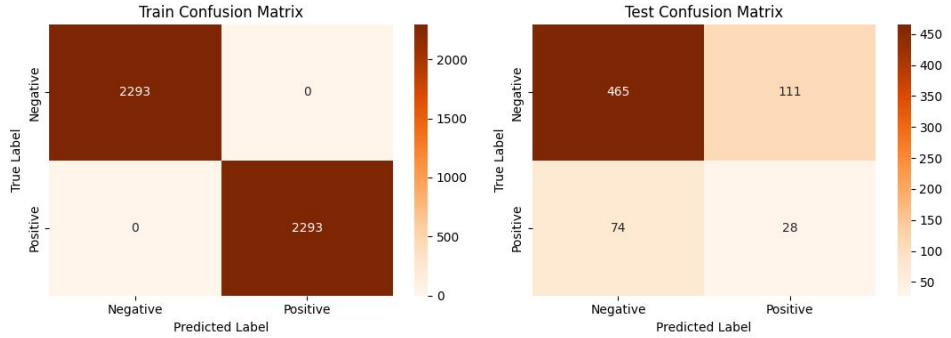
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.713455	0.749237	0.730908	2293
1.0	0.735996	0.699084	0.717066	2293
accuracy	0.72416	0.72416	0.72416	0.72416
macro avg	0.724726	0.72416	0.723987	4586
weighted avg	0.724726	0.72416	0.723987	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.90303	0.776042	0.834734	576
1.0	0.295082	0.529412	0.378947	102
accuracy	0.738938	0.738938	0.738938	0.738938
macro avg	0.599056	0.652727	0.606841	678
weighted avg	0.811569	0.738938	0.766164	678

Decision Tree



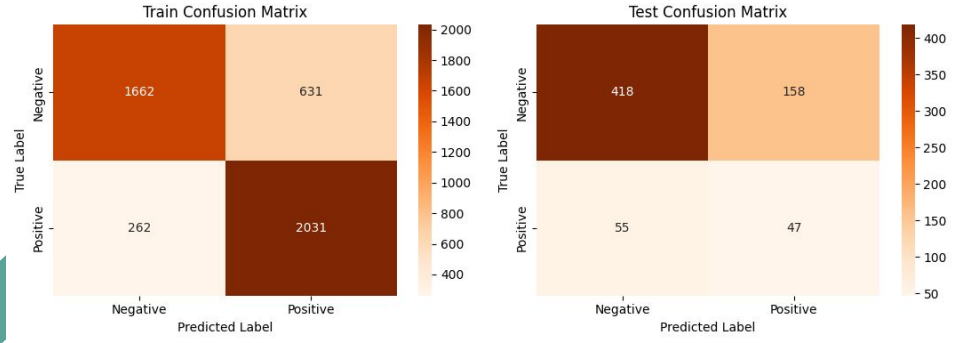
Train Classification Report:

	precision	recall	f1-score	support
0.0	1	1	1	2293
1.0	1	1	1	2293
accuracy	1	1	1	1
macro avg	1	1	1	4586
weighted avg	1	1	1	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.862709	0.807292	0.834081	576
1.0	0.201439	0.27451	0.232365	102
accuracy	0.727139	0.727139	0.727139	0.727139
macro avg	0.532074	0.540901	0.533223	678
weighted avg	0.763226	0.727139	0.743557	678

❖ After Tuned



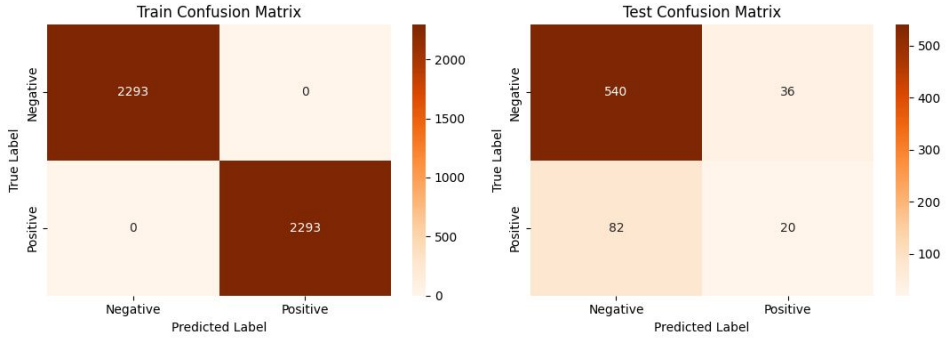
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.863825	0.724815	0.788238	2293
1.0	0.76296	0.885739	0.819778	2293
accuracy	0.805277	0.805277	0.805277	0.805277
macro avg	0.813393	0.805277	0.804008	4586
weighted avg	0.813393	0.805277	0.804008	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.883721	0.725694	0.796949	576
1.0	0.229268	0.460784	0.306189	102
accuracy	0.685841	0.685841	0.685841	0.685841
macro avg	0.556495	0.593239	0.551569	678
weighted avg	0.785263	0.685841	0.723118	678

Random Forest



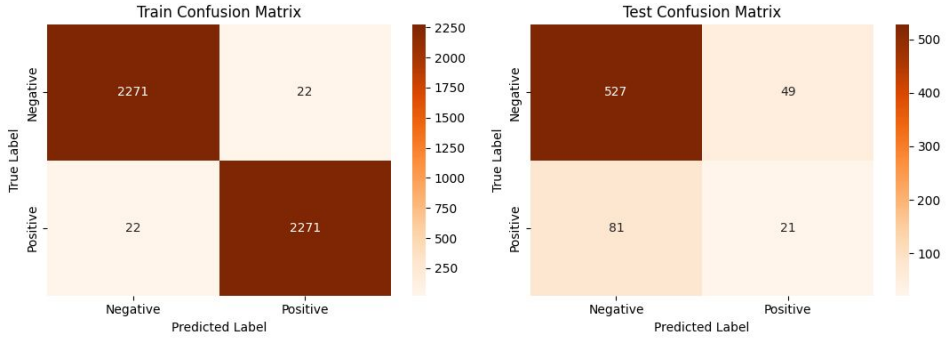
Train Classification Report:

	precision	recall	f1-score	support
0.0	1	1	1	2293
1.0	1	1	1	2293
accuracy	1	1	1	1
macro avg	1	1	1	4586
weighted avg	1	1	1	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.868167	0.9375	0.901503	576
1.0	0.357143	0.196078	0.253165	102
accuracy	0.825959	0.825959	0.825959	0.825959
macro avg	0.612655	0.566789	0.577334	678
weighted avg	0.791287	0.825959	0.803965	678

❖ After Tuned



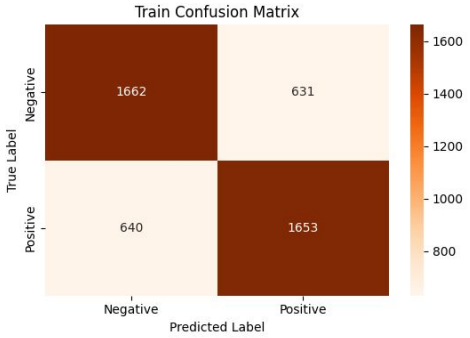
Train Classification Report:

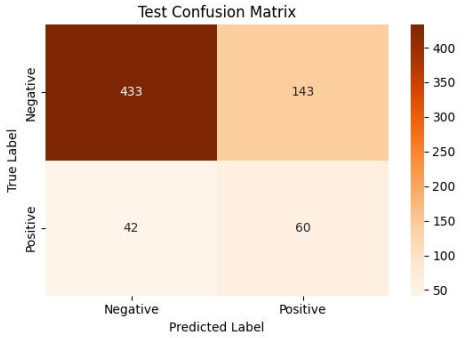
	precision	recall	f1-score	support
0.0	0.990406	0.990406	0.990406	2293
1.0	0.990406	0.990406	0.990406	2293
accuracy	0.990406	0.990406	0.990406	0.990406
macro avg	0.990406	0.990406	0.990406	4586
weighted avg	0.990406	0.990406	0.990406	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.866776	0.914931	0.890203	576
1.0	0.3	0.205882	0.244186	102
accuracy	0.80826	0.80826	0.80826	0.80826
macro avg	0.583388	0.560406	0.567194	678
weighted avg	0.781509	0.80826	0.793014	678

SVM (Support Vector Machine)





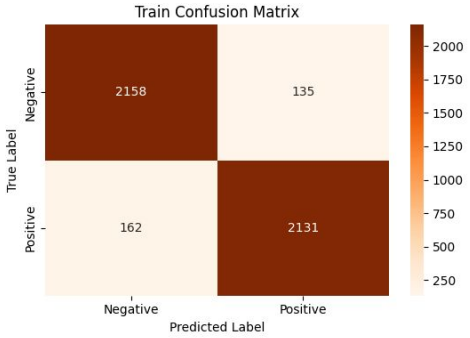
Train Classification Report:

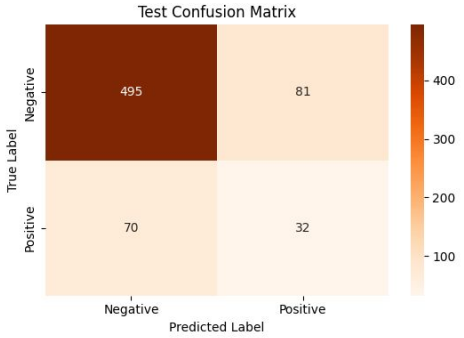
	precision	recall	f1-score	support
0.0	0.721981	0.724815	0.723395	2293
1.0	0.72373	0.72089	0.722307	2293
accuracy	0.722852	0.722852	0.722852	0.722852
macro avg	0.722856	0.722852	0.722851	4586
weighted avg	0.722856	0.722852	0.722851	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.911579	0.751736	0.823977	576
1.0	0.295567	0.588235	0.393443	102
accuracy	0.727139	0.727139	0.727139	0.727139
macro avg	0.603573	0.669986	0.60871	678
weighted avg	0.818905	0.727139	0.759206	678

❖ After Tuned





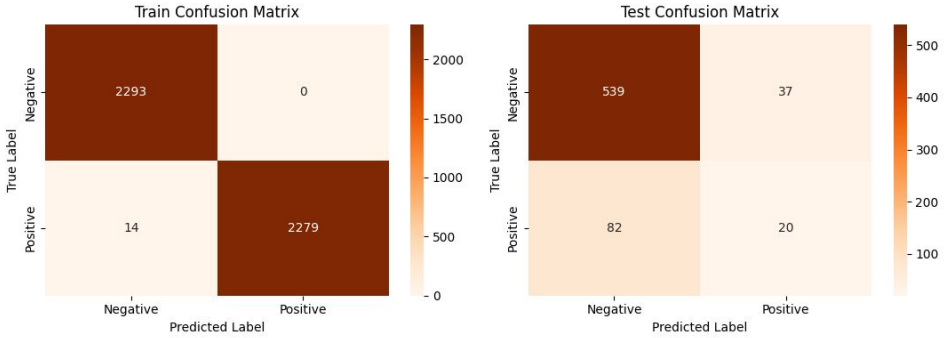
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.930172	0.941125	0.935617	2293
1.0	0.940424	0.92935	0.934854	2293
accuracy	0.935238	0.935238	0.935238	0.935238
macro avg	0.935298	0.935238	0.935235	4586
weighted avg	0.935298	0.935238	0.935235	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.876106	0.859375	0.86766	576
1.0	0.283186	0.313725	0.297674	102
accuracy	0.777286	0.777286	0.777286	0.777286
macro avg	0.579646	0.58655	0.582667	678
weighted avg	0.786906	0.777286	0.78191	678

Xtreme Gradient Boosting



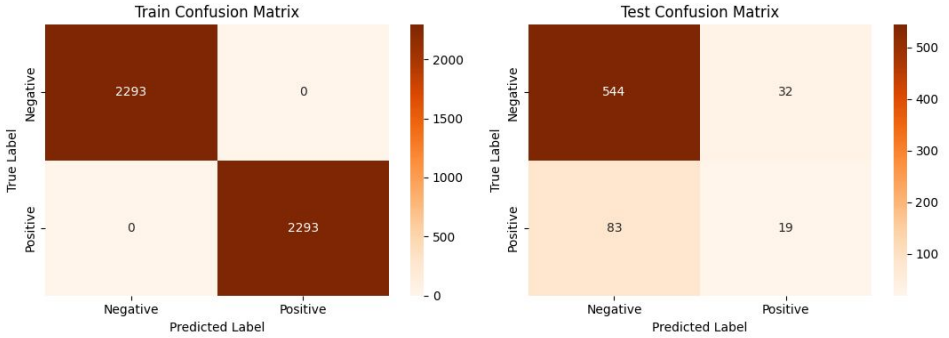
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.993932	1	0.996957	2293
1.0	1	0.993894	0.996938	2293
accuracy	0.996947	0.996947	0.996947	0.996947
macro avg	0.996966	0.996947	0.996947	4586
weighted avg	0.996966	0.996947	0.996947	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.867955	0.935764	0.900585	576
1.0	0.350877	0.196078	0.251572	102
accuracy	0.824484	0.824484	0.824484	0.824484
macro avg	0.609416	0.565921	0.576079	678
weighted avg	0.790164	0.824484	0.802946	678

After Tuned



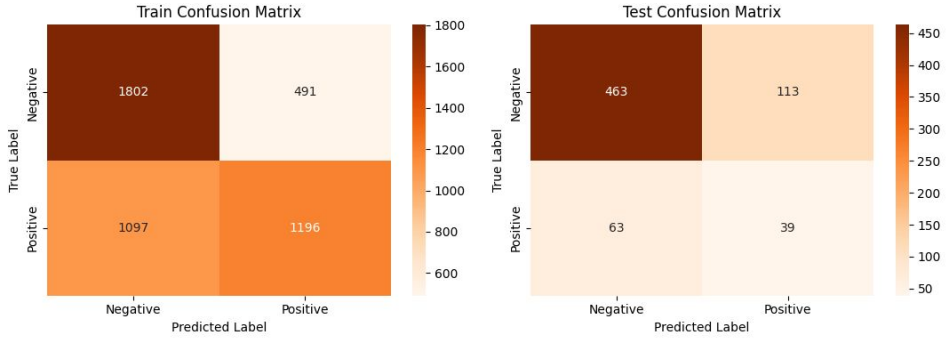
Train Classification Report:

	precision	recall	f1-score	support
0.0	1	1	1	2293
1.0	1	1	1	2293
accuracy	1	1	1	1
macro avg	1	1	1	4586
weighted avg	1	1	1	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.867624	0.944444	0.904406	576
1.0	0.372549	0.186275	0.248366	102
accuracy	0.830383	0.830383	0.830383	0.830383
macro avg	0.620086	0.565359	0.576386	678
weighted avg	0.793143	0.830383	0.805709	678

Naive Bayes



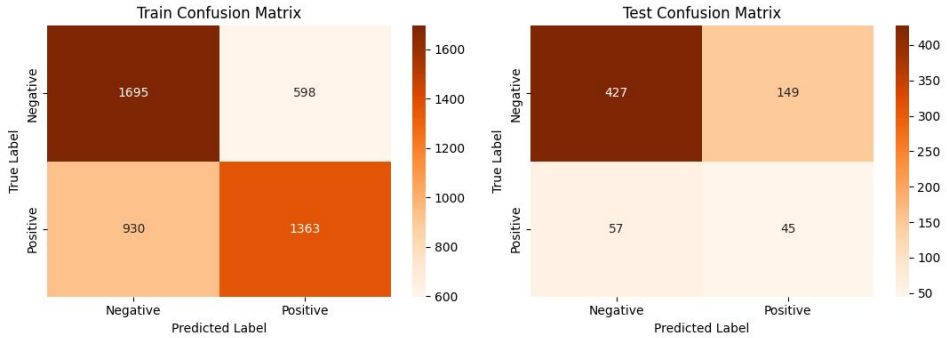
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.621594	0.78587	0.694145	2293
1.0	0.708951	0.521587	0.601005	2293
accuracy	0.653729	0.653729	0.653729	0.653729
macro avg	0.665272	0.653729	0.647575	4586
weighted avg	0.665272	0.653729	0.647575	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.880228	0.803819	0.84029	576
1.0	0.256579	0.382353	0.307087	102
accuracy	0.740413	0.740413	0.740413	0.740413
macro avg	0.568404	0.593086	0.573688	678
weighted avg	0.786405	0.740413	0.760074	678

❖ After Tuned



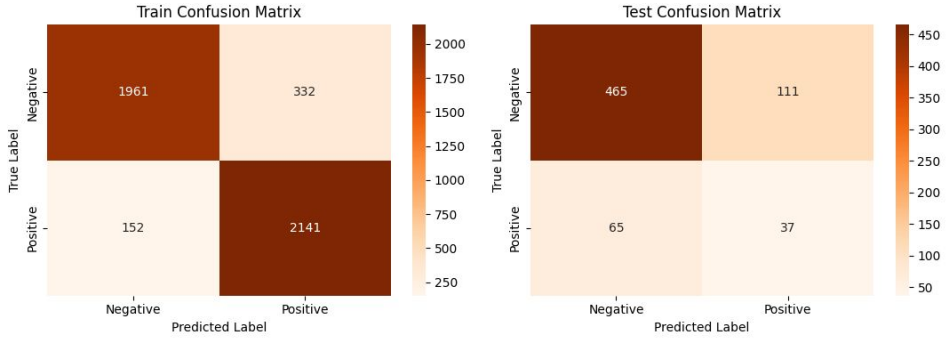
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.645714	0.739206	0.689305	2293
1.0	0.695054	0.594418	0.640809	2293
accuracy	0.666812	0.666812	0.666812	0.666812
macro avg	0.670384	0.666812	0.665057	4586
weighted avg	0.670384	0.666812	0.665057	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.882231	0.741319	0.80566	576
1.0	0.231959	0.441176	0.304054	102
accuracy	0.696165	0.696165	0.696165	0.696165
macro avg	0.557095	0.591248	0.554857	678
weighted avg	0.784403	0.696165	0.730197	678

Neural Network



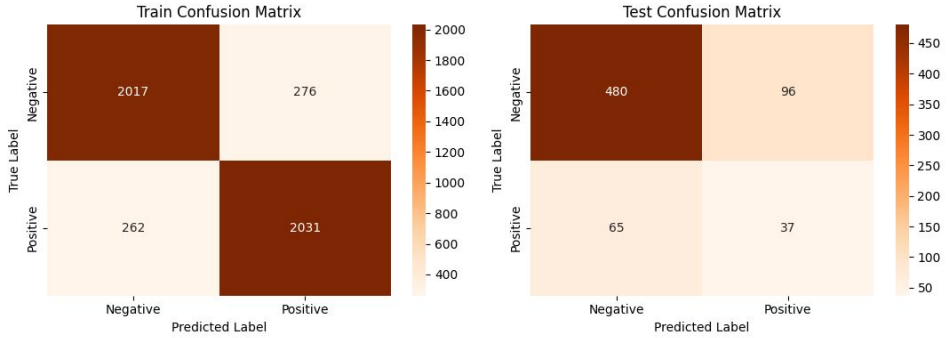
Train Classification Report:

	precision	recall	f1-score	support
0.0	0.928064	0.855212	0.89015	2293
1.0	0.86575	0.933711	0.898447	2293
accuracy	0.894461	0.894461	0.894461	0.894461
macro avg	0.896907	0.894461	0.894299	4586
weighted avg	0.896907	0.894461	0.894299	4586

Test Classification Report:

	precision	recall	f1-score	support
0.0	0.877358	0.807292	0.840868	576
1.0	0.25	0.362745	0.296	102
accuracy	0.740413	0.740413	0.740413	0.740413
macro avg	0.563679	0.585018	0.568434	678
weighted avg	0.782977	0.740413	0.758897	678

After Tuned



Train Classification Report:

	precision	recall	f1-score	support
0.0	0.885037	0.879634	0.882327	2293
1.0	0.880364	0.885739	0.883043	2293
accuracy	0.882686	0.882686	0.882686	0.882686
macro avg	0.882701	0.882686	0.882685	4586
weighted avg	0.882701	0.882686	0.882685	4586

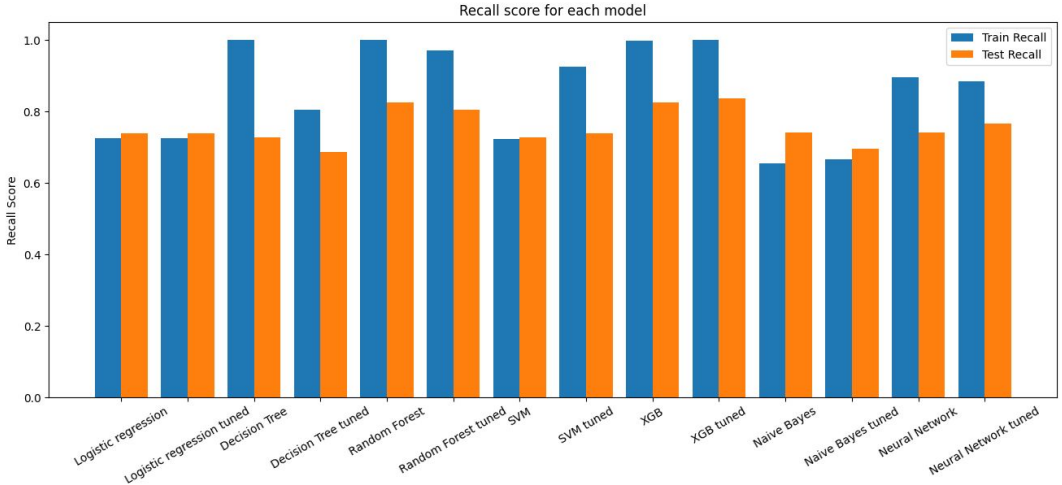
Test Classification Report:

	precision	recall	f1-score	support
0.0	0.880734	0.833333	0.856378	576
1.0	0.278195	0.362745	0.314894	102
accuracy	0.762537	0.762537	0.762537	0.762537
macro avg	0.579465	0.598039	0.585636	678
weighted avg	0.790087	0.762537	0.774916	678

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**.

Classification Model	Recall Train	Recall Test
Logistic regression	0.72416	0.738938
Logistic regression tuned	0.72416	0.738938
Decision Tree tuned	0.805277	0.685841
SVM	0.722852	0.727139
SVM tuned	0.920846	0.746313
Naive Bayes	0.653729	0.740413
Naive Bayes tuned	0.666812	0.696165
Neural Network	0.894461	0.740413
Neural Network tuned	0.894679	0.784661

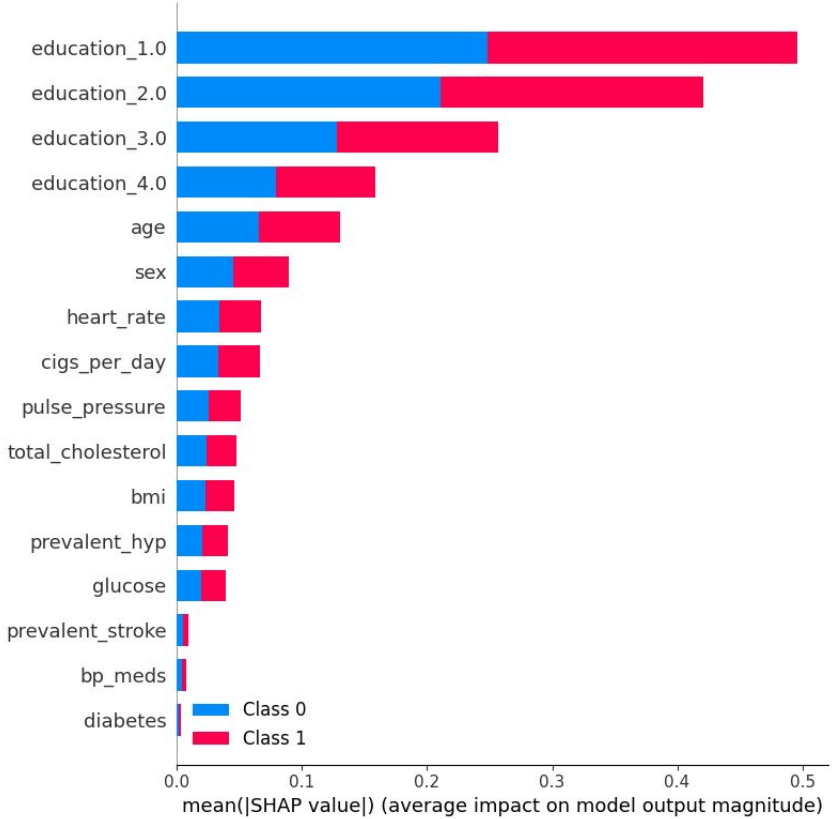


08

Model Interpretation

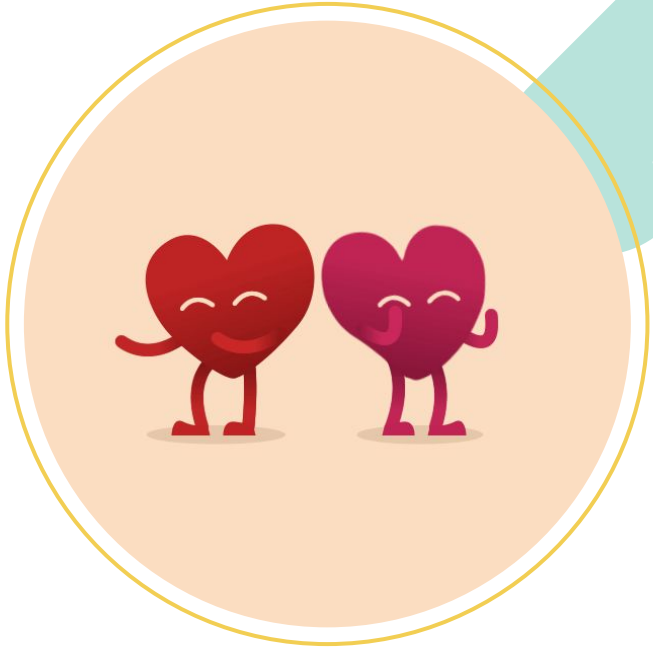
Important features by using model explainability tool

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!