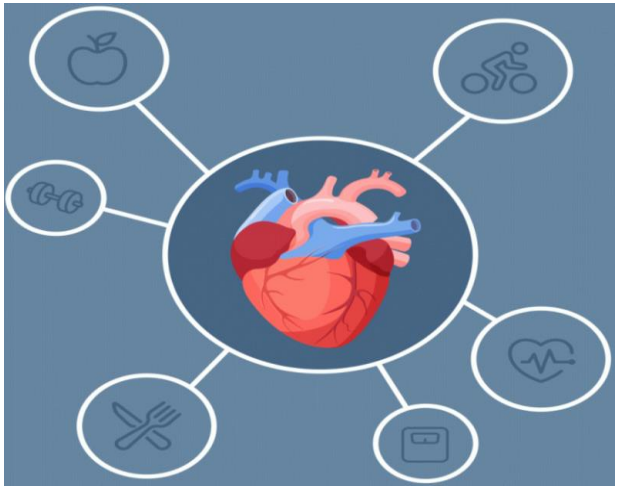
Capstone Project

Cardiovascular Risk Prediction



Submitted by
Vikas Chaudhary

Flow of the Presentation

AI

- Introduction
- Data Preparation
- EDA
- Algorithms Implementation
- Challenges
- Conclusion

Introduction

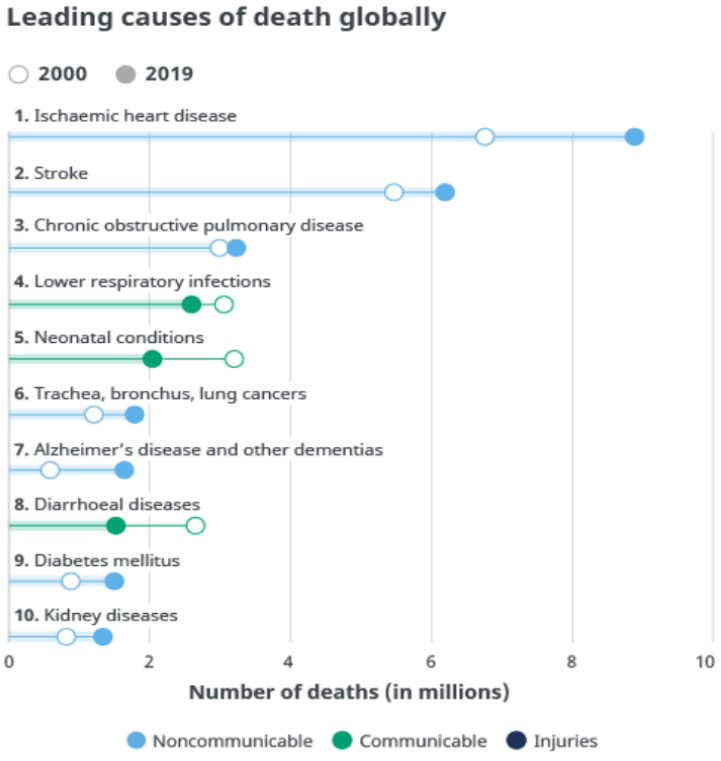
AI

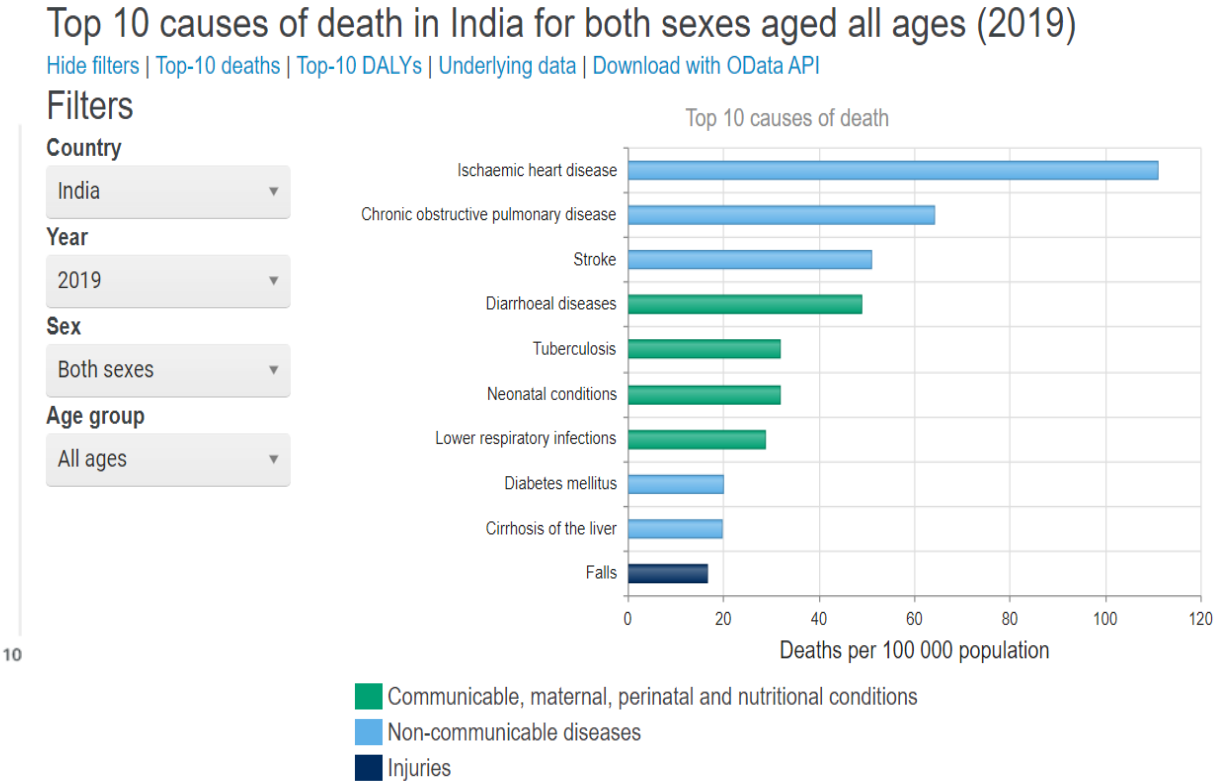
As per the World Health Organization (WHO), Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

Key facts

- Cardiovascular diseases (CVDs) are the leading cause of death globally.
- An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.
- Over three quarters of CVD deaths take place in low- and middle-income countries.
- Out of the 17 million premature deaths (under the age of 70) due to non-communicable diseases in 2019, 38% were caused by CVDs.
- Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.
- It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.

Leading causes of death





Out-Of-Pocket Expenditures (OOPE) and Poverty in India

- India has one-of-the highest level of Out-Of-Pocket Expenditures (OOPE) contributing directly to the high incidence of catastrophic expenditures and poverty, notes the Economic Survey 2020-21.
- It suggested an increase in public spending from 1% to 2.5-3% of GDP — as envisaged in the National Health Policy 2017 — can decrease the OOPE from 65% to 30% of overall healthcare spend.
- Household Out-Of-Pocket (OOP) expenses on healthcare services, especially medicines, continue to push more than 55 million people in India into poverty, with over 18 per cent of households incurring catastrophic levels of health expenditures annually.
- Retail health insurance penetrates only a meagre 3.2 per cent of the 138 crore population, leaving a huge part of it—38.8 per cent, which is about 56 crore individuals—unprotected from any sort of health insurance cover, making them prone to financial shocks and healthcare debt.

Universal Health Coverage 2030

As stated in the World Health Organization's sustainable development goal (SDG) 3.8, reaching universal health coverage and financial risk protection are important indicators to guarantee better healthy lives and higher well-being.

Objective: On the basis of the given dataset that contains previous records, we need to build a machine learning (ML) model to predict the risk of cardiovascular disease of any patient or individual over next 10 years.

Methodology: Machine Learning (ML) Classification (Supervised ML Model)

Database Summary:

•The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts (US).
•It has 3390 rows and 17 columns
• All the features of the database can be divided into three verticals, i.e. 1. Demographic, 2. Behavioral, 3. Medical Risk Factors (past and present).

Data Preparation

AI

In-dependent Features:

- id: a number (serial number) is given to each row.
- age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- education: it is a categorical feature contains 4 values in numerical form (1.0, 2.0, 3.0, 4.0)
- sex: male or female("M" or "F")
- 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.)
- 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)
- totChol: total cholesterol level (Continuous)
- sysBP: systolic blood pressure (Continuous)
- diaBP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- heartRate: 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)

Dependent Feature:

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

Overview of Dataset

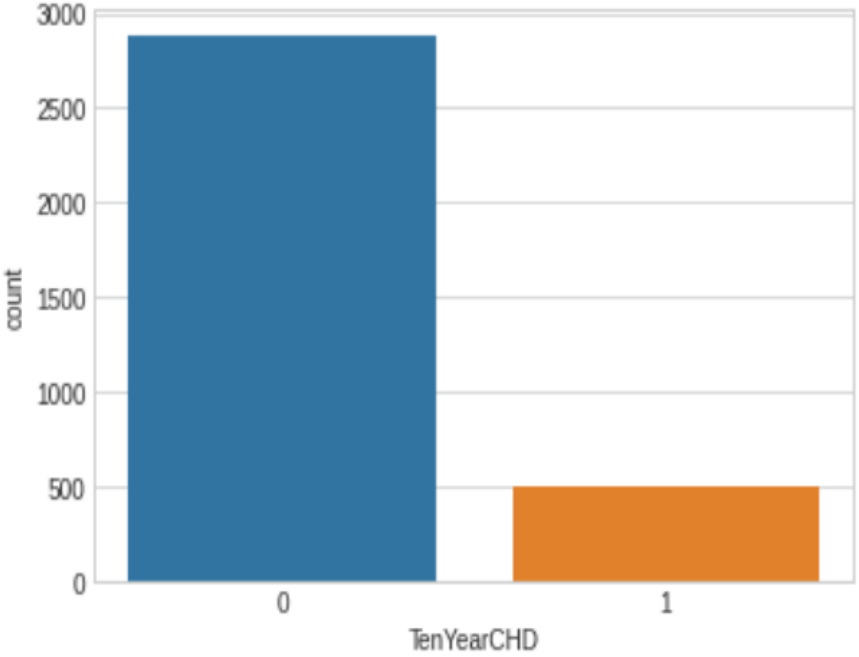
- It has ‘NaN’ or ‘Null’ values
- It has no duplicate values
- It contains the data of separate individuals
- ‘TenYearCHD’ is the dependent feature.
- The entire database has two types of features i.e. numerical and categorical.
- Different methods are being used to deal with null/NaN values in each feature.
- Every feature except ‘id’ has been used in the model.

```
#knowing if there is any NaN/Null values
df.isna().sum()
```

```
id                   0
age                  0
education           87
sex                  0
is_smoking           0
cigsPerDay          22
BPMeds              44
prevalentStroke      0
prevalentHyp         0
diabetes             0
totChol             38
sysBP                0
diaBP                0
BMI                 14
heartRate            1
glucose            304
TenYearCHD           0
```

EDA

Dependent Feature (TenYearCHD)

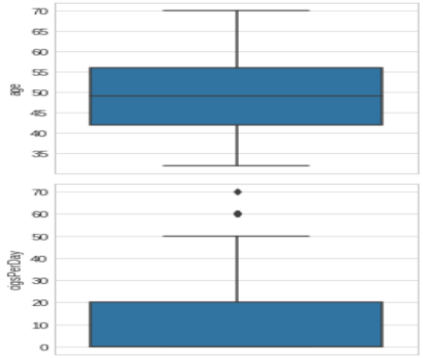


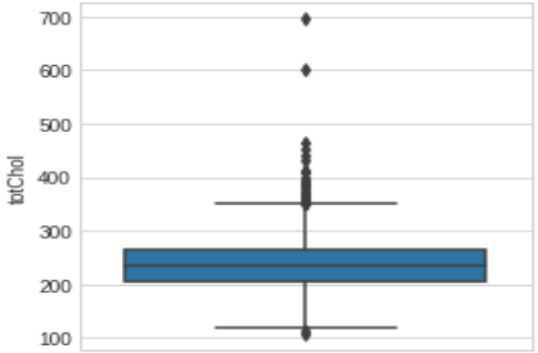
Out of 3390 records:

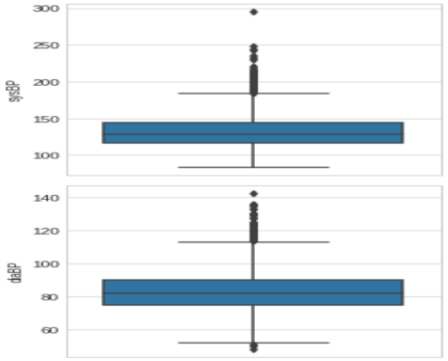
- 2879 have no cardiovascular risk
- 511 have the cardiovascular risk

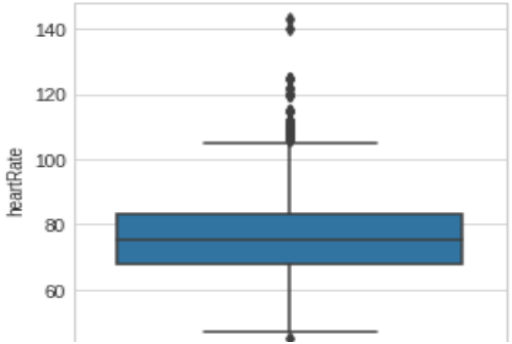
for next 10-years.

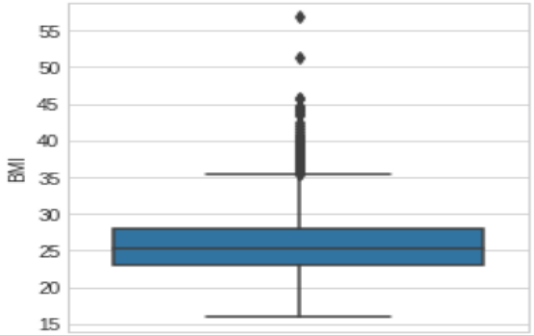
Box plots for Outliers

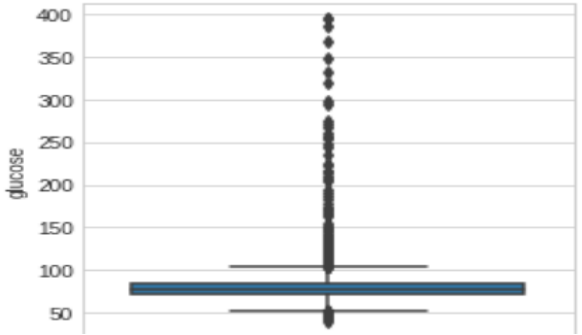




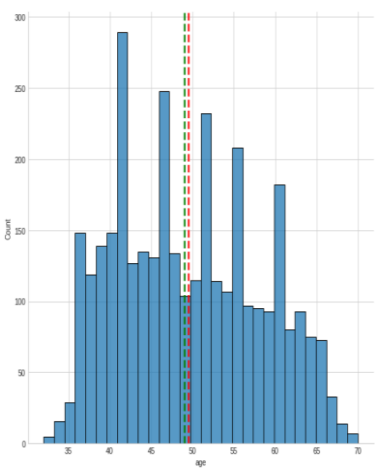


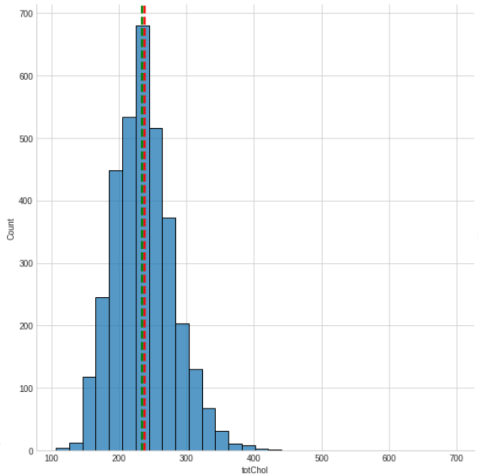


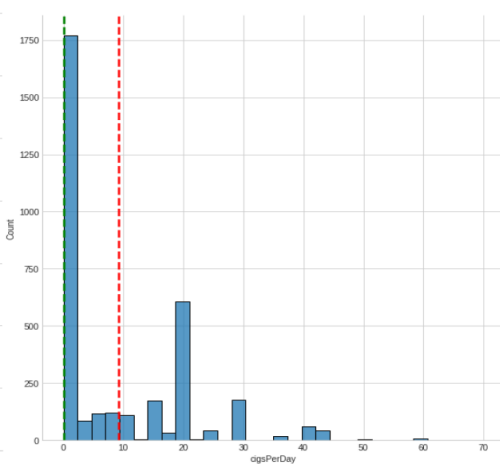


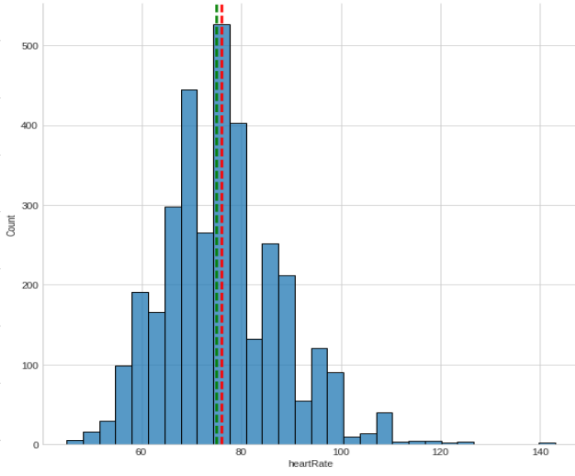


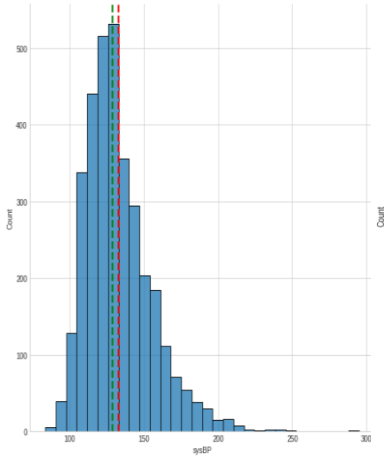
Numerical Features

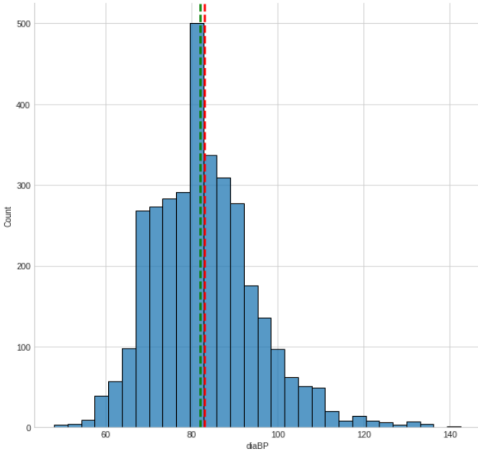


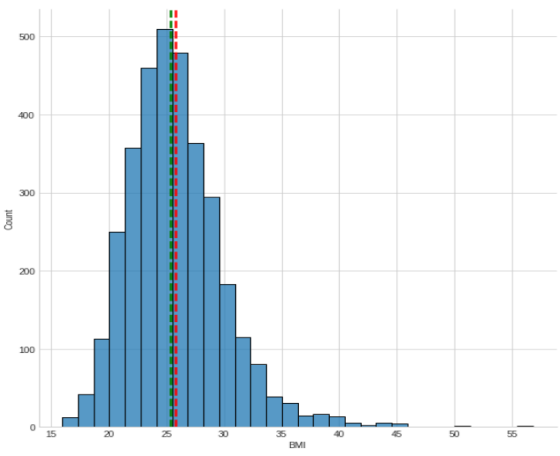


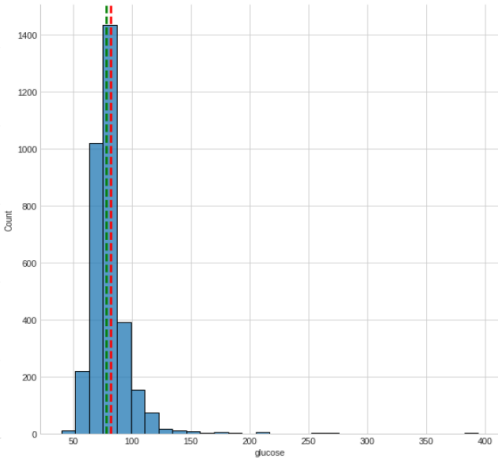




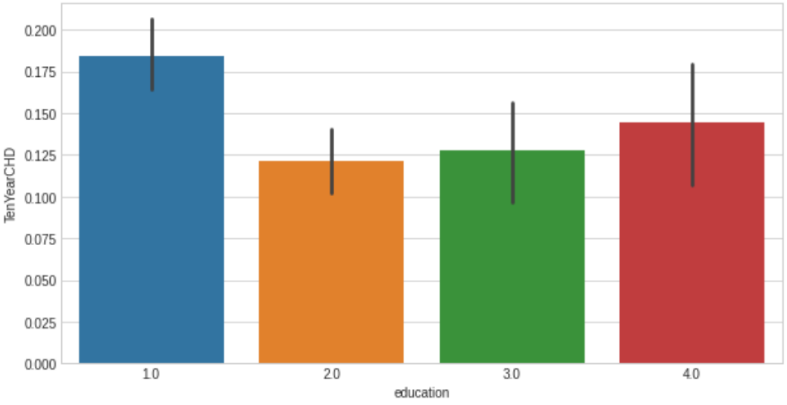


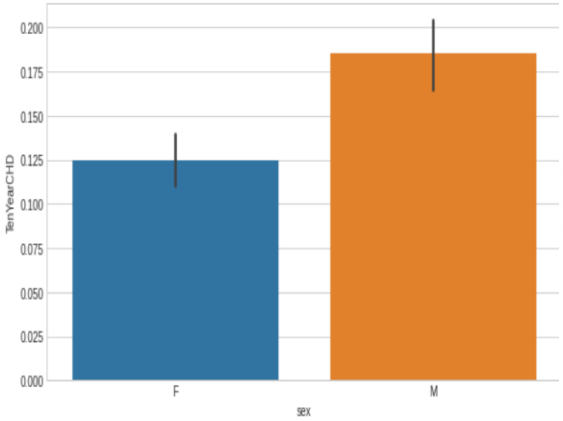


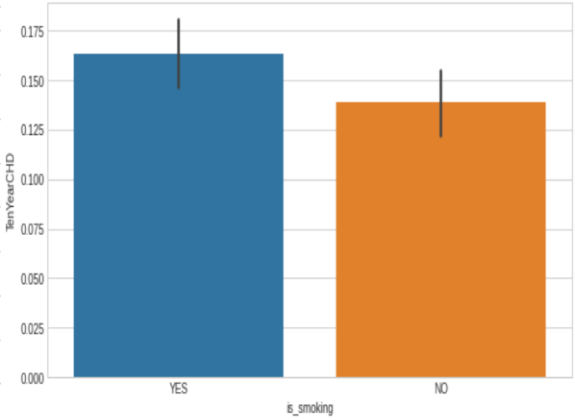


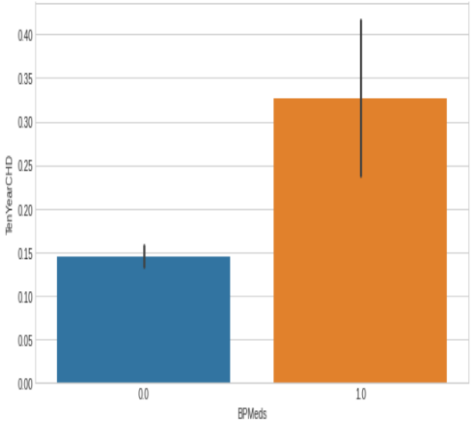


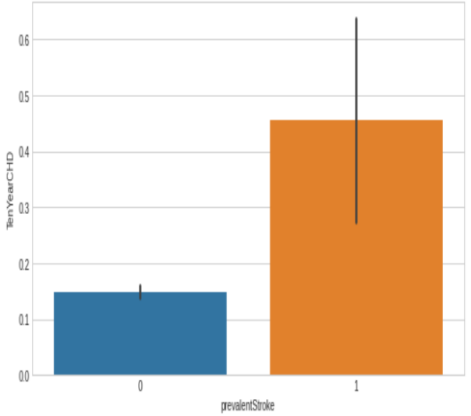
Categorical Features

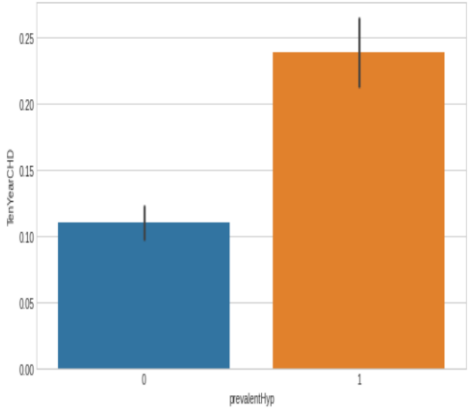


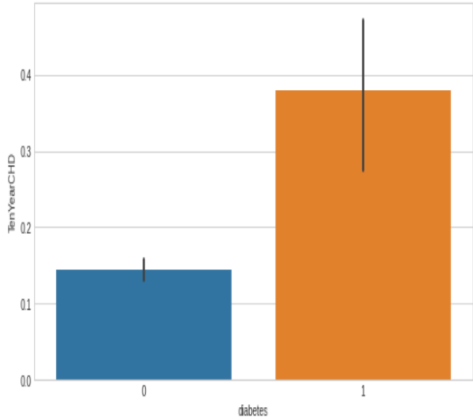




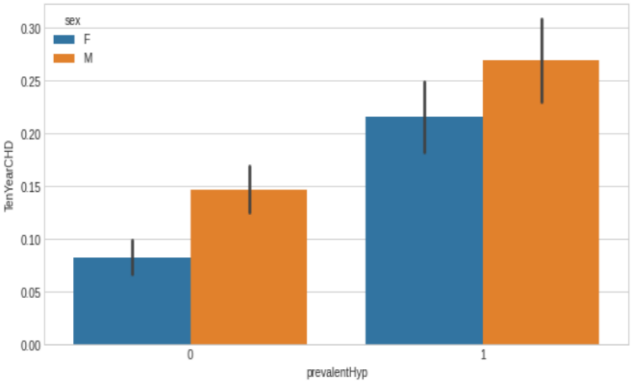


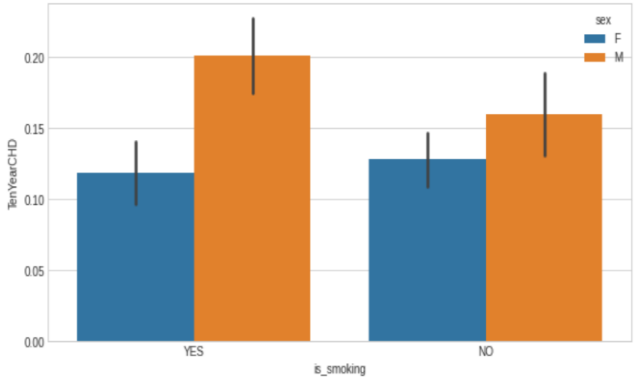


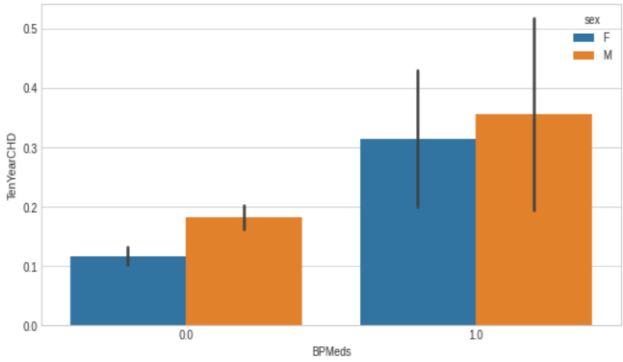


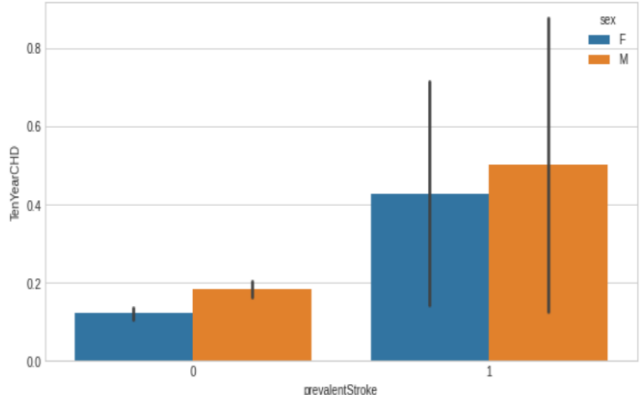


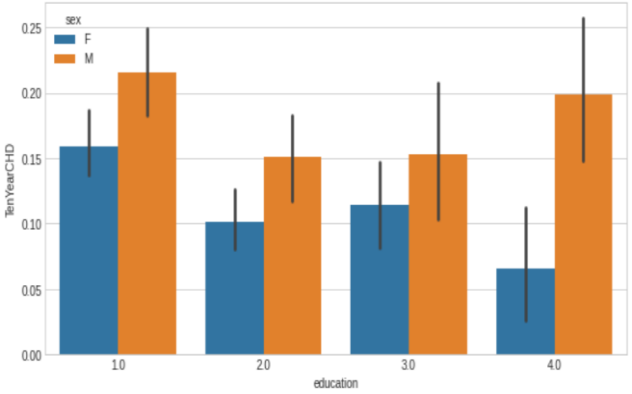
Male/Female

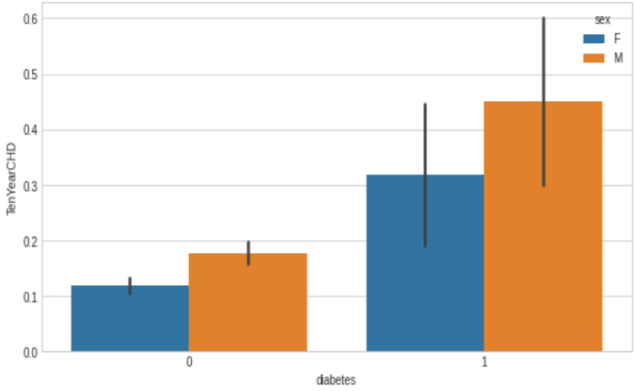

sex
F
M
TenYearCHD
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
prevalentHyp


sex
F
M
TenYearCHD
0.20
0.15
0.10
0.05
0.00
YES
NO
is_smoking

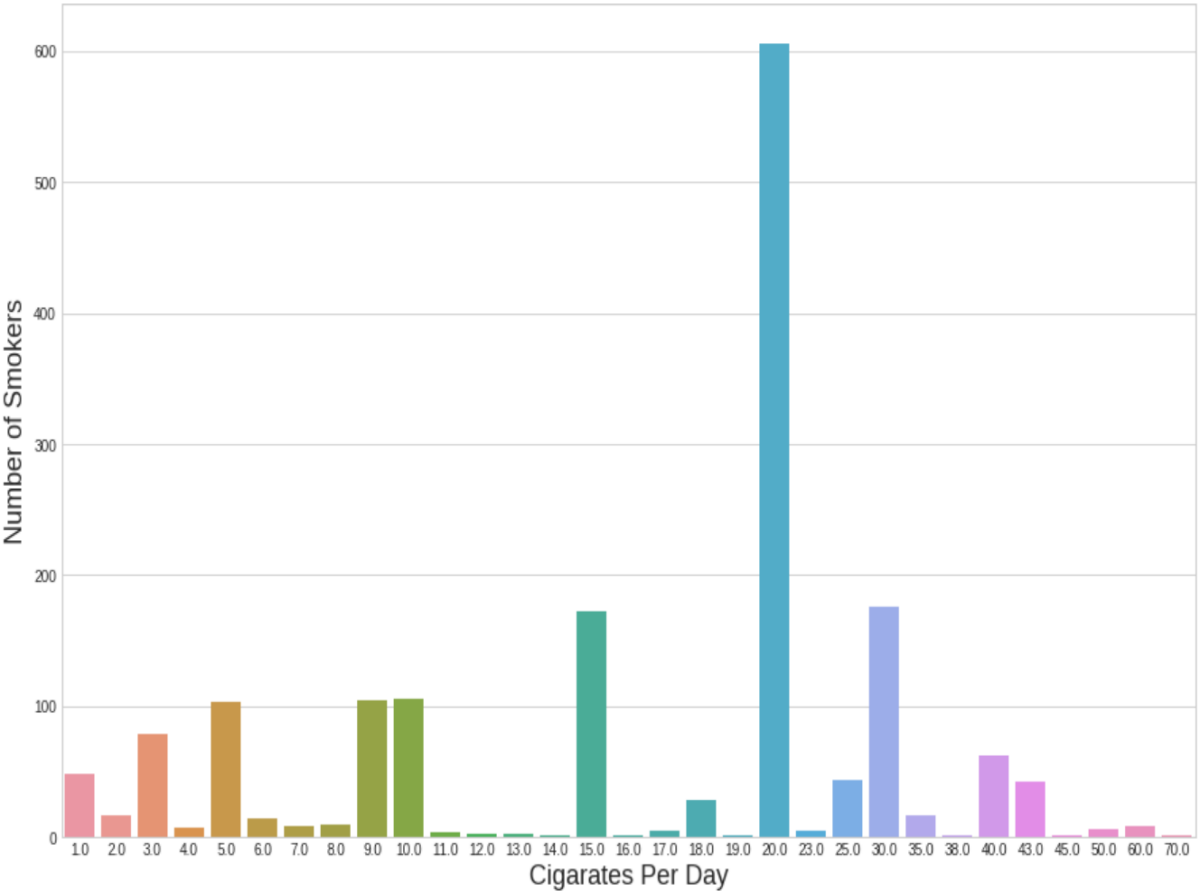

sex
F
M
TenYearCHD
0.5
0.4
0.3
0.2
0.1
0.0
0.0
1.0
BPMeds


sex
F
M
TenYearCHD
0.8
0.6
0.4
0.2
0.0
0
1
prevalentStroke


sex
F
M
TenYearCHD
0.25
0.20
0.15
0.10
0.05
0.00
1.0
2.0
3.0
4.0
education


sex
F
M
TenYearCHD
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
diabetes

Smoker's Graph



- The average cigarettes smoked by a smoker is: 18.34 per day.
- 20 cigarettes per day has the highest number of smokers.
- Number of cigarettes smoked by an individual goes to 70.

Who is/are at 10-year risk of coronary heart disease(CHD)?

	age_group	total_individuals	no_chd_total	yes_chd_total	risk_VS_noRisk_ratio
0	31-40	604	574	30	0.052265
1	41-50	1283	1144	139	0.121503
2	51-60	1041	834	207	0.248201
3	61-70	462	327	135	0.412844

Male or Female:

- 12.43 % Females
- 18.54 % Males

On BP Medication or Not on Medication

- 32.67 % of individuals on BP medication.
- 14.53 % of individuals not on BP medication.

Smoker or Non-Smoker

- 16.30 % Smokers
- 13.86 % Non-Smokers

Diabetic or Non-diabetic

- 37.93 % of diabetic individuals.
- 14.47 % of non-diabetic individuals.

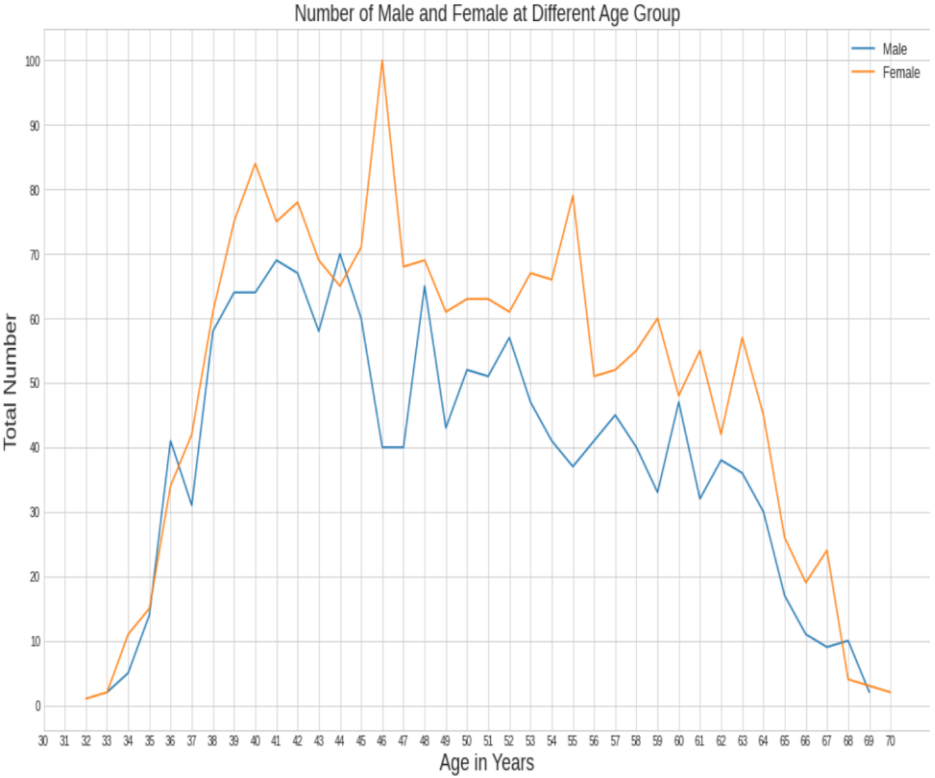
Prevalent Stroke condition or Not

- 45.45 % of individuals with prevalent stroke condition.
- 14.88 % of individuals with no prevalent stroke condition.

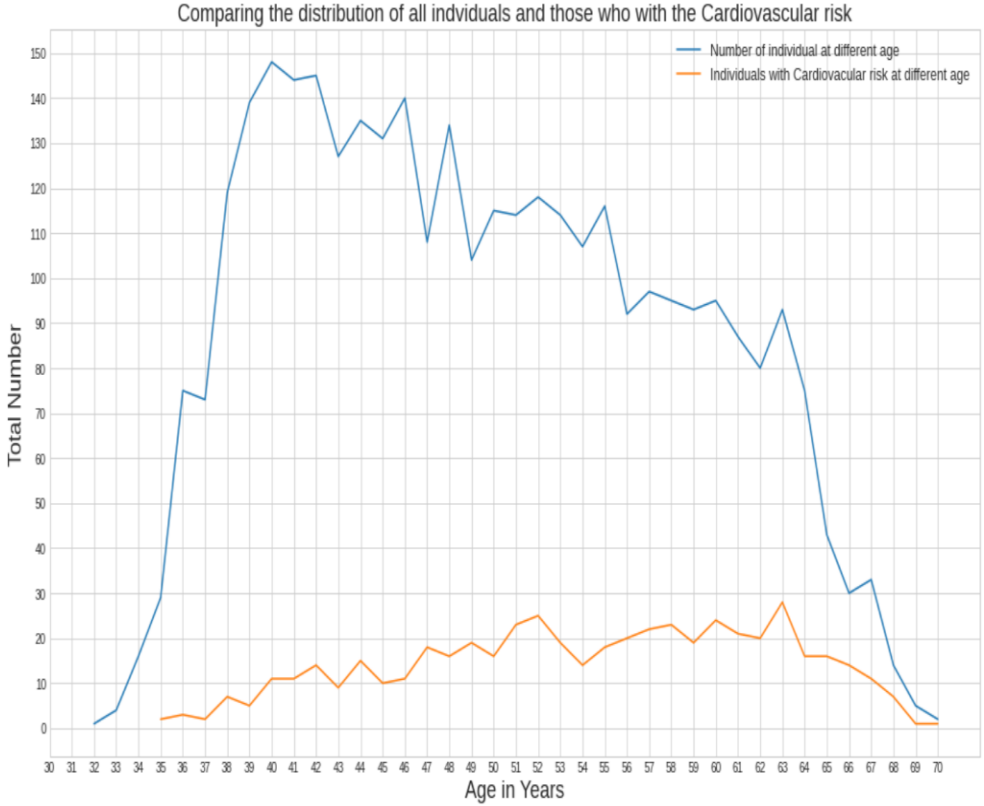
Prevalent Hypertension or Not

- 23.85 % of individuals with prevalent hypertension.
- 11.03 % of individuals with no prevalent hypertension.

Distribution of Data based on Age

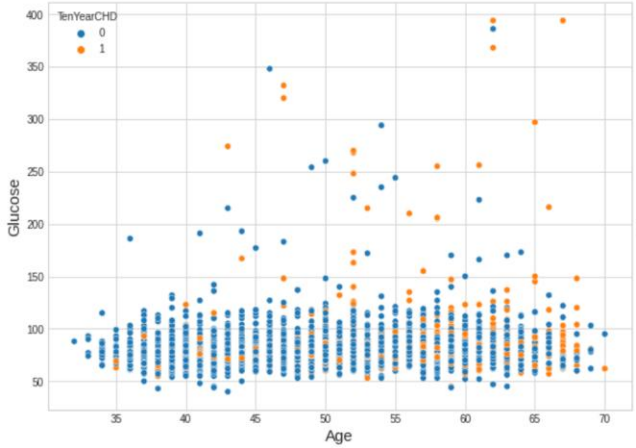


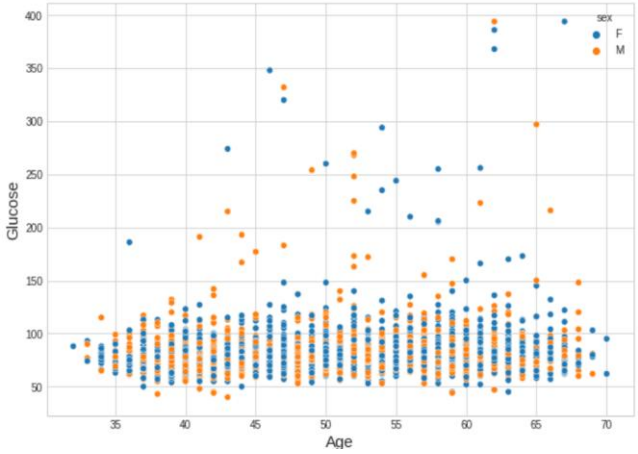
Medical record of males between the age 45-47 is highest, while of females it is not. Rest of the record for both Male and Female has almost same participation.

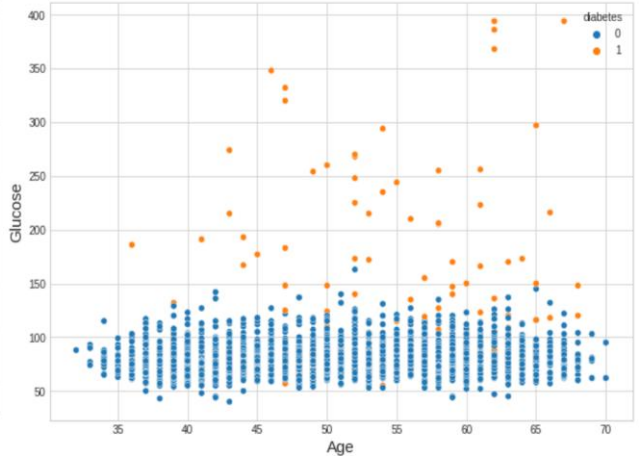


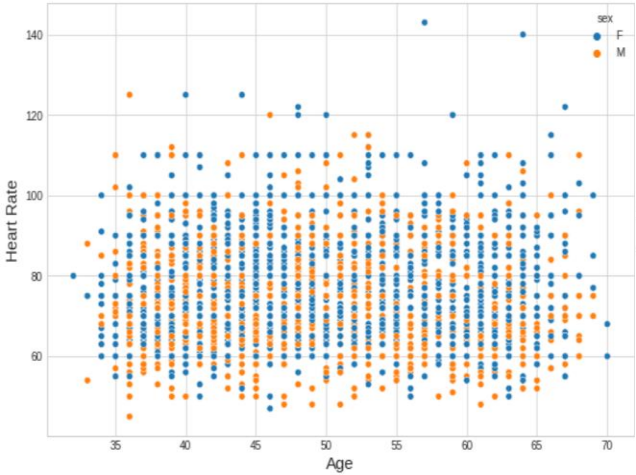
As the age increasing the gap between the both curves is decreasing.

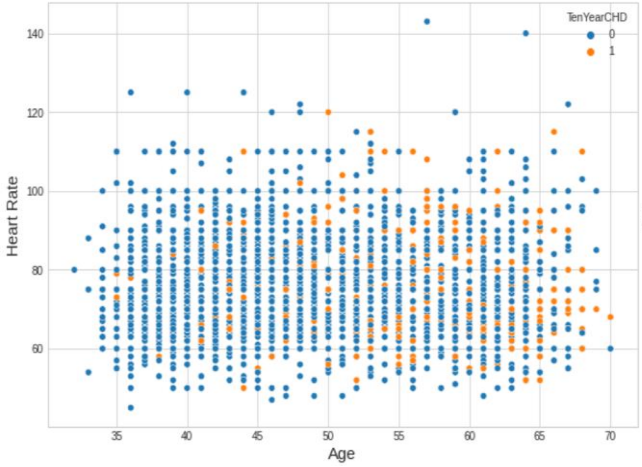
Scatter Plots to represent relation between different parameters with Age

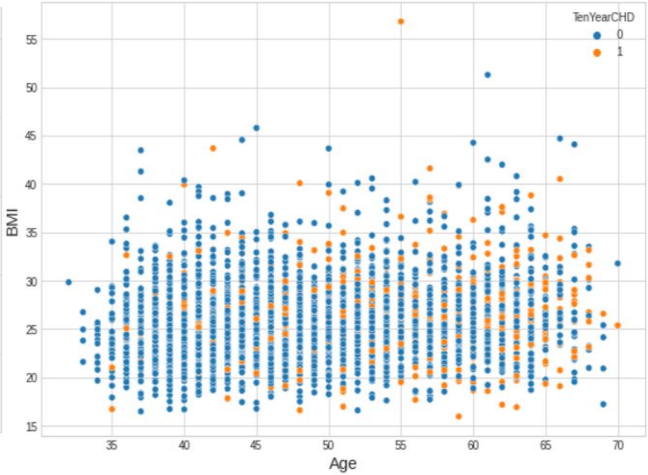




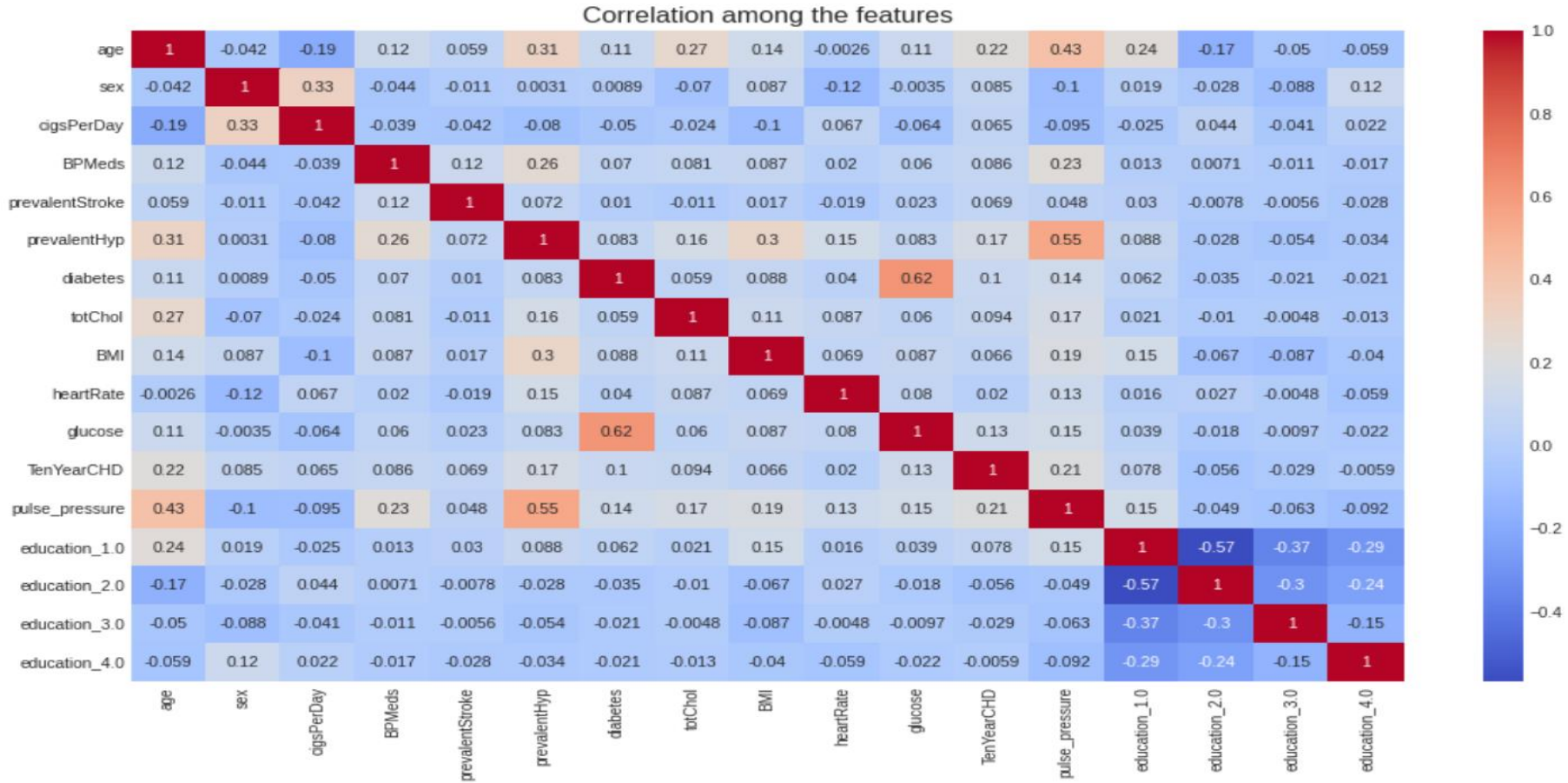








Heat Map



There is a slightly higher co-relation between ‘diabetes’ and ‘glucose’, ‘prevalentHyp’ and ‘pulse_pressure’.

Classification Algorithms Implementation

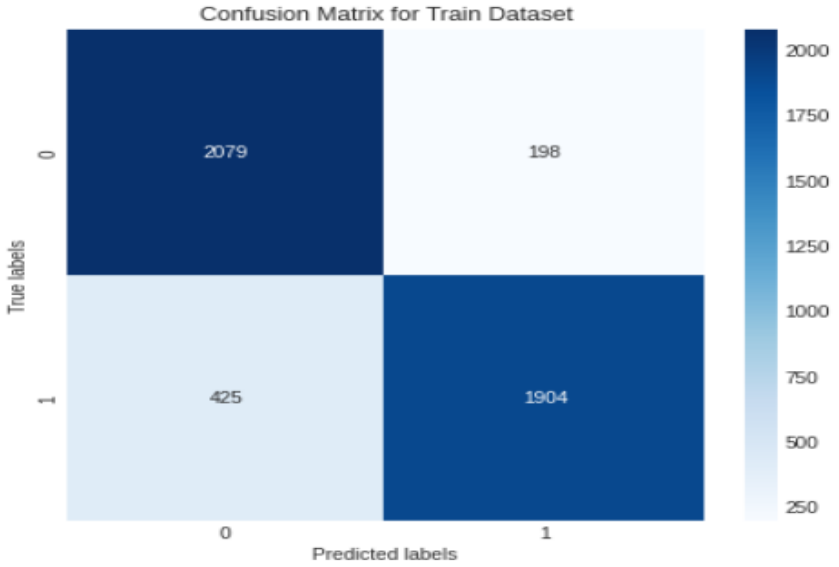
AI

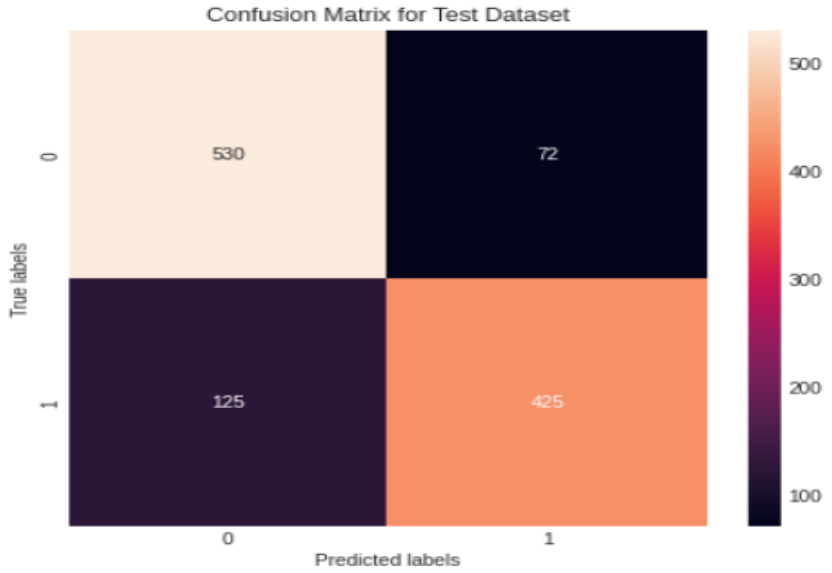
- K-Nearest Neighbors
- Decision Tress Classifier
- Logistic Regression Classifier
- Naïve Bayes Classifier
- Support Vector Classifier
- Random Forest Classifier

K-Nearest Neighbors

```
Train Accuracy = 0.8647416413373861
Test Accuracy = 0.8289930555555556
```

AI





Classification Report – Train Dataset

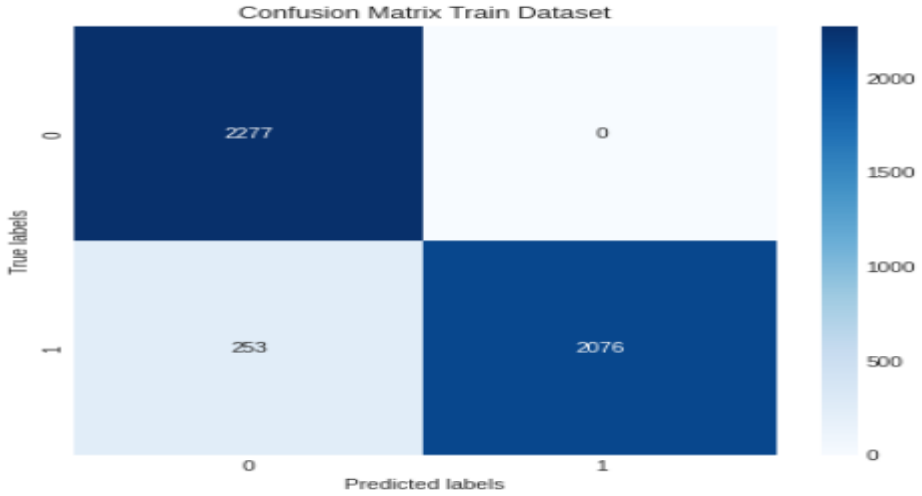
	precision	recall	f1-score	support
0	0.83	0.91	0.87	2277
1	0.91	0.82	0.86	2329
accuracy			0.86	4606
macro avg	0.87	0.87	0.86	4606
weighted avg	0.87	0.86	0.86	4606

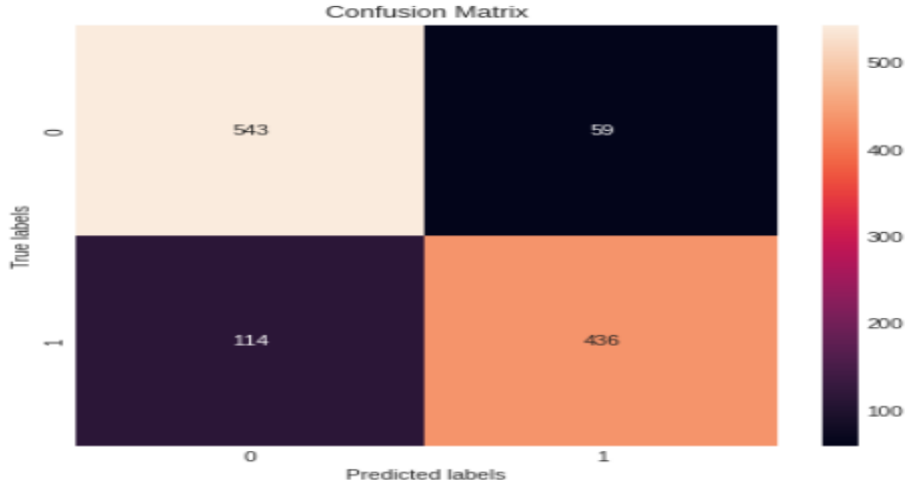
Classification Report – Test Dataset

	precision	recall	f1-score	support
0	0.81	0.88	0.84	602
1	0.86	0.77	0.81	550
accuracy			0.83	1152
macro avg	0.83	0.83	0.83	1152
weighted avg	0.83	0.83	0.83	1152

Hyper-parameter Tuning (K-Nearest Neighbors using GridSearchCV)

```
Train Accuracy = 0.9450716456795484
Test Accuracy = 0.8498263888888888
```





Classification Report – Train Dataset

	precision	recall	f1-score	support
0	0.90	1.00	0.95	2277
1	1.00	0.89	0.94	2329
accuracy			0.95	4606
macro avg	0.95	0.95	0.94	4606
weighted avg	0.95	0.95	0.94	4606

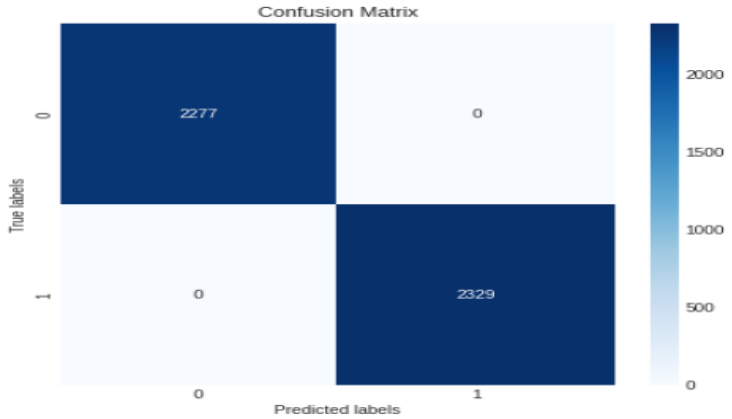
Classification Report – Test Dataset

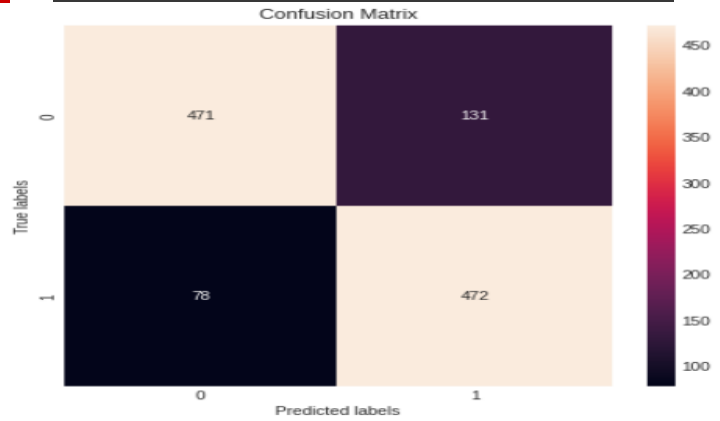
	precision	recall	f1-score	support
0	0.83	0.90	0.86	602
1	0.88	0.79	0.83	550
accuracy			0.85	1152
macro avg	0.85	0.85	0.85	1152
weighted avg	0.85	0.85	0.85	1152

Decision Tress Classifier

```
Train Accuracy = 1.0
Test Accuracy = 0.8185763888888888
```

AI



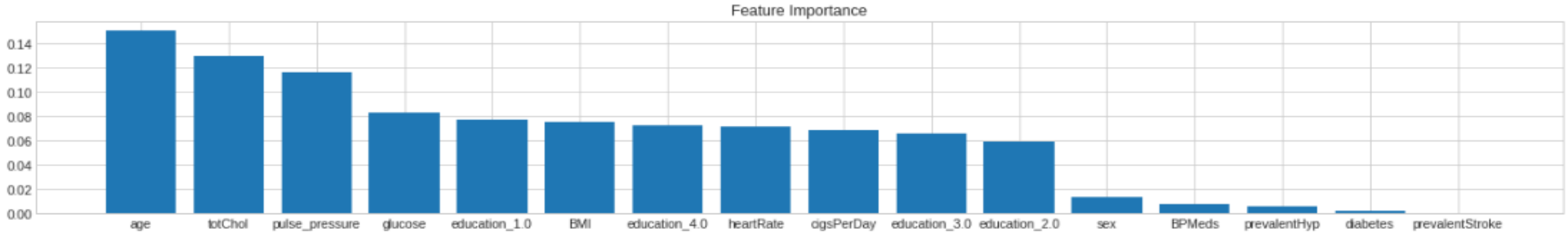


Classification Report – Train Dataset

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2277
1	1.00	1.00	1.00	2329
accuracy			1.00	4606
macro avg	1.00	1.00	1.00	4606
weighted avg	1.00	1.00	1.00	4606

Classification Report – Test Dataset

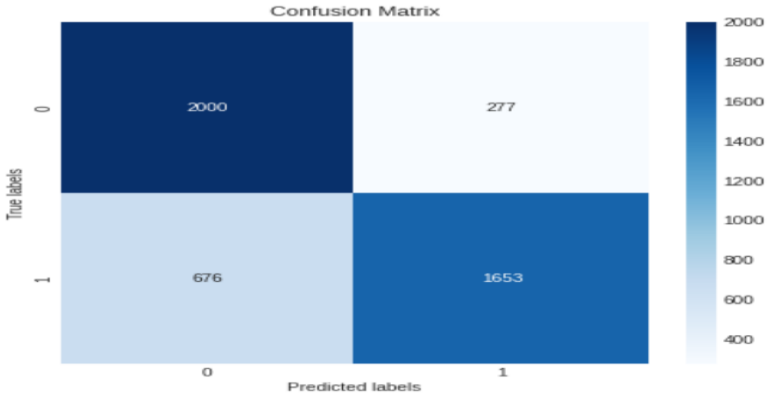
	precision	recall	f1-score	support
0	0.86	0.78	0.82	602
1	0.78	0.86	0.82	550
accuracy			0.82	1152
macro avg	0.82	0.82	0.82	1152
weighted avg	0.82	0.82	0.82	1152

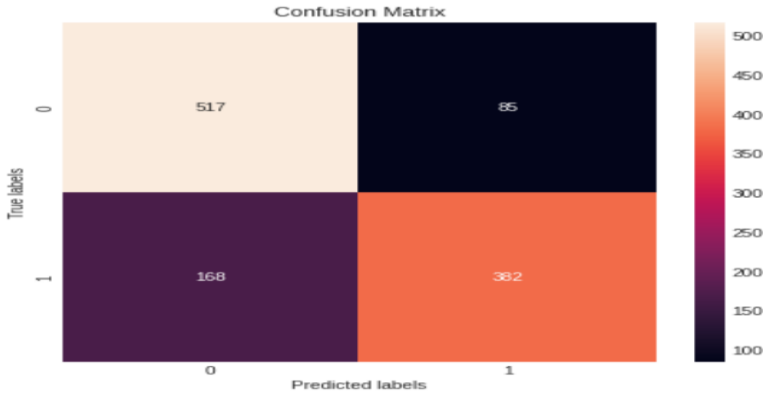


Logistic Regression Classifier

```
Train Accuracy = 0.7930959617889709
Test Accuracy = 0.7803819444444444
```

AI



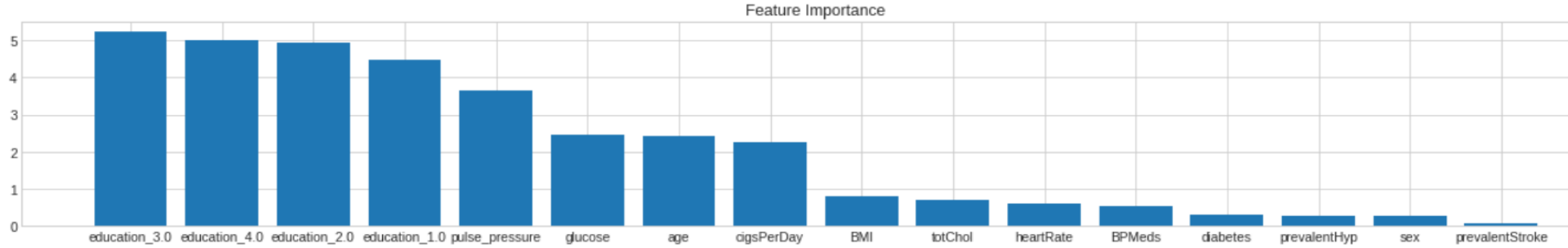


Classification Report – Train Dataset

	precision	recall	f1-score	support
0	0.75	0.88	0.81	2277
1	0.86	0.71	0.78	2329
accuracy			0.79	4606
macro avg	0.80	0.79	0.79	4606
weighted avg	0.80	0.79	0.79	4606

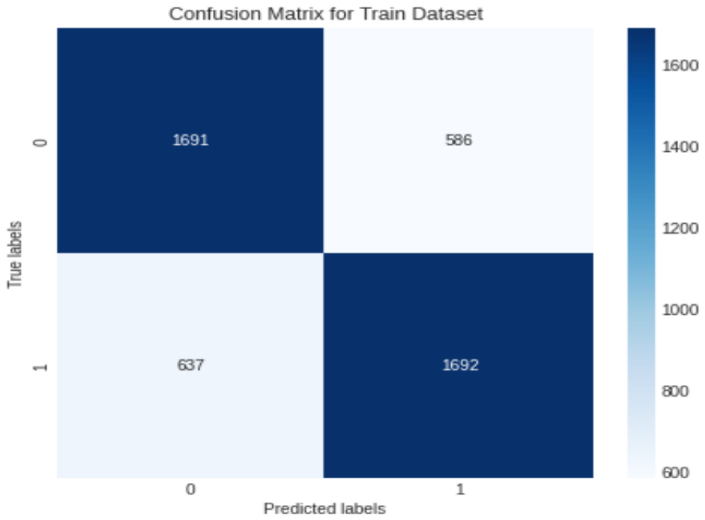
Classification Report – Test Dataset

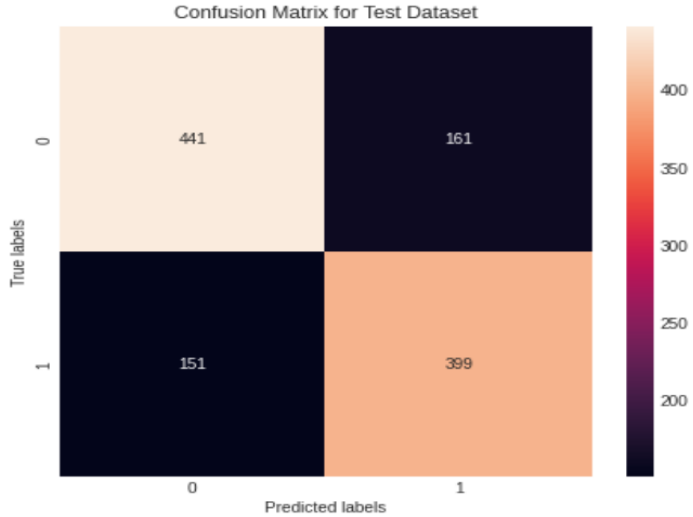
	precision	recall	f1-score	support
0	0.75	0.86	0.80	602
1	0.82	0.69	0.75	550
accuracy			0.78	1152
macro avg	0.79	0.78	0.78	1152
weighted avg	0.78	0.78	0.78	1152



Naive Bayes Classifier

```
Train Accuracy = 0.7344767694311767
Test Accuracy = 0.7291666666666666
```





Classification Report – Train Dataset

	precision	recall	f1-score	support
0	0.73	0.74	0.73	2277
1	0.74	0.73	0.73	2329
accuracy			0.73	4606
macro avg	0.73	0.73	0.73	4606
weighted avg	0.73	0.73	0.73	4606

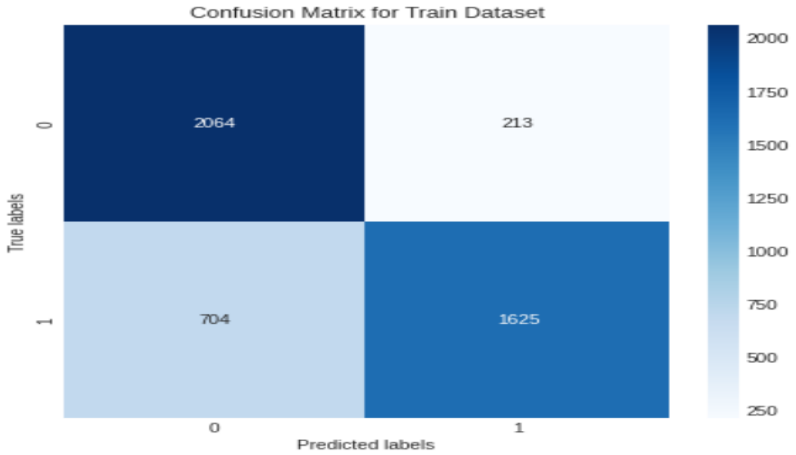
Classification Report – Test Dataset

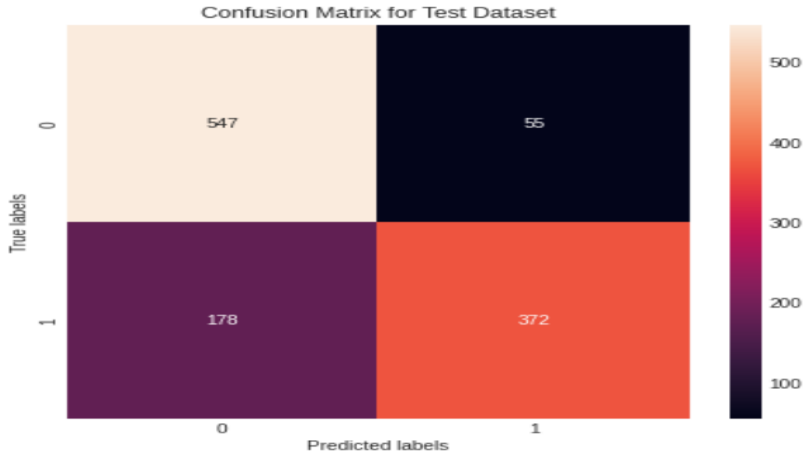
	precision	recall	f1-score	support
0	0.74	0.73	0.74	602
1	0.71	0.73	0.72	550
accuracy			0.73	1152
macro avg	0.73	0.73	0.73	1152
weighted avg	0.73	0.73	0.73	1152

Support Vector Classifier (SVC)

```
Train Accuracy = 0.8009118541033434
Test Accuracy = 0.7977430555555556
```

AI





Classification Report – Train Dataset

	precision	recall	f1-score	support
0	0.75	0.91	0.82	2277
1	0.88	0.70	0.78	2329
accuracy			0.80	4606
macro avg	0.81	0.80	0.80	4606
weighted avg	0.82	0.80	0.80	4606

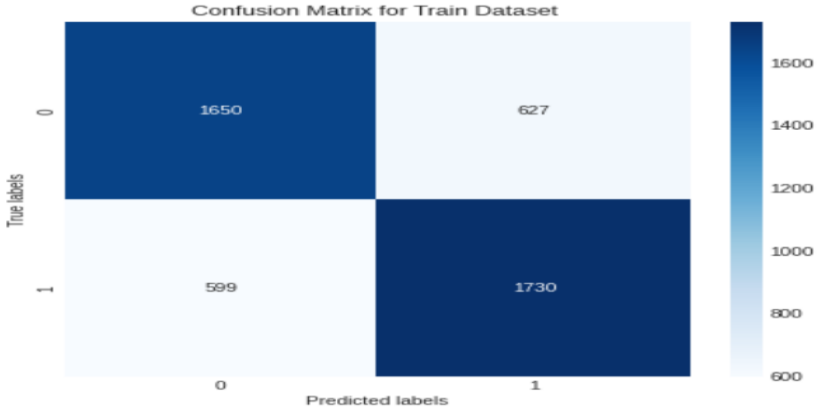
Classification Report – Test Dataset

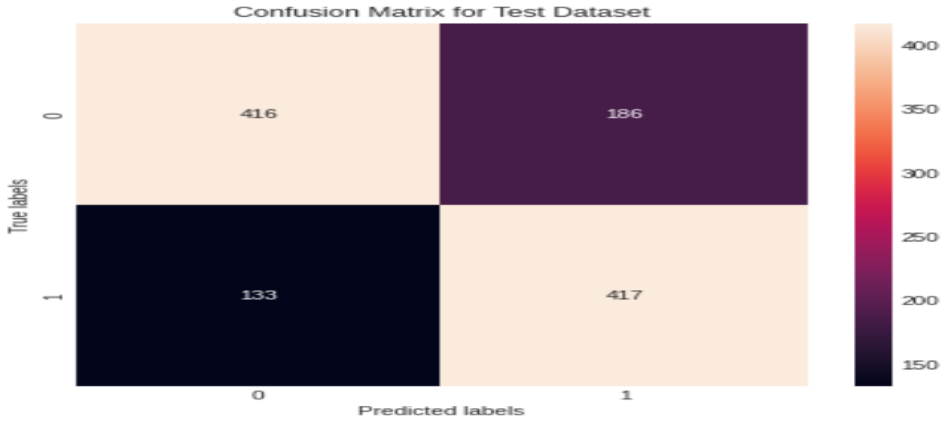
	precision	recall	f1-score	support
0	0.75	0.91	0.82	602
1	0.87	0.68	0.76	550
accuracy			0.80	1152
macro avg	0.81	0.79	0.79	1152
weighted avg	0.81	0.80	0.79	1152

Random Forest Classifier

```
Train Accuracy = 0.7338254450716457
Test Accuracy = 0.7230902777777778
```

AI



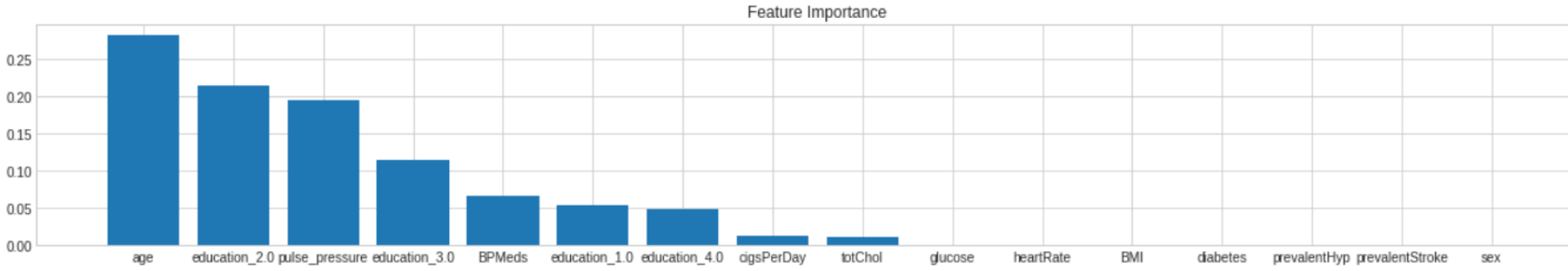


Classification Report – Train Dataset

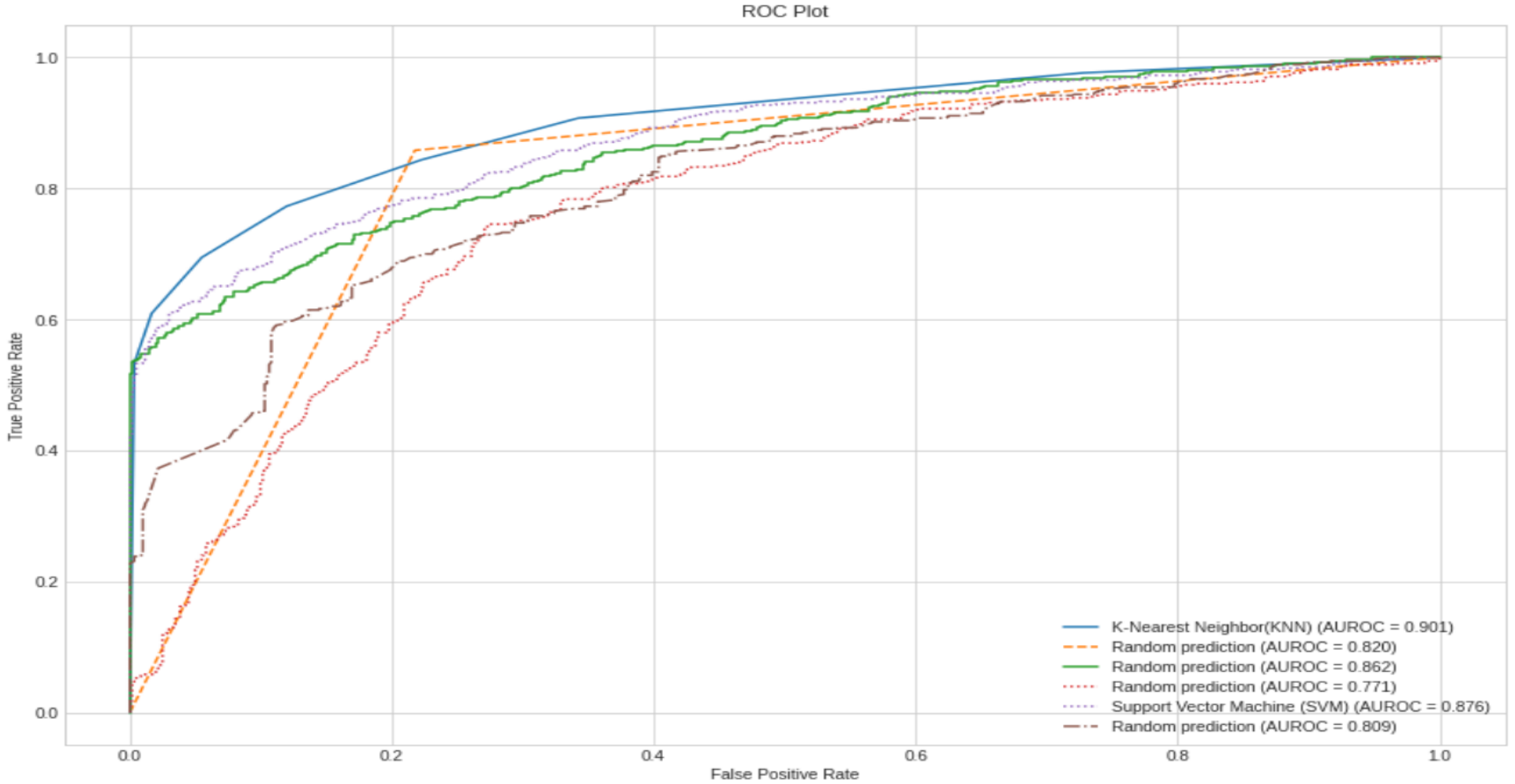
	precision	recall	f1-score	support
0	0.73	0.72	0.73	2277
1	0.73	0.74	0.74	2329
accuracy			0.73	4606
macro avg	0.73	0.73	0.73	4606
weighted avg	0.73	0.73	0.73	4606

Classification Report – Test Dataset

	precision	recall	f1-score	support
0	0.76	0.69	0.72	602
1	0.69	0.76	0.72	550
accuracy			0.72	1152
macro avg	0.72	0.72	0.72	1152
weighted avg	0.73	0.72	0.72	1152



ROC_AUC (Curve)



Challenges

AI

- Despite being a relatively small dataset, it has nearly 10% of rows with at least one null/nan value.
- The project requires good domain knowledge of medical field especially while dealing with outliers and while doing feature engineering.
- Some features are interconnected so the feature selection was challenging.
- There is information about ‘education’ column of the dataset, yet it is an important feature to train the model.
- Large graphical representation was required.
- The dataset has some rare data points e.g. high glucose level but no diabetes, high heart rate, BP (systolic and diastolic) but no prior and future cardiac risk associated with it, number of cigarettes smoked per day by an individual goes to 70.
- Different algorithms were giving different performance based on the different matrix.

Conclusion

- For covering maximum number of patients who could suffer from any cardiovascular risk in future, a high recall value is required for class 1 in the test dataset. But it will also include few such patient who may not face any risk under (false negative) prediction.
- Under high precision for class 1 several patients who don't require any treatment would be categorized as patients, while several actual potential patients will be left out.
- In the case of predicting Cardiovascular and such life threatening diseases model with high recall value is preferred.
- Best performance of Models on test dataset for class 1 (Risk of TenYearCHD):

1. Recall - Decision Tree
2. Precision - SVC
3. F1 Score - Decision Tree
4. Accuracy - K-Nearest Neighbor(KNN)

Best AUROC Score: K-Nearest Neighbor(KNN)

Thank You