Capstone Project-3

CARDIOVASCULAR RISK PREDICTION

by: BHARATH KUMAR A

Contents

- Problem Statement
- Data Summary
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Modelling Approach
- Model Comparison
- Conclusions

Problem Statement

▪ The dataset is from an ongoing cardiovascular study on residents of the town of 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 approx.4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Data Summary

Demographic:

- Sex: male or female("M" or "F")
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioral:

- is_smoking: whether or not the patient is a current smoker ("YES" or "NO")
- Cigs Per Day: the 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(previous history):

- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether a patient previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient has hypertension (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)

Data Summary

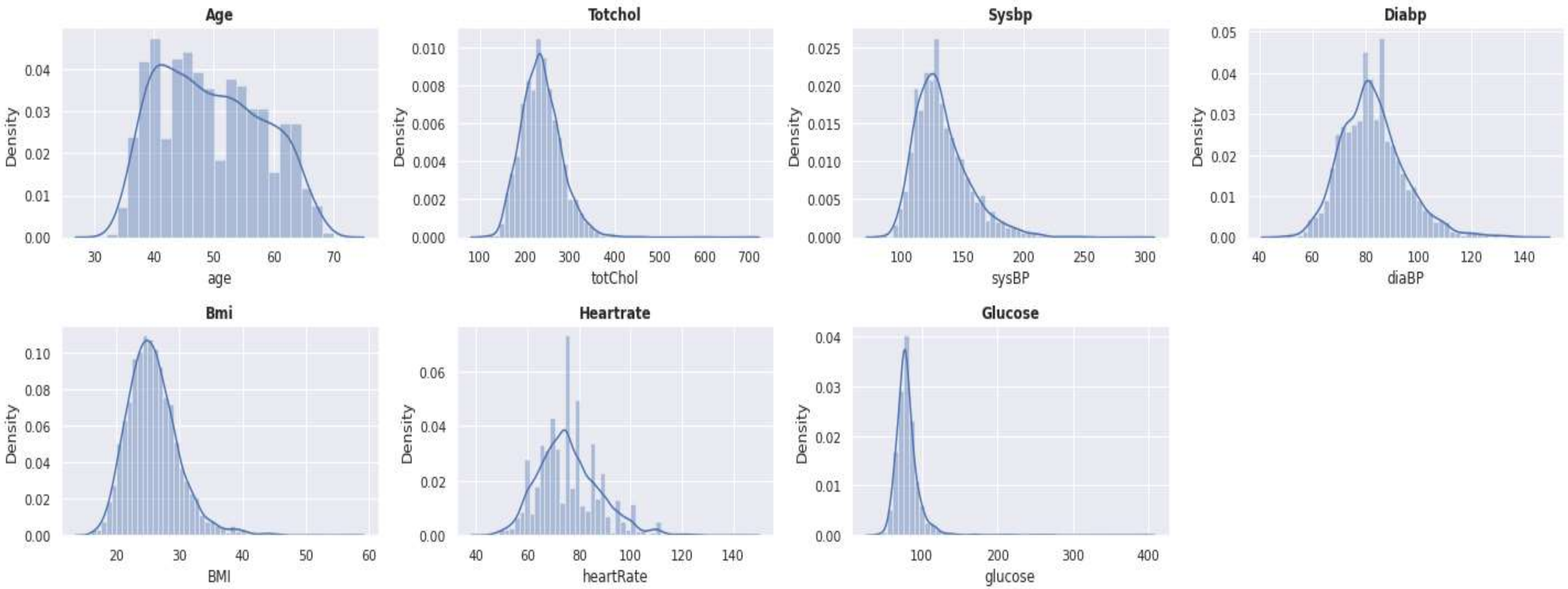
Medical(Current):

- 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):

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

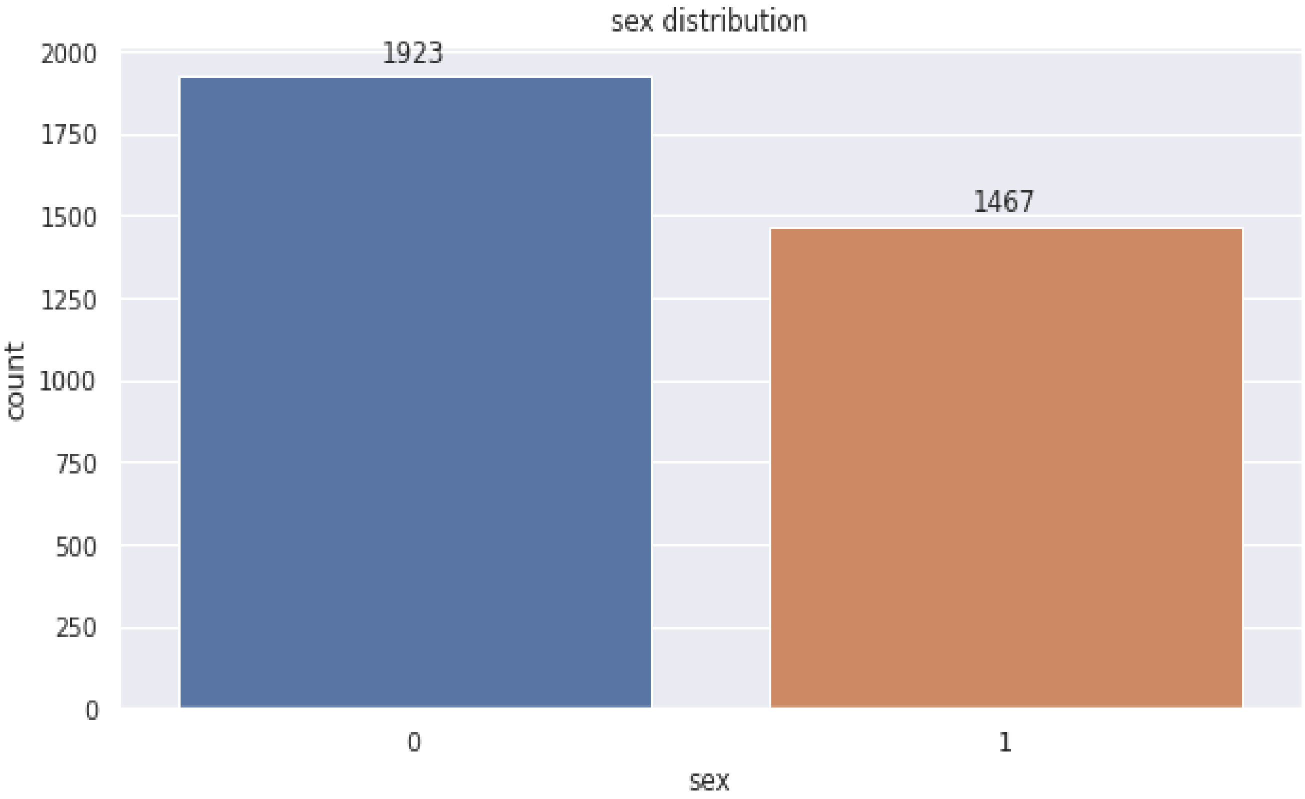
Exploratory Data Analysis (EDA)



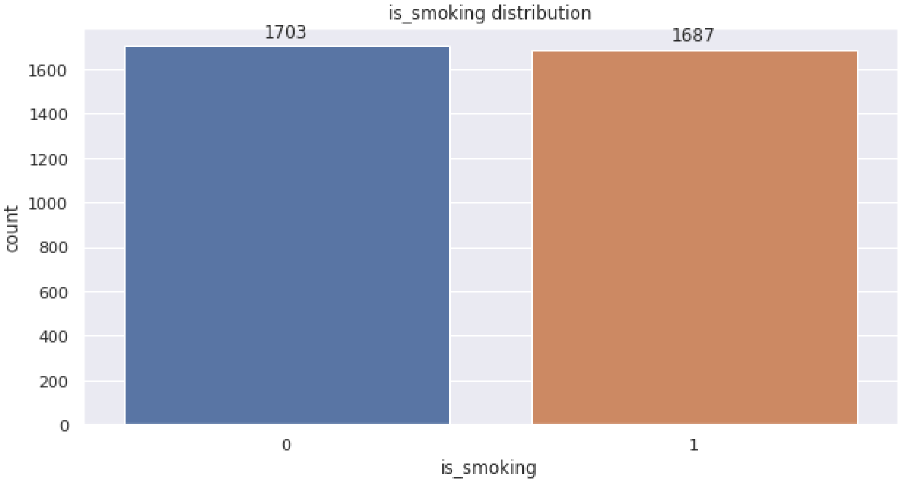
Distribution of continuous variables

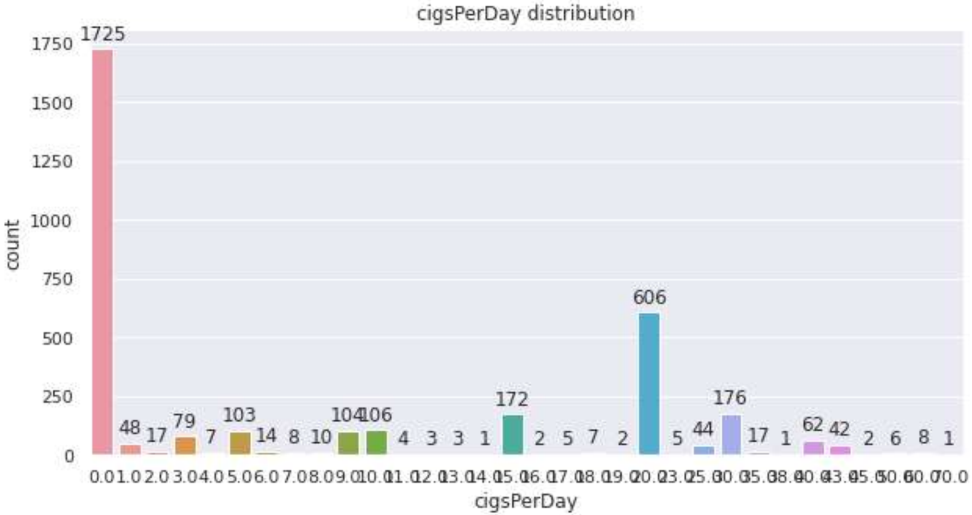
Gender Distribution

In this dataset there Are 1467 males and 1923 females. Clearly there are More females then males



Smoking and Cigerates per day

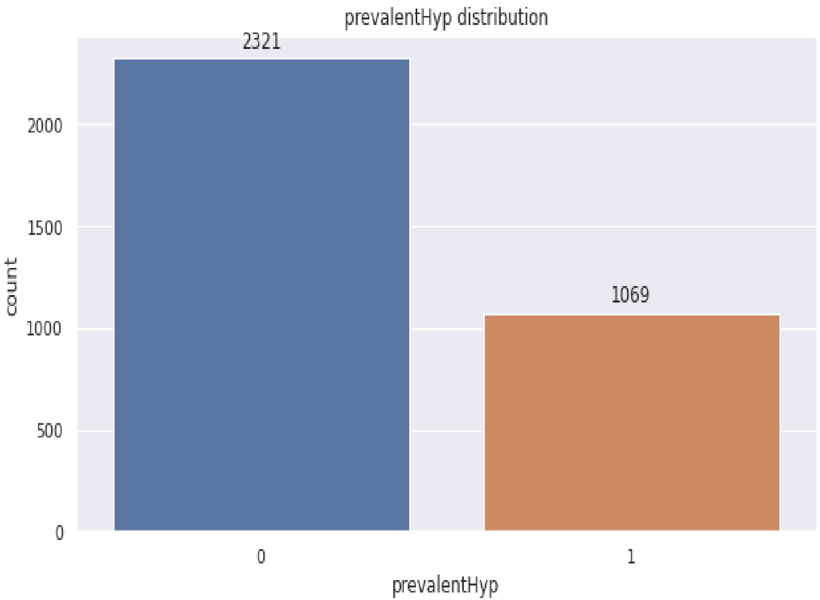


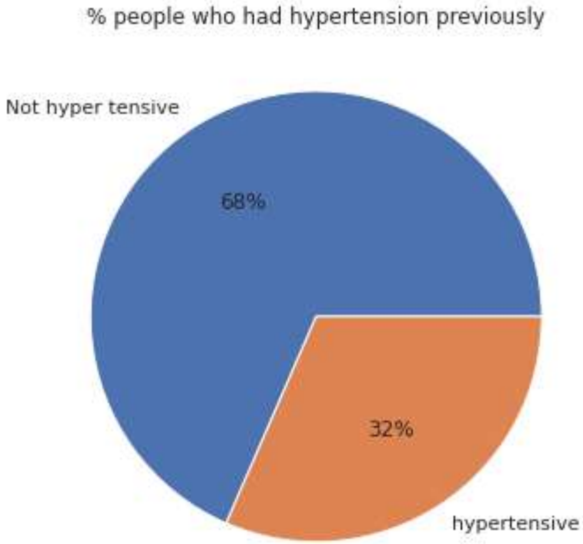


Total number of people smoking is 1687, and people who are not smoking are 1703

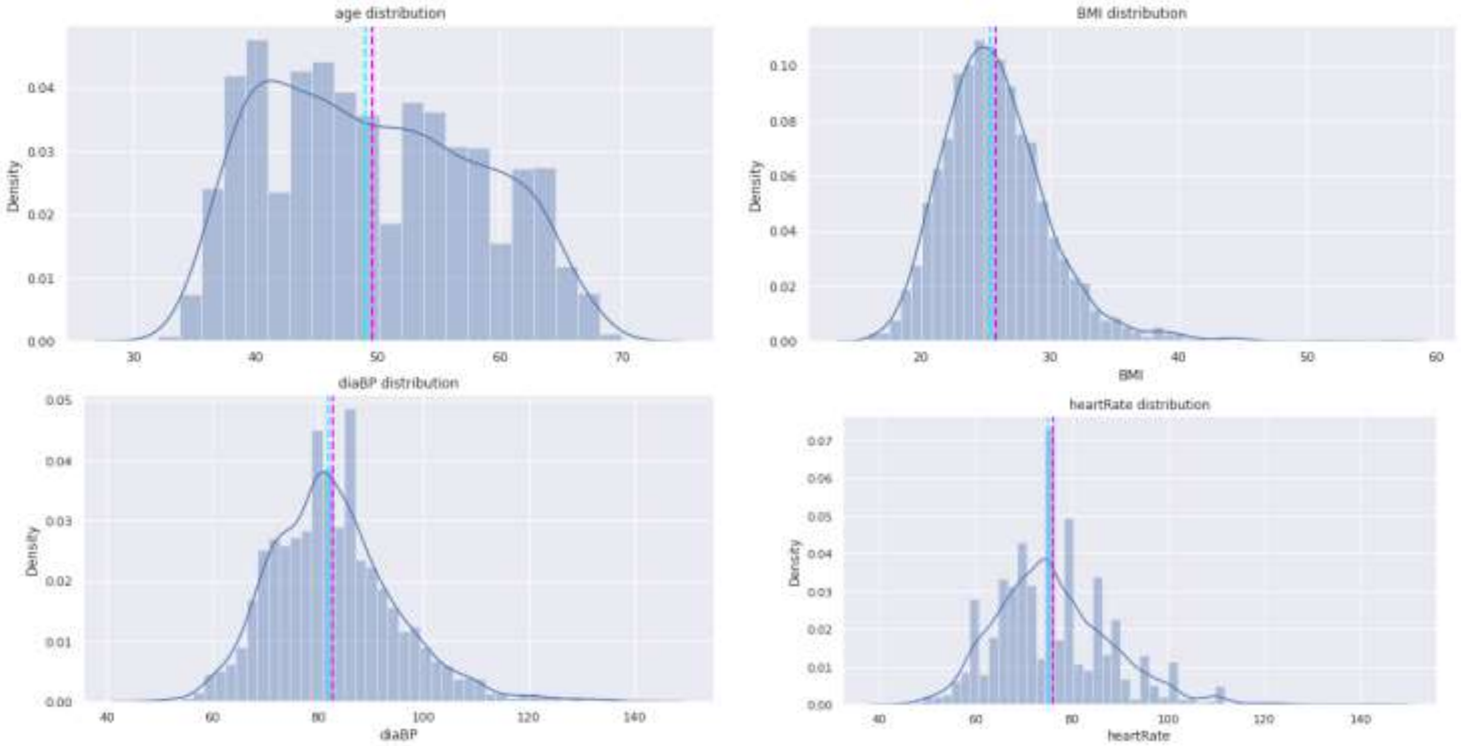
From the second graph we can observe that 1725 people smoke one cigerates per day and 606 people smoke 2 cigerated per day

Hyper tension



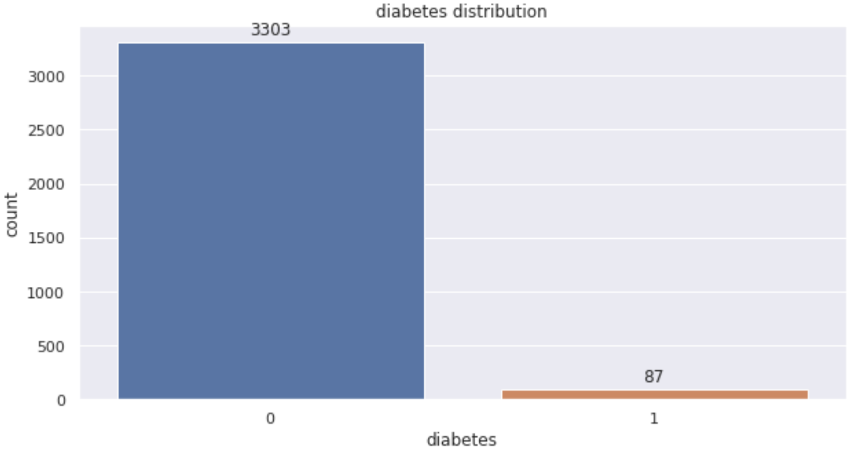


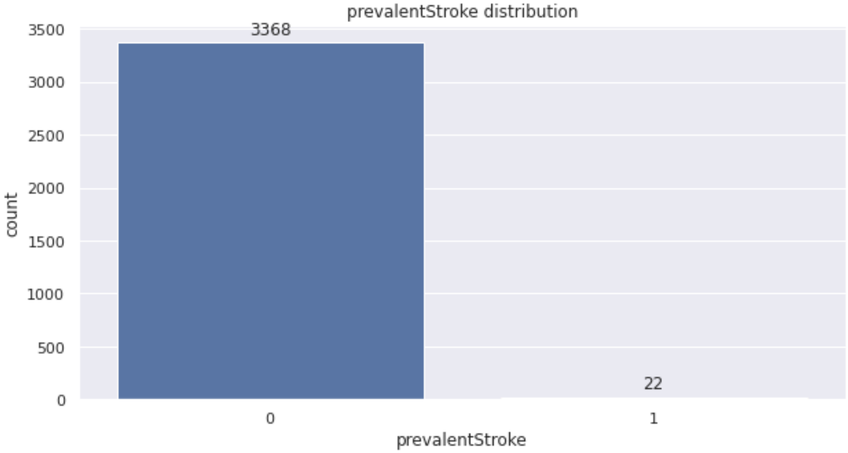
There the bargraph there are 2321 people with hypertension i.e, 68% from the data, 1069 people do not have hypertension i.e, 32%

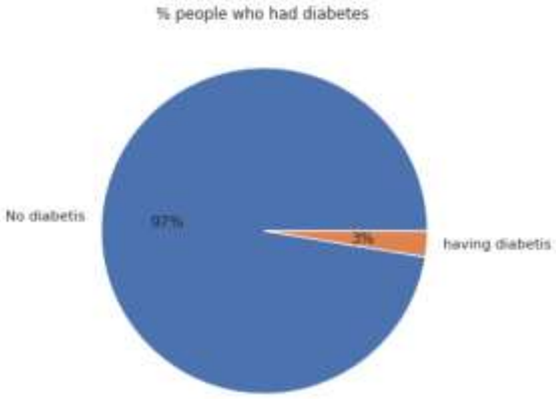


- Age distribution ranges from 32-70 years, bmi distribution ranges from 15-40,
- Bp distribution ranges from 60-140, heart rate distribution ranges from 45-120

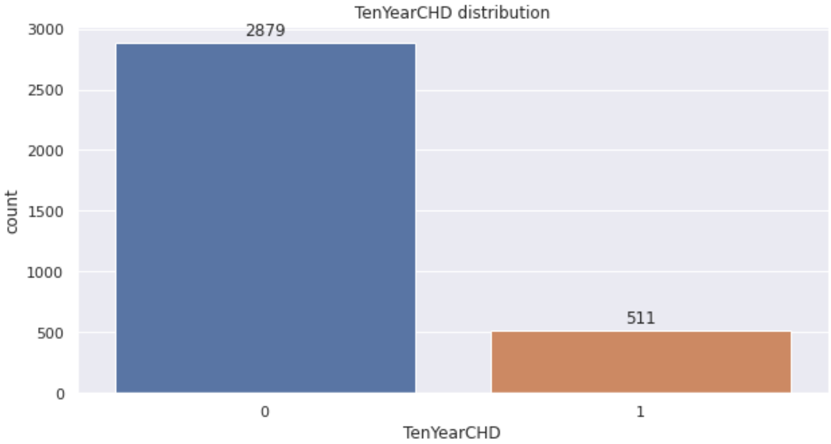
Diabetes and stroke

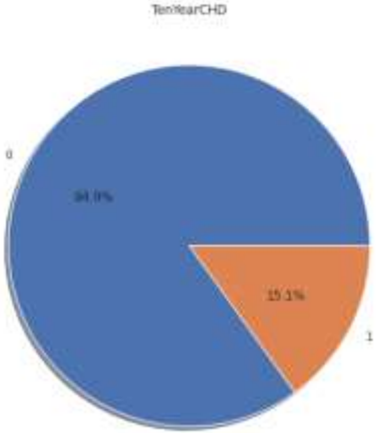


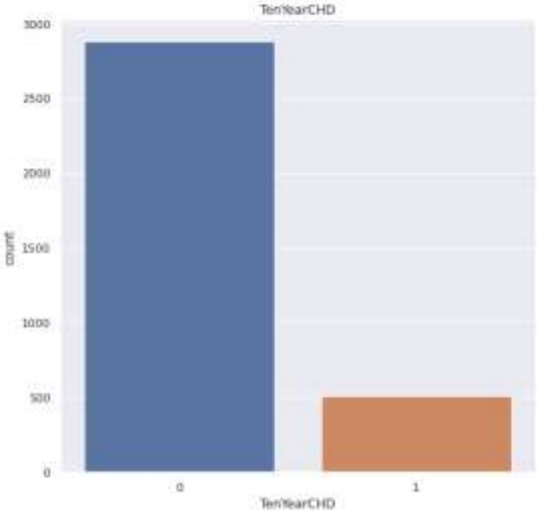


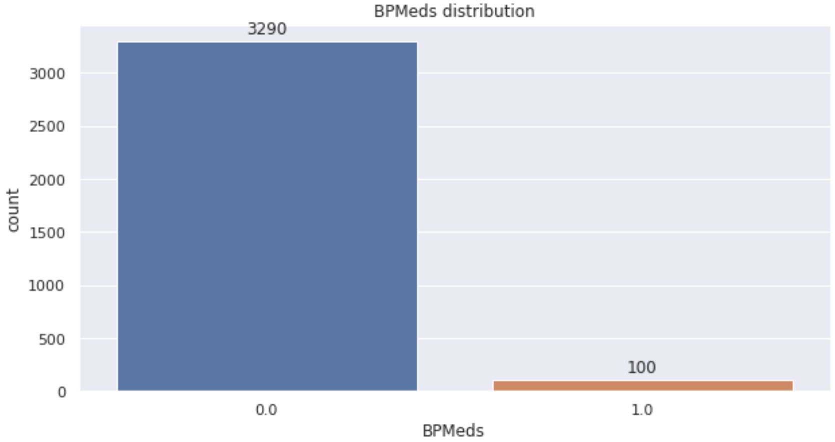


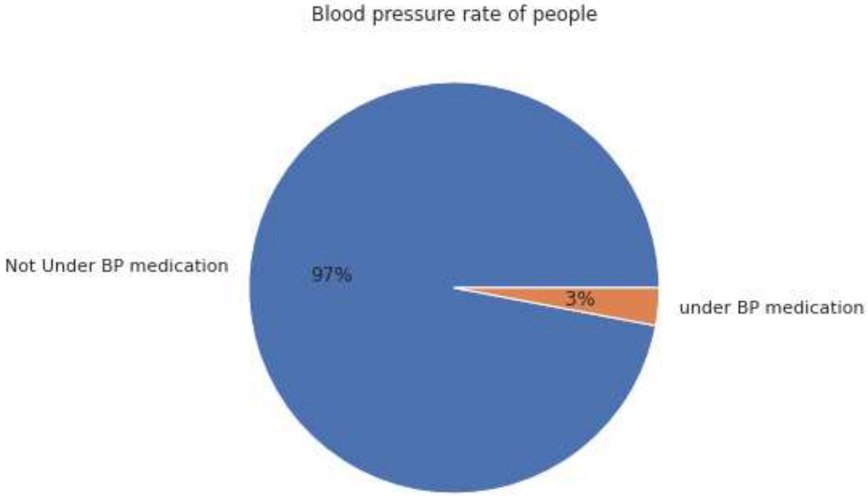
87 people with diabetes,22 people with stroke


TenYearCHD distribution
3000
2500
2000
1500
1000
500
0
count
2879
511
0
1
TenYearCHD

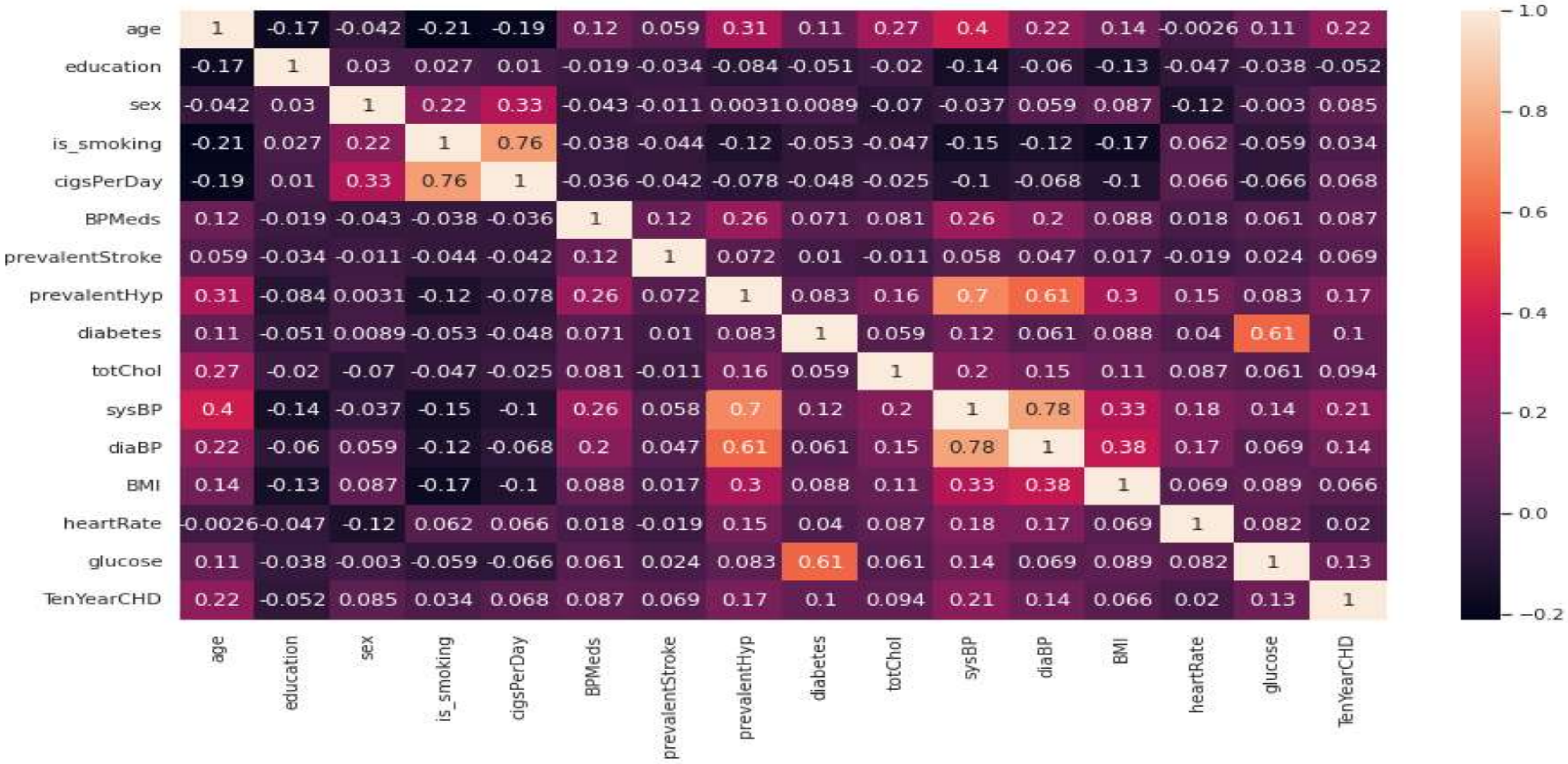

TenYearCHD
0
84.9%
15.1%
1


TenYearCHD
3000
2500
2000
1500
1000
500
0
count
0
1
TenYearCHD

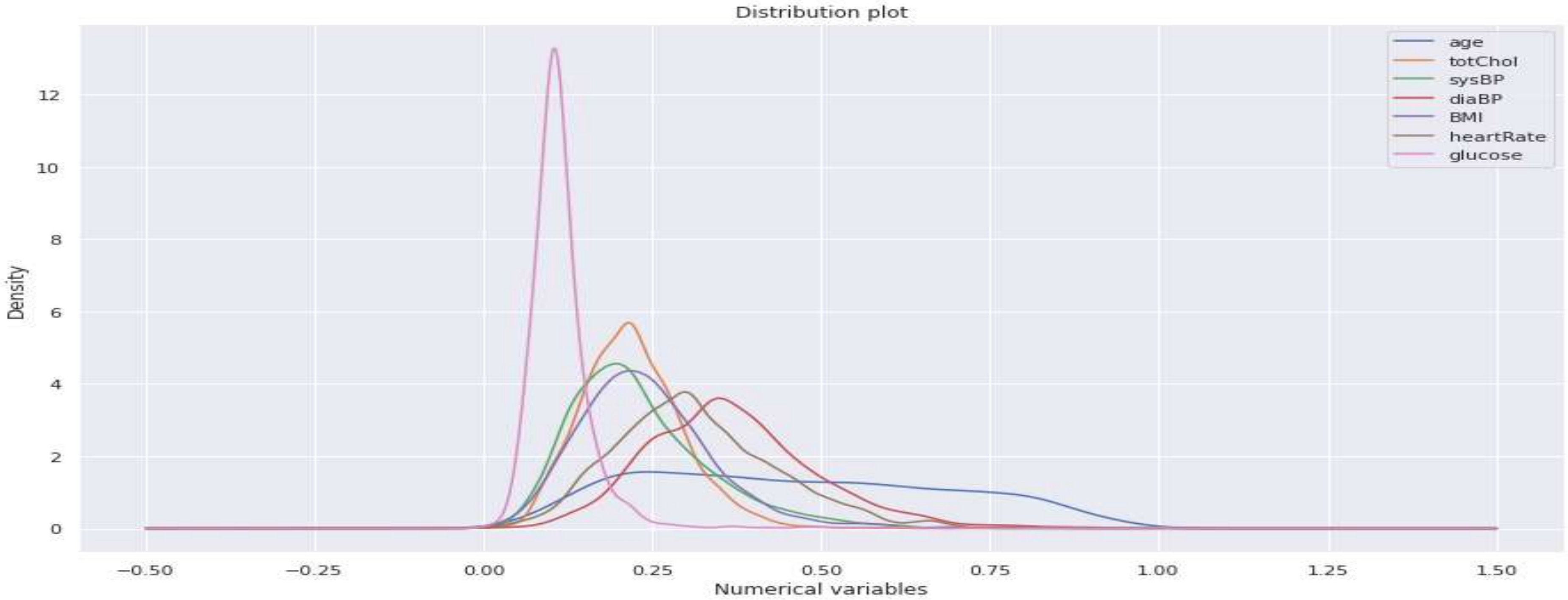

BPMeds distribution


Blood pressure rate of people

Correlation



	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	-0.17	-0.042	-0.21	-0.19	0.12	0.059	0.31	0.11	0.27	0.4	0.22	0.14	-0.0026	0.11	0.22
education	-0.17	1	0.03	0.027	0.01	-0.019	-0.034	-0.084	-0.051	-0.02	-0.14	-0.06	-0.13	-0.047	-0.038	-0.052
sex	-0.042	0.03	1	0.22	0.33	-0.043	-0.011	0.0031	0.0089	-0.07	-0.037	0.059	0.087	-0.12	-0.003	0.085
is_smoking	-0.21	0.027	0.22	1	0.76	-0.038	-0.044	-0.12	-0.053	-0.047	-0.15	-0.12	-0.17	0.062	-0.059	0.034
cigsPerDay	-0.19	0.01	0.33	0.76	1	-0.036	-0.042	-0.078	-0.048	-0.025	-0.1	-0.068	-0.1	0.066	-0.066	0.068
BPMeds	0.12	-0.019	-0.043	-0.038	-0.036	1	0.12	0.26	0.071	0.081	0.26	0.2	0.088	0.018	0.061	0.087
prevalentStroke	0.059	-0.034	-0.011	-0.044	-0.042	0.12	1	0.072	0.01	-0.011	0.058	0.047	0.017	-0.019	0.024	0.069
prevalentHyp	0.31	-0.084	0.0031	-0.12	-0.078	0.26	0.072	1	0.083	0.16	0.7	0.61	0.3	0.15	0.083	0.17
diabetes	0.11	-0.051	0.0089	-0.053	-0.048	0.071	0.01	0.083	1	0.059	0.12	0.061	0.088	0.04	0.61	0.1
totChol	0.27	-0.02	-0.07	-0.047	-0.025	0.081	-0.011	0.16	0.059	1	0.2	0.15	0.11	0.087	0.061	0.094
sysBP	0.4	-0.14	-0.037	-0.15	-0.1	0.26	0.058	0.7	0.12	0.2	1	0.78	0.33	0.18	0.14	0.21
diaBP	0.22	-0.06	0.059	-0.12	-0.068	0.2	0.047	0.61	0.061	0.15	0.78	1	0.38	0.17	0.069	0.14
BMI	0.14	-0.13	0.087	-0.17	-0.1	0.088	0.017	0.3	0.088	0.11	0.33	0.38	1	0.069	0.089	0.066
heartRate	-0.0026	-0.047	-0.12	0.062	0.066	0.018	-0.019	0.15	0.04	0.087	0.18	0.17	0.069	1	0.082	0.02
glucose	0.11	-0.038	-0.003	-0.059	-0.066	0.061	0.024	0.083	0.61	0.061	0.14	0.069	0.089	0.082	1	0.13
TenYearCHD	0.22	-0.052	0.085	0.034	0.068	0.087	0.069	0.17	0.1	0.094	0.21	0.14	0.066	0.02	0.13	1



1) From the above graph we can oberve that as age increases so does CHD

2) When increase in total cholostrol also results in increase in CHD

3) Increase in systolic blood pressure also increases in CHD

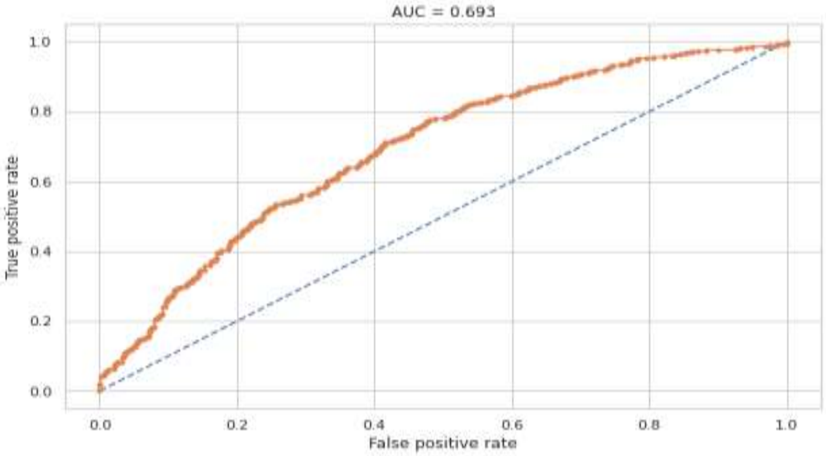
To summarize:

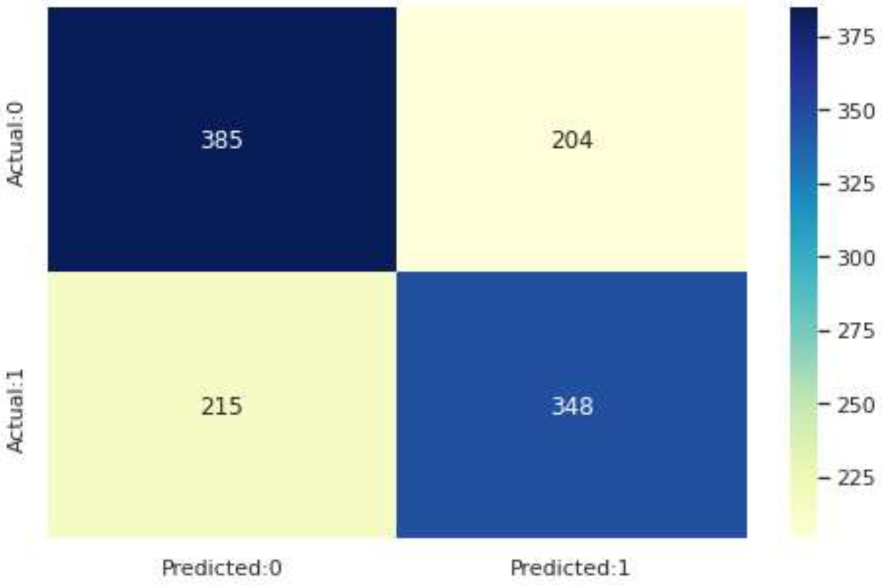
Age, total cholostrol and systolic blood are the top three factors for CHD

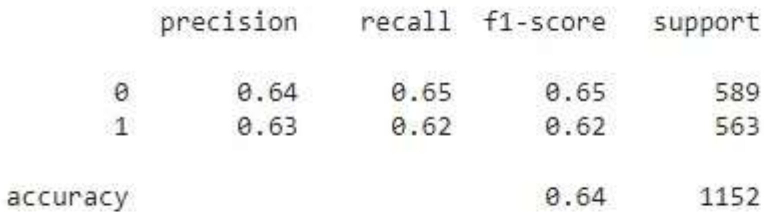
Logistic Regression

Evaluation metrics:

- Train Recall = 65%
- Test Recall = 62%
- Test Accuracy = 64%





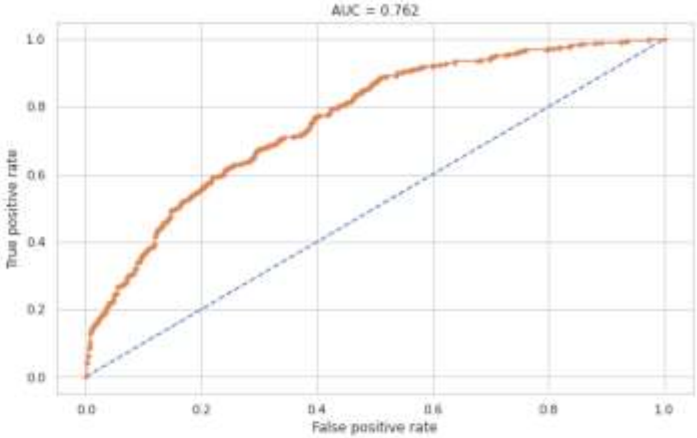


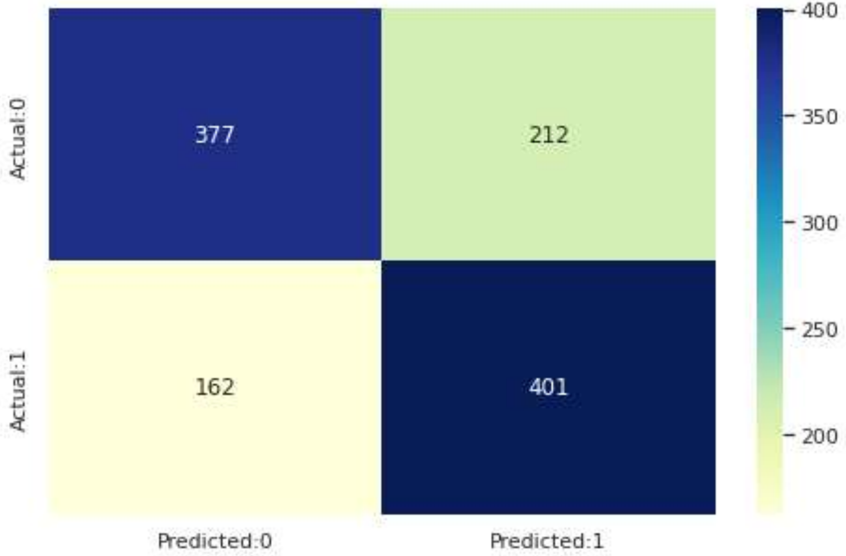
```
              precision    recall  f1-score   support

           0       0.64      0.65      0.65       589
           1       0.63      0.62      0.62       563

    accuracy                           0.64      1152
```

Random Forests

Parameters:

- max_depth = 8
- min_samples_leaf = 40
- min_samples_split = 50
- n_estimators = 100



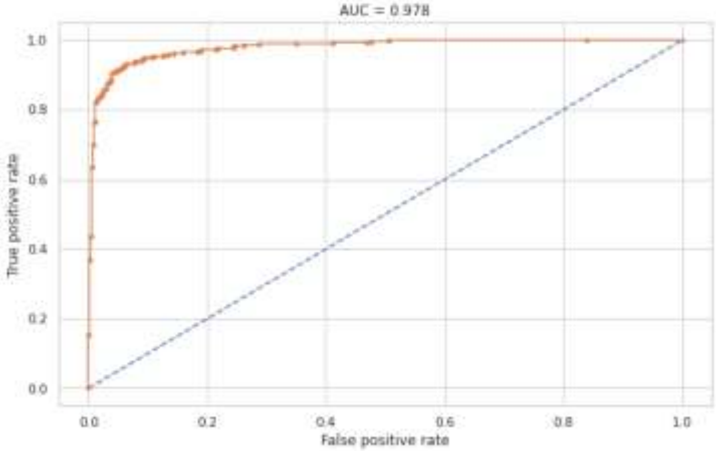


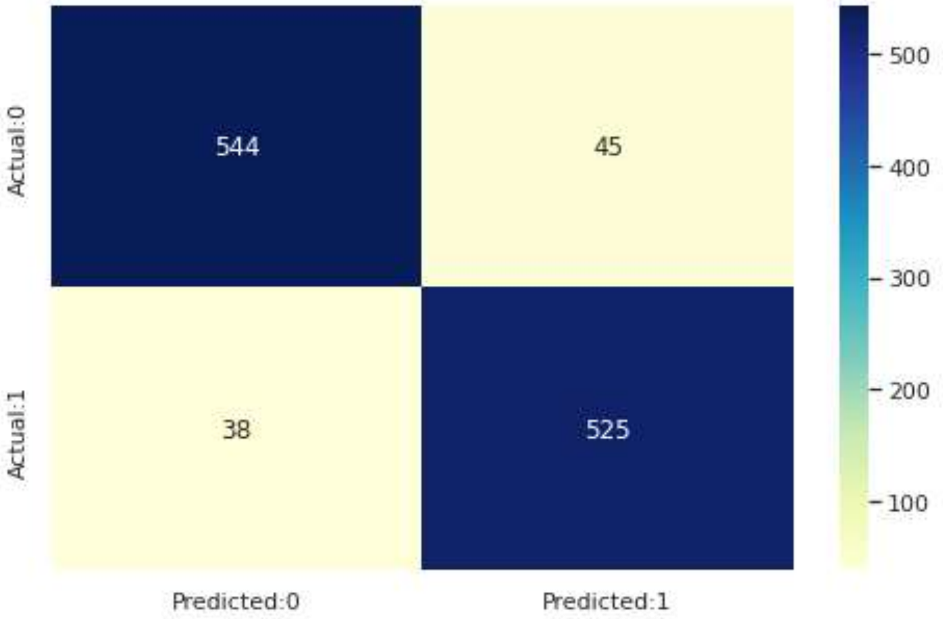
```
              precision    recall  f1-score   support

           0       0.70      0.64      0.67       589
           1       0.65      0.71      0.68       563

    accuracy                           0.68      1152
```

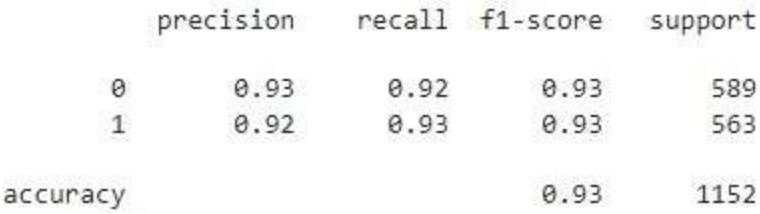
Support Vector Machines

Parameters:

- C = 1
- Gamma = 0.01
- Kernel = rbf

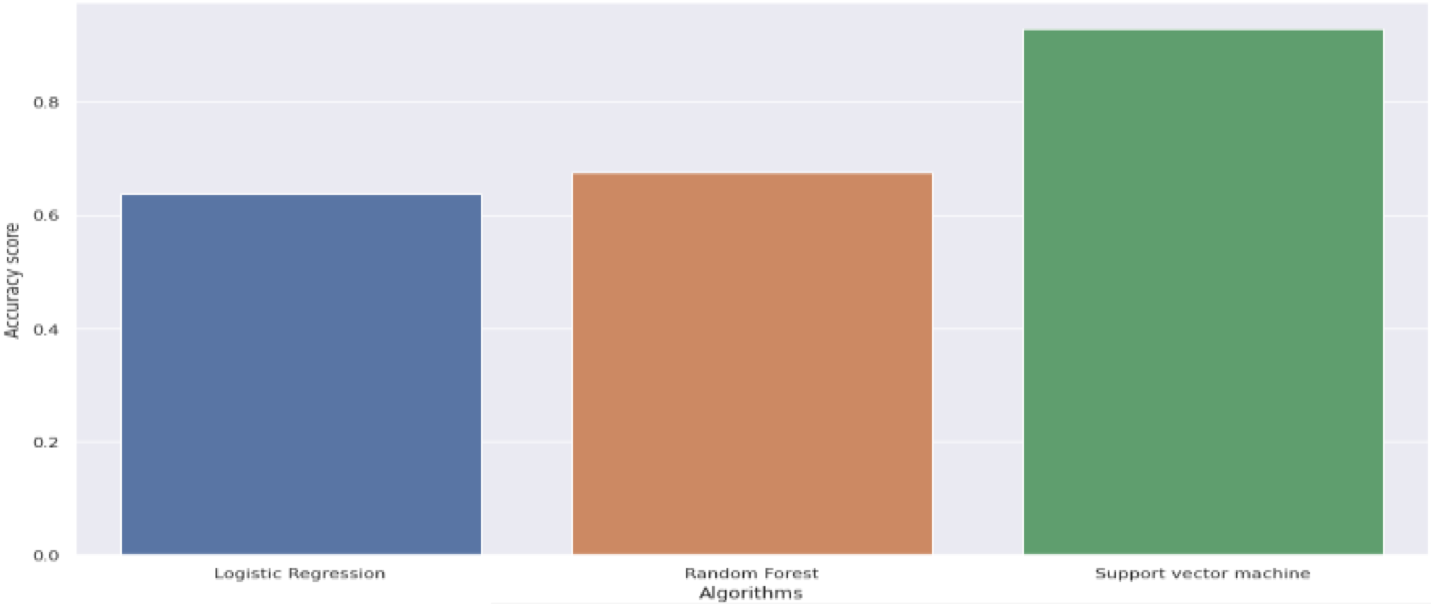


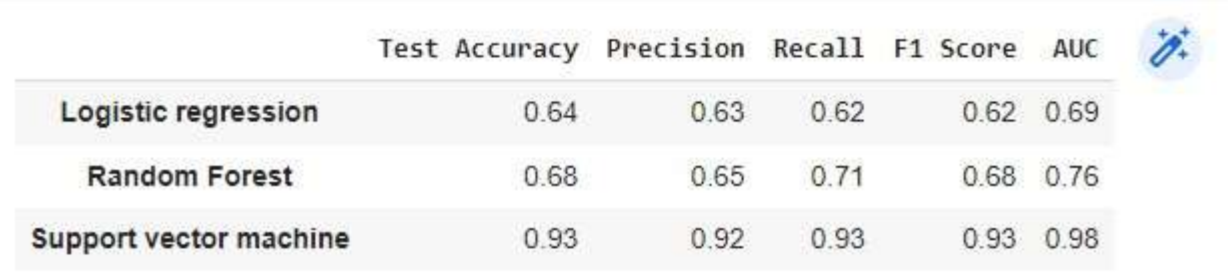




```
              precision    recall  f1-score   support

           0       0.93      0.92      0.93       589
           1       0.92      0.93      0.93       563

    accuracy                           0.93      1152
```

Models Comparision





	Test Accuracy	Precision	Recall	F1 Score	AUC
Logistic regression	0.64	0.63	0.62	0.62	0.69
Random Forest	0.68	0.65	0.71	0.68	0.76
Support vector machine	0.93	0.92	0.93	0.93	0.98

Conclusions

1) The top features in predicting the ten year risk of developing Cardiovasular Heart Disease are 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose’.

2) The people who have Cardiovascular heart disease is almost equal between smokers and non smokers.

3) 85% of people are not at risk of Cardio Vascular Risk. 3% of people who are at the risk of CHD are taking BP medication.

4) The Support vector machine with the radial kernel is the best performing model in terms of accuracy and the F1 score and Its high AUC-score shows that it has a high true positive rate.

Thank You