Capstone Project

‘Cardiovascular Risk prediction’

By-Deepak Kumar Dubey

INTRODUCTION

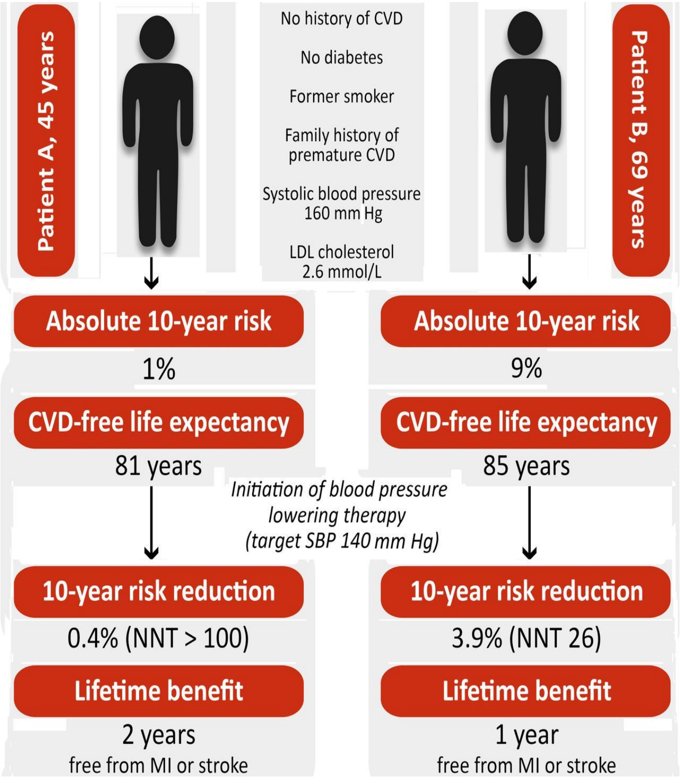
- Currently cardiovascular diseases (CVDs) are the main cause of death worldwide. Disease risk estimates can be used as prognostic information and support for treating CVDs.

- The commonly used Framingham risk score (FRS) for CVD prediction is outdated for the modern population, so FRS may not be accurate enough.

- In this project ,a CVD prediction system based on machine learning is proposed.

- 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 3,000 records and 15 attributes.

- Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors.



Patient A, 45 years

No history of CVD

No diabetes

Former smoker

Family history of premature CVD

Systolic blood pressure 160 mm Hg

LDL cholesterol 2.6 mmol/L

Patient B, 69 years

Patient A, 45 years	Patient B, 69 years
Absolute 10-year risk	Absolute 10-year risk
1%	9%
CVD-free life expectancy	CVD-free life expectancy
81 years	85 years
10-year risk reduction	10-year risk reduction
0.4% (NNT > 100)	3.9% (NNT 26)
Lifetime benefit	Lifetime benefit
2 years free from MI or stroke	1 year free from MI or stroke

Initiation of blood pressure lowering therapy (target SBP 140 mm Hg)

Data Introduction

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)

Behavioural

- 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)

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)

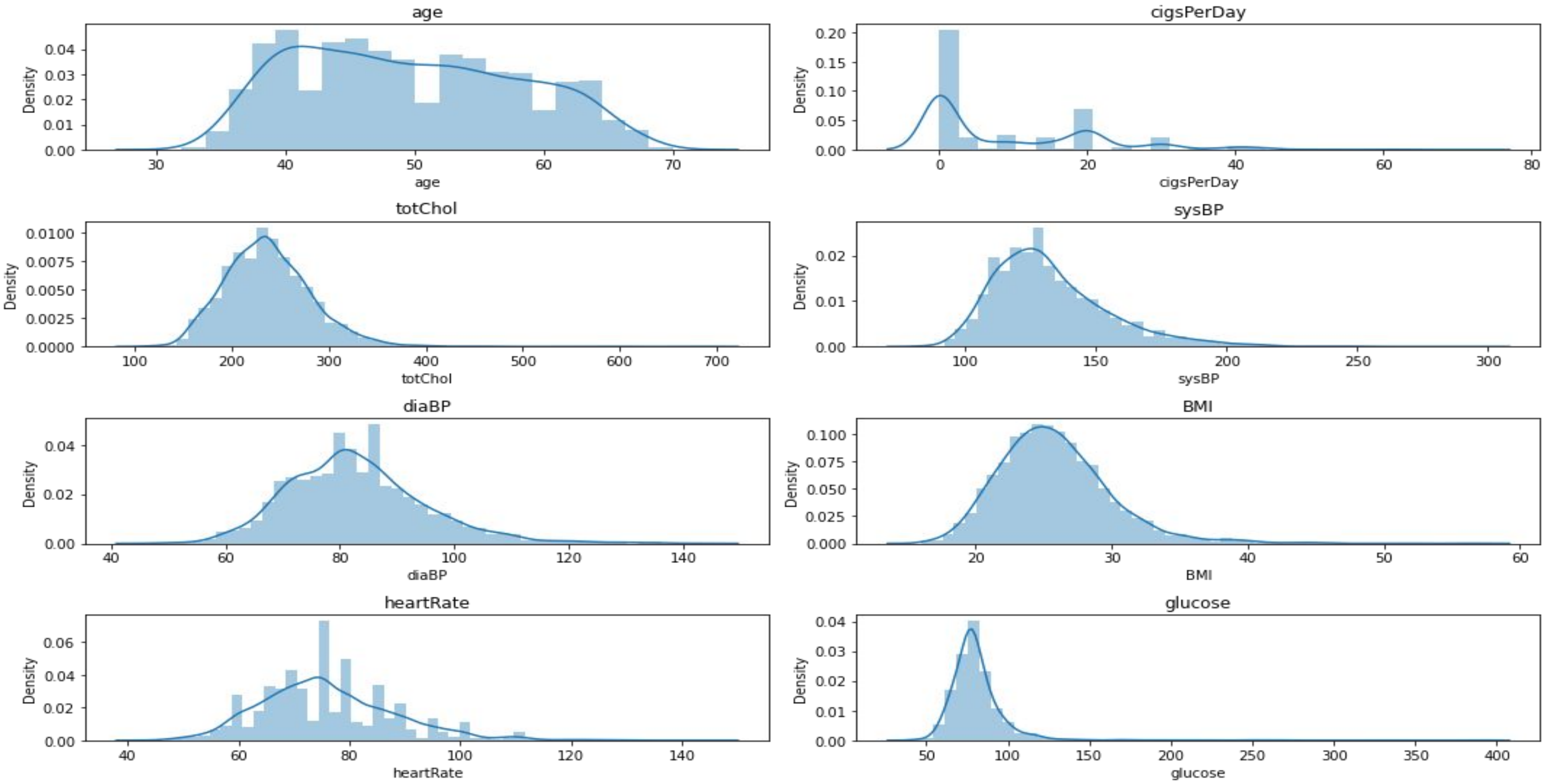
Dependent variable

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

CONTINUOUS	NOMINAL
AGE	IS_SMOKING
EDUCATION	SEX
CIGS_PER_DAY	DIABETES
TOT_CHOL	BP_MEDS
SYS_BP	PREVALENT_STROKE
DIA_BP	PREVALENT_HYP
BMI	DIABETES
HEART_RATE	
GLUCOSE	

Observations:3390 Features:17

Exploratory Data Analysis



Domain knowledge and feature distribution in our data

Age-we have data for people having age between 32 to 70 with mean age being around 49 years.

Cigarettes per day- most people don't smoke any cigarette in a day but those who smoke smoke from 1 to 70 cigarettes with mean smoked cigarettes being around 9.

Totchol- Total cholesterol value has mean value of 237 with minimu

Sys BP-Normal: Below 120

> Elevated: 120-129

> Stage 1 high blood pressure (also called hypertension): 130-139

> Stage 2 hypertension: 140 or more

> Hypertensive crisis: 180 or more

> In our data ,we have systolic bp distributed between 83.5 to 295 with mean being 132.

dia BP- This is what your diastolic blood pressure number means:

> Normal: Lower than 80

> Stage 1 hypertension: 80-89

> Stage 2 hypertension: 90 or more

> Hypertensive crisis: 120 or more

> In our data,We have diaBP values distributed between 48 to 142.5 with mean value being 83

.

BMI- Category BMI range - kg/m2

> Severe Thinness < 16

> Moderate Thinness 16 - 17

> Mild Thinness 17 - 18.5

> Normal 18.5 - 25

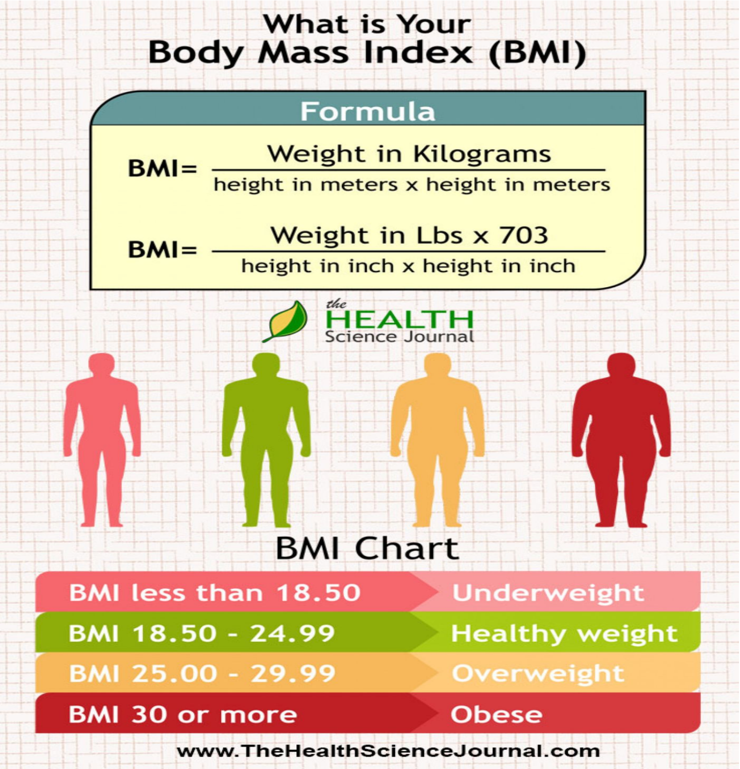
> Overweight 25 - 30

> Obese Class I 30 - 35

> Obese Class II 35 - 40

> Obese Class III > 40

We have BMI index values ranging between 15.96 to 57 with mean value being 25.79.



What is Your
Body Mass Index (BMI)

Formula

$$BMI = \frac{\text{Weight in Kilograms}}{\text{height in meters x height in meters}}$$

$$BMI = \frac{\text{Weight in Lbs x 703}}{\text{height in inch x height in inch}}$$

BMI Chart

BMI less than 18.50	Underweight
BMI 18.50 - 24.99	Healthy weight
BMI 25.00 - 29.99	Overweight
BMI 30 or more	Obese

www.TheHealthScienceJournal.com

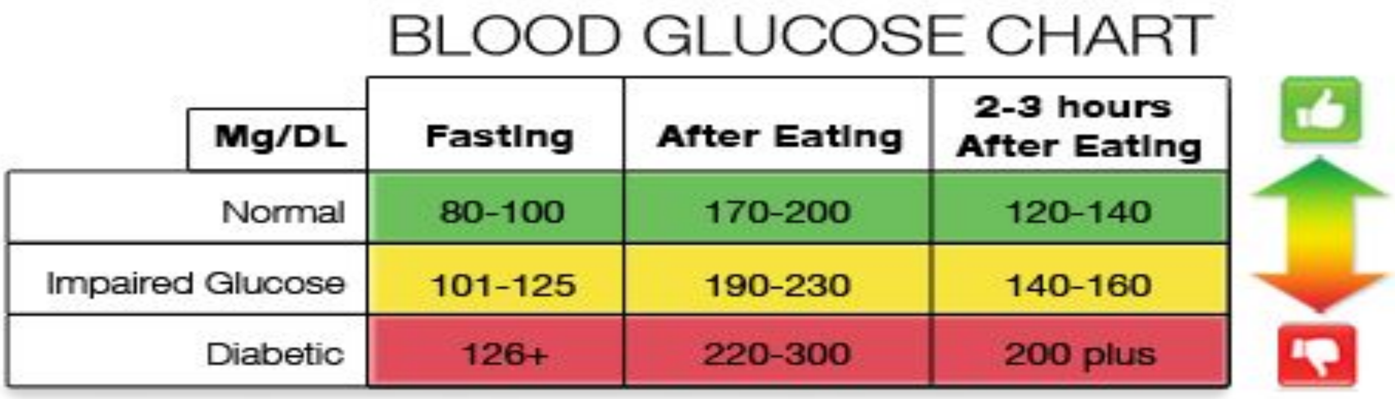
heart rate- A normal resting heart rate for adults ranges from 60 to 100 beats per minute.

> In our data we have heart rate values ranging from 45 to 143 with mean value being 76.

Glucose-

In our data, we have glucose ranging from 40 to 394 with mean value being 82.

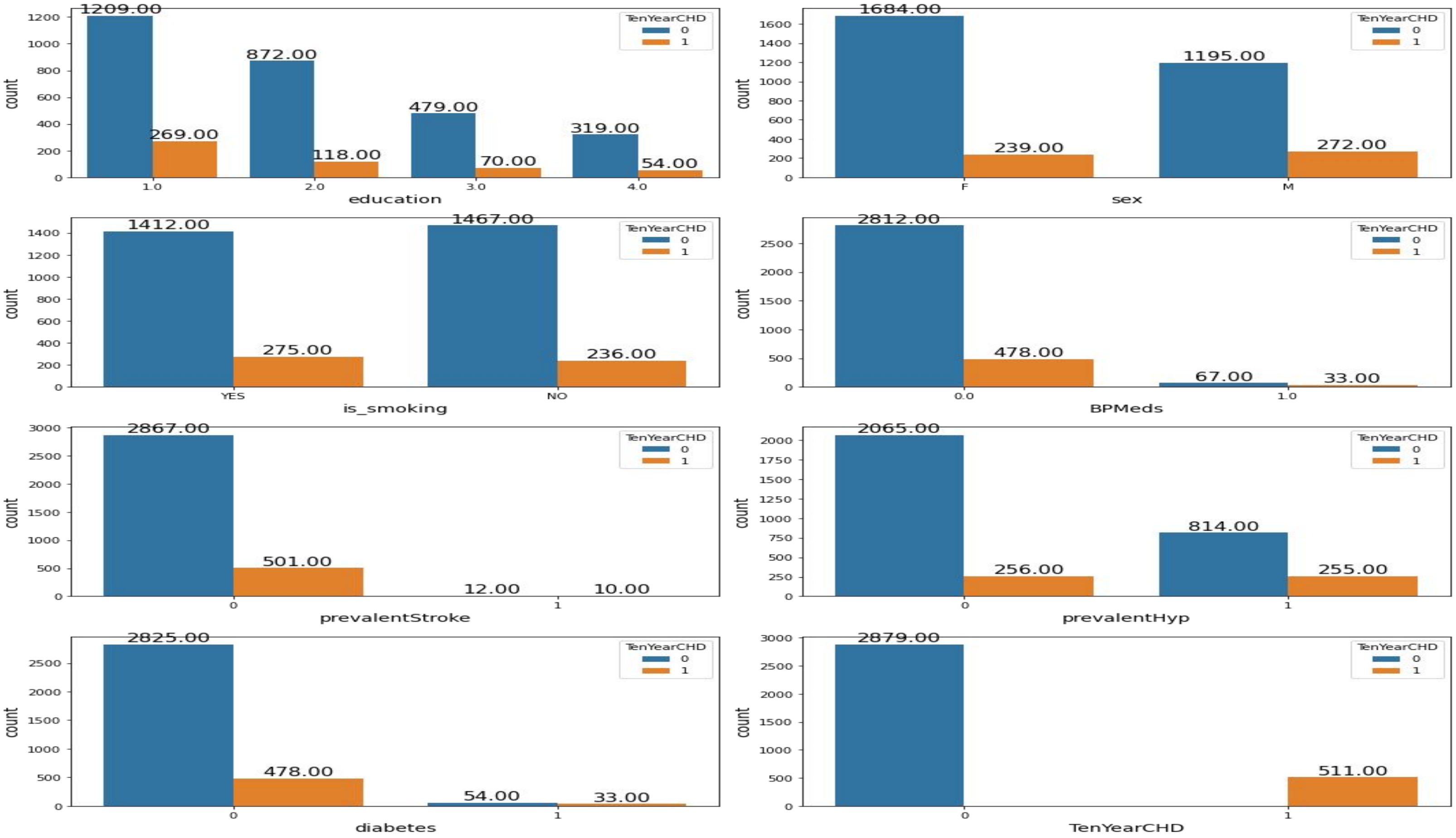
.



BLOOD GLUCOSE CHART

Mg/DL	Fasting	After Eating	2-3 hours After Eating
Normal	80-100	170-200	120-140
Impaired Glucose	101-125	190-230	140-160
Diabetic	126+	220-300	200 plus

Count plots for ordinal /nominal features



Conclusions based count plots

We can conclude following from the above countplots-

1)Males are more prone to CHD as compared to females.

2)Smokers are more prone CHD as compared to non smokers though it makes only a little difference.

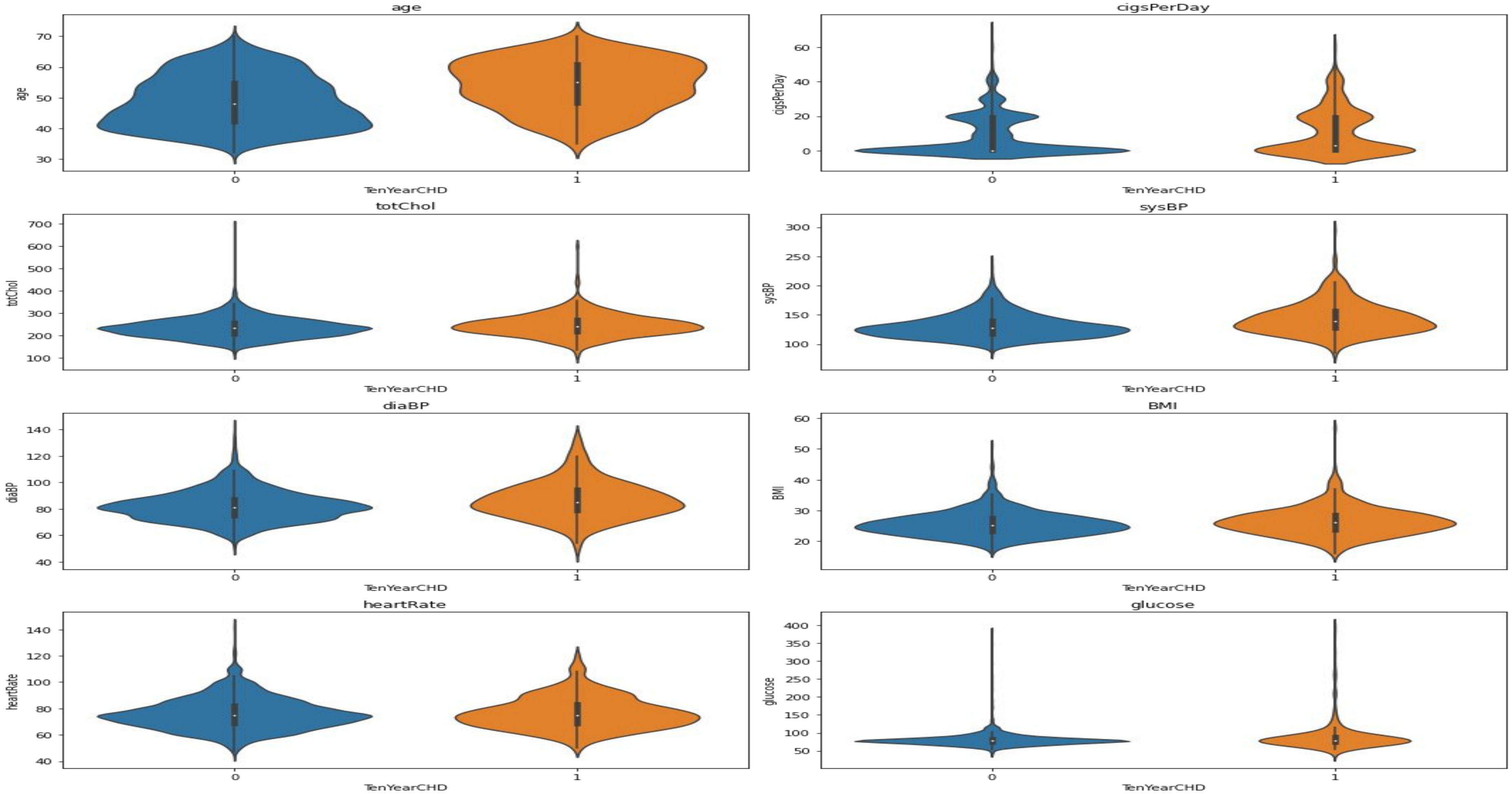
3)People on BP medications are more prone to CHD.

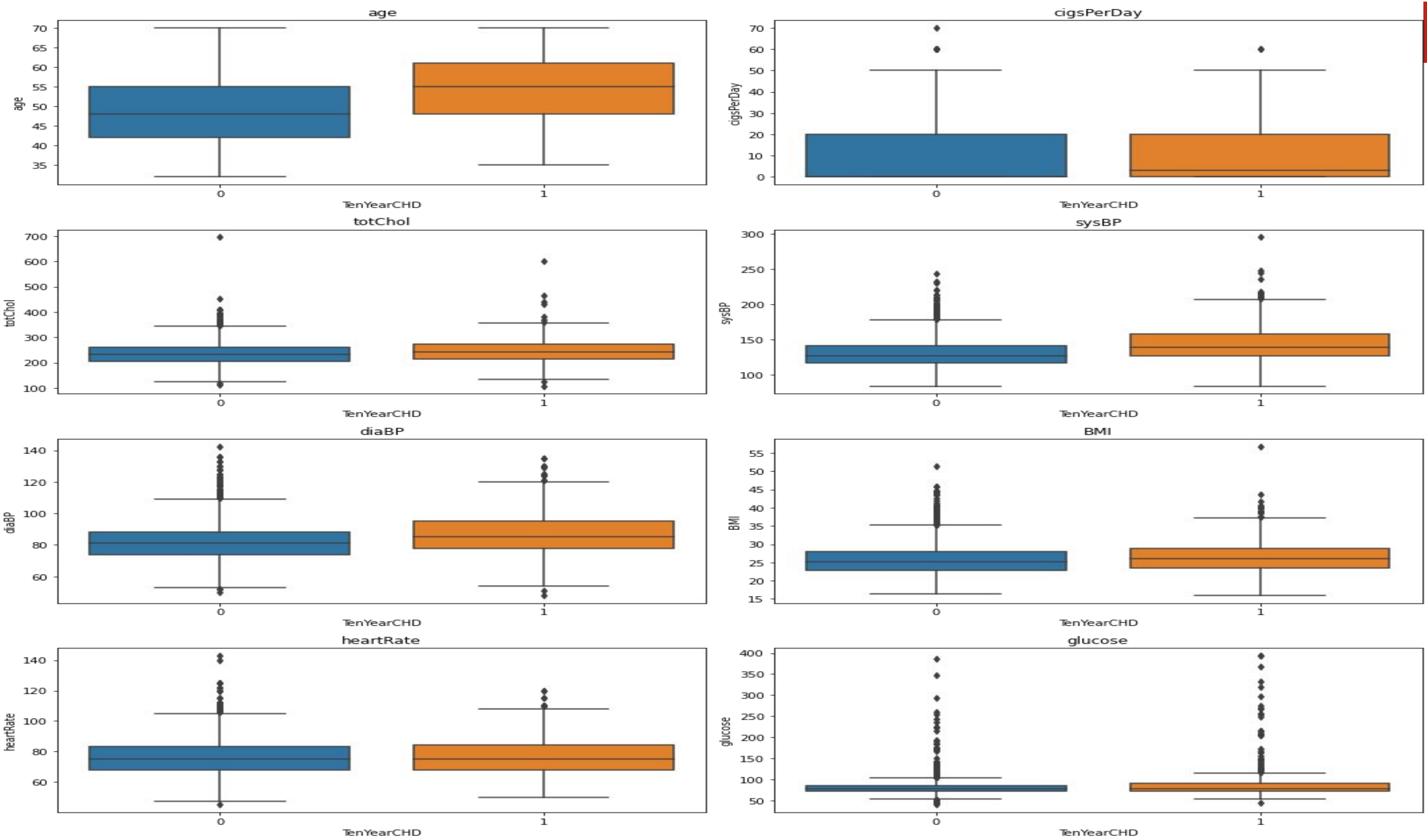
4)People who had any previous prevalent stroke are more prone to having CHD although people who had prevalent stroke they are very less.

5)Those who had prevalent hypertension are more prone to having CHD.

6)Those who have diabetes are also more prone to having CHD although number of people who have diabetes are very less.

Boxplot and violin plot for continuous/Discrete variables




age
age
70
65
60
55
50
45
40
35
0
1
TenYearCHD
cigsPerDay
cigsPerDay
70
60
50
40
30
20
10
0
0
1
TenYearCHD
totChol
totChol
700
600
500
400
300
200
100
0
1
TenYearCHD
sysBP
sysBP
300
250
200
150
100
0
1
TenYearCHD
diaBP
diaBP
140
120
100
80
60
0
1
TenYearCHD
BMI
BMI
55
50
45
40
35
30
25
20
15
0
1
TenYearCHD
heartRate
heartRate
140
120
100
80
60
0
1
TenYearCHD
glucose
glucose
400
350
300
250
200
150
100
50
0
1
TenYearCHD

Conclusions from above plots

We can conclude following from above graphs-

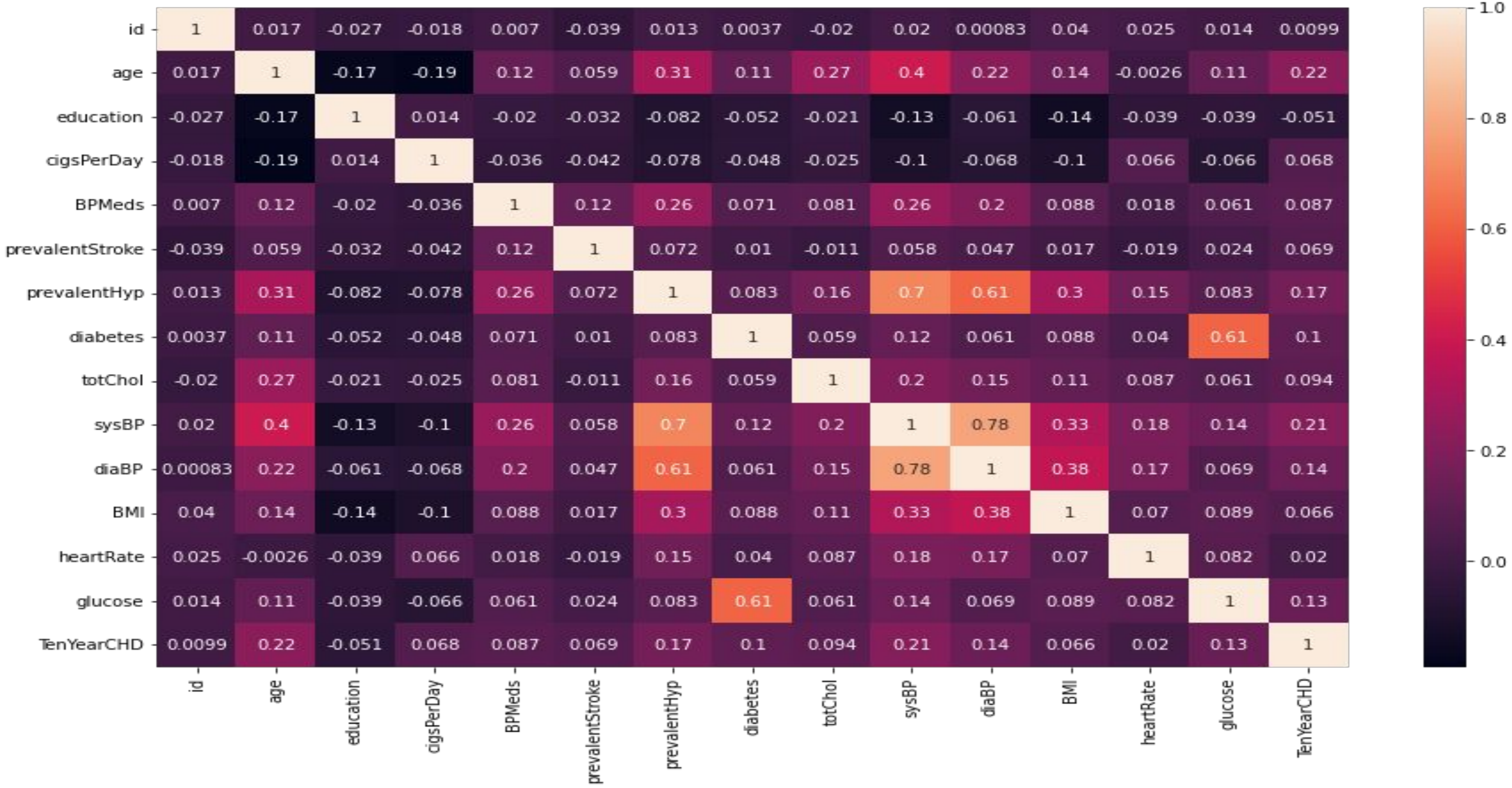
1)Median age of people who had CHD is higher as compared to people who don't have CHD.This implies that people having higher age are more prone to having CHD.

2)People who had CHD smoke more cigarettes as compared to people who had no CHD.

3)people who have CHD have more tot chol,sysBP,diaBP,BMI.

4)Heart rate and glucose levels have almost same median values for those who

Multicollinearity



	id	age	education	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
id	1	0.017	-0.027	-0.018	0.007	-0.039	0.013	0.0037	-0.02	0.02	0.00083	0.04	0.025	0.014	0.0099
age	0.017	1	-0.17	-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	-0.027	-0.17	1	0.014	-0.02	-0.032	-0.082	-0.052	-0.021	-0.13	-0.061	-0.14	-0.039	-0.039	-0.051
cigsPerDay	-0.018	-0.19	0.014	1	-0.036	-0.042	-0.078	-0.048	-0.025	-0.1	-0.068	-0.1	0.066	-0.066	0.068
BPMeds	0.007	0.12	-0.02	-0.036	1	0.12	0.26	0.071	0.081	0.26	0.2	0.088	0.018	0.061	0.087
prevalentStroke	-0.039	0.059	-0.032	-0.042	0.12	1	0.072	0.01	-0.011	0.058	0.047	0.017	-0.019	0.024	0.069
prevalentHyp	0.013	0.31	-0.082	-0.078	0.26	0.072	1	0.083	0.16	0.7	0.61	0.3	0.15	0.083	0.17
diabetes	0.0037	0.11	-0.052	-0.048	0.071	0.01	0.083	1	0.059	0.12	0.061	0.088	0.04	0.61	0.1
totChol	-0.02	0.27	-0.021	-0.025	0.081	-0.011	0.16	0.059	1	0.2	0.15	0.11	0.087	0.061	0.094
sysBP	0.02	0.4	-0.13	-0.1	0.26	0.058	0.7	0.12	0.2	1	0.78	0.33	0.18	0.14	0.21
diaBP	0.00083	0.22	-0.061	-0.068	0.2	0.047	0.61	0.061	0.15	0.78	1	0.38	0.17	0.069	0.14
BMI	0.04	0.14	-0.14	-0.1	0.088	0.017	0.3	0.088	0.11	0.33	0.38	1	0.07	0.089	0.066
heartRate	0.025	-0.0026	-0.039	0.066	0.018	-0.019	0.15	0.04	0.087	0.18	0.17	0.07	1	0.082	0.02
glucose	0.014	0.11	-0.039	-0.066	0.061	0.024	0.083	0.61	0.061	0.14	0.069	0.089	0.082	1	0.13
TenYearCHD	0.0099	0.22	-0.051	0.068	0.087	0.069	0.17	0.1	0.094	0.21	0.14	0.066	0.02	0.13	1

Removing Multicollinearity

We can see from the heatmap that following features are related to each other-

1)glucose and diabetes(corr coef-0.61)

2)sys and diaBP(corr coef-0.78)

3)prevalent hyp and sysBP(corr coef-0.7)

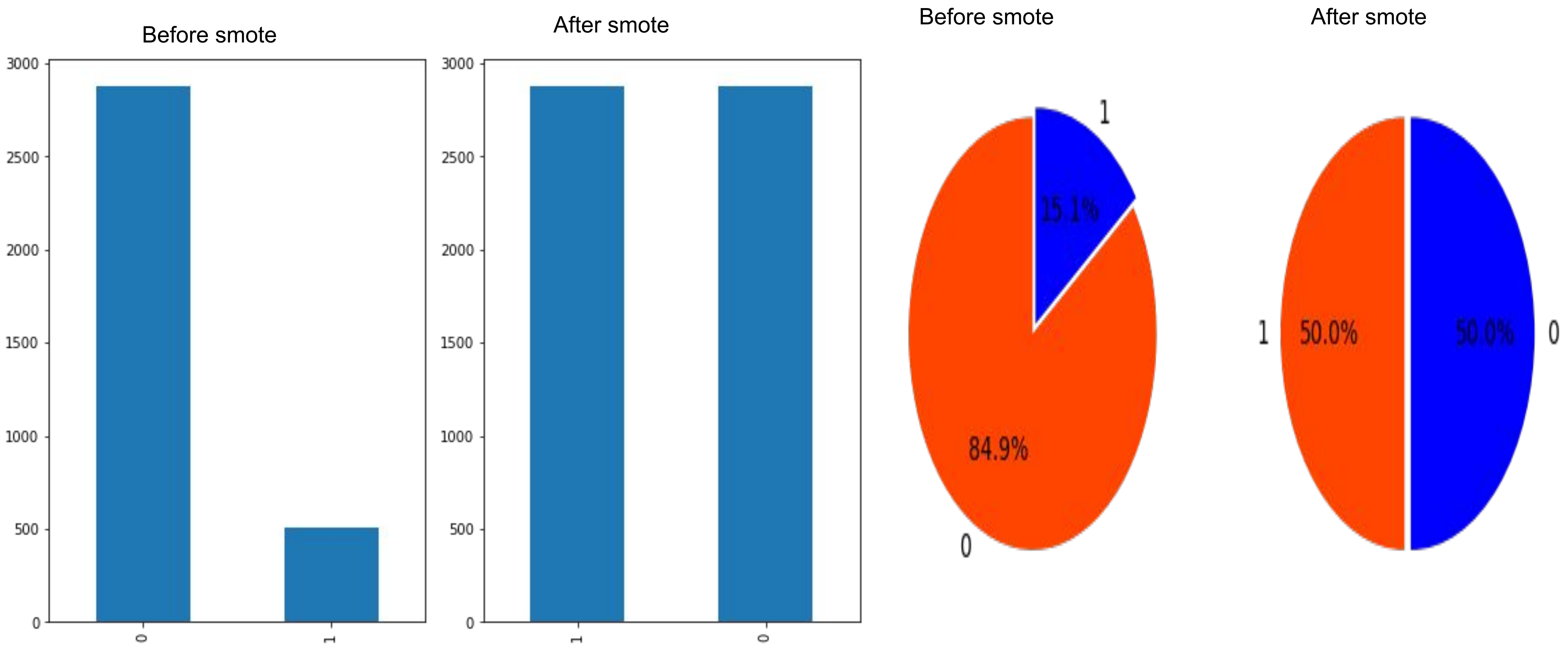
Linearly combine the independent variables, such as adding them together.

pulse rate=sys BP-dia BP

Remove some of the highly correlated independent variables.

Dropped dia BP,sys BP,diabetes and prevalent hypertension.

Data balancing of target variable using SMOTE


Before smote

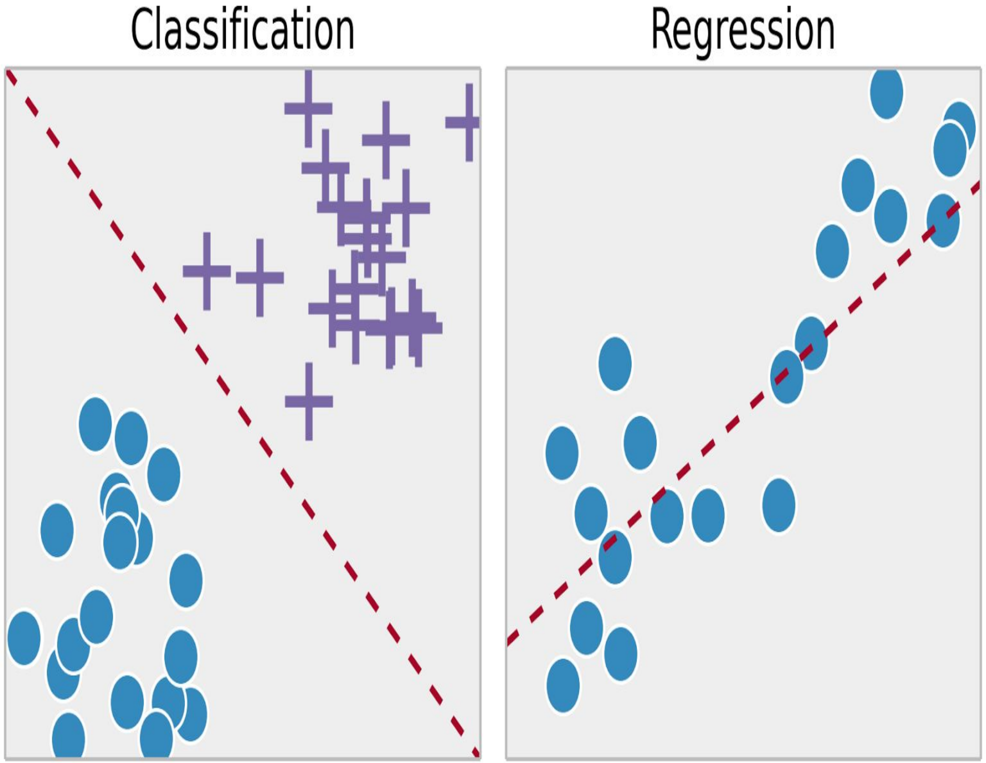
After smote

Before smote

After smote

Classification models used

- Logistic Regression
- K nearest neighbours Classifiers
- Support vector Classifier
- Decision Tree
- Random Forest Classifier
- Bagging Classifier
- Adaboost Classifier
- XG boost Classifier
- Catboost Classifier
- Stacking Classifier



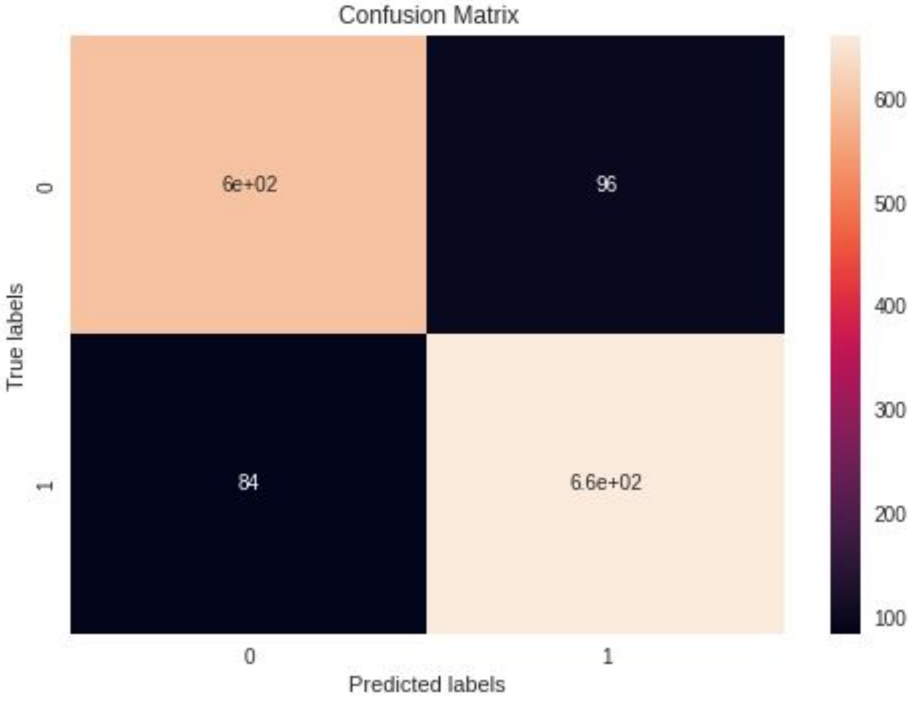
Best performing models and their parameters

Models	Parameters
Adaboost	GridSearchCV(estimator=AdaBoostClassifier(base_estimator=DecisionTreeClassifier()), n_jobs=-1,param_grid={'base_estimator__max_depth': [2, 4, 6, 8, 10], 'base_estimator__min_samples_leaf': [5, 10], 'learning_rate': [0.01, 0.1],'n_estimators': [10, 50, 250, 1000]},scoring='f1', verbose=3)
Stacking model	Base model-KNN,bagging,SVM Meta model-Logistic regression
Catboost classifier	{'depth': 4, 'learning_rate': 0.1, 'n_estimators': 1000}
Bagging classifier	GridSearchCV(cv=5,error_score=0,estimator=BaggingClassifier(random_state=0), n_jobs=-1,param_grid={'n_estimators': [50, 100, 500, 1000]}, scoring='roc_auc')
Random forest classifier	{'criterion': 'entropy', 'max_depth': 50, 'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 300}

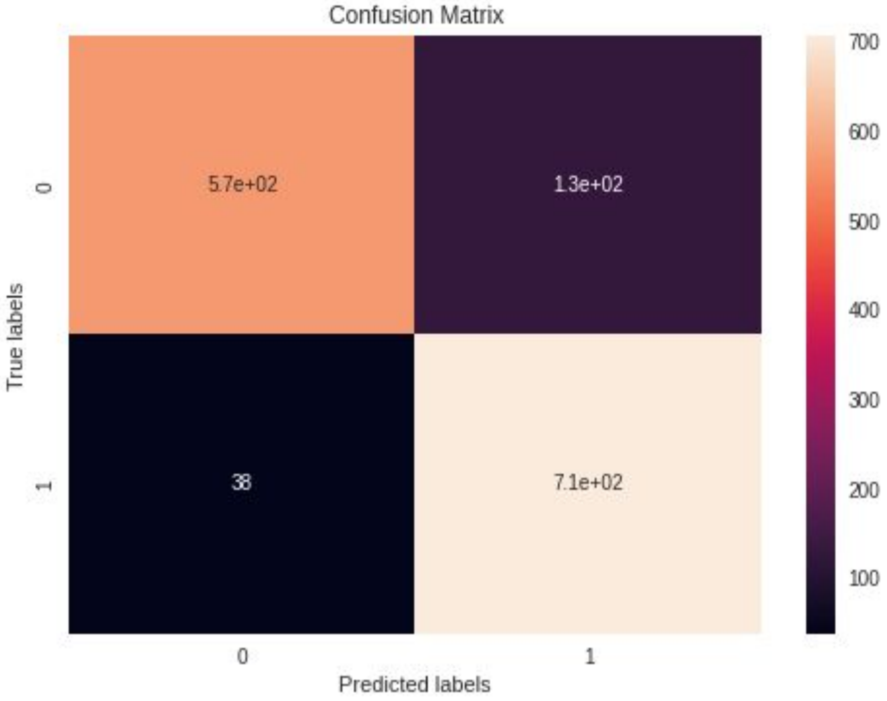
Evaluation Metrics for different models

index	model	Accuracy	Recall	Precision	ROC-AUC	F1 Score
1	Logistic	0.61	0.95	0.57	0.6	0.71
2	KNN	0.82	0.92	0.77	0.81	0.84
3	SVC	0.79	0.82	0.79	0.79	0.81
4	Decision tree	0.75	0.83	0.71	0.74	0.77
5	Random forest	0.83	0.94	0.78	0.83	0.85
6	Bagging	0.83	0.94	0.78	0.83	0.85
7	Adaboost	0.88	0.95	0.85	0.88	0.89
8	XG boost	0.87	0.89	0.87	0.87	0.88
9	Catboost	0.85	0.88	0.83	0.85	0.86
10	Stacking	0.87	0.9	0.86	0.87	0.88

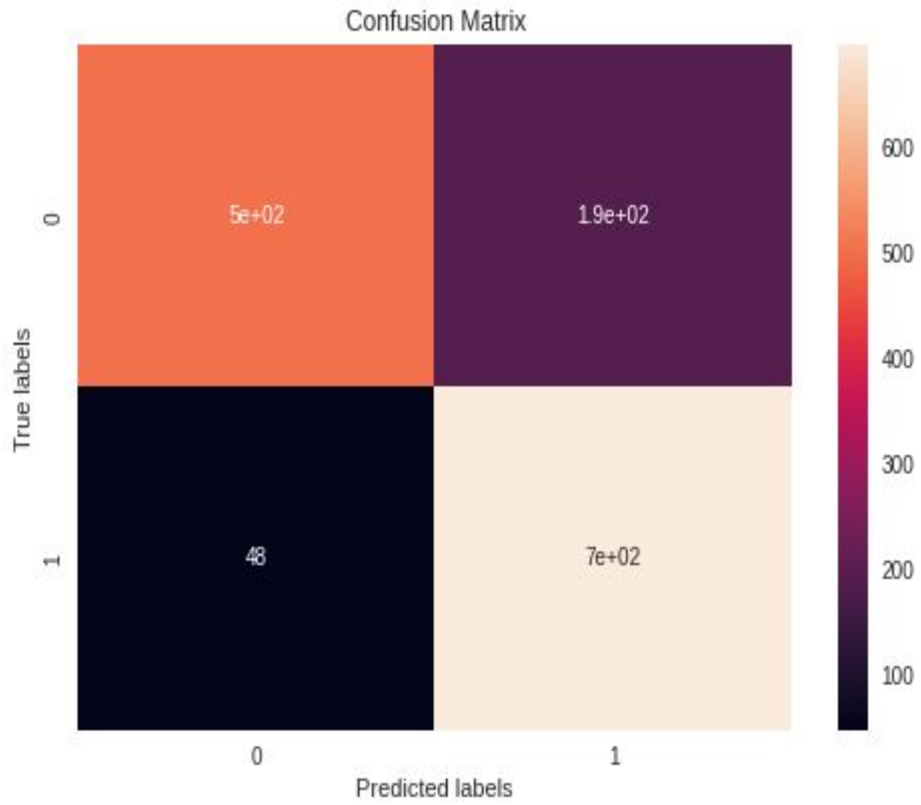
Confusion matrix for best performing model



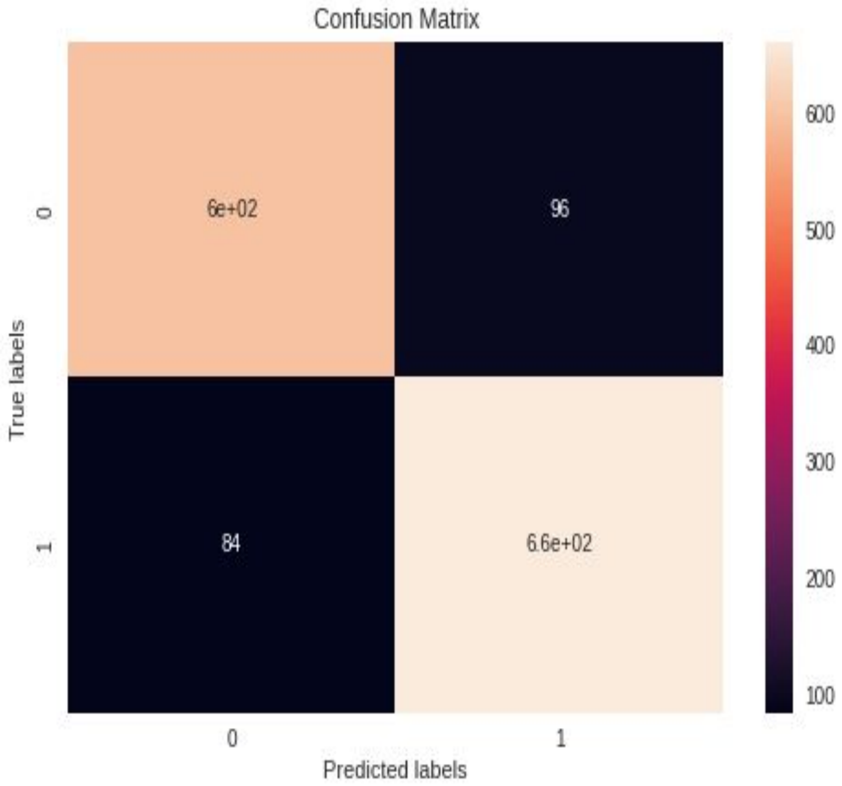
Stacking model



Adaboost model

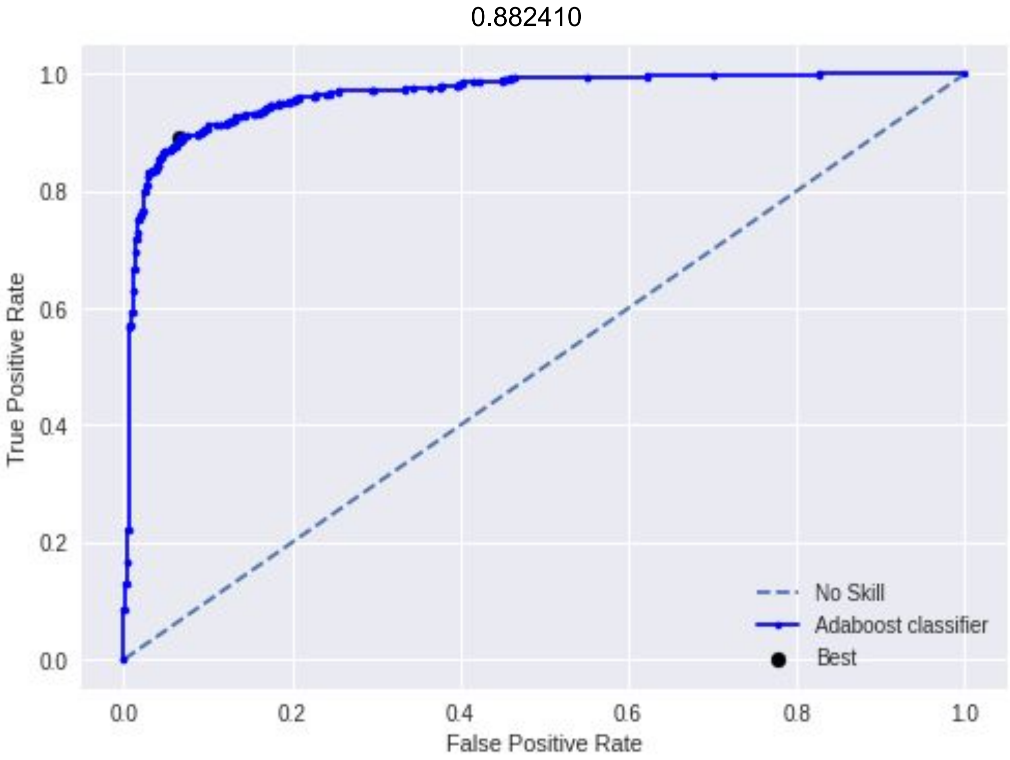


Bagging classifier



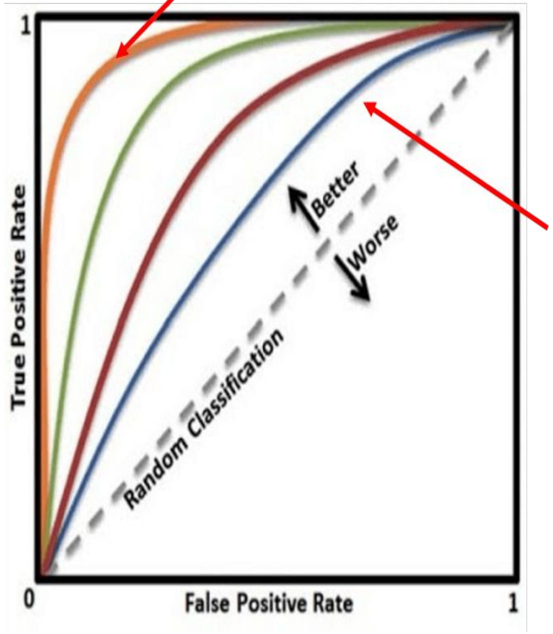
Cat boost classifier

ROC curve for best performing model

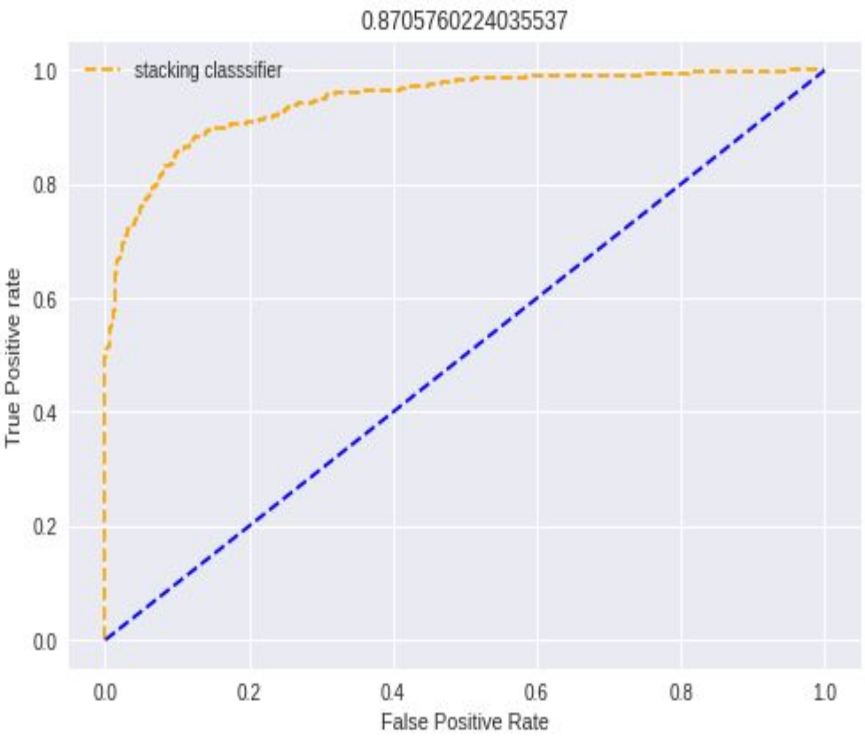


Adaboost model

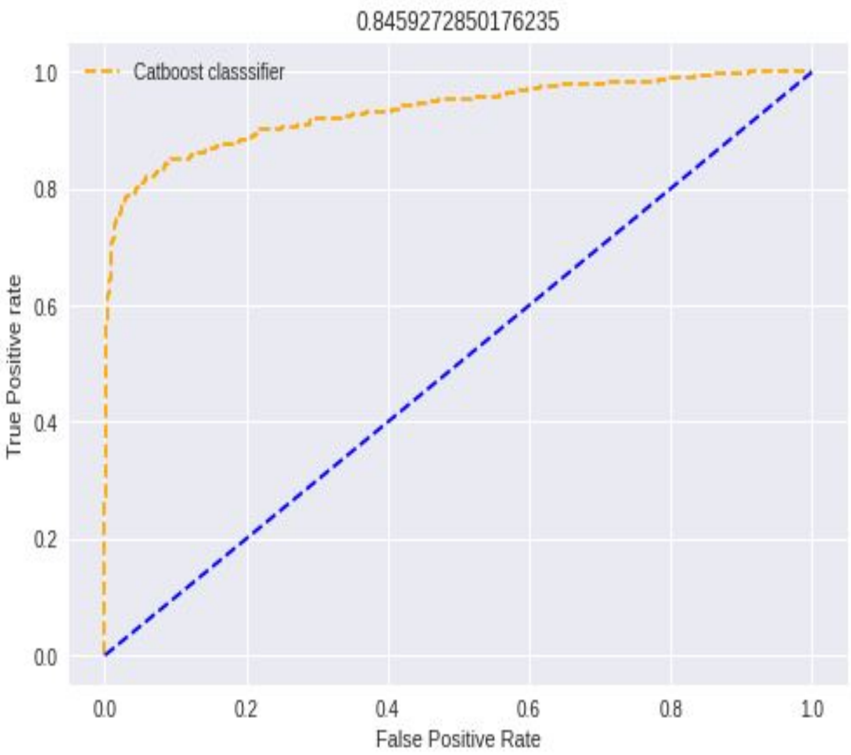
Good Model



Poor Model



Stacking model



Catboost model

Conclusions

1)Ensemble models are performing better as compared to other models with adaboost being the best performing model followed by stacking and bagging.

2)We can see from the evaluation metric table that best performing model is adaboost classifier because it has best accuracy and has highest recall as well.

3)In this type of problem our priority should be to reduce the number of False Negatives or find maximum Recall score. If we misclassify someone as having no risk to heart disease, it can be highly detrimental, it can lead to loss of life. Adaboost gives us an excellent Recall and at the same time doesn't compromise on Precision. If we require a model with more strict Recall values we can opt for KNN.

THANK YOU