Capstone Project

Cardiovascular Risk Prediction

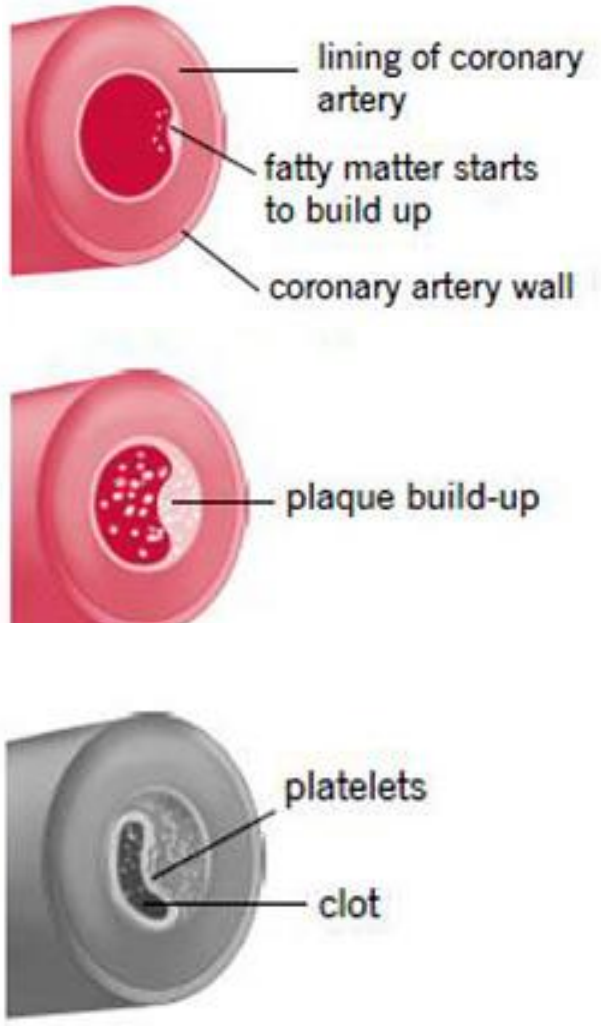
By:- Om Prakash Pradhan & Ruchika Nayak

Key Points

- ❑Introduction
- ❑Project Objectives
- ❑Data Summary
- ❑Methodology
- ❑Insights from EDA
- ❑Feature engineering
- ❑Data preparation for modelling
- ❑Model fitting and evaluation
- ❑Conclusion
- ❑Challenges

Introduction

- Coronary heart disease(CHD) is a narrowing or blockage of coronary arteries usually caused by the buildup of fatty material called plaque. Coronary heart disease is also called coronary artery disease, ischemic heart disease and heart disease.
- In some cases, when plaque breaks, a blood clot may block the supply to your heart muscle. This causes a heart attack.
- The damage may be caused by various factors including smoking, high blood pressure, high cholesterol, diabetes or insulin resistance, not being active (sedentary lifestyle) etc.



Project Objectives

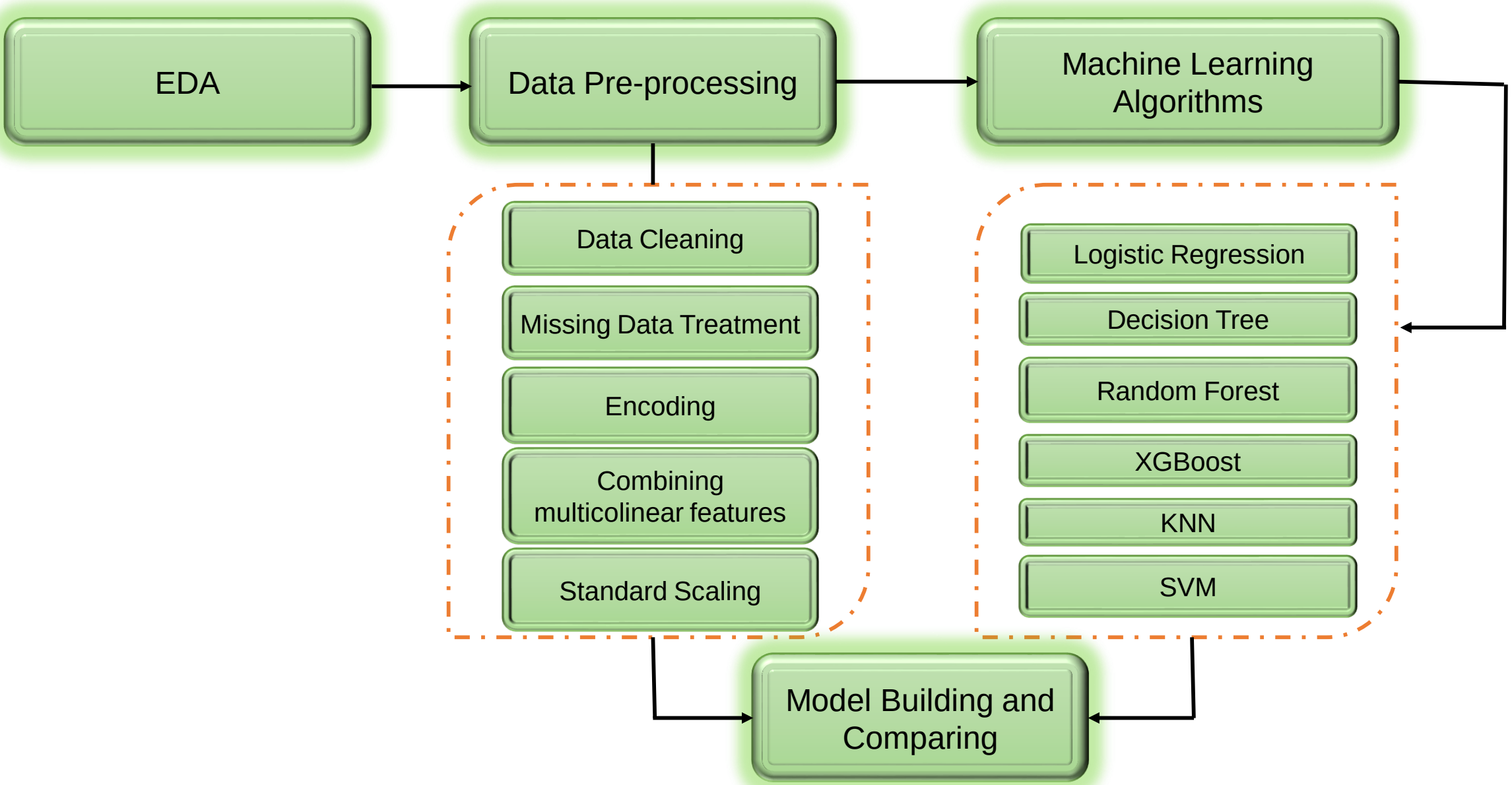
- The main goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).
- To highlight the main variables/factors influencing 10-year risk of future coronary heart disease (CHD).
- To compare the various classification models and find out the best model for the above task.

Data Summary

Id	age	education	sex	Is_smoking	cigsPerDay	BPMeds	Prevalent Stroke	Prevalent Hyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

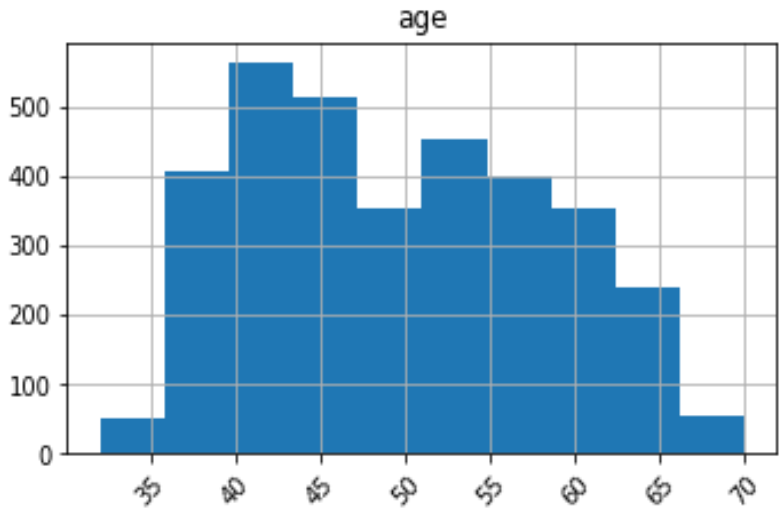
- The dataset provides the patients' information. It includes 15 attributes. Each attribute is a potential risk factor. There are demographic, behavioral, and medical risk factors.
- Dataset comprises of total 3390 rows and 17 columns and there are missing values in the education, cigsPerDay, BPMeds, totChol, BMI, heartRate and glucose columns. There are no duplicate values in the dataset.
- Out of all the features sex and is_smoking are categorical in nature.

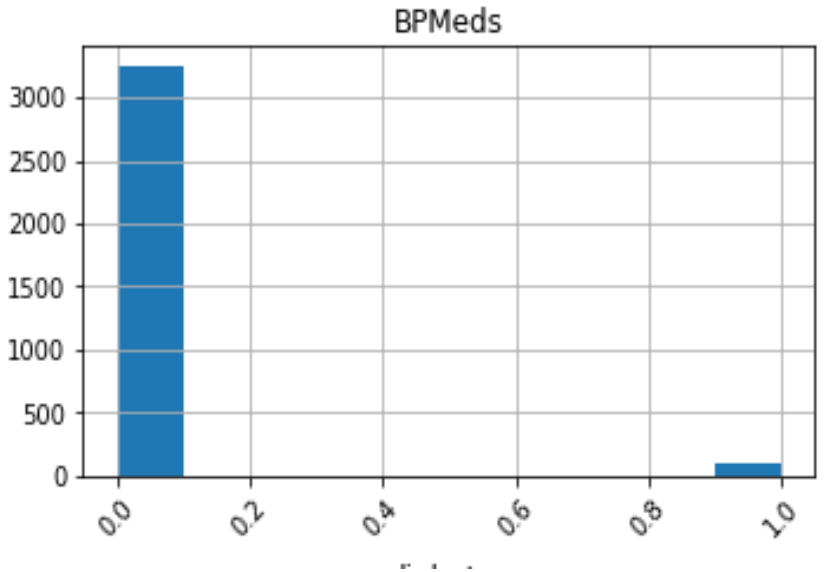
Methodology



Insights from EDA

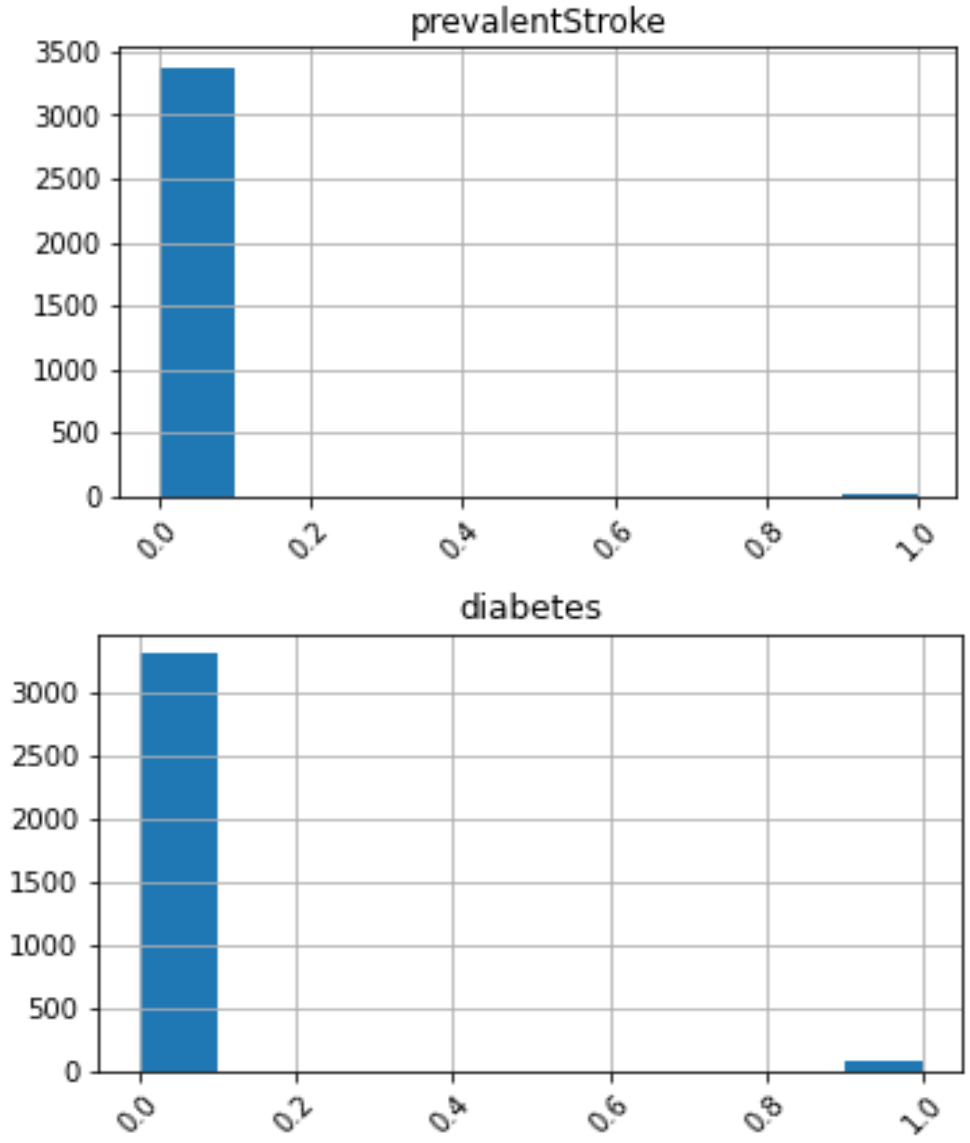
- Dataset contains mostly the data of middle aged patients.
- Most of the patients are not on blood pressure medication.





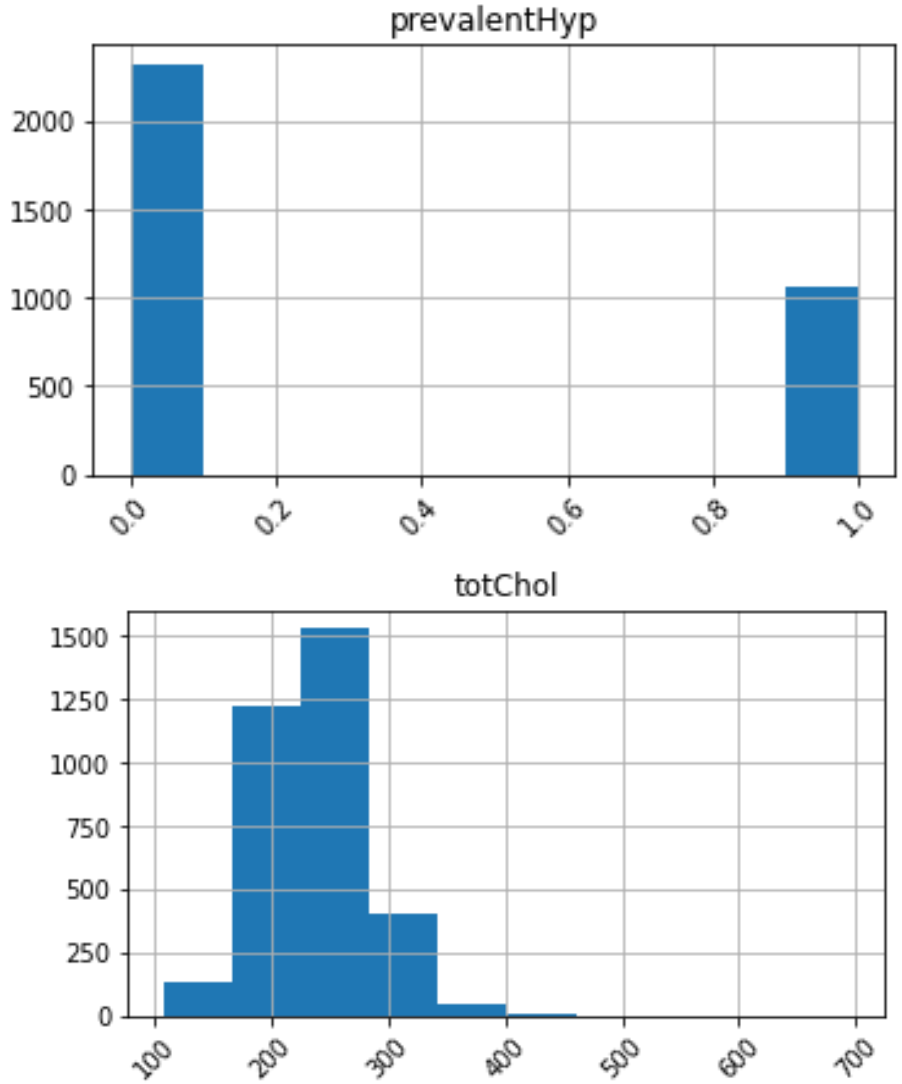
Insights from EDA

- Most of the patients don’t had any previous strokes or diabetes.



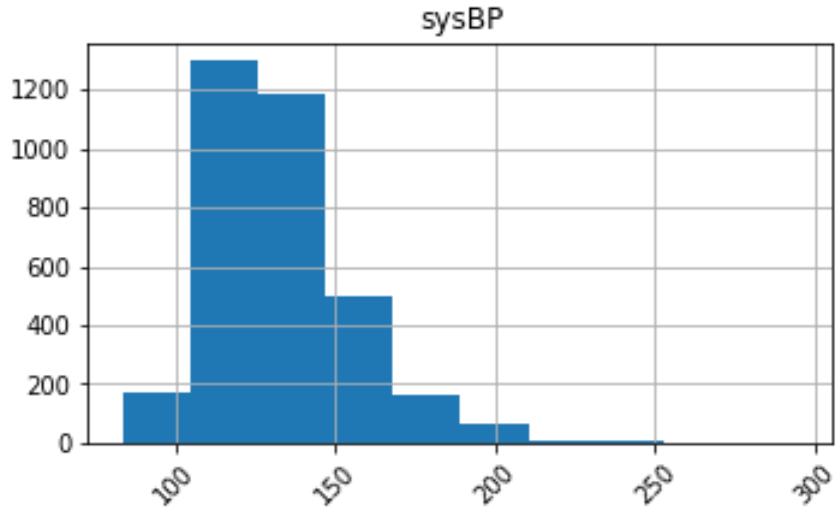
Insights from EDA

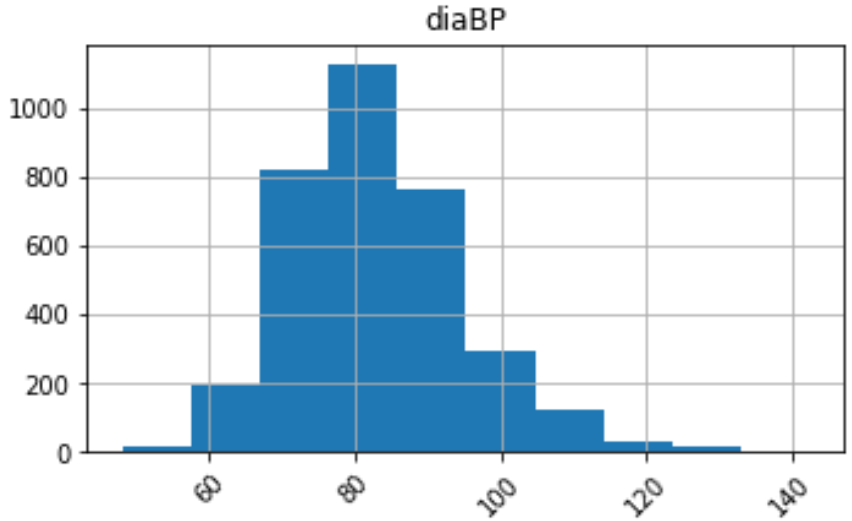
- More than 1000 patients were hypertensive.
- Most of the patients have total cholesterol level of 160 to 280.



Insights from EDA

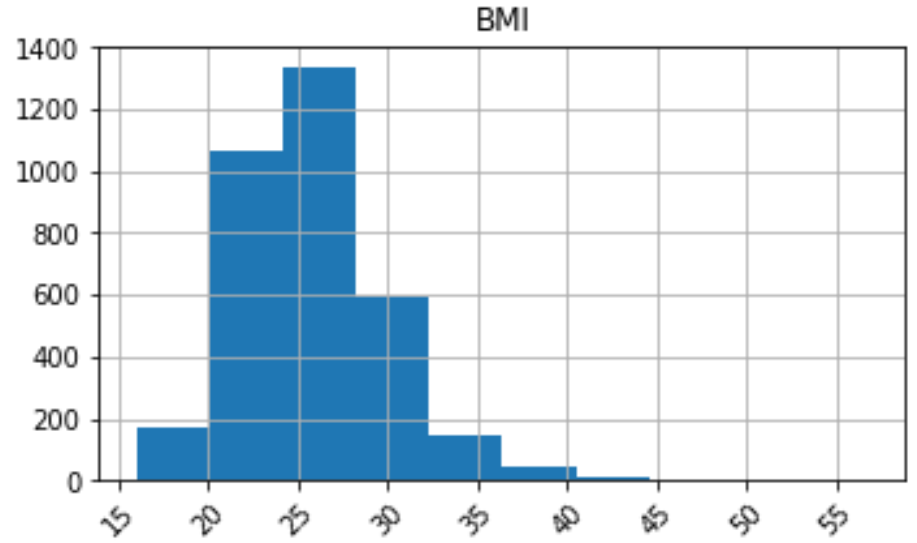
- Most of the patients have systolic blood pressure around 100 to 150 and diastolic blood pressure around 65 to 95.

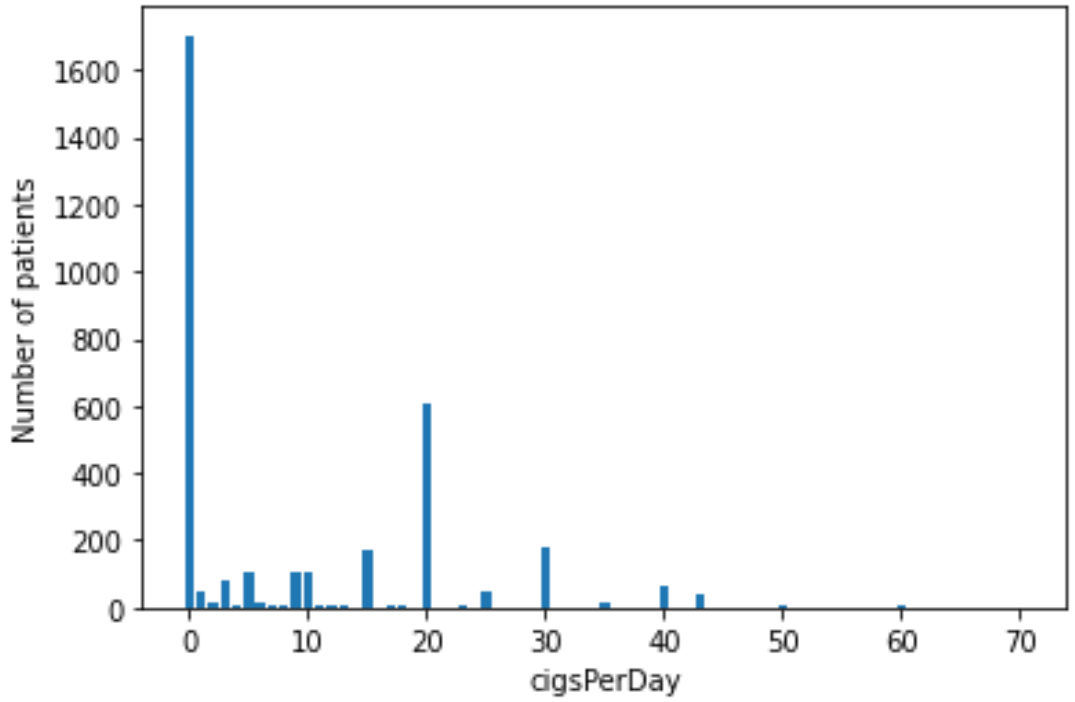




Insights from EDA

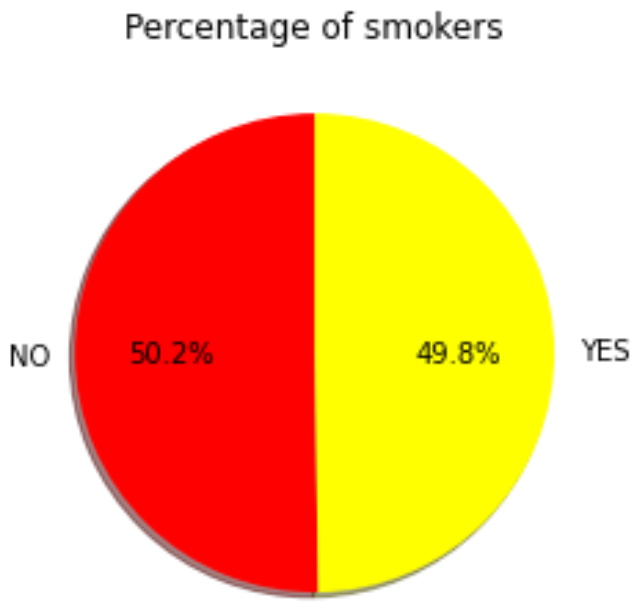
- BMI of most of the patients lies in the range of 20 to 30.
- More than 50% patients don’t smoke cigarette and there are more than 600 patients who smoke on an average 20 cigarettes per day.

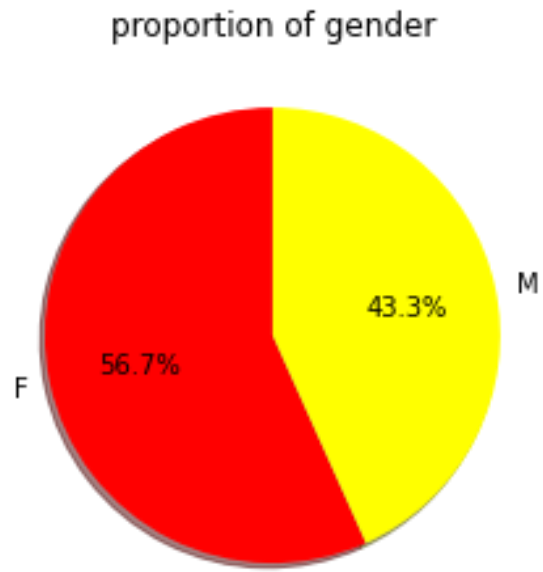




Insights from EDA

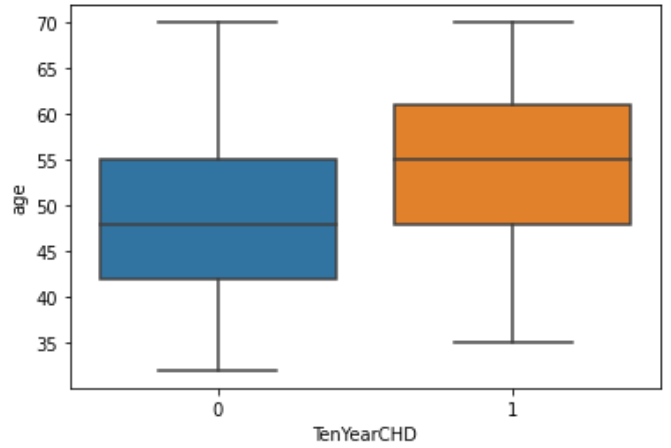
- The proportion of smokers and non-smokers are almost same.
- There are more female patients than male patients.

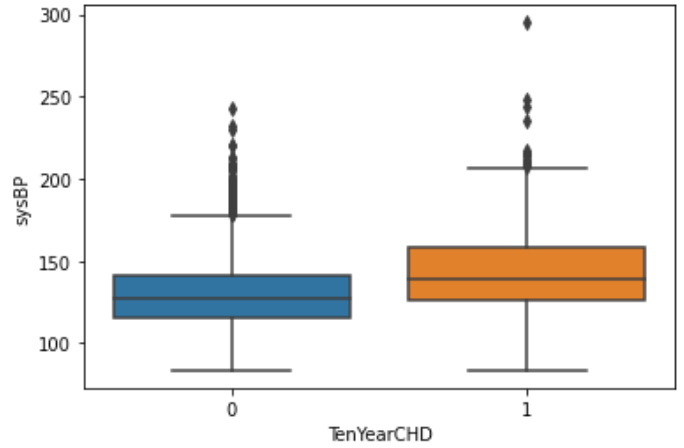


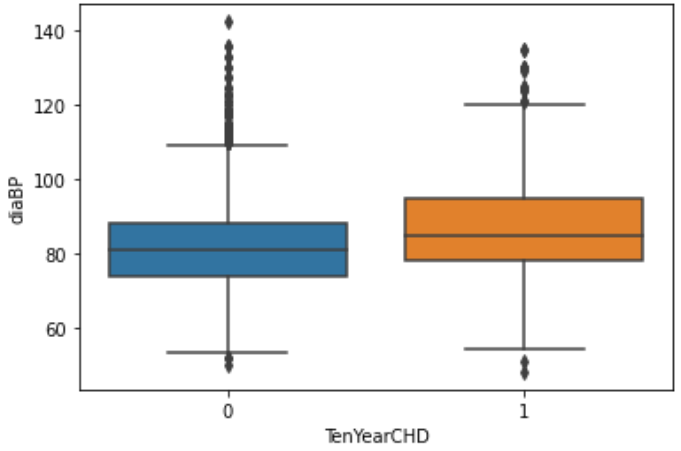


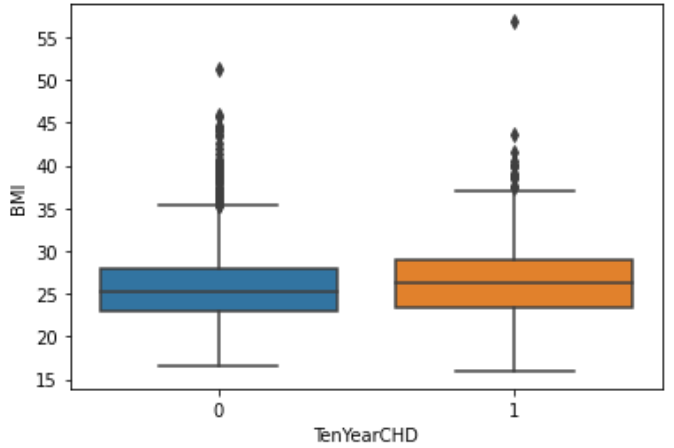
Insights from EDA

- The age is higher for the patients who have 10 year risk of CHD.
- Total cholesterol, sysBP, diaBP, BMI, and glucose are slightly higher in case of the patients who have 10 year risk of CHD.



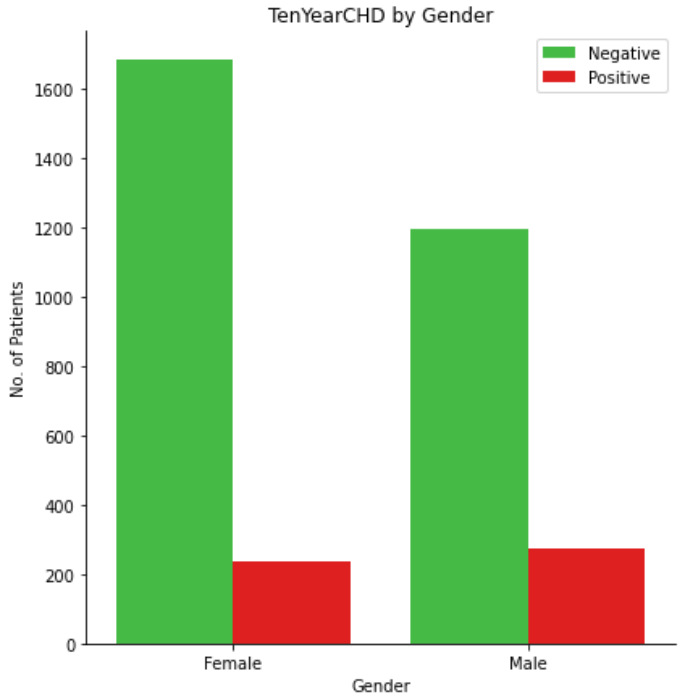


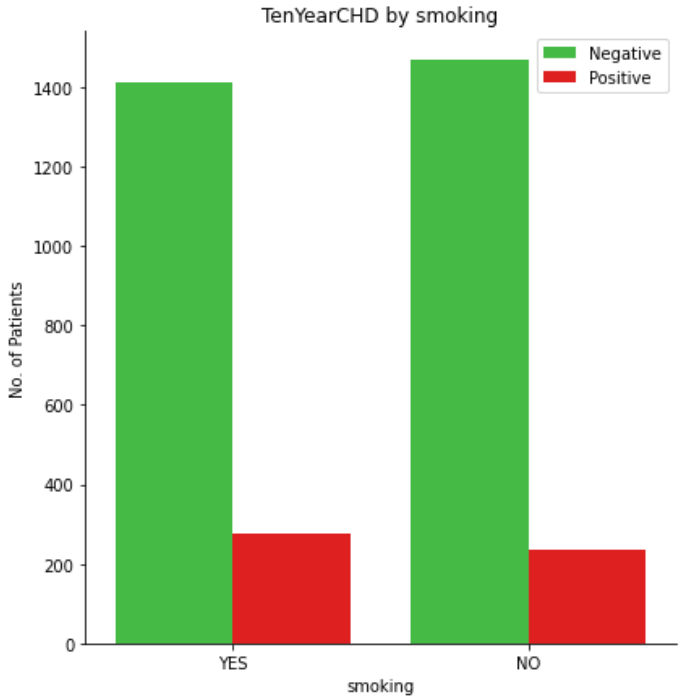




Insights from EDA

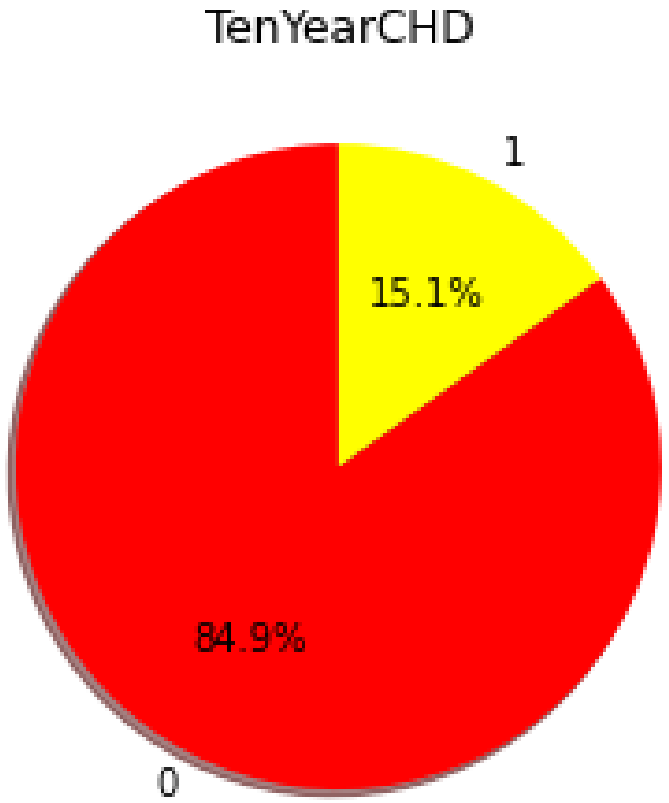
- 10 year risk of CHD is slightly more in case of male patients.
- 10 year risk of CHD is slightly more in case of patients who smoke.





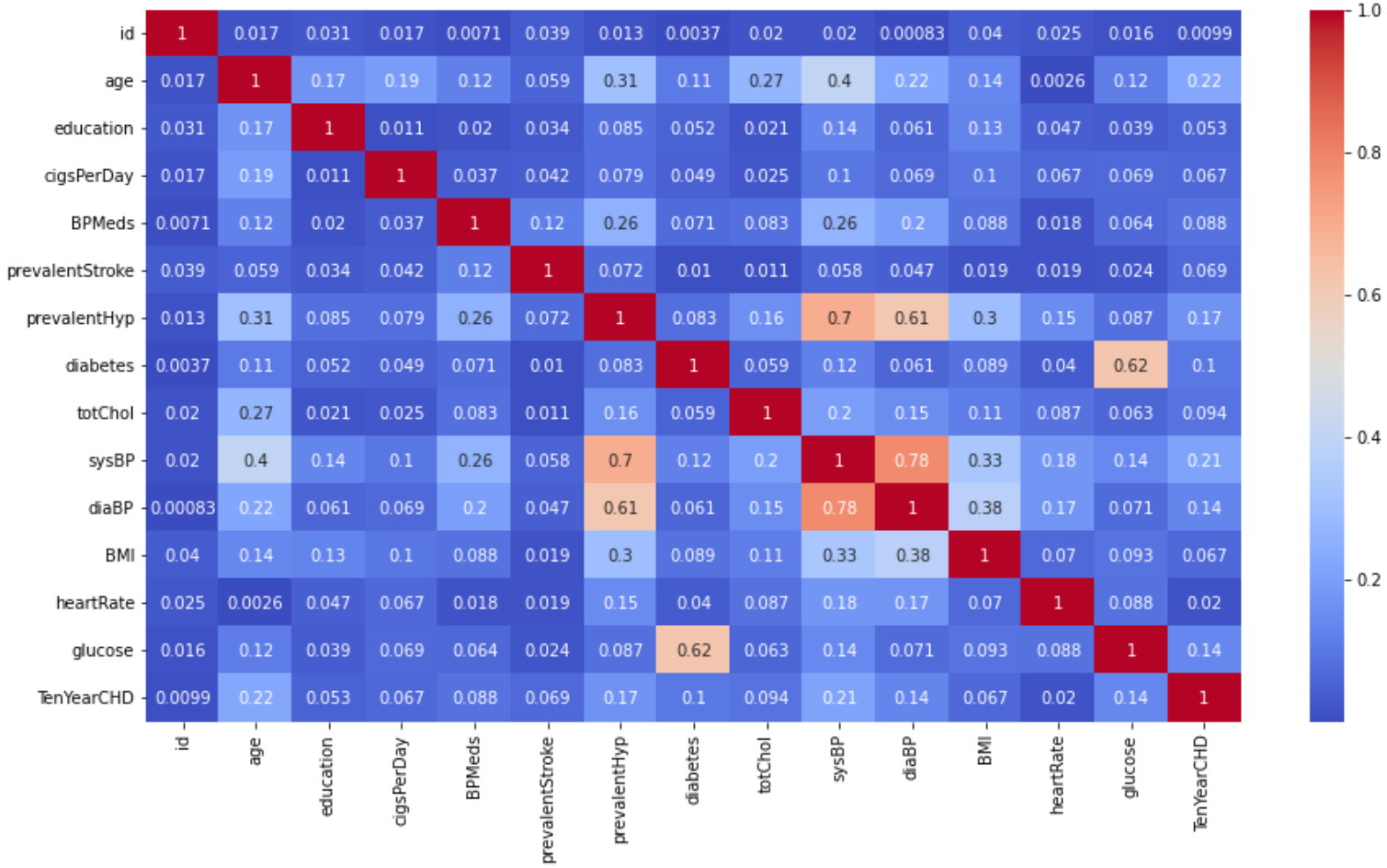
Dependent variable

- The dataset is heavily imbalanced.
- There are very less data (around 15%) for the patients who had 10-year risk of coronary heart disease.



Multicolinearity

- Features like sysBP and diaBP are highly correlated with each other. Also prevalentHyp is highly correlated with sysBP and diaBP.
- Glucose and diabetes are also highly correlated with each other.



Feature engineering

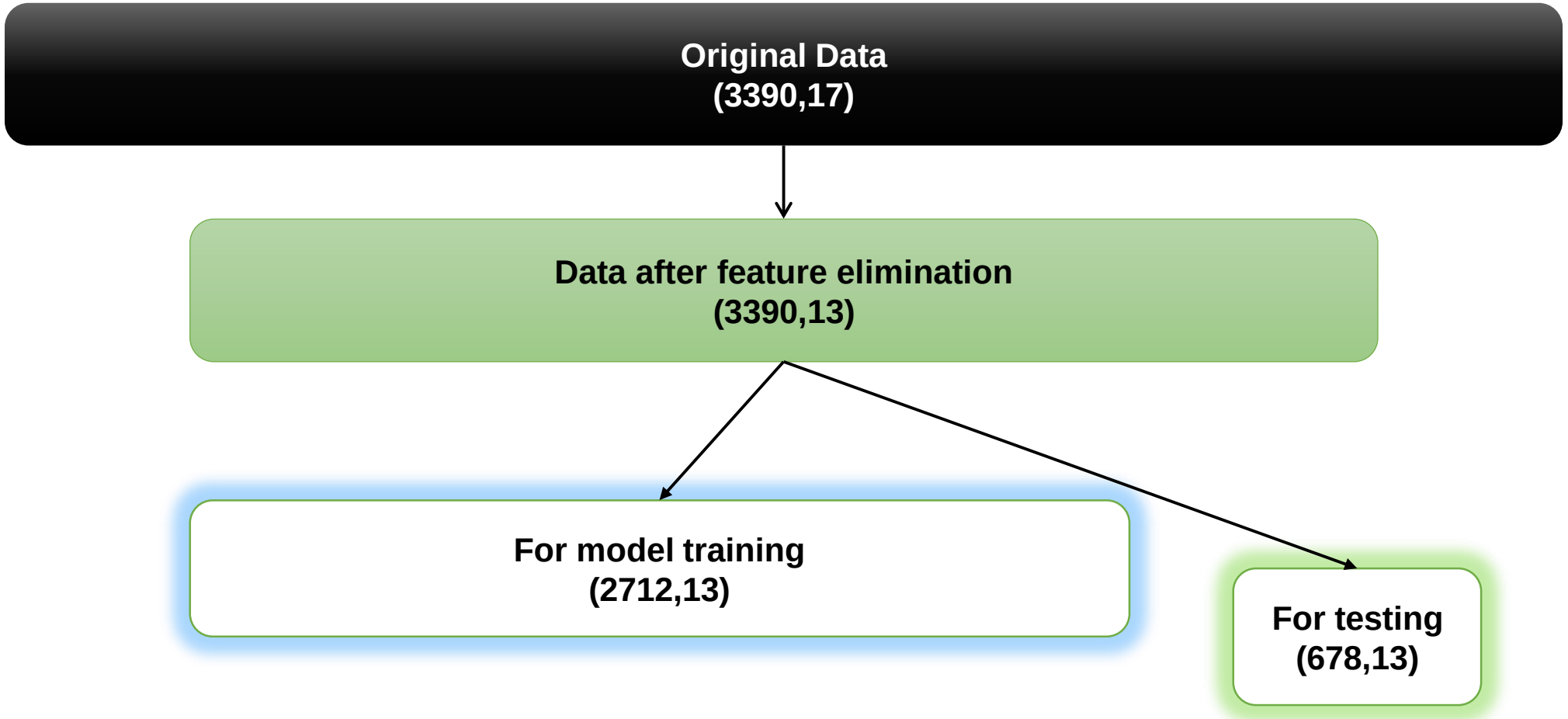
- We have created one new feature ‘BP’ by combining ‘sysBP’ and ‘diaBP’ features.
- We have applied label encoding for ‘sex’ and ‘is_smoking’ features as it comprises of only two distinct labels.
- We have used KNN imputer with n_neighbour = 1 to impute null values.
- We have dropped some features like id, education as it don’t impact the dependent variable.
- We have applied log transformation on all the continuous variable as distribution of these variables were right skewed.

Final Dataset

- After performing various operations like null value imputation, log transformation, encoding, feature elimination and feature combination the final dataset is like this.
- We will use this dataset to fit our model.

age	sex	cigsPerDay	BPMeds	Prevalent Stroke	Prevalent Hyp	diabetes	totChol	BP	BMI	heartRate	glucose	TenYearCHD
64.0	0.0	3.0	0.0	0.0	0.0	0.0	5.402677	116.50	3.299240	4.510860	4.394449	1.0
36.0	1.0	0.0	0.0	0.0	1.0	0.0	5.361292	133.00	3.426540	4.290459	4.330733	0.0
46.0	0.0	10.0	0.0	0.0	0.0	0.0	5.525453	93.50	3.061052	4.488636	4.553877	0.0
50.0	1.0	20.0	0.0	0.0	1.0	0.0	5.455321	123.00	3.376221	4.234107	4.553877	1.0
64.0	0.0	30.0	0.0	0.0	0.0	0.0	5.488938	110.75	3.311273	4.262680	4.356709	0.0

Data preparation for modelling

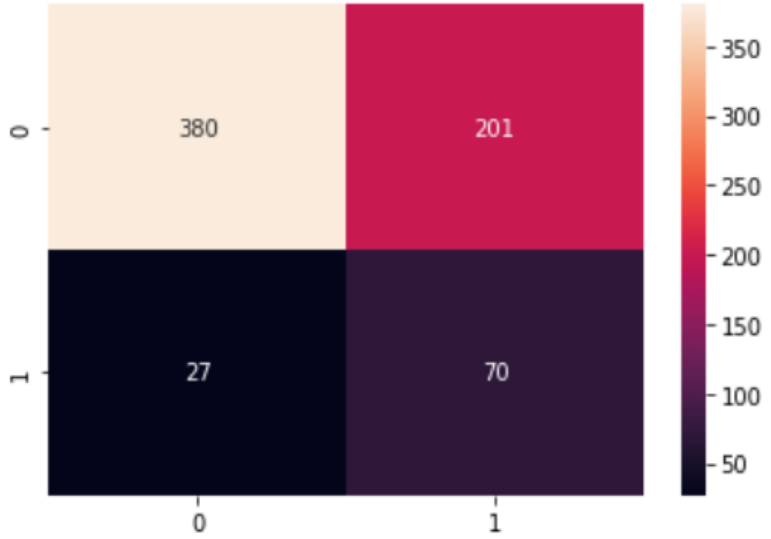


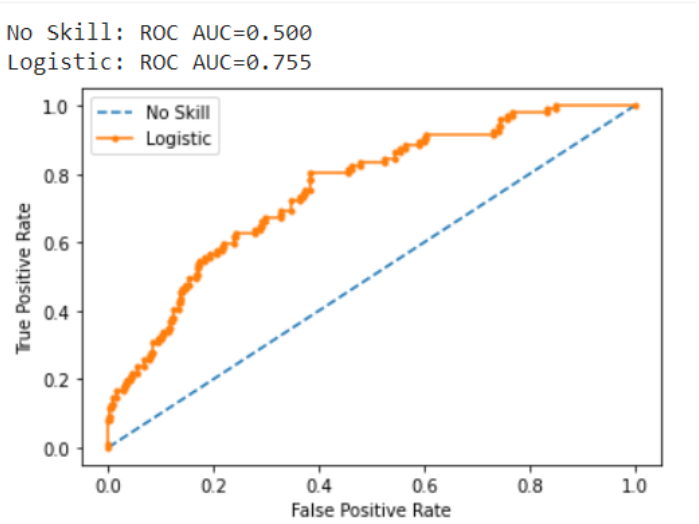
Selection of proper evaluation matrix

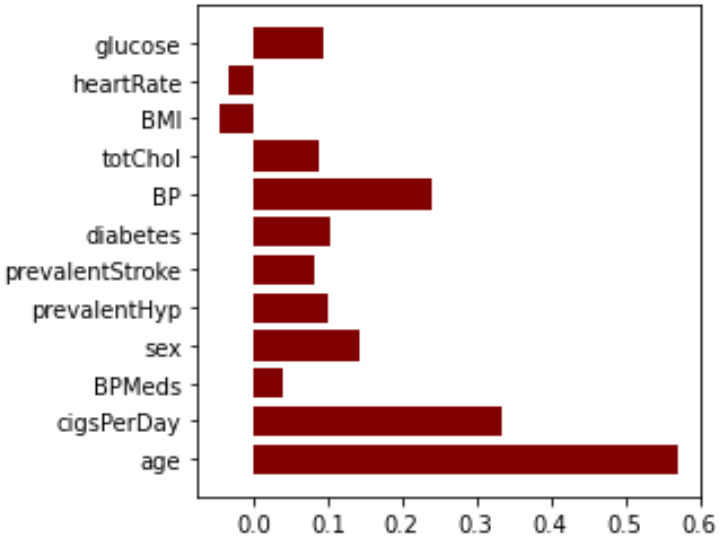
- The dataset is imbalanced with 85% of negative class. So ‘Accuracy’ will not be a good matrix to evaluate our model.
- As this is a health care domain project, falsely classified as negative should be our focus. So basically we need to reduce the false negative predictions.
- To summarize, recall and AUC ROC score will be our go to matrix for this project.

Logistic Regression

	precision	recall	f1-score	support
0.0	0.93	0.65	0.77	581
1.0	0.26	0.72	0.38	97
accuracy			0.66	678
macro avg	0.60	0.69	0.57	678
weighted avg	0.84	0.66	0.71	678



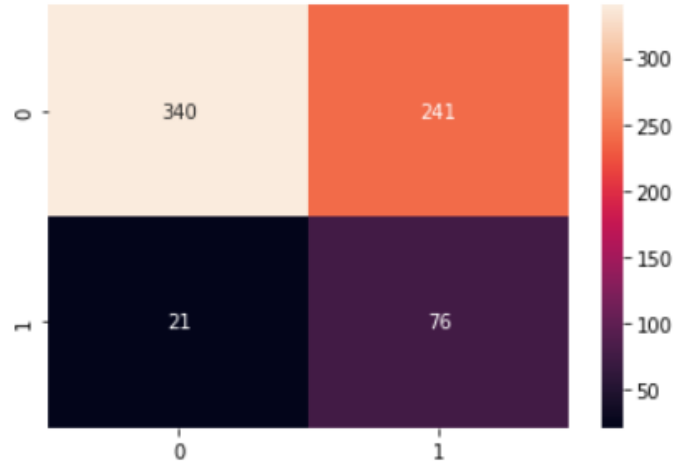




Logistic Regression

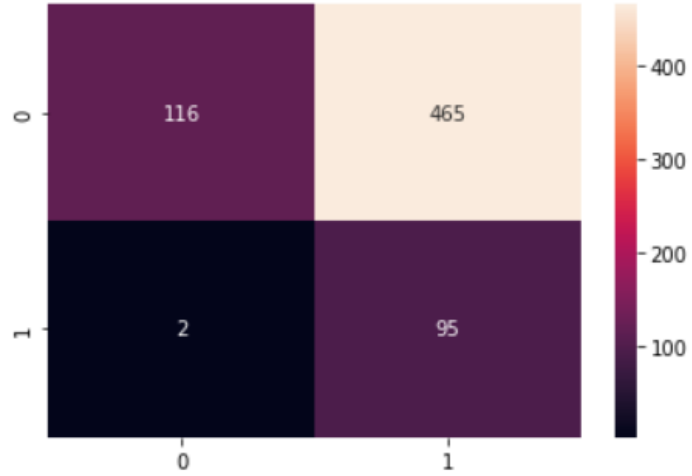
GridSearchCV

	precision	recall	f1-score	support
0.0	0.94	0.59	0.72	581
1.0	0.24	0.78	0.37	97
accuracy			0.61	678
macro avg	0.59	0.68	0.54	678
weighted avg	0.84	0.61	0.67	678



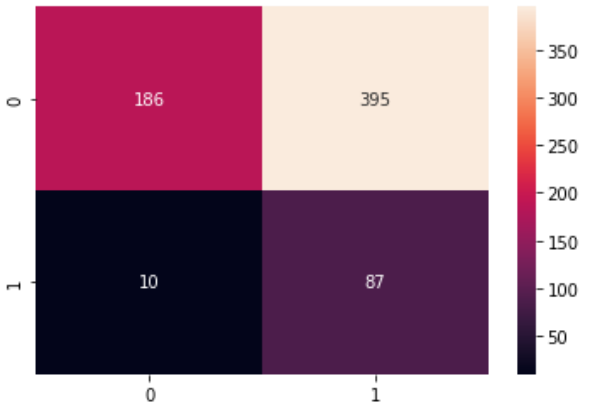
By changing default threshold

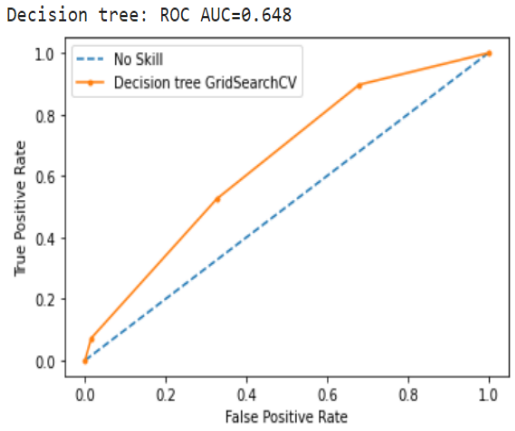
	precision	recall	f1-score	support
0.0	0.98	0.20	0.33	581
1.0	0.17	0.98	0.29	97
accuracy			0.31	678
macro avg	0.58	0.59	0.31	678
weighted avg	0.87	0.31	0.33	678



Decision tree

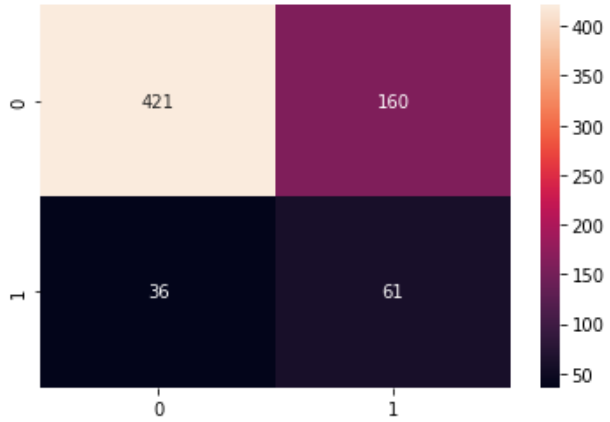
	precision	recall	f1-score	support
0.0	0.95	0.32	0.48	581
1.0	0.18	0.90	0.30	97
accuracy			0.40	678
macro avg	0.56	0.61	0.39	678
weighted avg	0.84	0.40	0.45	678

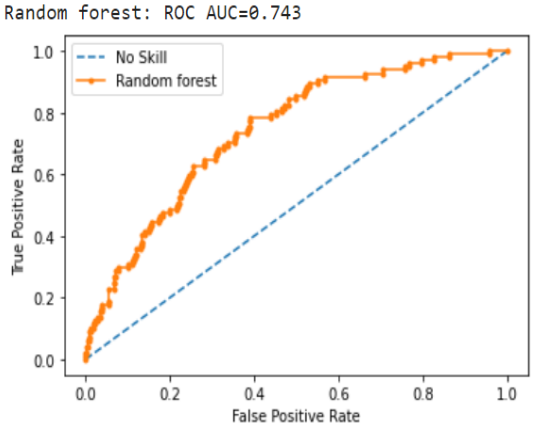




Random forest

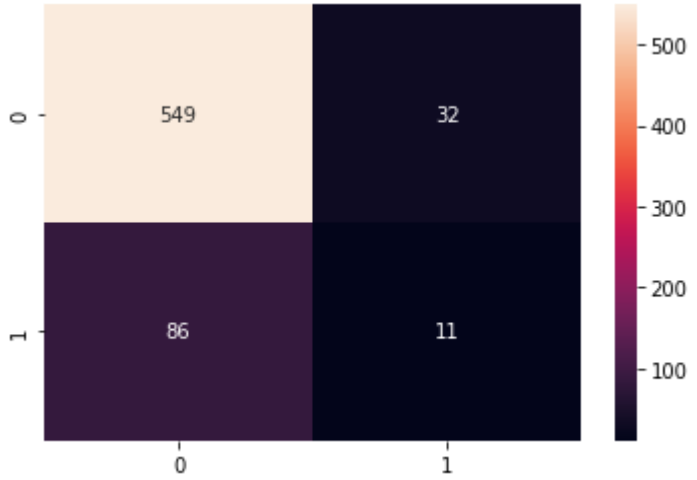
	precision	recall	f1-score	support
0.0	0.92	0.72	0.81	581
1.0	0.28	0.63	0.38	97
accuracy			0.71	678
macro avg	0.60	0.68	0.60	678
weighted avg	0.83	0.71	0.75	678





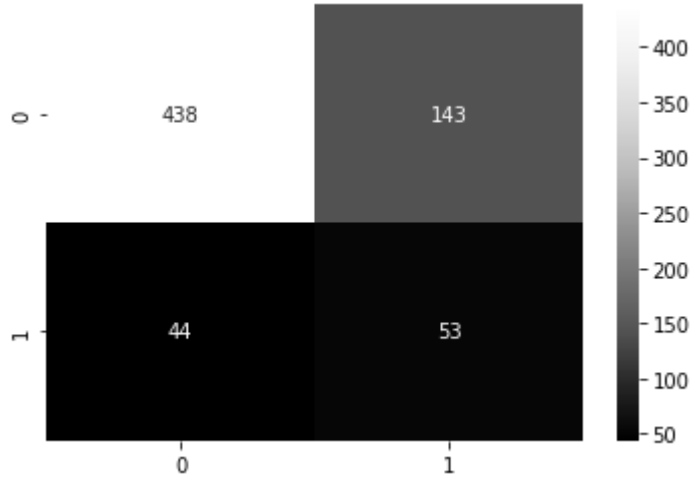
XGBoost

	precision	recall	f1-score	support
0.0	0.86	0.94	0.90	581
1.0	0.26	0.11	0.16	97
accuracy			0.83	678
macro avg	0.56	0.53	0.53	678
weighted avg	0.78	0.83	0.80	678



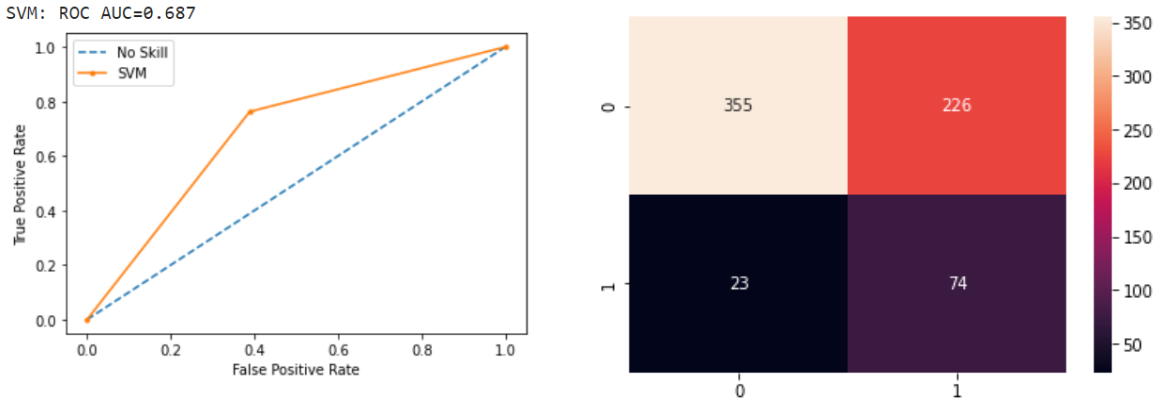
XGBoost with SMOTE

	precision	recall	f1-score	support
0.0	0.91	0.75	0.82	581
1.0	0.27	0.55	0.36	97
accuracy			0.72	678
macro avg	0.59	0.65	0.59	678
weighted avg	0.82	0.72	0.76	678



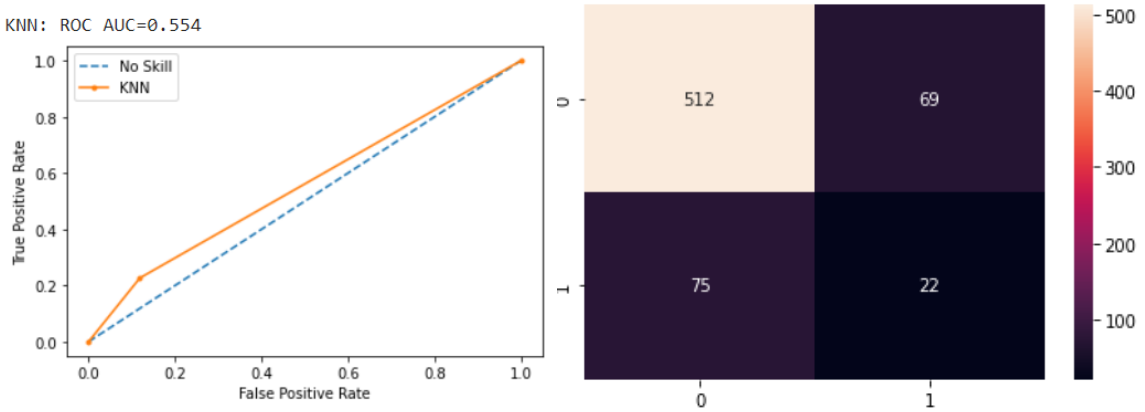
Support vector

	precision	recall	f1-score	support
0.0	0.94	0.61	0.74	581
1.0	0.25	0.76	0.37	97
accuracy			0.63	678
macro avg	0.59	0.69	0.56	678
weighted avg	0.84	0.63	0.69	678

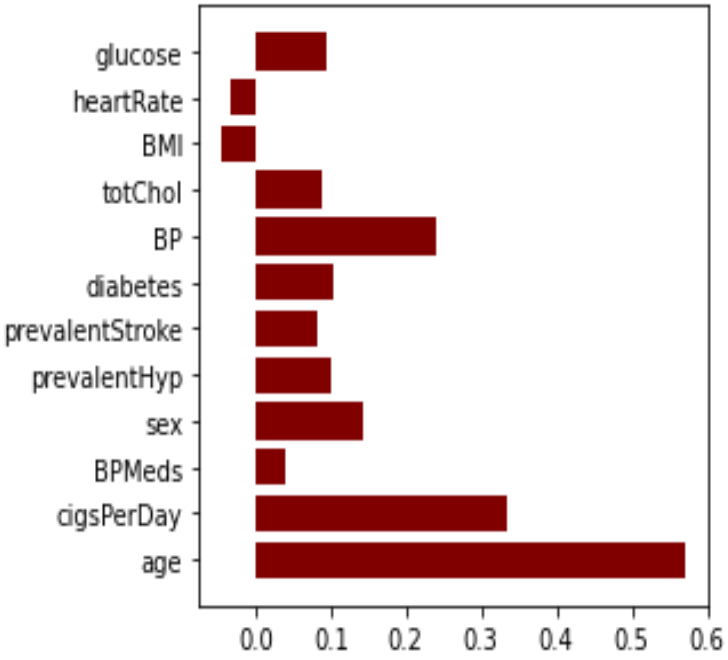


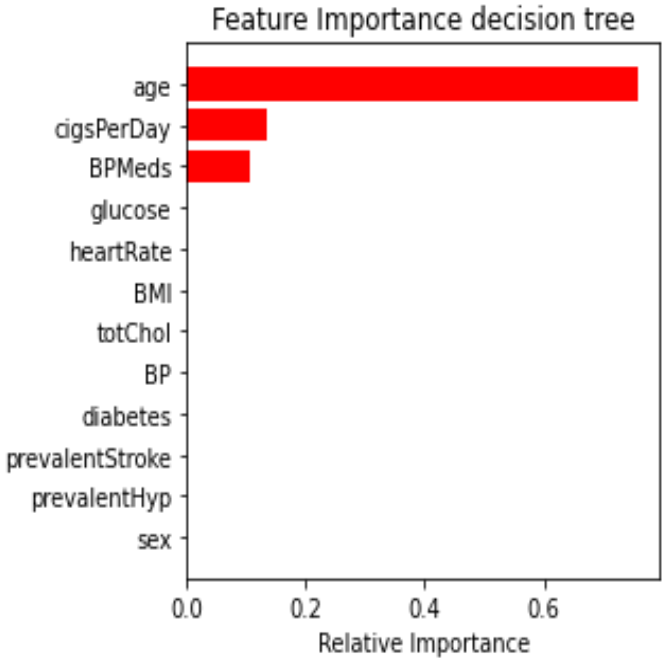
KNN

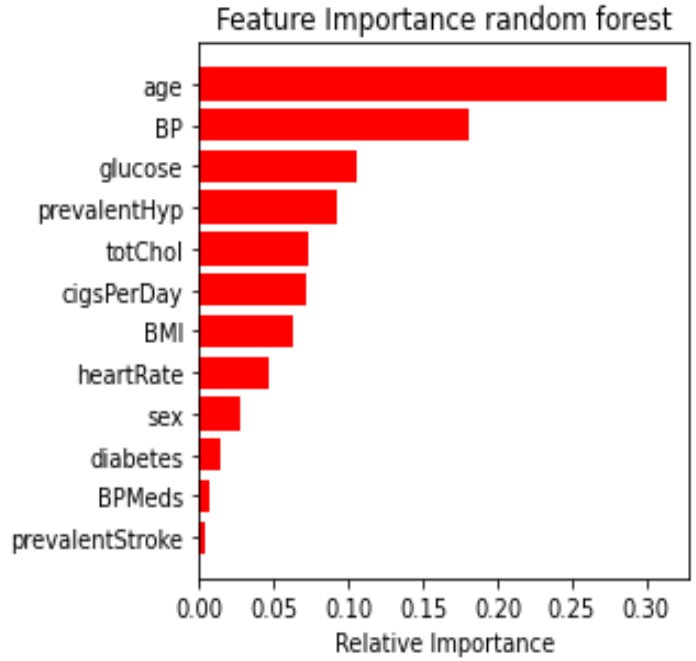
	precision	recall	f1-score	support
0.0	0.87	0.88	0.88	581
1.0	0.24	0.23	0.23	97
accuracy			0.79	678
macro avg	0.56	0.55	0.56	678
weighted avg	0.78	0.79	0.78	678



Feature importance





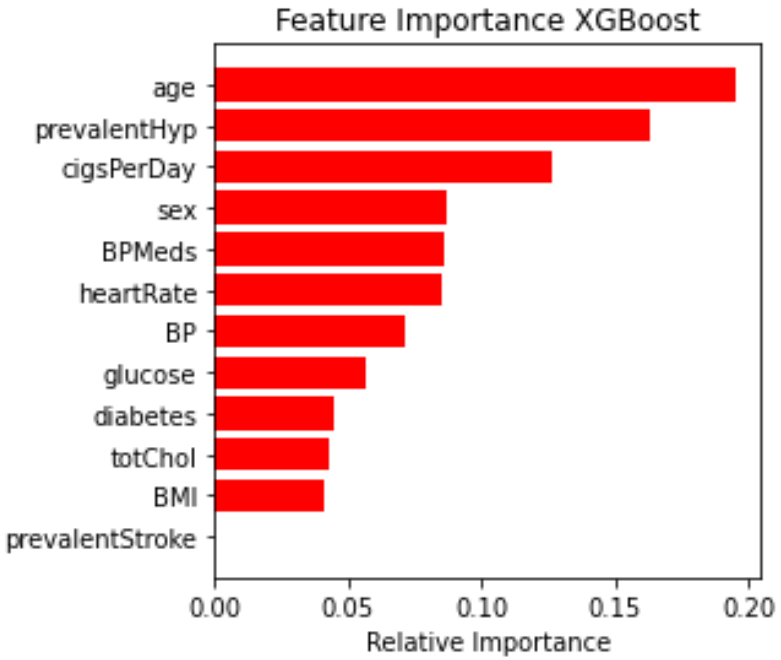


Logistic Regression

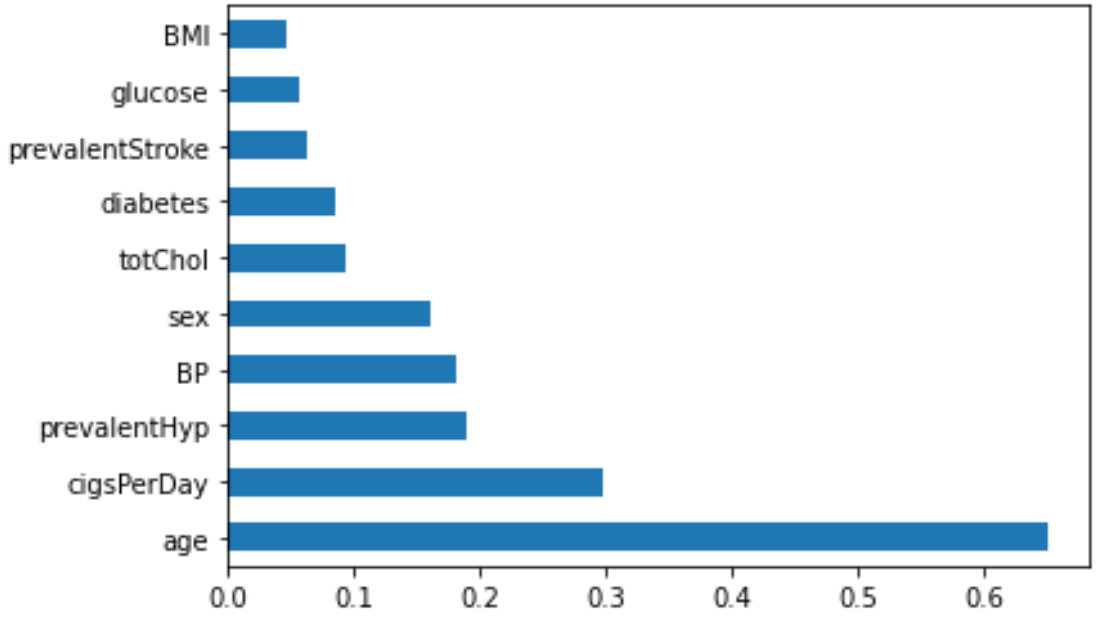
Decision tree

Random forest

Feature importance(Contd.)



XGBoost



Support vector

Conclusion

- Age and cigsPerDay are the two most important features given by most of the models.
- Logistic regression, random forest and support vector machine models are giving a good overall balanced result.
- Models like decision tree and logistic regression(by changing threshold) are giving very good recall score but they are certainly increasing the false positive predictions.

Challenges

- Handling null values.
- Dealing with multicolinearity.
- Selecting most relevant features.
- Selecting relevant set of hyper parameters for tuning.
- Computation time during GridSearchCV.

Thank
you