Heart Disease Prediction

Intro:

Heart disease is a leading cause of death globally and early detection is crucial in preventing the progression of the disease (W.H.O).

Goal:

The goal of the project is to predict if a person has a coronary heart disease based upon factors such as age, smoking habits, BMI. For this I have used logistic regression and Support vector machines and drew comparisons between both the models.

Exploratory Data Analysis:

The dataset I have used for this project is Framingham data set this dataset is a popular one with around 3000 peer reviewed scientific papers which have been published related to this. This study has been described as "on its way to becoming the gold standard for cardiovascular genetic epidemiology" -wikipedia

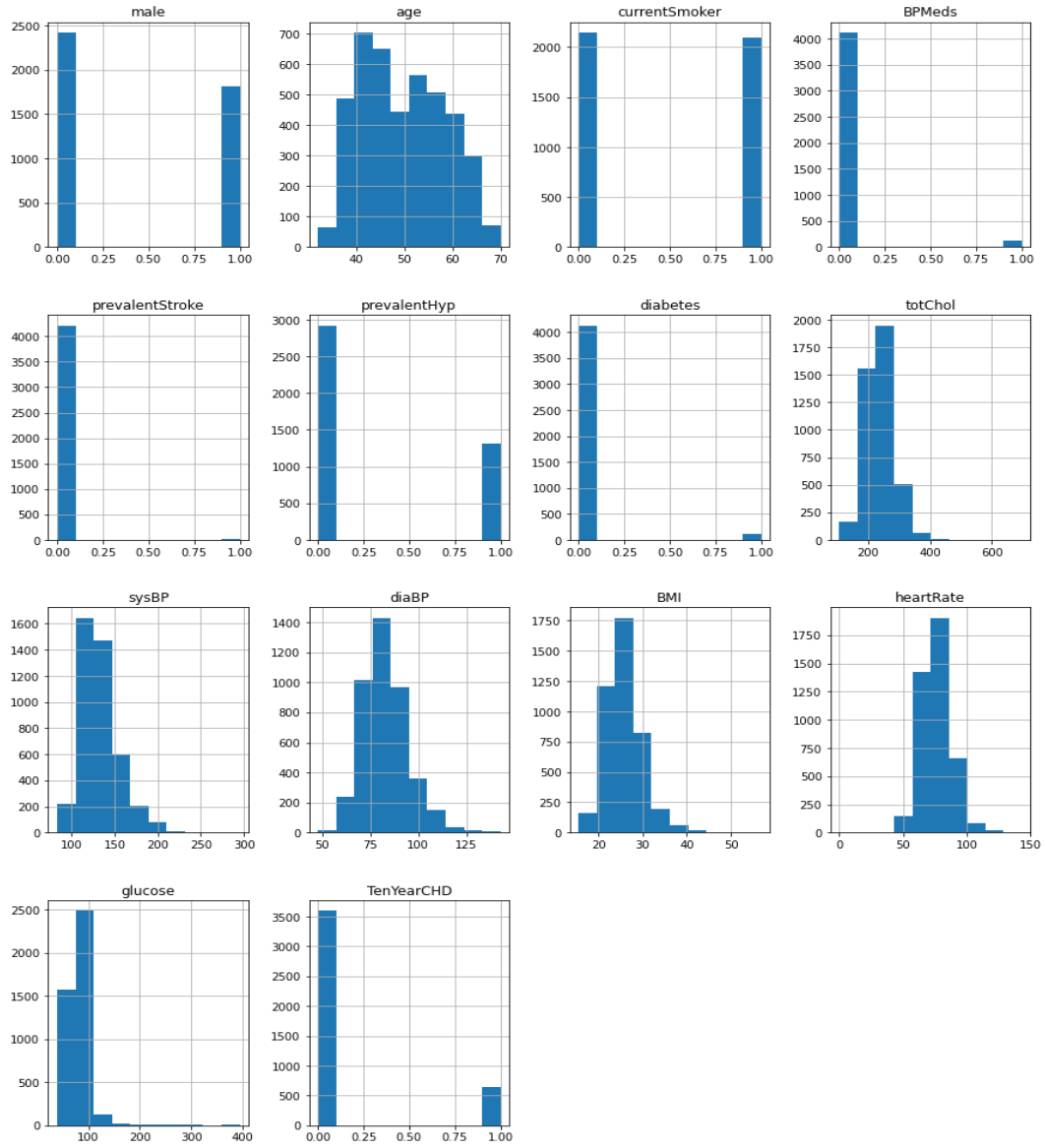
The data set consists of 4238 rows and 16 columns and the data is all numerical with no duplicated values, each column represents a feature the definitions for each feature are

framingham_heart_disease

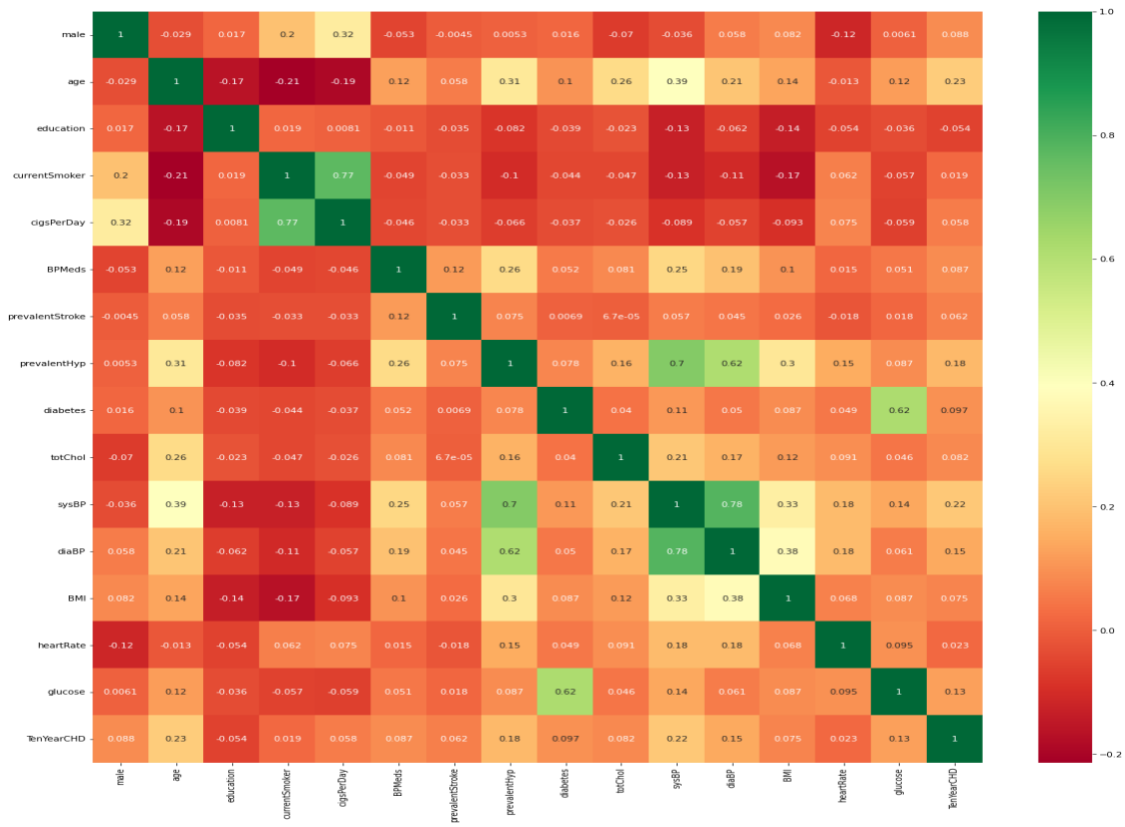
male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0

Feature	Description	Data collection type	Range
age	age of the patient	Continuous	20-99
education	education		0-1
currentSmoker	whether or not the patient is a current smoker	Nominal	0-1
cigsPerDay	the number of cigarettes that the person smoked on average in one day	Continuous	1+
BPMeds	whether or not the patient was on blood pressure medication	Nominal	0-1
prevalentStroke	whether or not the patient had previously had a stroke	Nominal	0-1
prevalentHyp	whether or not the patient was hypertensive	Nominal	0-1
diabetes	whether or not the patient had diabetes	Nominal	0-1
totChol	total cholesterol level	Continuous	0-1
sysBP	systolic blood pressure	Continuous	100-140
diaBP	diastolic blood pressure	Continuous	60-100
BMI	Body Mass Index	Continuous	20-40
heartRate	heart rate	Continuous	60-100
glucose	glucose level	Continuous	70-150
TenYearCHD	10 year risk of coronary heart disease	Continuous	0-1

Checking the distribution of data using histograms



Some observations from the histograms the graph. TenyearCHD which is our target variable has more bias with 520 cases with 1 and 3500+ cases with 0 similar cases with diabetes, BPmeds and prelevantstroke. Due to this the accuracy is going to be more biased towards 0.



Correlation graph

The row current smoker and cigs per day are highly corelated in which one can be ignored as both of them have similar data whenever cigs per day is 0 current smoker will be 0 .There is no influence of education on heart disease so that can be removed from the features.

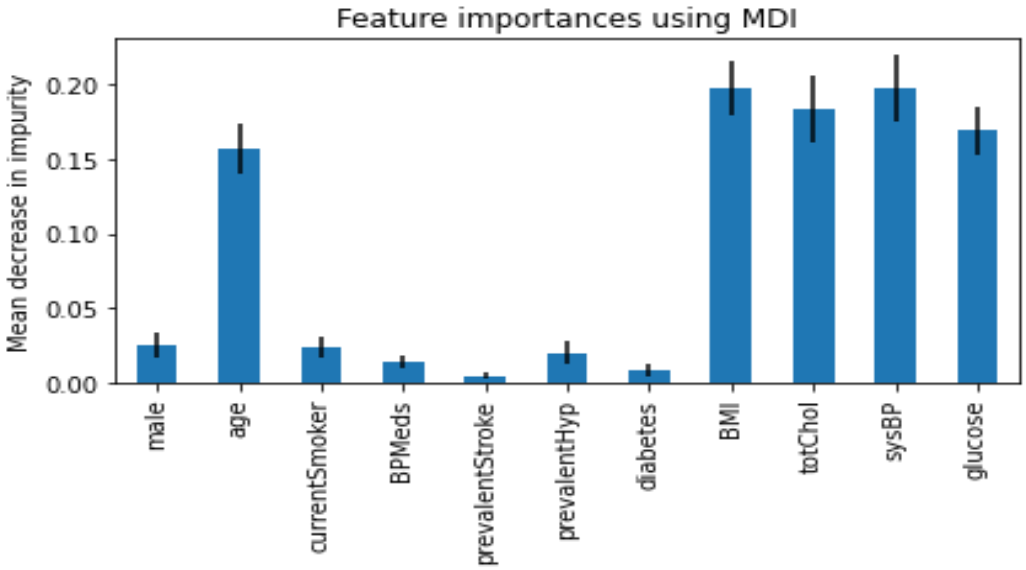
Feature Selection:

Feature selection process is one of the main components of a feature engineering process. This is how a model is developed by reducing the number of input variables.

Feature selection techniques are used to reduce the number of input variables by eliminating redundant or irrelevant features here in our dataset we have 14 columns currently in which not everything is useful for heart disease prediction.

So I used Feature importance based on mean decrease in impurity (MDI) MDI is a widely adopted measure of feature importance in random forests it is also known as gini importance it sums up the gain associated to all splits performed along a given variable.MDI relies on tree structure and heavily relies on trees it is used to calculate the variable importance

After applying feature importance, the result is plotted as a graph which is



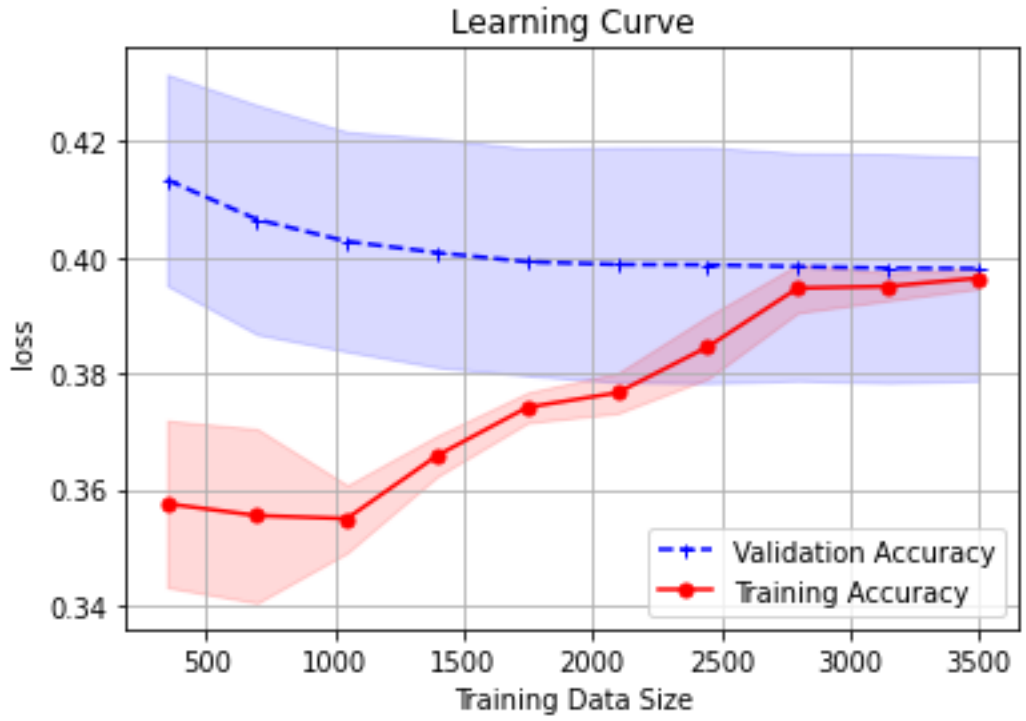
After sorting all the features based on their importance i received the following 5 as my most important features

- BMI
- SysBP
- TotChol
- Glucose
- Age

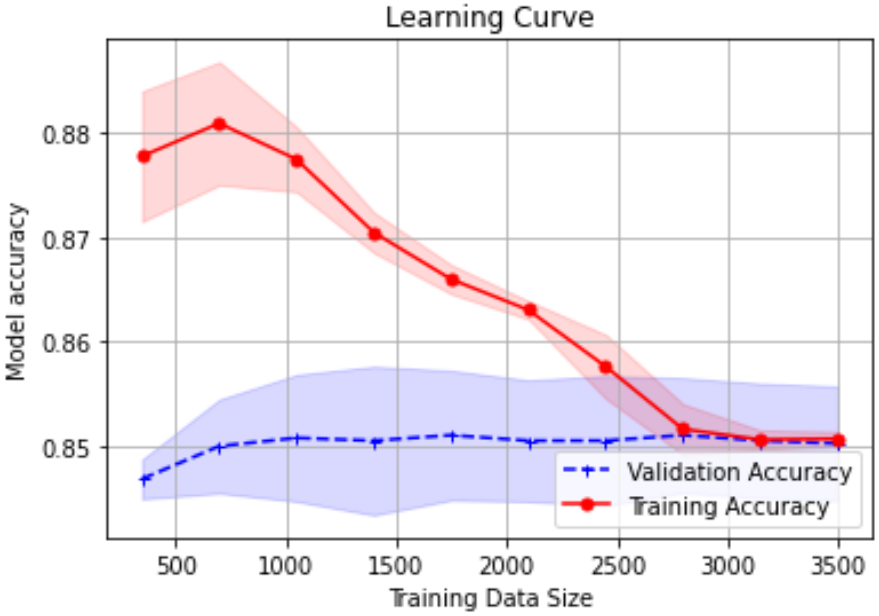
Logistic regression

It is a right regression model to use in case of binary output since we must predict If the patient has a heart disease or not, I choose Logistic regression as the right model for It. The dataset is splitted into Training 3814,5 and testing 424,5

The learning curve is made using sklearn.model_selection



The model is training and if trained on more datapoints the loss would converge



The start graph is little distorted. After using more that 1500 training examples model starts to perform well. Graph stabilizes around 0.85 and slowly decreases as we increase training examples

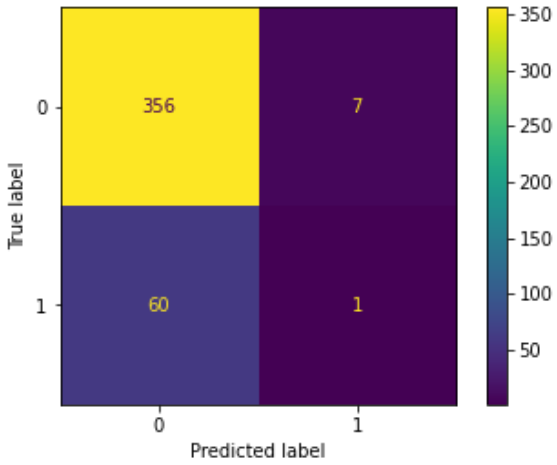
Classification report

	precision	recall	f1-score	support
0	0.86	0.98	0.92	363
1	0.45	0.08	0.14	61
accuracy			0.85	424
macro avg	0.66	0.53	0.53	424
weighted avg	0.81	0.85	0.81	424

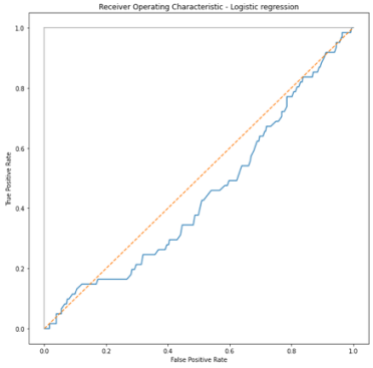
Precision: Out of all the CHD patients that the model predicted would have CHD, only **45%** actually did.

Recall: Out of all the CHD patients that actually did get CHD, the model only predicted this outcome correctly for **8%** patients.

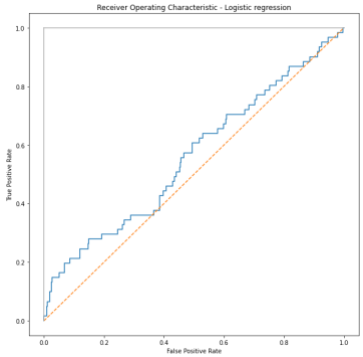
Confusion matrix



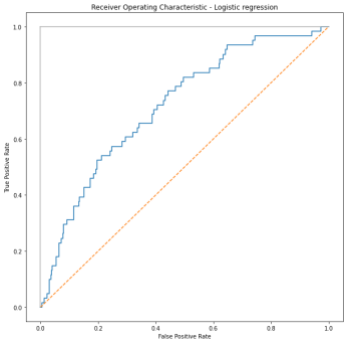
ROC curves with different c values c is Inverse of regularization strength



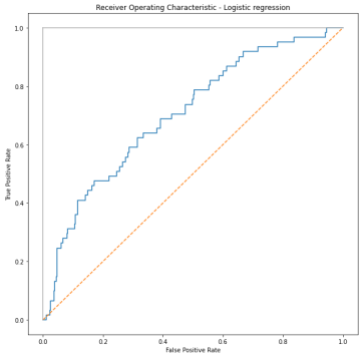
c=0.001



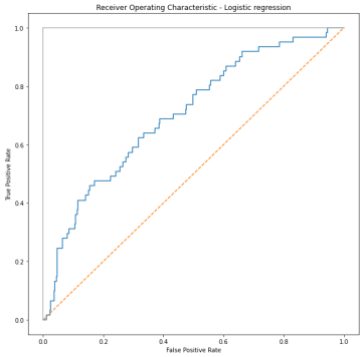
c=0.01



c=0.1



c=1



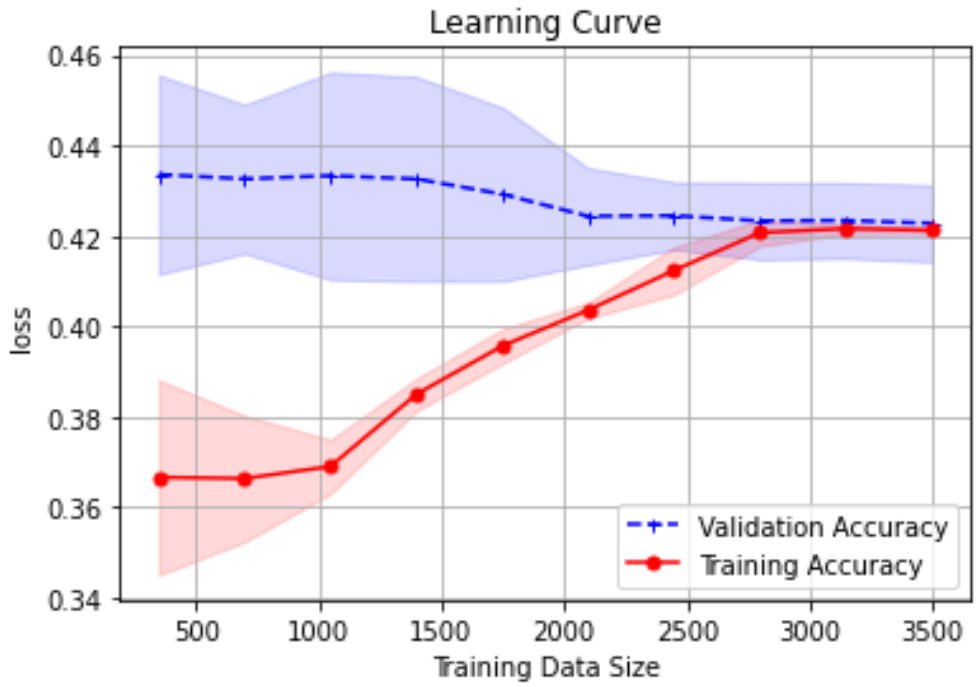
c=10

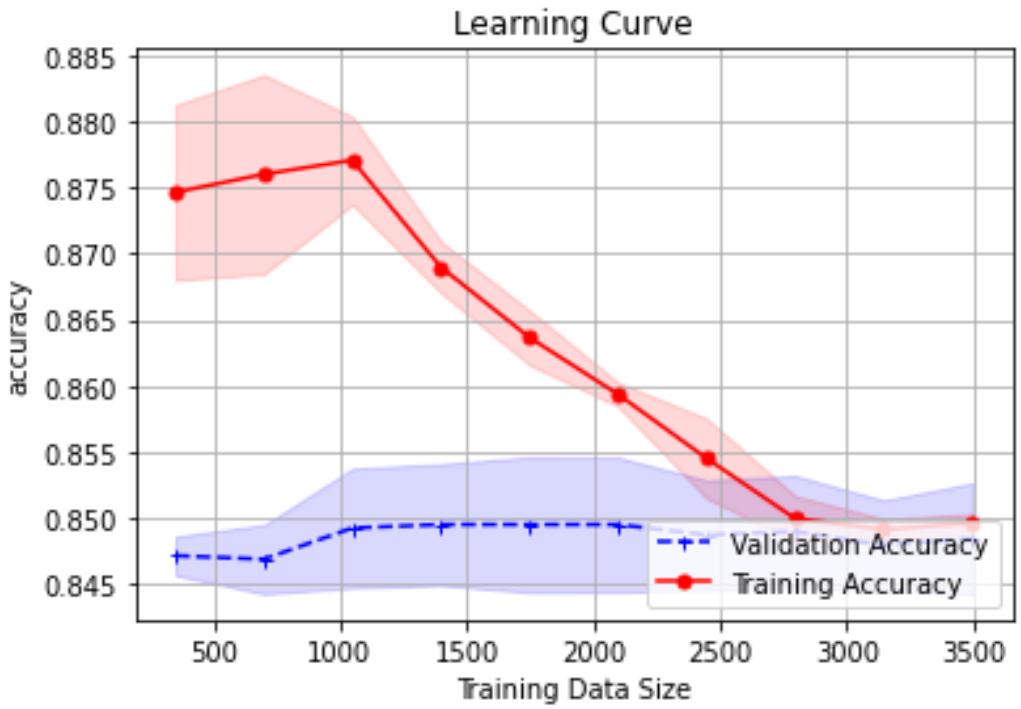
Hyperparameter tuning

I tried to change the values of c and checked the values of accuracy and roc_auc_score here are the results.

C	Accuracy score	Recall score	Roc_auc_score
0.001	0.856132	0.856132	0.5579
0.01	0.856132	0.856132	0.55796
0.1	0.853774	0.853773	0.71693
1	0.841981	0.841981	0.70482
10	0.841981	0.841981	0.70396

SVM (Support vector Machines):

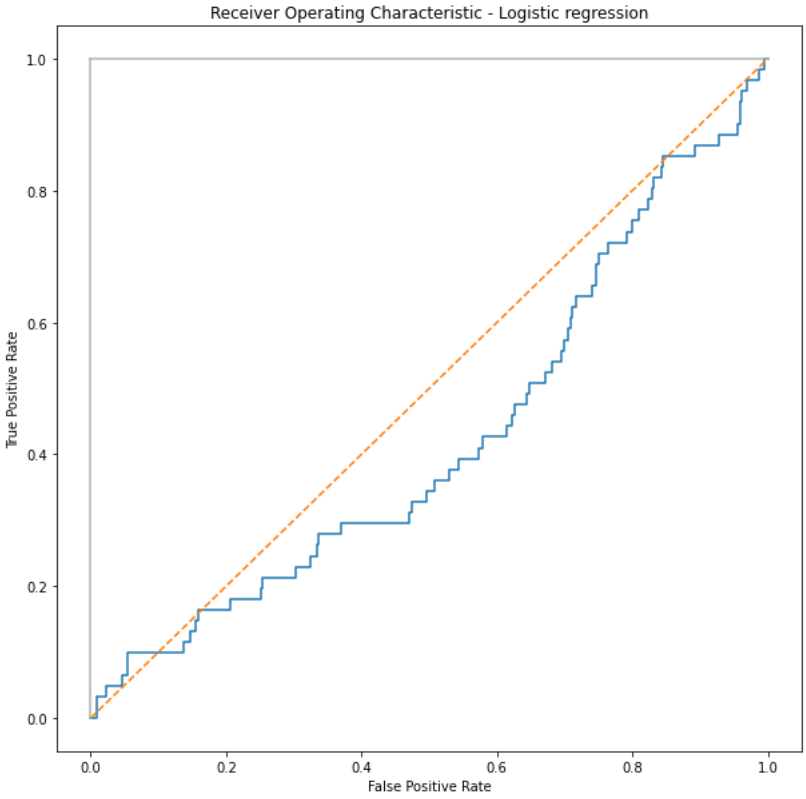




Classification report

	precision	recall	f1-score	support
0	0.86	1.00	0.92	363
1	0.00	0.00	0.00	61
accuracy			0.86	424
macro avg	0.43	0.50	0.46	424
weighted avg	0.73	0.86	0.79	424

ROC curve



accuracy score: 0.856132
Precision score: 0.8561320754716981
Recall score: 0.8561320754716981

Roc_auc_score for Logistic Regression: 0.4260940251998374

References:

https://www.ahajournals.org/doi/full/10.1161/01.CIR.103.9.1245
https://www.igi-global.com/article/prediction-of-heart-diseases-using-data-mining-techniques/223163