CAPSTONE PROJECT - 3

Cardiovascular Risk Prediction

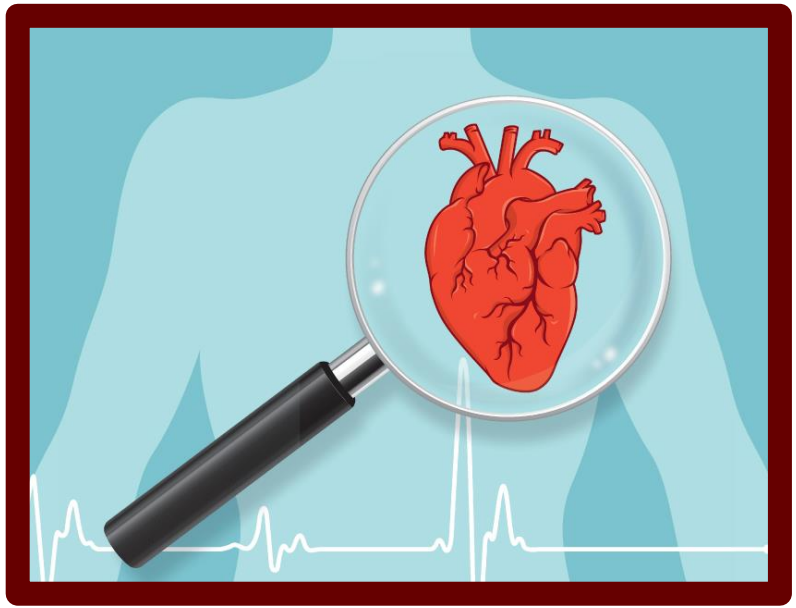
Created By : Sakshi Dhyani

Project Details

In this project, the dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The target is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

Steps performed

- Data cleaning
- Data visualizations
- Data preprocessing
- Model Implementation
- Evaluation metrics

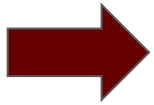


Data Summary

Independent Features

Sex :	Male or Female
Age :	Age of the patient
Is_smoking :	whether patient smokes or not
Cigs Per Day :	average no of cigarettes that person smoked on one day
BP Meds:	Whether patient is on blood pressure medication or not
Prevalent Stroke:	Whether patient previously had a stroke
Prevalent Hyp:	Whether patient was hypertensive or not
Diabetes:	Whether or not patient has diabetes
Tot Chol :	Total Cholesterol level
Sys BP :	Systolic Blood pressure
Dia BP :	Diastolic Blood pressure
BMI :	Body Mass Index
Heart Rate	
Glucose :	Glucose level

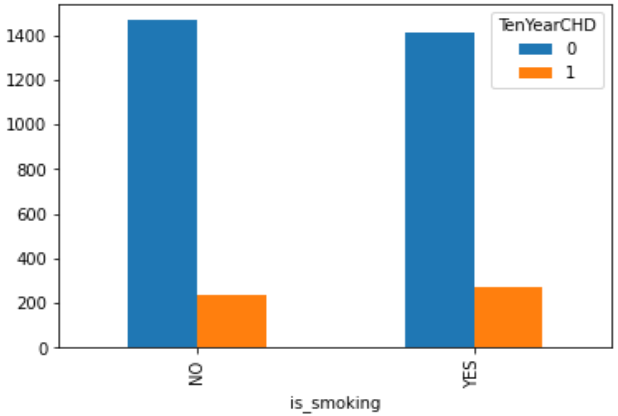
Dependent Feature

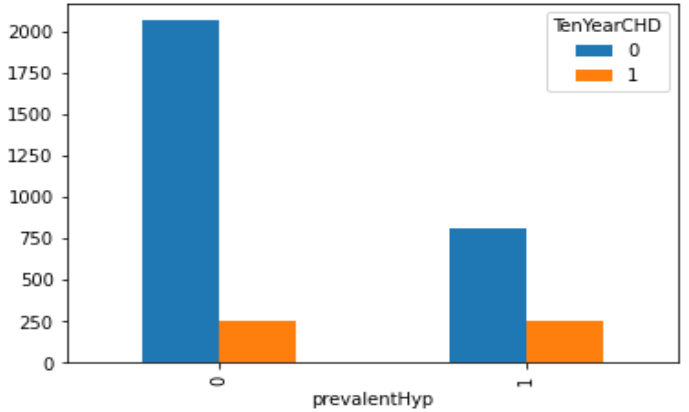


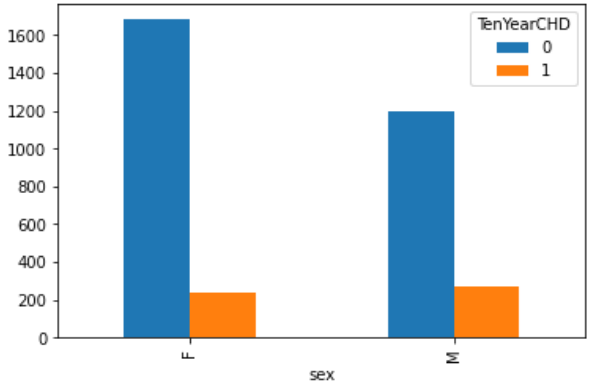
10- year risk of coronary heart disease (CHD)

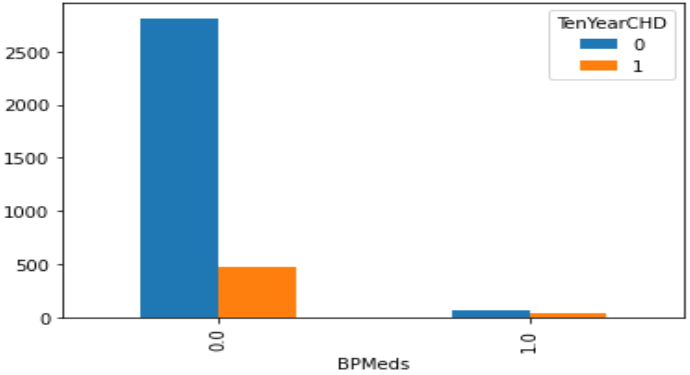
Exploratory Data Analysis

Bar Plots



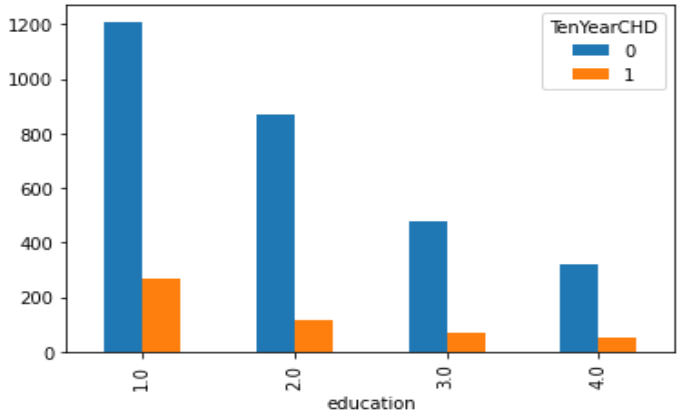


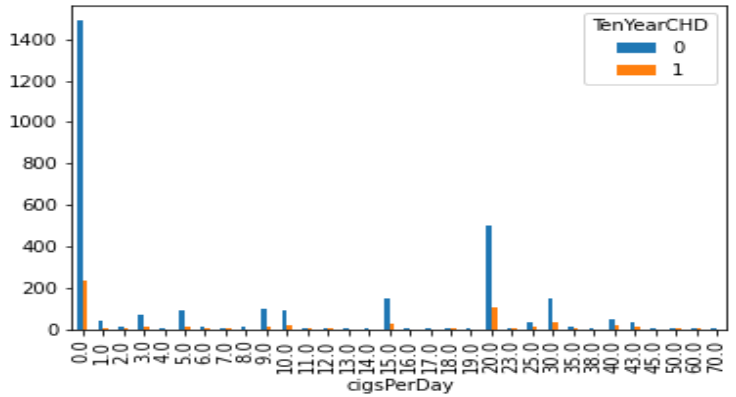


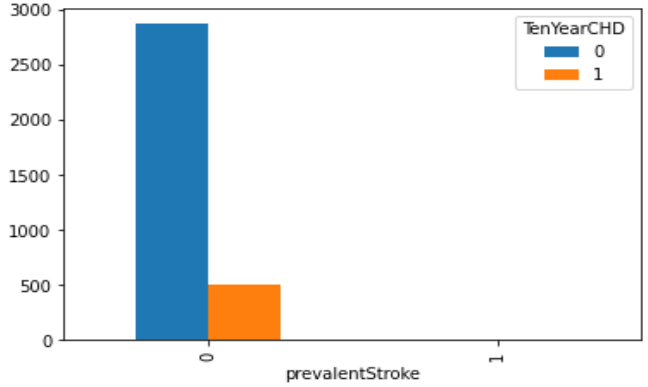


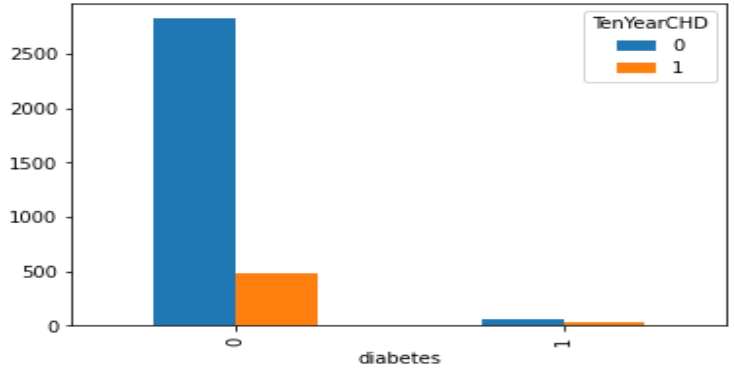
Exploratory Data Analysis

Bar Plots



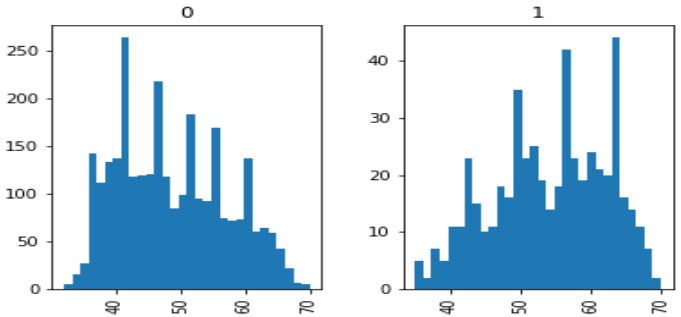




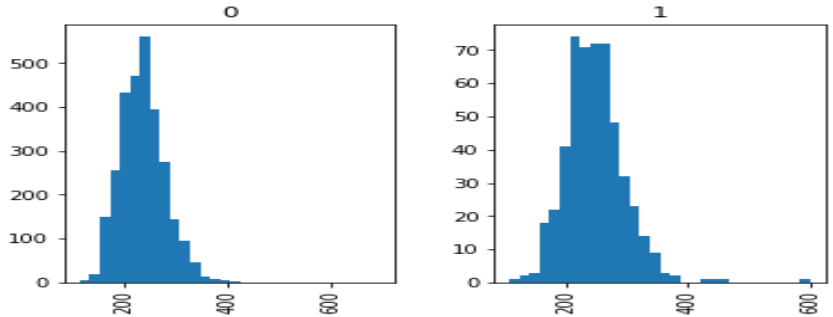


Distribution plots for different independent variables as per target variable labels

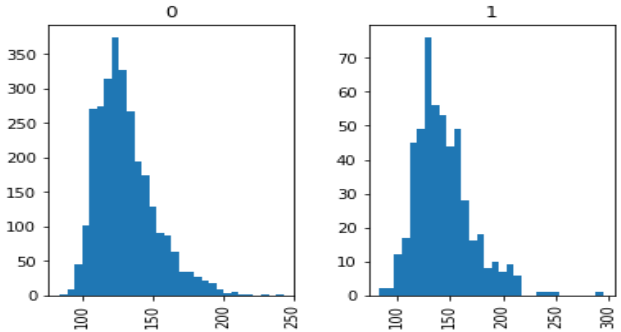
Ten Year CHD Values for different age values



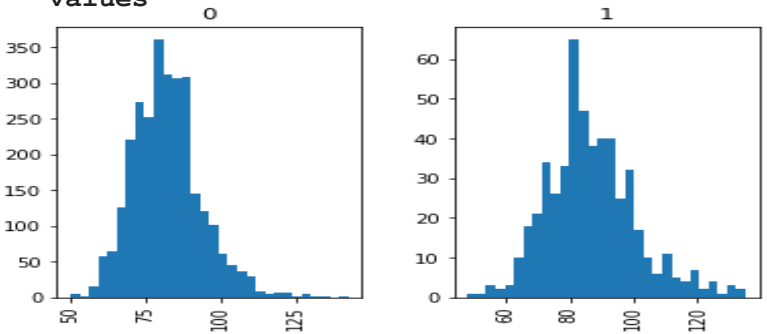
Ten Year CHD Values for different totChol values



Ten Year CHD Values for different sysBP values

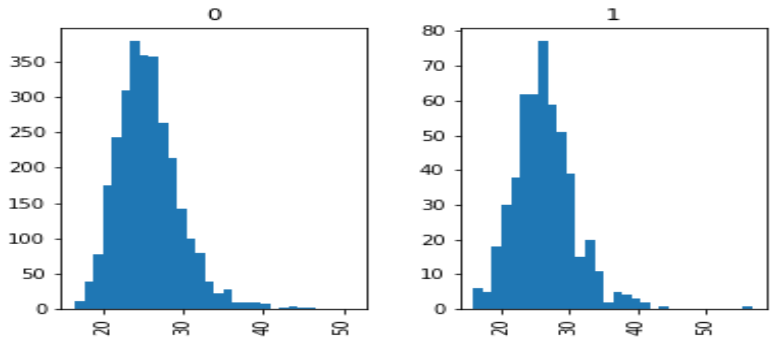


Ten Year CHD Values for different diaBP values

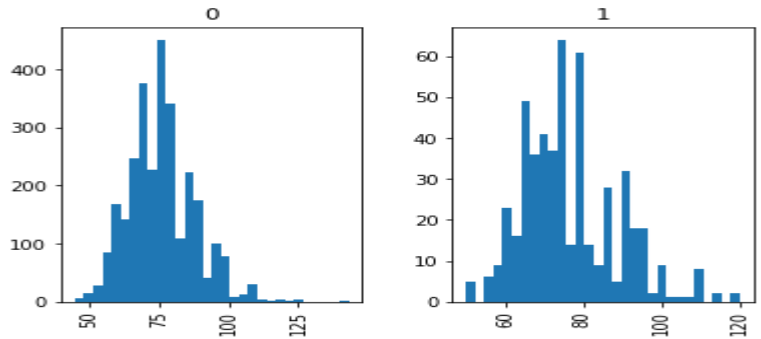


Distribution plots for different independent variables as per target variable labels

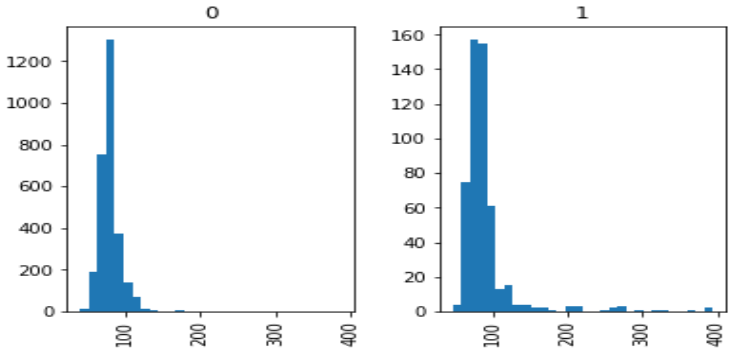
Ten Year CHD Values for different BMI values



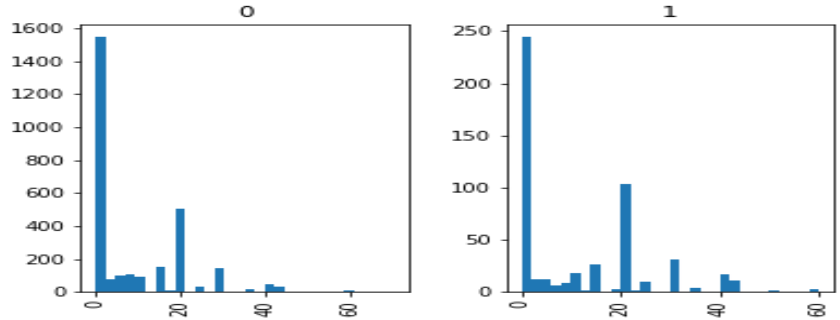
Ten Year CHD Values for different heart Rate values



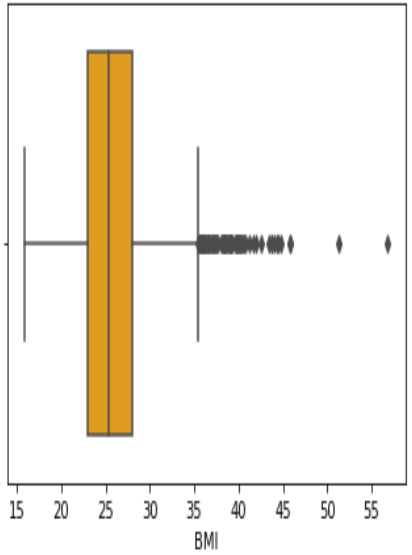
Ten Year CHD Values for different glucose values

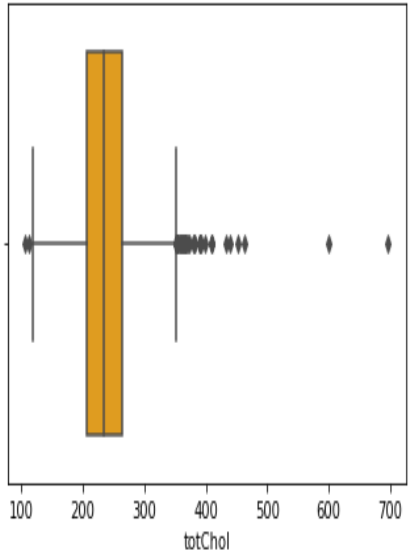


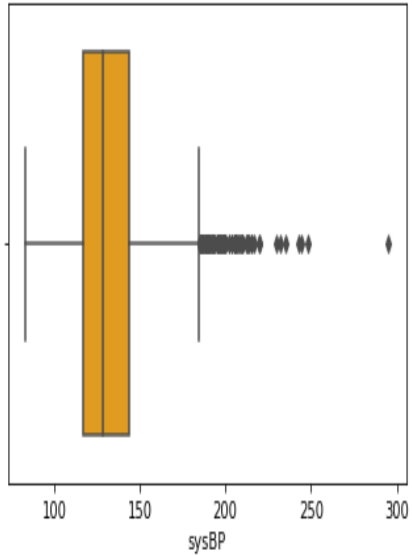
Ten Year CHD Values for different cigsPerDay values

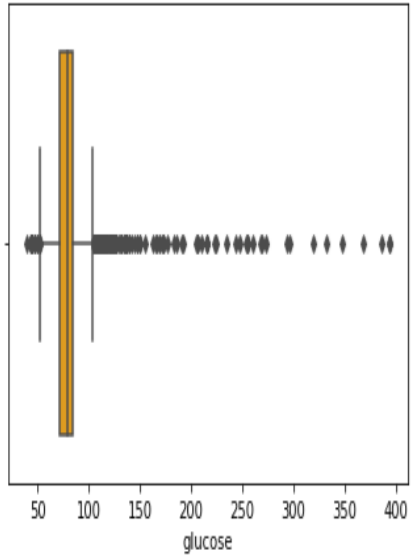


Outlier Detection using Box Plot

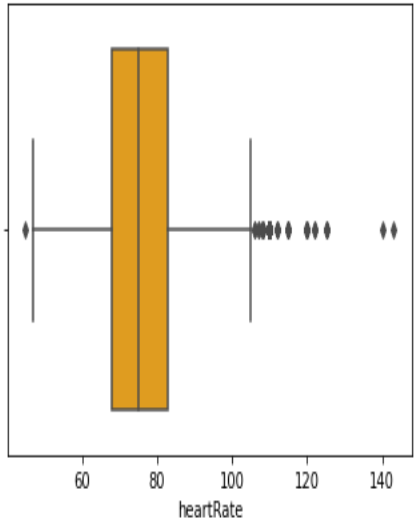


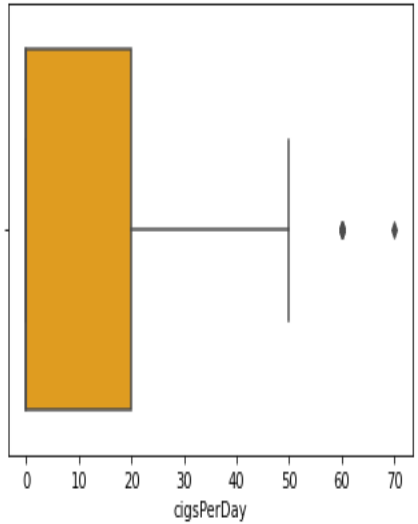


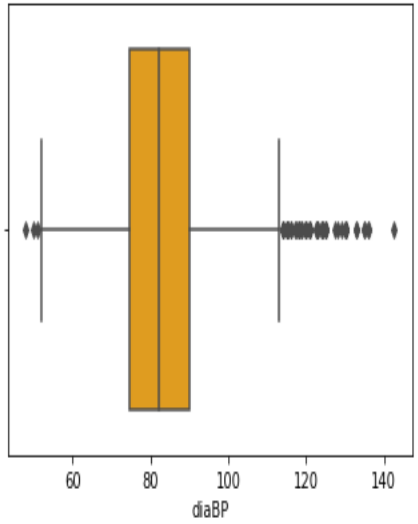


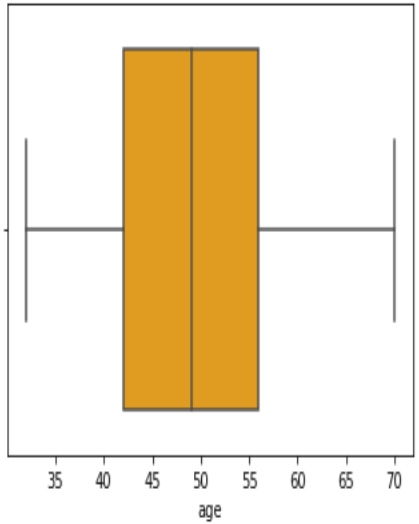


Outlier Detection using Box Plot









Data preparation

Initial Data Shape - 3390 Rows and 17 Columns (1 column is dependent)

After resampling dataset :

Training Data - 4606 Rows and 17 Columns (1 column dependent)

Test Data - 1152 Rows and 17 Columns (1 column dependent)

Dependent column will be predicted as that is the target variable named “Ten Year CHD.

Random Forest Classifier without resampling

Evaluation Metrics for test data

```
Accuracy score-> 0.8362831858407079

Precision score-> 0.1818181818181818

F1 Score-> 0.03478260869565218
```

Precision and F1 score are very less due to unbalanced dataset

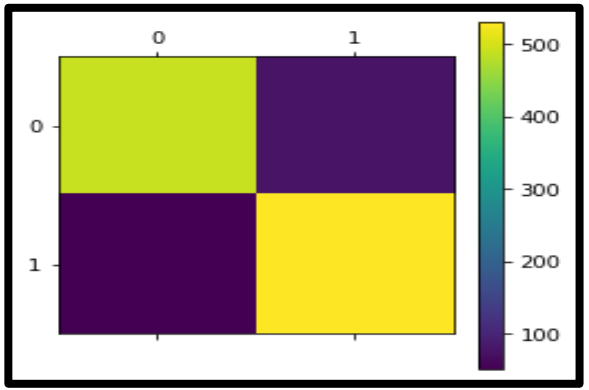
Significant features ranking using boruta selector

	Feature	Ranking
0	age	1
9	totChol	1
10	sysBP	1
11	diaBP	1
14	glucose	1
12	BMI	2
7	prevalentHyp	3
4	cigsPerDay	4
13	heartRate	5
2	sex	6
8	diabetes	6
1	education	8
6	prevalentStroke	8
3	is_smoking	10
5	BPMeds	11

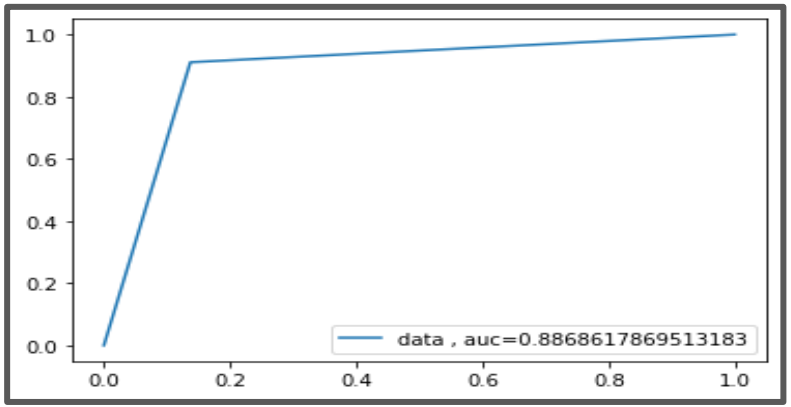
Random Forest Classifier after resampling

Evaluation Metrics for test data

```
Accuracy score-> 0.8871527777777778

Precision score-> 0.8719211822660099

F1 Score-> 0.8909395973154361
```



ROC Curve



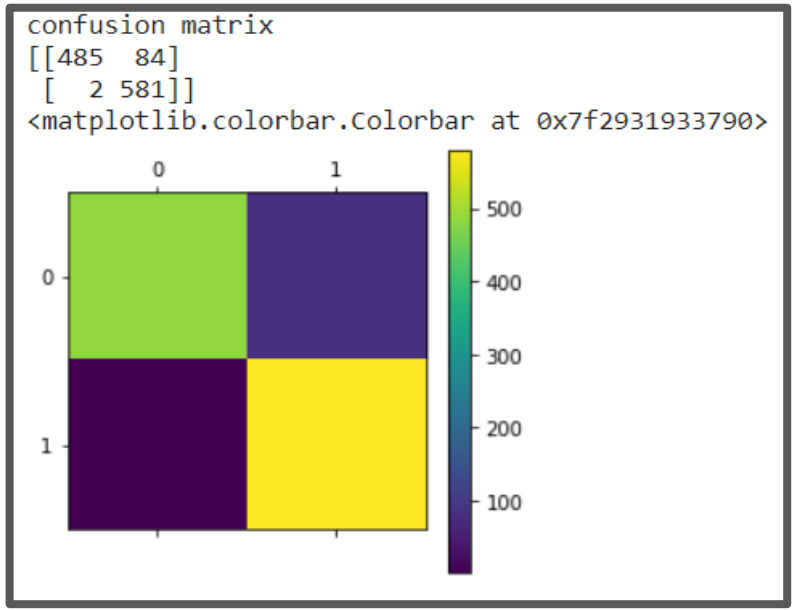

```
confusion matrix
[[491  78]
 [ 52 531]]
```

K Neighbour Classifier after resampling

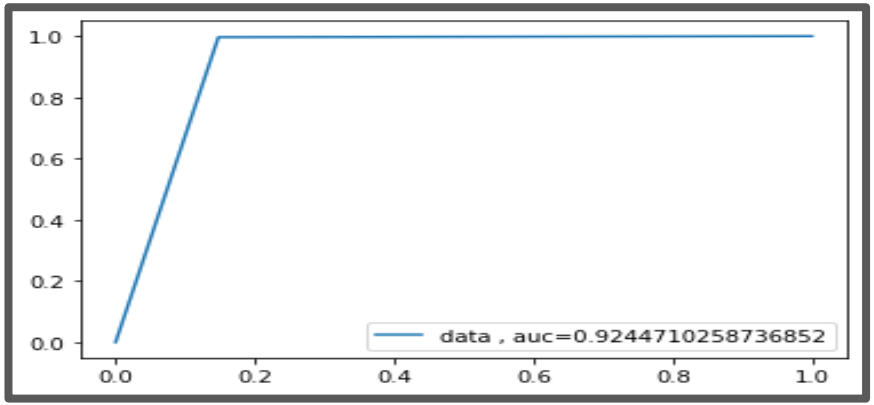
```
Evaluation Metric for test data

Accuracy Score :   0.9253472222222222

Precision Score :  0.8736842105263158

F1 Score :0.9310897435897436
```

Confusion Matrix

```
confusion matrix
[[485  84]
 [  2 581]]
<matplotlib.colorbar.Colorbar at 0x7f2931933790>
```



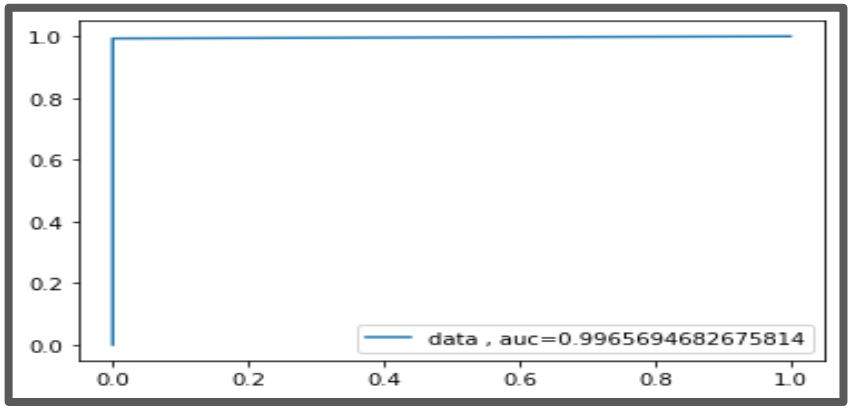
ROC Curve



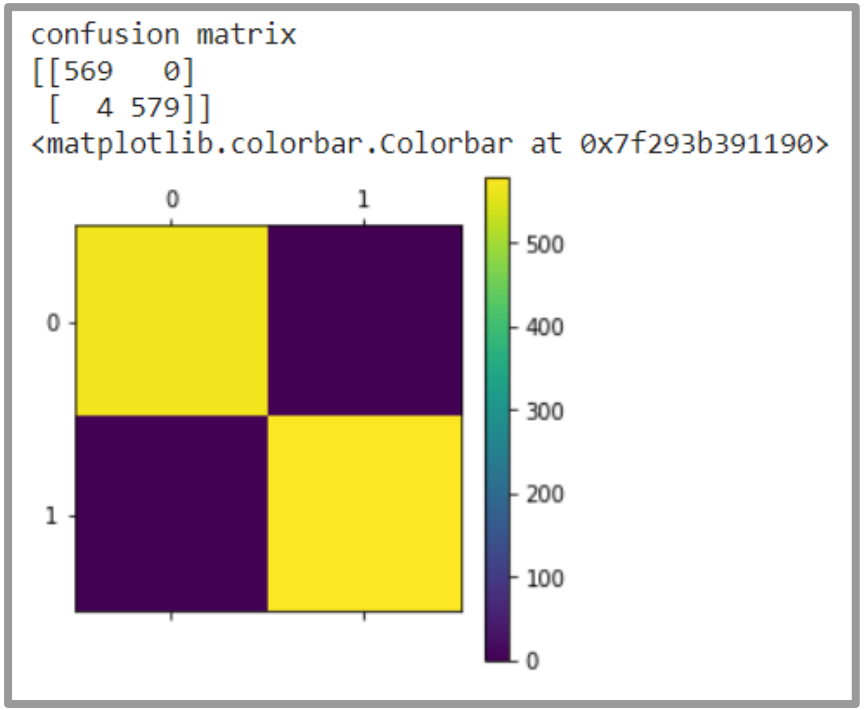
SVM Classifier after resampling

```
Evaluation Metric for test data

Accuracy Score :   0.9965277777777778

Precision Score : 1.0

F1 Score :0.9965576592082617
```

ROC Curve



Confusion Matrix



Classification Report for different Models

Random Forest Classifier

	precision	recall	f1-score	support
0	0.90	0.86	0.88	569
1	0.87	0.91	0.89	583
accuracy			0.89	1152
macro avg	0.89	0.89	0.89	1152
weighted avg	0.89	0.89	0.89	1152

K Neighbour Classifier

Classification Report			precision	recall	f1-score	support
0	1.00	0.85	0.92	569		
1	0.87	1.00	0.93	583		
accuracy			0.93	1152		
macro avg	0.93	0.92	0.92	1152		
weighted avg	0.93	0.93	0.92	1152		

SVM Classifier

Classification Report			precision	recall	f1-score	support
0	0.99	1.00	1.00	569		
1	1.00	0.99	1.00	583		
accuracy			1.00	1152		
macro avg	1.00	1.00	1.00	1152		
weighted avg	1.00	1.00	1.00	1152		

Conclusion

All metrics were evaluated for each model like accuracy score, precision score, f1 score, roc curve and confusion matrix. Resampling of data was performed as the data was not balanced. Imbalance data can give high accuracy but precision and F1 score needs to be taken care of in such cases.

Support Vector Classifier predicting the target variable for testing data more correctly as per all evaluation metrics like roc-auc curve, precision, accuracy, f1 score. Other Models like K neighbour Classifier and Random Forest Classifier are working well too.

Since it is a medical diagnosis cases, we would want the false negative values to be less. In Svm classifier the False negative value comes out to be 4 as per confusion matrix. In K Neighbour Classifier, the False negative value is only 2. For random forest, false negative values comes out to be 52. So preferably, K Neighbour classifier and Support vector classifier seems to be more perfect for classification in this case.

THANK YOU