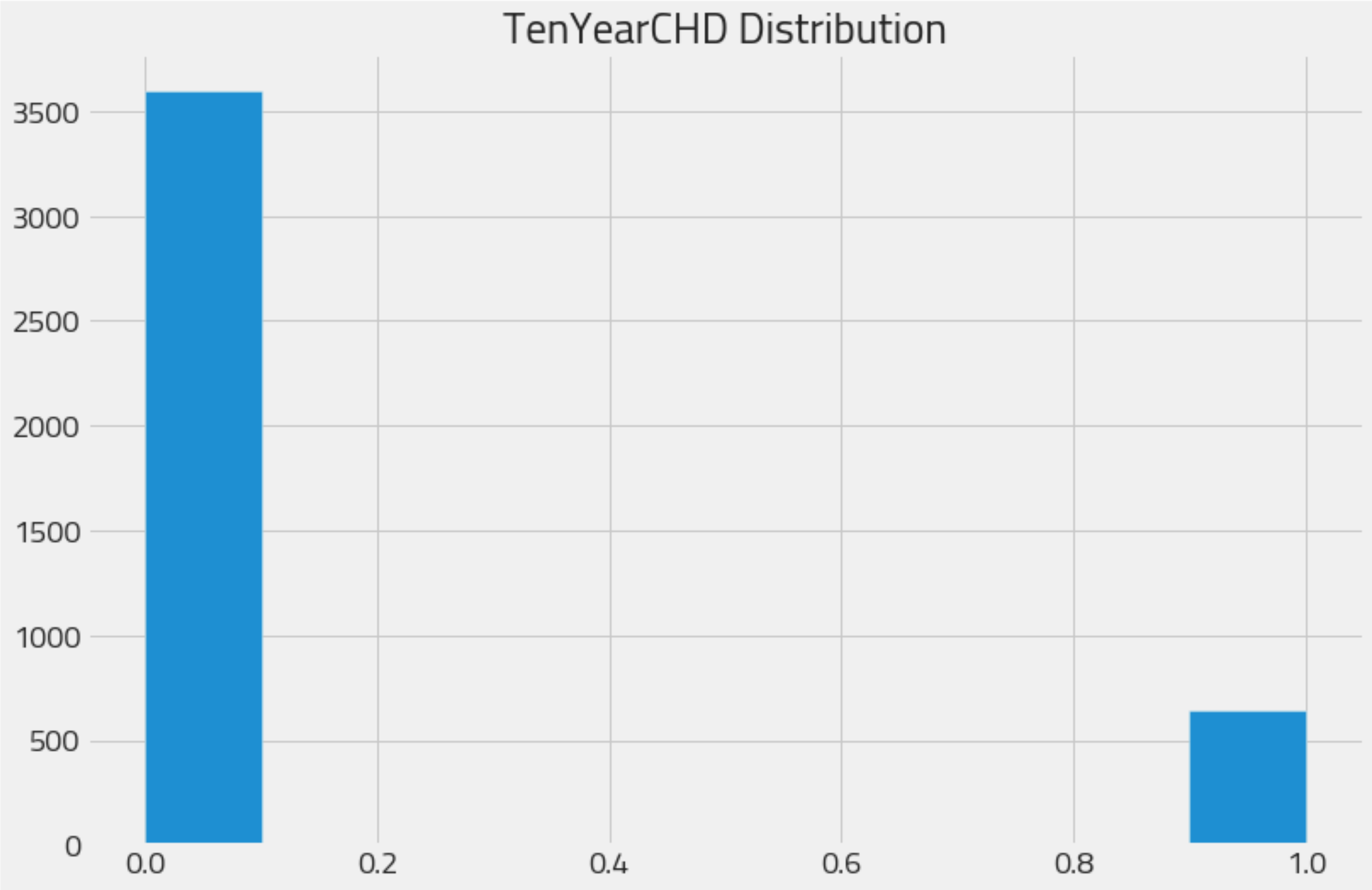
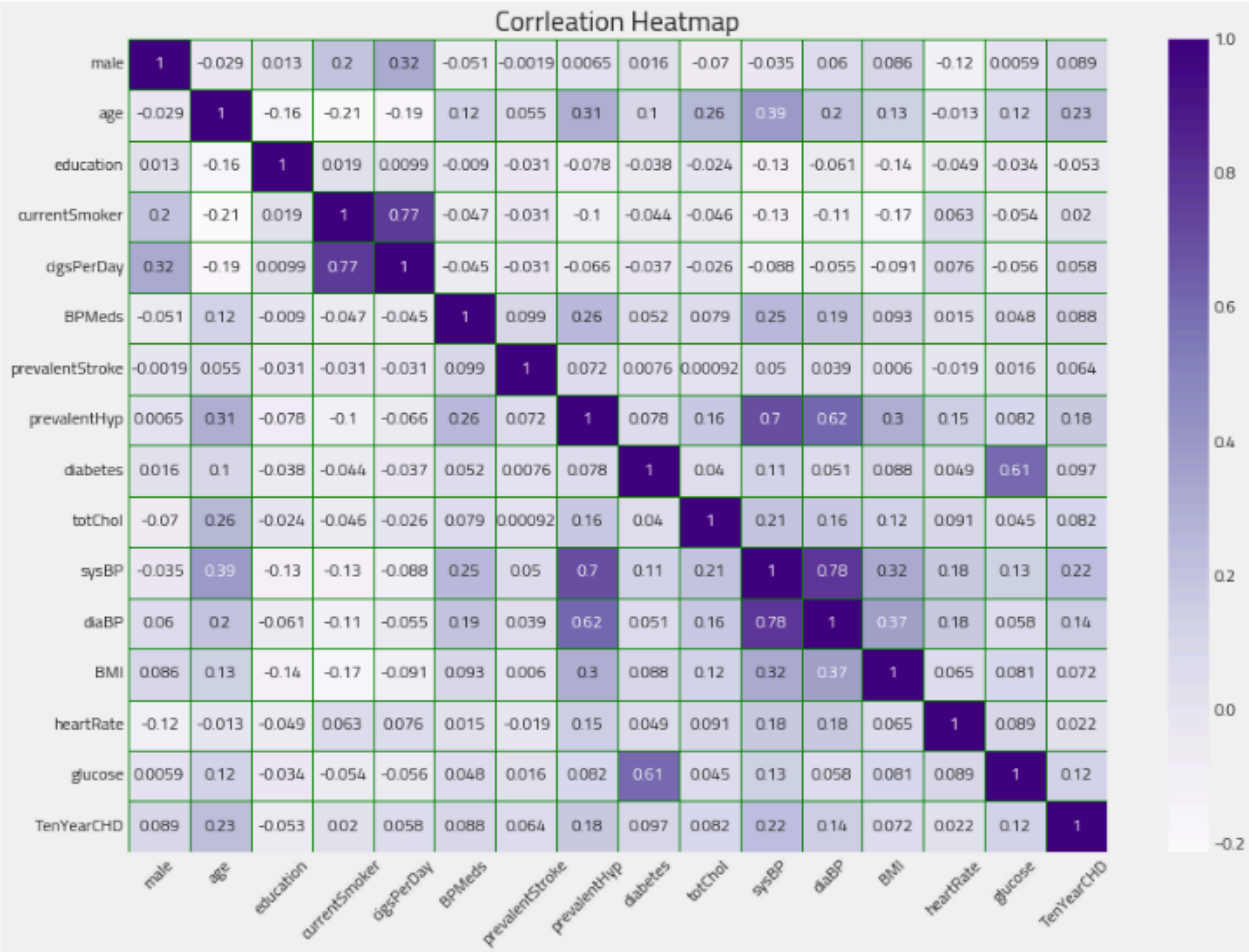
Heart Disease Prediction

Step 1: EDA

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0	22.00	84.0	86.0	0
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0	19.16	86.0	NaN	0
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80.0	107.0	0
4238	1	40	3.0	0	0.0	0.0	0	1	0	185.0	141.0	98.0	25.60	67.0	72.0	0

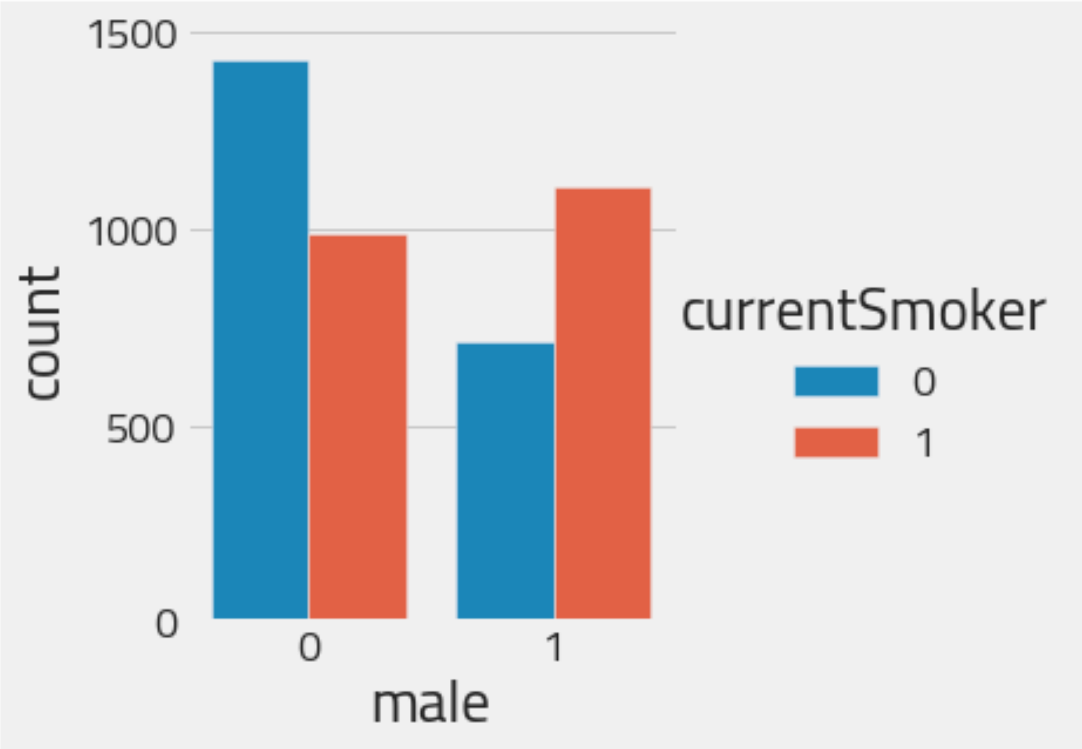
	count	mean	std	min	25%	50%	75%	max
male	4240.0	0.429245	0.495027	0.00	0.00	0.0	1.00	1.0
age	4240.0	49.580189	8.572942	32.00	42.00	49.0	56.00	70.0
education	4135.0	1.979444	1.019791	1.00	1.00	2.0	3.00	4.0
currentSmoker	4240.0	0.494104	0.500024	0.00	0.00	0.0	1.00	1.0
cigsPerDay	4211.0	9.005937	11.922462	0.00	0.00	0.0	20.00	70.0
BPMeds	4187.0	0.029615	0.169544	0.00	0.00	0.0	0.00	1.0
prevalentStroke	4240.0	0.005896	0.076569	0.00	0.00	0.0	0.00	1.0
prevalentHyp	4240.0	0.310613	0.462799	0.00	0.00	0.0	1.00	1.0
diabetes	4240.0	0.025708	0.158280	0.00	0.00	0.0	0.00	1.0
totChol	4190.0	236.699523	44.591284	107.00	206.00	234.0	263.00	696.0
sysBP	4240.0	132.354599	22.033300	83.50	117.00	128.0	144.00	295.0
diaBP	4240.0	82.897759	11.910394	48.00	75.00	82.0	90.00	142.5
BMI	4221.0	25.800801	4.079840	15.54	23.07	25.4	28.04	56.8
heartRate	4239.0	75.878981	12.025348	44.00	68.00	75.0	83.00	143.0
glucose	3852.0	81.963655	23.954335	40.00	71.00	78.0	87.00	394.0
TenYearCHD	4240.0	0.151887	0.358953	0.00	0.00	0.0	0.00	1.0


TenYearCHD Distribution
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0


Corrleation Heatmap
male
age
education
currentSmoker
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
TenYearCHD

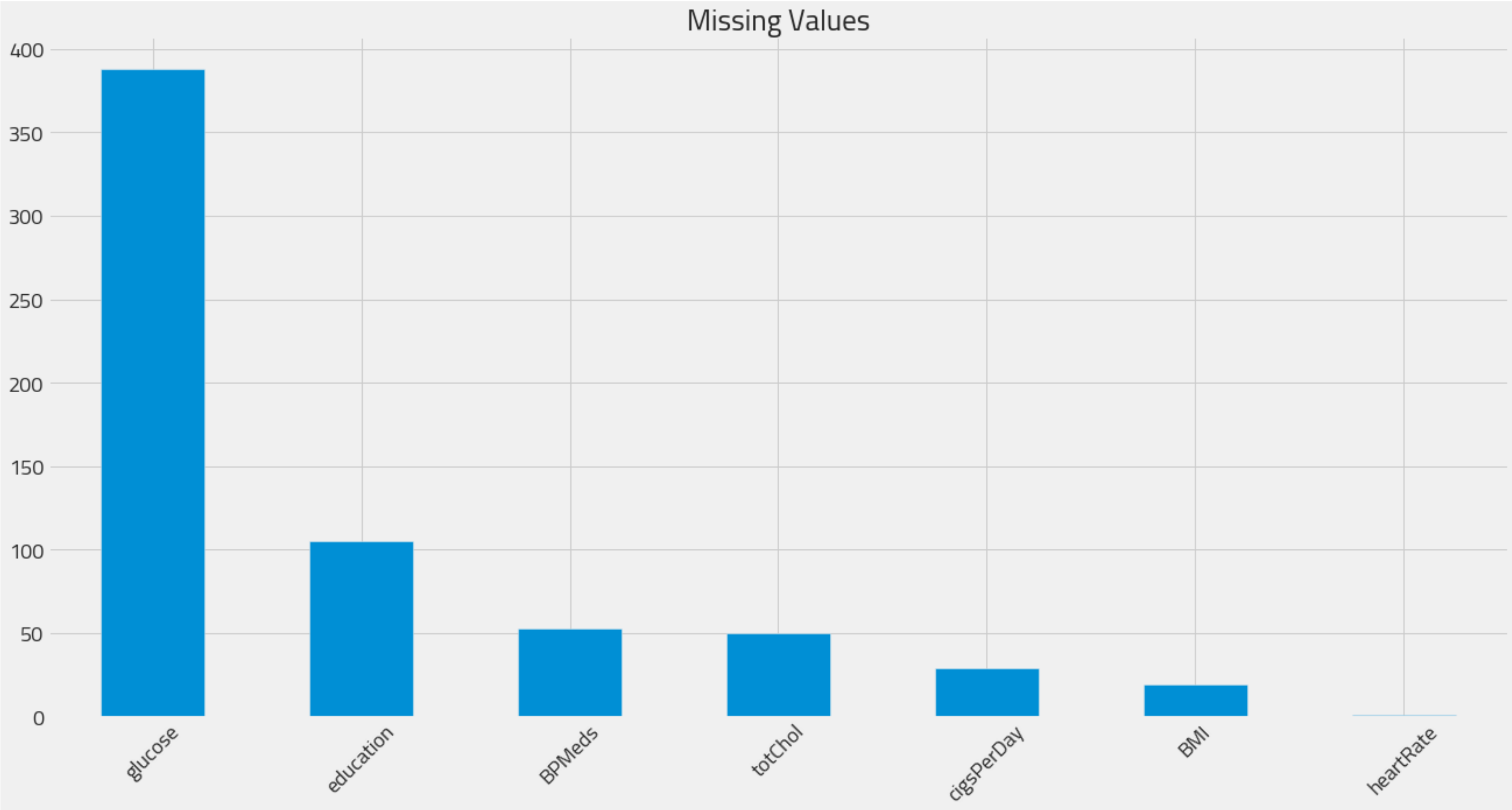
Countplot of people based on their sex and whether they are Current Smoker or not

In [664]:
```
sns.catplot(data=df, kind='count', x='male',hue='currentSmoker')
plt.show()
```



Step 2: Pre-Processing

Histogram of missing values:

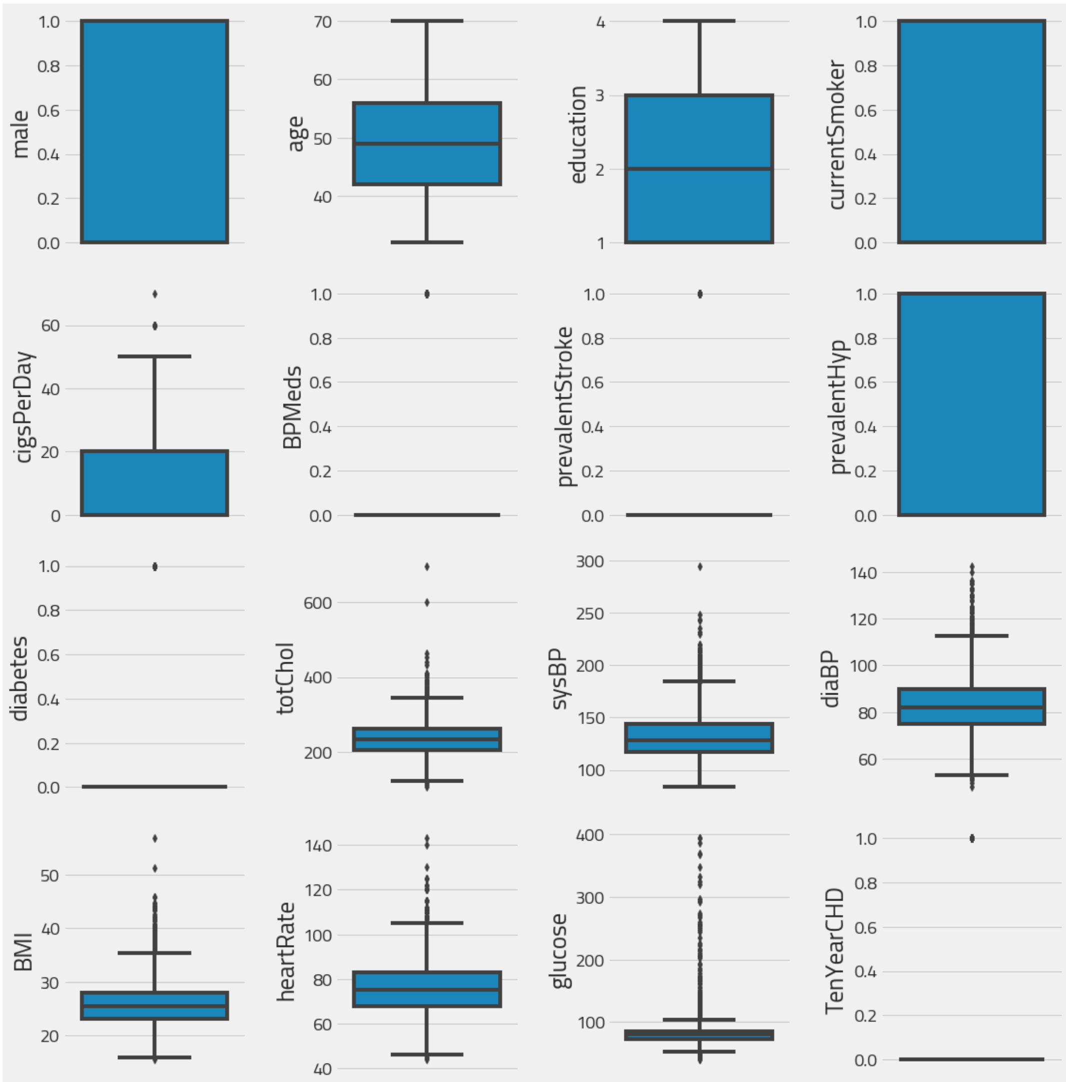


After fixing Nulls:

```
In [663]: print(f"Total Number of Nulls: {df.isna().sum().sum()}")

Total Number of Nulls: 0
```

Boxplot of numerical features distribution



In [282]:
```
df[df['BMI'] >=50]
```

Out[282]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
2657	0	55	1.0	0	0.0	0.0	0	1	0	208.0	190.0	130.0	56.80	90.0	86.0
3927	0	61	1.0	0	0.0	1.0	1	1	0	225.0	194.0	111.0	51.28	80.0	103.0

In [283]:
```
df.drop(df[df['BMI'] >=50].index,inplace = True)
df[df['BMI'] >=50]
```

Out[283]:

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYea

Glucose Level can be up to 400, and that indicates high sugar level in blood

Max of BMI is 50, it cannot be more than 50

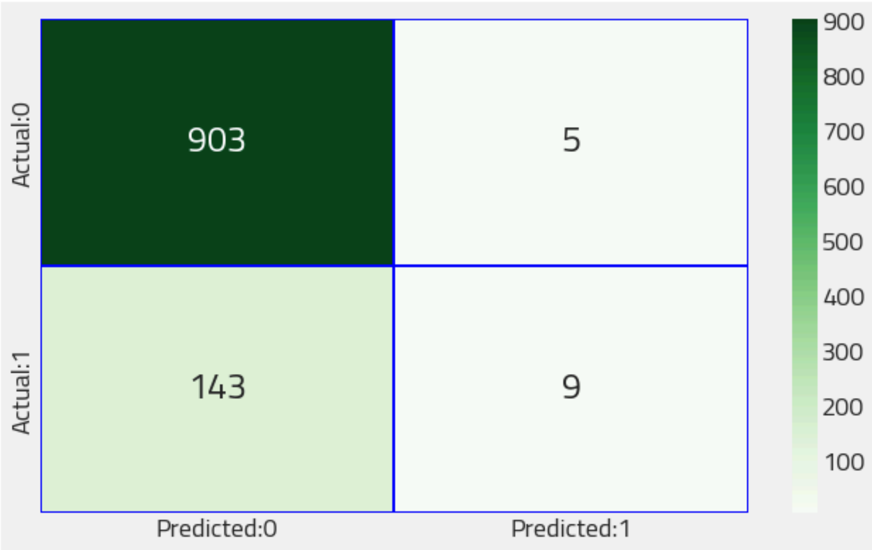
Heart Rate can increase to 160 and sometimes more, so outliers here are normal

cholLevel can increase to 800 and sometimes 1000, normal

diaBP is not considered outlier

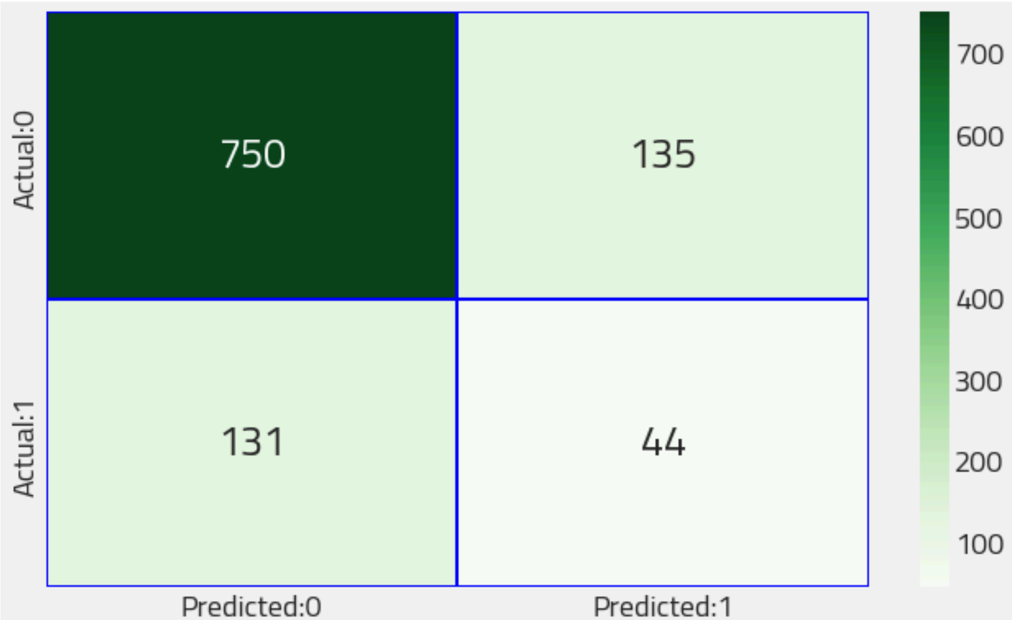
Step 3: Modelling & Evaluation-1

Logistic Regression:



	precision	recall	f1-score
0	0.86	0.99	0.92
1	0.64	0.06	0.11
accuracy			0.86
macro avg	0.75	0.53	0.52
weighted avg	0.83	0.86	0.81

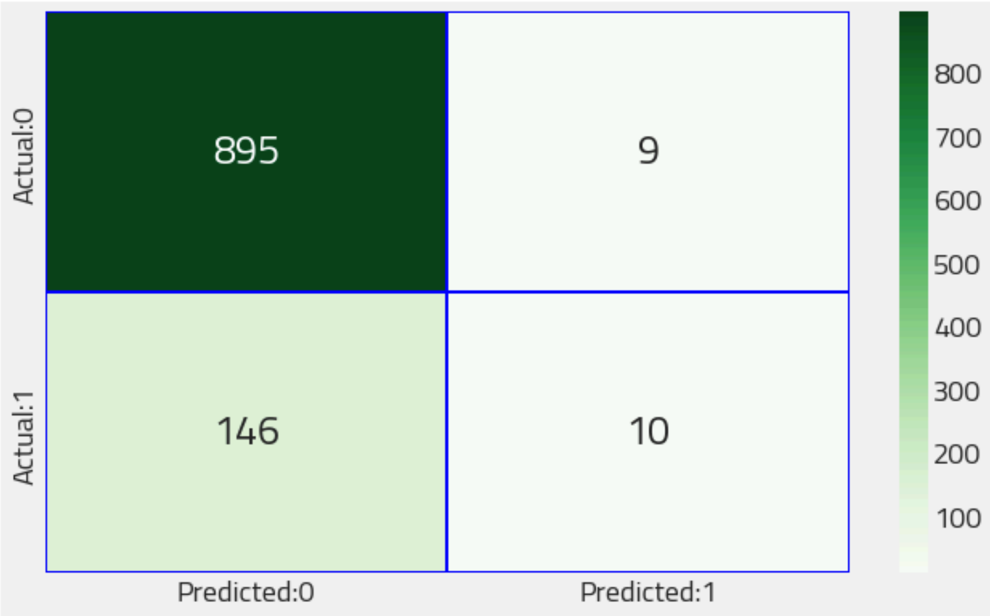
Decision Tree:



Classification Report:

	precision	recall	f1-score
0	0.85	0.85	0.85
1	0.25	0.25	0.25
accuracy			0.75
macro avg	0.55	0.55	0.55
weighted avg	0.75	0.75	0.75

Random Forest:

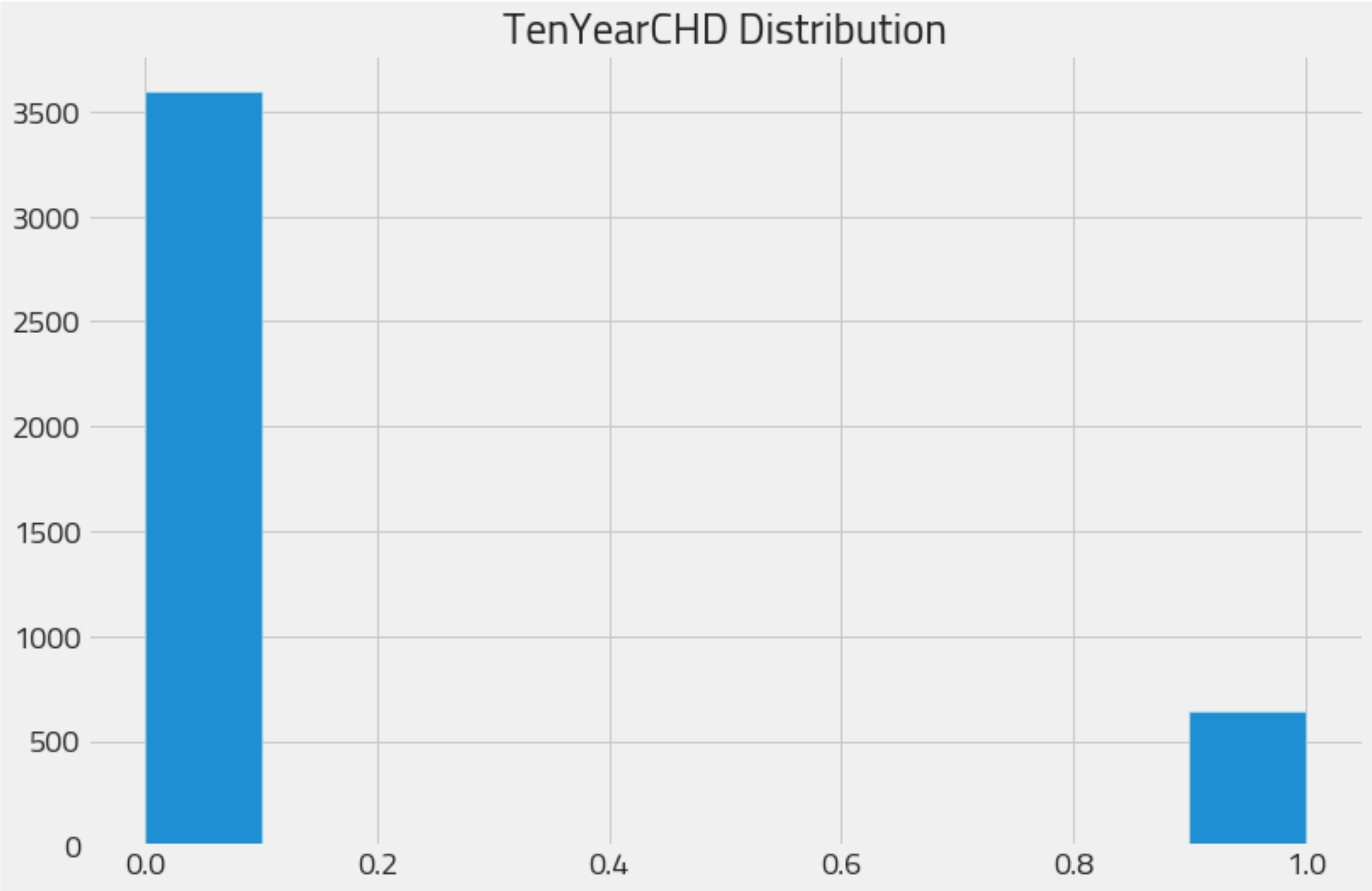


Classification Report:

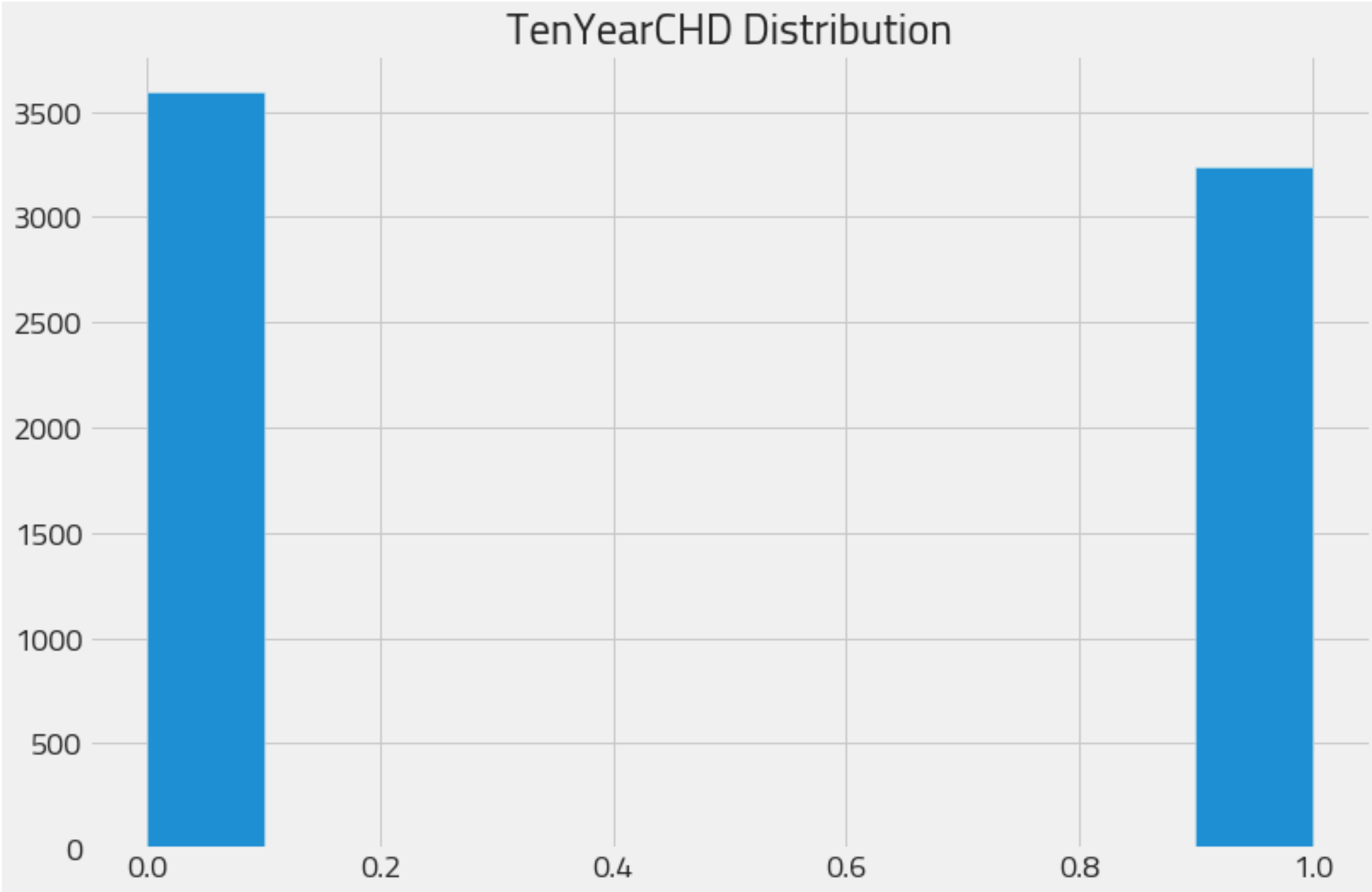
	precision	recall	f1-score
0	0.86	0.99	0.92
1	0.53	0.06	0.11
accuracy			0.85
macro avg	0.69	0.53	0.52
weighted avg	0.81	0.85	0.80

Recall is so low (max of 25%).

After investigation, we discovered that this happened because the data is imbalanced as shown:



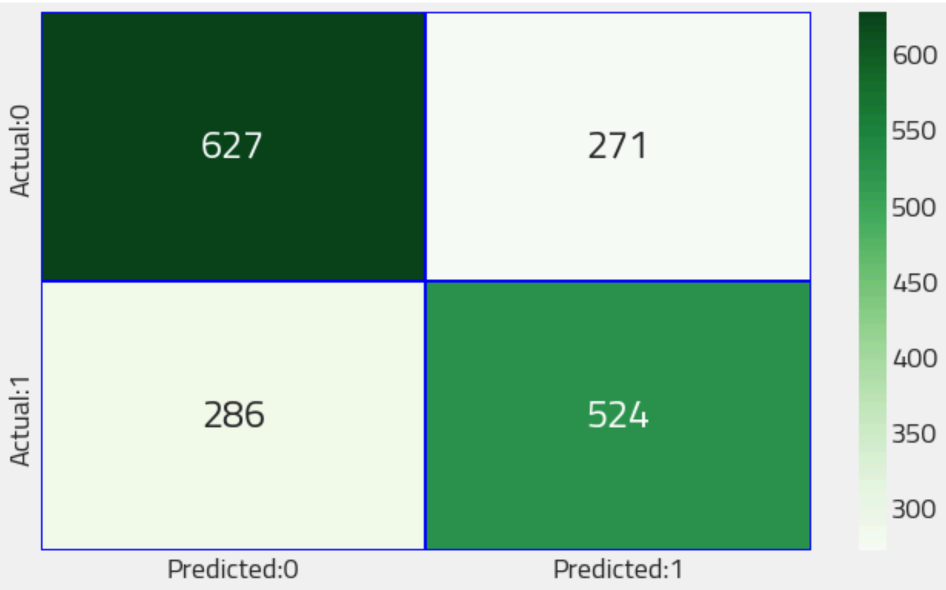
Using Smote Algorithm, we can oversample the data to have more values, and thus more values of class 1


TenYearCHD Distribution
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0

Step 3: Modelling & Evaluation-1

(w/oversampling)

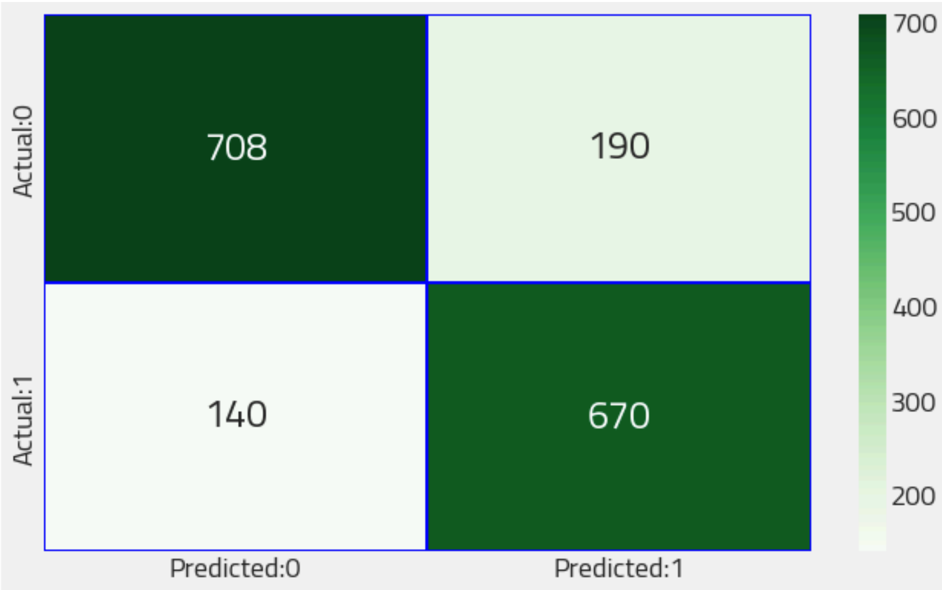
Logistic Regression:



Classification Report:

	precision	recall	f1-score
0	0.69	0.70	0.69
1	0.66	0.65	0.65
accuracy			0.67
macro avg	0.67	0.67	0.67
weighted avg	0.67	0.67	0.67

Decision Tree:



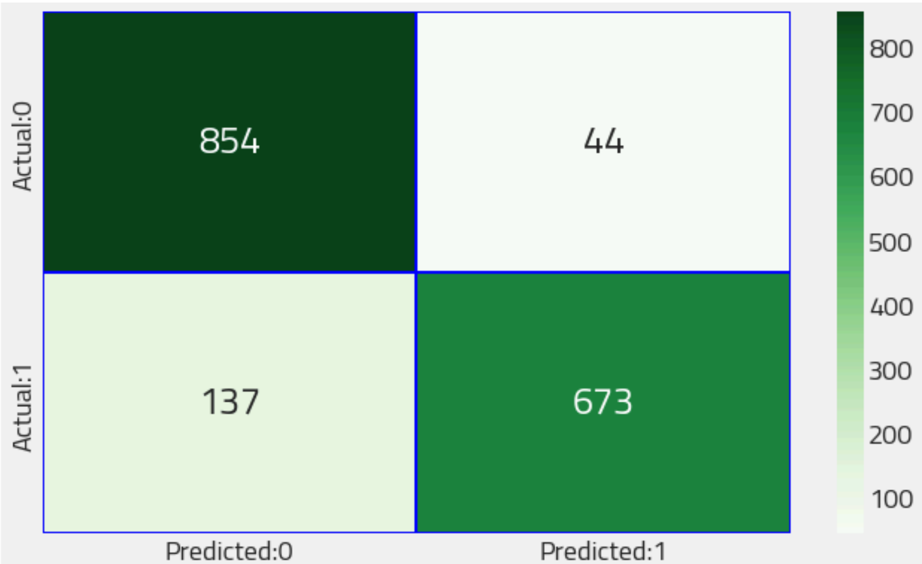

```
Classification Report:

              precision    recall  f1-score

           0       0.83      0.79      0.81
           1       0.78      0.83      0.80

    accuracy                           0.81
   macro avg       0.81      0.81      0.81
weighted avg       0.81      0.81      0.81
```

Random Forest:



Classification Report:

	precision	recall	f1-score
0	0.86	0.95	0.90
1	0.94	0.83	0.88
accuracy			0.89
macro avg	0.90	0.89	0.89
weighted avg	0.90	0.89	0.89

Summary

After Exploring the data and extracting information from, and having good info about the correlation and relationships between features, we started cleaning the data by fixing nulls and removing the outliers. Then tried modelling using 3 Algorithms: Logistic Regression, Decision Tree and Random Forest. And using the confusion matrix we could calculate accuracy, precision, recall and f1 score. There was a problem that the recall is so low. After investigation, we could determine the cause of this problem, which is the imbalance in the values of class label. Using oversampling technique (Smote Algorithm), we could add more data and achieve the balance in the data. After running the model on the enhanced data, we could have achieved accuracy of 89% (using Random Forest), which is a pretty good number for such data.