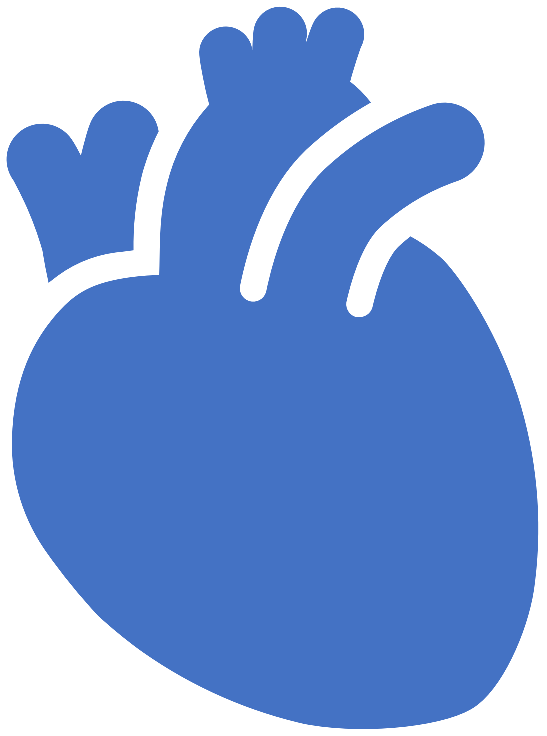
Heart Disease Analysis & Prediction



Problem Statement:

a) Perform Detail Analysis of Heart Disease.

b) Create Appropriate Machine Learning Model for Disease Prediction.

Checking for Missing Values

```
df.isnull().sum()
```

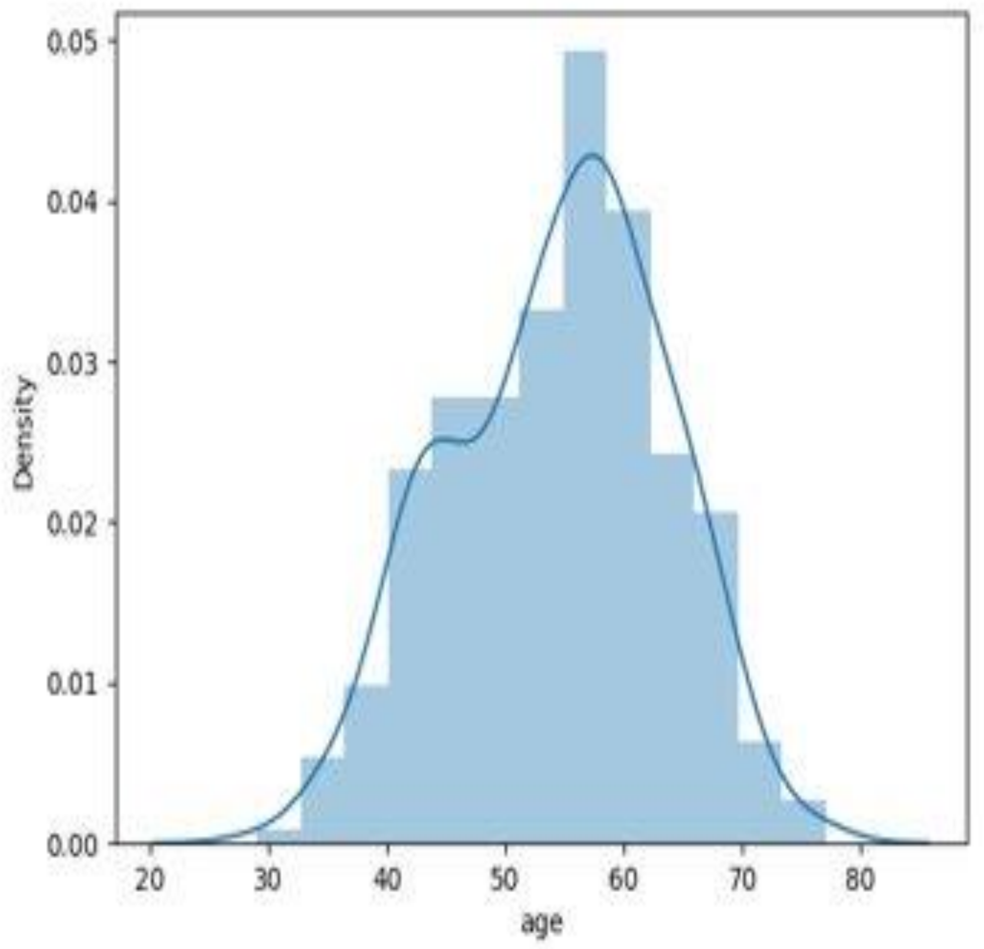
```
age              0
sex              0
ChestPainType    0
BloodPressure    0
cholestoral      0
Bloodsugar       0
ECG              0
Max_heartrate    0
Ex_Pain          0
oldpeak          0
slope            0
No_of_vassels    0
Thalassemia      0
Target           0
dtype: int64
```

DATA

- The dataset consists of 303 individuals data instances.
- Data is distributed across 14 columns.
- There is no any Missing Values in DataFrame.

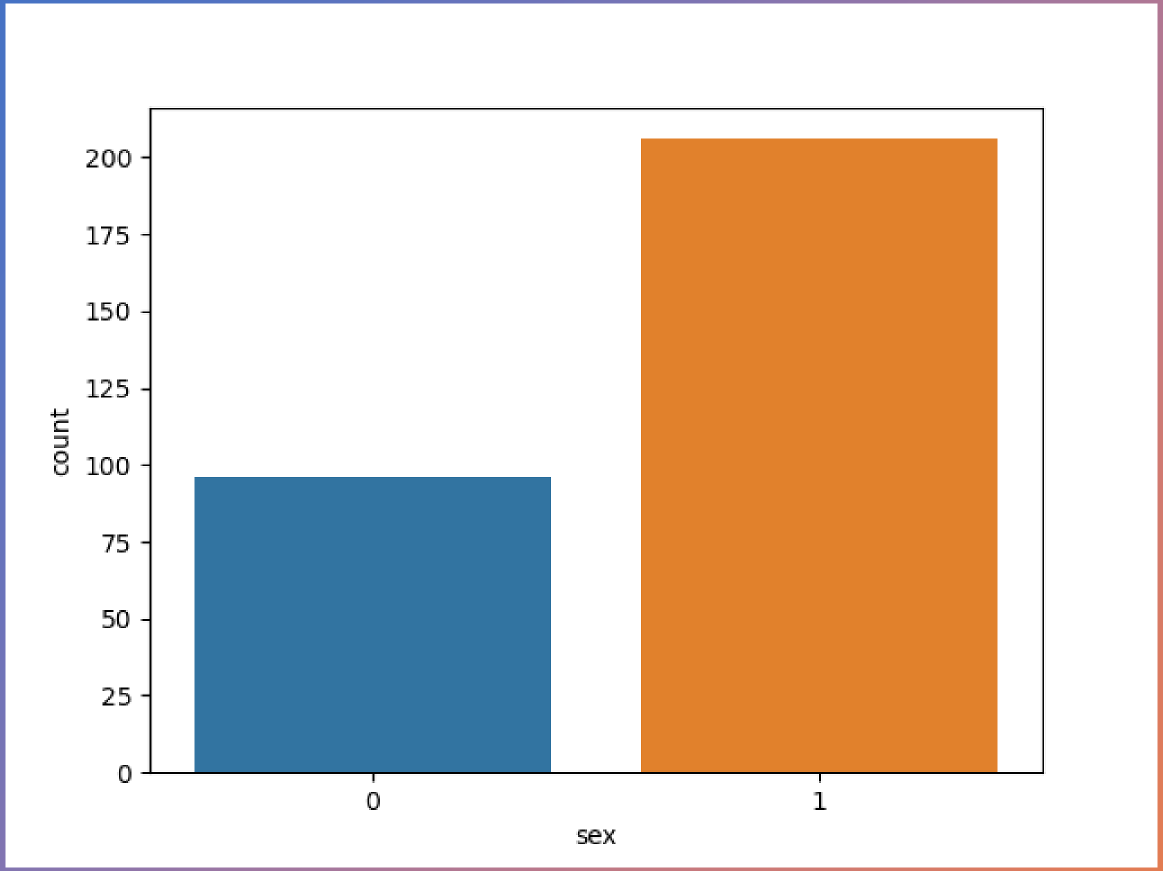
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   age            303 non-null    int64
 1   sex            303 non-null    int64
 2   ChestPainType  303 non-null    int64
 3   BloodPressure  303 non-null    int64
 4   cholestoral    303 non-null    int64
 5   Bloodsugar     303 non-null    int64
 6   ECG            303 non-null    int64
 7   Max_heartrate  303 non-null    int64
 8   Ex_Pain        303 non-null    int64
 9   oldpeak        303 non-null    float64
 10  slope          303 non-null    int64
 11  No_of_vassels  303 non-null    int64
 12  Thalassemia    303 non-null    int64
 13  Target         303 non-null    int64
dtypes: float64(1), int64(13)
```



UNIVARIATE ANALYSIS

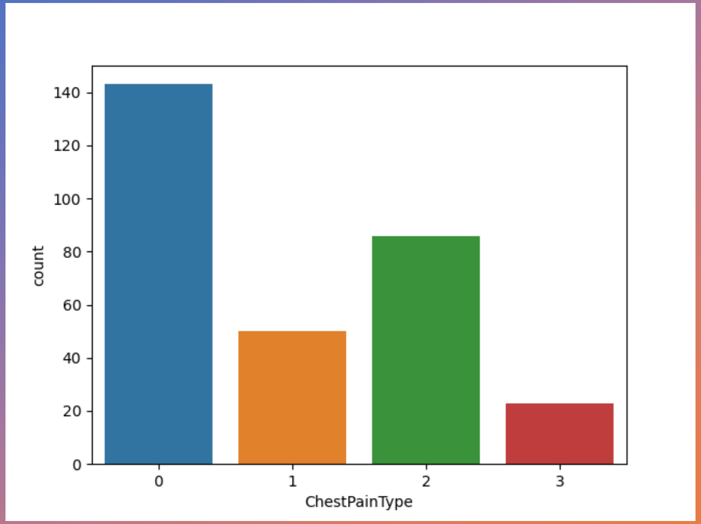
- **AGE**
- data is normally distributed
- does not have outliers
- skewness coefficient is around -0.2, no need of transformation




```
[23] df.sex.value_counts(normalize = True)

1    0.682119
0    0.317881
Name: sex, dtype: float64
```

UNIVARIATE ANALYSIS

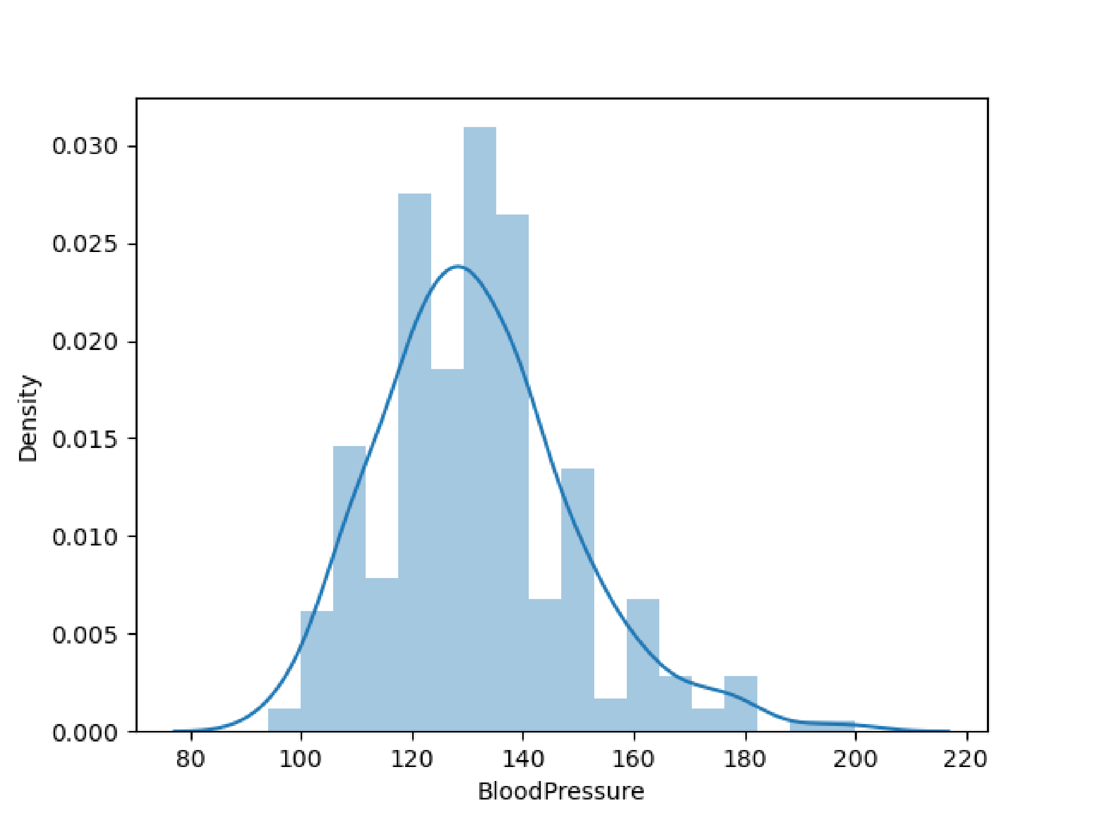
- **SEX**
- data contains details of 206 male and 96 female




```
[26] df['ChestPainType'].value_counts(normalize = True)

0    0.473510
2    0.284768
1    0.165563
3    0.076159
Name: ChestPainType, dtype: float64
```

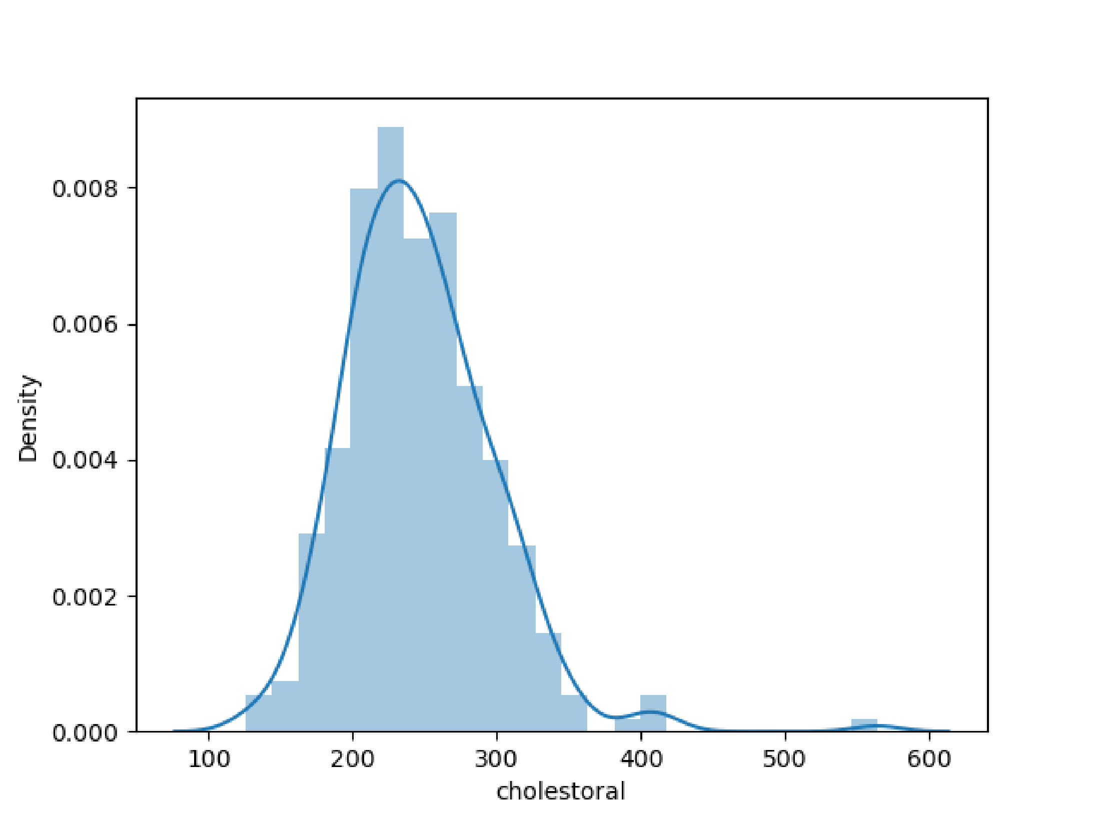
UNIVARIATE ANALYSIS

- **ChestPainType**
- 143 patients have **typical angina.**
- 86 patients have **atypical angina.**
- 50 patients have **non-anginal pain.**
- 23 patients have **asymptomatic** pain



UNIVARIATE ANALYSIS

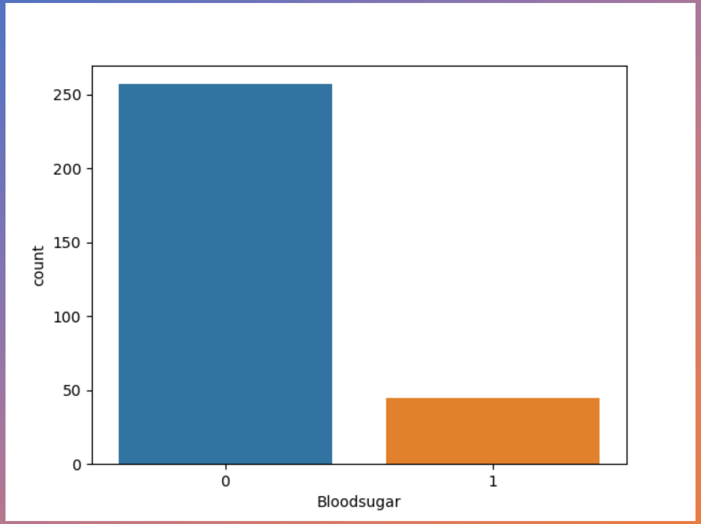
- **Blood Pressure**
- data is normally distributed
- has outliers
- skewness coefficient is around 0.7, no need of transformation



UNIVARIATE ANALYSIS

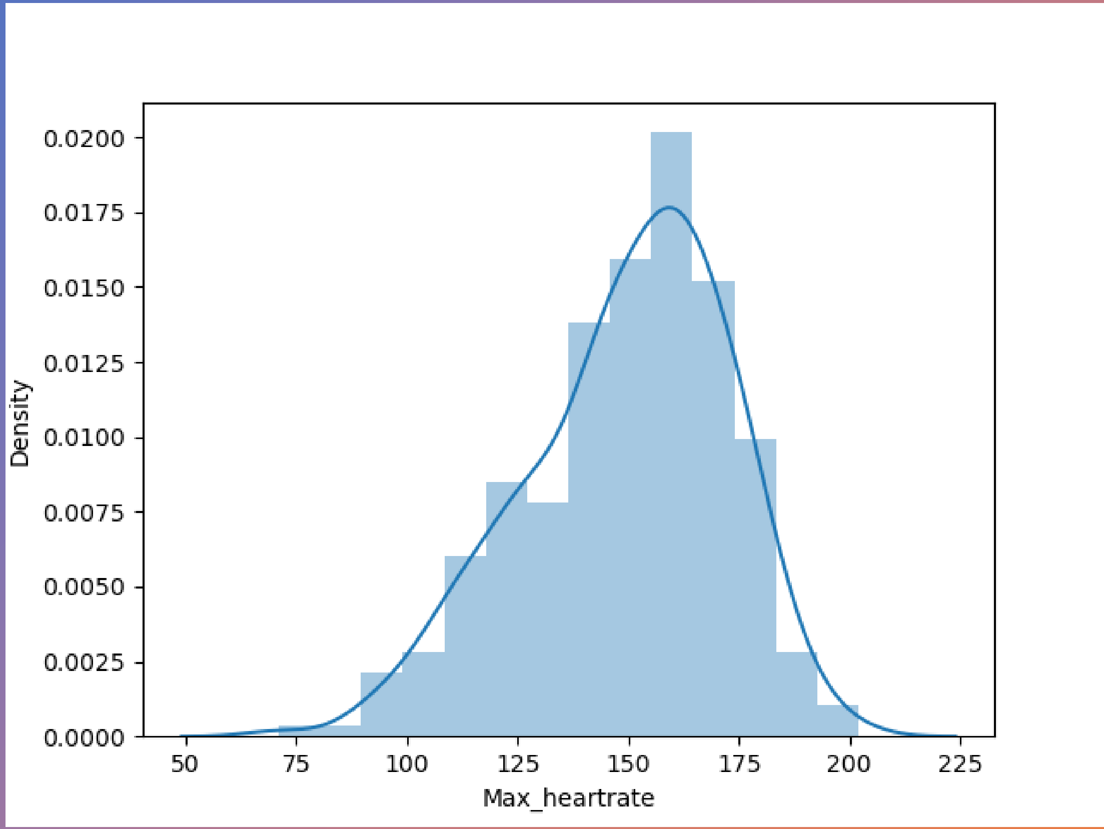
- **Cholestoral**
- **data is normally distributed and skewed rightwards**
- **has outliers**
- **skewness coefficient is around 1.15, needs transformation**

UNIVARIATE ANALYSIS



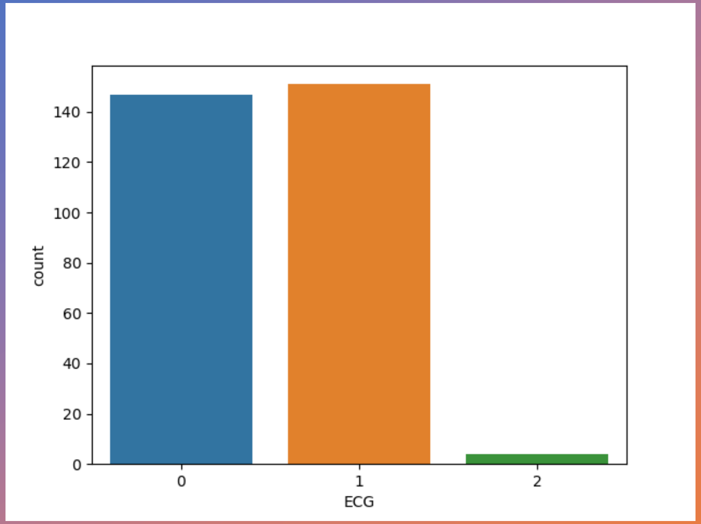

```
[34] df.Bloodsugar.value_counts()

0    257
1     45
Name: Bloodsugar, dtype: int64
```

- **Bloodsugar**
- **257 patients have Bloodsugar**
- **45   patients do not have Bloodsugar**



UNIVARIATE ANALYSIS

- **Max_heartrate**
- data is normally distributed and skewed leftwards
- has outliers
- skewness coefficient is around - 0.5, do not need transformation

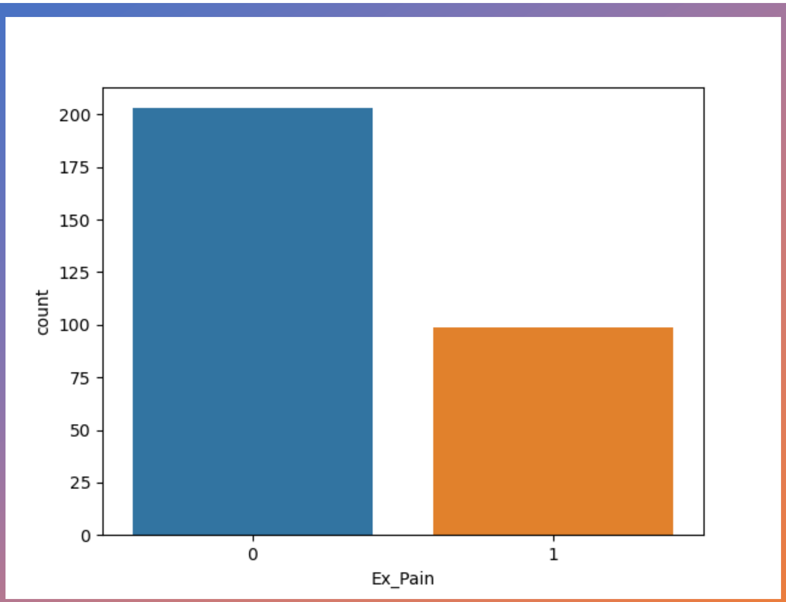



```
[36] df.ECG.value_counts()

1    151
0    147
2      4
Name: ECG, dtype: int64
```

UNIVARIATE ANALYSIS

- **ECG**

- **147 patients have Normal ECG**
- **151 patients have ST-T wave abnormality in ECG**
- **4 patients have probable or definite left ventricular hypertrophy**

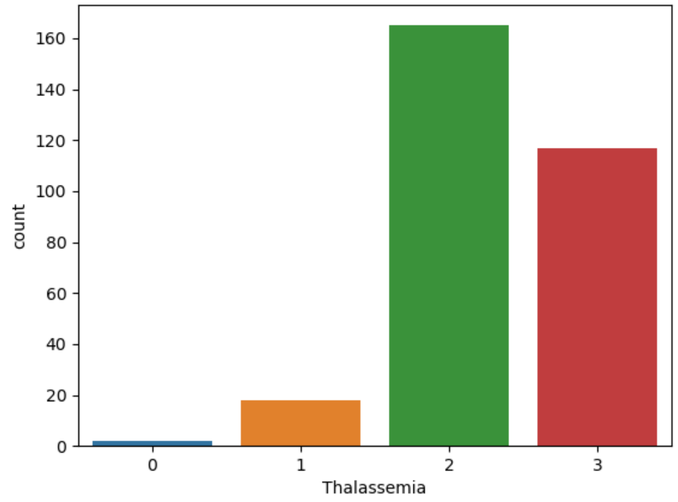



```
[41] df['Ex_Pain'].value_counts()

0    203
1     99
Name: Ex_Pain, dtype: int64
```

UNIVARIATE ANALYSIS

- **Ex - Pain**
- 99 patients have Exercise induced Angina
- 203 patients do not have Exercise induced Angina

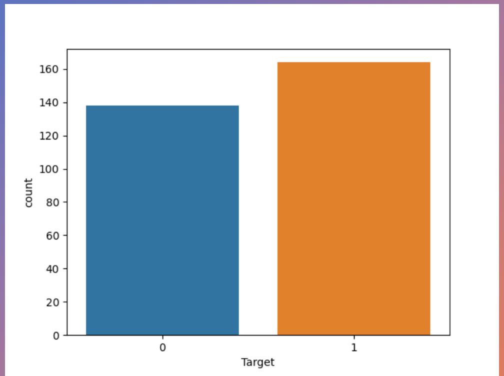
```
[45] df['Thalassemia'].value_counts()

2    165
3    117
1     18
0      2
Name: Thalassemia, dtype: int64
```



UNIVARIATE ANALYSIS

- **Thalassemia**
- **165 patients have normal blood flow**
- **117 patients have reversible defect (a blood flow is observed but it is not normal)**
- **18 patients have fixed defect (no blood flow in some part of the heart)**

UNIVARIATE ANALYSIS



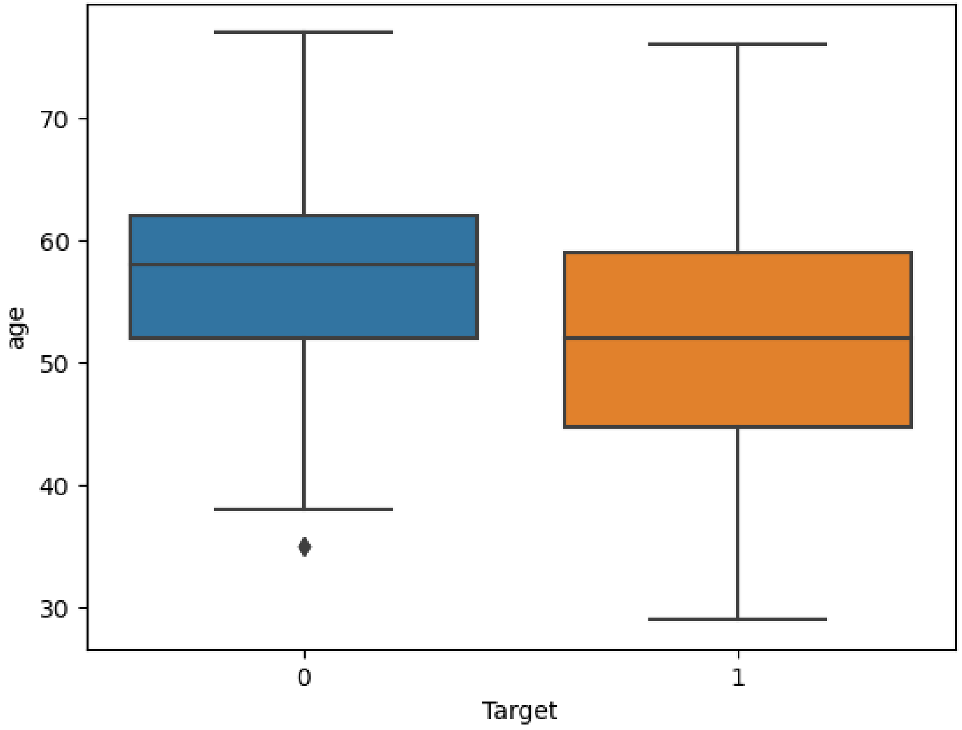

```
[47] df.Target.value_counts()

1    164
0    138
Name: Target, dtype: int64
```

- **Target**
- **138 patients are suffering from heart disease**
- **164 patients do not have heart disease**

BIVARIATE ANALYSIS

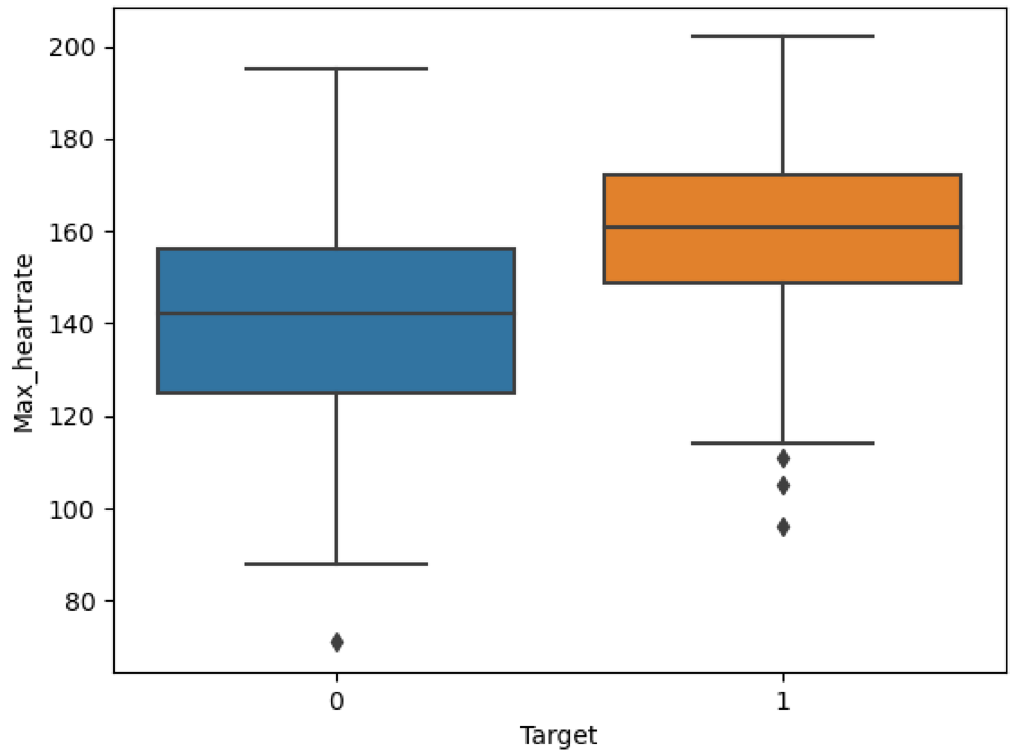
- **Target Vs Age**

- - There exists a significant relationship among them
- - ANOVA
- - statistic=10675.801467178899, p-value=0.0
- * Very high value suggests significant relationship and reject the Null Hypothesis
- - Point Biserial correlation
- - SignificanceResult(statistic=-0.221, p-value=0.00)
- - Boxplot has shown the variation in mean accross categories



BIVARIATE ANALYSIS

- **Target Vs Max_heartrate**
- There exists a significant relationship among them
- ANOVA
- - statistic=12779.77, p-value=0.0
- * Very high value suggests significant relationship and reject the Null Hypothesis
- Point Biserial correlation
- SignificanceResult (statistic= 0.41, p-value= 0.0)
- Boxplot has shown the variation in mean across categories



BIVARIATE ANALYSIS

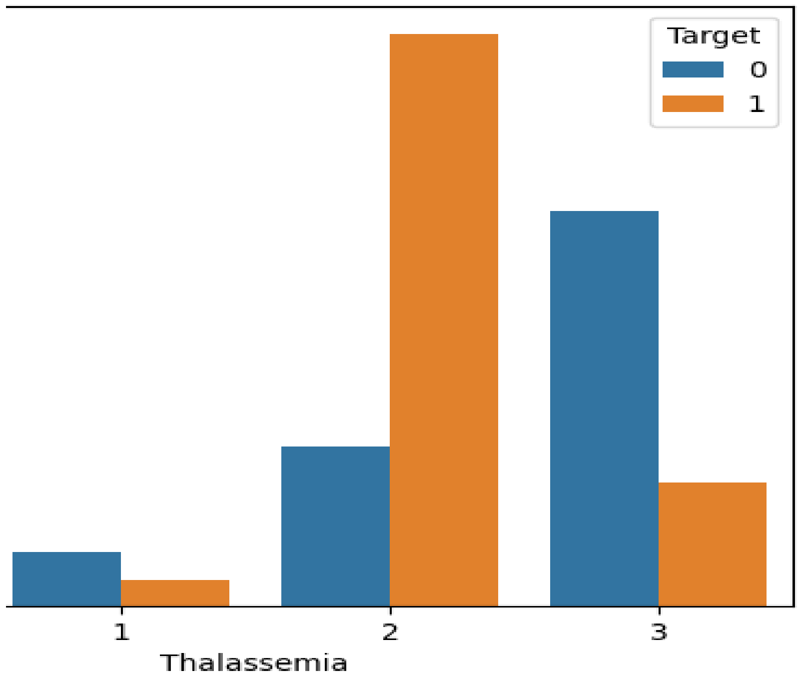
```
[61] chi2, p
```

```
(84.61031794685029, 3.1462951
```

```
contingency_table
```

Thalassemia	0	1	2	3
Target				
0	1	12	36	89
1	1	6	129	28

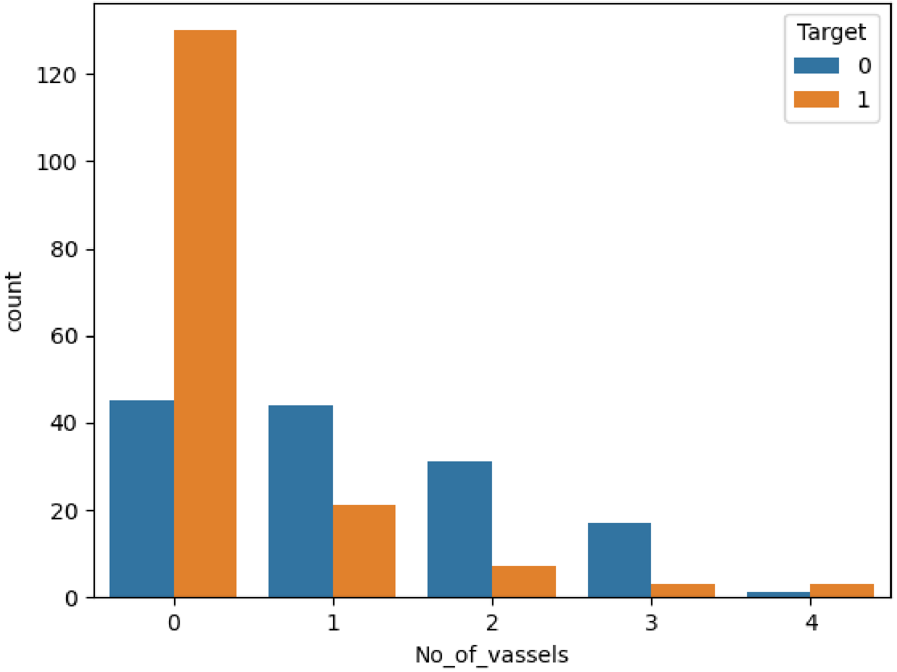
- **Target Vs Thalassemia**
- There exists a significant relationship among them
- - chi2 = 84.6
- - pvalue = 0.0
- * High chi2 value with pvalue less than significance level indicated association among them



BIVARIATE ANALYSIS

• Target Vs No_of_vassels

- There exists a significant relationship among them
- - chi2 = 73.6
- - p-value = 0.0
- * High chi2 value with p-value less than significance level indicated association among them




```
chi2, p
```

```
(73.68984583164412, 3.771038067427657e-15)
```

```
[65] contingency_table
```

No_of_vassels	0	1	2	3	4
Target					
0	45	44	31	17	1
1	130	21	7	3	3

BIVARIATE ANALYSIS

• Target Vs slope

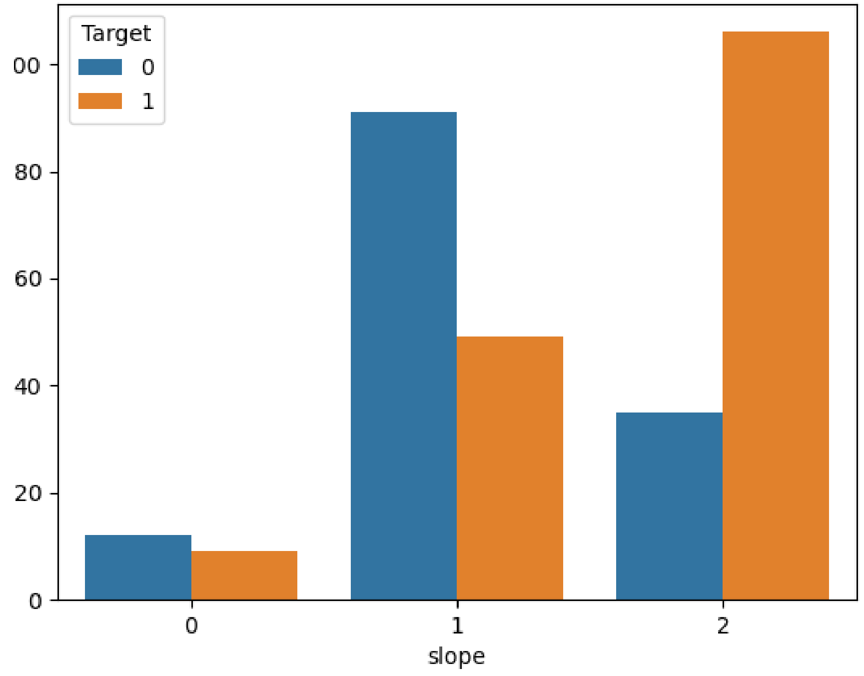
- There exists a significant relationship among them
-     - chi2 = 46.8
-     - p-value = 0.0
- * High chi2 value with p-value less than significance level indicated association among them

```
5  chi2, p
```

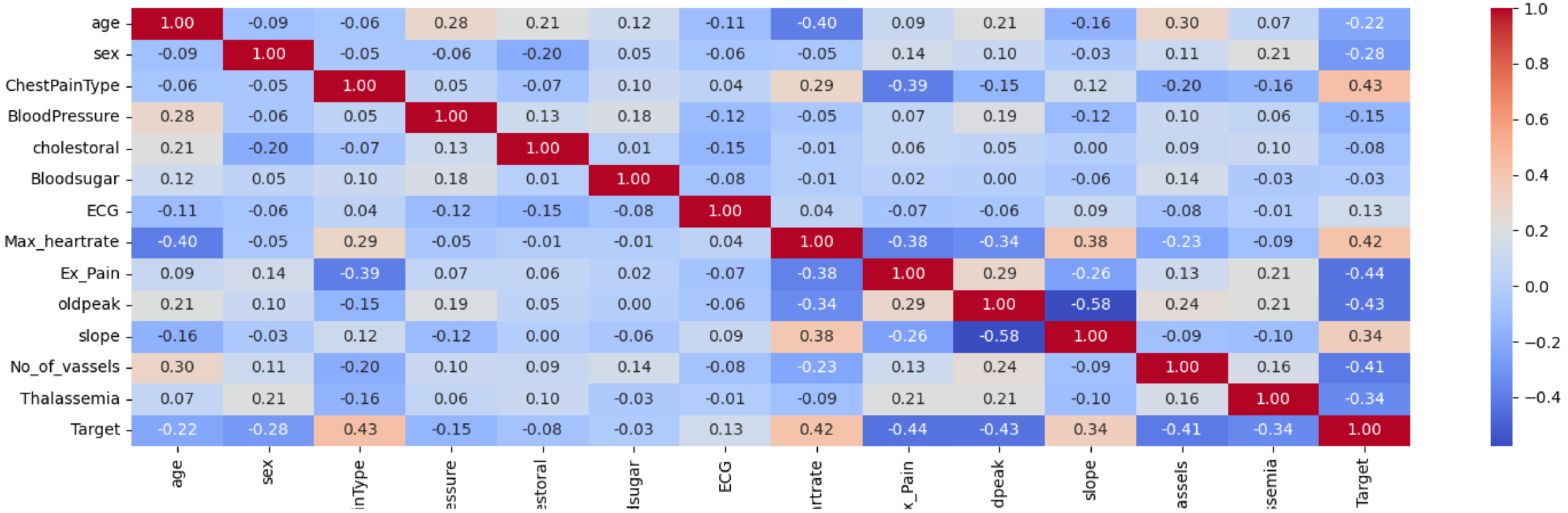
(46.88947660161814, 6.5777827609179e-11)

```
1  contigency_table
```

slope	0	1	2
Target			
0	12	91	35
1	9	49	106



Correlation Matrix



HYPOTHESIS

- **Performing Hypothesis Testing to analyze the relationship among Dependent and Independent Variables.**
- **H(O):There is no any relationship among Dependent and Independent Variables.**
- **H(A):There exists a strong relationship among Dependent and Independent Variables.**
- **Taking Level of Significance (ALPHA) = 0.05**
- **I will reject the Null Hypothesis for those variables having p-value less than alpha, signifying association existing among them.**

OLS Regression Results

Dep. Variable:	Target	R-squared:	0.519
Model:	OLS	Adj. R-squared:	0.497
Method:	Least Squares	F-statistic:	23.88
Date:	Sun, 14 Jan 2024	Prob (F-statistic):	1.48e-38
Time:	23:34:14	Log-Likelihood:	-107.62
No. Observations:	302	AIC:	243.2
Df Residuals:	288	BIC:	295.2
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8156	0.293	2.785	0.006	0.239	1.392
age	-0.0004	0.003	-0.145	0.885	-0.006	0.005
sex	-0.1965	0.047	-4.172	0.000	-0.289	-0.104
ChestPainType	0.1108	0.022	4.941	0.000	0.067	0.155
BloodPressure	-0.0021	0.001	-1.664	0.097	-0.005	0.000
cholestoral	-0.0003	0.000	-0.773	0.440	-0.001	0.001
Bloodsugar	0.0218	0.060	0.365	0.715	-0.096	0.139
ECG	0.0478	0.040	1.197	0.232	-0.031	0.126
Max_heartrate	0.0030	0.001	2.662	0.008	0.001	0.005
Ex_Pain	-0.1444	0.051	-2.814	0.005	-0.245	-0.043
oldpeak	-0.0572	0.023	-2.494	0.013	-0.102	-0.012
slope	0.0790	0.042	1.866	0.063	-0.004	0.162
No_of_vassels	-0.1075	0.023	-4.771	0.000	-0.152	-0.063
Thalassemia	-0.1175	0.036	-3.296	0.001	-0.188	-0.047

Variables which are having statistically significant relationships are as follows:

1)sex
2)ChestPainType
3)Max_heartrate
4)Ex_Pain
5)oldpeak
6)No_of_vassels
7)Thalassemia

Machine Learning Model: Logistic Regression

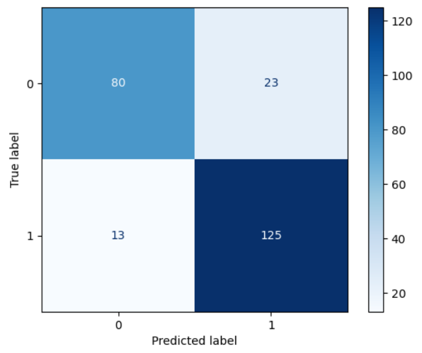
- **In this problem, the dependent variable ('Target') has binary values.**
- **Performing Binary Classification Using Logistic Regression.**

Model Training

- **Training Accuracy: 85%**
- **Confusion Matrix**
- **Classification Report:**
  - **Recall for Positive Class: 91%**
  - **Precision for Positive Class: 84%**

```
In [86]: accuracy_score(y_train,y_pred)
Out[86]: 0.8506224066390041
```

```
In [87]: print(confusion_matrix(y_train,y_pred))
[[ 80  23]
 [ 13 125]]
```



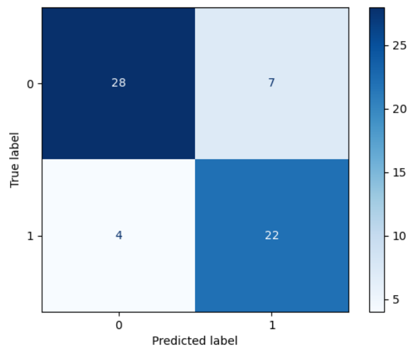

```
In [89]: print(classification_report(y_train,y_pred))
              precision    recall  f1-score   support

           0       0.86      0.78      0.82       103
           1       0.84      0.91      0.87       138

    accuracy                           0.85       241
   macro avg       0.85      0.84      0.85       241
weighted avg       0.85      0.85      0.85       241
```

Model Validation

- **Training Accuracy: 81.9 %**
- **Confusion Matrix**
- **Classification Report:**
  - **Recall for Positive Class: 85 %**
  - **Precision for Positive Class: 76 %**

```
In [91]: accuracy_score(y_test,y_pred)
Out[91]: 0.819672131147541
```

```
In [92]: print(confusion_matrix(y_test,y_pred))

[[28  7]
 [ 4 22]]
```




```
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.88      0.80      0.84        35
           1       0.76      0.85      0.80        26

    accuracy                           0.82        61
   macro avg       0.82      0.82      0.82        61
weighted avg       0.83      0.82      0.82        61
```

Sensitivity and Specificity of the Model

- **Sensitivity : 87.5 %**
- **Specificity : 75.8 %**

Checking for Sensitivity and Specificity

```
confusion_matrix = cm
total=sum(sum(confusion_matrix))

sensitivity = confusion_matrix[0,0]/(confusion_matrix[0,0]+confusion_matrix[1,0])
print('Sensitivity : ', sensitivity )

specificity = confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[0,1])
print('Specificity : ', specificity)
```

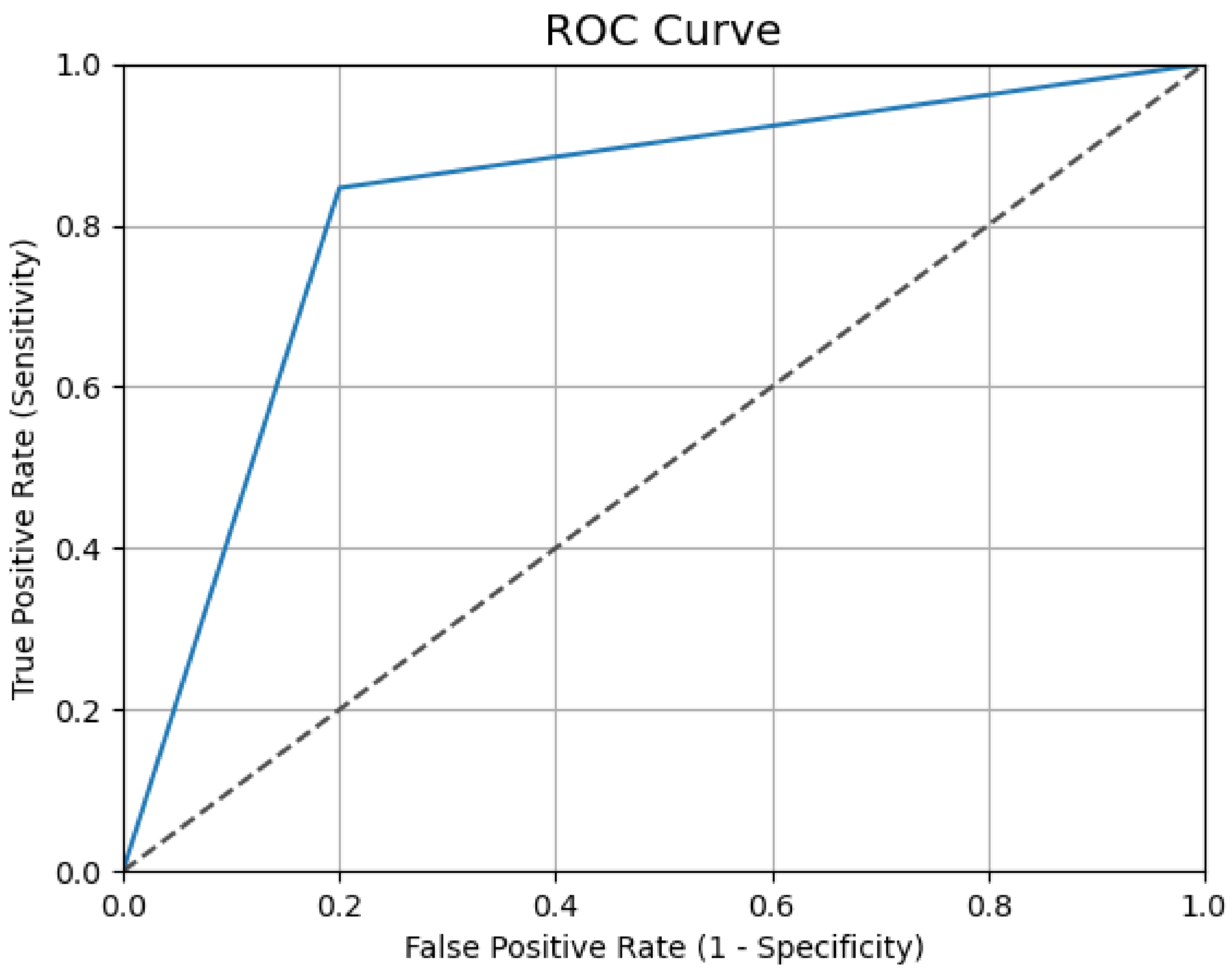
```
Sensitivity :  0.875
Specificity :  0.7586206896551724
```

ROC-Curve

- ROC is plot of TPR Vs FPR
- True Positive Rate (Sensitivity)

  Vs

- False Positive Rate (1 - Specificity)
- Area Under the Curve measures the performance of the Model



Conclusion:

- Heart is the vital organ of Human being, and Heart Problems are very frequent and one of the major concerns for society today.
- It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data.