HEART DISEASES

Data analytics project

Name	ID	sec
Shrouk mohsen ibrahim	20201700401	Sec5
Habiba mohamed adel	20201701191	Sec3
Kholoud mohamed elfateh	20201700239	Sec3
Eman ahmed abdelfatah	20201701061	Sec2
Ibrahim saeed saleh	20201701197	Sec1

The Heart Disease Project detects whether or not that person has heart disease , based on some attributes.

we have the dataset contains 1009 rows and 12 Column.

The dataset contains information about individuals and their heart health, with

the following attributes:

Age: Age of the individual in years (numerical).

Sex: Gender of the individual (categorical: 'M' for male, 'F' for female).

ChestPainType: The type of chest pain the individual is experiencing (categorical: '**ATA**' for atypical angina, '**NAP**' for non-anginal pain, '**ASY**' for asymptomatic, '**TA**' for typical angina).

RestingBP: Resting blood pressure of the individual (numerical).

Cholesterol: Serum cholesterol level of the individual in mg/dl (numerical).

FastingBS: Fasting blood sugar of the individual, where 1 means fasting blood sugar > 120 mg/dl and 0 otherwise (binary).

RestingECG: Resting electrocardiographic results of the individual (categorical: 'Normal', 'ST', 'T').

MaxHR: Maximum heart rate achieved by the individual (numerical).

ExerciseAngina: Whether the individual has angina induced by exercise, where 'Y' means yes and 'N' means no (binary).

Oldpeak: ST depression induced by exercise relative to rest (numerical).

ST_Slope: The slope of the peak exercise ST segment, where 'Up' is upsloping, 'Flat' is flat, and 'Down' is downsloping (categorical).

HeartDisease: The presence of heart disease, where 1 has heart disease and 0 not have heart disease (binary).

Steps:

- Preprocessing

-we read dataset from csv file and display info of this dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1008 entries, 0 to 1007
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             1008 non-null   int64
 1   Sex             1007 non-null   object
 2   ChestPainType   1007 non-null   object
 3   RestingBP       1000 non-null   float64
 4   Cholesterol     1002 non-null   float64
 5   FastingBS       1007 non-null   float64
 6   RestingECG      1002 non-null   object
 7   MaxHR           1003 non-null   float64
 8   ExerciseAngina  1001 non-null   object
 9   Oldpeak         1008 non-null   float64
 10  ST_Slope        1008 non-null   object
 11  HeartDisease    1008 non-null   int64
dtypes: float64(5), int64(2), object(5)
```

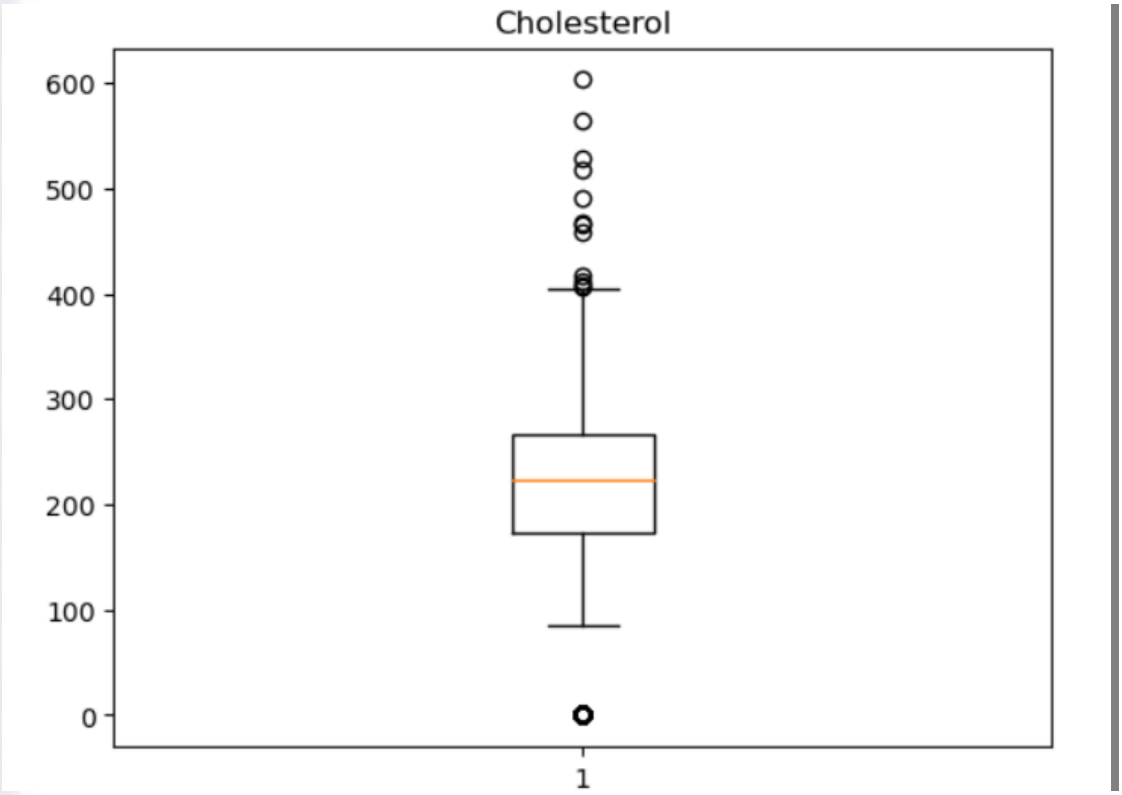
- Found 8 rows with null values and They were dropped.

--checked about duplicate rows and find 81 duplicated and They were dropped.

- detect outliers in some attributes:

(Cholesterol , RestingBP, MaxHR)

Detect outliers in ‘cholesterol’ attribute by Boxplot:



Q1 (25%) = 174.0 Q3 (75%) = 267.0 IQR = 93.0

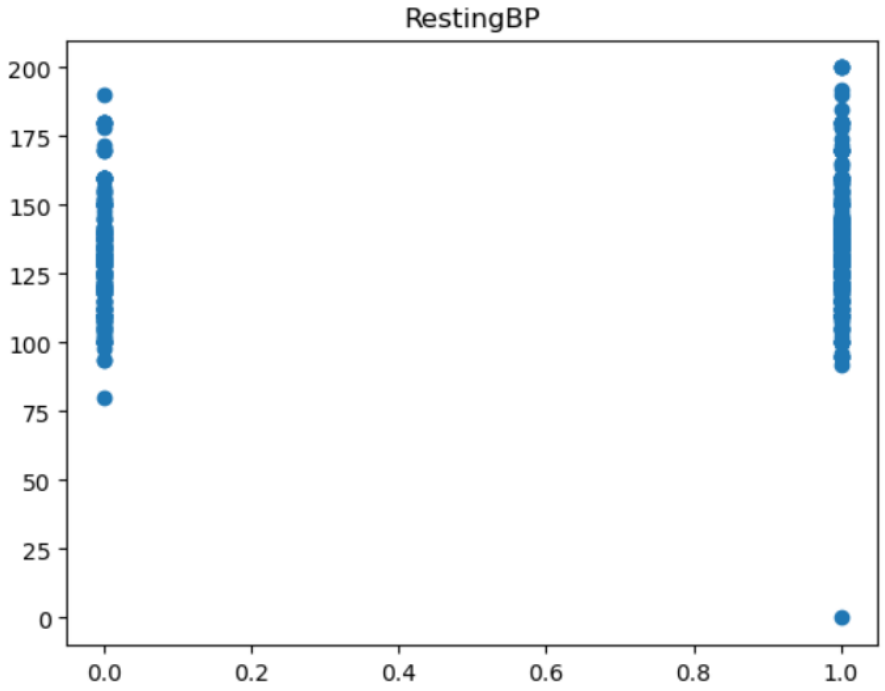
lower = 34.5 upper = 406.5

Number Of Outlier Element : 183 elements.

Number Of Non Outlier Element : 734 elements.

Detect outliers in ‘RestingBP’ attribute by

Scatter plot:



Q1 (25%) = 120.0 Q3 (75%) = 140.0 IQR = 20.0

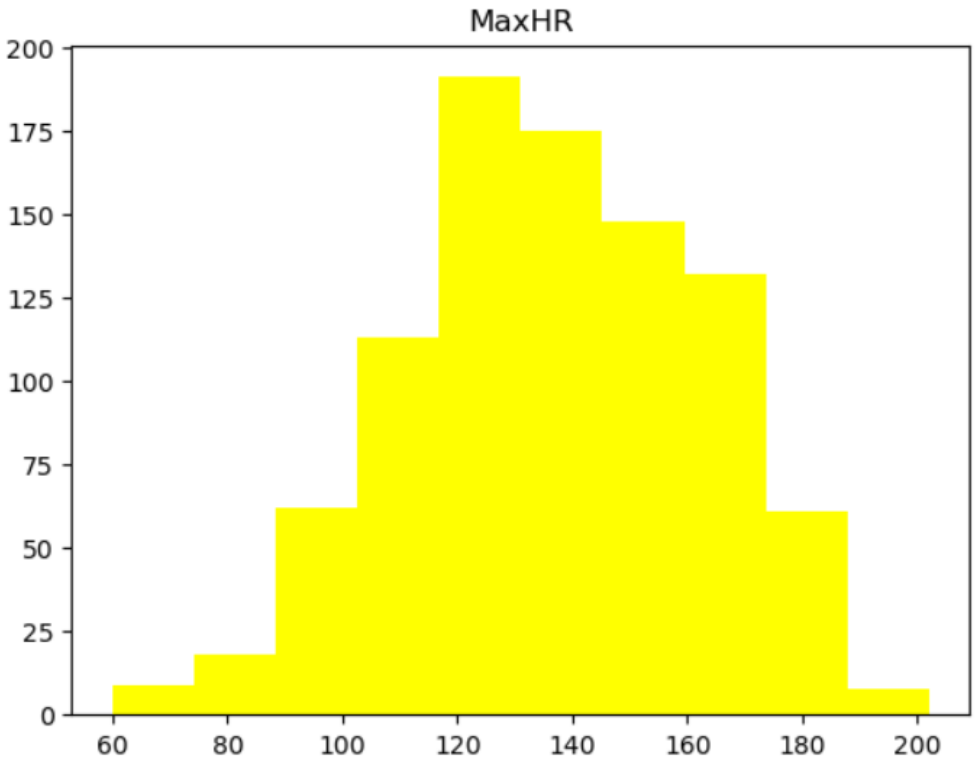
lower = 90.0 upper = 170.0

Number Of Outlier Element : 28 elements

Number Of Non Outlier Element : 889 elements

The range of outliers is [80:200] and it is acceptable in Resting BP so we didn't removed.

Detect outliers in ‘MaxHR’ attribute by Histogram: (not outliers in MaxHR)



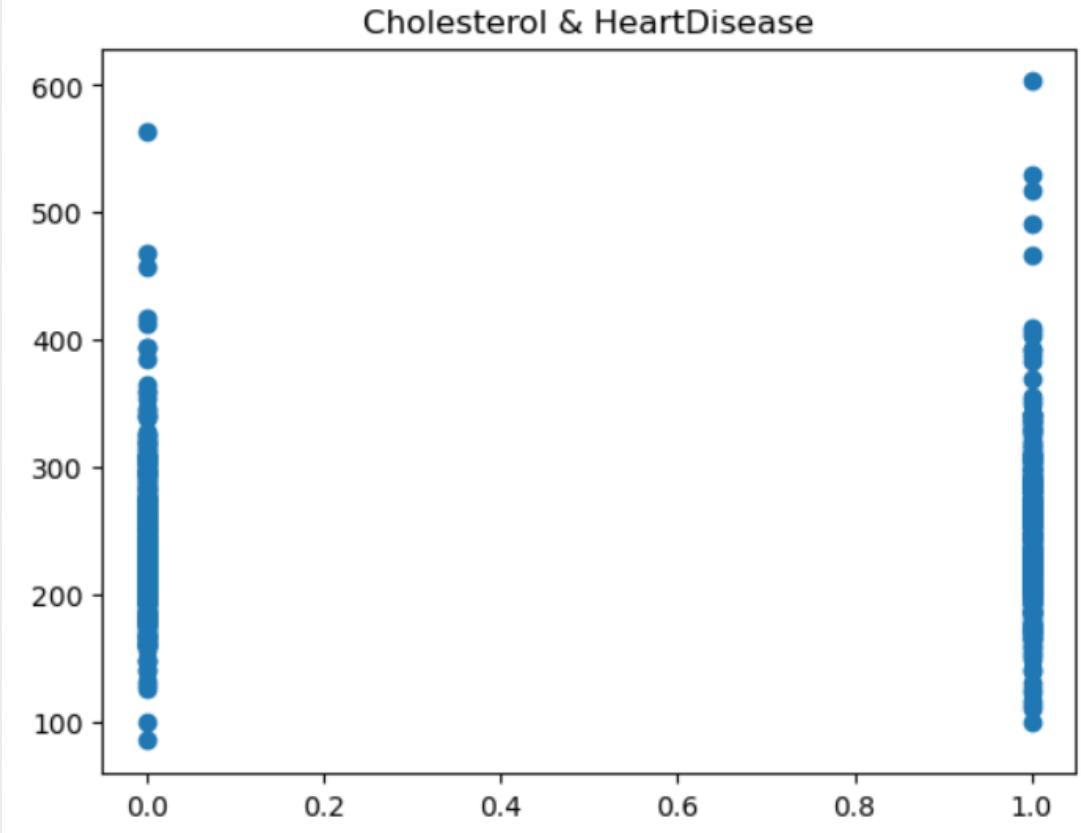
- Convert categorical data to numeric:

```
data["ST_Slope"]=data["ST_Slope"].replace({"Flat":1, "Up":2,"Down":0}, inplace=False)
data["ExerciseAngina"]=data["ExerciseAngina"].replace({"N": 0, "Y":1}, inplace=False)
data["RestingECG"]=data["RestingECG"].replace({"Normal":0, "ST":1,"LVH":2}, inplace=False)
data["Sex"]=data["Sex"].replace({"M": 1, "F":0}, inplace=False)
data["ChestPainType"]=data["ChestPainType"].replace({"ATA": 0, "TA":1,"NAP": 2, "ASY":3}, inplace=False)
```

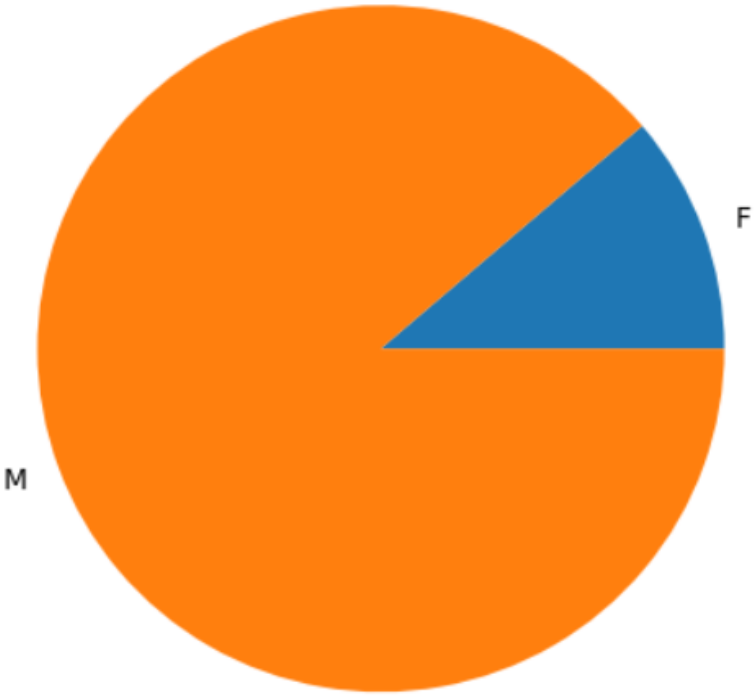
- data visualization charts:

  - scatter plot chart :

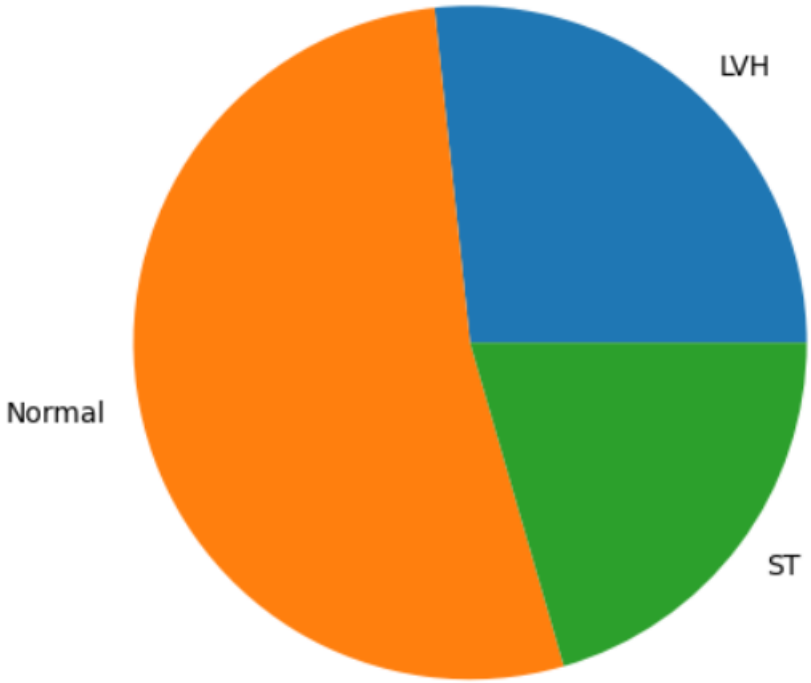
  Cholesterol above 400 indicates heart disease.



-pie chart: The percentage of men with heart disease is more than that of women.

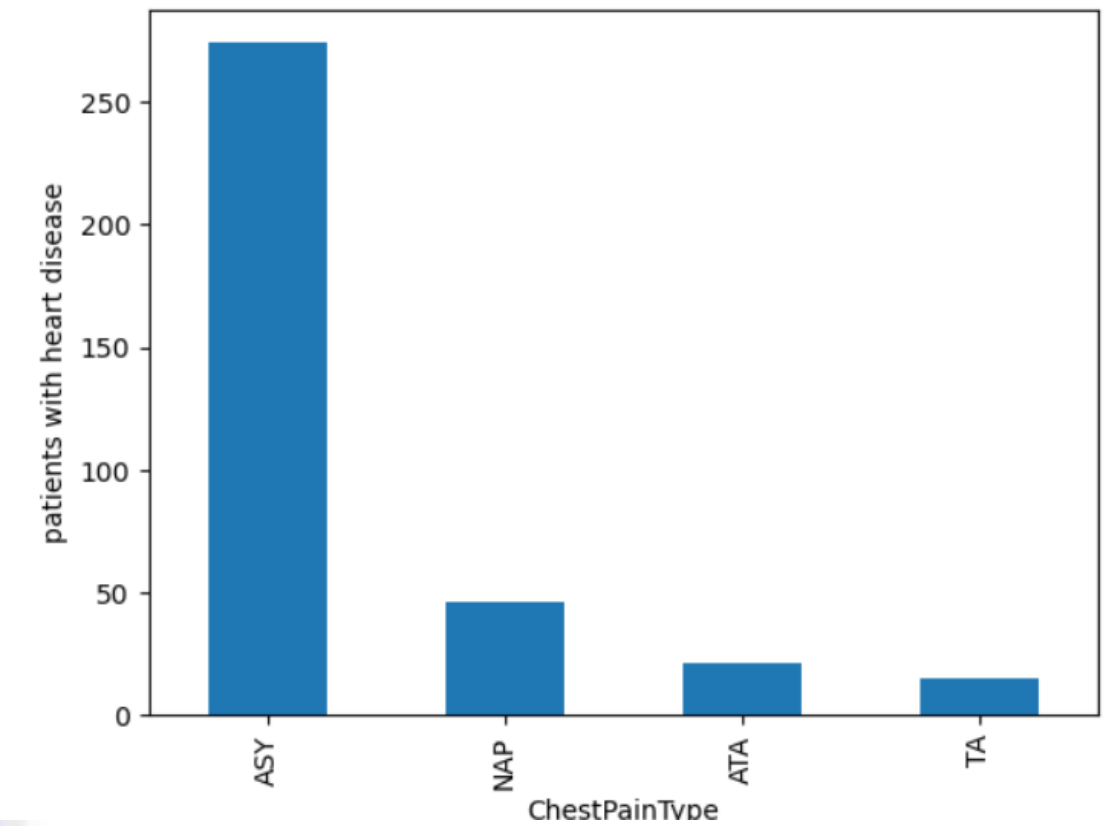


-More than half of people with heart disease have an regular heartbeat



- Most of those who feel TA or ATA (typical angina) do not have heart disease.

-Most people who feel the pain of angina asymptomatic (ASY), has a heart disease

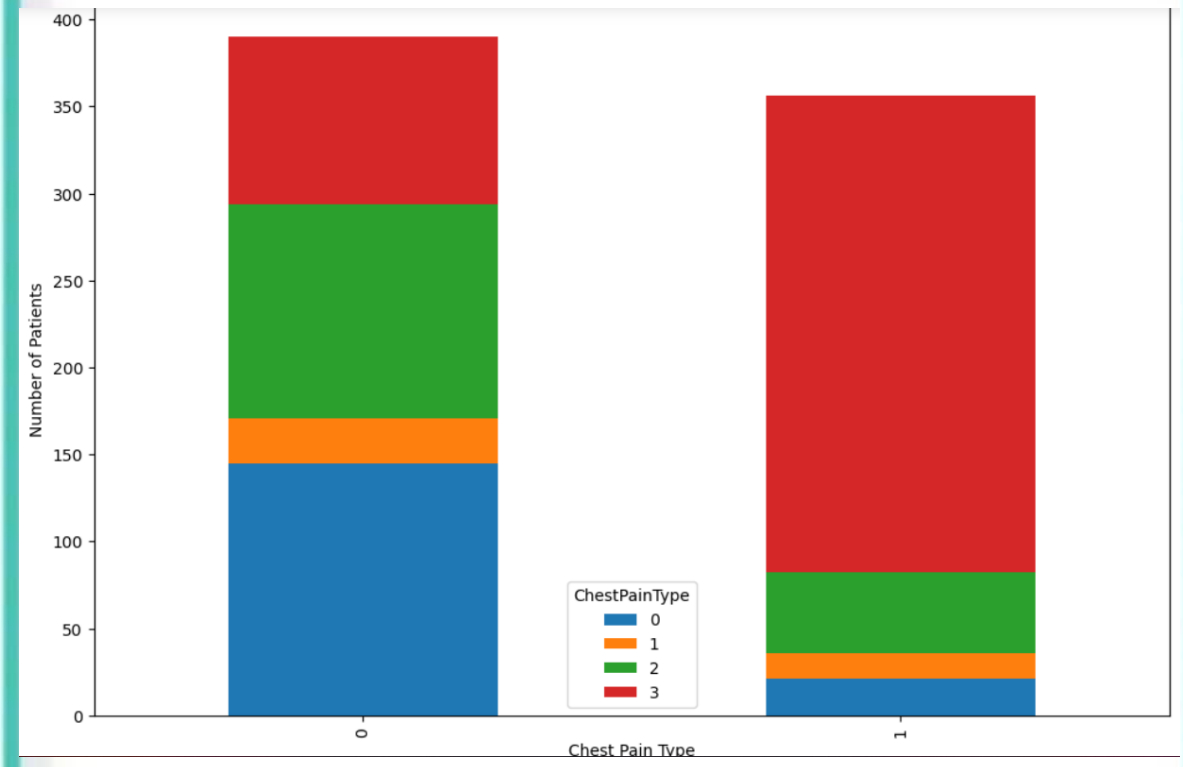


- Stacked Column Chart of Chest Pain Type and Heart Disease:

Blue: ATA ,Orande: TA ,Green: NAP , Red: ASY

Most of those who feel ATA (typical angina) do not have heart disease .

Most of those who feel ASY(Angina pain is not accompanied by symptoms) have heart disease .



- classification models:

implement GNB Classifier

```
from sklearn.metrics import accuracy_score
predict=nv.predict(x_test)
accuracy_score(y_test,predict)
```

```
0.84
```

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,predict)
```

```
array([[48, 29],
       [27, 46]], dtype=int64)
```

implement knn Classifier

```
accuracy_score(y_test,predict)
```

```
0.6266666666666667
```

```
confusion_matrix(y_test,predict)
```

```
array([[45, 31],
       [25, 49]], dtype=int64)
```

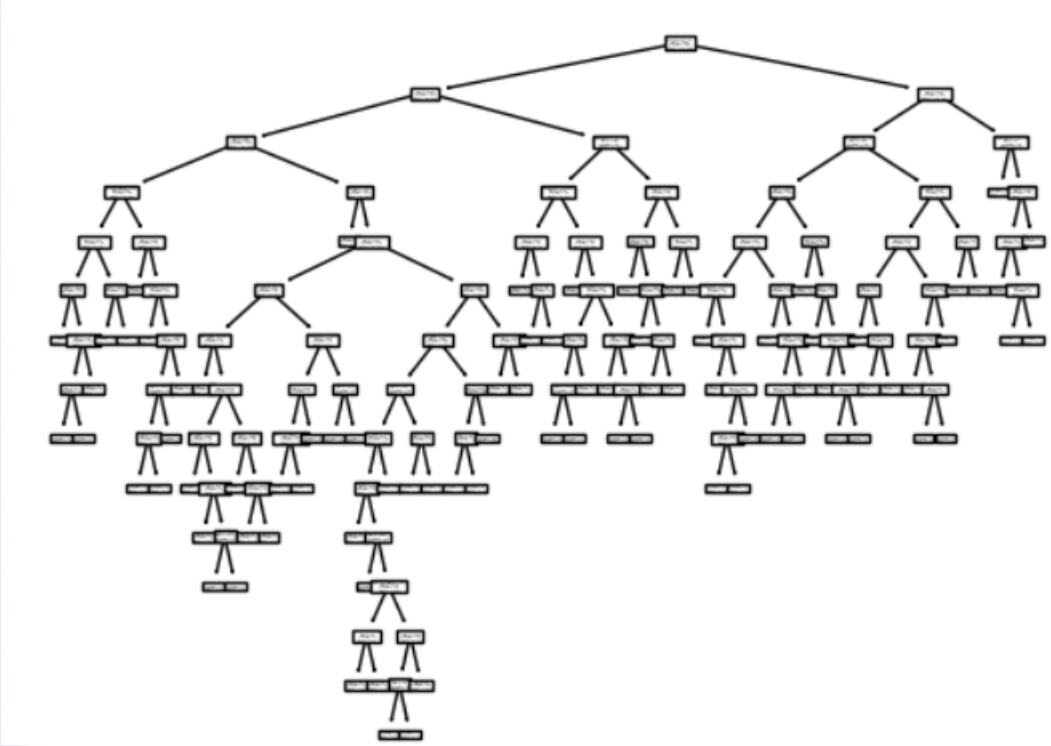
implement Decision Tree Classifier Classifier

```
predict=dtree.predict(x_test)
accuracy_score(y_test,predict)
```

```
0.8
```

```
confusion_matrix(y_test,predict)
```

```
array([[64, 12],
       [18, 56]], dtype=int64)
```



Concolusion:

Best classifier to predict the new person has heart disease or not based on some attributes is GNB Classifier with accuracy 0.84