Heart Disease Diagnostic Analysis

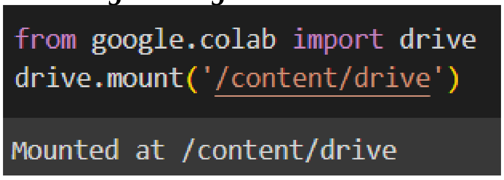
Problem Statement:

Health is real wealth in the pandemic time we all realized the brute effects of covid-19 on all irrespective of any status. You are required to analyze this health and medical data for betterfuture preparation.

Attribute Information:

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholesterol in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiograph results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- old-peak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by fluoroscope
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

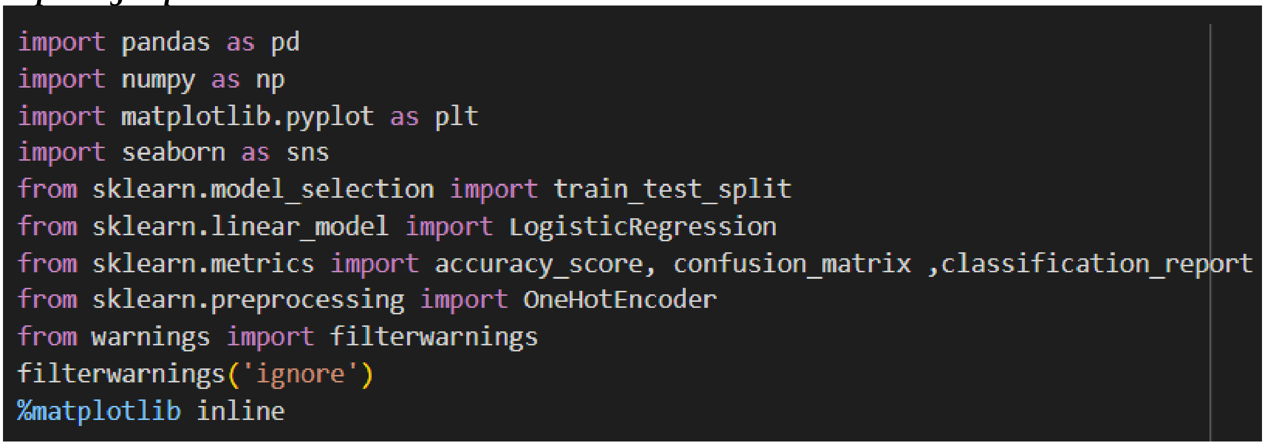
Mounting to Google Drive



```
from google.colab import drive
drive.mount('/content/drive')
```

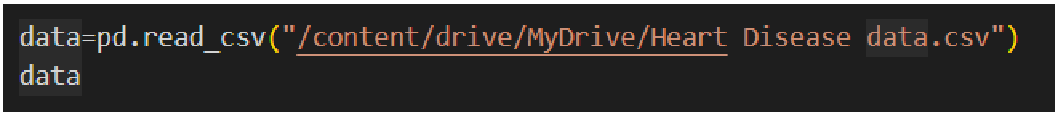
```
Mounted at /content/drive
```

Importing required libraries

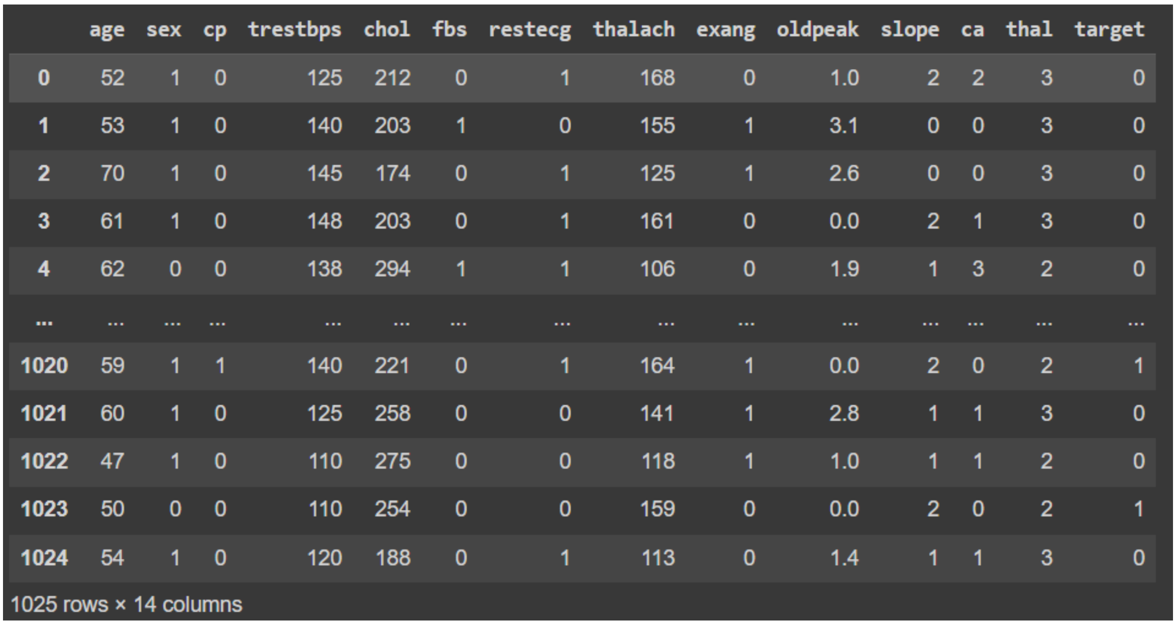


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix ,classification_report
from sklearn.preprocessing import OneHotEncoder
from warnings import filterwarnings
filterwarnings('ignore')
%matplotlib inline
```

Reading\Loading the Data set



```
data=pd.read_csv("/content/drive/MyDrive/Heart Disease data.csv")
data
```



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

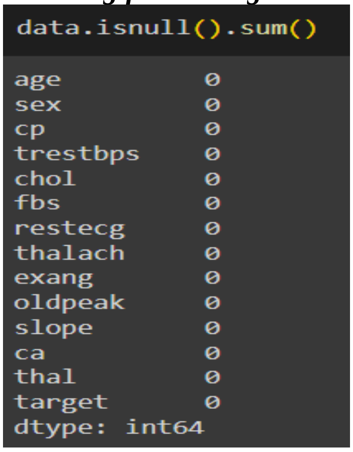
1025 rows × 14 columns

There are 1025 rows and 14 columns present in the Data set record.

Data Checks to perform

- Checking Missing values
- Checking Duplicates
- Checking data type
- Checking the number of unique values of each column
- Checking statistics of data set

Checking for missing values in data

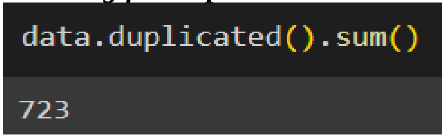


```
data.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

No Missing Values were found in the data.

Checking for Duplicate values in data

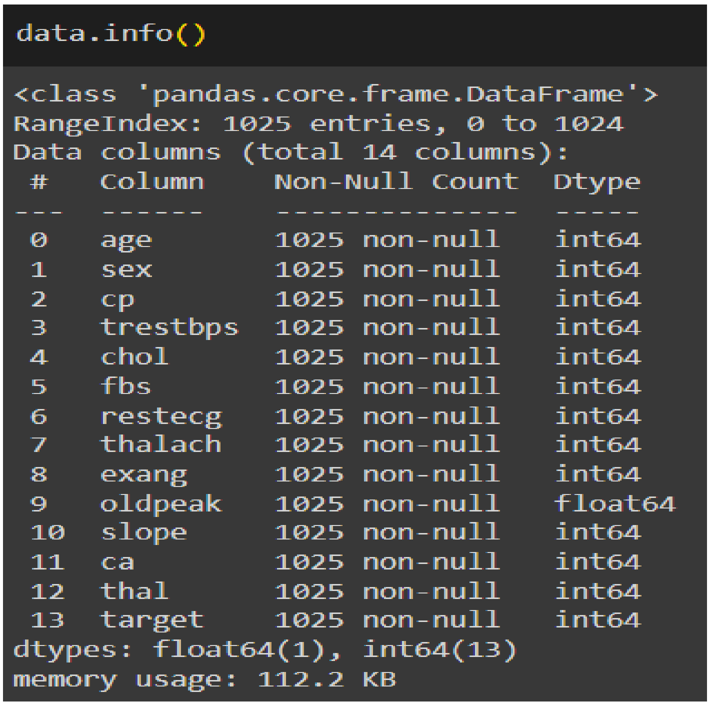


```
data.duplicated().sum()
```

```
723
```

723 Duplicate values were found in the given data set.

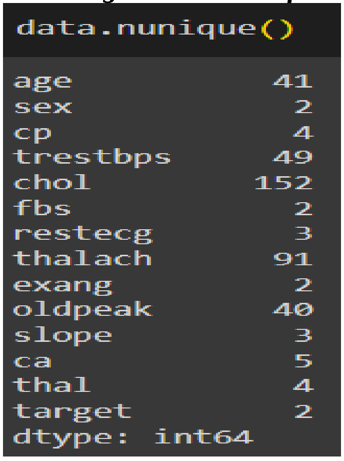
Checking the Data Types



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

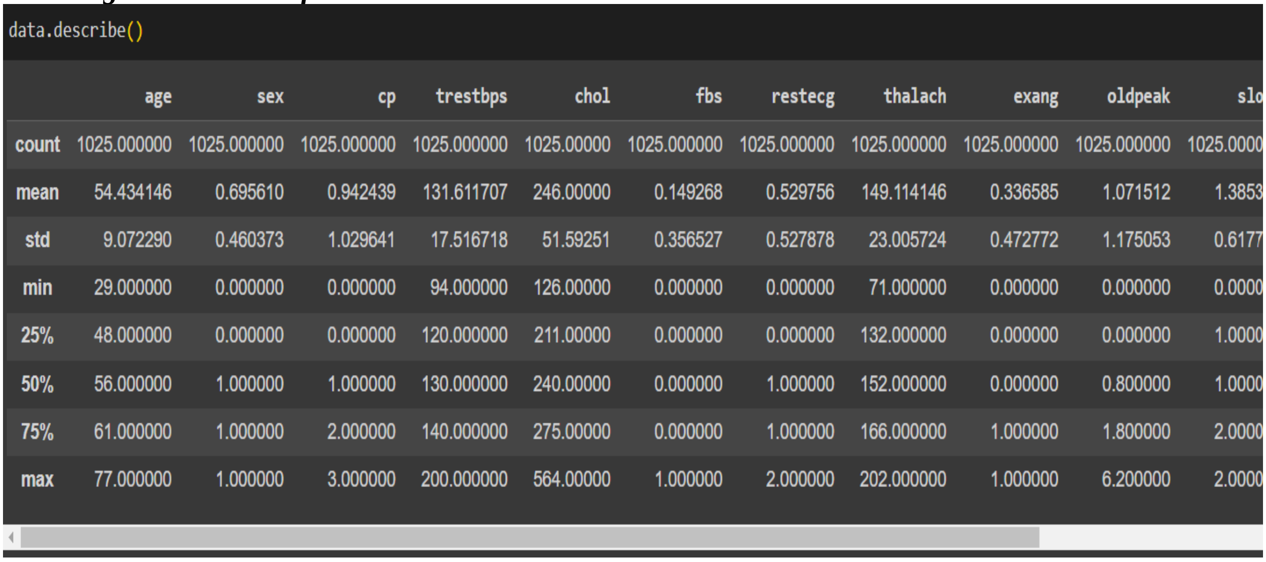
Checking the number of unique values of each columns



```
data.nunique()
```

```
age          41
sex           2
cp            4
trestbps     49
chol        152
fbs           2
restecg       3
thalach      91
exang         2
oldpeak      40
slope         3
ca            5
thal          4
target        2
dtype: int64
```

Checking the Statistics of the Data set



```
data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slo
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.0000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.3853
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.6177
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.0000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.0000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.0000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.0000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.0000

Insight 1: AGE & Heart Disease

•Age Distribution: The majority of the participants in the sample are between the ages of 48 and 61, with an average age of about 54. The age range has a maximum of 77 years and a minimum of 29 years.

•Impact on Heart Disease: Heart disease is more common in older people. The fact that people with heart disease tend to be older on average than people without the condition lends credence to this. Age-related increases in blood pressure and cholesterol are two heart disease risk factors that tend to rise with age.

Insight 2: Gender & Heart Disease

•Gender Distribution: The mean sex of the individuals is roughly 0.695, meaning that 69.5% of them are male.

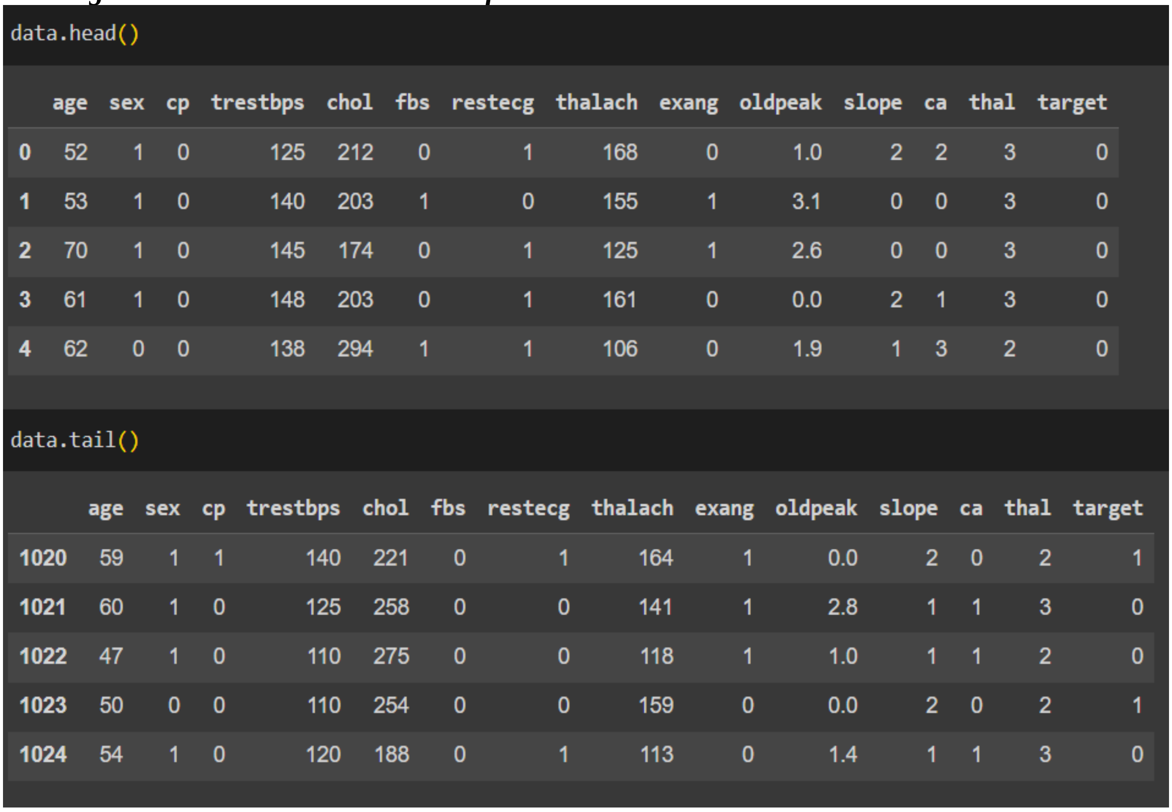
•Heart Disease Prevalence: Compared to women, men are more likely to suffer from heart disease. There could be a combination of genetic, lifestyle, and behavioral factors contributing to the higher frequency in men. For focused health interventions and awareness campaigns, this is essential.

Insight 3: Cholesterol Levels

•Distribution of Cholesterol: The data set's average cholesterol level is roughly 246 mg/dl, with a standard deviation of about 51.6 mg/dl. The range of cholesterol concentrations is 126 mg/dl to 564 mg/dl.

•Impact on Heart Disease: One of the main risk factors for heart disease is high cholesterol. Heart disease is more common in people with higher cholesterol levels. The data set makes this clear, showing that those with heart disease typically had greater cholesterol levels than people without the condition.One of the main goals of preventative health care initiatives should be cholesterol management.

Printing the First 5 rows & Last 5 rows from the Data set.



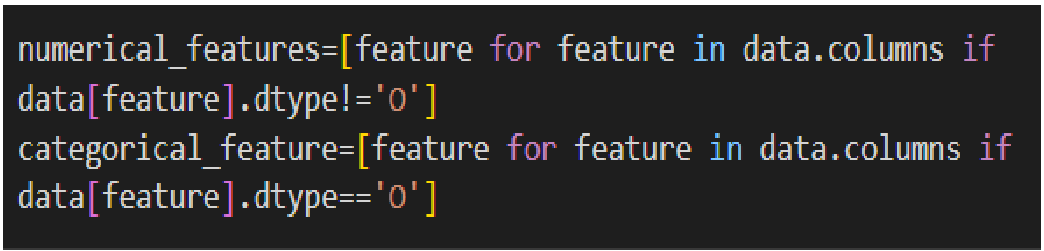
```
data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

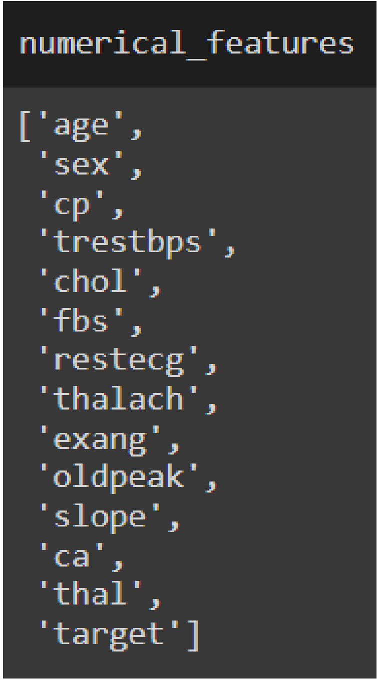
```
data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Separating Numerical and Categorical Columns

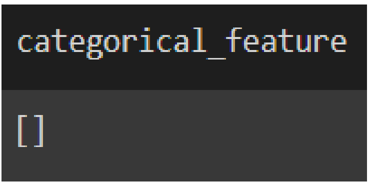


```
numerical_features=[feature for feature in data.columns if
data[feature].dtype!='O']
categorical_feature=[feature for feature in data.columns if
data[feature].dtype=='O']
```



```
numerical_features
```

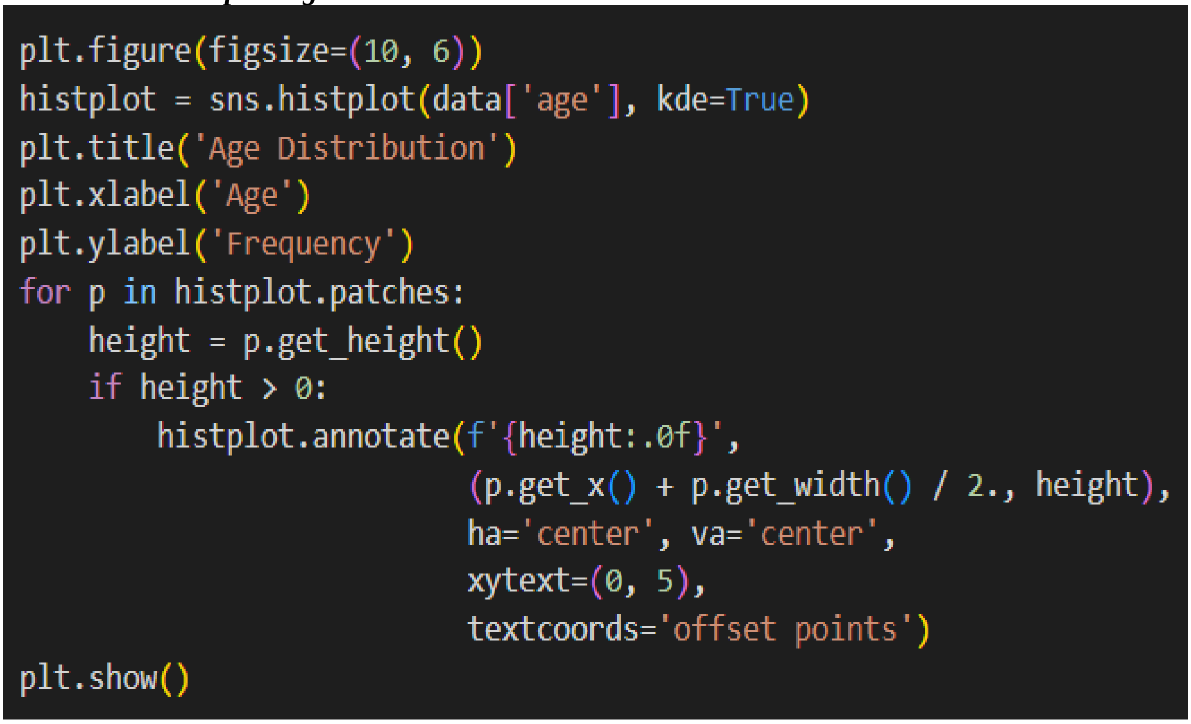
```
['age',
 'sex',
 'cp',
 'trestbps',
 'chol',
 'fbs',
 'restecg',
 'thalach',
 'exang',
 'oldpeak',
 'slope',
 'ca',
 'thal',
 'target']
```



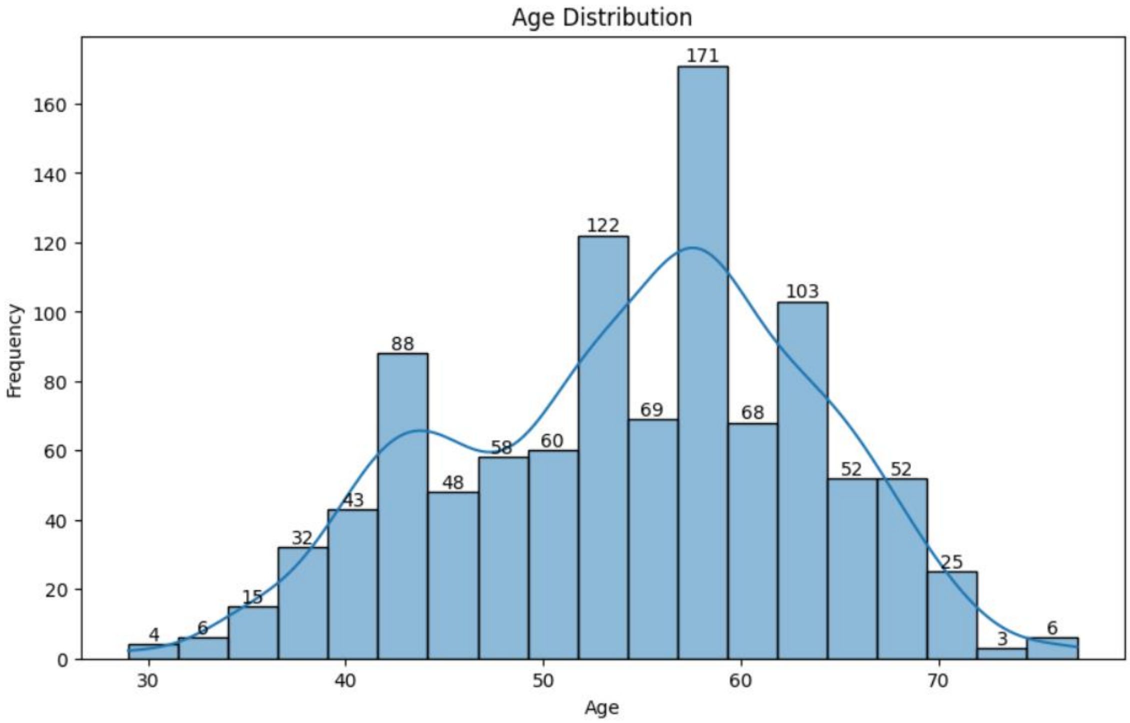
```
categorical_feature
```

```
[]
```

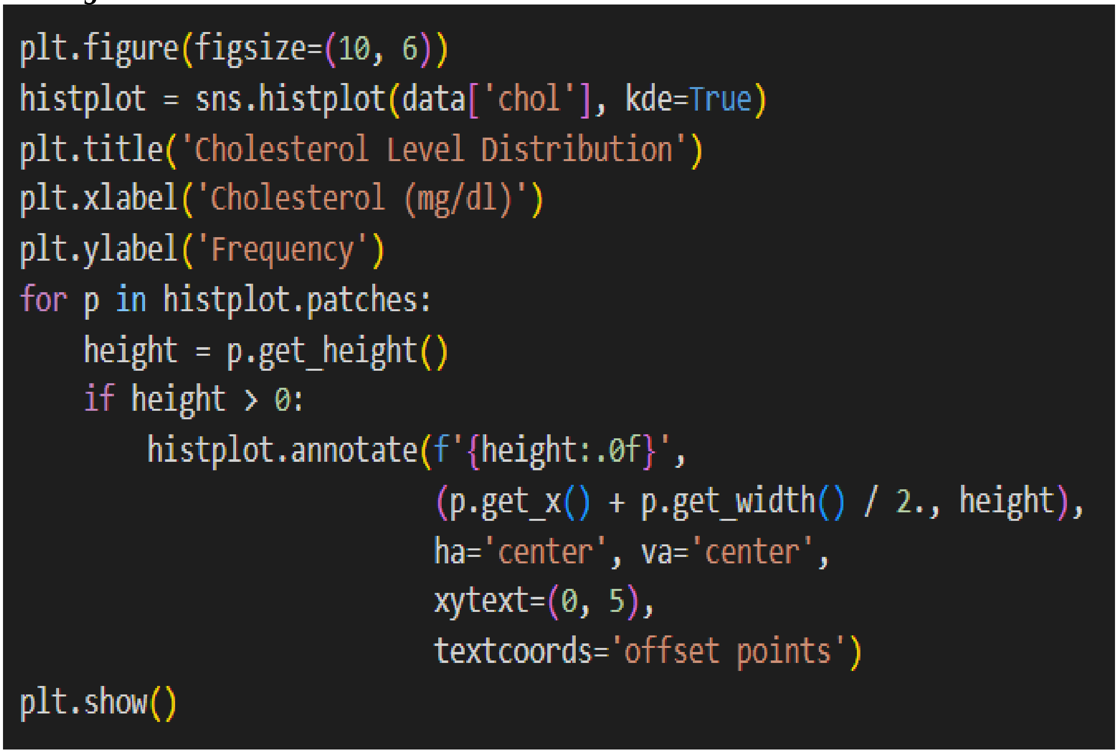
AGE Distribution plotting



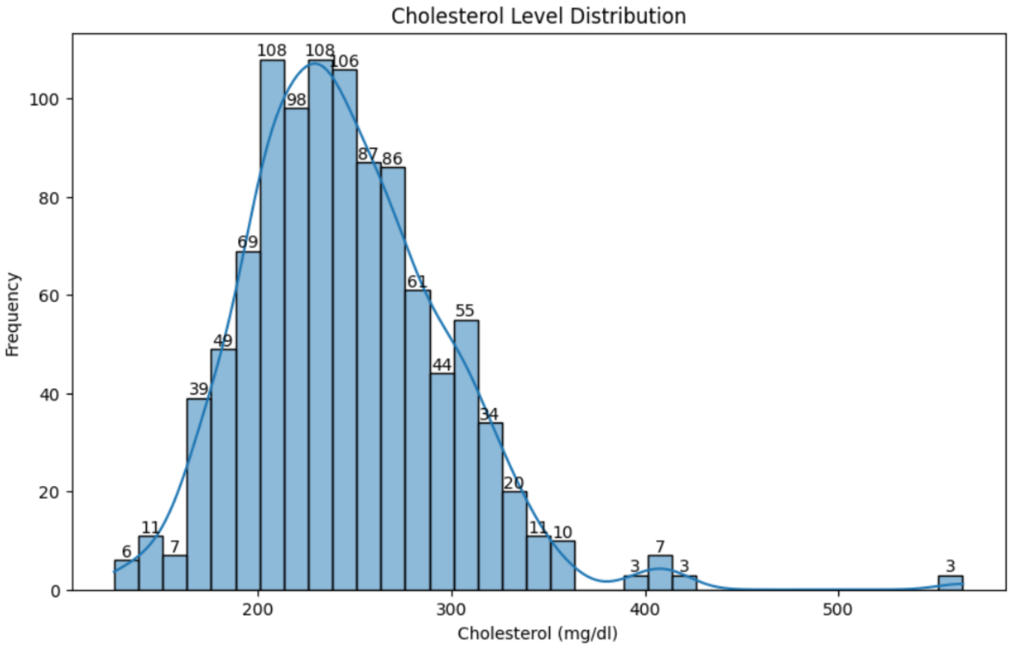
```
plt.figure(figsize=(10, 6))
histplot = sns.histplot(data['age'], kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
for p in histplot.patches:
    height = p.get_height()
    if height > 0:
        histplot.annotate(f'{height:.0f}',
                          (p.get_x() + p.get_width() / 2., height),
                          ha='center', va='center',
                          xytext=(0, 5),
                          textcoords='offset points')
plt.show()
```



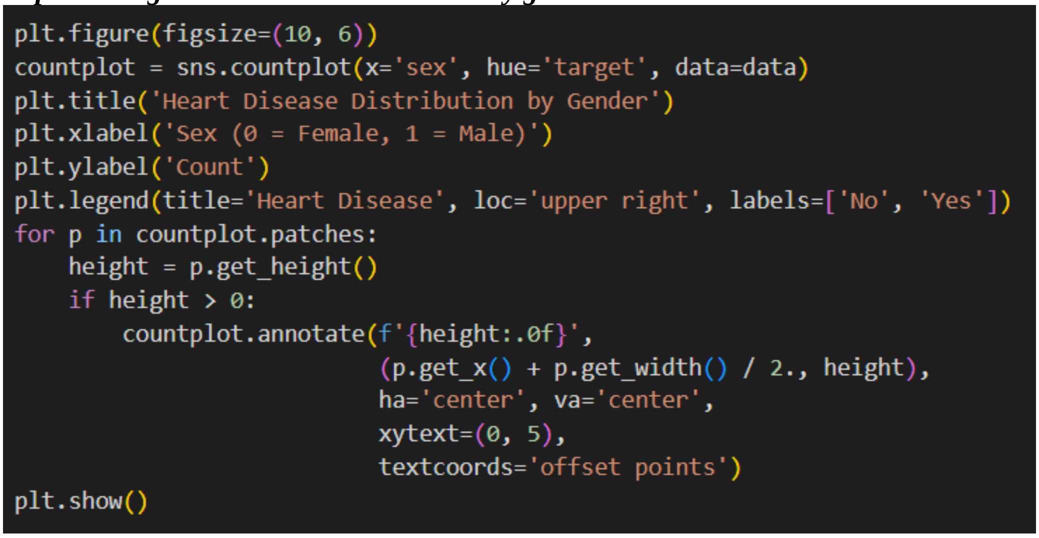
Plotting cholesterol level distribution



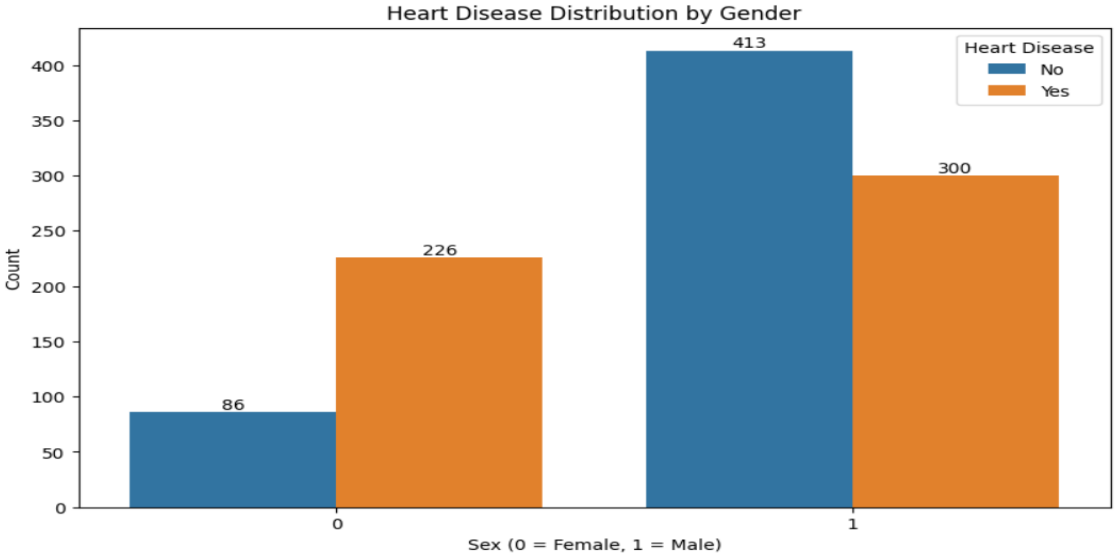
```
plt.figure(figsize=(10, 6))
histplot = sns.histplot(data['chol'], kde=True)
plt.title('Cholesterol Level Distribution')
plt.xlabel('Cholesterol (mg/dl)')
plt.ylabel('Frequency')
for p in histplot.patches:
    height = p.get_height()
    if height > 0:
        histplot.annotate(f'{height:.0f}',
                          (p.get_x() + p.get_width() / 2., height),
                          ha='center', va='center',
                          xytext=(0, 5),
                          textcoords='offset points')
plt.show()
```



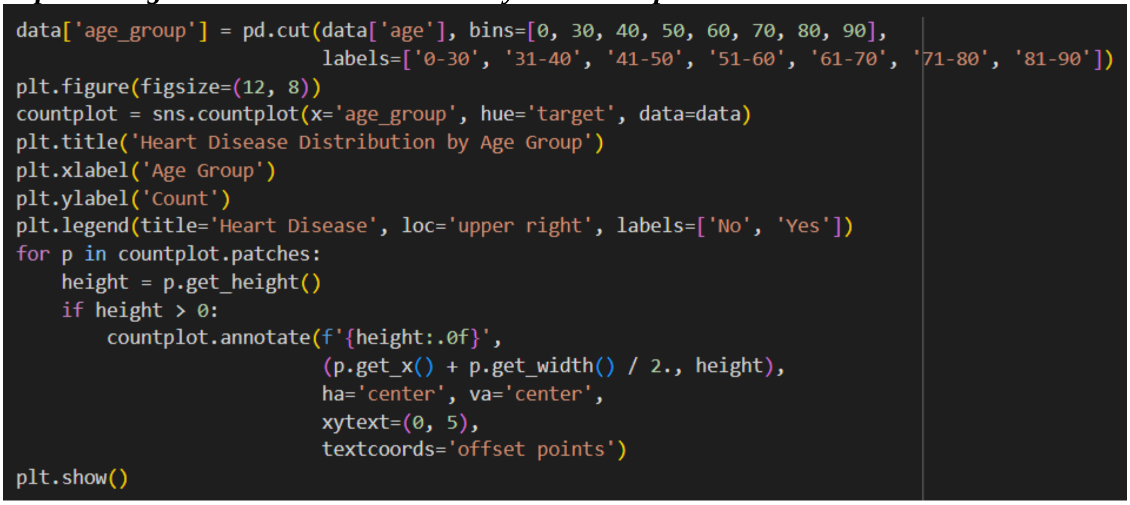
Representing Heart disease distribution by gender



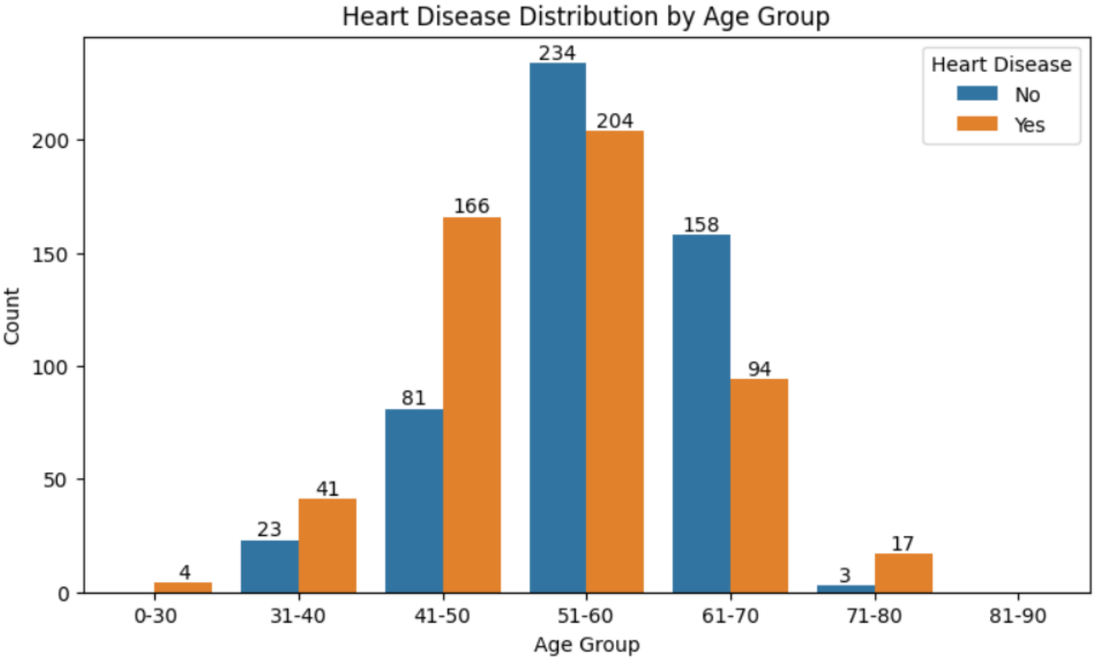
```
plt.figure(figsize=(10, 6))
countplot = sns.countplot(x='sex', hue='target', data=data)
plt.title('Heart Disease Distribution by Gender')
plt.xlabel('Sex (0 = Female, 1 = Male)')
plt.ylabel('Count')
plt.legend(title='Heart Disease', loc='upper right', labels=['No', 'Yes'])
for p in countplot.patches:
    height = p.get_height()
    if height > 0:
        countplot.annotate(f'{height:.0f}',
                           (p.get_x() + p.get_width() / 2., height),
                           ha='center', va='center',
                           xytext=(0, 5),
                           textcoords='offset points')
plt.show()
```



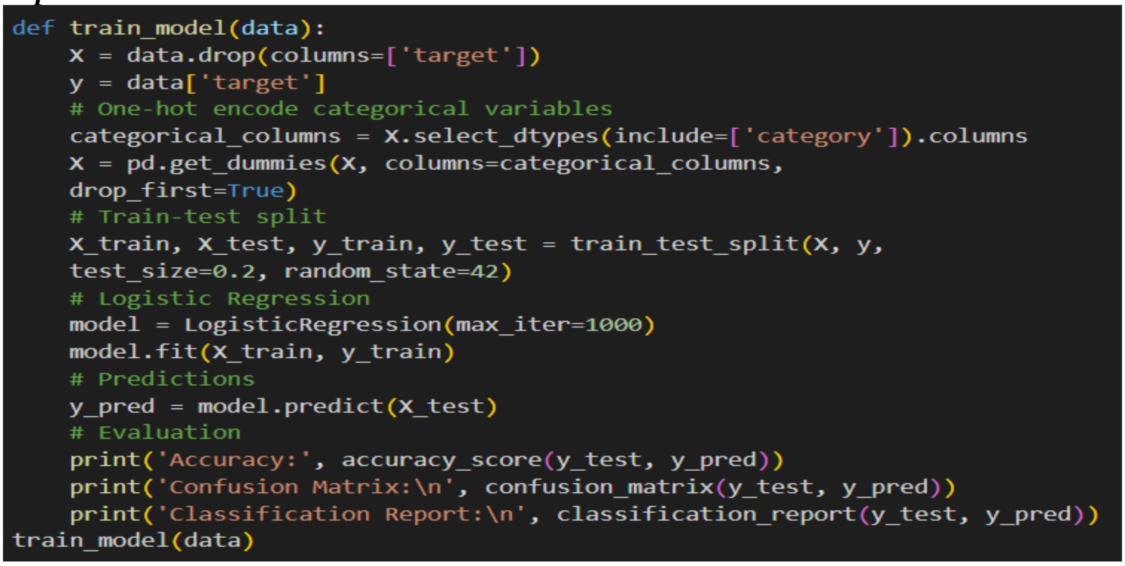
Representing Heart disease distribution by AGE Group



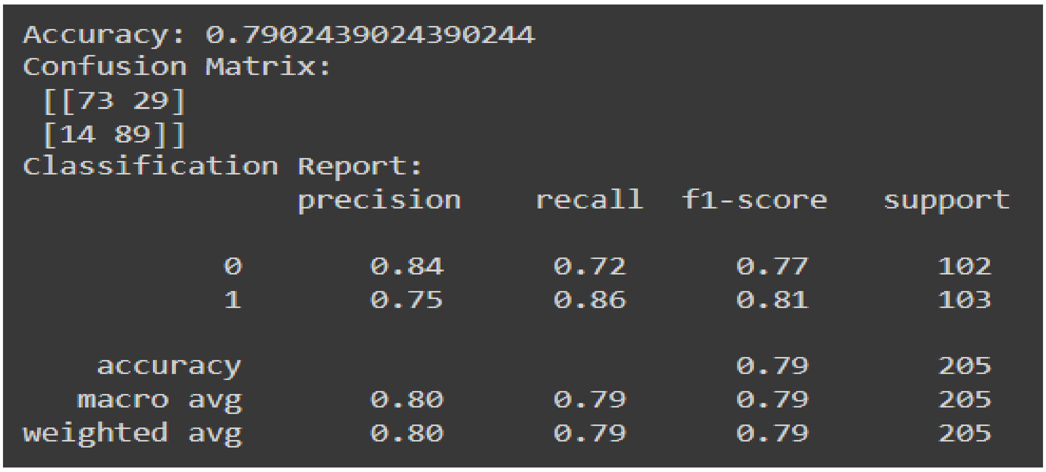
```
data['age_group'] = pd.cut(data['age'], bins=[0, 30, 40, 50, 60, 70, 80, 90],
                           labels=['0-30', '31-40', '41-50', '51-60', '61-70', '71-80', '81-90'])
plt.figure(figsize=(12, 8))
countplot = sns.countplot(x='age_group', hue='target', data=data)
plt.title('Heart Disease Distribution by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.legend(title='Heart Disease', loc='upper right', labels=['No', 'Yes'])
for p in countplot.patches:
    height = p.get_height()
    if height > 0:
        countplot.annotate(f'{height:.0f}',
                           (p.get_x() + p.get_width() / 2., height),
                           ha='center', va='center',
                           xytext=(0, 5),
                           textcoords='offset points')
plt.show()
```



Using Machine Learning Models to find Accuracy, Confusion Matrix & Classification Report



```
def train_model(data):
    X = data.drop(columns=['target'])
    y = data['target']
    # One-hot encode categorical variables
    categorical_columns = X.select_dtypes(include=['category']).columns
    X = pd.get_dummies(X, columns=categorical_columns,
    drop_first=True)
    # Train-test split
    X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=0.2, random_state=42)
    # Logistic Regression
    model = LogisticRegression(max_iter=1000)
    model.fit(X_train, y_train)
    # Predictions
    y_pred = model.predict(X_test)
    # Evaluation
    print('Accuracy:', accuracy_score(y_test, y_pred))
    print('Confusion Matrix:\n', confusion_matrix(y_test, y_pred))
    print('Classification Report:\n', classification_report(y_test, y_pred))
train_model(data)
```



```
Accuracy: 0.7902439024390244
Confusion Matrix:
 [[73 29]
 [14 89]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.72      0.77       102
           1       0.75      0.86      0.81       103

    accuracy                           0.79       205
   macro avg       0.80      0.79      0.79       205
weighted avg       0.80      0.79      0.79       205
```

Conclusions

Age and Distributions of Cholesterol: Heart disease is more common in those in their mid-50s. The risk of heart disease is increased by raised cholesterol levels, which vary widely.

Disparities by Gender: Men are more likely than women to get heart disease.

Analysis of Correlation: Age and the maximum heart rate reached are negatively correlated. There is a strong correlation between the types of chest pain, the prevalence of heart disease, and exercise-induced angina.

Analysis of Age Groups: The 51–60 and 61–70 age groups have higher rates of heart disease.

Model Performance: With an accuracy of 79.02%, the logistic regression model accurately predicted heart disease with good precision and recall.

-----------------------Thank You-----------------------