Unified Mentor Data Analytics Internship.

Name : - Ashutosh Kumar

project : Heart Disease Diagnostic Analysis

1) 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 better future preparation.

2) Data Collection

- Dataset Source - Heart Disease data.csv(Given)
- The data consists of 14 column and 1024 rows.

3) Attribute Information:

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

```
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
```

```
## Create DataFrame And read the dataset using pandas
data = pd.read_csv('Heart Disease data.csv')
data.head()
```

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

   ca  thal  target  
0   2     3       0  
1   0     3       0  
2   0     3       0  
3   1     3       0  
4   3     2       0  
```

```
data.shape
```

```
(1025, 14)
```

3. Data Checks to perform

- Check Missing values
- Check Duplicates
- Check data type
- Check the number of unique values of each column
- Check statistics of data set
- Check various categories present in the different categorical column

```
## Check missing values
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
```

Insights or Observation

There are no missing values

```
data.isna().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
```

```
## Check Duplicates
data.duplicated().sum()
```

```
723
```

There are 722 duplicates values in the dataset

```
## check datatypes
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
```

```
## 3.1 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
```

```
## Check the statistics of the dataset
data.describe()
```

```
               age          sex           cp     trestbps        chol
\
count  1025.000000  1025.000000  1025.000000  1025.000000  1025.00000

mean     54.434146     0.695610     0.942439   131.611707   246.00000

std       9.072290     0.460373     1.029641    17.516718    51.59251

min      29.000000     0.000000     0.000000    94.000000   126.00000

25%      48.000000     0.000000     0.000000   120.000000   211.00000

50%      56.000000     1.000000     1.000000   130.000000   240.00000

75%      61.000000     1.000000     2.000000   140.000000   275.00000

max      77.000000     1.000000     3.000000   200.000000   564.00000
```

```
               fbs     restecg     thalach       exang     oldpeak  \
count  1025.000000  1025.000000  1025.000000  1025.000000  1025.000000

mean      0.149268     0.529756   149.114146     0.336585     1.071512

std       0.356527     0.527878    23.005724     0.472772     1.175053

min       0.000000     0.000000    71.000000     0.000000     0.000000

25%       0.000000     0.000000   132.000000     0.000000     0.000000

50%       0.000000     1.000000   152.000000     0.000000     0.800000

75%       0.000000     1.000000   166.000000     1.000000     1.800000

max       1.000000     2.000000   202.000000     1.000000     6.200000


             slope           ca         thal       target
count  1025.000000  1025.000000  1025.000000  1025.000000
mean      1.385366     0.754146     2.323902     0.513171
std       0.617755     1.030798     0.620660     0.500070
min       0.000000     0.000000     0.000000     0.000000
25%       1.000000     0.000000     2.000000     0.000000
50%       1.000000     0.000000     2.000000     1.000000
75%       2.000000     1.000000     3.000000     1.000000
max       2.000000     4.000000     3.000000     1.000000
```

Insight 1: Age and Heart Disease

- Age Distribution: The average age of individuals in the dataset is approximately 54 years, with the majority falling between 48 to 61 years old. The maximum age is 77 years, and the minimum is 29 years.

- Impact on Heart Disease: Older individuals are more likely to have heart disease. This is supported by the higher mean age of individuals with heart disease compared to those without it. As age increases, the risk factors associated with heart disease, such as higher cholesterol levels and increased blood pressure, also tend to increase.

Insight 2: Gender and Heart Disease

- Gender Distribution: About 69.5% of the subjects are male (sex mean is approximately 0.695).

- Heart Disease Prevalence: Males are more affected by heart disease compared to females. The higher prevalence among males may be related to a combination of genetic, lifestyle, and behavioral factors. This is crucial for targeted health interventions and awareness programs.

Insight 3: Cholesterol Levels

- Cholesterol Distribution: The average cholesterol level in the dataset is around 246 mg/dl, with a standard deviation of approximately 51.6 mg/dl. The cholesterol levels range from 126 mg/dl to 564 mg/dl.

- Impact on Heart Disease: High cholesterol is a significant risk factor for heart disease. Individuals with higher cholesterol levels are more likely to develop heart disease. This is evident from the dataset where those with heart disease tend to have higher cholesterol levels on average compared to those without heart disease. Cholesterol management should be a key focus in preventive healthcare strategies.

```
## Explore more info about the data
data.head()

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

   ca  thal  target  
0   2     3       0  
1   0     3       0  
2   0     3       0  
3   1     3       0  
4   3     2       0  

data.tail()

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

      slope  ca  thal  target  
```

```
1020       2   0     2        1
1021       1   1     3        0
1022       1   1     2        0
1023       2   0     2        1
1024       1   1     3        0
```

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

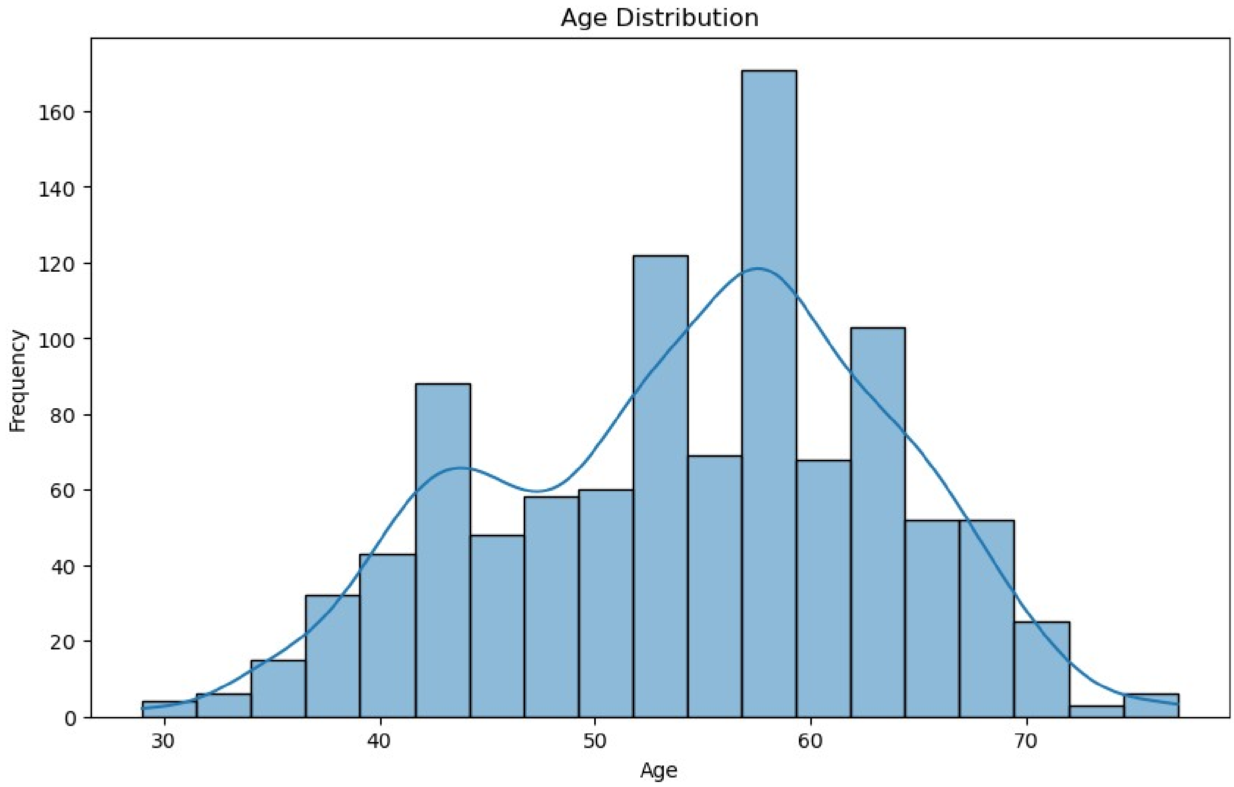
```
[]
```

```
#segrregate numerical and categorical features
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']
```

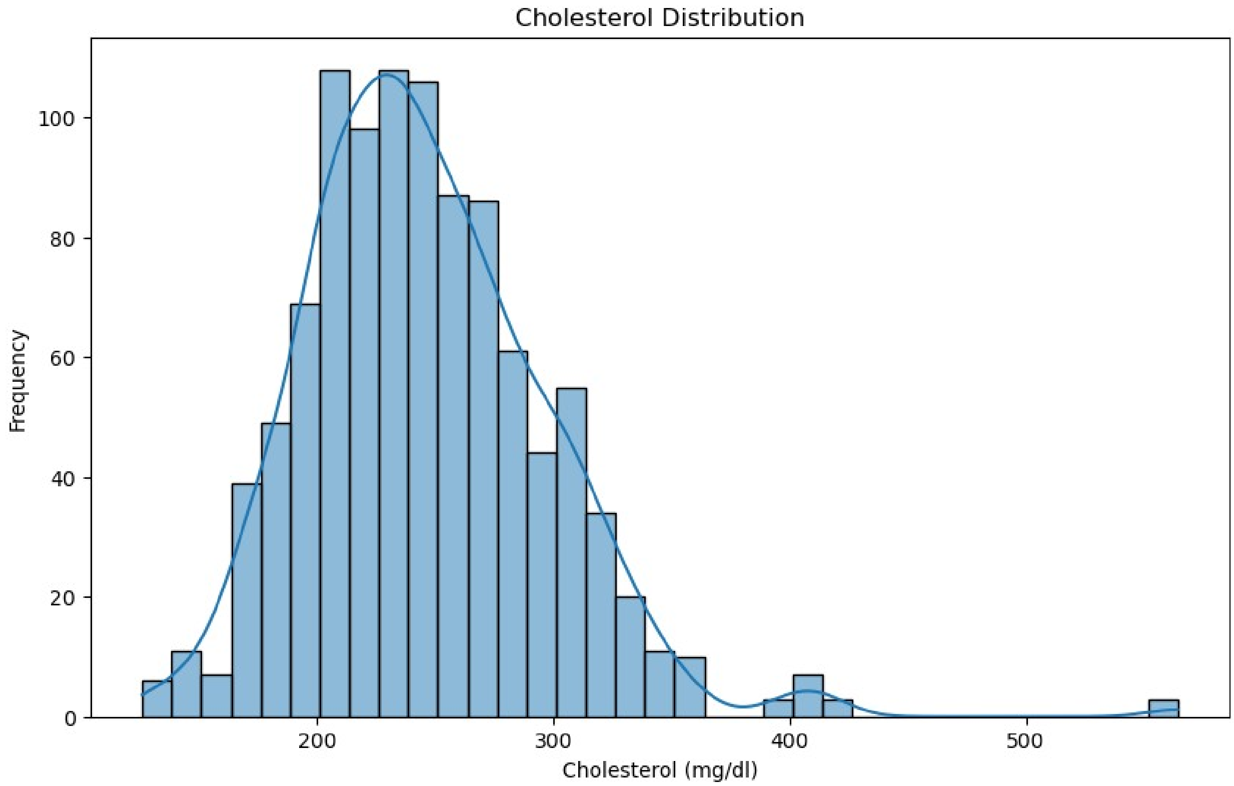
```
numerical_features
```

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

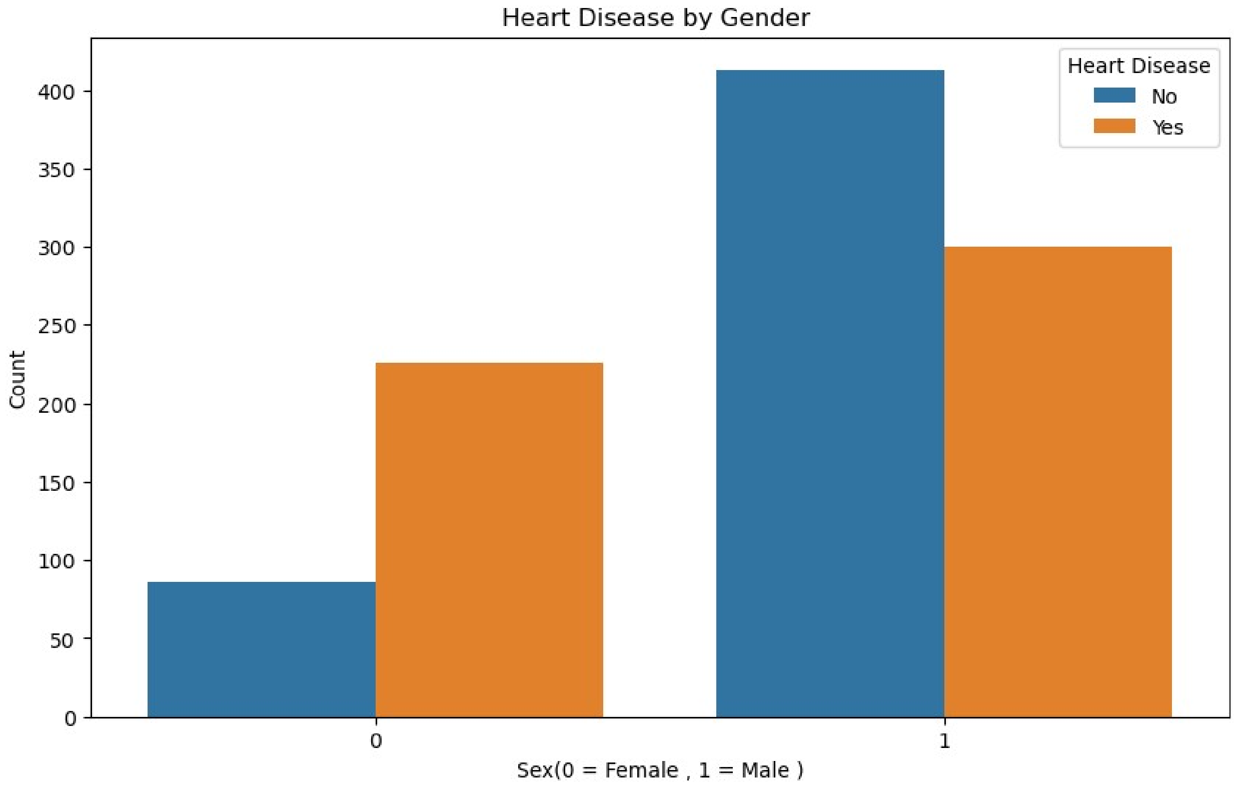
```
## Age Distribution
plt.figure(figsize=(10 , 6))
sns.histplot(data['age'], kde= True )
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



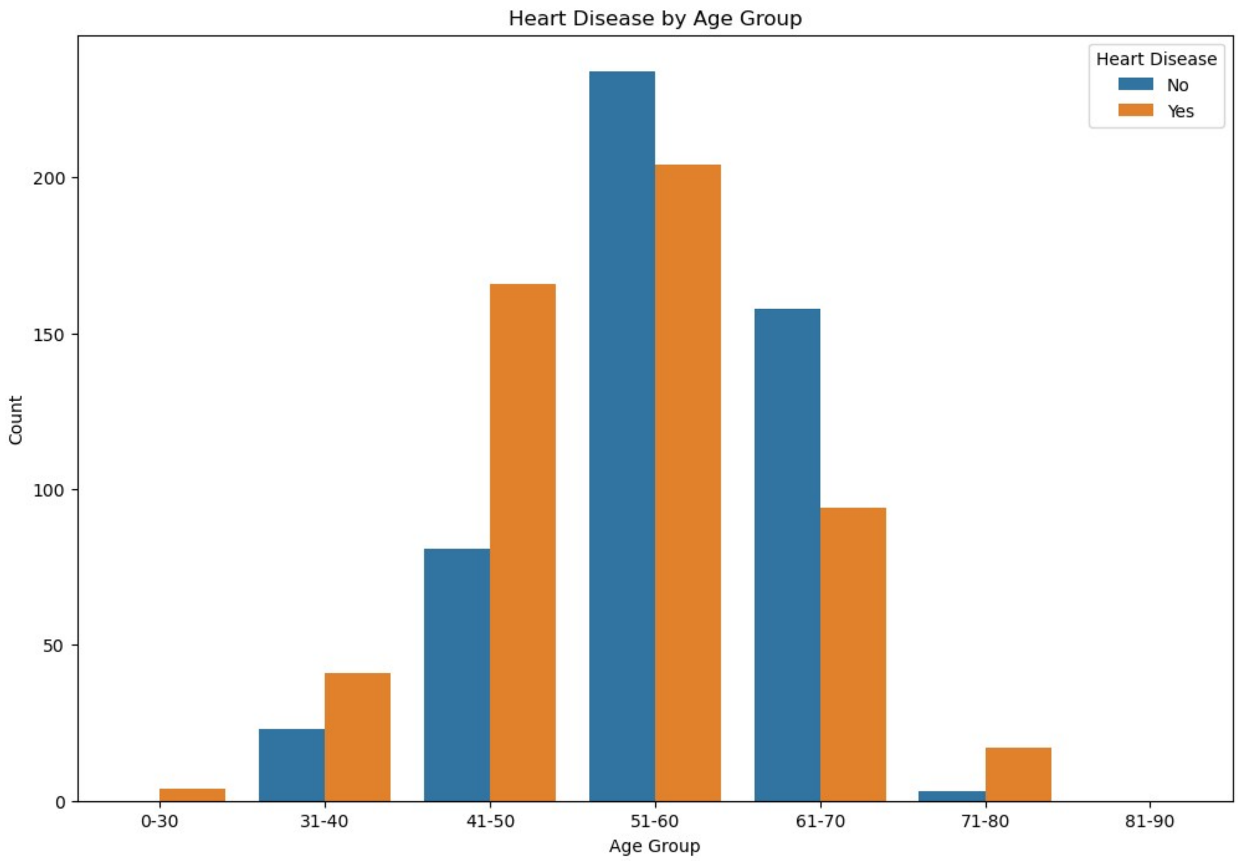
```
## Cholesterol distribution
plt.figure(figsize=(10,6))
sns.histplot(data['chol'], kde = True )
plt.title('Cholesterol Distribution')
plt.xlabel('Cholesterol (mg/dl)')
plt.ylabel('Frequency')
plt.show()
```



```
## Heart Disease by Gender
plt.figure(figsize=(10 ,6))
sns.countplot(x = 'sex' , hue = 'target' , data = data )
plt.title('Heart Disease by Gender ')
plt.xlabel('Sex(0 = Female , 1 = Male )')
plt.ylabel ('Count')
plt.legend(title = 'Heart Disease' , loc = 'upper right' , labels =
['No' , 'Yes'])
plt.show()
```



```
# Heart Disease 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))
sns.countplot(x='age_group', hue='target', data=data)
plt.title('Heart Disease by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.legend(title='Heart Disease', loc='upper right', labels=['No',
'Yes'])
plt.show()
```



Machine Learning Model

```
def train_model(data):
    # Select features and target
    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: 0.7853658536585366
Confusion Matrix:
 [[72 30]
 [14 89]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.71      0.77       102
           1       0.75      0.86      0.80       103

    accuracy                           0.79       205
   macro avg       0.79      0.78      0.78       205
weighted avg       0.79      0.79      0.78       205
```

Summary of Findings

-Age and Cholesterol Distributions:

Heart disease is more prevalent among individuals in their mid-50s. Cholesterol levels vary widely, with high levels contributing to heart disease risk. Gender Differences:

Males are more likely to suffer from heart disease compared to females. Correlation Analysis:

Negative correlation between age and maximum heart rate achieved. Strong correlation between exercise-induced angina, chest pain types, and heart disease occurrence. Age Group Analysis:

Higher prevalence of heart disease in age groups 51-60 and 61-70. Model Performance:

Logistic regression model achieved an accuracy of 78.54%, effectively predicting heart disease with good precision and recall.