Heart Disease Risk Analysis

Mohit Thaakar
(SUID: 450607855)
Electrical Engineering and Computer Science Department
Syracuse University
Syracuse, United States

Kavya Shah
(SUID: 330126500)
Electrical Engineering and Computer Science Department
Syracuse University
Syracuse, United States

Sohan Thakur
(SUID: 578607587)
Electrical Engineering and Computer Science Department
Syracuse University
Syracuse, United States

***Abstract*— Cardiovascular diseases pose a significant health concern that necessitates a holistic approach to understand its causes and risk at an early age. Our research aims to analyze the impact of various lifestyle factors on heart disease risk among patients. The overall goal is to provide valuable insights into the complex interplay between lifestyle choices and heart disease risk.**

Keywords—Heart Disease risk analysis, Data Preprocessing, Data Visualization, Exploratory Data Analysis, Logistic Regression, Predictive modeling

I. Introduction

Heart Disease is one of the leading causes of death in the United States that leads to thousands of lives every year. The impact of this is that it affects healthcare and individuals that leads to high socio-economic costs. There could be multiple factors leading to heart diseases which range from demographic factors such as age, gender, family history and genetics to some of the modifiable factors such as increased blood pressure, high cholesterol, sedentary lifestyle, high consumption of alcohol and regular smoking. The effect of this leads to higher risk of developing a heart disease and can have serious impacts on an individual. There are various organizations such as American Heart Association (AHA), the American College of Cardiology (ACC) and the European Society of Cardiology that work day in and day out to give us updated information and guidelines on the heart related topics. As we dive deep into the journey of unraveling the complexities of the heart, it becomes very important to have a good understanding of it. Our research paper dives into the intricate details of heart disease, how the factors increase the risk and proposing strategies to alleviate the risk.

II. Background

Cardiovascular diseases are one of the leading factors of death globally. Moreover, they also lead to a lot of premature deaths which also cause problems in the healthcare infrastructures. There are two types of factors that cause heart diseases, modifiable and non-modifiable factors. Non modifiable factors include age, gender, family history, genetics. Modifiable factors include hypertension, hyperlipidemia, diabetes, obesity, smoking, alcohol consumption, etc.

In-depth research has let to providing guidelines by various organizations that advocate lifestyle changes such as exercising, meditation, having a good balanced diet, quit smoking.

The relationship among these factors call for having a good understanding of probability and statistics. The analysis of predictive models help utilize demographic and lifestyle data that gives us immense knowledge. It also calls for using preventive measures to the patients and the models can also help in the early detection of heart diseases that can save an individual's life and improve the overall healthcare system..

III. Methodology

We make use of Python programming and other robust libraries that contribute to our analysis on heart disease. We first import all the necessary libraries and load the heart disease dataset from Kaggle into a panda DataFrame. This dataset has diverse patient attributes such as smoking habits, RestingBP, Cholesterol levels, blood pressure, age, chest pain type, etc.

The initial step in our analysis is to review the dataset's dimensions and shape, then followed by an examination for missing values. Identified missing entries are removed to ensure the integrity and accuracy of our analysis. For example, for some data, we have Cholesterol equal to zero, so we have replaced it with the average value of existing non zero Cholesterol values, to make sure we have consistency across the data.

We apply data preprocessing techniques after data cleansing. This includes label encoding, which converts categorical data into a numerical format. For example, chest pain types are mapped to numerical values based on their severity, facilitating quantitative analysis.

In order to visualize the distribution of data in our dataset, we have conducted exploratory data analysis. There are various tools such as histograms and bar plots that we used to explore and understand both numerical and categorical variables. This helped us to understand the Categorical Data Columns and the Frequency Distribution between them. We used box plot to identify outlier's numerical variables between individuals having and not having heart diseases. Then, we computed a correlation matrix to understand how various attributes are

related to one another, for example chest pain type, cholesterol levels, and the risk of heart disease so we will display the correlation coefficients among different variables [2]. This is a very important tool in statistics and data analysis that helps us to understand the relationships between variables and and also help analyze their linear association. In a correlation matrix, each row and column signify a feature, and the cells of the matrix have the correlation coefficients among pairs of features. We visualize these correlations using a heatmap, that provides us with a clear graphical representation of the data interconnections.

We make use of logistic regression for predictive analysis. Logistic Regression is a statistical technique employed to predict the probability of a binary outcome, relying on one or more predictive variables. So, in our case, we used logistic regression to analyze heart disease risk, as our analysis categorizes the likelihood of individuals developing heart disease into 0 or 1, that is binary format. The method meets our needs because we are predicting our outcome in binary format. The coefficients indicate the impact of each independent variable and the amount of contribution that the variable has towards the outcome. For example, a positive coefficient for age suggests that an increase in age corresponds to a higher risk of cardiovascular disease. Furthermore, the dataset is divided into training and testing sets, and a logistic regression model is trained after scaling our data. Upon testing, the model provided is accurate heart disease prediction with an accuracy of 83%, that indicates that our model is robust and is capable of identifying patients at risk of heart disease.

This methodology allows us to methodically dissect the factors that influence heart disease, helping in the development of effective predictive models and also guiding complex healthcare decisions.

IV. Results And Analysis

4.1. Dataset Collection:

The heart disease dataset is taken from kaggle. This dataset comprises various patient attributes vital for cardiovascular risk assessment. This dataset amalgamates data from five distinct sources, namely Cleveland, Hungarian, Switzerland, Long Beach VA, and staglog (Heart) datasets, totalling 918 unique observations after eliminating duplicates. This dataset contains 12 attributes and multiple characteristics including integers, categorical and real values. Table 2. contains the description of the dataset.

Table 2: Dataset Attributes Description [1]

No.	Features	Description	Value
1.	Age	Age of the patient	Years
2.	Sex	Gender of the patient	Female = False, Male = True
3.	ChestPainType	Type of chest pain experienced by the patient	Atypical Angina(ATA):0, Non-Anginal Pain (NAP): 1 ,Asymptomatic(ASY): 2, Typical Angina (TA): 3
4.	RestingBP	The patient's blood pressure at rest	Integer value
5.	Cholesterol	Serum Cholesterol[1]	Either integer or float value
6.	FastingBS	Fasting Blood sugar > 120 mg/dL [1]	0 = False, 1 = True
7.	RestingECG	ECG when patient is at rest	Normal: 0, ST-T wave abnormality: 1, Left Ventricular Hypertrophy (LVH): 2
8.	MaxHR	Maximum Heart Rate of the patient	Integer value
9.	ExerciseAngina	Distress caused by Exercise	Y = 1, N = 0
10.	OldPeak	Exercise induced ST depression compared to rest [1]	Float value
11.	ST_Slope	Slope of exercise ST segment	Up: 0, Flat: 1 Down: 2
12.	HeartDisease	whether the patient has heart disease or not	0 = False, 1 = True

4.2. Analysis And Results:

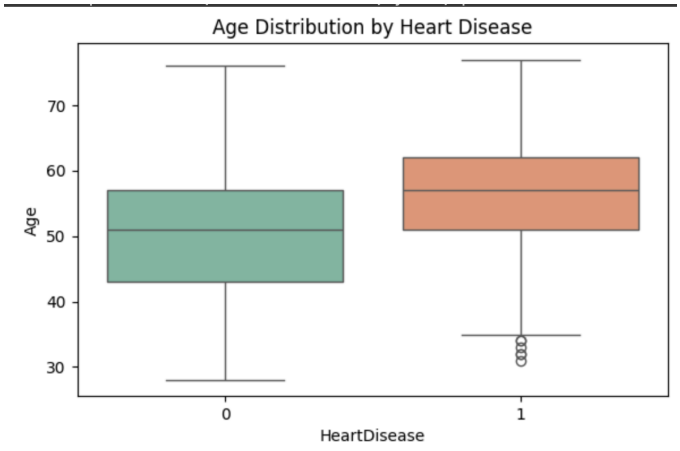
In this section, we have plotted the feature of the heart disease dataset vs. count(distribution count / predictive attribute) for data visualization. We have performed Exploratory Data Analysis (EDA) , which is a technique used for analyzing datasets to summarize their main characteristics, which is frequently accomplished through the use of statistical graphics and other data visualization methods [1].

a) Analyzing Age Distribution

We can see in the figure 4.2.1 that the median age of individuals with heart disease are higher than those without.

This shows that the elderly people have higher chances of getting a heart disease than the young ones. The x-axis denotes whether the individual has heart disease or not and the y-axis denotes age.

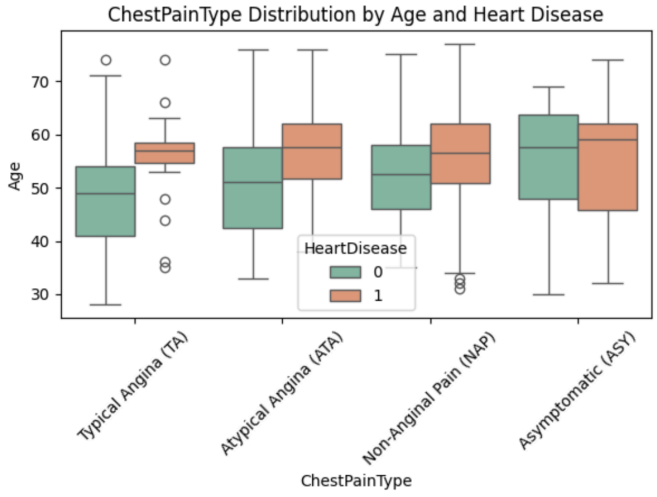
Fig: 4.2.1 Age Distribution By Heart Disease



b) Analyzing Chest Pain

Patients with heart disease may experience chest pain[1]. As we can see in the figure 4.2.2, those with asymptotic chest pain between the ages of 43 and 60 are more likely to experience heart disease followed by Non-anginal Pain and others.

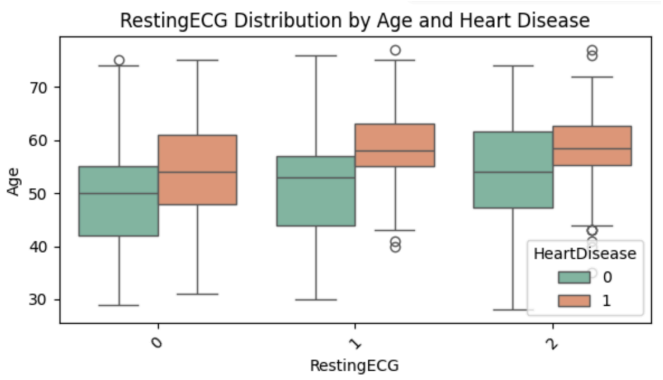
Fig: 4.2.2 Chest Pain Distribution by heart disease



c) Analyzing Resting Electrocardiographic (ECG)

In the figure 4.2.3., x-axis shows the categories of findings observed in resting electrocardiographic (ECGs). These categories denote diagnostic indicators that help in the interpretation of the heart's electrical activity and structure. Specifically, the x-axis denotes three distinct classifications: 'Normal (0)', 'Abnormal(ST-T wave abnormality) (1)', and 'Left Ventricular Hypertrophy (LVH) (2)'. We may infer from the graph that most people's ECGs are normal. Those with abnormalities in their ECG who are older than 55 years of age or older may be more susceptible to heart disease.

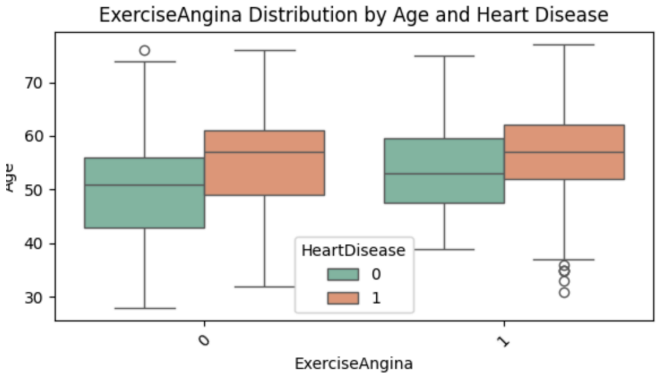
Fig: 4.2.3 Resting ECG distribution by age and heart disease



d) Analyzing Exercise Angina

In the figure 4.2.4., we can see that there is a difference between people who have experienced exercise induced angina (meaning distress) among individuals aged 45 to 60. Some people show a significant distress experienced during physical activity that indicates a potential association with underlying heart disease, a majority of them reported no symptoms.

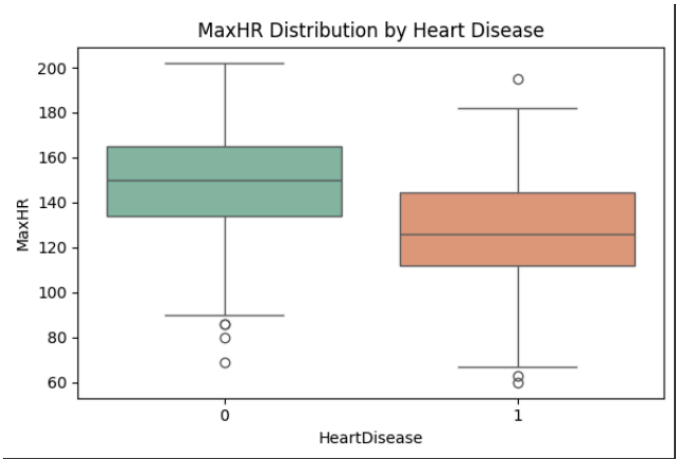
Fig: 4.2.4 Exercise angina distribution by age and heart disease

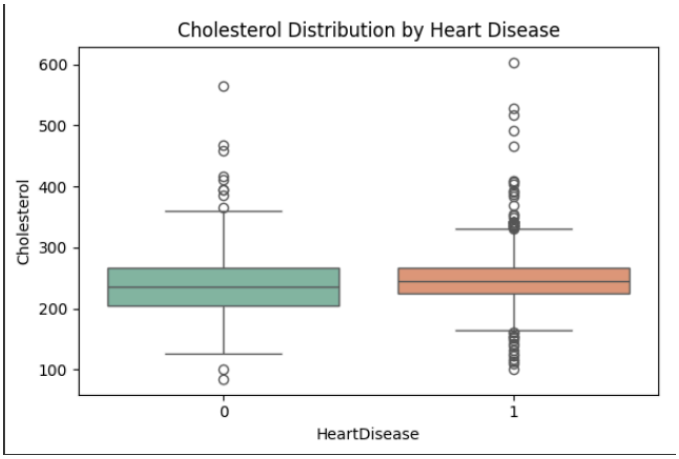


e) Analyzing Maximum Heart Rate

In the figure 4.2.5., we can see the trend of the relationship between maximum heart rate and whether one has a heart disease or not. We can see that individuals having higher maximum heart rates, especially between the range of 135 to 165, show a lower risk of developing a heart disease. Whereas, on the other hand, those who have lower maximum heart rate show higher risk of getting a heart disease, suggesting closer monitoring.

Fig: 4.2.5. MaxHR distribution by Heart Disease

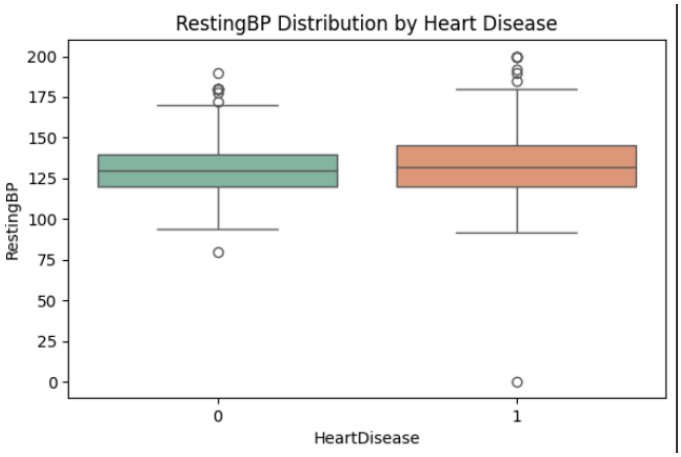




f) Analyzing Resting Blood Pressure

In the figure given below, we can see that the median Blood pressure of approximately 126 or nearby suggests that there is a minimal risk of heart disease, whereas if BP slightly increases, around 130, it shows an increase in the chances of getting the heart disease. The risk range of heart disease appears if BP ranges from 126 to 145.

Fig: 4.2.6. RestingBP Distribution by Heart Disease



g) Analyzing Cholesterol

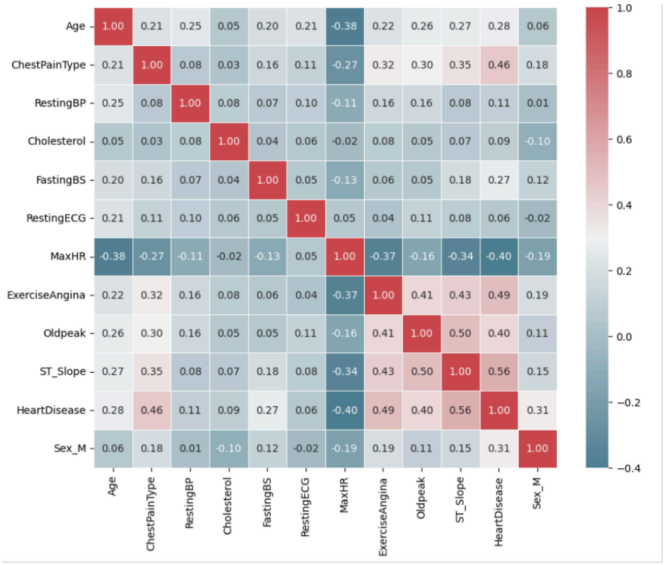
In the figure 4.2.7., we can see that people that do not have heart disease show a median cholesterol level of approximately 240. On the other hand those who have heart disease show a higher median cholesterol level, around 250. This suggests that if the cholesterol level of an individual increases there are high chances of them getting a heart disease.

Fig: 4.2.7. Cholesterol Distribution By Heart Disease

4.3 *Correlation Matrix*

Correlation is a statistical feature that describes the strength and route of a linear relationship among two quantitative variables[1].In our analysis, there are several key factors that show a notable relationship with the risk of getting a heart disease. We can see that age shows a positive correlation (0.282) with heart disease, indicating that as individuals get old, their risk of having a heart disease tends to increase. Additionally, chest pain type (0.459), exercise-induced angina (0.494), and ST slope (0.559) also shows a strong positive correlations with heart disease, that suggests individuals that experience certain types of chest pain or distress during exercise, as well as those with specific ST segment slopes, may be at higher risk of developing heart disease. On the other hand, maximum heart rate (MaxHR) shows a negative correlation (-0.4000) with heart disease, which states that individuals with lower maximum heart rates may have a higher risk. These observations demonstrate the nature of heart disease risk, where various demographic and clinical factors contribute to overall chances of getting the disease.

Fig: 4.3.1. Correlation Matrix



4.4 *Prediction and Evaluation Metric*

Logistic Regression is a statistical method used for binary classification tasks that makes it suitable for prediction of heart disease based on the patient attributes[1]. In this section, we evaluate the accuracy of our logistic regression model in predicting heart disease risk, aiming to provide insights into the model's performance and its utility in clinical decision-making. The accuracy is calculated using "TP - True Positive, TN - True Negative, FN - False Negative and FP - False Positive," rate[1]. All true positives and negatives predictions are split into all positive and negative predictions[1].

Accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \qquad [1]$$

Our logistic regression model achieved an accuracy of 82.61% on the testing data set, indicating its efficacy in accurately predicting heart disease risk.

V. Conclusions and Future Scope

The model that we used in this study helped us gain insights on assessing the risk of heart disease based on various attributes of the patient. Age, High Cholesterol Levels, Excess consumption of alcohol, elevated blood sugar levels, severe chest pain, and high resting blood pressure (BP) are all critical indicators of higher risk of heart problems that might require immediate medical attention. Conversely, having lower levels of cholesterol, young age, low blood pressure and higher maximum heart rates are possible indicators of a healthy heart that suggests the patient will be at a lower risk of having a heart disease. Our findings indicate that attributes such as Chest Pain Type, ST_Slope and OldPeak are substantial contributors towards heart disease as they have higher correlation values. These attributes serve clinicians as essential information for risk assessment, diagnosis and helping the patients with suspected heart diseases. Incorporating these methods into the healthcare decision making process can help to enhance the patients health, optimize resources and overall improve the healthcare system.

The future scope for this research would be to gather more data in order to improve our prediction accuracy. We could also expand our model to incorporate environmental factors such as diet, physical activity, family history in order to improve our prediction. Poor mental health and sleep patterns have also been independently linked towards heart diseases and hence these factors should also be considered in the future. Moreover, we could monitor an individual's healthcare access, how frequently they use medical services that would be beneficial in early detection of heart disease.

References

[1] Hassan CAU, Iqbal J, Irfan R, Hussain S, Algarni AD, Bukhari SSH, Alturki N, Ullah SS. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors (Basel). 2022 Sep 23;22(19):7227. doi: 10.3390/s22197227. PMID: 36236325; PMCID: PMC9573101.

[2] M. B. Dkhil, B. Rabbouch and F. Saadaoui, "Risk Factor Prediction of Heart Disease using Machine Learning Approaches," 2023 International Conference on Cyberworlds (CW), Sousse, Tunisia, 2023, pp. 298-305, doi: 10.1109/CW58918.2023.00053.

[3] Johnson, K, Torres Soto, J, Glicksberg, B. et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 2018 Jun, 71 (23) 2668–2679.

[4] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[5] https://towardsdatascience.com/how-to-choose-the-best-evaluation-metric-for-classification-problems-638e845da334

[6] https://lost-stats.github.io/Presentation/Figures/heatmap_colored_correlation_matrix.html

[7] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[8] https://en.wikipedia.org/wiki/Logistic_regression

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9573101/

[10] https://www.heart.org/en/health-topics/high-blood-pressure/why-high-blood-pressure-is-a-silent-killer/know-your-risk-factors-for-high-blood-pressure

Hardware Specifications:

1. Macbook Air (M1 chip, 13 inch)
2. Macbook Air (M2 chip, 13 inch)
3. Dell inspiron (Windows, 12th Gen Intel(R) Core(TM), i7)

Software Specifications:

1. Google Colab (jupyter notebook).
2. Python programming language.
3. Python libraries: matplotlib, seaborn, pandas, sklearn.

Appendix

#Python code

```
# Import modules
import pandas as pd
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split

# Read Dataset and Print it
df = pd.read_csv('heart.csv')
print(f"Dataset: \n\n{df.head()}")
print(f"\n\nShape of the dataset: {df.shape}")
print(f"\n\nTotal Null Values in dataset{df.isnull().sum()}")

# Replacing the zero - value of Cholestrol coulmn with the
average
print("Total zero-values: ",(df['Cholesterol'] == 0).sum())

cholesterol = df['Cholesterol']

# Calculate the average of non-zero values
non_zero_cholesterol = cholesterol[cholesterol != 0]
average_cholesterol = non_zero_cholesterol.mean()
```

```
df['Cholesterol'] = df['Cholesterol'].replace(0,
average_cholesterol)
# Printing Uniques Values of categorical data columns And
counting for each of them
print("Categorical Data Columns & Frequency Distribution:
\n")

# 1. Sex Column:
print("Sex: ")
print("  - Male (M):", (df['Sex'] == 'M').sum())
print("  - Female (F):", (df['Sex'] == 'F').sum())
print()

# 2. Chest Paint Type Column:
print("Chest Pain Type: ")
print("  - Typical Angina (TA):", (df['ChestPainType'] ==
'TA').sum())
print("  - Atypical Angina (ATA):", (df['ChestPainType'] ==
'ATA').sum())
print("  - Non-Anginal Pain (NAP):", (df['ChestPainType'] ==
'NAP').sum())
print("  - Asymptomatic (ASY):", (df['ChestPainType'] ==
'ASY').sum())
print()

# 3. Resting ECG Type Column:
print("Resting ECG: ")
print("  - Normal:", (df['RestingECG'] == 'Normal').sum())
print("  - Abnormal (ST-T wave abnormality):",
(df['RestingECG'] == 'ST').sum())
print("  - Left Ventricular Hypertrophy (LVH):",
(df['RestingECG'] == 'LVH').sum())
print()

# 4. Exercise Angina Column:
print("Exercise-Induced Angina: ")
print("  - No:", (df['ExerciseAngina'] == 'N').sum())
print("  - Yes:", (df['ExerciseAngina'] == 'Y').sum())
print()

# 5. ST Slope Column:
print("ST Segment Slope: ")
print("  - Upsloping:", (df['ST_Slope'] == 'Up').sum())
print("  - Flat:", (df['ST_Slope'] == 'Flat').sum())
print("  - Downsloping:", (df['ST_Slope'] == 'Down').sum())


# Performing Label Encoding to convert the categorical data
columns:

# Perform label encoding for Sex column
df = pd.get_dummies(df, columns=['Sex'], drop_first=True)

# Map chest pain types to numerical values based on severity
# TA (Typical Angina): Common heart-related chest pain.
# ATA (Atypical Angina): Chest discomfort that's not typical
angina.
# NAP (Non-Anginal Pain): Pain not related to the heart.ASY
(Asymptomatic): Absence of chest pain, but hints at potential
heart issues.

chest_pain_map = {'TA': 3, 'ATA': 0, 'NAP': 1, 'ASY': 2}
df['ChestPainType'] =
df['ChestPainType'].map(chest_pain_map)

# Map resting electrocardiogram results to numerical values
based on severity
resting_ecg_map = {'Normal': 0, 'ST': 1, 'LVH': 2}
df['RestingECG'] = df['RestingECG'].map(resting_ecg_map)

# Map ExerciseAngina to numerical values
exercise_angina_map = {'Y': 1, 'N': 0}
df['ExerciseAngina'] =
df['ExerciseAngina'].map(exercise_angina_map)

# Map ST_Slope to numerical values based on severity
st_slope_map = {'Up': 0, 'Flat': 1, 'Down': 2}
df['ST_Slope'] = df['ST_Slope'].map(st_slope_map)


# Data Visualization

# Histograms for numerical variables
plt.figure(figsize=(12, 8))
for i, col in enumerate(['Age', 'RestingBP', 'Cholesterol',
'MaxHR']):
    plt.subplot(2, 2, i+1)
    sns.histplot(df[col], kde=True)
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()

# Bar plots for categorical variables
plt.figure(figsize=(12, 8))
for i, col in enumerate(['ChestPainType', 'RestingECG',
'ExerciseAngina', 'ST_Slope']):
    plt.subplot(2, 2, i+1)
    sns.countplot(data=df, x=col, palette='Set2')
    plt.title(f'Frequency of {col}')
plt.tight_layout()
plt.show()

plt.figure(figsize=(18, 12))

# Box plots for numerical variables
numerical_variables = ['Age', 'RestingBP', 'Cholesterol',
'MaxHR']
for i, col in enumerate(numerical_variables):
    plt.subplot(2, 4, i+1)
    sns.boxplot(data=df, x='HeartDisease', y=col,
palette='Set2')
    plt.title(f'{col} Distribution by Heart Disease')
```

```
# Box plots for categorical variables
categorical_variables = ['ChestPainType', 'RestingECG',
'ExerciseAngina', 'ST_Slope']
for i, col in enumerate(categorical_variables):
    plt.subplot(2, 4, i+len(numerical_variables)+1)
    sns.boxplot(data=df, x=col, y='Age', hue='HeartDisease',
palette='Set2')
    plt.title(f'{col} Distribution by Age and Heart Disease')

plt.tight_layout()
plt.show()


# Compute Correlation Matrix:
corr_matrix = df.corr(method = 'pearson')
print("\n-------Correlation Matrix------\n\n")
print(corr_matrix)

# Correlation Graph Plot:
fix, ax = plt.subplots(figsize = (10,8))
plt.title("Correlation Coefficient Heatmap \n", fontsize = 20)
print()
cmap = sns.diverging_palette(220, 10, as_cmap=True)
# Generating heat maps:
heatmap = sns.heatmap(corr_matrix, annot = True,
annot_kws={"fontsize": 10}, fmt = '.2f',
                      linewidths = 0.5, cmap=cmap)
plt.show()

# Separate features and target
X = df.drop(columns=['HeartDisease'])
y = df['HeartDisease']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

from sklearn.preprocessing import StandardScaler

# Create a scaler
scaler = StandardScaler()

# Fit and transform the training data
X_train_scaled = scaler.fit_transform(X_train)

# Transform the test data using the same scaler
X_test_scaled = scaler.transform(X_test)


# Train a Logistic Regression model on scaled data
model = LogisticRegression()
model.fit(X_train_scaled, y_train)

# Make predictions
y_pred = model.predict(X_test_scaled)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred) * 100
print(f'Accuracy: {accuracy:.2f}%')
```