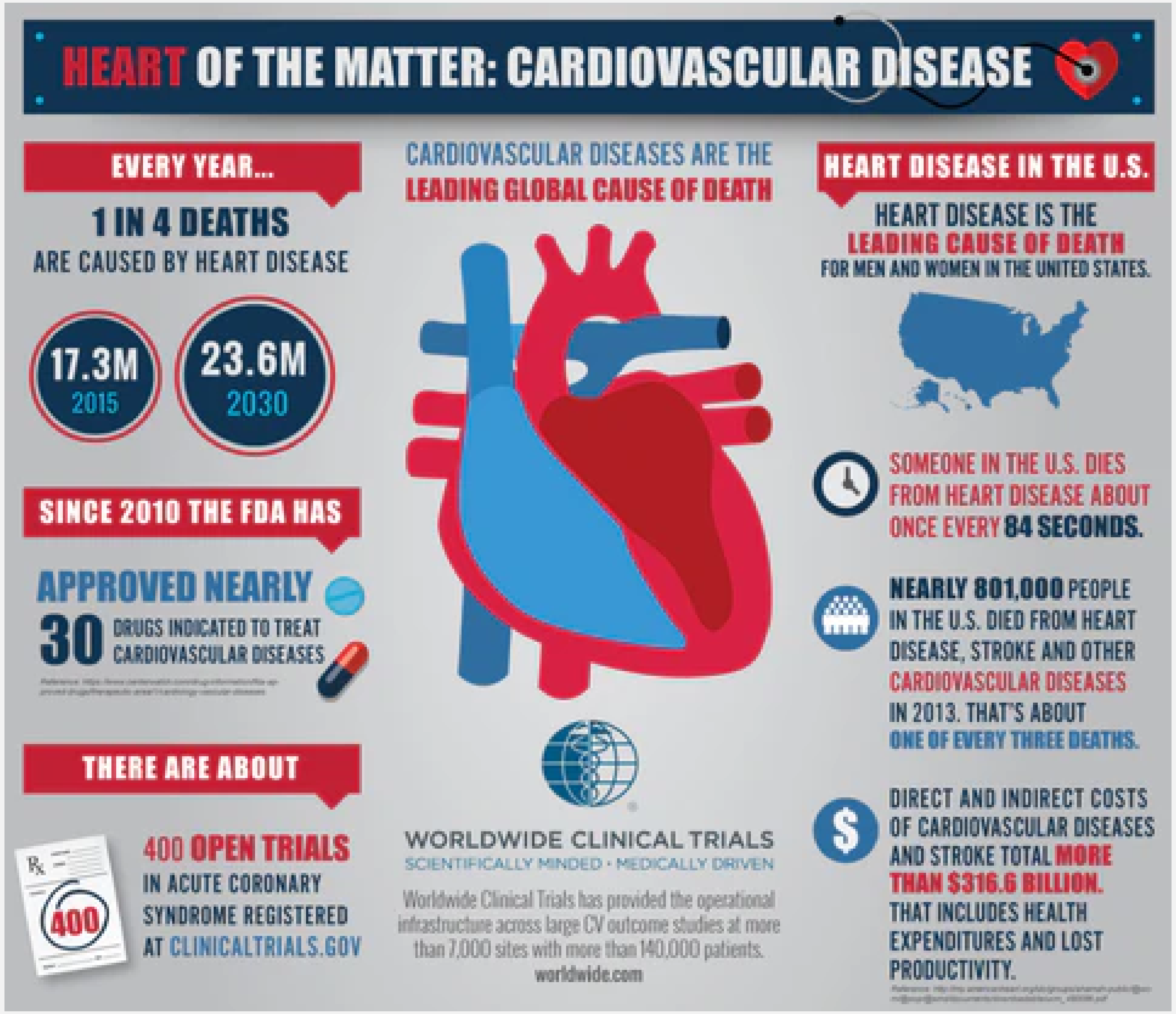
Understanding Heart Health

Data Analysis on Heart

INSIGHTS FOR A HEALTHIER LIFE

SAI VINAY BHIMANA, PREM MALDE, RASIKA PANDE, DEVIKA RAHEJA



HEART OF THE MATTER: CARDIOVASCULAR DISEASE

EVERY YEAR...

1 IN 4 DEATHS ARE CAUSED BY HEART DISEASE

17.3M 2015

23.6M 2030

SINCE 2010 THE FDA HAS

APPROVED NEARLY 30 DRUGS INDICATED TO TREAT CARDIOVASCULAR DISEASES

THERE ARE ABOUT

400 OPEN TRIALS IN ACUTE CORONARY SYNDROME REGISTERED AT CLINICALTRIALS.GOV

CARDIOVASCULAR DISEASES ARE THE LEADING GLOBAL CAUSE OF DEATH

WORLDWIDE CLINICAL TRIALS

SCIENTIFICALLY MINDED • MEDICALLY DRIVEN

Worldwide Clinical Trials has provided the operational infrastructure across large CV outcome studies at more than 7,000 sites with more than 140,000 patients.

worldwide.com

HEART DISEASE IN THE U.S.

HEART DISEASE IS THE LEADING CAUSE OF DEATH FOR MEN AND WOMEN IN THE UNITED STATES.

SOMEONE IN THE U.S. DIES FROM HEART DISEASE ABOUT ONCE EVERY 84 SECONDS.

NEARLY 801,000 PEOPLE IN THE U.S. DIED FROM HEART DISEASE, STROKE AND OTHER CARDIOVASCULAR DISEASES IN 2013. THAT'S ABOUT ONE OF EVERY THREE DEATHS.

DIRECT AND INDIRECT COSTS OF CARDIOVASCULAR DISEASES AND STROKE TOTAL MORE THAN $316.6 BILLION. THAT INCLUDES HEALTH EXPENDITURES AND LOST PRODUCTIVITY.

INTRODUCTION

Globally, cardiovascular diseases (CVDs) are the leading cause of death, claiming an estimated 17.9 million lives annually, accounting for 32% of all global deaths. Of these, 85% are due to heart attacks and strokes.

WHO BENEFITS FROM THIS DATA?

Healthcare Providers

Pharmaceutical Companies

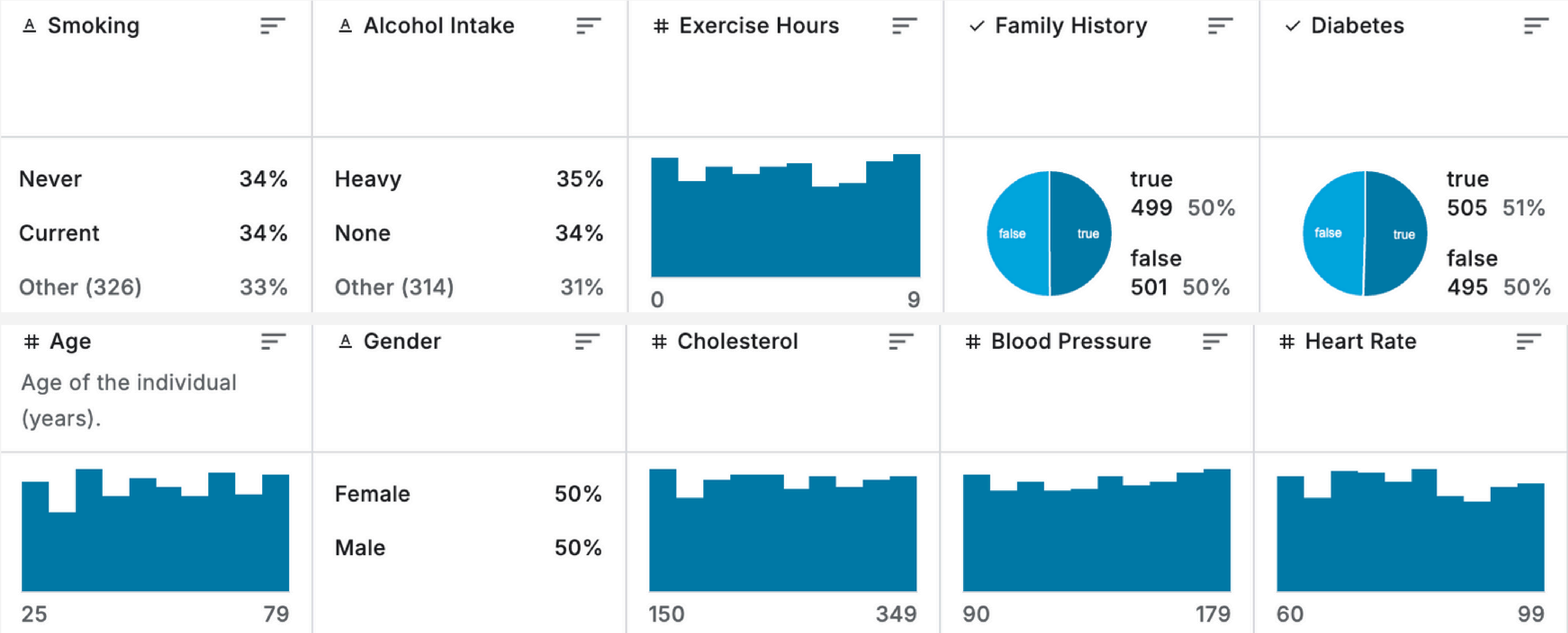
Insurance Companies

Individuals

DATASET OVERVIEW

Numerical: Age, Cholesterol, Blood Pressure, Heart Rate, Exercise Hours, Blood Sugar

Categorical: Gender, Smoking, Alcohol Intake, Family History, Diabetes, Obesity, Stress Level, Exercise-Induced Angina, Chest Pain Type, Heart Disease



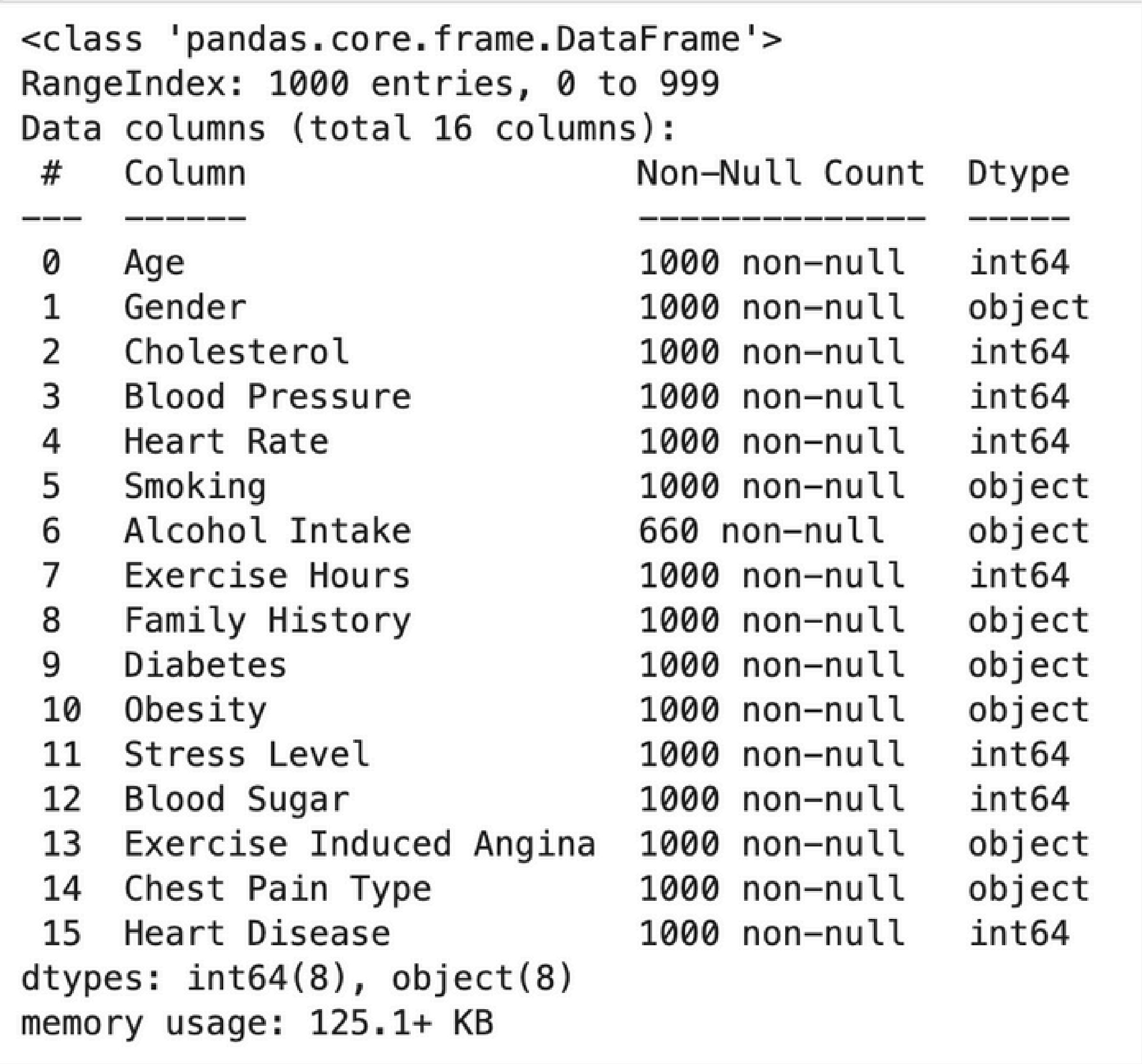
A Smoking	A Alcohol Intake	# Exercise Hours	✓ Family History	✓ Diabetes
Never 34% Current 34% Other (326) 33%	Heavy 35% None 34% Other (314) 31%	0 9	true 499 50% false 501 50%	true 505 51% false 495 50%

# Age	A Gender	# Cholesterol	# Blood Pressure	# Heart Rate
Age of the individual (years).				
25 79	Female 50% Male 50%	150 349	90 179	60 99

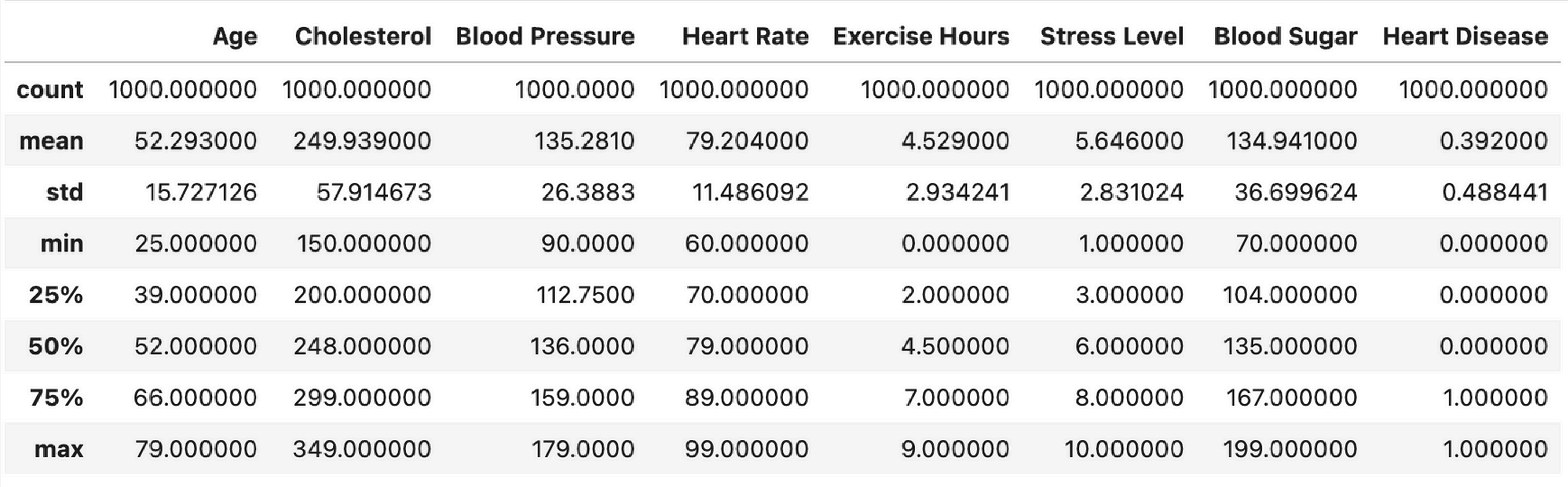
DATA UNDERSTANDING

Shape
The shape is 1000,16

Null Values
“None’ being interpreted as null values in ‘Alcohol Intake’

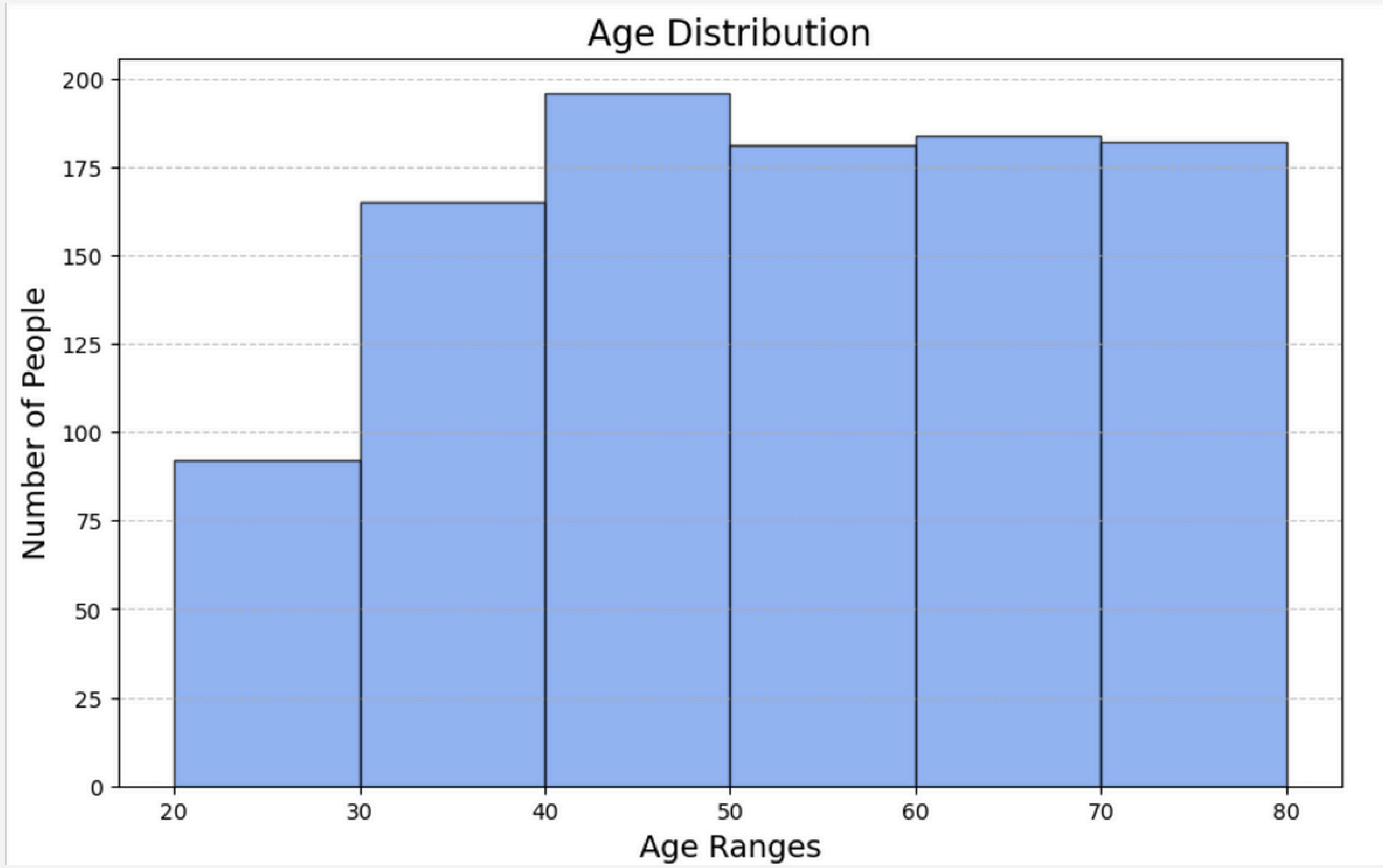


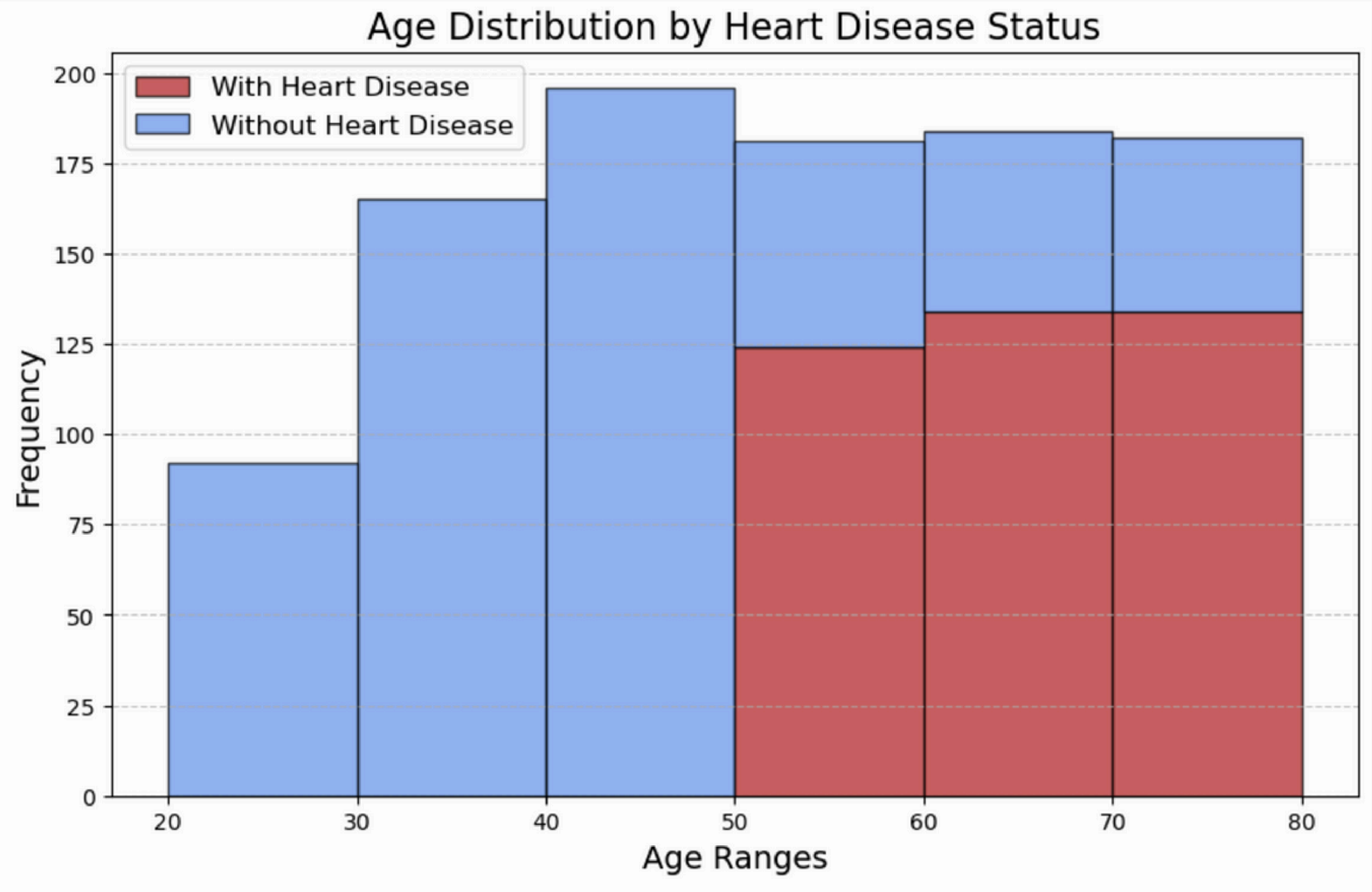
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Age                      1000 non-null   int64 
 1   Gender                   1000 non-null   object
 2   Cholesterol              1000 non-null   int64 
 3   Blood Pressure           1000 non-null   int64 
 4   Heart Rate               1000 non-null   int64 
 5   Smoking                  1000 non-null   object
 6   Alcohol Intake           660 non-null    object
 7   Exercise Hours           1000 non-null   int64 
 8   Family History           1000 non-null   object
 9   Diabetes                 1000 non-null   object
 10  Obesity                  1000 non-null   object
 11  Stress Level             1000 non-null   int64 
 12  Blood Sugar              1000 non-null   int64 
 13  Exercise Induced Angina  1000 non-null   object
 14  Chest Pain Type          1000 non-null   object
 15  Heart Disease            1000 non-null   int64 
dtypes: int64(8), object(8)
memory usage: 125.1+ KB
```



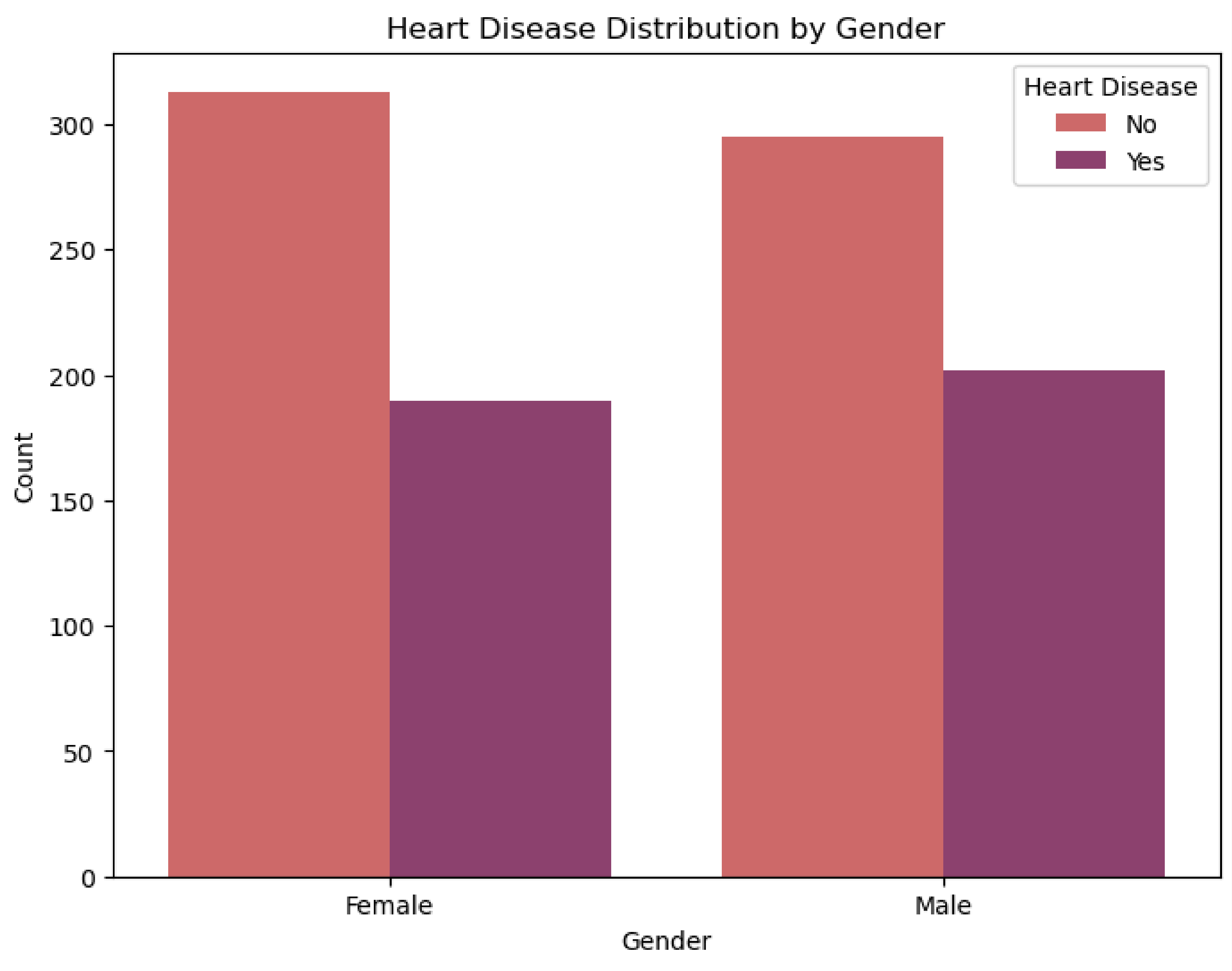
	Age	Cholesterol	Blood Pressure	Heart Rate	Exercise Hours	Stress Level	Blood Sugar	Heart Disease
count	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	52.293000	249.939000	135.2810	79.204000	4.529000	5.646000	134.941000	0.392000
std	15.727126	57.914673	26.3883	11.486092	2.934241	2.831024	36.699624	0.488441
min	25.000000	150.000000	90.0000	60.000000	0.000000	1.000000	70.000000	0.000000
25%	39.000000	200.000000	112.7500	70.000000	2.000000	3.000000	104.000000	0.000000
50%	52.000000	248.000000	136.0000	79.000000	4.500000	6.000000	135.000000	0.000000
75%	66.000000	299.000000	159.0000	89.000000	7.000000	8.000000	167.000000	1.000000
max	79.000000	349.000000	179.0000	99.000000	9.000000	10.000000	199.000000	1.000000

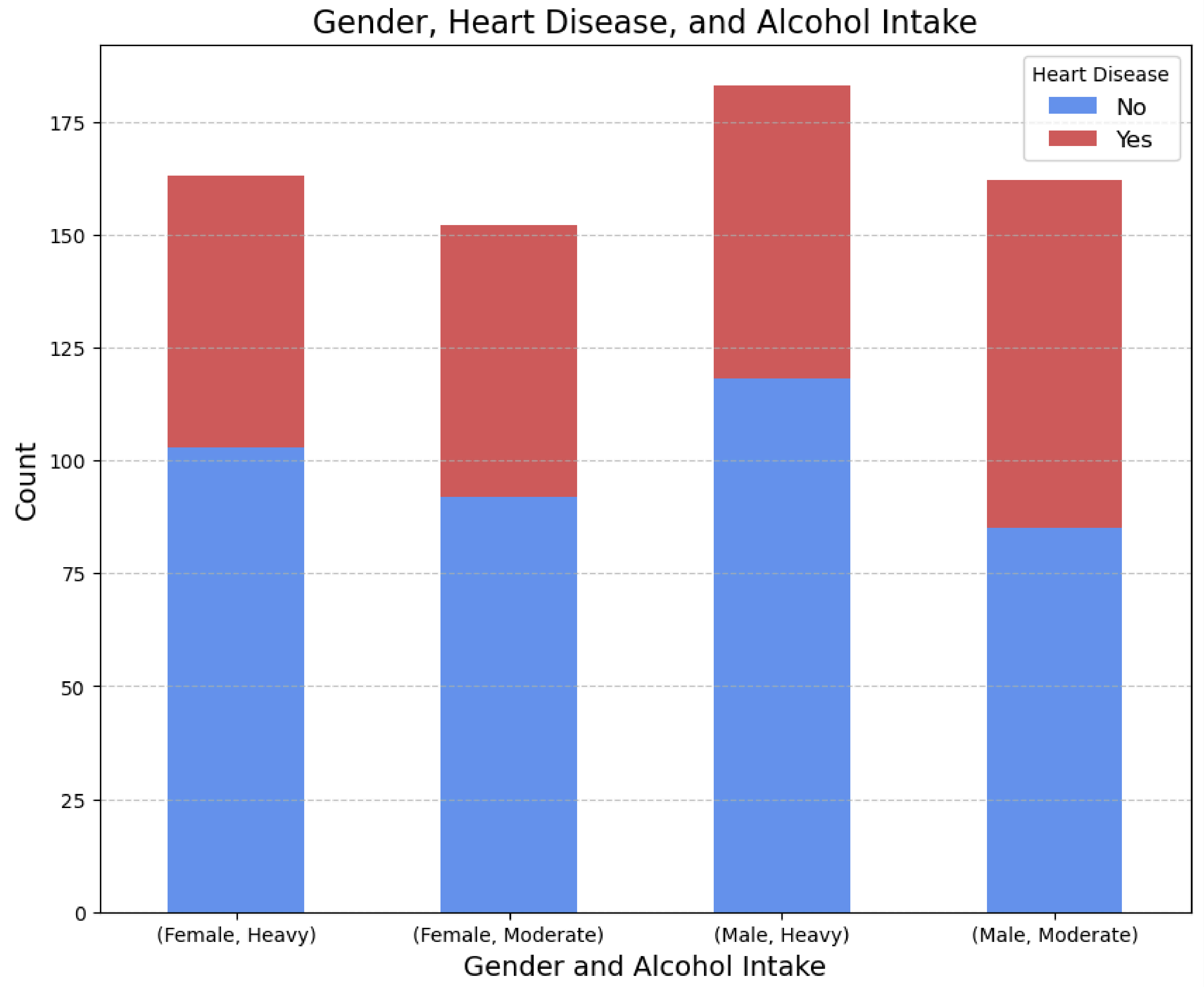
EXPLORING AGE





EXPLORING GENDER

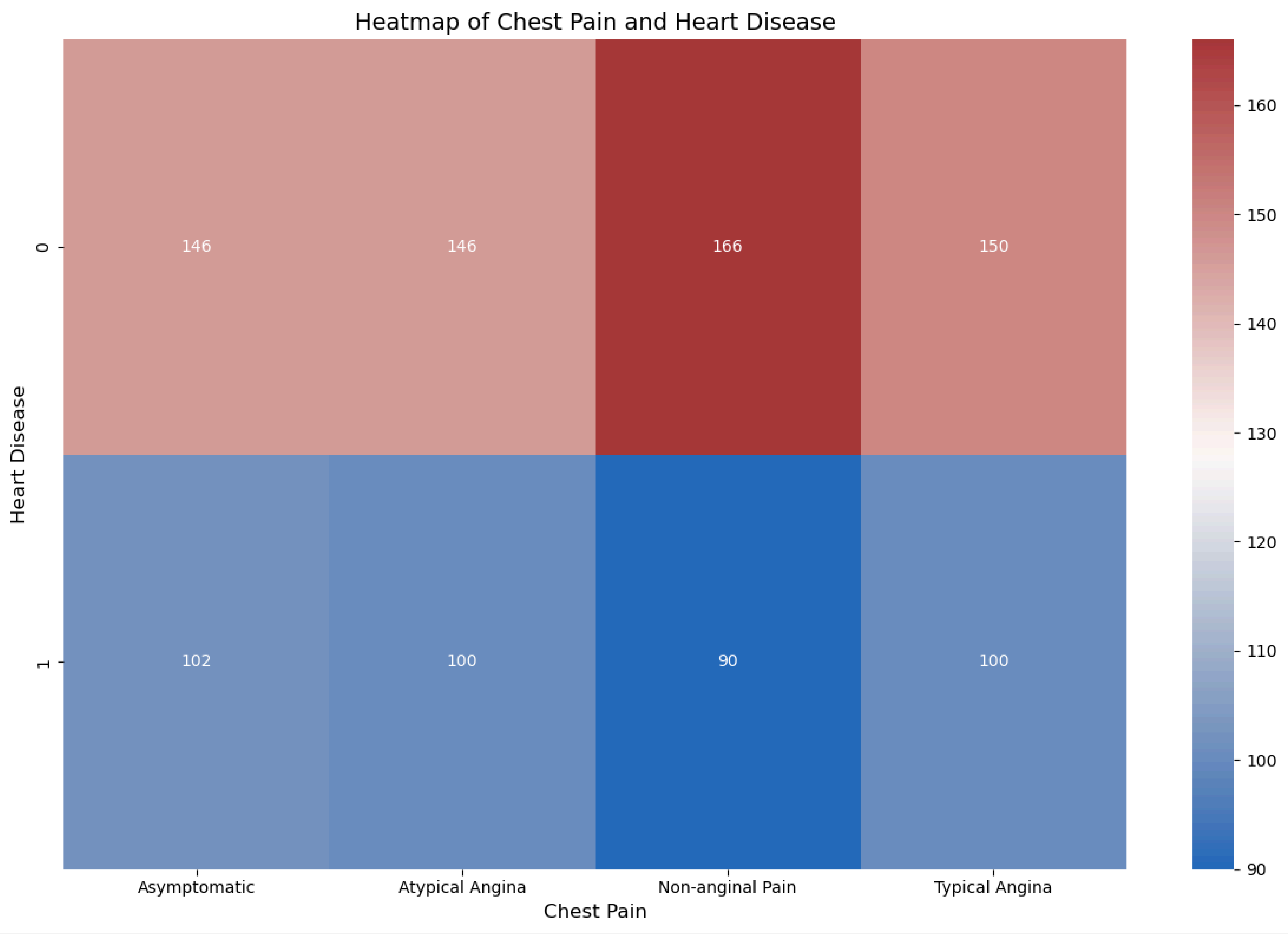




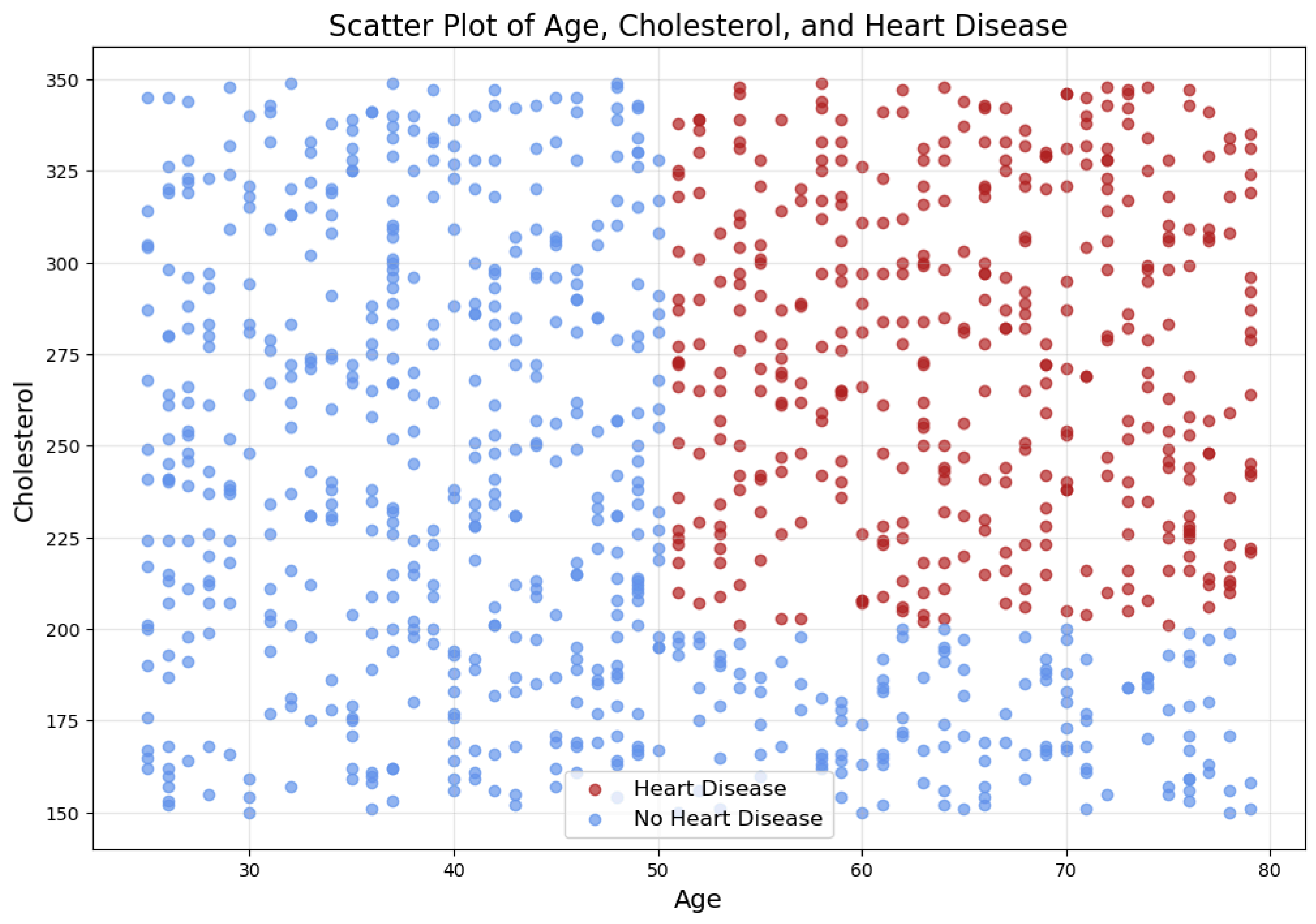
EXPLORING CHEST PAIN

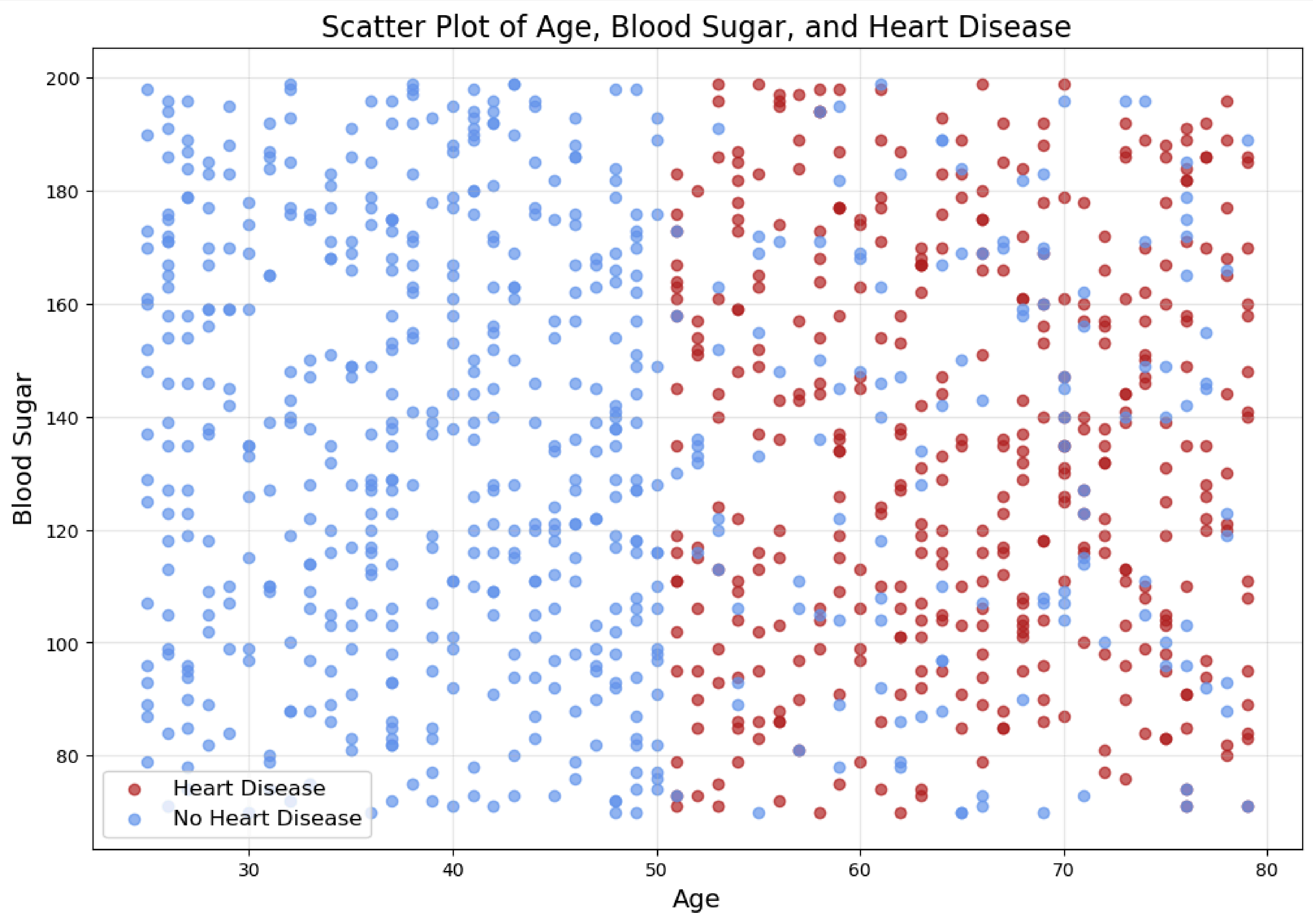
- Asymptomatic and Atypical Angina have relatively balanced distributions of heart disease cases, which may warrant deeper analysis into other co-factors.

- Typical Angina, while a classic symptom, does not have the strongest association with heart disease in this dataset, suggesting potential underdiagnosis in cases with other pain types.

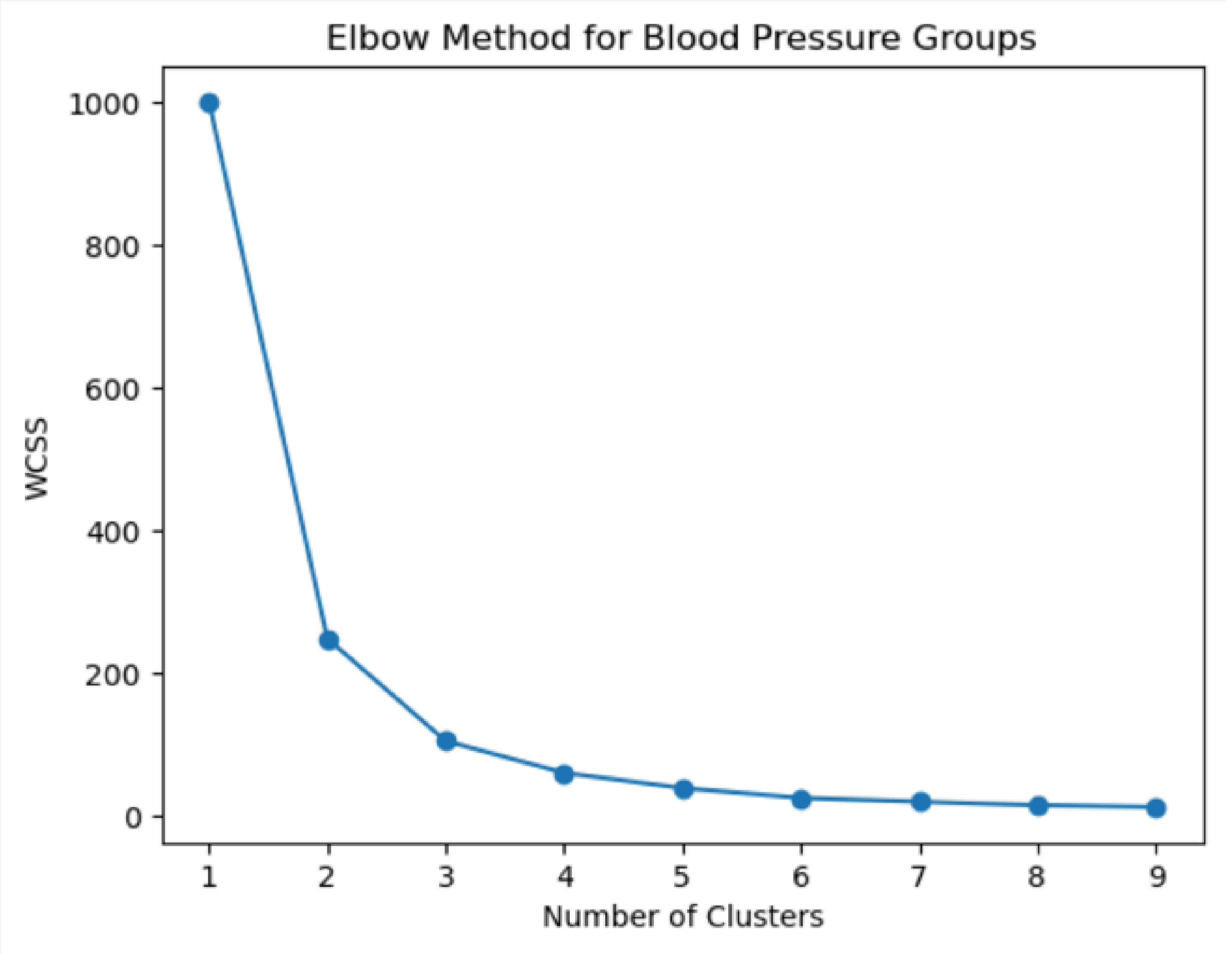


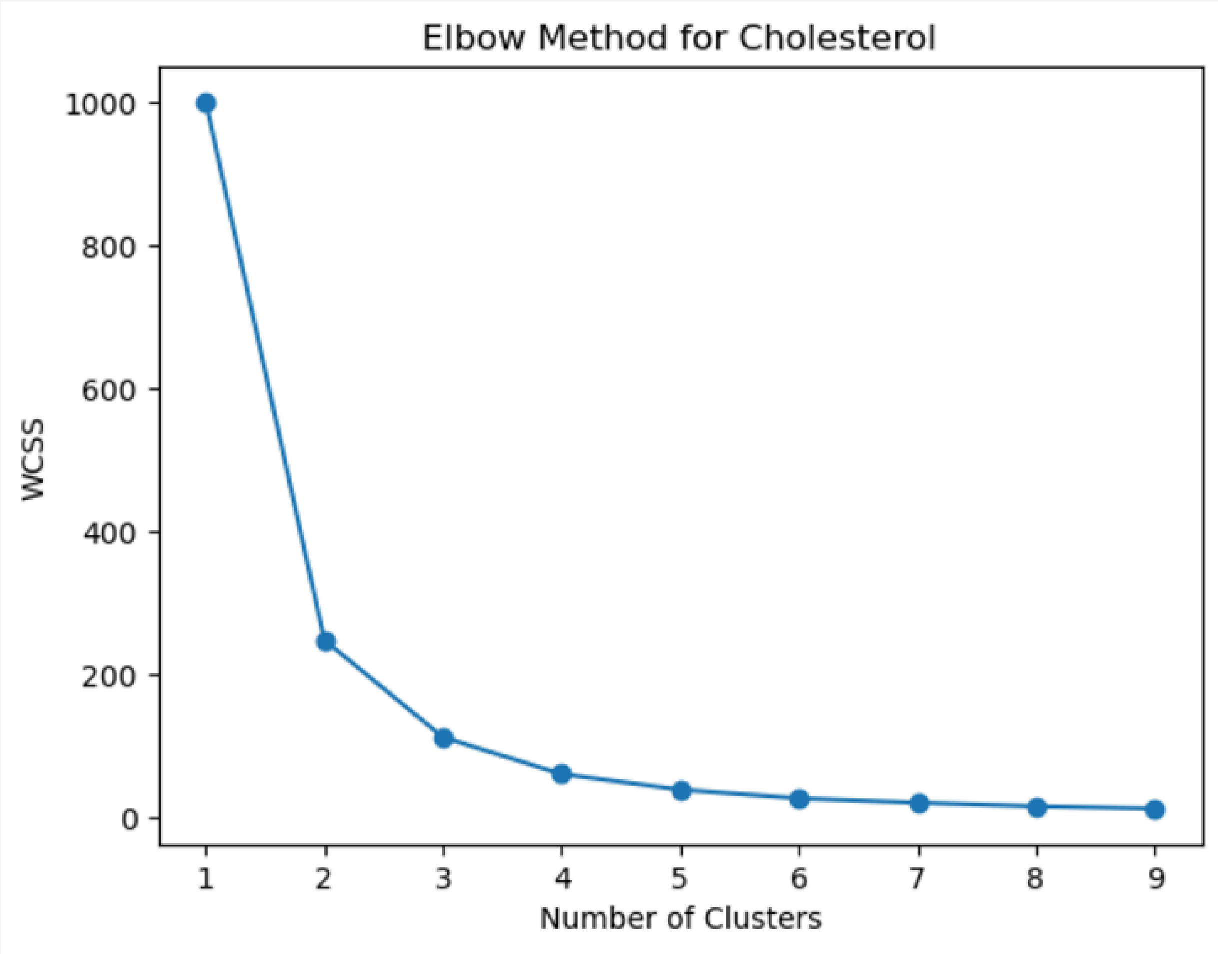
EXPLORING CHOLESTEROL & BLOOD SUGAR

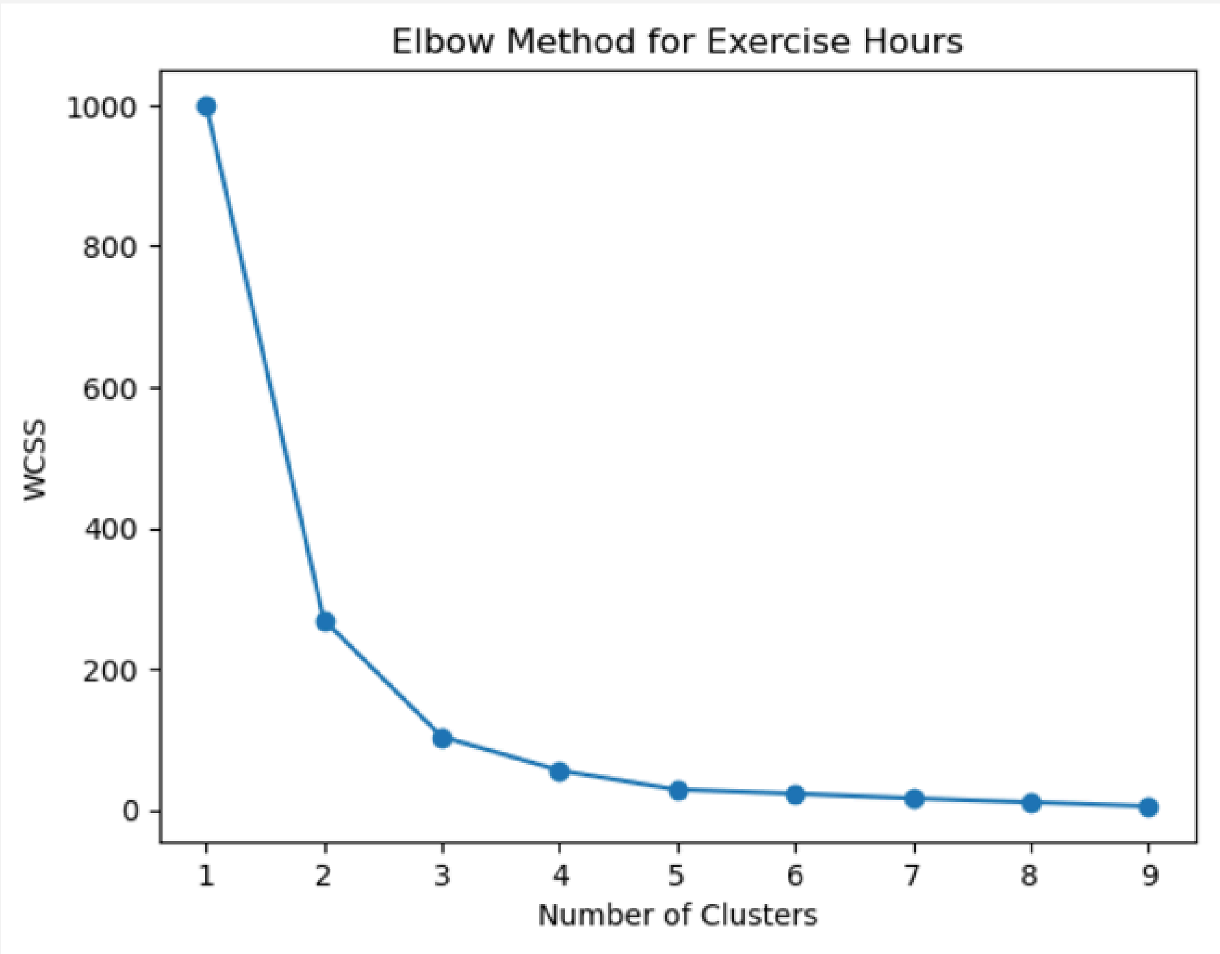




CLUSTER ANALYSIS

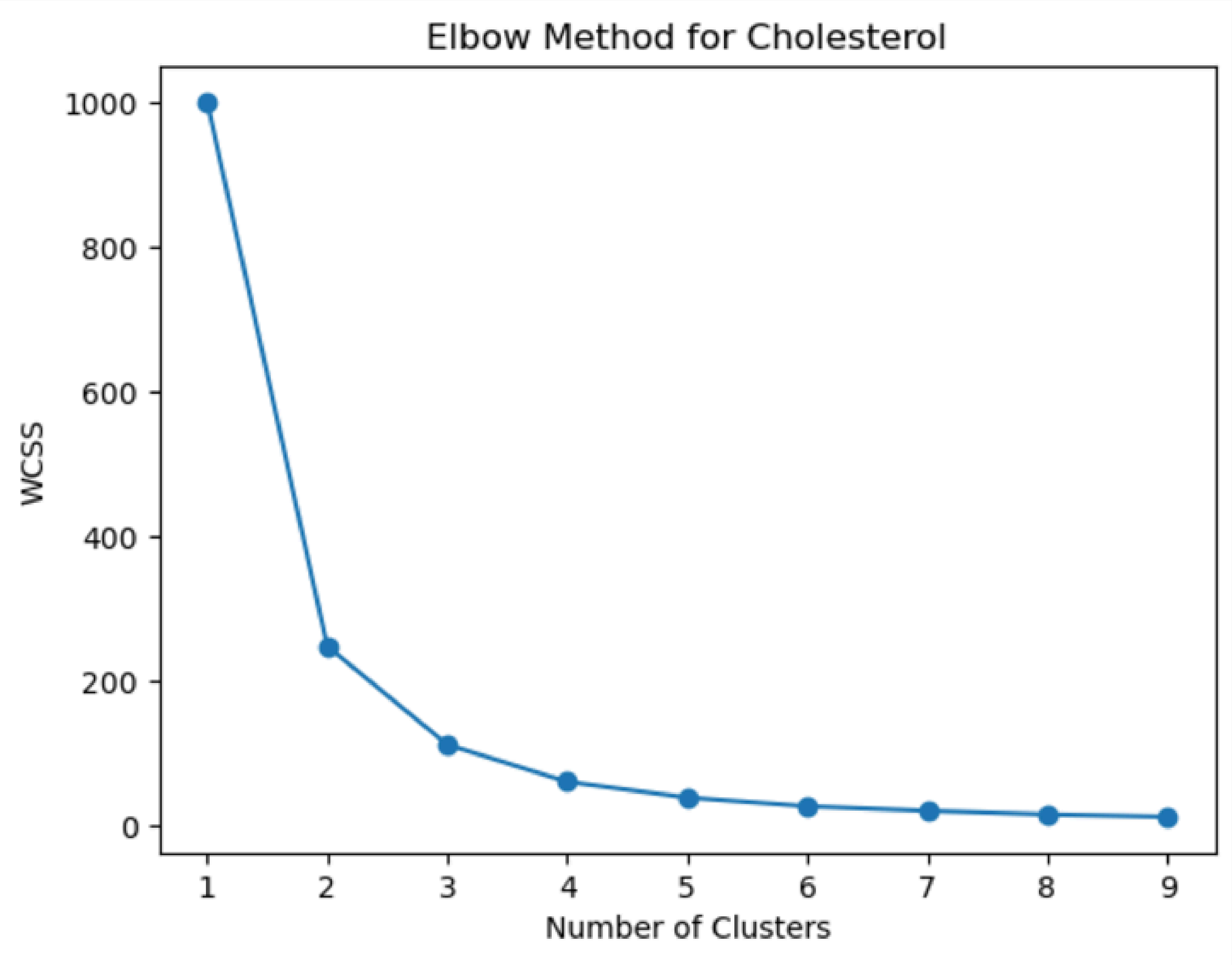

Elbow Method for Blood Pressure Groups

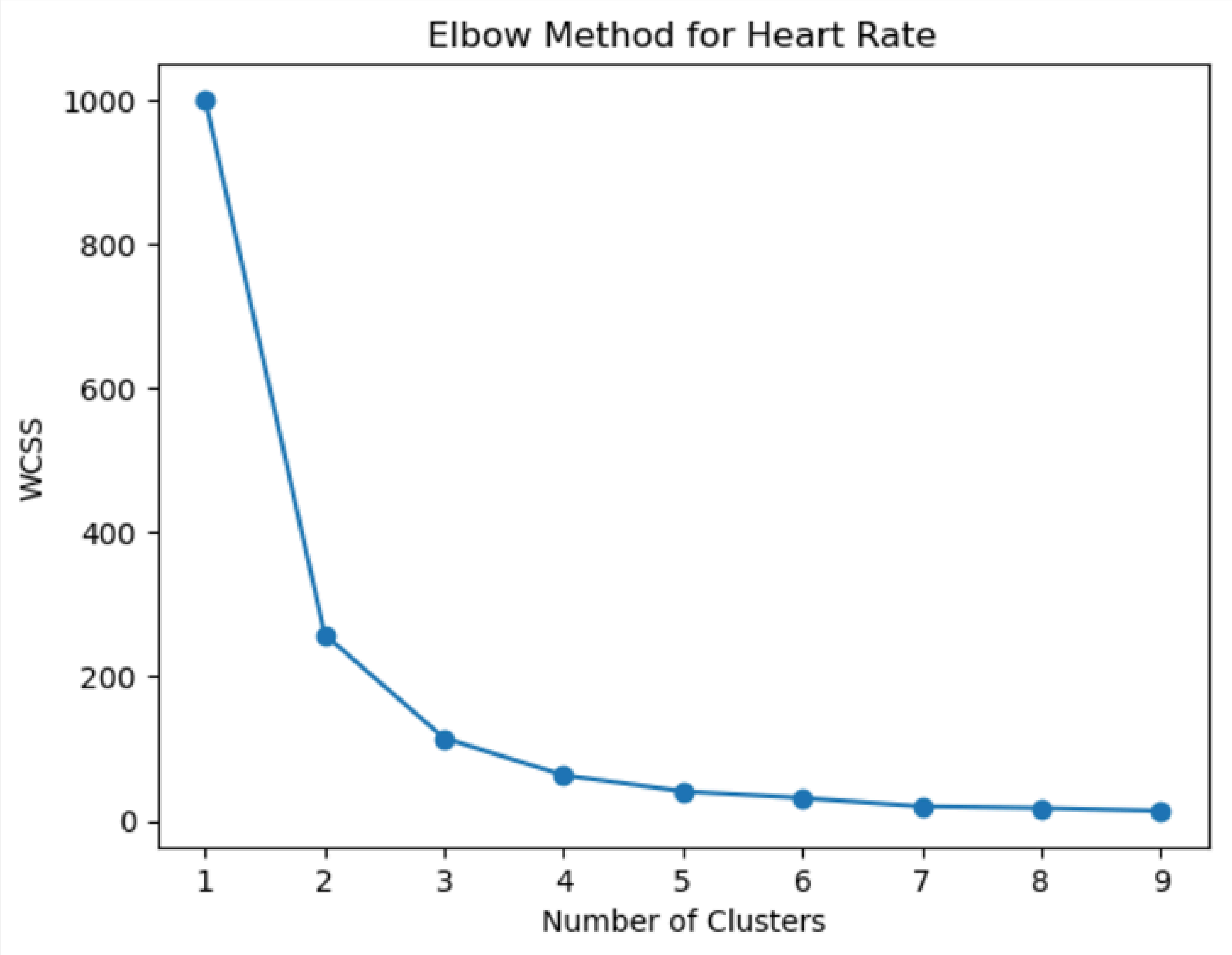

Elbow Method for Cholesterol


Elbow Method for Exercise Hours

- Elbow Curve forms when number of clusters are 3 and 4
- Same trend was observed for almost all the variables affecting heart diseases

CLUSTER ANALYSIS



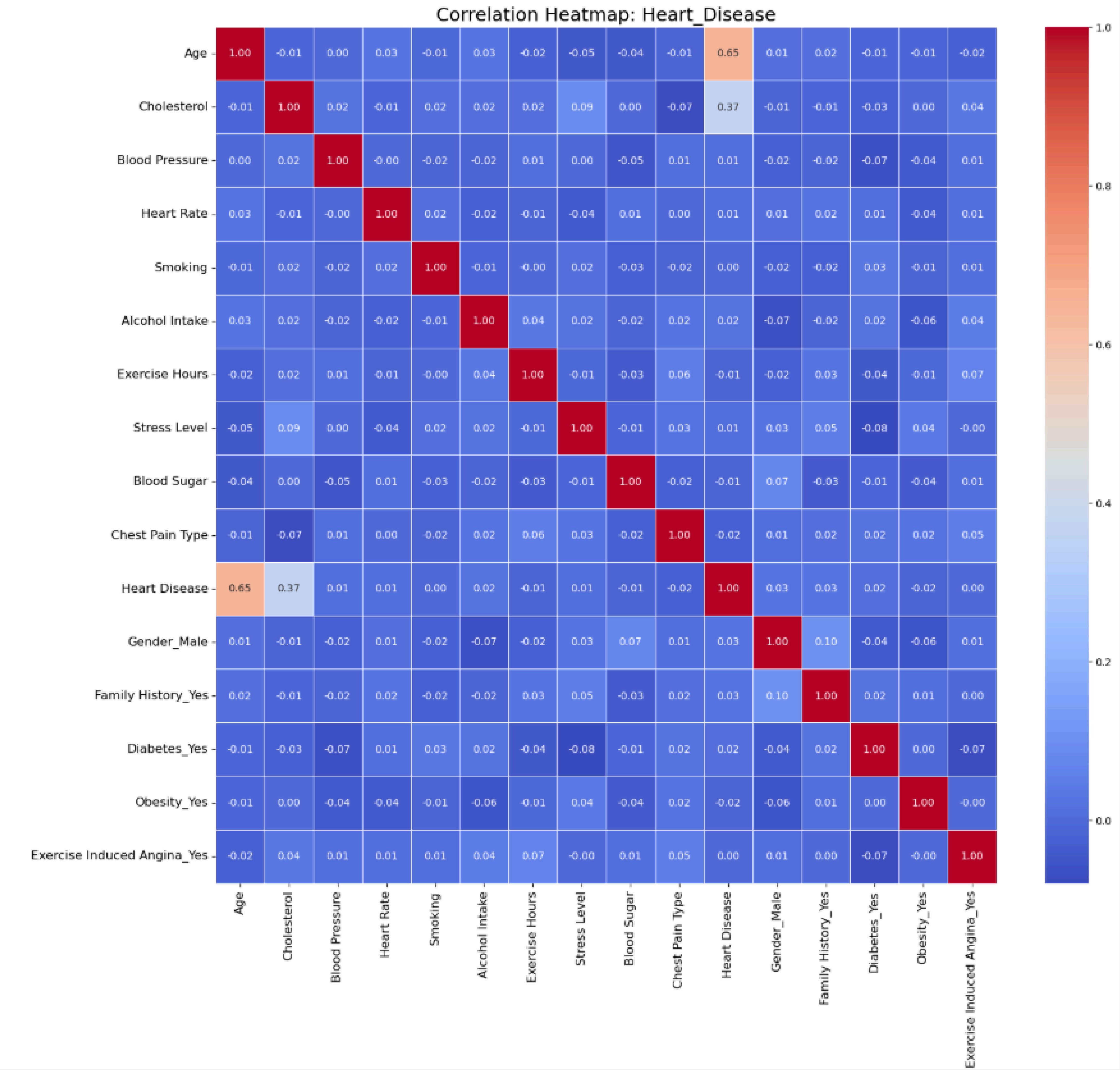


Cholesterol Levels	Heart disease patient count
150-218	47
219-285	171
286-349	174

Heart Rate Ranges	Heart disease patient count
60-73	138
74-87	141
88-99	113

HEATMAP

- Visual representation on how various features correlate with heart disease
- No Multicollinearity



Correlation Heatmap: Heart_Disease

	Age	Cholesterol	Blood Pressure	Heart Rate	Smoking	Alcohol Intake	Exercise Hours	Stress Level	Blood Sugar	Chest Pain Type	Heart Disease	Gender_Male	Family History_Yes	Diabetes_Yes	Obesity_Yes	Exercise Induced Angina_Yes
Age	1.00	-0.01	0.00	0.03	-0.01	0.03	-0.02	-0.05	-0.04	-0.01	0.65	0.01	0.02	-0.01	-0.01	-0.02
Cholesterol	-0.01	1.00	0.02	-0.01	0.02	0.02	0.02	0.09	0.00	-0.07	0.37	-0.01	-0.01	-0.03	0.00	0.04
Blood Pressure	0.00	0.02	1.00	-0.00	-0.02	-0.02	0.01	0.00	-0.05	0.01	0.01	-0.02	-0.02	-0.07	-0.04	0.01
Heart Rate	0.03	-0.01	-0.00	1.00	0.02	-0.02	-0.01	-0.04	0.01	0.00	0.01	0.01	0.02	0.01	-0.04	0.01
Smoking	-0.01	0.02	-0.02	0.02	1.00	-0.01	-0.00	0.02	-0.03	-0.02	0.00	-0.02	-0.02	0.03	-0.01	0.01
Alcohol Intake	0.03	0.02	-0.02	-0.02	-0.01	1.00	0.04	0.02	-0.02	0.02	0.02	-0.07	-0.02	0.02	-0.06	0.04
Exercise Hours	-0.02	0.02	0.01	-0.01	-0.00	0.04	1.00	-0.01	-0.03	0.06	-0.01	-0.02	0.03	-0.04	-0.01	0.07
Stress Level	-0.05	0.09	0.00	-0.04	0.02	0.02	-0.01	1.00	-0.01	0.03	0.01	0.03	0.05	-0.08	0.04	-0.00
Blood Sugar	-0.04	0.00	-0.05	0.01	-0.03	-0.02	-0.03	-0.01	1.00	-0.02	-0.01	0.07	-0.03	-0.01	-0.04	0.01
Chest Pain Type	-0.01	-0.07	0.01	0.00	-0.02	0.02	0.06	0.03	-0.02	1.00	-0.02	0.01	0.02	0.02	0.02	0.05
Heart Disease	0.65	0.37	0.01	0.01	0.00	0.02	-0.01	0.01	-0.01	-0.02	1.00	0.03	0.03	0.02	-0.02	0.00
Gender_Male	0.01	-0.01	-0.02	0.01	-0.02	-0.07	-0.02	0.03	0.07	0.01	0.03	1.00	0.10	-0.04	-0.06	0.01
Family History_Yes	0.02	-0.01	-0.02	0.02	-0.02	-0.02	0.03	0.05	-0.03	0.02	0.03	0.10	1.00	0.02	0.01	0.00
Diabetes_Yes	-0.01	-0.03	-0.07	0.01	0.03	0.02	-0.04	-0.08	-0.01	0.02	0.02	-0.04	0.02	1.00	0.00	-0.07
Obesity_Yes	-0.01	0.00	-0.04	-0.04	-0.01	-0.06	-0.01	0.04	-0.04	0.02	-0.02	-0.06	0.01	0.00	1.00	-0.00
Exercise Induced Angina_Yes	-0.02	0.04	0.01	0.01	0.01	0.04	0.07	-0.00	0.01	0.05	0.00	0.01	0.00	-0.07	-0.00	1.00

WHY PREDICTIVE MODELING?

- To accurately assess heart disease risk based on multiple health and lifestyle factors, enabling early detection and prevention

- Enhances decision-making, supports personalized care, and improves public health by focusing on high-risk groups

Logistic regression

Ideal for binary classification tasks like predicting heart disease presence

Random Forest

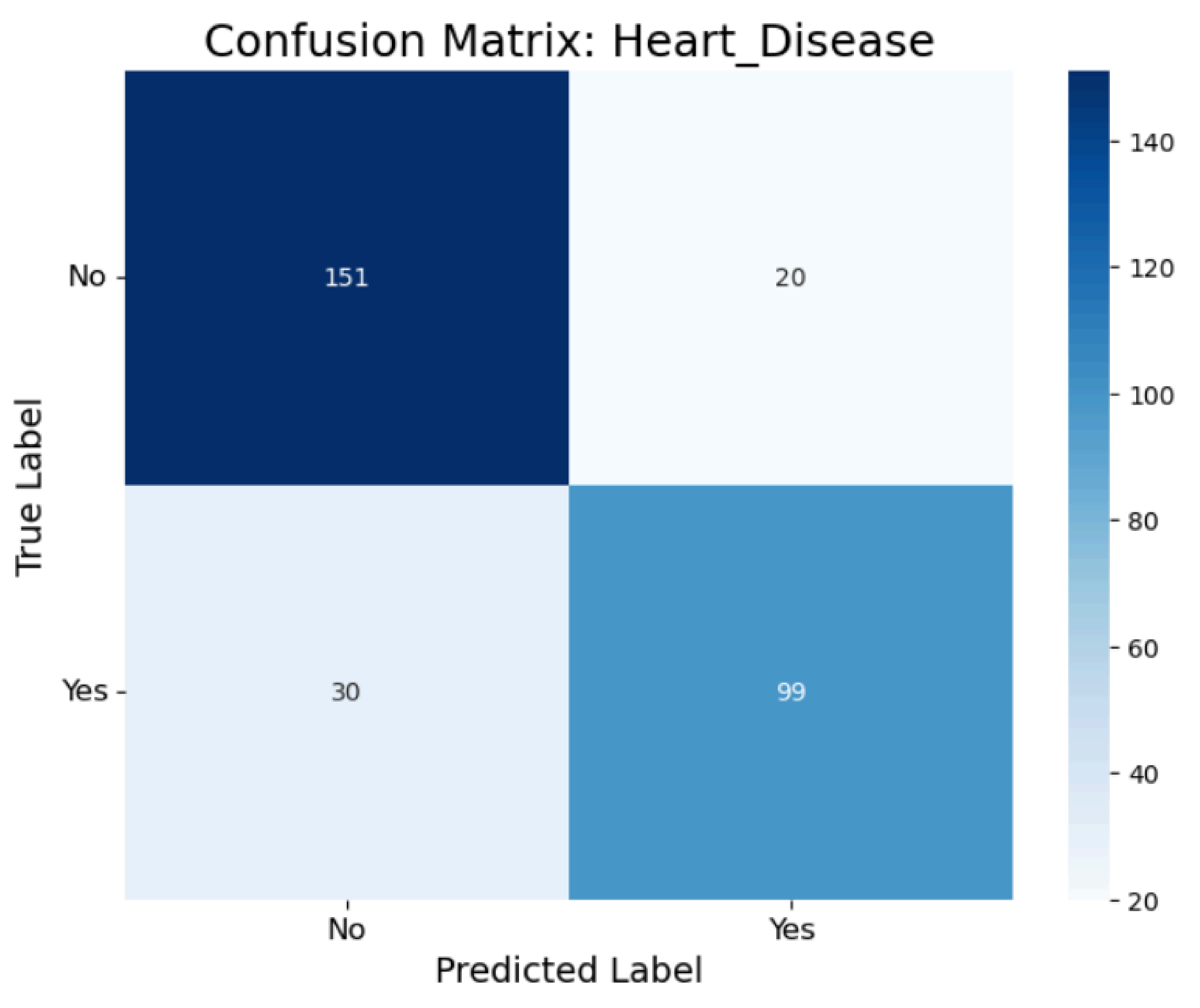
Robust model that handles complex relationships.

K-Nearest Neighbors

Effective for making predictions based on the similarity between data points

CONFUSION MATRIX

- ***True Positives:*** 76.4% of the actual heart disease cases were correctly identified
- ***True Negatives:*** 83.8% of the healthy individuals were correctly identified as not having heart disease
- ***False Positives:*** 13.2% of individuals were incorrectly predicted to have heart disease
- ***False Negatives:*** 19.1% of heart disease cases were missed by the model, indicating potential areas for improvement



LOGISTIC REGRESSION

Accuracy Score: 83.33

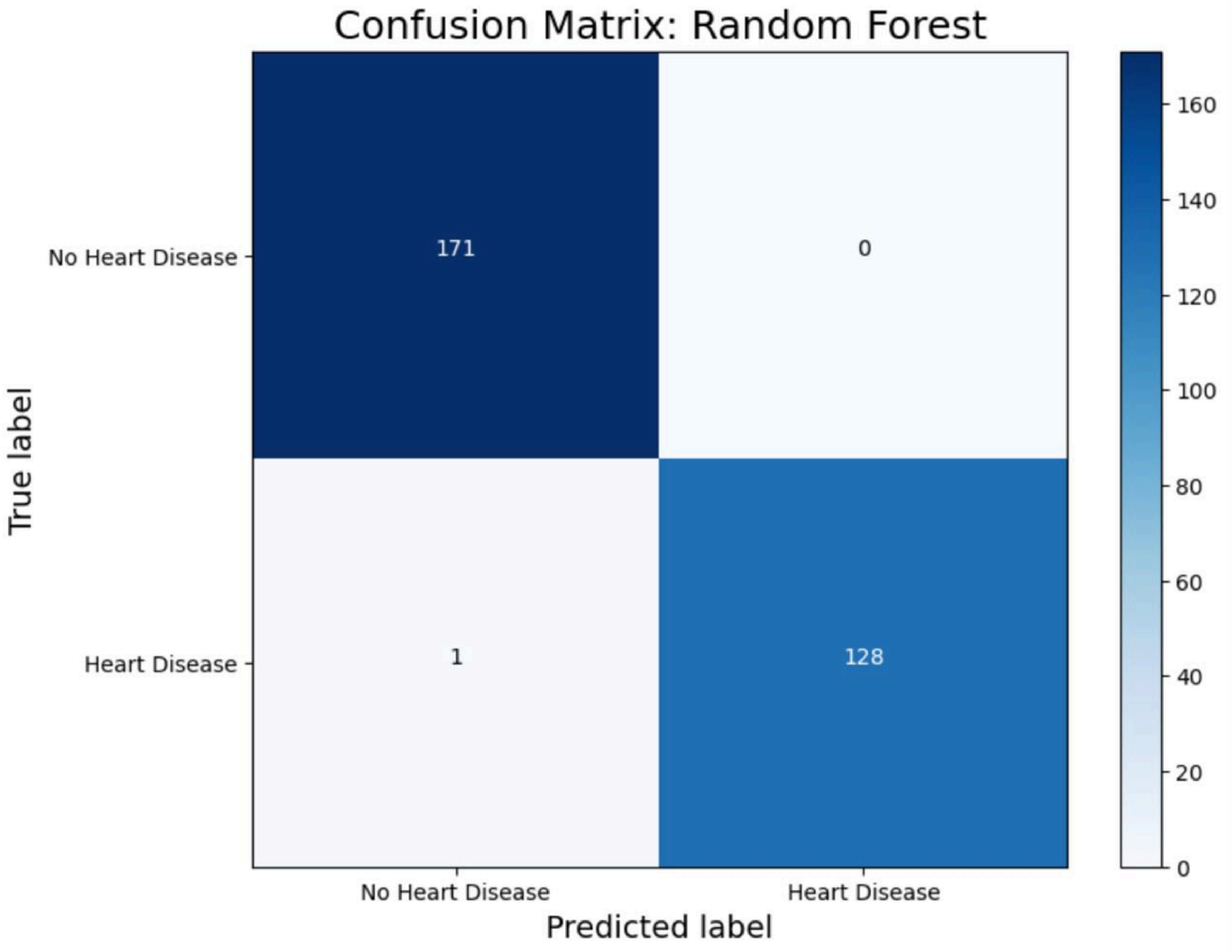
Logistic Regression	Precision	Recall	F1-score	Support
No Heart Disease	0.83	0.88	0.86	171
Heart Disease	0.83	0.77	0.70	129

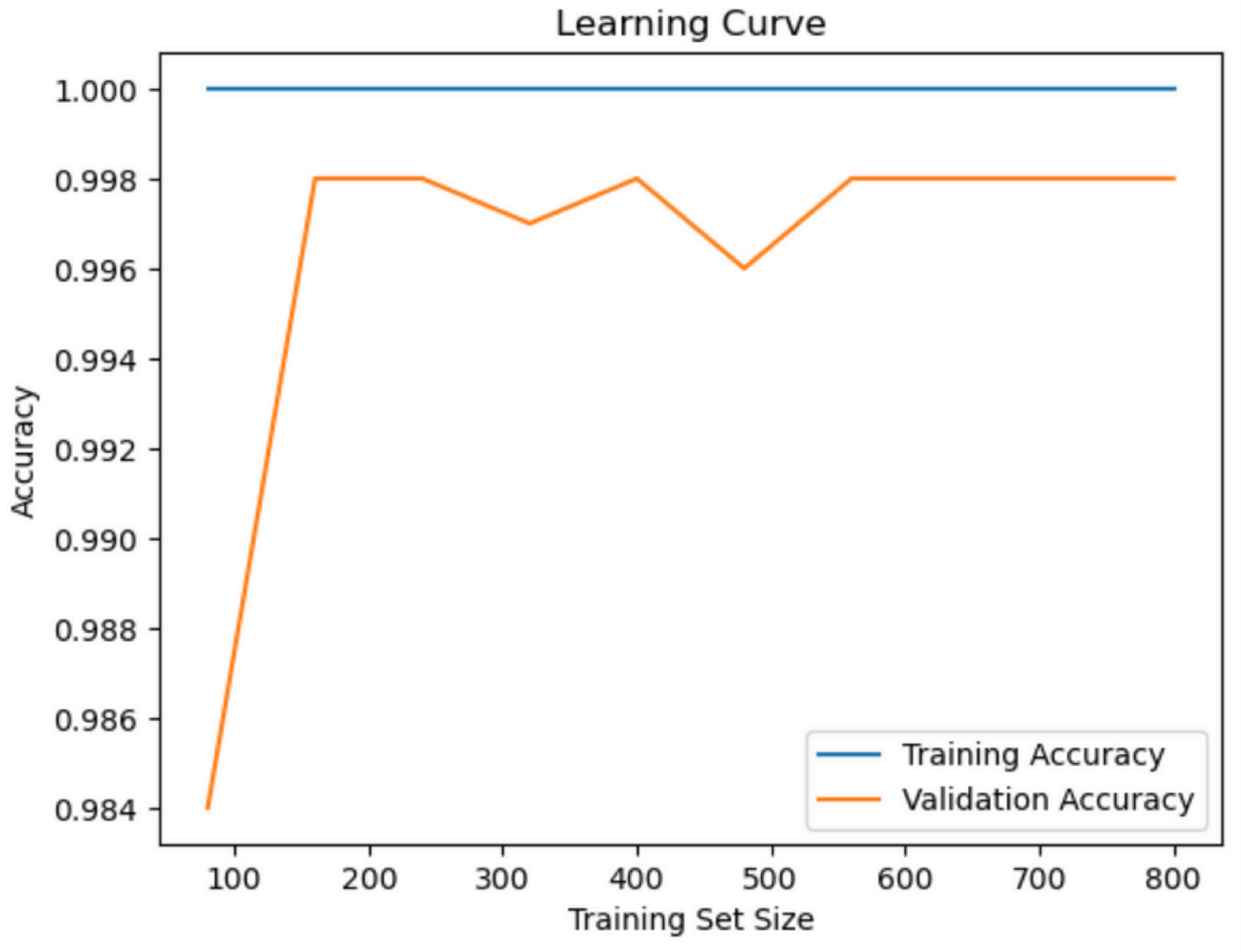
K-NEAREST NEIGHBORS

Accuracy Score: 87.33

KNN	Precision	Recall	F1-score	Support
No Heart Disease	0.87	0.92	0.89	171
Heart Disease	0.88	0.81	0.85	129

RANDOM FOREST





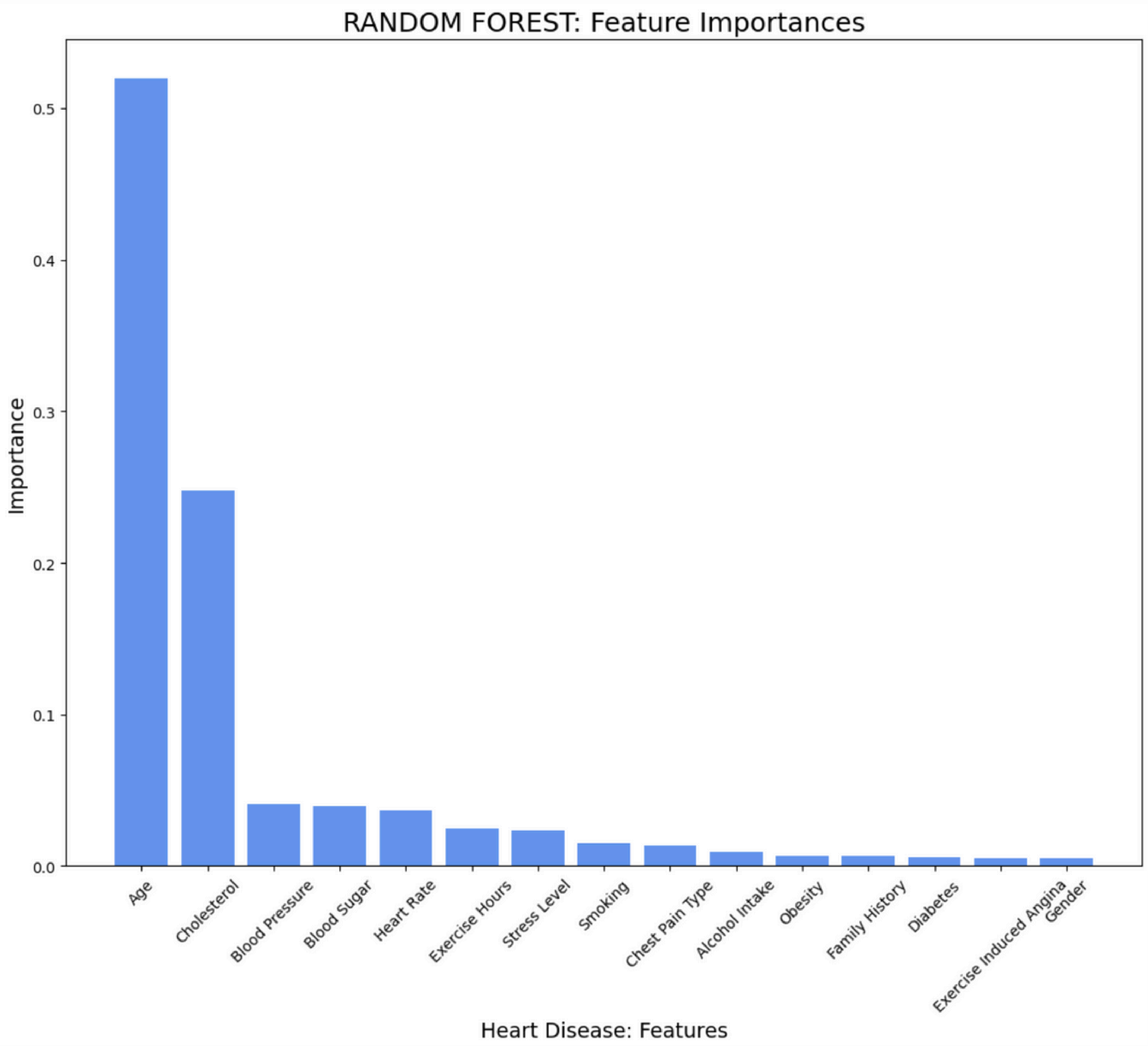
Cross Validation Accuracy: 99.8%

RANDOM FOREST

- *Accuracy:* 99.9%

- Age and Cholesterol are the most influential factors in predicting heart disease, significantly outweighing all other features.

- Lifestyle and demographic factors like Smoking, Stress Level, and Gender have minimal impact in this model.



INSIGHTS

- People above the age of 50 are more susceptible to heart disease
- Cholesterol levels above 218 may significantly increase chances of developing a heart disease.
- Typical Angina did not show any strong correlations in this data suggesting potential under diagnosis of other pain symptoms
- Moderate alcohol intake increases the possibility of heart disease

RECOMMENDATIONS

- Increase sample size of data to improve representation of people less than the age of 40
- Healthcare providers should evaluate atypical/ typical angina’s and other chest pain types to prevent misdiagnosis
- Insurance companies can target a better coverage plan for people above the age of 50
- Individuals should understand the risks of cholesterol and intake of alcohol
- Startup’s can leverage this data to build and leverage products that track cholesterol

Q
A