Heart Disease Prediction

Submitted By:

Hari Prabha N

hariii2598@gmail.com

1. Abstract

Heart disease cases are escalating rapidly, making early prediction and diagnosis increasingly critical. Effective prediction models can significantly enhance preventive measures and treatment outcomes. We have developed a heart disease prediction system that utilises the patient's medical history to ascertain the probability of heart disease.

Our approach involves deploying different machine learning algorithms, including ***Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM),Gradient Boosting*** and ***Multi-layer Perceptron (MLP)*** to predict and classify patients at risk of heart disease. These algorithms were chosen for their ability to handle classification tasks effectively and provide high accuracy. The system is designed to optimise prediction accuracy, particularly focusing on improving the identification of individuals at risk of a heart attack.This improvement in prediction accuracy alleviates some of the pressures faced by healthcare professionals in diagnosing heart disease, ensuring more reliable and timely predictions.

By implementing this heart disease prediction system, we aim to enhance medical care quality and reduce associated costs. The system provides significant insights that aid in the accurate prediction of heart disease in patients. The model is implemented in a Jupyter Notebook (.pynb) format, facilitating ease of use and integration into healthcare systems. This project underscores the potential of machine learning in transforming heart disease prediction and improving patient outcomes.

Introduction

Machine learning is a powerful tool that helps us find useful information in data that we might not see otherwise. It's a growing field that's becoming more important every day, especially in healthcare. Machine learning includes various methods to analyze data and make predictions. In our project, we use machine learning to predict heart disease, which can save lives by catching the disease early.

Cardiovascular diseases (CVDs) are heart-related conditions that are very common today. These conditions include coronary artery disease, heart attacks, heart failure, and arrhythmias. According to the World Health Organization, CVDs cause about 17.9 million deaths worldwide each year, making them the leading cause of death among adults . This is why predicting heart disease early is so crucial. Early diagnosis and treatment can significantly reduce the risk of severe complications and improve the quality of life for patients.

Our Heart Disease Prediction System (HDPS) helps identify people who are likely to develop heart disease by analyzing their medical history. By doing this, we can help doctors diagnose the disease earlier with fewer tests and provide effective treatments sooner. Traditional methods of diagnosing heart disease can be invasive, expensive, and time-consuming. Using machine learning can streamline the process, making it more efficient and accessible.

We use a dataset from the UCI Machine Learning Repository, which includes 14 medical attributes such as age, gender, chest pain type, and fasting blood sugar levels . We analyze this data using six different machine learning models:

- **Logistic Regression:** A model that predicts the likelihood of an outcome based on input data.
- **Decision Tree Classifier:** A model that uses a tree-like structure to make decisions based on the data.
- **Random Forest Classifier:** An advanced model that uses many decision trees to improve prediction accuracy. This model achieved the highest accuracy of 98% in our project.
- **Support Vector Machine (SVM):** A model that finds the best boundary to separate different classes in the data.
- **Gradient Boosting Classifier:** A model that builds a series of simple models to gradually improve prediction accuracy.
- **Multi-Layer Perceptron (MLP) Neural Network:** A type of artificial neural network that can model complex relationships in the data.

By using these models, we can classify patients based on their risk of developing heart disease. The Random Forest Classifier, in particular, proved to be the most effective, with an accuracy of 98%. This high accuracy means our system can help doctors make better predictions and decisions, ultimately improving patient care and reducing healthcare costs.

The risk factors for heart disease include high blood pressure, high cholesterol, diabetes, physical exercises , and few other details related to heart disease . By incorporating these factors into our prediction models, we can provide more accurate assessments of a patient's risk. This allows for earlier interventions and personalised treatment plans, which can prevent the progression of heart disease and improve patient outcomes.

Our project shows how machine learning can transform heart disease prediction, making it more accurate and efficient.

Project Objectives

Heart disease is a leading cause of death worldwide, making early detection and intervention critical for improving patient outcomes and reducing healthcare costs. This project aims to develop a machine learning-based heart disease prediction system that addresses several important business needs. By analyzing medical data, the system can identify patients at risk of developing heart disease before symptoms become severe, allowing for timely and potentially life-saving interventions. The machine learning models used in this project—such as logistic regression, decision tree classifier, random forest classifier, SVM, gradient boosting classifier, and MLP neural network—provide improved accuracy over traditional diagnostic methods, reducing the risk of misdiagnosis and ensuring patients receive the appropriate care.

The benefits of this system extend beyond individual patient care. Healthcare providers, including doctors and nurses, can use the system as a reliable tool to assess patient risk, supporting clinical decision-making and prioritizing patients who need immediate attention. Healthcare institutions, such as hospitals and clinics, can integrate the prediction system into their electronic health records (EHR) to streamline patient management and improve service efficiency. Additionally, this project can significantly reduce healthcare costs by preventing heart disease or managing it more effectively in its early stages, thus reducing the need for expensive procedures and long-term care.

For the health tech industry, this system represents a valuable business opportunity. Companies can offer advanced diagnostic tools to healthcare providers, opening up new markets for predictive analytics in healthcare and leading to the development of additional services and products. Patients, too, benefit by becoming more informed about their health risks, enabling them to take proactive steps in consultation with their healthcare providers. This empowerment can lead to better adherence to preventive measures and treatment plans.

Overall, the heart disease prediction system aims to enhance early detection, improve diagnostic accuracy, reduce healthcare costs, enhance patient care, and ensure efficient resource allocation. By addressing these business needs, the project can significantly impact the healthcare industry, benefiting healthcare providers, institutions, patients, and companies in the health tech sector.

Methodology

4.1. Data Collection

In this study, the primary dataset used is sourced from the UCI Machine Learning Repository, a well-regarded resource for research in predictive analytics [4, 7]. This dataset is specifically tailored for predicting heart disease, making it ideal for our analysis. Upon retrieval, the dataset is split into two subsets: a training set comprising 70% of the data and a testing set containing the remaining 30%. This division ensures that our models are trained on a sufficient amount of data while preserving a separate portion for unbiased evaluation.

4.2. Dataset and Attributes

The dataset consists of 14 attributes that provide crucial insights into the health profile of individuals at risk of heart disease. These attributes include demographic details such as age and sex, clinical indicators like chest pain type (cp), resting blood pressure (trestbps), serum cholesterol levels (chol), and physiological responses during exercise (exang, oldpeak). Each attribute plays a vital role in the predictive models' ability to discern patterns indicative of heart disease. To optimise model performance, attribute selection and feature engineering techniques, such as correlation analysis, are employed to identify the most influential predictors.

Detailed Dataset Overview:

- Age: Ranges from 29 to 77 years, with an average of 54 years.
- Sex: Approximately 75% of the dataset comprises males.
- Chest Pain (cp): Represents varying levels of chest pain, averaging around 0.94 out of 3.
- Resting Blood Pressure (trestbps): Average resting blood pressure is 131.61 mm Hg, ranging from 94 mm Hg to 200 mm Hg.
- Serum Cholesterol (chol): Average serum cholesterol level is 246 mg/dL, ranging from 126 mg/dL to 564 mg/dL.
- Fasting Blood Sugar (fbs): About 15% of individuals have fasting blood sugar levels above 120 mg/dL.
- Resting Electrocardiographic Results (restecg): Indicates a mix of normal and abnormal resting ECG results.
- Maximum Heart Rate Achieved (thalach): Average maximum heart rate achieved is 149.11 beats per minute, ranging from 71 to 202 beats per minute.
- Exercise-Induced Angina (exang): Approximately 33% of individuals experience exercise-induced angina.
- ST Depression Induced by Exercise (oldpeak): Average ST depression induced by exercise is 1.07, ranging from 0 to 6.2.
- Slope of the Peak Exercise ST Segment (slope): Average slope is 1.38, ranging from 0 to 2.
- Number of Major Vessels Colored by Fluoroscopy (ca): On average, 0.75 major vessels are colored by fluoroscopy, ranging from 0 to 4.
- Thalassemia (thal): Predominantly includes individuals with normal or fixed defect thalassemia types.

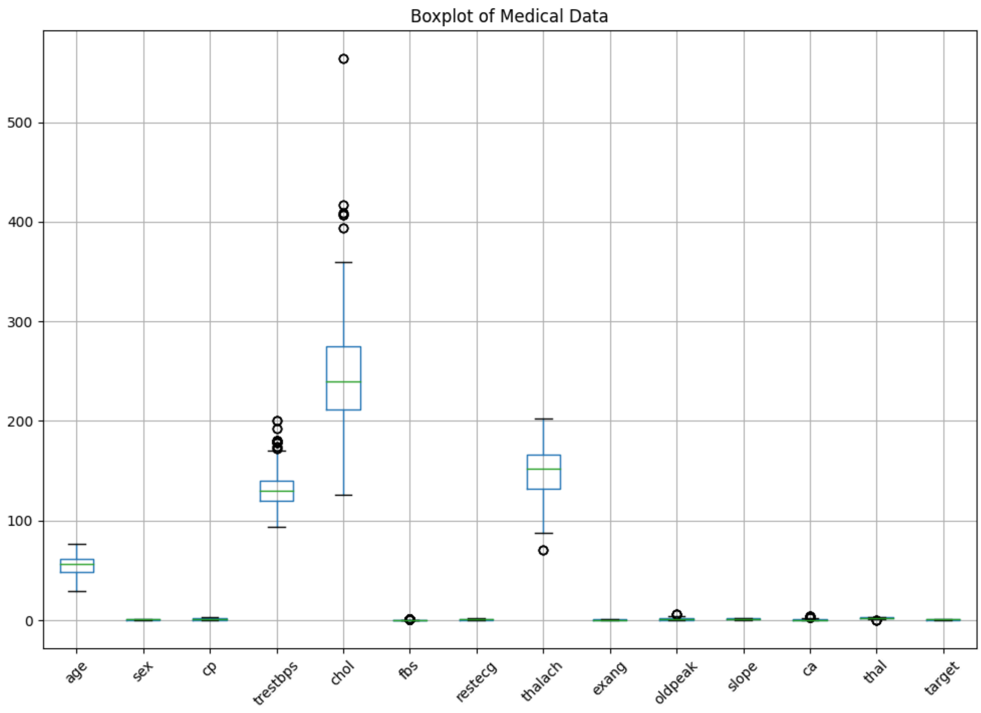
- Target Variable (target): About 75% of the dataset shows signs of heart disease (target = 0), while the remaining 25% do not have heart disease (target = 1).

	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.4341463415	9.0722902332	29.0	48.0	56.0	61.0	77.0
sex	1025.0	0.6956097561	0.4603733241	0.0	0.0	1.0	1.0	1.0
cp	1025.0	0.9424390244	1.0296407436	0.0	0.0	1.0	2.0	3.0
trestbps	1025.0	131.6117073171	17.5167180054	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.0000000000	51.5925102062	126.0	211.0	240.0	275.0	564.0
fbs	1025.0	0.1492682927	0.3565266897	0.0	0.0	0.0	0.0	1.0
restecg	1025.0	0.5297560976	0.5278775669	0.0	0.0	1.0	1.0	2.0
thalach	1025.0	149.1141463415	23.0057237460	71.0	132.0	152.0	166.0	202.0
exang	1025.0	0.3365853659	0.4727723760	0.0	0.0	0.0	1.0	1.0
oldpeak	1025.0	1.0715121951	1.1750532552	0.0	0.0	0.8	1.8	6.2
slope	1025.0	1.3853658537	0.6177552672	0.0	1.0	1.0	2.0	2.0
ca	1025.0	0.7541463415	1.0307976650	0.0	0.0	0.0	1.0	4.0
thal	1025.0	2.3239024390	0.6206602381	0.0	2.0	2.0	3.0	3.0
target	1025.0	0.5131707317	0.5000704981	0.0	0.0	1.0	1.0	1.0

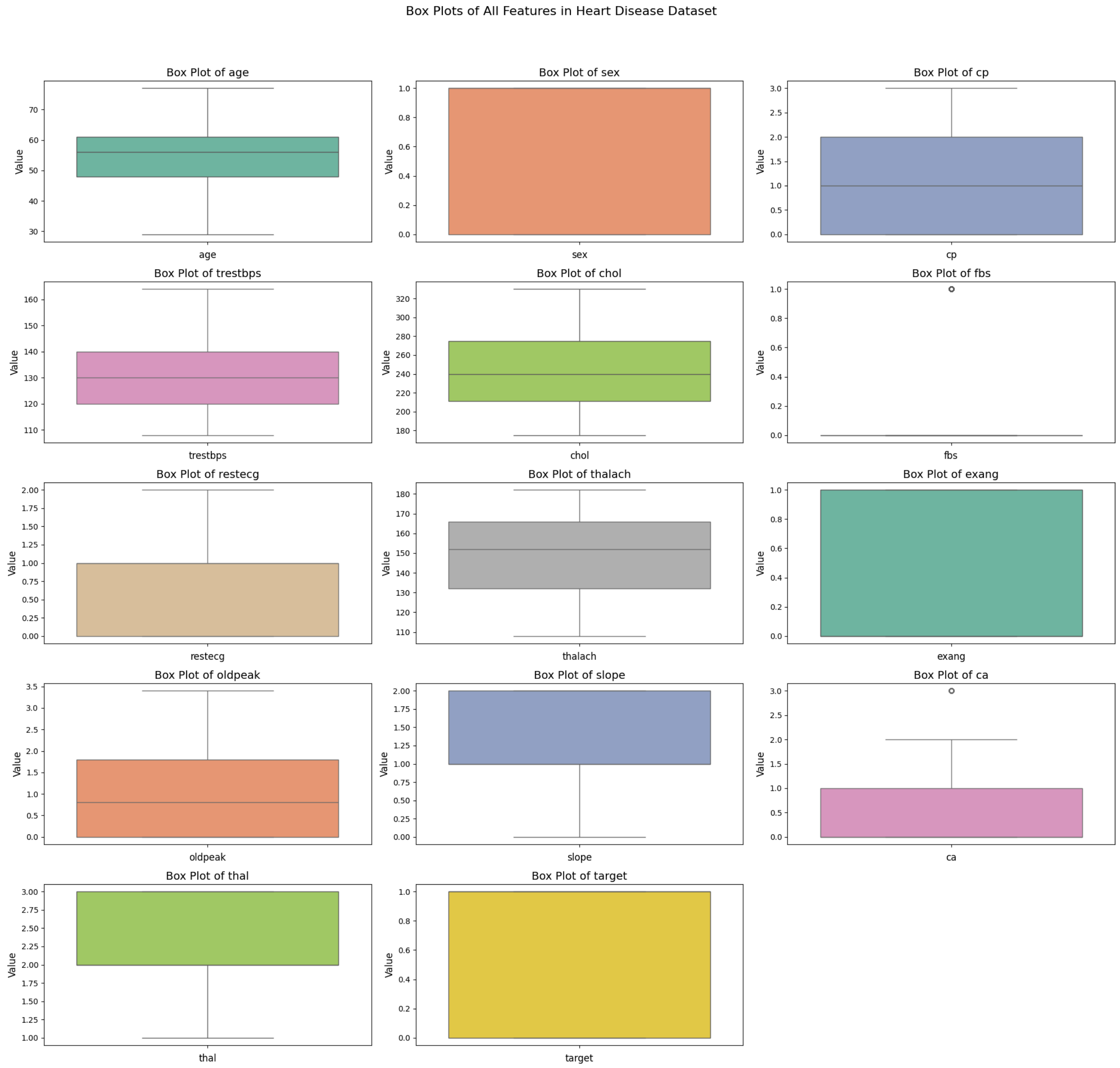
4.3. Pre-processing of Data

Before training the models, the dataset undergoes thorough pre-processing to ensure it is clean and ready for analysis. This process begins with handling missing data, though in this case, there are no missing values in the dataset. Next, all attributes are scaled using the StandardScaler method to ensure numerical values are standardised. This prevents any single attribute from dominating the model's learning process.

Outliers are present in several attributes, including 'trestbps' (Resting Blood Pressure), 'chol' (Cholesterol), 'fbs' (Fasting Blood Sugar), 'thalach' (Maximum Heart Rate), 'oldpeak' (ST Depression), 'ca' (Number of Major Vessels Coloured by Fluoroscopy), and 'thal' (Thalassemia). These outliers are managed using the Winsorization method, which replaces extreme values with the nearest values within defined bounds. This approach ensures that outliers do not skew the predictive models, maintaining the accuracy of the predictions by preventing overly influential outliers from misleading the analysis. Accurate predictions rely on patterns and trends in the data, and managing outliers is crucial to maintaining the reliability of heart disease risk or severity predictions.

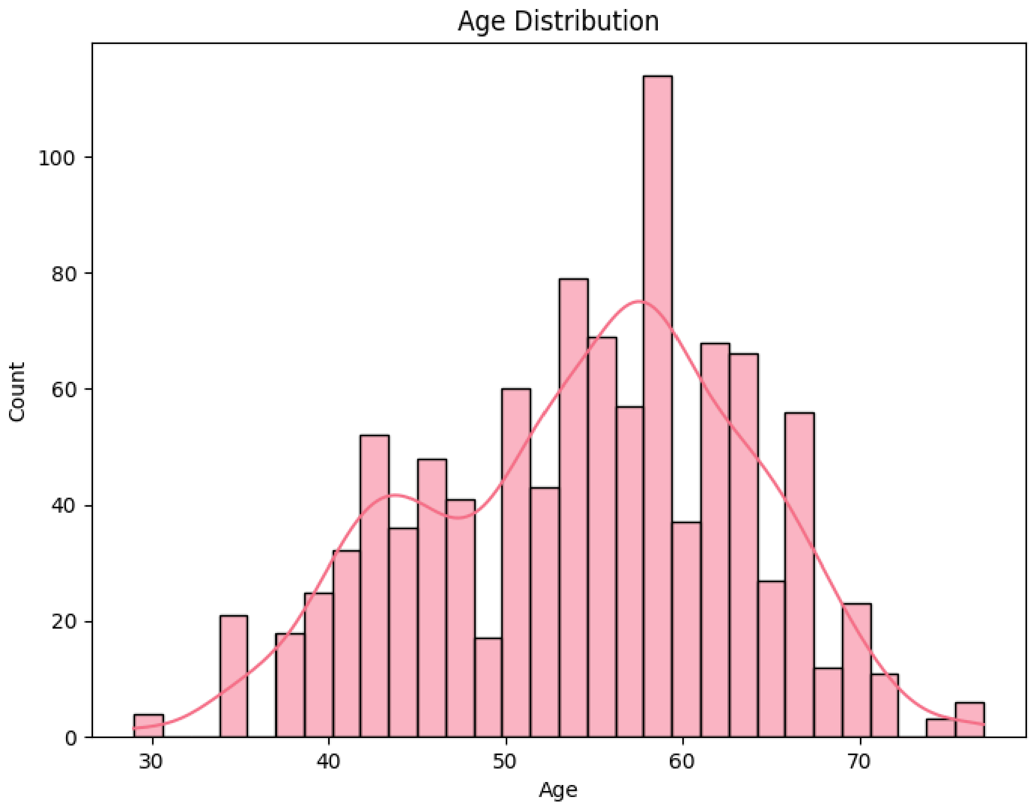


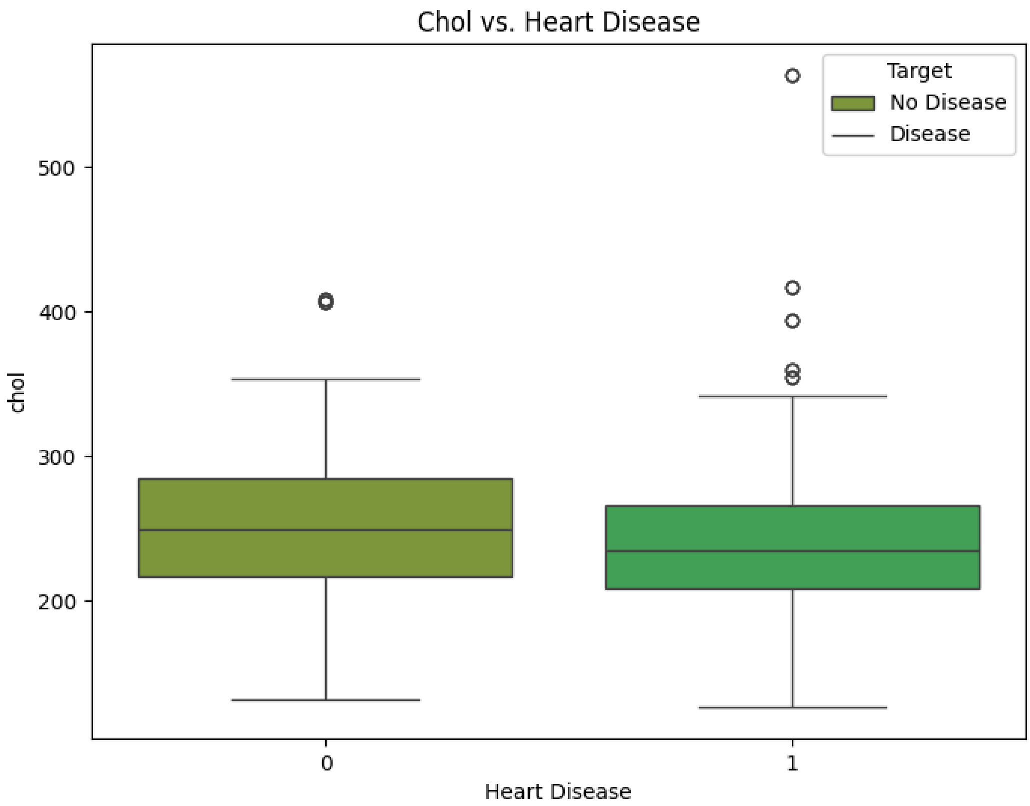
The Box Plot after handling with the outliers

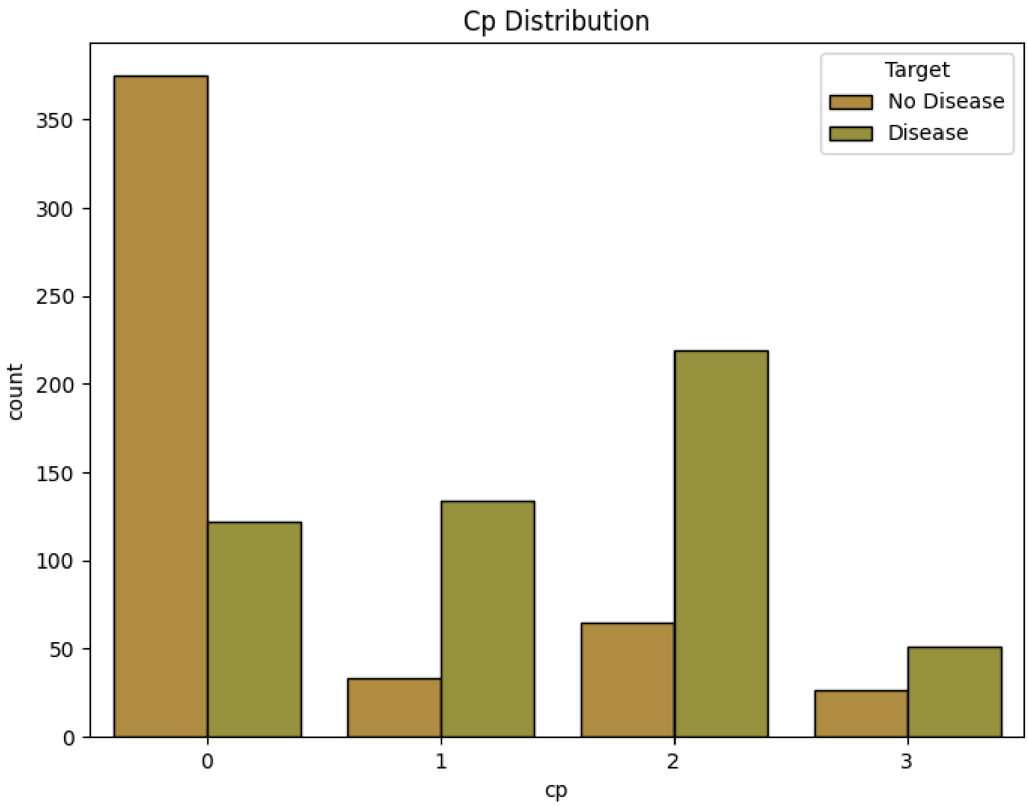


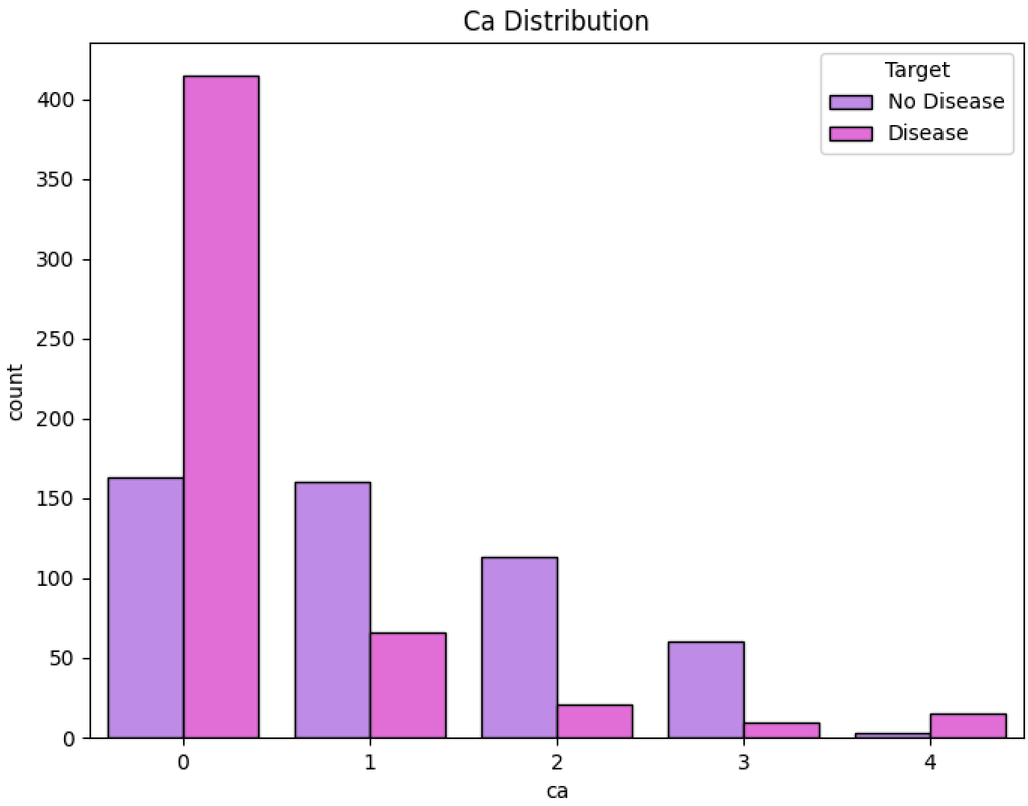
4.4. Data Visualization

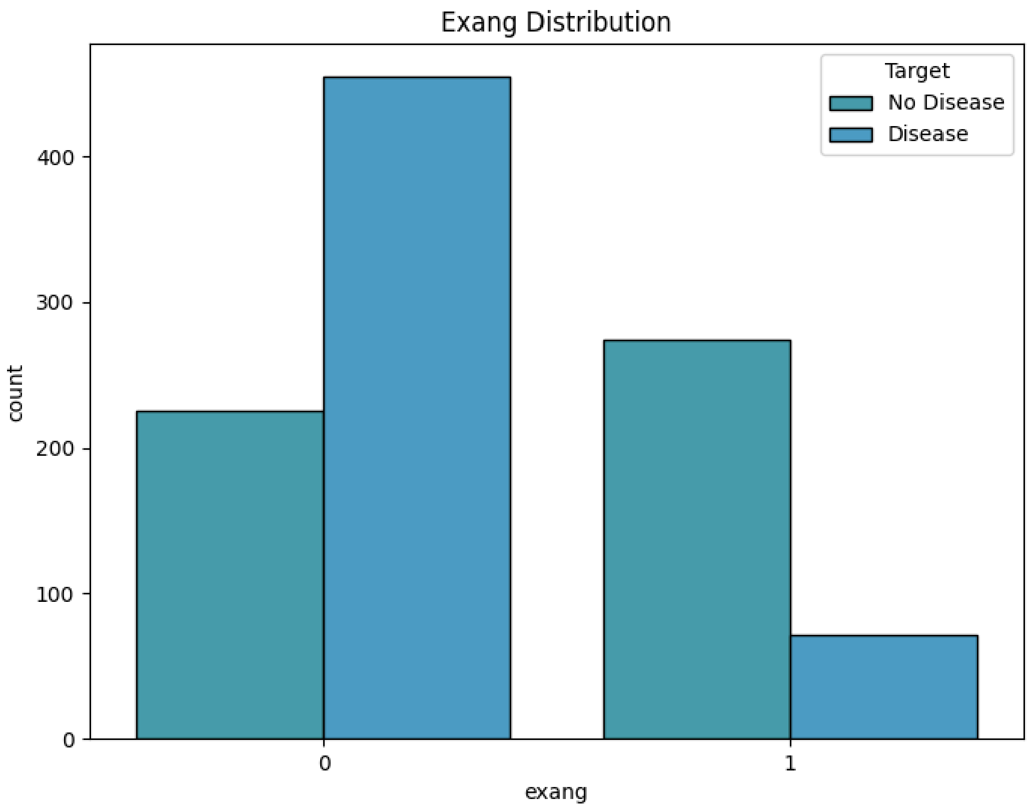
We begin with exploratory data analysis (EDA) to understand the data distribution and identify patterns. Visualisation techniques such as histograms, box plots, scatter plots, and correlation heatmaps are used to gain insights.

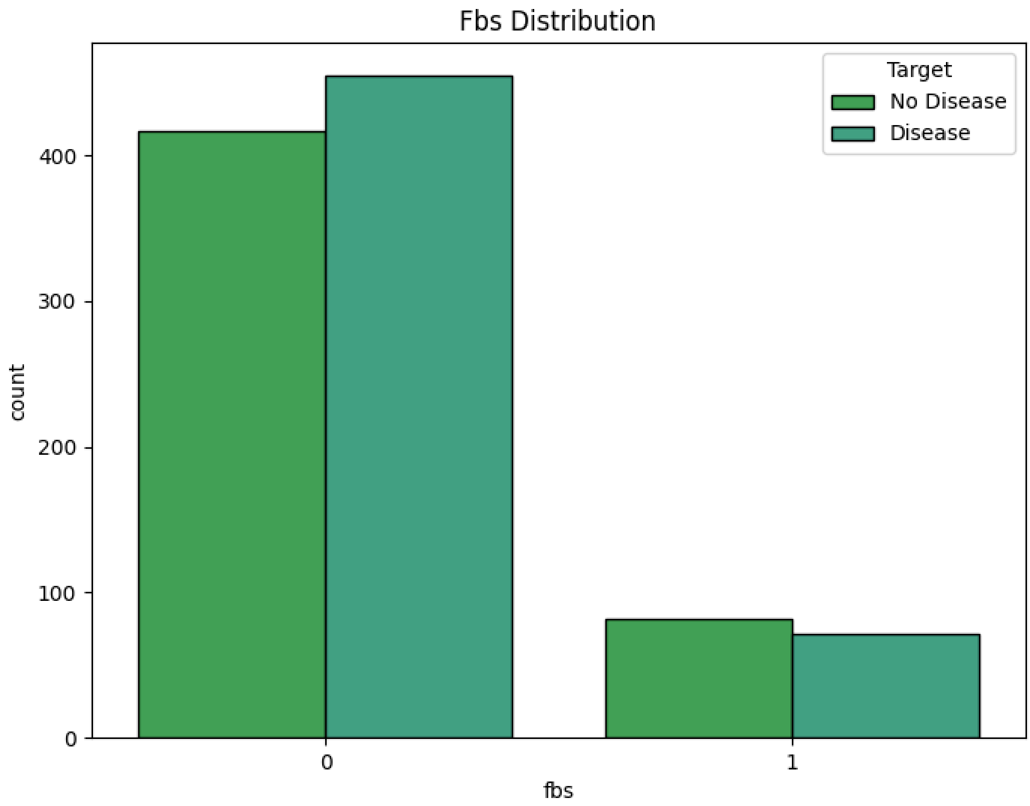

Age Distribution
Count
Age
0
20
40
60
80
100
30
40
50
60
70

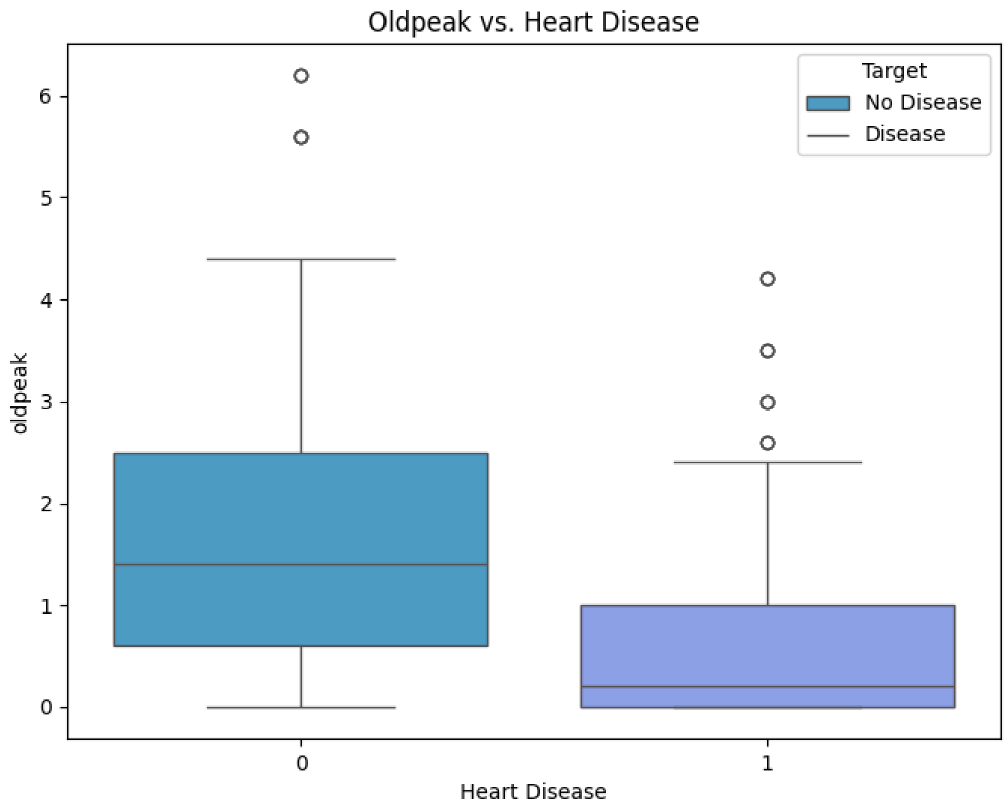

Chol vs. Heart Disease
Target
No Disease
Disease
chol
Heart Disease
200
300
400
500
0
1

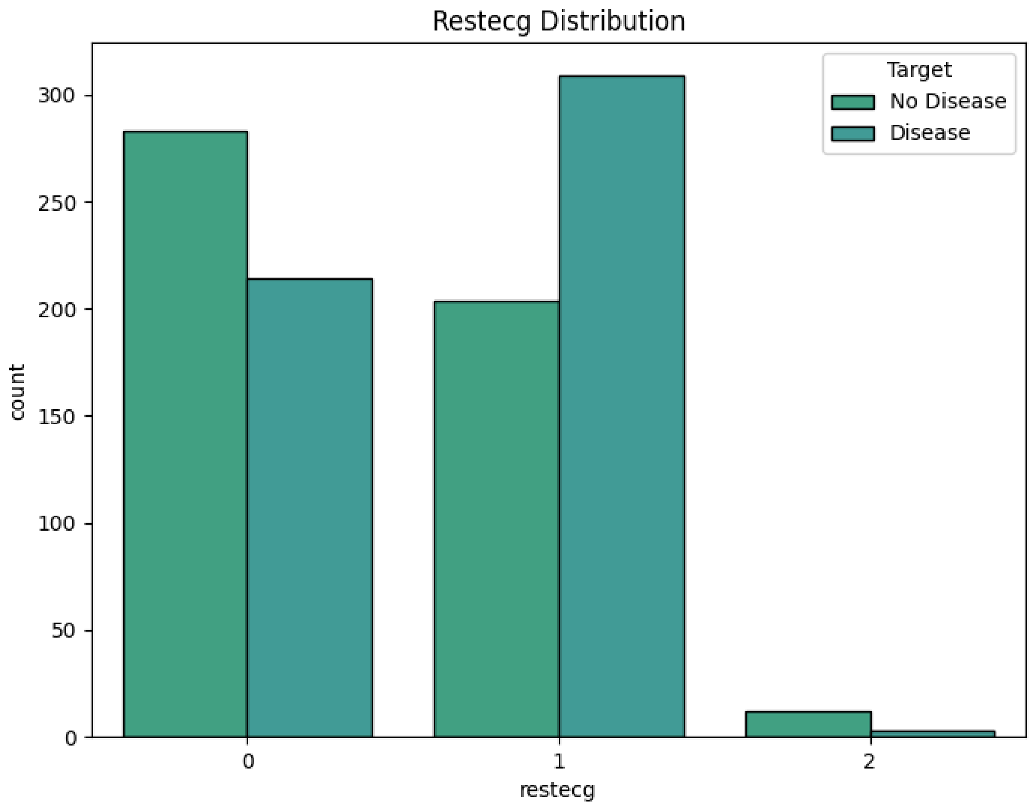

Cp Distribution
Target
No Disease
Disease
count
cp
0
50
100
150
200
250
300
350
0
1
2
3

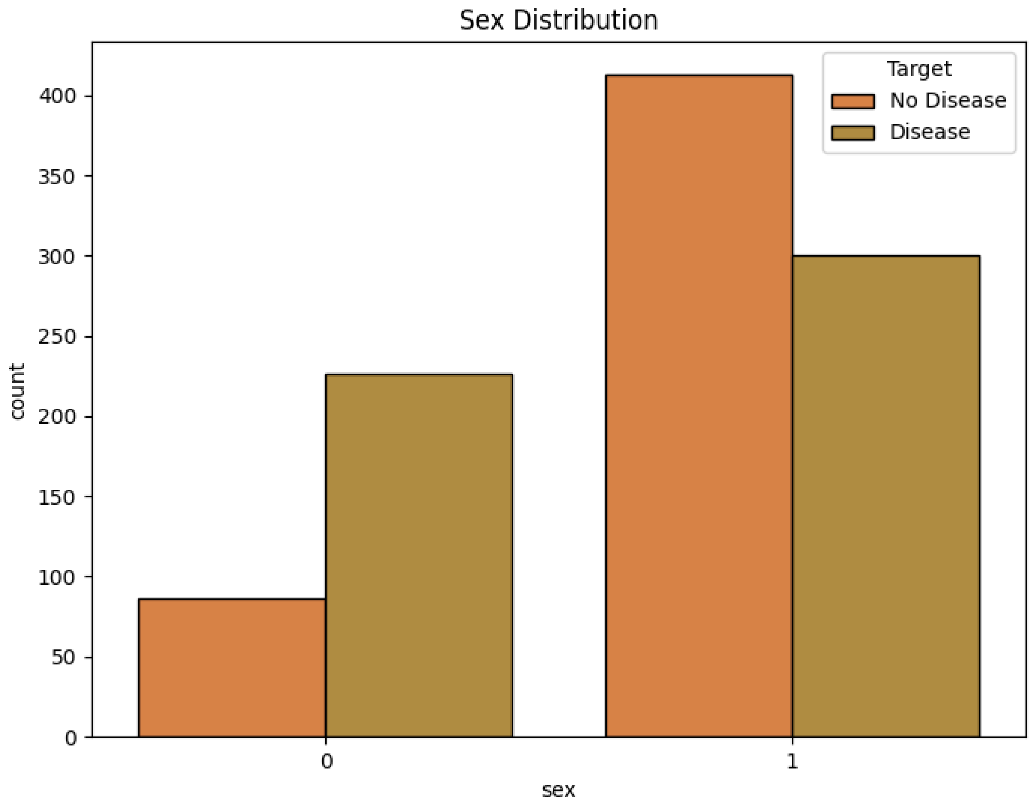

Ca Distribution
Target
No Disease
Disease
count
400
350
300
250
200
150
100
50
0
0
1
2
3
4
ca

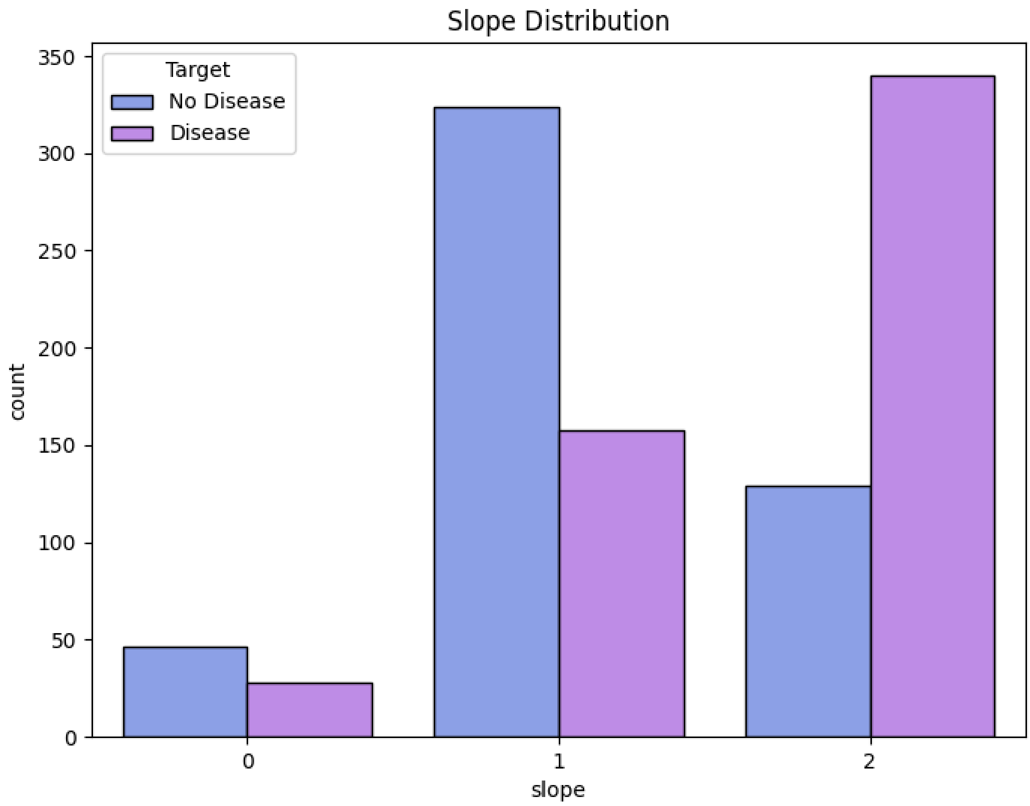

Exang Distribution
Target
No Disease
Disease
count
400
300
200
100
0
0
1
exang

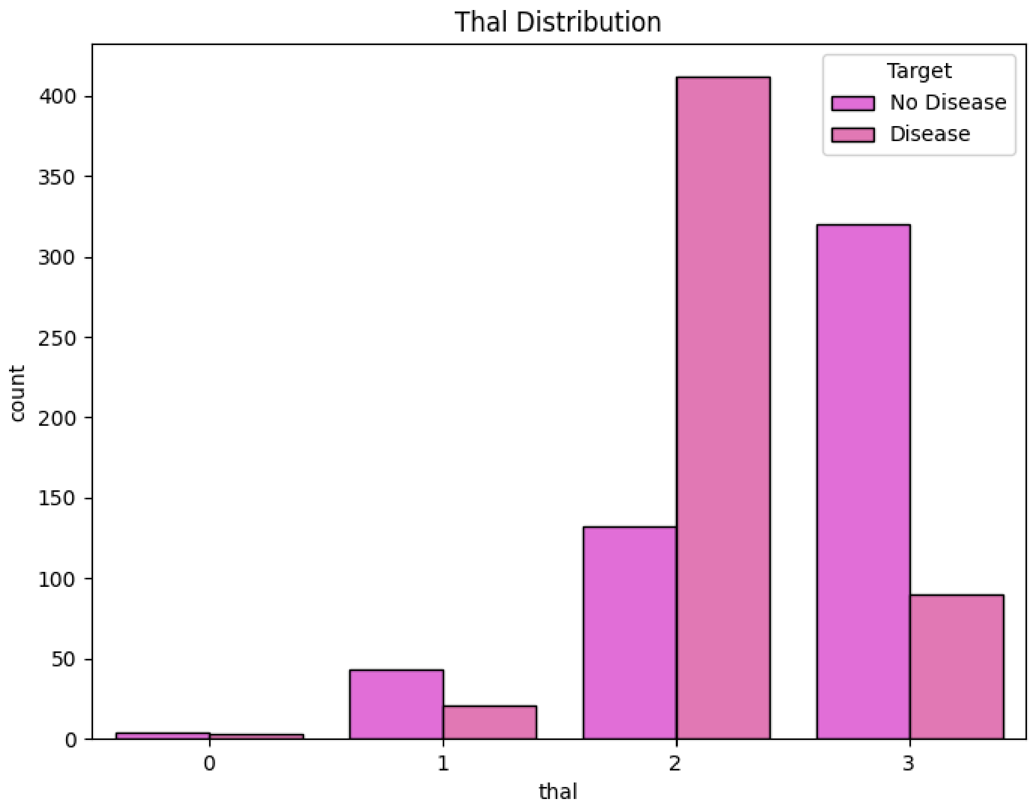

Fbs Distribution
Target
No Disease
Disease
400
300
200
100
0
count
0
1
fbs

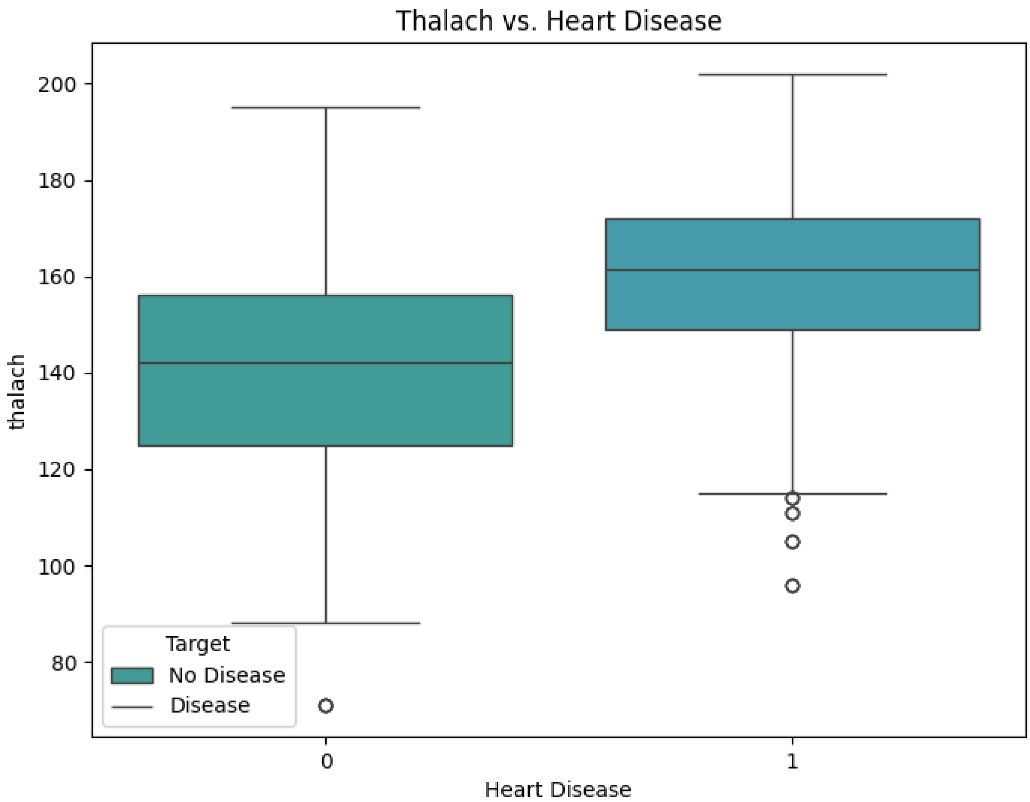

Oldpeak vs. Heart Disease
Target
No Disease
Disease
6
5
4
3
2
1
0
oldpeak
0
1
Heart Disease

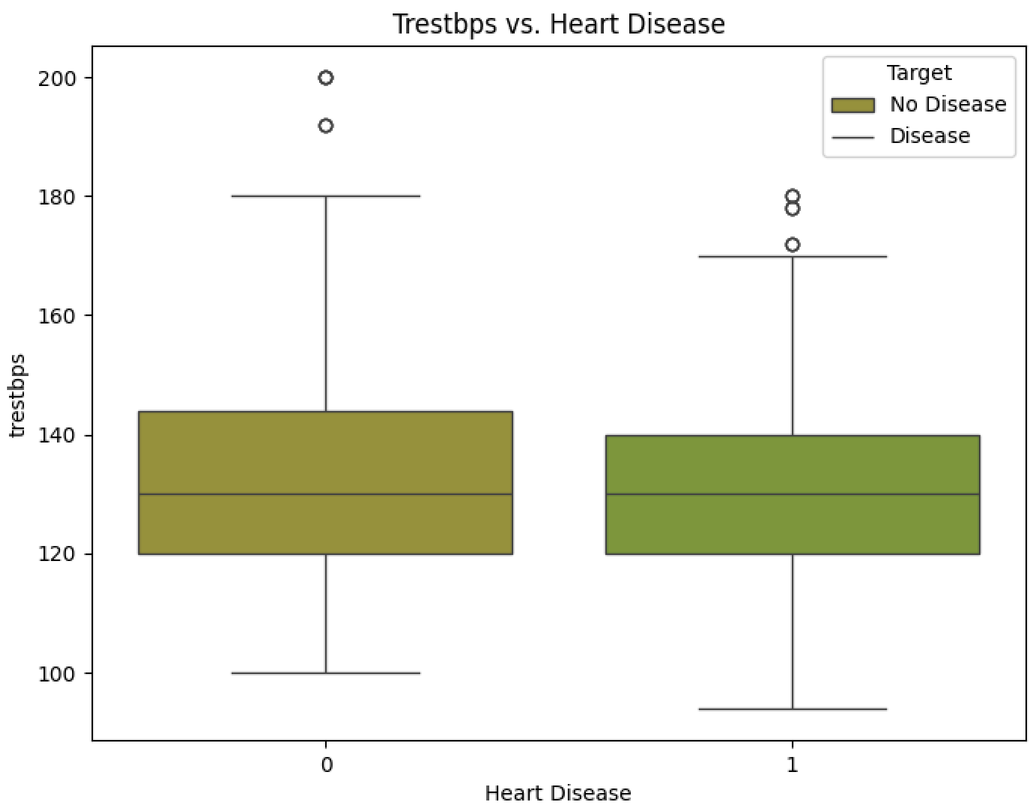

Restecg Distribution
Target
No Disease
Disease
count
300
250
200
150
100
50
0
0
1
2
restecg


Sex Distribution
Target
No Disease
Disease
count
400
350
300
250
200
150
100
50
0
0
1
sex


Slope Distribution
Target
No Disease
Disease
count
slope
350
300
250
200
150
100
50
0
0
1
2


Thal Distribution
Target
No Disease
Disease
count
thal
400
350
300
250
200
150
100
50
0
0
1
2
3


Thalach vs. Heart Disease
200
180
160
140
120
100
80
thalach
Target
No Disease
Disease
0
1
Heart Disease


Trestbps vs. Heart Disease
200
180
160
140
120
100
trestbps
Target
No Disease
Disease
0
1
Heart Disease

1. AGE

The age distribution appears to be in bell-shaped, with a peak around the age of 60. Most individuals in the dataset are between the ages of 40 and 70, with a higher concentration in the 50-60 age range.

2. SEX

The plot shows a higher number of males with heart disease compared to females, it suggests that males have a higher probability of getting heart disease in the dataset.

3. Chest Pain Type

- Typical Angina is more common with individuals without heart disease.
- Atypical Angina, Non-Anginal Pain, and Asymptomatic chest pain types are more frequent in individuals with heart disease, suggesting they are stronger indicators for the condition.

4. Resting Blood Pressure

The box plot shows the distribution of resting blood pressure for people with and without heart disease. We can see that people with heart disease tend to have higher resting blood pressure than people without heart disease

5. Serum Cholesterol Levels

People with heart disease tend to have higher cholesterol levels, with a wider range of values. The outliers indicate some individuals have exceptionally high cholesterol levels. This suggests a potential relationship between high cholesterol and heart disease risk.

6. Fasting Blood Sugar

The graph compares Resting Blood Pressure (mm Hg) to a target. The target appears to be around to be above 100 mm Hg, with a mention of Heart Disease. The graph suggests that blood pressure above the target range may be associated with Heart Disease.

7. Resting Electrocardiographic Results

- Resting ECG Result 0 (Normal): This result is more common in individuals without heart disease, indicating that normal ECG results are associated with a lower likelihood of having heart disease.
- Resting ECG Result 1 (Abnormality): This result is more common in individuals with heart disease, suggesting that ECG abnormalities are an indicator of heart disease.
- Resting ECG Result 2 (Left ventricular hypertrophy): This result is rare but slightly more prevalent in individuals with heart disease, indicating a heart disease.

8. Maximum Heart Rate Achieved

The plot shows that people with heart disease tend to have a higher maximum heart rate achieved, than those without heart disease.

This suggests that a higher maximum heart rate achieved could be a risk factor for heart disease

9. Exercise-Induced Angina

The people with exercise-induced angina are more likely to have heart disease. Specifically, people who experience exercise-induced angina are more likely to have heart disease.

10. ST Depression Induced by Exercise

The median ST depression for patients with heart disease is lower than the median for patients without heart disease.

There are also more outliers among the patients with heart disease, suggesting a greater variability in ST depression among this group.

11. Slope of Peak Exercise ST Segment

Flat ST Segment (value 1): The majority of patients with a flat ST segment during peak exercise do not have heart disease. This suggests a flat ST segment may be more prevalent in individuals without heart disease.

Downsloping ST Segment (value 2): There's a significant increase in the number of patients with heart disease when the slope of the ST segment is downsloping during peak exercise. This implies a downsloping ST segment may be a stronger indicator of heart disease.

Upsloping ST Segment (value 0): This category appears to have fewer individuals compared to the other two categories, indicating it might be less common.

12. Number of Major Vessels Colored by Fluoroscopy

The number of major vessels colored by fluoroscopy is generally lower in people with heart disease.

There is a large number of people with heart disease who have 0 major vessels colored by fluoroscopy, and a smaller number of people without heart disease who have 4 major vessels colored by fluoroscopy.

13. Thalassemia

The distribution of thalassemia in patients with and without heart disease shows that a higher percentage of people with heart disease have thalassemia (value 2) compared to people without heart disease.

Correlation

The correlation matrix helps to understand the relationships between different variables. In this dataset, some variables like cp, thalach, oldpeak, and exang show strong correlations with the target variable, indicating their potential importance in predicting heart disease.

Correlation between the columns

Strong Positive Correlations

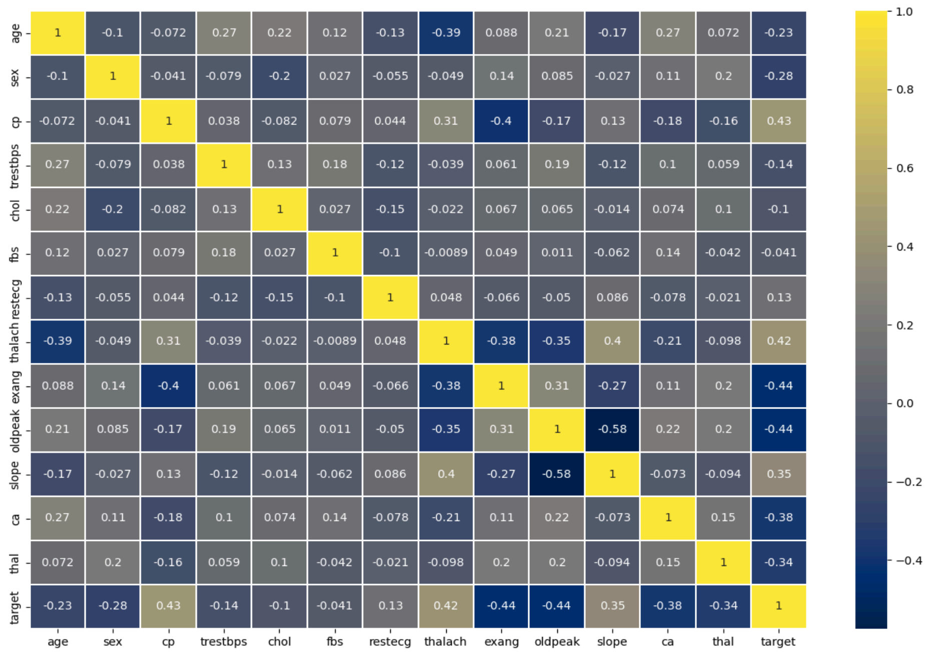
cp and target (0.43): Higher chest pain type is associated with the *presence of heart disease.*

Strong Negative Correlations:

thalach and target (-0.42)*:* Higher maximum heart rate achieved is associated with the *absence of heart disease.*

oldpeak and target (-0.44): Higher ST depression is associated with the *presence of heart disease.*

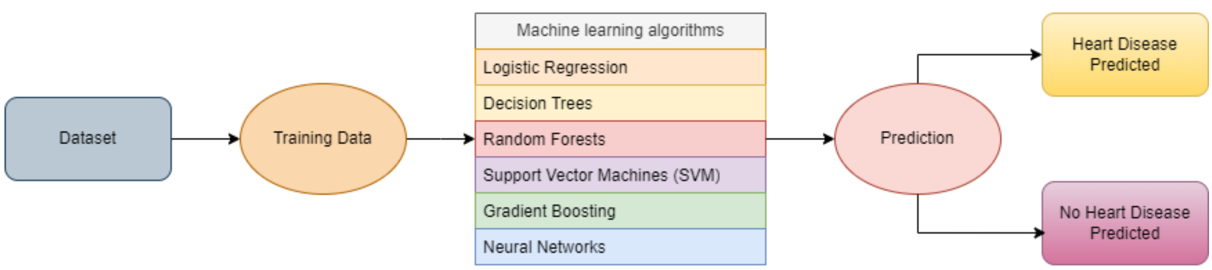
exang and target (-0.44): Exercise-induced angina is associated with the *presence of heart disease.*



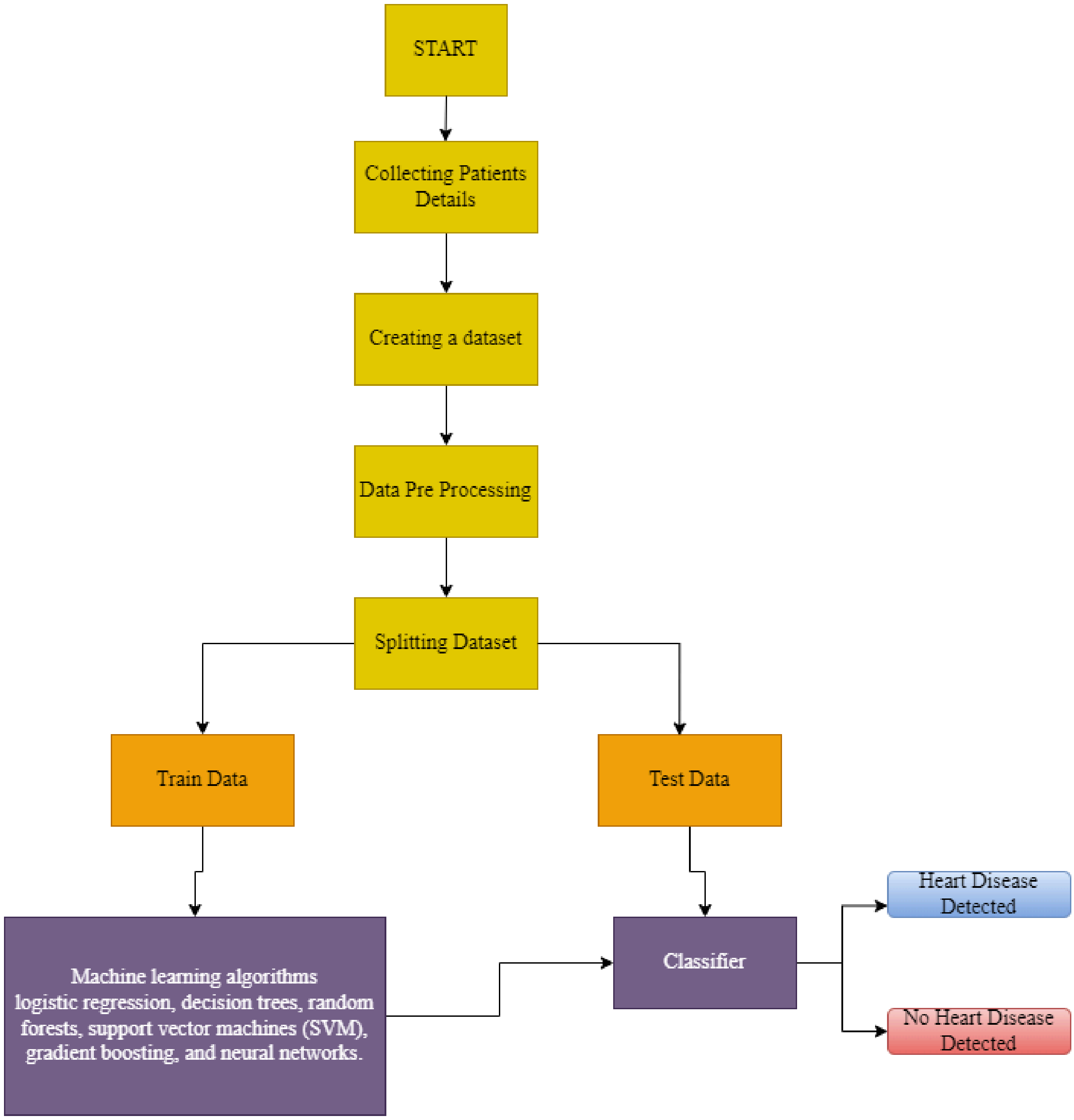
4.5. Prediction of Disease

With the pre-processed and balanced dataset, multiple machine learning algorithms are employed to train predictive models. These algorithms include logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting, and neural networks. The models are trained on the training dataset and evaluated using the testing dataset to assess their accuracy, precision, recall, and F1-score. The goal is to select the best-performing model based on these metrics, ensuring it meets clinical standards for accuracy and reliability in predicting heart disease.

This comprehensive methodology ensures that the developed predictive models are robust, accurate, and clinically relevant, aiming to improve early detection and treatment outcomes for heart disease.



Flow Diagram:

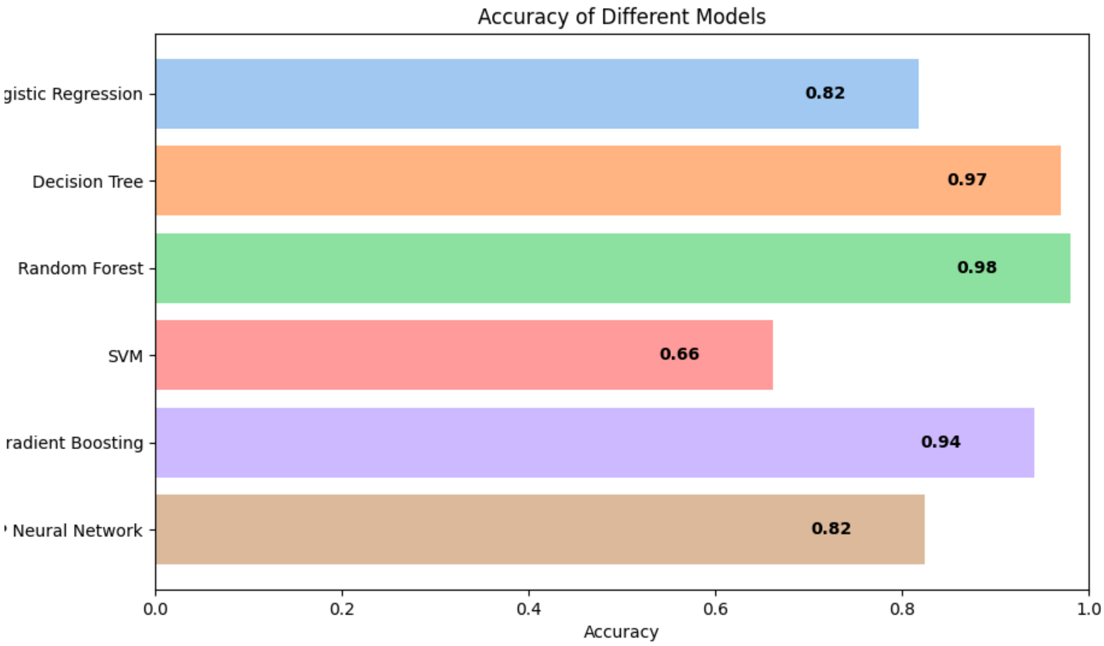

START
Collecting Patients Details
Creating a dataset
Data Pre Processing
Splitting Dataset
Train Data
Test Data
Machine learning algorithms logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting, and neural networks.
Classifier
Heart Disease Detected
No Heart Disease Detected

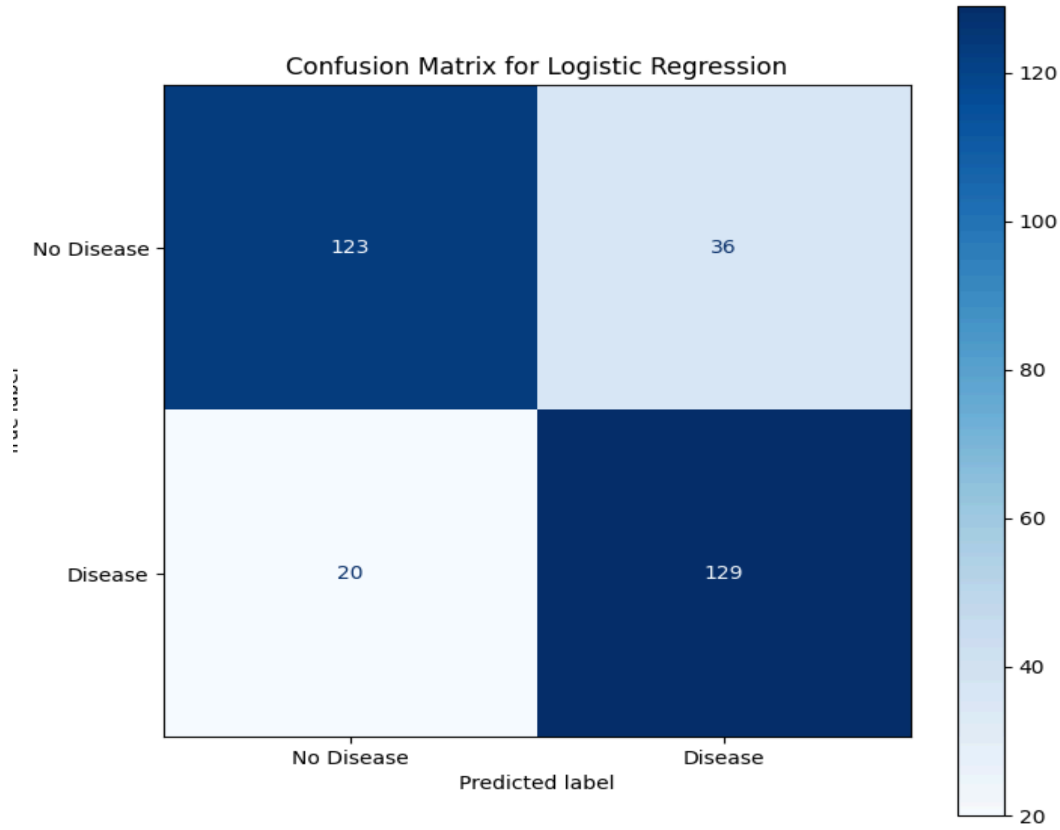
Models Comparison & Selection

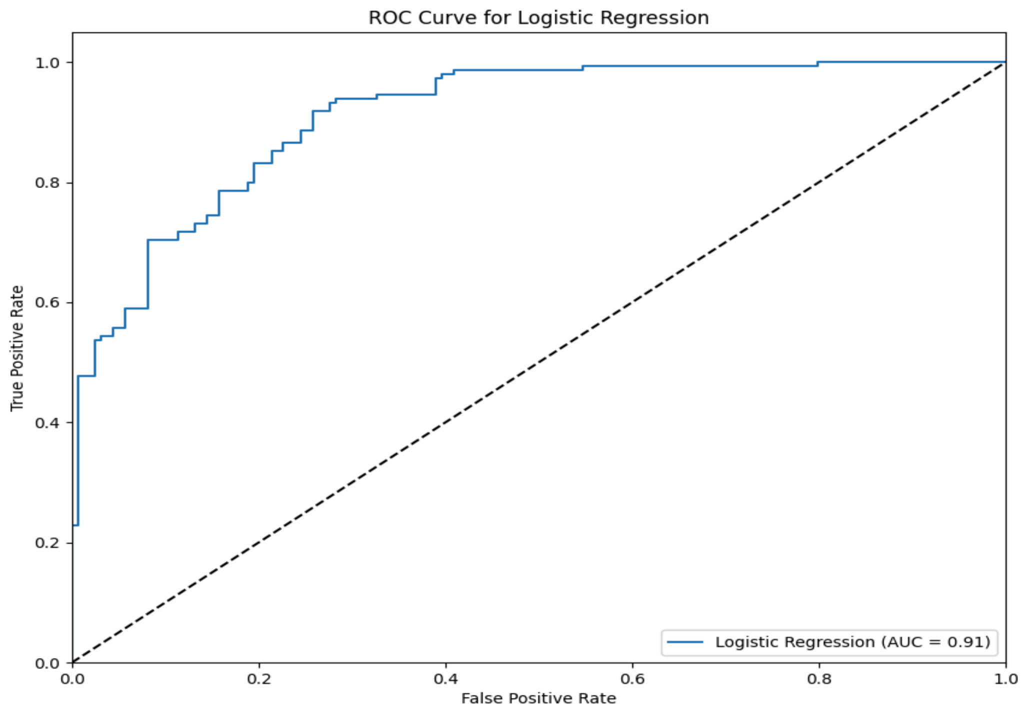
Several machine learning models are evaluated, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and ensemble methods like Gradient Boosting. The models are compared based on metrics such as accuracy, precision, recall, and F1-score.

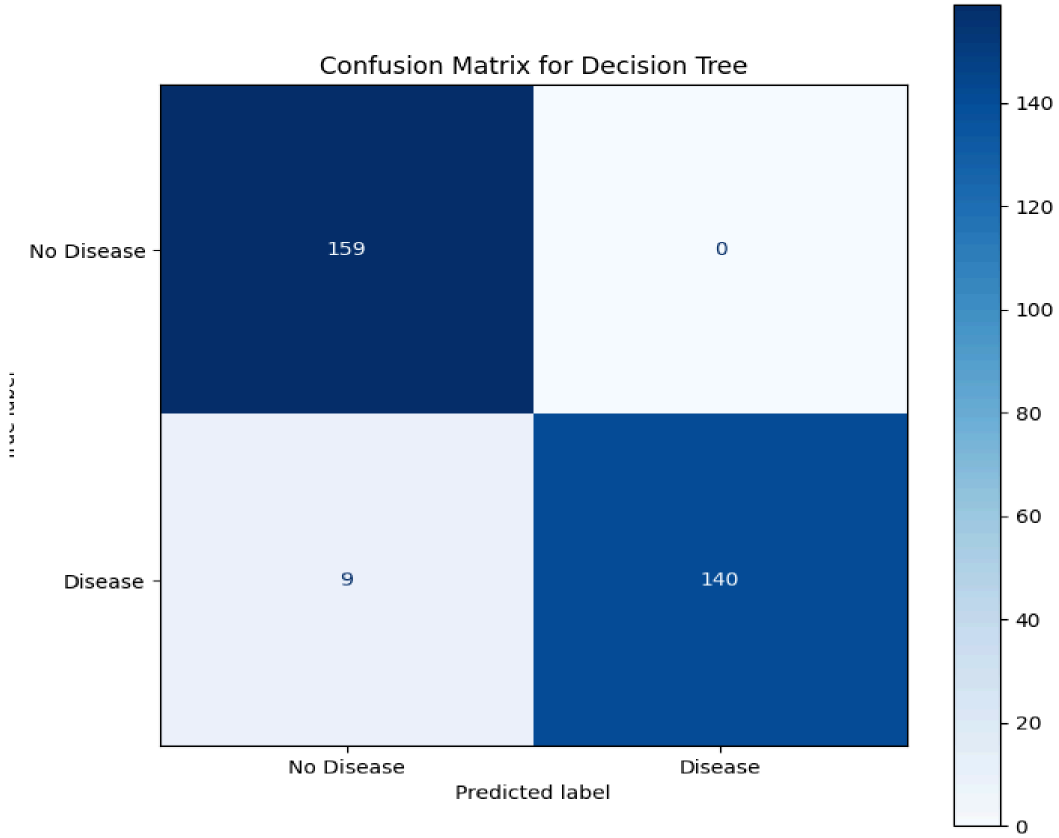
For choosing an appropriate machine learning algorithm for Heart Disease Prediction tasks, I'm using the mentioned models such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, ensemble methods like Gradient Boosting , neural network - MLP These models are commonly used in predictive analytics and have been shown effective in various healthcare applications such as Diabetes Prediction,Disease Outcome Prediction, and etc

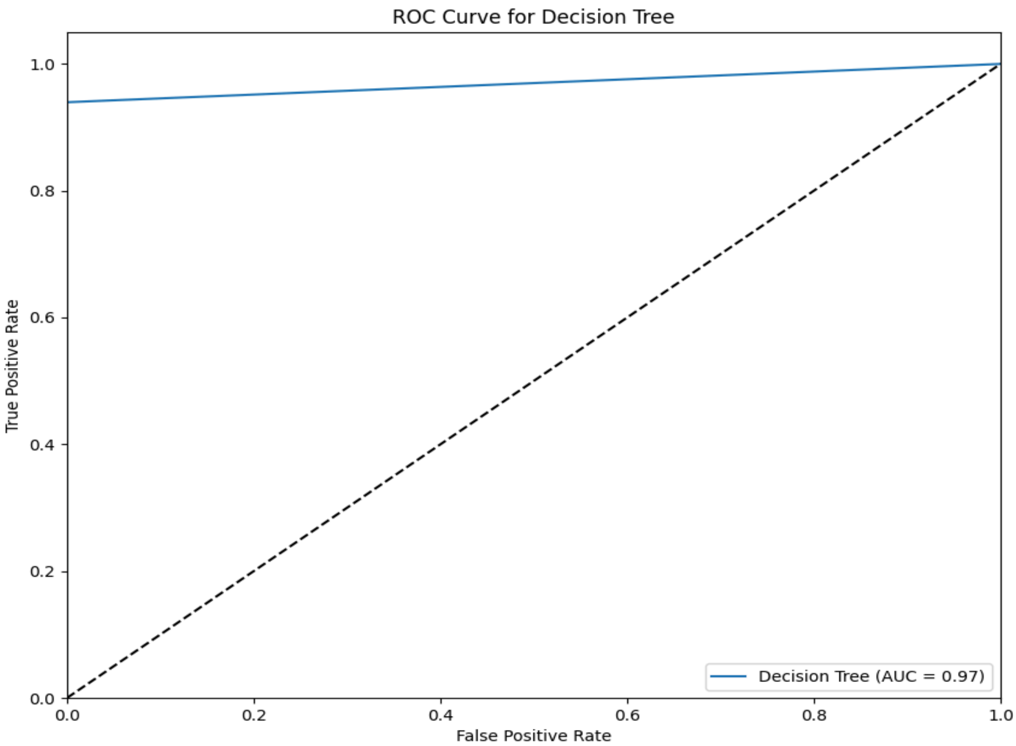
- Logistic Regression: Simple and interpretable, often used for binary classification tasks like predicting heart disease.
- Decision Trees: Can capture complex interactions and are easy to interpret.
- Random Forests: Improve upon decision trees by reducing overfitting and providing robust predictions.
- Support Vector Machines (SVM): Effective in high-dimensional spaces, good at handling complex relationships in data.
- Gradient Boosting: Builds models sequentially, focusing on correcting errors of previous models, leading to high accuracy.
- Multi-layer Perceptron (MLP): Neural networks capable of learning complex patterns, suitable for tasks with non-linear relationships.

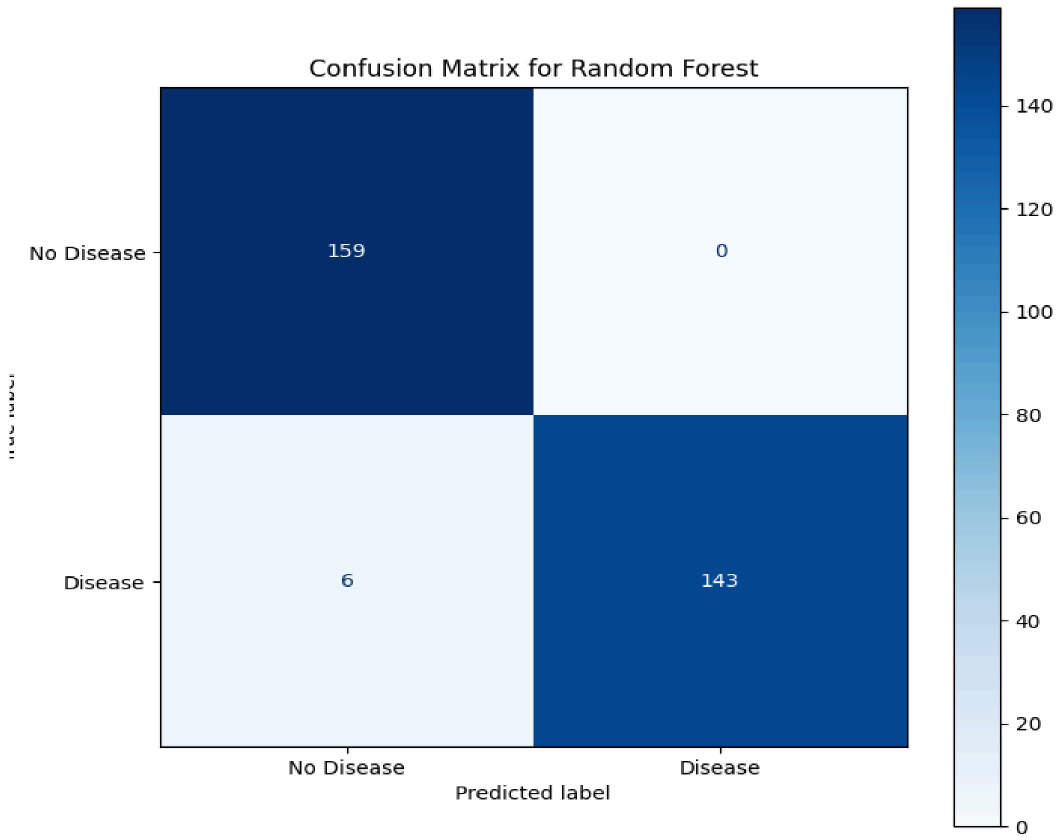


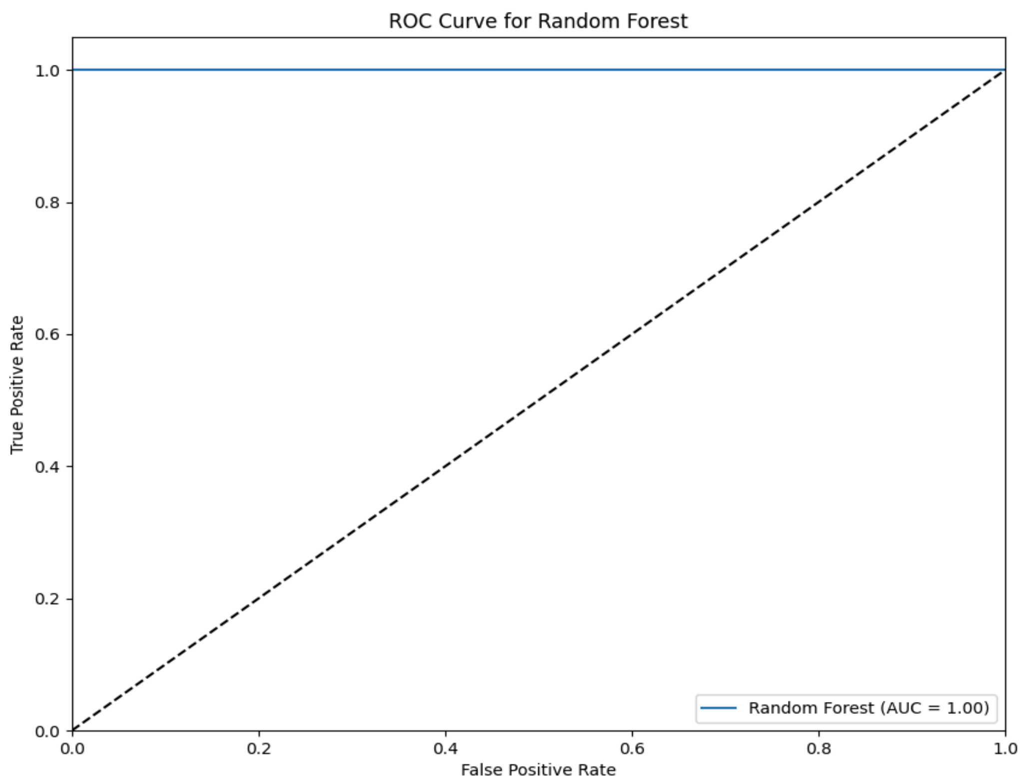

Confusion Matrix for Logistic Regression
No Disease
Disease
123
36
20
129
No Disease
Disease
Predicted label
120
100
80
60
40
20

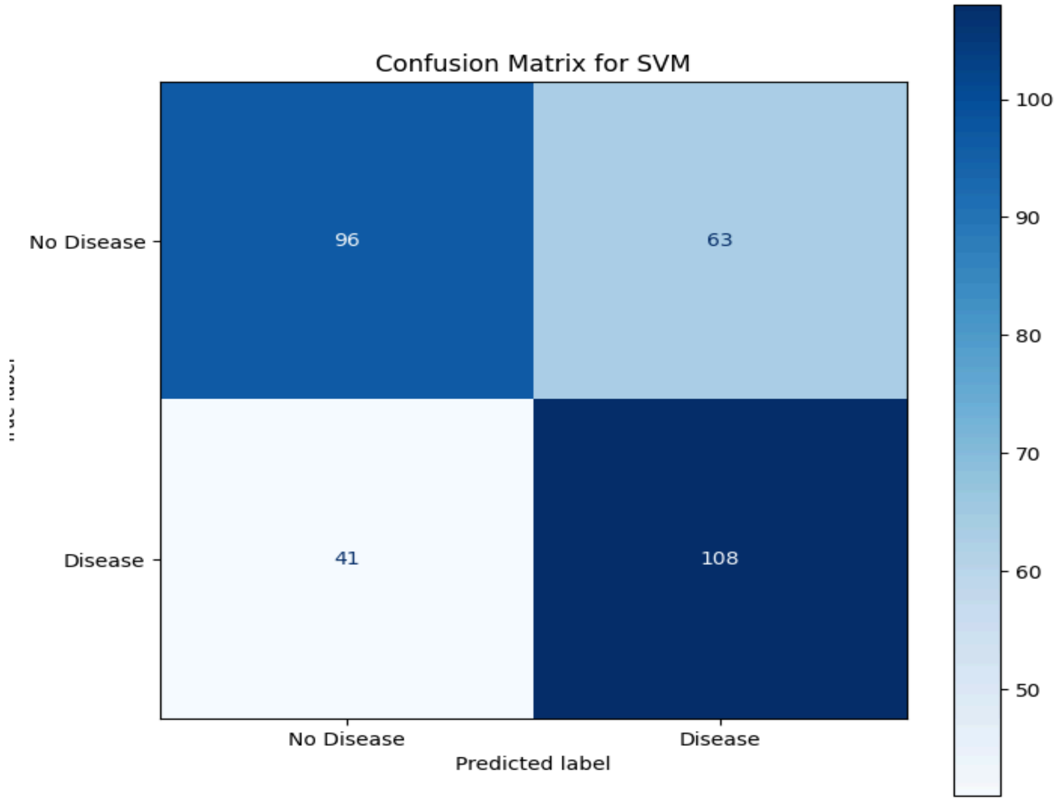

ROC Curve for Logistic Regression
True Positive Rate
False Positive Rate
Logistic Regression (AUC = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0

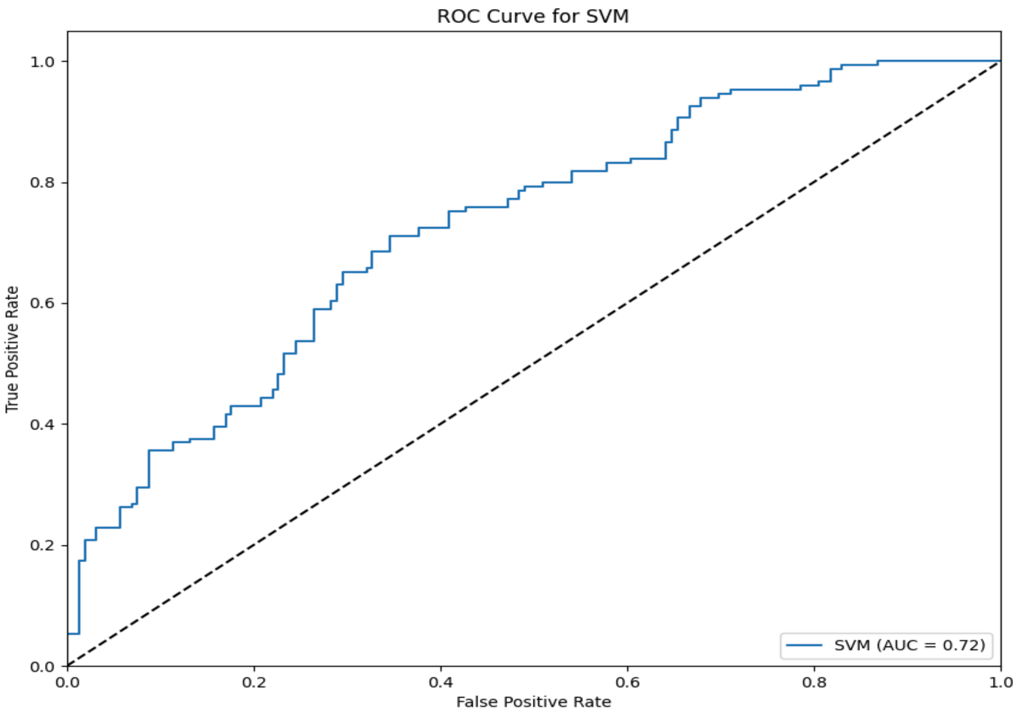

Confusion Matrix for Decision Tree
No Disease
Disease
159
0
9
140
No Disease
Disease
Predicted label
140
120
100
80
60
40
20
0

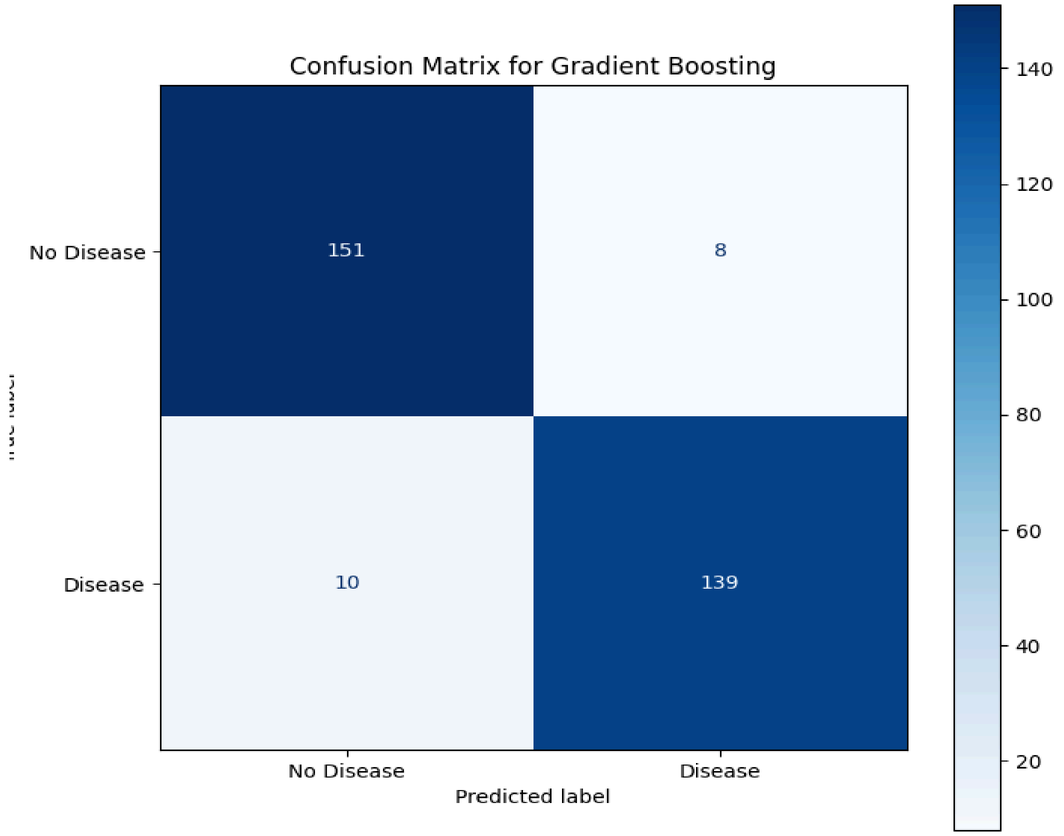

ROC Curve for Decision Tree
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Decision Tree (AUC = 0.97)

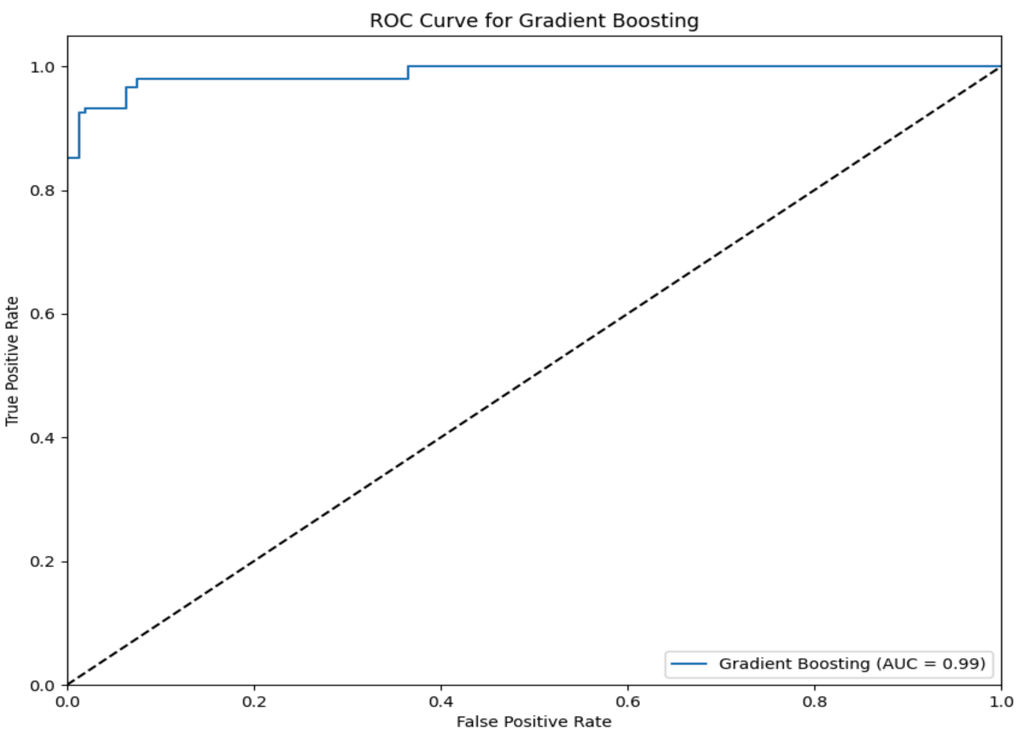

Confusion Matrix for Random Forest
No Disease
Disease
159
0
6
143
No Disease
Disease
Predicted label
140
120
100
80
60
40
20
0

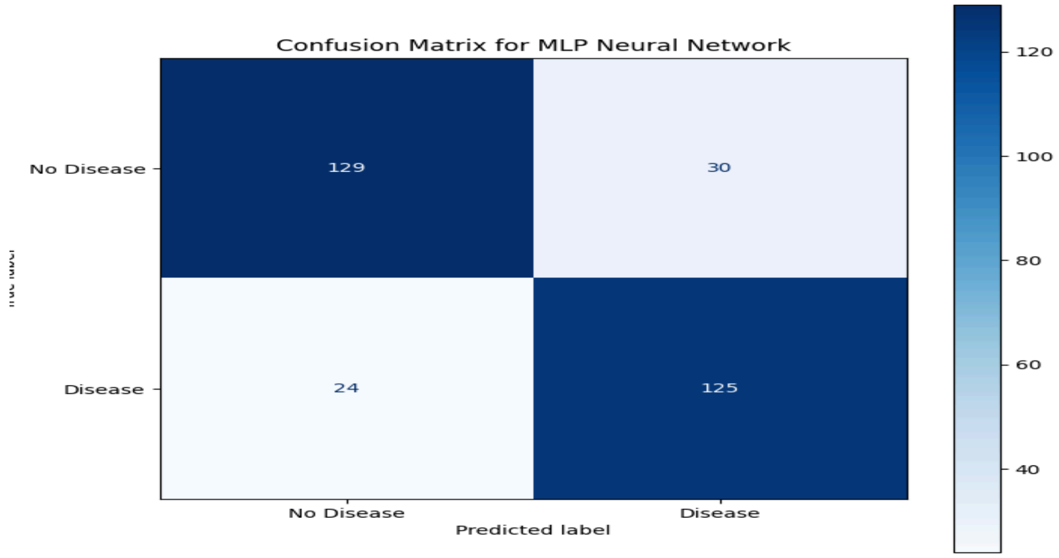

ROC Curve for Random Forest
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Random Forest (AUC = 1.00)

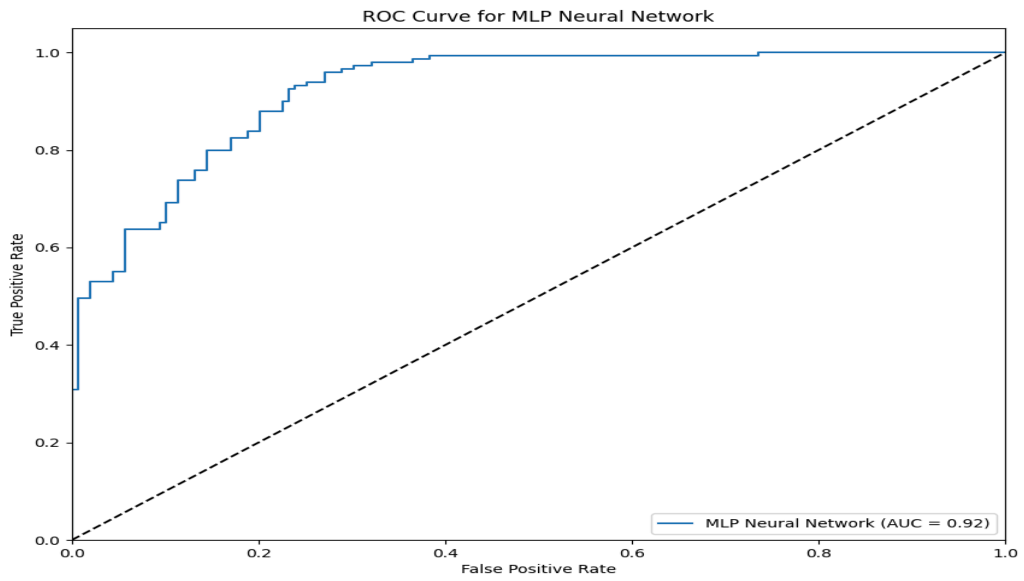

Confusion Matrix for SVM
No Disease
Disease
96
63
41
108
No Disease
Disease
Predicted label
100
90
80
70
60
50

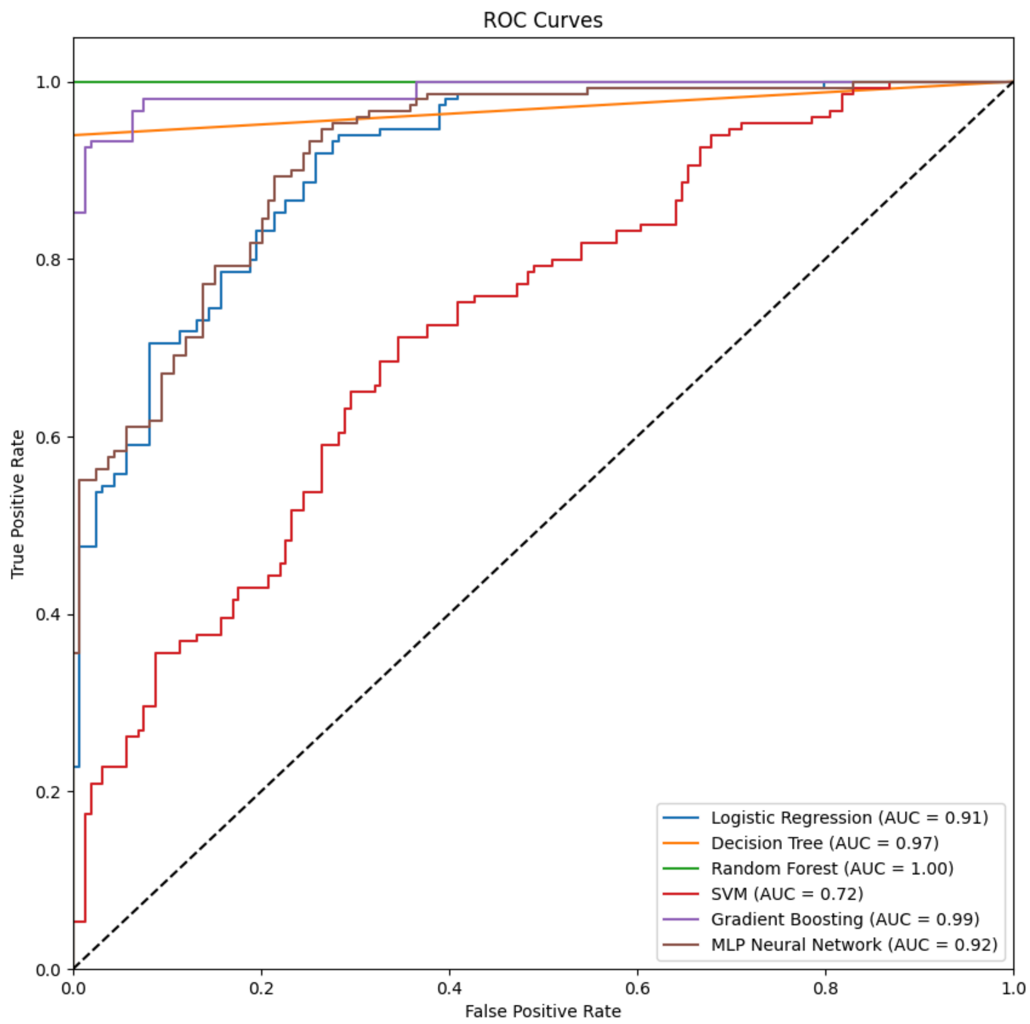

ROC Curve for SVM
True Positive Rate
False Positive Rate
SVM (AUC = 0.72)
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix for Gradient Boosting
No Disease
Disease
151
8
10
139
No Disease
Disease
Predicted label
140
120
100
80
60
40
20


ROC Curve for Gradient Boosting
True Positive Rate
False Positive Rate
Gradient Boosting (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix for MLP Neural Network
No Disease
Disease
129
30
24
125
No Disease
Disease
Predicted label
120
100
80
60
40





Logistic Regression:

- Accuracy: 0.82 , F1-score (weighted avg): 0.818
- Insights: Logistic Regression performs decently with balanced precision and recall but has lower overall accuracy compared to other models.

Decision Tree:

- Accuracy: 0.96,F1-score (weighted avg): 0.971
- Insights: Decision Tree shows very high accuracy and F1-scores, especially with perfect precision for class 0 and high recall for both classes. It indicates strong performance and simplicity.

Random Forest:

- Accuracy: 0.99,F1-score (weighted avg): 0.980
- Insights: Random Forest achieves near-perfect accuracy and F1-scores, demonstrating robust performance with perfect precision and high recall for both classes. It indicates excellent predictive capability.

Support Vector Machine (SVM):

- Accuracy: 0.66, F1-score (weighted avg): 0.661
- Insights: SVM shows lower accuracy and F1-scores compared to other models, particularly in recall for class 0, suggesting suboptimal performance for this dataset.

Gradient Boosting:

- Accuracy: 0.94,F1-score (weighted avg): 0.942
- Insights: Gradient Boosting performs well with balanced precision and recall, although slightly lower than Random Forest in overall metrics but still robust.

MLP Neural Network:

- Accuracy: 0.82 , F1-score (weighted avg): 0.850
- Insights: MLP Neural Network shows competitive performance with good recall for class 1 but slightly lower precision compared to Random Forest and Decision Tree.

Based on the highest accuracy and F1-score, Random Forest suits to be the best fit for this heart disease prediction task. It provides near-perfect accuracy, precision, and recall, indicating reliable predictions across both classes.

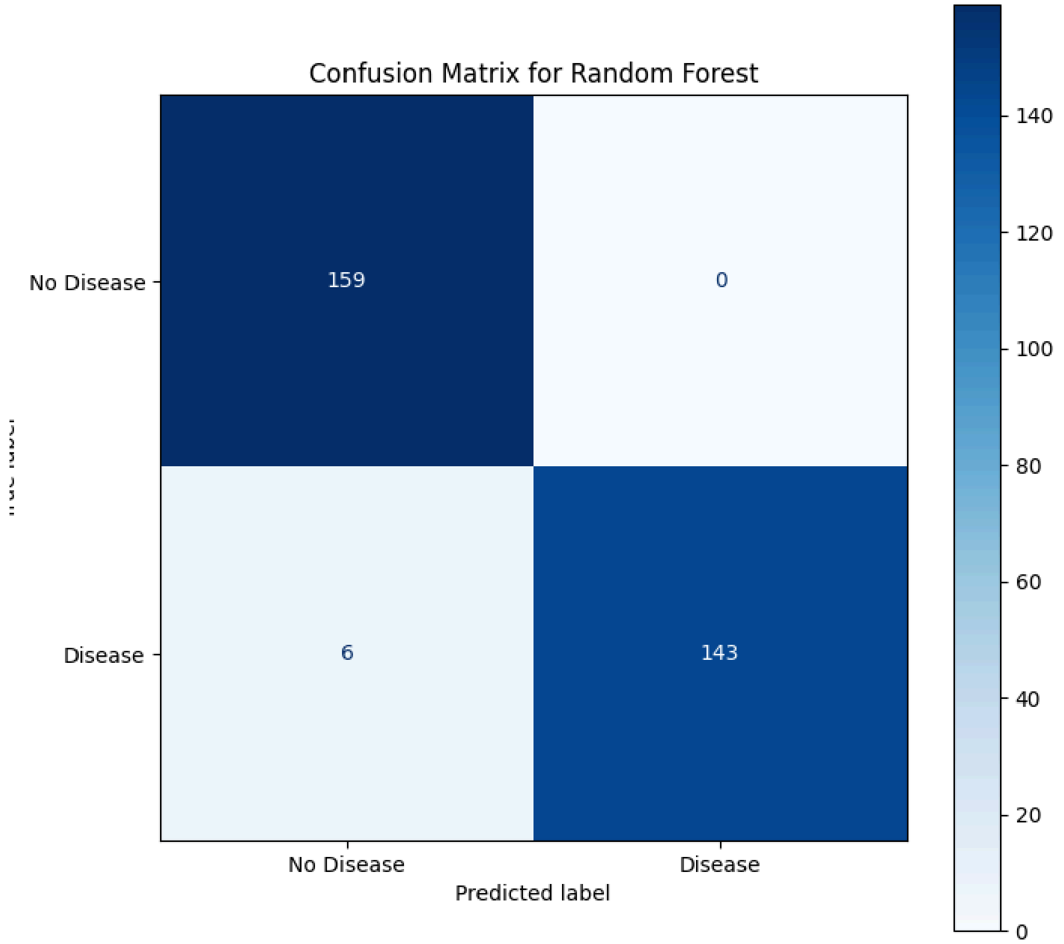
Therefore, ***Random Forest*** *is used as the optimal model for heart disease prediction based on the evaluation metrics.*

Model Training

The Random Forest model's best parameters include using the Gini criterion, 300 estimators, default auto setting for max features, minimum samples per leaf of 1, and minimum samples per split of 2. Achieving a cross-validation score of 0.97 highlights its robust performance in predicting heart disease based on input features. This model configuration emphasizes comprehensive exploration of feature importance and ensemble learning for reliable classification outcomes.

- Accuracy for Random Forest classifier: **0.9805194805194806**
- Precision for Random Forest classifier: **1.0**
- Recall for Random Forest classifier: **0.959731543624161**
- F1-Score for Random Forest classifier: **0.9794520547945206**
- ROC-AUC for Random Forest classifier: **0.9999999999999999**

● Confusion Matrix
○ True Positives (TP): The value at the top-left corner (159) represents the number of correctly classified positive cases. The model correctly predicted these cases as positive.
○ True Negatives (TN): The value at the bottom-right corner (143) represents the number of correctly classified negative cases. The model correctly predicted these cases as negative.
○ False Positives (FP): The value in the top-right corner (0) represents the number of incorrectly classified cases. The model predicted as there is no negative cases as positive (Type I error).
○ False Negatives (FN): The value in the bottom-left corner (6) represents the number of missed positive cases. The model predicted these positive cases as negative (Type II error).
● ROC Curve:
The ROC curve for your Random Forest classifier indicates outstanding performance, with an **AUC of 1**. This suggests that the model perfectly distinguishes between patients with and without heart disease.



Model Deployment

After training and evaluation, the best-performing machine learning model for predicting heart disease is selected and prepared for deployment in a web application. This model deployment process is crucial as it aims to make the predictive capabilities accessible and user-friendly to healthcare professionals and individuals concerned about their heart health. Deploying the best-performing heart disease prediction model in a Streamlit web application ensures that advanced predictive capabilities are made accessible, user-friendly, and impactful in improving health outcomes and patient care.

After deploying the model, we can create a localtunnel and make a connection to the streamlit and predict the values.

Example Input (using streamlit)

Example 1:

- Age: 55 years

- Sex: Female

- Chest Pain Type (cp): 3 (Non-anginal pain)

- Resting Blood Pressure (trestbps): 140 mm Hg

- Serum Cholesterol (chol): 240 mg/dL

- Fasting Blood Sugar (fbs): Yes (> 120 mg/dL)

- Resting ECG Results (restecg): 0 (Normal)

- Maximum Heart Rate Achieved (thalach): 150 bpm

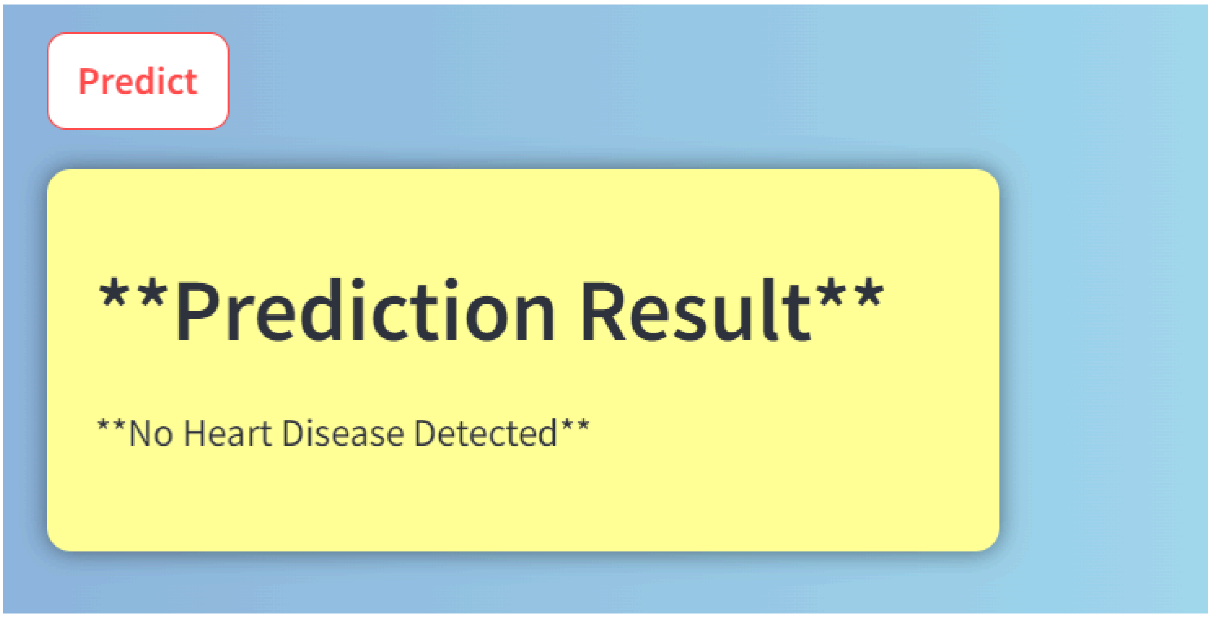
- Exercise Induced Angina (exang): No

- ST Depression Induced by Exercise (oldpeak): 1.0

- Slope of the Peak Exercise ST Segment (slope): 1 (Flat)

- Number of Major Vessels Colored by Fluoroscopy (ca): 0

- Thalassemia (thal): 3 (Normal)



Example 2:

- Age: 60 years

- Sex: Male

- Chest Pain Type (cp): 1 (Typical angina)

- Resting Blood Pressure (trestbps): 160 mm Hg

- Serum Cholesterol (chol): 280 mg/dL

- Fasting Blood Sugar (fbs): Yes (> 120 mg/dL)

- Resting ECG Results (restecg): 2 (Showing probable or definite left ventricular hypertrophy)

- Maximum Heart Rate Achieved (thalach): 130 bpm

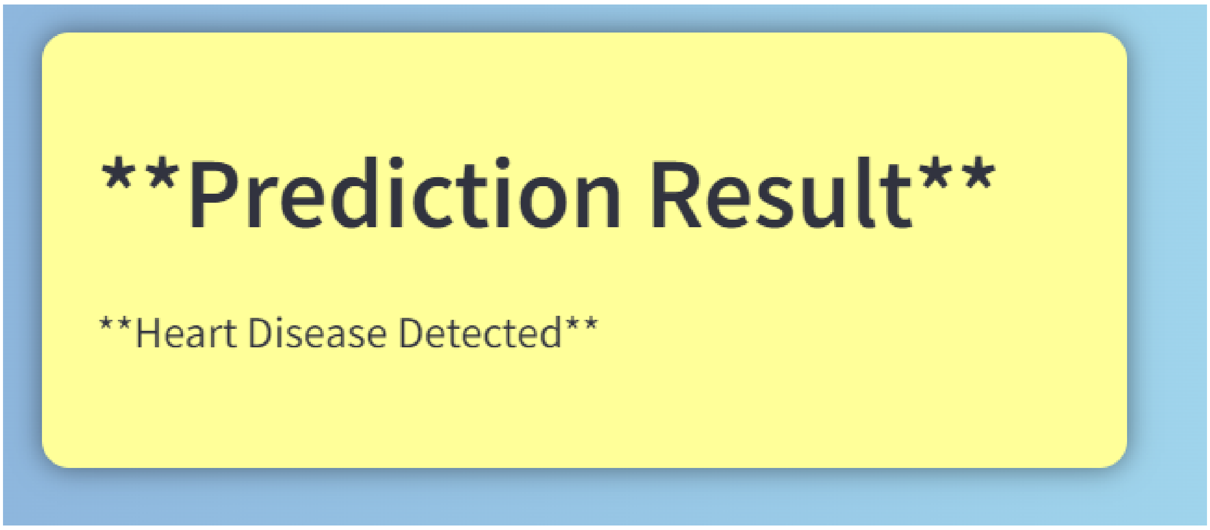
- Exercise Induced Angina (exang): Yes

- ST Depression Induced by Exercise (oldpeak): 3.0

- Slope of the Peak Exercise ST Segment (slope): 0 (Upsloping)

- Number of Major Vessels Colored by Fluoroscopy (ca): 2

- Thalassemia (thal): 6 (Fixed defect)


Prediction Result
Heart Disease Detected

Conclusion and Future Scope

The heart is a vital organ, but heart disease remains a significant global concern due to its increasing prevalence. Early detection and accurate prediction are crucial for effective management and treatment. This study focuses on developing a machine learning model that can predict the initial stages of heart disease, aiming to reduce uncertainty and healthcare costs associated with diagnosis.

Importance of Heart Disease Prediction

The primary goal of this research is to create a highly accurate machine learning model capable of identifying early signs of heart disease. By leveraging data-driven insights, such as those obtained from the UCI Machine Learning Repository, we can enhance diagnostic capabilities and support timely interventions.

Performance Evaluation

Several machine learning algorithms, including Random Forest, were evaluated based on metrics such as accuracy, precision, recall, and F1-score. Random Forest emerged as the top performer, achieving an impressive accuracy rate of 98%. This indicates its potential as a reliable tool for early detection of heart disease.

Future Directions

Moving forward, there are several avenues for expanding and improving this research:

- **Integration of Additional Data:** Incorporating more diverse datasets and medical attributes could further refine the model's predictive accuracy.
- **Exploration of Advanced Techniques:** Investigating ensemble methods or deep learning algorithms could enhance the model's ability to detect subtle patterns indicative of heart disease.
- **Application to Other Diseases:** The methodologies developed here can be extended to predict other health conditions such as cardiovascular diseases, diabetes, and various types of cancer.
- **Longitudinal Analysis:** Analyzing longitudinal patient data over time could provide insights into disease progression and personalized treatment strategies.

References:

- American Heart Association
- Framingham Heart Study
- SCORE Project
- ATP III Guidelines
- ACC/AHA Risk Calculator
- Machine Learning in Heart Disease Prediction
- Deep Learning for Cardiovascular Diseases
- Predictive Modeling in Cardiovascular Disease
- Machine Learning Techniques for Clinical Decision Support in Cardiovascular Disease
- ESC Clinical Practice Guidelines
- World Health Organization - Cardiovascular diseases (CVDs)
- UCI Machine Learning Repository - Heart Disease Data Set
- Centers for Disease Control and Prevention - Heart Disease Risk Factors