Heart Disease Prediction Project Report

1. Introduction

The objective of this project is to predict the presence of heart disease based on several health-related features using machine learning techniques like Logistic Regression and Decision Tree. Predicting heart disease is crucial in the medical field as it enables early detection and prevention. This report includes data preprocessing, exploratory data analysis (EDA), and model evaluation to assess the prediction accuracy.

2. Dataset Overview

The Heart Disease dataset contains 1025 observations with 14 features. The target variable is 'target', which indicates whether a patient has heart disease (1: Yes, 0: No). Below are the key features used in the analysis.

- Age: Age of the patient
- Sex: Gender of the patient (1 = male, 0 = female)
- CP: Chest pain type
- Trestbps: Resting blood pressure
- Chol: Serum cholesterol in mg/dl
- FBS: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- Restecg: Resting electrocardiographic results
- Thalach: Maximum heart rate achieved
- Exang: Exercise induced angina (1 = yes, 0 = no)
- Oldpeak: ST depression induced by exercise relative to rest
- Slope: Slope of the peak exercise ST segment
- Ca: Number of major vessels (0-3) colored by fluoroscopy
- Thal: Thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect)

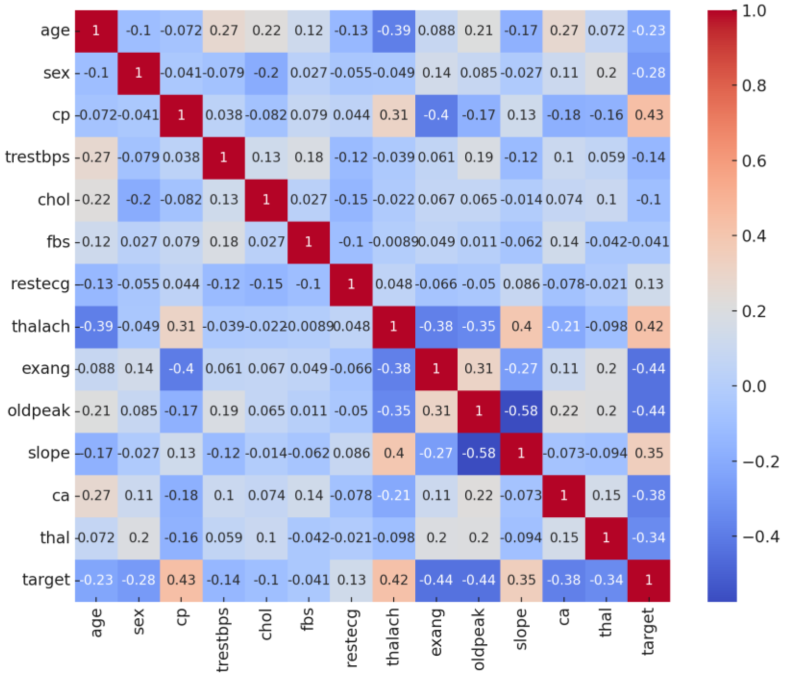
3. Data Preprocessing

Before training the model, the dataset underwent several preprocessing steps:
3.1 Handling Missing Values: No missing values were found in the dataset.
3.2 Encoding Categorical Variables: The categorical variables were encoded using one-hot encoding to convert them into numerical form. Variables like 'cp', 'restecg', and 'thal' were transformed into dummy variables.
3.3 Feature Scaling: Numerical variables such as age, cholesterol, and resting blood pressure were standardized using StandardScaler.

4. Exploratory Data Analysis (EDA)

4.1 Distribution of Target Variable: A count plot was used to visualize the distribution of heart disease cases in the dataset.
4.2 Correlation Matrix: A correlation heatmap was generated to understand the relationships between features. This helped identify important features influencing heart disease.
4.3 Age and Cholesterol Analysis: Box plots were created to analyze the distribution of age and cholesterol levels across patients with and without heart disease.

Below is the correlation matrix for the dataset:



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.1	-0.072	0.27	0.22	0.12	-0.13	-0.39	0.088	0.21	-0.17	0.27	0.072	-0.23
sex	-0.1	1	-0.041	-0.079	-0.2	0.027	-0.055	-0.049	0.14	0.085	-0.027	0.11	0.2	-0.28
cp	-0.072	-0.041	1	0.038	-0.082	0.079	0.044	0.31	-0.4	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.079	0.038	1	0.13	0.18	-0.12	-0.039	0.061	0.19	-0.12	0.1	0.059	-0.14
chol	0.22	-0.2	-0.082	0.13	1	0.027	-0.15	-0.022	0.067	0.065	-0.014	0.074	0.1	-0.1
fbs	0.12	0.027	0.079	0.18	0.027	1	-0.1	-0.0089	0.049	0.011	-0.062	0.14	-0.042	-0.041
restecg	-0.13	-0.055	0.044	-0.12	-0.15	-0.1	1	0.048	-0.066	-0.05	0.086	-0.078	-0.021	0.13
thalach	-0.39	-0.049	0.31	-0.039	-0.022	-0.0089	0.048	1	-0.38	-0.35	0.4	-0.21	-0.098	0.42
exang	0.088	0.14	-0.4	0.061	0.067	0.049	-0.066	-0.38	1	0.31	-0.27	0.11	0.2	-0.44
oldpeak	0.21	0.085	-0.17	0.19	0.065	0.011	-0.05	-0.35	0.31	1	-0.58	0.22	0.2	-0.44
slope	-0.17	-0.027	0.13	-0.12	-0.014	-0.062	0.086	0.4	-0.27	-0.58	1	-0.073	-0.094	0.35
ca	0.27	0.11	-0.18	0.1	0.074	0.14	-0.078	-0.21	0.11	0.22	-0.073	1	0.15	-0.38
thal	0.072	0.2	-0.16	0.059	0.1	-0.042	-0.021	-0.098	0.2	0.2	-0.094	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.1	-0.041	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1

5. Logistic Regression Model

Logistic regression was chosen as it is a widely used classification algorithm for binary problems such as heart disease prediction. The model was trained on 80% of the data and tested on 20%.
The logistic regression model achieved an accuracy of 82% on the test data. Below are the key metrics and visualizations for the model's performance.

The confusion matrix and ROC curve were used to evaluate the model's performance, demonstrating good prediction results.

6. Decision Tree Model

A Decision Tree model was also applied to the dataset. Although the model achieved 100% accuracy, this indicates potential overfitting. Decision trees can easily overfit the data, especially when there are many features or small datasets. The performance of the Decision Tree is visualized using a confusion matrix and ROC curve.

7. Conclusion

This project successfully demonstrated the application of Logistic Regression and Decision Tree models to predict heart disease. While the logistic regression model showed promising results with an accuracy of 82%, the Decision Tree model overfitted the data. Future improvements could involve using regularization techniques or ensemble methods to improve model robustness.