Machine Learning Analysis of Heart Disease Data Report

Task	Task Description	Team Member
Task 1	Data Acquisition and Project Setup	Mohamad Hassan
Task 2	Data Preprocessing and Exploratory Data Analysis	Wassem
Task 3	Model Implementation and Evaluation	Mohamad Khalil
Task 4	Reporting and Documentation	Youssef E. ElKalla

Heart Disease Dataset Report

1. Introduction

This project applies machine learning techniques to a medical dataset related to heart disease. The main objective is to analyze patient data and solve three different machine learning tasks: regression, classification, and clustering.

First, a regression model is used to predict the age of patients based on clinical features. Second, a classification model is used to predict whether a patient has heart disease or not. Third, a clustering model is applied to identify natural groups of patients without using predefined labels.

The Heart Disease Dataset contains medical and demographic information that helps in understanding patterns associated with heart conditions. This project demonstrates how different machine learning approaches can be applied to the same dataset to extract meaningful insights.

2. Data Preprocessing and Exploratory Data Analysis

2.1 Dataset Description

The dataset contains 270 patient records with 14 features. These features include age, sex, blood pressure, cholesterol level, heart rate, and other medical measurements related to heart health.

Two target variables are used in this project. Age is used as the target variable for the regression task, while heart disease status is used as the target variable for the classification task.

2.2 Data Cleaning

The dataset was inspected for missing values. Although no significant missing data was observed, a median imputation step was applied as a safety preprocessing measure to ensure data robustness and avoid potential training issues.

2.3 Categorical Feature Encoding

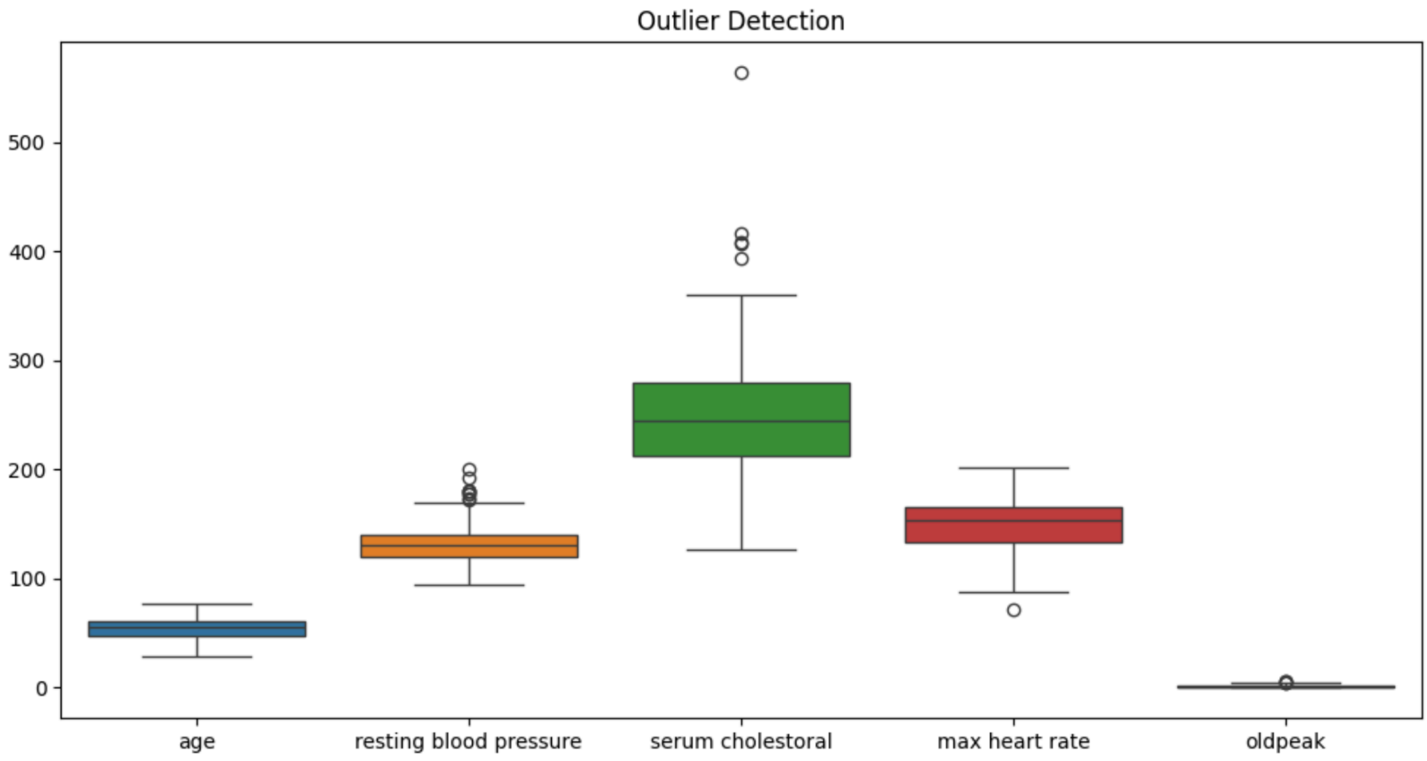
Categorical features such as sex and chest pain type were converted into numerical values using one-hot encoding.

This step was necessary because machine learning algorithms require numerical input to perform calculations and make predictions.

2.4 Outlier Analysis

Boxplots were used to visually detect potential outliers in important numerical features, including age, resting blood pressure, serum cholesterol, maximum heart rate, and oldpeak.

Some extreme values were observed, particularly in serum cholesterol. These values were not removed, as they may represent valid medical conditions and removing them could reduce the realism of the dataset.



Boxplot showing potential outliers in key numerical features

2.5 Feature Scaling and Data Splitting

All features were scaled using StandardScaler to ensure equal contribution of each feature during model training. The dataset was then split into 80% training data and 20% testing data. Stratified splitting was applied for classification tasks to preserve class distribution.

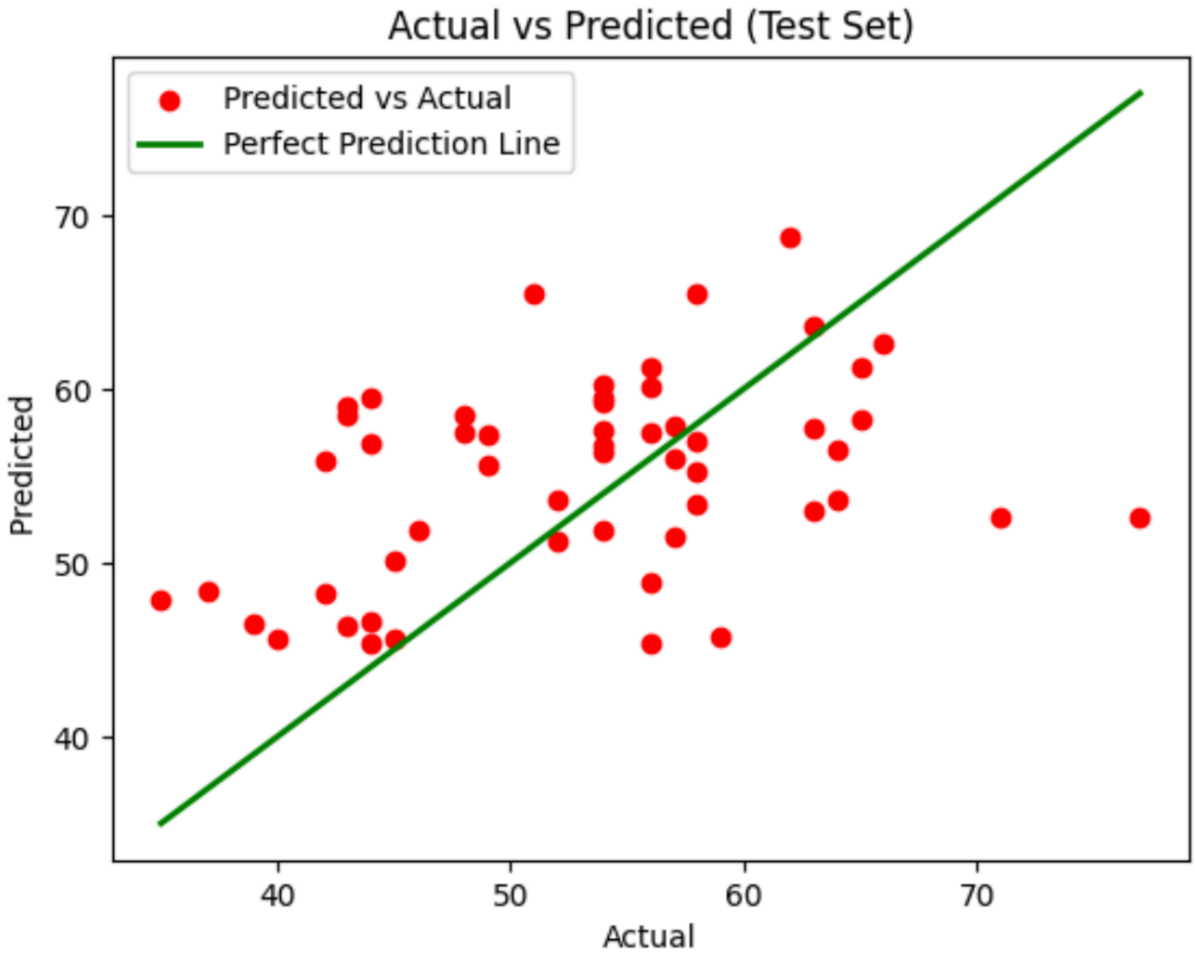
---

3. Modeling and Results

3.1 Regression: Linear Regression

A Linear Regression model was trained to predict patient age using the remaining clinical features. Prior to training, all features were standardized using StandardScaler.

The model was evaluated using RMSE and R² score. The results show a relatively high RMSE and a low R² value, indicating that age is not strongly determined by the available clinical measurements. This behavior is expected and reflects realistic medical relationships rather than model overfitting

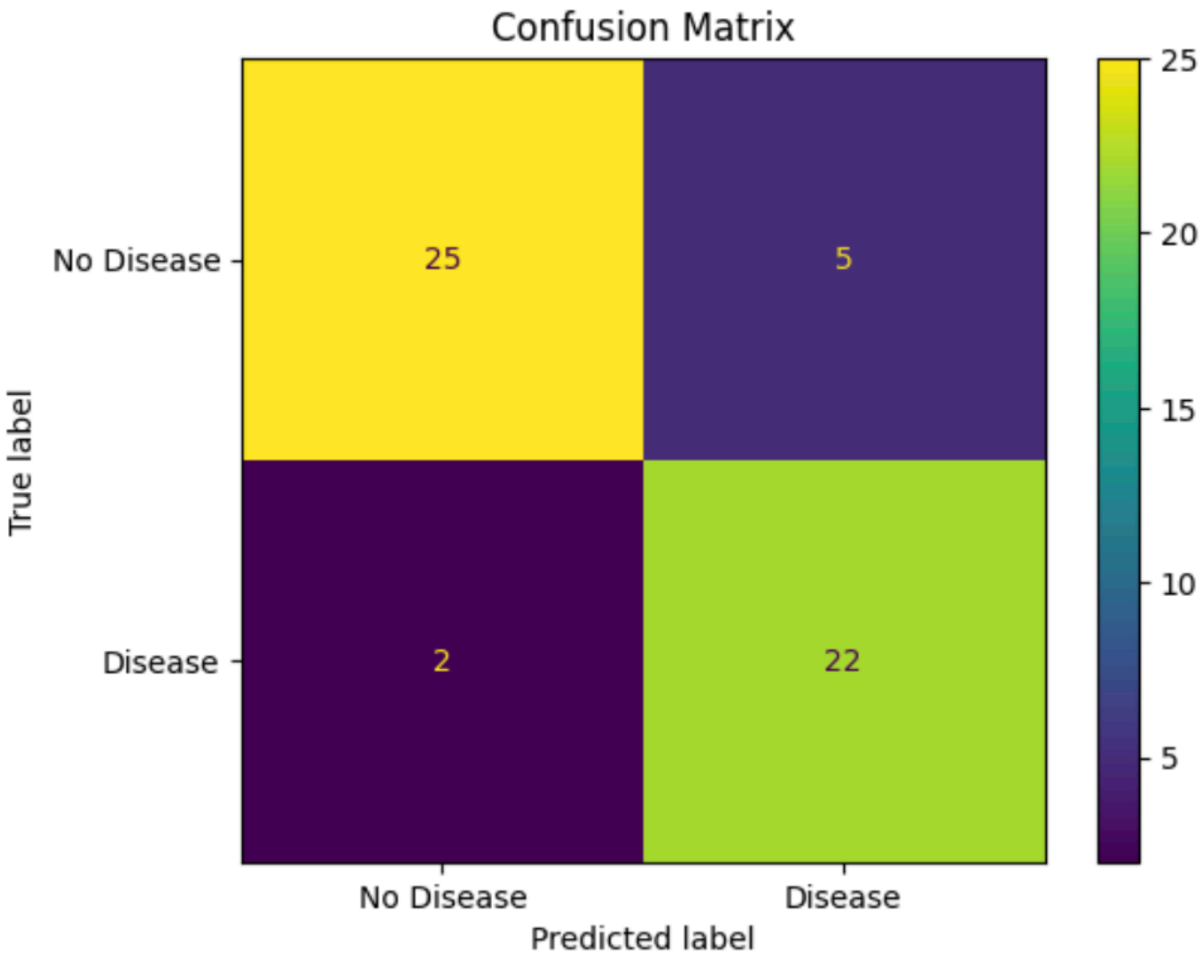


Actual vs predicted age values for the test set

3.2 Classification: Logistic Regression

Logistic Regression was used to classify patients based on the presence of heart disease. The target variable was mapped into a binary format where 0 represents no heart disease and 1 represents heart disease.

Feature scaling was applied before training the model. The model achieved high accuracy, precision, and recall, indicating strong classification performance. The confusion matrix further confirms the model's ability to correctly distinguish between healthy and diseased patients



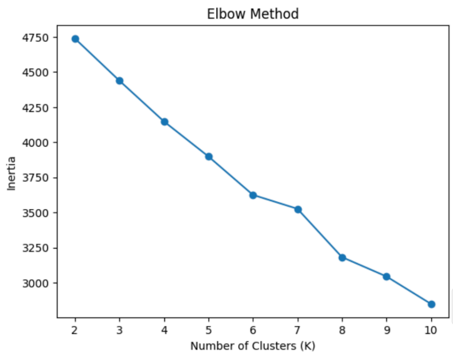
Confusion matrix of the Logistic Regression model.

3.3 Clustering: K-Means

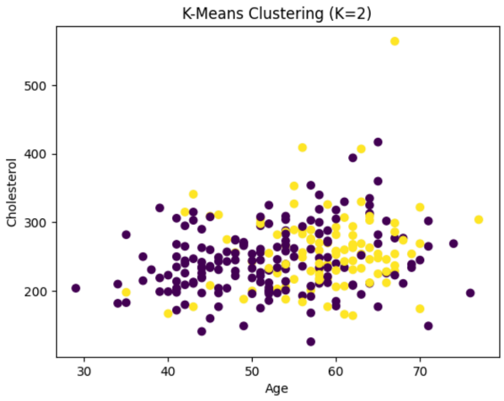
K-Means clustering was applied to the dataset after removing the target variable. All features were scaled using StandardScaler to ensure fair distance calculations.

The Elbow Method was used to determine the optimal number of clusters, and K = 2 was selected. The clustering quality was evaluated using the silhouette score, which showed moderate cluster separation.

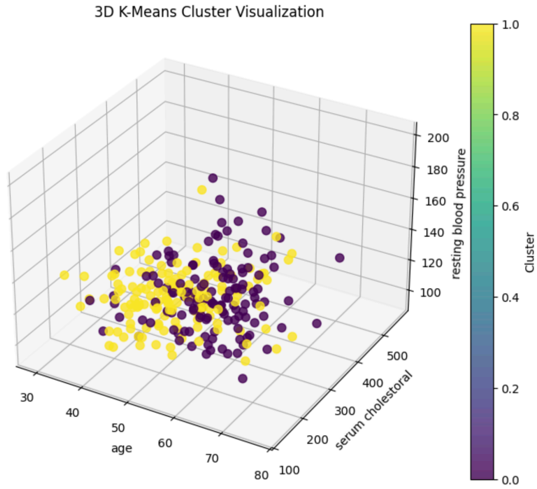
Two-dimensional visualization using age and serum cholesterol was used to display the clusters. Additionally, a three-dimensional visualization based on age, serum cholesterol, and resting blood pressure was created to better interpret patient groupings across multiple features



Elbow Method used to determine the optimal number of clusters.



2D visualization of K-Means clusters using age and serum cholesterol.



3D visualization of K-Means clustering using age, serum cholesterol, and resting blood pressure.

• Conclusion

This project applied basic machine learning methods to a heart disease dataset. Data preprocessing and exploratory analysis were performed before modeling.

Linear Regression was used to predict age and showed limited performance, which is expected because age is not strongly related to the other clinical features. Logistic Regression performed well in predicting heart disease, showing good accuracy and reliability. K-Means clustering identified reasonable patient groups, with some overlap, which is normal in medical data.

Overall, the results are realistic and suitable for an academic machine learning project.

• Future Work

In the future, more advanced models such as Random Forest or Gradient Boosting can be tested. Feature selection or dimensionality reduction methods like PCA may also be applied. Using a larger dataset could further improve the results.

• Google Colab Link

👉 https://colab.research.google.com/drive/1DdymlbnEZhgytZW4MXqvZRQqnnAktbDV?usp=sharing