Heart Disease Prediction Project Report

1. Objective

The objective of this project was to build a machine learning model capable of predicting whether a patient is at risk of heart disease. The model uses patient health data to make predictions and is intended to aid in early detection and preventative healthcare.

2. Dataset Description

We used the Heart Disease dataset from the UCI Machine Learning Repository. It contains 303 records and 14 features including age, sex, chest pain type, cholesterol, fasting blood sugar, ECG results, exercise-induced angina, and more. After removing duplicates, we were left with 302 records. The target column indicates whether a person is at risk (1) or not (0).

3. Data Preprocessing

We checked for missing values and found none. Duplicate records (723) were removed to ensure data quality. Categorical features were encoded to prepare them for modeling. Outliers were visualized using boxplots to better understand data distribution.

4. Exploratory Data Analysis (EDA)

EDA was performed using various visualizations to understand trends and patterns in the data. Count plots, histograms, and correlation heatmaps revealed that features like chest pain type, oldpeak, ECG abnormalities, and vessel count have noticeable differences between healthy and at-risk groups.

5. Model Training

We applied Logistic Regression to build a binary classification model. The data was split into training and testing sets. The model was trained using the training set and evaluated on the testing set.

6. Model Evaluation

The Logistic Regression model achieved an accuracy of 82% and an AUC score of 0.87, indicating strong

performance in distinguishing between patients at risk and not at risk. A confusion matrix and ROC curve further validated the model's reliability.

7. Feature Importance

We analyzed the logistic regression coefficients to determine which features had the greatest impact on predictions. The most influential features included chest pain type (typical angina), vessel count, thalassemia type, sex, oldpeak, and ECG results.

8. Final Insights

This project shows that interpretable models like logistic regression can provide valuable insights in the medical domain. Features such as chest pain type, ECG results, and vessel count significantly contribute to predicting heart disease risk. The model achieved solid performance and helps support early diagnostic decisions.