Name	**Abdul Naafeh**
Roll Number	**DHC-673**
Subject	**AI/Machine Learning**
Date of Submission	**05/09/2025**

WEEK 1: Disease Prediction Using Patient Data

Disease Prediction Using Machine Learning (Heart Disease Dataset)

1. Introduction

The objective of this task is to predict the presence of heart disease in patients using machine learning techniques. This involves data preprocessing, exploratory data analysis (EDA), and model building using two algorithms: Logistic Regression and Random Forest. The primary metric for evaluating model performance is accuracy.

2. Dataset Details

Dataset Name: Heart Disease Dataset
Number of Records: 303
Number of Features: 14 (including target)
Target Variable: Target (1 = Disease, 0 = No Disease)

- **Key Features:**
  - Age: Patient's age in years
  - Sex: 1 = male; 0 = female
  - ChestPain: Type of chest pain (categorical)
  - RestBP: Resting blood pressure (mm Hg)
  - Chol: Serum cholesterol (mg/dl)
  - Fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
  - Thal: Thalassemia (categorical: normal, fixed, reversable)

3. Data Preprocessing

- **Handling Missing Values:**
  Thal column had 2 missing values; filled using mode (most frequent value). No missing values in numeric columns.
- **Encoding:**
  Categorical columns (ChestPain, Thal) were label-encoded.
- **Feature Scaling:**
  All numeric features normalized to [0, 1] using MinMaxScaler.

4. Exploratory Data Analysis

- Statistical Summary:
  After preprocessing, descriptive statistics were generated using describe().
  Key observations:
  * Age ranged between 29 and 77 years.
  * Cholesterol values showed slight skewness.
- Correlation Analysis:
  A correlation heatmap was plotted to understand relationships between features and the target variable.

  Features like ChestPain, Ca, and Thal had higher correlation with heart disease.

5. Model Building

Two models were implemented:

- **Logistic Regression:** A linear model suitable for binary classification.
- **Random Forest Classifier:** An ensemble model using multiple decision trees for better performance.

6. Model Evaluation

Evaluation Metric: Accuracy (on the test set).

Model	**Accuracy**
Logistic Regression	**86.89%**
Random Forest	**91.80%**

Confusion Matrices: Both models performed well, but Random Forest provided slightly better accuracy and balance in predictions.

7. Results & Conclusion

Random Forest achieved the highest accuracy (91.80%) and is the better model for this dataset. Logistic Regression also performed well (86.89%), making it a good baseline model.

Future Improvements:

- Apply cross-validation for better generalization.
- Experiment with hyperparameter tuning and feature selection.