Project Report

Repository: noor12155221/FINAL-ML-Project
Google Colab Notebook: Another copy of Yet another copy of Final_ML_project.ipynb - Colab

---

I. Introduction

Project Goal

The objective of this project is to apply **Regression**, **Classification**, and **Clustering** techniques on the **Heart Disease dataset** to analyze relationships between medical features and heart disease outcomes.

Specifically, the goals are:

- **Regression:** Predict the cholesterol level (chol) based on age using Linear Regression.
- **Classification:** Classify whether a patient has heart disease using Logistic Regression.
- **Clustering:** Group patients into clusters based on their medical features using K-Means clustering.

Data Description

The dataset used in this project is the **Heart Disease dataset**.

- **Source:** Kaggle / UCI Machine Learning Repository
- **Size:** ~300 samples
- **Target Variable (Classification):** target (0 = No disease, 1 = Disease)
- **Features:** Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting Electrocardiographic Results, Maximum Heart Rate Achieved, Exercise-Induced Angina, ST Depression Induced by Exercise Relative to Rest, Slope of the Peak Exercise ST Segment, Number of Major Vessels Colored by Fluoroscopy, Thalassemia, Presence of Heart Disease.

II. Data Preprocessing & Exploratory Data Analysis (EDA)

Data Cleaning and Preprocessing

The following preprocessing steps were applied:

- Loaded the dataset from the GitHub repository inside Google Colab.
- Checked dataset shape and data types to understand feature structure.
- Split the dataset into **training and testing sets (80/20)**.
- Handled missing numerical values using **mean imputation**.
- Handled missing categorical values using **mode imputation**.
- Applied **One-Hot Encoding** for categorical variables.
- Used **StandardScaler** to normalize features for Regression, Classification, and Clustering tasks.

Key EDA Findings

- Age shows a weak to moderate relationship with cholesterol levels.
- Cholesterol values are widely distributed across patients.
- Some medical features show noticeable differences between patients with and without heart disease.

III. Modeling and Results

A. Regression Results

- **Model Used:** Linear Regression
- **Features Used:** Age
- **Target Variable:** Cholesterol (chol)
- **Evaluation Metrics:** MSE, RMSE, R^2 Score

Final Results:

- Mean Squared Error (MSE): **3362.13**
- Root Mean Squared Error (RMSE): **57.98**
- R^2 Score: **0.0176**

Interpretation:
The R^2 score of **0.0176** indicates that age alone explains only about **1.76%** of the variance in cholesterol levels, which confirms that cholesterol is influenced by many additional medical factors beyond age. The relatively high RMSE reflects this limited predictive power.

Residual plots show that errors are centered around zero but widely spread, suggesting that a **single-feature linear model is insufficient** for accurate cholesterol prediction.

B. Classification Results

- **Model Used:** Logistic Regression
- **Target Variable:** Heart Disease (target)
- **Evaluation Metrics:** Accuracy, Precision, Recall, Confusion Matrix

Final Results:

- Accuracy: **0.7951 (79.5%)**
- Precision: **0.7563 (75.6%)**
- Recall: **0.8738 (87.4%)**

Confusion Matrix Analysis:

- True Positives (TP): **90** patients correctly identified as having heart disease
- True Negatives (TN): **73** patients correctly identified as healthy
- False Positives (FP): **29** patients incorrectly classified as having disease
- False Negatives (FN): **13** patients with disease incorrectly classified as healthy

The model shows a **high recall**, which is especially important in medical applications, as it minimizes the number of missed disease cases while maintaining reasonable precision.

C. Clustering Results

- **Model Used:** K-Means Clustering
- **Data Used:** Medical features without target labels
- **Data Shape:** (1025, 13)

Choosing the Number of Clusters

- The **Elbow Method** was applied for K values from 1 to 10.
- Based on the elbow curve, the optimal number of clusters was chosen as **K = 2**.

Evaluation Metric

- **Silhouette Score:** 0.3972
- Interpretation:
    - The score ranges from 0 to 1; the closer to 1, the better the separation between clusters.
    - A score of ~0.4 indicates **moderate separation**, meaning there is some overlap between patients with similar medical profiles, but the clusters are still meaningful and useful for pattern analysis.

Cluster Interpretation

- **Cluster 0:** Patients with relatively lower-risk medical profiles.
- **Cluster 1:** Patients with higher-risk medical characteristics.

IV. Conclusion

Best Performing Models

- **Regression:** Linear Regression was used to predict cholesterol levels based on age.
- **Classification:** Logistic Regression achieved good accuracy, precision, and recall in predicting heart disease.
- **Clustering:** K-Means with K = 2 produced interpretable patient clusters.

Challenges Encountered

- Limited predictive power when using a single feature for regression
- Feature scaling consistency across different tasks
- Selecting the optimal number of clusters

Future Work

- Use multiple features for regression instead of a single variable
- Try advanced models such as **Random Forest**
- Apply **PCA** before clustering to improve separation
- Perform hyperparameter tuning for Logistic Regression

References & Resources

- Kaggle Datasets
- UCI Machine Learning Repository
- Google Colab **Project Report**