Machine Learning Capstone Project Report

Heart Disease Prediction & Clustering

Team Members

- Shaza Abdulnaser
- Zizi Mostafa
- Naira Ahmed
- Sarah Hassan
- Sama Mohamed Tawfik
- Alaa Ali Omar

I. Introduction

The objective of this project is to implement a complete machine learning pipeline, starting from data acquisition and preprocessing, to model training, evaluation, and result interpretation.

This project addresses three core machine learning tasks:

- **Regression:** Predicting a continuous value using Linear Regression.
- **Classification:** Predicting the presence or absence of heart disease.
- **Clustering:** Discovering hidden patterns in patient data using unsupervised learning.

The project uses a real-world medical dataset and follows best practices to ensure reproducibility, clarity, and meaningful evaluation.

II. Data Description

- **Dataset Source:**

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

- **Dataset Type:** Structured tabular medical data
- **Number of Records:** 1025
- **Number of Features:** 14

Features Include:

- Age
- Sex
- Chest Pain Type (cp)
- Resting Blood Pressure (trestbps)
- Cholesterol (chol)
- Fasting Blood Sugar (fbs)
- Resting ECG (restecg)
- Maximum Heart Rate (thalach)
- Exercise Induced Angina (exang)
- ST Depression (oldpeak)
- Slope of ST (slope)
- Number of Vessels (ca)
- Thalassemia (thal)
- Heart Disease (target)

Target Variable (Classification):

- `target`
  - 0 → No heart disease
  - 1 → Presence of heart disease

The dataset is commonly used for evaluating machine learning models in medical diagnosis tasks.

III. Data Preprocessing & Exploratory Data Analysis (EDA)

1. Data Loading & Inspection

The dataset was loaded using the Pandas library. Initial inspection was conducted to:

- Check the dataset shape
- Identify feature data types
- Detect missing values
- Separate features and target variables

The target variable distribution was as follows:

- 526 samples with heart disease
- 499 samples without heart disease

This indicates a relatively balanced dataset.

2. Handling Missing Values

Missing values were intentionally introduced for experimental purposes.
Numerical features were filled using the median, while categorical features were filled using the mode.
No rows were removed to preserve the dataset size.

3. Feature Encoding

Categorical features were converted into numerical format to ensure compatibility with machine learning algorithms.

- Label Encoding was applied to binary and categorical features.

4. Feature Scaling

Feature scaling was applied using **StandardScaler**.
This step was essential for:

- Logistic Regression
- K-Means Clustering

Scaling ensured that all features contributed equally during model training.

5. Exploratory Data Analysis (EDA)

EDA techniques included:

- Histograms to analyze feature distributions
- Correlation heatmaps to observe relationships between variables
- Box plots to detect potential outliers

Key observations:

- Some features showed noticeable correlation with the target variable.
- Feature distributions justified the need for scaling.

IV. Modeling and Results

A. Regression Task – Linear Regression

Objective

To build a baseline regression model using Linear Regression.

Target Variable

The regression task used the **target column**, treated as a continuous variable, to evaluate baseline regression performance.

Model Used

Linear Regression (scikit-learn)

Methodology

- Dataset split into **80% training** and **20% testing** sets
- Model trained on training data
- Predictions generated on test data

Evaluation Metrics

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Results

- **MSE:** 0.246
- **RMSE:** 0.496
- **R^2 Score:** 0.016

Interpretation

The low R^2 score indicates that Linear Regression explains only a small portion of the variance in the target variable.
This suggests that the relationship between features and the target is likely non-linear and that more advanced models could achieve better performance.

B. Classification Task – Logistic Regression

Objective

To predict whether a patient has heart disease using a supervised classification model.

Model Used

Logistic Regression

Data Preparation

- Target variable defined as binary (0 or 1)
- Features scaled using StandardScaler
- Dataset split into **80% training** and **20% testing** sets

Evaluation Metrics

- Accuracy
- Precision
- Recall
- Confusion Matrix

Results

- **Accuracy:** 0.80

Class	Precision	Recall	F1-score
0 (No Disease)	0.88	0.70	0.78
1 (Disease)	0.76	0.90	0.83

Analysis

The Logistic Regression model achieved strong performance with approximately **80% accuracy**.
Recall for patients with heart disease reached **90%**, which is particularly important in medical diagnosis, as minimizing false negatives helps avoid missing critical cases.

C. Clustering Task – K-Means Clustering

Objective

To identify natural groupings among patients without using class labels.

Methodology

- Target variable removed before clustering
- Elbow Method used to determine the optimal number of clusters
- K-Means model trained using the selected value of K

Evaluation

- Elbow Method to analyze inertia
- Silhouette Score to evaluate cluster quality

Results

The chosen value of K provided a balanced trade-off between compactness and separation of clusters.
Cluster visualization showed meaningful grouping of patients based on selected medical features.

V. Conclusion

This project successfully demonstrated the implementation of a complete machine learning workflow on a real-world medical dataset.

Summary of Results

- Linear Regression provided baseline regression performance.
- Logistic Regression achieved reliable classification results.
- K-Means clustering revealed meaningful patient groupings.

Challenges

- Feature selection and preprocessing decisions
- Choosing the optimal number of clusters

Future Work

- Applying advanced models such as Random Forest or Gradient Boosting
- Using PCA for dimensionality reduction
- Performing hyperparameter tuning to improve performance

VI. Google Colab Notebook Link

- https://colab.research.google.com/drive/1vgjdLKsU-A0JY0XClbinbuBkMP-8PETw?usp=sharing (Heart_Disease Prediction)

- https://colab.research.google.com/drive/1g-9qde5xukySy9bALZ4G53GNaCIVXcGH?usp=sharing (Heart Kmeans)