Name: Ishaan Khan
Class: D15C
Roll No: 29
Batch: B

MLDL Practical 6

Aim: To apply K-Means and Hierarchical Clustering on a real-world medical dataset and analyze clustering performance.

Dataset Source

Dataset Name: Heart Disease Dataset
Platform: Kaggle
Dataset Link:
https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

This dataset contains medical attributes used to predict the presence of heart disease.

Dataset Description

The Heart Disease dataset is a real-world healthcare dataset used for classification and clustering experiments.

Dataset Characteristics

- Number of instances: 1,025
- Number of features: 13 numerical features
- Target variable (for evaluation only): target
    - 1 → Presence of heart disease
    - 0 → Absence of heart disease

Important Features

- age – Age of patient
- sex – Gender
- cp – Chest pain type
- trestbps – Resting blood pressure
- chol – Cholesterol level
- thalach – Maximum heart rate achieved
- oldpeak – ST depression induced by exercise

Mathematical Formulation

K-Means Clustering

K-Means partitions data into K clusters by minimizing the Within-Cluster Sum of Squares (WCSS).

Objective Function

Minimize:

$J = \Sigma\ \Sigma\ ||\ x_i - \mu□\ ||^2$

Where:

- x_i = data point
- $\mu□$ = centroid of cluster j
- K = number of clusters

The algorithm iteratively updates cluster assignments and centroids until convergence.

Hierarchical Clustering

Hierarchical clustering builds nested clusters using an agglomerative (bottom-up) approach.

Euclidean Distance

$d(x, y) = \sqrt{}\ \Sigma\ (x_i - y_i)^2$

Linkage Method

Ward linkage minimizes variance within clusters.

A dendrogram is used to visualize the merging of clusters.

Algorithm Limitations

K-Means Limitations

- Requires predefined number of clusters (K)
- Sensitive to feature scaling
- Sensitive to outliers
- Assumes spherical cluster structure

Hierarchical Clustering Limitations

- Computationally expensive for large datasets
- Cannot reverse cluster merges
- Memory intensive

Methodology / Workflow

Steps Followed

1. Load dataset using KaggleHub
2. Remove target column for clustering
3. Apply feature scaling using StandardScaler
4. Determine optimal K using Elbow Method
5. Apply K-Means clustering
6. Apply Agglomerative Hierarchical Clustering
7. Plot dendrogram
8. Compare clusters with actual heart disease labels
9. Compute Silhouette Score

Performance Analysis

Since clustering is unsupervised, traditional accuracy is not optimized during training. However, performance was analyzed using:

- Elbow Method
- Silhouette Score
- PCA-based visualization
- Comparison with actual disease labels

Observations

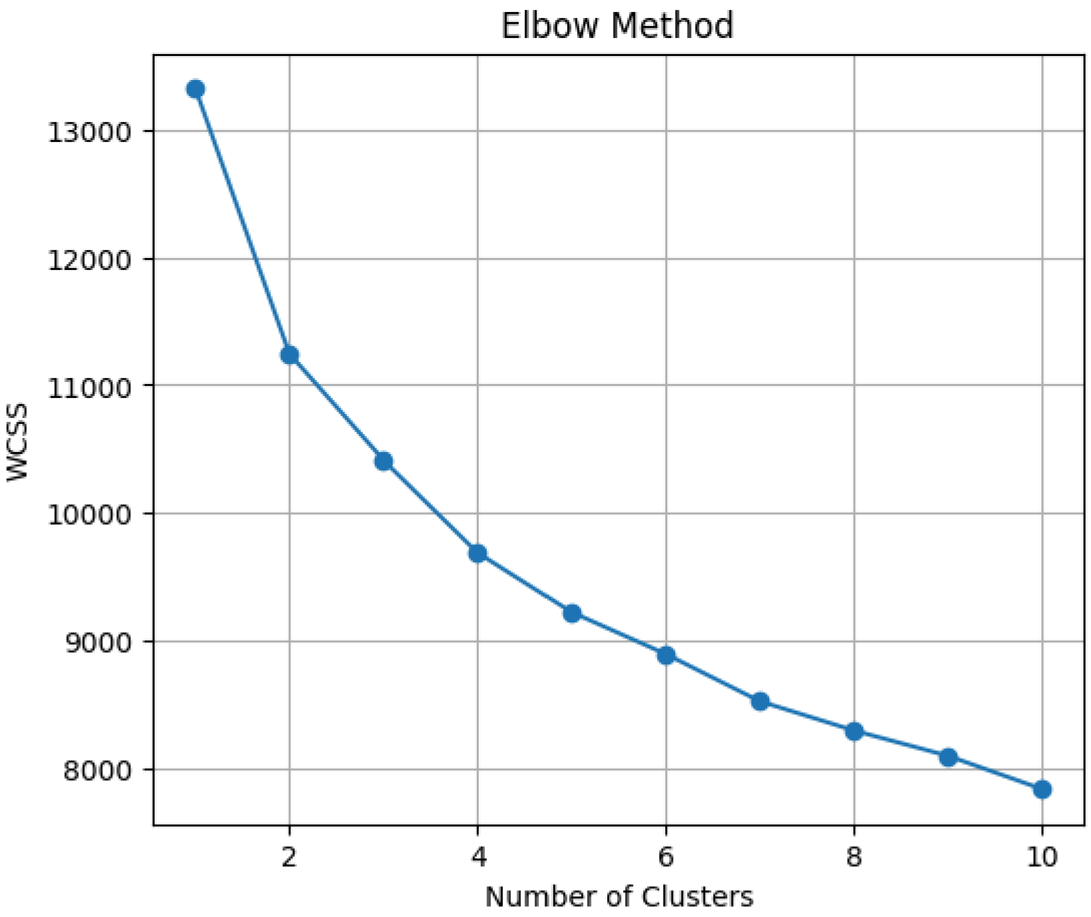
- The Elbow Method suggested K = 2
- K-Means successfully formed two primary clusters
- Hierarchical clustering showed similar grouping patterns
- Silhouette score indicated moderate cluster separation

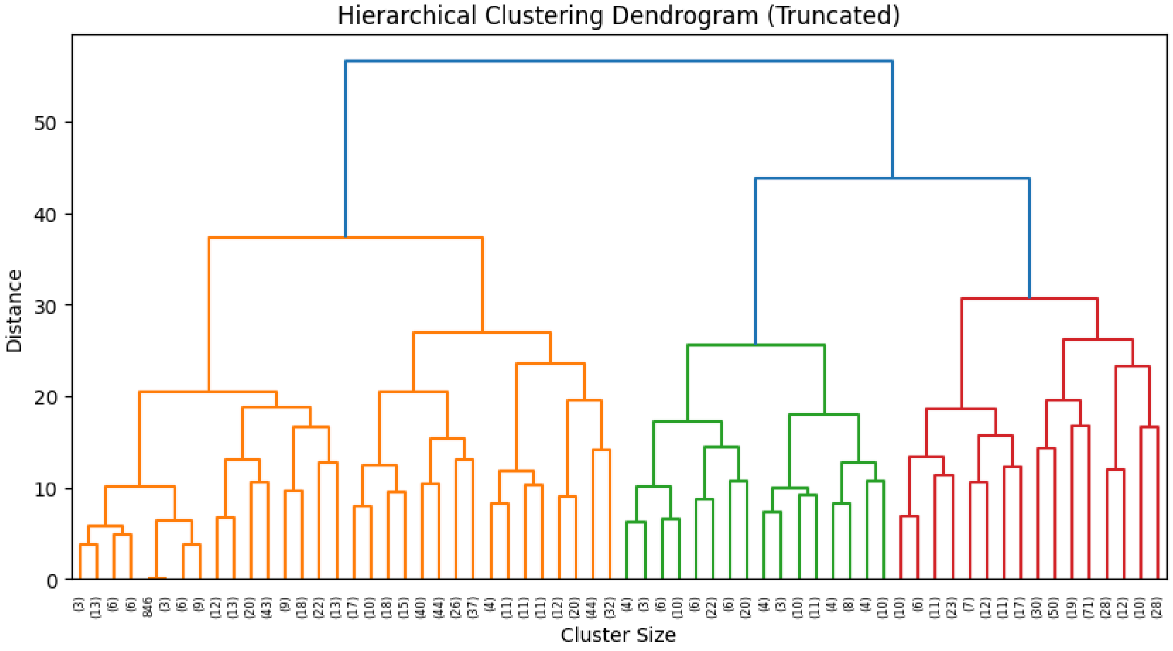
When compared with actual labels, clusters corresponded closely to:

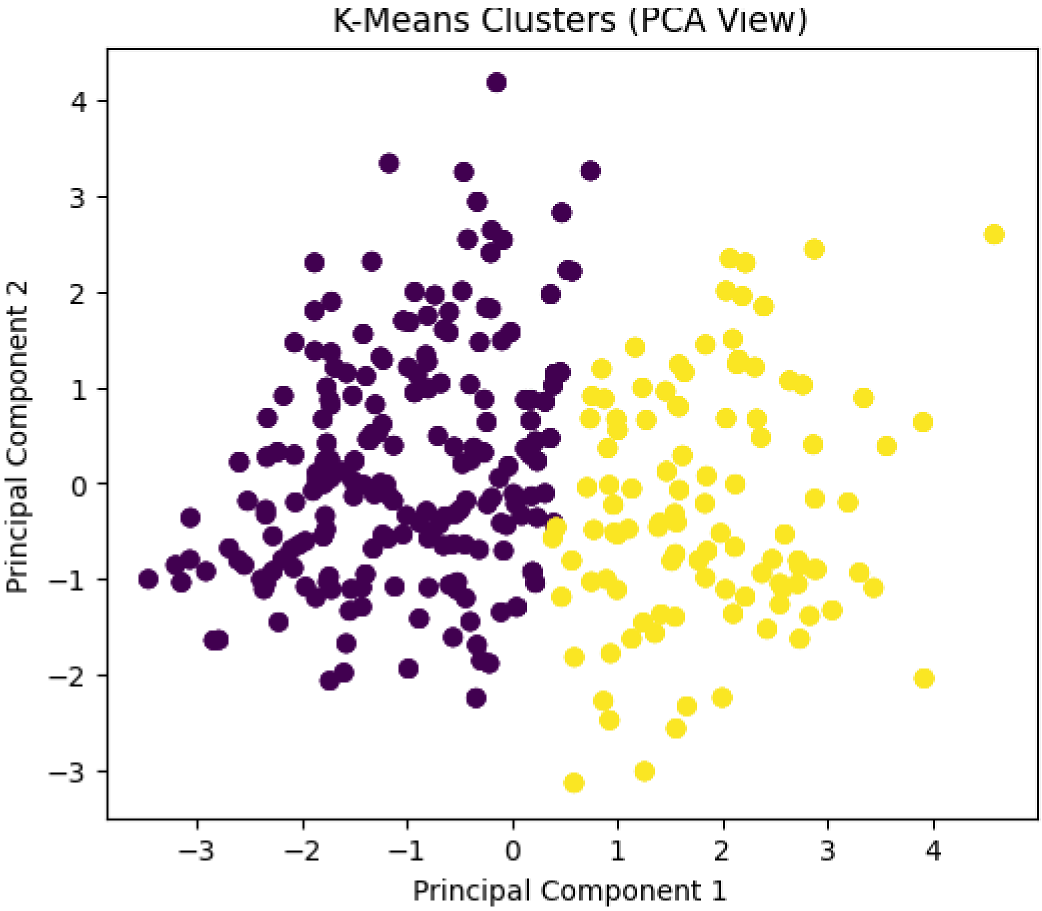
- Patients with heart disease
- Patients without heart disease

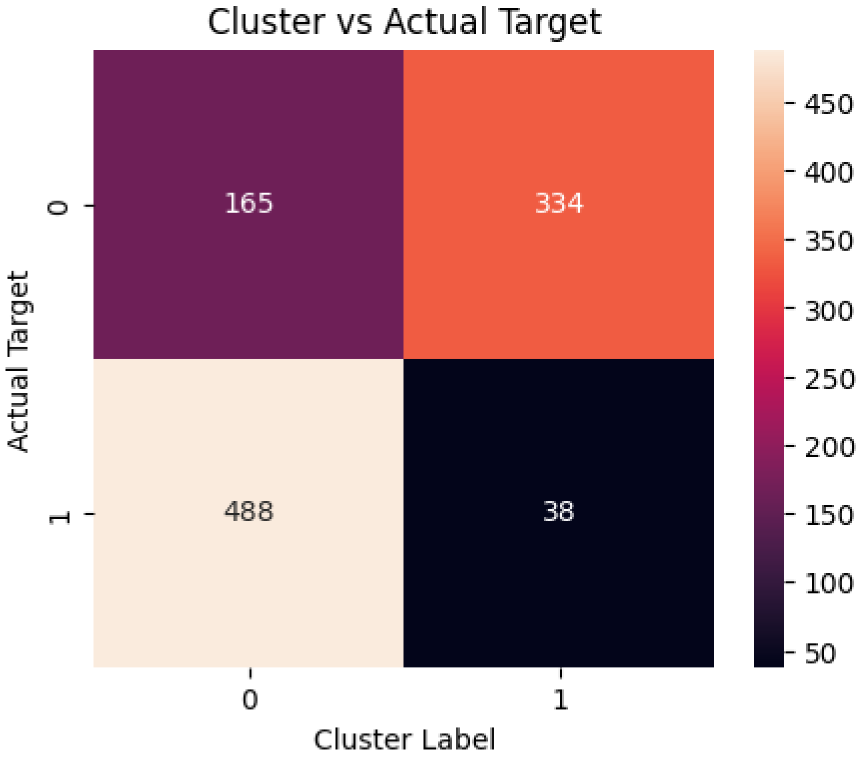
Feature scaling significantly improved clustering performance.

Output


Elbow Method
WCSS
13000
12000
11000
10000
9000
8000
2
4
6
8
10
Number of Clusters


Hierarchical Clustering Dendrogram (Truncated)
Distance
50
40
30
20
10
0
(3) (13) (6) (6) 846 (3) (6) (9) (12) (13) (20) (43) (9) (18) (22) (13) (17) (10) (18) (15) (40) (44) (26) (37) (4) (11) (11) (11) (12) (20) (44) (32) (4) (3) (6) (10) (6) (22) (6) (20) (4) (3) (10) (11) (4) (8) (4) (10) (10) (6) (11) (23) (7) (12) (11) (17) (30) (50) (19) (71) (28) (12) (10) (28)
Cluster Size


K-Means Clusters (PCA View)
Principal Component 1
Principal Component 2
−3
−2
−1
0
1
2
3
4
−3
−2
−1
0
1
2
3
4


Cluster vs Actual Target
165
334
488
38
Actual Target
Cluster Label
0
1
0
1
450
400
350
300
250
200
150
100
50

Conclusion

In this experiment, K-Means and Hierarchical Clustering were successfully applied to the Heart Disease dataset.

K-Means proved efficient and scalable for partitioning medical records into clusters, while Hierarchical Clustering provided better interpretability through dendrogram visualization.

This experiment demonstrates:

- The importance of preprocessing in unsupervised learning
- The ability of clustering algorithms to discover hidden patterns
- The applicability of clustering techniques in medical data analysis

Clustering can assist in identifying patient risk groups and supporting early disease detection strategies.