EXPERIMENT 6

Apply K-Means and Hierarchical Clustering on sample datasets

Aim of the Experiment

To implement **K-Means Clustering** and **Hierarchical Clustering** on a real-world dataset and analyze customer segmentation using appropriate evaluation metrics and hyperparameter tuning.

Theory

Clustering is an **unsupervised learning technique** used to group similar data points together without predefined labels.

K-Means Clustering

K-Means partitions the dataset into **K clusters** by minimizing the within-cluster variance.

It works by:

- Selecting K centroids
- Assigning points to nearest centroid
- Updating centroids
- Repeating until convergence

Hierarchical Clustering

Hierarchical clustering builds clusters in a tree-like structure.

Two types:

- Agglomerative (Bottom-Up)
- Divisive (Top-Down)

In this experiment, **Agglomerative Clustering** was used with Ward linkage.

Dataset Source

Dataset: **Mall Customer Segmentation Dataset**
Source: Kaggle

Link:
https://www.kaggle.com/datasets/shwetabh123/mall-customers

2️⃣ Dataset Description

- Total Records: **200**
- Total Features: **5**

Features:

Feature	Description
CustomerID	Unique ID
Gender	Male/Female
Age	Age of customer
Annual Income (k$)	Income in thousand dollars
Spending Score (1–100)	Score assigned based on spending behavior

For clustering, we used:

- Annual Income
- Spending Score

Target Variable: None (Unsupervised learning)

Mathematical Formulation of the Algorithm

K-Means Objective Function

K-Means minimizes:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

Where:

- C_i = Cluster i
- μ_i = Centroid of cluster i
- $||x - \mu_i||^2$ = Euclidean distance

Hierarchical Clustering (Ward Method)

Ward method minimizes increase in variance:

$$\Delta E = \sum (x - \bar{x})^2$$

It merges clusters that cause minimum increase in total within-cluster variance.

Algorithm Limitations

K-Means Limitations

- Requires predefining K
- Sensitive to initial centroid selection
- Works poorly with non-spherical clusters
- Sensitive to outliers

Hierarchical Limitations

- Computationally expensive for large datasets
- Once merged, clusters cannot be split
- Memory intensive

Methodology / Workflow

Step 1: Data Collection

Download dataset from Kaggle.

Step 2: Data Preprocessing

- Selected relevant features
- Applied StandardScaler for normalization

Step 3: K-Means Implementation

- Applied Elbow Method
- Selected optimal K = 5
- Generated cluster visualization

Step 4: Hierarchical Clustering

- Generated dendrogram
- Applied Agglomerative clustering
- Compared linkage methods

Step 5: Performance Evaluation

- Used Silhouette Score
- Compared both algorithms

Performance Analysis

K-Means

Silhouette Score = **0.5546**

Hierarchical

Silhouette Score = **0.5538**

Interpretation:

- Score > 0.5 indicates good cluster separation.
- Both algorithms performed almost equally.
- K-Means slightly better.

Ward linkage performed best in hierarchical clustering.

Hyperparameter Tuning

K-Means

Tuned number of clusters (K from 2 to 9)

Best K = 5
Highest Silhouette Score = 0.5546

Hierarchical

Tested linkage methods:

- Ward (Best)
- Complete
- Average
- Single (Worst)

Ward linkage gave highest silhouette score.

Observation

- Dataset contains **200 customers**.
- Features used: **Annual Income** and **Spending Score**.
- Elbow Method shows optimal clusters at **K = 5**.

- Highest Silhouette Score for K-Means = **0.5546 (at K=5)**.
- Hierarchical Clustering Silhouette Score = **0.5538**.
- Ward linkage performs best in hierarchical clustering.

Result Analysis

- Both **K-Means and Hierarchical Clustering** perform almost equally.
- Optimal number of clusters = **5**.
- Silhouette score > 0.5 indicates **good cluster separation**.
- K-Means is slightly better and more efficient.
- Customer segmentation into 5 meaningful groups is achieved.

Conclusion

- Both K-Means and Hierarchical Clustering successfully segmented customers into 5 meaningful clusters.
- Silhouette scores confirm good clustering structure.
- K-Means is computationally efficient and slightly better.
- Hierarchical clustering provides better visualization through dendrogram.
- Customer segmentation can help in targeted marketing and business decision-making.