Name: Rushikesh Kolekar
Roll No: 282020
Batch: B3

Practical 5

Problem Statement:

a) Apply clustering algorithms (K-Means and Hierarchical Clustering) to group customers based on their spending behavior.

b) Visualize the resulting customer segments and assess clustering performance using metrics like the Silhouette Score.

c) Perform cross-validation or alternative validation methods to test clustering stability.

Dataset:

Download Mall Customer data from the following link:
Mall Customers Dataset – Kaggle

This dataset contains customer demographics and behavioral data collected from a shopping mall. The attributes include Customer ID, Gender, Age, Annual Income, and Spending Score. The Spending Score is a value between 1 and 100 assigned by the mall based on customer behavior and spending nature.

Objectives:

1. Implement data preprocessing including encoding and normalization.
2. Apply K-Means Clustering and Agglomerative Hierarchical Clustering.
3. Visualize clusters and analyze customer segments.
4. Evaluate clustering performance using the Silhouette Score.
5. Validate the consistency of clustering using different subsets or initializations.

Resources Used:

- **Software:** Jupyter Notebook, Visual Studio Code
- **Libraries:** Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib, Scipy

Theory:

1. Clustering

Clustering is an unsupervised learning method that groups similar data points based on feature similarity. It helps in customer segmentation, pattern detection, and discovering hidden structures in the data.

2. K-Means Clustering

K-Means is a partition-based algorithm that divides the dataset into K clusters. Each point belongs to the cluster with the nearest mean, and centroids are updated iteratively to minimize the variance within clusters.

3. Hierarchical Clustering

This algorithm builds a hierarchy of clusters using a bottom-up (agglomerative) or top-down (divisive) approach. It does not require a predefined number of clusters and is visualized through a dendrogram.

Methodology:

1. Data Preprocessing

- Load the dataset using Pandas
- Encode categorical features like Gender using Label Encoding
- Scale numerical data using StandardScaler or MinMaxScaler for better clustering results

2. Model Implementation

- Apply **K-Means Clustering** and use the Elbow Method to choose the optimal number of clusters
- Apply **Hierarchical Clustering** and visualize the dendrogram to determine the cluster count

3. Visualization

- Plot clusters using scatter plots (e.g., Spending Score vs. Annual Income)
- Visualize the dendrogram for Hierarchical Clustering

4. Performance Evaluation

- Compute the **Silhouette Score** to assess the compactness and separation of clusters
- Use visualization techniques to interpret clustering results and identify profitable customer segments

5. Validation

- Perform clustering on different data subsets or with varying initial centroids
- Check the consistency of cluster assignments across runs

Conclusion:

- Successfully applied K-Means and Hierarchical Clustering to group customers into meaningful segments
- Visualized clusters to identify high-value and low-value customer groups
- Evaluated clustering quality using the Silhouette Score
- Identified potential customer segments that can be targeted with specific marketing strategies
- Future improvements could include using additional behavioral features or applying DBSCAN for density-based clustering