Name: Vivek Mule
Roll No: 281072
Batch: A3

Practical 5

Problem Statement:

a) Use clustering algorithms (K-Means and Hierarchical Clustering) to categorize customers based on their spending patterns.
b) Visualize the formed customer segments and measure clustering effectiveness using the Silhouette Score.
c) Conduct validation through cross-validation or alternative techniques to ensure clustering reliability.

Dataset:

Download the Mall Customer dataset from:
Mall Customers Dataset – Kaggle (https://www.kaggle.com/shwetabh123/mall-customers)

This dataset captures demographic and behavioral information of mall visitors. It includes fields such as Customer ID, Gender, Age, Annual Income, and Spending Score (ranging from 1 to 100, reflecting customer spending behavior and engagement).

Objectives:

1. Perform data preprocessing including encoding of categorical variables and normalization of numeric features.
2. Implement both K-Means and Agglomerative Hierarchical Clustering algorithms.
3. Generate visual representations of the resulting clusters and interpret customer groups.
4. Evaluate cluster quality using the Silhouette Score.
5. Assess the robustness of clustering through different initializations or subsets.

Resources Used:

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

Theory:

1. Clustering

Clustering is an unsupervised machine learning technique that groups data points with similar characteristics. It's commonly used for customer segmentation and uncovering patterns in data.

2. K-Means Clustering

K-Means is a centroid-based algorithm that partitions data into K groups. Each point is assigned to the cluster with the closest centroid, and centroids are updated to reduce intra-cluster variation.

3. Hierarchical Clustering

This approach creates a nested cluster structure using either a bottom-up (agglomerative) or top-down (divisive) strategy. The resulting clusters are represented using a dendrogram, which helps in choosing the appropriate number of groups.

Methodology:

1. Data Preprocessing

- Load the dataset using Pandas
- Encode categorical variables like Gender using Label Encoding
- Normalize numerical features using MinMaxScaler or StandardScaler

2. Clustering Implementation

- Apply K-Means and determine the best number of clusters using the Elbow Method
- Perform Agglomerative Hierarchical Clustering and use a dendrogram to select cluster count

3. Visualization

- Display clusters with scatter plots, particularly using Spending Score and Annual Income
- Plot the dendrogram for a visual understanding of hierarchical clustering

4. Performance Evaluation

- Use the Silhouette Score to measure how well the data fits into the clusters
- Interpret cluster visualizations to identify high-value customer segments

5. Validation

- Run clustering on various subsets or with different initializations
- Compare results to evaluate clustering stability across different runs

Conclusion:

- Implemented both K-Means and Hierarchical Clustering to segment customers into meaningful groups
- Visualized customer clusters to identify potential high-value targets
- Used Silhouette Score to evaluate the clustering effectiveness
- Gained insights into customer behavior which can support strategic marketing
- Future enhancements can involve including more behavioral features or applying advanced methods like DBSCAN for improved segmentation