Name: Dev Bhangale
Roll No: 281059
Batch: A3

Assignment 5

Problem Statement:

Write a program to do following:

Data Set: https://www.kaggle.com/shwetabh123/mall-customers

This dataset givesthe data of Income and money spent by the customers visiting a shopping mall.

The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, asa mall owner you need to find the group of people who are the profitable customers for the mallowner. Apply at least two clustering algorithms (based on Spending Score) to find the group ofcustomers.

a) Apply Data pre-processing

b) Perform data-preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

d) Evaluate the model

e) Apply Cross-validation and Evaluate the model

Dataset:

Mall Customer Dataset
The dataset includes the following features:

- Customer ID
- Gender
- Age
- Annual Income (k$)
- Spending Score (1–100)

This dataset represents customer demographics and behavior in a mall. The goal is to group customers in such a way that helps the mall management identify high-value or profitable segments.

Objectives:

1. Perform necessary data preprocessing steps such as Label Encoding and normalization.
2. Prepare the dataset, including train-test split if applicable.
3. Apply clustering algorithms like K-Means and Hierarchical Clustering to group customers.
4. Evaluate clustering performance using visualization and clustering scores like Silhouette Score.
5. Perform cross-validation where applicable to validate clustering stability.

Resources Used:

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

Theory:

Clustering
Clustering is an unsupervised machine learning technique used to group similar data points together based on feature similarity. It is useful in identifying customer segments, detecting anomalies, and pattern discovery.

K-Means Clustering
K-Means partitions the data into K distinct clusters based on feature similarity. Each data point is assigned to the cluster with the nearest centroid. The algorithm iteratively updates centroids to minimize intra-cluster variance.

Hierarchical Clustering
This method builds a hierarchy of clusters either through an agglomerative (bottom-up) or divisive (top-down) approach. It doesn't require specifying the number of clusters in advance and is visualized using dendrograms.

Methodology:

1. **Data Preprocessing:**
    - Load the dataset using Pandas.
    - Encode categorical columns like Gender using Label Encoding.
    - Scale numerical features using StandardScaler or MinMaxScaler for better clustering.
2. **Train-Test Split (if required):**
    - While clustering typically uses the full dataset, a split may be used for validation purposes.
    - Can use 80% of data for clustering and 20% to validate cluster stability.
3. **Applying Clustering Algorithms:**
    - Apply K-Means clustering and determine the optimal number of clusters using the Elbow Method.
    - Perform Agglomerative Hierarchical Clustering and plot dendrogram to choose the number of clusters.
4. **Model Evaluation:**
    - Visualize clusters using scatter plots for Spending Score vs. Annual Income.
    - Use the Silhouette Score to evaluate clustering performance.
5. **Cross-Validation (if applicable):**
    - Use different initializations or subsets to test clustering consistency.
    - Evaluate how stable the clusters are across different runs.

Conclusion:

- Both K-Means and Hierarchical Clustering effectively grouped customers into meaningful segments.
- Visualization helped understand which groups are likely to be the most profitable.
- Silhouette Scores indicated the compactness and separation of clusters.
- These insights can help mall management create personalized marketing strategies for high-value customer groups.

- Further improvements can involve clustering on additional behavioral data or applying DBSCAN for density-based clustering.