Name: Adhija Bachhav

Roll No: 281053

PRN: 22311656

Batch: A-3

ML Assignment 5

Problem Statement:
Write a program to do following: Data Set: https://www.kaggle.com/shwetabh123/mall-customers. This dataset gives the 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, as a mall owner you need to find the group of people who are the profitable customers for the mall owner.
Apply at least two clustering algorithms (based on Spending Score) to find the group of customers.
a) Apply Data pre-processing
b) Perform data-preparation (Train-Test Split)
c) Apply Machine Learning Algorithm
d) Evaluate Model.
e) Apply Cross-Validation and Evaluate Mode

Introduction:
This project aims to segment customers into different groups based on their income and spending behaviour using unsupervised machine learning algorithms, particularly clustering techniques. Customer segmentation helps mall owners or marketers identify profitable customers, targeted advertising, and personalized services, leading to better customer satisfaction and increased revenue.

Objectives:

- To analyze and preprocess the Mall Customers dataset.
- To visualize the distribution and relationships in customer data through EDA.
- To apply KMeans and Agglomerative Clustering based on Spending Score and Income.
- To evaluate clustering using Silhouette Score.
- To use dendrograms to determine optimal clusters in hierarchical clustering.
- To perform cross-validation to check clustering robustness.

Theory:

Clustering: Clustering is an unsupervised machine learning technique used to group data points based on similarity. It does not require labeled data. The goal is to group customers

such that those in the same cluster are more similar to each other than to those in other clusters.

KMeans Clustering

- Partitional clustering algorithm.
- Divides the dataset into *K* clusters by minimizing **intra-cluster variance**.
- Requires specification of the number of clusters.
- Uses centroids to represent each cluster.

Agglomerative Hierarchical Clustering

- A bottom-up approach to hierarchical clustering.
- Initially, each data point is a single cluster.
- Pairs of clusters are merged based on linkage criteria until all points belong to one cluster or desired cluster count is reached.

Dendrogram

- A dendrogram is a tree-like diagram used to determine the optimal number of clusters in hierarchical clustering.
- Y-axis represents the distance between clusters being merged.
- The point where the longest vertical line is not crossed by any horizontal cut suggests the ideal number of clusters.

Silhouette Score

A metric to evaluate the quality of clusters.

- Ranges from -1 to +1.
- A higher score means better-defined and well-separated clusters.

Cross-Validation in Clustering

Unlike supervised learning, clustering lacks true labels. However, we can perform custom validation (like repeated silhouette scoring) on different folds of the dataset to check consistency and stability of clustering results.

Steps Performed in the assignment:

a) Data Preprocessing

- Removed irrelevant columns (e.g., CustomerID).
- Converted categorical variables (Gender) into numerical.
- Standardized the features using **StandardScaler** for uniformity.

b) Data Preparation

- Selected relevant features: **Annual Income** and **Spending Score**.

- Split the dataset into **training and testing sets**.

c) Clustering Algorithms Applied

- **KMeans Clustering** with predefined number of clusters (e.g., K=5).
- **Agglomerative Clustering** with dendrogram analysis to determine optimal number of clusters.

d) Model Evaluation

- Used **Silhouette Score** to evaluate clustering performance.
- Visualized clusters using scatter plots and color coding.

e) Cross-Validation

- Implemented custom 5-fold silhouette scoring to assess clustering stability and robustness.

Conclusion:

By applying clustering algorithms on the Mall Customers dataset, we successfully segmented the customers into different behavioural groups.
The Agglomerative Clustering with dendrogram helped in choosing the number of clusters more intuitively, while KMeans provided faster computation.
These insights can be highly valuable for mall management to target profitable customers, design personalized marketing campaigns, and improve customer engagement.