Name: Prachee Prasad
Class: CSE-AI SY A
PRN: 22311748
Roll no.: 281**60**

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 Model.

e) Apply Cross-Validation and Evaluate Mode

Dataset

The dataset used in this assignment was downloaded from:
https://www.kaggle.com/shwetabh123/mall-customers

It contains the following columns:

- Customer ID
- Gender
- Age
- Annual Income (in thousands)
- Spending Score (1–100)

The goal is to segment customers based on their **Spending Score** using clustering algorithms to identify profitable groups for business targeting.

Objective

To apply clustering techniques to segment mall customers and identify groups of customers who are potentially more profitable based on their spending behavior. This helps in making informed decisions about marketing and product placements.

Software and Libraries Used

Software Requirements

- Operating System: Windows/Linux
- Programming Language: Python
- Environment: Jupyter Notebook / Google Colab

Python Libraries

- pandas – for data manipulation
- numpy – for numerical operations
- matplotlib and seaborn – for data visualization
- sklearn.cluster – for KMeans and Agglomerative Clustering

- sklearn.model_selection – for cross-validation and splitting
- scipy.cluster.hierarchy – for hierarchical clustering

Theory

What is Clustering?

Clustering is an unsupervised machine learning technique used to group data points into clusters based on similarity. Unlike supervised learning, clustering does not use labeled data. It is commonly used for customer segmentation, anomaly detection, and image compression.

What is K-Means Clustering?

K-Means is a partition-based clustering algorithm that groups the data into *k* clusters.

- It randomly selects *k* initial centroids
- Assigns each point to the nearest centroid
- Recalculates centroids as the average of assigned points
- Repeats the process until convergence

Applications of K-Means:

- Customer segmentation
- Market basket analysis
- Document classification
- Image segmentation

What is Agglomerative Clustering?

Agglomerative Clustering is a hierarchical clustering method that builds nested clusters by merging or splitting them successively.

- Initially, each data point is considered as an individual cluster
- At each step, the two closest clusters are merged
- The process continues until all points belong to a single cluster

Applications of Agglomerative Clustering:

- Hierarchical document classification
- Gene expression data analysis
- Social network analysis

Methodology

1. **Data Collection**
   Imported the mall customer dataset using pandas. Focused on two key features: **Annual Income** and **Spending Score**.

2. **Data Preprocessing**

- Checked for missing values and nulls
- Removed unnecessary columns like Customer ID
- Scaled data if required for clustering algorithms

3. **Train-Test Split**
   Though clustering is unsupervised and doesn't need a target variable, the dataset can still be split for model evaluation using methods like silhouette score and Davies-Bouldin Index.

4. **Model Building**

   a) **K-Means Clustering**

- Chose the number of clusters using the Elbow Method
- Applied KMeans from sklearn.cluster
- Visualized clusters using scatter plots

5. b) **Agglomerative Clustering**
   - Applied AgglomerativeClustering from sklearn.cluster
   - Used dendrograms to decide optimal cluster count
   - Compared results with K-Means

6. **Model Evaluation**
   Used the following evaluation metrics:

- **Silhouette Score** – measures how similar a point is to its own cluster vs other clusters
- **Davies-Bouldin Index** – lower values indicate better clustering
- **Inertia (for K-Means)** – sum of squared distances between points and their centroid

7. **Cross-Validation**

- Applied cross-validation on clustering results using silhouette and Davies-Bouldin metrics
- Ensured the consistency of cluster quality across multiple iterations

Conclusion

Clustering techniques like K-Means and Agglomerative Clustering successfully segmented the mall customers into distinct groups based on their spending behavior. These clusters can help the mall management in identifying high-value customers, improving marketing strategies, and enhancing customer service. While K-Means is faster and efficient, Agglomerative Clustering offers more interpretable hierarchical relationships between customers.