**Name:** Janhvi Katakdhond

**Roll no:** 282067

**Batch:** B3

**ASSIGNMENT 5**

**Problem Statement:**

Write a program to do following: Data Set:

https://www.kaggle.com/shwetabh123/mall-customers This dataset gives thedata 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 of customers.

1. Apply Data pre-processing
2. Perform data-preparation (Train-Test Split)
3. Apply Machine Learning Algorithm

**Objective:**

The objective of this assignment is to analyze customer data from a shopping mall to identify distinct groups of customers based on their spending behavior. By applying data preprocessing and preparation techniques, and using at least two clustering algorithms—such as K-Means and Hierarchical Clustering—the goal is to segment customers according to their Spending Score and Annual Income. This segmentation will help in identifying profitable customer groups, enabling better-targeted marketing strategies and business decisions for the mall owner.

**S/W Packages and H/W apparatus used:**

Software used:

1. Python 3.x

2. Google Colab

Libraries and packages used: NumPy, Pandas

**Theory:**

In the field of machine learning, clustering is an **unsupervised learning technique** used **to group similar data points together based on their features**. Unlike supervised learning, clustering does not rely on labeled output data. Instead, it seeks to uncover the inherent structure or patterns in the data by forming distinct groups, or clusters, where members of each cluster are more similar to each other than to those in other clusters.

**Purpose of Clustering in Business**

Clustering is widely used in customer segmentation to understand different consumer behaviors. For businesses like shopping malls, identifying groups of customers based on their demographics and spending habits can help in:

* Designing personalized marketing strategies
* Improving customer satisfaction
* Enhancing revenue generation by targeting profitable segments

**Clustering Algorithms**

1. **K-Means Clustering**

K-Means is one of the most popular and simple clustering algorithms. It partitions the data into *k* clusters by minimizing the variance within each cluster. The algorithm works iteratively to assign data points to one of the *k* centroids and then updates those centroids based on the points assigned to them.

* Efficient and scalable
* Requires pre-specifying the number of clusters
* Sensitive to initial placement of centroids

1. **Hierarchical Clustering**

Hierarchical clustering builds a tree-like structure (called a dendrogram) that represents data grouped into nested clusters. It can be:

* **Agglomerative** (bottom-up): Starts with each point as its own cluster and merges them step-by-step
* **Divisive** (top-down): Starts with all points in one cluster and divides them recursively This method does not require specifying the number of clusters in advance and is useful for visualizing the structure of clusters

**Data Preprocessing and Preparation**

Before applying clustering algorithms, it is essential to perform:

* **Data Cleaning**: Handling missing values, correcting inconsistencies
* **Feature Selection**: Choosing relevant variables (e.g., Annual Income and Spending Score)
* **Scaling**: Normalizing or standardizing data to bring features to a similar scale
* **Train-Test Split**: Although clustering is unsupervised, splitting data may still be useful in hybrid approaches or validation scenarios

**Conclusion:**

In this assignment, customer data from a shopping mall was analyzed using unsupervised machine learning techniques to identify meaningful customer segments. After performing necessary data preprocessing and preparation, two clustering algorithms—**K-Means Clustering** and **Hierarchical Clustering**—were applied using features such as Annual Income and Spending Score.

The results revealed distinct customer groups with varying spending behaviors and income levels. These segments provide valuable insights for the mall management to target profitable customers more effectively. For instance, high-income and high-spending customers could be prioritized for premium services and targeted marketing campaigns, while low-spending groups might benefit from engagement strategies to increase their value.

Overall, this clustering approach demonstrated the practical importance of customer segmentation in enhancing business decision-making, personalizing marketing strategies, and ultimately improving customer satisfaction and revenue generation.