**Name:** Chandrakant Dattatrey Thakare

**Roll no:** 282014  **PRN:** 22310303

**Class :** SY CSE AI**Batch:** B1

**Assignment 5: Customer Segmentation using Clustering**

**Problem Statement:**

Analyze the **Mall Customers Dataset** available at:  
<https://www.kaggle.com/shwetabh123/mall-customers>

This dataset provides information on **Customer ID**, **Gender**, **Age**, **Annual Income**, and **Spending Score** of individuals visiting a mall.

As a mall owner, the objective is to identify and segment groups of customers to determine who are the most **profitable customers**.

Apply at least **two clustering algorithms** based on **Spending Score** to perform customer segmentation.

Tasks include:  
a) Data Preprocessing  
b) Data Preparation (Train-Test Split)  
c) Applying Machine Learning Algorithms  
d) Evaluating the Model  
e) Applying Cross-Validation and Evaluating the Model

**Objective**

* Understand the importance of clustering in customer segmentation and marketing strategy.
* Apply unsupervised learning algorithms to identify patterns in consumer spending behavior.
* Explore how machine learning can aid decision-making for business optimization.

**Tools and Resources**

* **Software Used**: Google Colab / Jupyter Notebook
* **Libraries Used**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

**Key Algorithms and Techniques Used**

* **K-Means Clustering**: Groups customers based on similarity in spending habits.
* **Hierarchical Clustering**: Builds a tree-like structure to group similar customers.
* StandardScaler, LabelEncoder: Used for preprocessing numerical and categorical data.
* train\_test\_split: Prepares the dataset for model evaluation.
* cross\_val\_score, Silhouette Score, Elbow Method: Evaluate clustering performance and stability.

**Methodology**

**1. Data Preprocessing**

* Load the dataset using Pandas.
* Handle missing values (if any).
* Encode categorical columns like **Gender** using LabelEncoder.
* Standardize numerical columns like **Income** and **Spending Score** using StandardScaler.

**2. Data Preparation**

* Although clustering is unsupervised, train\_test\_split is used to test cluster stability across subsets.
* Identify input features for clustering (e.g., **Annual Income** and **Spending Score**).

**3. Machine Learning Application**

* Apply **K-Means Clustering**:
  + Use the **Elbow Method** to determine the optimal number of clusters.
  + Fit the model and assign each customer to a cluster.
* Apply **Hierarchical Clustering**:
  + Generate dendrograms to visualize the linkage.
  + Cut the tree at an appropriate level to form distinct customer segments.

**4. Model Evaluation**

* Visualize clusters using scatter plots for better interpretation.
* Evaluate the clustering performance using **Silhouette Score**.

**5. Cross-Validation**

* Apply cross-validation techniques to validate cluster robustness.
* Assess clustering quality on different data splits to ensure consistency in results.

**Advantages of Clustering for Business**

* Identifies **high-value customer segments**.
* Helps in **targeted marketing** and personalized service delivery.
* Informs strategic business decisions like offers, discounts, and store layout.

**Challenges**

* Determining the **optimal number of clusters**.
* Scaling and encoding data properly for accurate clustering.
* Interpreting clusters when dimensionality increases.

**Conclusion**

This assignment provided hands-on experience in **unsupervised learning** and **customer segmentation** using clustering techniques. I learned to:

* Preprocess and prepare real-world datasets.
* Apply clustering algorithms like **K-Means** and **Hierarchical Clustering**.
* Evaluate and validate models using visual and quantitative metrics.
* Extract business insights from data using machine learning techniques.