**Name:** Shubham Dahane  
**Class:** CSE(AI) B Div  
**Roll No.:** 282015  
**PRN No.:** 22310312  
**Subject:** Machine Learning

**Assignment 5 – Customer Segmentation Using Clustering**

**Objective**

The goal of this assignment is to segment customers based on their **Annual Income** and **Spending Score** using various **clustering algorithms**. This helps businesses like malls identify **profitable customer groups** for better marketing and personalized service strategies.

**Dataset Used**

**Source:** [Kaggle - Mall Customers Dataset](https://www.kaggle.com/shwetabh123/mall-customers)

**Attributes in the dataset:**

* CustomerID: Unique identifier for each customer
* Gender: Male/Female
* Age: Age of the customer
* Annual Income (k$): Annual income in thousand dollars
* Spending Score (1-100): Score assigned by the mall based on customer behavior

**a) Data Preprocessing**

* The dataset was loaded and inspected using .info(), .head(), .describe(), and .isnull().sum() to understand structure and cleanliness.
* The column Genre was renamed to Gender for clarity.
* The Gender column was **label encoded** (Male = 1, Female = 0).
* Selected features for clustering:  
  Annual Income (k$) and Spending Score (1–100)
* Features were **standardized** using StandardScaler to ensure equal weight during clustering.

**b) Data Preparation**

* The dataset was **split** into **training and testing sets** in an 80:20 ratio using train\_test\_split() to simulate real-world data generalization.
* Shapes of the train and test sets were printed for confirmation.

**c) Machine Learning Algorithms Applied**

**1. K-Means Clustering**

* **K=5** was selected based on the Elbow Method (not shown in code but considered standard).
* K-Means was trained on the **training set** and predictions were made for the **entire dataset**.
* Cluster labels were stored in a new column KMeans\_Cluster.

**2. DBSCAN (Density-Based Spatial Clustering)**

* Unlike K-Means, DBSCAN does not require the number of clusters in advance and is robust to noise.
* Applied DBSCAN with eps=0.5 and min\_samples=5.
* Cluster labels (including noise points labeled as -1) were stored in DBSCAN\_Cluster.

**d) Model Evaluation**

* **Silhouette Score** is typically used to evaluate clustering quality.  
  *(In practice, it can be calculated using sklearn.metrics.silhouette\_score)*
* DBSCAN can result in some points marked as noise (-1), which can lower its score but helps identify outliers.
* The **distribution of clusters** was printed to check how customers are segmented.

**e) Cross-Validation (Demonstrated on Example Datasets)**

Though clustering is unsupervised, **cross-validation** concepts were explored using dummy datasets:

* **K-Fold**
* **Leave-One-Out**
* **Leave-P-Out**
* **Stratified K-Fold**
* **Repeated K-Fold**  
  These techniques were shown for better understanding of model reliability and generalization in supervised learning.

**Cluster Visualization**

**K-Means Clusters:**

* Scatter plot using Seaborn:
  + X-axis: Annual Income (k$)
  + Y-axis: Spending Score (1–100)
  + Colored clusters show distinct customer segments.

**DBSCAN Clusters:**

* Similar scatter plot for DBSCAN:
  + Clusters formed based on density.
  + Noise/outliers identified (colored differently or excluded).

**Insights Gained**

* **K-Means** identified five distinct groups of customers such as:
  + High income – high spenders
  + Low income – low spenders
  + Average income – varying spending behavior
* **DBSCAN** helped in detecting **outliers** that don't belong to any major segment.
* **Standardization** greatly improved cluster separation, especially for DBSCAN.

A graph with colored dots

AI-generated content may be incorrect.

**A chart with many colored dots

AI-generated content may be incorrect.**

**Conclusion**

Clustering techniques like **K-Means** and **DBSCAN** successfully helped segment customers based on income and spending behavior.  
These insights can assist mall management in:

* **Tailoring marketing campaigns**
* **Offering exclusive deals to high-spending segments**
* **Detecting unusual customer patterns or outliers**

Unsupervised learning proves to be a powerful tool in understanding and leveraging customer behavior for business growth.