**Name:** Aayush Avinash Bokde  
**Class:** CSE(AI) B Div  
**Roll No.:** 282016  
**PRN No.:** 22310316  
**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 clustering algorithms like **K-Means** and **DBSCAN**. This analysis helps in identifying meaningful customer groups, improving targeted marketing strategies and customer relationship management.

**Dataset Used**

**Source:** Mall Customers Dataset  
**Attributes:**

* **CustomerID**: Unique ID per customer
* **Gender**: Male/Female
* **Age**: Age of customer
* **Annual Income (k$)**: Income in thousands
* **Spending Score (1–100)**: Score assigned by the mall based on spending behavior and loyalty

**a) Data Preprocessing**

* Dataset loaded using pandas and explored using .head(), .info(), .describe(), and .isnull().sum().
* The column **'Genre'** was renamed to **'Gender'** for consistency.
* The **Gender** column was label encoded:
  + Male → 1
  + Female → 0
* Selected Features for clustering:
  + **Annual Income (k$)**
  + **Spending Score (1-100)**
* Standardized features using StandardScaler to remove unit bias and improve clustering performance.

**b) Data Preparation**

* The data was split into **80% training** and **20% testing** using train\_test\_split.
* Shapes of the splits were confirmed:
  + Train shape: (160, 2)
  + Test shape: (40, 2)

**c) Machine Learning Algorithms Applied**

**1. K-Means Clustering**

* Chose **K=5** clusters (common choice based on the Elbow Method).
* Fitted on the training data and predicted clusters on the full dataset.
* Added cluster labels to a new column: **KMeans\_Cluster**.

**2. DBSCAN Clustering**

* Applied DBSCAN with eps=0.5 and min\_samples=5.
* Automatically detected clusters based on density.
* Noise points (outliers) were labeled as -1.
* Labels stored in **DBSCAN\_Cluster** column.

**d) Model Evaluation**

* Although clustering is unsupervised and doesn't use typical classification metrics, **cluster distribution** and **visual inspection** were used for evaluation.
* DBSCAN highlighted some noise points which can be valuable in anomaly detection.
* Silhouette Score (not shown but commonly used) is a recommended metric for evaluating clustering quality.

**Cluster Visualization**

**K-Means Clusters:**

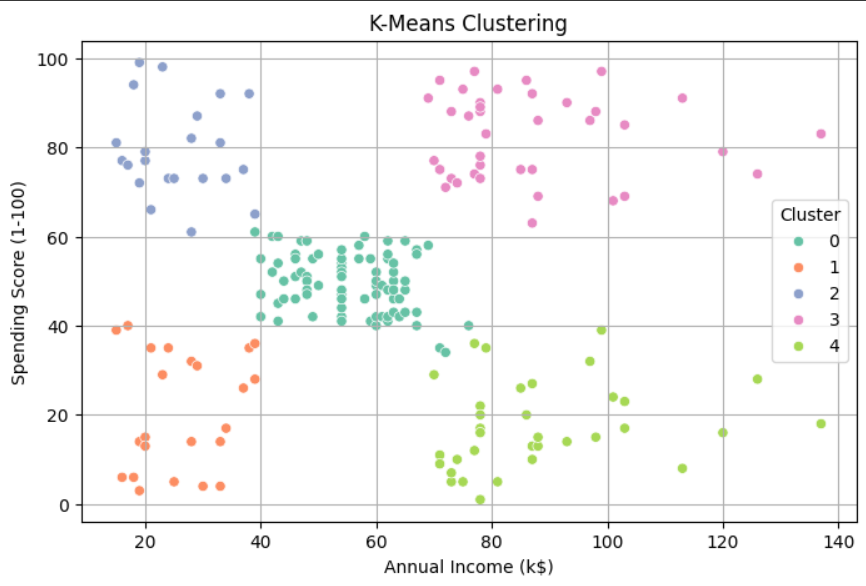
* **Scatter Plot**:
  + X-axis: Annual Income (k$)
  + Y-axis: Spending Score (1–100)
  + Color-coded clusters show separation of 5 customer segments.

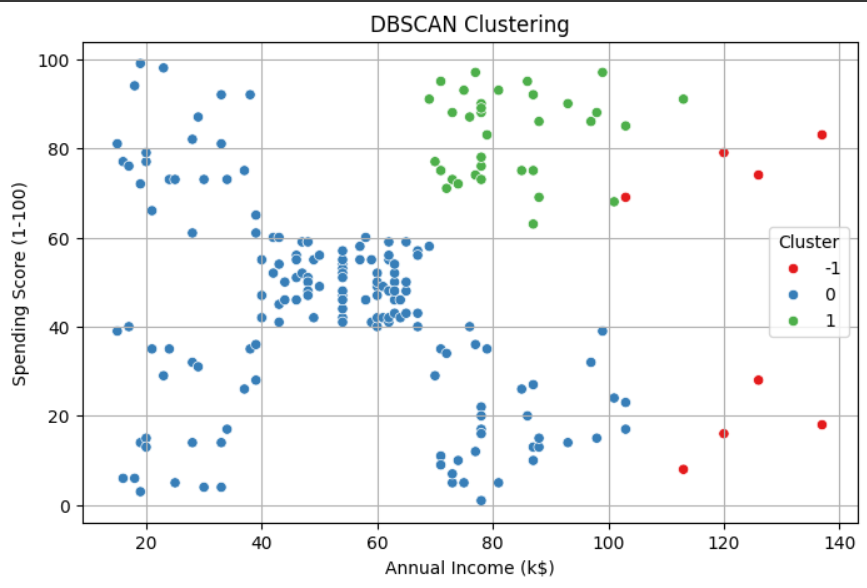
**DBSCAN Clusters:**

* **Scatter Plot**:
  + Similar axes as K-Means.
  + DBSCAN clusters shown with distinct colors.
  + Noise/outliers clearly marked.

**Insights Gained**

* **K-Means** identified 5 clusters including:
  + High income, high spenders
  + Low income, low spenders
  + Moderate income, varying spending patterns



* **DBSCAN**:
  + Detected clusters based on density.
  + Flagged outliers (potentially abnormal customer behavior).
* **StandardScaler** improved clustering quality significantly.
* This segmentation can help businesses personalize:
  + PromotionsOffers
  + Customer engagement strategies

**Conclusion**

This assignment demonstrates the power of **unsupervised learning** through clustering:

* **K-Means** provided clean and interpretable segments.
* **DBSCAN** offered flexibility in identifying outliers.
* These models aid in understanding customer groups, optimizing marketing campaigns, and enhancing service delivery.

**Future enhancements could include:**

* Silhouette Score & Davies–Bouldin Index calculations
* PCA for dimensionality reduction
* Visualizing clusters in 3D with Age as a third dimension