Name: Prachi Shedge
Roll no: 281050
Batch: A2
Assignment 5

Statement:

In this assignment, we have to perform:
a) Data pre-processing on customer data from a shopping mall.
b) Data preparation including train-test split.
c) Application of clustering algorithms to identify profitable customer segments.

Objective:

1. This assignment aims to introduce clustering techniques for customer segmentation.
2. It familiarizes users with data standardization and evaluation of clustering algorithms.
3. Enhances skills in visualizing and interpreting clusters for business insights.

Resources Used:

- **Software used:** Visual Studio Code
- **Libraries used:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

Introduction to Clustering:

Clustering is an unsupervised machine learning technique used to group similar data points together. In this assignment, we apply two clustering algorithms:

- **K-means:** Partitions data into K clusters by minimizing variance within clusters.
- **DBSCAN:** Density-based clustering that identifies clusters as high-density areas separated by low-density areas.

Methodology:

1. **Data Collection and Exploration:**
    - Load the mall customers dataset and explore its structure.
    - Select relevant features (Annual Income and Spending Score).
2. **Data Preprocessing:**
    - Handle categorical variables (Gender) by encoding.
    - Standardize numerical features for clustering.
3. **Model Training and Evaluation:**
    - Apply K-means clustering with optimal K determined by the Elbow Method.
    - Apply DBSCAN clustering with tuned parameters.
    - Evaluate clusters using silhouette score and visualize results.
4. **Business Interpretation:**

- Identify profitable customer segments based on cluster centroids.

Advantages:

1. K-means is computationally efficient for large datasets.
2. DBSCAN does not require pre-specifying the number of clusters and handles noise well.
3. Visualizations (e.g., scatter plots) provide intuitive insights into customer behavior.

Disadvantages:

1. K-means requires choosing K and is sensitive to initial centroids.
2. DBSCAN performance heavily depends on parameter selection (eps, min_samples).
3. Feature scaling is critical for both algorithms.

Conclusion:

This assignment demonstrated the application of clustering algorithms (K-means and DBSCAN) for customer segmentation. We identified profitable customer groups based on spending patterns, which can help mall owners tailor marketing strategies. The project reinforced skills in data preprocessing, model evaluation, and business interpretation of machine learning results.