Assignment 5

Name: Yashraj Nimbalkar

Roll No: 281039
Batch: A2

Problem Statement

The goal of this assignment is to:
a) Utilize data preprocessing and clustering techniques to categorize mall customers based on their purchasing patterns.
b) Implement two clustering algorithms—KMeans and DBSCAN—to identify valuable customer segments.
c) Assess clustering effectiveness using:
• Silhouette Score
• Davies-Bouldin Index
d) Visualize the resulting customer clusters to enhance interpretability.

Objective

1. Gain a solid understanding of unsupervised machine learning, with a focus on clustering.
2. Use clustering methods to classify customers in a way that supports business strategies.
3. Evaluate the clustering models with internal performance metrics.
4. Visualize clusters to extract practical insights.

Tools and Resources

• Software: Jupyter Notebook / Visual Studio Code
• Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

Clustering Overview

Clustering is a core unsupervised learning method that groups data points so that those within the same cluster are more alike than those in other clusters. In this task, clustering is applied to segment mall customers using their Annual Income and Spending Score. This helps the management discover profitable customer profiles.

Libraries Utilized

1. Pandas & NumPy – For managing and manipulating data
2. Matplotlib & Seaborn – For visual representation of clusters

3. Scikit-learn – For implementing KMeans, DBSCAN, and evaluating clustering models

Methodology

1. Data Collection and Preprocessing

- Dataset: Mall_Customers.csv (includes attributes like Gender, Age, Income, and Spending Score)
- Process:
    - Loaded the data using Pandas
    - Renamed columns for clarity
    - Removed unnecessary fields (e.g., CustomerID)
    - Encoded the 'Gender' column
    - Standardized numerical columns using StandardScaler

2. Preparing Data for Clustering

- As clustering is unsupervised, no train-test split was necessary.
- Only Annual Income and Spending Score were used for segmentation.

3. Model Implementation

a) KMeans Clustering

- The Elbow Method was used to identify the optimal number of clusters (k = 5)
- The model was trained with the fit_predict() method
- Scatter plots helped illustrate distinct customer groups

b) DBSCAN (Density-Based Spatial Clustering)

- DBSCAN was applied using eps = 0.5 and min_samples = 5
- Identified core points, border points, and outliers
- Visualized the results using cluster labels

---

4. Evaluation of Models

The clustering performance was assessed with the following metrics:

Metric	KMeans (Example)	DBSCAN (Example)
Silhouette Score	~0.55	~0.48
Davies-Bouldin Index	~0.56	~0.68

Note: Results may vary with different datasets or DBSCAN parameters.

---

5. Visualization

- Scatter plots were used to visualize the KMeans and DBSCAN clusters.
- X-axis: Annual Income, Y-axis: Spending Score
- Cluster labels distinguished customer groups
- Helped in spotting profitable clusters like high-spending, high-income individuals

---

Benefits of Clustering

1. Enables customer segmentation based on behavioral patterns
2. KMeans is computationally efficient for large datasets
3. DBSCAN handles arbitrary cluster shapes and identifies noise

Limitations

1. KMeans assumes spherical clusters and is sensitive to outliers
2. DBSCAN's performance heavily depends on parameter tuning
3. Absence of labeled data makes evaluation subjective

Conclusion

This assignment demonstrated how KMeans and DBSCAN can be applied to segment mall customers based on income and spending habits. Evaluation through internal metrics and visualizations highlighted key customer segments, especially high-value ones. These insights can significantly support targeted marketing and business optimization.