Assignment - 5

Name: Yashashree Nimbalkar
Roll No: 281040
Batch: A2
PRN: 22311423

Statement

In this assignment, we aim to:
a) Apply clustering algorithms (K-Means and DBSCAN) to segment customers from the Mall Customer Segmentation dataset.
b) Use the Elbow Method and Silhouette Score to determine the optimal number of clusters.
c) Evaluate the performance of clustering models.
d) Visualize the customer clusters using appropriate plots.

Objective

1. Understand the concept and applications of clustering in machine learning.
2. Implement unsupervised learning algorithms for real-world data.
3. Analyze customer segments for business insights.
4. Compare clustering techniques (K-Means vs DBSCAN).
5. Visualize clusters for meaningful interpretation.

Resources Used

- Software: VS Code
- Libraries: Pandas, NumPy, Matplotlib, Scikit-learn

Introduction to Clustering

Clustering is an unsupervised learning technique used to group similar data points based on features. It helps discover hidden patterns or groupings in unlabeled data. In this assignment, we use clustering to segment mall customers based on their annual income and spending score, allowing businesses to tailor services and strategies for each group.

Key Libraries Used

1. **Pandas & NumPy** – For data handling and preprocessing.

2. **Matplotlib** – For data visualization.

3. **Scikit-learn** – For implementing clustering models (KMeans, DBSCAN) and evaluation metrics.

Methodology

1. Data Collection and Preprocessing

● Dataset Used: *Mall_Customers.csv*
● Data Source: Kaggle
● Initial Steps:
○ Loaded the dataset using Pandas.
○ Renamed columns for consistency.
○ Converted categorical 'Gender' column into numerical values.
○ Selected relevant features: Annual Income and Spending Score.
○ Standardized the data using StandardScaler for better clustering results.

---

2. Clustering Techniques

A. K-Means Clustering
● Applied the Elbow Method to find the optimal number of clusters.
● Found that `k=5` gives the best clustering.
● Fitted KMeans and added predicted cluster labels to the dataset.
● Visualized the clusters and centroids on a scatter plot.

B. DBSCAN Clustering

- Applied DBSCAN to detect dense regions in data.
- Identified noise points and number of clusters.
- Compared results with KMeans.

---

3. Evaluation

- **Silhouette Score** was used to evaluate the quality of clustering:
  - Higher score indicates well-defined clusters.
- Compared the Silhouette Score of KMeans and DBSCAN.
- Performed cross-validation using silhouette as the scoring metric.

---

4. Visualization

- Plotted KMeans clusters and centroids.
- Compared actual clusters visually.
- Used scatter plots to differentiate customer groups.

Advantages of Clustering

1. Helps in identifying distinct customer groups.
2. No prior labels are required for training (unsupervised).
3. Improves decision-making in marketing and product design.

Disadvantages

1. KMeans is sensitive to initial centroids and outliers.
2. DBSCAN struggles with varying density clusters.
3. Choosing optimal parameters (like `k`, `eps`) requires tuning.

Conclusion

This assignment focused on applying clustering algorithms to real-world customer data. KMeans and DBSCAN were implemented to group customers based on spending behavior and income. The optimal number of clusters was determined using the Elbow Method and evaluated using Silhouette Score. Visualizing the clusters helped understand the distribution of customer segments. Such techniques are essential in customer segmentation, market research, and targeted marketing strategies.