Name: Sejal Sampat Godbole

Roll No: 281030

Batch: A2

Assignment 5

Problem Statement:

Write a program to do following:

1. Apply Data pre-processing
2. Perform data-preparation (Train-Test Split).
3. Apply Machine Learning Algorithm.
4. Evaluate Model.
5. Apply Cross-Validation and Evaluate Mode

Objectives:

1. To preprocess the dataset by handling missing values, encoding categorical variables, and standardizing numerical features.
2. To explore and visualize customer segments using spending score and annual income through clustering techniques.
3. To apply at least two clustering algorithms (e.g., KMeans and DBSCAN) for identifying distinct customer groups.
4. To evaluate clustering performance using metrics such as silhouette score and visualizations.
5. To apply cross-validation techniques to assess the consistency and robustness of the clustering models.

Resources used:

1. **Software used:** Google colab
2. **Libraries used:** Pandas, Matplotlib, SKLearn

Theory:

1. **Clustering:**
   Clustering is an unsupervised learning technique that groups similar data points into clusters. It helps in customer segmentation, anomaly detection, and more.
   Clustering Methods
   I. **K-Means Clustering**

- **Type:** Centroid-based
- **Input:** Number of clusters (K)
- **Works by:** Minimizing distance between points and centroids
- **Pros:** Fast, simple
- **Cons:** Sensitive to outliers, needs K

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

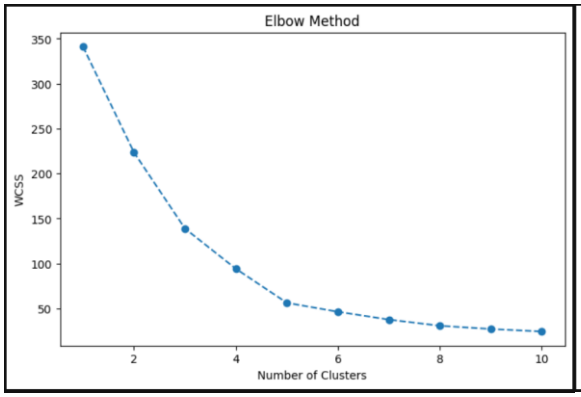
- **Type:** Density-based
- **Input:** eps (neighborhood radius), min_samples
- **Works by:** Grouping points with high density, marking sparse points as noise
- **Pros:** Detects arbitrary shapes, handles outliers
- **Cons:** Sensitive to eps and min_samples

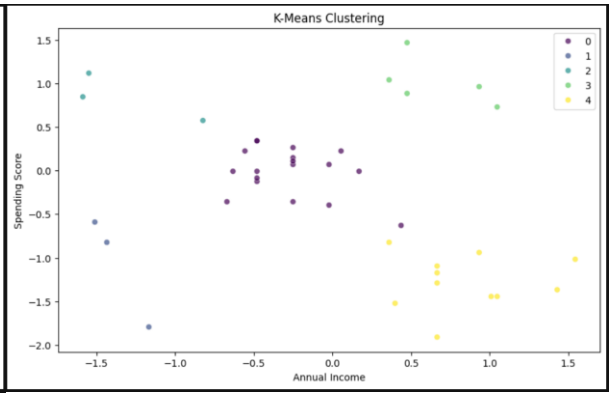
Methodology:

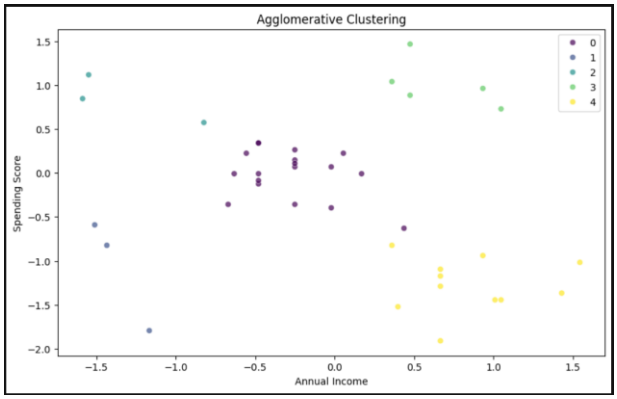
1. Concise Methodology:

1. Load and preprocess the dataset by handling missing values, encoding categorical data, and standardizing numerical features.
2. Perform exploratory data analysis to understand distributions and relationships between features.
3. Select relevant features (Annual Income, Spending Score) for clustering.
4. Apply KMeans and DBSCAN clustering algorithms to segment customers.
5. Evaluate cluster quality using Silhouette Score and visualizations.
6. Apply cross-validation to assess clustering consistency and robustness.
7. Interpret the results to identify profitable customer segments for strategic decision-making.

Results:







Conclusion:

Clustering techniques applied to the Mall Customers dataset effectively identified distinct customer segments based on income and spending behavior. KMeans and DBSCAN revealed profitable customer groups, supported by silhouette scores and visualizations. The results offer valuable insights for targeted marketing and strategic business decisions.