Unsupervised Learning

Exploring Unsupervised Learning Techniques for Customer Segmentation

Objective: In this mini project, you will perform customer segmentation using unsupervised learning techniques. You will apply K-Means, Hierarchical Clustering, PCA and t-SNE on real-world data.

Dataset:

- Use the **Online Retail dataset** from the UCI Machine Learning Repository (or a similar dataset). This dataset contains transactions occurring between 01/12/2010 to 09/12/2011 for a UK-based online retailer.

Part 1: Data Preparation

- **Preprocessing:**
  - Clean the data (handle missing values, remove outliers).
  - Create **Recency, Frequency, and Monetary (RFM) metrics** for each customer.
  - Standardize the data using StandardScaler to ensure each metric contributes equally to the clustering process

Part 2: Clustering Techniques

i) K-Means Clustering:

- Apply K-Means to segment customers.
- Use the **Elbow Method** to determine the optimal number of clusters.
- Visualize clusters and cluster centers in 2D.

ii) Hierarchical Clustering:

- Apply **Agglomerative Clustering** and visualize the dendrogram.
- Experiment with different linkage methods (single, complete, ward).
- Visualize the clusters formed by hierarchical clustering.
- Evaluate the clusters using the **Silhouette Score**.

Part 3: Dimensionality Reduction with PCA

- Apply PCA to reduce the dimensionality of the data.
- Visualize the clusters in the PCA-reduced space.
- Analyze the **explained variance ratio** and discuss how much variance is retained.

t-SNE for Dimensionality Reduction

- Apply **t-SNE** to reduce the dataset to 2 dimensions and visualize the clusters.
- Compare **t-SNE** with **PCA** in terms of:
    1. **Cluster Separation**: Which technique provides clearer separation of clusters?
    2. **Computational Time**: Report the time taken to run t-SNE vs PCA.
    3. **Interpretability**: Discuss which method makes it easier to interpret the clusters.

Part 4: Evaluation and Comparison

- **Compare the Clustering Methods**:
    - Use the **Silhouette Score** and **Davies-Bouldin Index** to evaluate the clustering performance (K-Means, Hierarchical Clustering).
    - Summarize the results and performance of each algorithm in a table.
    - Discuss which algorithm performed best for this dataset and why.