Syed Darain Hyder Kazmi (a.k.a Sawab_e_Darain)

TASK # 22

Unsupervised Learning & Clustering

1. Introduction

Unsupervised learning works with unlabeled data to find hidden structures or relationships. It's challenging because there's no predefined truth to evaluate results. Issues like choosing the right features, scaling, and interpreting clusters make it tricky. Still, it's valuable for exploration, anomaly detection, and feature learning.

Real-world examples:

- Customer segmentation – group buyers by purchase behavior.
- Topic modeling – find themes in articles or reviews.
- Fraud detection – detect irregular spending patterns.
- Image segmentation – cluster pixels into regions.

Assessment 1: MovieLens Project – cluster users or movies to improve recommendations.

Assessment 2: FoodHub – segment customers or group recipes using order and menu data.

---

2. Clustering

Clustering groups data points based on similarity when no labels exist. It's useful for structure discovery, summarizing datasets, and input to recommendation or anomaly detection systems.

---

3. K-means Clustering

K-means assumes spherical, equal-sized clusters and needs a predefined number of clusters (K). Choosing the wrong K can create meaningless or overlapping groups. The algorithm cycles through initialization, assignment, update, and convergence to minimize within-cluster variance.

Cluster Evaluation Methods:

- Silhouette Score – higher means better separation.
- Davies–Bouldin Index – lower indicates better clusters.
- Elbow Method – helps pick suitable K.

- Domain interpretation – real-world sense check.

---

4. Beyond K-means

A cluster depends on the distance metric and data shape. Metrics like cosine similarity, Manhattan distance, or Mahalanobis distance suit different data. DBSCAN or hierarchical clustering handle irregular or unknown cluster counts better than K-means. Scaling and dimensionality reduction help improve performance.

---

5. Beyond Clustering

- Dimensionality reduction (PCA, t-SNE, UMAP).
- Topic modeling (LDA) for text data.
- Anomaly detection via distance-based outliers.
- Representation learning using embeddings or autoencoders.

---

6. Case Study: Geographic Clustering with Socio-Economic Data

Objective: Cluster regions using income, education, employment, and healthcare metrics to reveal inequality patterns.

Method: Standardized features, applied K-means (K=4), evaluated with Silhouette score 0.67.

Findings:

- Cluster 1 – high income, urban zones.
- Cluster 2 – middle income, developing regions.
- Cluster 3 – low income, limited healthcare.
- Cluster 4 – rural, agriculture-dominant areas.

Insight: Helps policymakers target underdeveloped areas and allocate resources efficiently.
