UE23CS352A: Machine Learning Lab   **Week 13: Clustering**
NAME:- DRISHTI GOLCHHA   **SRN:- PES2UG23CS185**

1.Dimensionality Justification:

Dimensionality reduction was necessary because the dataset had many correlated features that carried overlapping information.

PCA helped remove redundancy and made visualization easier.

The **first two principal components** captured about **60–70% of the total variance**, which means they represent most of the meaningful information in fewer dimensions.

2. Optimal Clusters

From the **elbow curve**, the bend appeared around **k = 3**, where inertia stopped decreasing sharply.

The **silhouette score** was also highest near **3 clusters**, indicating clear separation and compactness.

Hence, **3 is the optimal number of clusters**, supported by both metrics.

3. Cluster Characteristics

Some clusters are larger because they represent more common customer behaviors or profiles (e.g., average income and loan patterns).

Smaller clusters may represent niche or outlier groups, like high-balance or frequent campaign responders.

This shows that customer segments vary in size and behavior frequency.

4. Algorithm Comparison

The **Bisecting K-Means** algorithm usually gave a slightly **higher silhouette score** than standard K-Means.

This happened because it recursively splits clusters, leading to more balanced and well-separated groups.

So, **Bisecting K-Means performed a bit better** for this dataset.

5. Business Insights

The clusters indicate **different customer types**, such as:

- **Cluster 0:** Younger customers with low balances
- **Cluster 1:** Middle-aged or stable-income customers
- **Cluster 2:** Older or high-balance customers
  These insights help the bank target specific groups for marketing — for example, offering loans to Cluster 0 and investment products to Cluster 2.

6. Visual Pattern Recognition

The **turquoise, yellow, and purple** regions in the PCA plot represent distinct customer groups with different financial behaviors.

Sharp boundaries indicate strong differences between groups (e.g., income or loan status), while **diffuse boundaries** suggest overlap — meaning some customers share mixed traits between segments.

Explained Variance by Components

Data Distribution in PCA Space

✅ Shape after PCA: (45211, 2)

Inertia Plot (Elbow Method)

Silhouette Score vs Number of Clusters

K-Means Cluster Sizes

Silhouette Distribution per Cluster

Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39