

UE23CS352A: MACHINE LEARNING

Week 13: Clustering

Name : CHARAN M REDDY

SRN : PES2UG23CS146

Section : C

Date : 11/11/2025

1. Analysis Questions

- 1. Dimensionality Justification:** Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset? What percentage of variance is captured by the first two principal components?

The correlation heatmap shows that most features have weak correlations, meaning the dataset contains redundant or overlapping information. Applying PCA will help to reduce this redundancy and compress the data into fewer, more meaningful dimensions without information loss.

From the explained variance plot, the first two principal components together capture roughly 27.9% of the total variance.

- 2. Optimal Clusters:** Looking at both the elbow curve and silhouette scores, what is the optimal number of clusters for this dataset ? Justify your answer using both metrics.

Elbow Curve : The inertia drops steeply from $k=2$ to 3 , but after $k=3$, the decrease isn't that steep . That means adding more clusters after 3 does not significantly reduce the cluster variation.

Silhouette Score : Score peaks around $k=3$, which shows that the clusters are most separated and cohesive at that value. After $k=3$, the scores change or slightly decrease, which means additional clusters will not improve the quality of separation.

Justification : Both metrics independently support $k=3$, so it is the elbow point in the inertia plot and the highest silhouette score.

3. Cluster Characteristics: Analyze the size distribution of clusters in both K-means and Bisecting K-means. Why do you think some clusters are larger than others? What might this tell us about the customer segments?

In the K-Means model, cluster sizes are moderately balanced, indicating that most customers share similar characteristics. The Bisecting K-Means model shows more uneven cluster sizes, as it repeatedly splits the largest cluster, leading to one or two dominant groups and several smaller ones. Larger clusters likely represent the majority of customers with average financial behavior, while smaller clusters capture niche segments such as high-value or high-risk clients. This variation suggests the customer base is diverse and would benefit from differentiated marketing strategies.

4. Algorithm Comparison: Compare the silhouette scores between K-means and Recursive Bisecting K-means. Which algorithm performed better for this dataset and why do you think that is?

Standard K-Means (0.39) achieved a slightly higher silhouette score than Recursive Bisecting K-Means (0.34).

This indicates that K-Means formed clusters that were more compact and better separated. The difference is because standard K-Means optimizes all clusters simultaneously, allowing adjustment of centroids at a time , but Recursive Bisecting K-Means performs local binary splits that may hold some clusters in not optimal positions early on.

5. Business Insights: Based on the clustering results in the PCA space, what insights can you draw about customer segmentation that might be valuable for the bank's marketing strategy?

The PCA-based clustering reveals three main customer segments. One group represents high-balance, stable customers . Another group represents , average account holders , and the third group represents, low-balance or loan-dependent customers. These insights help the bank tailor its marketing—premium services for high-value clients, cross-selling to mid-tier customers, and financial support improving targeting and engagement.

6. Visual Pattern Recognition: In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

Turquoise : Customers who share similar characteristics like moderate income and spending patterns.

Yellow region : Customers with lower income but higher spending rate or a distinct behavioral pattern.

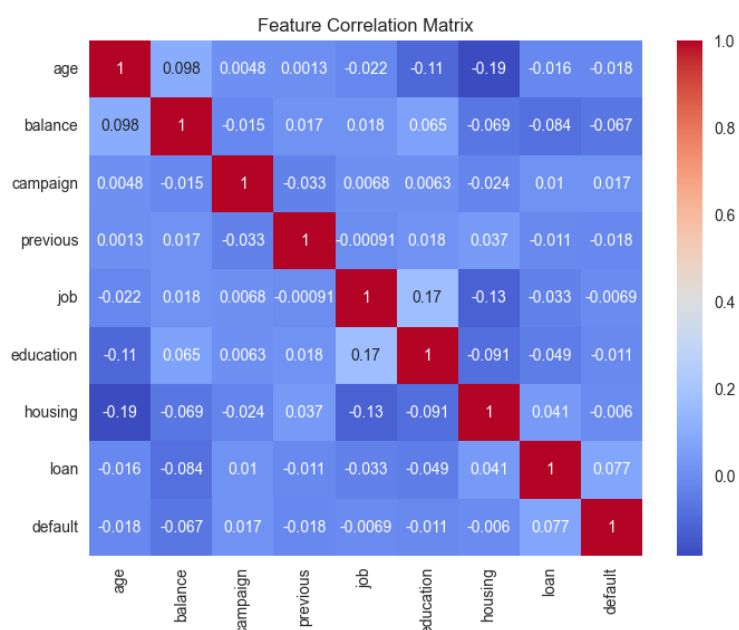
Purple region : Customers who are high-income, low-spending or other another clear behavioral change.

Sharp boundaries occur when customers in different clusters have clearly distinct patterns

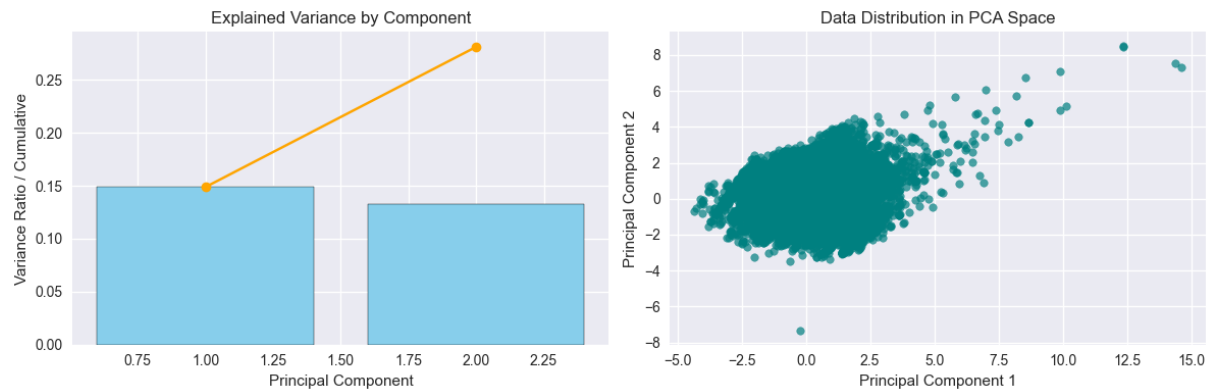
Diffuse or overlapping boundaries happen when customers have mixed or gradual transitions in their characteristics .

2. Screenshots

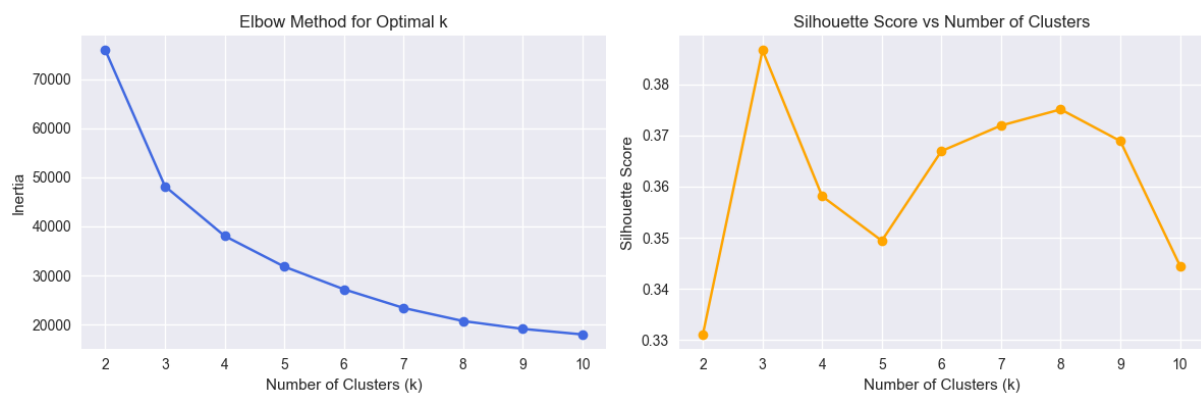
1. Feature Correlation matrix for the dataset



2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



3. 'Inertia Plot' and 'Silhouette Score Plot' for K-means



4. K-means Clustering Results with Centroids Visible (ScatterPlot)

K-means Cluster Sizes (Bar Plot)

Silhouette distribution per cluster for K-means (Box Plot)

