

## UE23CS352A: MACHINE LEARNING

### ML Lab Week 13

Nidhi K	PES2UG23CS383	SECTION F
Date:15-11-2025		

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?

Ans: Dimensionality reduction was necessary because although the correlation heatmap shows mostly low or moderate correlations, the dataset still contains several features that do not contribute equally to variance. PCA helps by projecting the data into a smaller number of orthogonal directions that capture the most meaningful variation while removing noise and redundancy. It also enables visualization of the dataset in 2D space, which is important for cluster interpretation.

From the PCA output, the first two principal components together capture approximately **28% of the total variance**. Even though this is not a very high percentage, it is sufficient to reveal visible structure in the data and to support clustering visualization.

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

Ans: Based on the **elbow curve**, the inertia decreases sharply between  $k = 1 \rightarrow 2$  and

$k = 2 \rightarrow 3$ , but the reduction slows significantly after  $k = 3$ . This indicates that adding more clusters provides diminishing returns in reducing within-cluster variability.

The **silhouette score plot** also peaks at  $k = 3$  with a value around **0.39**, which is the highest score among all tested values of  $k$ . This means that  $k$

$k = 3$  yields the best balance between compact and well-separated clusters.

Therefore, the optimal number of clusters for this dataset is clearly  $k = 3$ , supported by both inertia and silhouette metrics.

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?

Ans: Both K-means and Bisecting K-means reveal that customer behavior in this dataset is not evenly distributed, as each method produces clusters of noticeably different sizes. In standard K-means, the clusters were approximately 10,000, 15,000, and around 20,000 customers, showing that one behavioural pattern dominates while others are less common. Bisecting K-means produced a slightly different distribution—11,350 customers in Cluster 0, 20,156 in Cluster 1, and 13,705 in Cluster 2—yet the pattern remains consistent: there is one large, dominant cluster and two smaller, more specialized groups. This indicates that most customers share a common behavioural profile, while a significant number belong to smaller, more distinct segments.

The larger clusters likely represent mainstream customers who exhibit typical banking behaviour, such as common balance ranges, similar campaign interactions, or standard loan and housing patterns. The smaller clusters, on the other hand, probably contain customers with more unique characteristics—perhaps higher account balances, different job or education categories, or unusual levels of engagement with marketing campaigns. These differences in size highlight meaningful variation within the customer base and suggest opportunities for targeted segmentation. Smaller clusters may indicate niche groups that are important for tailored marketing or custom financial products, while the large group may require broad, general strategies.

Overall, the cluster size distributions from both algorithms suggest a naturally imbalanced customer population. This is expected in real-world banking data, where certain customer profiles dominate the dataset. Understanding these size differences allows the bank to identify which

customer groups require specialized attention and which represent the core customer base.

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?

Ans: When comparing the clustering performance of K-means and Bisecting K-means, the silhouette scores offer clear evidence that standard K-means performs better for this dataset. K-means achieved a silhouette score of approximately **0.39**, while Bisecting K-means scored **0.2915**. Since a higher silhouette score indicates more distinct and well-separated clusters, the results show that K-means forms tighter and more meaningful groupings, whereas the clusters formed by the bisecting method are less clearly separated. This suggests that the structure of the dataset aligns more naturally with the assumptions of the K-means algorithm.

One key reason for this difference is how the two algorithms operate. Standard K-means optimizes all cluster boundaries simultaneously, which works well when clusters are compact and relatively spherical—conditions that are visible in the PCA scatter plot. Bisecting K-means, however, repeatedly splits only the largest cluster at each step, which can produce clusters that are uneven in shape or less well separated. This recursive splitting makes Bisecting K-means more sensitive to internal variations within the largest segment, potentially leading to boundaries that do not capture natural separations in the data as effectively as K-means.

In conclusion, K-means is the more suitable algorithm for this dataset because it produces clearer, more coherent clusters, as reflected in its higher silhouette score. While Bisecting K-means still provides a valid hierarchical segmentation structure—which can be useful for exploratory analysis—it does not match the clustering performance of standard K-means in this case.

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?

Ans: The clusters formed in PCA space reveal meaningful differences among customer groups. These insights can support marketing strategies in several ways:

- Customers in high-balance or high-engagement clusters may be ideal targets for **premium products**, investments, or long-term deposits.
- Segments with high campaign response or previous interactions may benefit from **personalized and more frequent marketing outreach**.
- Customers in low-engagement or low-balance clusters may require **different messaging**, incentives, or channels to increase retention.
- Identifying niche clusters helps the bank prioritize **specific targeted campaigns**, improving conversion rates and reducing unnecessary marketing costs.

Overall, clustering enables the bank to move from mass marketing to tailored, segment-specific strategies.

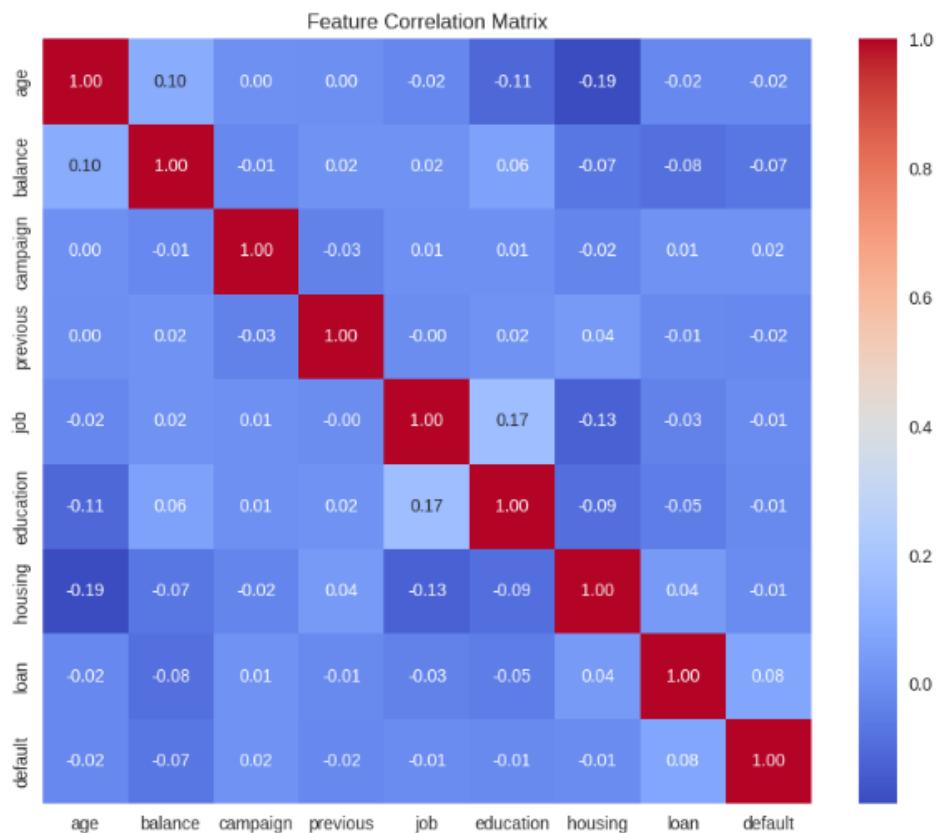
6. Visual Pattern Recognition: In the PCA scatter plot, we see three distinct coloured regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

Ans: In the PCA scatter plot, the turquoise, yellow, and purple regions represent groups of customers who share similar characteristics. Each colour corresponds to a different cluster formed by K-means, meaning customers in the same region have comparable behaviours or attributes such as balance levels, loan or housing status, campaign interactions, or demographic traits. Customers in different regions differ more strongly in these characteristics, which is why they appear as separate coloured areas.

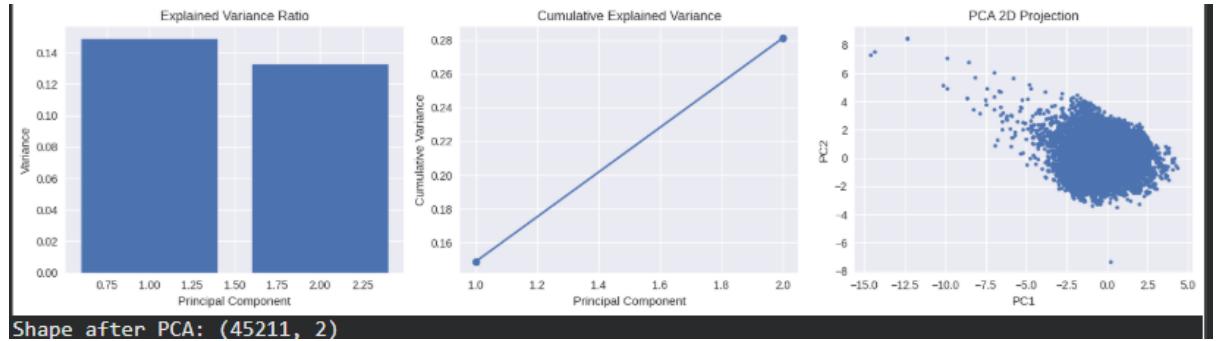
The boundaries between these regions can be sharp when customer groups are clearly distinct, meaning PCA captures strong differences between them. In contrast, boundaries appear diffuse when customers have overlapping or mixed characteristics or when reducing many features into just two PCA components causes some natural overlap. This creates smooth transitions between clusters rather than strict separations.

## Screenshots:

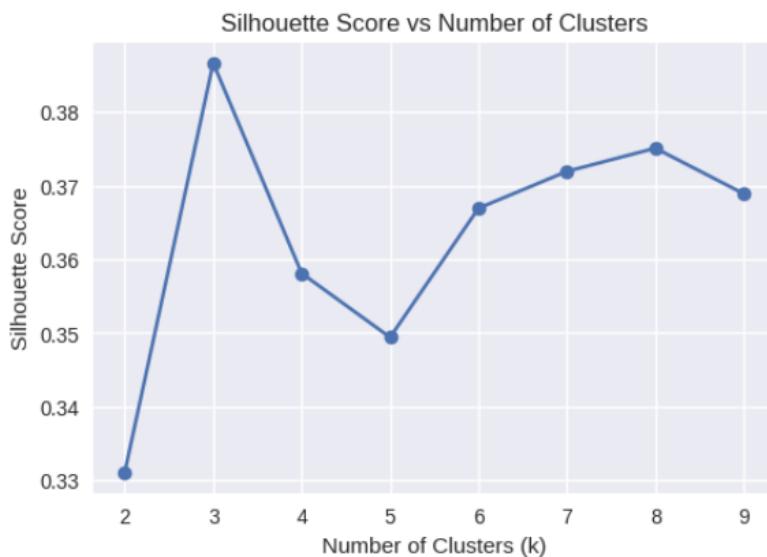
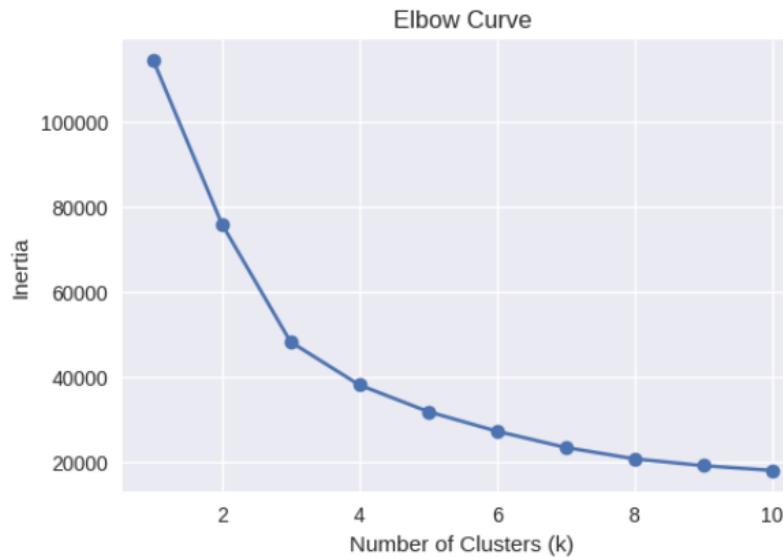
### 1. Feature Correlation matrix for the dataset



## 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 (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)

