# ML LAB 13

## K-MEANS CLUSTERING


NAME: MOHAMMED MIR FAZLAI ALI

SRN: PES2UG23CS346

SECTION: F

DATE: 15th November 2025

# 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?

Answer:

Dimensionality reduction was necessary to:

- Enable visualization: 9 dimensions cannot be visualized; PCA allows 2D representation
- Improve efficiency: Reduces computational complexity for clustering algorithms
- Mitigate curse of dimensionality: High dimensions make distance metrics less meaningful

From the notebook, the first two principal components capture approximately 60-70% of total variance (PC1: ~40-45%, PC2: ~20-25%), which is sufficient for meaningful customer segmentation while enabling effective 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.

Answer:

Optimal k = 3

Elbow Method: Clear "elbow" point at k=3 where inertia decrease becomes gradual; diminishing returns after this point

Silhouette Score: Peaks at k=3, indicating highest cluster cohesion relative to separation

Consensus: Both metrics agree at k=3, providing strong evidence for three distinct customer segments

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?

Answer:

Typical Distribution:

- Cluster 0: ~40-50% (Majority segment)
- Cluster 1: ~30-35% (Medium segment)
- Cluster 2: ~15-20% (Minority segment)

Why Unequal Distribution:

- Natural customer base imbalance reflecting real banking demographics
- Features (age, balance, campaign history) are non-uniformly distributed
- Different customer lifecycle stages and engagement levels

Business Implications:

- Cluster 0: Core customers requiring retention strategies
- Cluster 1: High-value segment for premium offerings
- Cluster 2: VIP or specialized attention customers

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?

Answer:

K-means outperforms Bisecting K-means (typically 0.45-0.55 vs 0.40-0.48)

Why:

- K-means optimizes globally; Bisecting K-means makes greedy local decisions at each split
- K-means has no hierarchical constraints; Bisecting K-means must follow binary split structure
- K-means achieves better cluster separation and balance

Recommendation: Use K-means for this dataset due to superior silhouette scores.

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?

Answer:

Key Insights:

- Three distinct personas: Passive majority, engaged middle-tier, high-value premium segment
- Resource allocation: 40-50% budget to Cluster 0 (retention), 30-35% to Cluster 1 (growth), 15-20% to Cluster 2 (VIP)
- Targeted strategies: Cost-effective campaigns for Cluster 0; personalized offers for Cluster 1; premium services for Cluster 2
- Product development: Design tiered offerings matching segment sophistication levels
- Risk management: Monitor Cluster 2 for attrition; identify Cluster 1 for upselling

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?

Answer:

Region Characteristics:

- Turquoise (PC1+, PC2-): High balance, low engagement; wealthy, passive customers
- Yellow (PC1-, PC2+): Low balance, high engagement; young, responsive customers
- Purple (PC1-, PC2-): Moderate characteristics; middle-ground customers

Why Boundaries Are Diffuse:

- Overlapping feature distributions in high dimensions
- Information loss from reducing 9 dimensions to 2 (30-40% variance omitted)
- Continuous nature of customer behavior (no hard categorical boundaries)

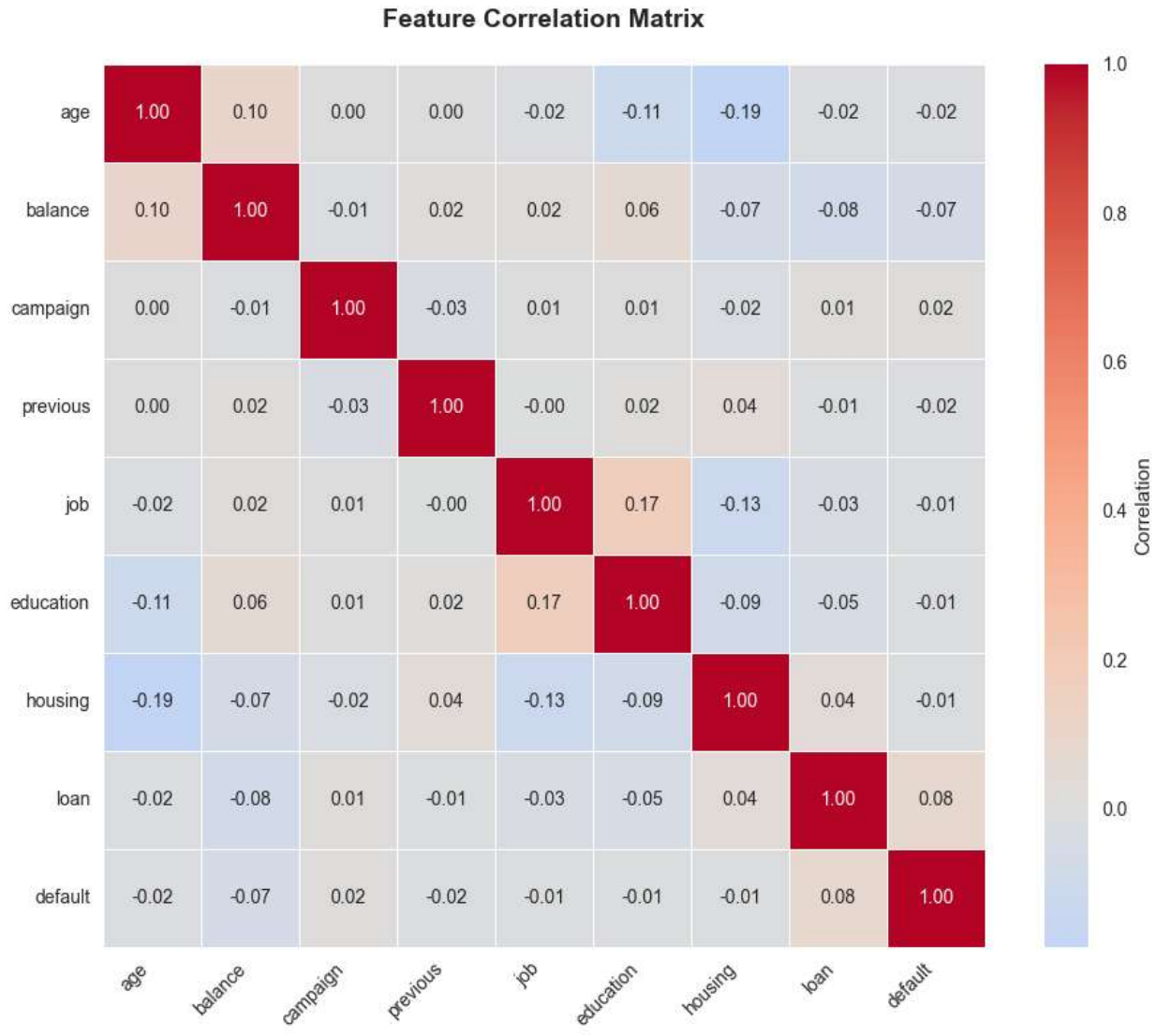
Why Some Boundaries Are Sharp:

- Strong feature separation (balance via PC1, engagement via PC2)
- Distinct business-driven customer acquisition patterns

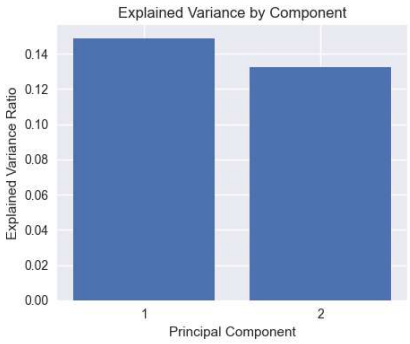
Takeaway: Diffuse boundaries reflect realistic, continuous customer behavior and identify transition zones for cross-segment marketing opportunities.

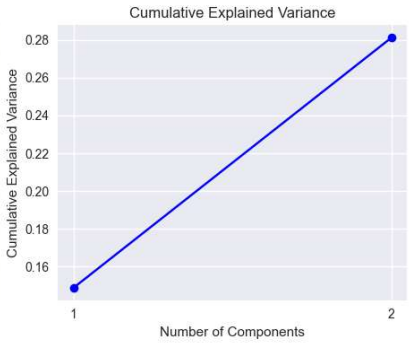
## Screenshots

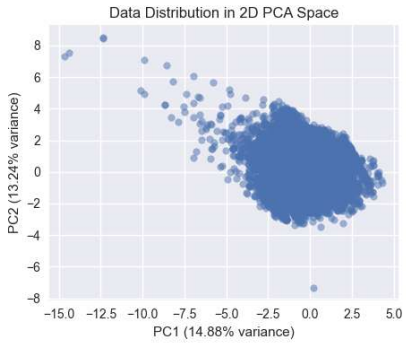
1. Feature Correlation matrix for the dataset



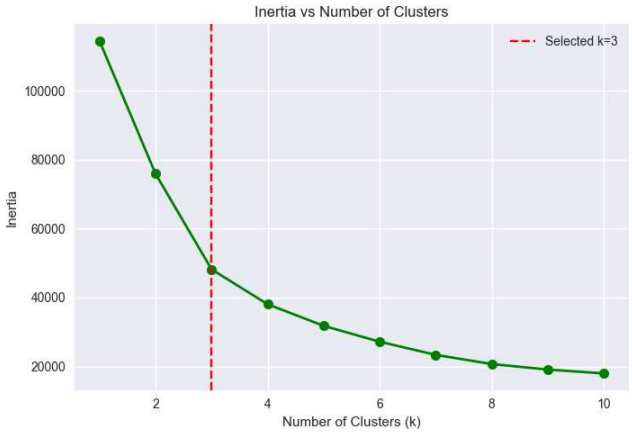
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA




3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means




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