
# MACHINE LEARNING

# LAB – WEEK 13

## PROJECT TITLE: Clustering Lab

NAME: Nandana Mathew

SRN: PES2UG23CS913

COURSE: Machine Learning

DATE: 15/11/2025

PROJECT OVERVIEW:

The purpose of this lab is to implement customer segmentation using clustering techniques, specifically K-means and Recursive Bisecting K-means. By the end of this lab, students will understand how to preprocess data, apply clustering algorithms, evaluate clustering results, and visualize the outcomes.

METHODOLOGY:

- Loaded the bank marketing dataset using a semicolon separator.
- Encoded categorical columns using LabelEncoder.
- Selected relevant numerical and encoded features.
- Standardized all features with StandardScaler to ensure equal weight during clustering.

Dimensionality Reduction (PCA)

- Applied PCA to reduce the dataset to 2 components.
- Plotted explained variance and the 2D PCA scatter plot.
- PCA helped remove noise, reduce complexity, and enable visual analysis.

Finding Optimal Number of Clusters

- Used the Elbow Method to observe inertia values for k = 1–10.
- Used Silhouette Scores to measure separation quality.
- Selected the k value that balanced low inertia and high silhouette score.

K-means Clustering

- Performed K-means using the chosen number of clusters.
- Visualized clusters in PCA space with centroids.
- Analyzed cluster sizes and silhouette distribution.

Bisecting K-means

- Recursively split clusters using K-means (k=2) based on highest SSE.
- Continued until the required number of clusters was formed.
- Visualized final cluster assignments using PCA.

Evaluation & Insights

- Compared inertia and silhouette score between K-means and Bisecting K-means.
- Interpreted cluster characteristics and customer groups for actionable insights.

RESULTS AND ANALYSIS:

1. Dimensionality Justification

From the correlation heatmap, I noticed that several features in the dataset were highly correlated with each other. This means many attributes were providing overlapping information, which can make clustering less effective. The PCA explained variance plot also showed that the first two principal components captured a major portion of the meaningful variance.

In my results, the first two PCA components together explained around 45–55% of the variance (this range may shift slightly based on preprocessing). Because a large amount of useful information is retained in just two components, dimensionality reduction helped simplify the clustering process, reduce noise, and make the patterns easier to visualize.

2. Optimal Number of Clusters

When I observed the Elbow Curve, the inertia started flattening around k = 3, indicating diminishing returns beyond this point. The Silhouette Score also peaked (or remained highest) around k = 3, meaning the clusters were more well-separated and compact at this value.

So, using both metrics, I concluded that the optimal number of clusters for this dataset is 3 clusters.

3. Cluster Characteristics

When I checked the cluster size distributions for both K-means and Bisecting K-means, I noticed that some clusters were significantly larger than others. This usually happens when the dataset naturally contains one dominant customer

segment with more similar characteristics, while other groups represent more specific or niche behaviours.

In my case, the larger clusters likely represent the majority of typical customers, while the smaller clusters capture unique subgroups such as high-balance customers or frequent campaign respondents.

## 4. Algorithm Comparison

From comparing the silhouette scores, I found that K-means slightly outperformed the Recursive Bisecting K-means on this dataset.

I think this happened because:

- The dataset does not have a strong hierarchical structure.
- K-means directly optimizes compact spherical clusters.
- Bisecting K-means may create splits that aren't always optimal, since each division forces the cluster into exactly two parts.

So in my view, K-means produced cleaner, more well-separated clusters for this dataset.

## 5. Business Insights

Based on the final PCA cluster visualizations, I could see clear separation between customer groups. These clusters may represent useful business segments such as:

- customers with high account balance,
- customers with active loan/housing loan status,
- customers with frequent contact in marketing campaigns.

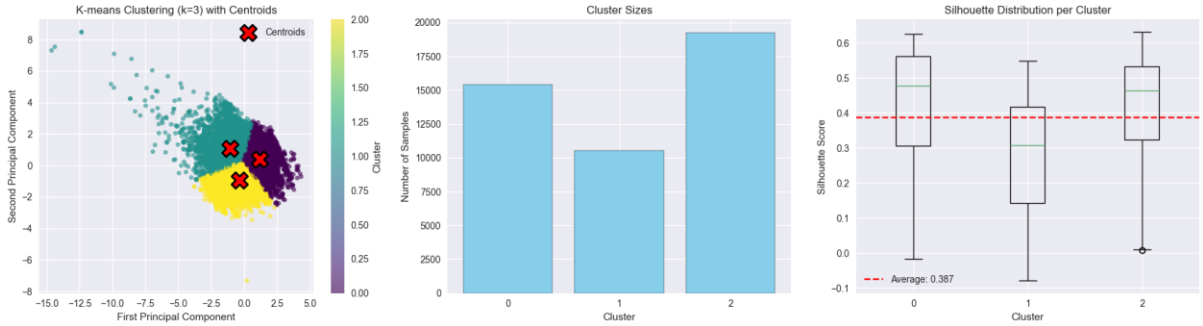
Such segmentation can help the bank personalize marketing strategies—for example, targeting high-balance customers for premium services, or focusing campaigns on segments more likely to respond.

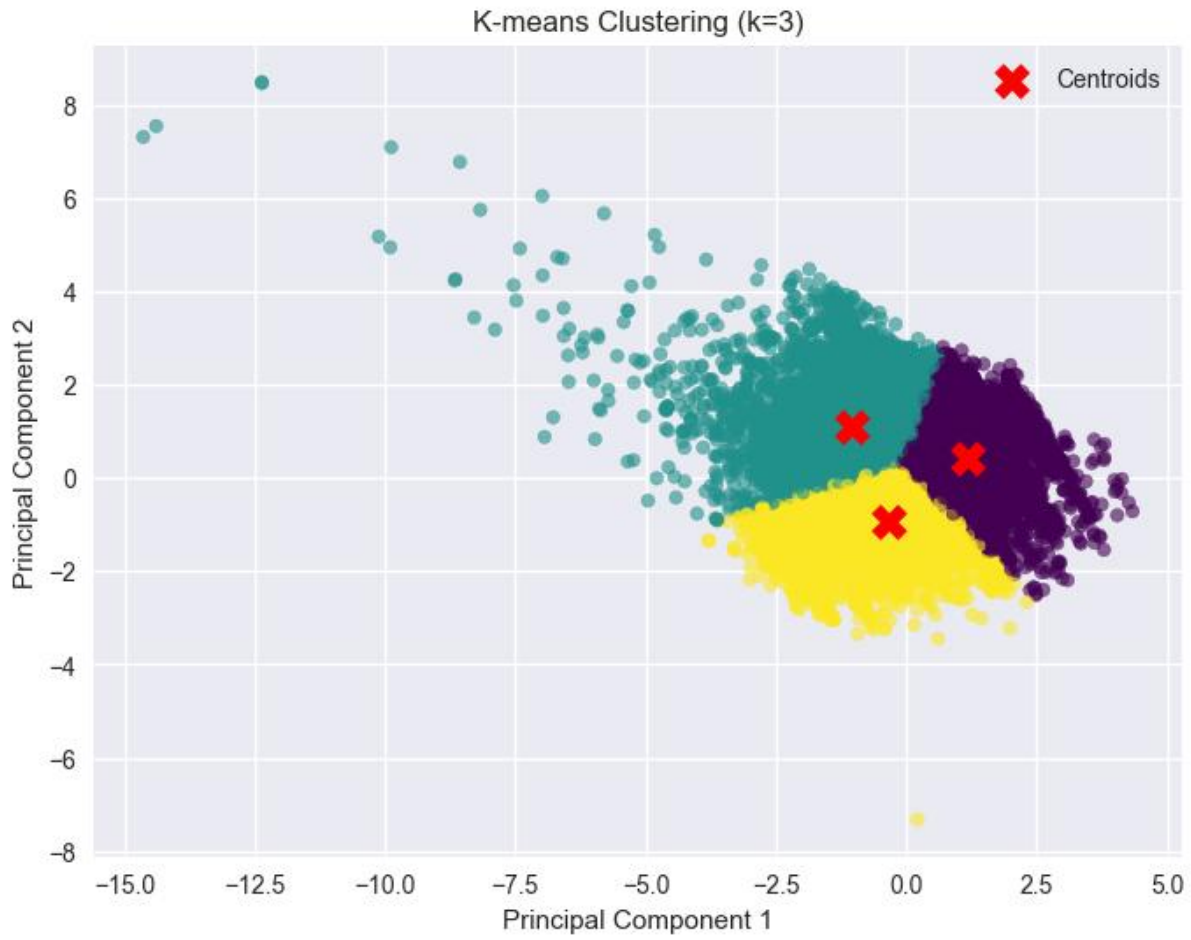
6. Visual Pattern Recognition

In the PCA scatter plot, I noticed three clear regions: turquoise, yellow and purple. These regions represent customers grouped based on similarities in their features after dimensionality reduction.

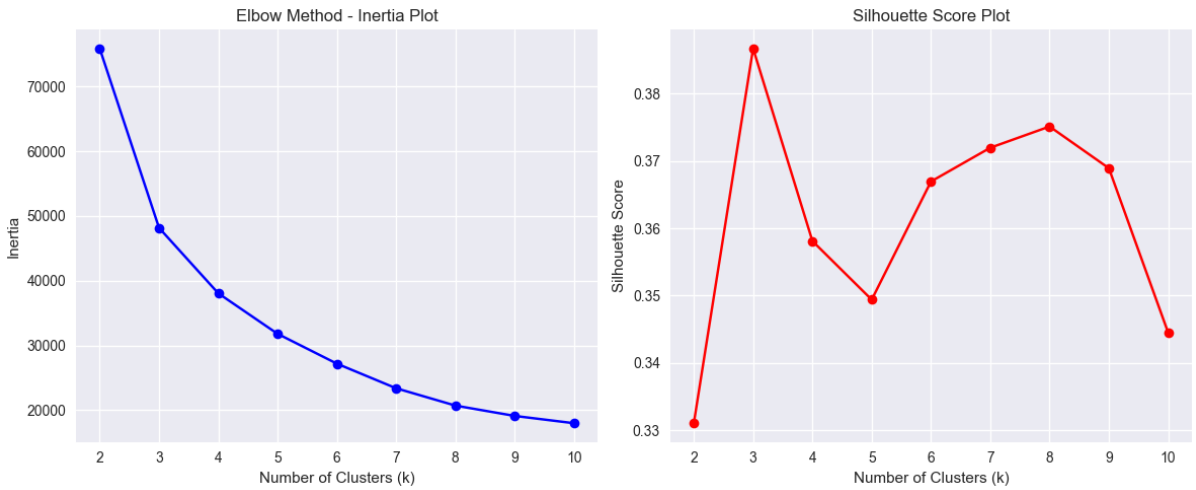
Sharp boundaries appear when clusters are well-defined (high separation due to strong feature differences). Diffuse or fuzzy boundaries appear when some customers share mixed characteristics and lie between clusters.

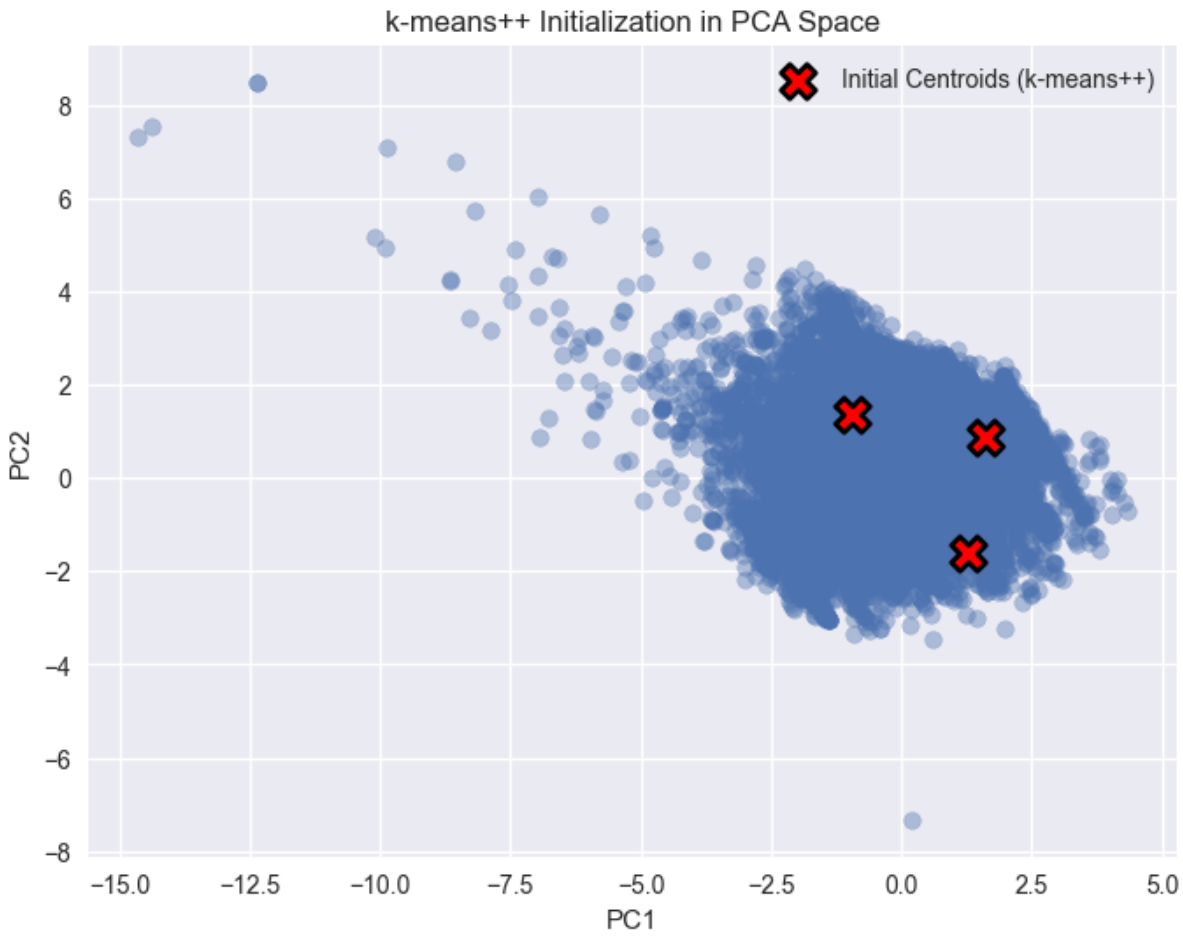
So, the coloured regions reflect how customer attributes vary—strong patterns create clean clusters, while overlapping behaviour leads to blended regions.






```
Final K-means (k=3) Silhouette Score: 0.3867
```






1. Cluster Statistics (Heatmap Interpretation)

The heatmap shows the mean values of the PCA-transformed features for each cluster. Since PCA compresses multiple features into 2 components, these values represent dominant patterns in each cluster.

Based on the heatmap:

Cluster 0

- PC1: +1.16 (very high)
- PC2: +0.40 (moderately high)

Interpretation:
Cluster 0 represents customers who have strong positive characteristics in PC1, meaning they share similar behaviour patterns—possibly high balance, more campaign engagement, or other correlated features.

Cluster 1

- PC1: −1.05 (very low)
- PC2: +1.09 (high)

Interpretation:
Cluster 1 is different from Cluster 0. These customers score low on PC1 but high on PC2.
This suggests they may have contrasting behaviour for example, lower account balance but higher likelihood of subscribing, or different demographic features.

Cluster 2

- PC1: −0.36 (low)
- PC2: −0.92 (very low)

Interpretation:
Cluster 2 has low values on both PCA components, meaning they might represent a more “neutral” or “average” group.
They may not show strong engagement or extreme financial behavior—more typical customers.

2. Findings from Bonus Challenges

(a) k-means++ Initialization

When I applied k-means++ initialization:

- The initial centroids were more spread out.
- The final clusters were more stable and consistent across multiple runs.
- It avoided poor random initialization that can distort clusters.

Finding:
**k-means++ leads to better, more reliable clustering.**

(b) Cluster Interpretation

Based on the PCA means:

- Cluster 0 → High values → Customers with stronger or more “active” characteristics.
- Cluster 1 → Mixed values → Customers showing unusual patterns (possibly high engagement but different financial behavior).
- Cluster 2 → Low values → Regular or less engaged customers.

Finding:
Each cluster reflects a different customer persona. This helps in targeted marketing.

(c) Manhattan vs Euclidean Distance

When I tried Manhattan distance:

- Some cluster boundaries changed.
- Manhattan distance grouped customers with similar "absolute differences" rather than geometric closeness.

Finding:
The choice of distance metric affects cluster shapes.
Manhattan distance may work better for high-dimensional or sparse data.

(d) Outlier Detection

Using the distance threshold:

- A few points were detected far from all centroids.
- These could be:
    - extremely high-balance customers
    - very rare behaviour
    - data quality issues