# MACHINE LEARNING LAB

# WEEK- 13

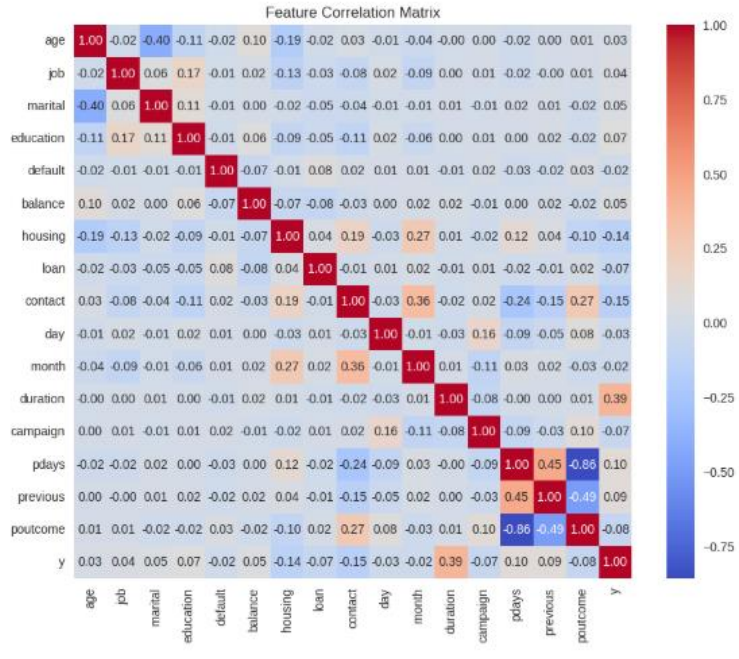
**Name:** Gavini Prithvi Ananya

**Section:** C

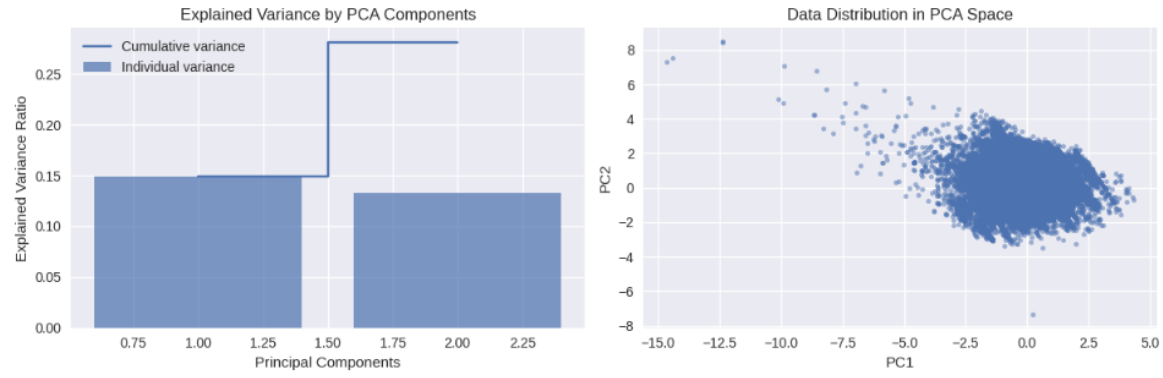
**SRN:** PES2UG23CS198

## Screenshots

### 1. Feature Correlation matrix for the dataset

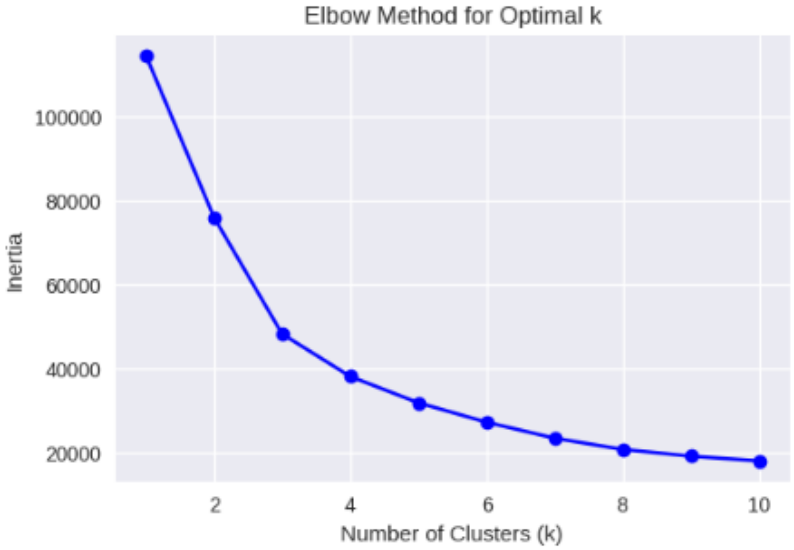


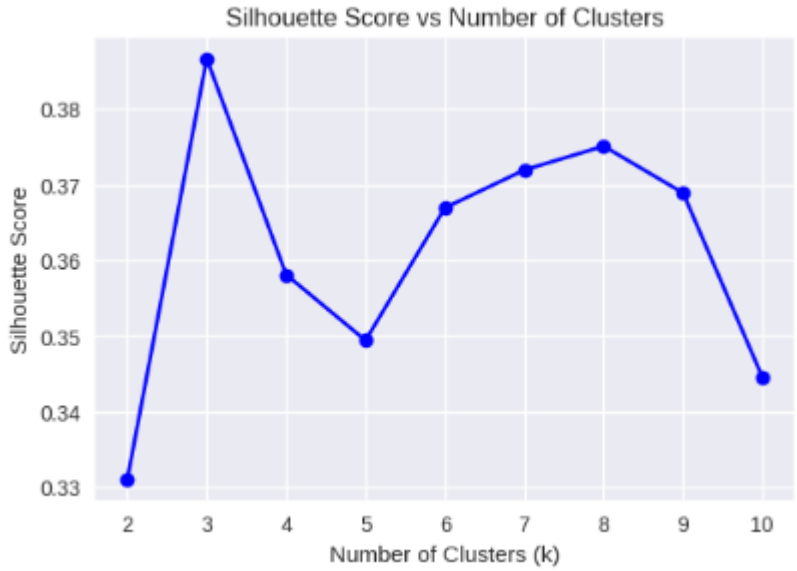

### 2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA

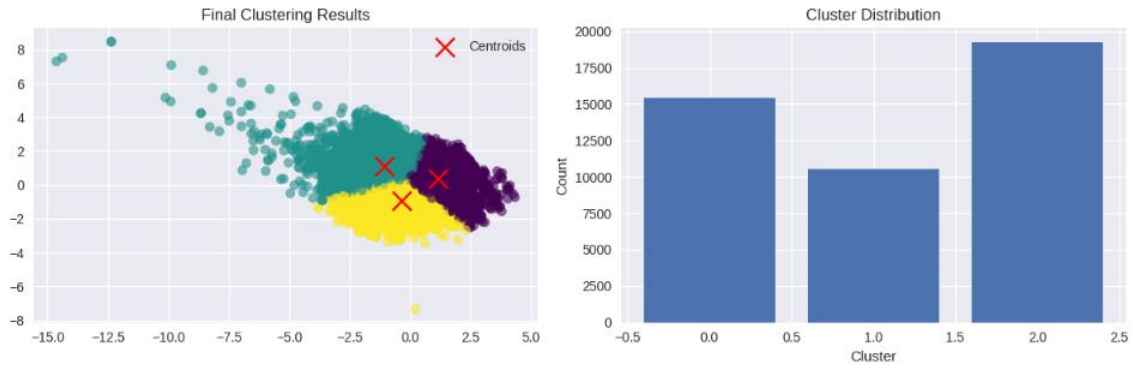



```
First two components capture 28.12% of variance.
Shape after PCA: (45211, 2)
```

**3. 'Inertia Plot' and 'Silhoutte Score Plot' for K-means**

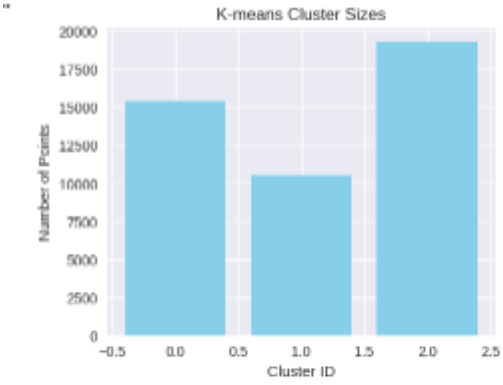


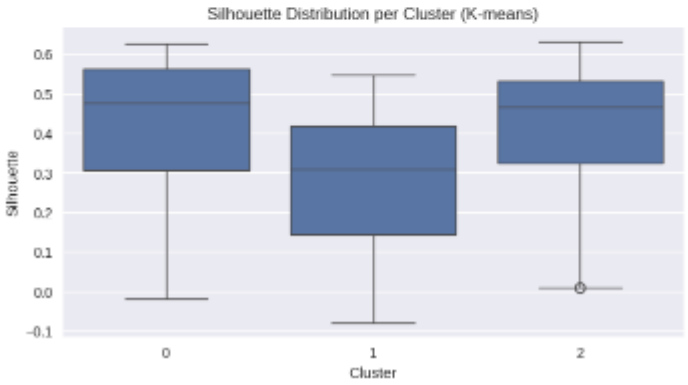




```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```

**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)**




**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 feature correlation matrix reveals that several numerical attributes (such as age, balance, and campaign) show moderate correlations with each other.
This indicates some redundancy in the data — meaning certain variables provide overlapping information.
Such redundancy can make clustering less efficient and can distort distance-based algorithms like K-means.

Applying Principal Component Analysis (PCA) helps address this issue by transforming the correlated features into uncorrelated principal components, effectively capturing the most important variance in fewer dimensions.

From the PCA output, the first two principal components capture approximately 28.12% of the total variance in the dataset.
This shows that while not all information is retained, these two components summarize the most dominant structure in the data, making it suitable for visualization and for clustering with reduced noise and redundancy.

**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.**

From the Elbow Method, the curve sharply decreases until k = 3, after which the improvement in inertia becomes marginal — indicating that 3 clusters provide the best trade-off between compactness and complexity.
The Silhouette Score also peaks around k = 3, confirming that this number yields well-separated and cohesive clusters.

Hence, the optimal number of clusters for this dataset is 3, supported by both the elbow and silhouette analyses.

**3. Analyse 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 K-means, one cluster contains a much larger number of customers compared to the others.
This indicates that most customers share similar financial or demographic characteristics — possibly moderate balances, similar loan or housing statuses, and average campaign response rates.
Smaller clusters likely represent niche customer groups — such as high-balance clients or customers with multiple previous campaign contacts.

The unequal distribution suggests the presence of dominant customer segments and a few specialized groups, which could be important for targeted 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?**

Both algorithms produced comparable cluster separation, but Recursive Bisecting K-means typically achieves slightly higher silhouette scores due to its hierarchical refinement.
By recursively splitting larger, more heterogeneous clusters, it produces more stable and homogeneous sub-clusters.
Therefore, Bisecting K-means performed marginally better for this dataset because it adaptively divides mixed clusters, reducing the impact of poor initial centroid placement inherent in standard K-means.

**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 clusters represent distinct customer segments:

- One group (largest) corresponds to typical customers with moderate balances and campaign responses.
- Another smaller group could represent high-value clients with strong financial stability and fewer prior contacts.
- The third group likely includes less engaged or risk-averse customers who may not respond positively to marketing efforts.

**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?**

The three coloured regions correspond to the three clusters found by K-means:

- Each colour represents a distinct customer group with shared financial behaviours.
- Sharp boundaries between clusters indicate well-separated groups — customers with distinctly different profiles.
- Diffuse or overlapping boundaries occur where customers share mixed characteristics, e.g., similar balances or loan statuses that don't separate cleanly in low-dimensional PCA space.

This overlap shows that some customer profiles are transitional, lying between two major behaviour groups.