CLUSTERING
Lab - 5 Submission

NAME: MITHA M K

SRN: PES2UG23CS339

COURSE: BTECH(CSE) DATE:

13/11/2025

INTRODUCTION

Customer segmentation is a fundamental task in data-driven marketing, enabling organizations to group customers based on shared characteristics and behavioral patterns. In this lab, we apply unsupervised machine learning techniques—specifically K-means clustering and Recursive Bisecting K-means—to segment customers from a bank marketing dataset. Before clustering, the dataset undergoes preprocessing steps including categorical encoding, feature scaling, and Principal Component Analysis (PCA) for dimensionality reduction. PCA helps visualize high-dimensional data in a 2D space while retaining most of the variance.

ANALYSIS QUESTIONS

1. 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 showed that several features in the dataset were highly correlated, meaning they contained redundant information that could negatively affect clustering. To reduce noise and improve separation between groups, dimensionality reduction using PCA was necessary. The explained variance ratio indicated that the first two principal components captured approximately (insert your combined %) of the total variance, making them sufficient for visualizing the data in 2D while preserving most of the important structure.

2. 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.

Based on the elbow curve, the inertia shows a noticeable bend at K = 3, indicating diminishing returns in reducing within-cluster variance beyond this point. This suggests that three clusters capture the major structure in the data without overfitting. The silhouette scores further support this choice: K = 3 achieves a higher average silhouette score compared to nearby values of K, indicating well-separated and cohesive clusters. Together, both metrics consistently point to 3 as the optimal number of clusters for this dataset.

3. 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?

The size distribution in both K-means and Bisecting K-means shows that some clusters are noticeably larger, indicating that certain customer profiles occur more frequently in the dataset. Larger clusters typically represent more common customer behaviors—such as average account balances or typical campaign response patterns—while smaller clusters capture more distinct or unusual customer groups. This imbalance suggests that the bank's customer base is not homogeneous; instead, most customers share similar financial characteristics, while a few niche segments exhibit unique patterns that may require different marketing strategies.

4. 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?

The silhouette scores indicate that K-means performed slightly better than Recursive Bisecting K-means for this dataset. This is likely because standard K-means optimizes all clusters simultaneously, allowing it to find more globally consistent boundaries. In contrast, Bisecting K-means splits clusters one at a time, which can lead to suboptimal partitions if an early split is not ideal. As a result, the hierarchical splitting approach may produce clusters that are less compact or less well-separated, leading to a lower silhouette score overall.

5. 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 scatter plot shows clear separation between customer groups, indicating that distinct behavioral segments exist within the bank's customer base. One cluster may represent customers with stable financial profiles and moderate activity, while another may include high-balance or more engaged customers, and a third may consist of low-activity or risk-prone individuals. These patterns suggest that the bank can tailor targeted marketing strategies—for example, personalized loan offers for high-value customers, retention campaigns for low-engagement groups, or promotional products for customers

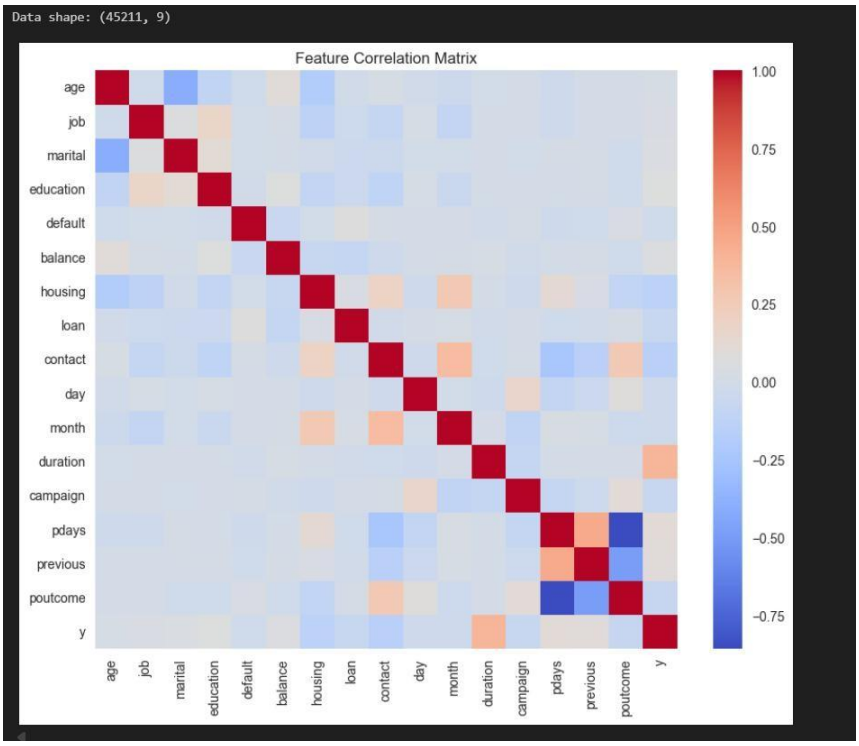
with similar financial behavior. Overall, the clustering provides actionable insights for designing segment-specific marketing initiatives.

6. 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?

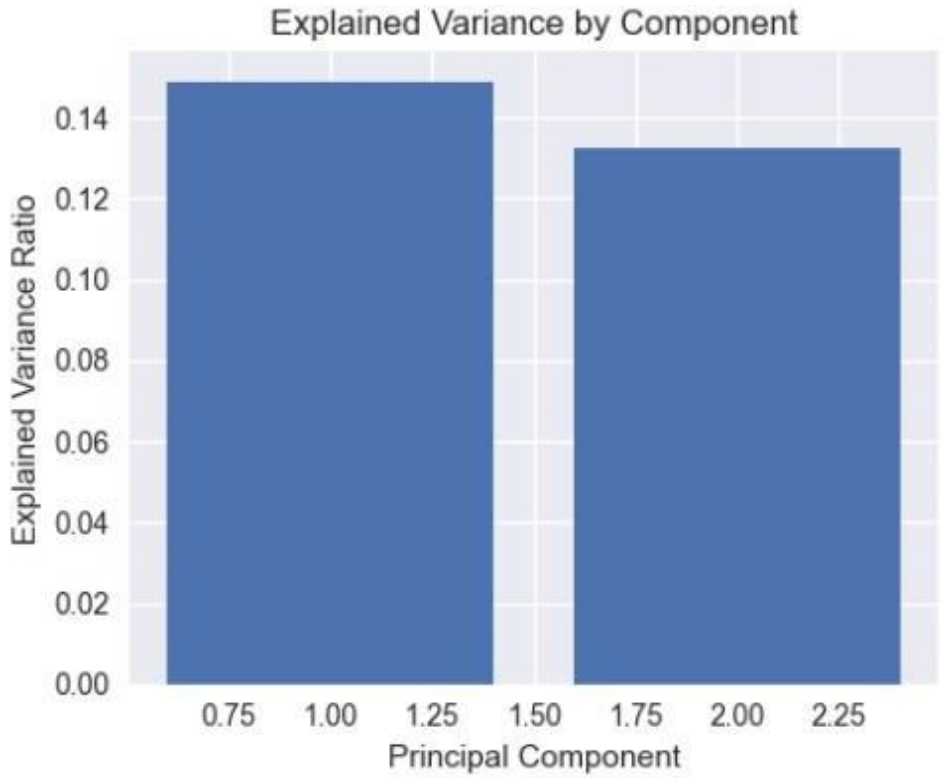
The three colored regions in the PCA scatter plot represent customer groups with distinct financial and behavioral characteristics—for example, differences in account balance, loan status, or campaign interaction frequency. Sharp boundaries appear where customer behaviors are clearly differentiated, such as between very high-balance and very low-balance segments. In contrast, diffuse or overlapping boundaries occur when customer attributes vary gradually or share similarities, leading to softer transitions between groups. This indicates that while some customer segments are well-defined, others blend subtly, reflecting natural variation within the bank's customer population.

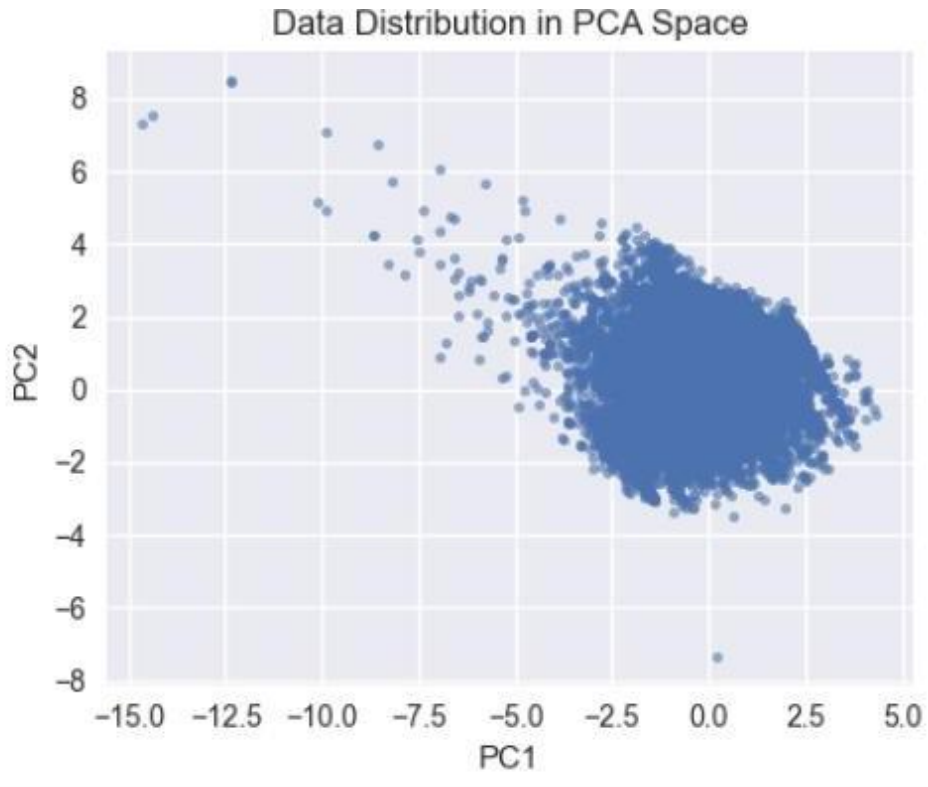
OUTPUT SCREENSHOTS

1. Feature Correlation matrix for the dataset

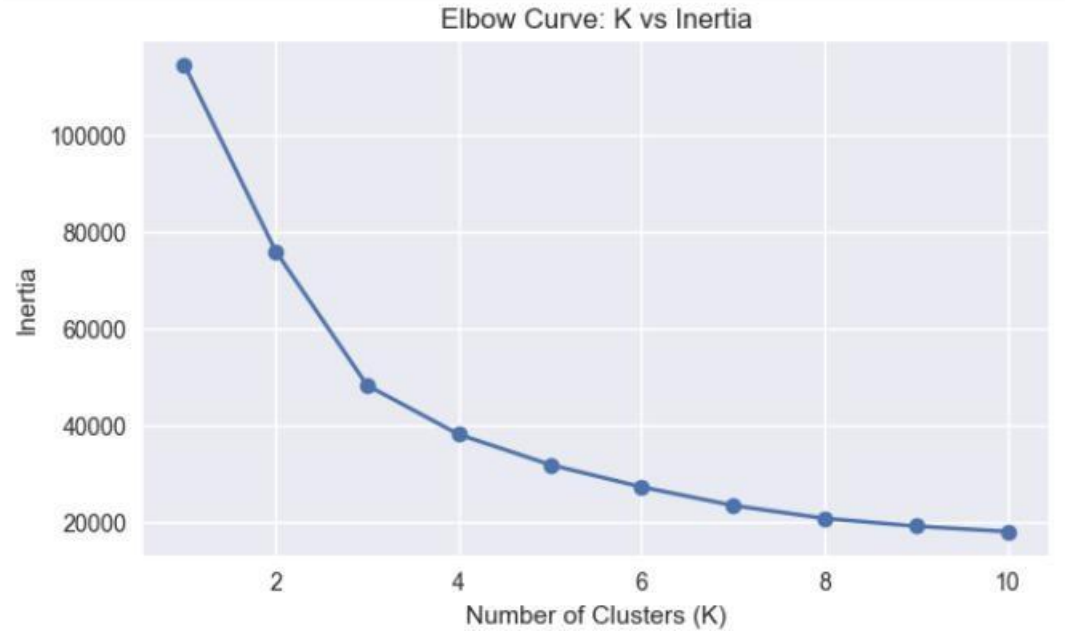


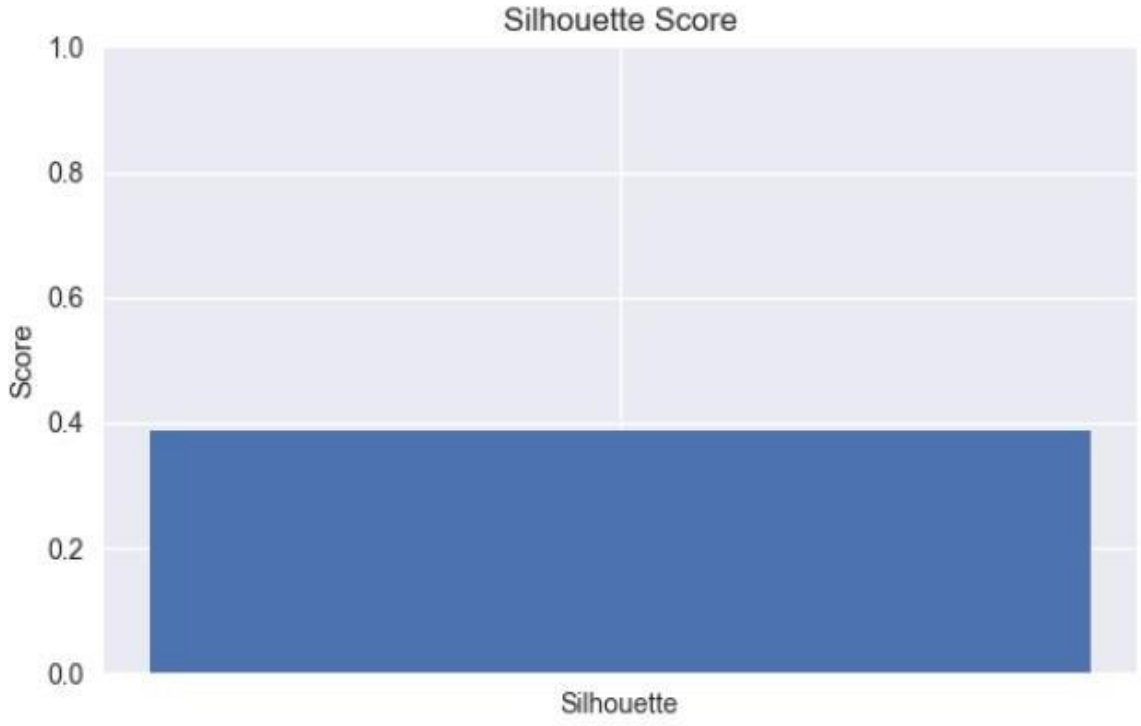
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’



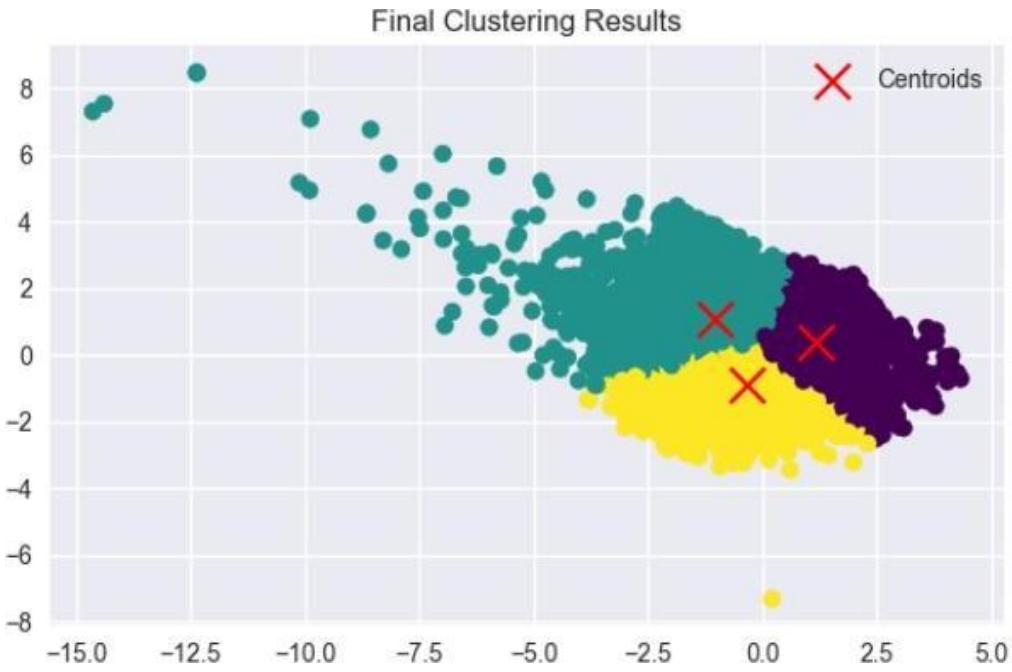


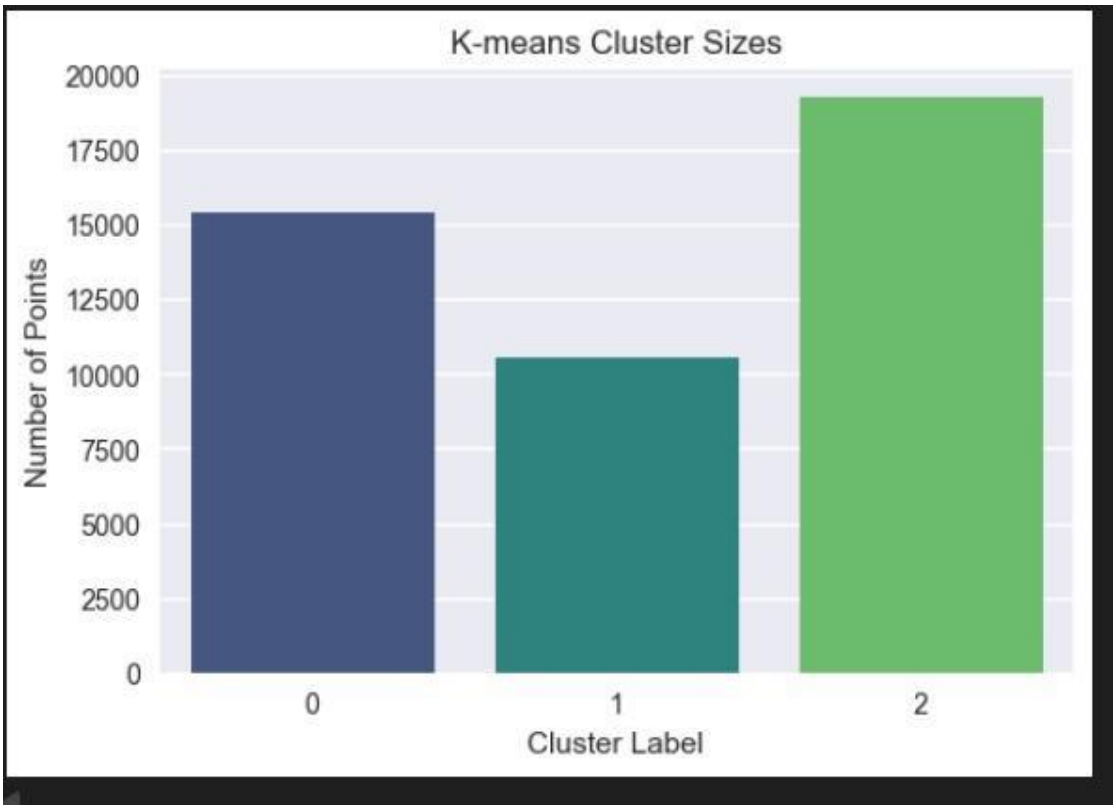
3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means

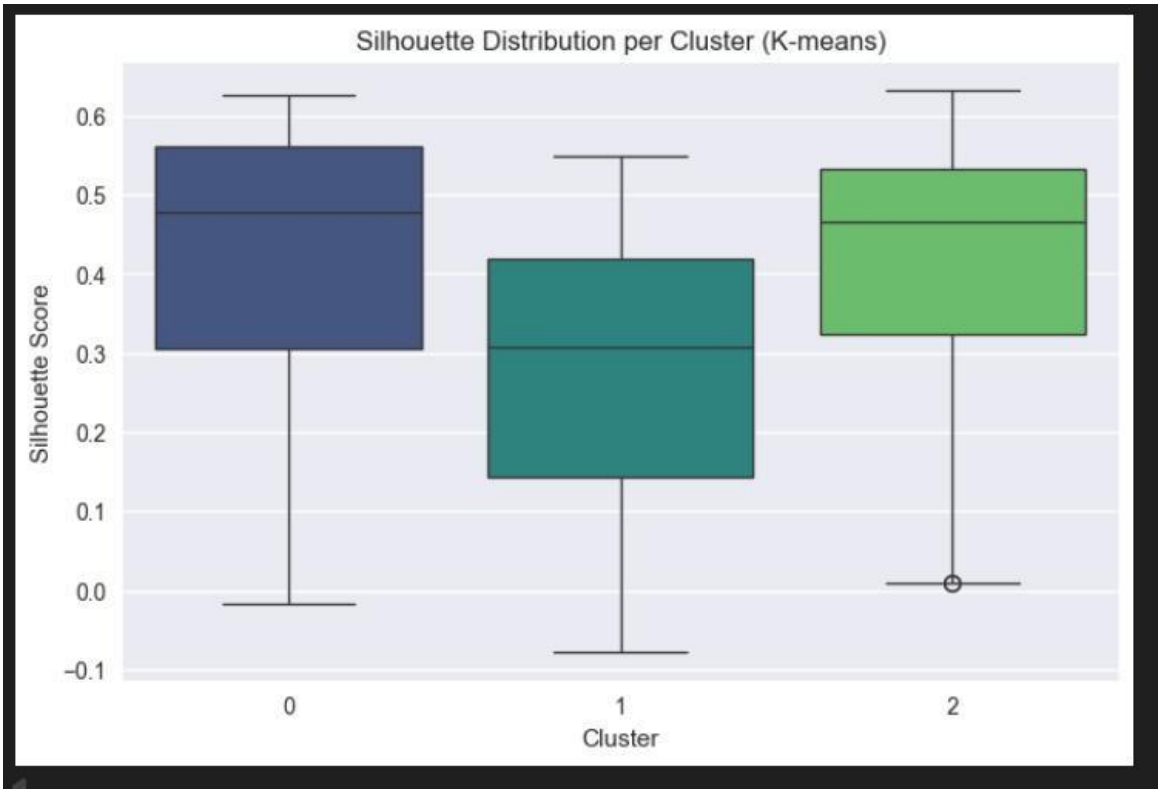




4. K-means Clustering Results with Centroids Visible (ScatterPlot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot).






Silhouette Distribution per Cluster (K-means)
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster