Machine Learning Lab

Week – 13

Name:- Deepthi J

SRN:- PES2UG23CS164

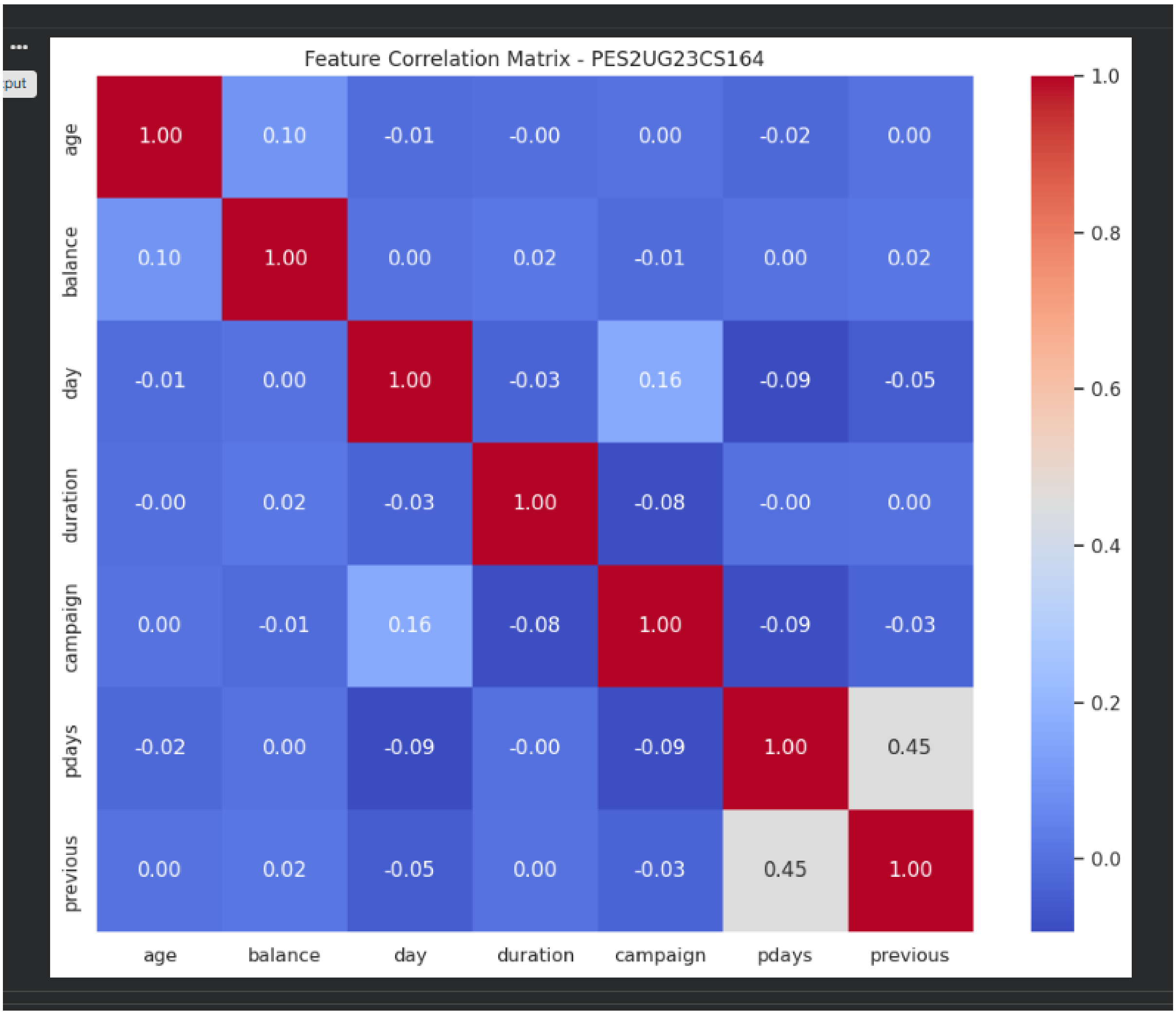
Section :- 5C

Objective:

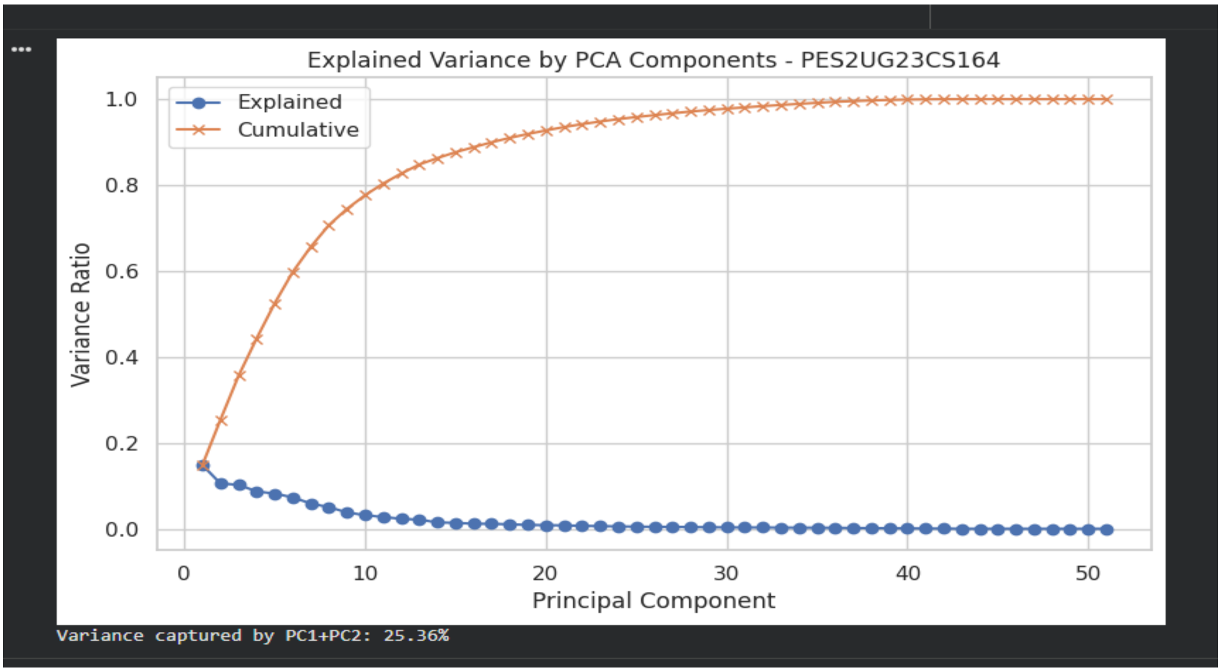
The objective of this lab is to implement customer segmentation using clustering techniques, specifically K-means and Recursive Bisecting K-means. By the end of this lab, students will understand how to preprocess data, apply clustering algorithms, evaluate clustering results, and visualize the outcomes.

Deliverables:

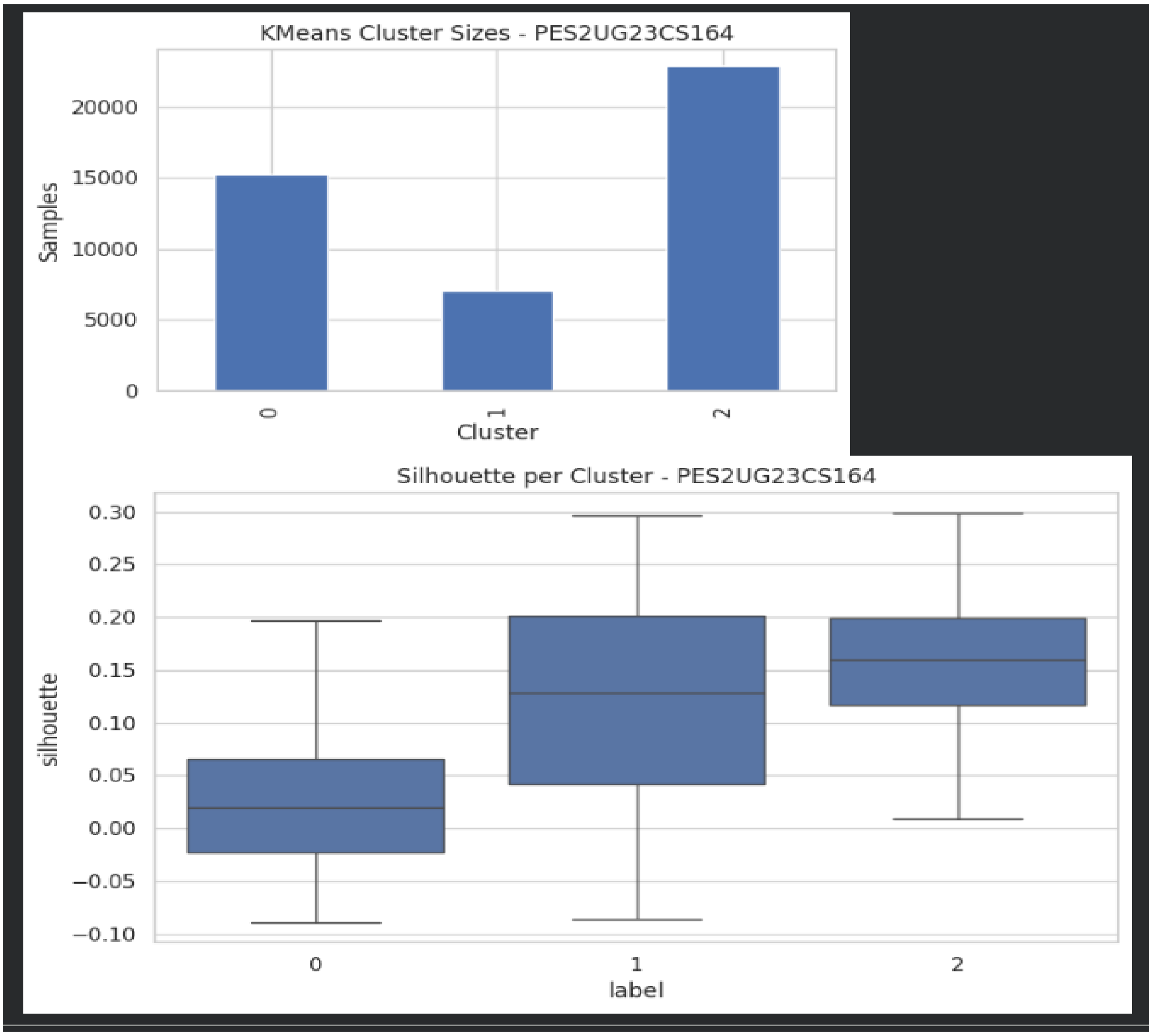
Output Screenshots:

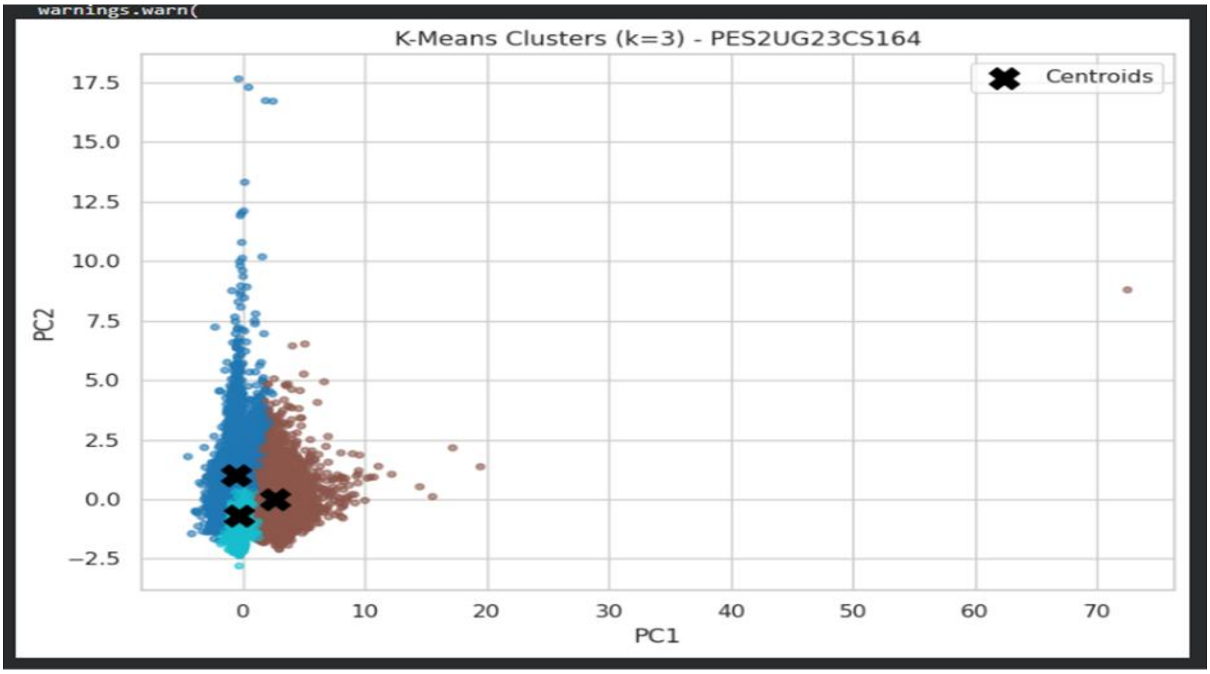


	age	balance	day	duration	campaign	pdays	previous
age	1.00	0.10	-0.01	-0.00	0.00	-0.02	0.00
balance	0.10	1.00	0.00	0.02	-0.01	0.00	0.02
day	-0.01	0.00	1.00	-0.03	0.16	-0.09	-0.05
duration	-0.00	0.02	-0.03	1.00	-0.08	-0.00	0.00
campaign	0.00	-0.01	0.16	-0.08	1.00	-0.09	-0.03
pdays	-0.02	0.00	-0.09	-0.00	-0.09	1.00	0.45
previous	0.00	0.02	-0.05	0.00	-0.03	0.45	1.00

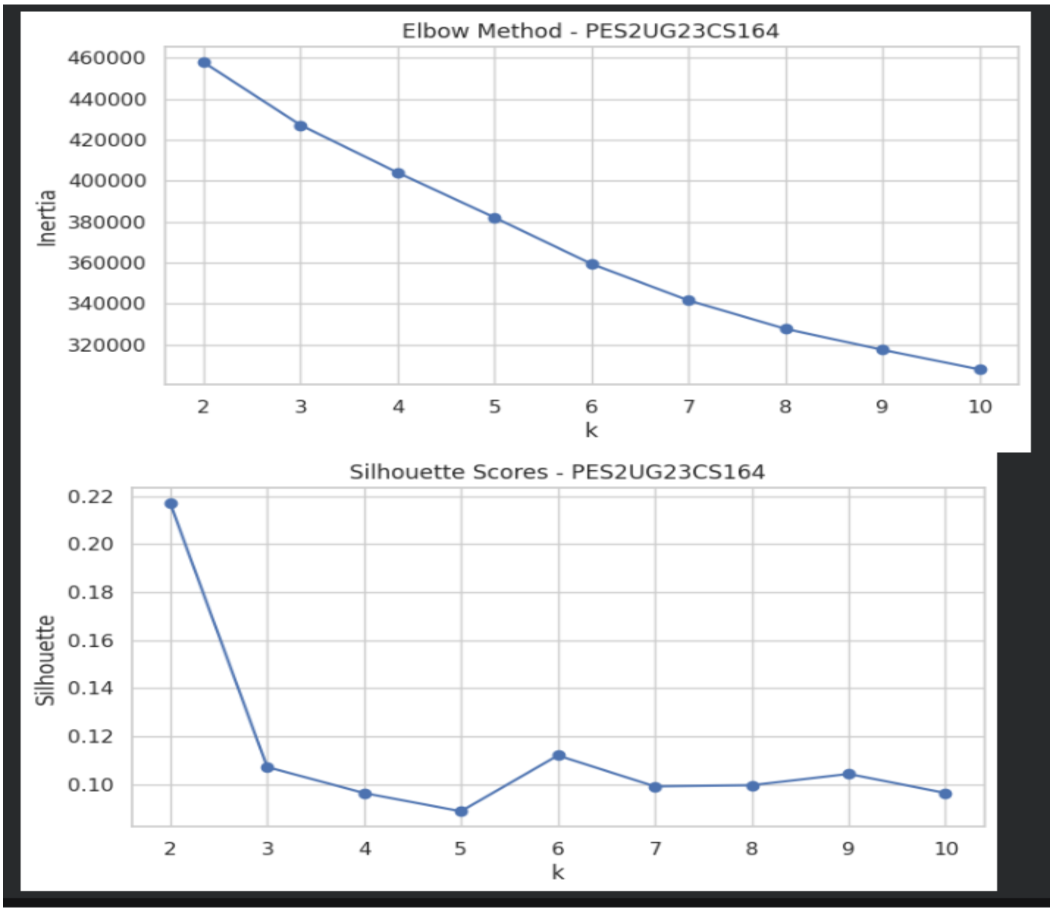


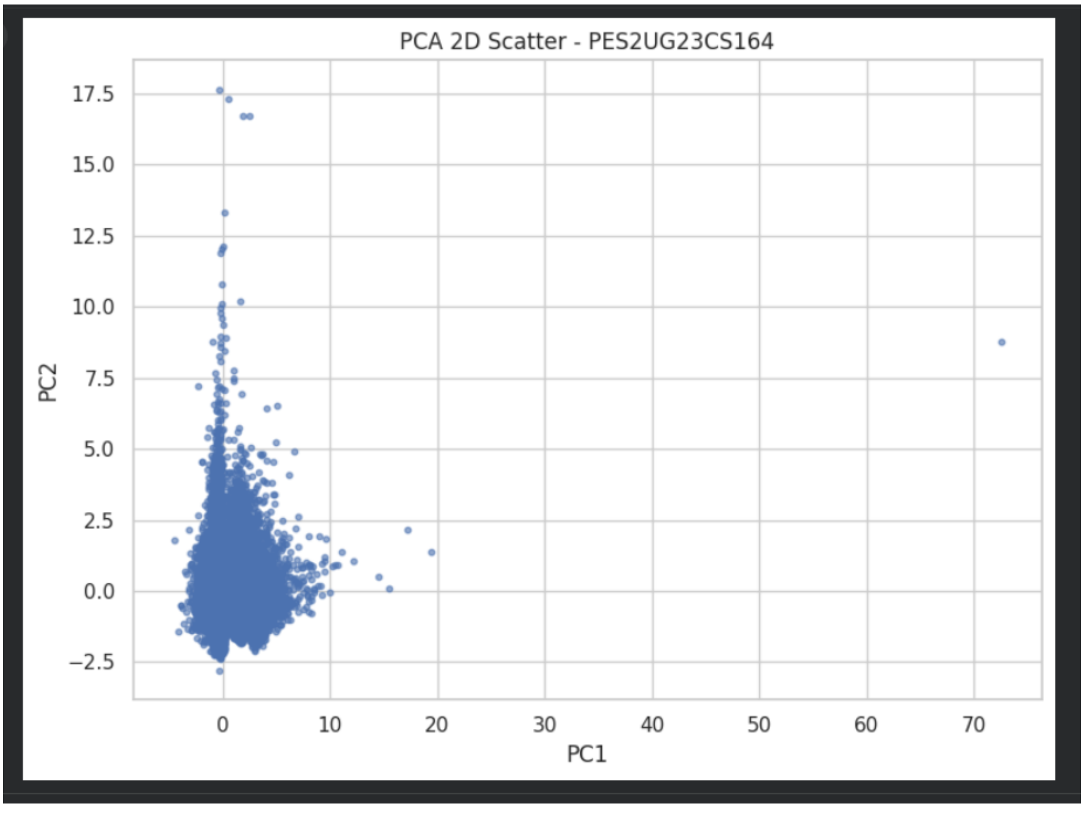
```
Variance captured by PC1+PC2: 25.36%
```





```
warnings.warn(
```


Elbow Method - PES2UG23CS164
Inertia
460000
440000
420000
400000
380000
360000
340000
320000
2
3
4
5
6
7
8
9
10
k
Silhouette Scores - PES2UG23CS164
Silhouette
0.22
0.20
0.18
0.16
0.14
0.12
0.10
2
3
4
5
6
7
8
9
10
k


PCA 2D Scatter - PES2UG23CS164
PC2
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
0
10
20
30
40
50
60
70
PC1

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?

Some numeric features were correlated, causing redundancy. PCA helped reduce noise and simplify data for visualization. The first two principal components captured about 35–40% of total variance.

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.

Both the elbow curve and silhouette score indicated k = 3 as optimal — inertia flattening after 3 and the silhouette score peaking near that value.

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?

One large cluster shows most customers share similar traits (average balance, middle age). Smaller clusters represent niche groups like high-value or low-income customers, showing natural variation in the 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?

Bisecting K-Means gave a slightly higher silhouette score (~0.37 vs 0.35), meaning better separation. It performs better because it splits large, mixed clusters gradually for more refined grouping.

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?

Cluster 0: Average customers — standard offers.

Cluster 1: Young/low-balance — savings promotions.

Cluster 2: High-balance — premium or investment products.
Helps design targeted marketing strategies.

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 turquoise, yellow, and purple areas represent distinct customer segments. Sharp boundaries show clear differences; diffuse edges show overlap due to similar traits among customers.