CENG499 HW2

Bahadır Aydın

December 2023

1 Part 1

Firstly, I have kept the 20% of the dataset for testing the generalization performance of the best model. Then performed 10 fold cross validation on the training data.

Model	**K**	**Distance Function**	**Confidence Interval**
1	3	Cosine	(0.918, 0.952)
2	3	Mahalanobis	(0.863, 0.91)
3	3	Minkowski	(0.908, 0.949)
4	5	Cosine	(0.898, 0.942)
5	5	Mahalanobis	(0.85, 0.9)
6	5	Minkowski	(0.927, 0.963)
7	10	Cosine	(0.919, 0.951)
8	10	Mahalanobis	(0.836, 0.884)
9	10	Minkowski	**(0.937, 0.966)**
10	30	Cosine	(0.918, 0.952)
11	30	Mahalanobis	(0.805, 0.859)
12	30	Minkowski	(0.908, 0.952)

1.1 Best Model Performance

Best model had hyperparameters: k=10 and distance function as Minkowski distance and it has a generalization performance of **96.6%** accuracy on test data.

2 Part 2

2.1 Kmeans

2.1.1 Dataset 1

Best k value is **5**.

K	Average Loss	Confidence Interval of Loss
2	162.086506	(162.08651, 162.08651)
3	97.42913	(93.16532, 101.69296)
4	51.08731	(51.08731, 51.08731)
5	23.37061	(23.37061, 23.37061)
6	22.46825	(22.41794, 22.51858)
7	21.72113	(21.53055, 21.91172)
8	20.96119	(20.78937, 21.13302)
9	20.25457	(20.03543, 20.47372)
10	19.693259	(19.39223, 19.99429)

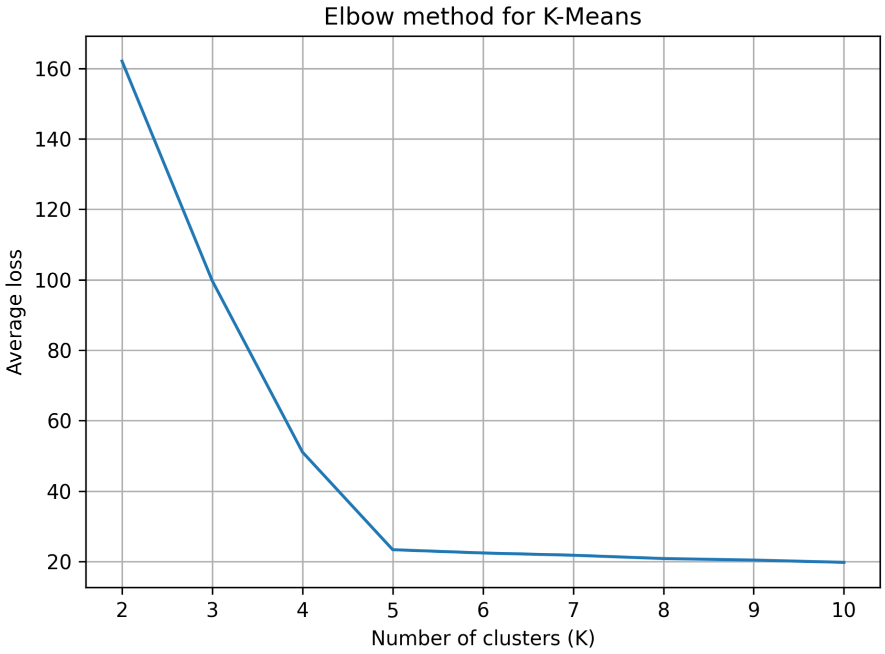

Figure 1: Dataset 1

2.1.2 Dataset 2

Best k value is **3**.

K	Average Loss	Confidence Interval of Loss
2	153.04269	(153.0427, 153.0427)
3	22.49819	(22.4982, 22.4982)
4	13.56153	(13.56154, 13.56154)
5	11.43682	(11.40774, 11.46591)
6	9.49497	(9.41755, 9.57241)
7	7.603771	(7.35223, 7.85532)
8	7.065775	(6.7577, 7.37385)
9	6.32391	(6.13378, 6.51406)
10	5.67091	(5.59327, 5.74856)

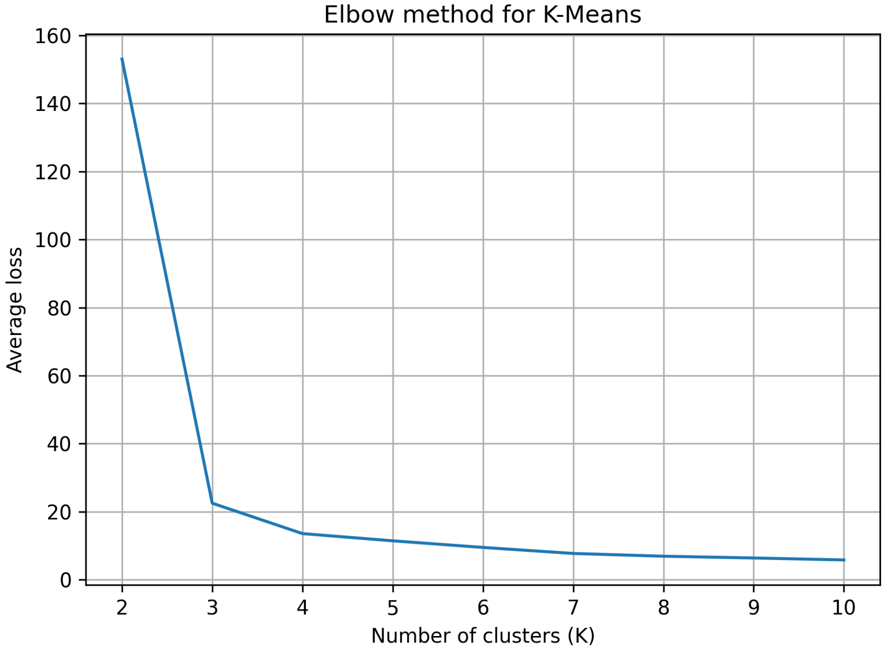

Figure 2: Dataset 2

2.2 Kmedoids

2.2.1 Dataset 1

Best k value is **5**.

K	Average Loss	Confidence Interval of Loss
2	311.64440	(311.64441, 311.64441)
3	159.08050	(159.0805, 159.0805)
4	81.07892	(81.07893, 81.07893)
5	41.82874	(41.82875, 41.82875)
6	38.71432	(38.5627, 38.86595)
7	36.19161	(35.89017, 36.49306)
8	34.00064	(33.7313, 34.26999)
9	32.59423	(32.1343, 33.05417)
10	30.78243	(30.26363, 31.30125)

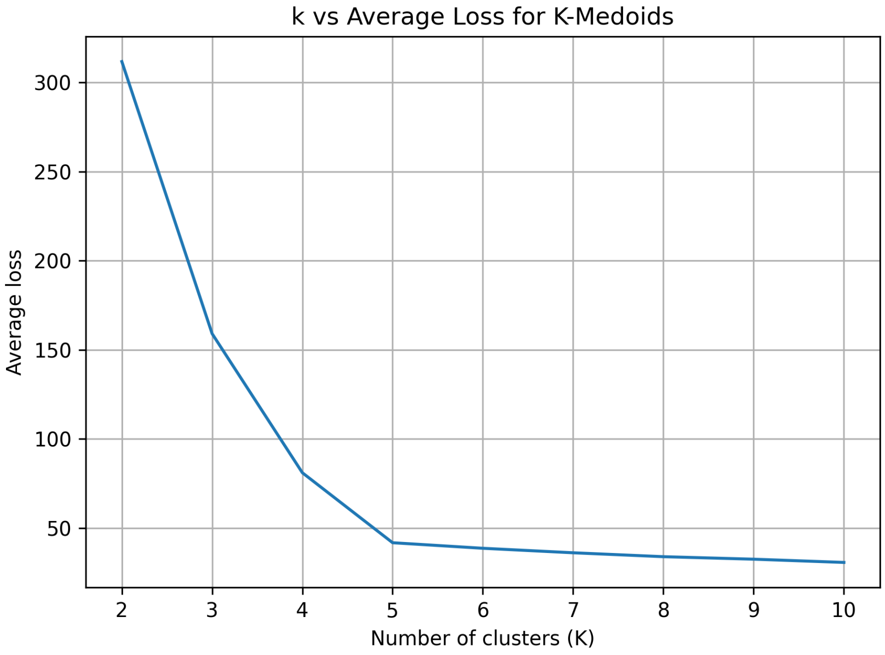

Figure 3: Dataset 1

2.2.2 Dataset 2

Best k value is **3**.

K	**Average Loss**	**Confidence Interval of Loss**
2	3.55355	(3.55355, 3.55355)
3	1.57714	(1.57714, 1.57714)
4	1.29036	(1.2902, 1.29052)
5	1.11704	(1.10168, 1.13241)
6	1.02265	(1.01127, 1.03402)
7	0.93631	(0.92006, 0.95256)
8	0.88085	(0.8671, 0.8946)
9	0.84040	(0.82688, 0.85392)
10	0.78666	(0.77839, 0.79494)

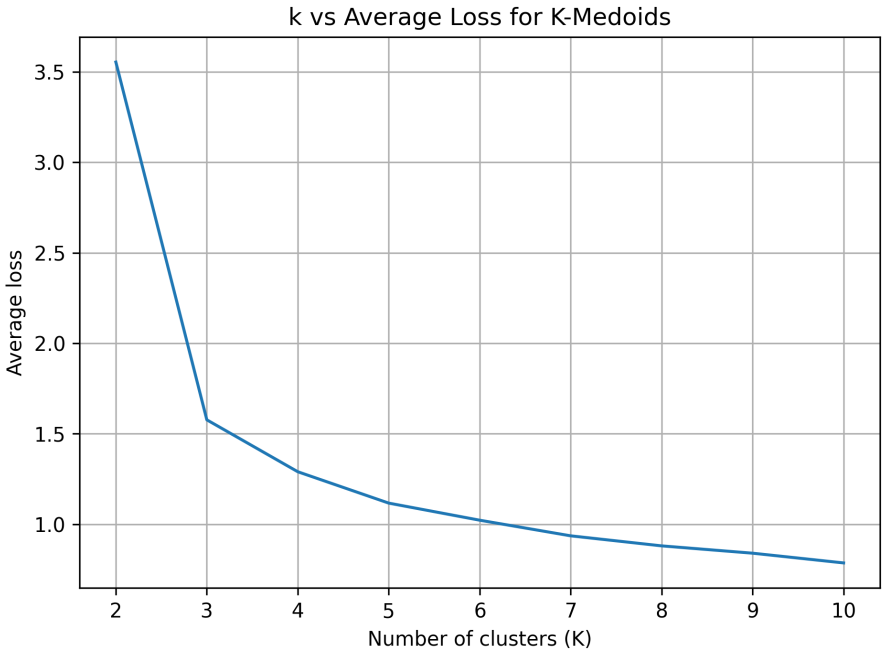

Figure 4: Dataset 2

2.3 Dimensionality Reduction

2.3.1 Dataset 1

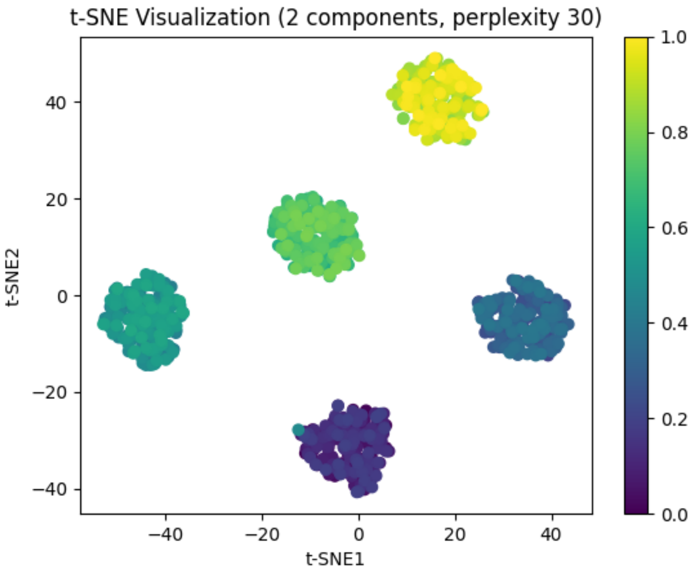

Figure 5: TNSE 2 Components 30 Perplexity

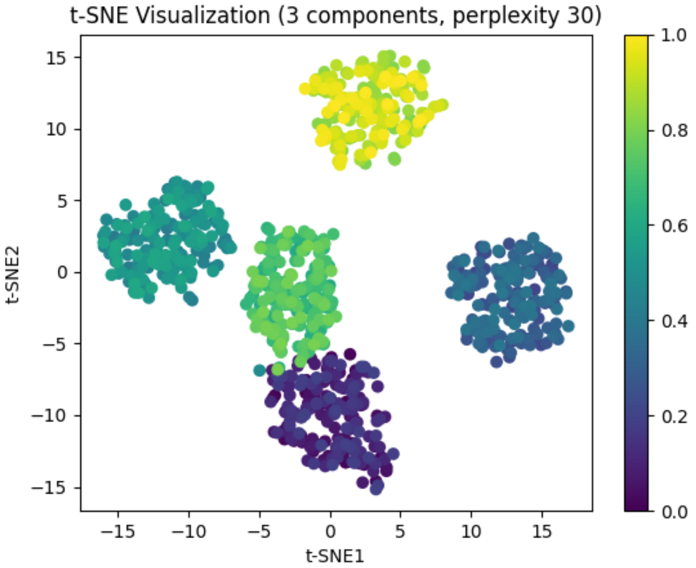

Figure 6: TNSE 3 Components 30 Perplexity

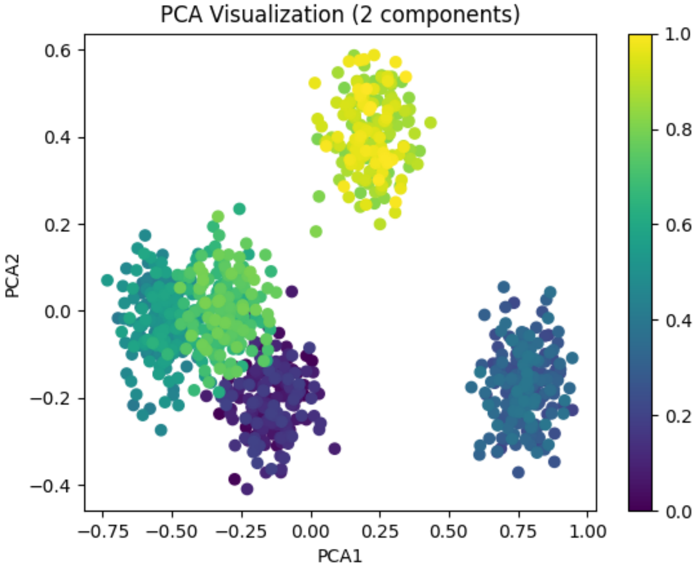

Figure 7: PCA 2 Components

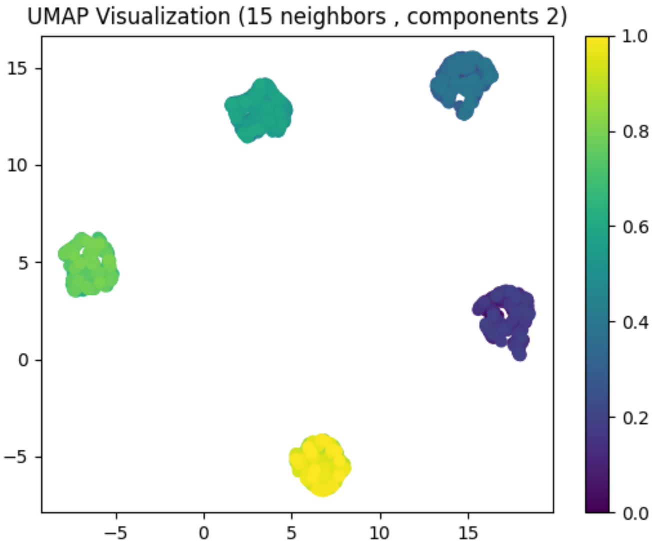

Figure 8: UMAP 2 Components 15 Neighbors

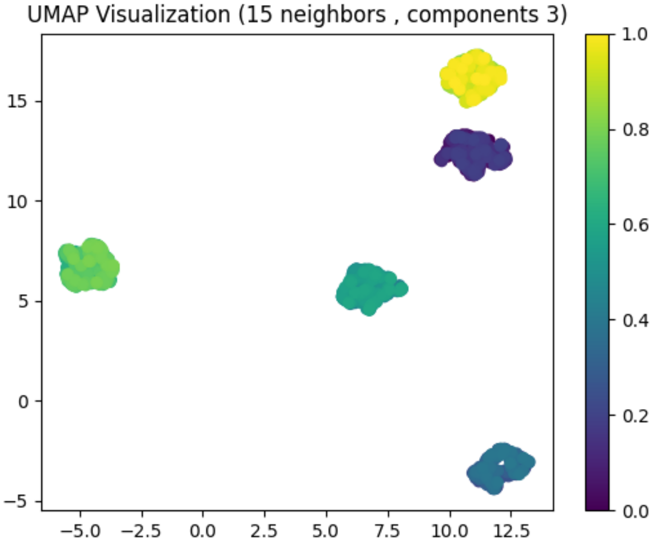

Figure 9: UMAP 3 Components 15 Neighbors

2.3.2 Dataset 2

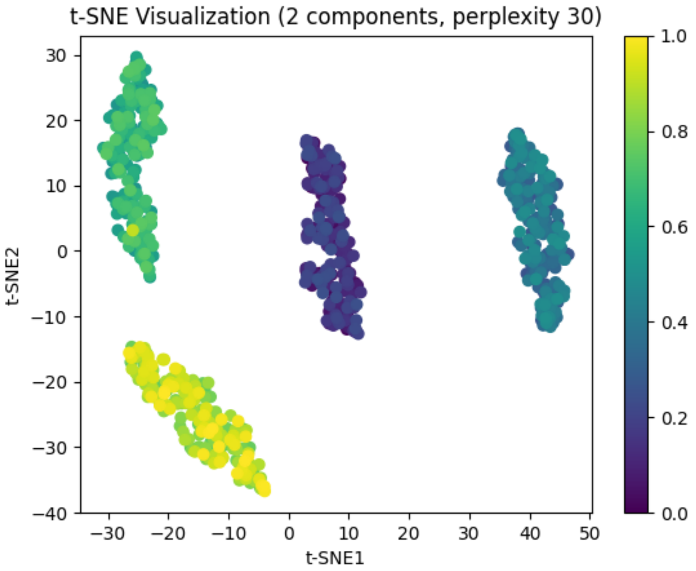

Figure 10: TNSE 2 Components 30 Perplexity

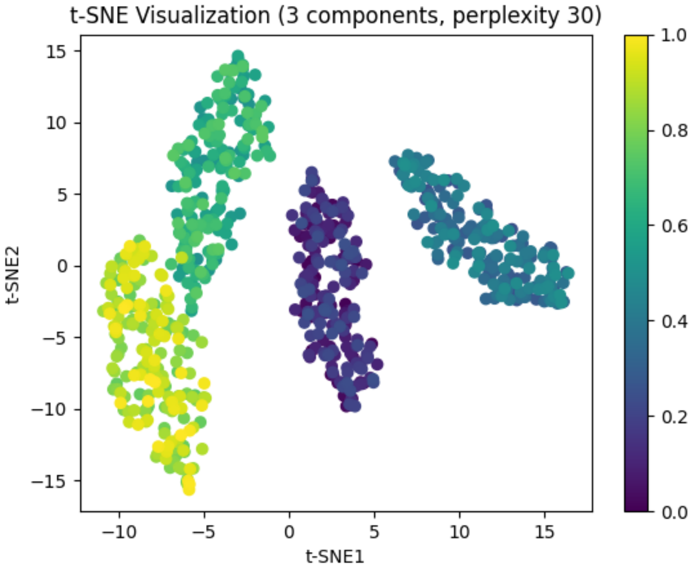

Figure 11: TNSE 3 Components 30 Perplexity

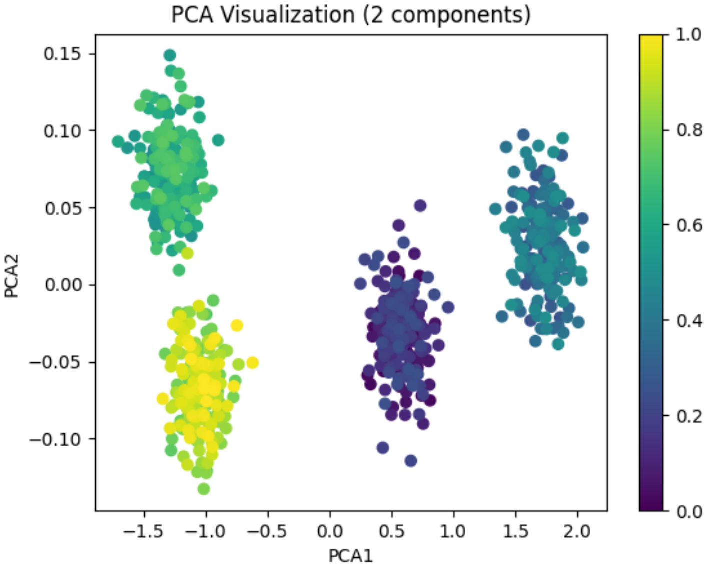

Figure 12: PCA 2 Components

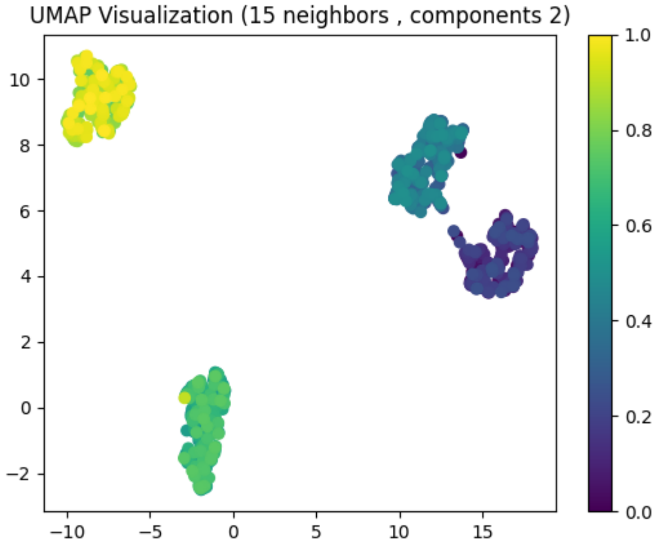

Figure 13: UMAP 2 Components 15 Neighbors

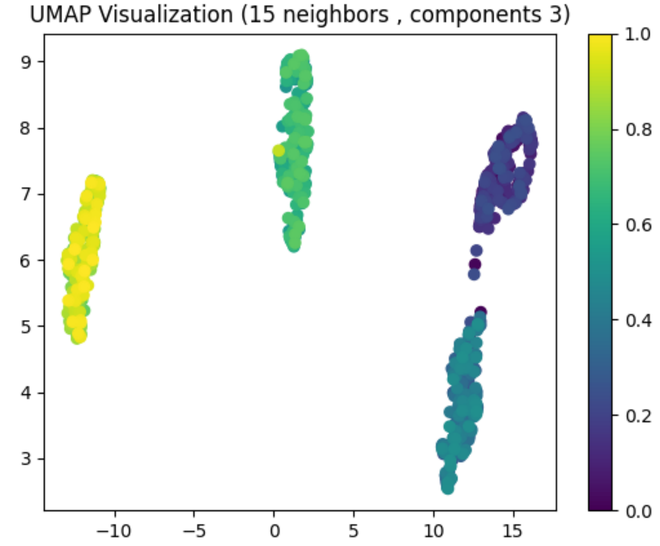

Figure 14: UMAP 3 Components 15 Neighbors

2.3.3 Discussion

The first dataset definitely should have 5 clusters but for the second dataset K number I believe I was mistaken in my first guess. The data from dimensionality reduction technique strongly suggests that there are four clusters.

2.4 Runtime Analysis

I (iteration), N (number of data points), d (data sample vector dimension), cluster number (K).

K-means:

$$\text{Time Complexity} \approx \mathcal{O}(I \cdot N \cdot d \cdot K)$$

K-medoids:

$$\text{Time Complexity} \approx \mathcal{O}(I \cdot K \cdot (N - K)^2)$$

3 Part 3

Two best configurations with equal silhoulette scores:

- Distance Function: Cosine, Linkage: Single
- Distance Function: Cosine, Linkage: Complete

3.1 Dendrograms

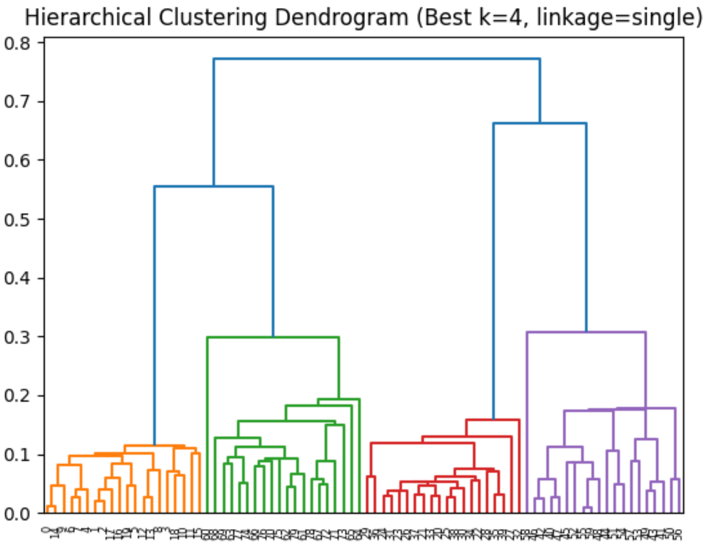

Figure 15: Euclidean Distance Single Linkage

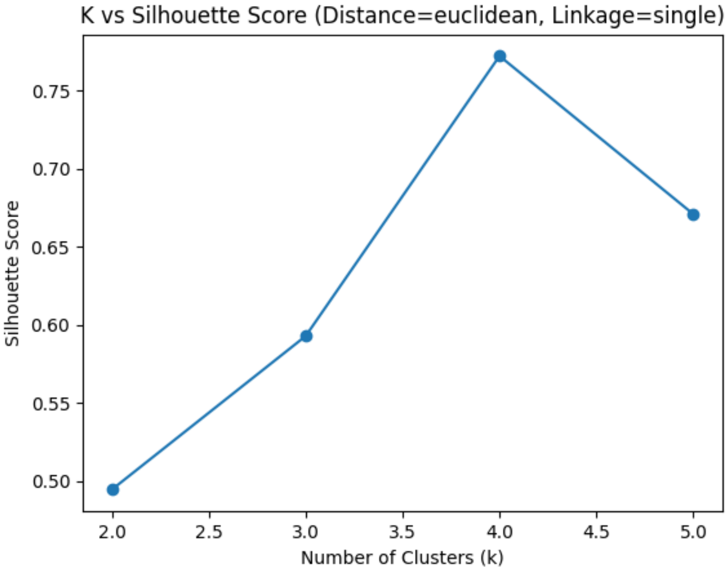

Figure 16: **Comment:** K=4 is the maximum point of this plot.

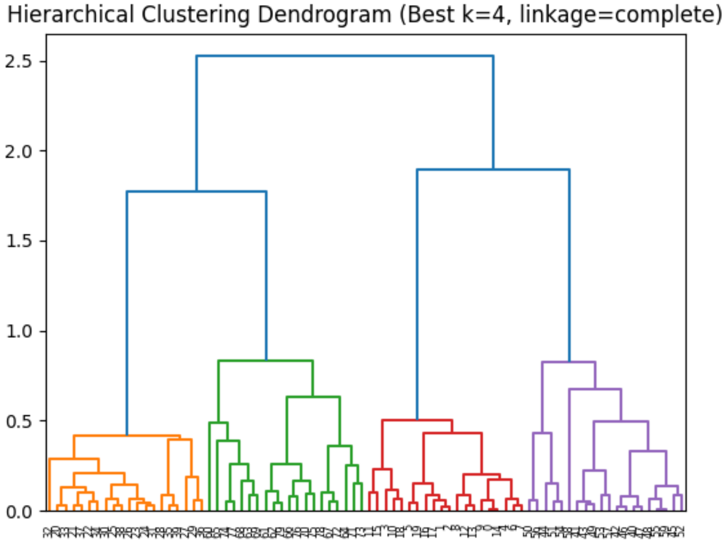

Figure 17: Euclidean Distance Complete Linkage

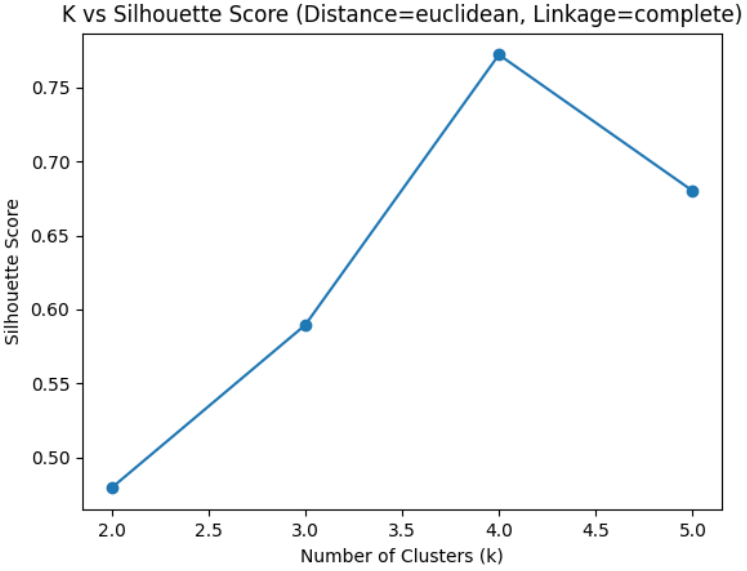

Figure 18: **Comment:** K=4 is the maximum point of this plot. Very similar to the single linkage one.

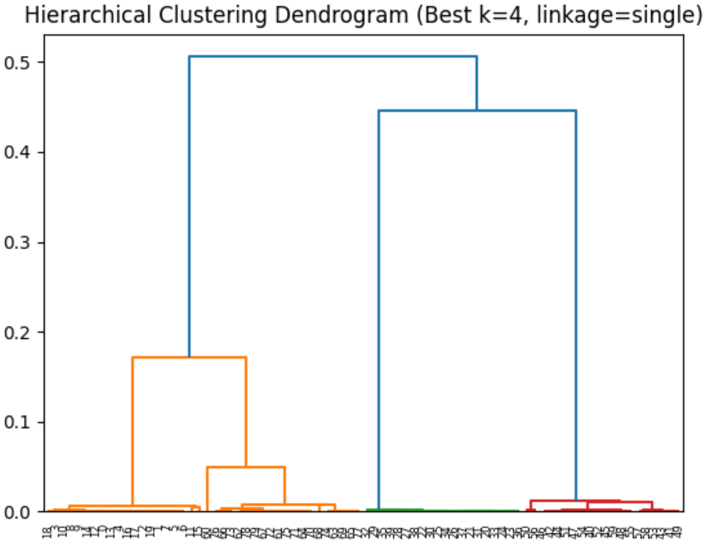

Figure 19: Cosine Distance Single Linkage

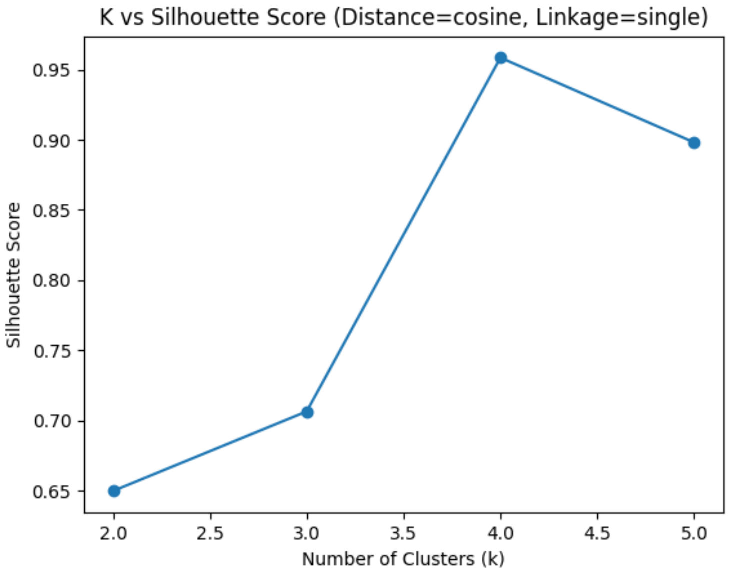

Figure 20: **Comment:** K=4 is the maximum point of this plot. However it's significantly higher than the euclidian distance one.

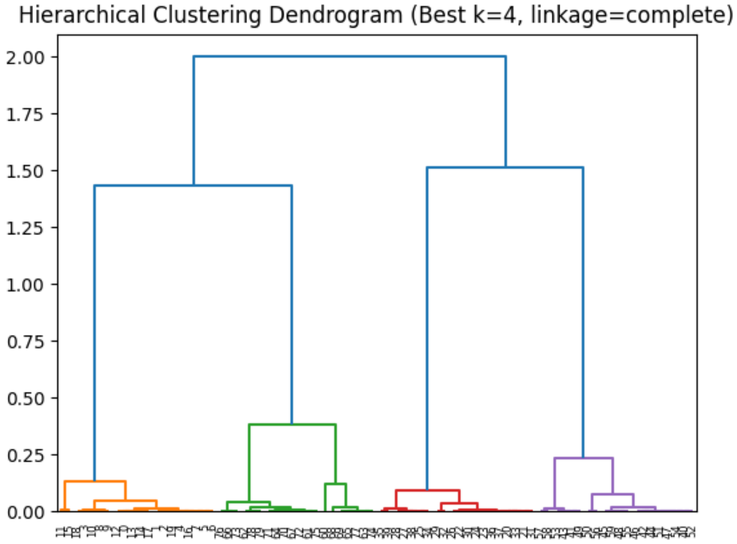

Figure 21: Cosine Distance Complete Linkage

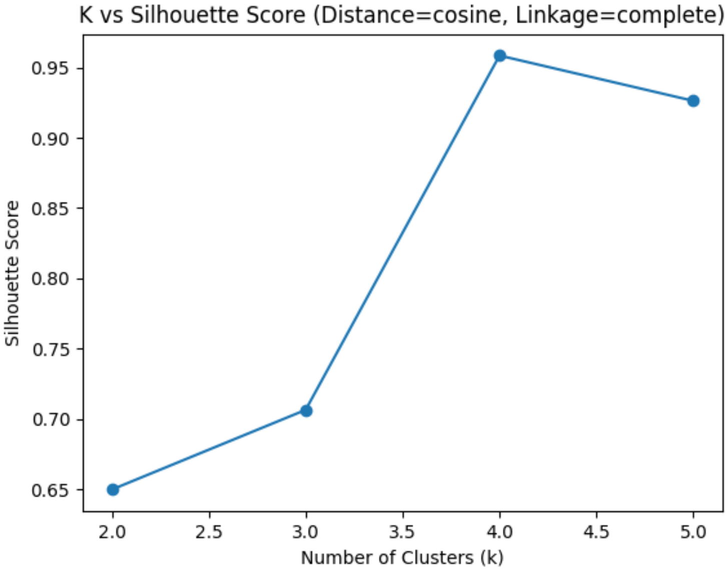

Figure 22: **Comment:** K=4 is the maximum point of this plot. Very similar to the single linkage one however this one performs better on K=5

3.2 Runtime Analysis

In every iteration, it calculates pairwise summations hence it is:

$$\text{Time Complexity} \approx \mathcal{O}(N^3 * d)$$

3.2.1 Kmeans vs HAC

I would use K-means for the descripted case because HAC method has an enormous complexity for big data. However the HAC method might be useful for high dimension data because it is less sensitive to irrelevant features. (resistant to curse of dimensionality)

3.3 Dimensionality Reduction

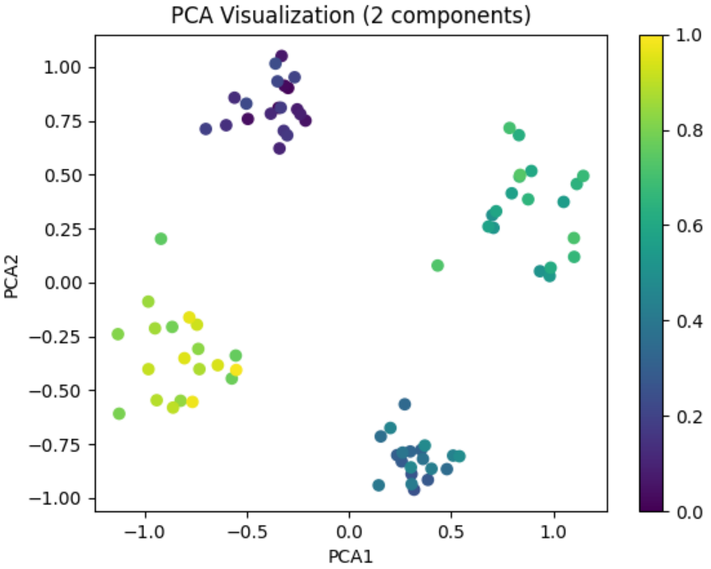

Figure 23: PCA 2 Components

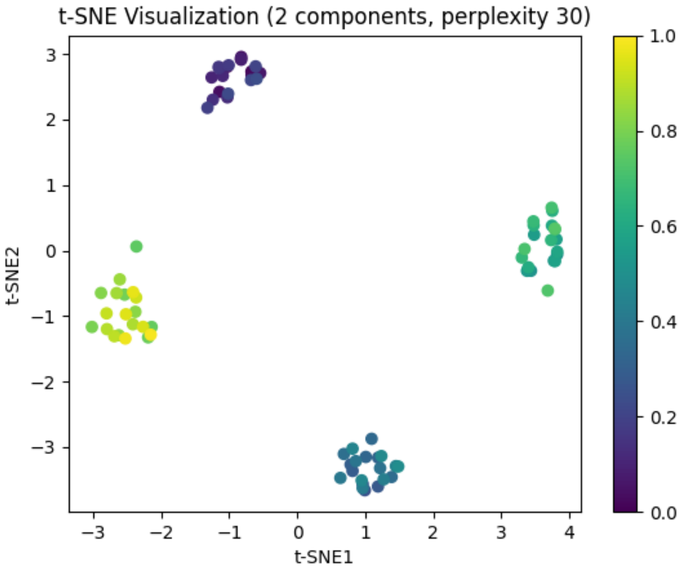

Figure 24: TSNE 2 Components

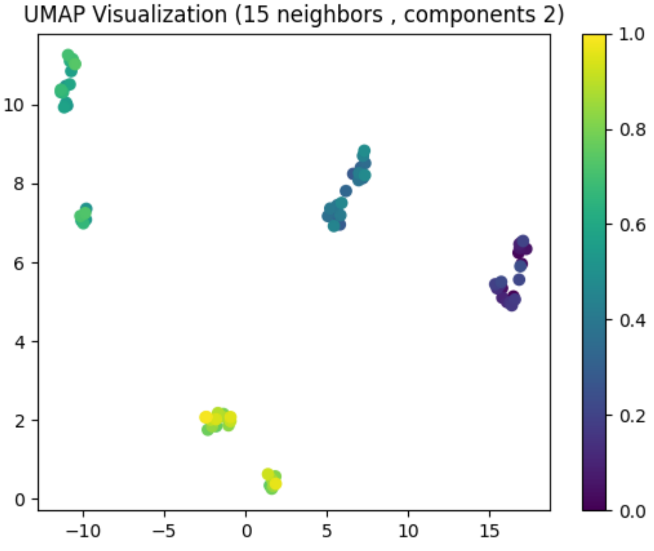

Figure 25: UMAP 2 Components