HW4(A) - Iris

Introduction
We ran the K-means algorithm on the Iris dataset for K values ranging from 2 to 20. The silhouette score was used to evaluate the performance of each model. The silhouette score ranges from -1 to 1, with values closer to 1 indicating better clustering.

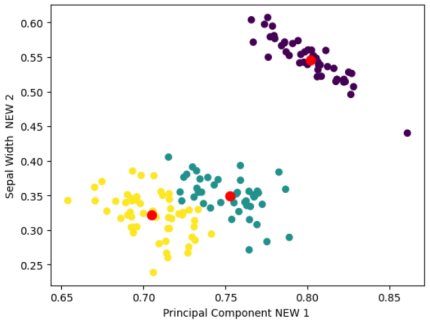
Pseudo Code (KMeansImpl) **:**

We initialize the centroids by selecting random data points from our dataset.

Main(X_Data,no_of_clusters):
Run For loop for400 .

1. Calculates the Euclidean distance between the data points in X_Data and the centroids in Centroid and stores the result in distValue.
2. Classify the data into clusters based on initial centroids.
3. Recompute the centroids to the mean of the points assigned to that cluster, and return True if the new centroids match the current centroids. If centroids return true then break as we have found convergence.
4. We calculate the inertia by calculating the sum of squared distances from X_Data to Centroids.

Our Clustering classifies as follows:

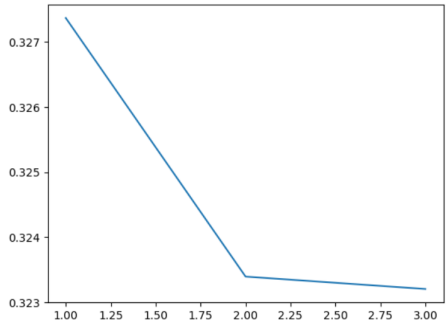


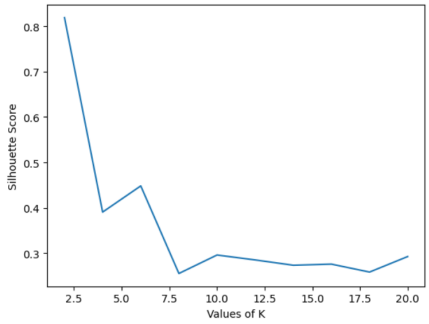
The results of our experiment are summarized in the following table:

K Value	Silhouette Score (Normalized)	Silhouette Score (UnNormalized - PCA)
2	0.8188	0.7055
4	0.3907	0.5589

6	0.4483	0.4496
8	0.2555	0.5241
10	0.2962	0.4183
12	0.2854	0.3761
14	0.2737	0.3898
16	0.2761	0.3606
18	0.2586	0.3422
20	0.2928	0.3624

The normalized Silhouette Scores show that K=2 has the highest score of 0.8188, indicating that the data can be well-clustered into two groups. As K increases, the Silhouette Score decreases, suggesting that the clusters become less well-defined.
The unnormalized Silhouette Scores with PCA show a slightly different pattern, with K=4 having the highest score of 0.5589. However, the general trend of decreasing Silhouette Scores as K increases still holds.
Below is the Inertia and Silhouette plot for the 3 clusters:





Based on our metrics our clustering seems to be moderate to good quality as the elbow curve points out the optimal number of K's to be chosen.

Dimensionality Reduction:
I have chosen normalization because it reduces the impact of differences in the scale of variables, while PCA is a technique for reducing the dimensionality of data by identifying the underlying structure of the data.

Distance Measure(Consistent Across Both Datasets):
Since we are reducing the dimensions, the magnitude of the distance between data points becomes more crucial, and the high dimensionality and different scaling of the features are eliminated. Therefore, we have used Euclidean distance as our distance measure.

V measure (miner- deeppp15, rank- 45, score- 0.95):

A V Measure of 0.95, indicates that the clustering solution has high homogeneity and completeness. For a clustering solution to have high intra-cluster similarity and low inter-cluster similarity (as reflected in the Silhouette Score), but still have high homogeneity and completeness (as reflected in the V Measure) if the clusters are well-separated and highly distinct.

HW4(B) - Image

Dimensionality Reduction:

- On our data set I have used Gaussian Blur function from OpenCV as it smoothes out the image and reduces the noise input in our Kmeans.
- On our Dataset t-SNE is more suitable because our data is complex with a non-linear structure, where the relationships between data points are not well-represented by linear transformations. Whereas, PCA is more suited for linear transformations.

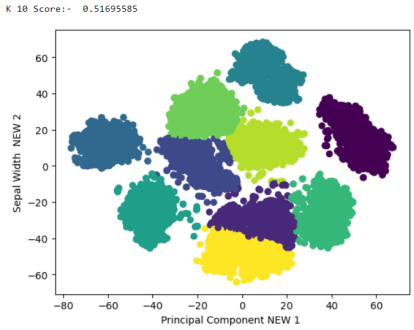
V measure (miner- deeppp15 , rank- 127 ,score- 0.83):
A V Measure of 0.83 suggests that the clustering solution has moderate homogeneity and completeness. Overall, the high Silhouette Score indicates that the clustering solution is worth further exploration and refinement to achieve a higher V Measure.

Results:

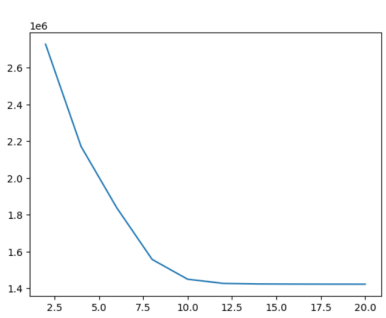
K Value	Silhouette Score (Normalized)	Silhouette Score (UnNormalized - PCA)
2	0.3471	0.3667
4	0.4359	0.3610
6	0.4698	0.3420
8	0.4967	0.3337
10	0.5168	0.3405
12	0.4840	0.3401
14	0.4421	0.3308
16	0.4834	0.3301
18	0.4037	0.3285

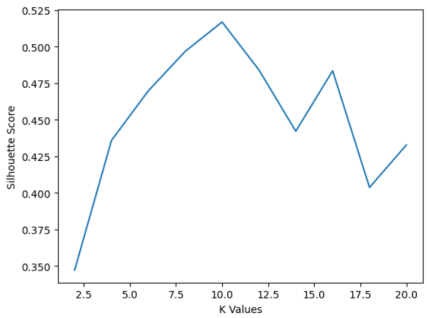
20	0.4328	0.3362

Our Clustering classifies as follows:



Below is the Inertia and Silhouette plot for the 3 clusters:





From the graph, we can observe that the Silhouette Score initially increases as K value increases and then starts to decrease. This is a common pattern in clustering, where an optimal number of clusters exists that maximizes the clustering quality. In this case, the optimal number of clusters seems to be around 10, as it has the highest Silhouette Score for both normalized and unnormalized PCA data. The Elbow curve is also meeting at 10 which signifies optimal clustering number.

Conclusion

In this report, we applied the K-means clustering algorithm to the Iris dataset and evaluated its performance using the silhouette score. We found that the best K value for the iris dataset is 2, which achieved a silhouette score of 0.8188 with a v- measure of 0.95 and the best K value for the image dataset is 10, which achieved a silhouette score of 0.5168 with a v- measure of 0.83.

Hence, by having a silhouette score of 0.51 we can say that our model is an average performer.