IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – VII
Clustering

Student's Name: Vikas Dangi **Mobile No: 9406661661**

Roll Number: B20238 **Branch: EE**

1

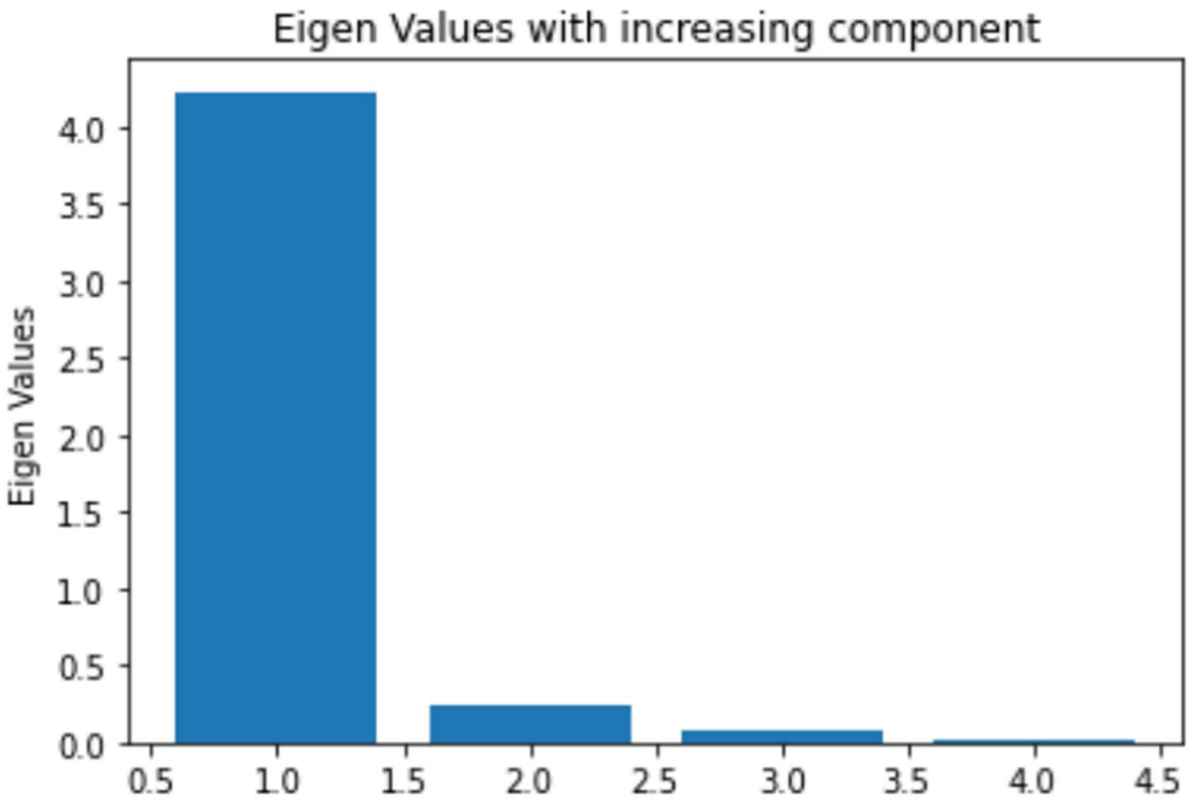

Figure 1 Eigenvalue vs. components

Inferences:

1. With each component increase, the eigenvalue decreases. Until component = 2, there is a significant drop. Following that, the reduction is less.
2. This is to be expected, given eigenvalues are a measure of variance along the principal components, and PCA is done in such a way that data variance diminishes as principal components are added.

2 a.

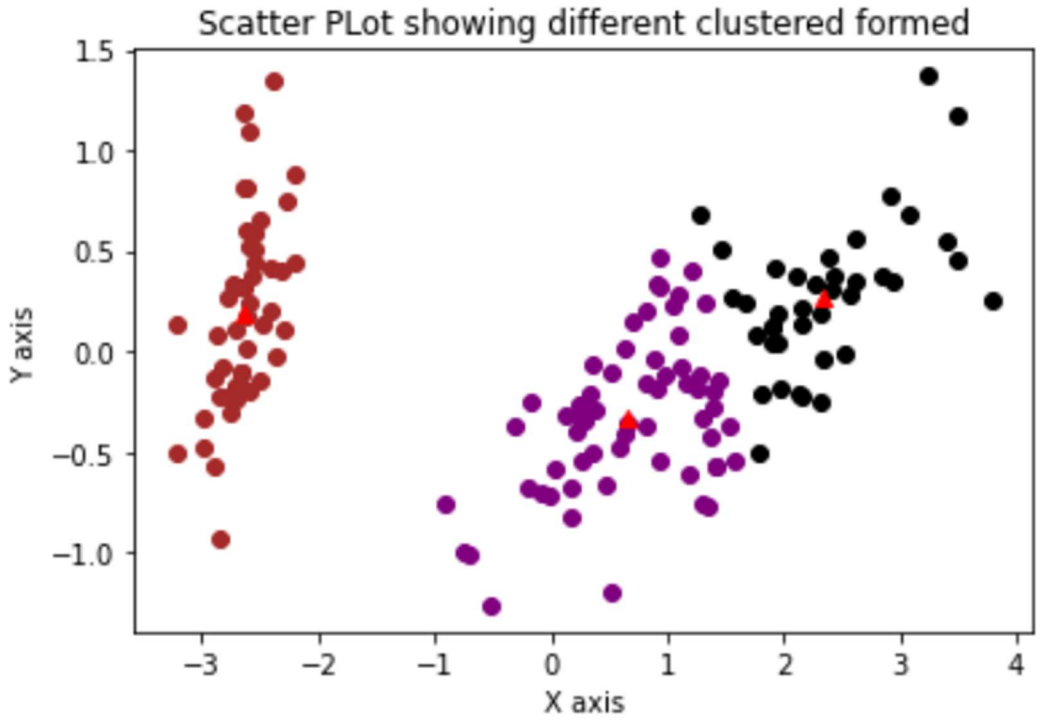

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The clusters appear to be well defined and circular in shape, with significant densities of datapoints in each cluster centre.
2. Yes, the boundaries seem to be circular.

b. The value for distortion measure is 63.874

c. The purity score after examples are assigned to the clusters is 88.7%

3

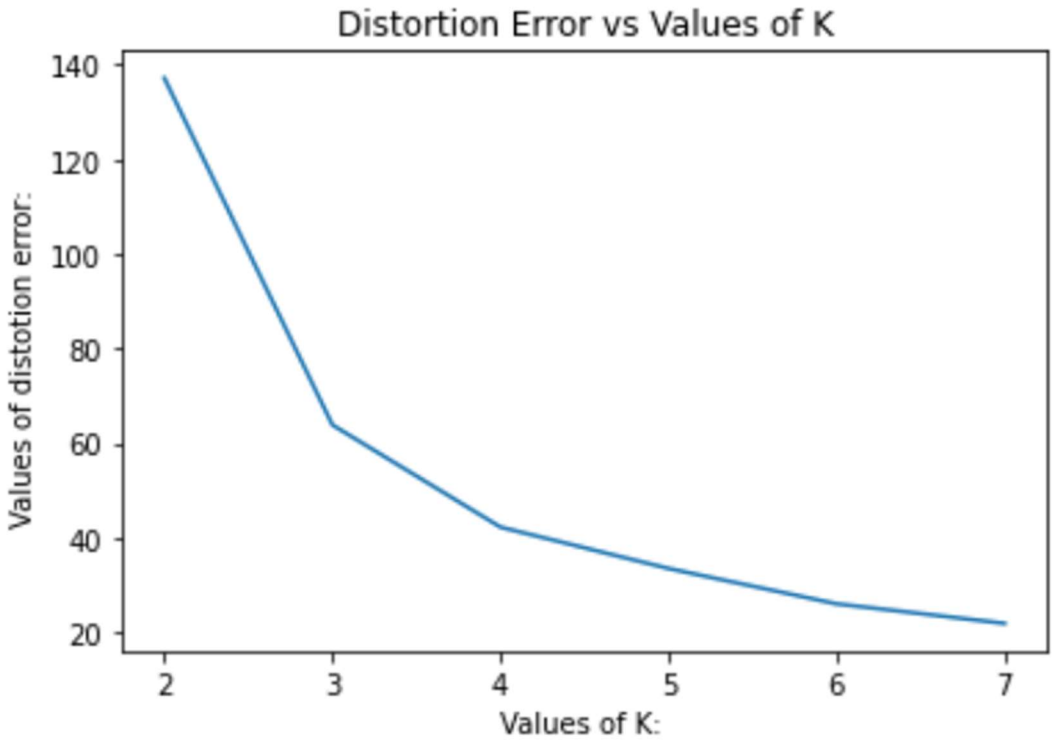

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with an increase in K.
2. More centers are developed when more clusters are formed. Because the distortion measure is the total of the distance between cluster points and their centers, it decreases as K increases.
3. We can see from the graph that K=3 is the best value for best results when using the elbow method. And the outcome is exactly what we expected, with K=3 having the greatest purity score of all the Ks.

Table 1 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	66.667
3	88.667
4	68.667
5	67.333
6	50.667
7	50.667

Inferences:

1. The highest purity score is obtained with K =3
2. Purity initially increases as the k value rises. Then it starts to go downhill. At the ideal k value, purity is at its greatest.
3. This is because as we get closer to the actual number of clusters in the dataset ie.3, more datapoints are correctly classified, and purity score rises; however, once we reach the maximum value at K=3, we start making more clusters than are actually present in the dataset, and thus begin incorrectly classifying datapoints, lowering our purity score.
4. They have a similar tendency once the purity score's k value reaches its maximum k; both begin to decline after that. In addition, the elbow point of distortion measurement yields the k value, which corresponds to the highest purity score.

4 a.

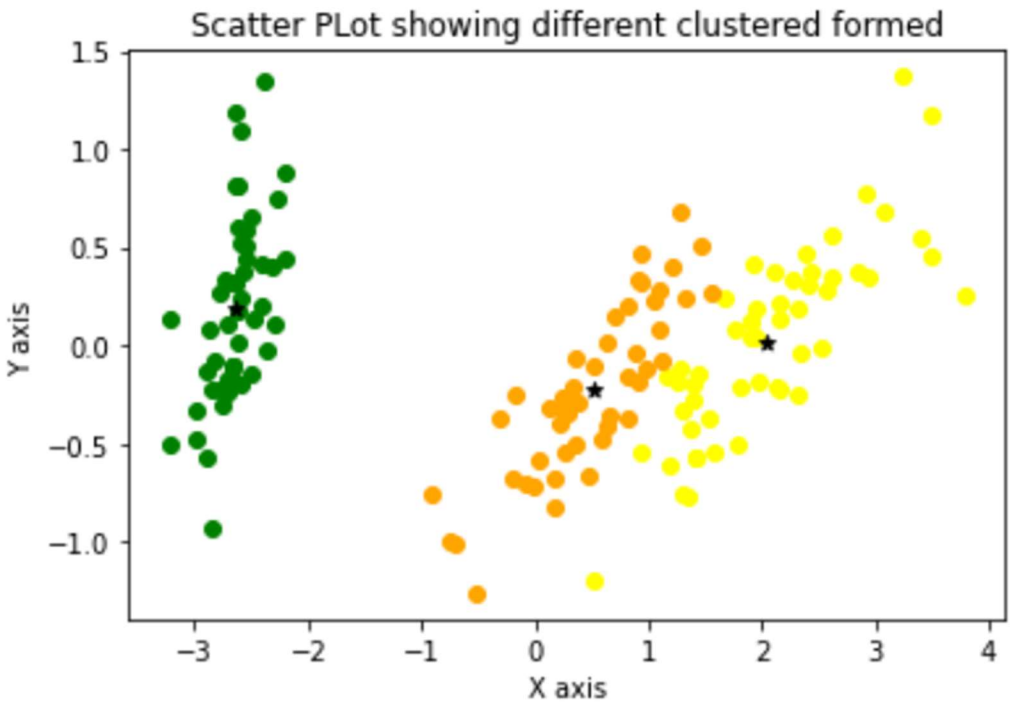

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. The clusters formed in the above plot are elliptical in shape.
2. GMM algorithm assumes cluster boundaries to be elliptical in 2D. From the output, the boundary seems to be elliptical.
3. The GMM has elliptical cluster boundaries while the K means has circular.

b. The value for distortion measure is -280.87

c. The purity score after examples are assigned to the clusters is 0.98

5

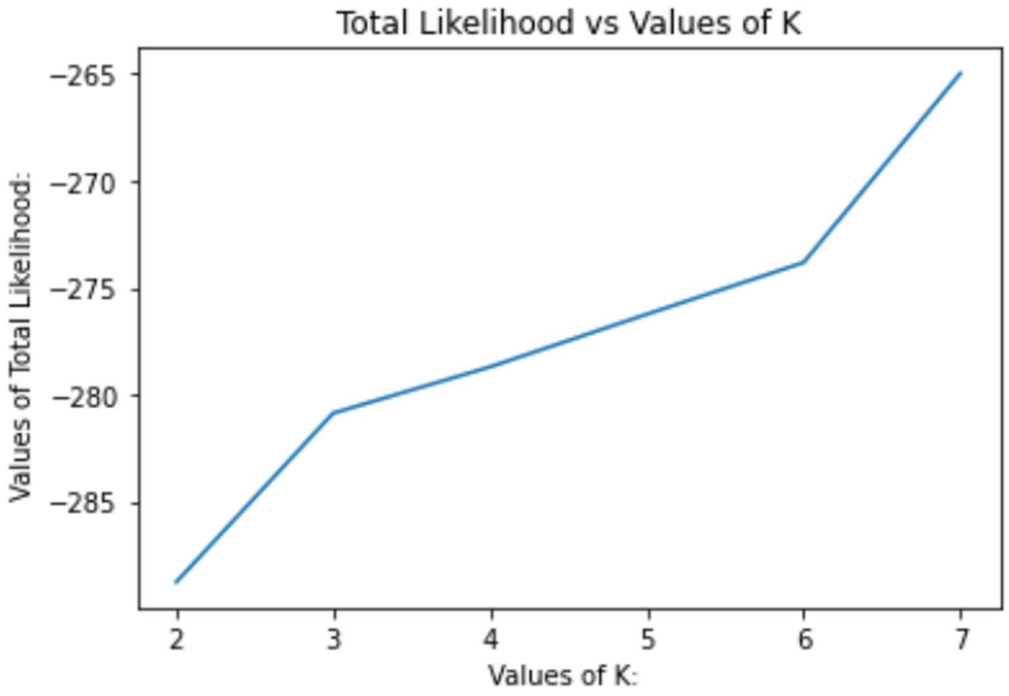

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure increase with an increase in K. Justify the observed trend.

2. As the number of species in the given dataset is, we can intuitively say that optimum clusters is
3. Yes, the elbow method allows us to follow intuition.

Table 2 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	66.67
3	98
4	83.33
5	76.67
6	64.0
7	62.27

Inferences:

1. The highest purity score is obtained with K =3
2. Purity initially increases as the k value rises. Then it starts to go downhill. At the ideal k value, purity is at its greatest.
3. This is because, as we get closer to the actual number of clusters in the dataset , a greater number of datapoints are correctly classified, and the purity score rises; however, once we reach the maximum value at K=3, we begin to make more clusters than are actually present in the dataset, and thus begin incorrectly classifying datapoints, lowering our purity score.
4. They have a similar tendency once the purity score's k value reaches its maximum k; both begin to decline after that. In addition, the elbow point of distortion measurement yields the k value, which corresponds to the highest purity score.
5. The GMM model appears to better suit the data than the K-Means model. This is to be expected, given that GMM is a probabilistic soft classifier.

6

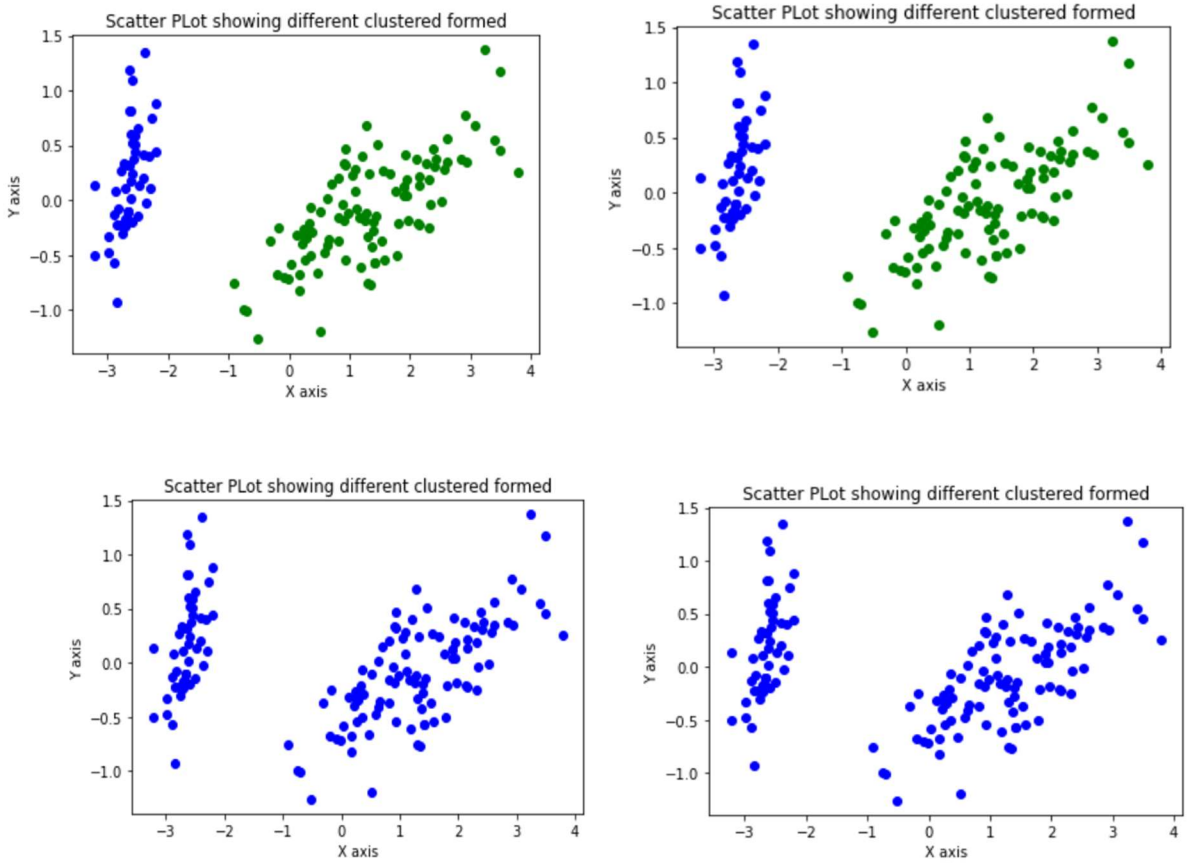

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. We are not getting a good accuracy the reason maybe that we are not taking an appropriate value of radius and Min_sample.
2. The DBCSAN is not able to separate the two clusters in the right as they don't seem to have a common point where their frequency differ. GMM has elliptical boundary, K means has circular and DBSCAN's boundary can take any shape.

b.

Eps	Min_samples	Purity Score
1	4	66.67
	10	66.67
4	4	33.33
	10	33.33

Inferences:

1. Min_samples doesn't affect purity scores value.
2. For the same min_samples, increasing eps value decreased the purity score

THANK YOU