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

Student's Name: Bharat Kumar **Mobile No: 6299862503**

Roll Number: B20090 **Branch:** CSE

1

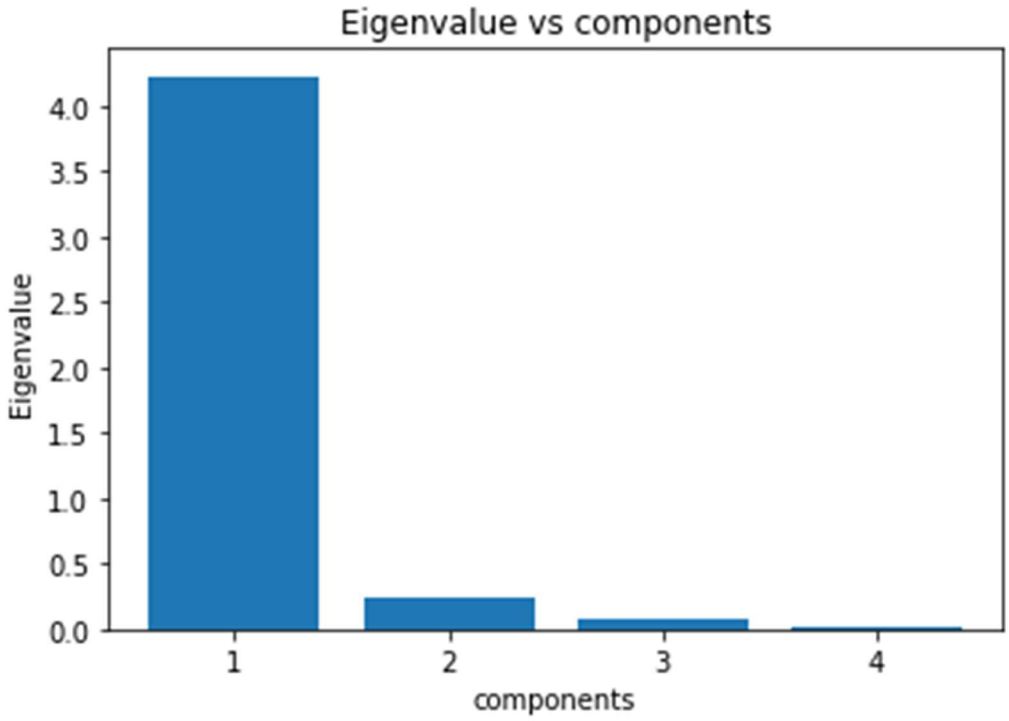

Figure 1 Eigenvalue vs. components

Inferences:

1. Eigenvalues decreases as the number of components increases.
2. Since, they are covering less variances. So eigen value are decreasing.

2

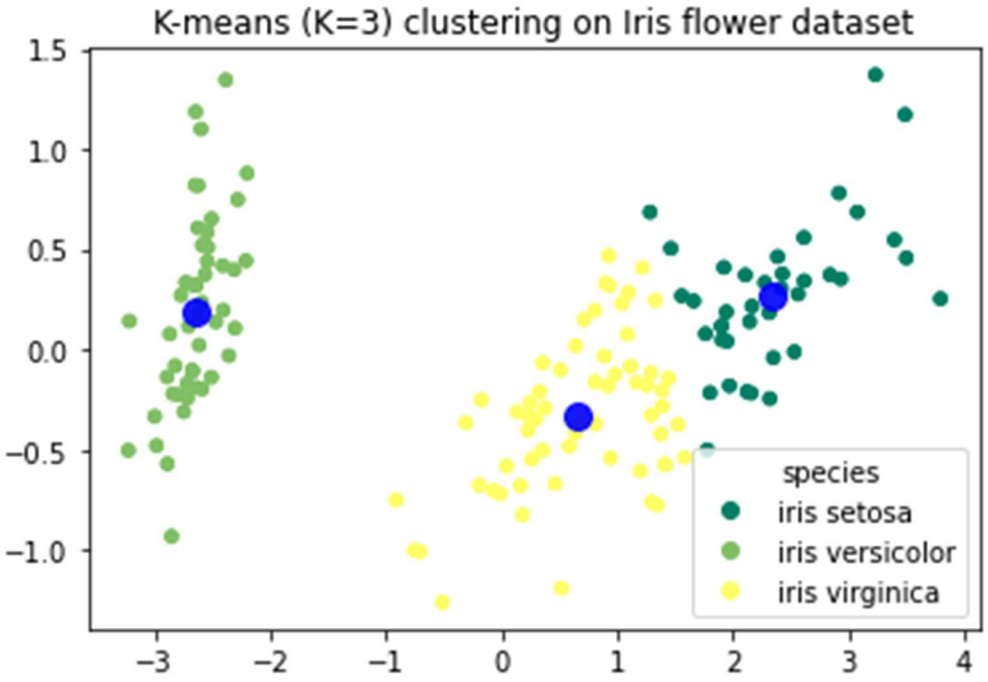

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. It is able to cluster data according to their lables with less errors.
2. The boundaries are not exactly circular. But it resembles to that at a descent level.

b. The value for distortion measure is 63.874

c. The purity score after examples are assigned to the clusters is 0.887

3

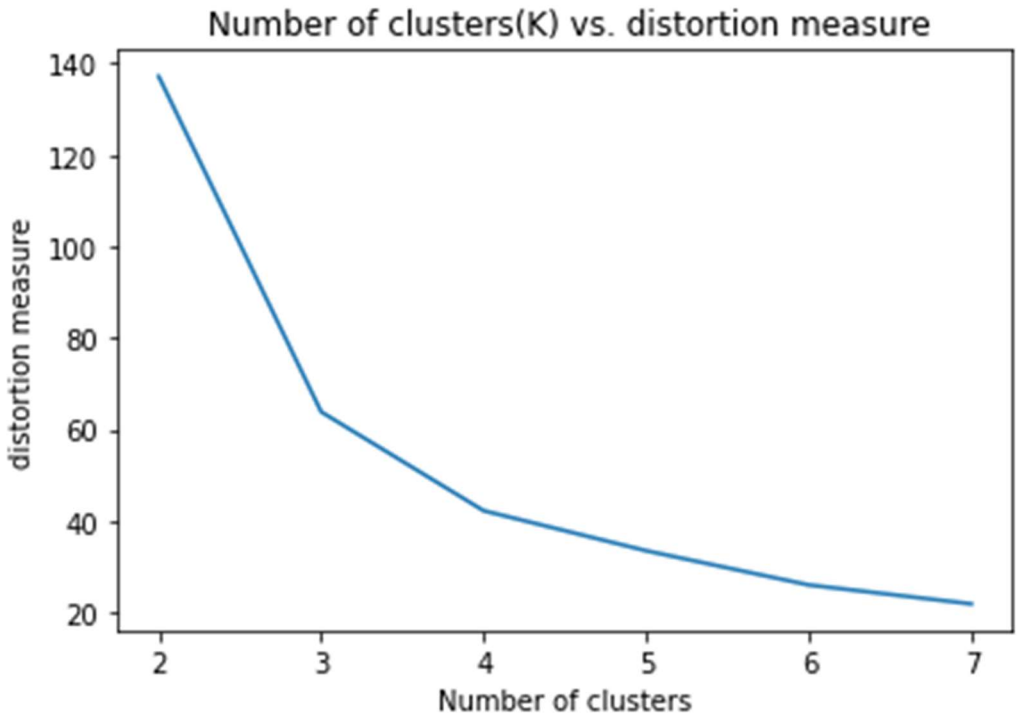

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measure decreases with increase in number of clusters.
2. When there are more clusters, the data are closer to their cluster centers. So, the value of distortion measure decreases.
3. From the number of species in the given dataset, intuitively 3 should be the number of optimum clusters. Yes, the elbow and distortion measure plot follows the intuition.

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

	no. of clusters	purity score
0	2.0	0.666667
1	3.0	0.886667
2	4.0	0.693333
3	5.0	0.680000
4	6.0	0.506667
5	7.0	0.506667

Inferences:

1. The highest purity score is obtained with K =3.
2. Increasing the value of K here first increases the purity score and then decreases.
3. When it is moving closer to optimum value of K it increases and when the number of clusters increases from optimum value of K then it decreases.
4. No observable trend between purity score and distortion measure is found here.

4 a.

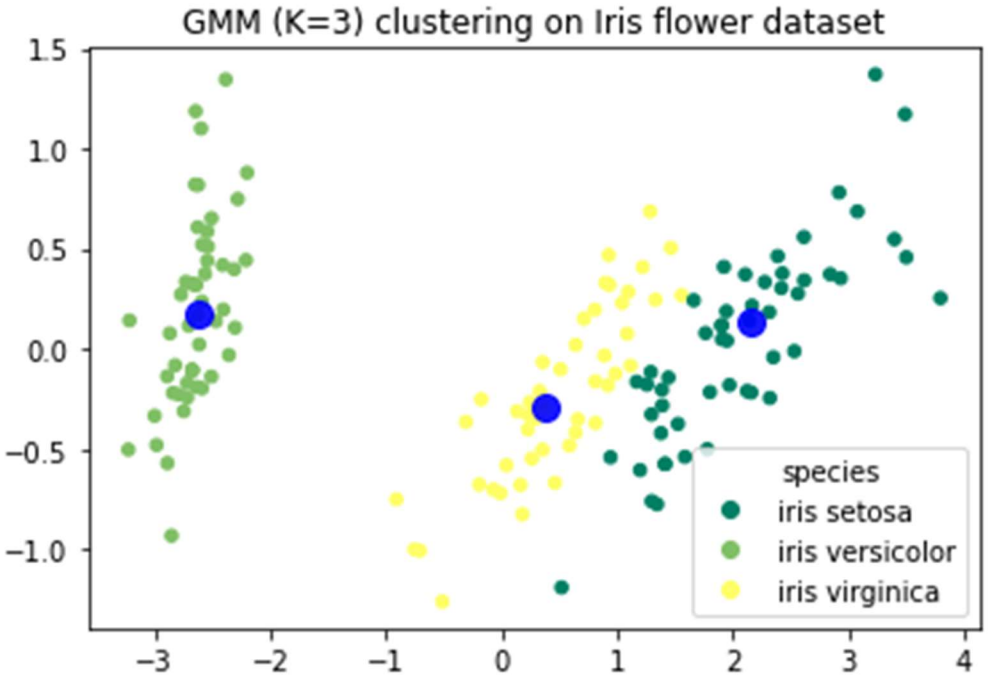

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. It is able to cluster data according to their labels with very less errors.
2. Its boundaries are not exactly elliptical but are similar to that.
3. It is clustering the data in an elliptical fashion while K means is classifying the data in a circular fashion.

b. The value for distortion measure is -280.870.

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

5

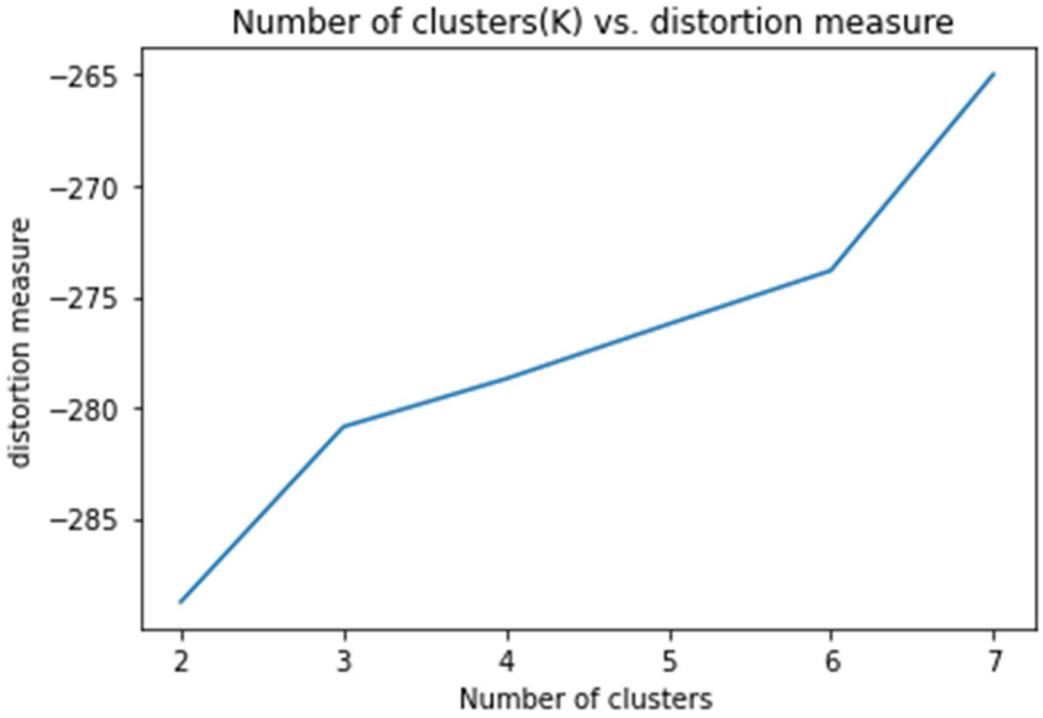

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The magnitude of distortion measure increases with an increase in K.
2. When there are more clusters, the data are closer to their cluster centers. So, the value of distortion measure decreases.
3. From the number of species in the given dataset, intuitively 3 should be the number of optimum clusters. Yes, the elbow and distortion measure plot follows the intuition.

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

	no. of clusters	purity score
0	2.0	0.666667
1	3.0	0.980000
2	4.0	0.833333
3	5.0	0.766667
4	6.0	0.640000
5	7.0	0.626667

Inferences:

1. The highest purity score is obtained with K =.3.
2. When it is moving closer to optimum value of K it increases and when the number of clusters increases from optimum value of K then it decreases.
3. No observable trend between purity score and distortion measure is found here.

6

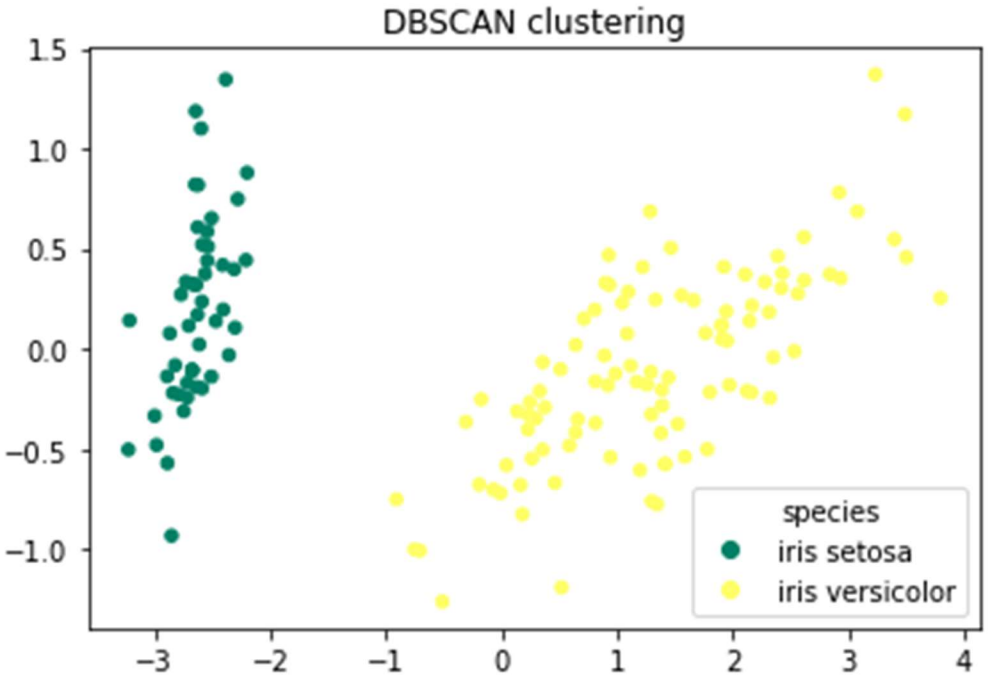

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. It is not able to make clusters as efficiently as the previous techniques were able to.
2. The difference between these clustering techniques is that they have different shape of clusters and different purity score.

b.

	eps	minimum_samples	purity score
0	1	4	0.666667
1	1	10	0.666667
2	5	4	0.333333
3	5	10	0.333333

Inferences:

1. For the same eps value, does increasing min_samples does not change purity score.s
2. For the same min_samples, does increasing eps value decrease purity score.