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

Student's Name: Vishal Sharma

Mobile No: 9540140310

Roll Number: B20239

Branch: Data Science and Engineering

1

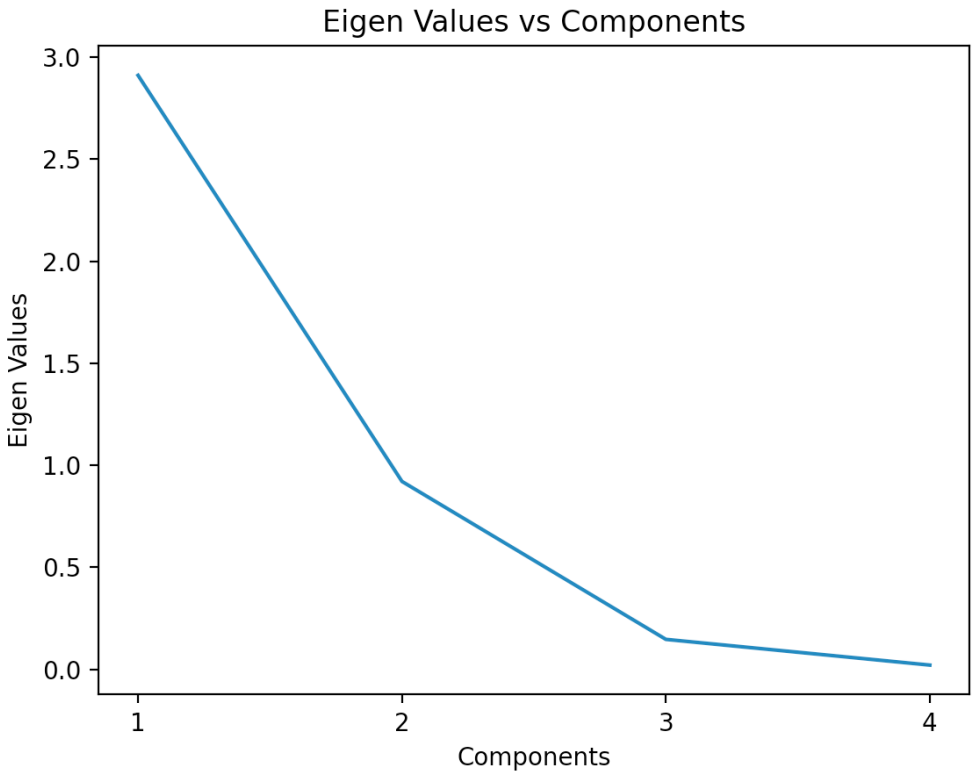

Figure 1 Eigenvalue vs. components

Inferences:

1. The eigen value decreases with increase in the no of components
2. It is because attributes are more dependent on the components having greater eigen value.

2 a.

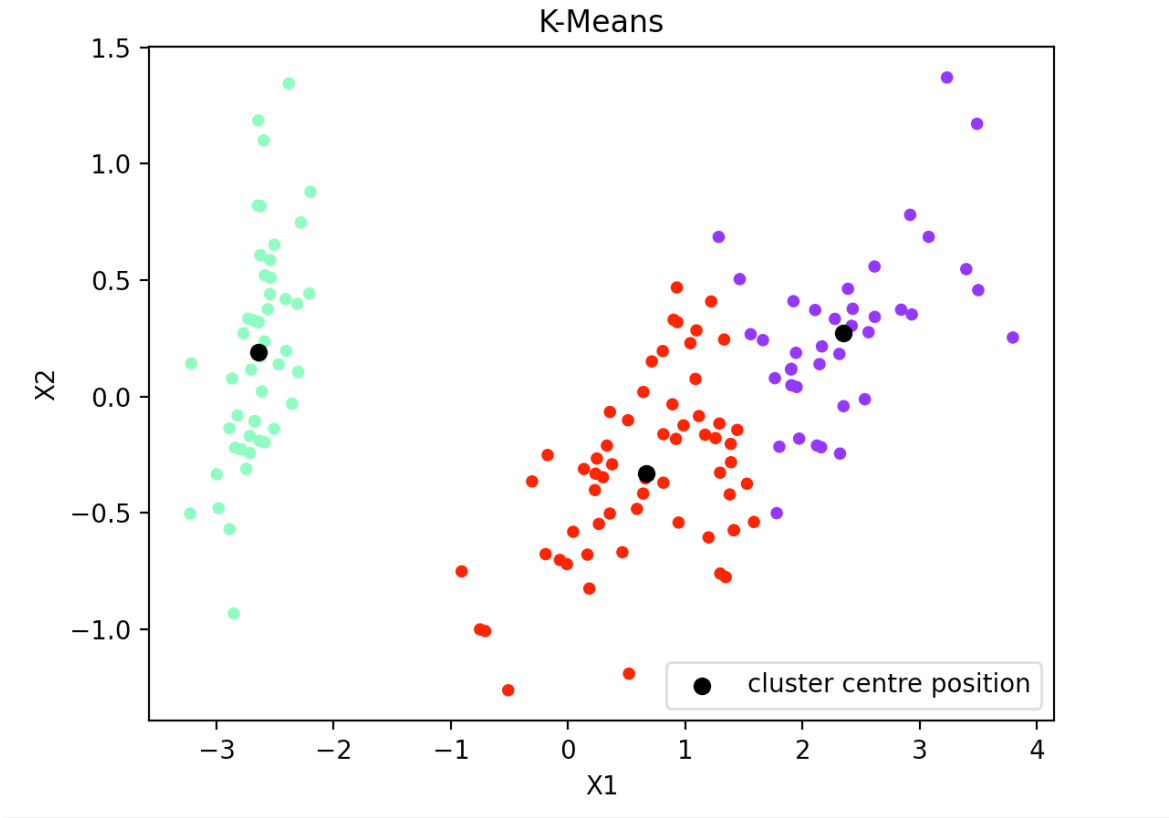

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. K-means clustering algorithm is an unsupervised learning algorithm and judging from the plot, it has clustered points quite well.
2. K-means algorithm clusters data in a circle. Judging from the shape of the clusters, they are not perfect circle, but seems to be forming a circle.

b. The value for distortion measure is 63.874

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

3

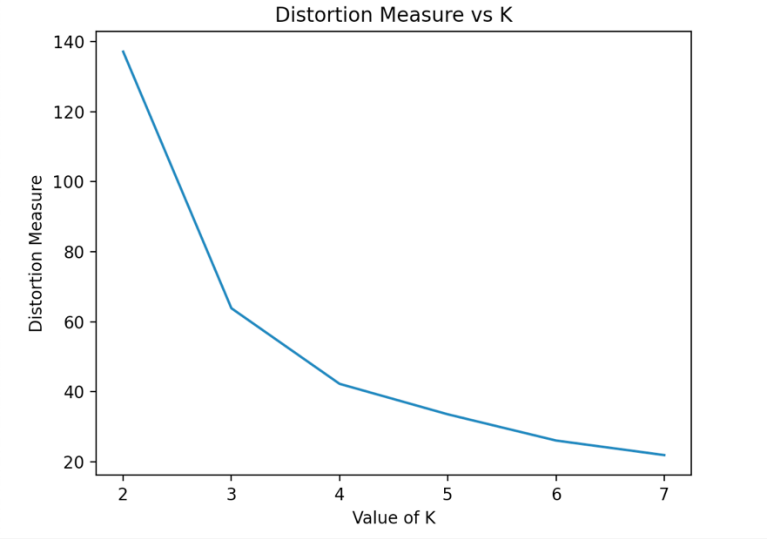

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with increase in value of K.
2. The number of different species in the given data is 3. Thus, distortion decreases drastically from K=2 to K=3
3. Judging from the above plot, K=3 will give appropriate clusters. It will also follow the elbow vs distortion measure plot.

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

K value	Purity score
2	0.667
3	0.887
4	0.693
5	0.68
6	0.507
7	0.507

Inferences:

1. The highest purity score is obtained with K =.3
2. Purity scores increase from K=2 to K=3 but decreases afterwards.
3. As the given data has 3 different species, K=3 will cluster the data with most accuracy.
4. Yes, except for K=3, value of purity score increases with decrease in distortion measure.

4 a.

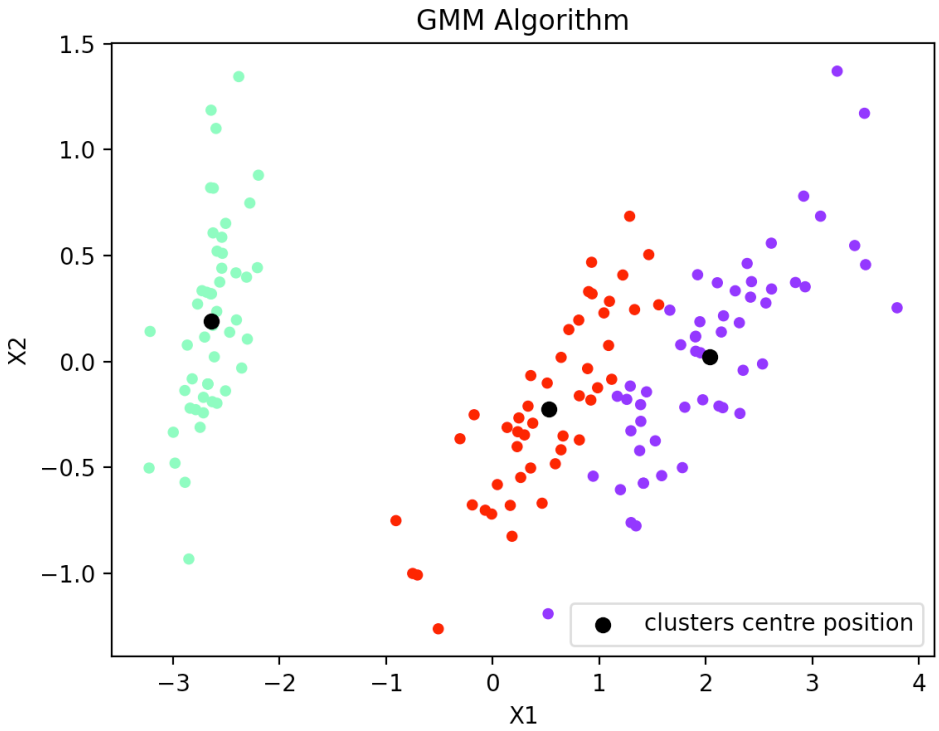

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. As the predicted values are quite close to the real value, we can say that the GMM model is quite accurate.
2. From the above plot, the clusters seem to take the shape of ellipse, which is true for GMM to take an elliptical shape in 2-d.
3. Yes, from the graph of K-Means, the clusters were circular, while in GMM they seems to take the shape of ellipse.

b. The value for distortion measure is -280.87

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

5

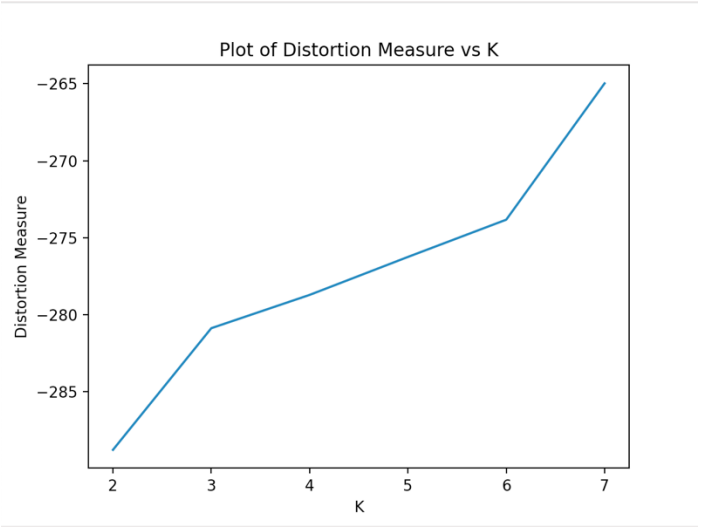

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measure increases with increase in value of K.
2. As there are only 3 species the given data, there is greater slope from K =2 to K=3, but it becomes gradual afterwards and increase abruptly after K = 6.
3. Judging from the above plot, K=3 will give appropriate clusters. It will also follow the elbow vs distortion measure plot.

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

K value	Purity score
2	0.667
3	0.980
4	0.833
5	0.767
6	0.64
7	0.627

Inferences:

1. The highest purity score is obtained with K =3.
2. Purity scores increase from K=2 to K=3 but decreases afterwards.
3. As the given data has 3 different species, K=3 will cluster the data with most accuracy.
4. Yes, except for K=3, value of purity score increases with decrease in distortion measure.
5. From the values of purity score, GMM model can predict data more accurately than K-Means Model.

6

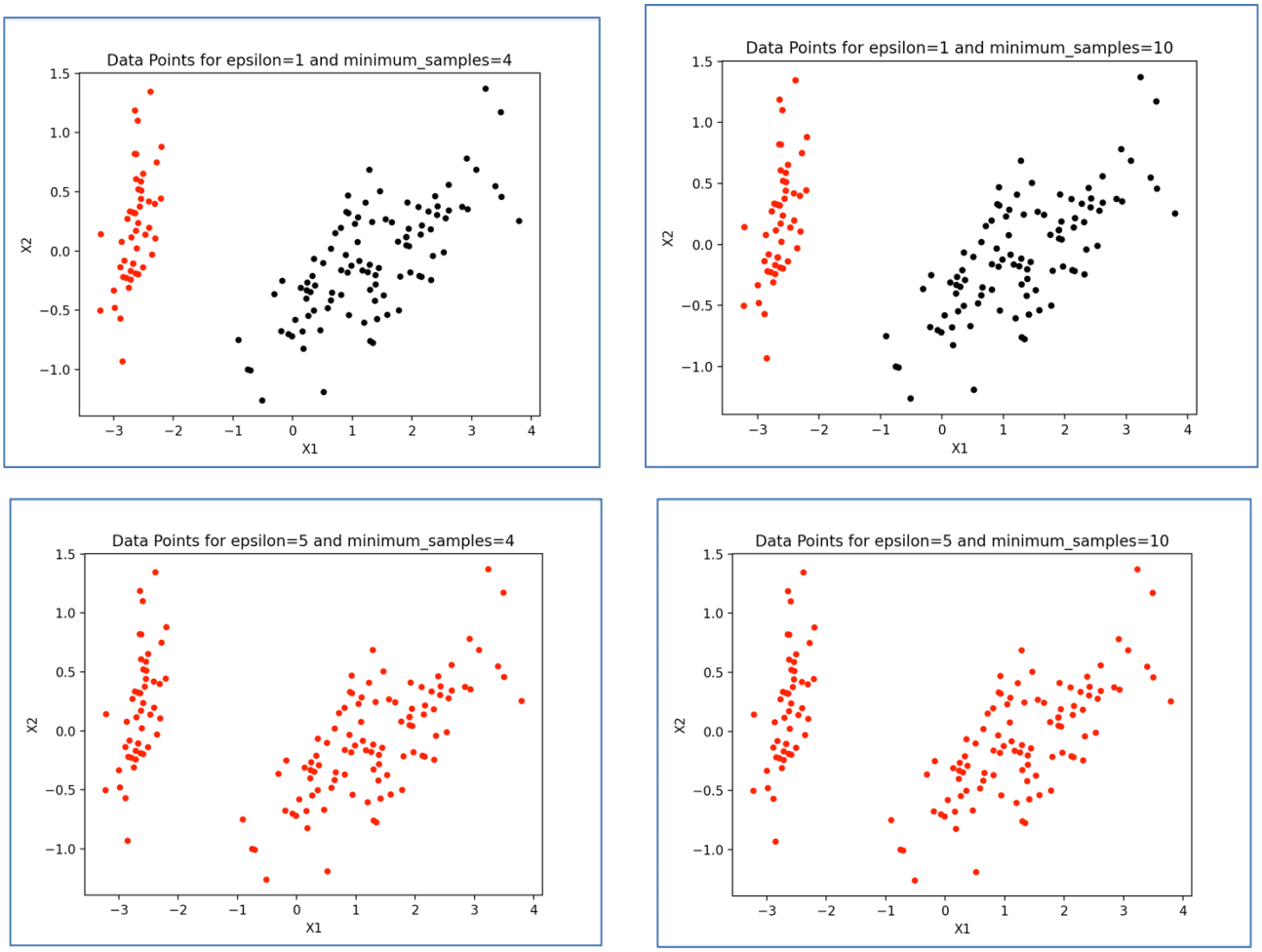

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. From the above scatter plots, it seems that the accuracy is not good. There can be many factors, like the value chosen for epsilon and min_points is not appropriate.
2. The boundary Is not clear in the DBSCAN model, also the number of clusters are also less.

b.

Eps	Min_samples	Purity Score
1	5	0.667
	10	0.667
4	5	0.333
	10	0.333

Inferences:

1. For same value of epsilon, the purity score does nto change with increase in number of min_points.
2. For same value of min_points, the value of purity score decreases with increase in value of epsilon.