

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

---

Student's Name: Gaurav Kumar

Mobile No: 8529143452

Roll Number: B20197

Branch:EE

---

1

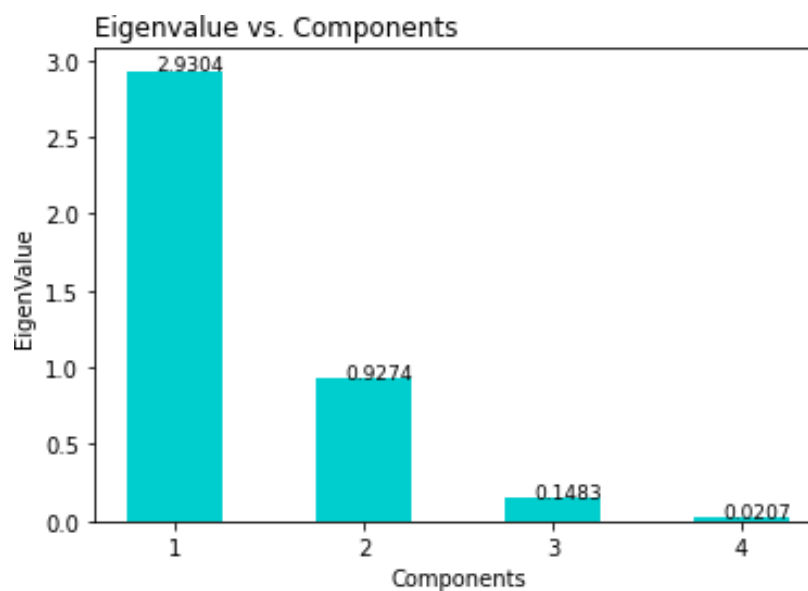


Figure 1 Eigenvalue vs. components

**Inferences:**

1. Eigen Value is decreasing with every successive increase in component value.
2. Eigen value represent variance of components so, it's natural that some component will have cover more variance compared to other. Highest eigen value represents whole dataset better compared to others.

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

2 a.

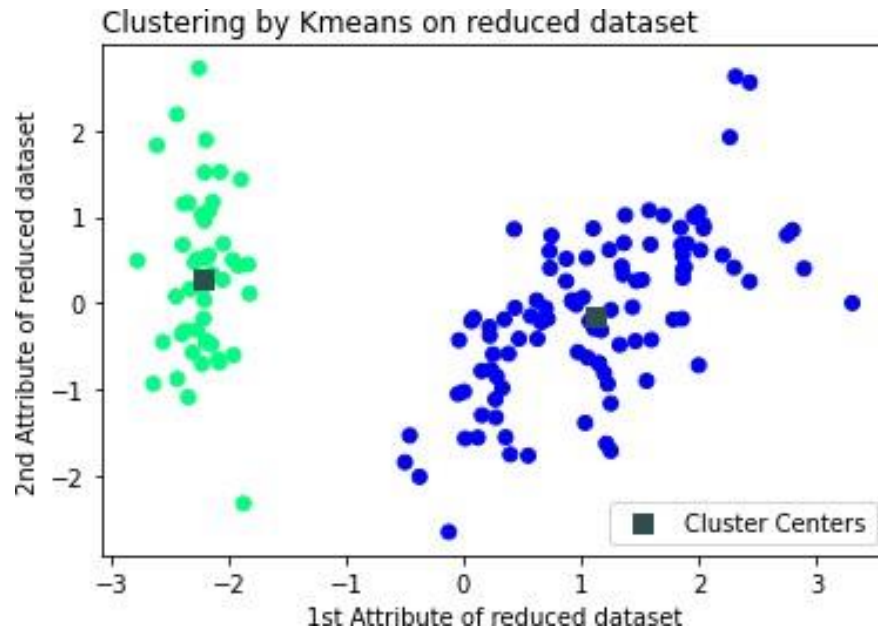


Figure 2 K-means (K=3) clustering on Iris flower dataset

**Inferences:**

1. Clustering prowess of the algorithm is too good.
2. **No**, boundaries of clustering are more in a straight line.

**b.** The value for distortion measure is 145.525.

**c.** The purity score after examples are assigned to the clusters is 0.667.

## IC 272: DATA SCIENCE - III

### LAB ASSIGNMENT – VII

#### Clustering

3

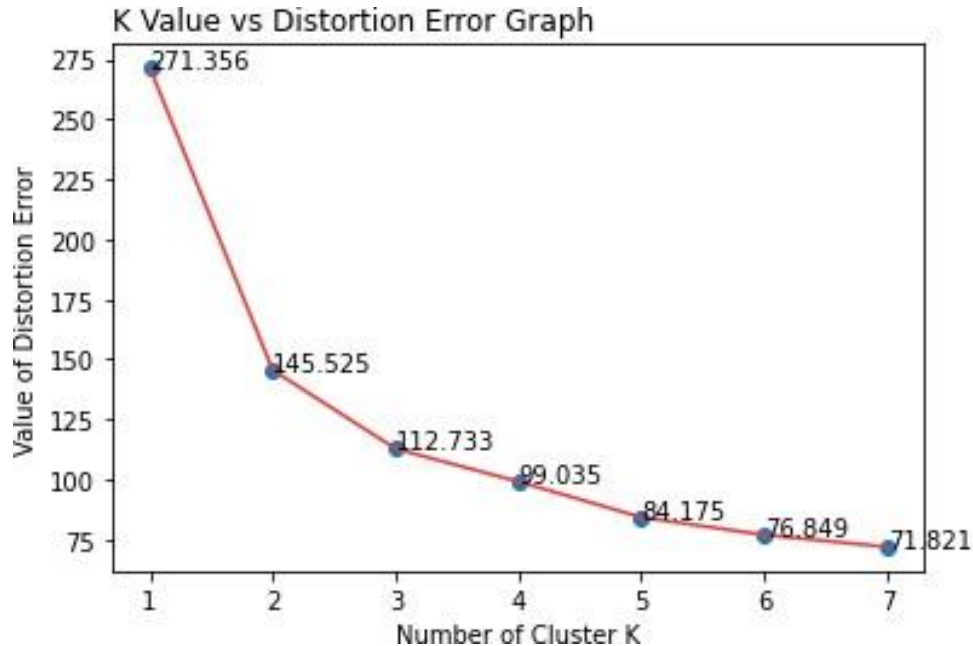


Figure 3 Number of clusters(K) vs. distortion measure

#### Inferences:

1. Distortion measure decreases with the increase in value of K.
2. Distortion measure is the sum of the square of the distance of each example to its assign cluster center. So, if number of clusters are less than some points will be more far away from their respective cluster centers in comparison to the case in which clusters are more. Hence, distortion measure will be high for a smaller number of clusters.
3. Intuitively, there should be only **3** clusters in the dataset. No, because Kmeans is an unsupervised clustering algorithm means it didn't use the labels of data points and there are only two visible clusters so, it gives 2 as optimal number of clusters.

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

K value	Purity score
2	0.667
3	0.833
4	0.667
5	0.547
6	0.507
7	0.453

## IC 272: DATA SCIENCE - III

### LAB ASSIGNMENT – VII

#### Clustering

#### Inferences:

1. The highest purity score is obtained with  $K=3$ .
2. By increasing the value of  $K$  decreases the purity score.
3. Since, in original dataset there are only 3 labels. So, if we increase the value of  $k$  more than 3 than some points will get assigned to which they don't which leads in decrease in purity score.
4. Yes, after maximum value of purity score, its value decreases with the increase in  $K$ .

4 a.

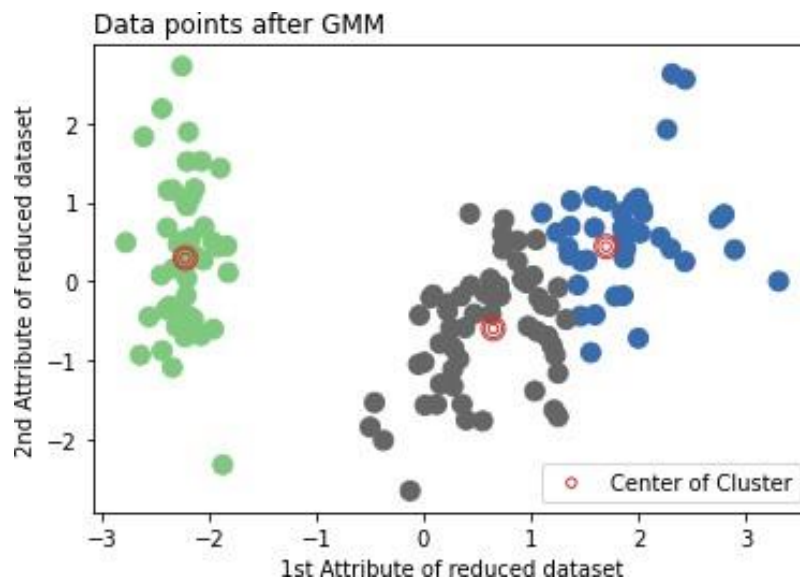


Figure 4 GMM ( $K=3$ ) clustering on Iris flower dataset

#### Inferences:

1. Clustering power of algorithm is very good.
2. No.
3. No.

b. The value for distortion measure is -379.923

c. The purity score after examples are assigned to the clusters is 0.867.

## IC 272: DATA SCIENCE - III

### LAB ASSIGNMENT – VII

#### Clustering

5

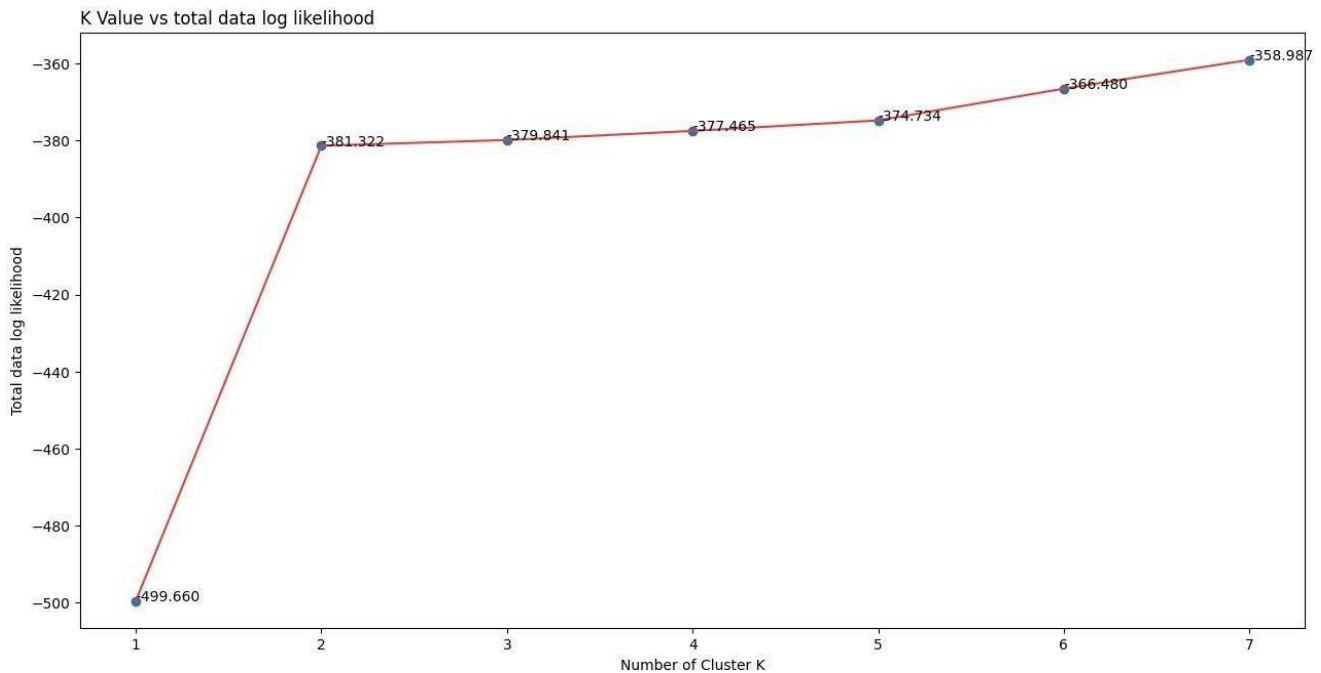


Figure 5 Number of clusters(K) vs. distortion measure

#### Inferences:

1. Distortion measure decreases with an increase in K value.
2. From the scatter plot of dataset there are only two visible clusters and by using elbow method we get the optimal value of clusters is 2. So, after  $K = 2$  decrease in distortion measure becomes linear.
3. Intuitively, there should be **3** clusters in the dataset. No, because GMM is an unsupervised clustering algorithm means it didn't use the labels of data points and there are only two visible clusters so, it gives 2 as optimal number of clusters.

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

K value	Purity score
2	0.667
3	0.893
4	0.740
5	0.647

## IC 272: DATA SCIENCE - III

### LAB ASSIGNMENT – VII

#### Clustering

6	0.673
7	0.493

#### Inferences:

1. The highest purity score is obtained with  $K = 3$ .
2. Increasing the value of  $K$  decreases the purity score.
3. Same as explained above in the Kmeans.
4. Yes, after maximum value of purity score, its value decreases with the increase in  $K$ .

6

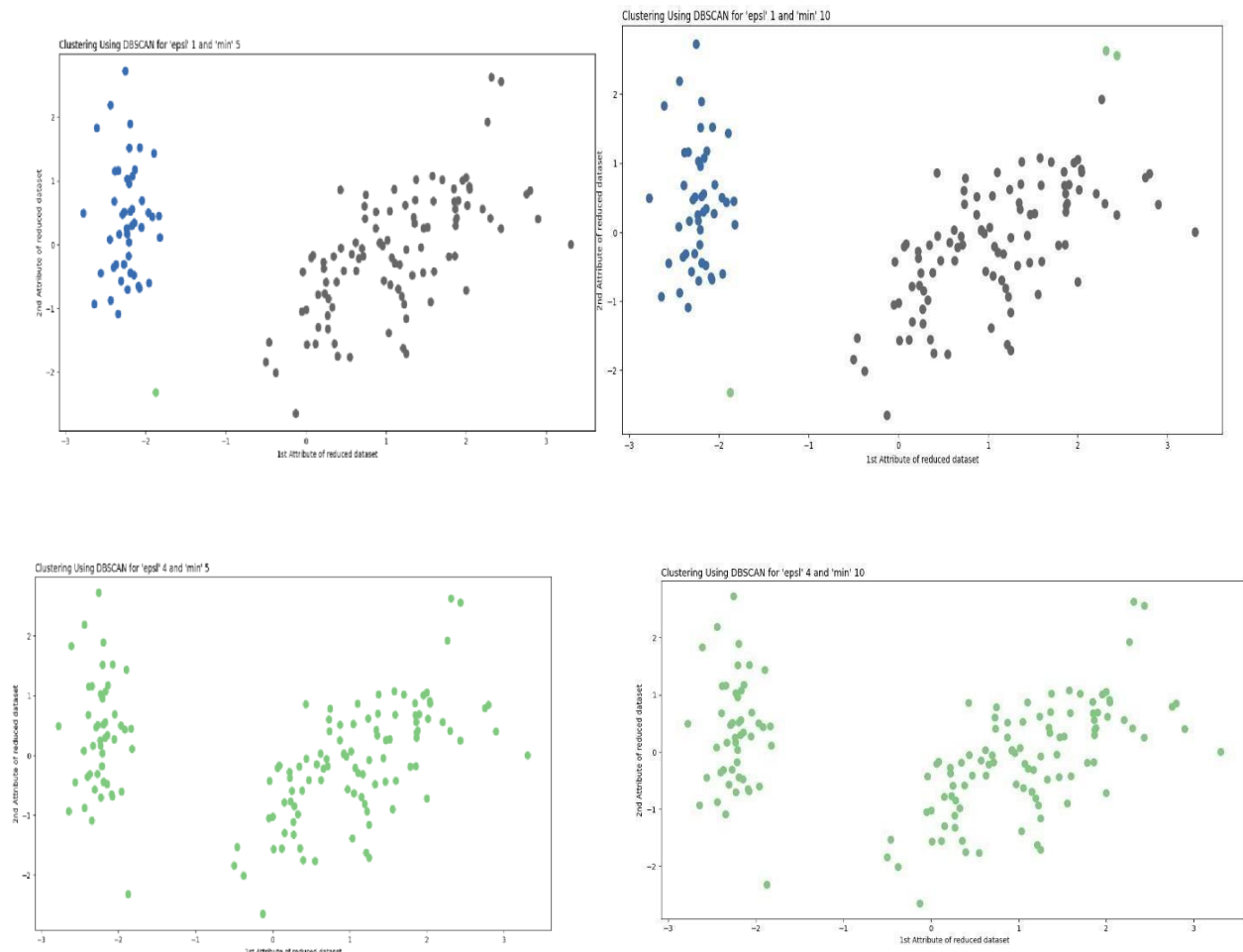


Figure 6 DBSCAN clustering on Iris flower dataset

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

---

**Inferences:**

1. If we choose optimal number of epsl and min\_points then clustering prowess is very good.
2. Yes, in 2.a and 4.a we have given the value of K. But in 6.a number of clusters is decided by the algorithm itself and since, it's an unsupervised technique due to which means it will cluster datapoints according to similarity not by seeing whether a point belongs to certain label or not. So, it results in the formation of two clusters.

**b.**

Eps	Min_samples	Purity Score
1	5	0.660
	10	0.673
4	5	0.333
	10	0.333

**Inferences:**

1. For the same eps value, does increasing min\_samples increase purity score.
2. For the same min\_samples, does increasing eps value decrease purity score.