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

Student's Name: Kuldeep Jain Dugar **Branch:**

Roll Number: B20112 **CSE**

Mobile No: 8986388665

1

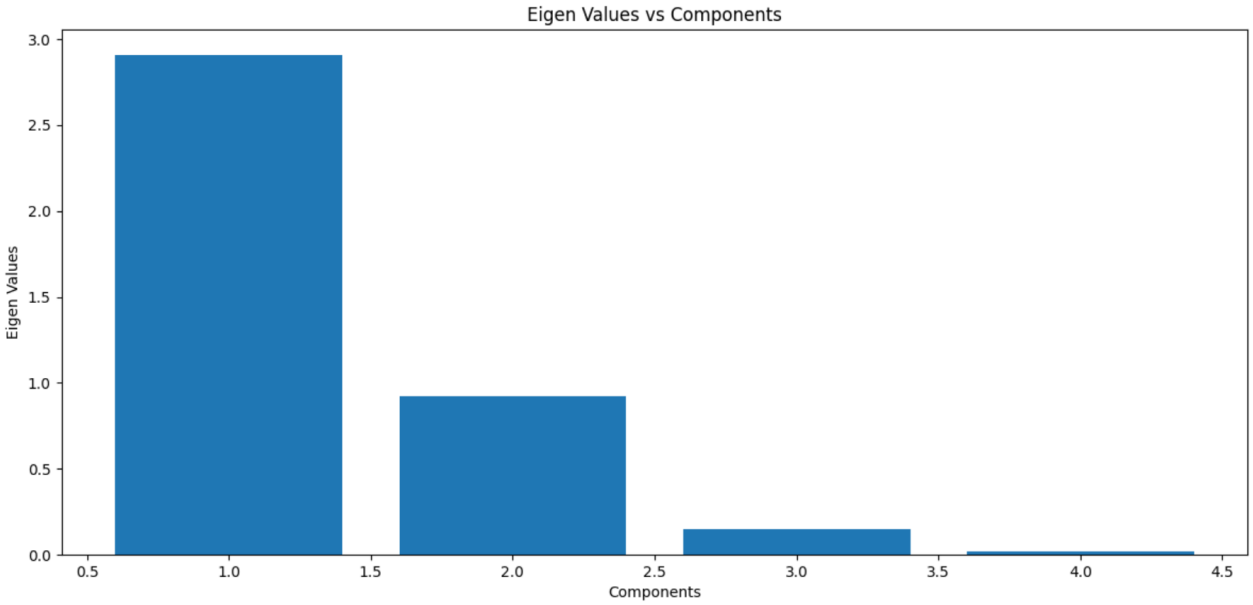


Figure 1 Eigenvalue vs. components

Inferences:

1. Does the eigenvalue increase or decrease corresponding to each component increase or decrease successively?- It DECREASES
2. They represent variance of components, so some will have more and the other will have less.

2 a.

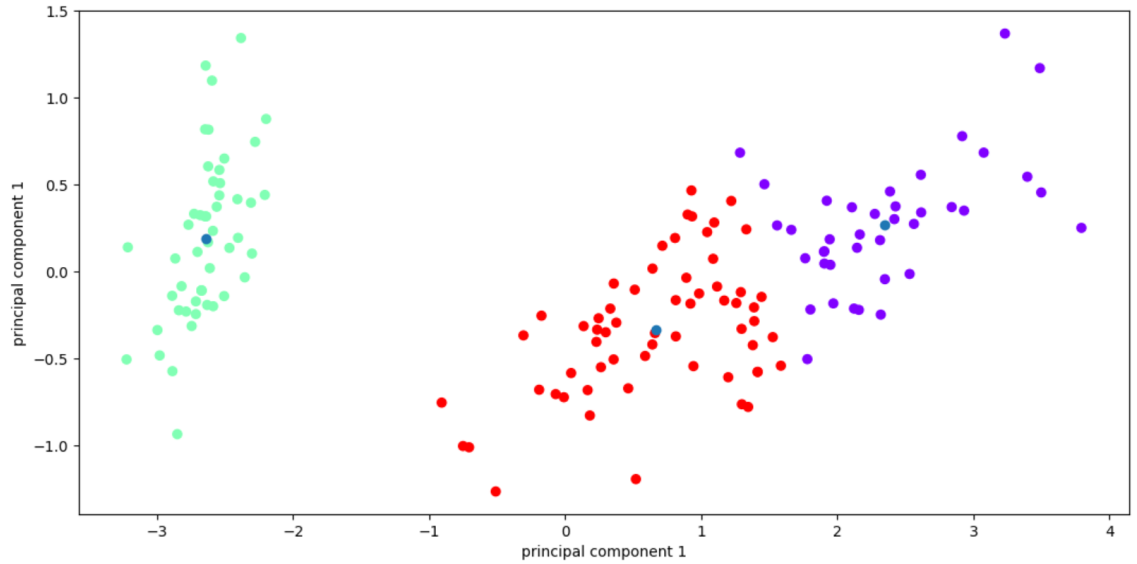


Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. Good Clustering algorithm
2. K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, does the boundary seem to be circular?- No , its more a straight line

b. The value for distortion measure is 63.873

c. The purity score after examples are assigned to the clusters is 0.886

3

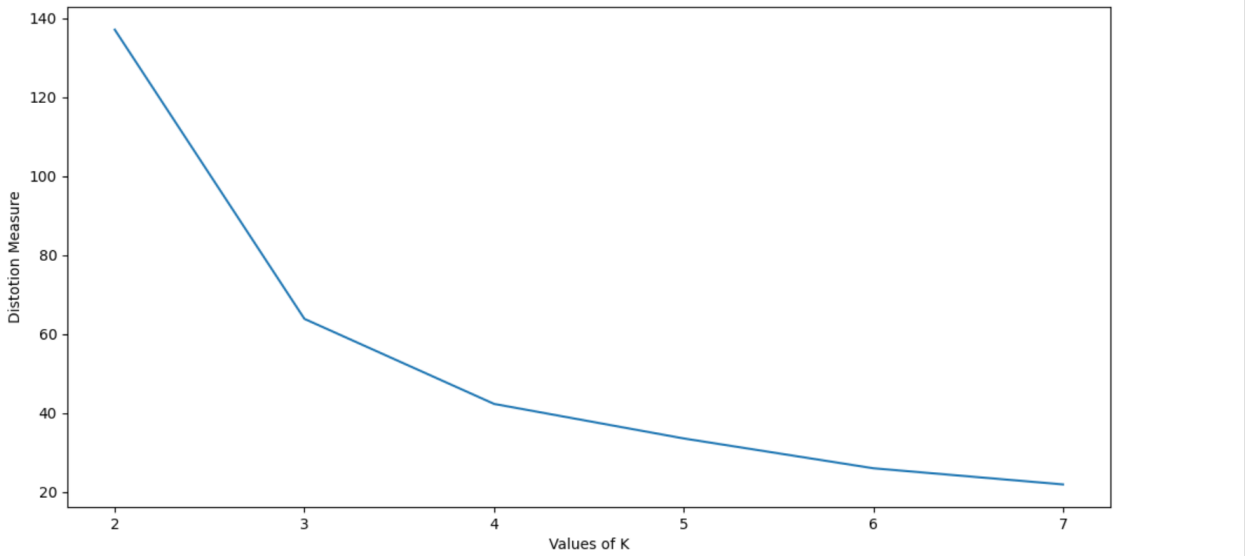


Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. the distortion measure decreases with an increase in K?
2. As we increase no. of cluster the cluster centers will be more spreaded out and distance of individual points with their centers will be less
3. From the number of species in the given dataset, intuitively what should be the number of optimum clusters? – 3
4. Does the elbow and distortion measure plot follow the intuition? – NO, elbow method suggests there should be 2.
   Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Label x-axis as distortion measure and y-axis as number of clusters (K).

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

K value	Purity score
2	0.667
3	0.887
4	0.687
5	0.667
6	0.52
7	0.51

Inferences:

1. The highest purity score is obtained with K = 3
2. increasing the value of K first increase then decrease the purity score.
3. Since the real data has only 3 labels so on
4. Is there any observable relationship between purity score and distortion measure?- yes till k =3, More distortion less purity

4 a.

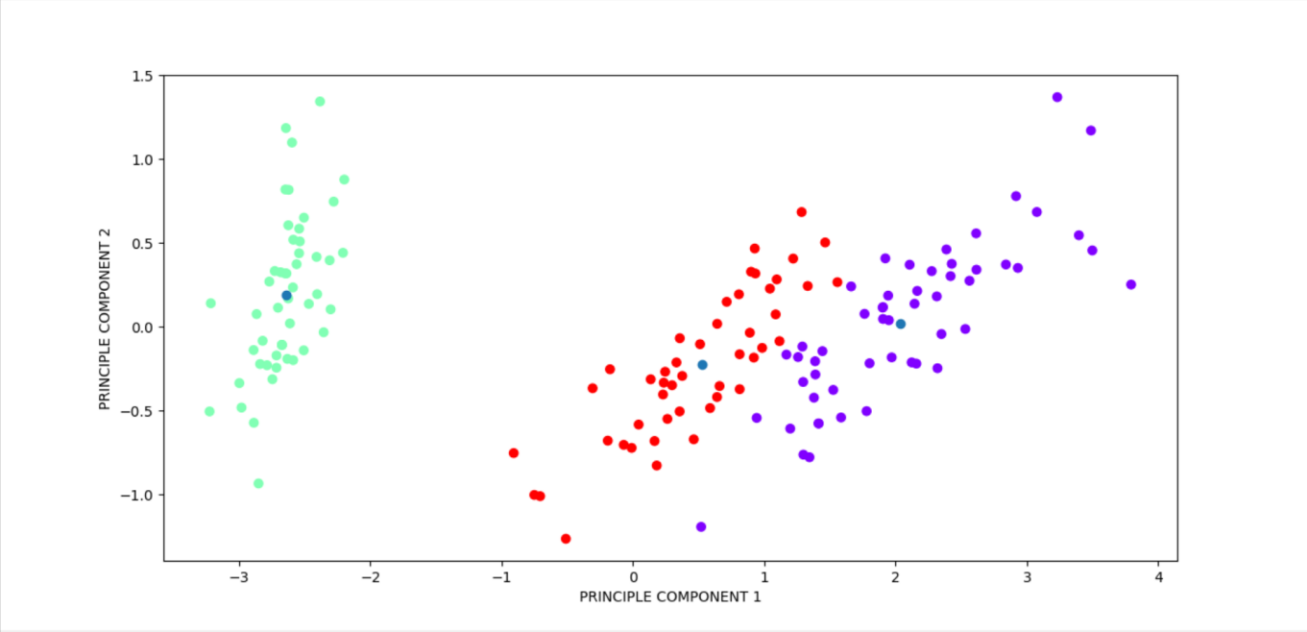


Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. Clustering process of the algorithm is very good
2. GMM algorithm assumes cluster boundaries to be elliptical in 2D. From the output, does the boundary seem to be circular**?- no**
3. Is there any observable difference between clusters formed using K-means in 2.a and GMM in 4.a?- **no**

b. The value for distortion measure is -280.87

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

5

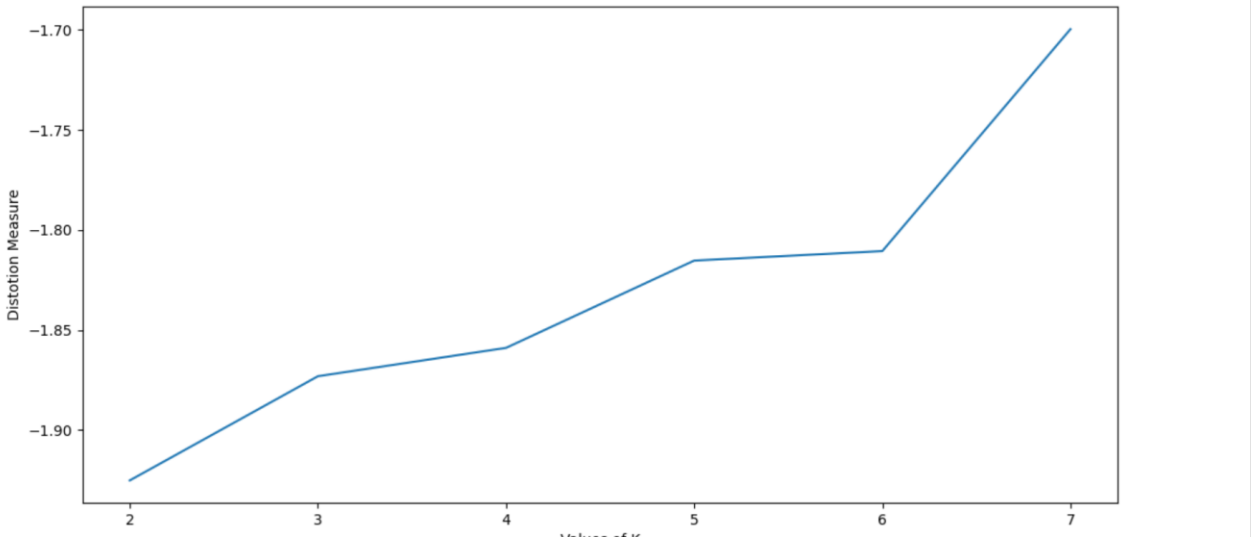


Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. Does the distortion measure decrease in magnitude with an increase in K?
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. Intutively 3 clusters must be formed. But elbow method suggests 2.

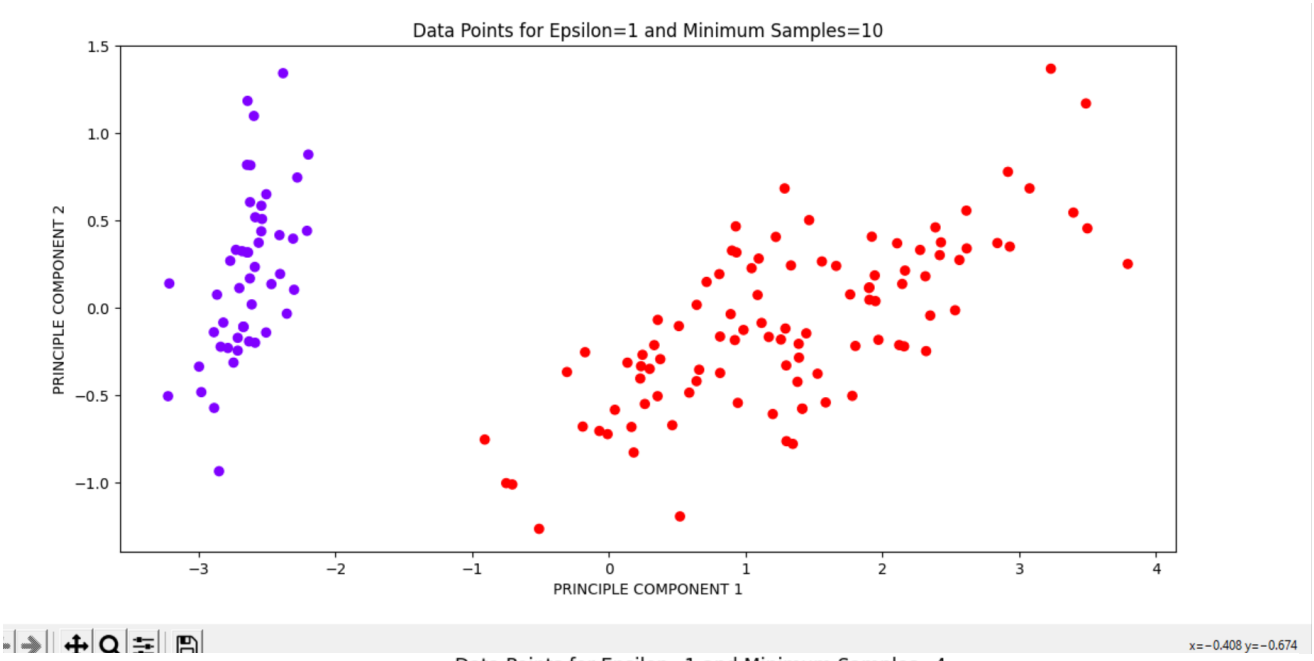
Table 2 Purity score for K value = 2,3,4,5,6 & 7

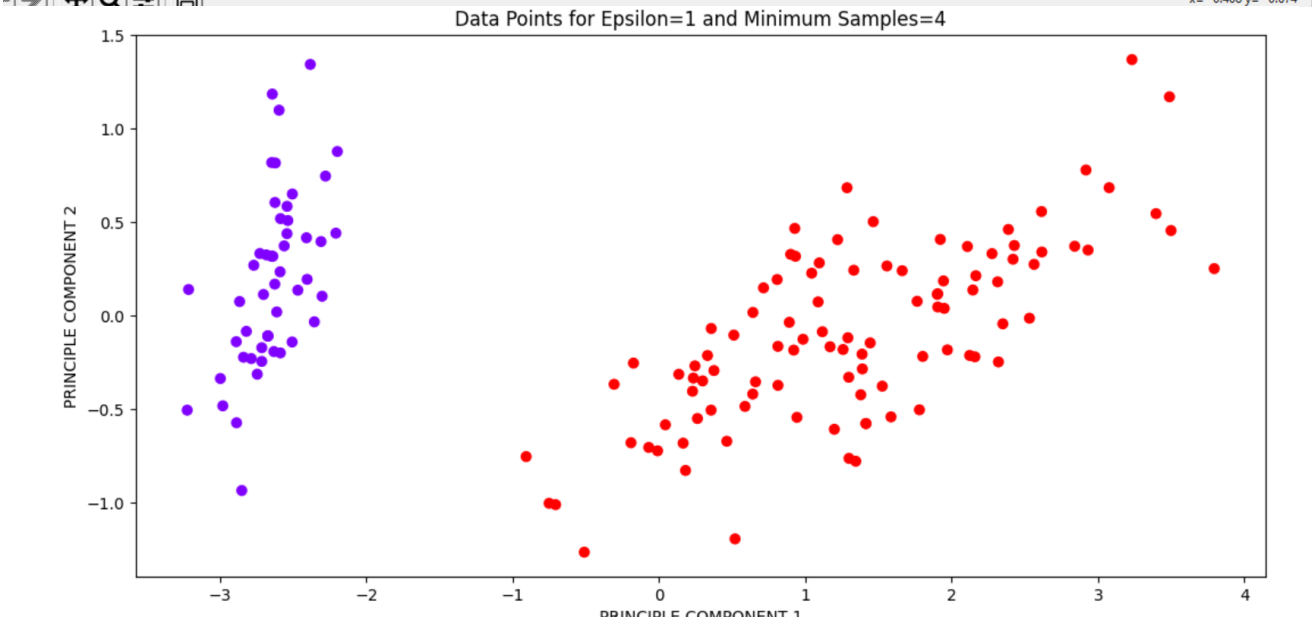
K value	Purity score
2	0.667
3	0.98
4	0.833
5	0.773
6	0.693
7	0.647

Inferences:

1. The highest purity score is obtained with K = 3
2. On increasing the value of K decreases the purity score.
3. Because in the data there is only 3 clusters .
4. Yes, after maximum value of purity score, its value decreases with the increase in K.
5. Compare K-means and GMM based on inferences in Q3 and Q5.- Both have max purity score =3

6


Data Points for Epsilon=1 and Minimum Samples=10
PRINCIPLE COMPONENT 2
PRINCIPLE COMPONENT 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
4


x=−0.408 y=−0.674
Data Points for Epsilon=1 and Minimum Samples=4
PRINCIPLE COMPONENT 2
1.5
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
4

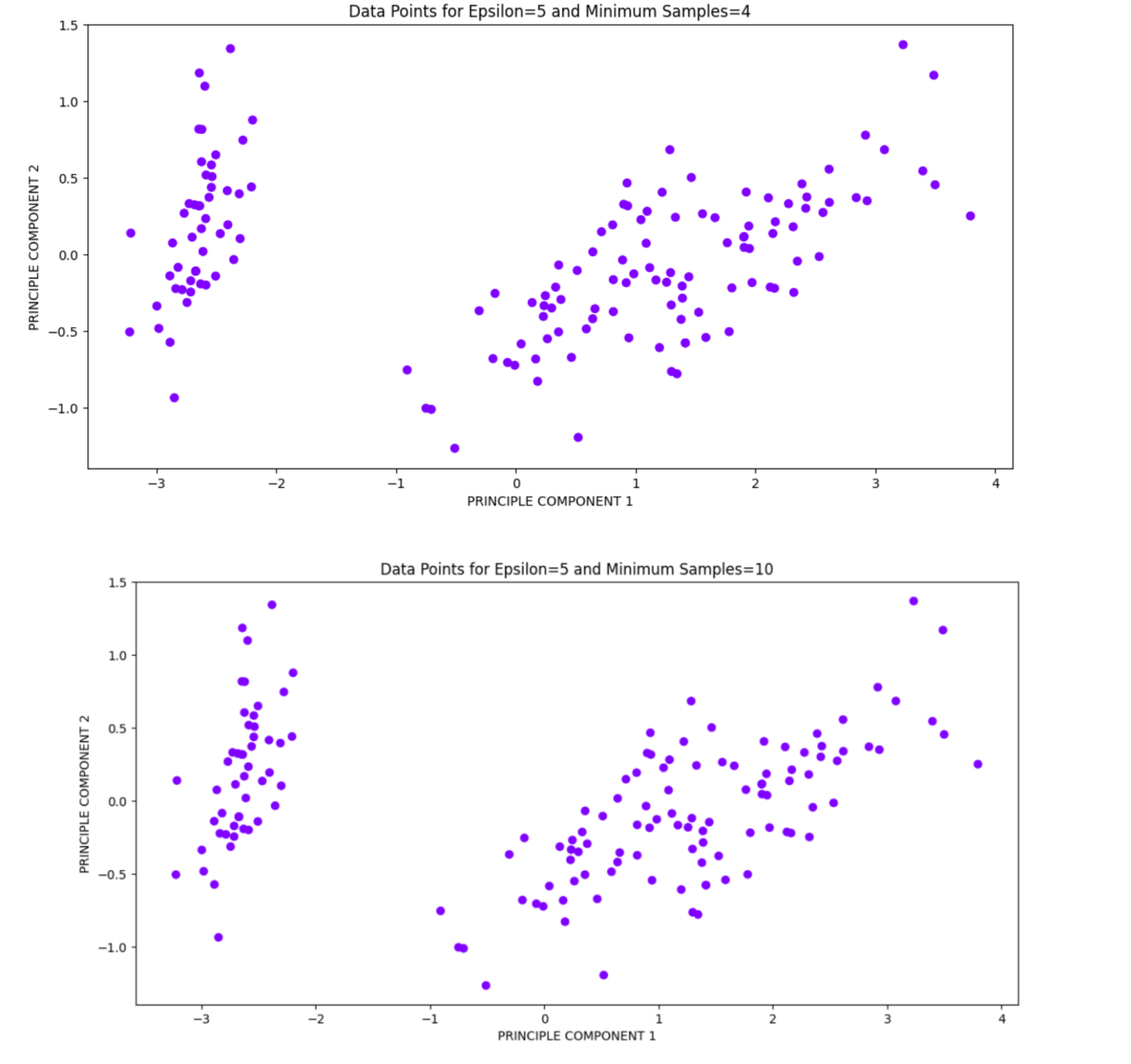


Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, comment on the clustering prowess of the algorithm. – It is very good
2. In 2.a and 4.a , the number of cluster is given but in this case the algorithm decides the number of cluster

b.

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