CSE – 454
DATA MINING

HOMEWORK 1 REPORT

OĞUZHAN SEZGİN

SUMMARY

I implemented DBSCAN algorithm in python language. I used the **math** library to get the square and square root of the numbers when calculating the distance , I used **sklearn** library to get data for test my dbscan implementation, I used **matplotlib** library to to visualize clusters.

I tested 3 different **Eps** variable and 3 different **MinPts** variable.

CASE 1: Below are 3 plot with fixed **MinPts** and changed **Eps** values.

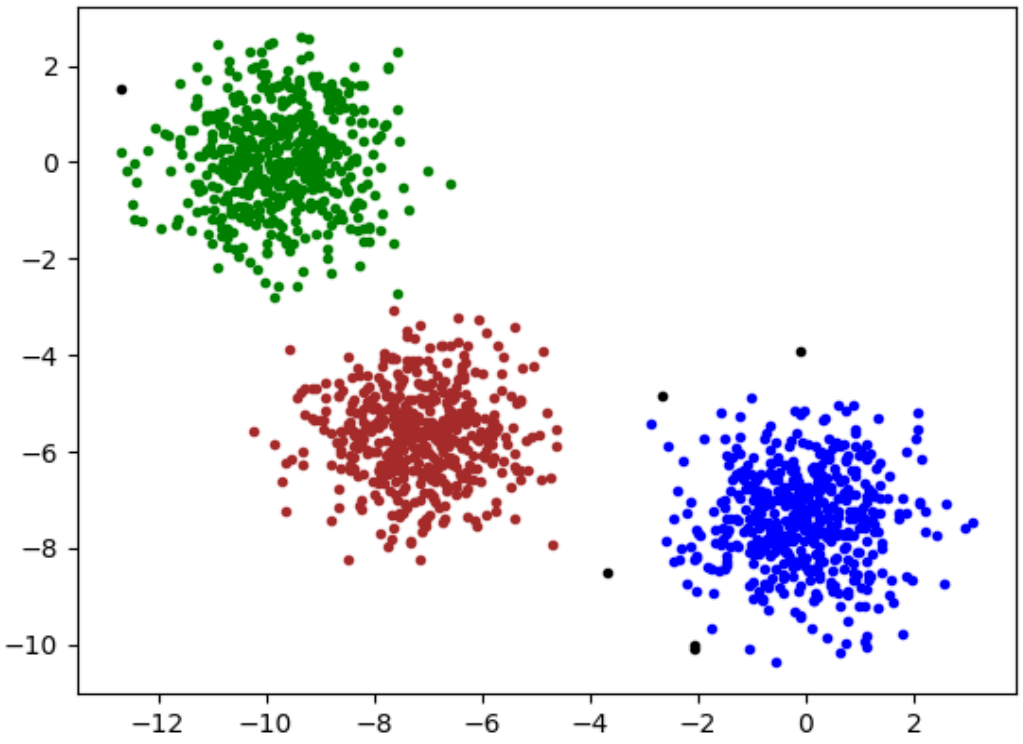

Figure 1. *Eps=1 , MinPts=10*

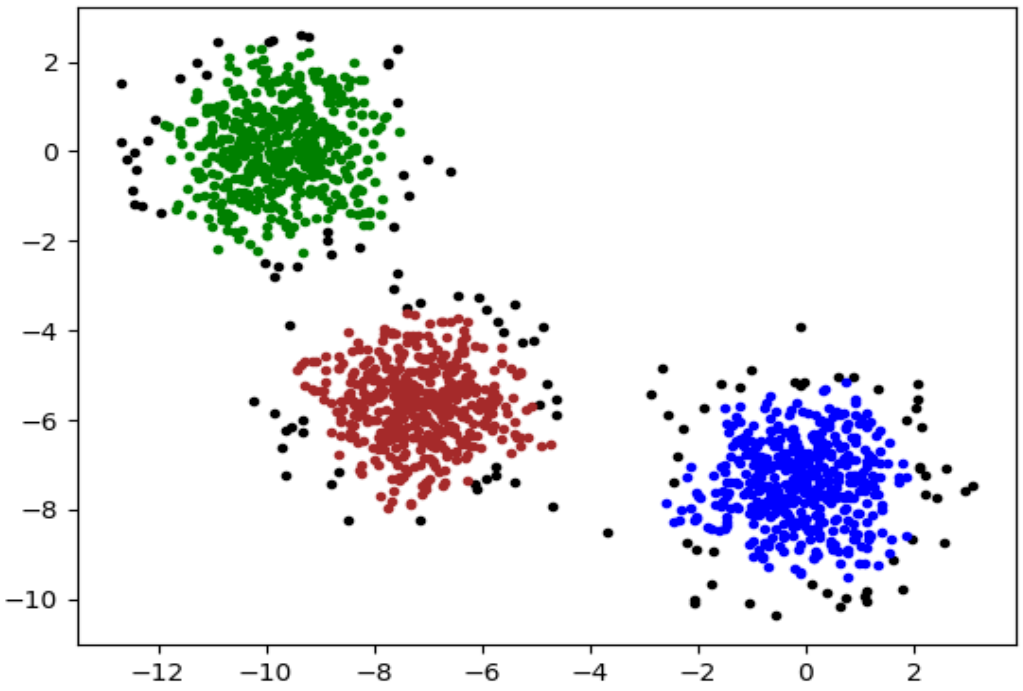

Figure 2. *Eps=0.5 , MinPts=10*

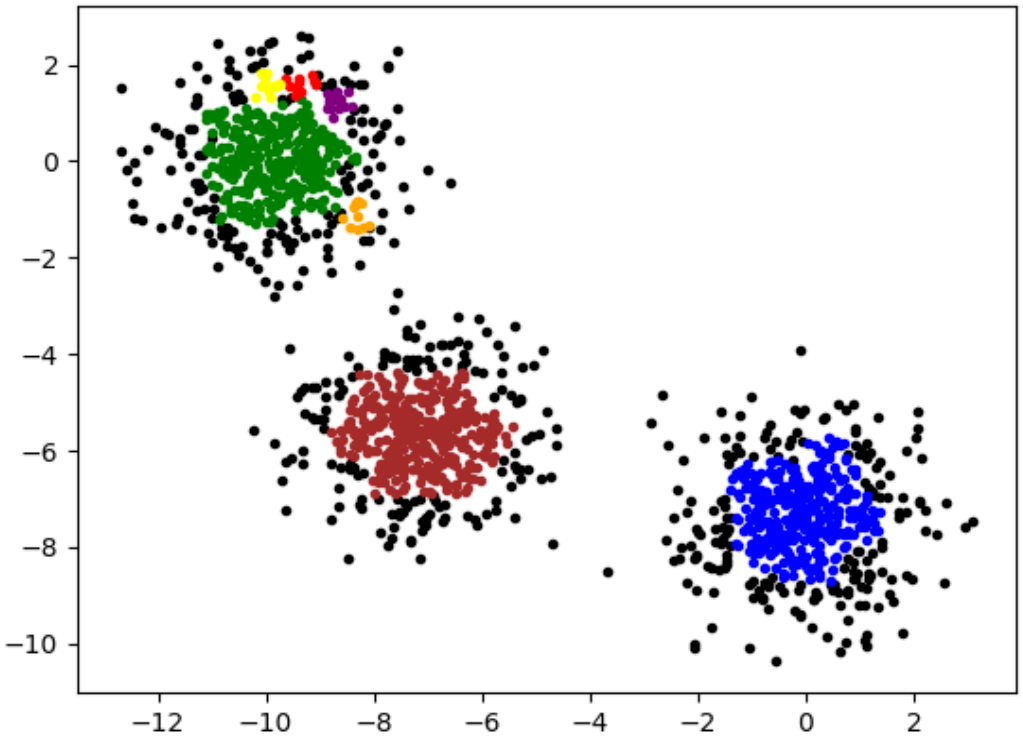

Figure 3. *Eps=0.3 , MinPts=10*

RESULT 1 : As we can see in the case-1 figures, the lower the **Eps** value, the smaller the clusters and new clusters are formed. In addition, as the eps value decreases, the number of noisy points increases.

For this reason, **Eps** values should be kept as large as possible when performing a more general data analysis. Thus, more data is combined in a cluster and a more comprehensive generalization is made.

When a more detailed data analysis is made, the **Eps** value is kept small, so that the data are more separated from each other and the clusters are composed of more similar data.

CASE 2 : Below are 3 plot with fixed **Eps** and changed **MinPts** values.

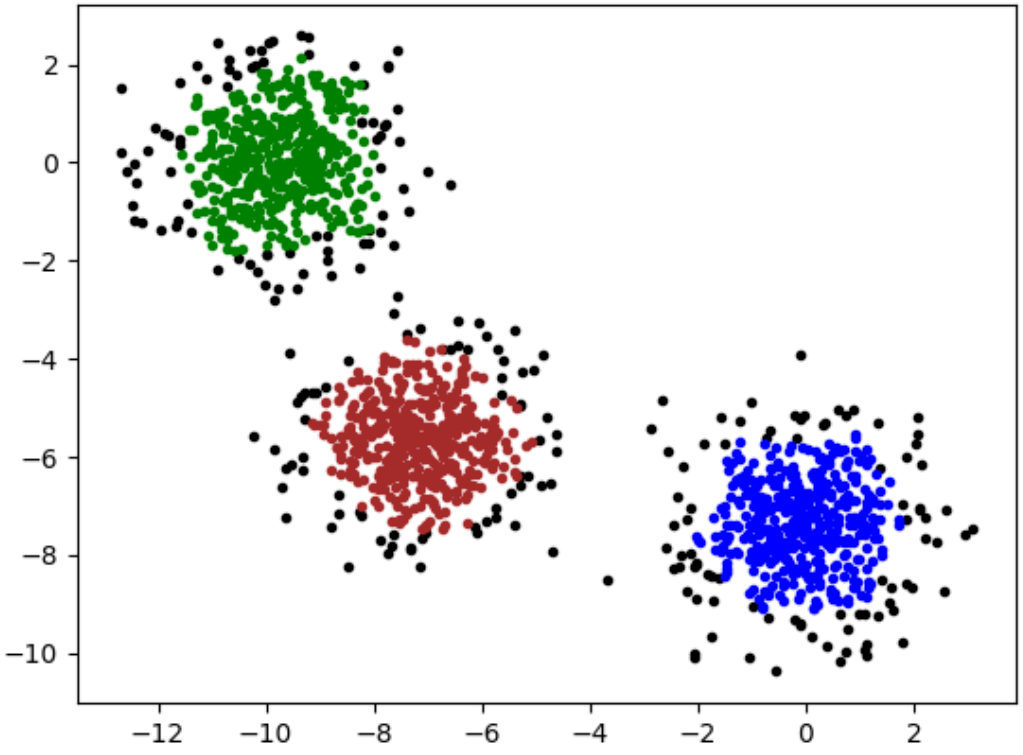

Figure 4. *Eps=0.5 , MinPts=20*

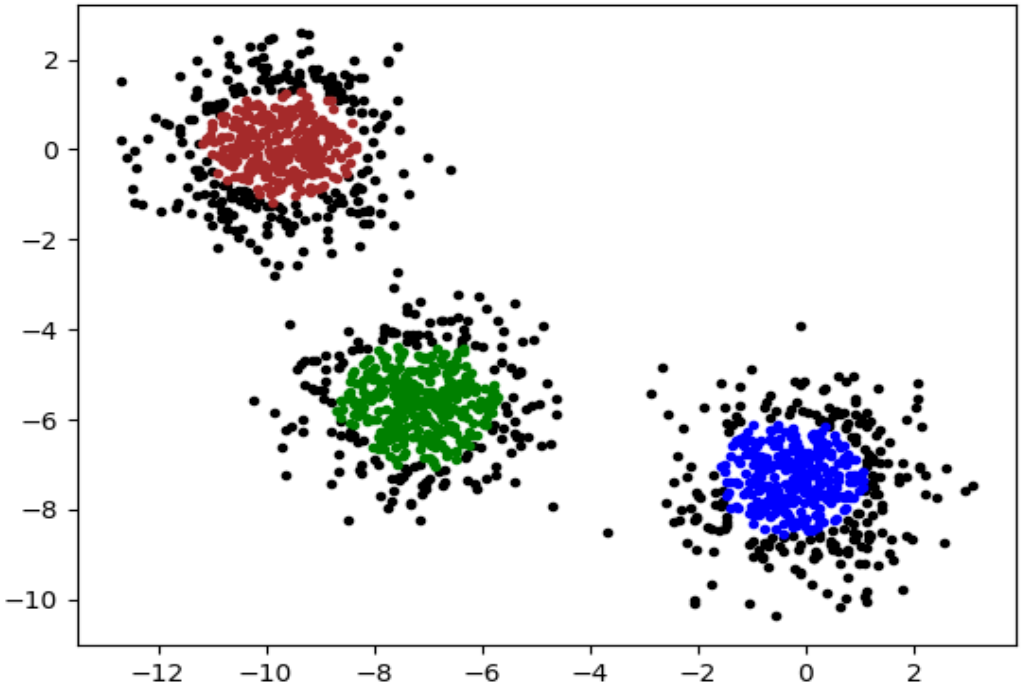

Figure 5. *Eps=0.5 , MinPts=40*

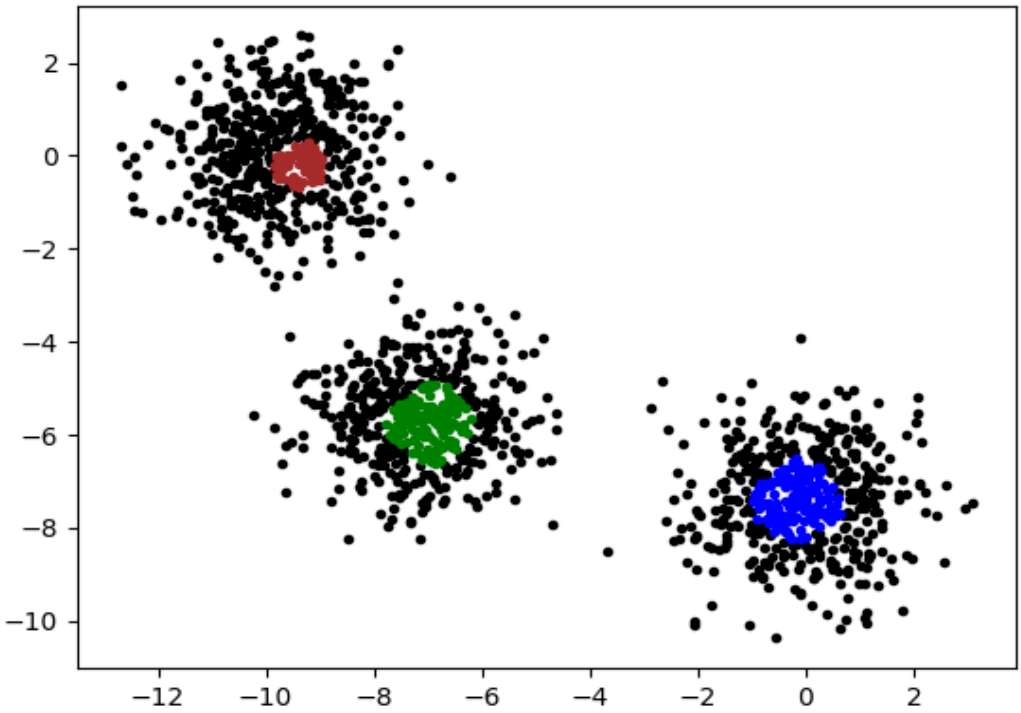

Figure 6. *Eps=0.5 , MinPts=60*

RESULT 2 : As we can understand from case-2 figures, when the **MinPts** value is increased, the number of clusters in the data set does not change. Besides, there is an increased in the number of noisy points in the data set. In this case, more similar values remain in clusters, and they do not occur in new clusters. If low mps value this means that build more cluster from noisy data.

This status can only be used when grouping for certain properties. Thus, unwanted situations do not form new clusters, and a small number of clusters will consist of the most similar data.

OPTIMAL PARAMETERS

The method proposed here consists of calculating k-nearest neighbor distances.

The distance of the K points to the nearest neighbor point is determined and then these distances are averaged. The K value indicates the mps value in the algorithm. The k value is increased and the average values corresponding to these values are reflected on the graph.

An elbow occurs in the graph. The formed elbow point is accepted as the threshold value. The average distance at the threshold value is determined as the eps value, the k value is determined as the mps value.