T.C.

GEBZE TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

DATA MINING
CSE-454

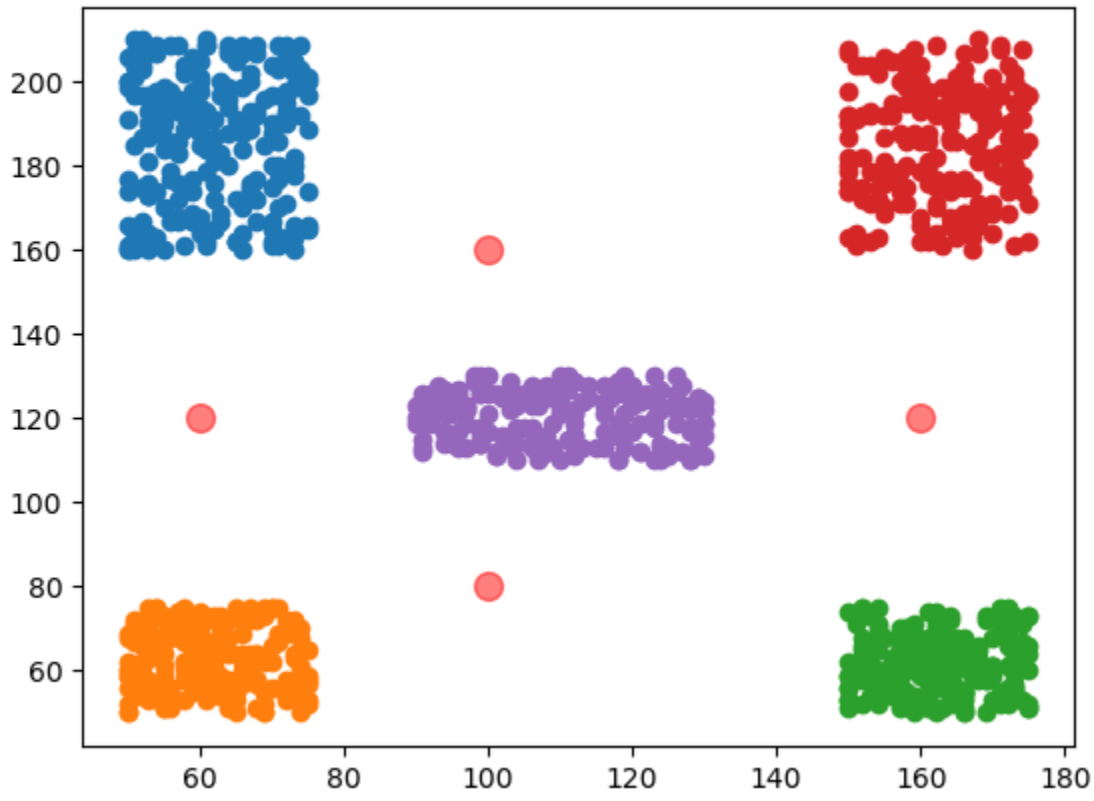
HOMEWORK 1

Öğrencinin Adı ve Soyadı: Samet GÜLMEZ

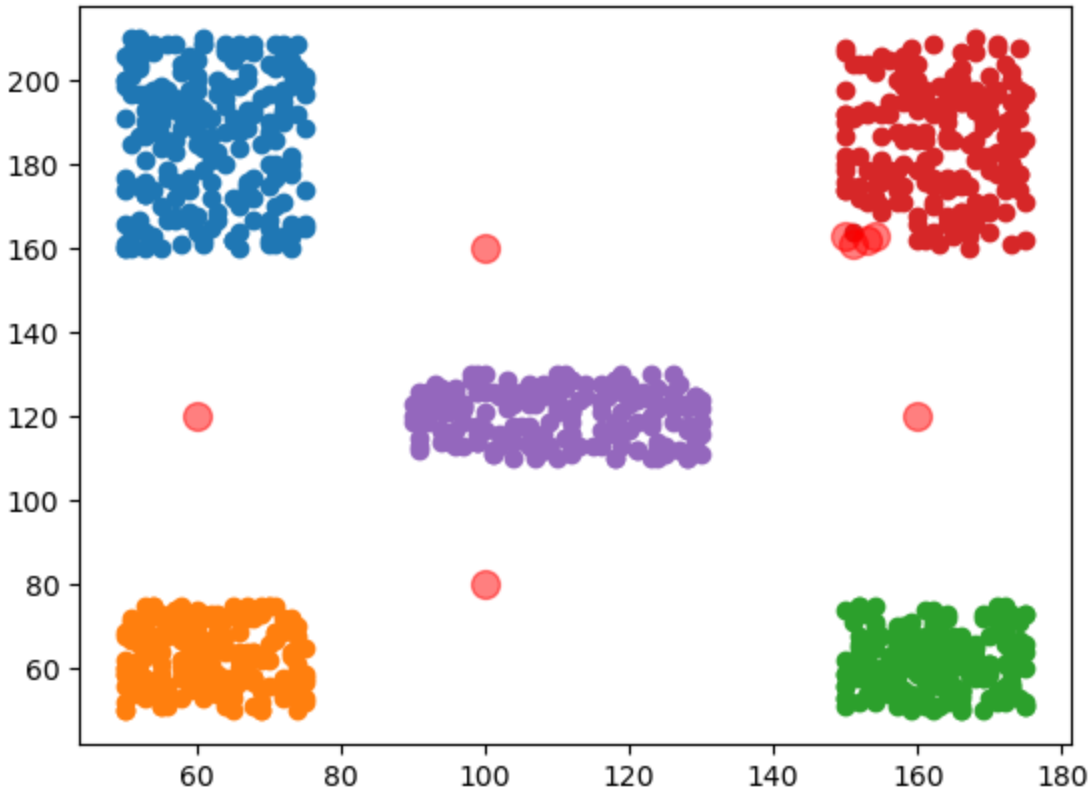
Öğrencinin Numarası : 161044110

1- Showing extracted clusters for at least 3 values of each parameter.

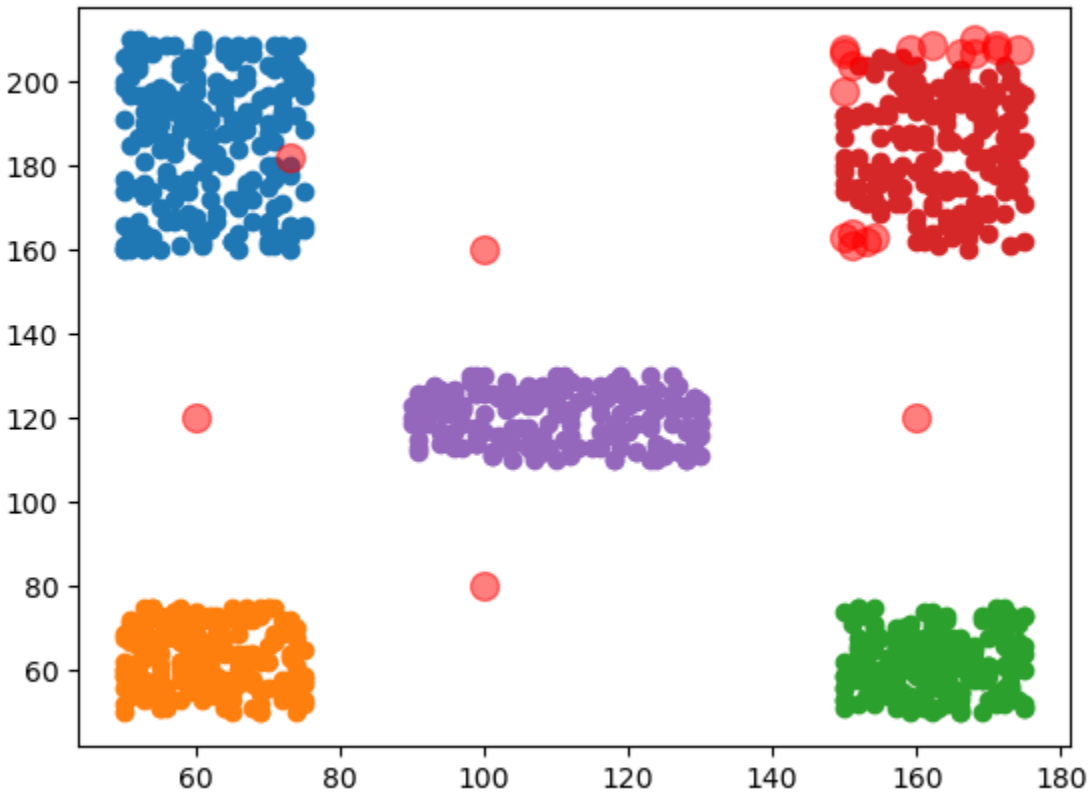
MinPoints = 2 Radius = 5



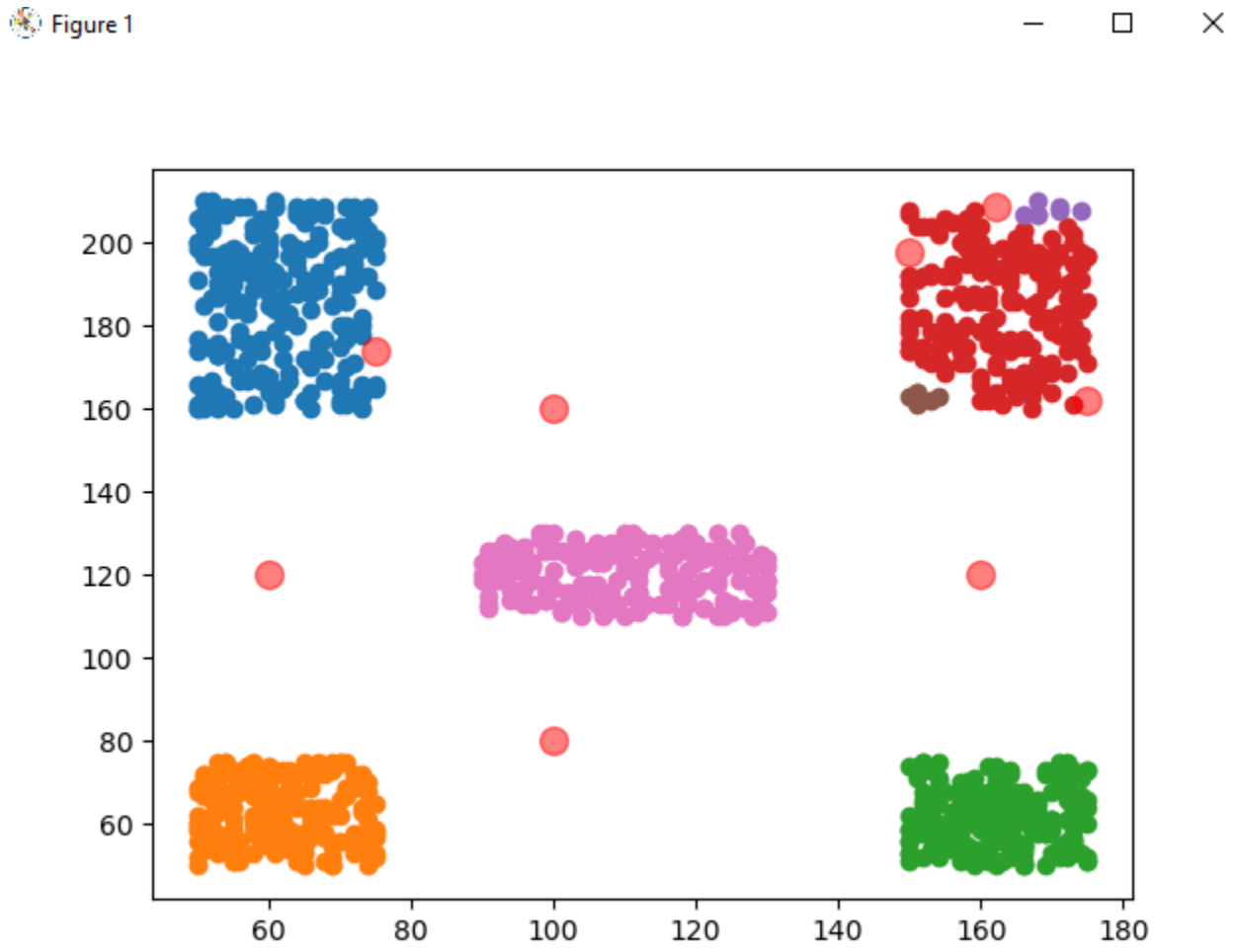
MinPoints = 8 Radius = 5



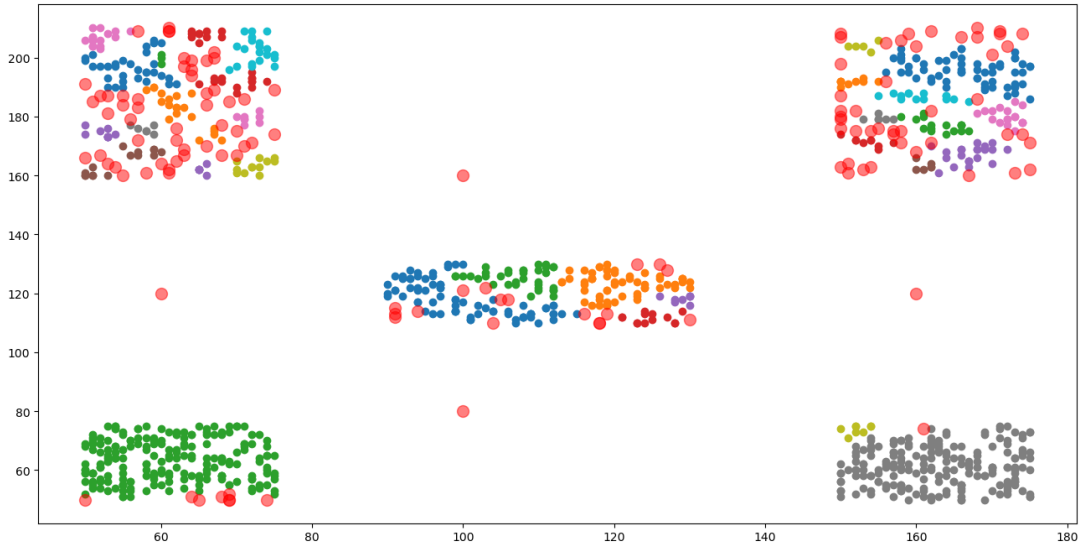
MinPoints = 15 Radius = 5



MinPoints = 4 Radius = 3



MinPoints = 4 Radius = 2



2- Discussion about how the parameters effect the results.

In second example where MinPoints = 2 Radius = 4, my clusters are being created correctly. When I increase MinPoints, some points in my clusters appear to be outside the cluster. The number of points required for a point to form a set has increased.Therefore, it cannot find enough points for the cluster and some points are outside of the cluster.

In the 3rd example, when minPoints are increased more, it is observed that the points outside the set increase even more. It must have more close points to form a cluster. Unable to meet this, more points are out of the cluster.

In example 4, my MinPoints remained constant and the radius was decreased. When the radius is reduced, the points need to be closer to form clusters. When the radius value is large, the points in the same set have started to form

different clusters because they can no longer meet the required distance. and some points were left out of the cluster because they did not have the required distance for the cluster.

In my 5th example, my radius value has been further reduced. After I decreased my value further, more different clusters started to emerge as the distance distance decreased. As the radius got smaller, different clusters began to form.

3- Give a technique to automatically decide on the parameters of DB-Dcan?

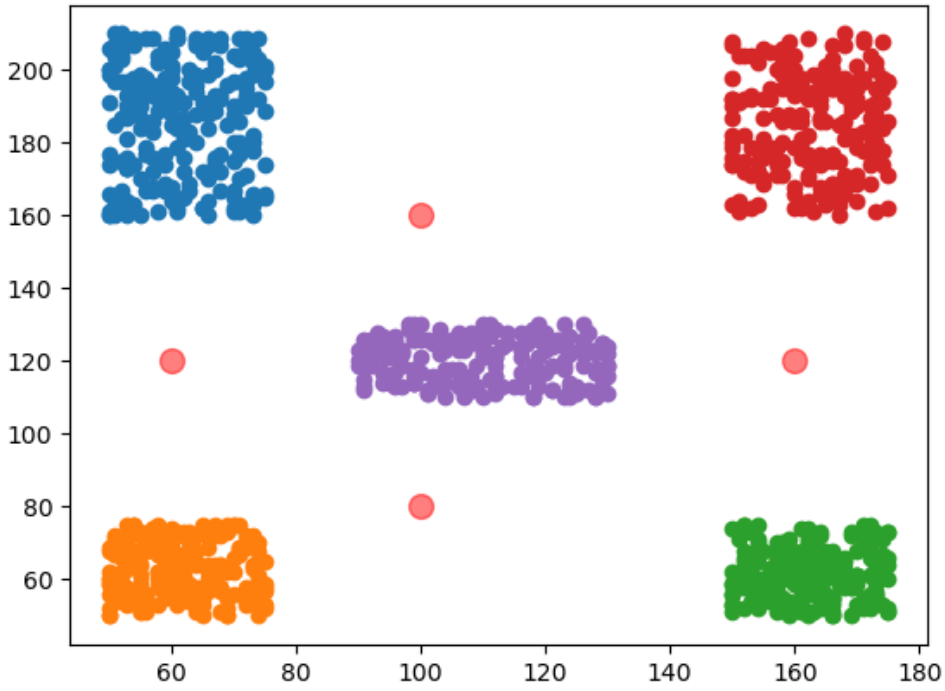
Points can form clusters with points close to them. If we give the best radius value for these points, the clustering will be more successful. At this point, an automatic radius value is

generated for the best clustering. I developed an application like this for this. A number is determined according to the number of points on the chart.

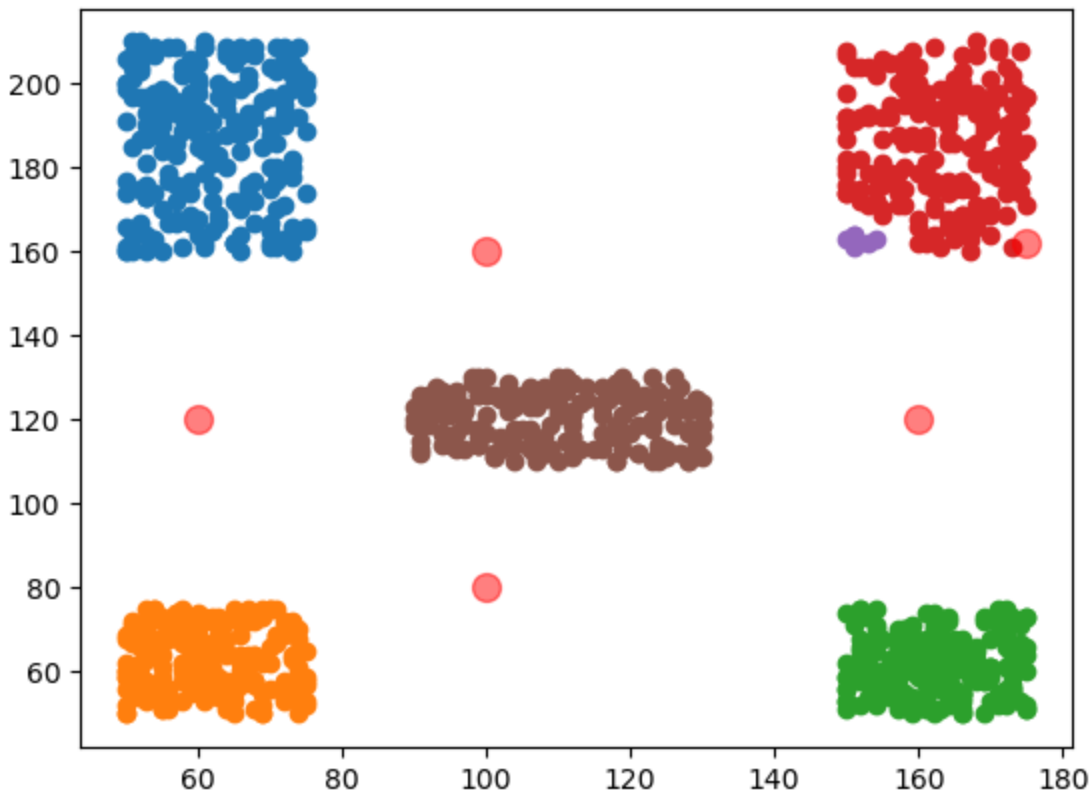
This number is the number of points closest to each point itself. We find the points closest to each point and then we get an average of these points. Thus, we find the average radius value we need to form a cluster from the given point. Then, the general averages of the averages found for each point are taken. As a result, this value we get is the radius value that is required for clustering. The more the total amount of dots in the graph increases, the number of dots closest to a point should increase. Otherwise, we cannot get exact values.

For example :

There is 1021 points. So I find 30 points close to one point and take the average of their distance. Then the end after finding the overall average is 5.15. When I give the radius value 5.15 to my application, the clustering is done properly.



If I find the closest 20 points, my mean will be closer and the cluster will not work properly.
Conclusion 4.17


200
180
160
140
120
100
80
60
60
80
100
120
140
160
180