BE20B020
Krithikaa

BT3041 – Analysis and Interpretation of Biological Data
Assignment 1

DBSCAN algorithm is implemented through a custom function dbscan

Algorithm:

- Each point is initialized as unvisited
- Upon visiting each point, number of points at a radius of epsilon is determined. If the number of points is greater than the minimum points, it is considered as a core point and a cluster is formed.
- Unvisited neighbours of the border points are assessed in a similar fashion and assigned to the cluster (if they have eps-neighbours greater than minimum point)
- The neighbours are determined from the distance matrix calculated using scipy.

Final values of epsilon and minimum points are determined by running the algorithm iteratively until the desired clusters are obtained.

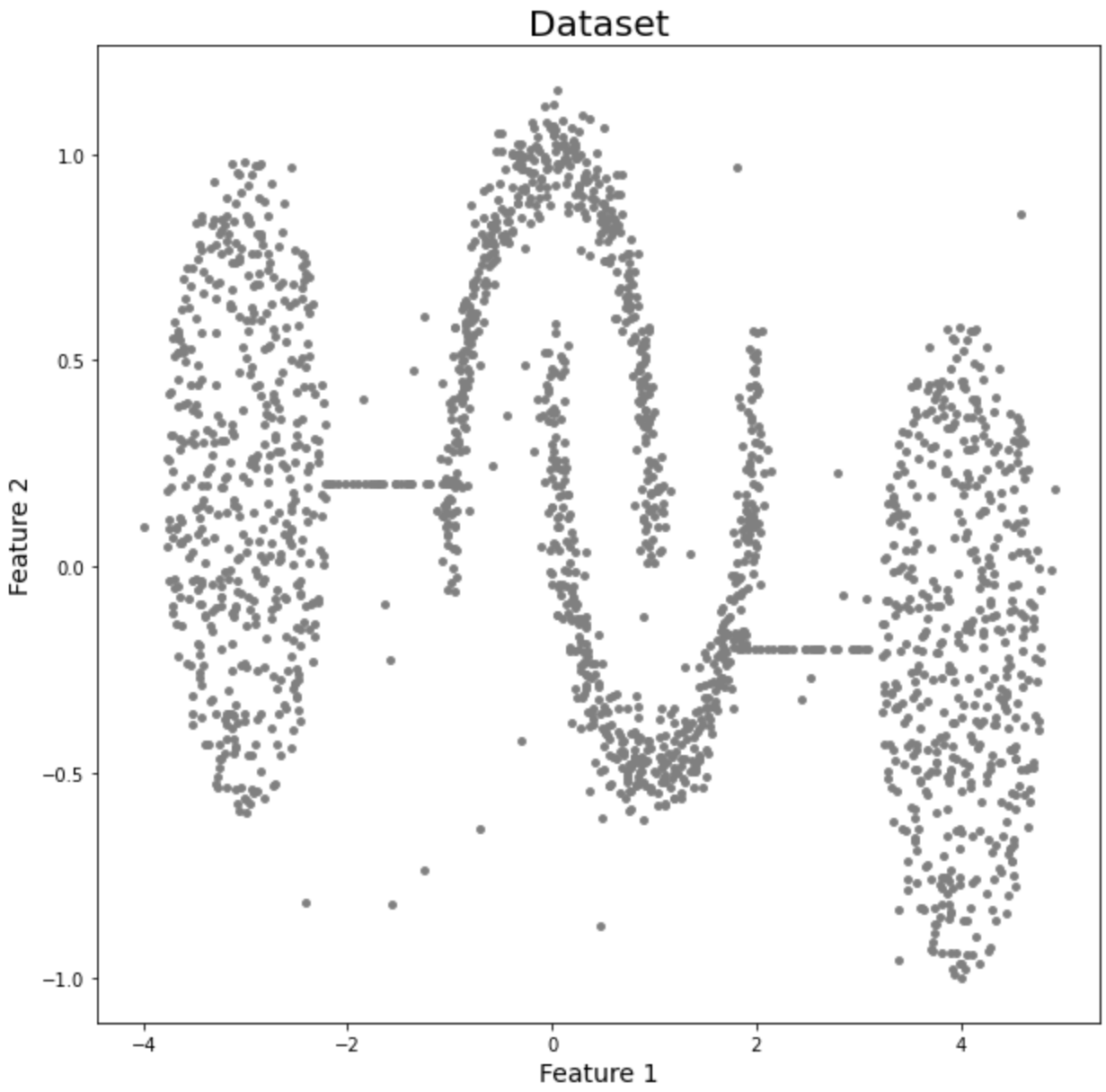
The clusters are plotted using the scatter function of matplotlib

1.

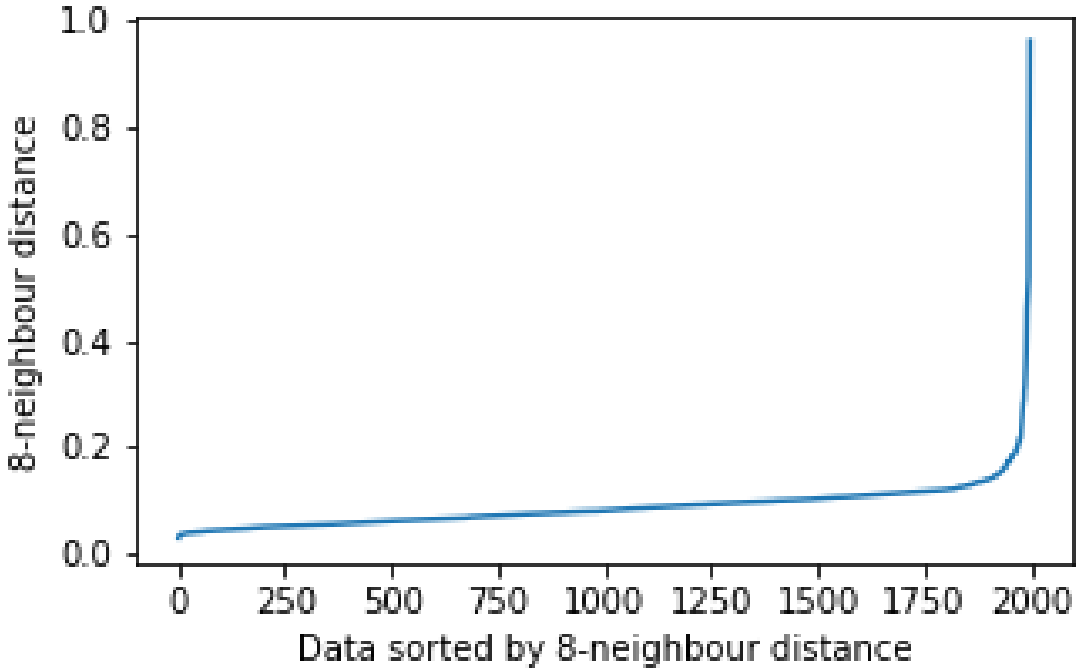
epsilon = 0.15
minimum points = 13

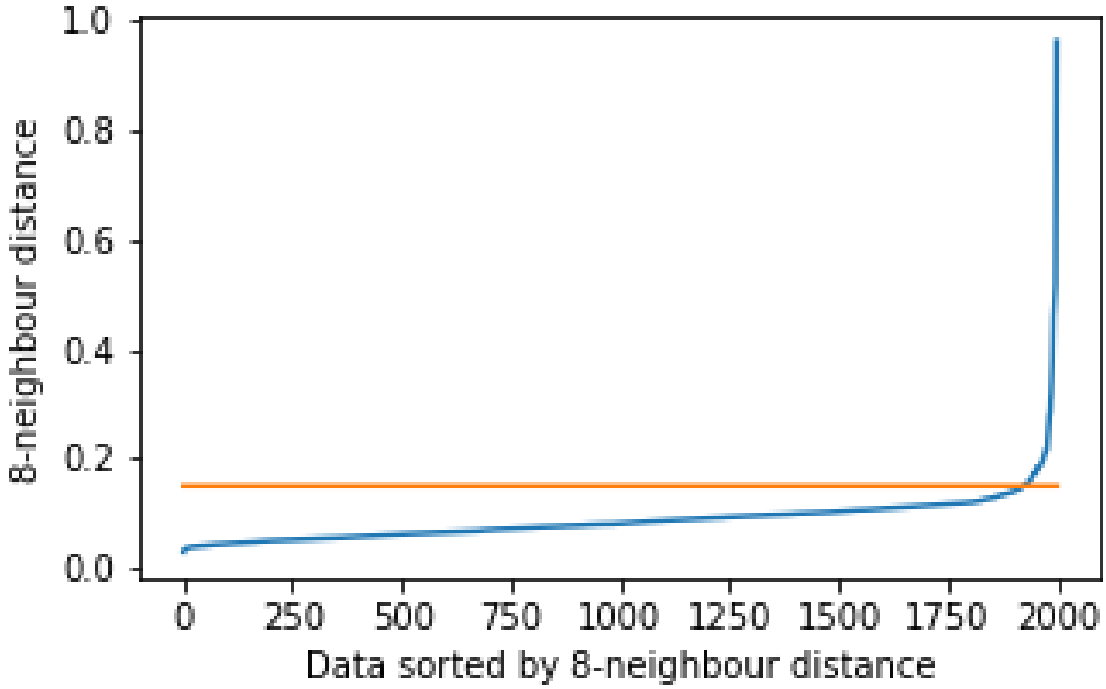
*Number of Clusters obtained: **4***

Given .mat file was imported using scipy.io module.
Given data is first analysed to determine if there are any null values.



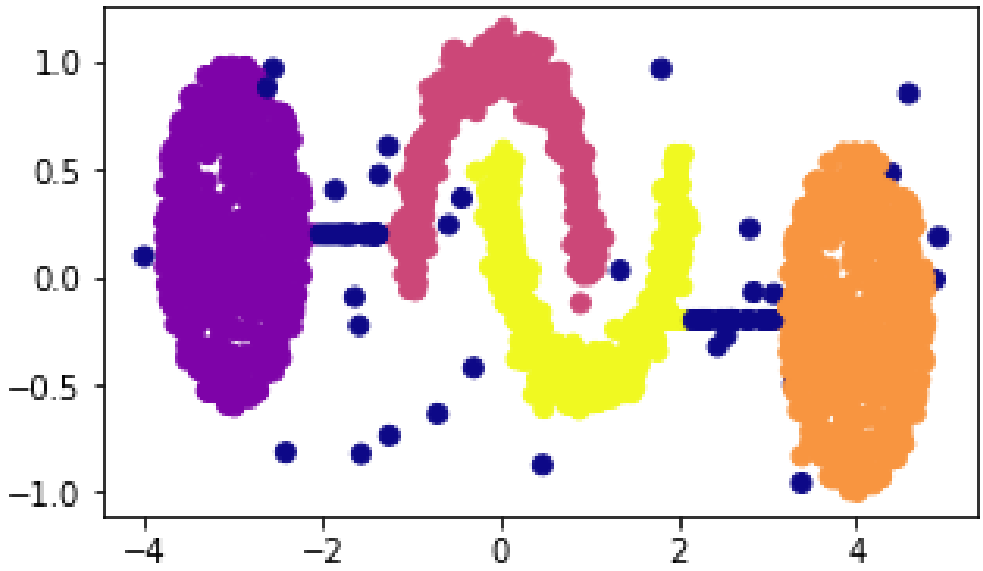
The values of epsilon and minimum points can be determined by making an initial estimate. K-th neighours are found using the package sklearn. Distance of the k-th neighbour is plotted in a sorted fashion. From the graph, the epsilon can be estimated as 0.15 for minimum points 8.





The dbscan algorithm is implemented (from scratch) using epsilon as 0.15 and minimum points as 8, intially. Then through iteration, it was found minimum points 13 give the optimal clusters for epsilon 0.15

It can be observed that there are four clusters for epsilon 0.15 and minimum points 13. The outlier points are colored in dark blue.
It can be noted that multiple combinations are possible for getting four clusters, some of these combinations include the horizontal set of points as border points (unlike what is observed at eps=0.15 and minPts=13 in the graph below).

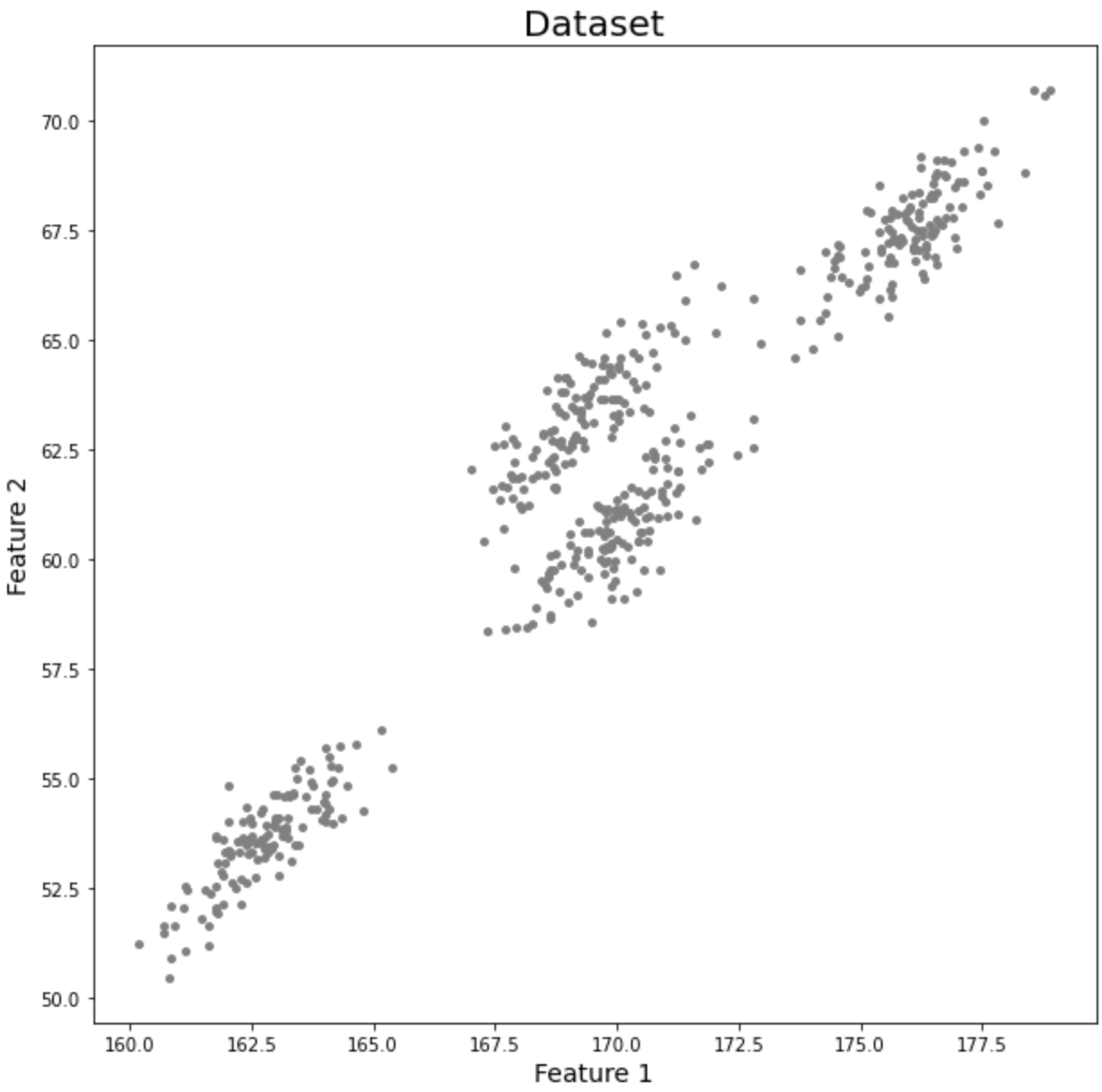


2.

epsilon = 0.5
minimum points = 10

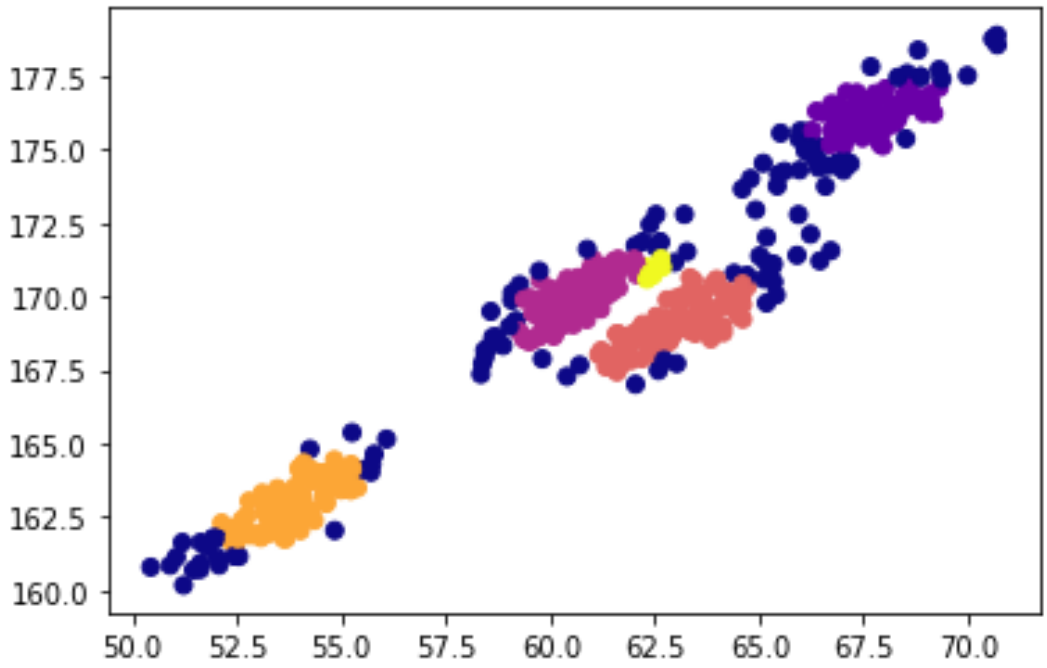
Number of Clusters obtained: ***5***

Given .csv file was imported using read_csv function of pandas.
Given data is first analysed to determine if there are any null values.



The dbscan algorithm is implemented (from scratch) using epsilon as 0.5 and minimum points as 10.

It can be observed that there are five clusters. The outlier points are colored in dark blue.



Upon on studying the dataset, it can be seen that there are four clusters. This indicates that the chosen parameters arent the ideal values. This can be further verified by plotting the 10th-neighbour distance of each point. It is evident that 0.5 distance doesnt intersect the graph at the elbow point.