Assignment 1

BT3041

Yogesh Vijay Deokar EE17B006

1 Algorithm for DBSCAN

1. For each point in the dataset, its distance to all the points were calculated and the points within an *epsilon* distance were classified as neighbouring points for that particular point. The neighbouring points for all points in dataset were stored in an array $all\_neighbours$.

2. Now for each point, if it had number of neighbours $\geq minimum\_points$ then it will be classified as core point. In this way all core points are found out.

3. For the remaining points, if any of its neighbours is a core point, then it will be classified as border points. In this way, all the border points were obtained.

4. Rest all the points were classified as noise points.

5. Now we start with a core point A and label it as $cluster_1$. All its neighbours are also labelled as $cluster_1$.

6. We create an empty list $traverse\_list$. This list will be used to store nodes which we have to travel to. If no neighbours of core point A are core points then this will complete the cluster. But if core point A has neighbours who are also core, these will be stored in $traverse_list$.

7. Now we move to the first point in $traverse_list$ and classify all its neighbours as $cluster_1$. If any neighbours are core points, it will be put in $traverse\_list$. Now the first will be removed from traverse list and we

will move to new first point in *traverse_list* and same procedure will be followed.

8. Step 7 continues till the *traverse_list* is empty. Once the traverse list becomes empty, $cluster_1$ is completed.

9. Now we move to the next core point that is not yet classified into any cluster and repeat the same procedure from step 5 on wards with the only difference that points will now be labelled as *cluster_2*.

10. This will continue till all core points are classified into clusters.

2 Question 1

2.1 Plot of Given dataset

Figure 1 shows the plot of given dataset for question 1.

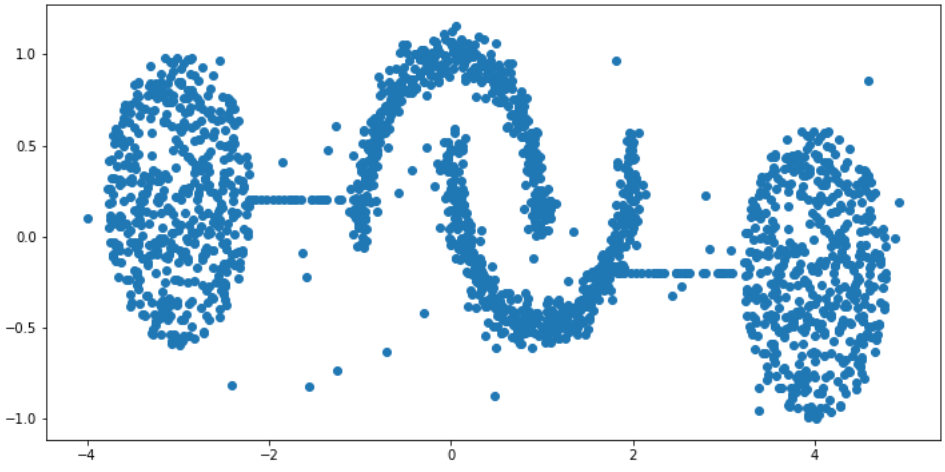

Figure 1: Given Data for Q1

2.2 Classifying points as core, border and noise

Figure 2 shows the scatter plot for points classified as core, border and noise points. The points in green colour are core points (denoted by number -2 in legend). The points in orange are border points (denoted by number -1 in legend). The points in blue are noise points(denoted by number 0 in legend).

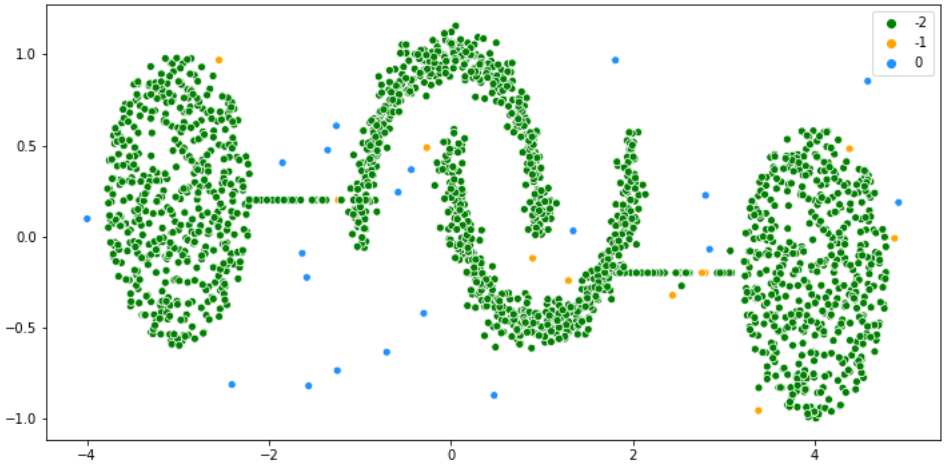

Figure 2: Note : Core = -2, Border = -1, Noise = 0

2.3 Clustering the given dataset

Figure 3 shows the clustered plot for Question 1 dataset. In the figure the noise points are denoted by green colour (number 0 in legend) and the clusters are denoted by blue, brown, black and orange colors (denote by numbers 1,2,3 and 4 in legend).
The parameters used to obtain 4 clusters are EPSILON = 0.15 and MINIMUM NUMBER OF POINTS = 5.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Noise
Number of Points	528	457	515	480	20

Table 1: Table representing number of points in each cluster

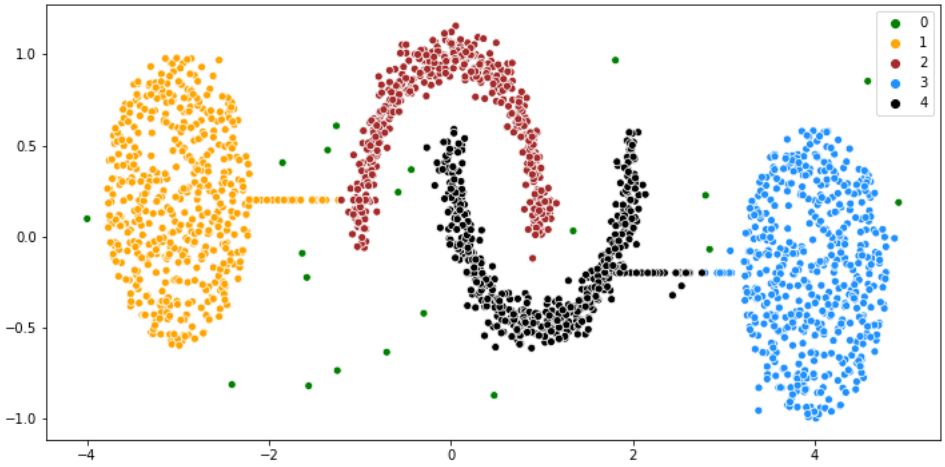

Figure 3: Note: 0 = noise and 1,2,3,4 are clusters

2.4 Conclusion for Question 1

As we see in figure 3, we have obtained 4 clusters using paramaters **EPSILON = 0.15 and MINIMUM NUMBER OF POINTS = 5.**

3 Question 2 using paramaters EPSILON =0.5 and MINIMUM POINTS = 10

3.1 Plot of Given dataset

Figure 4 shows the plot of given dataset for question 2.

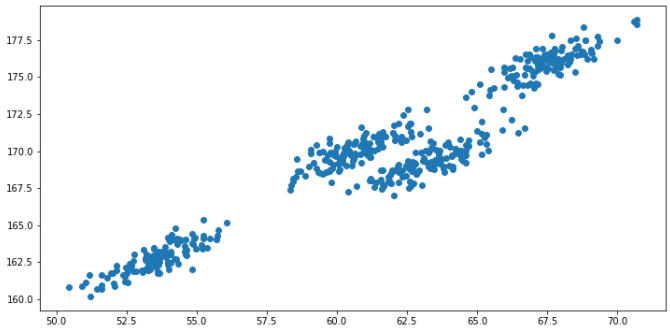

Figure 4: Given Data for Q2

3.2 Classifying points as core, border and noise

Figure 5 shows the scatter plot for points classified as core, border and noise points. The points in green colour are core points (denoted by number -2 in legend). The points in orange are border points (denoted by number -1 in legend). The points in blue are noise points(denoted by number 0 in legend).

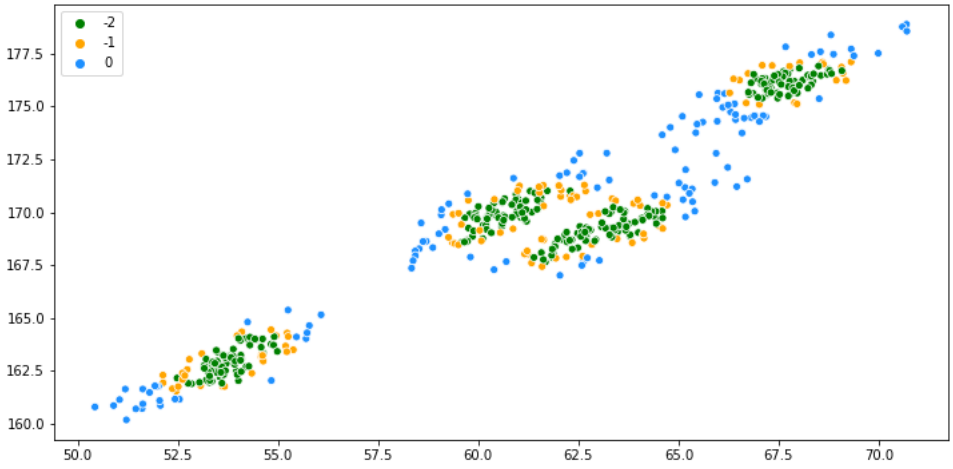

Figure 5: Note : Core = -2, Border = -1, Noise = 0

3.3 Clustering the given dataset

Figure 6 shows the clustered plot for the Question 2 dataset.

	C1	C2	C3	C4	C5	Noise
Number of Points	85	91	102	100	7	115

Table 2: Table representing number of points in each cluster

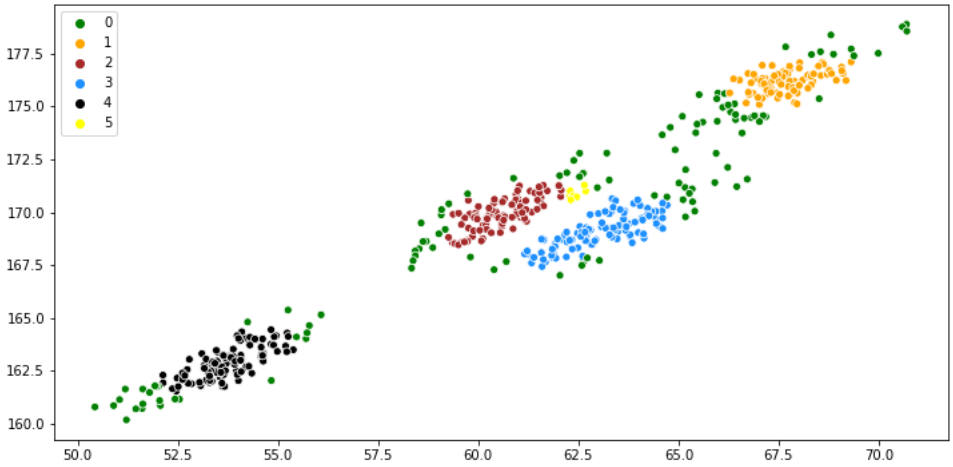

Figure 6: Note: 0 = noise and 1,2,3,4 and 5 are clusters

4 Question 2 Using paramaters EPSILON=0.63 and MINIMUM POINTS=4

4.1 Classifying points as core, border and noise

Figure 7 shows the scatter plot for points classified as core, border and noise points.

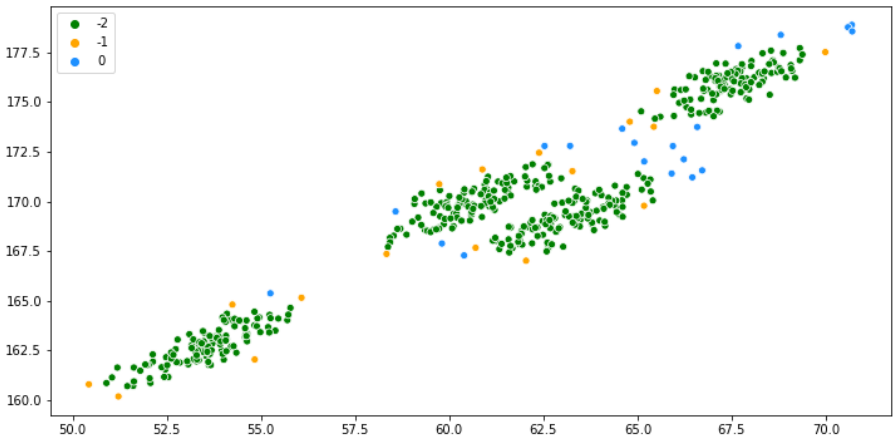

Figure 7: Note : Core = -2, Border = -1, Noise = 0

4.2 Clustering the given dataset

Figure 6 shows the clustered plot for the Question 2 dataset.

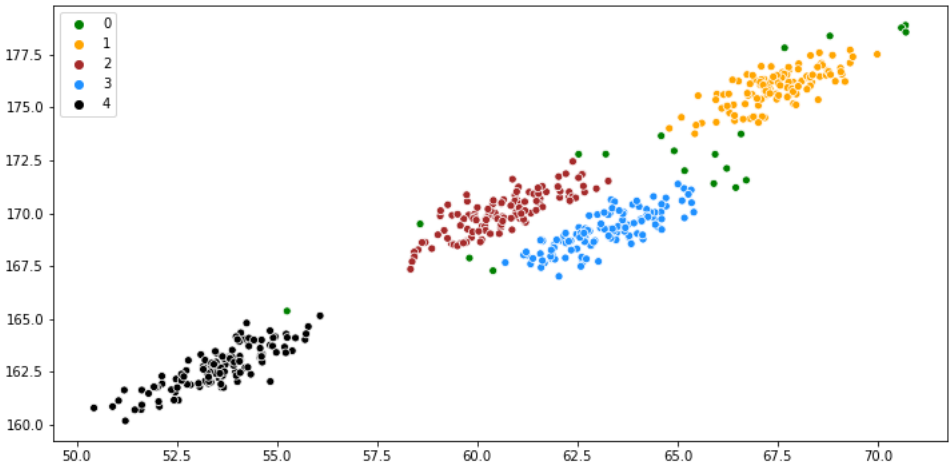

Figure 8: Note: 0 = noise and 1,2,3,4 are clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Noise
Number of Points	117	121	118	124	20

Table 3: Table representing number of points in each cluster

4.3 Conclusion for Question 2

As we can see from above figures paramaters **EPSILON =0.63, MINIMUM POINTS = 4** gave a better clustering than **EPSILON =0.5, MINIMUM POINTS = 10**. This is because there were not many points who had 10 neighbours in radius of 0.5. Therefore we see low number of noise points (Table 3) for the former case. We observe the area corresponding to 5th cluster in Figure 6. Observing this same exact area in Figure 5, we notice that there is one core point in green surrounded by all the border points. This means that there were no core points in between to connect this point to the 2nd cluster. Hence we get the fifth cluster. But no such problem arises in second case, as minimum number of points required to classify as core point is only 5 (not as high as 10 in the initial case).