VIPIN KUMAR SETH
111701030

For the given dataset I used the DBSCAN to find the outliers because of the following advantages:

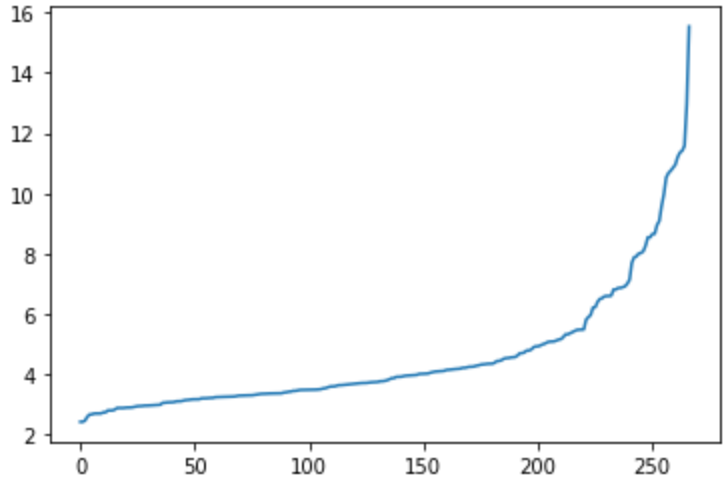
- In spectral clustering, we have to specify the number of clusters, but in this case, we don't know how many clusters we want. We just want to detect outliers.
- An outliner can be defined as the point which is very less dense, and DBSCAN works on the property of density on points. So it's natural to use DBSCAN.
- Further DBSCAN and easily detect the outliers.

So, moving with DBSCAN
I normalized the data as we are using Euclidean distance in DBSCAN.
Now it's very important to choose two very important parameters: epsilon and min_points.

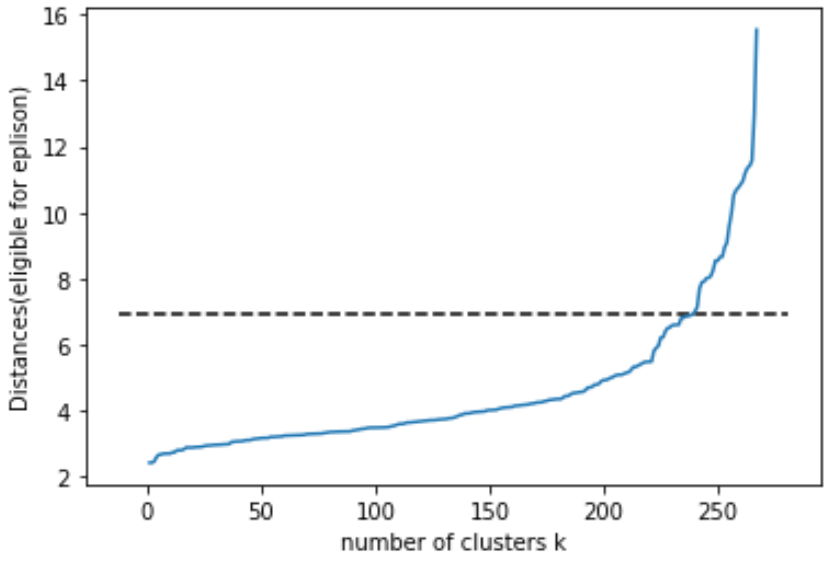
- *MinPts*: As a rule of thumb, a minimum *minPts* can be derived from the number of dimensions *D* in the data set, as *minPts* ≥ *D* + 1. As a rule of thumb, *minPts* = 2·*dim* can be used, but it may be necessary to choose larger values for very large data, for noisy data or for data that contains many duplicates.
- ε: The value for ε can then be chosen by using a k-distance graph, plotting the distance to the *k* = *minPts*-1 nearest neighbor ordered from the largest to the smallest value. Good values of ε are where this plot shows an "elbow". if ε is chosen much too small, a large part of the data will not be clustered; whereas for a too high value of ε, clusters will merge and the majority of objects will be in the same cluster. In general, small values of ε are preferable, and as a rule of thumb, only a small fraction of points should be within this distance of each other.

  Ref : (https://en.wikipedia.org/wiki/DBSCAN) (Parameter estimation)

Hence choosing the min_points = 2*(Dimension of features = 44) = 88
Now for choosing ε as stated above I am using k-distance graph, plotting the distances



Next is to find the elbow point. An elbow is a point with the maximum absolute second derivative. Here I am using a python library kneed.KneeLocator. Hence the elbow point is at 239



Optimal value for epilson = 6.91856275687389

Now performing the DBSCAN with the parameters estimated.

After fitting the labels obtained =

```
[ 0 -1  0  0 -1  0  0  0  0  0  0  0  0  0 -1 -1  0  0 -1  0  0  0  0  0
 -1 -1  0  0  0  0  0 -1 -1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0 -1  0  0 -1  0 -1 -1  0  0  0  0 -1  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1 -1  0  0  0  0
  0  0  0  0 -1  0  0  0  0 -1  0  0  0  0  0  0  0  0  0 -1  0  0 -1
  0  0 -1  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0
  0 -1  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0 -1  0 -1  0  0  0 -1  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0 -1  0
  0  0  0  0 -1  0  0  0 -1 -1  0  0 -1  0 -1  0  0  0  0  0  0  0  0  0
  0  0 -1]
```

Here in the label -1 represent the outliers.

Hence the total number of outliers = 36.