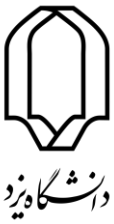
In the name of Allah



Computer Engineering Faculty of Yazd University
Artificial Intelligence

Pattern Recognition Course

Homework 1
(Clustering)

Instructor: Dr. Yazdian

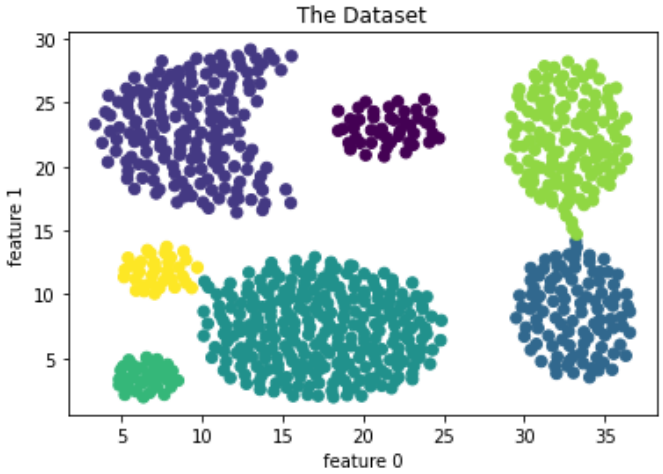
Shiva Zeymaran

Student ID: 40109434

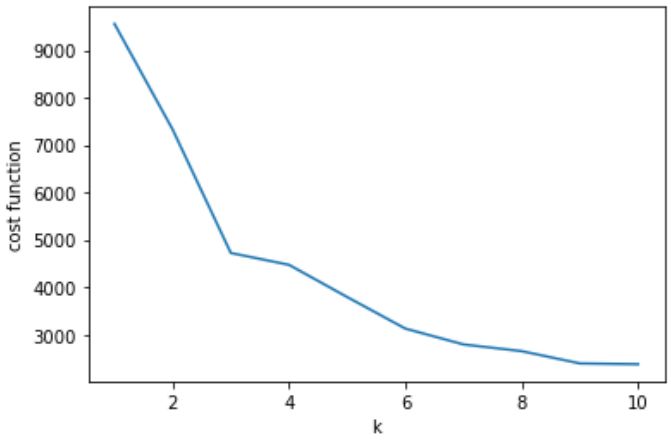
Winter 2023

Part A – Distance-Based (k-means)

1. Visualizing the dataset:

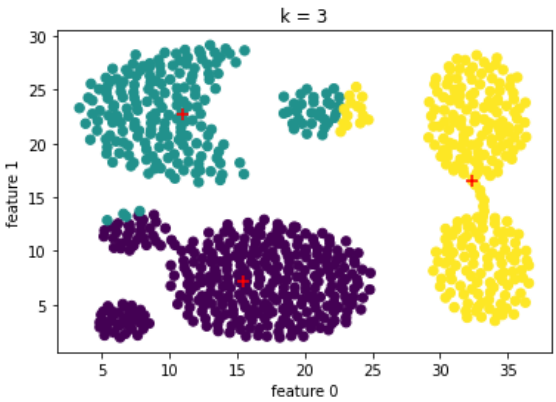


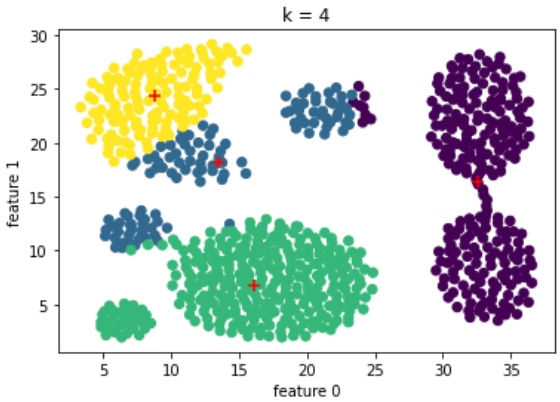
2. Plot the cost function values by changing values of parameter k in k-means clustering:

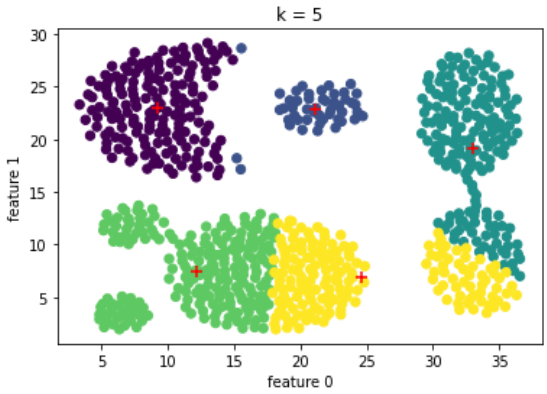


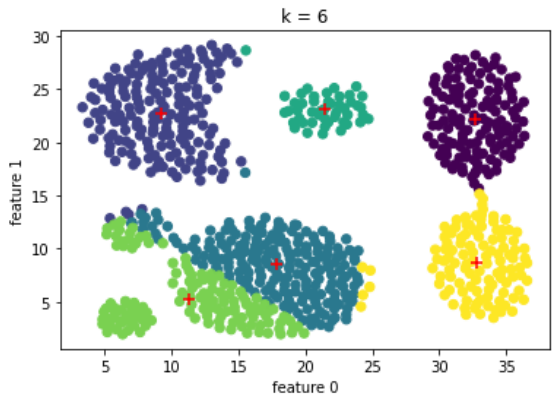
3. The optimal k value is perhaps 6. According to the Elbow Method, by changing k value until 6, the cost function dramatically decreases but after that by increasing the value of k, there is not a big change in the value of cost function. However, this plot can change by running the code again (because of random initial centroids) but we can say that k=6 in average is the better value for k.

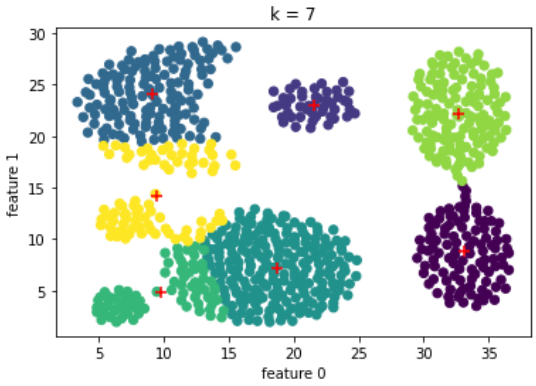
4. We plot the classification result for 5 models with k = 3,4,5,6,7 as follows:







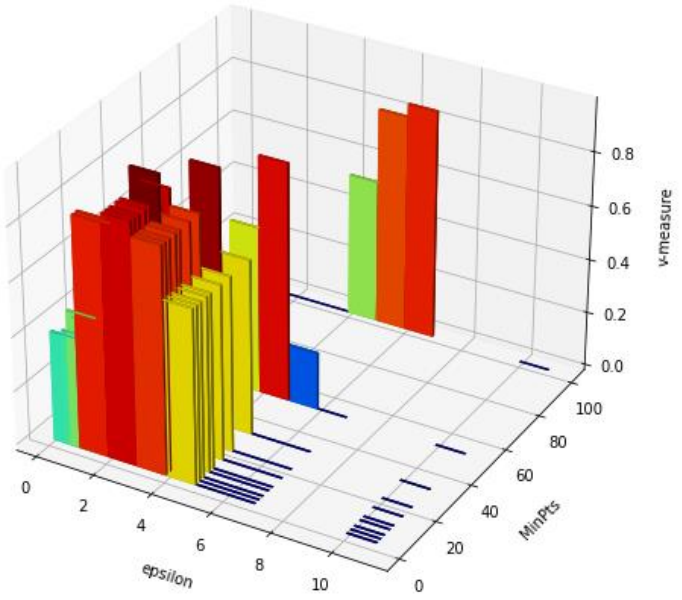




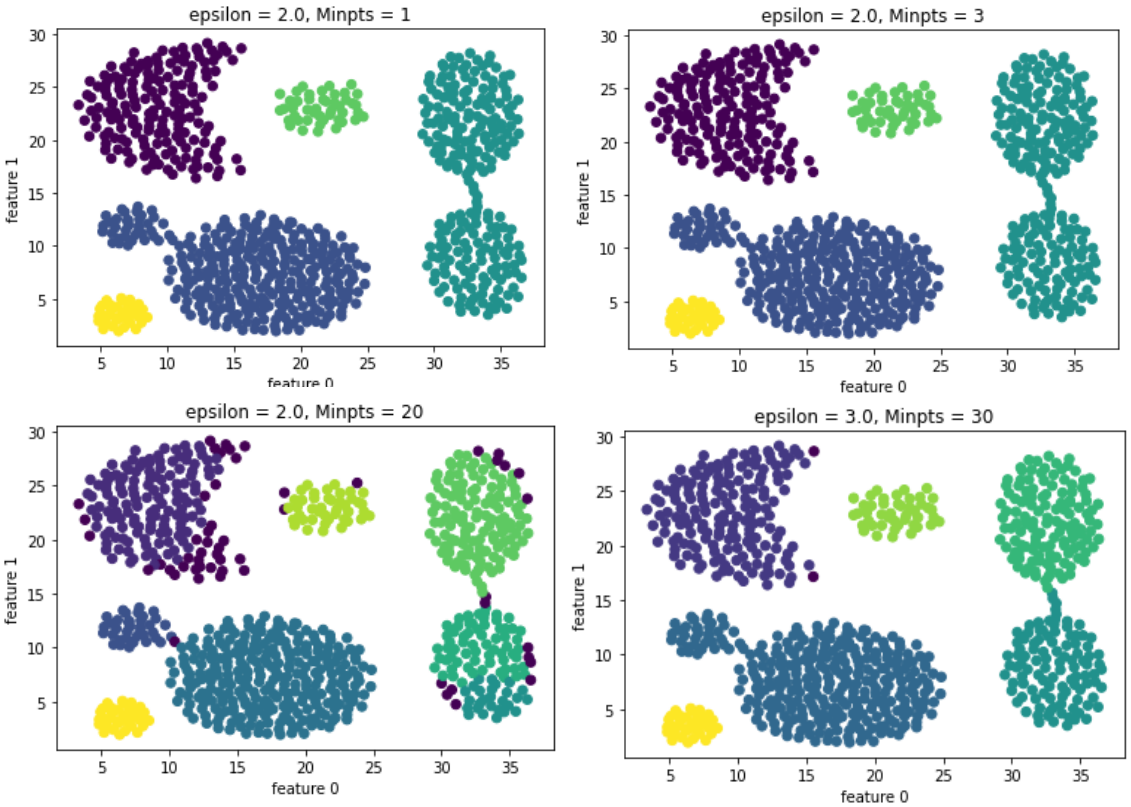
From above, we understand that by increasing the value of k, clusters can be separated more accurate. Then in k=6, clusters are identified properly. After that, in k=7 again we can see some problems in identifying clusters truly.

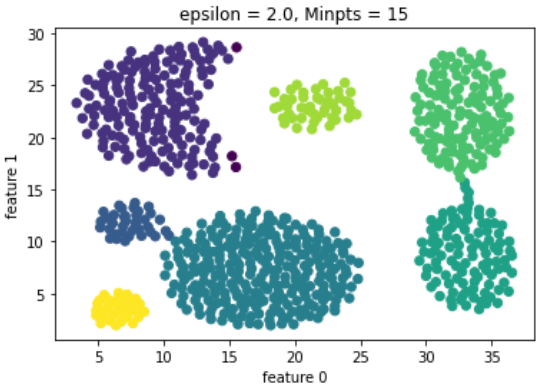
Part B – Density-Based (DBSCAN)

1. Plot a 3D diagrams: MinPts, e, MSE



2. We plot the classification result for 5 models with (epsilon, minpts) = (2,1), (2,3), (2,20), (3,30), (2,15) as follows:



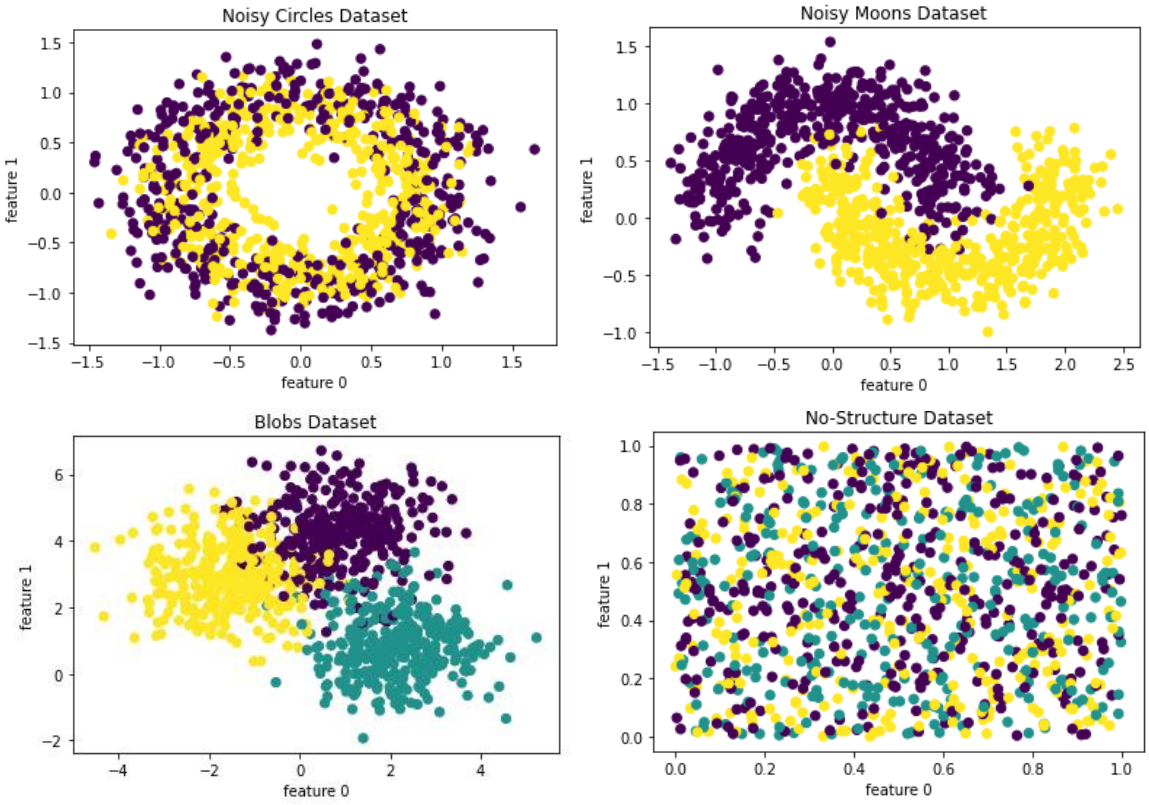


To find the best 5 models, we compare 100 models with their v-measure values. The higher the v-measure value, the better is the model. Therefore, we find the top-5 v-measure values and their corresponding epsilon and minpts.

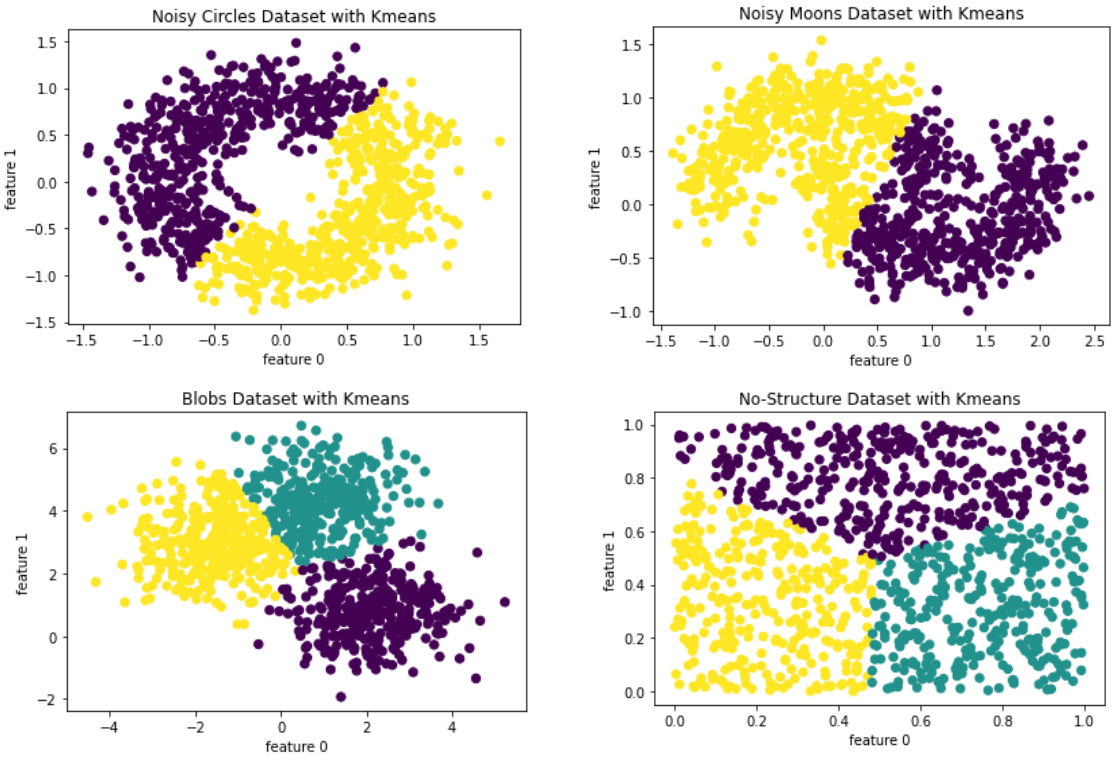
3. From above, we understand that the last model which is epsilon = 2 and minpts = 15 is the best model (highest v-measure) that identified 7 clusters properly.
   In comparison with the best model obtained form k-means, DBSCAN performs better. In k-means, clusters and their borders are not distinguishing properly as in DBSCAN.
   We saw that when k-means performs clustering with k=7 (which is the true number of our groups of data and obtained from DBSCAN), the clusters are mixed and k-means becomes overfitted.

Part C – K-means, DBSCAN and Agglomerative Clustering

1. Visualize each dataset:

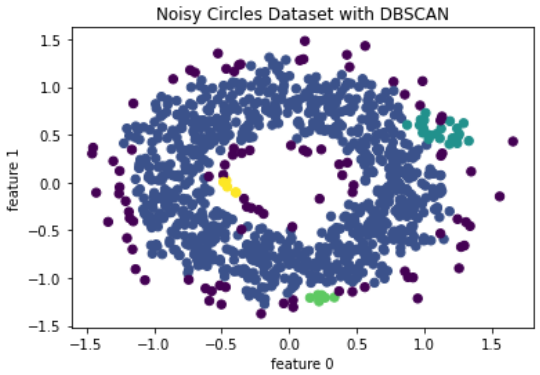


2. Plot the results of training k-means on each dataset:

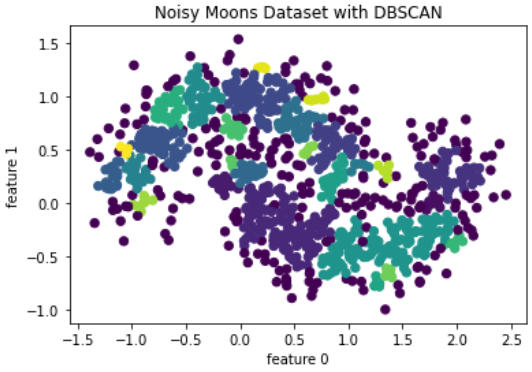


3. Plot the results of training DBSCAN on each dataset (with best epsilon value for each):

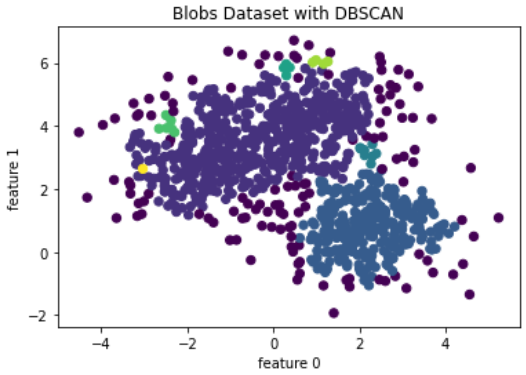
```
best epsilon is: 0.1
```



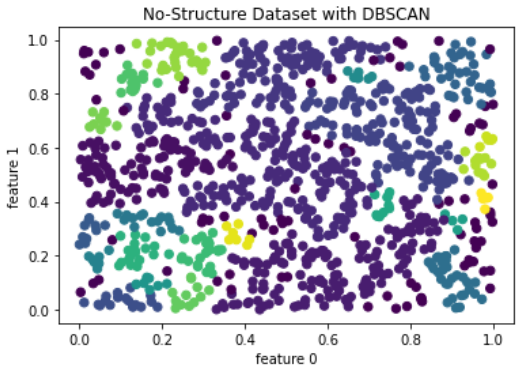

```
best epsilon is: 0.08
```



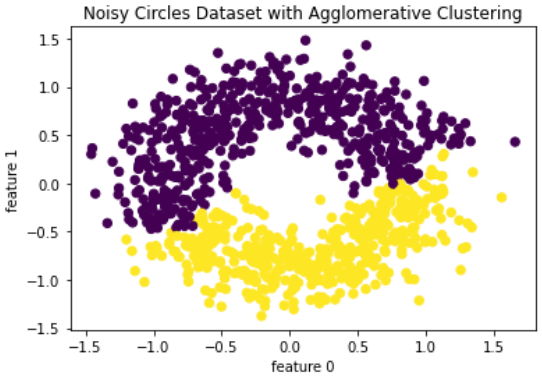

```
best epsilon is: 0.28
```

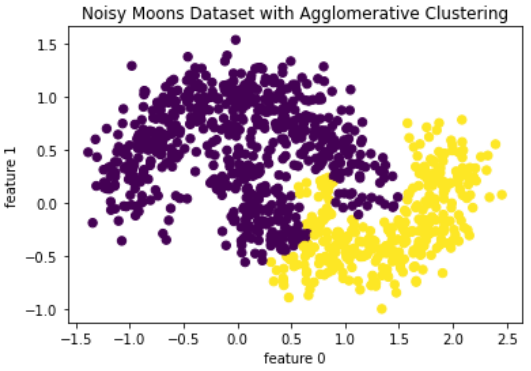


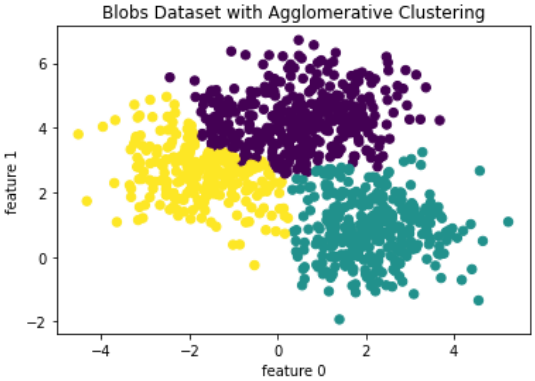

```
best epsilon is: 0.04
```

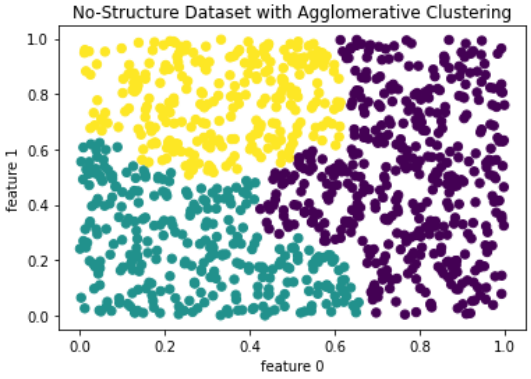


4. Plot the results of training Agglomerative clustering on each dataset:









5. Comparing 3 methods on different datasets:

   K-means: This method does not work well for Noisy Circles and Noisy Moons datasets; since this method can find convex shapes for clusters. However, this works well for Blobs dataset as it is clear from the plot. It also find 3 convex-shape clusters for no-structure dataset.

   DBSCAN: It seems that this method does not work well for any of our datasets (even for their best epsilon values). The reason possibly is that in this method we do not set the number of clusters as a parameter and we see that there are lots of clusters identified in each dataset.

   Agglomerative: This method does not work well for Noisy Circles dataset and can not find clusters inside each other. However, it works good for Noisy Moons and Blobs dataset and finds each of clusters properly. Also, it gives us 3 clusters for no-structure dataset.