Ceng499 Hw2 Report

Giray Keskin

15/05/2021

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

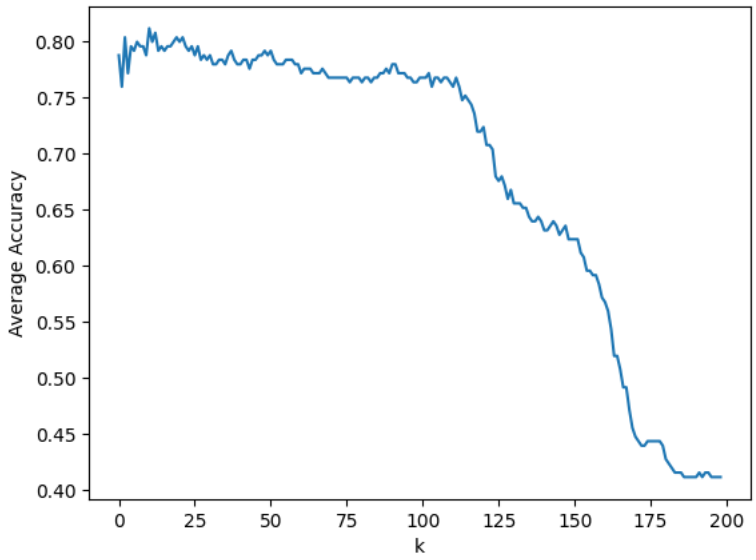

Figure 1: Average Accuracy vs k in K-fold Cross Validation

1.2 Accuracy drops with very large k values

While checking k nearest datas, majority voting fails due to large k.

1.3 Accuracy on test set with the best k

My best accuracy was calculated when k was 10, and the test accuracy is 80%.

2 Part 2: K-means Clustering

2.1 Elbow method

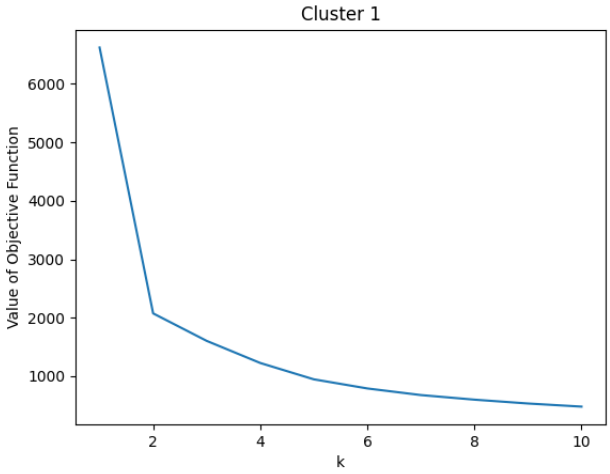

Figure 2: Value of Objective Function vs k in cluster 1 of K-means

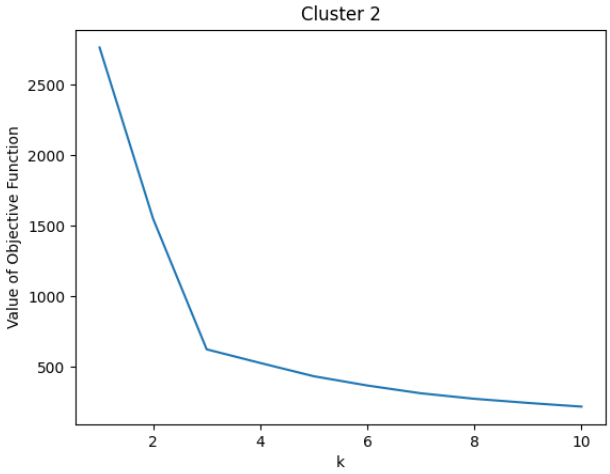

Figure 3: Value of Objective Function vs k in cluster 2 of K-means

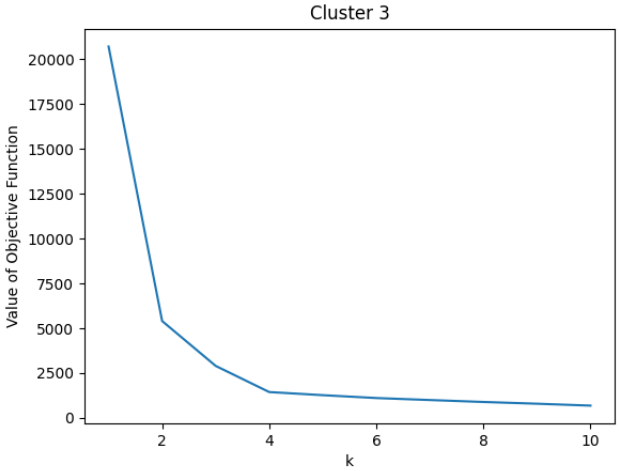

Figure 4: Value of Objective Function vs k in cluster 3 of K-means

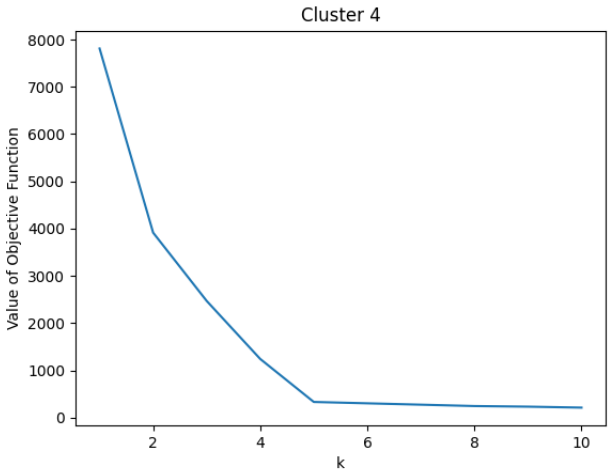

Figure 5: Value of Objective Function vs k in cluster 4 of K-means

For the first cluster, k = 2 is an obvious choice from the elbow method.
For the second cluster, k = 3 is an obvious choice from the elbow method.
For the third cluster, k = 4 is my choice from the elbow method and since we have 4 clusters in the data.
For the fourth cluster, k = 5 is my choice from the elbow method and since we have 5 clusters in the data.

2.2 Resultant Clusters

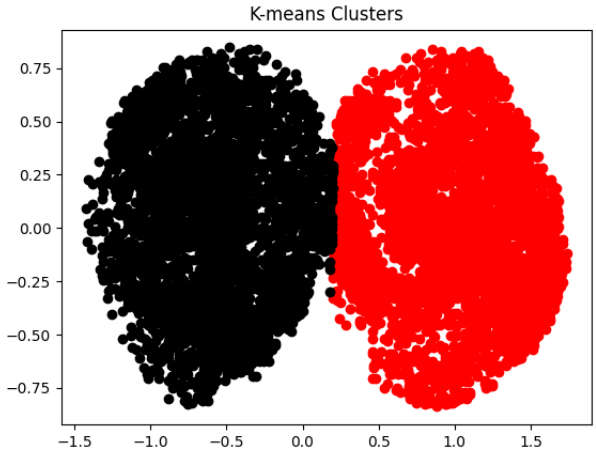

Figure 6: First data's cluster

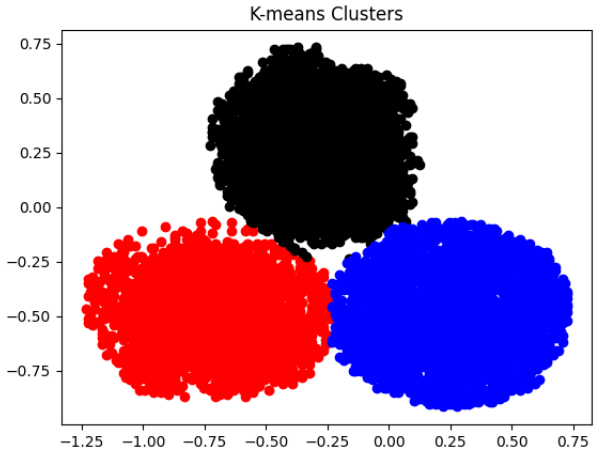

Figure 7: Second data's cluster

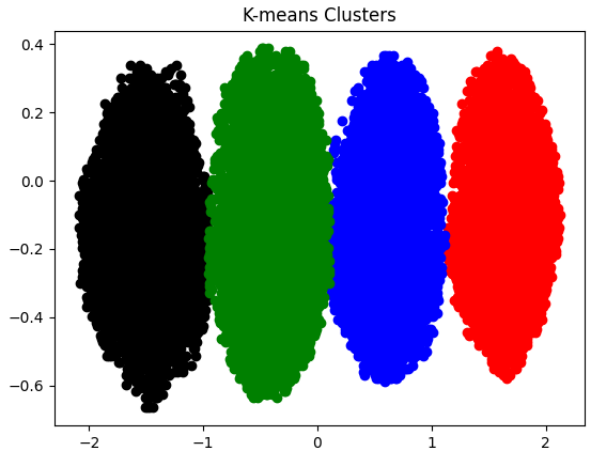

Figure 8: Third data's cluster

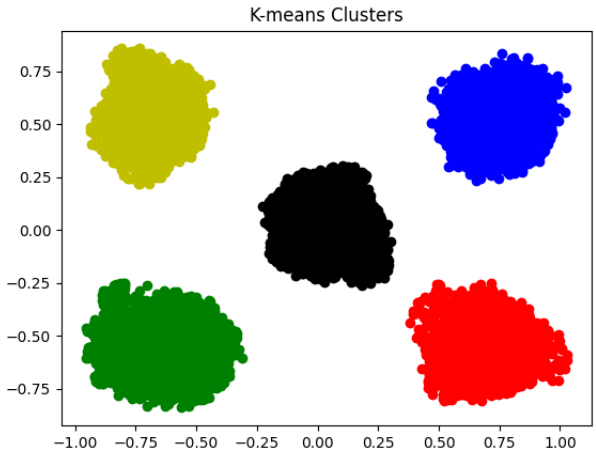

Figure 9: Fourth data's cluster

3 Part 3: Hierarchical Agglomerative Clustering

3.1 data1

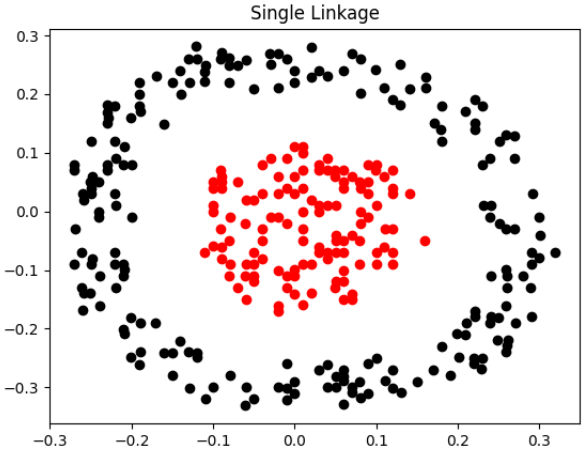

Figure 10: First data's cluster using single linkage

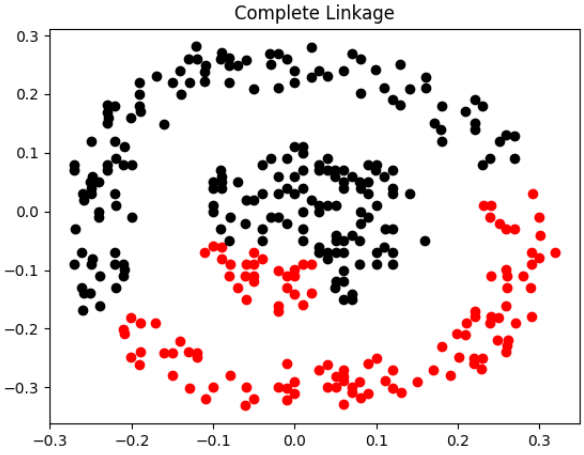

Figure 11: First data's cluster using complete linkage

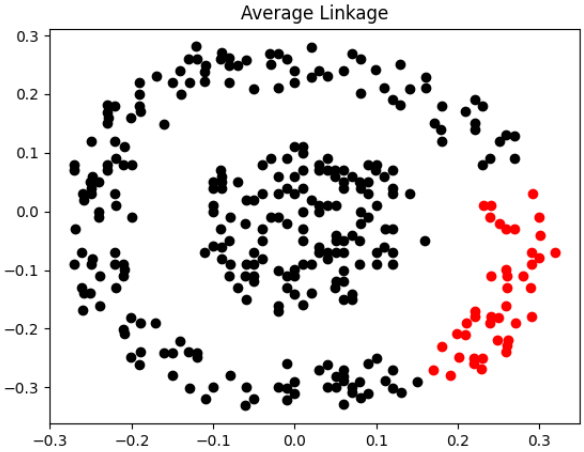

Figure 12: First data's cluster using average linkage

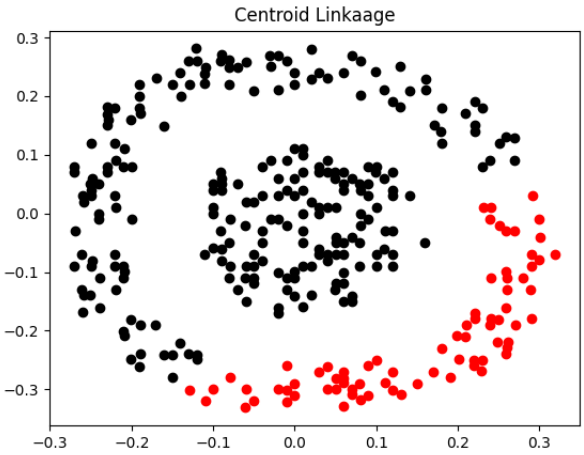

Figure 13: First data's cluster using centroid linkage

Single linkage is the best here since it takes the nearest data to the cluster and the two groups here dont have any close datas.
Complete linkage fails since it takes the farthest data from the cluster to calculate distance.
Average linkage fails since average of the distances are similar at the outer circle.
Centroid linkage fails since centers are very near.

3.2 data2

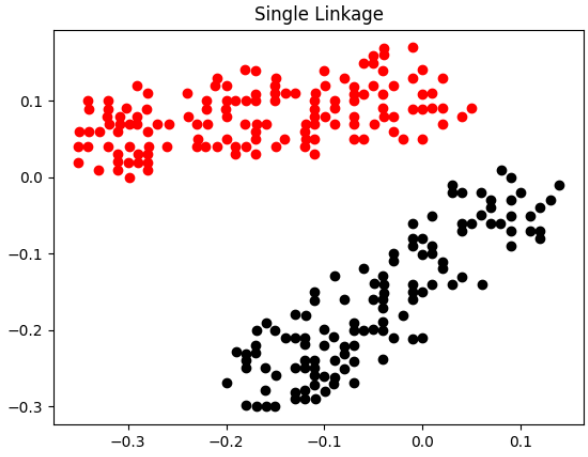

Figure 14: Second data's cluster using single linkage

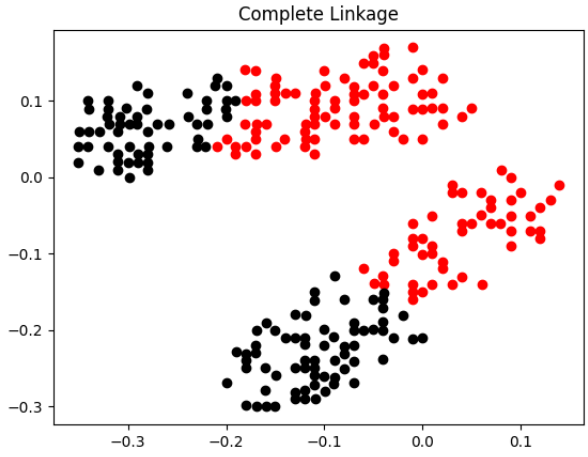

Figure 15: Second data's cluster using complete linkage

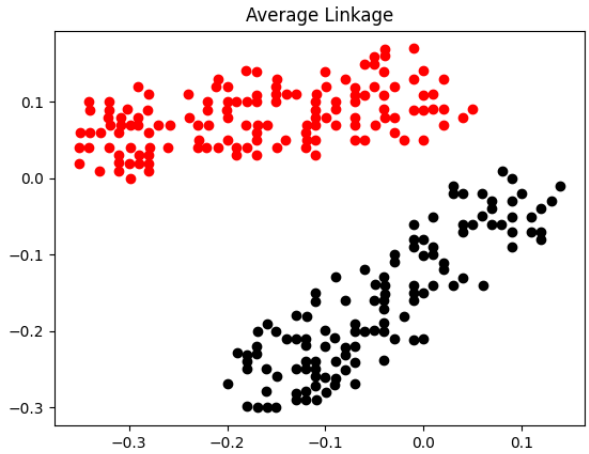

Figure 16: Second data's cluster using average linkage

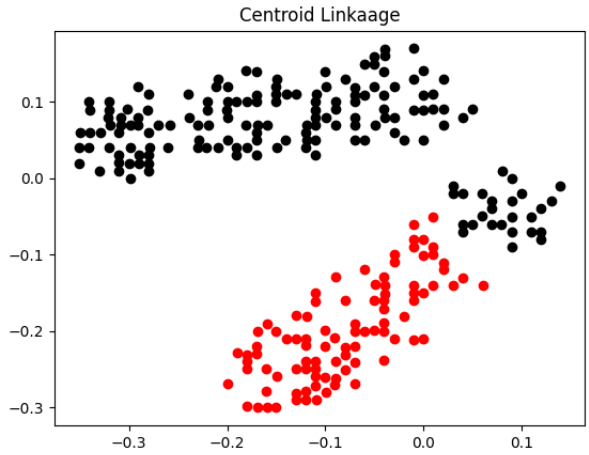

Figure 17: Second data's cluster using centroid linkage

Single linkage is one of the best here since it takes the nearest data to the cluster and the two groups here dont have any close datas.
Complete linkage fails since it takes the farthest data from the cluster to calculate distance.
Average linkage is one of the best here since average distance to the both clusters are very distinct.
Centroid linkage fails since center of the top cluster is near some of the other

cluster's data.

3.3 data3

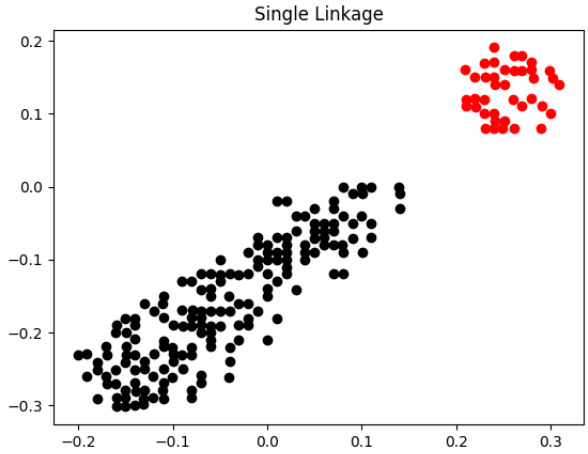

Figure 18: Third data's cluster using single linkage

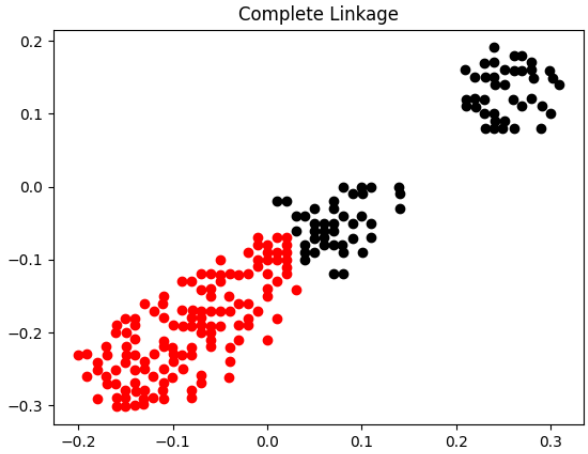

Figure 19: Third data's cluster using complete linkage

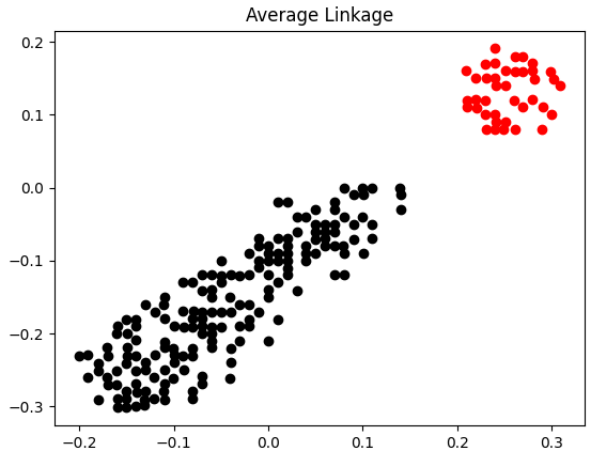

Figure 20: Third data's cluster using average linkage

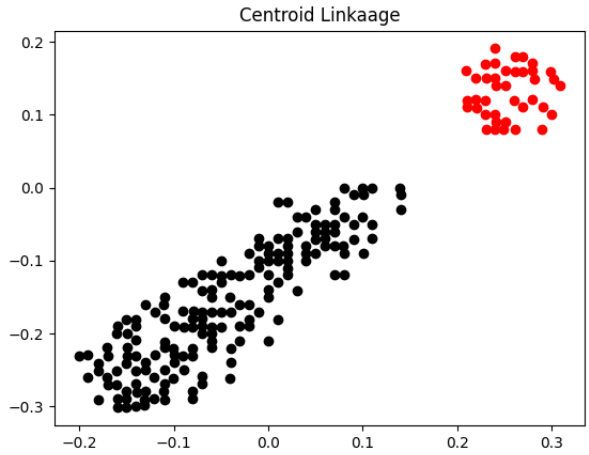

Figure 21: Third data's cluster using centroid linkage

Single linkage is one of the best here since it takes the nearest data to the cluster and the two groups here dont have any close datas.
Complete linkage fails since it takes the farthest data from the cluster to calculate distance.
Average linkage is one of the best here since average distance to the both clusters are very distinct.
Centroid linkage is one of the best since centers of the clusters relative to datas

are very distinct.

3.4 data4

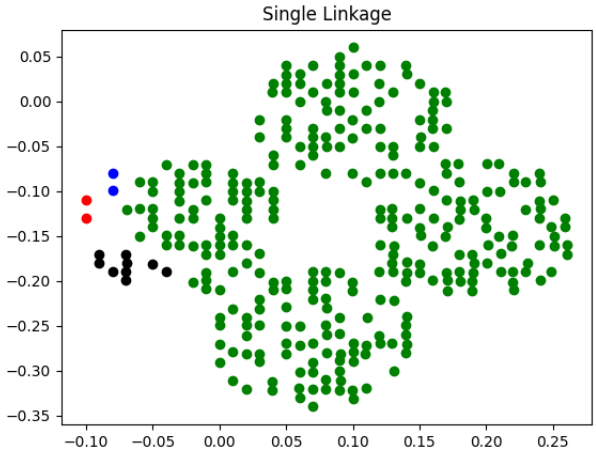

Figure 22: Fourth data's cluster using single linkage

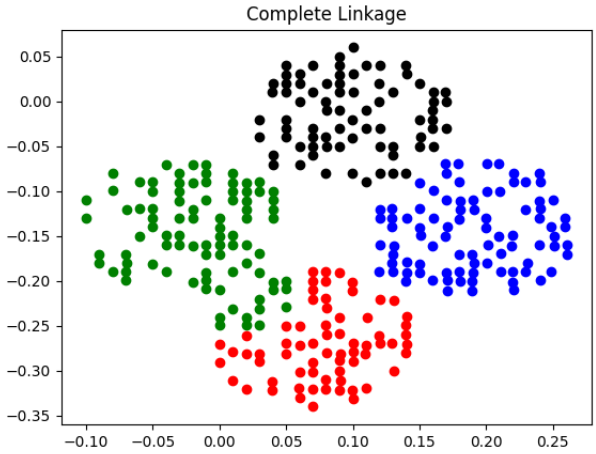

Figure 23: Fourth data's cluster using complete linkage

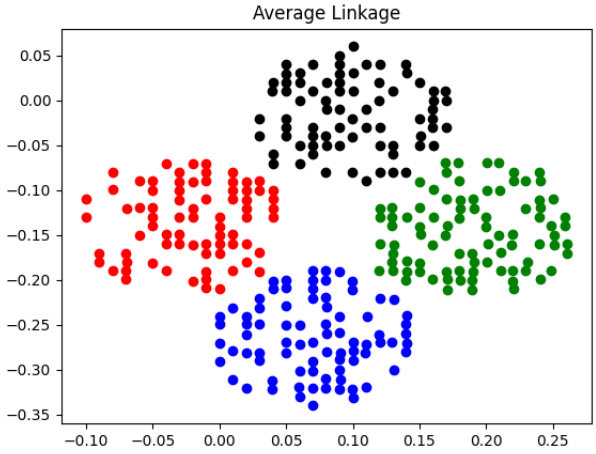

Figure 24: Fourth data's cluster using average linkage

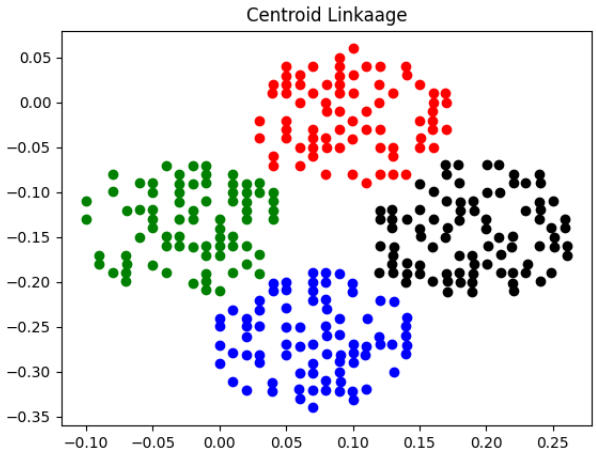

Figure 25: Fourth data's cluster using centroid linkage

Single linkage is the worst here since clusters are next to each other.
Complete linkage is sufficient since farthest points are very distinct.
Average linkage is one of the best here since average distance to the both clusters are very distinct.
Centroid linkage is one of the best since centers of the clusters relative to datas are very distinct.