Middle East Technical University Northern Cyprus Campus
Computer Engineering Program

CNG 409 Special Topics: Introduction to Machine Learning
Fall 2021-2022

Homework 2
K-Nearest Neighbors Algorithm, K-Means Clustering, and Hierarchical Agglomerative Clustering

2243392 Pınar Dilbaz

Chapter 1

INTRODUCTION

Thanks to this assignment, I had the opportunity to work with K-Nearest Neighbors Algorithm(KNN), K-Means Clustering and Hierarchical Agglomerative Clustering(HAC) algorithms and to code the algorithms separately. It was very useful for me to understand how to use these algorithms that we learned theoretically in the lessons. I think that while working for these algorithms, I gain practicality and understand the logic of the algorithms better. We were expected to code the KNN algorithm for the first part, and then I added a main part and observed how the average accuracy changed with increasing k value. In the second part, I wrote the K-means algorithm and added the main part. In this way, I examined the graphics using the elbow method and found which k value was the most suitable for each data separately, and I clustered with this k value. At the same time, I had the opportunity to create graphics for all data and visually examine them. In addition, finally I wrote a code for the HAC algorithm in the third part. I used 4 different methods for each dataset and I had a chance to visualize them and understand which distance finding method is best for which dataset.

Chapter 2

K-Nearest Neighbors Algorithm

2.1 K-Fold Cross Validation

2.1.1 Manhattan Distance

Below, you can examine the average accuracy and k-value graph with 10-fold CV applied to the training set.

- KNN = 1, 2, 3, ..., 179
- Manhattan (L1) distance

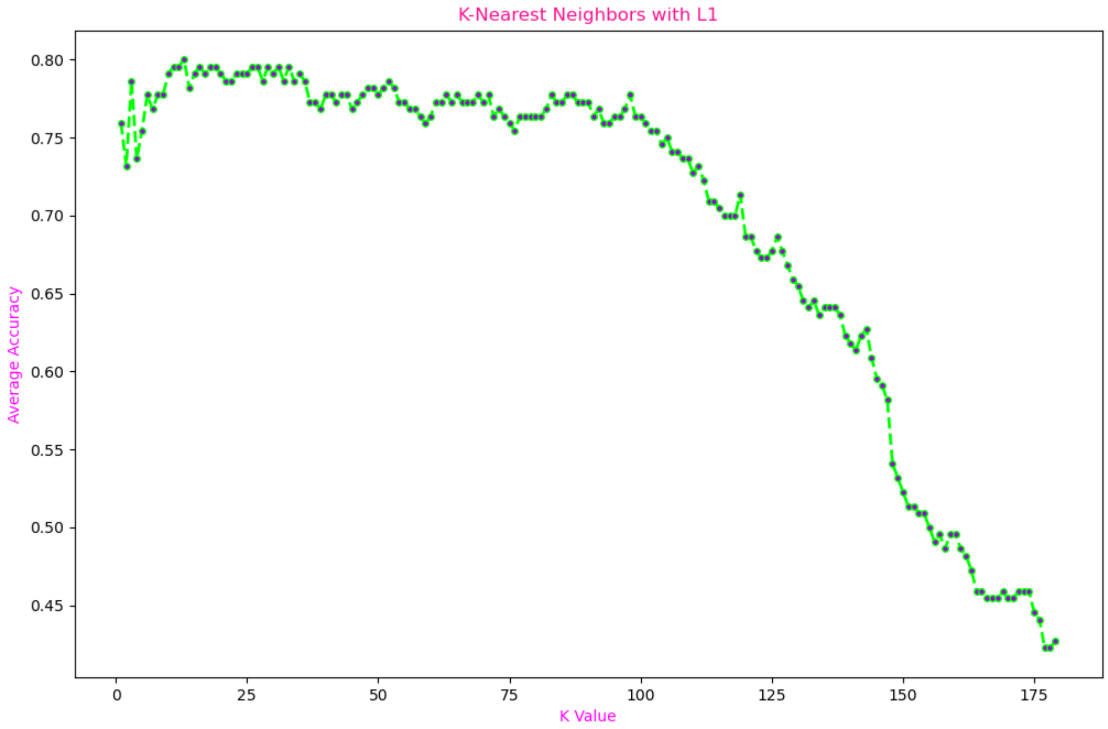

Figure 2.1: 10-Fold CV with Manhattan distance

The Test Accuracy for the best k value

- The best k-value is **13**
- The test accuracy is **0.833**

2.1.2 Euclidean Distance

Below, you can examine the average accuracy and k-value graph with 10-fold CV applied to the training set.

- KNN = 1, 2, 3, ..., 179
- Euclidean (L2) distance

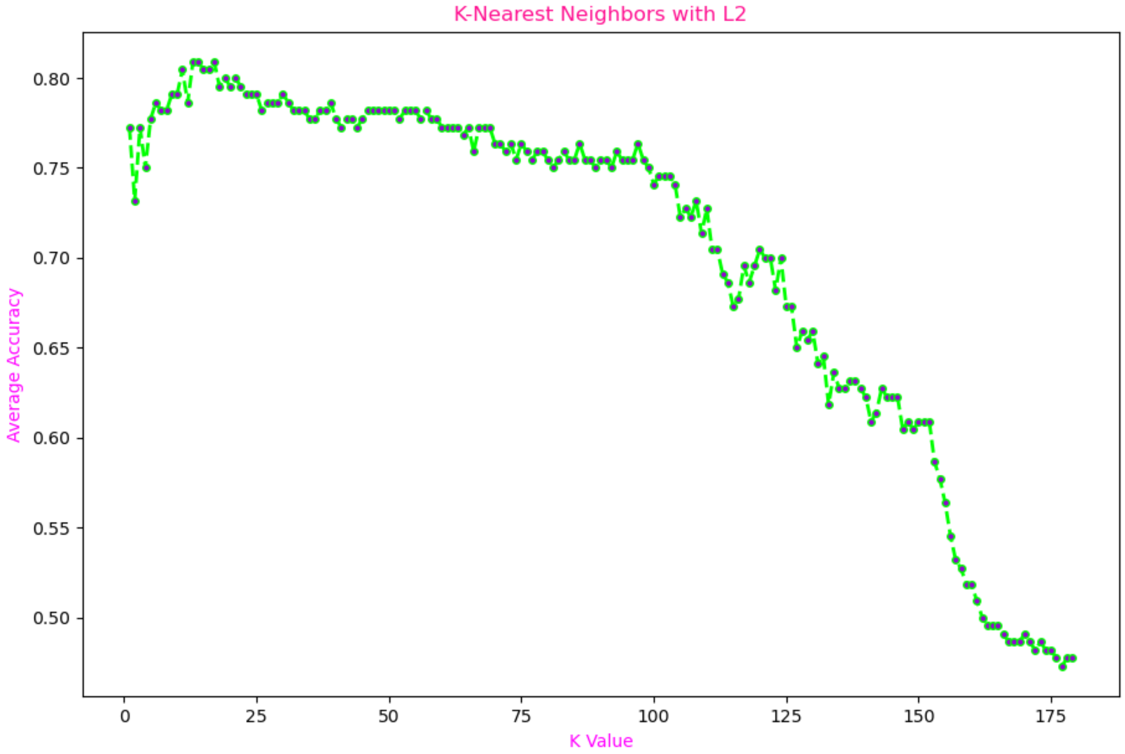

Figure 2.2: 10-Fold CV with Euclidean distance

The Test Accuracy for the best k value

- The best k-value is **14**
- The test accuracy is **0.806**

When we examine the graphics, we can clearly see that the average accuracy decreases when the K value increases too much. The reason for this is overfitting with a large k value, and therefore our average accuracy drops considerably.

Chapter 3

K-Means Clustering

3.1 Dataset-1

Elbow Method for Dataset-1

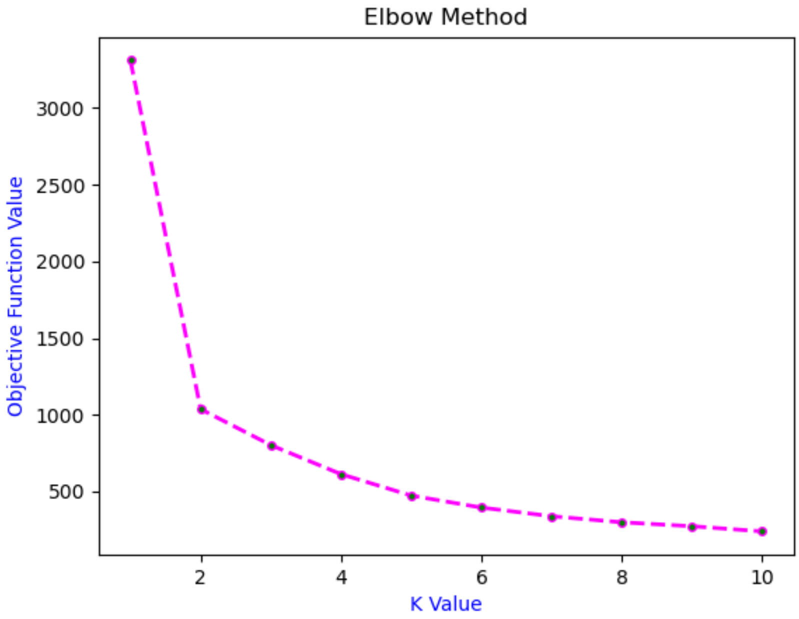

Figure 3.1: Elbow Method for Dataset-1

The best k value is **2** by using the elbow method. Below we can see colorization of the dataset-1 after it has been clustered with the best k value

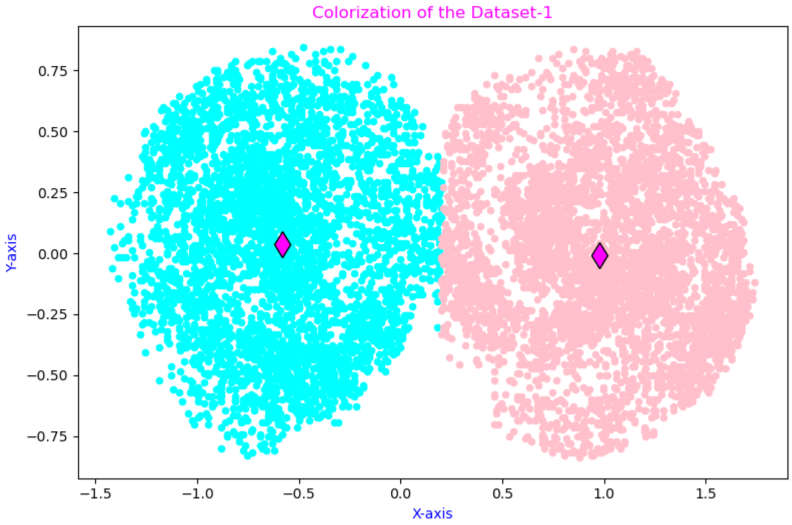

Figure 3.2: Colorization of the Dataset-1 with k = 2

3.2 Dataset-2

Elbow Method for Dataset-2

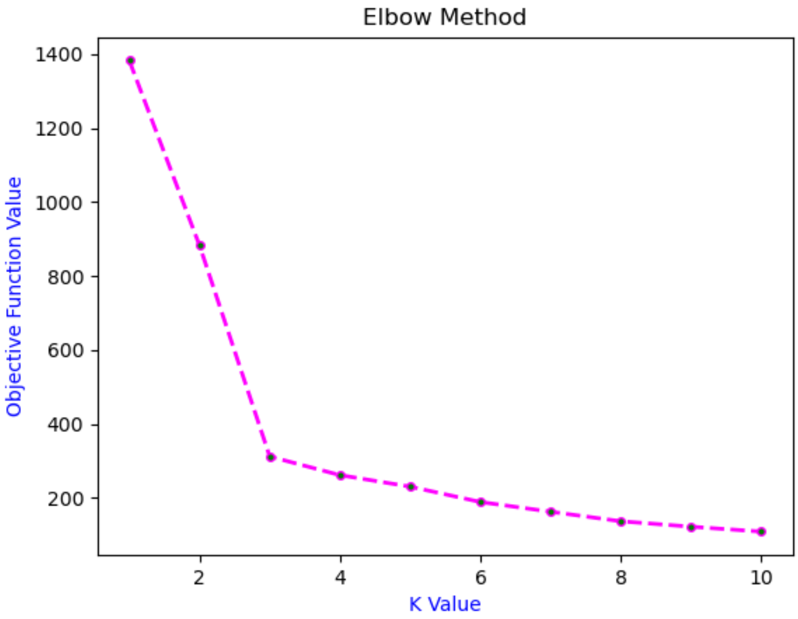

Figure 3.3: Elbow Method for Dataset-2

The best k value is **3** by using the elbow method. Below we can see colorization of the dataset-2 after it has been clustered with the best k value

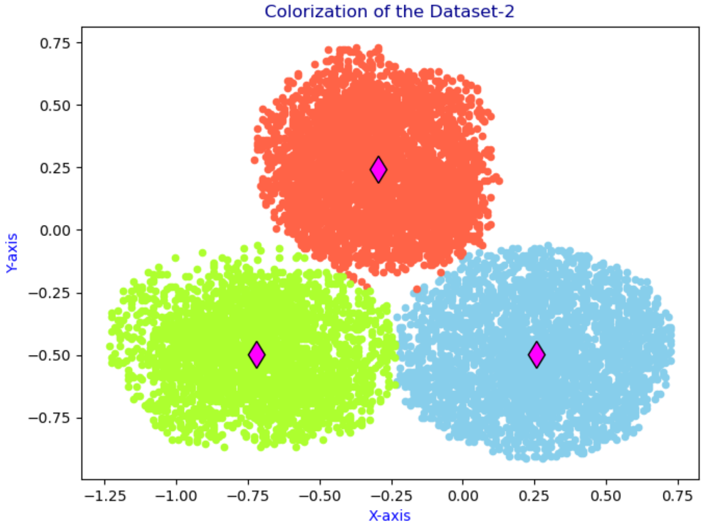

Figure 3.4: Colorization of the Dataset-2 with k = 3

3.3 Dataset-3

Elbow Method for Dataset-3

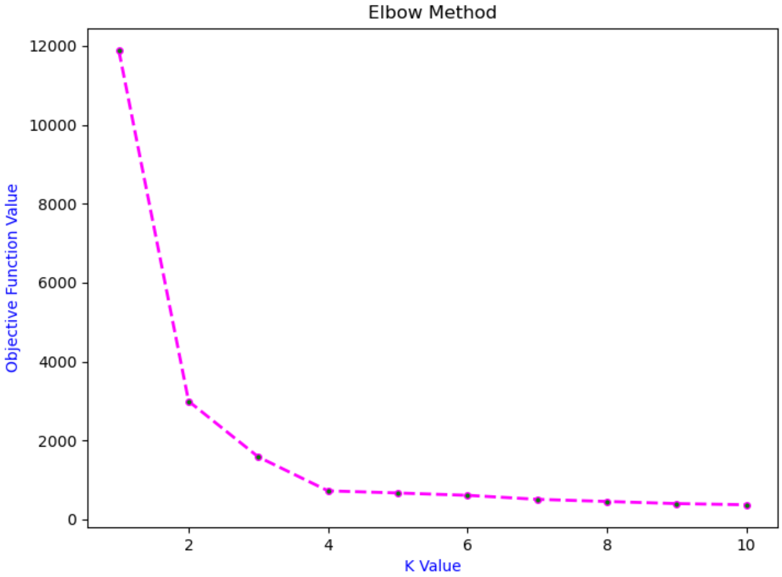

Figure 3.5: Elbow Method for Dataset-3

The best k value is **4** by using the elbow method. Below we can see colorization of the dataset-3 after it has been clustered with the best k value

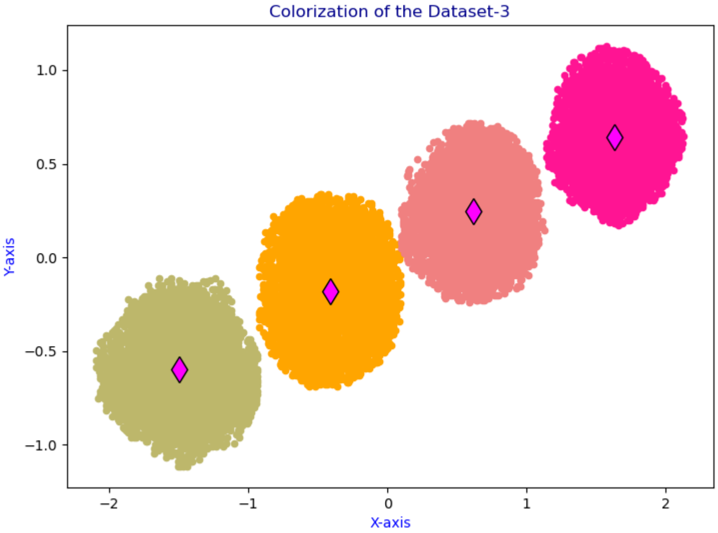

Figure 3.6: Colorization of the Dataset-3 with k = 4

3.4 Dataset-4

Elbow Method for Dataset-4

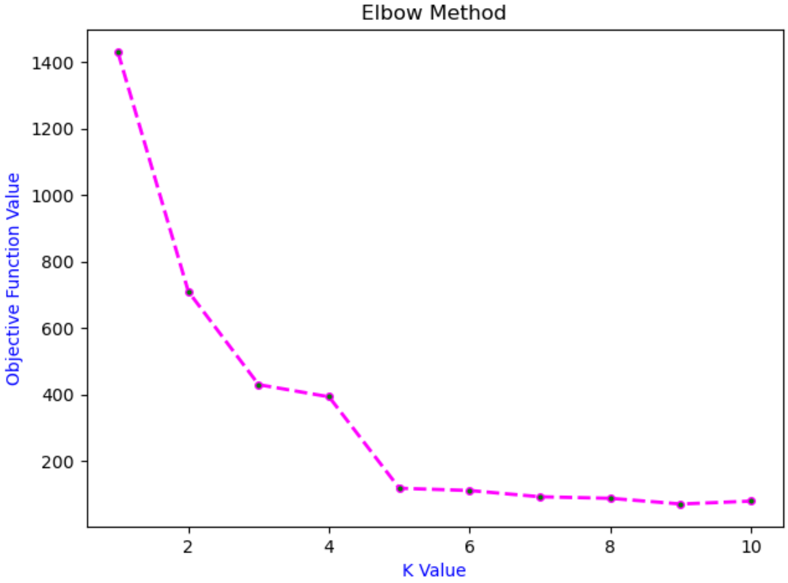

Figure 3.7: Elbow Method for Dataset-4

The best k value is **4** by using the elbow method. Below we can see colorization of the dataset-4 after it has been clustered with the best k value

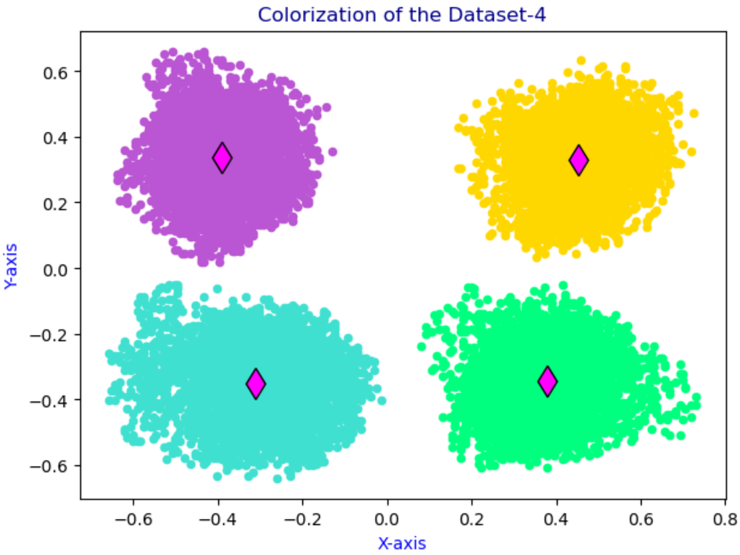

Figure 3.8: Colorization of the Dataset-4 with k = 4

Chapter 4

Hierarchical Agglomerative Clustering

4.1 Dataset-1

Colorization the resulting clusters

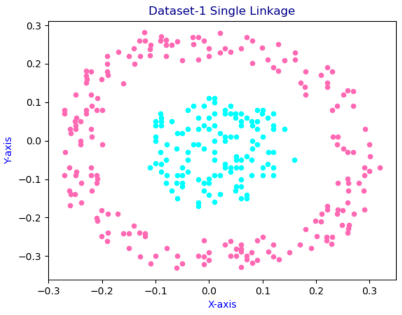

Figure 4.1: 1-single

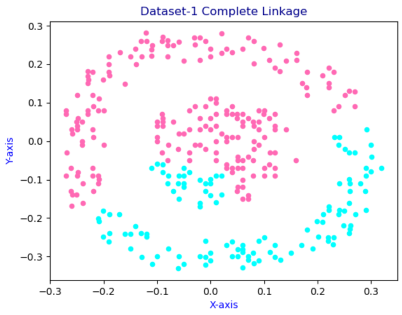

Figure 4.2: 1-complete

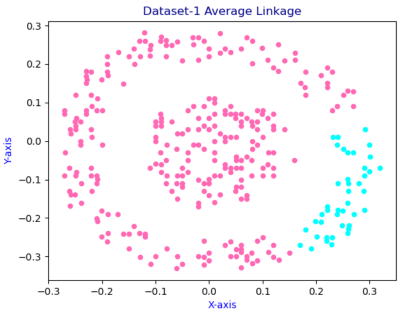

Figure 4.3: 1-average

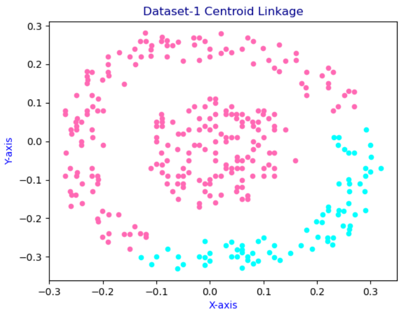

Figure 4.4: 1-centroid

1. **The Single Linkage Criterion:** When the single linkage criterion is used, we get the best clustering result because of the distribution of the dataset's data. Because the distance between the center circle and the outer circle is sufficient for this.

2. **The Complete Linkage Criterion:** The complete linkage criterion does not seem appropriate for this dataset because some data are too close to each other.

3. **The Average Linkage Criterion:** Average linkage criterion is not suitable for this dataset, just like complete linkage.
4. **The Centroid Linkage Criterion:** Since the centers of both clusters are very close, the centroid linkage does not seem very suitable for this dataset.

- *The best result:* The Single Linkage Criterion

4.2 Dataset-2

Colorization the resulting clusters

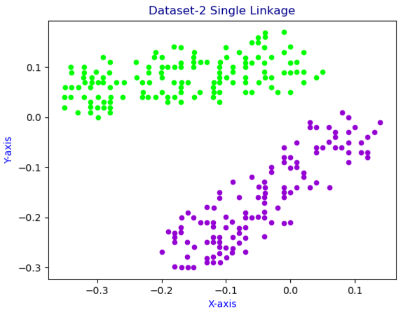

Figure 4.5: 2-single

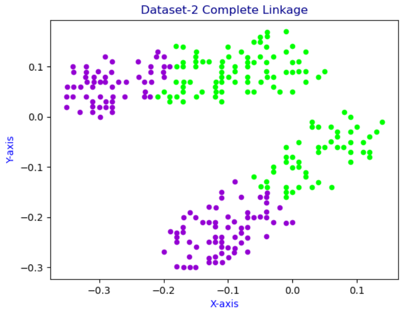

Figure 4.6: 2-complete

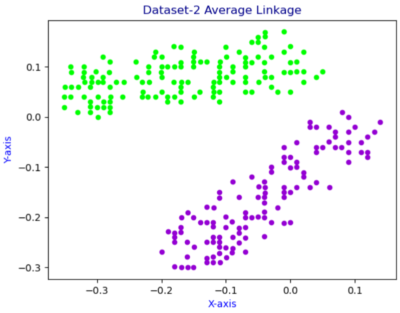

Figure 4.7: 2-average

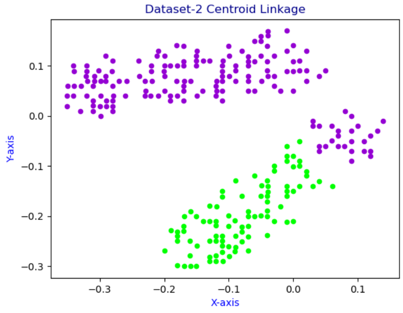

Figure 4.8: 2-centroid

1. **The Single Linkage Criterion:** It gave a fairly good clustering result because the data of the two clusters are far enough from each other.
2. **The Complete Linkage Criterion:** Not suitable for this dataset, clustering is not done properly.
3. **The Average Linkage Criterion:** Just like single linkage, the clustering process has been successfully completed, therefore, a suitable criterion for this dataset.
4. **The Centroid Linkage Criterion:** Where data belonging to different clusters are close to each other, incorrect clustering has taken place and therefore it is not suitable for our dataset.

- *The best results:* The Single Linkage Criterion and The Average Linkage Criterion

4.3 Dataset-3

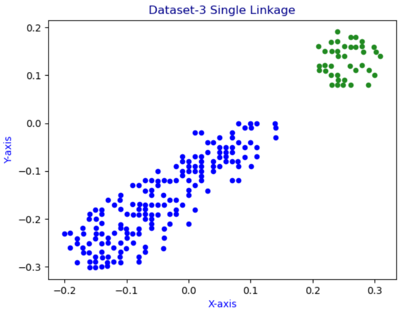

Figure 4.9: 3-single

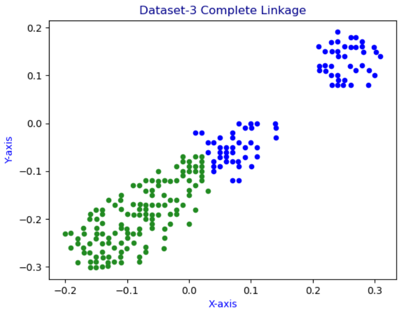

Figure 4.10: 3-complete

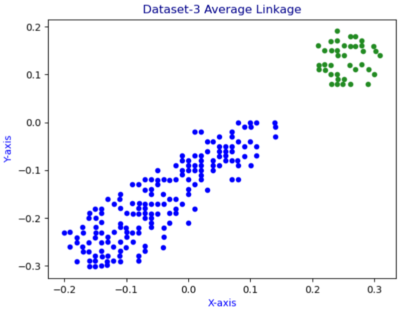

Figure 4.11: 3-average

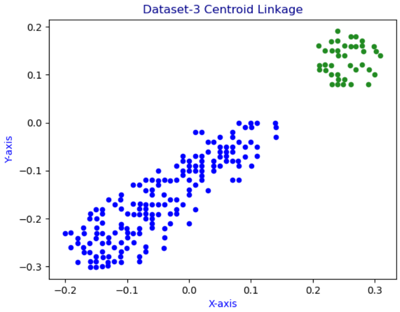

Figure 4.12: 3-centroid

1. **The Single Linkage Criterion:** The Single Linkage criterion is a suitable method for this dataset and it was able to separate the two sets from each other properly because of the distance of the data from each other.

2. **The Complete Linkage Criterion:** This is the only unsuitable method for the dataset. Some data are closer to the data that does not have its own cluster and therefore could not make the appropriate clustering.

3. **The Average Linkage Criterion:** Another of the best methods for this dataset is the Average Linkage Criterion because it was able to separate the data clearly.

4. **The Centroid Linkage Criterion:** Likewise, Centroid Linkage Criterion is one of the best methods for this dataset.

- *The best results:* The Single Linkage Criterion and The Average Linkage Criterion and The Centroid Linkage Criterion

4.4 Dataset-4

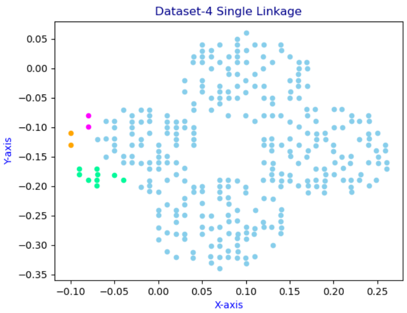

Figure 4.13: 4-single

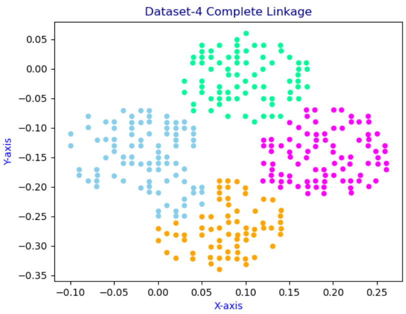

Figure 4.14: 4-complete

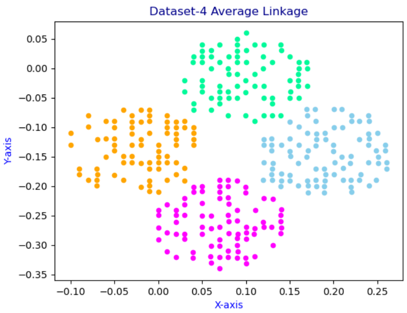

Figure 4.15: 4-average

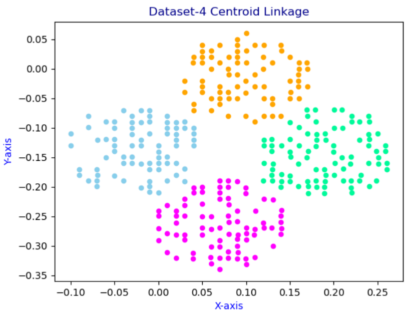

Figure 4.16: 4-centroid

1. **The Single Linkage Criterion:** As can be seen, this is not suitable for the dataset because the distance between the clusters is not far enough and therefore the correct clustering could not be made.

2. **The Complete Linkage Criterion:** A better result was obtained than Singe Linkage, but we still cannot say the best method for this dataset. Because some data seems to belong to the wrong cluster.

3. **The Average Linkage Criterion:** It is one of the methods that gives the best result. Due to the shape of the dataset, the clustering process has yielded very successful result.

4. **The Centroid Linkage Criterion:** Just like Average Linkage, Centroid Linkage Criterion is our best result. Same way, we achieved a successful result due to the shape of the clusters.

- *The best results:* The Average Linkage Criterion and The Centroid Linkage Criterion