UTM
UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF
COMPUTING SEMESTER
2/20232024

SECJ3563-06 COMPUTATIONAL INTELLIGENCE (KEPINTARAN KOMPUTER)

ASSIGNMENT 2

LECTURER: DR SHAFAATUNNUR HASAN

NAME	MATRIC NO
GAN HENG LAI	A21EC0176
NG KAI ZHENG	A21EC0101
LEW CHIN HONG	A21EC0044
YEO CHUN TECK	A21EC0148

Data reference

Original dataset: https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Reduced size dataset: Google Drive

Data Visualization and Pre-processing

In this dataset, the focus lies on the development of a classification model to differentiate between seven distinct registered varieties of dry beans. By utilizing the latest developments in computer vision technology, a large dataset including high-resolution camera images was assembled. A total of 16 characteristics were obtained by going through a variety of processing stages such as segmentation and feature extraction, with each image representing a single dried bean grain.

The dataset's multidimensional nature enables a detailed examination of the various attributes associated with the dry bean grains. With 13,611 instances and 16 important features per instance, the dataset offers ample opportunities for thorough analysis and model development (In this case, we are only taking half of the instances in each class which totally 6824 instances since the size of the original dataset is too large for Weka). The features those was taken in dataset like Area, Perimeter, MajorAxisLength, MinorAxisLength, AspectRation, Eccentricity, ConvexArea, EquivDiameter, Extent, Solidity, roundness, Compactness. Below are the visualization of this dataset:

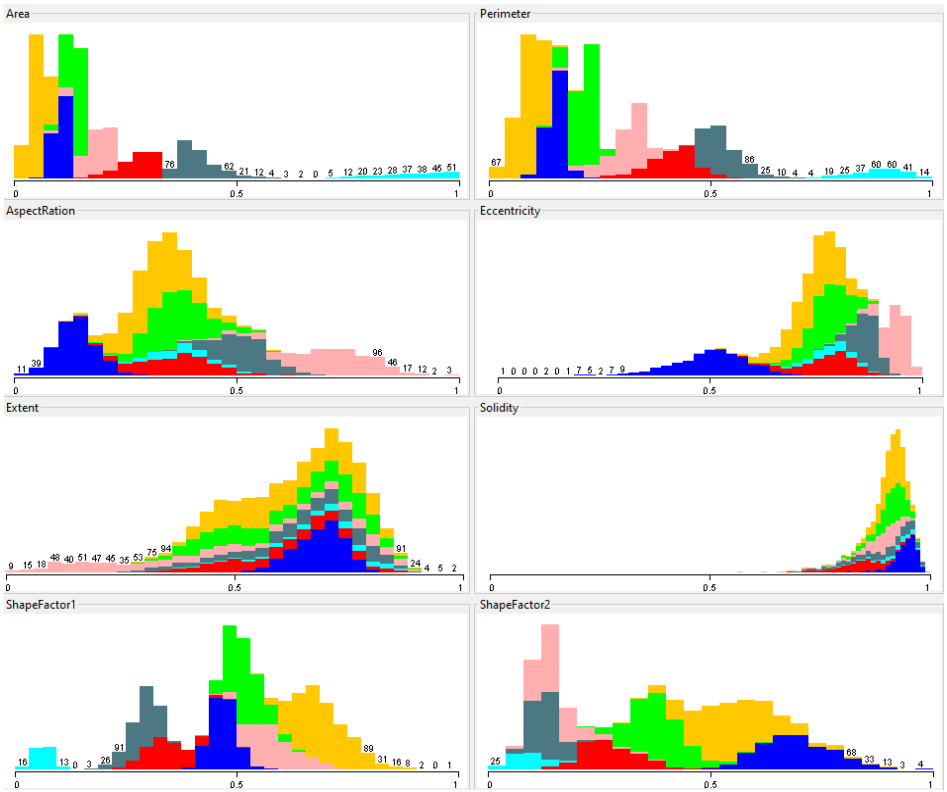

Figure 1 First Part of Data Visualization

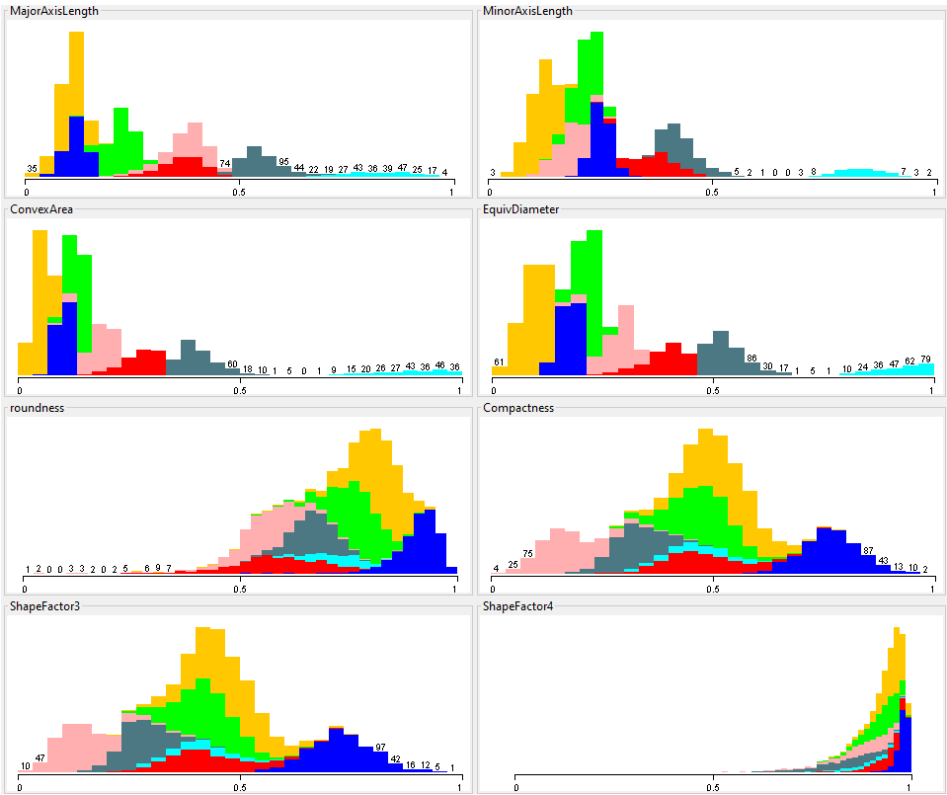

Figure 2 Second Part of Visualization

Pre-processing Dataset (Using Normalization)

Important Attributes	Before Pre-processing	After Pre-processing (Normalize)
Area	Minimum: 20420 Maximum: 171423 Mean: 49343.477 Standard Deviation: 27465.864	Minimum: 0 Maximum: 1 Mean: 0.192 Standard Deviation: 0.182
Perimeter	Minimum: 524.736 Maximum: 1631.451 Mean: 821.628 Standard Deviation: 212.155	Minimum: 0 Maximum: 1 Mean: 0.268 Standard Deviation: 0.192
MajorAxisLength	Minimum: 183.601 Maximum: 738.86 Mean: 320.142 Standard Deviation: 85.694	Minimum: 0 Maximum: 1 Mean: 0.274 Standard Deviation: 0.190
MinorAxisLength	Minimum: 122.513 Maximum: 408.343 Mean: 195.529 Standard Deviation: 44.904	Minimum: 0 Maximum: 1 Mean: 0.255 Standard Deviation: 0.157
AspectRatio	Minimum: 1.025 Maximum: 2.43 Mean: 1.569 Standard Deviation: 0.252	Minimum: 0 Maximum: 1 Mean: 0.387 Standard Deviation: 0.179
Eccentricity	Minimum: 0.219 Maximum: 0.911 Mean: 0.744 Standard Deviation: 0.097	Minimum: 0 Maximum: 1 Mean: 0.758 Standard Deviation: 0.140
ConvexArea	Minimum: 20684 Maximum: 175736 Mean: 50016.964 Standard Deviation: 27868.803	Minimum: 0 Maximum: 1 Mean: 0.189 Standard Deviation: 0.180
EquivDiameter	Minimum: 161.244 Maximum: 467.186 Mean: 243.579	Minimum: 0 Maximum: 1 Mean: 0.269

	Standard Deviation: 59.125	Standard Deviation: 0.193
Extent	Minimum: 0.572 Maximum: 0.866 Mean: 0.751 Standard Deviation: 0.049	Minimum: 0 Maximum: 1 Mean: 0.609 Standard Deviation: 0.165
Solidity	Minimum: 0.919 Maximum: 0.995 Mean: 0.987 Standard Deviation: 0.005	Minimum: 0 Maximum: 1 Mean: 0.898 Standard Deviation: 0.061
Roundness	Minimum: 0.557 Maximum: 0.991 Mean: 0.876 Standard Deviation: 0.0059	Minimum: 0 Maximum: 1 Mean: 0.737 Standard Deviation: 0.136
Compactness	Minimum: 0.641 Maximum: 0.987 Mean: 0.804 Standard Deviation: 0.064	Minimum: 0 Maximum: 1 Mean: 0.471 Standard Deviation: 0.184
ShapeFactor1	Minimum: 0.003 Maximum: 0.010 Mean: 0.007 Standard Deviation: 0.001	Minimum: 0 Maximum: 1 Mean: 0.503 Standard Deviation: 0.168
ShapeFactor2	Minimum: 0.001 Maximum: 0.004 Mean: 0.002 Standard Deviation: 0.001	Minimum: 0 Maximum: 1 Mean: 0.388 Standard Deviation: 0.214
ShapeFactor3	Minimum: 0.41 Maximum: 0.975 Mean: 0.650 Standard Deviation: 0.103	Minimum: 0 Maximum: 1 Mean: 0.425 Standard Deviation: 0.182
ShapeFactor4	Minimum: 0.948 Maximum: 1 Mean: 0.995 Standard Deviation: 0.004	Minimum: 0 Maximum: 1 Mean: 0.913 Standard Deviation: 0.081

Data Training

- **Euclidean Distance with default parameter**

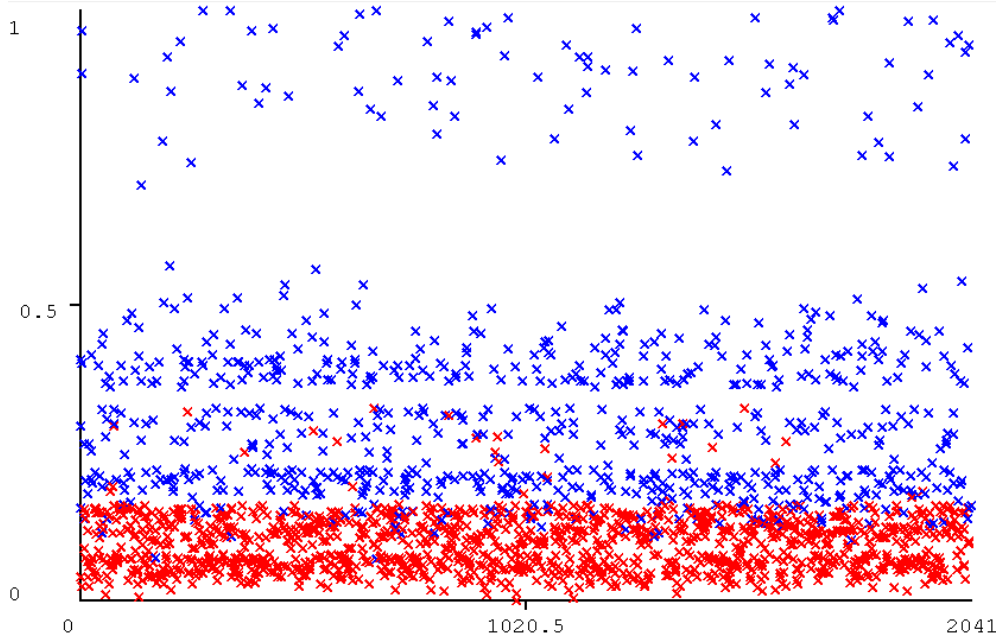

Figure 3 K-means clustering with Default Parameter

After applying the K-means algorithm with default parameters to the dataset, the resulting output revealed a plot depicting two distinct types of dry beans. On closer inspection, it was clear that the cluster distribution was unbalanced. In particular, a large percentage of the grains from cluster 1 seemed to be grouped together near the bottom of the plot, but a small percentage of the grains from cluster 0 were plotted at the upper part of the graph. There are only 789 instances that belong to the first cluster while 1253 instances belong to the second cluster.

The dataset may have underlying patterns or qualities that are indicated by this imbalance in the cluster distribution. It implies that some characteristics, or combinations of characteristics, may be more common or unique among grains assigned to cluster 0 (39%) than among those in cluster 1 (61%).

```
Class attribute: Class
Classes to Clusters:

    0    1  <-- assigned to cluster
 1013    0 | SEKER
   80  581 | BARBUNYA
    0  261 | BOMBAY
    0  815 | CALI
   32  932 | HOROZ
 1279   39 | SIRA
 1773    0 | DERMASON

Cluster 0 <-- DERMASON
Cluster 1 <-- HOROZ

Incorrectly clustered instances :       4100.0   60.2498 %
```

Figure 4 Cluster Evaluation of Default Parameter

After having the evaluation assignment, the incorrectly clustered instances get high to 4100 (60.25%) from the overall dry bean. The first cluster was given by the Dermason class while the second cluster was given by the Horoz class. Additional investigation is necessary to fully understand the observed clustering behavior.

- **Euclidean Distance with parameter**

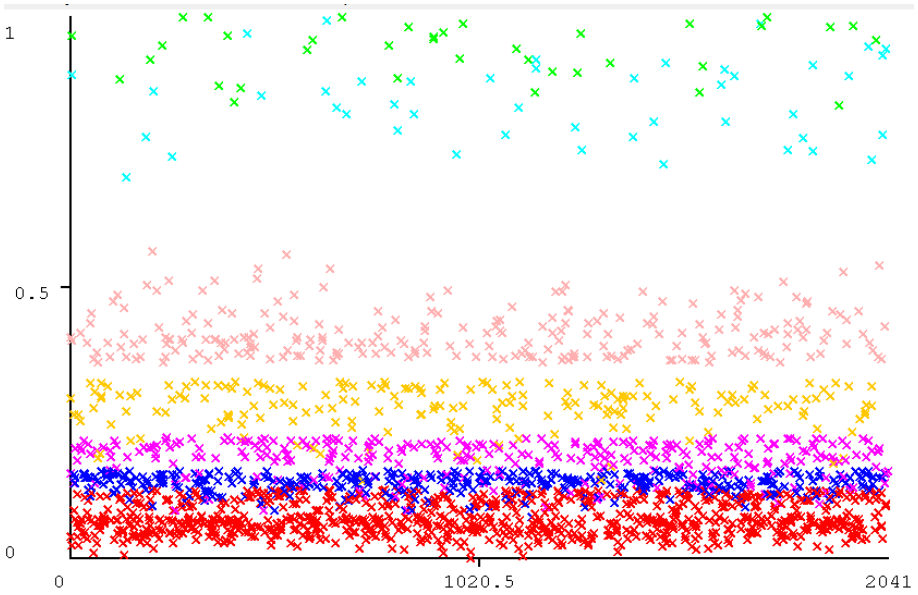

Figure 5 K-means clustering with Euclidean

After applying the K-means algorithm with the parameter specifying 7 clusters, K-means++ as initialization method and utilizing Euclidean Distance as the distance metric, the resulting classification of the dry beans was clearly delineated in the graph. This selection of seven clusters corresponds to the number of unique dry bean varieties found in the dataset, allowing for a one-to-one mapping between bean varieties and clusters.

The plot's visual examination revealed that every cluster had a distinct concentrated horizontal scale, indicating the uniqueness of the beans within each cluster. This shows that the algorithm was able to efficiently capture the intrinsic heterogeneity across various bean kinds by grouping the dried bean grains according to their unique traits. In contrast to the more densely packed formations seen in the other clusters, two clusters stood out due to its dispersed dispersion across the plotting graph. The number of instances in each cluster are 386 (19%) for Cluster 0, 842 (41%) for Cluster 1, 37 (2%) for Cluster 2, 45 (2%) for Cluster 3, 242 (12%) for Cluster 4, 401 (15%) for Cluster 5, and 189 (9%) for Cluster 6.

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
  992    0    0   11    0    0   10 | SEKER
    6    5  594    0   24    0   32 | BARBUNYA
    0    0    0    0    0  261    0 | BOMBAY
    0    0    3    0  812    0    0 | CALI
    0  911    4    0    0    0   49 | HOROZ
   16   14    1   24    0    0 1263 | SIRA
   62    5    0 1685    0    0   21 | DERMASON

Cluster 0 <-- SEKER
Cluster 1 <-- HOROZ
Cluster 2 <-- BARBUNYA
Cluster 3 <-- DERMASON
Cluster 4 <-- CALI
Cluster 5 <-- BOMBAY
Cluster 6 <-- SIRA

Incorrectly clustered instances :       287.0     4.2175 %
```

Figure 6 Cluster Evaluation of Euclidean Distance

After applying the cluster's evaluation, we can find the incorrectly clustered instances reduced to 287 which is 4.22% from the overall dry bean. The Cluster 0 assigned by Seker, Cluster 1 is Horoz, Cluster 2 is Barbunya, Cluster 3 is Dermason, Cluster 4 is Cali, Cluster 5 is Bombay and Cluster 6 is Sira.

- **Manhattan Distance**

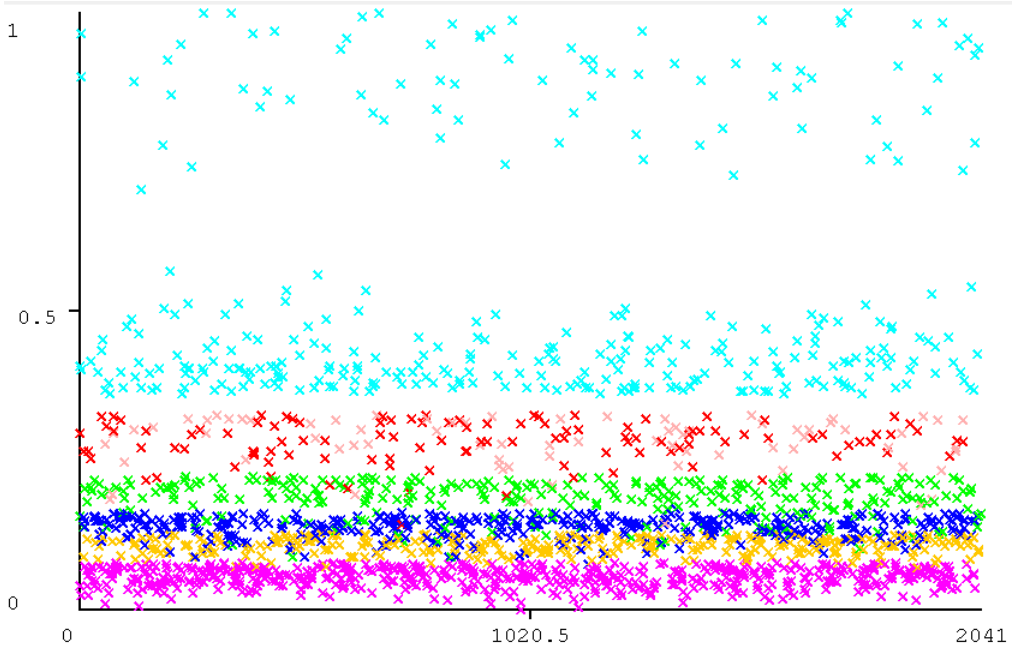

Figure 7 K-means clustering with Manhattan

This is the visualization of cluster assignment using Manhattan Distance with the same parameter with type of distance. It shows the similar plotting outcome in Euclidean Distance. However, there are disparities between the two distance metrics. The finding that some clusters had lower density sections at the top of the plotting graph was very notable. This suggests that the Manhattan Distance measure may have had a different impact on the clustering findings than the closely packed formations shown in the Euclidean Distance figure.

The second difference between both types of distances is the number of instances belonging to each cluster. The number of instances in each cluster are 392 (19%) for Cluster 0, 113 (6%) for Cluster 1, 295 (14%) for Cluster 2, 324 (16%) for Cluster 3, 75 (4%) for Cluster 4, 546 (27%) for Cluster 5, and 297 (15%) for Cluster 6.

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
  996    0    0    9    0    0    8 | SEKER
    6    0    7   35  597   16    0 | BARBUNYA
    0  261    0    0    0    0    0 | BOMBAY
    0    0    0    0    3  812    0 | CALI
    0    0  916   45    3    0    0 | HOROZ
   19    0   13 1255    0    0   31 | SIRA
   36    0    5    0    0    0 1732 | DERMASON

Cluster 0 <-- SEKER
Cluster 1 <-- BOMBAY
Cluster 2 <-- HOROZ
Cluster 3 <-- SIRA
Cluster 4 <-- BARBUNYA
Cluster 5 <-- CALI
Cluster 6 <-- DERMASON

Incorrectly clustered instances :       236.0     3.468  %
```

Figure 8 Cluster Evaluation of Manhattan Distance

After applying the cluster's evaluation, we can find the incorrectly clustered instances reduced to 236 which is 3.468% from the overall dry bean. The cluster 0 assigned by Seker, Cluster 1 is Bombay, Cluster 2 is Horoz, Cluster 3 is Sira, Cluster 4 is Barbunya, Cluster 5 is Caliand and Cluster 6 is Dermason.

Experimental Results and Algorithm Performances of K-means

Cycles	Iterations	Initialization Method	No of cluster	Type of Distance Measures	
				Euclidean	Manhattan
1	23	K-means++	7	4.22% Incorrect	
2	23	K-means++	7	4.22% Incorrect	
3	23	K-means++	7	4.22% Incorrect	
			Average	4.22% Incorrect	
4	26	K-means++	7		3.74% Incorrect
5	26	K-means++	7		3.74% Incorrect
6	26	K-means++	7		3.74% Incorrect
			Average		3.74% Incorrect
7	12	Random	2	60.25% Incorrect	
8	12	Random	2	60.25% Incorrect	
9	12	Random	2	60.25% Incorrect	
			Average	60.25% Incorrect	

Table 1 K-means Clustering Algorithms Performance Measure

Hierarchical Cluster Algorithm

- **Euclidean Distance**:

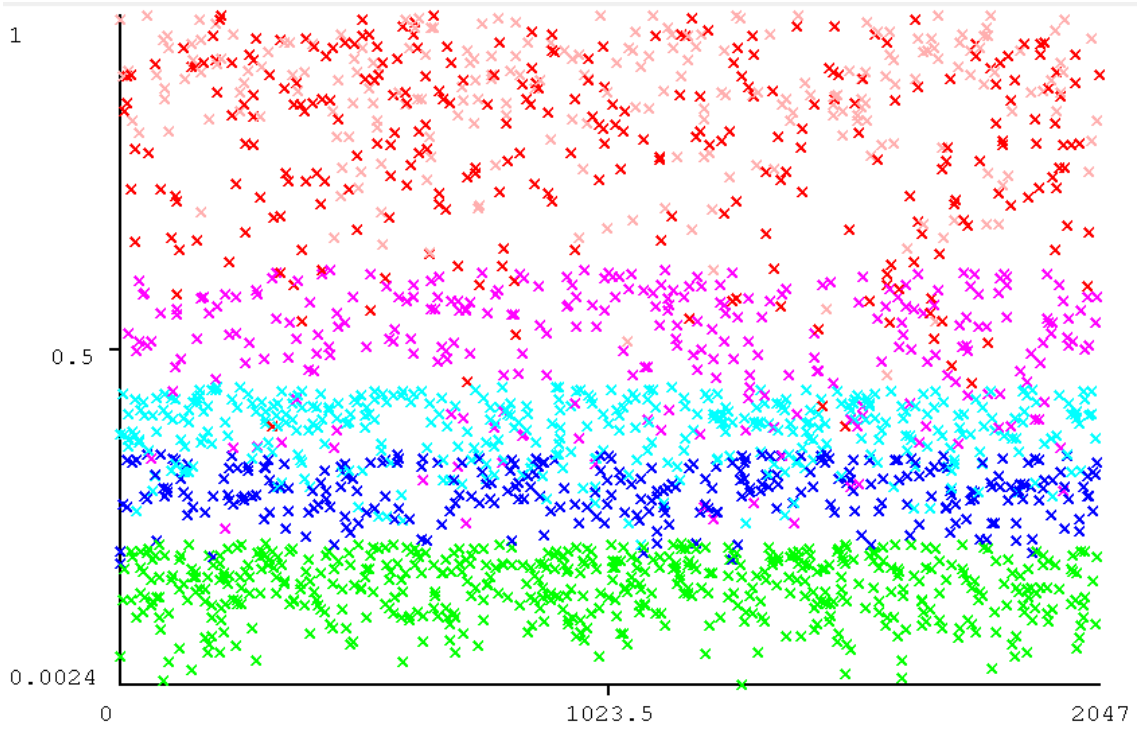

Figure 9 Hierarchical Clustering with Euclidean

Upon applying the hierarchical cluster algorithm with the parameter specifying 7 clusters and utilizing Euclidean Distance as the distance metric, the resulting classification of the dry beans exhibited some ambiguity in the graph. This selection of seven clusters corresponds to the number of unique dry bean varieties found in the dataset, but the 70 percent split training data resulted in the formation of six clusters: Cluster 0 - 82 instances (4%), Cluster 1 - 242 (12%), Cluster 2 - 546 (27%), Cluster 3 - 687 (34%), Cluster 4 - 296 (14%) and Cluster 5 - 189 (9%) .

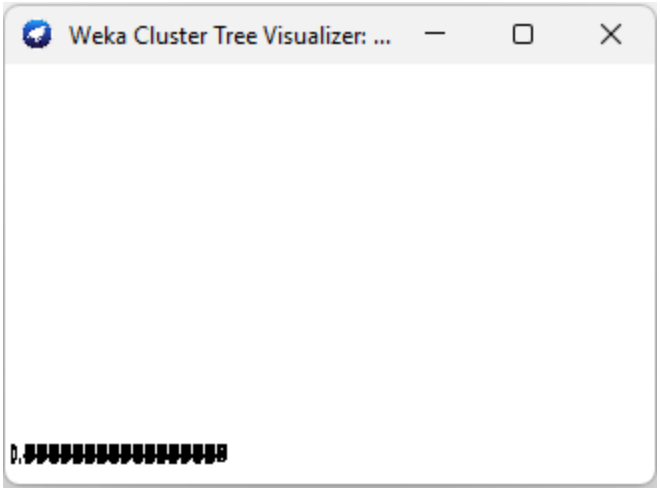

Figure 10 Tree Visualization

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
 1012    1    0    0    0    0    0 | SEKER
  661    0    0    0    0    0    0 | BARBUNYA
  260    0    1    0    0    0    0 | BOMBAY
  815    0    0    0    0    0    0 | CALI
  960    0    0    1    2    1    0 | HOROZ
 1318    0    0    0    0    0    0 | SIRA
 1772    0    0    0    0    0    1 | DERMASON

Cluster 0 <-- DERMASON
Cluster 1 <-- SEKER
Cluster 2 <-- BOMBAY
Cluster 3 <-- No class
Cluster 4 <-- HOROZ
Cluster 5 <-- No class
Cluster 6 <-- No class

Incorrectly clustered instances :       5029.0   73.9015 %
```

Figure 11 Cluster Evaluation of Euclidean Distance (Hierarchical)

After applying the cluster evaluation, the incorrectly clustered instances rise to 5029 instances which is almost 74% from the overall number of the dry bean instances. The Cluster 0, Cluster 1, Cluster 2 and Cluster 4 were assigned by Dermason, Seker, Bombay and Horoz respectively. The rest three clusters are not assigned with any class data. These unassigned clusters indicate that there may have been uncertainties in the dataset that confused the clustering algorithm and caused clusters without distinct class relationships to arise.

- **Manhattan Distance**

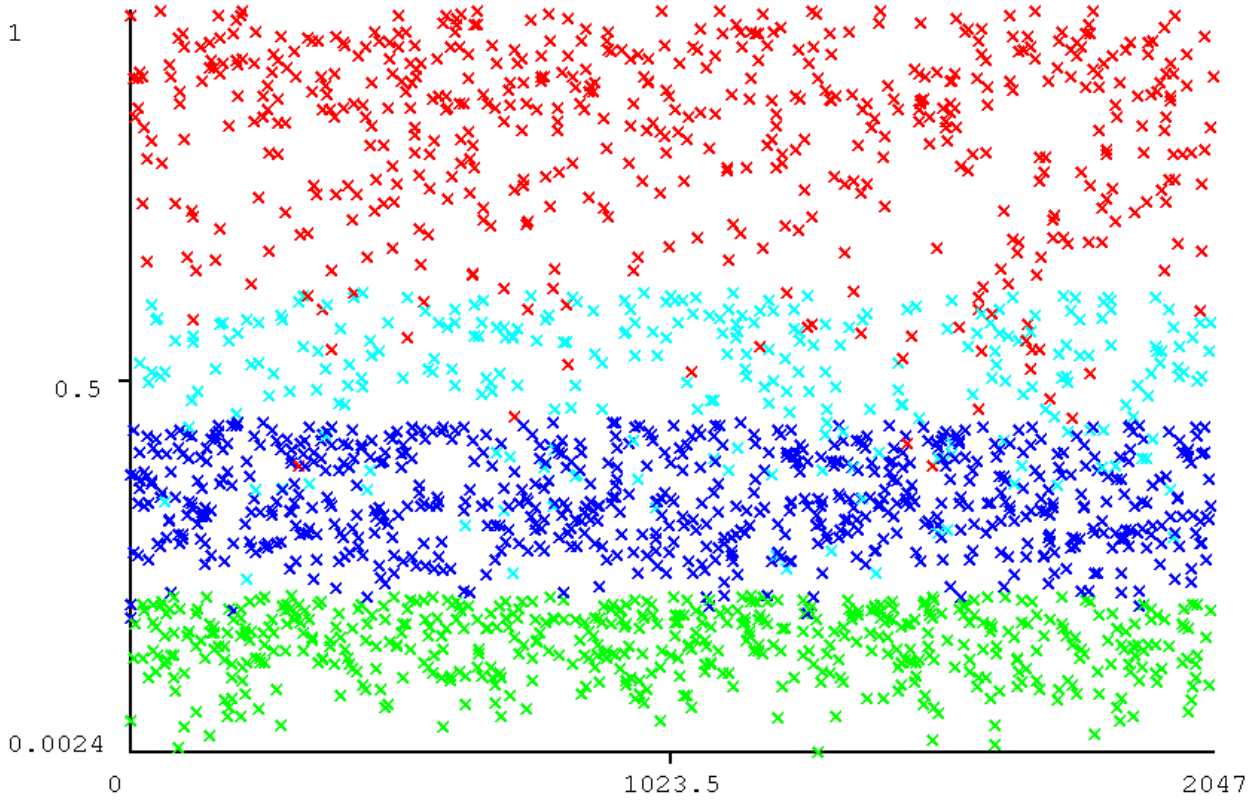

Figure 12 Hierarchical Clustering with Manhattan

Above shown the application of the hierarchical cluster algorithm with the parameter specifying 7 clusters and utilizing Manhattan Distance as the distance metric, the resulting classification of the dry beans exhibited some ambiguity in the graph. The hierarchical clustering algorithm produced only five distinct groups when trained on 70% of the data, despite the specification of seven clusters, which corresponded to the number of unique dry bean types discovered in the dataset.

These clusters were labeled as cluster 0, Cluster 1, Cluster 2, Cluster 3 and Cluster 4, comprising 82 instances (4%), 242 instances (12%), 546 instances (27%), 876 instances (43%) and 296 instances (14%) respectively.

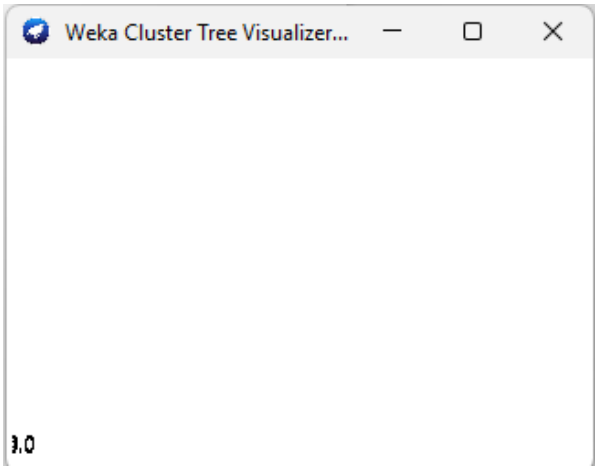

Figure 13 Tree Visualization

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
 1012    1    0    0    0    0    0 | SEKER
  661    0    0    0    0    0    0 | BARBUNYA
  259    0    1    1    0    0    0 | BOMBAY
  814    0    0    0    1    0    0 | CALI
  962    0    0    0    0    1    1 | HOROZ
 1318    0    0    0    0    0    0 | SIRA
 1773    0    0    0    0    0    0 | DERMASON

Cluster 0 <-- DERMASON
Cluster 1 <-- SEKER
Cluster 2 <-- No class
Cluster 3 <-- BOMBAY
Cluster 4 <-- CALI
Cluster 5 <-- No class
Cluster 6 <-- HOROZ

Incorrectly clustered instances :       5028.0   73.8868 %
```

Figure 14 Cluster Evaluation of Manhattan Distance (Hierarchical)

After applying the cluster evaluation, the incorrectly clustered instances maintain at 5029 instances which is 73.89% from the overall number of the dry bean instances. The Cluster 0, Cluster 1, Cluster 3, Cluster 4 and Cluster 6 were assigned by Dermason, Seker, Bombay, Cali and Horoz respectively. The other two clusters are not assigned with any class data. These unassigned clusters show that the dataset may have contained errors that misled the clustering algorithm and led to the emergence of clusters without clear class links.

- **Chebyshev Distance**

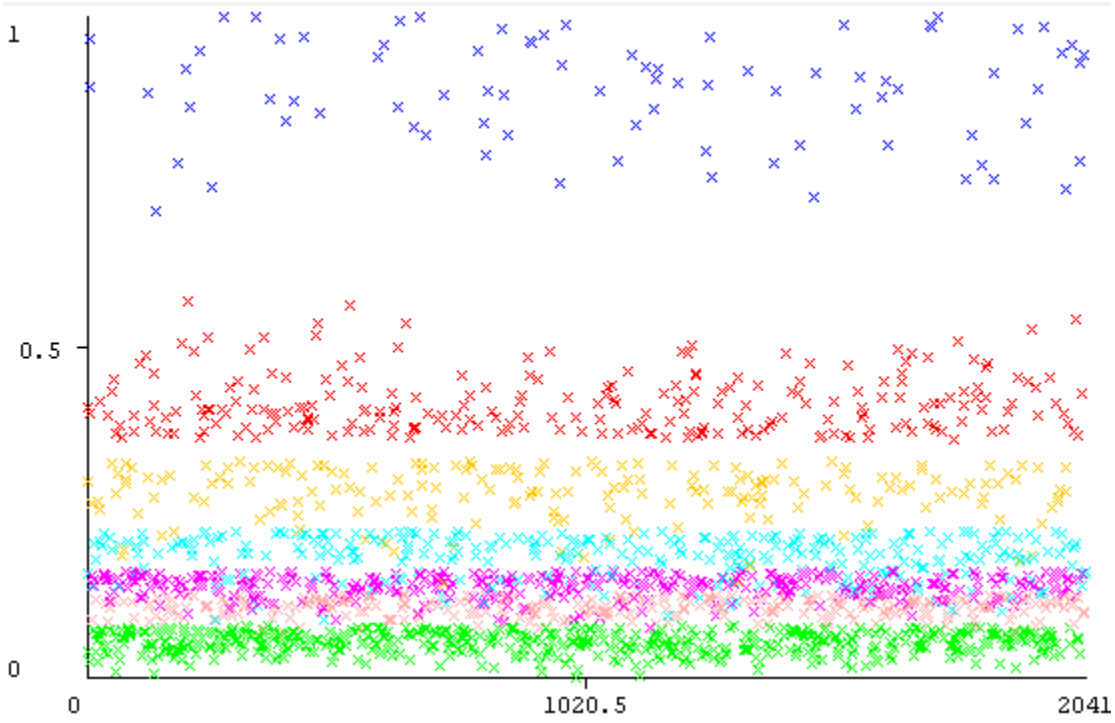

Figure 15 Hierarchical Clustering with Chebyshev

After applying the hierarchical cluster algorithm with the parameter specifying 7 clusters and utilizing Chebyshev Distance as the distance metric, the resulting classification of the dry beans exhibited some ambiguity in the graph. This selection of seven clusters corresponds to the number of unique dry bean varieties found in the dataset, Cluster 0 - 82 instances (4%), Cluster 1 - 242 (12%), Cluster 2 - 546 (27%), Cluster 3 - 301 (15%), Cluster 4 - 296 (14%), Cluster 5 - 386 (19%) and Cluster 6 - 189 (9%).

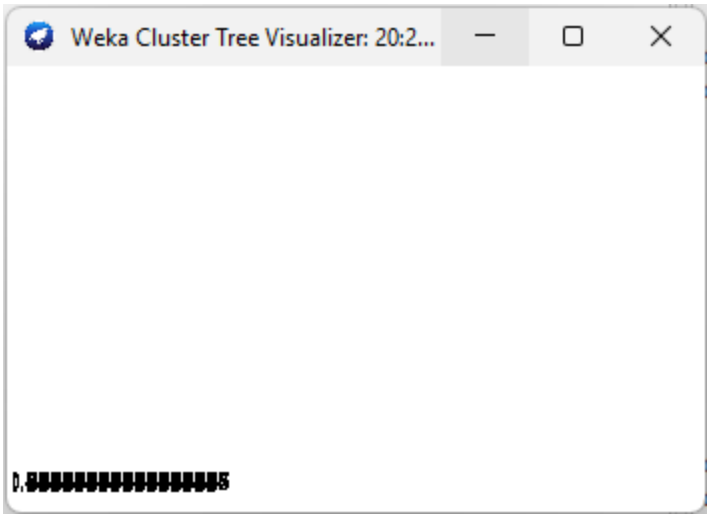

Figure 16 Tree Visualization

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
 1012    1    0    0    0    0    0 | SEKER
  661    0    0    0    0    0    0 | BARBUNYA
  261    0    0    0    0    0    0 | BOMBAY
  815    0    0    0    0    0    0 | CALI
  956    0    3    3    1    1    0 | HOROZ
 1318    0    0    0    0    0    0 | SIRA
 1772    0    0    0    0    0    1 | DERMASON

Cluster 0 <-- DERMASON
Cluster 1 <-- SEKER
Cluster 2 <-- No class
Cluster 3 <-- HOROZ
Cluster 4 <-- No class
Cluster 5 <-- No class
Cluster 6 <-- No class

Incorrectly clustered instances :       5029.0   73.9015 %
```

Figure 17 Cluster Evaluation of Chebyshev Distance (Hierarchical)

After applying the cluster evaluation, the incorrectly clustered instances still maintain at 5029 instances which is 73.90% from the overall number of the dry bean instances. There are only three assigned by class: Cluster 0 (Dermason), Cluster 1 (Seker), Cluster 3 (Horoz). The rest 4 clusters are not assigned with any class data. These unassigned clusters show that the dataset may have contained errors that misled the clustering algorithm and led to the emergence of clusters without clear class links.

- **Minkowski Distance**

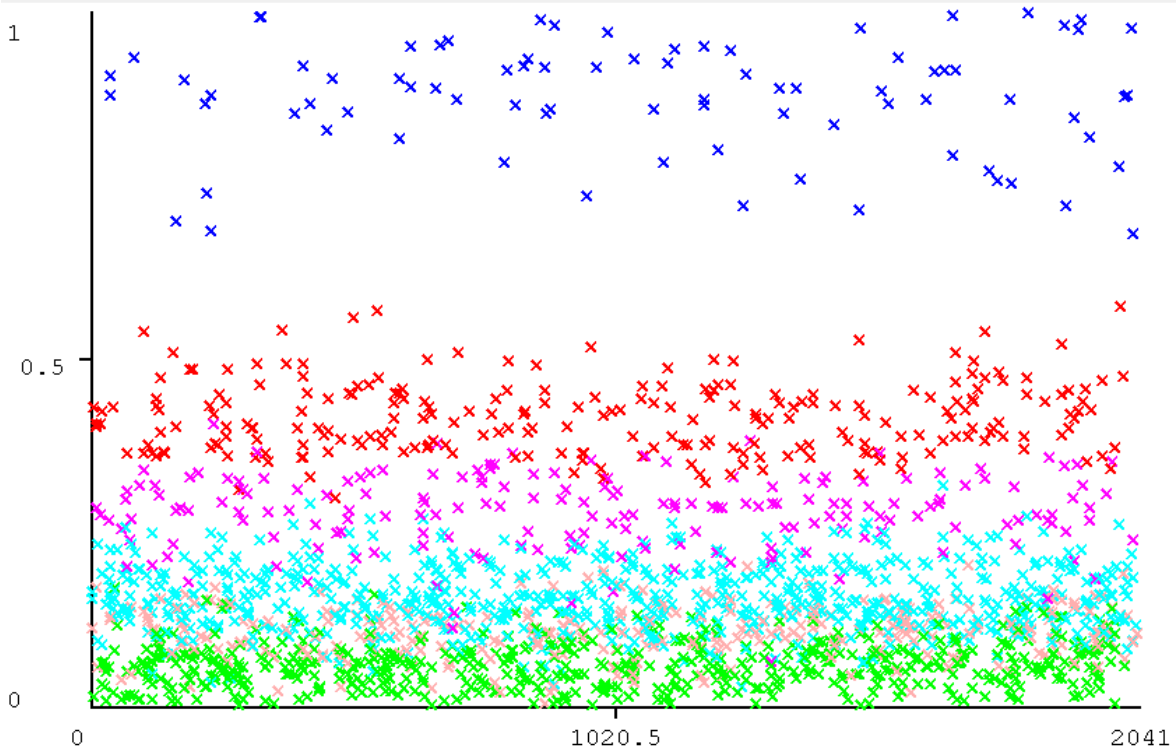

Figure 18 Hierarchical Clustering with Minkowski

After applying the hierarchical cluster algorithm with the parameter specifying 7 clusters and utilizing Minkowski Distance as the distance metric, the resulting classification of the dry beans exhibited some ambiguity in the graph. This selection of seven clusters corresponds to the number of unique dry bean varieties found in the dataset, but the splitting training data method resulted in the formation of six clusters: cluster 0 - 82 instances (4%), Cluster 1 - 242 (12%), Cluster 2 - 546 (27%), Cluster 3 - 687 (34%), Cluster 4 - 296 (14%) and Cluster 5 - 189 (9%).

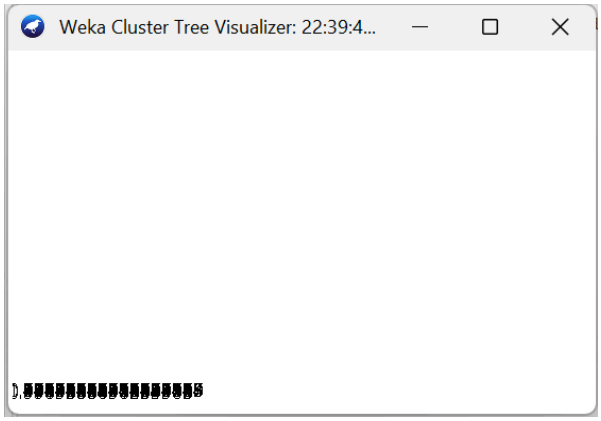

Figure 19 Tree Visualization

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
 1012    1    0    0    0    0    0 | SEKER
  661    0    0    0    0    0    0 | BARBUNYA
  260    0    1    0    0    0    0 | BOMBAY
  815    0    0    0    0    0    0 | CALI
  960    0    0    1    2    1    0 | HOROZ
 1318    0    0    0    0    0    0 | SIRA
 1772    0    0    0    0    0    1 | DERMASON

Cluster 0 <-- DERMASON
Cluster 1 <-- SEKER
Cluster 2 <-- BOMBAY
Cluster 3 <-- No class
Cluster 4 <-- HOROZ
Cluster 5 <-- No class
Cluster 6 <-- No class

Incorrectly clustered instances :       5029.0   73.9015 %
```

Figure 20 Cluster Evaluation of Minkowski Distance (Hierarchical)

After applying the cluster evaluation, the incorrectly clustered instances still maintain at 5029 instances which is 73.90% from the overall number of the dry bean instances. This result is similar to the cluster evaluation by using the Euclidean Distance. There are only 4 assigned by class: Cluster 0 (Dermason), Cluster 1 (Seker), Cluster 2 (Bombay), and Cluster 4 (Horoz). The remaining 3 clusters are not assigned with any class data. These unassigned clusters indicate that there might have been mistakes in the dataset that tricked the clustering algorithm and caused clusters without obvious class relationships to form.

- **Filtered Distance**

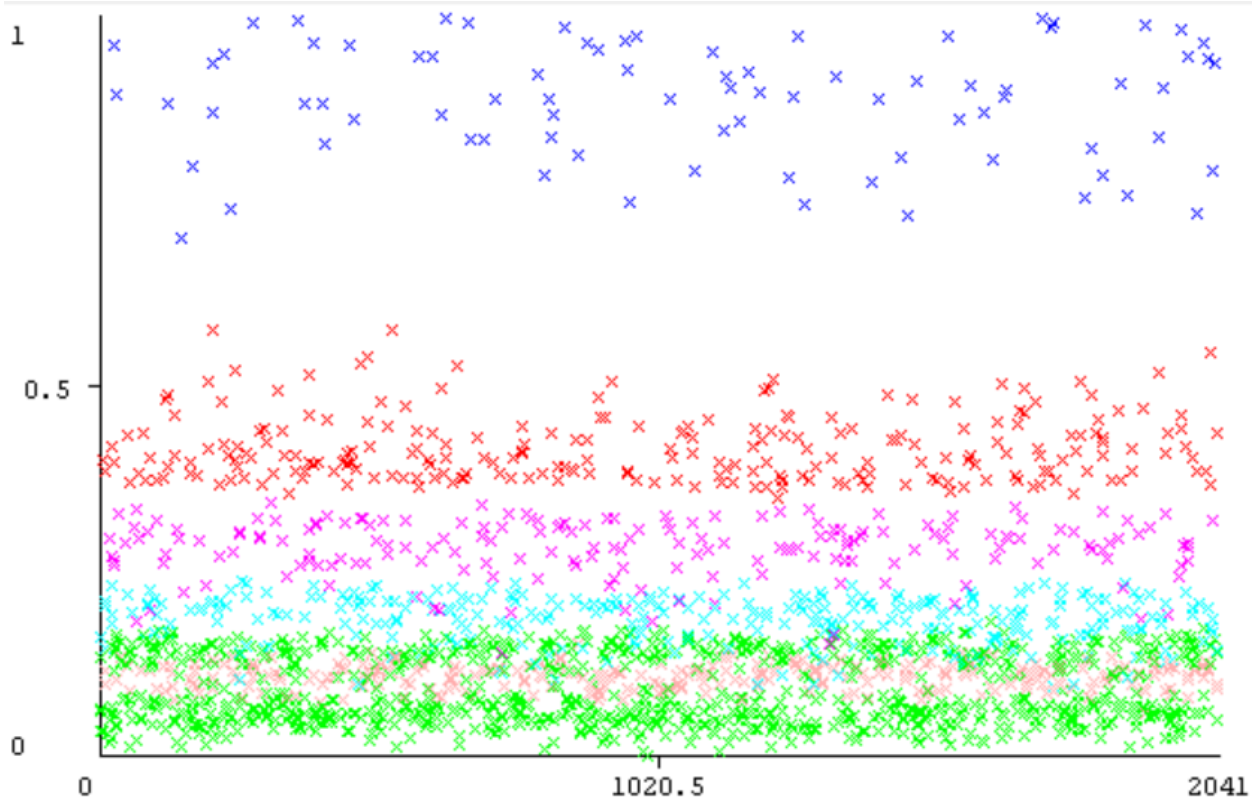

Figure 21 Hierarchical Clustering with Filtered

After applying the hierarchical cluster algorithm with the parameter specifying 7 clusters and utilizing Filtered Distance as the distance metric, the resulting classification of the dry beans exhibited some ambiguity in the graph. This selection of seven clusters corresponds to the number of unique dry bean varieties found in the dataset, but the splitting training data method resulted in the formation of six clusters: cluster 0 - 82 instances (4%), Cluster 1 - 242 (12%), Cluster 2 - 932 (46%), Cluster 3 - 301 (15%), Cluster 4 - 296 (14%) and Cluster 5 - 189 (9%).

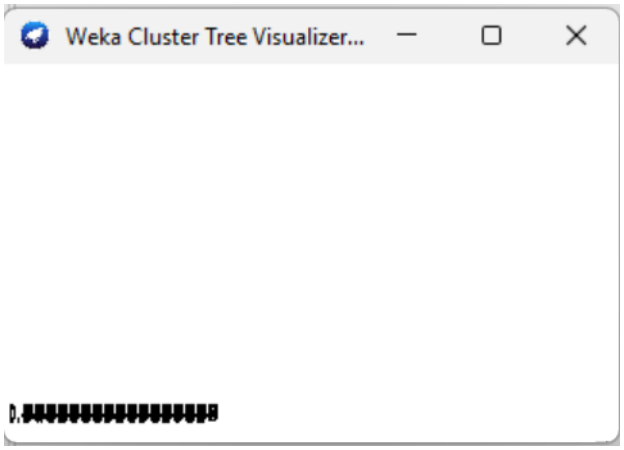

Figure 22 Tree Visualization

```
Class attribute: Class
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
 1010    1    1    1    0    0    0 | SEKER
  661    0    0    0    0    0    0 | BARBUNYA
  261    0    0    0    0    0    0 | BOMBAY
  815    0    0    0    0    0    0 | CALI
  960    0    0    0    3    1    0 | HOROZ
 1318    0    0    0    0    0    0 | SIRA
 1772    0    0    0    0    0    1 | DERMASON

Cluster 0 <-- DERMASON
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- SEKER
Cluster 4 <-- HOROZ
Cluster 5 <-- No class
Cluster 6 <-- No class

Incorrectly clustered instances :       5029.0   73.9015 %
```

Figure 20 Cluster Evaluation of Filtered Distance (Hierarchical)

After applying the cluster evaluation, the incorrectly clustered instances still maintain at 5029 instances which is 73.90% from the overall number of the dry bean instances. This result is similar to the cluster evaluation by using the Euclidean Distance. There are only 4 assigned by class: Cluster 0 (Dermason), Cluster 1 (Seker), Cluster 2 (Bombay), and Cluster 4 (Horoz). The remaining 3 clusters are not assigned with any class data. These unassigned clusters indicate that there might have been mistakes in the dataset that tricked the clustering algorithm and caused clusters without obvious class relationships to form.

Experimental Results and Algorithm Performances of Hierarchical

Cycles	No of cluster	Type of Distance Measures				
		Euclidean	Manhattan	Chebyshev	Minkowski	Filtered
1	7	73.90% Incorrect				
2	7	73.90% Incorrect				
3	7	73.90% Incorrect				
	Average	73.90% Incorrect				
4	7		73.89% Incorrect			
5	7		73.89% Incorrect			
6	7		73.89% Incorrect			
	Average		73.89% Incorrect			
7	7			73.90% Incorrect		
8	7			73.90% Incorrect		
9	7			73.90% Incorrect		
	Average			73.90% Incorrect		
10	7				73.90% Incorrect	
11	7				73.90% Incorrect	

12	7				73.90% Incorrect	
	Average				73.90% Incorrect	
13	7					73.90% Incorrect
14	7					73.90% Incorrect
15	7					73.90% Incorrect
	Average					73.90% Incorrect

Table 2 K-means Clustering Algorithms Performance Measure

Comparison Results

Algorithms	Type of Distance	Performance	Ranking
K-means	Manhattan	3.74% Incorrect	1
K-means	Euclidean	4.22% Incorrect	2
Hierarchical	Manhattan	73.89% Incorrect	3
K-means	Euclidean (Default)	60.25% Incorrect	4
Hierarchical	Euclidean	73.90% Incorrect	5
Hierarchical	Chebyshev	73.90% Incorrect	6
Hierarchical	Minkowski	73.90% Incorrect	7
Hierarchical	Filtered	73.90% Incorrect	8

Table 3 Table of Comparison

Above table is the collection of the both k-means and hierarchical clustering results and provides insightful information on how well each performed in clustering the dry bean dataset. Therefore, we can conclude that the better algorithm to use is K-means clustering with Manhattan Distance using K-means initialization method. The Manhattan Distance measure seems to more fully represent the intrinsic structure of the data with its focus on distance along axes.

Conversely, the algorithms with worst performance have 4 which are Euclidean, Chebyshev, Minkowski and Filtered in hierarchical clustering. These algorithms utilizing yielded clusters with varying degrees of overlap and unclear boundaries.