Evaluation of Clustering Algorithms for Inferring Inversion Genotypes

Matthew Aleck[1], Ronald J. Nowling[1]

[1] Electrical Engineering and Computer Science, Milwaukee School of Engineering

Abstract

Genotypes of large polymorphic inversions can be inferred from single nucleotide polymorphism (SNP) data. Recombination is repressed in inversion regions, allowing alleles to remain private to a single inversion orientation. For large inversions, there are sufficient numbers of private alleles such that samples segregate according to inversion genotypes.

In principal component analysis (PCA), samples form groups according to their inversion genotypes. Clustering algorithms can be used to determine the distinct groups and their members. Depending on the particular data set, however, inversions cause different patterns. For example, *D. melanogaster* samples form three distinct, circular clusters with only a handful of outliers when analyzing SNPs on the 2L chromosome arm, while the origin separates *An. gambiae* samples by their 2Rb inversion genotypes but the clusters are not clearly defined since the intra- and inter-cluster variances are nearly equal.

Our aim was to characterize the different cluster patterns caused by large polymorphic inversions and determine which clustering algorithm is best suited for each type of pattern. We performed PCA of SNPs from chromosome arms of *An. gambiae*, *An. coluzzii*, and *D. melanogaster*. We developed a set of descriptive criteria which we applied to characterize the observed patterns for each inversion. We then applied several clustering algorithms to the data sets and evaluated their predictions against the known genotypes for the samples. We use our characterizations and clustering results to recommend the most appropriate clustering algorithms for each type of pattern.

Clustering Algorithms

DBSCAN

- User enters epsilon and minimum points
- Epsilon is the maximum distance another point can be from the current point to be in the same cluster.
- Minimum points is the minimum number of points that are required for a cluster

K-Means

- User enters the number of desired clusters, n
- Partitions the global space into n regions of approximately equal size

Cluster Classifications

- **Density**: Are the points within the same cluster close together (Dense) or far apart (Loose)?
- **Separation:** Is the distance between cluster substantial (Well Separated) or is the distance between cluster negligible (Not Well Separated)?
- **Even Size**: Do the clusters have a similar number of points (Evenly Sized) or a dissimilar number of points (Not Evenly Sized)?
- **Circular / Not Circular**: Is the cluster (Circular) or (Not Circular)?

Conclusions

Density: K-Means performed best on Dense clusters and DBSCAN performed well on both Dense and Loose Clusters

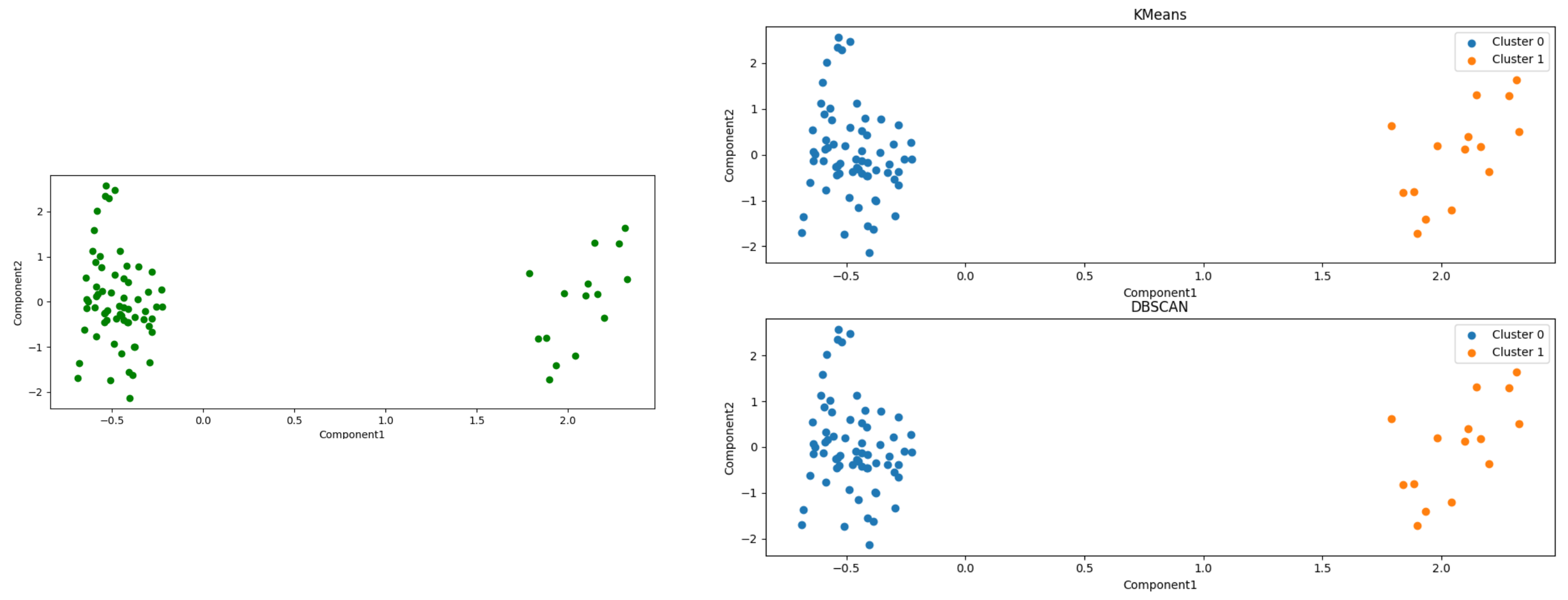
Separation: Both K-Means and DBSCAN performed well on clusters that were Well Separated, but K-Means also performed well on clusters that were Not Well Separated.

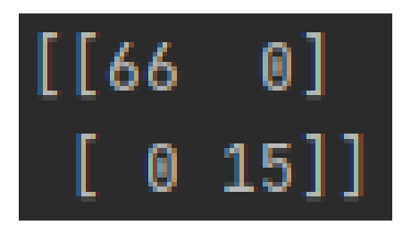
Even Size: Both K-Means as DBSCAN performed well regardless if the clusters were Even Sized or Not Even Sized.

Circular / Not Circular: K-Means performed best on Circular clusters and DBSCAN performed well on both Circular and Not Circular clusters.

Given the clusters are Well Separated, DBSCAN performs the same or better than K-Means for the other cluster classifications. In conclusion, DBSCAN is the more appropriate clustering algorithm for classifying inversion genotypes.

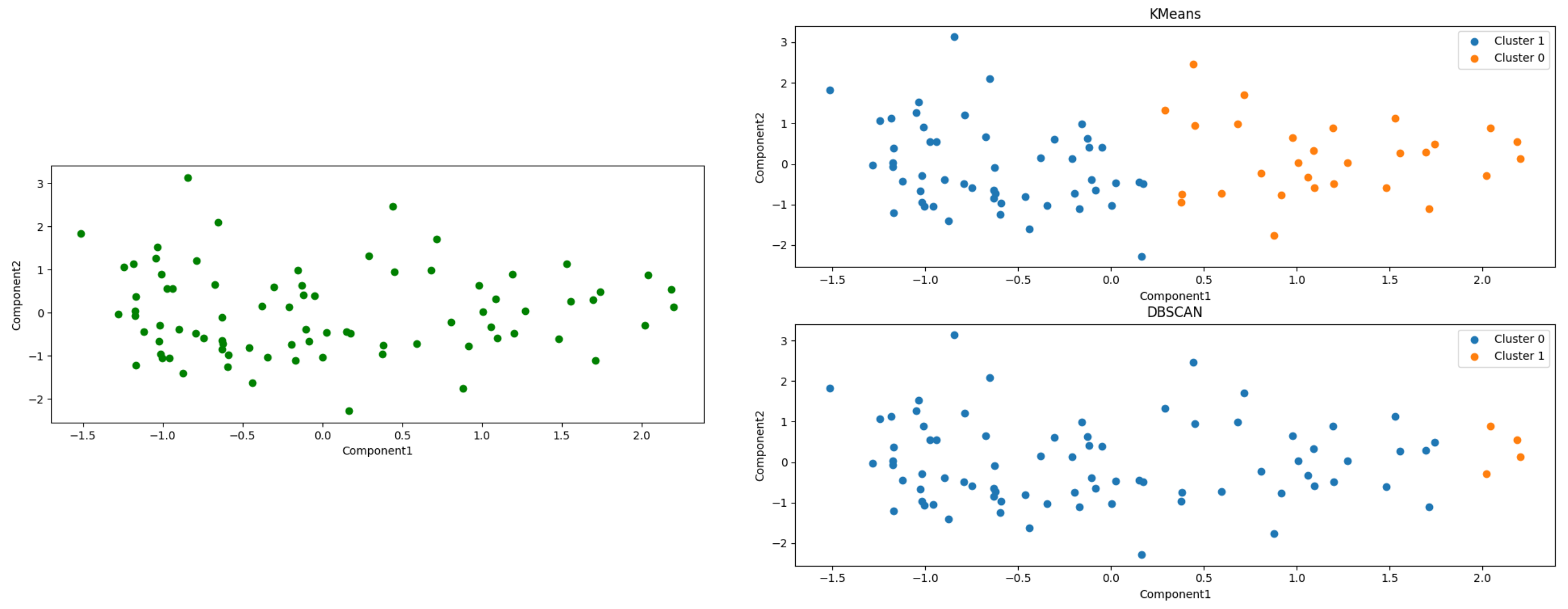
Results

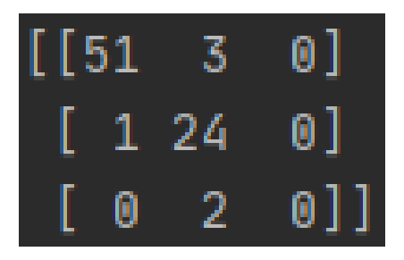




```
[[66  0]
 [ 0 15]]
```

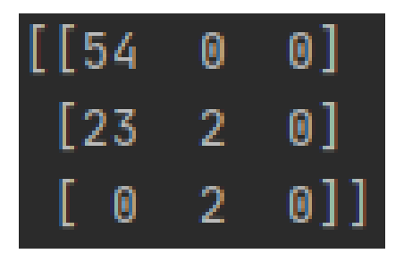
Accuracy: 100%





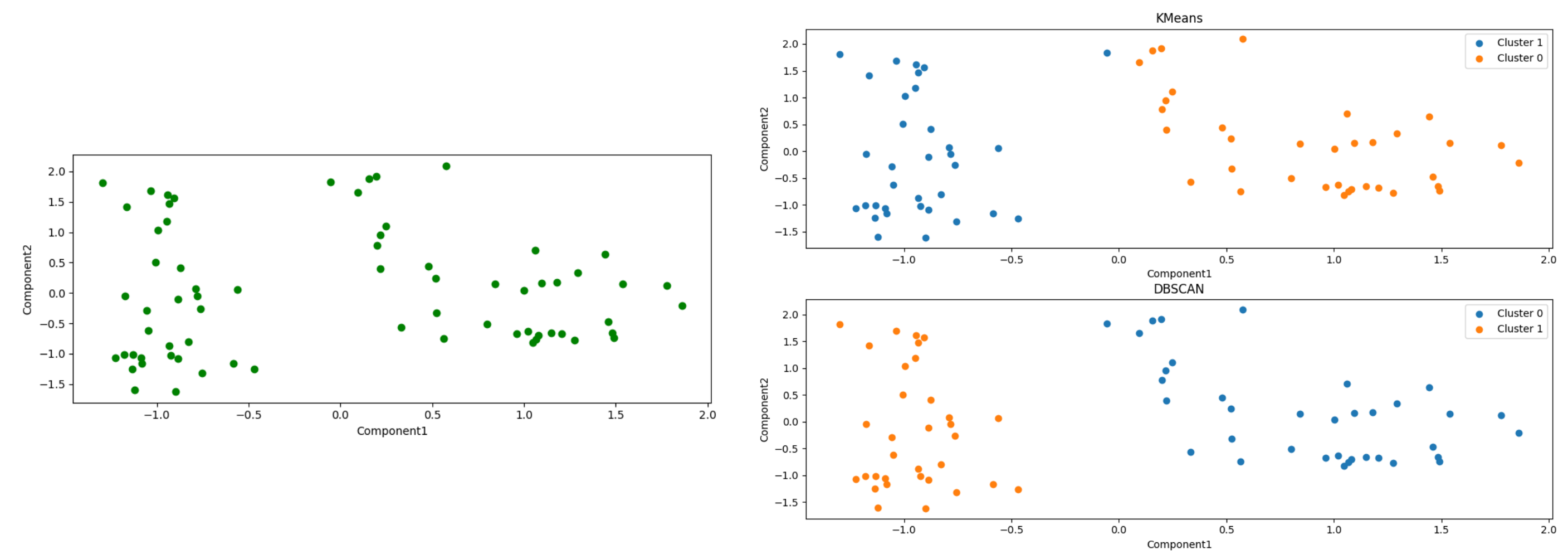
```
[[51  3  0]
 [ 1 24  0]
 [ 0  2  0]]
```

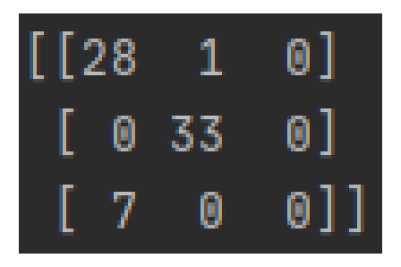
Accuracy: 92.6%



```
[[54  0  0]
 [23  2  0]
 [ 0  2  0]]
```

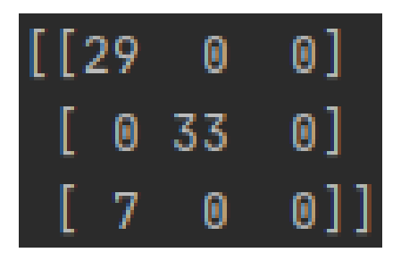
Accuracy: 69.1%





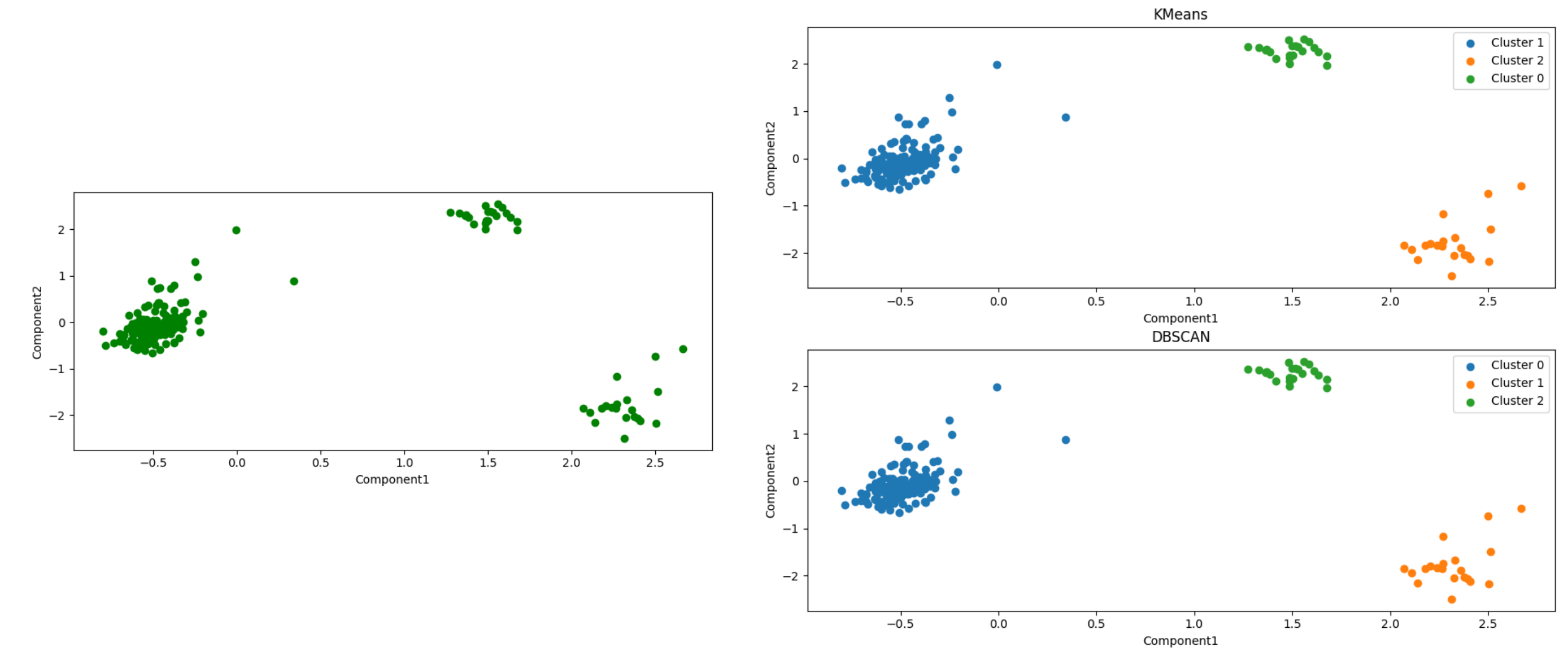
```
[[28  1  0]
 [ 0 33  0]
 [ 7  0  0]]
```

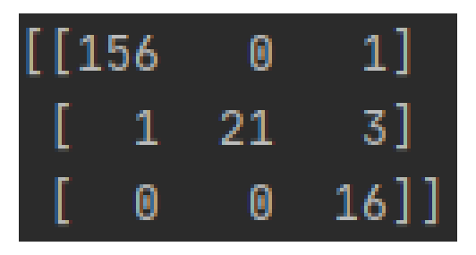
Accuracy: 88.4%



```
[[29  0  0]
 [ 0 33  0]
 [ 7  0  0]]
```

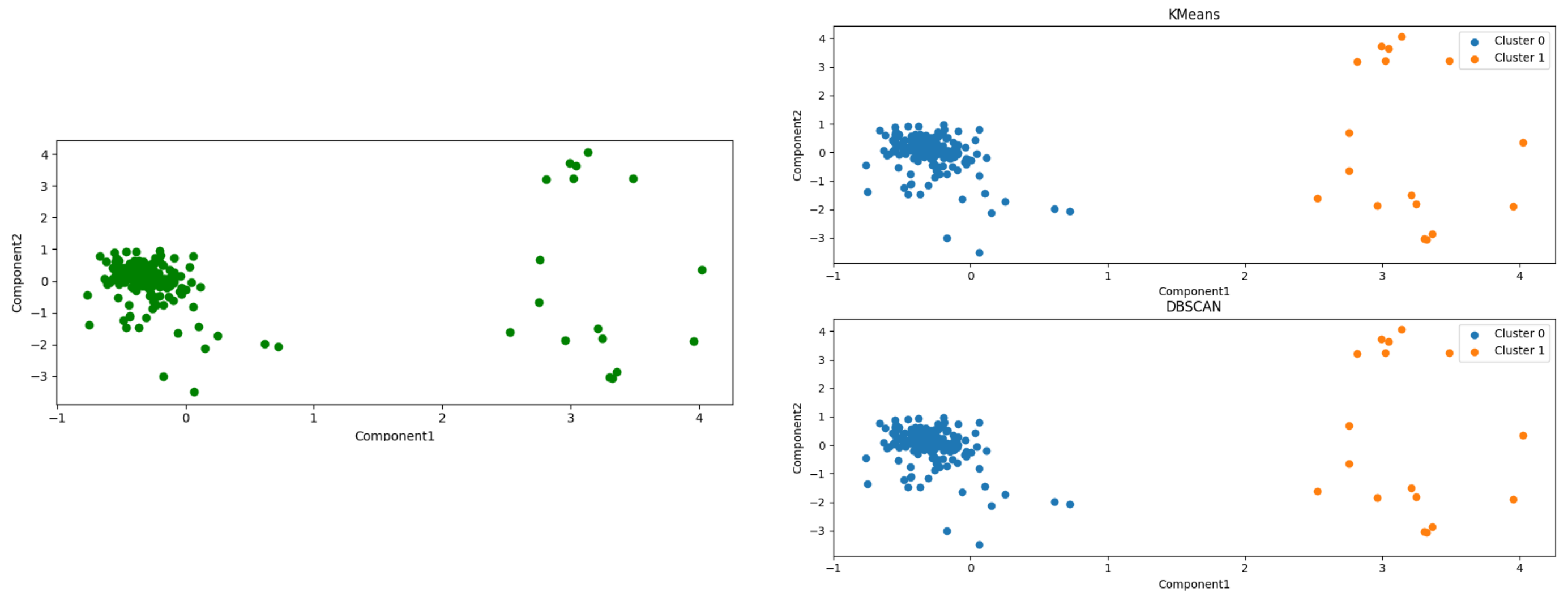
Accuracy: 89.9%

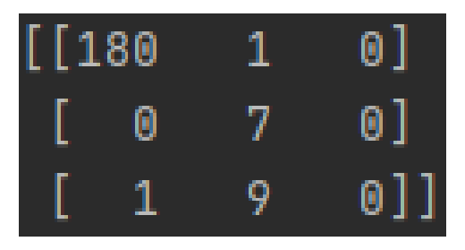




```
[[156   0   1]
 [  1  21   3]
 [  0   0  16]]
```

Accuracy: 97.4%





```
[[180   1   0]
 [  0   7   0]
 [  1   9   0]]
```

Accuracy: 94.4%

Discussion

2L Gambiae

Pre-Clustering (Left Plot)

Classifications: Well Separated and Not Evenly Sized

Cluster 0: Dense and Not Circular, **Cluster 1**: Loose and Circular

Post-Clustering (Right Plot)

K-Means: Able to accurately classify clusters 0 and 1 because they are Well Separated.

Parameters: Number of Clusters (n) = 2

DBSCAN: Able to accurately classify clusters 0 and 1 because they are Well Separated.

Parameters: Epsilon = 0.4, Minimum Points = 4

2R Gambiae

Pre-Clustering (Left Plot)

Classifications: Not Well Separated and Not Evenly Sized

Cluster 0: Loose and Not Circular

Post-Clustering (Right Plot)

K-Means: Able to more accurately classify clusters 0 and 1 than DBSCAN because despite the clusters being Not Well Separated it divided the space in half, which happened to be an accurate classification. Didn't detect cluster 2 at all.

Parameters: Number of Clusters (n) = 2

DBSCAN: Unable to accurately classify clusters 0 and 1 because the clusters are Not Well Separated, so setting an appropriate distance between points to classify the clusters is difficult. Didn't detect cluster 2 at all. Epsilon = 0.25, Minimum Points = 4

2R Coluzzii

Pre-Clustering (Left Plot)

Classifications: Well Separated and Evenly Sized

Cluster 0: Loose and Not Circular, **Cluster 1**: Loose and Not Circular

Post-Clustering (Right Plot)

K-Means: Able to accurately classify clusters 0 and 1 because they are Well Separated. Didn't detect cluster 2 at all.

Parameters: Number of Clusters (n) = 2

DBSCAN: Able to more accurately classify clusters 0 and 1 because they are Well Separated and classifies clusters based on distance between points. Didn't detect cluster 2 at all.

Parameters: Epsilon = 0.4, Minimum Points = 4

2L DGRP2

Pre-Clustering (Left Plot)

Classifications: Well Separated and Not Evenly Sized

Cluster 0: Dense and Circular, **Cluster 1**: Dense and Circular, **Cluster 2**: Dense and Circular

Post-Clustering (Right Plot)

K-Means: Able to accurately classify clusters 0 and 1 because they are Well Separated.

Parameters: Number of Clusters (n) = 2

DBSCAN: Able to accurately classify clusters 0 and 1 because they are Well Separated.

Parameters: Epsilon = 0.35, Minimum Points = 4

2R DGRP2

Pre-Clustering (Left Plot)

Classifications: Well Separated and Not Evenly Sized

Cluster 0: Dense and Circular, **Cluster 1**: Loose and Not Circular

Post-Clustering (Right Plot)

K-Means: Able to accurately classify clusters 0 and 1 because they are Well Separated. Didn't detect cluster 2 at all.

Parameters: Number of Clusters (n) = 2

DBSCAN: Able to accurately classify clusters 0 and 1 because they are Well Separated. Didn't detect cluster 2 at all.

Parameters: Epsilon = 0.55, Minimum Points = 4