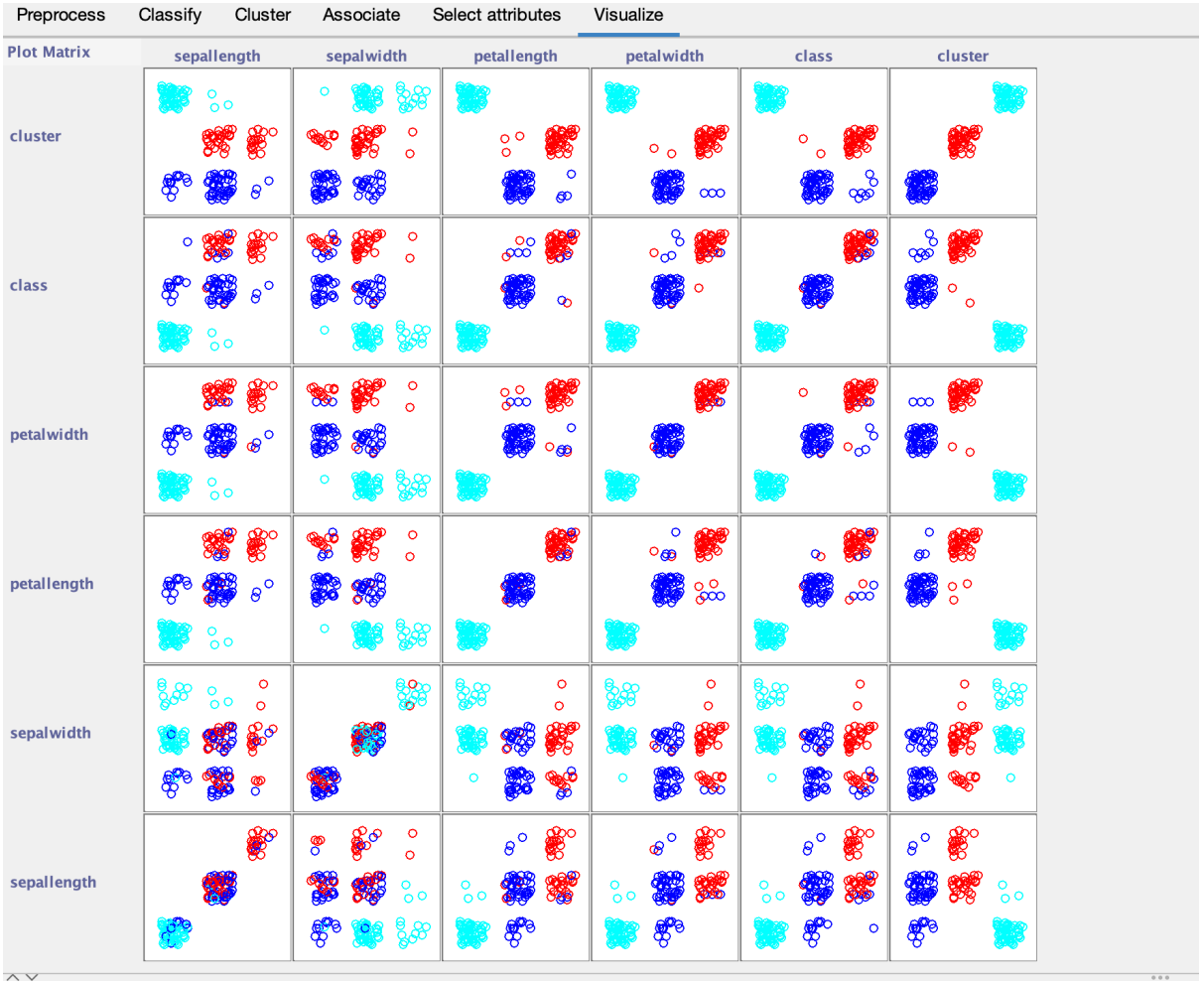
Data Mining Lab 2

Siddhesh Sreedar (sidsr770)
Marijn (marja987)

Clustering

Below is the output using SimpleKMeans with three clusters and seed value ten.

```
Number of iterations: 3
Within cluster sum of squared errors: 96.0

Initial starting points (random):

Cluster 0: '\'(5.5-6.7]\'','\'(2.8-3.6]\'','\'(2.966667-4.933333]\'','\'(0.9-1.7]\'',cluster1
Cluster 1: '\'(6.7-inf)\'','\'(2.8-3.6]\'','\'(4.933333-inf)\'','\'(1.7-inf)\'',cluster2
Cluster 2: '\'(-inf-5.5]\'','\'(3.6-inf)\'','\'(-inf-2.966667]\'','\'(-inf-0.9]\'',cluster3

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                Full Data                   0                   1                   2
                           (150.0)                (55.0)              (45.0)              (50.0)
=============================================================================================================
sepallength                        '(5.5-6.7]'          '(5.5-6.7]'          '(5.5-6.7]'          '(-inf-5.5]'
sepalwidth                         '(2.8-3.6]'          '(-inf-2.8]'         '(2.8-3.6]'          '(2.8-3.6]'
petallength              '(2.966667-4.933333]' '(2.966667-4.933333]'      '(4.933333-inf)'     '(-inf-2.966667]'
petalwidth                         '(0.9-1.7]'          '(0.9-1.7]'          '(1.7-inf)'          '(-inf-0.9]'
cluster                              cluster1             cluster1             cluster2             cluster3



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       55 ( 37%)
1       45 ( 30%)
2       50 ( 33%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0  0 50 | Iris-setosa
 48  2  0 | Iris-versicolor
  7 43  0 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-virginica
Cluster 2 <-- Iris-setosa

Incorrectly clustered instances :       9.0       6      %
```



Based on the association, we are only able to find rules for cluster three, some of them are:

- petalwidth = (-inf-0.9] → cluster 3
- petallength = (-inf-2.966667] → cluster 3
- petallength = (-inf-2.966667] and petalwidth = (-inf-0.9] → cluster 3

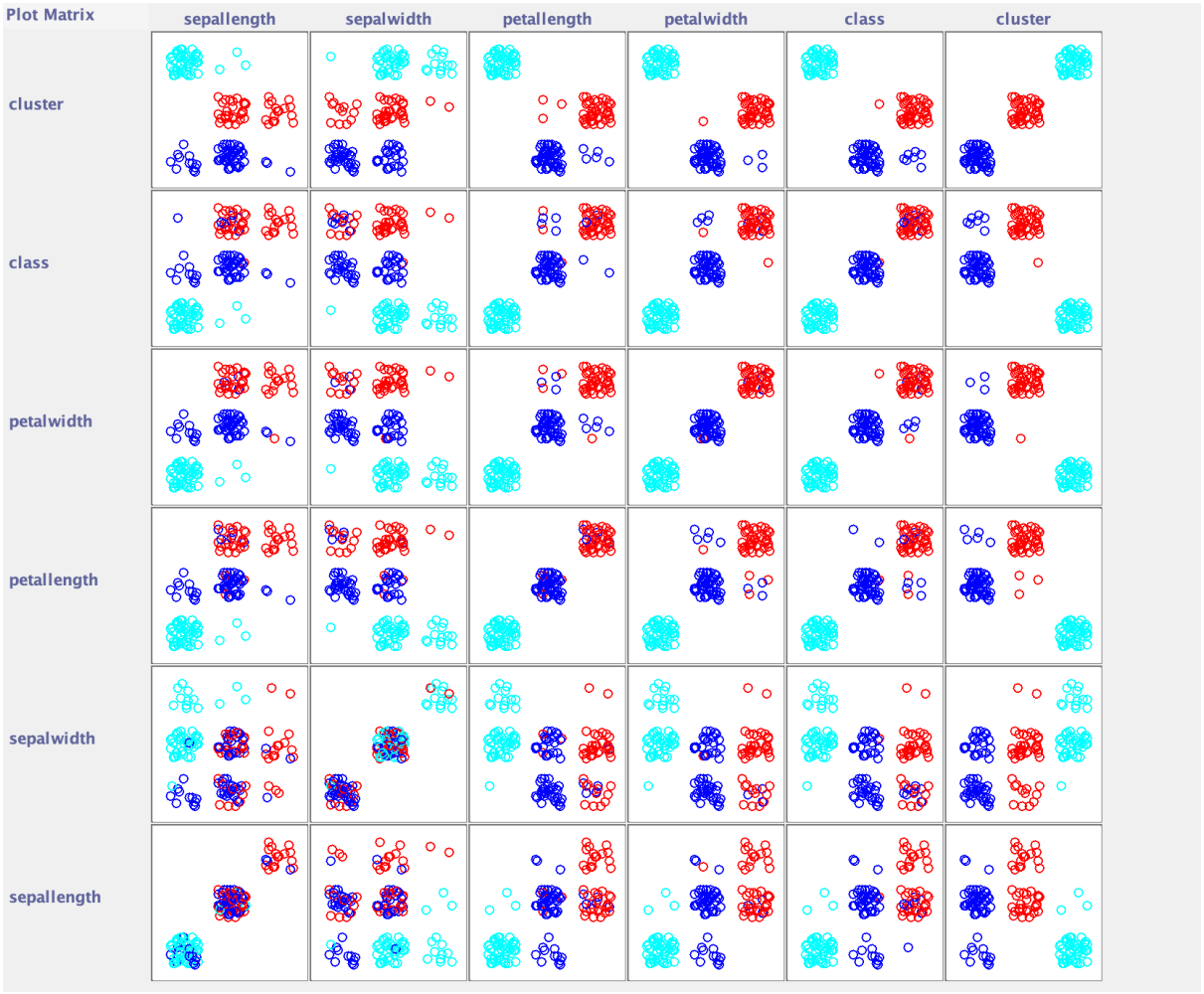
More rules were found, but we have added only 3 here.

Now, we change only the clustering algorithm:
We choose the “Make density-based clustering” algorithm.

Below is the output for it:

```
Number of iterations: 3
Within cluster sum of squared errors: 105.0

Initial starting points (random):

Cluster 0: '\'(5.5-6.7]\'','\'(2.8-3.6]\'','\'(2.966667-4.933333]\'','\'(0.9-1.7]\'',Iris-versicolor
Cluster 1: '\'(6.7-inf)\'','\'(2.8-3.6]\'','\'(4.933333-inf)\'','\'(1.7-inf)\'',Iris-virginica
Cluster 2: '\'(-inf-5.5]\'','\'(3.6-inf)\'','\'(-inf-2.966667]\'','\'(-inf-0.9]\'',Iris-setosa

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                    Full Data                0                   1                   2
                               (150.0)             (57.0)              (43.0)              (50.0)
=============================================================================================================
sepallength                 '(5.5-6.7]'          '(5.5-6.7]'         '(5.5-6.7]'         '(-inf-5.5]'
sepalwidth                  '(2.8-3.6]'          '(-inf-2.8]'        '(2.8-3.6]'         '(2.8-3.6]'
petallength     '(2.966667-4.933333]' '(2.966667-4.933333]'  '(4.933333-inf)'  '(-inf-2.966667]'
petalwidth                  '(0.9-1.7]'          '(0.9-1.7]'         '(1.7-inf)'         '(-inf-0.9]'
class                       Iris-setosa      Iris-versicolor      Iris-virginica          Iris-setosa



Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3791

Attribute: sepallength
Discrete Estimator. Counts =  13 43 4  (Total = 60)
Attribute: sepalwidth
Discrete Estimator. Counts =  35 24 1  (Total = 60)
Attribute: petallength
Discrete Estimator. Counts =  1 53 6  (Total = 60)
Attribute: petalwidth
Discrete Estimator. Counts =  1 54 5  (Total = 60)
Attribute: class
Discrete Estimator. Counts =  1 51 8  (Total = 60)

Cluster: 1 Prior probability: 0.2876

Attribute: sepallength
Discrete Estimator. Counts =  1 27 18  (Total = 46)
Attribute: sepalwidth
Discrete Estimator. Counts =  13 30 3  (Total = 46)
Attribute: petallength
Discrete Estimator. Counts =  1 3 42  (Total = 46)
Attribute: petalwidth
Discrete Estimator. Counts =  1 2 43  (Total = 46)
Attribute: class
Discrete Estimator. Counts =  1 1 44  (Total = 46)
```

```
Cluster: 2 Prior probability: 0.3333

Attribute: sepallength
Discrete Estimator. Counts =  48 4 1  (Total = 53)
Attribute: sepalwidth
Discrete Estimator. Counts =  2 37 14  (Total = 53)
Attribute: petallength
Discrete Estimator. Counts =  51 1 1  (Total = 53)
Attribute: petalwidth
Discrete Estimator. Counts =  51 1 1  (Total = 53)
Attribute: class
Discrete Estimator. Counts =  51 1 1  (Total = 53)


Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       54 ( 36%)
1       46 ( 31%)
2       50 ( 33%)


Log likelihood: -2.83633


Class attribute: cluster
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 53  3  0 | cluster1
  1 43  0 | cluster2
  0  0 50 | cluster3

Cluster 0 <-- cluster1
Cluster 1 <-- cluster2
Cluster 2 <-- cluster3

Incorrectly clustered instances :       4.0       2.6667 %
```



Based on the association, we are able to find rules for cluster 3 and cluster 1 but not any for cluster 2, some of them are:

For cluster 1:

- petallength = (2.966667-4.933333] and petalwidth = (0.9-1.7] → cluster 1

Only 1 rule was found.

For cluster 3:

- sepallength = (-inf-5.5] and petallength = (-inf-2.966667] → cluster 3
- petallength = (-inf-2.966667] and petalwidth = (-inf-0.9] → cluster 3
- sepallength = (-inf-5.5] and petalwidth = (-inf-0.9] → cluster 3

More rules were found, but we have added only three here.

Since we found a rule for cluster 1, we shall build on the “Make density-based clustering” algorithm. We will now change the number of discretized bins from three to five.

Below is the output for it:

```
Number of iterations: 5
Within cluster sum of squared errors: 217.0

Initial starting points (random):

Cluster 0: '\'(5.74-6.46]\'','\'(2.48-2.96]\'','\'(4.54-5.72]\'','\'(1.06-1.54]\'',Iris-versicolor
Cluster 1: '\'(5.74-6.46]\'','\'(2.48-2.96]\'','\'(3.36-4.54]\'','\'(1.06-1.54]\'',Iris-versicolor
Cluster 2: '\'(6.46-7.18]\'','\'(2.96-3.44]\'','\'(4.54-5.72]\'','\'(2.02-inf)\'',Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data              0              1              2
                          (150.0)         (59.0)         (41.0)         (50.0)
================================================================================
sepallength          '(5.74-6.46]'   '(5.74-6.46]'   '(5.02-5.74]'   '(-inf-5.02]'
sepalwidth           '(2.96-3.44]'   '(2.96-3.44]'   '(2.48-2.96]'   '(2.96-3.44]'
petallength          '(-inf-2.18]'   '(4.54-5.72]'   '(3.36-4.54]'   '(-inf-2.18]'
petalwidth           '(-inf-0.58]'   '(1.54-2.02]'   '(1.06-1.54]'   '(-inf-0.58]'
class                  Iris-setosa  Iris-virginica Iris-versicolor     Iris-setosa



Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3922

Attribute: sepallength
Discrete Estimator. Counts =  2 3 27 20 12  (Total = 64)
Attribute: sepalwidth
Discrete Estimator. Counts =  2 22 35 4 1  (Total = 64)
Attribute: petallength
Discrete Estimator. Counts =  1 1 3 42 17  (Total = 64)
Attribute: petalwidth
Discrete Estimator. Counts =  1 1 8 30 24  (Total = 64)
Attribute: class
Discrete Estimator. Counts =  1 10 51  (Total = 62)

Cluster: 1 Prior probability: 0.2745

Attribute: sepallength
Discrete Estimator. Counts =  4 19 16 6 1  (Total = 46)
Attribute: sepalwidth
Discrete Estimator. Counts =  10 25 9 1 1  (Total = 46)
Attribute: petallength
Discrete Estimator. Counts =  1 4 33 7 1  (Total = 46)
Attribute: petalwidth
Discrete Estimator. Counts =  1 8 35 1 1  (Total = 46)
Attribute: class
Discrete Estimator. Counts =  1 42 1  (Total = 44)

Cluster: 2 Prior probability: 0.3333
```

```
Attribute: sepallength
Discrete Estimator. Counts =  29 22 2 1 1  (Total = 55)
Attribute: sepalwidth
Discrete Estimator. Counts =  2 2 28 18 5  (Total = 55)
Attribute: petallength
Discrete Estimator. Counts =  51 1 1 1 1  (Total = 55)
Attribute: petalwidth
Discrete Estimator. Counts =  50 2 1 1 1  (Total = 55)
Attribute: class
Discrete Estimator. Counts =  51 1 1  (Total = 53)


Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       58 ( 39%)
1       42 ( 28%)
2       50 ( 33%)


Log likelihood: -4.44178


Class attribute: cluster
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 57  7  0 | cluster1
  1 33  0 | cluster2
  0  2 50 | cluster3

Cluster 0 <-- cluster1
Cluster 1 <-- cluster2
Cluster 2 <-- cluster3

Incorrectly clustered instances :       10.0      6.6667 %
```

Based on the association, we were able to find rules for all the three clusters:

Cluster 1:

- sepallength = (5.74-6.46] and petallength = (4.54-5.72] → cluster 1
- petallength = (4.54-5.72] → cluster 1

Only two rules found.

Cluster 2:

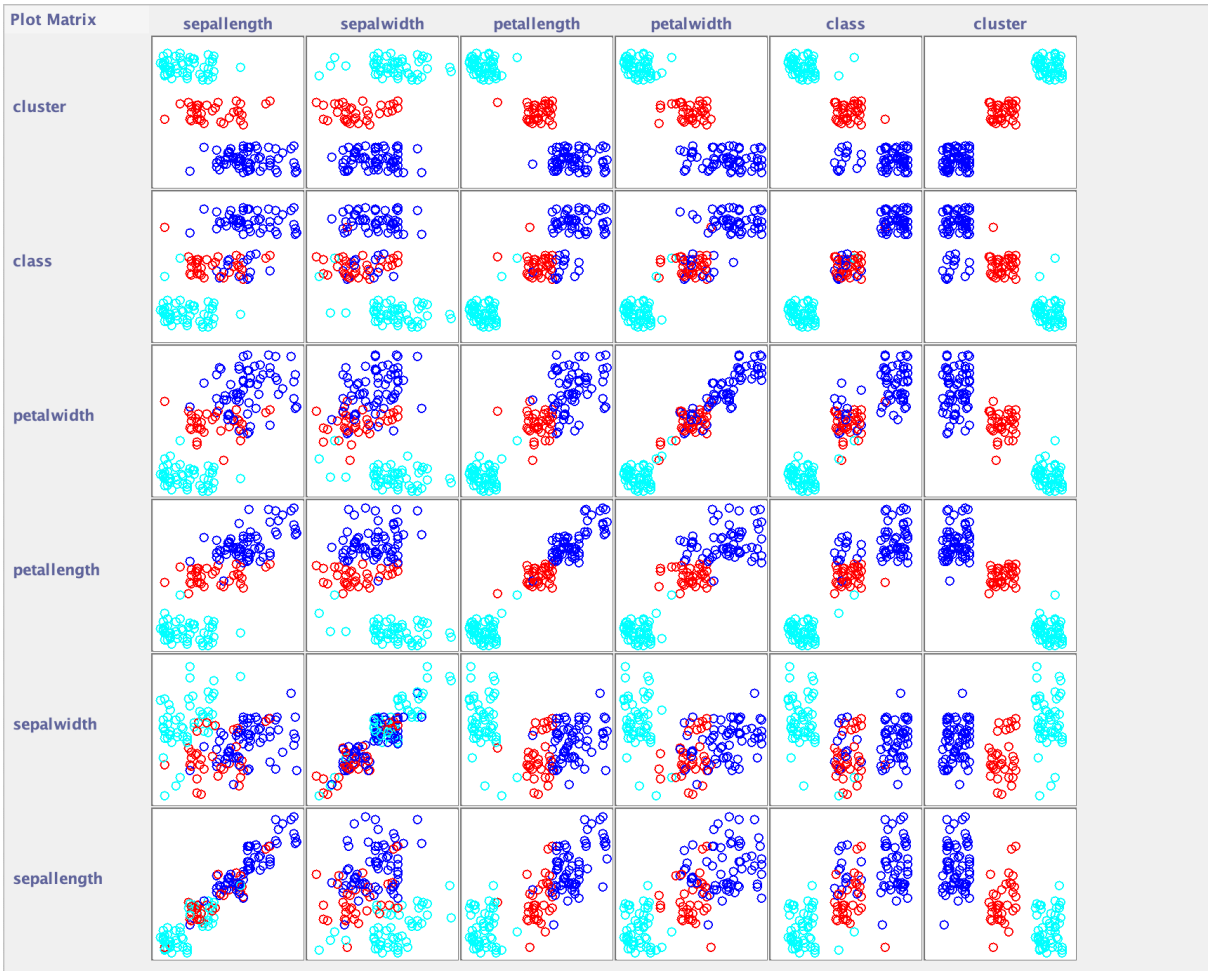
- petallength = (3.36-4.54] and petalwidth = (1.06-1.54] → cluster 2

Only one rule was found.

Cluster 3:

- petallength = (-inf-2.18] → cluster 3
- petalwidth = (-inf-0.58] → cluster 3
- petallength = (-inf-2.18] and petalwidth = (-inf-0.58] → cluster 3

More rules were found but added three.



Explanation

We first used the SimpleKMeans algorithm using three bins to discretize the variables and were only able to find rules for one cluster. We then expanded to the density-based clustering algorithm and were able to find rules for two clusters. We further built on this by expanding the number of bins to discretize the variables from three to five and found rules for all three clusters. We hypothesize that the density-based algorithm performed better because it is more robust to varying cluster shapes while SimpleKMeans requires

well-separated clusters. Further, having more bins on the variables allows for a more complex representation of the true data, which in turn helps create better clusters and more specific rules.