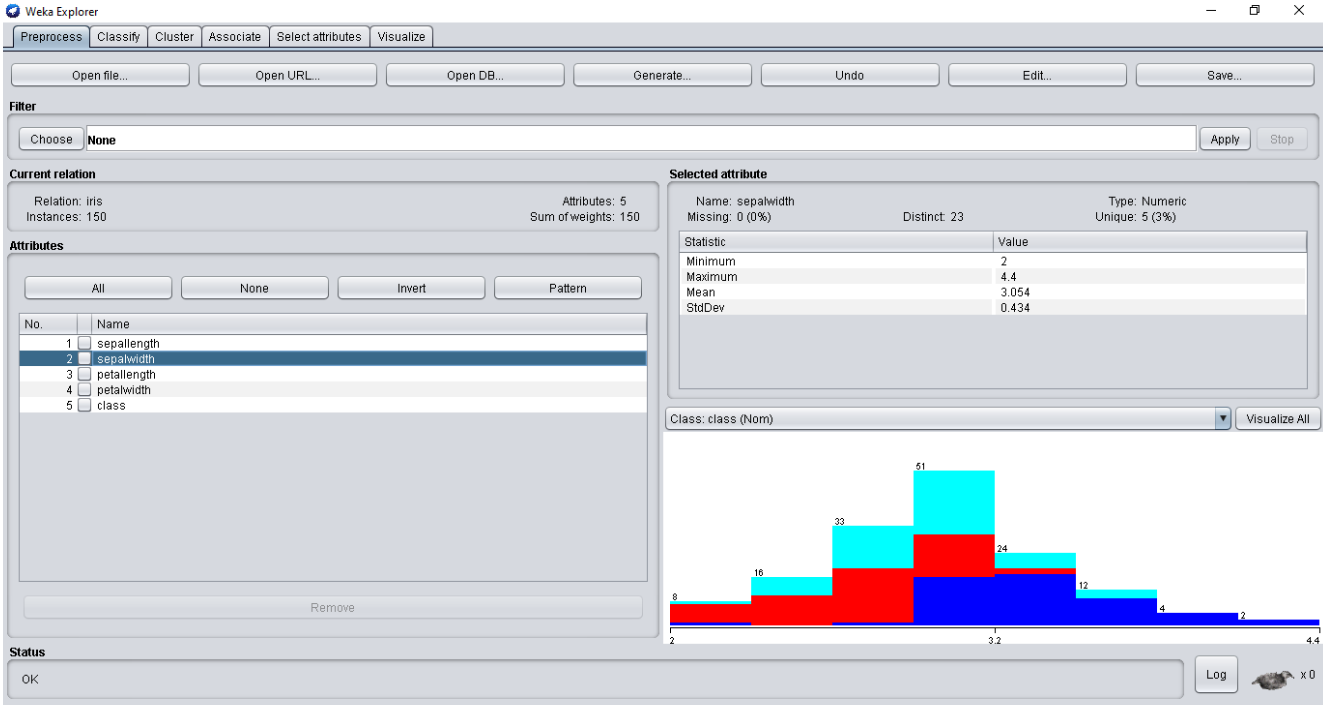
CO328 - Machine Learning
Weka Lab - Part 2
E/16/319 Rathnayake R.P.V.N

01. Load iris.arff



Filter: None

Current relation

Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

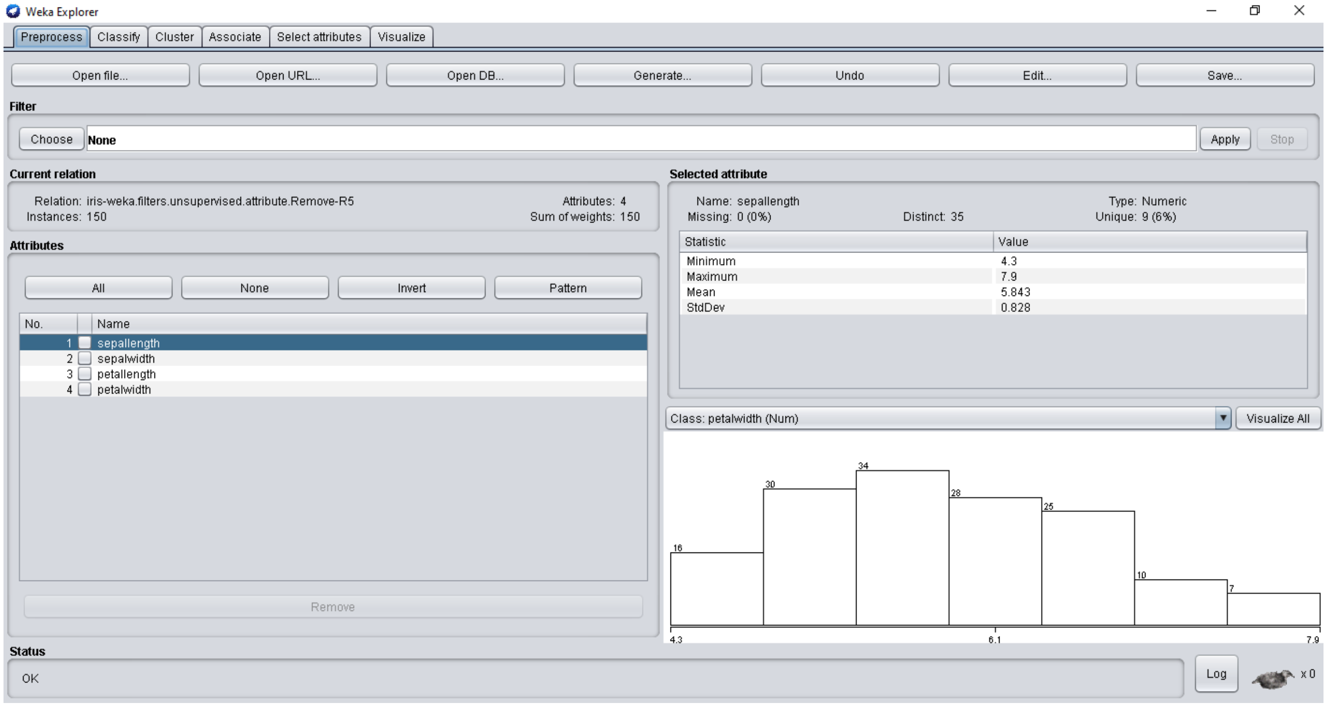
Name: sepalwidth
Missing: 0 (0%)
Distinct: 23
Type: Numeric
Unique: 5 (3%)

Statistic	Value
Minimum	2
Maximum	4.4
Mean	3.054
StdDev	0.434

Class: class (Nom)

Status

OK



Filter: None

Current relation

Relation: iris-weka.filters.unsupervised.attribute.Remove-R5
Instances: 150
Attributes: 4
Sum of weights: 150

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth

Selected attribute

Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: petalwidth (Num)

Status

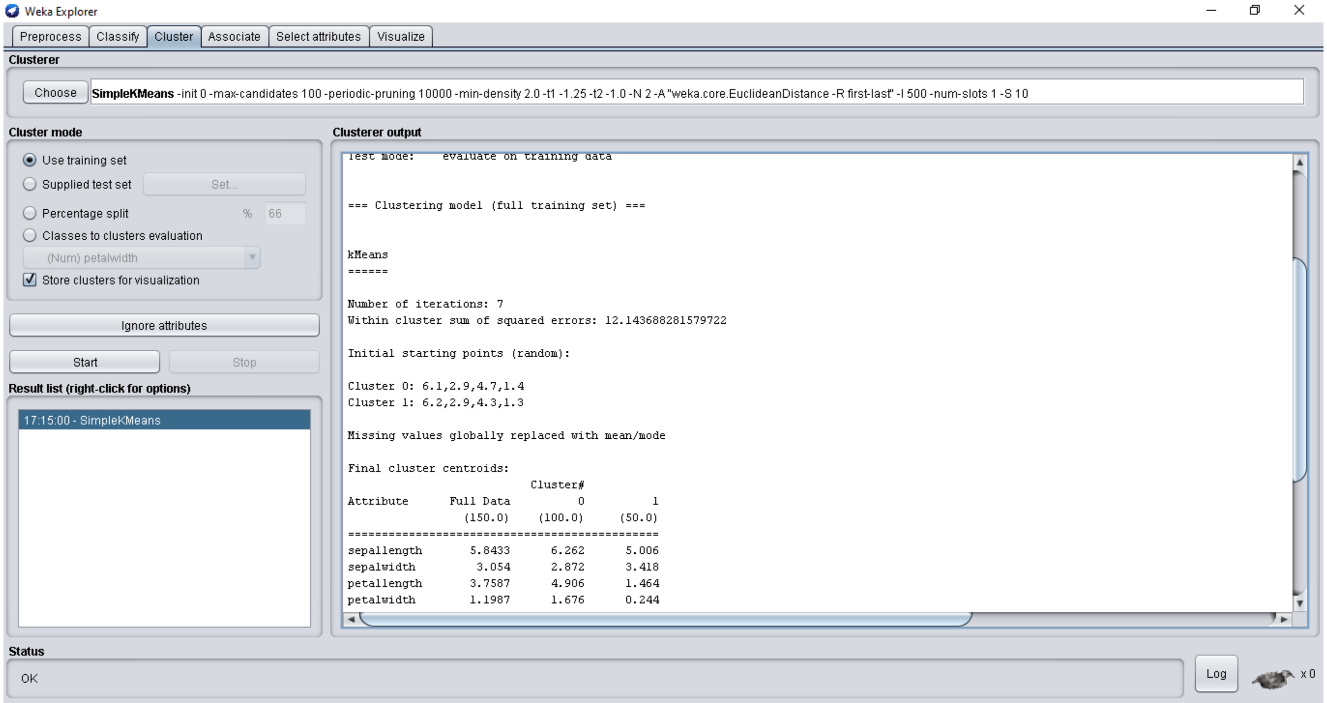
OK

02. Build C4.5 decision tree - training set

03.

Seed is just a random numbers seed. Once seed is fixed, even a randomized algorithm behaves deterministically. KMeans is not deterministic, so if you want repeatable results - you fix a seed. However there is completely no relation between exact value of the seed and the results of KMeans clustering.

04.



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) petalwidth
☑ Store clusters for visualization

17:15:00 - SimpleKMeans

```
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 12.143688281579722

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1
                 (150.0)    (100.0)     (50.0)
==============================================
sepallength       5.8433      6.262      5.006
sepalwidth         3.054      2.872      3.418
petallength       3.7587      4.906      1.464
petalwidth        1.1987      1.676      0.244
```

Status: OK

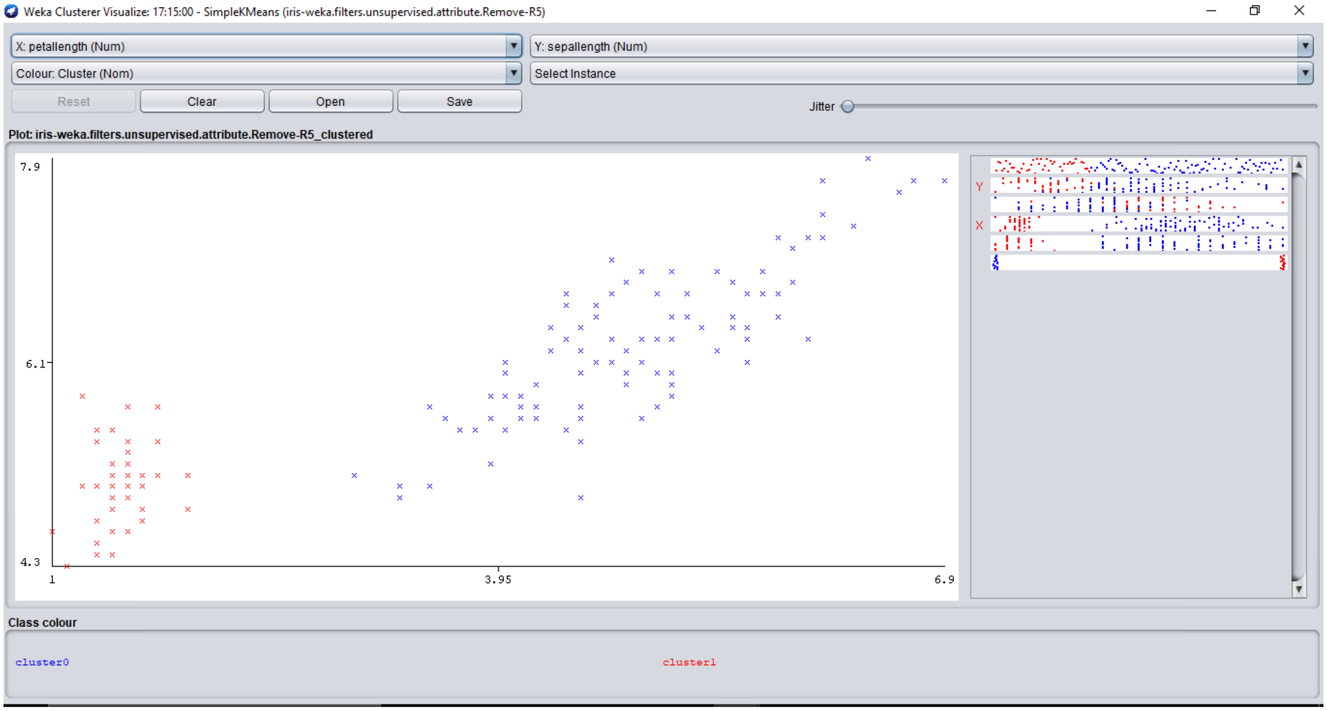
Within cluster sum of squared errors: 12.143688281579722

Clustered Instances

0	100 (67%)
1	50 (33%)

Mean absolute error	0.0464
Root mean squared error	0.1965
Total Number of Instances	20

05.

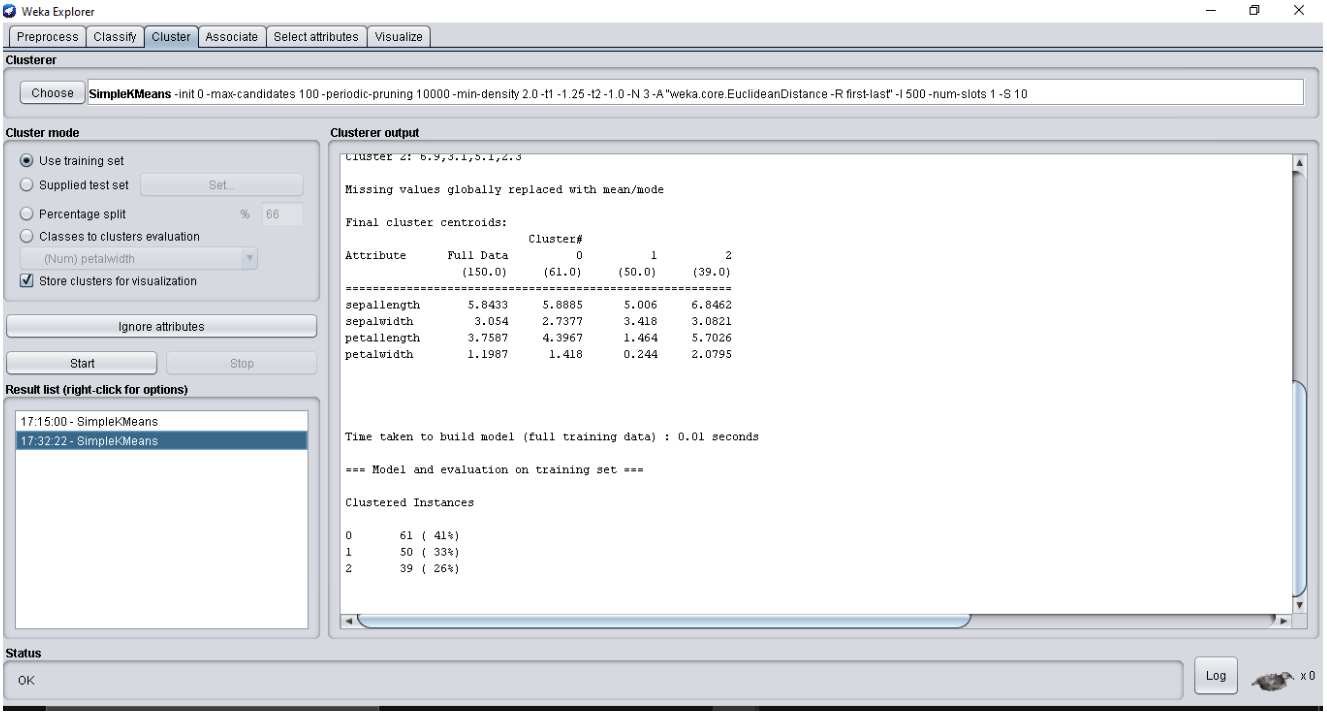

Weka Clusterer Visualize: 17:15:00 - SimpleKMeans (iris-weka.filters.unsupervised.attribute.Remove-R5)
X: petallength (Num)
Y: sepallength (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: iris-weka.filters.unsupervised.attribute.Remove-R5_clustered
7.9
6.1
4.3
1
3.95
6.9
Y
X
Class colour
cluster0
cluster1

In this use patalleangth as X and sapalleangth as Y color as cluster

06.

ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

07



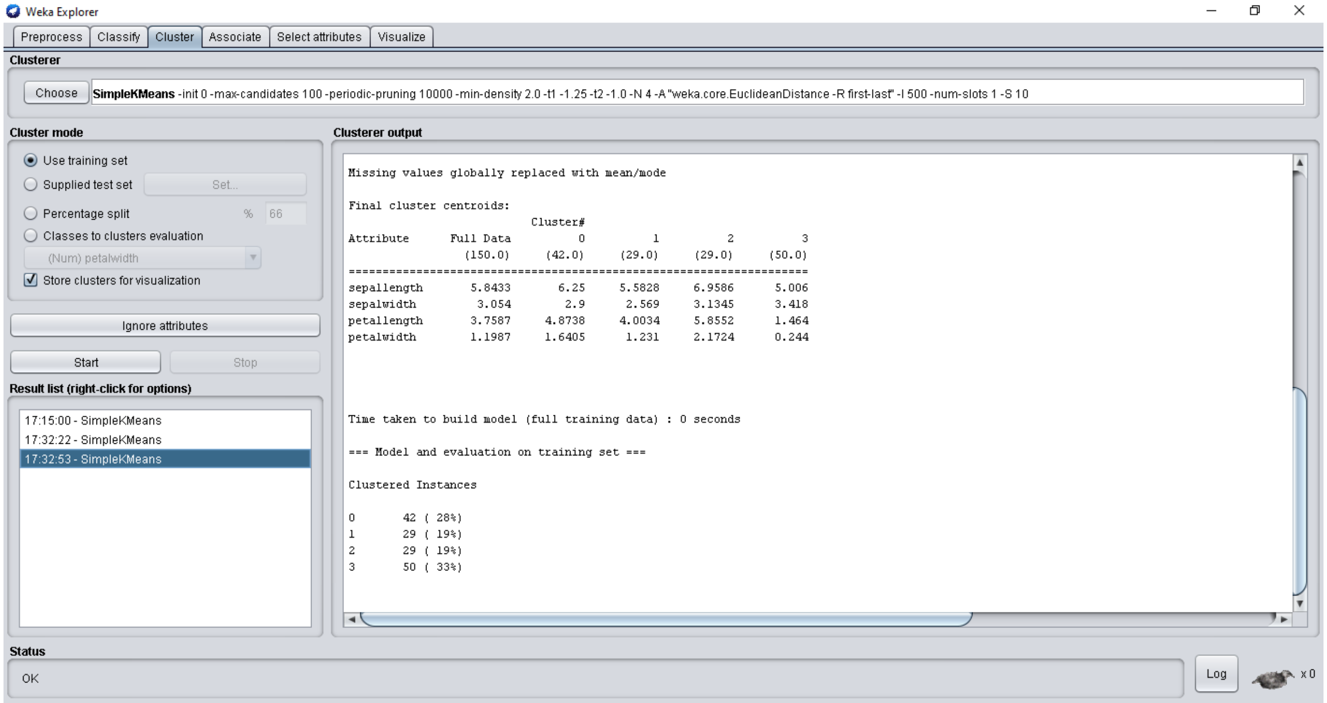
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) petalwidth
☑ Store clusters for visualization

17:15:00 - SimpleKMeans
17:32:22 - SimpleKMeans

```
Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
=========================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795




Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

OK



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) petalwidth
☑ Store clusters for visualization

Result list (right-click for options)

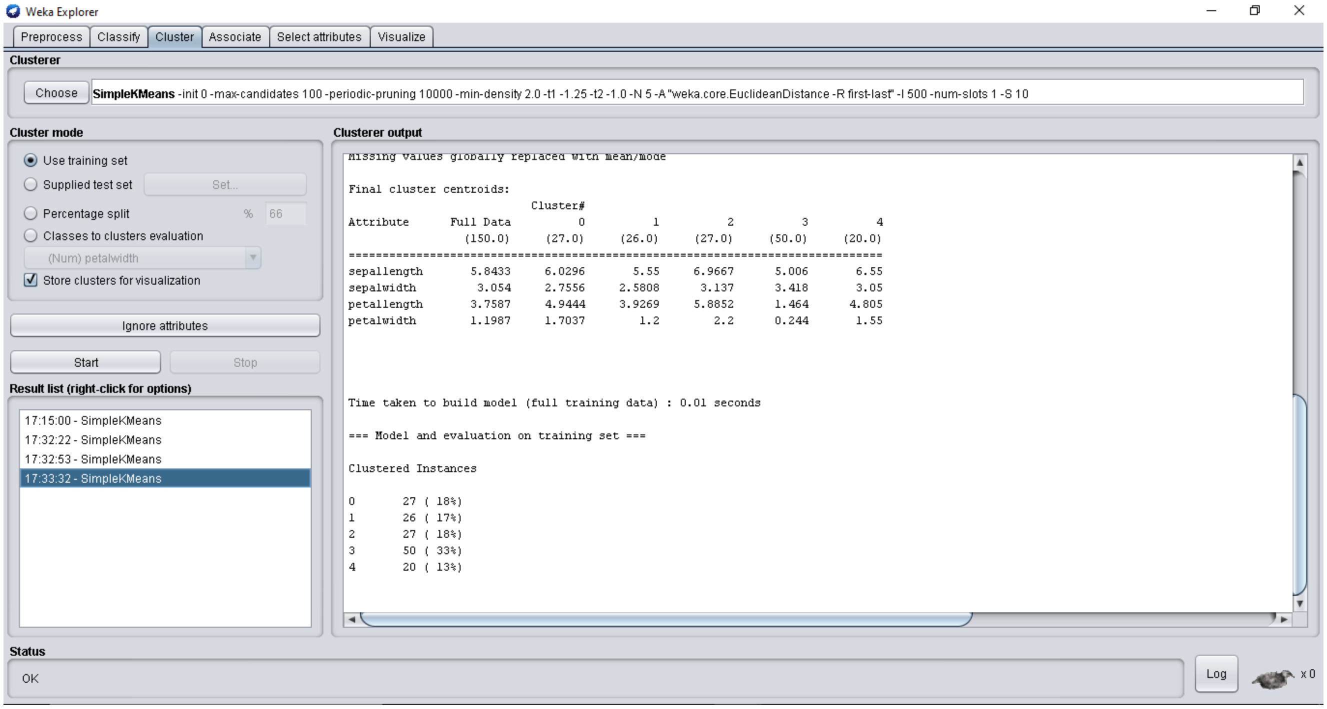
17:15:00 - SimpleKMeans
17:32:22 - SimpleKMeans
17:32:53 - SimpleKMeans

Clusterer output

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2          3
                 (150.0)     (42.0)     (29.0)     (29.0)     (50.0)
====================================================================
sepallength       5.8433       6.25     5.5828     6.9586      5.006
sepalwidth         3.054        2.9      2.569     3.1345      3.418
petallength       3.7587     4.8738     4.0034     5.8552      1.464
petalwidth        1.1987     1.6405      1.231     2.1724      0.244




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       42 ( 28%)
1       29 ( 19%)
2       29 ( 19%)
3       50 ( 33%)
```

Status

OK



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) petalwidth
☑ Store clusters for visualization

Result list (right-click for options)

17:15:00 - SimpleKMeans
17:32:22 - SimpleKMeans
17:32:53 - SimpleKMeans
17:33:32 - SimpleKMeans

Clusterer output

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2          3          4
                 (150.0)     (27.0)     (26.0)     (27.0)     (50.0)     (20.0)
===============================================================================
sepallength       5.8433     6.0296       5.55     6.9667      5.006       6.55
sepalwidth         3.054     2.7556     2.5808      3.137      3.418       3.05
petallength       3.7587     4.9444     3.9269     5.8852      1.464      4.805
petalwidth        1.1987     1.7037        1.2        2.2      0.244       1.55




Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       27 ( 18%)
1       26 ( 17%)
2       27 ( 18%)
3       50 ( 33%)
4       20 ( 13%)
```

Status

OK

When k= 5 lowest mean square error obtained therefore k=5 is most durable