**Prajakta Deokule**

**3330**

**A1**

## ASSIGNMENT 2

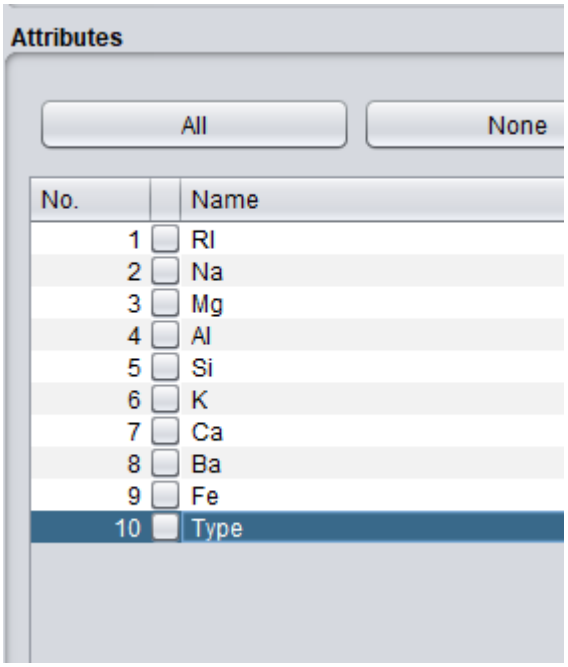
Load any one dataset in Weka and observe the following : List the attribute names and their types, Number of records in each dataset, class attribute (if any), Plot Histogram, Determine the number of records for each class, Visualize the data in various dimensions; Apply various pre-processing tasks; Apply classification OR clustering algorithms on the chosen dataset and observe the results.

### Dataset taken- glass.arff

Instances:214

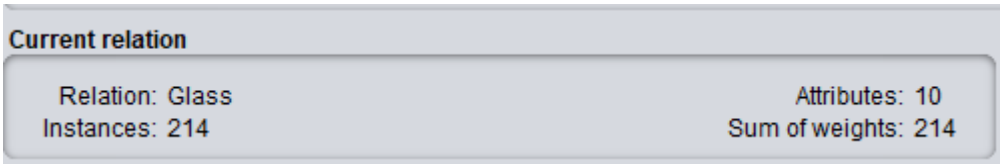
Attributes:10

### List the attribute name and their types




All the attributes are numeric type.

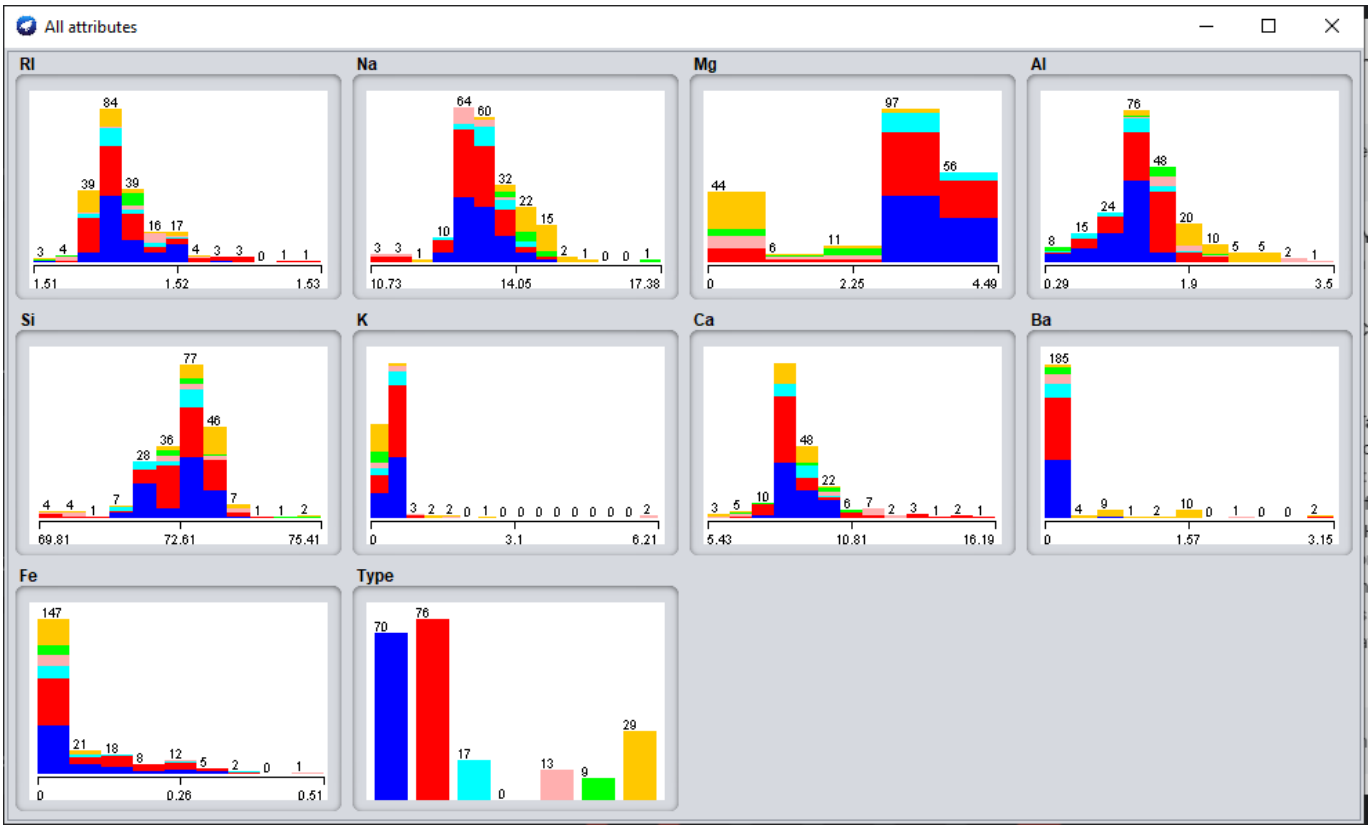
### Number of records in each dataset




Class attribute: **Type attribute**

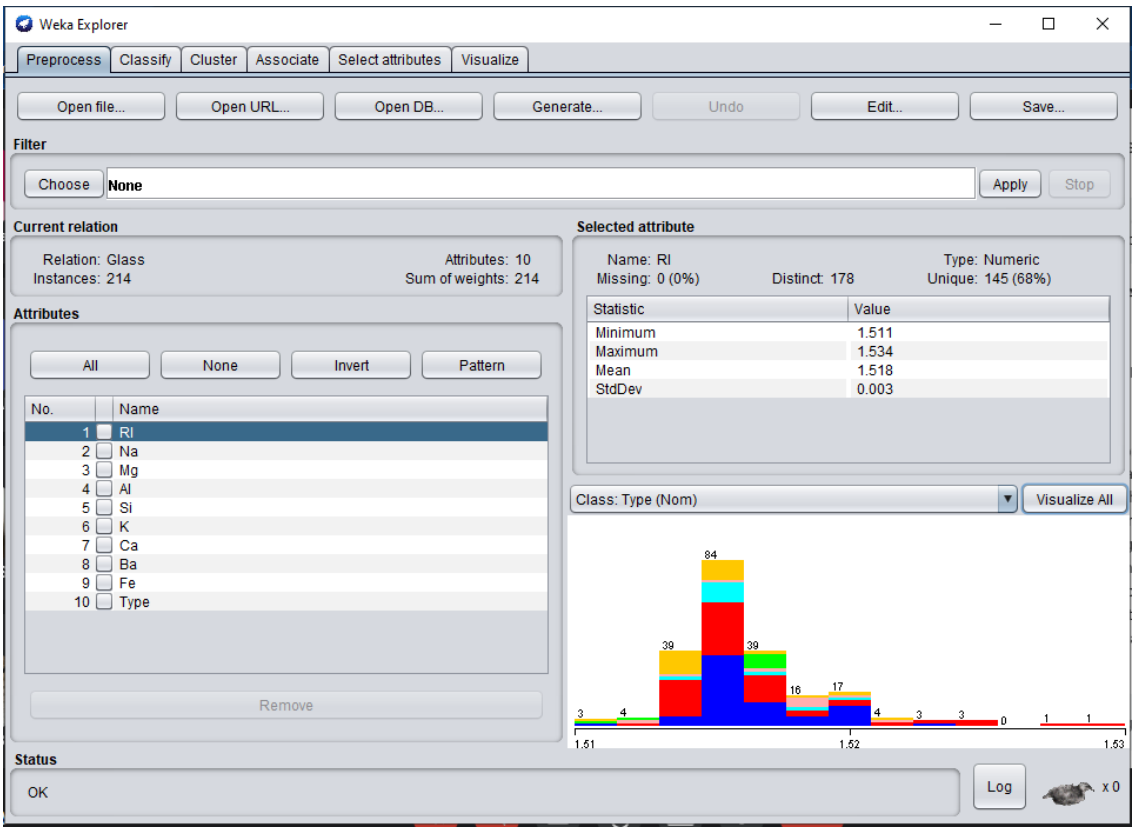
## Plot Histogram

Visualize all attributes:

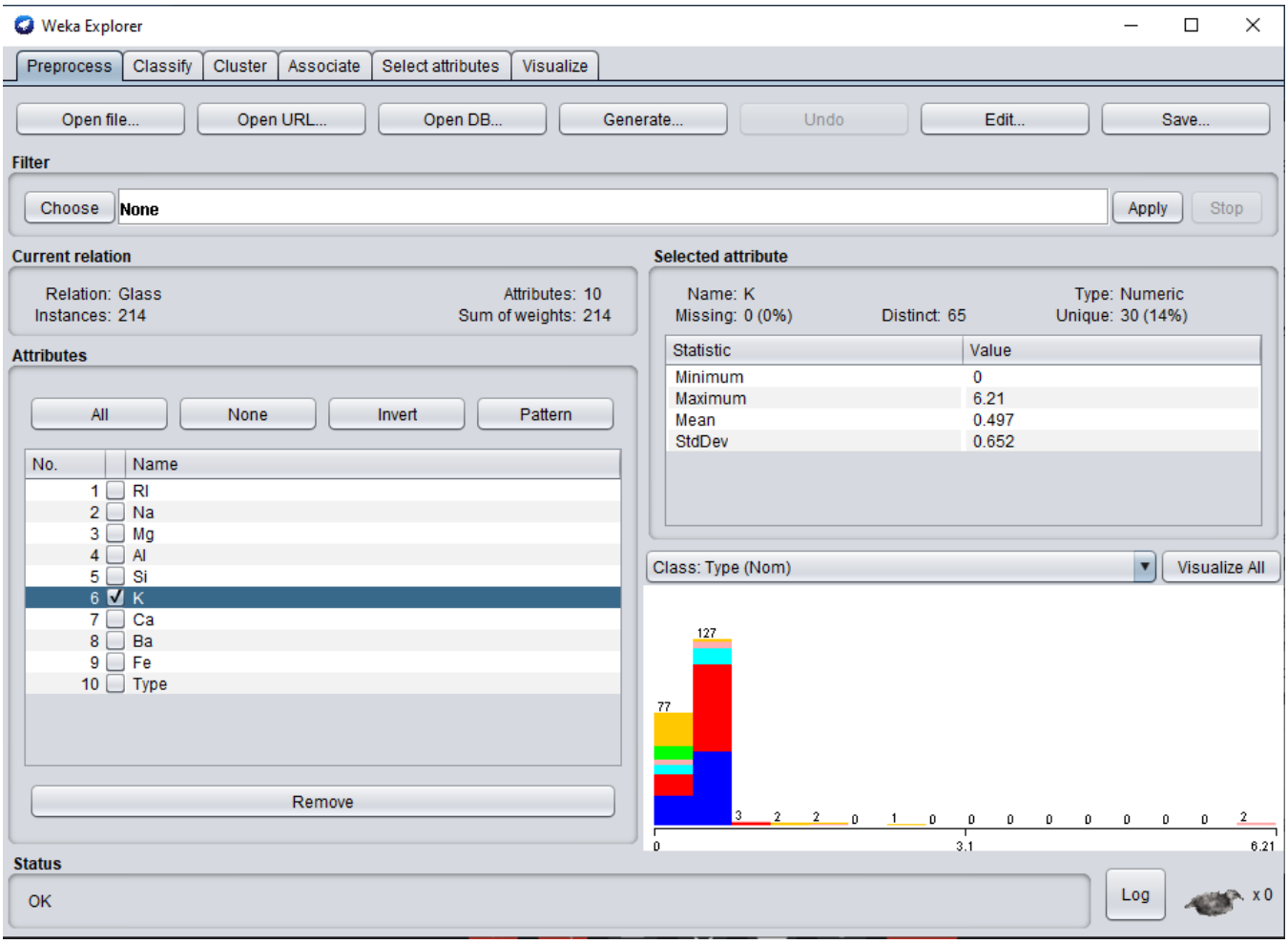



## Plotting the histogram for selected attribute

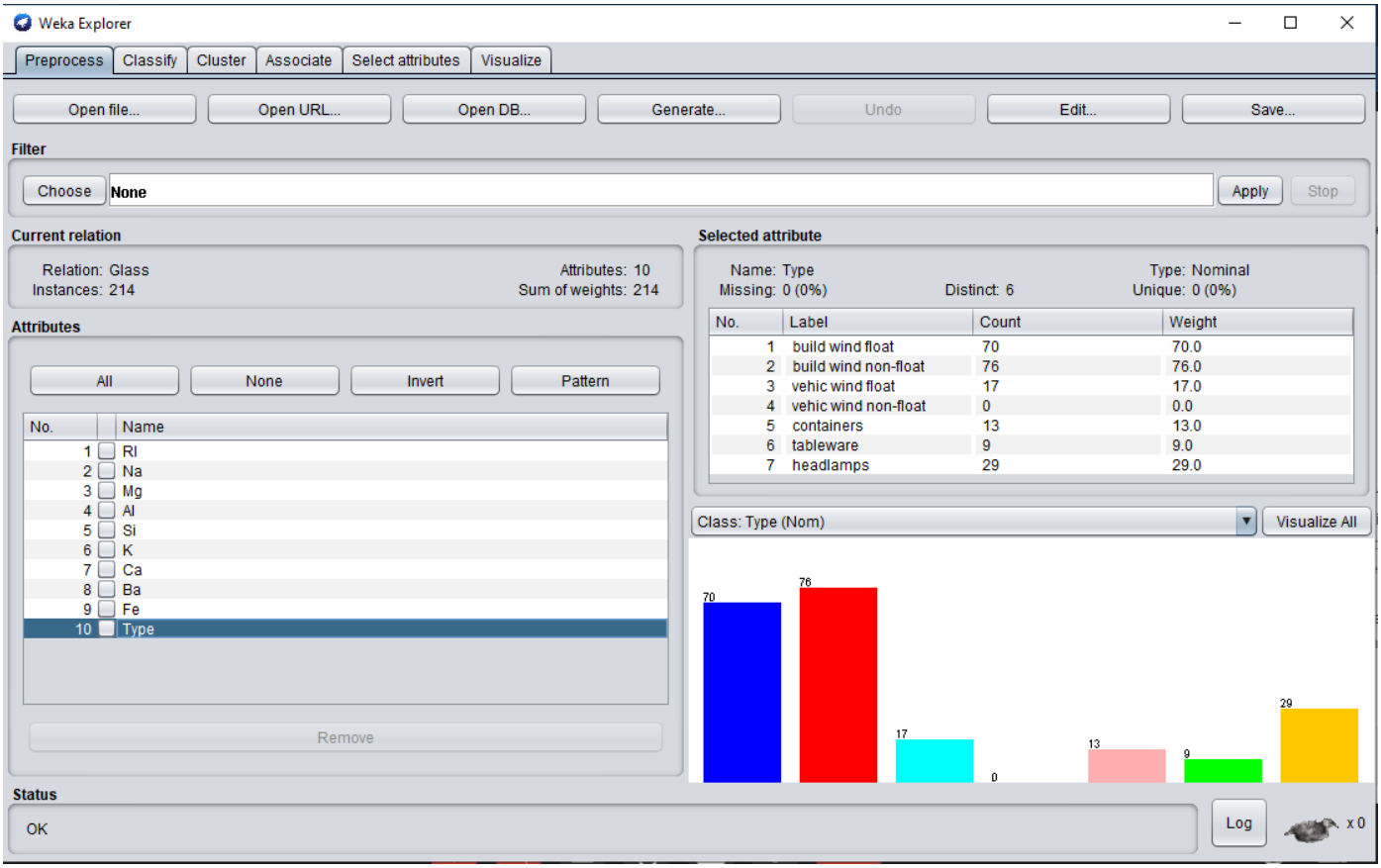
Selected attribute RI



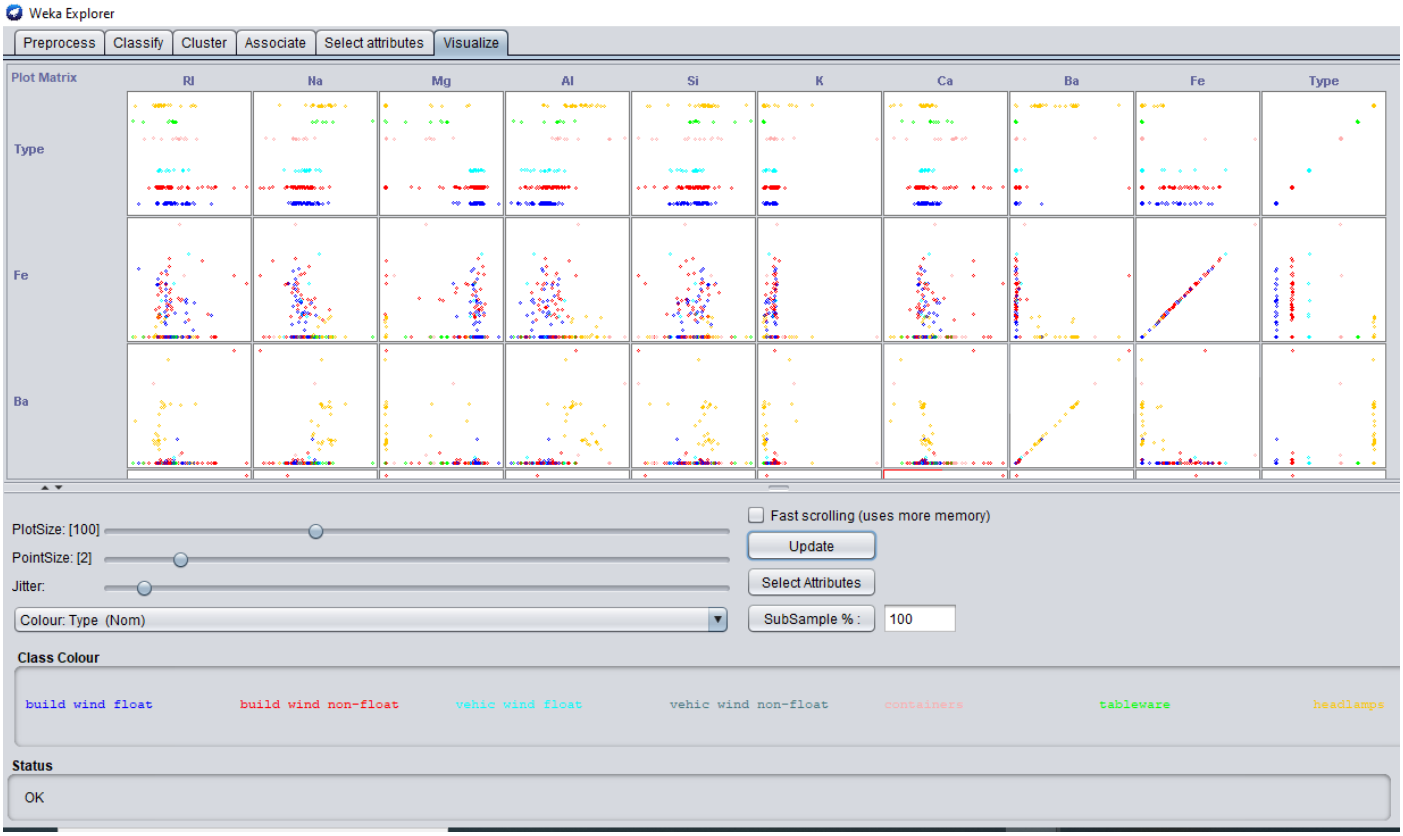
Selected attribute: K



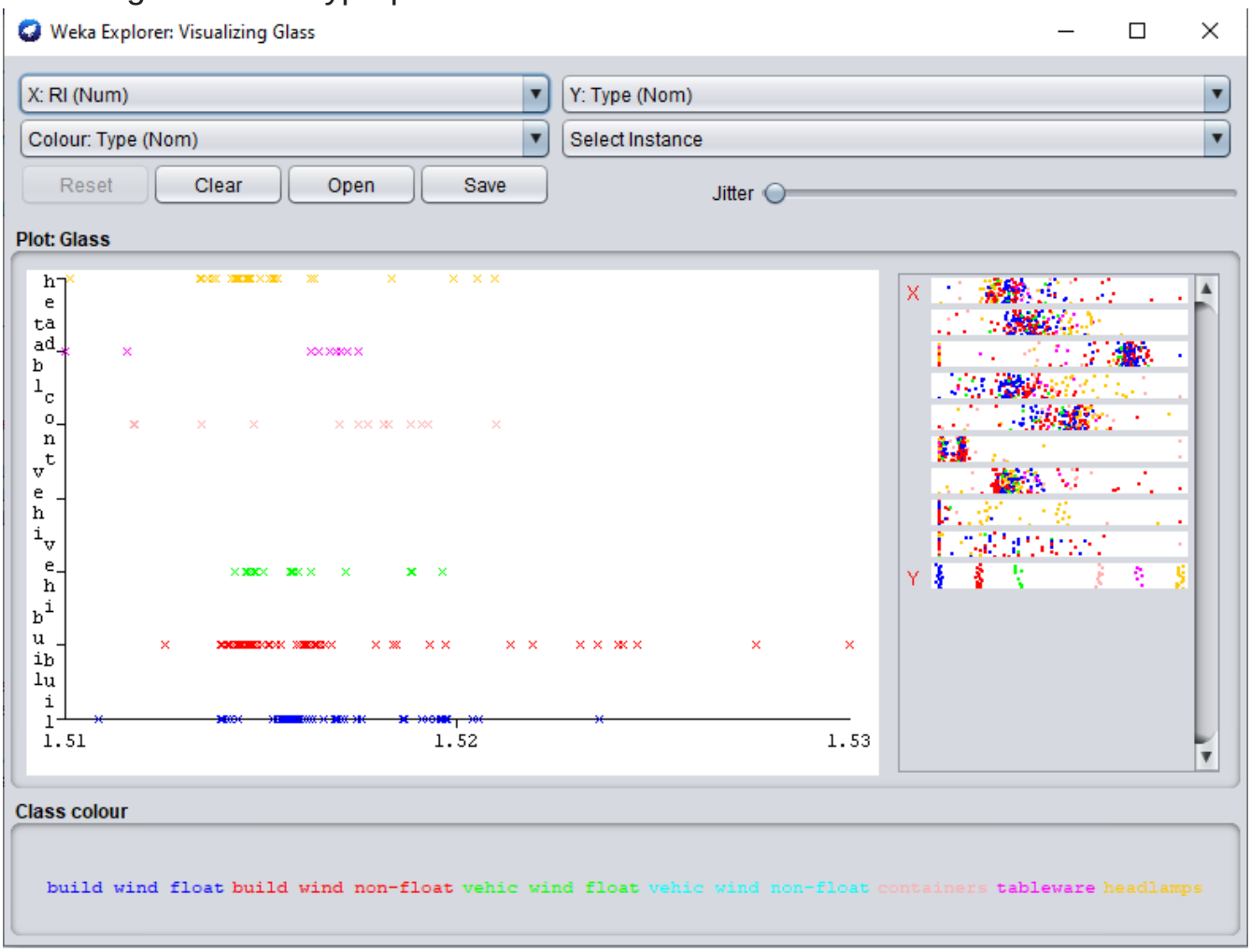

Selected attribute: Type



## Visualize the data in various dimensions



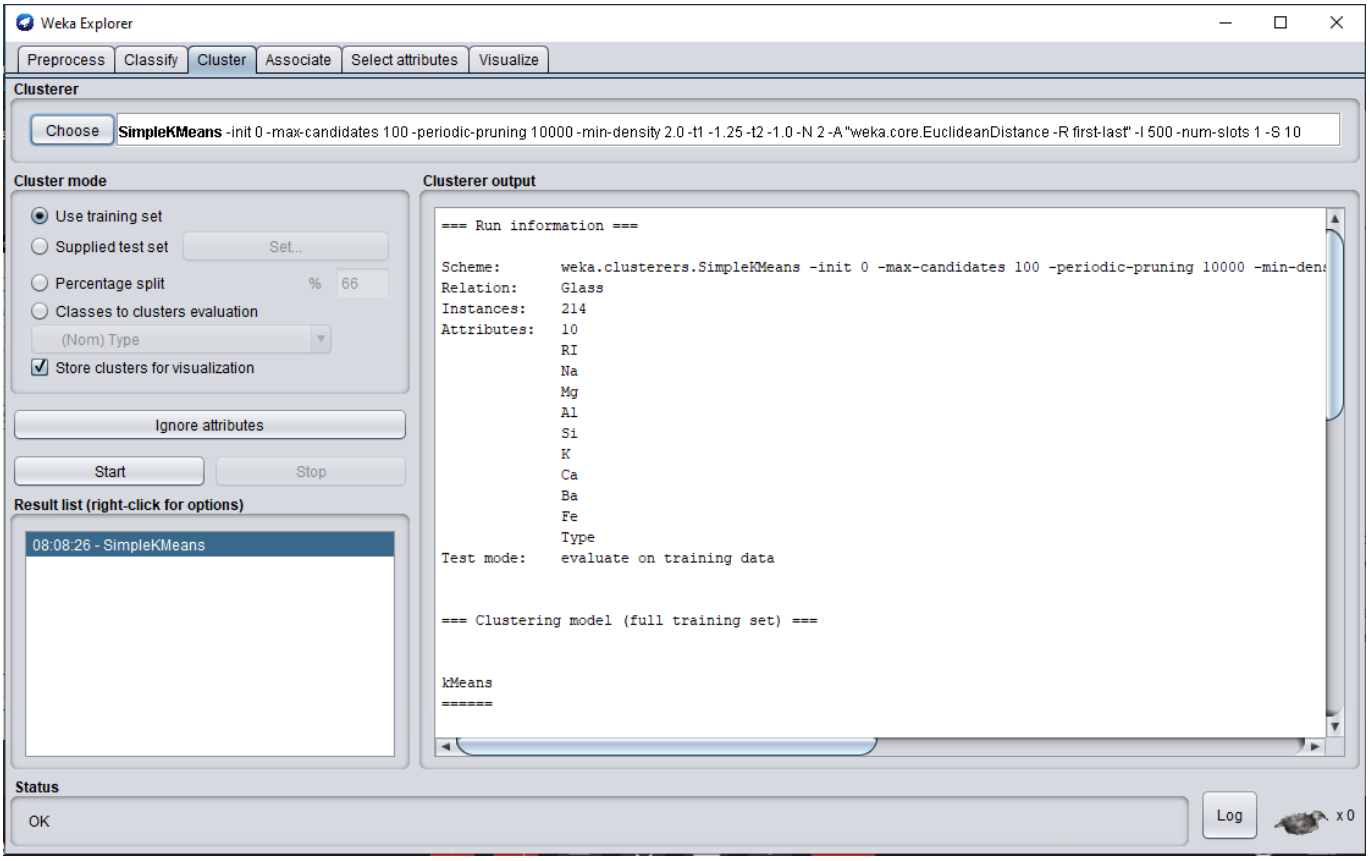

Selecting the RI vs Type plot:

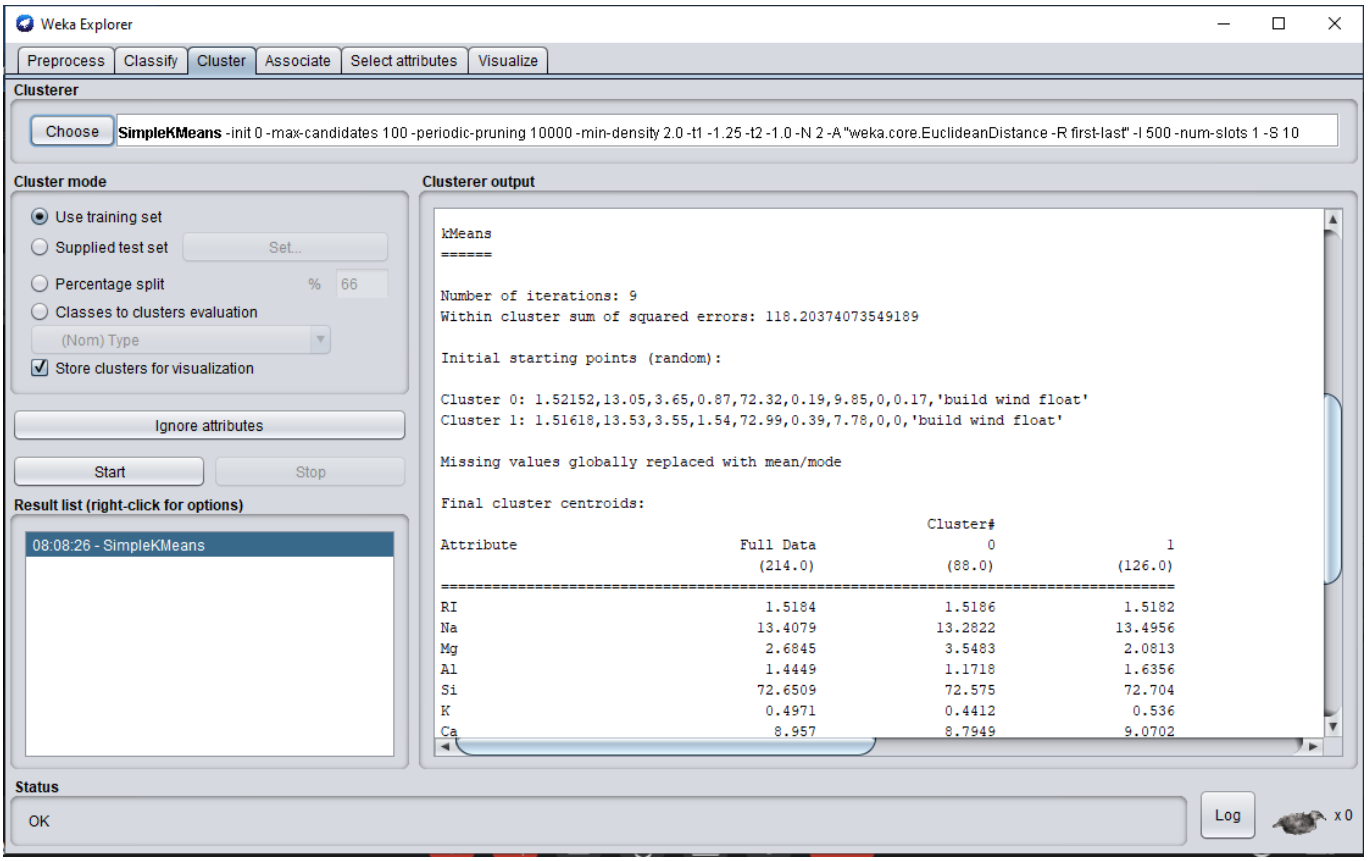


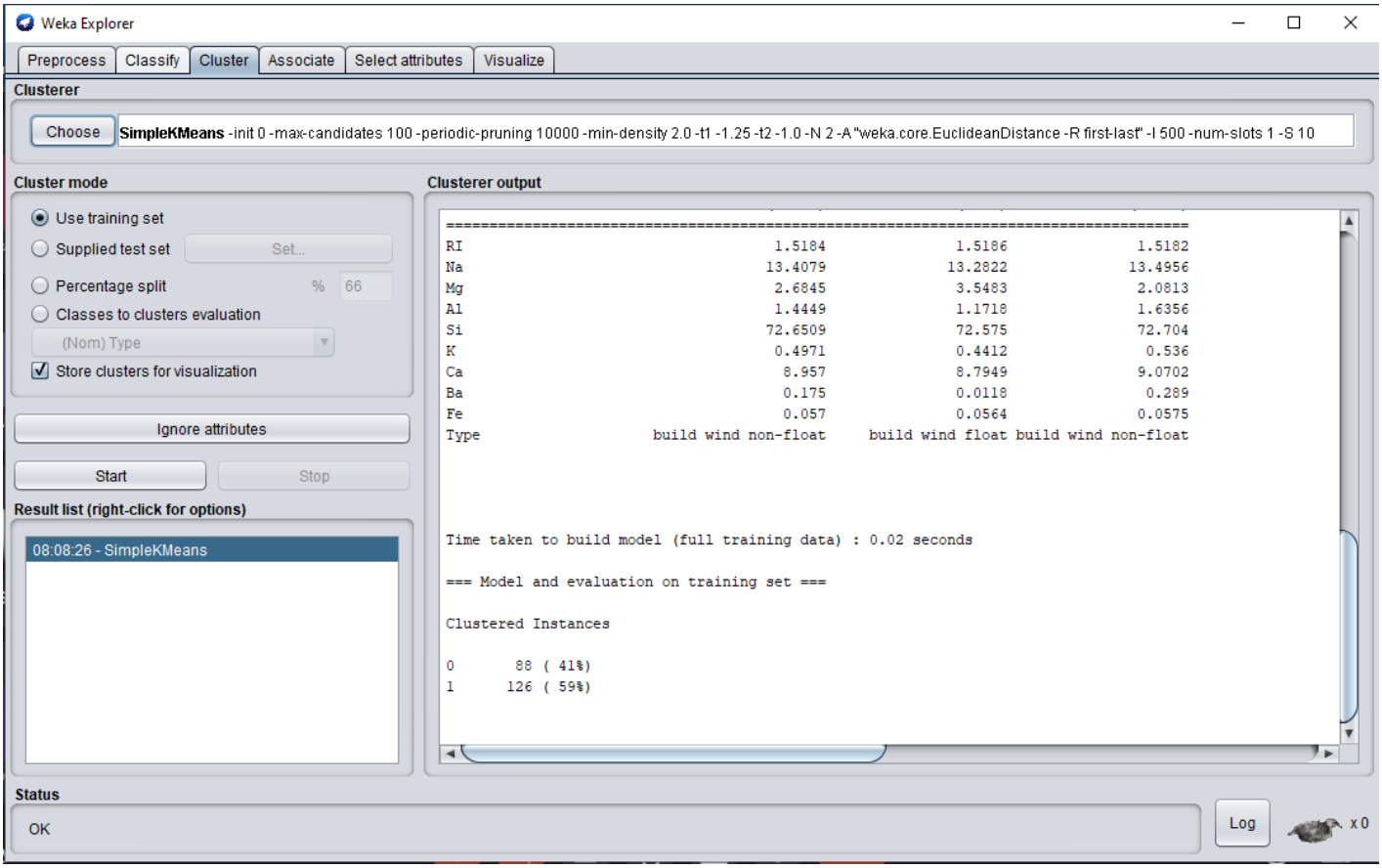
## Clustering-Applying K Means

k-Means Clustering is an unsupervised learning algorithm that is used for clustering .It takes a bunch of unlabelled points and tries to group them into “k” number of clusters. The “k” in k-means denotes the number of clusters you want to have in the end.

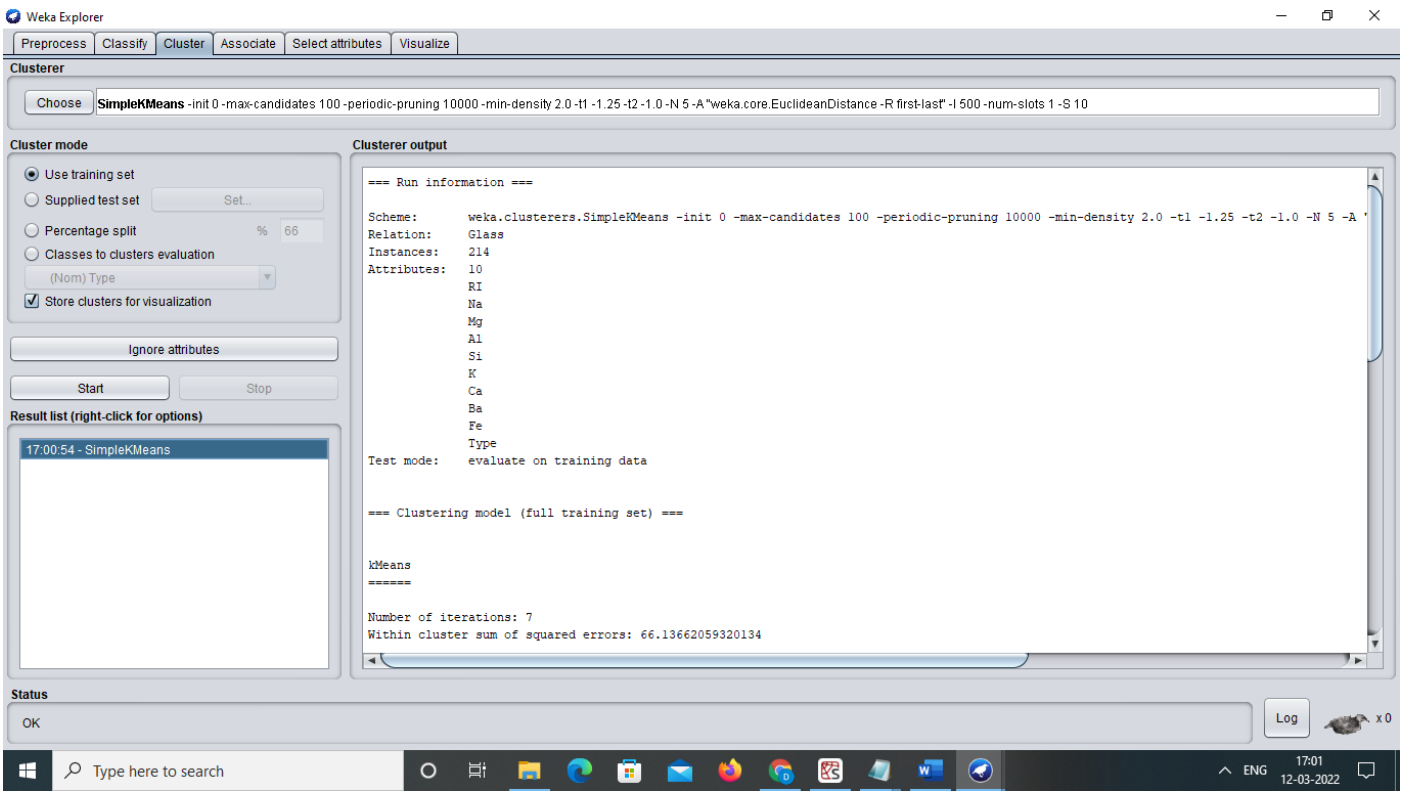
### No of Clusters=2 ( By default)

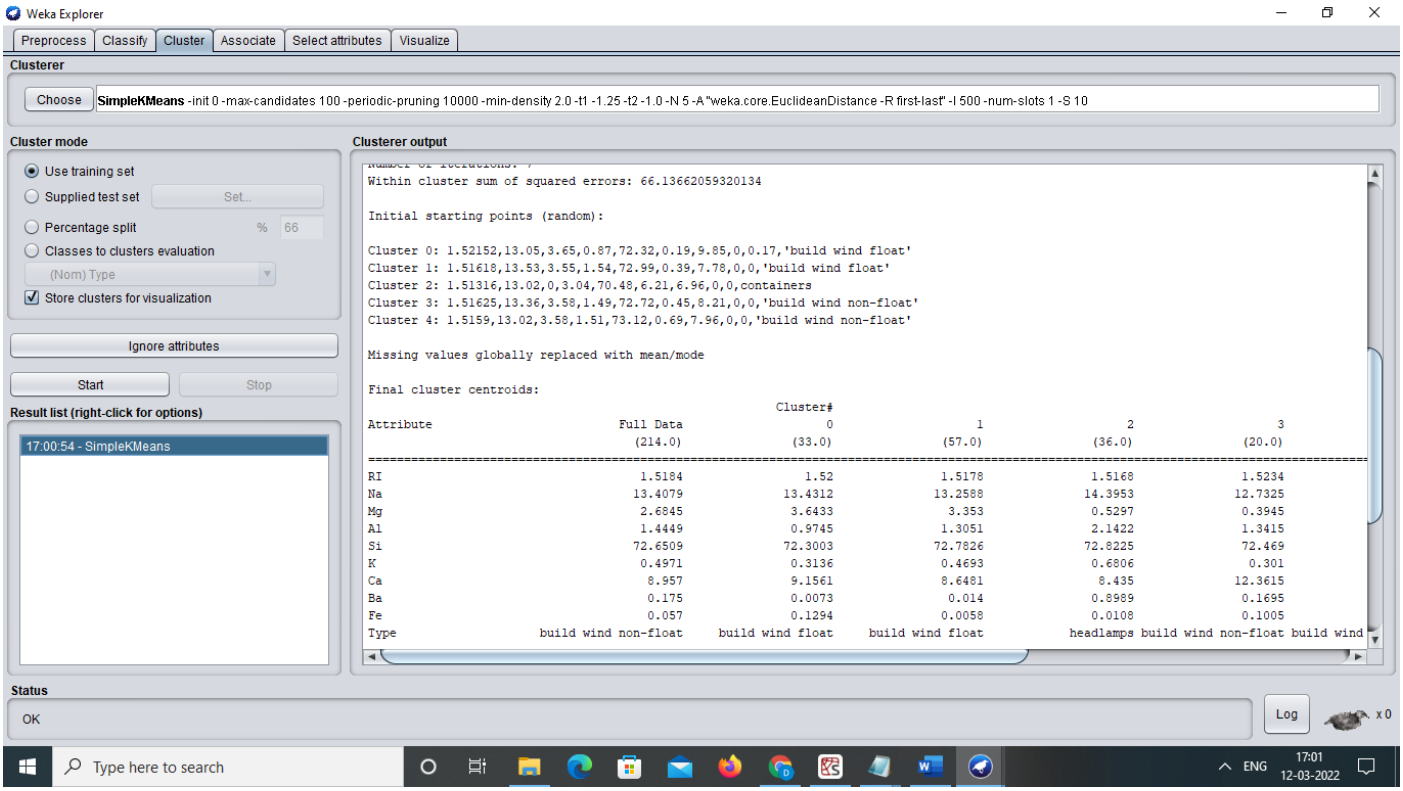


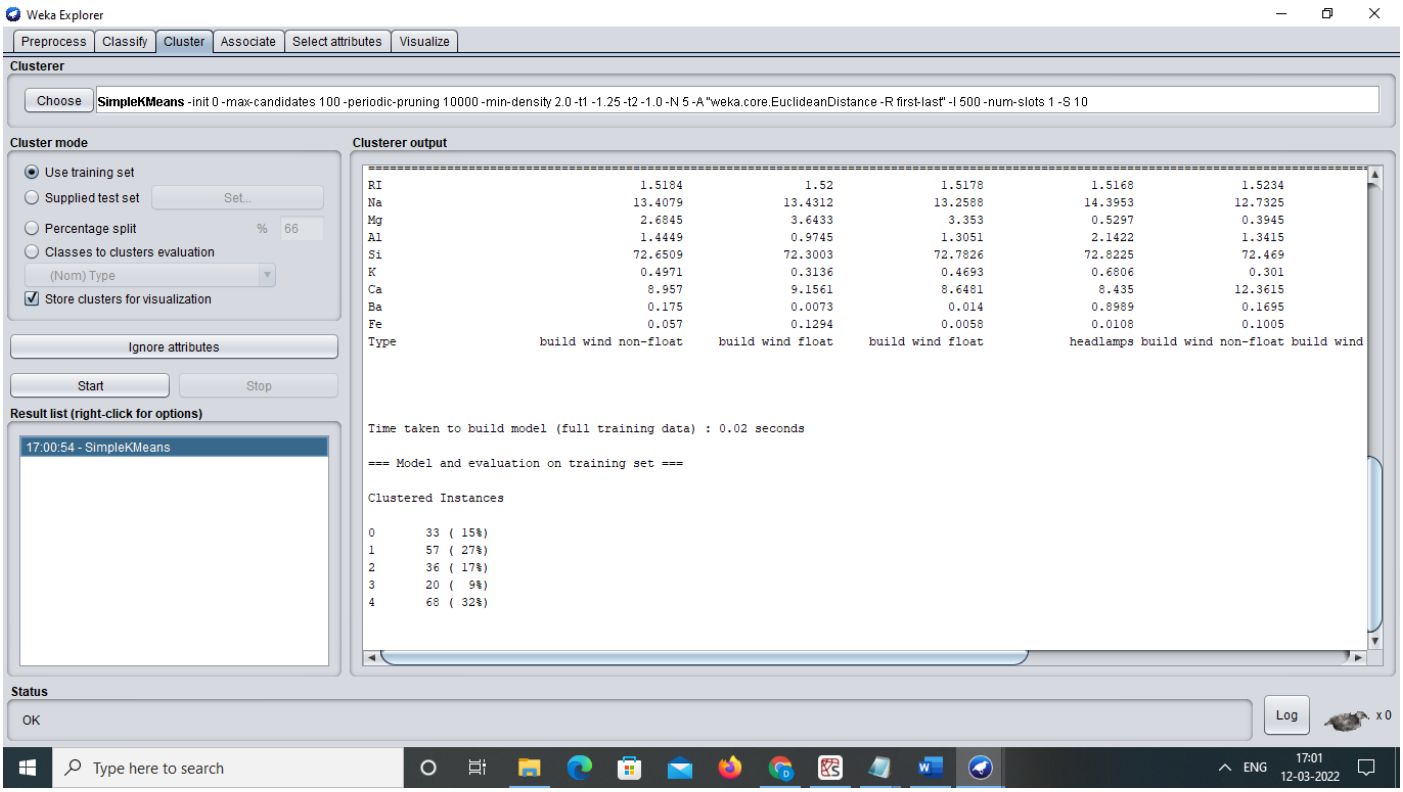






## No of Clusters=5

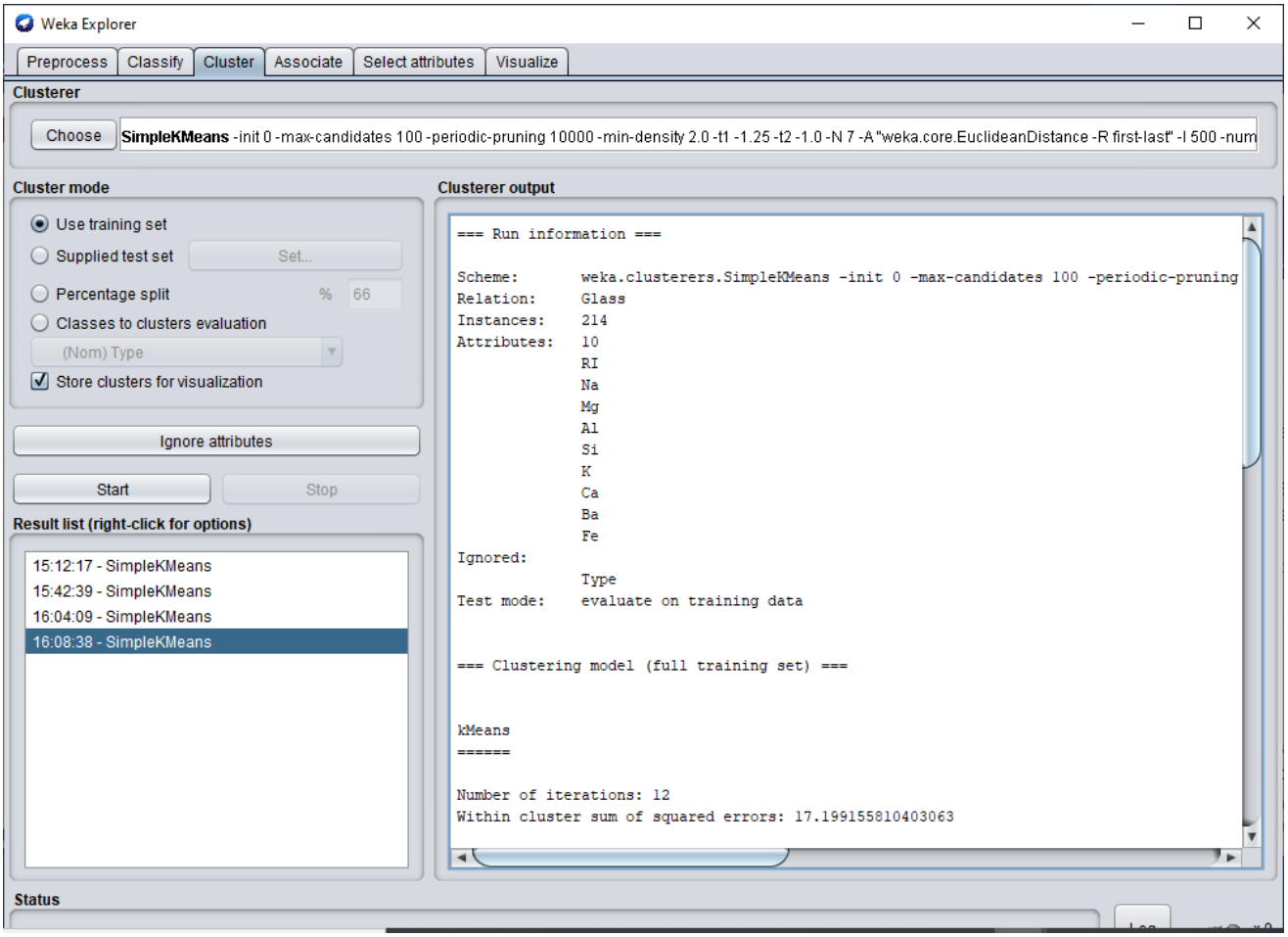


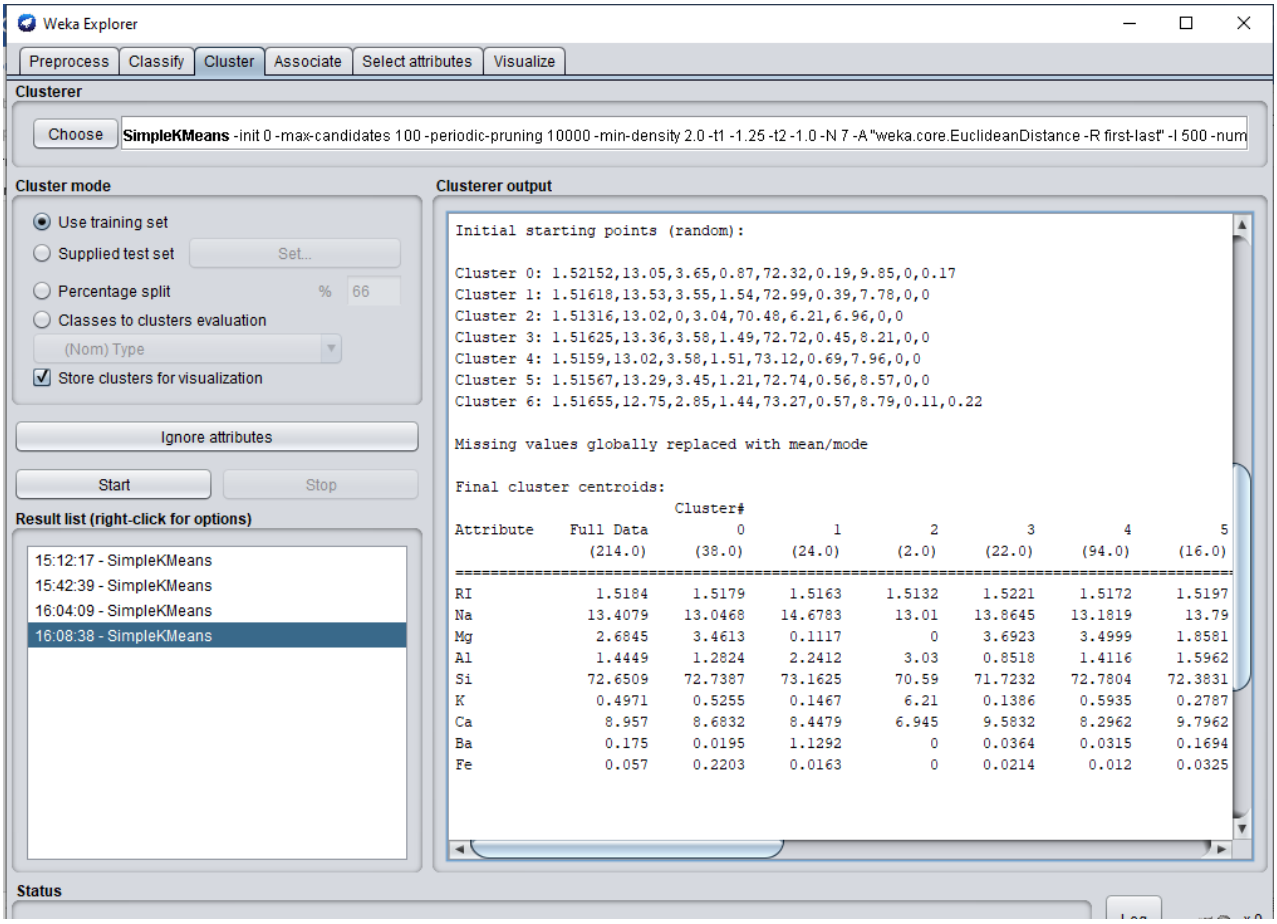


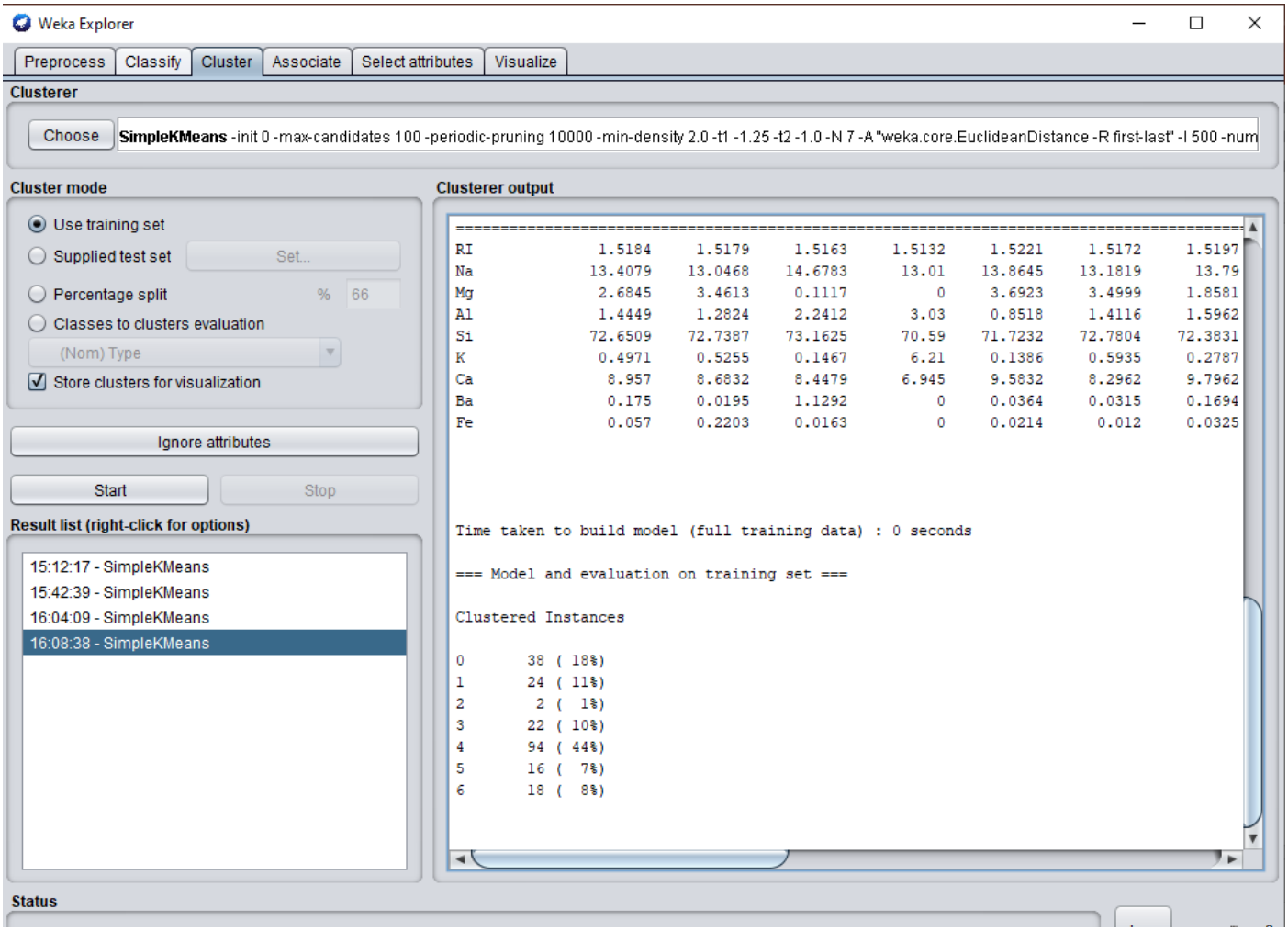




# No of clusters=7

# Here type attribute has been ignored


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 7 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Type
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
15:12:17 - SimpleKMeans
15:42:39 - SimpleKMeans
16:04:09 - SimpleKMeans
16:08:38 - SimpleKMeans
Clusterer output
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning
Relation: Glass
Instances: 214
Attributes: 10
RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Ignored:
Type
Test mode: evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 12
Within cluster sum of squared errors: 17.199155810403063
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 7 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Type
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
15:12:17 - SimpleKMeans
15:42:39 - SimpleKMeans
16:04:09 - SimpleKMeans
16:08:38 - SimpleKMeans
Clusterer output
Initial starting points (random):
Cluster 0: 1.52152,13.05,3.65,0.87,72.32,0.19,9.85,0,0.17
Cluster 1: 1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0,0
Cluster 2: 1.51316,13.02,0,3.04,70.48,6.21,6.96,0,0
Cluster 3: 1.51625,13.36,3.58,1.49,72.72,0.45,8.21,0,0
Cluster 4: 1.5159,13.02,3.58,1.51,73.12,0.69,7.96,0,0
Cluster 5: 1.51567,13.29,3.45,1.21,72.74,0.56,8.57,0,0
Cluster 6: 1.51655,12.75,2.85,1.44,73.27,0.57,8.79,0.11,0.22
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1 2 3 4 5
(214.0) (38.0) (24.0) (2.0) (22.0) (94.0) (16.0)
RI 1.5184 1.5179 1.5163 1.5132 1.5221 1.5172 1.5197
Na 13.4079 13.0468 14.6783 13.01 13.8645 13.1819 13.79
Mg 2.6845 3.4613 0.1117 0 3.6923 3.4999 1.8581
Al 1.4449 1.2824 2.2412 3.03 0.8518 1.4116 1.5962
Si 72.6509 72.7387 73.1625 70.59 71.7232 72.7804 72.3831
K 0.4971 0.5255 0.1467 6.21 0.1386 0.5935 0.2787
Ca 8.957 8.6832 8.4479 6.945 9.5832 8.2962 9.7962
Ba 0.175 0.0195 1.1292 0 0.0364 0.0315 0.1694
Fe 0.057 0.2203 0.0163 0 0.0214 0.012 0.0325
Status



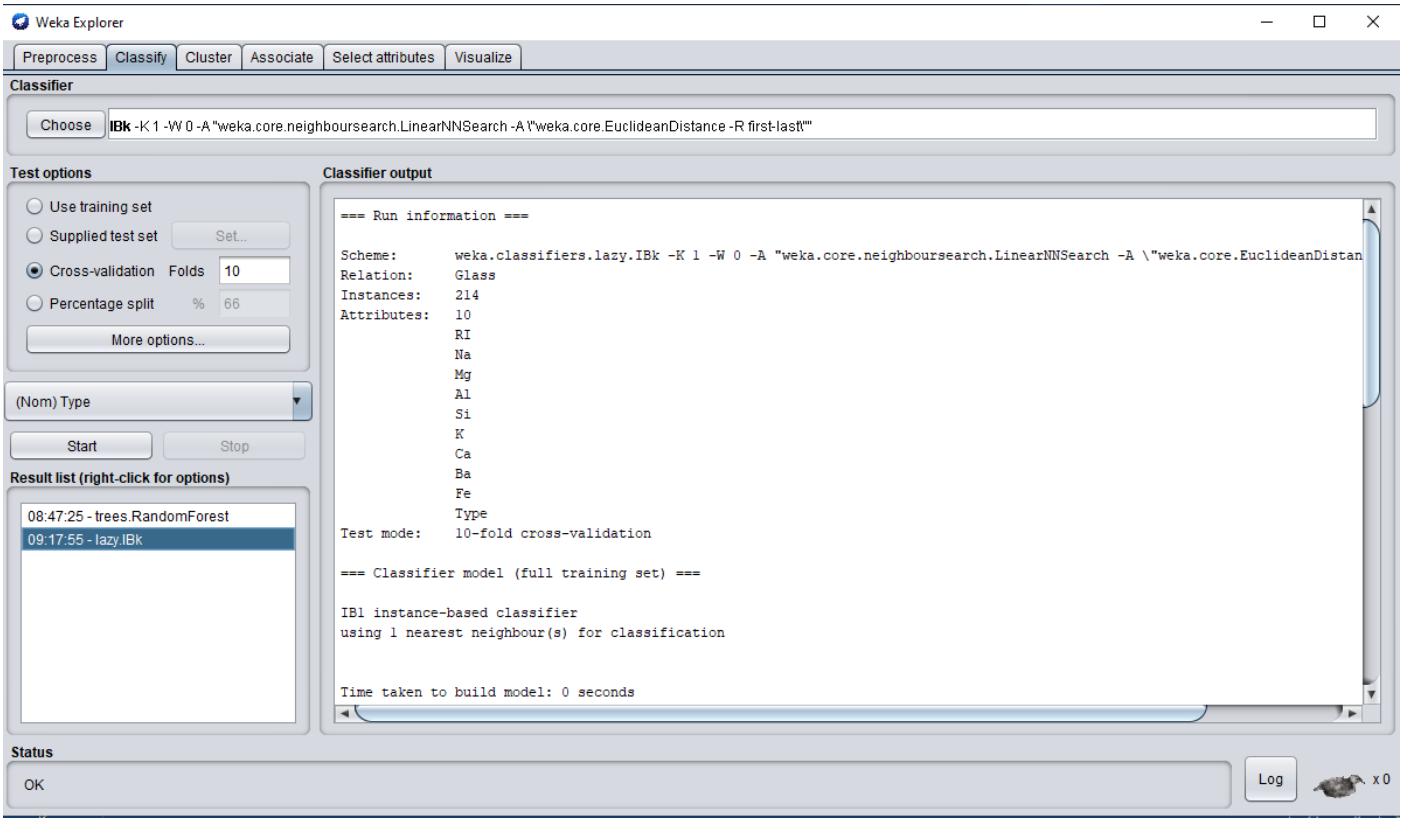

## Classification

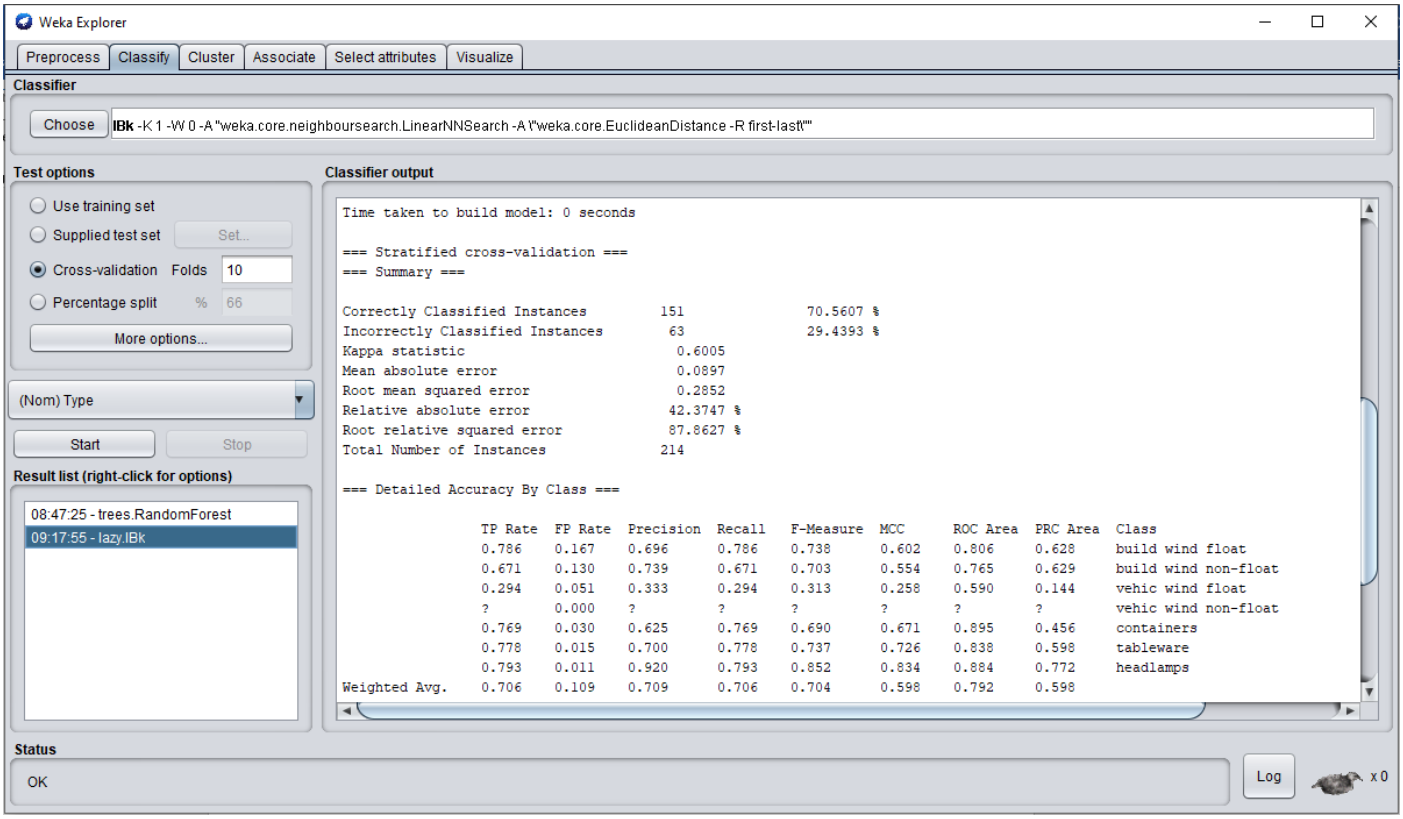
### Classify Using Nearest Neighbour Algorithm for Classifiation

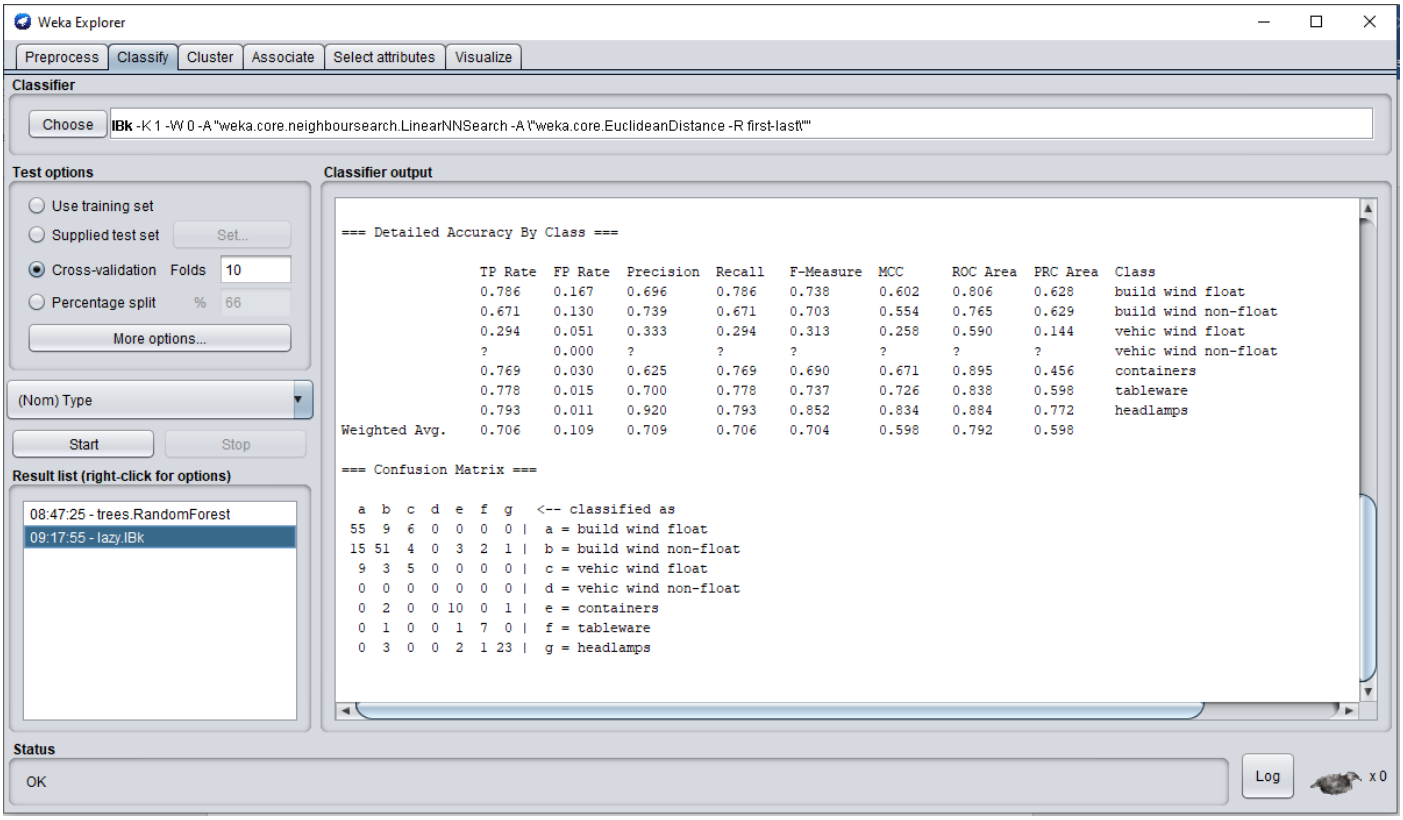
KNN Algorithm is a supervised classification algorithm

For KNN use lazy IBk under classify

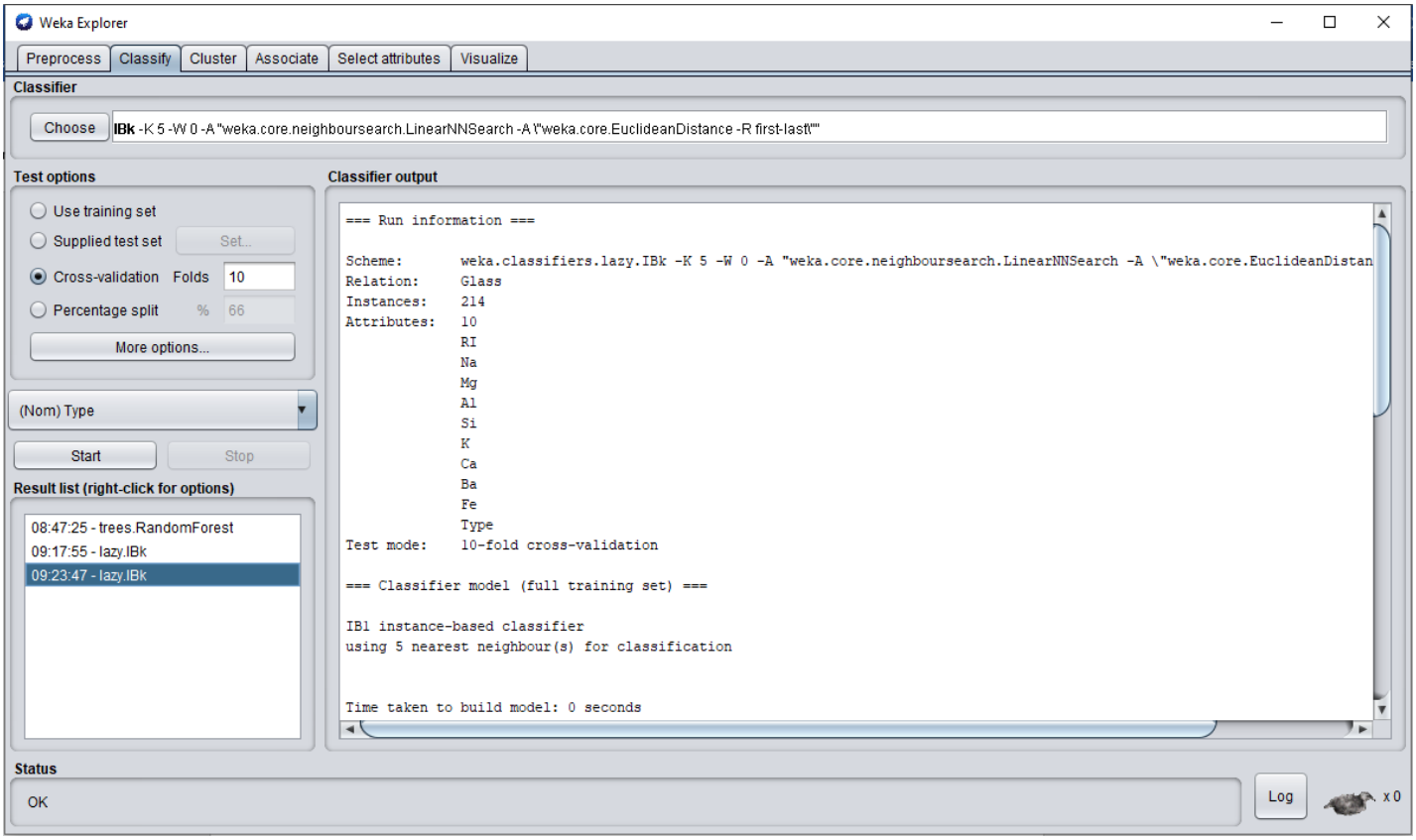
**For KNN=1(by default)**

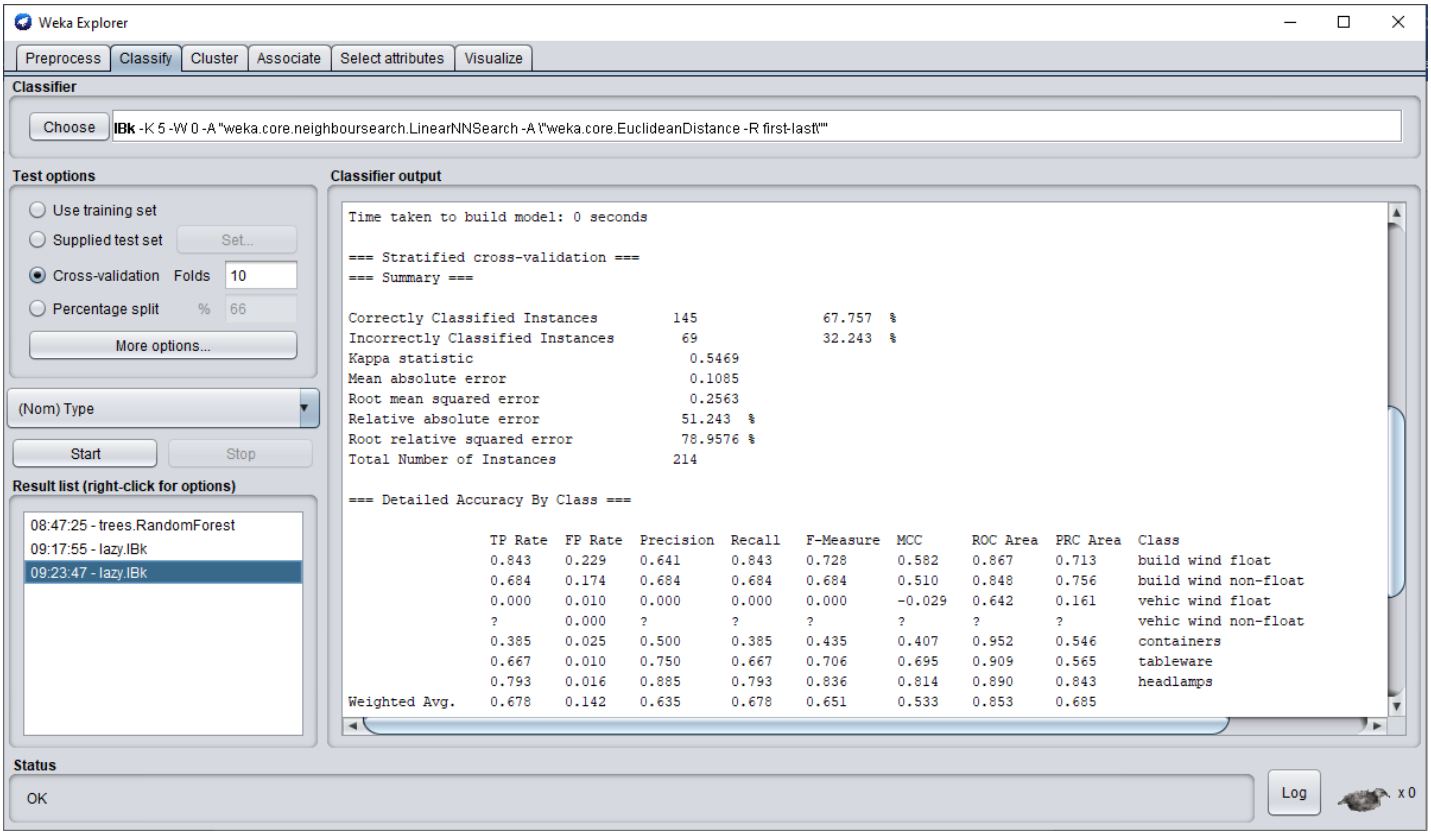

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
08:47:25 - trees.RandomForest
09:17:55 - lazy.IBk
Classifier output
=== Run information ===
Scheme: weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistan
Relation: Glass
Instances: 214
Attributes: 10
RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
Status
OK
Log
x 0

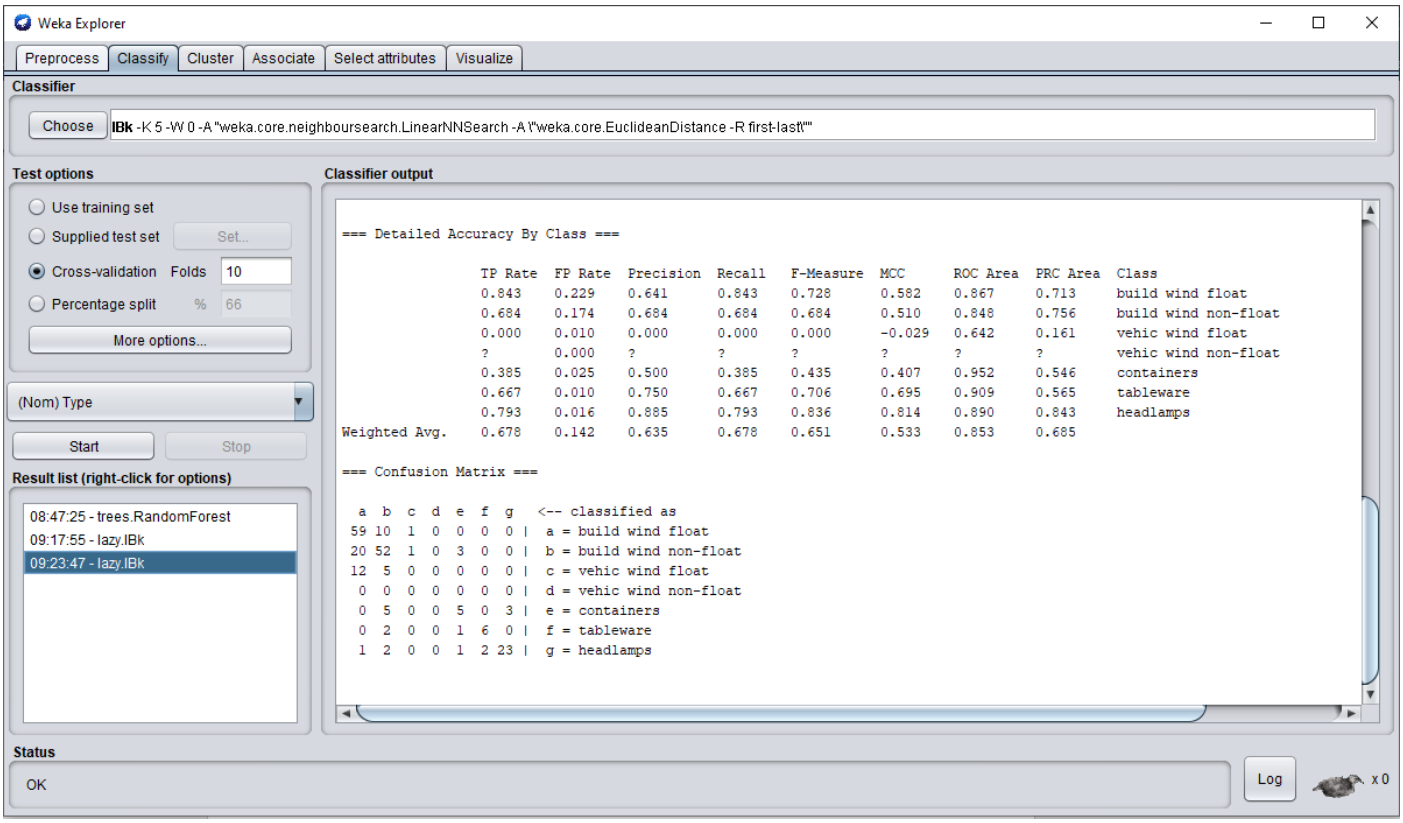

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
08:47:25 - trees.RandomForest
09:17:55 - lazy.IBk
Classifier output
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 151 70.5607 %
Incorrectly Classified Instances 63 29.4393 %
Kappa statistic 0.6005
Mean absolute error 0.0897
Root mean squared error 0.2852
Relative absolute error 42.3747 %
Root relative squared error 87.8627 %
Total Number of Instances 214
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.786 0.167 0.696 0.786 0.738 0.602 0.806 0.628 build wind float
0.671 0.130 0.739 0.671 0.703 0.554 0.765 0.629 build wind non-float
0.294 0.051 0.333 0.294 0.313 0.258 0.590 0.144 vehic wind float
? 0.000 ? ? ? ? ? ? vehic wind non-float
0.769 0.030 0.625 0.769 0.690 0.671 0.895 0.456 containers
0.778 0.015 0.700 0.778 0.737 0.726 0.838 0.598 tableware
0.793 0.011 0.920 0.793 0.852 0.834 0.884 0.772 headlamps
Weighted Avg. 0.706 0.109 0.709 0.706 0.704 0.598 0.792 0.598
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
08:47:25 - trees.RandomForest
09:17:55 - lazy.IBk
Classifier output
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.786 0.167 0.696 0.786 0.738 0.602 0.806 0.628 build wind float
0.671 0.130 0.739 0.671 0.703 0.554 0.765 0.629 build wind non-float
0.294 0.051 0.333 0.294 0.313 0.258 0.590 0.144 vehic wind float
? 0.000 ? ? ? ? ? ? vehic wind non-float
0.769 0.030 0.625 0.769 0.690 0.671 0.895 0.456 containers
0.778 0.015 0.700 0.778 0.737 0.726 0.838 0.598 tableware
0.793 0.011 0.920 0.793 0.852 0.834 0.884 0.772 headlamps
Weighted Avg. 0.706 0.109 0.709 0.706 0.704 0.598 0.792 0.598
=== Confusion Matrix ===
a b c d e f g <-- classified as
55 9 6 0 0 0 0 | a = build wind float
15 51 4 0 3 2 1 | b = build wind non-float
9 3 5 0 0 0 0 | c = vehic wind float
0 0 0 0 0 0 0 | d = vehic wind non-float
0 2 0 0 10 0 1 | e = containers
0 1 0 0 1 7 0 | f = tableware
0 3 0 0 2 1 23 | g = headlamps
Status
OK
Log
x 0

For KNN=5


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
08:47:25 - trees.RandomForest
09:17:55 - lazy.IBk
09:23:47 - lazy.IBk
Classifier output
=== Run information ===
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistan
Relation: Glass
Instances: 214
Attributes: 10
RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 5 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
08:47:25 - trees.RandomForest
09:17:55 - lazy.IBk
09:23:47 - lazy.IBk
Classifier output
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 145 67.757 %
Incorrectly Classified Instances 69 32.243 %
Kappa statistic 0.5469
Mean absolute error 0.1085
Root mean squared error 0.2563
Relative absolute error 51.243 %
Root relative squared error 78.9576 %
Total Number of Instances 214
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.843 0.229 0.641 0.843 0.728 0.582 0.867 0.713 build wind float
0.684 0.174 0.684 0.684 0.684 0.510 0.848 0.756 build wind non-float
0.000 0.010 0.000 0.000 0.000 -0.029 0.642 0.161 vehic wind float
? 0.000 ? ? ? ? ? ? vehic wind non-float
0.385 0.025 0.500 0.385 0.435 0.407 0.952 0.546 containers
0.667 0.010 0.750 0.667 0.706 0.695 0.909 0.565 tableware
0.793 0.016 0.885 0.793 0.836 0.814 0.890 0.843 headlamps
Weighted Avg. 0.678 0.142 0.635 0.678 0.651 0.533 0.853 0.685
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
08:47:25 - trees.RandomForest
09:17:55 - lazy.IBk
09:23:47 - lazy.IBk
Classifier output
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.843 0.229 0.641 0.843 0.728 0.582 0.867 0.713 build wind float
0.684 0.174 0.684 0.684 0.684 0.510 0.848 0.756 build wind non-float
0.000 0.010 0.000 0.000 0.000 -0.029 0.642 0.161 vehic wind float
? 0.000 ? ? ? ? ? ? vehic wind non-float
0.385 0.025 0.500 0.385 0.435 0.407 0.952 0.546 containers
0.667 0.010 0.750 0.667 0.706 0.695 0.909 0.565 tableware
0.793 0.016 0.885 0.793 0.836 0.814 0.890 0.843 headlamps
Weighted Avg. 0.678 0.142 0.635 0.678 0.651 0.533 0.853 0.685
=== Confusion Matrix ===
a b c d e f g <-- classified as
59 10 1 0 0 0 0 | a = build wind float
20 52 1 0 3 0 0 | b = build wind non-float
12 5 0 0 0 0 0 | c = vehic wind float
0 0 0 0 0 0 0 | d = vehic wind non-float
0 5 0 0 5 0 3 | e = containers
0 2 0 0 1 6 0 | f = tableware
1 2 0 0 1 2 23 | g = headlamps
Status
OK
Log
x 0