Assignment 1

Classification and Clustering using Weka Explorer

Hafara Firdausi 05111950010040

Description

- Find dataset in UCI machine learning, Kaggle, etc, to be processed.
- Do **classification** using the following methods (supervised learning):
  - Naive Bayes
  - K-Nearest Neighbor (KNN)
  - Support Vector Machine (SVM)
- Do **clustering** using the following methods (unsupervised learning):
  - K-Means

Tools

- Weka Explorer 3.8.2

Dataset Description

- Title :

  Forest Covertype data

- Description :

  Forest Cover Type (FC) data contains **tree observations** from four wilderness areas located in the **Roosevelt National Forest of northern Colorado**. All observations are cartographic variables (no remote sensing) from 30 meter x 30 meter sections of forest. This dataset includes information on tree type, shadow coverage, distance to nearby landmarks (roads etcetera), soil type, and local topography.

  This dataset is part of the **UCI Machine Learning Repository** (here), but covertype dataset that I use comes from **Kaggle** that can be found here. The original database owners are **Jock A. Blackard, Dr. Denis J. Dean, and Dr. Charles W. Anderson** of the Remote Sensing and GIS Program at Colorado State University.

- Details :

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer
Associated Tasks	Classification
Number of Instances	581012
Number of Attributes	54
Missing Values?	No

Date Donated	1998-08-01

- Attributes:

Column Name	Number of Columns
Elevation	1
Aspect	1
Slope	1
Horizontal_Distance_To_Hydrology	1
Vertical_Distance_To_Hydrology	1
Hillshade_9am	1
Hillshade_Noon	1
Hillshade_3pm	1
Horizontal_Distance_To_Fire_Points	1
Wilderness_Area	4
Soil_Type	40
Cover_Type	1

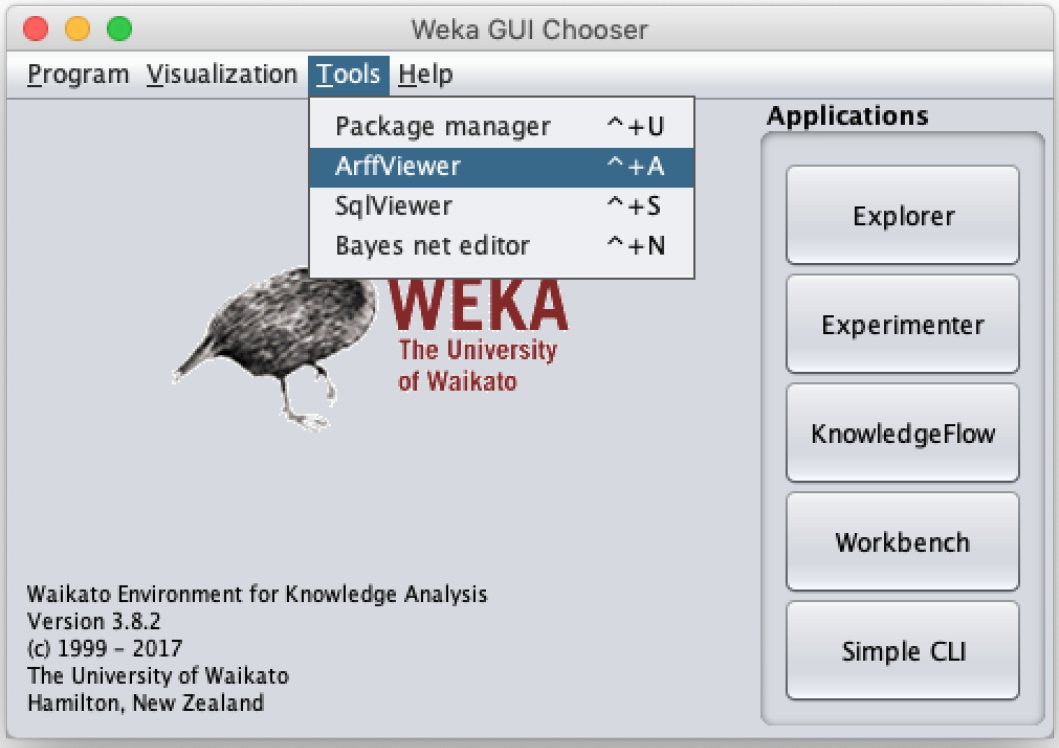
- There are 7 classes (covertype) here, that is:

    1. Spruce/Fir
    2. Lodgepole Pine
    3. Ponderosa Pine
    4. Cottonwood/Willow
    5. Aspen
    6. Douglas-fir
    7. Krummholz

Report

Data Preparation

1. Convert dataset file **covertype.csv** into **covertype.arff** using **ArffViewer** (it can be found in the Weka's main menu > Tools > ArffViewer) because Weka Explorer just open data with **.arff** extension (it's like usual .csv file with header information).



2. Reduce the number of data become **10000 rows** as a sample because **too much data** cause some processes in Weka to get stuck (source: here | data reduction process: here).

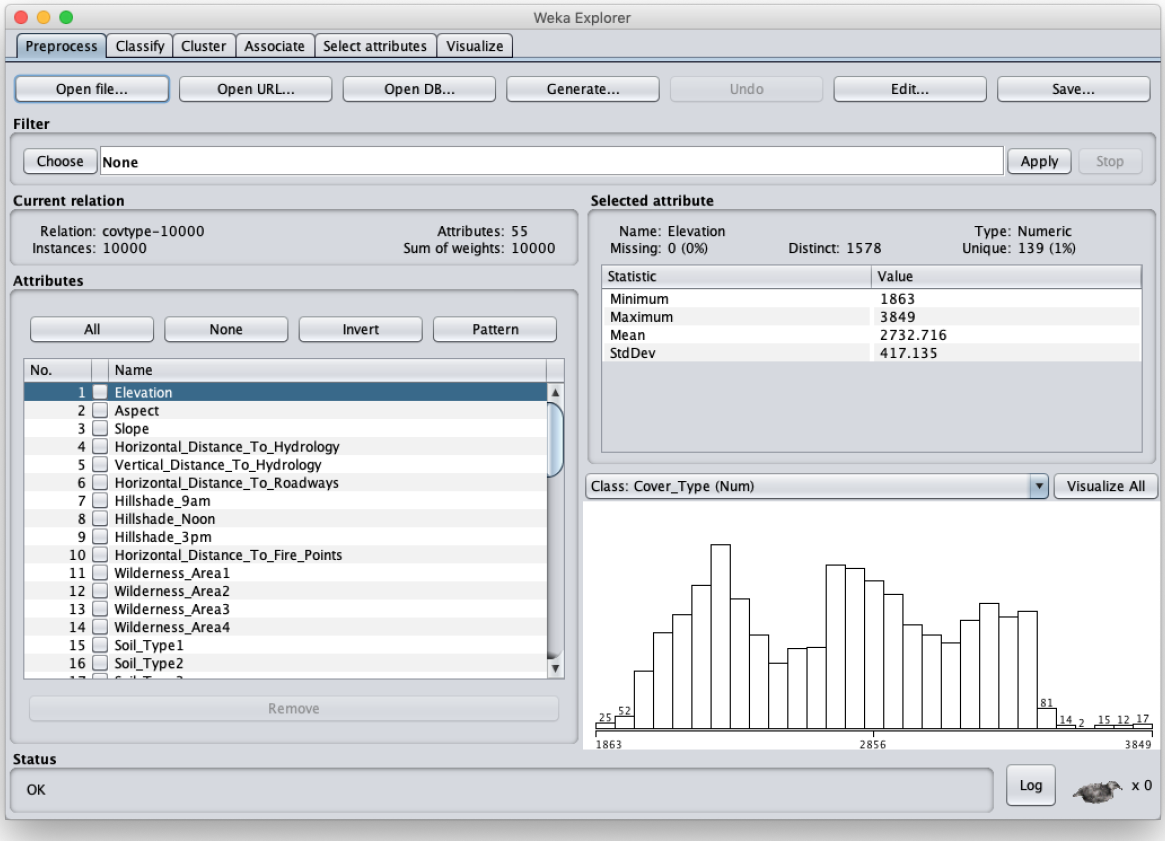
- Here's the **arff** file.

```
@attribute Elevation numeric
@attribute Aspect numeric
@attribute Slope numeric
@attribute Horizontal_Distance_To_Hydrology numeric
@attribute Vertical_Distance_To_Hydrology numeric
@attribute Horizontal_Distance_To_Roadways numeric
@attribute Hillshade_9am numeric
@attribute Hillshade_Noon numeric
@attribute Hillshade_3pm numeric
@attribute Horizontal_Distance_To_Fire_Points numeric
@attribute Wilderness_Area1 numeric
@attribute Wilderness_Area2 numeric
@attribute Wilderness_Area3 numeric
@attribute Wilderness_Area4 numeric
@attribute Soil_Type1 numeric
@attribute Soil_Type2 numeric
@attribute Soil_Type3 numeric
@attribute Soil_Type4 numeric
@attribute Soil_Type5 numeric
@attribute Soil_Type6 numeric
@attribute Soil_Type7 numeric
```

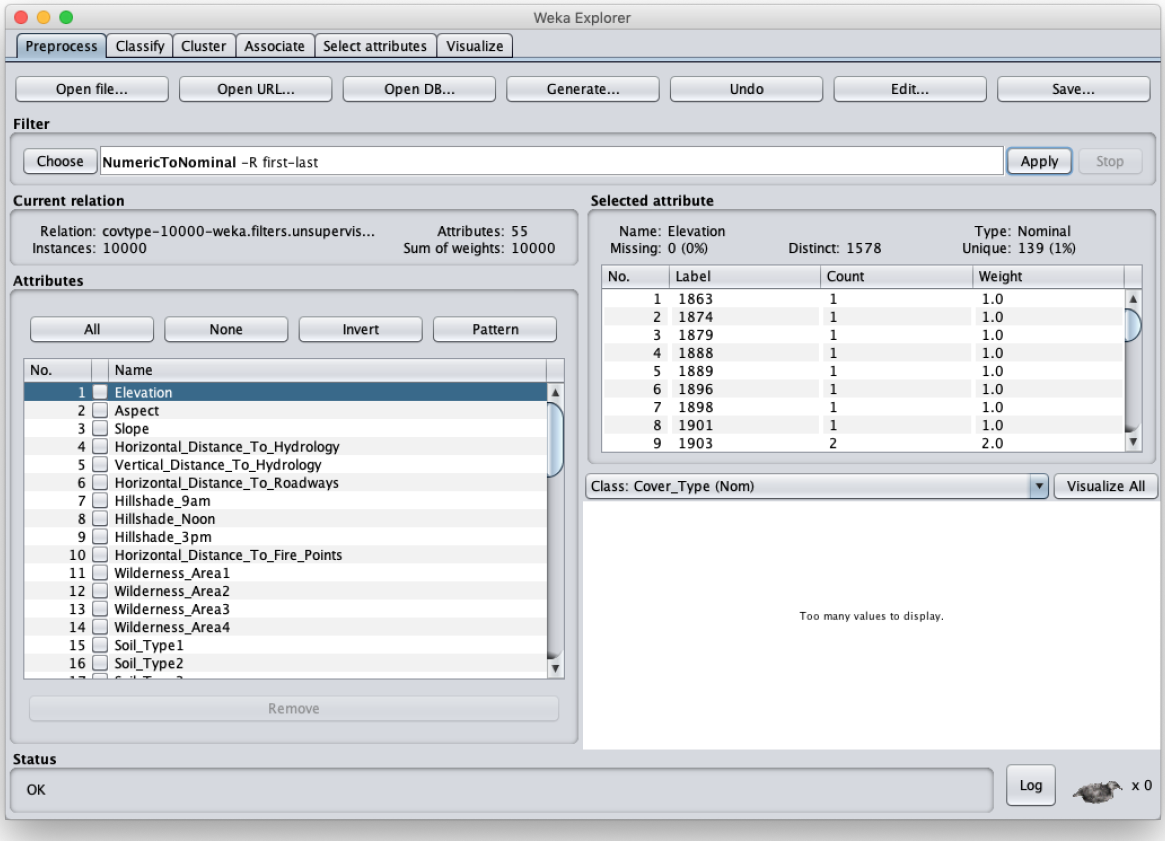
```
@attribute Soil_Type8 numeric
@attribute Soil_Type9 numeric
@attribute Soil_Type10 numeric
@attribute Soil_Type11 numeric
@attribute Soil_Type12 numeric
@attribute Soil_Type13 numeric
@attribute Soil_Type14 numeric
@attribute Soil_Type15 numeric
@attribute Soil_Type16 numeric
@attribute Soil_Type17 numeric
@attribute Soil_Type18 numeric
@attribute Soil_Type19 numeric
@attribute Soil_Type20 numeric
@attribute Soil_Type21 numeric
@attribute Soil_Type22 numeric
@attribute Soil_Type23 numeric
@attribute Soil_Type24 numeric
@attribute Soil_Type25 numeric
@attribute Soil_Type26 numeric
@attribute Soil_Type27 numeric
@attribute Soil_Type28 numeric
@attribute Soil_Type29 numeric
@attribute Soil_Type30 numeric
@attribute Soil_Type31 numeric
@attribute Soil_Type32 numeric
@attribute Soil_Type33 numeric
@attribute Soil_Type34 numeric
@attribute Soil_Type35 numeric
@attribute Soil_Type36 numeric
@attribute Soil_Type37 numeric
@attribute Soil_Type38 numeric
@attribute Soil_Type39 numeric
@attribute Soil_Type40 numeric
@attribute Cover_Type numeric
```

Data Preprocessing

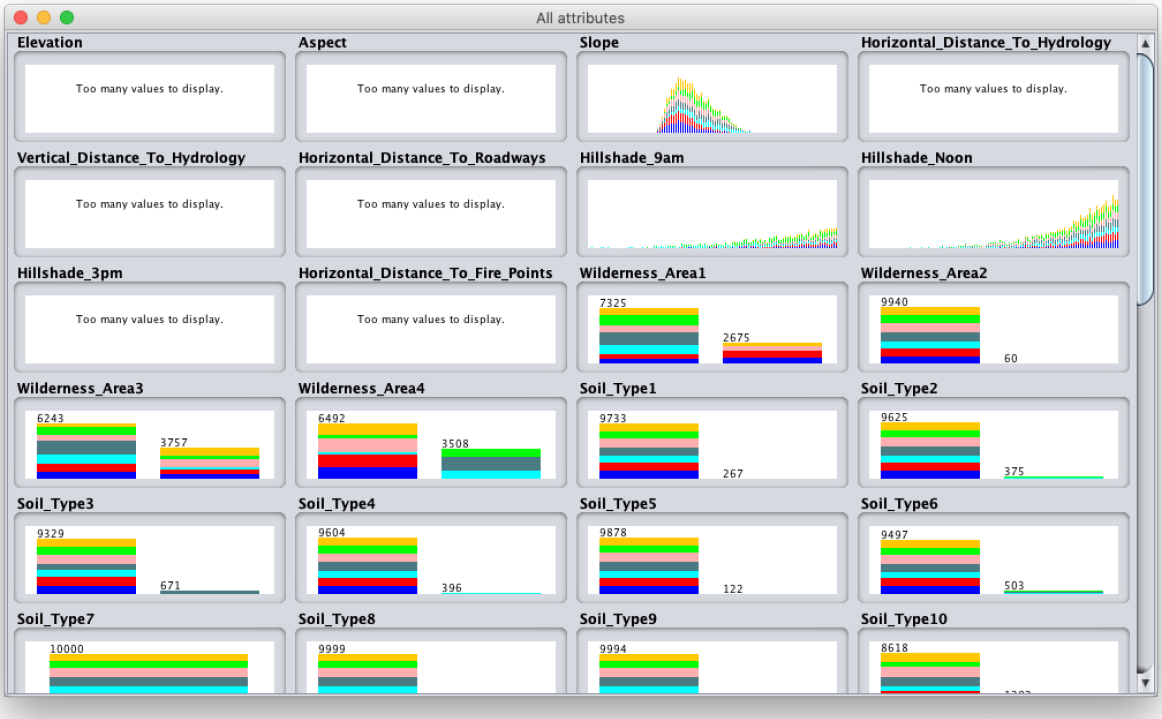
1. Open the dataset with **.arff** extension in the Weka.

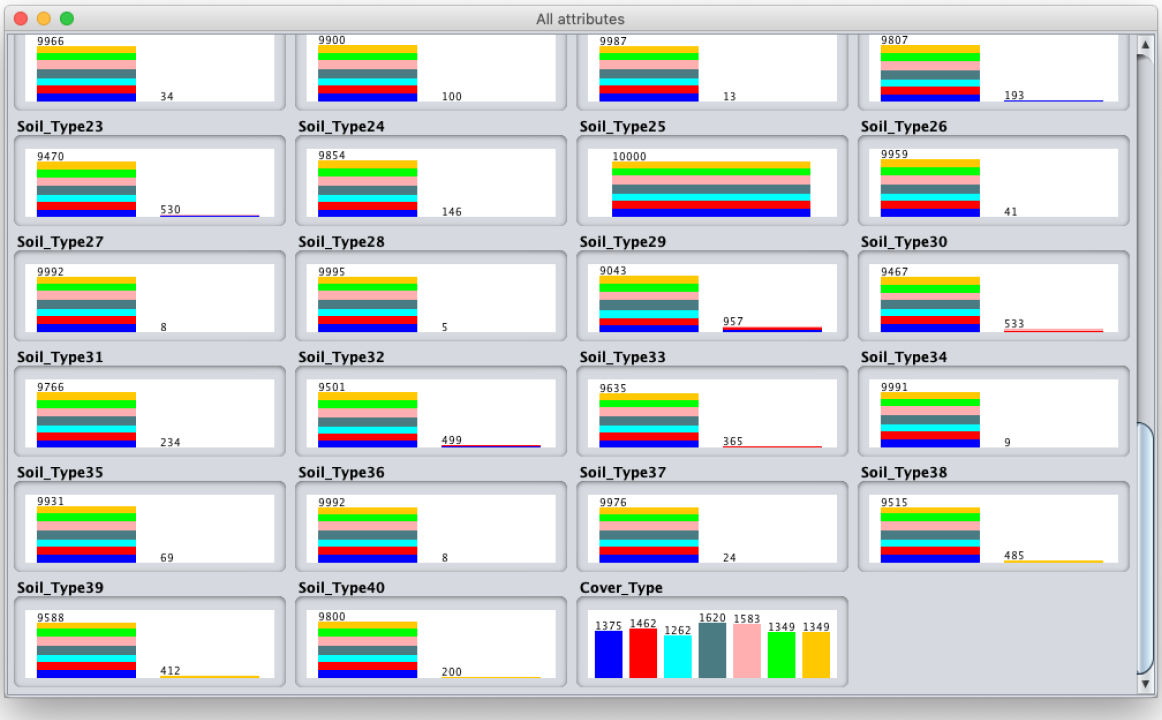


2. Some algorithms must use data with **Nominal** type. So, we must preprocess the data first by using filter **NumericToNominal**.



3. Click **Visualize all** to visualize all attributes.

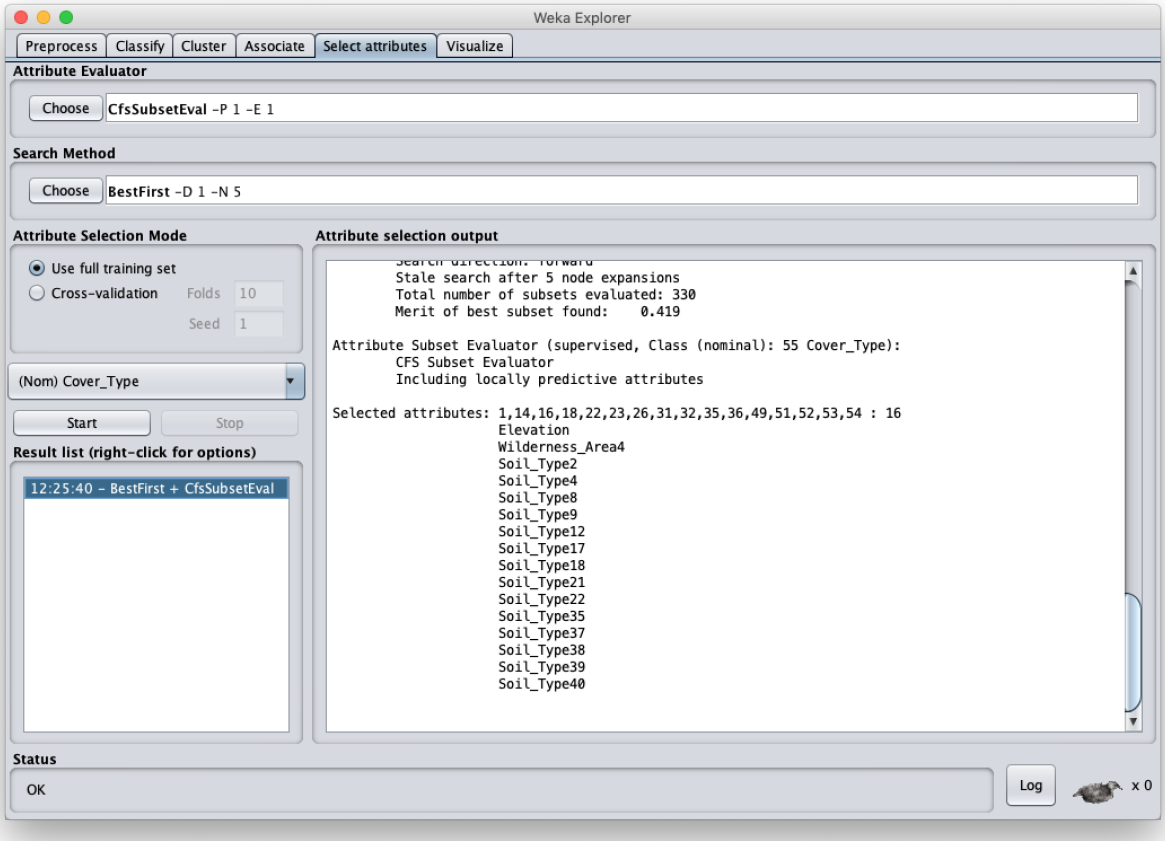




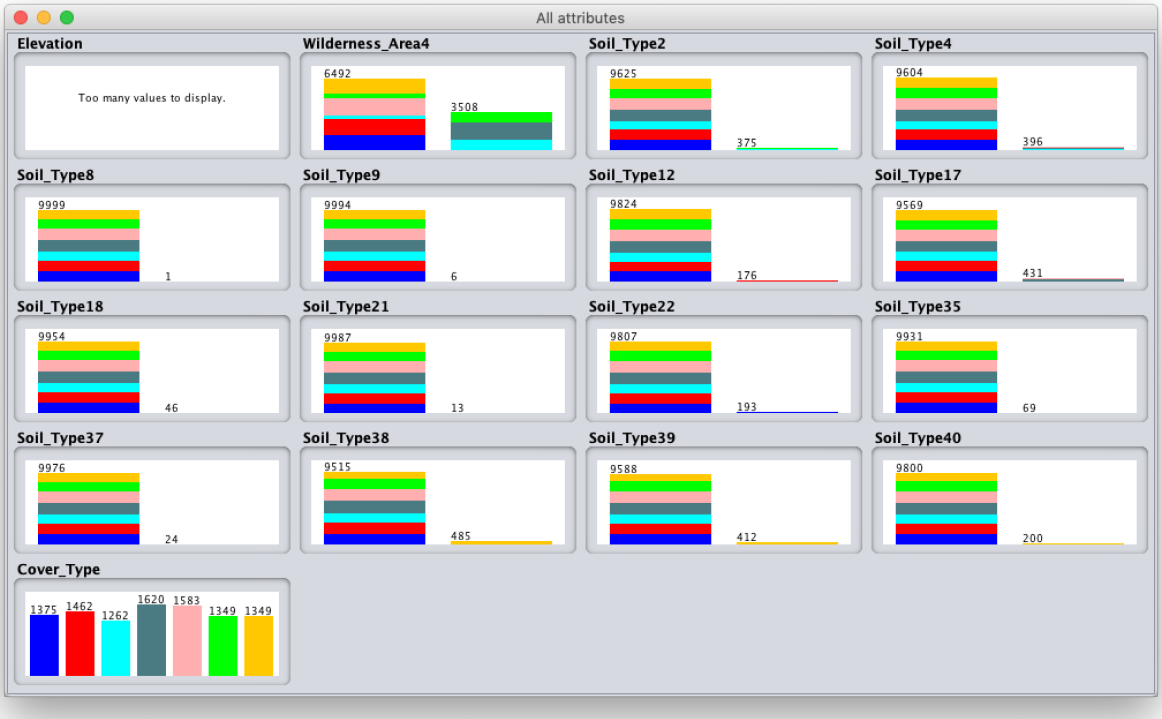
Select Attributes

1. Use default attribute evaluator (**CfsSubsetEval**) and search method (**BestFirst**).

2. Here's the selected attributes.



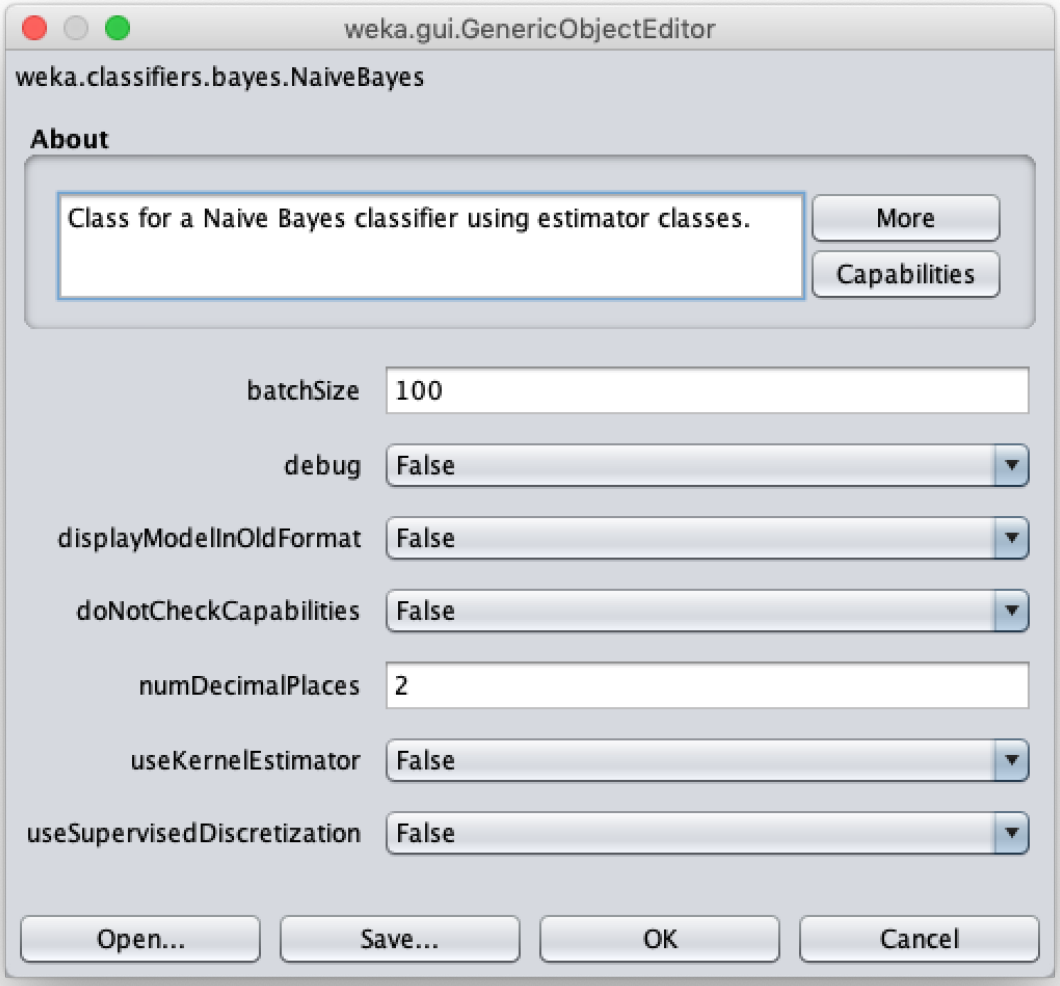
3. Visualize attributes.



Classification - Naive Bayes

1. Click **Classify** tab.

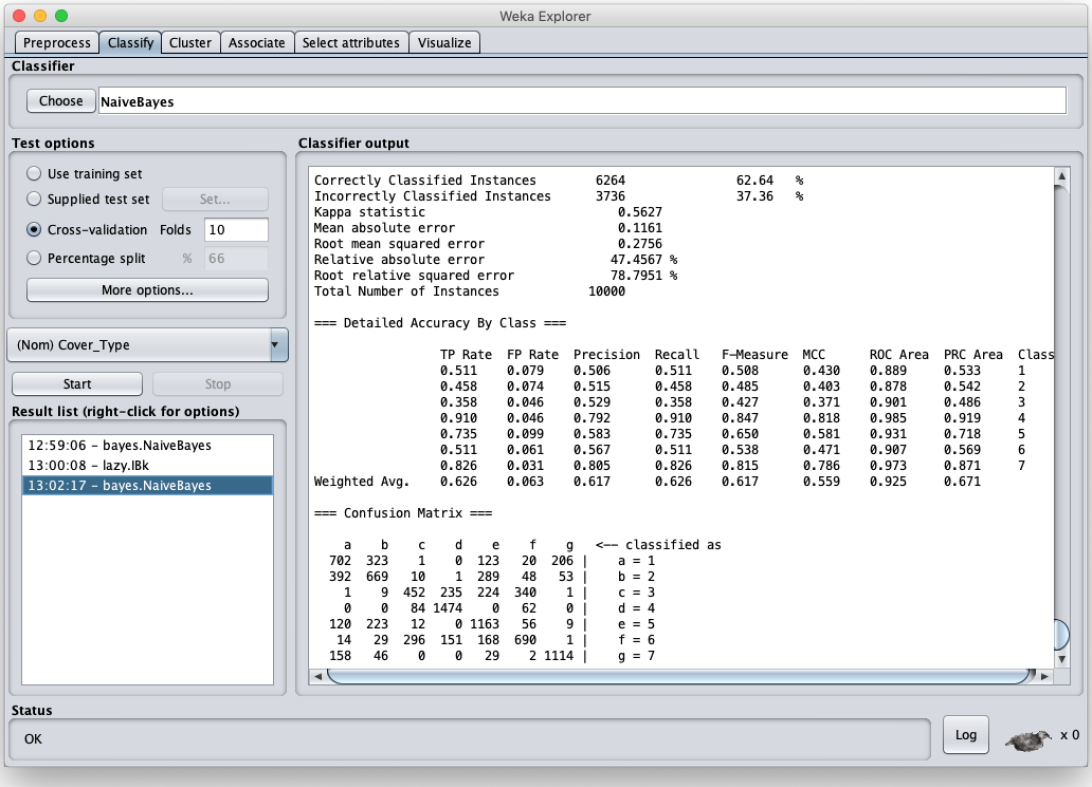
2. Click **Choose** button and select **NaiveBayes** under the **bayes** group.



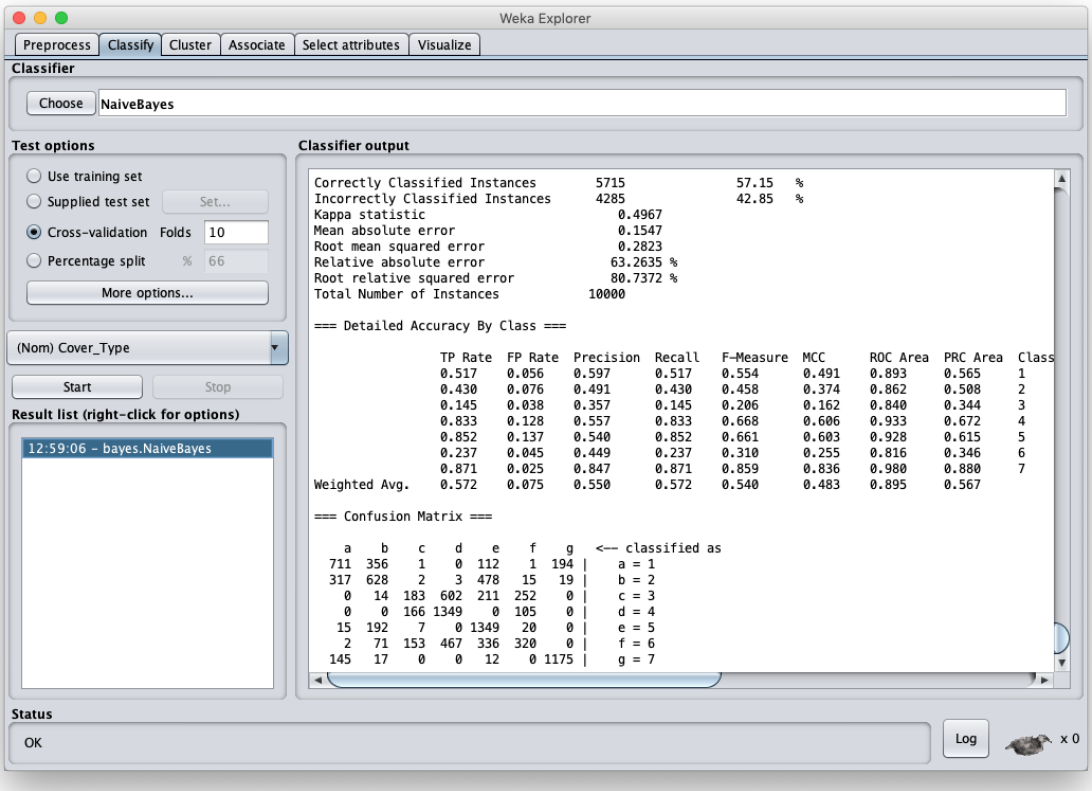
3. Click the **Start** button to run the algorithm on the Covertype dataset.

4. Here's the result.

    - Origin Dataset



- After Attributes Selection

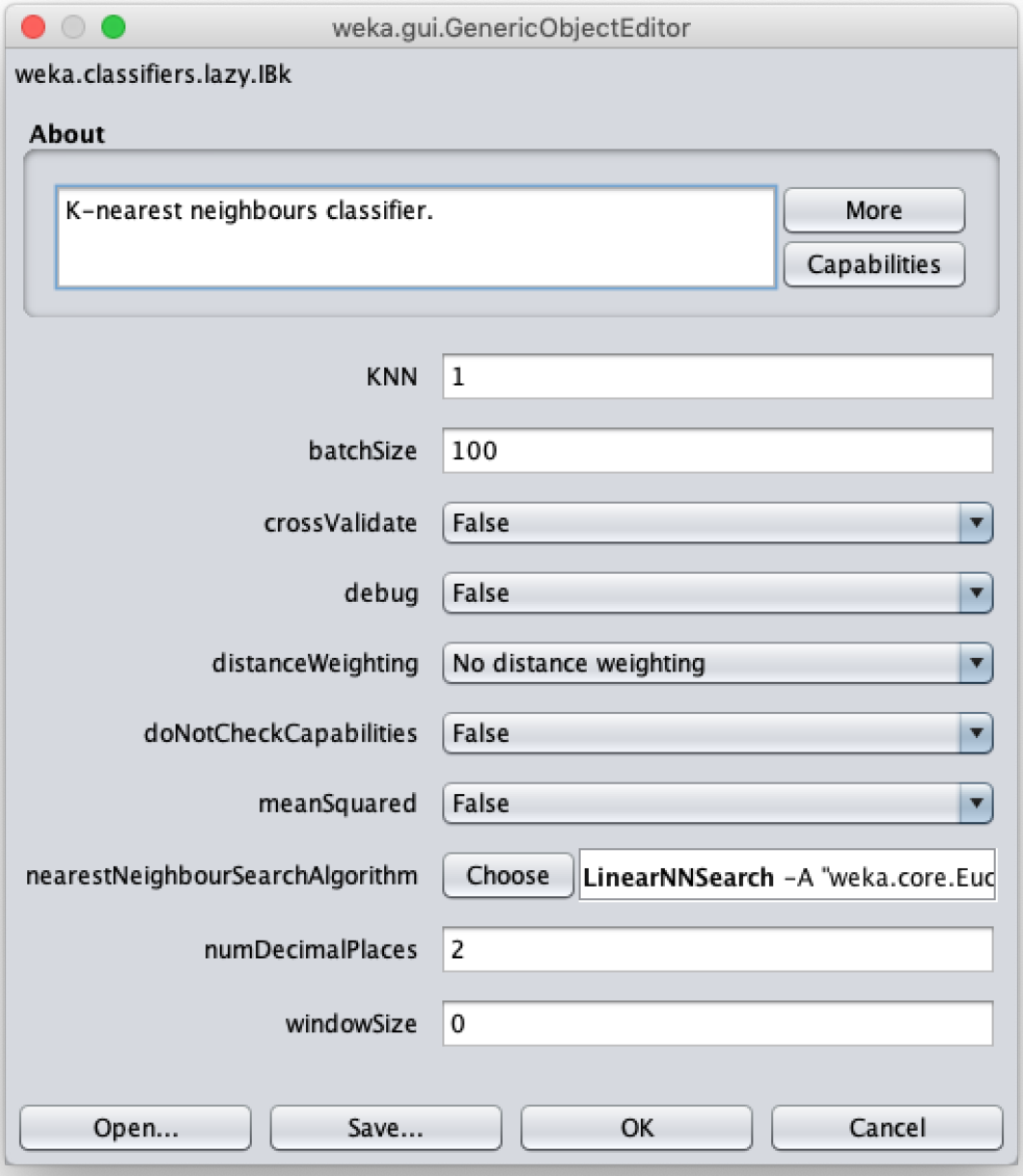


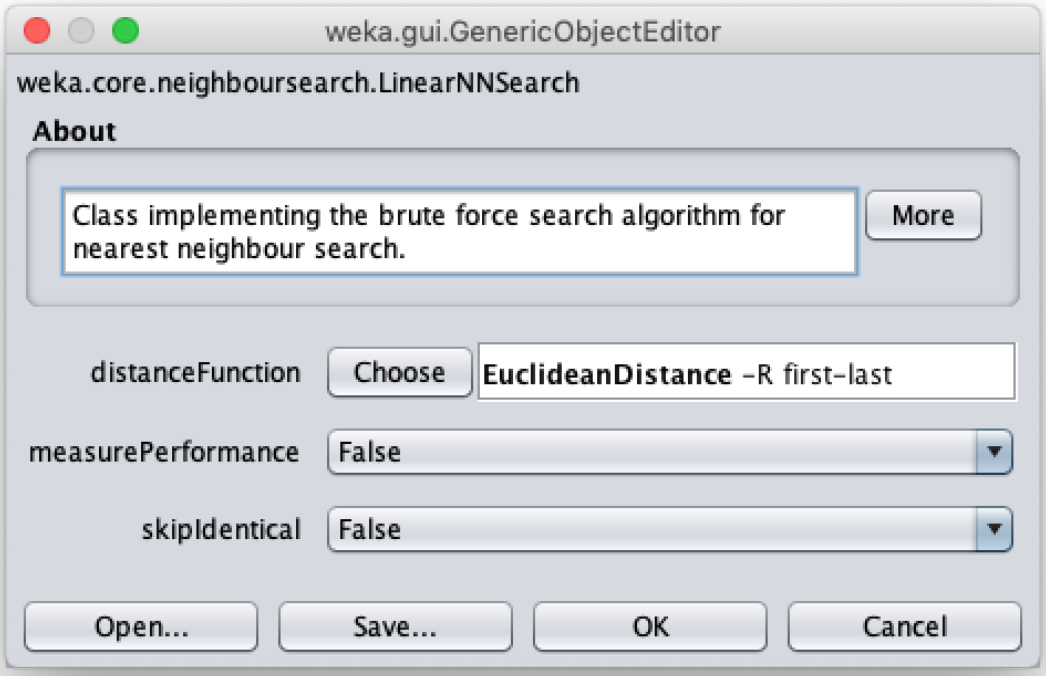
Based on the result above, **origin dataset has better accuracy** than reduced dataset after attribute selection.

Classification - K-Nearest Neighbor (KNN)

1. Still in the **Classify** tab.

2. Click the **Choose** button and select **IBk** under the **lazy** group.

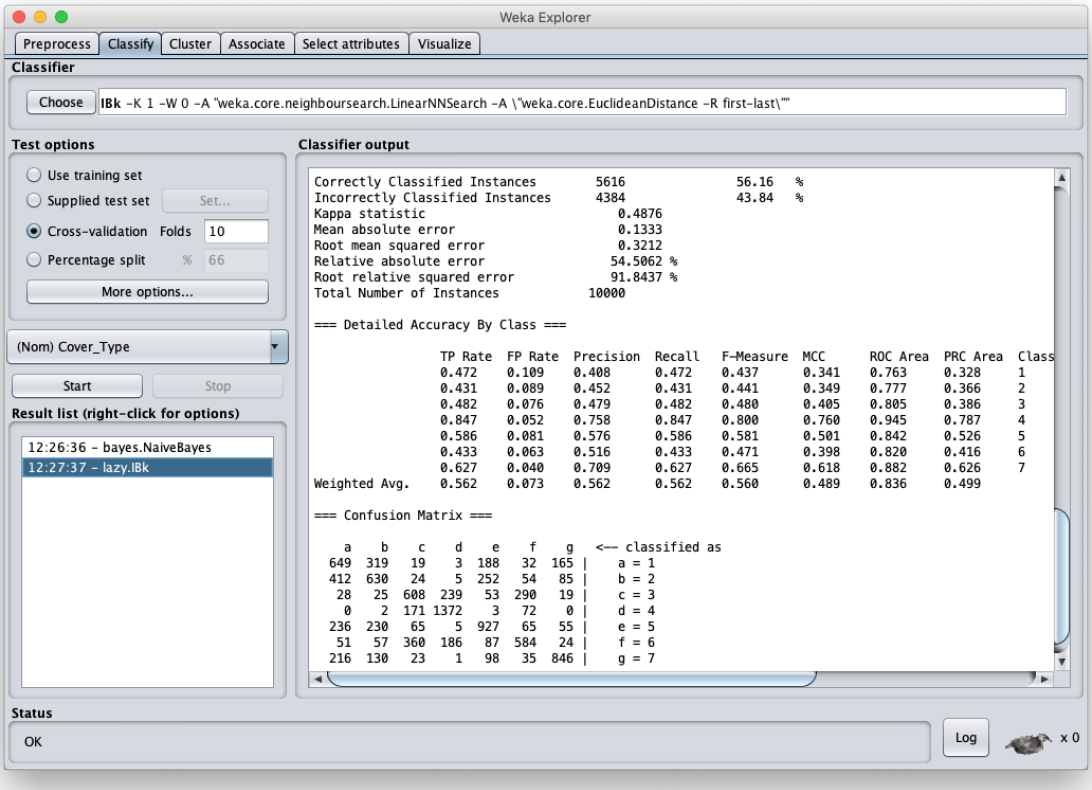




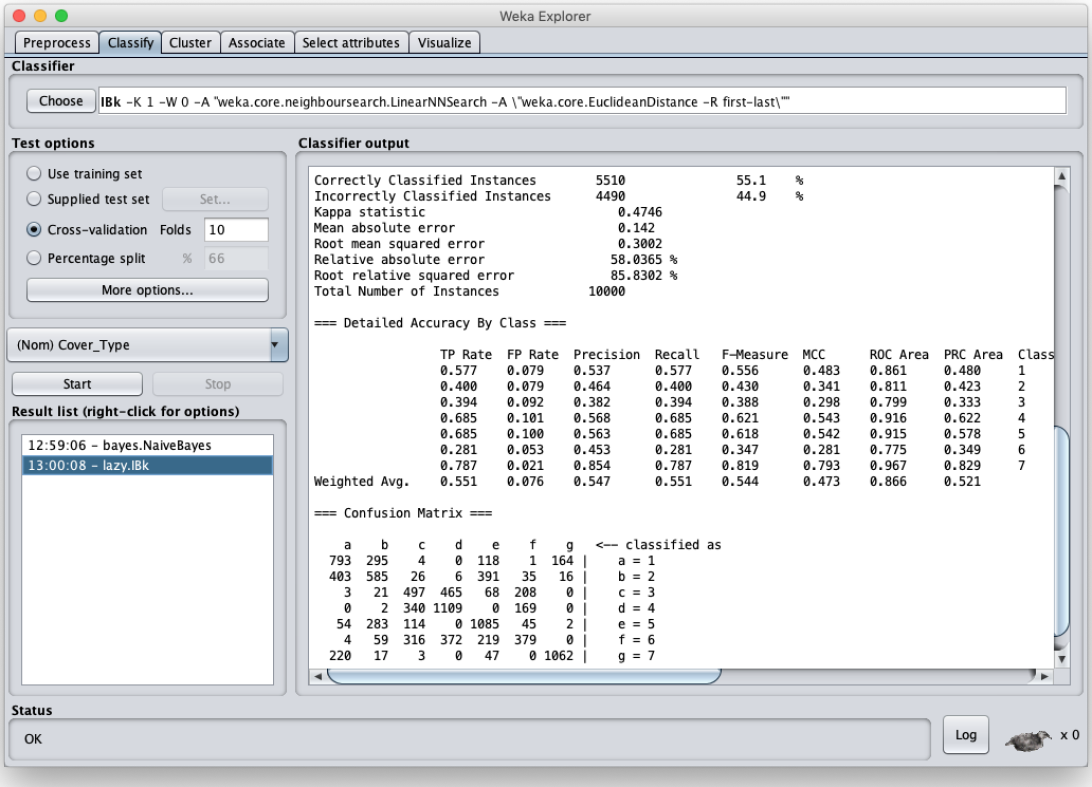
3. Click the **Start** button to run the algorithm on the Covertype dataset.

4. Here's the result using `k = 1`

    - Origin Dataset

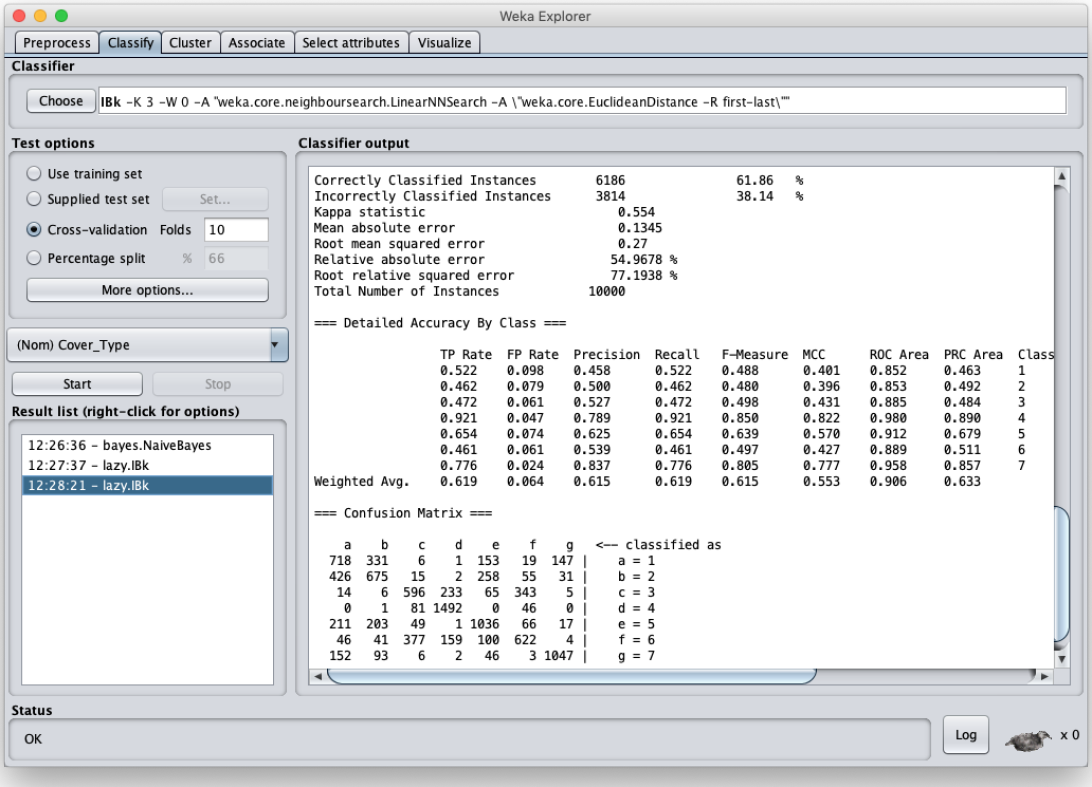


- After Attributes Selection

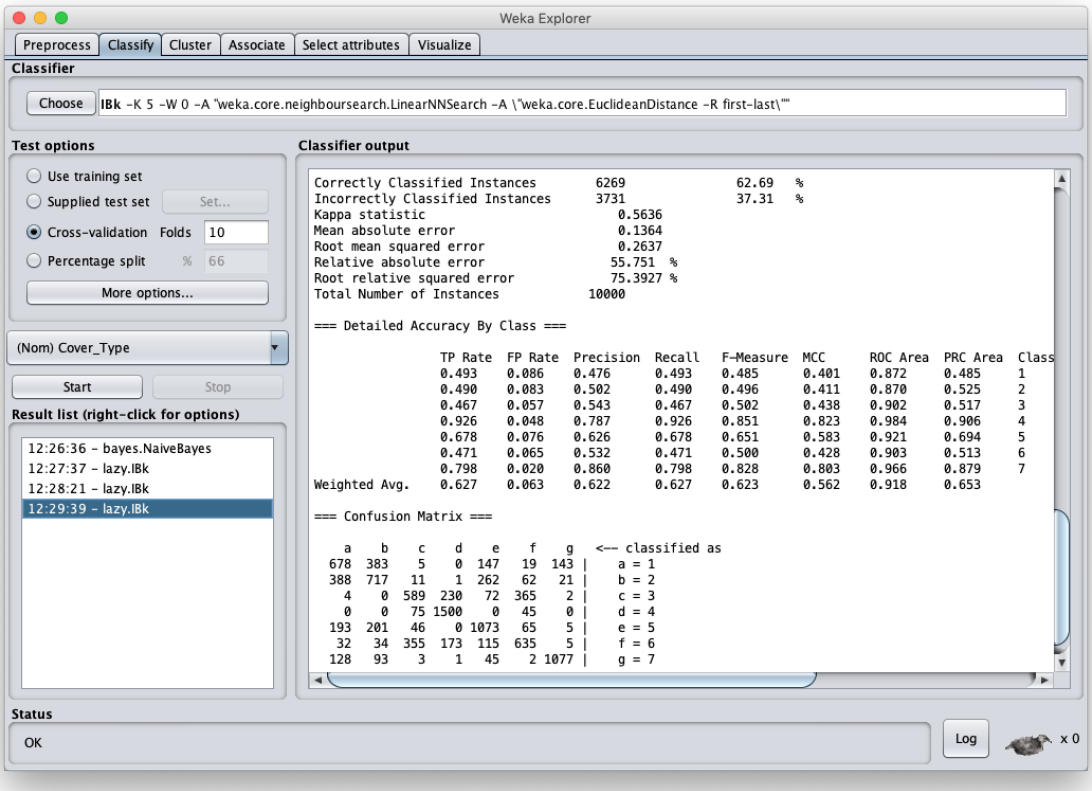


Based on the result above, **origin dataset still has better accuracy** than reduced dataset after attribute selection. So, the reduced dataset will not be used for tests anymore.

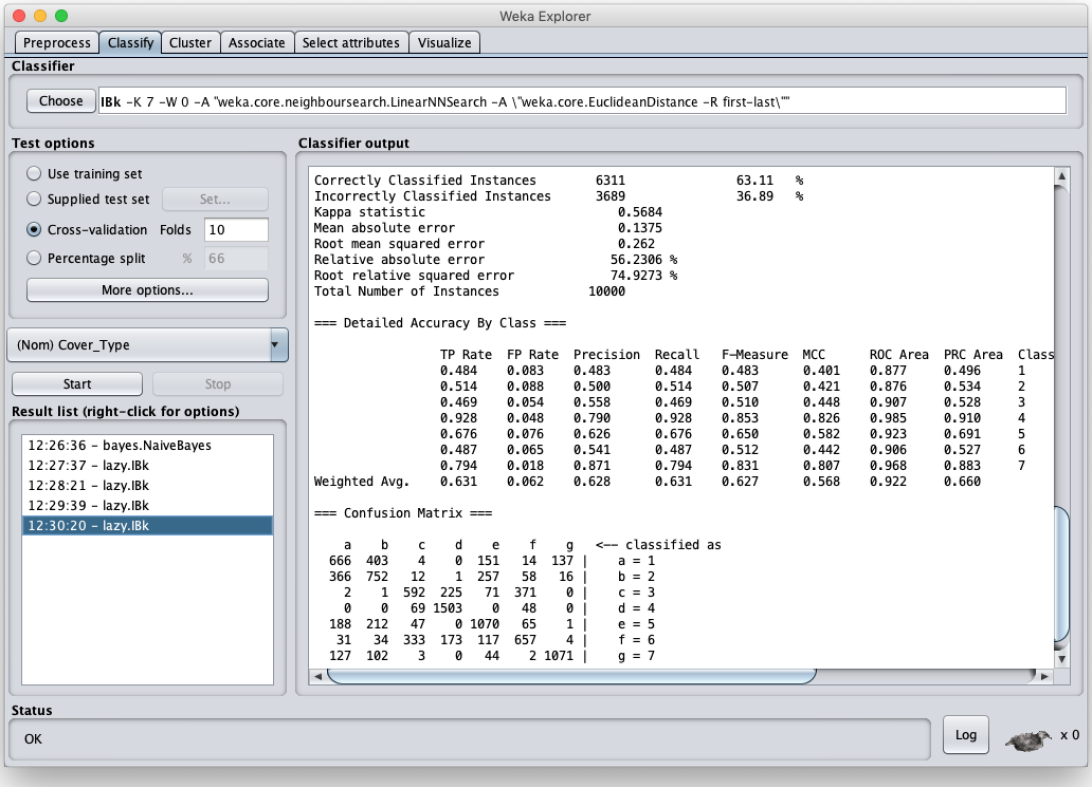
- k = 3


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 3 -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) Cover_Type
Start Stop
Result list (right-click for options)
12:26:36 - bayes.NaiveBayes
12:27:37 - lazy.IBk
12:28:21 - lazy.IBk
Classifier output
Correctly Classified Instances 6186 61.86 %
Incorrectly Classified Instances 3814 38.14 %
Kappa statistic 0.554
Mean absolute error 0.1345
Root mean squared error 0.27
Relative absolute error 54.9678 %
Root relative squared error 77.1938 %
Total Number of Instances 10000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.522 0.098 0.458 0.522 0.488 0.401 0.852 0.463 1
0.462 0.079 0.500 0.462 0.480 0.396 0.853 0.492 2
0.472 0.061 0.527 0.472 0.498 0.431 0.885 0.484 3
0.921 0.047 0.789 0.921 0.850 0.822 0.980 0.890 4
0.654 0.074 0.625 0.654 0.639 0.570 0.912 0.679 5
0.461 0.061 0.539 0.461 0.497 0.427 0.889 0.511 6
0.776 0.024 0.837 0.776 0.805 0.777 0.958 0.857 7
Weighted Avg. 0.619 0.064 0.615 0.619 0.615 0.553 0.906 0.633
=== Confusion Matrix ===
a b c d e f g <-- classified as
718 331 6 1 153 19 147 | a = 1
426 675 15 2 258 55 31 | b = 2
14 6 596 233 65 343 5 | c = 3
0 1 81 1492 0 46 0 | d = 4
211 203 49 1 1036 66 17 | e = 5
46 41 377 159 100 622 4 | f = 6
152 93 6 2 46 3 1047 | g = 7
Status
OK
Log
x 0

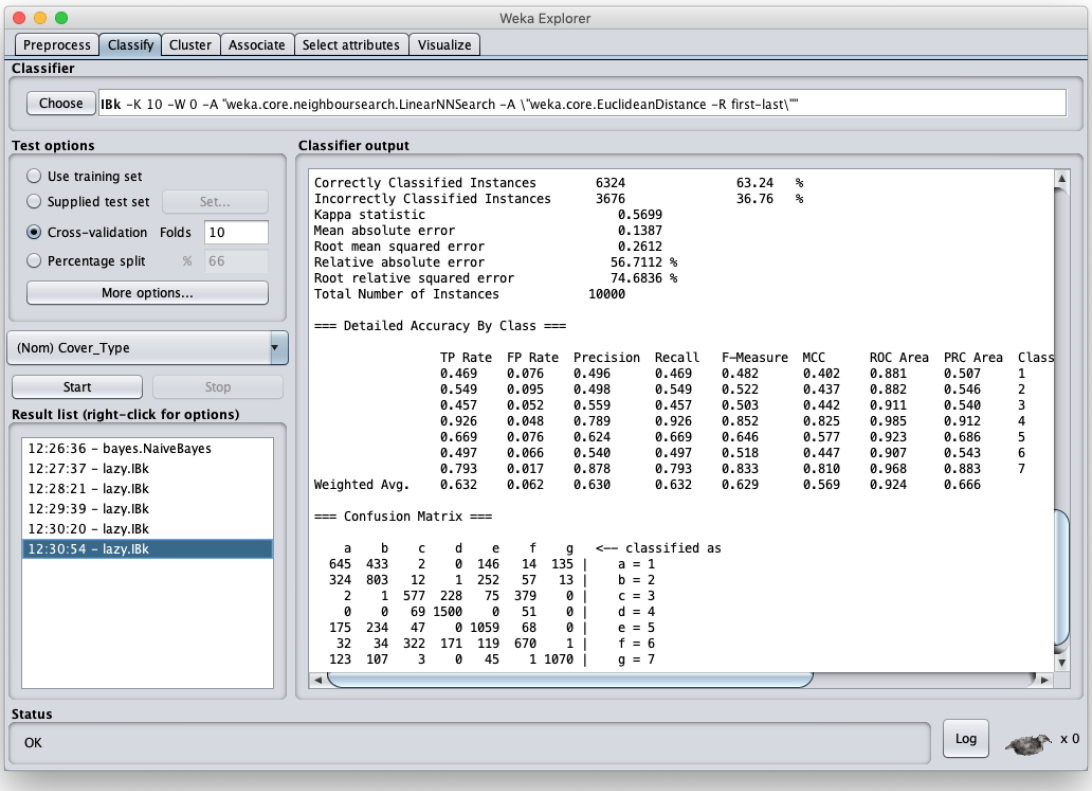
- k = 5


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) Cover_Type
Start Stop
Result list (right-click for options)
12:26:36 - bayes.NaiveBayes
12:27:37 - lazy.IBk
12:28:21 - lazy.IBk
12:29:39 - lazy.IBk
Classifier output
Correctly Classified Instances 6269 62.69 %
Incorrectly Classified Instances 3731 37.31 %
Kappa statistic 0.5636
Mean absolute error 0.1364
Root mean squared error 0.2637
Relative absolute error 55.751 %
Root relative squared error 75.3927 %
Total Number of Instances 10000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.493 0.086 0.476 0.493 0.485 0.401 0.872 0.485 1
0.490 0.083 0.502 0.490 0.496 0.411 0.870 0.525 2
0.467 0.057 0.543 0.467 0.502 0.438 0.902 0.517 3
0.926 0.048 0.787 0.926 0.851 0.823 0.984 0.906 4
0.678 0.076 0.626 0.678 0.651 0.583 0.921 0.694 5
0.471 0.065 0.532 0.471 0.500 0.428 0.903 0.513 6
0.798 0.020 0.860 0.798 0.828 0.803 0.966 0.879 7
Weighted Avg. 0.627 0.063 0.622 0.627 0.623 0.562 0.918 0.653
=== Confusion Matrix ===
a b c d e f g <-- classified as
678 383 5 0 147 19 143 | a = 1
388 717 11 1 262 62 21 | b = 2
4 0 589 230 72 365 2 | c = 3
0 0 75 1500 0 45 0 | d = 4
193 201 46 0 1073 65 5 | e = 5
32 34 355 173 115 635 5 | f = 6
128 93 3 1 45 2 1077 | g = 7
Status
OK
Log
x 0

- k = 7



- k = 10

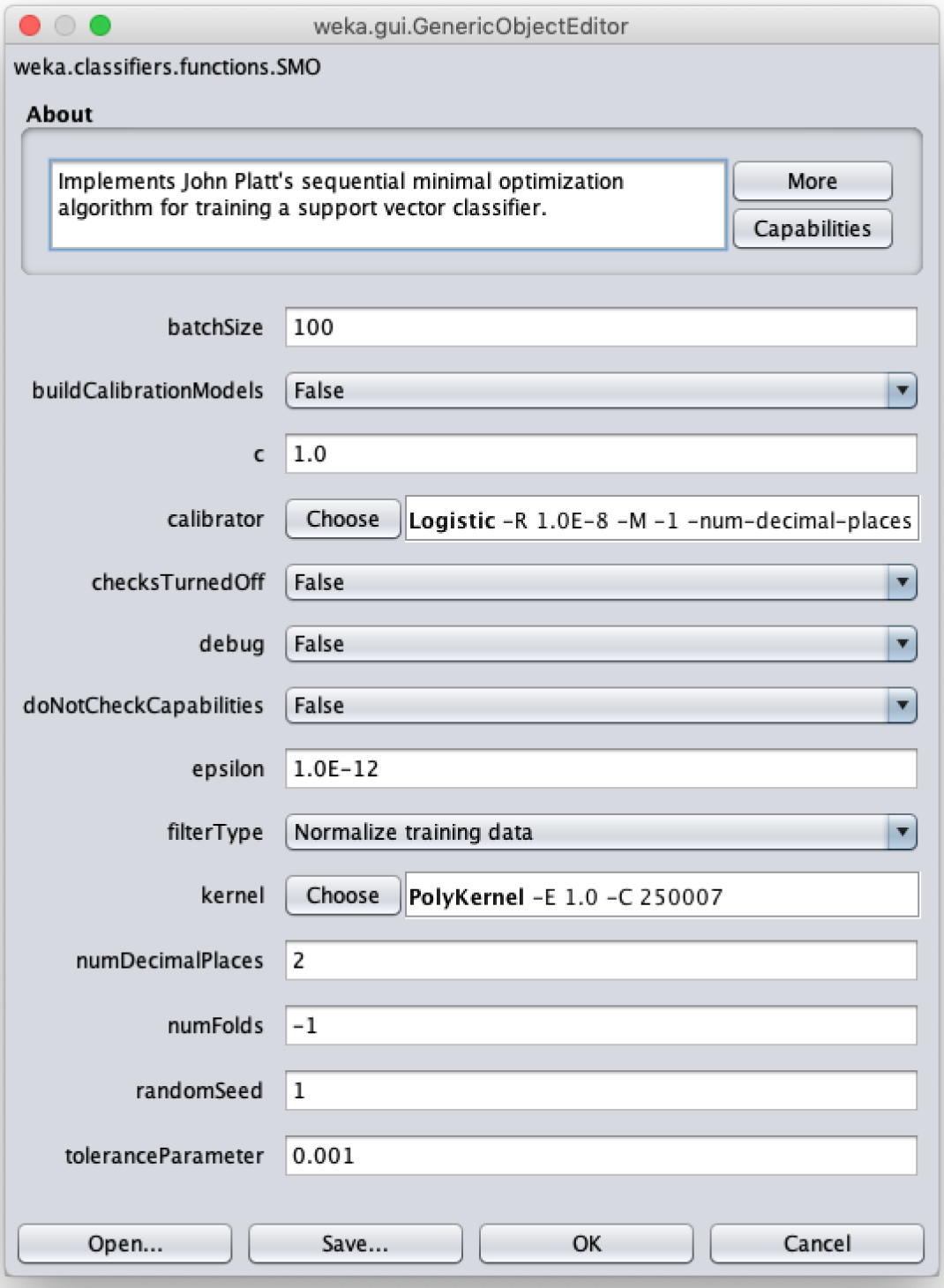


Based on the result above, **the greater k used, the better accuracy result obtained**.

Classification - Support Vector Machine (SVM)

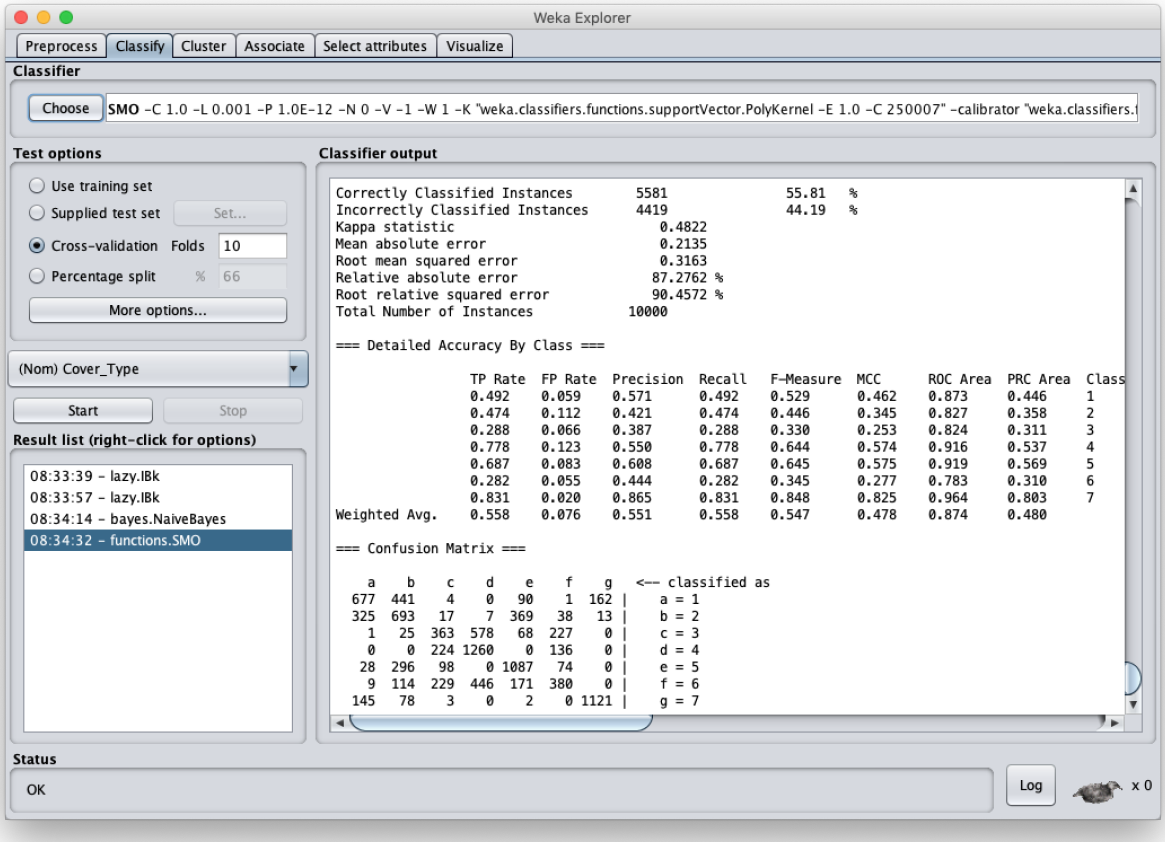
1. Still in the **Classify** tab.

2. Click the **Choose** button and select **SMO** under the **function** group.



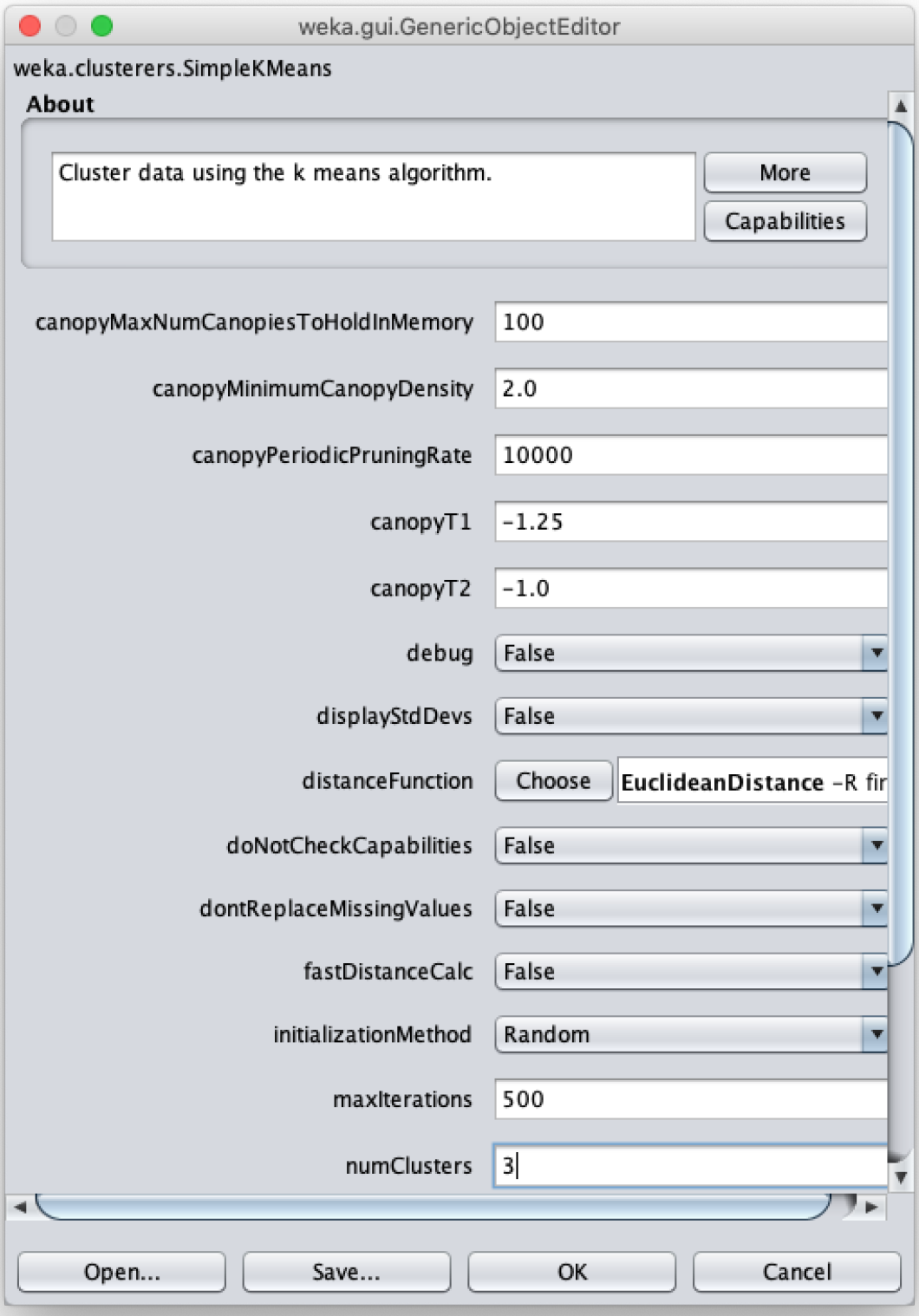
3. Click the **Start** button to run the algorithm on the Covertype dataset.

4. Here's the result.



Clustering - K-Means

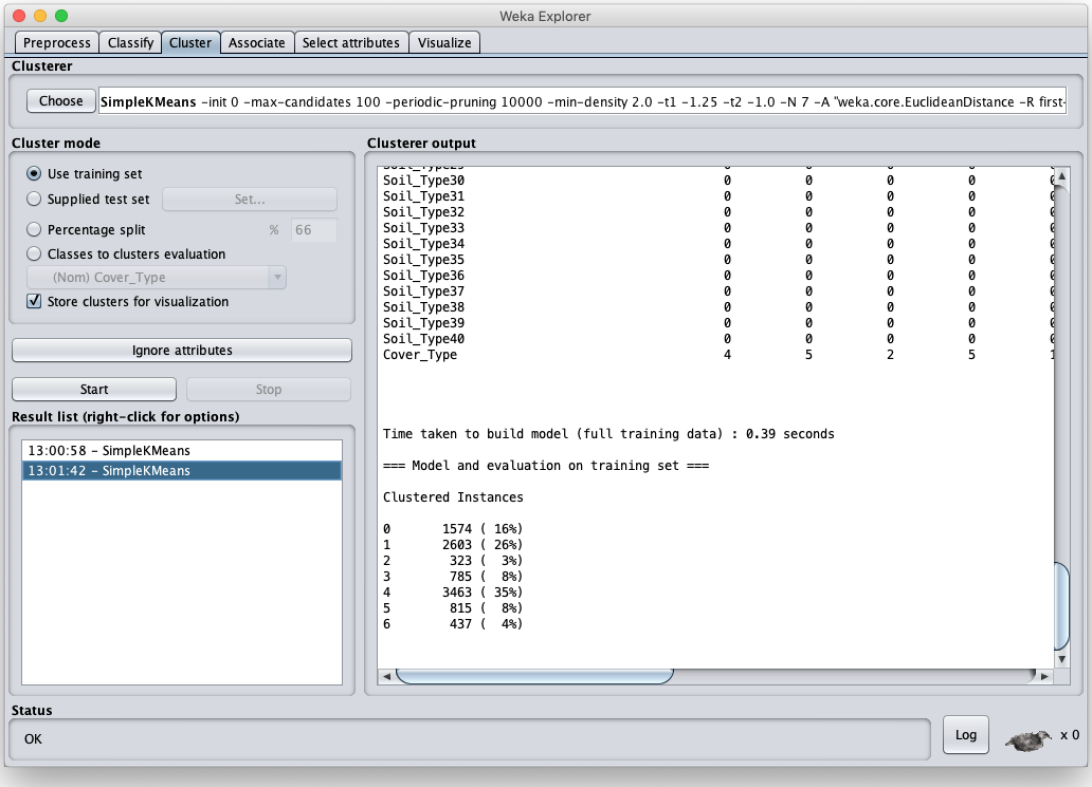
1. Click **Cluster** tab.
2. Click the Clusterer “Choose” button and select .
3. Click the **Choose** button and select **SimpleKMeans** under the **clusterers** group.



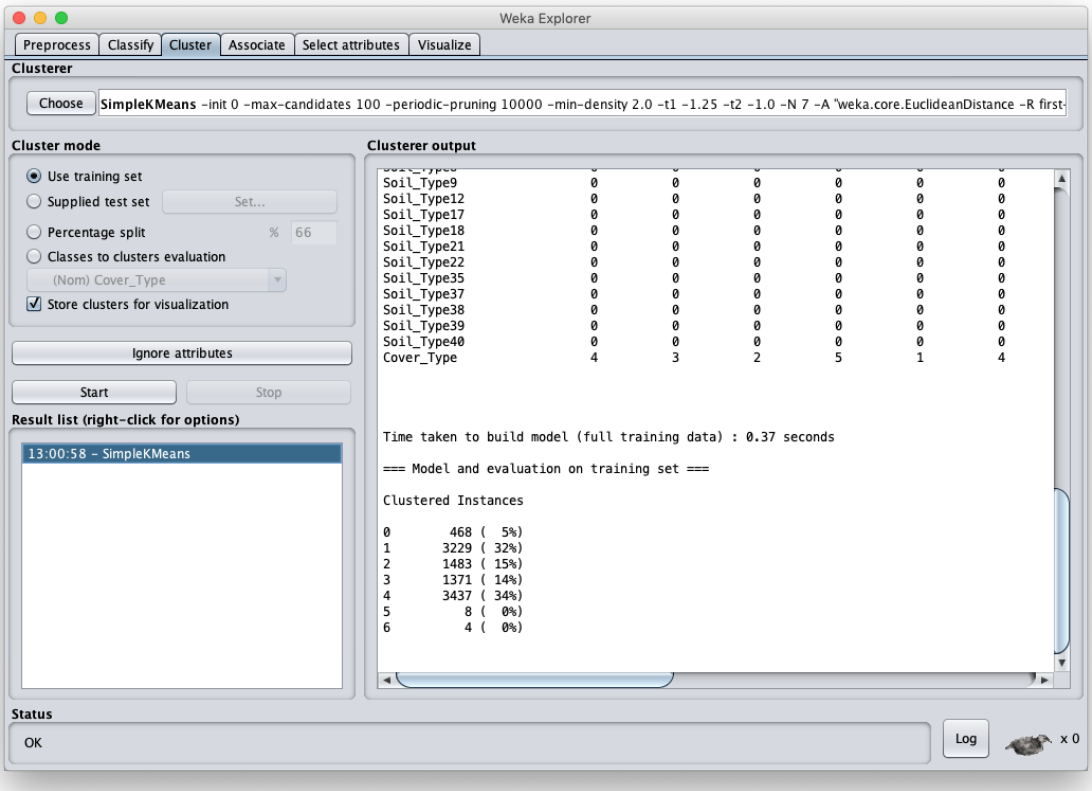
4. Click the **Start** button to run the algorithm on the Covertype dataset.

5. Here's the result using $k = 7$ because there are 7 classes in this dataset.

    - Origin Dataset



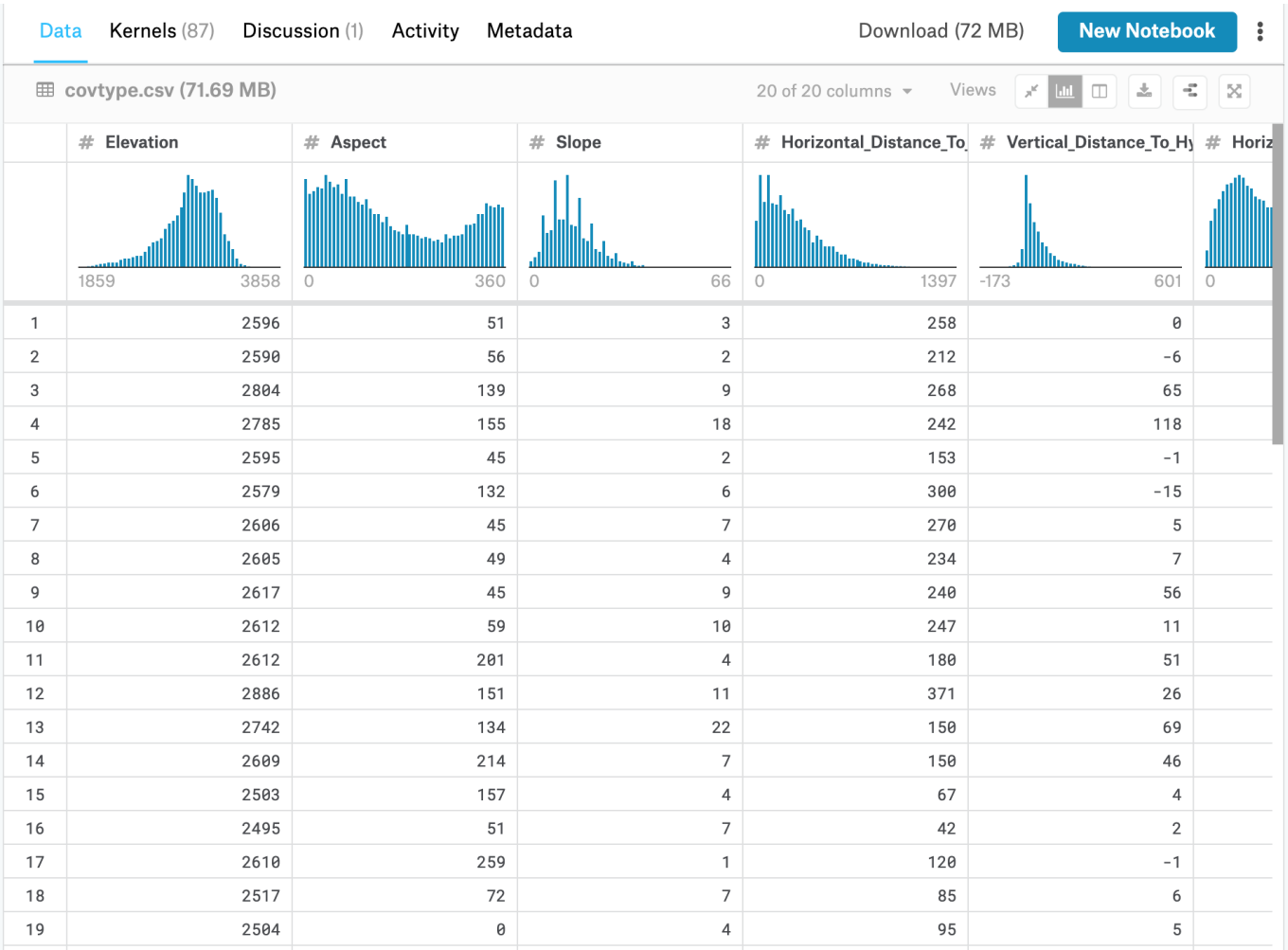
- After Attributes Selection



Conclusion

The best classification method for this dataset in this experiment is **K-Nearest Neighbour** with `k = 10` because it has the highest accuracy result. However if it's using `k = 1` it will get worse accuracy than **Naive Bayes**. The classification method that got the worst accuracy results is **SVM**, so this method is not suitable to classify this dataset.

In fact, this data processing is still bad, because the range of values for each feature in this dataset is too different. So, I think this data must be normalized first.



	Elevation	Aspect	Slope	Horizontal_Distance_To	Vertical_Distance_To_Hy	Horiz
1	2596	51	3	258	0	
2	2590	56	2	212	-6	
3	2804	139	9	268	65	
4	2785	155	18	242	118	
5	2595	45	2	153	-1	
6	2579	132	6	300	-15	
7	2606	45	7	270	5	
8	2605	49	4	234	7	
9	2617	45	9	240	56	
10	2612	59	10	247	11	
11	2612	201	4	180	51	
12	2886	151	11	371	26	
13	2742	134	22	150	69	
14	2609	214	7	150	46	
15	2503	157	4	67	4	
16	2495	51	7	42	2	
17	2610	259	1	120	-1	
18	2517	72	7	85	6	
19	2504	0	4	95	5	

References

- https://archive.ics.uci.edu/ml/datasets/covertype
- https://www.kaggle.com/uciml/forest-cover-type-dataset/