

Data Mining for Scientific Applications
Course No. CSE-40770

Laboratory Assignment II:

To download additional .arff data sets go to:

<http://repository.seasr.org/Datasets/UCI/arff/>

or alternatively

<http://www.hakank.org/weka/>

1. Use the following learning schemes to analyze the zoo data (in zoo.arff):

OneR - weka.classifiers.OneR

Decision table - weka.classifiers.DecisionTable -R

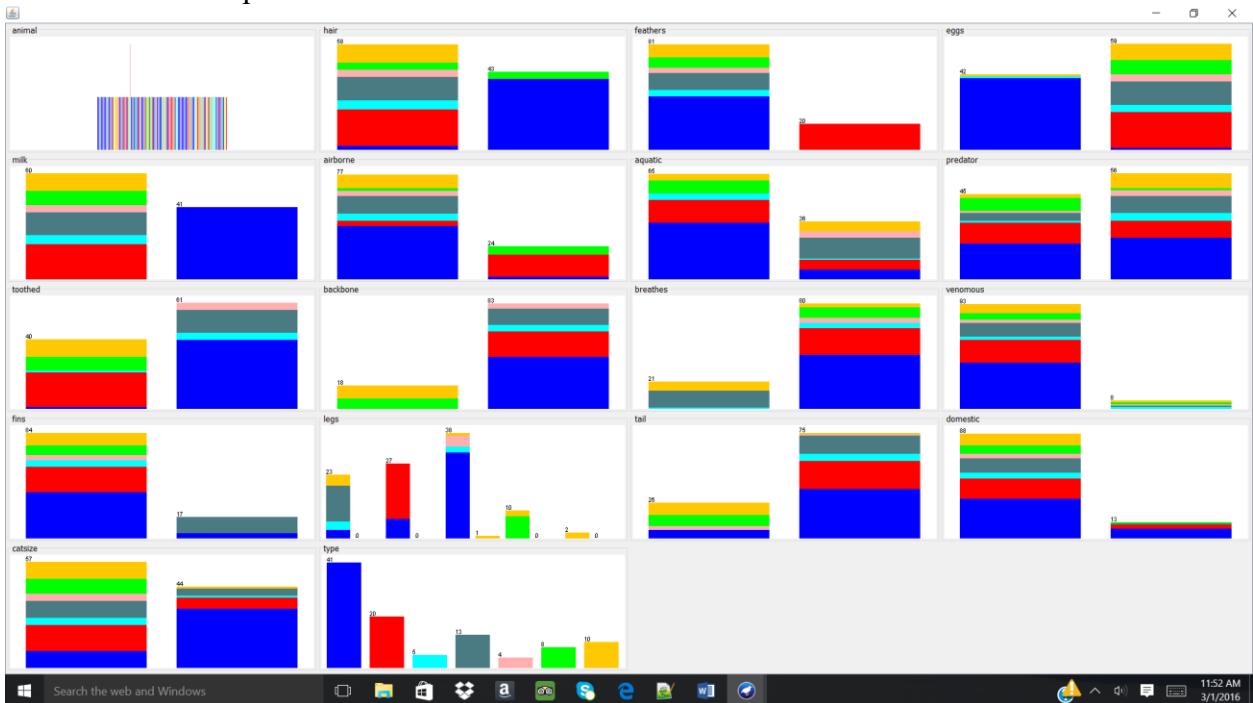
C4.5 - weka.classifiers.j48.J48

K-means - weka.clusterers.SimpleKMeans

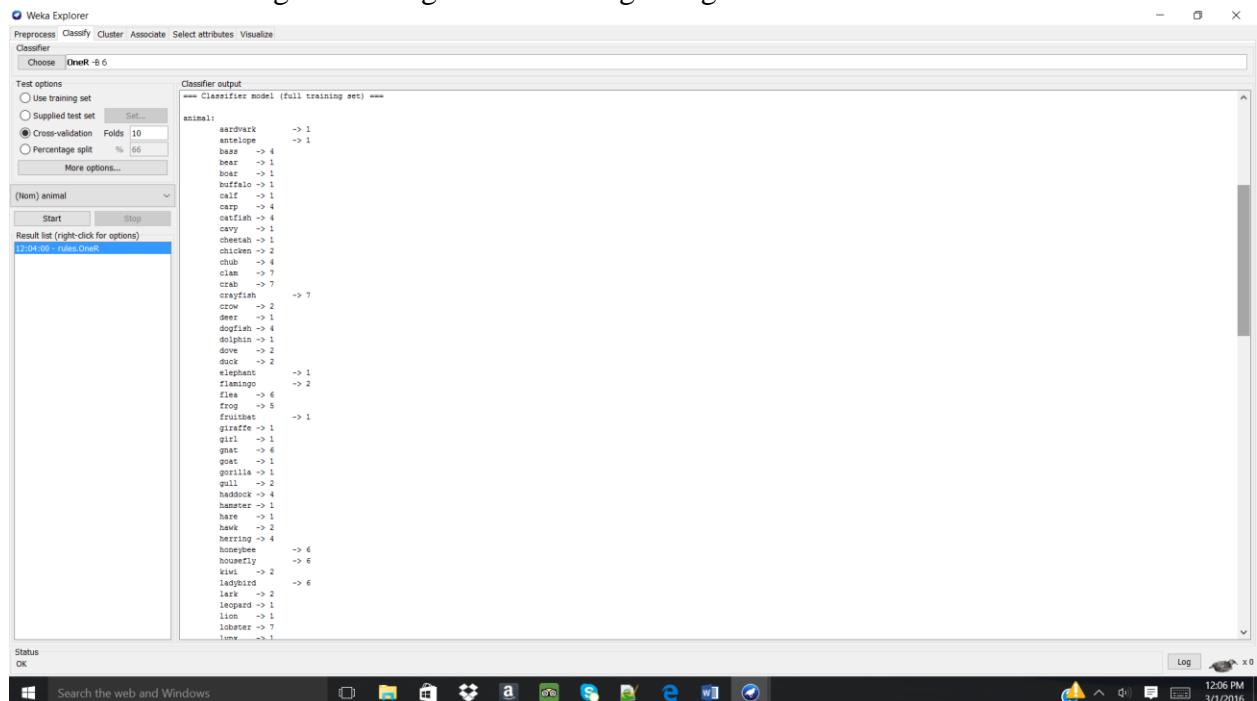
Try using reduced error pruning for the C4.5. Did it change the produced model? Why?

For K-means, for the first run, set $k=10$. Adjust as needed. What was the final number of k ? Why?

A. The zoo.arff dataset was opened in Weka and “Visualize All” was used to visualize the instance distributions for each of the 18 attribute values. Note: all attribute values are nominal values, and attribute “animal” provides a name to each individual instance.



B. Classify/Classifier/rules/OneR was selected. Under “Test options”, “Cross-validation” at “Fold 10” was selected in order to accomplish the most believable evaluations where the data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing. The following was generated:



The screenshot shows the Weka Explorer interface with the "Classify" tab selected. Under "Classifier", "OneR" is chosen. In the "Test options" section, "Cross-validation" is selected with "Folds" set to 10. The "Result list" pane displays the generated rule list:

```

animal:
sandmark    -> 1
antelope     -> 1
bass         -> 4
bee           -> 1
bear          -> 1
buffalo      -> 1
calf          -> 1
carp          -> 4
crocodile   -> 4
cavy          -> 1
cheetah       -> 1
chicken      -> 2
clerk         -> 4
clint         -> 7
crab          -> 7
crayfish     -> 7
crow          -> 2
duck          -> 1
dolphin      -> 4
dove          -> 2
duck          -> 2
elephant     -> 1
flamingo     -> 2
flea          -> 6
frog          -> 5
fruitbat     -> 1
giraffe       -> 1
girl          -> 1
gnat          -> 6
goat          -> 1
grizzly       -> 1
hail          -> 2
haddock     -> 4
hamster       -> 1
hare          -> 1
hawk          -> 2
herring      -> 4
honeybee     -> 6
housefly    -> 2
ladybird     -> 6
lark          -> 2
leopard      -> 1
lion          -> 1
lobster      -> 7
luna          -> 1

```

Note: the class attribute is “type”. The algorithm creates a model based on a single attribute, to determine the classes for the datasets. The attribute “animal” is selected as providing the most information gain for the data and the OneR algorithm uses numeric restrictions in determining which instances fall under the class “type”. This rule maps 43 instances correctly, but 58 incorrectly with a mean absolute error of 0.1641. Although more information is learned from the OneR rule in comparison to the ZeroR rule, error still exists in the system showing that a single rule based on one attribute cannot map out all of the instances properly, therefore a more complex rule/model needs to be used.

C. Classify/Classifier/rules/DecisionTable was selected. Under “Test options”, “Cross-validation” at “Fold 10” was selected in order to accomplish the most believable evaluations where the data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing. The following was generated:

Orysya Stus

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose: DecisionTable -x 1-S "weka.attributeSelection.BestFirst-D 1-N 5"
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)
12:04:00 - rules.OneR
12:12:01 - rules.DecisionTable
Time taken to build model: 0.04 seconds

Decision Table:
Number of training instances: 101
Number of folds: 10
Run matches correctly by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Number of nodes after node expansions: 17
Total number of subsets evaluated: 121
Merit of best subset found: 93.069
Evaluation (for feature selection): CV (leave one out)
Feature set: 5,13,14,15,16

Correctly Classified Instances      89      88.1188 %
Incorrectly Classified Instances   12      11.8812 %
Kappa statistic                   0.8978
Mean absolute error               0.1231
Root mean squared error          0.205
Relative absolute error           56.1596 %
Root relative squared error      62.1351 %
Total Number of Instances        101

Detailed Accuracy By Class:
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
1       0.133    0.837    0.911    0.994    1
1       0         1         1         1         1         2
0       0         0         0         0         0.884    3
1       0.011    0.929    1         0.963    0.998    4
0.75   0.01     0.75     0.75     0.982    0.982    5
1       0.011    0.929    1         0.963    0.998    6
0.4    0.011    0.8     0.4      0.833    0.896    7
Weighted Avg. 0.881    0.058    0.837    0.881    0.849    0.98

Confusion Matrix:
a b c d e f g <- classified as
41 0 0 0 0 0 0 1 a = 1
0 20 0 0 0 0 0 1 b = 2
2 0 0 1 0 0 1 1 c = 3
0 0 0 13 0 0 0 1 d = 4

```

17 rules were generated with 89/101 instances classified correctly. The mean absolute error was lower compared to the OneR, at 0.1231.

D. Classify/classifiers/trees/J48 was selected. Under “Test options”, “Cross-validation” at “Fold 10” was selected. Default settings were applied and the following was generated:

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose: J48 -C 0.25 -M 2
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)
12:04:00 - rules.OneR
12:12:01 - rules.DecisionTable
12:26:20 - trees.J48
Classifier output
J48 pruned tree
feathers false
| milk = true
| | backbone = false
| | | airbone = false
| | | | predator = false
| | | | | legs = 0: 7 (0.0)
| | | | | legs = 1: 6 (0.0)
| | | | | legs = 2: 6 (0.0)
| | | | | legs = 3: 6 (0.0)
| | | | | legs = 4: 6 (0.0)
| | | | | legs = 5: 6 (0.0)
| | | | | legs = 6: 6 (0.0)
| | | | | legs = 7: 6 (0.0)
| | | | | legs = 8: 6 (0.0)
| | | | | | legs = 9: 6 (0.0)
| | | | | | | species = trout: 5 (0.0)
| | | | | | | airbone = trout: 4 (0.0)
| | | | | | | backbone = true
| | | | | | | fins = false
| | | | | | | tail = true: 5 (3.0)
| | | | | | | tail = false: 0 (0.0/1.0)
| | | | | | | fins = trout: 4 (13.0)
| | | | | | | milk = trout: 1 (41.0)
| | | | | | | feathers = true: 2 (20.0)

Number of Leaves : 17
Size of the tree : 25

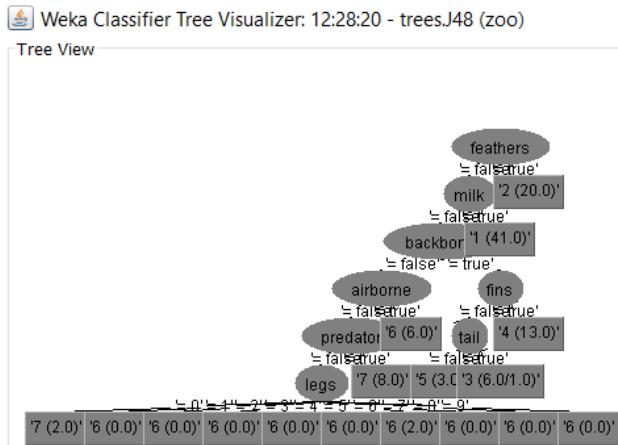
Time taken to build model: 0.02 seconds

Detailed Accuracy By Class:
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
93      7.9205 %
Incorrectly Classified Instances   8      0.8955
Kappa statistic                   0.0225
Mean absolute error               0.1275
Root mean squared error          10.2478 %
Relative absolute error           41.6673 %
Root relative squared error      Total Number of Instances 101

Confusion Matrix:
a b c d e f g <- classified as
41 0 0 0 0 0 0 1 a = 1
0 20 0 0 0 0 0 1 b = 2
2 0 0 1 0 0 1 1 c = 3
0 0 0 13 0 0 0 1 d = 4

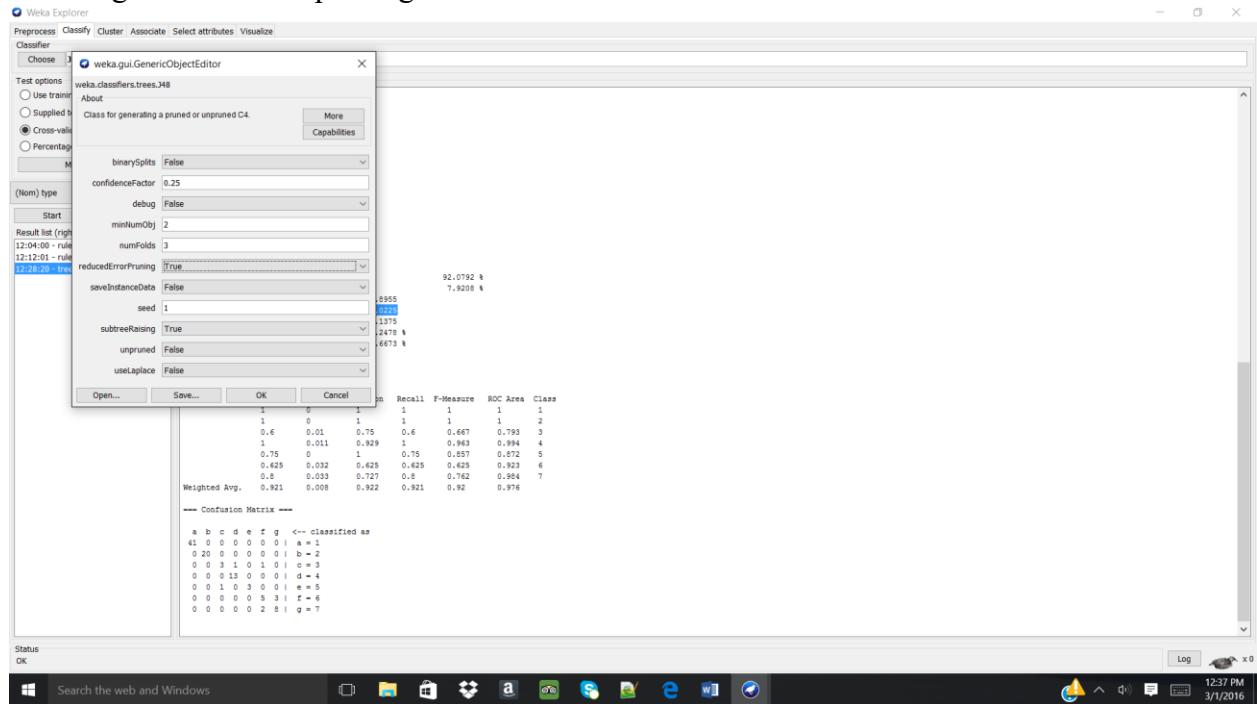
```

Orysya Stus



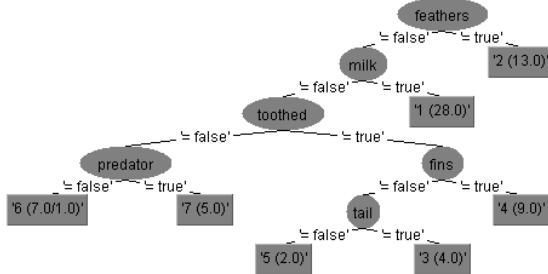
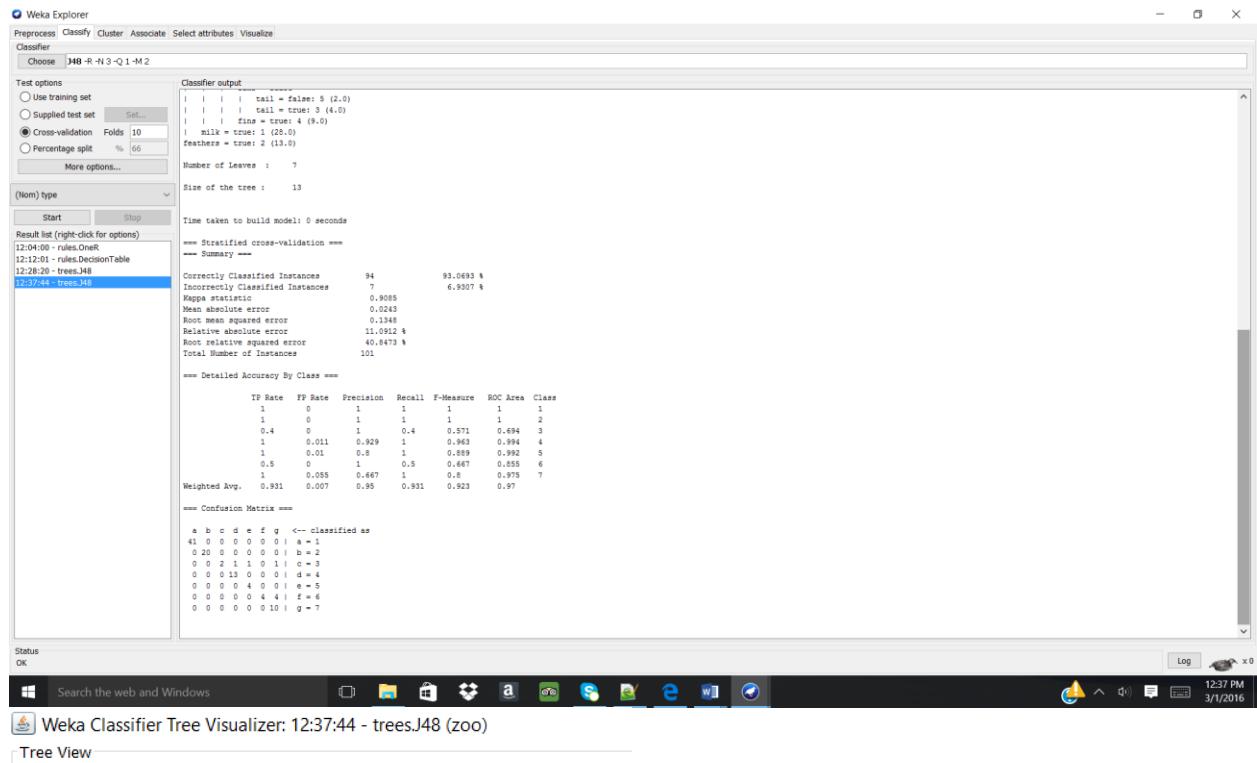
Here, the attribute “feathers” provides the most information gain in determining the class attribute “type”. The tree generated was able to classify 93/101 instances correctly, with a mean absolute error of 0.0225. Thus, the generated decision tree provides the most accurate information gain in classifying the data set.

Selecting reduced error pruning to “True”:



Generated the following decision tree:

Orysya Stus

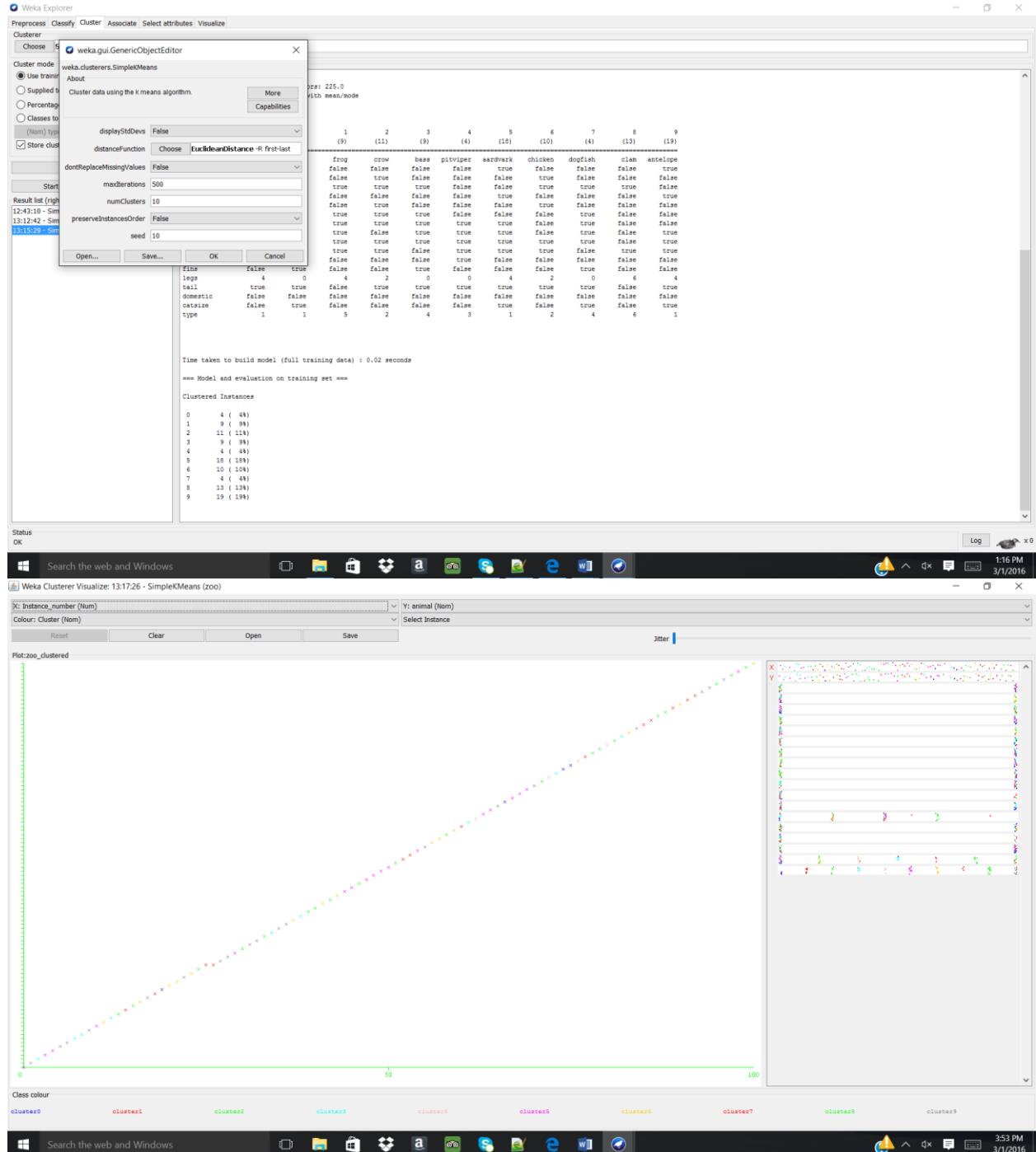


Compared to the tree generated above, selecting reduced error pruning creates a cleaner model which actually correlates to 94/101 classified correctly with a mean absolute error of 0.0243. Although the error is slightly larger, the model works well.

E. Cluster/SimpleKMeans was selected. Under test options, “Full training data” was selected in order to use a substantial amount of training data to generate a model which can be tested upon.

Orysya Stus

The following was generated with k=10:

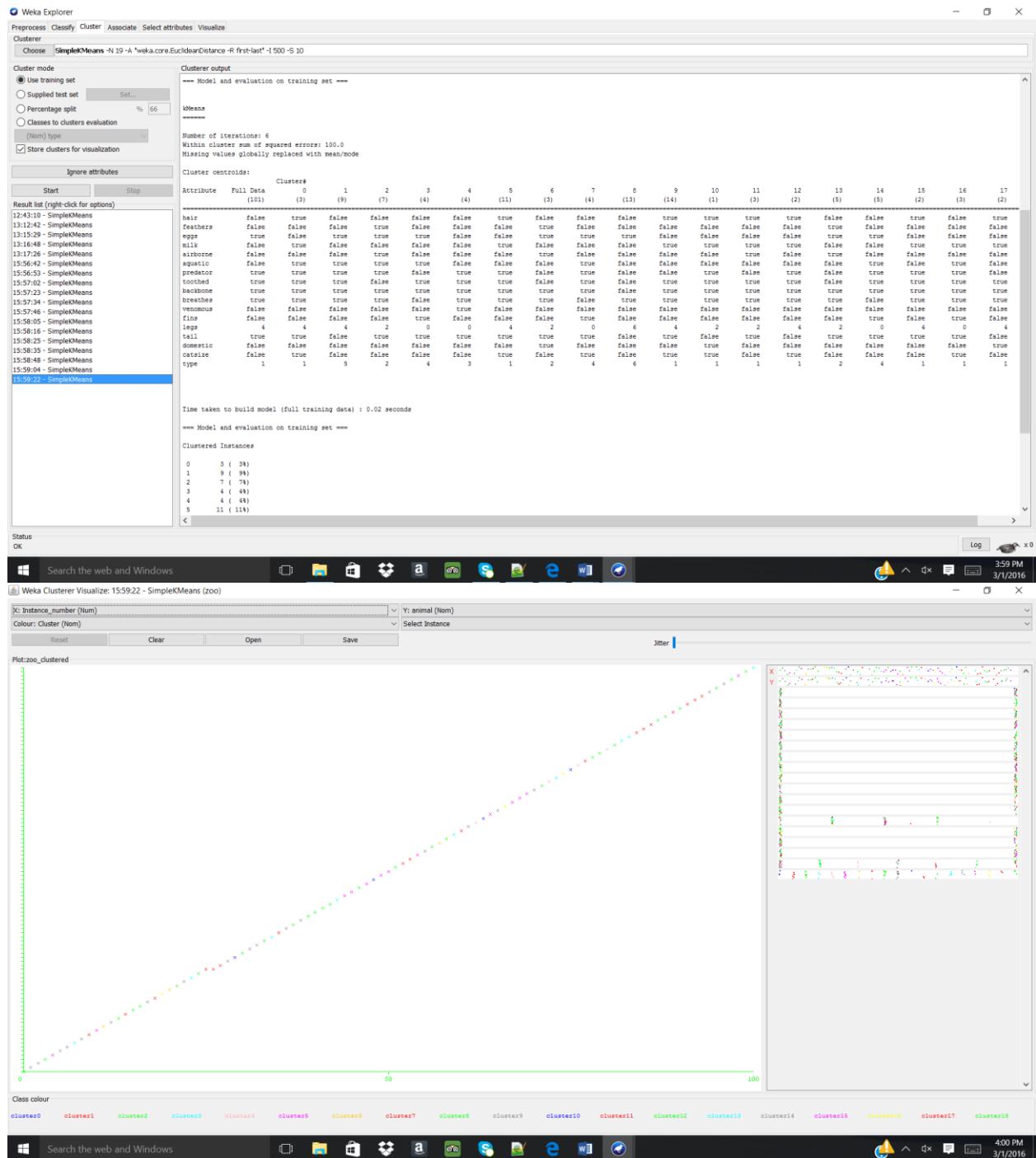


When working with k-means, choosing the best number of clusters, k, can be completed by varying k through multiple iterations and determining which value of k results in the lowest error.

When:

- | | |
|--------|-------------------------------|
| k = 1 | sum of squared errors = 134.0 |
| k = 19 | sum of squared errors = 100.0 |

Orysya Stus



K= 19 was chosen for the k-clusters since in the data set there are 19 divisions of the attribute “animal”.

2. Use the following learning schemes to analyze the Automobile (<http://archive.ics.uci.edu/ml/datasets/Automobile>) data set.

Linear regression

- weka.classifiers.LinearRegression

M5'

- weka.classifiers.M5'

Regression Tree

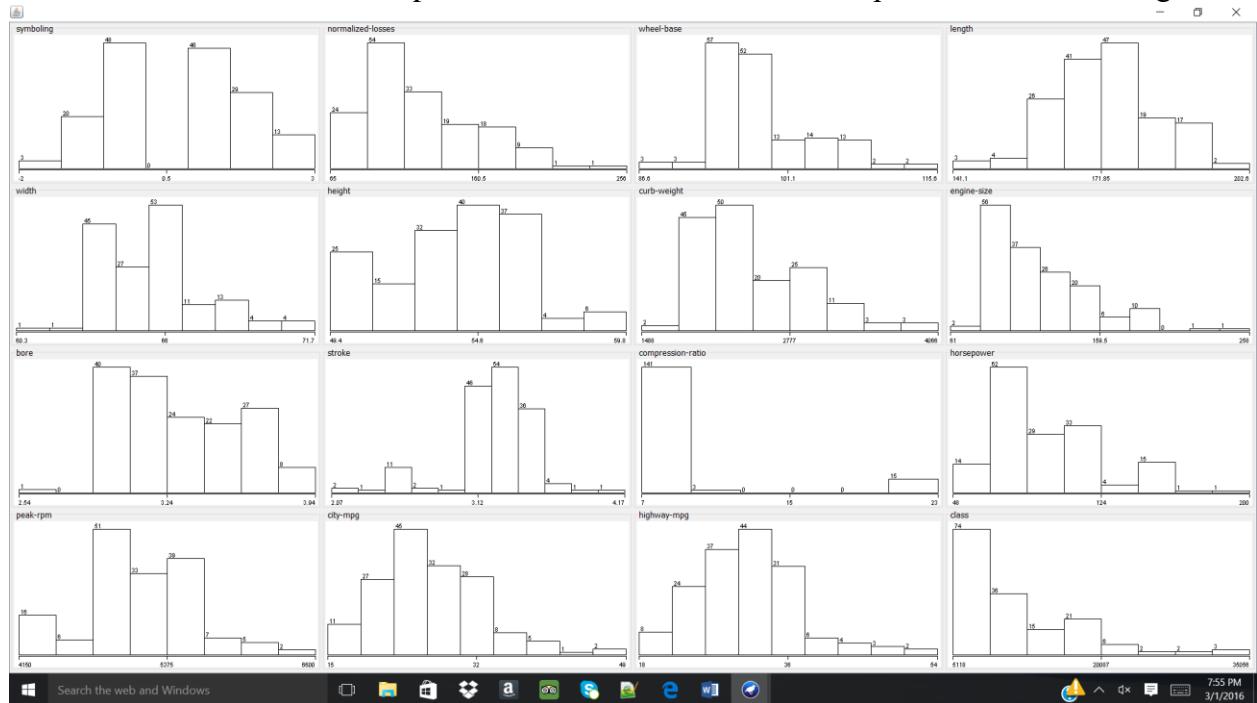
- weka.classifiers.M5'

K-means clustering - weka.clusterers.SimpleKMeans

A) How many leaves did the Model tree produce? Regression Tree? What happens if you change the pruning factor?

How many clusters did you choose for the K-means method? Was that a good choice? Did you try a different value for k ?

A. The file autoPrice.arff was opened in Weka and “Visualize All” produced the following:



Note: all attribute values are numeric and the class attribute “class” correlates to the price of the vehicle.

B. Classify/functions/LinearRegression was selected. The “test option”, “Cross-validation Fold 10” was selected. Further default values were maintained, and the following was generated:

Orysya Stus

The screenshot shows the Weka Explorer interface with the "Classify" tab selected. Under "Test options", "Cross-validation" is chosen with "Folds 10". The "Classifier output" pane displays the generated linear regression equation:

```

Linear Regression Model

class =
  178.5892 * wheel-base +
  -59.5456 * length +
  786.4855 * width +
  5.1527 * curb-weight +
  52.0096 * engine-size +
  -204.6104 * horsepower +
  -1529.0394 * stroke +
  110.0562 * compression-ratio +
  27.1139 * peak-rpm +
  0.4612 * peak-rpm +
  -55823.3154

```

Below the equation, it says "Time taken to build model: 0.02 seconds". The "Correlation coefficient" is listed as 0.8733.

The generated equation assigns different weights to the attribute values in combination in order to determine the value of the class, this equation has a correlation coefficient of 0.8733.

C. Classify/rules/M5Rules was selected with “Cross-validation Fold 10” as the test option. The following 5 rules were generated:

The screenshot shows the Weka Explorer interface with the "Classify" tab selected. Under "Test options", "Cross-validation" is chosen with "Folds 10". The "Classifier output" pane displays five generated rules:

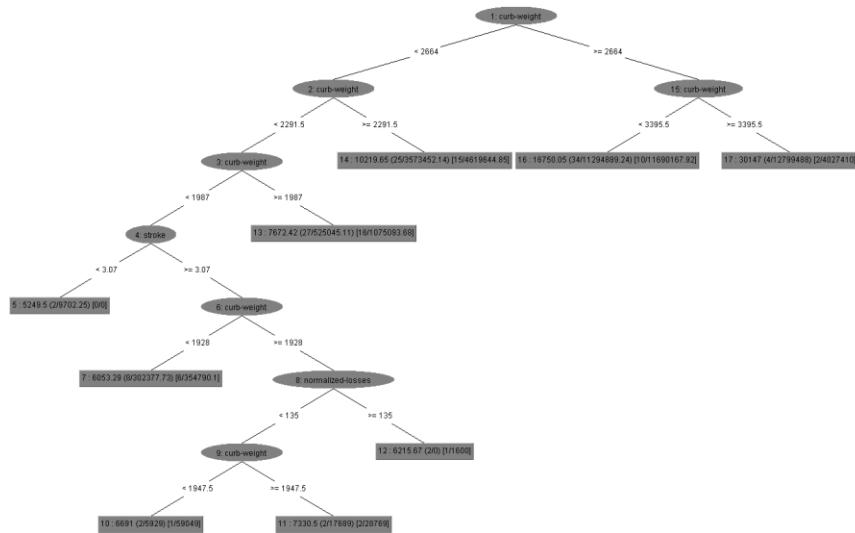
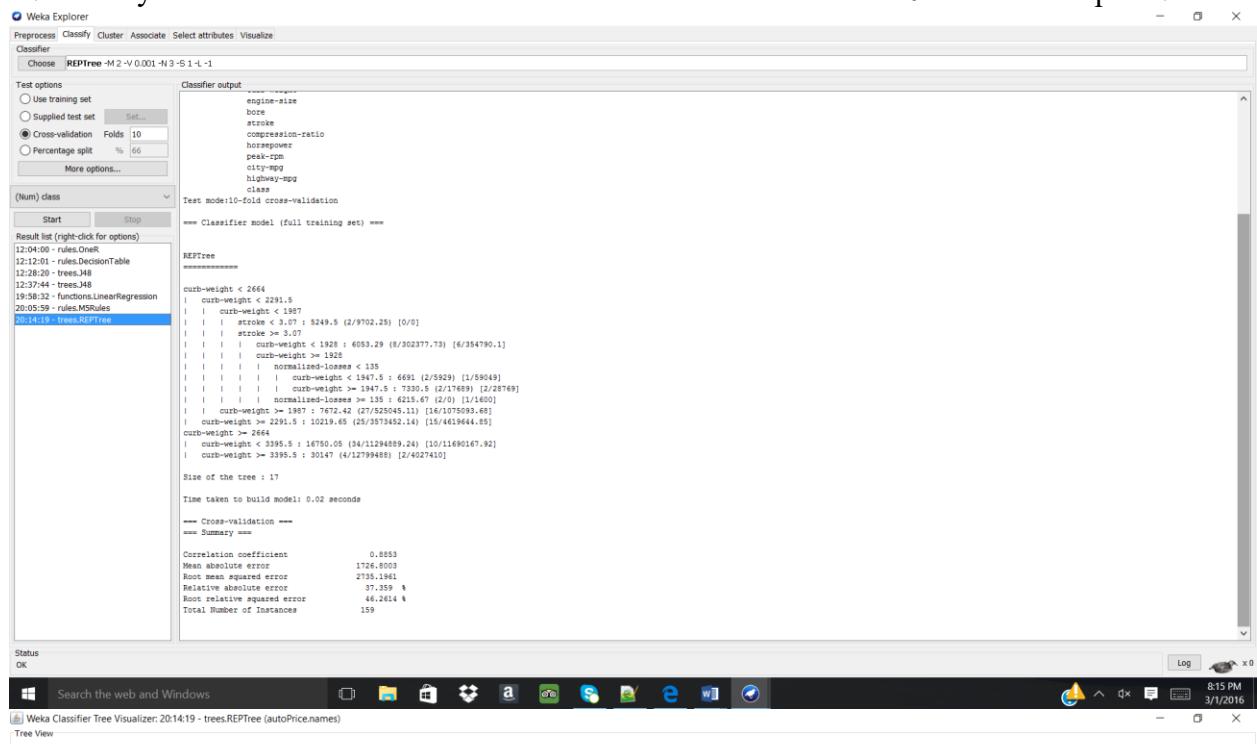
- Rule 1:** IF curb-weight <= 2664 AND curb-weight <= 2291.5 THEN class = 8.4769 * normalized-losses + 8.5012 * wheel-base + 12.9746 * length + 103.5554 * width + 17.2865 * height + 4.6161 * curb-weight + 24.4627 * engine-size + 24.4627 * horsepower + 61.9356 * city-mpg + 71.1605 * highway-mpg + 1552.4175 [69/10.764%]
- Rule 2:** IF curb-weight > 2664 AND width <= 68.85 AND horsepower <= 158 THEN class = -119.7339 * symboling + 27.3115 * normalized-losses + 42.4906 * length + 499.305 * width + 94.3851 * height + 112.9927 * curb-weight + 172.9927 * engine-size + 35.6559 * horsepower + 285.8787 * city-mpg + 21301.4771 [20/42.076%]
- Rule 3:** IF curb-weight > 2733 THEN class = 660.9393 * symboling + 26.7564 * normalized-losses + 802.7044 * width + 12.4721 * curb-weight + 70140.1854 [22/24.502%]
- Rule 4:** IF length <= 176.4 THEN class = -169.6013 * symboling - 8.1042 * normalized-losses + 43.0443 * engine-size + 35.175 * horsepower - 3305.8355 [31/46.357%]
- Rule 5:** class = 91.0447 * normalized-losses - 1096.5785 [9/34.711%]

At the bottom, it says "Time taken to build model: 0.12 seconds". The "Correlation coefficient" is listed as 0.8335.

These rules are generated using “perfect” rules that use combinations of attribute values in order to find the value of “class”. The correlation coefficient is slightly lower compared to that of the linear regression equation, at 0.8335.

Orysya Stus

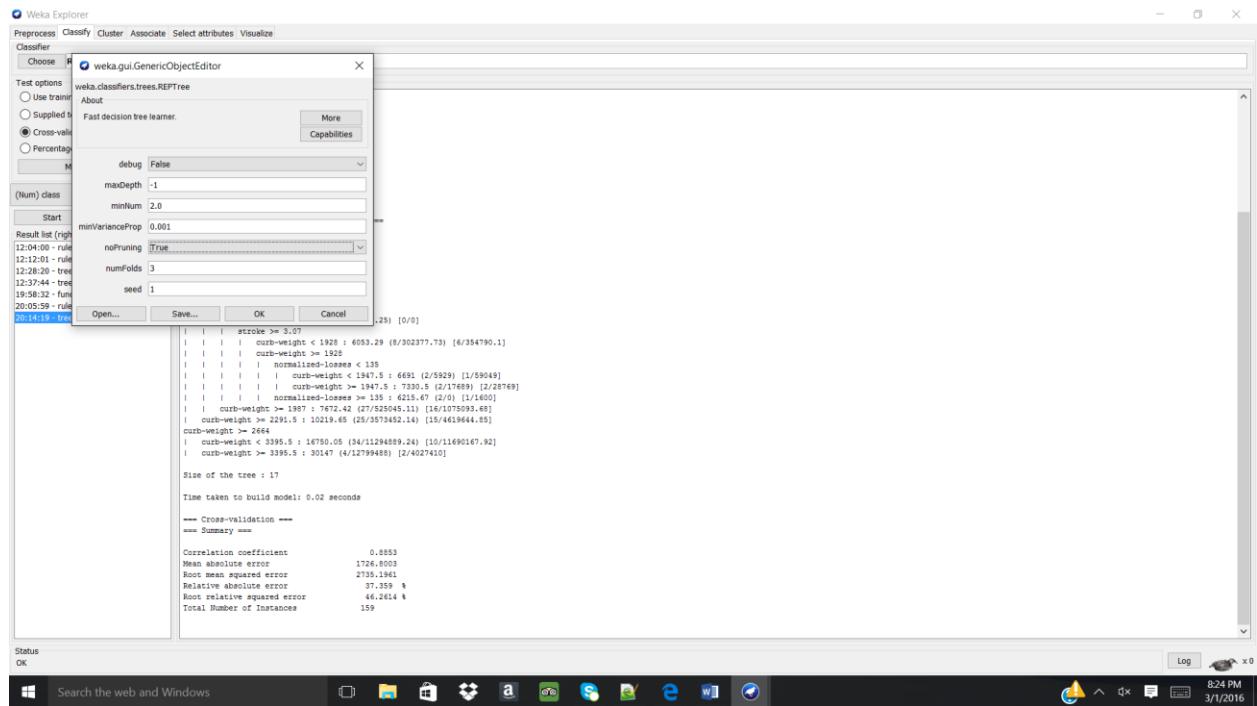
D. Classify/trees/REPTree was selected with “Cross-validation Fold 10” as the test option.



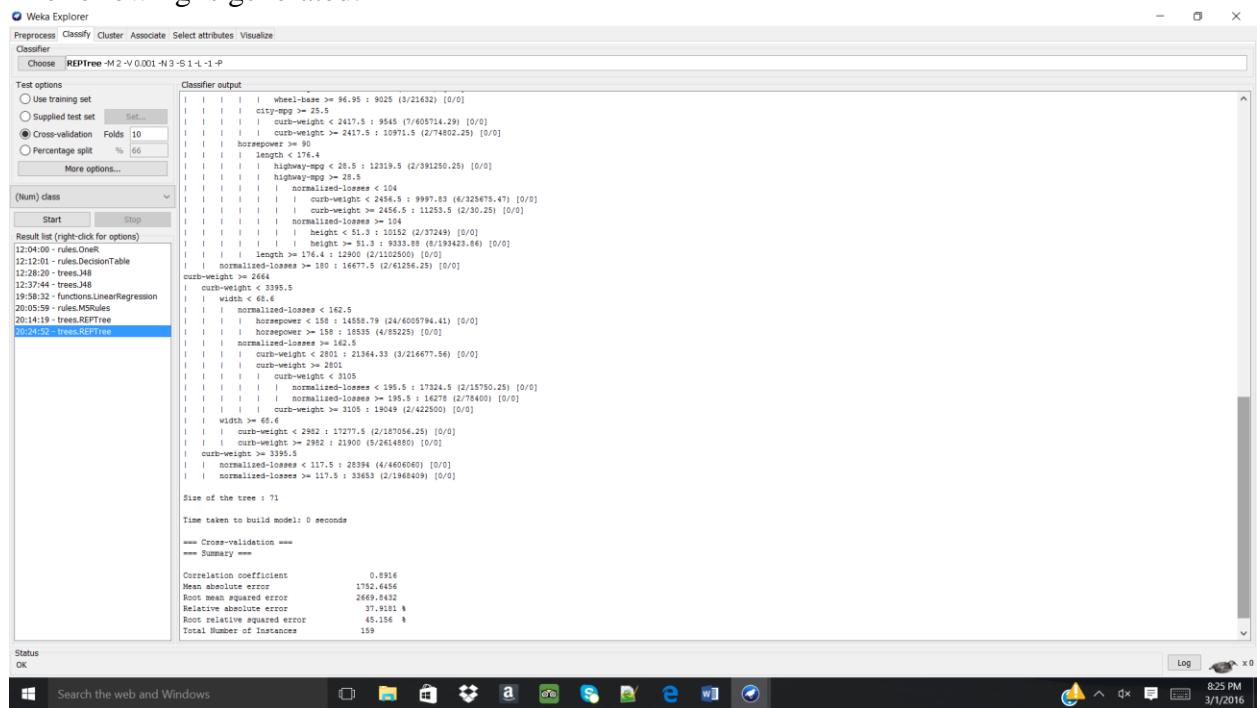
The tree generated has a higher correlation coefficient compared to the above learning schemes, at 0.8853. The model tree has 9 nodes.

When the pruning factor is changed where “nopruning” is changed from false to true:

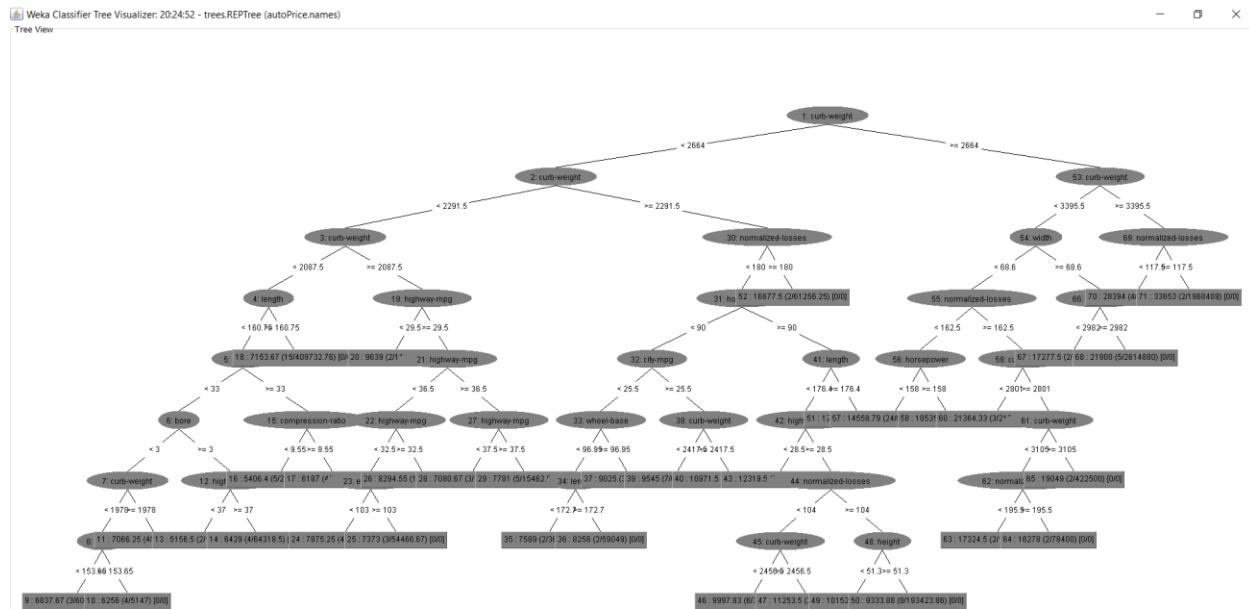
Orysya Stus



The following is generated:



Orysya Stus



The correlation coefficient falls down to 0.8916 and the number of leaves greatly increases.

E. Cluster/clusterers/SimplekMeans was selected with the entire training set being used to generate the model. K = 10 was used for the number of clusters since this provided the lowest error in the learning scheme.

The Weka Explorer interface shows the following details for the SimplekMeans clustering process:

- Cluster mode:** Use training set
- Number of iterations:** 6
- Within cluster sum of squared errors:** 30.8934859519556447
- Missing values globally replaced with mean/mode**
- Cluster centroids:**

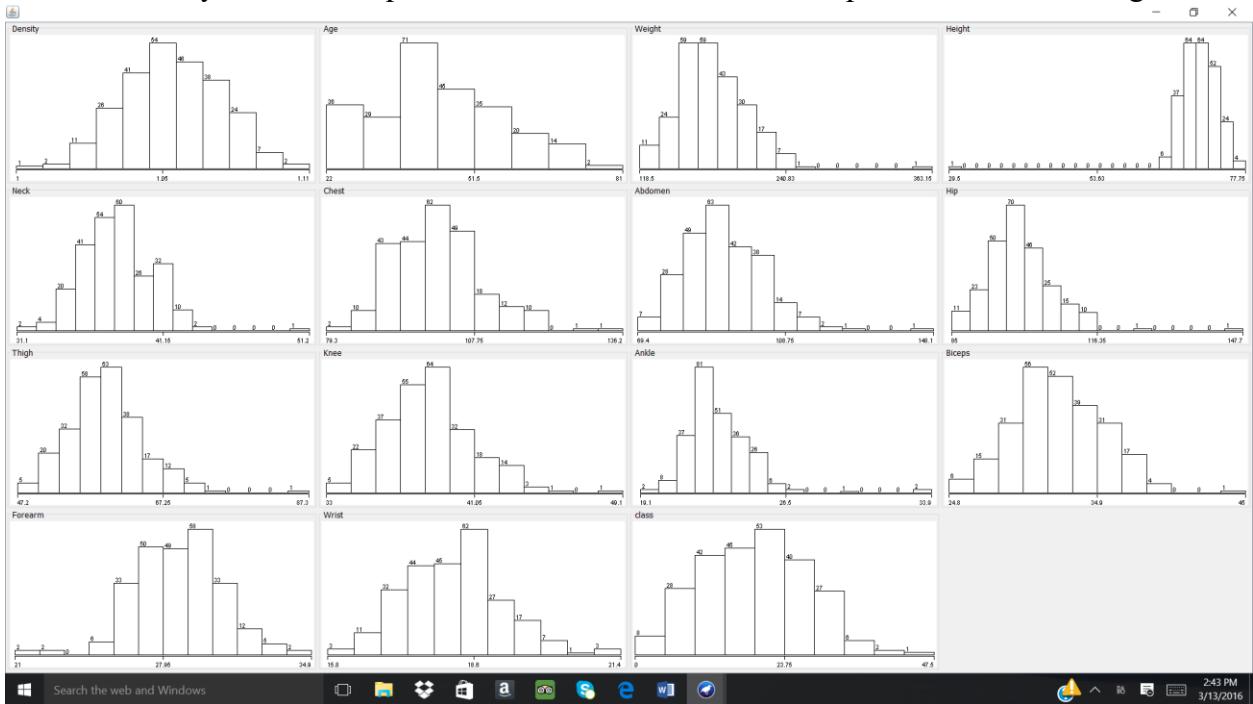
Attribute	Full Data	Cluster# 0	1	2	3	4	5	6	7	8	9
	(159)	(18)	(16)	(23)	(12)	(19)	(33)	(10)	(9)	(8)	(11)
symboling	0.3556	-0.2776	1.4375	-0.5217	0.5	1.2632	0.4465	2.1	-0.4444	2.75	1.2727
normalized-losses	121.1321	88.3889	129.5	121.6957	92.25	128.4737	102.8182	147.4	121.6667	185.75	163.8182
wheel-base	90.2442	99.1779	97.425	105.1522	96.175	95.4515	99.62	109.7889	96.825	94.1273	102.1273
length	171.0	173.0	174.0	171.0	169.0	164.0	154.0	162.0	160.0	170.0	164.0
width	65.4675	65.3722	65.3725	67.6826	64.99	65.7516	64.3394	66.12	69.5559	68.025	63.8909
height	51.8994	55.9333	51.9188	55.8334	53.75	51.4895	54.3333	55.38	56.0889	50.5375	51.7364
curb-weight	2461.1384	2421.2228	2487.3125	3006.8261	2316.25	1894.7695	2031.4545	2642.4	3501.5555	3076.875	2158.9091
engine-size	113.2264	113.3889	127.25	146.2174	107.0	89.3158	98.5158	118.7	176.7775	178.25	98.1818
stroke	3.856	3.856	3.856	3.856	3.856	3.856	3.856	3.856	3.856	3.856	3.856
stroke	3.2364	3.3739	3.4813	3.1522	2.4167	3.2489	3.2482	2.982	3.4322	3.3273	3.2309
compression-ratio	10.4511	11.7833	8.6958	8.7224	8.8167	10.1005	9.9879	9.081	20.0111	8.65	8.7565
horsepower	99.8685	79.3889	104.8125	128.5685	86.28	66.7895	72.2121	117.9	117.6667	160.428	87.7273
peak-rpm	5111.1865	4662.125	5727.125	5727.125	5727.125	5727.125	5727.125	5404.4375	5727.125	5520.5555	5727.125
city-mpg	28.5556	24.1213	18.5522	26.1625	28.5557	32.3538	32.3538	21	23.6667	18.5556	27.1518
highway-mpg	32.0218	34.0556	30.5425	24.9565	30.79	41.0526	36.4364	27.4	27.2222	23.75	32.5453
class	11445.7796	9271.1113	10213.625	12783.138	8541.28	6317.7895	7513.0604	15699.5	24038.4444	19736.375	8296.8182
- Time taken to build model (full training data):** 0.02 seconds
- Model and evaluation on training set**
- Clustered Instances:**

0	18 (11)
0	18 (11)
1	14 (10)
2	23 (14)
3	12 (8)
4	19 (12)
5	33 (23)
6	10 (4)
7	9 (4)
8	8 (5)
9	11 (7)

Taken together, the highest correlation coefficient belonged to the pruned model tree and allowed for prediction of the dataset the best. The attribute “curb_weight” was used provided the most information gain to this dataset.

B) Now perform the same analysis on the bodyfat.arff data set.

A. The file bodyfat.arff was opened in Weka and “Visualize All” produced the following:



Note: all attribute values are numeric and the class attribute “class” correlates to the numeric value of the individual’s BMI.

B. Classify/functions/LinearRegression was selected. The “test option”, “Cross-validation Fold 10” was selected. Further default values were maintained, and the following was generated:

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose: LinearRegression -S 0 -R 1.0E-8
Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
More options...
(Num) class
Start Stop
Result list (right-click for options)
14:57:06 - Functions.LinearRegression
Classifier output
Run information
Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation: bodyfat.names
Instances: 252
Attributes: 15
Attributes:
Density
Age
Height
Hip
Neck
Chest
Abdomen
Hip
Thigh
Knee
Ankle
Biceps
Forearm
Wrist
class
Test mode:10-fold cross-validation
Classifier model (full training set)

Linear Regression Model
Class =
-410.2167 * Density +
0.0121 * Age +
0.0001 * Chest +
0.0314 * Abdomen +
446.1513

Time taken to build model: 0.03 seconds
Cross-validation
Summary

Correlation coefficient 0.9962
Mean absolute error 0.519
Root mean squared error 1.384
Relative absolute error 7.5247 %
Root relative squared error 16.4053 %
Total Number of Instances 252

```

Orysya Stus

The generated equation assigns different weights to the attribute values in combination in order to determine the value of the class, this equation has a correlation coefficient of 0.9862.
class = -410.2167 * Density + 0.0124 * Age + 0.0253 * Chest + 0.0314 * Abdomen + 446.1513

C. Classify/rules/M5Rules was selected with “Cross-validation Fold 10” as the test option. The following 6 rules were generated:

The image displays two windows of the Weka Explorer interface, both titled "Weka Explorer". Both windows show the "Classify" tab selected. In the top bar, "Classifier" is chosen, and "Choose: M5Rules - M 4.0" is selected.

Left Window (Classifier output):

- Test options:** Cross-validation, Folds 10.
- Classifier output:** Shows the generated rules:
 - Rule: 1
IF Density <= 1.056
THEN
class = -455.1528 * Density + 0.0055 * Chest + 498.7782 [130/2.3089]
 - Rule: 2
IF Density > 1.066
Density <= 1.078
THEN
class = -374.8046 * Density - 0.0107 * Weight + 411.7526 [44/31.5644]
 - Rule: 3
IF Density <= 1.072
THEN
class = -419.5503 * Density + 0.0082 * Weight + 460.2068 [43/0.6278]
 - Rule: 4
IF Density <= 1.084
THEN
class = -296.0056 * Density + 0.0013 * Weight + 322.668 [15/1.1058]
 - Rule: 5
IF Density <= 1.092
THEN
class = -199.8723 * Density + 0.0054 * Weight + 214.0779 [14/0.7078]
 - Rule: 6
class = 0.1673 * Weight - 20.6973 [4/14.6758]
- Status:** OK

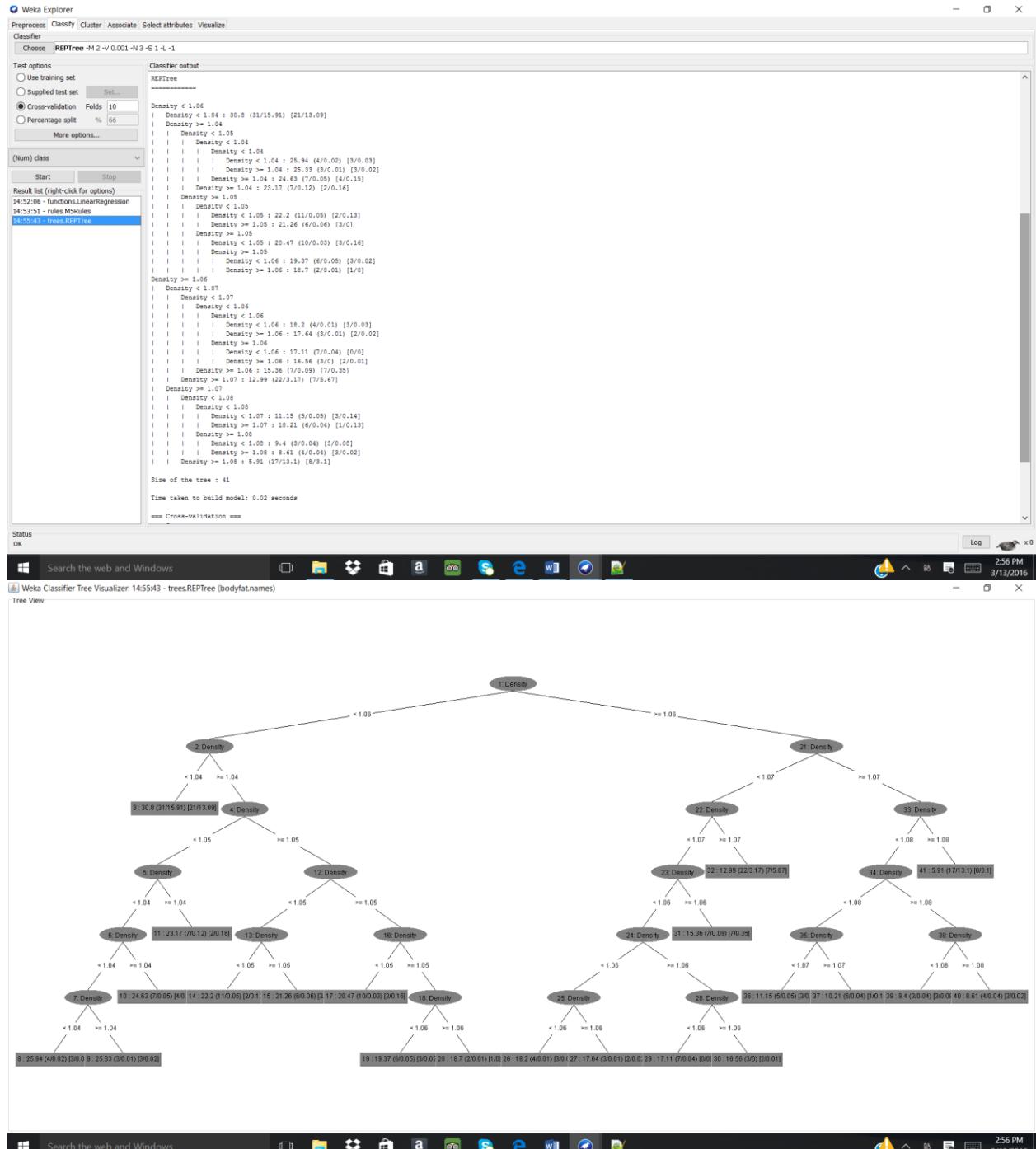
Right Window (Classifier output):

- Test options:** Cross-validation, Folds 10.
- Classifier output:** Shows the generated rules and summary statistics:
 - Time taken to build model: 0.25 seconds
 - Correlation coefficient: 0.986
 - Mean absolute error: 0.4172
 - Root mean squared error: 1.3973
 - Relative absolute error: 6.0481 %
 - Root relative squared error: 16.4436 %
 - Total Number of Instances: 252
- Status:** OK

These rules are generated using “perfect” rules that use combinations of attribute values in order to find the value of “class”. The correlation coefficient is slightly lower compared to that of the linear regression equation, at 0.986.

D. Classify/trees/REPTree was selected with “Cross-validation Fold 10” as the test option.

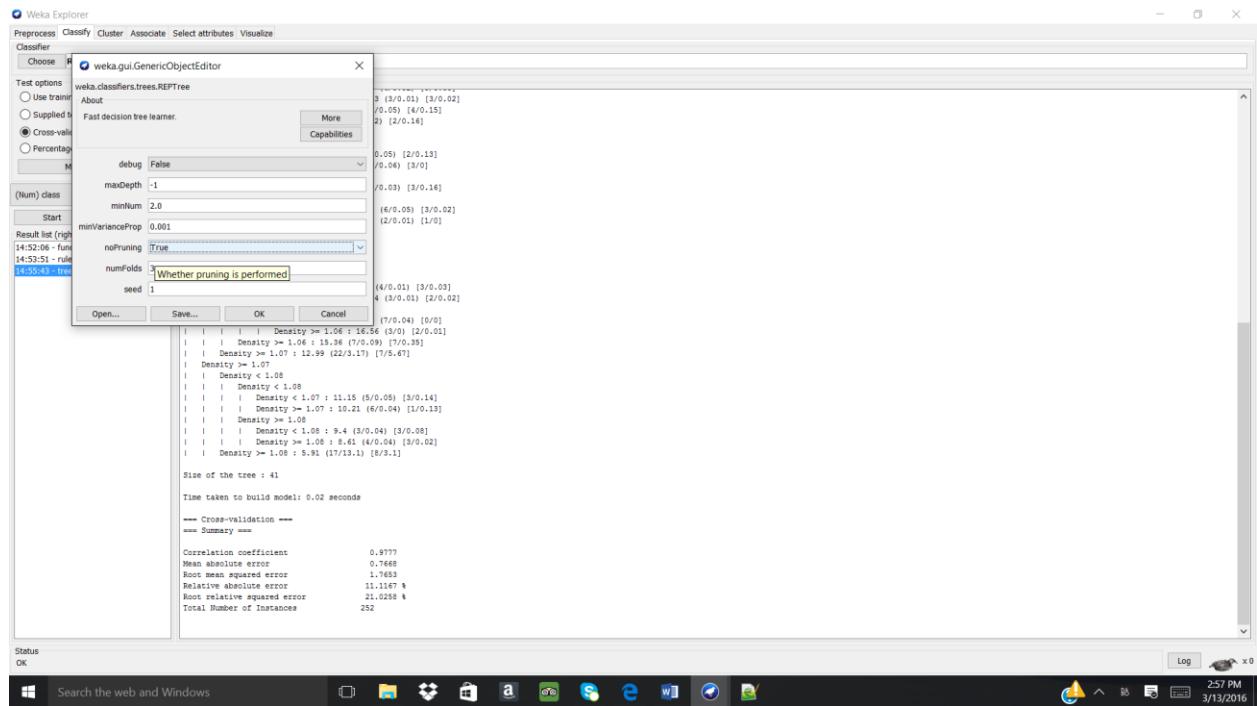
Orysya Stus



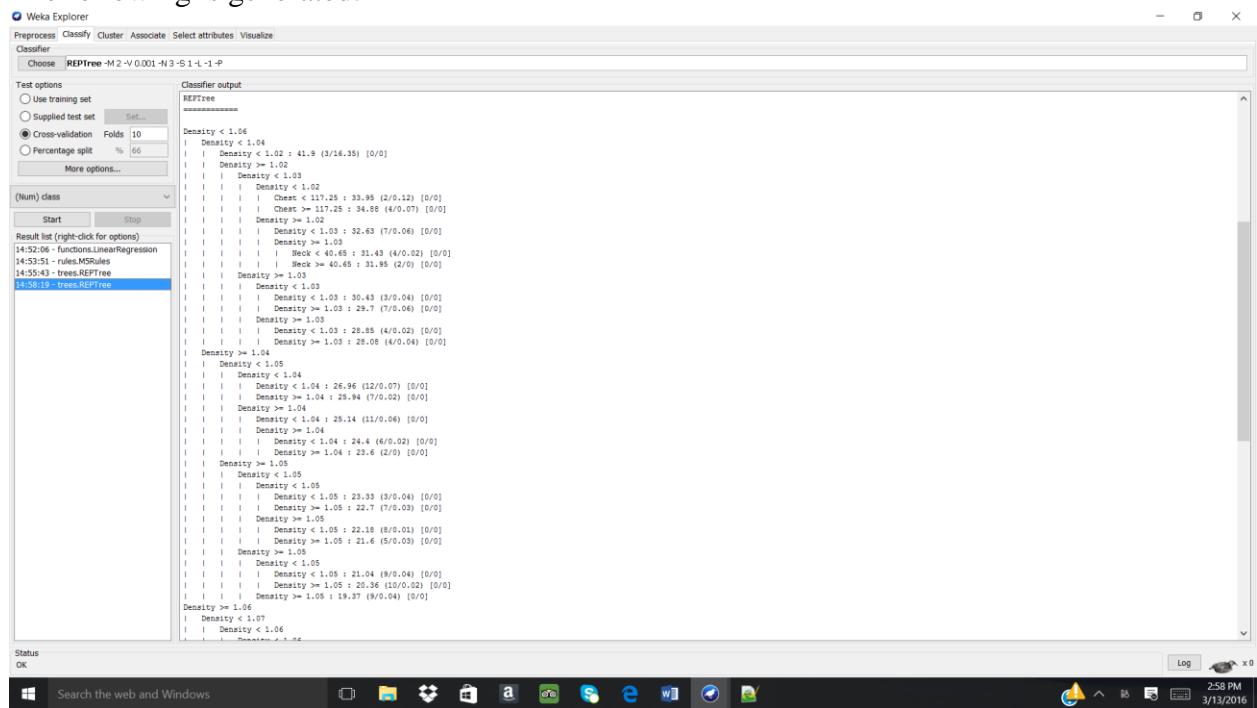
The tree generated has a lower correlation coefficient compared to the above learning schemes, at 0.9777. The model tree has many nodes.

When the pruning factor is changed where “nopruning” is changed from false to true:

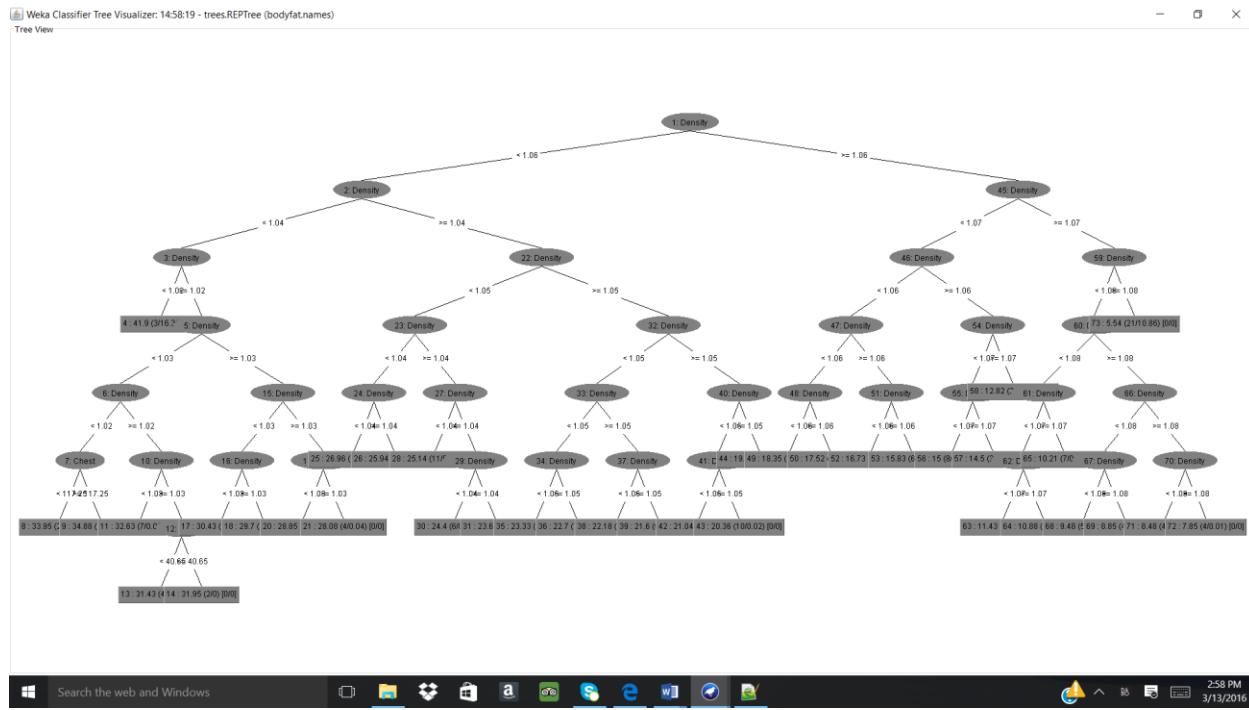
Orysya Stus



The following is generated:



Orysya Stus



The correlation coefficient increases to 0.9826 and the number of leaves greatly increases.

E. Cluster/clusterers/SimplekMeans was selected with the entire training set being used to generate the model. K = 10 was used for the number of clusters since this provided the lowest error in the learning scheme.

The screenshot shows the Weka Explorer interface with the following details:

- Preprocess**: Classify, Cluster, Associate, Select attributes, Visualize.
- Clusterer**: Choose **SimplekMeans** - N 10 - A "weka.core.EuclideanDistance-R first-last"-I 500 - S 10
- Cluster mode**: Use training set
- Supplied test set**: Set... (25)
- Percentage split**: % 66
- Classes to clusters evaluation**: (None) class
- Store clusters for visualization**:
- Clusterer output**:
 - Means**
 - Number of iterations: 28
 - Within cluster sum of squared errors: 27.694207311226837
 - Missing values globally replaced with mean mode
- Attribute centroids:**

	Cluster#	0	1	2	3	4	5	6	7	8	9
Density	1.0556	1.0725	1.0664	1.0297	1.0268	1.0551	1.0481	1.0499	1.0768	1.0331	1.0515
Age	49.8489	29.8571	56.9459	44.6975	42	49.5	33.2857	43	41.0952	63.0831	46.2333
Weight	176.9841	150.3382	164.2179	197.1137	255.4565	179.4	213.5101	166.3861	143.3981	199.3641	193.325
Height	70.4488	71.5554	69.7939	68.8444	71.5555	72.5555	69.4477	68.5552	68.5552	70.5555	70.5555
Neck	37.9921	37.8755	37.427	38.5938	43.1	37.6	40.4524	36.9295	35.1667	40.0417	39.2933
Chest	100.0242	98.1643	97.5649	108.4	120.75	98.7	108.1286	97.5591	91.0524	109.5917	104.39
Abdomen	92.556	86.7393	87.9469	104.7063	118.621	89.2	101.8146	89.9559	79.7107	104.2	94.4167
Hip	98.9481	99.9058	96.454	101.5037	118.525	98.5	106.1146	91.9559	91.9559	104.2	103.3333
Thigh	59.406	56.6794	63.5125	59.375	54.1	64.5419	58.2614	51.1024	60.9917	61.0667	
Knee	38.9095	39.4679	38.1541	39.6063	43.1625	38.25	41.8429	37.4102	35.6476	40.4093	39.6133
Ankle	23.1024	23.3143	22.0108	23.2	25.975	33.8	24.7952	22.2114	21.6643	23.5042	23.5467
Biceps	32.7794	32.4714	30.9479	33.9938	37.5125	32.45	36.0238	31.0205	28.8643	34.1205	33.8767
Pectoras	24.9569	24.8569	25.2062	29.2965	27.1	23.1	21.0205	20.9559	29.312	24.9569	24.9569
Wrist	18.2289	18.3036	18.5027	17.9188	19.9125	18.3	19.0429	17.9455	17.1119	19.1467	18.7867
Class	19.1505	11.5429	14.3459	32.6438	31.95	19.25	22.0619	21.5045	9.6405	29.1417	21.3933
- Time taken to build model (full training data) : 0.09 seconds**
- Model and evaluation on training set**
- Clustered Instances**

0	26	(15%)
1	37	(18%)
2	16	(8%)
3	8	(4%)
4	2	(1%)
5	21	(11%)
6	44	(17%)
7	42	(17%)
8	24	(10%)
9	30	(12%)
- Status**: OK
- Log**: 3:00 PM
- Search the web and Windows**

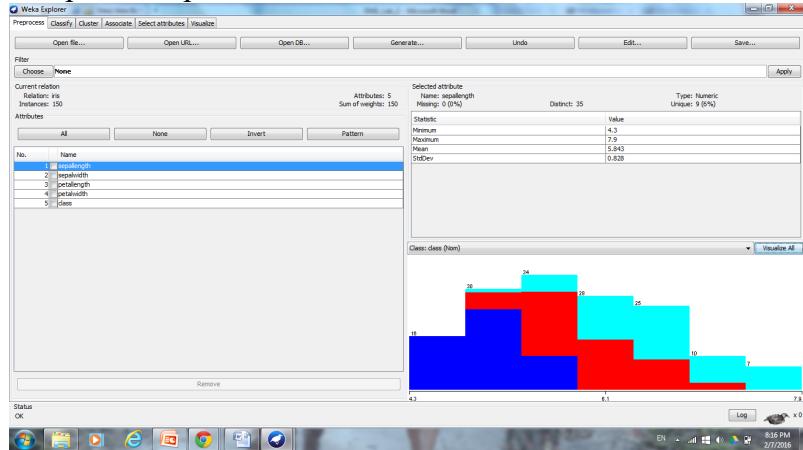
Taken together for this dataset, the highest correlation coefficient belonged to linear regression model and the M5 rules, showing that these two models would be the best to use in order to

Orysya Stus

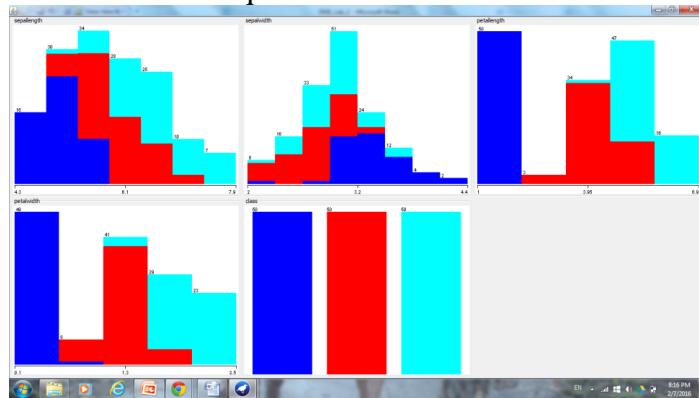
predict other BMIs. The attribute “Density” was provided the most information gain to this dataset in classifying the “class” of the BMI.

3. Use a k-means clustering technique to analyze the iris data set. What did you set the k value to be? Try several different values. What was the random seed value? Experiment with different random seed values. How did changing of these values influence the produced models?

A. The iris.arff was uploaded into Weka, there was no need to further filter the data during the “Preprocess” phase. Note: 5 attributes exist in this data set.

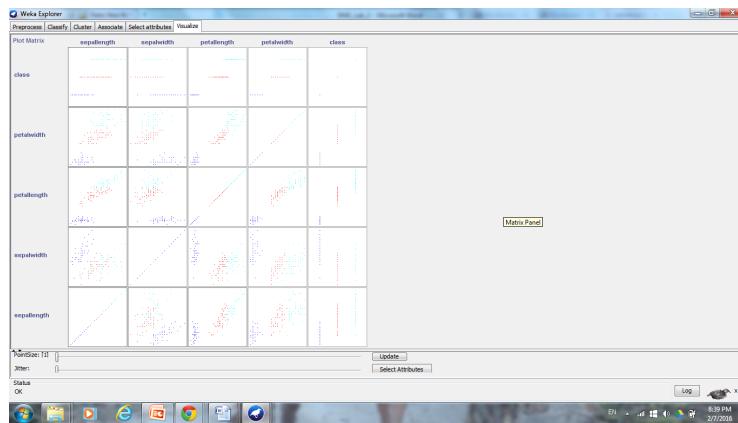


B. “Visualize all” was used to visual the distributions for each class. Where, attributes “sepalwidth” and “sepallength” follow positive skewed Gaussian distributions, the attributes “petallength” and “petalwidth” follow no concrete distribution, and the “class” attribute distributions are equal.

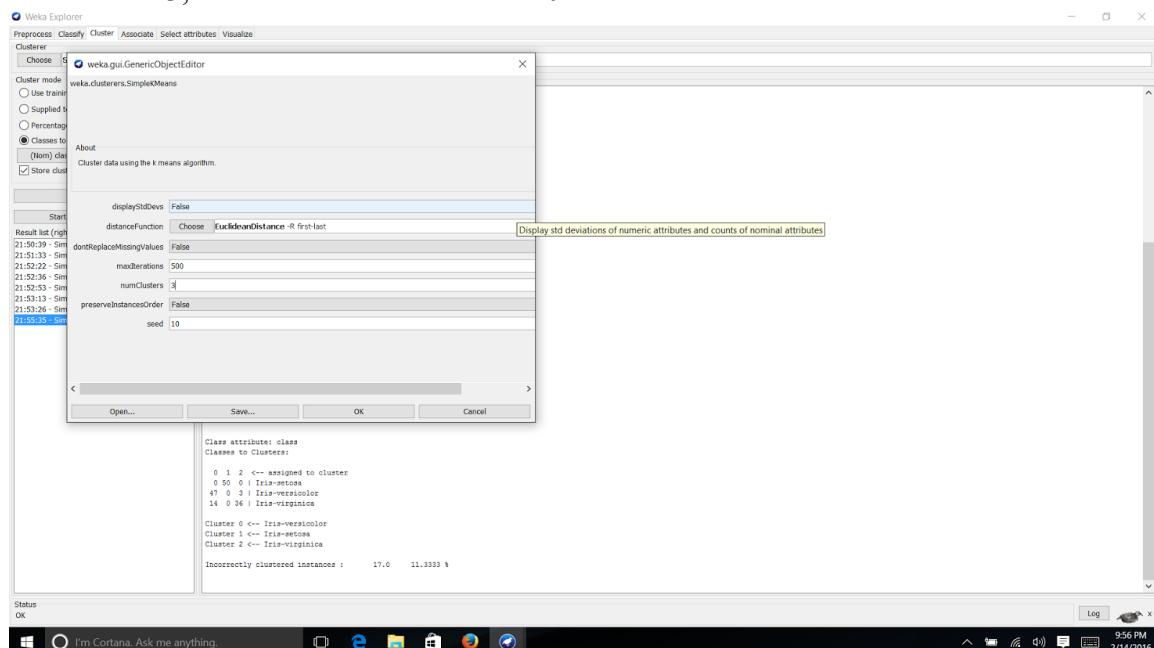


C. Utilizing the “Visualize” tab, clusters for attribute values can be seen with the most apparent clusters seen for petalwidth vs. sepalwidth/petallength and petallength vs. sepallength/sepalwidth/petalwidth. These attributes will be utilized in order to determine the class (Iris-setosa, Iris-versicolor, or Iris-virginica).

Orysya Stus

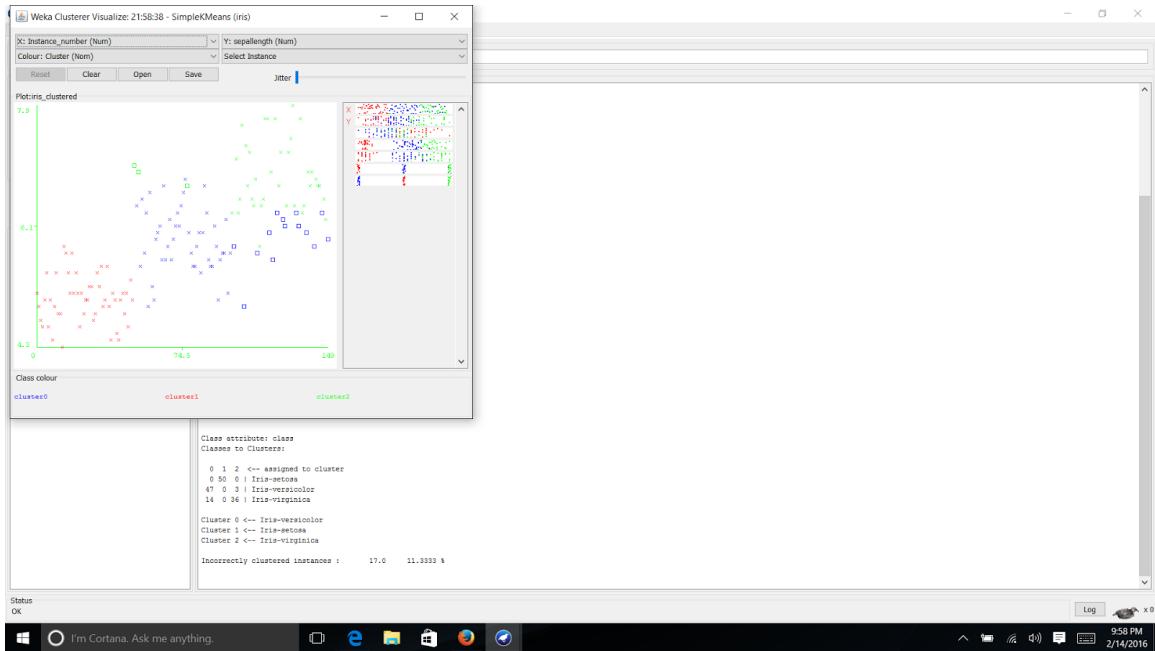


D. To apply the k-means clustering technique, under the “Cluster” tab Clusterers/SimpleKMeans is selected. At the pop-up window, “numClusters” was changed to 3 since the number of class attributes is 3, “seed” was maintained at 10.



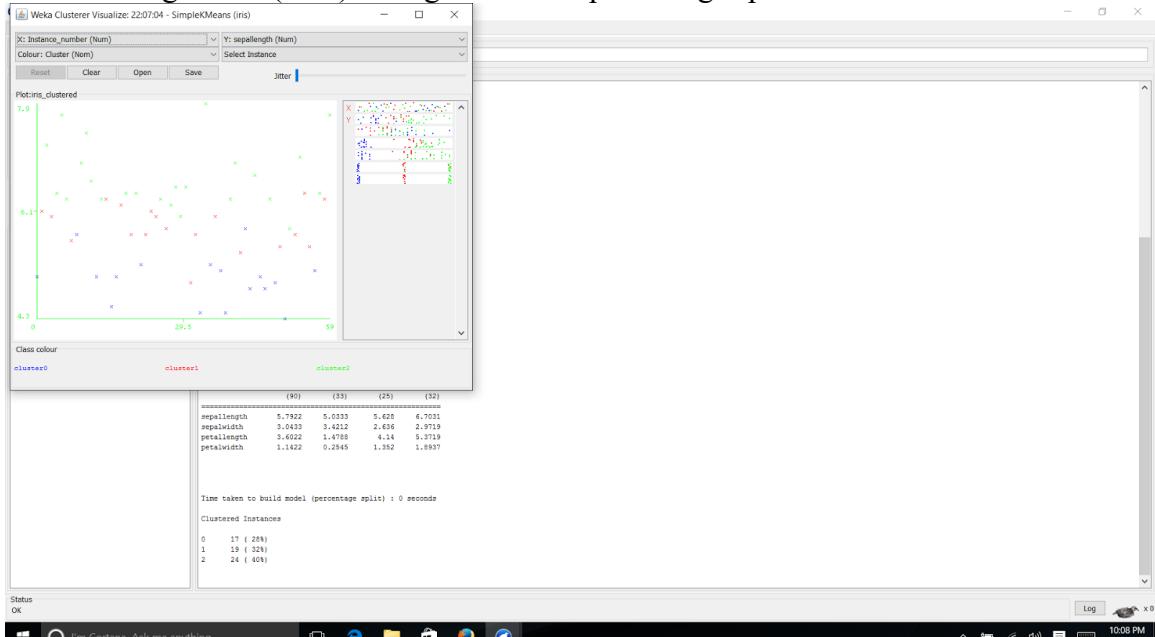
Note: for the first means of analysis, the class attribute was not ignored and the “classes to cluster evaluation” was set to “class (nom)”.

Orysya Stus



Three clusters were created with the “Clustered Instances” containing 61 (41%) in cluster 0, 50 (33%) in cluster 1, and 39 (26%) in cluster 2. Note: in the class distribution in Part B, Iris-setosa, Iris-versicolor, and Iris-virginica had equal distributions.

But as clustering is unsupervised, next the class attribute was ignored and only the four attributes - sepallength, sepalwidth, petallength, and petalwidth were considering in clustering to compare the clustering. Class (nom) was ignored and “percentage split” at “60%” was used.

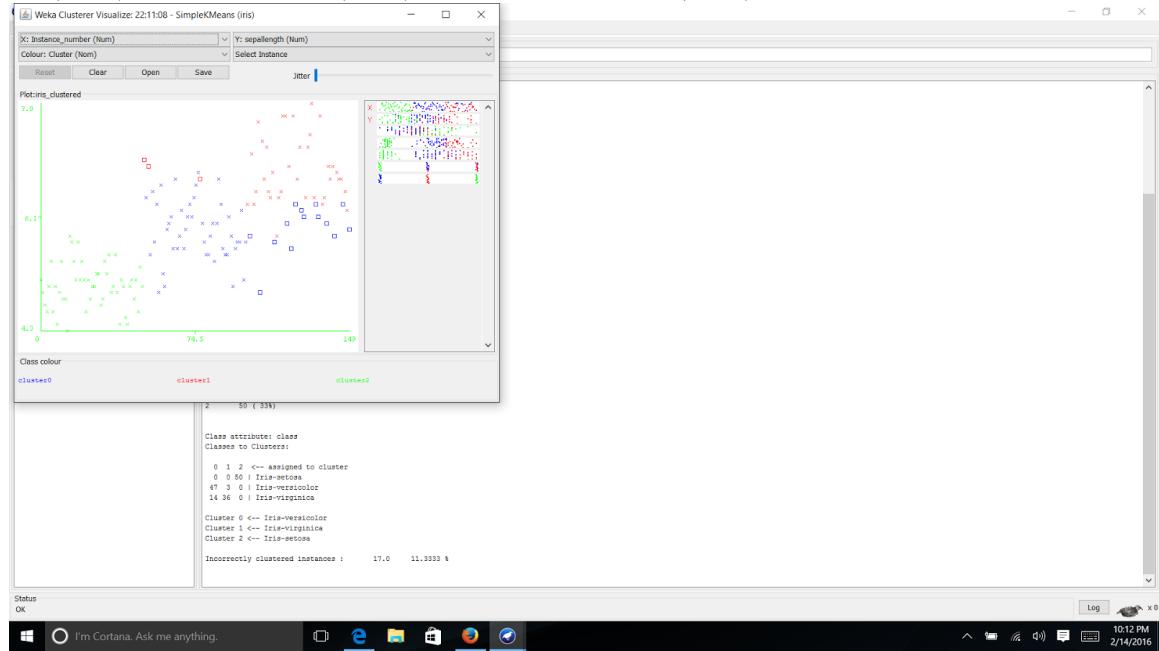


Making the above adjustments, changed the clustering vastly in comparison with the cluster centroids being located in different places then before.

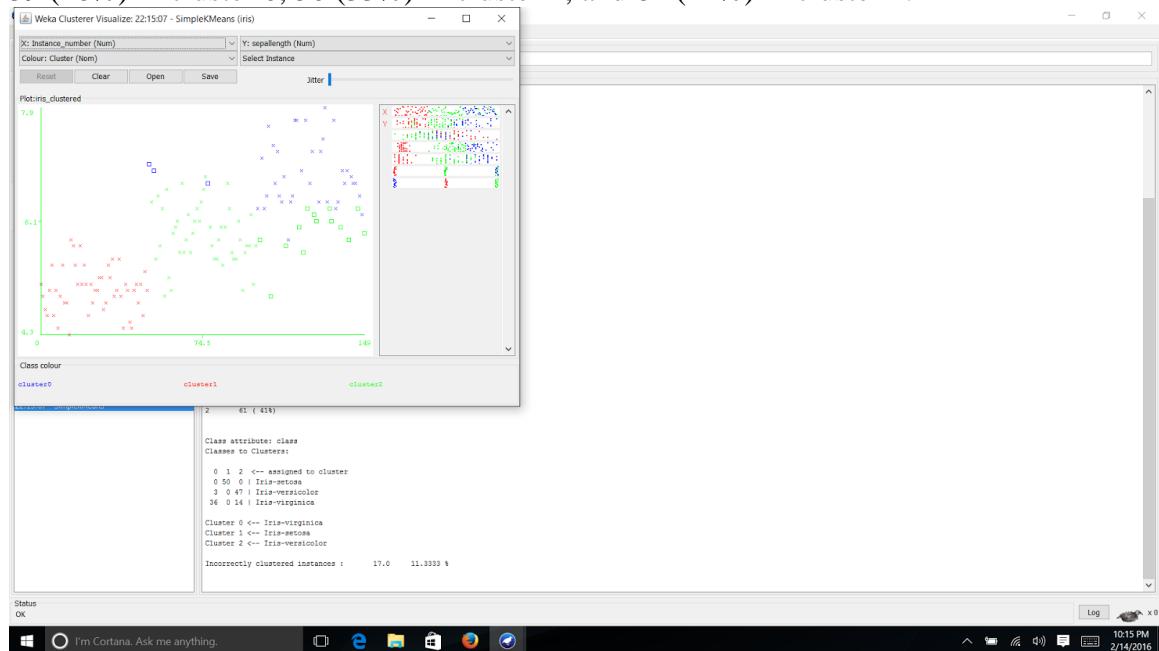
Orysya Stus

Furthermore, parameter variations for the first cluster produced, where “class (nom)” is not ignored were made.

When increasing the “seed” from 10 to 50, the clusters looked similar to that shown above with 61 (41%) in cluster 0, 39 (26%) in cluster 1, and 50 (33%) in cluster 2.



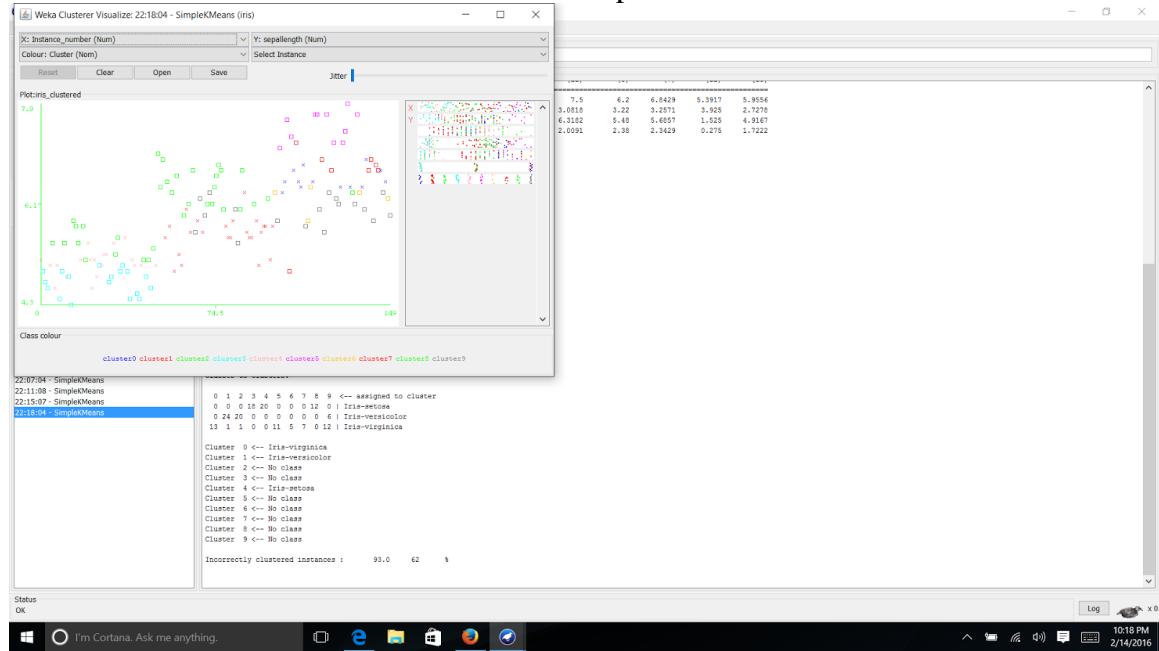
When decreasing the “seed” from 10 to 5, the clusters looked similar to that shown above with 39 (26%) in cluster 0, 50 (33%) in cluster 1, and 61 (41%) in cluster 2.



Note: It is interesting that the ratio 39:50:61 is conserved when changing the “seed” number. The seed number (any integer) is the randomization for your initial K points. K represents the number of clusters. Because K-means is sensitive to initial points, you will have to try experimentation

on the stability of your clusters with different seeds. However, K is user defined which could also be guided by domain knowledge and other practical factors.

Changing the number of clusters from 3 to 10, does not create nice clean clusters for all. Nicer clusters can be created after recursive measures of visualizing the clustering and determining how much variation between instances are acceptable.

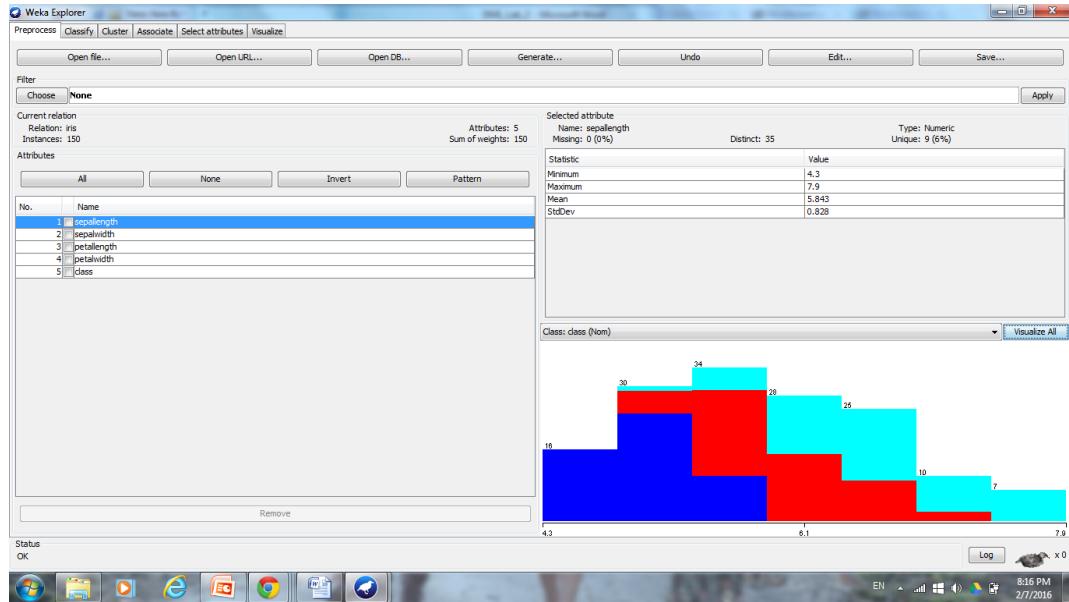


4. Produce a hierarchical clustering (COBWEB) model for iris data. How many clusters did it produce? Why? Does it make sense? What did you expect?

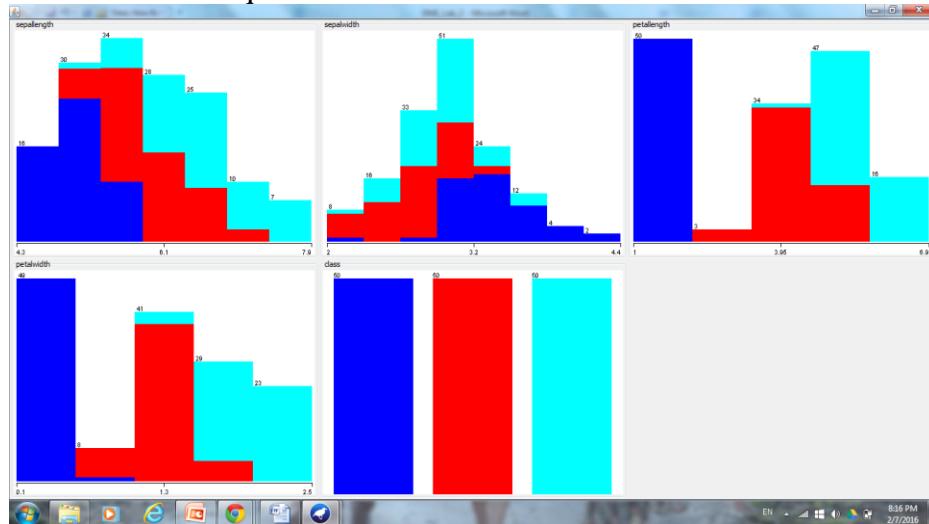
Change the acuity and cutoff parameters in order to produce a model similar to the one obtained in the book. Use the classes to cluster evaluation – what does that tell you?

- The iris.arff was uploaded into Weka, there was no need to further filter the data during the “Preprocess” phase. Note: 5 attributes exist in this data set.

Orysya Stus

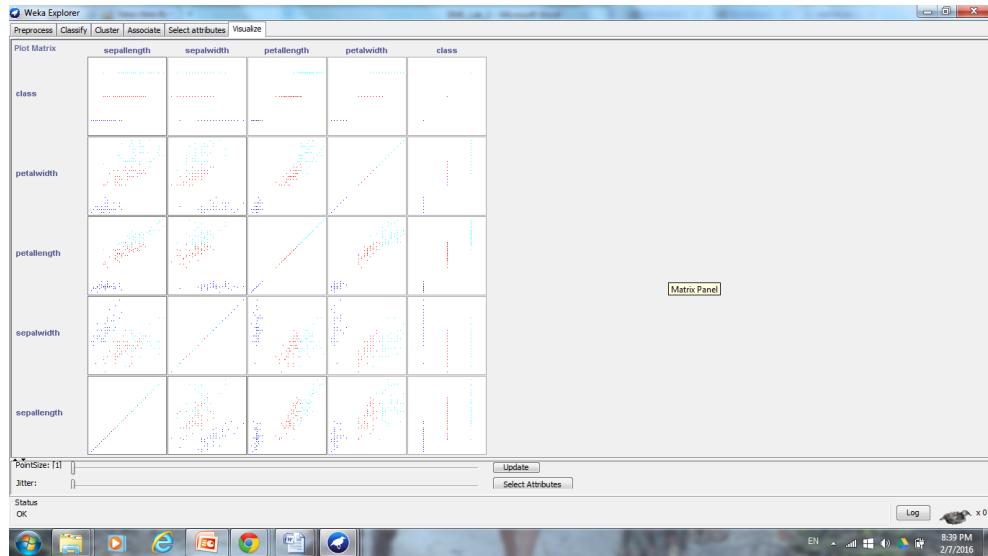


B. “Visualize all” was used to visual the distributions for each class. Where, attributes “sepallength” and “sepalwidth” follow positive skewed Gaussian distributions, the attributes “petallength” and “petalwidth” follow no concrete distribution, and the “class” attribute distributions are equal.



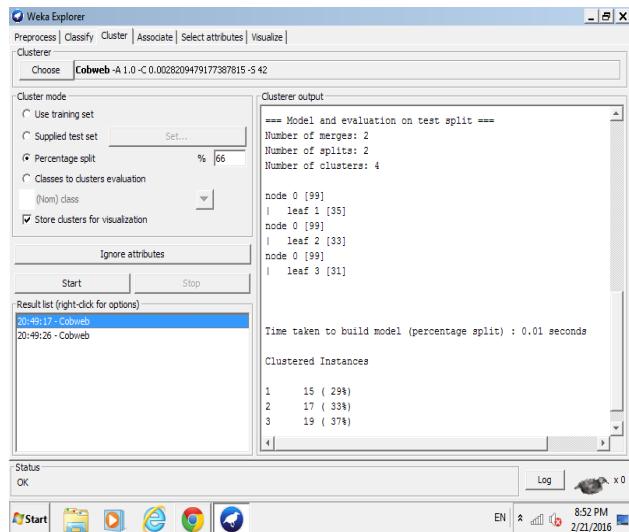
C. Utilizing the “Visualize” tab, clusters for attribute values can be seen with the most apparent clusters seen for petalwidth vs. sepalwidth/petallength and petallength vs. sepallength/sepalwidth/petalwidth. These attributes will be utilized in order to determine the class (Iris-setosa, Iris-versicolor, or Iris-virginica).

Orysya Stus

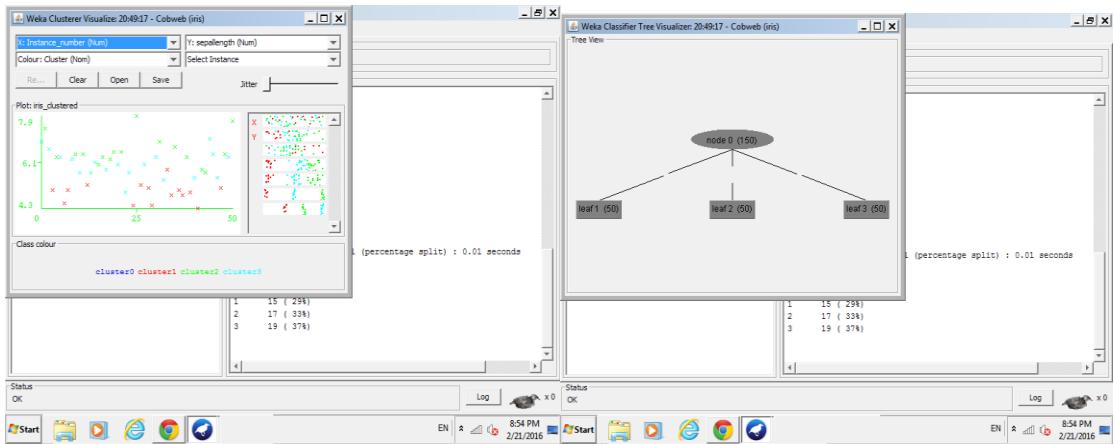


D. Under Cluster/COBWEB, the “Percentage split” at 66% was used to test the data set. COBWEB, utilizes incremental clustering on nominal values. All of the values in the iris data set are nominal. The analysis was run first with the “class” attribute as well as ignoring the attribute “class”:

With the “class” attribute:

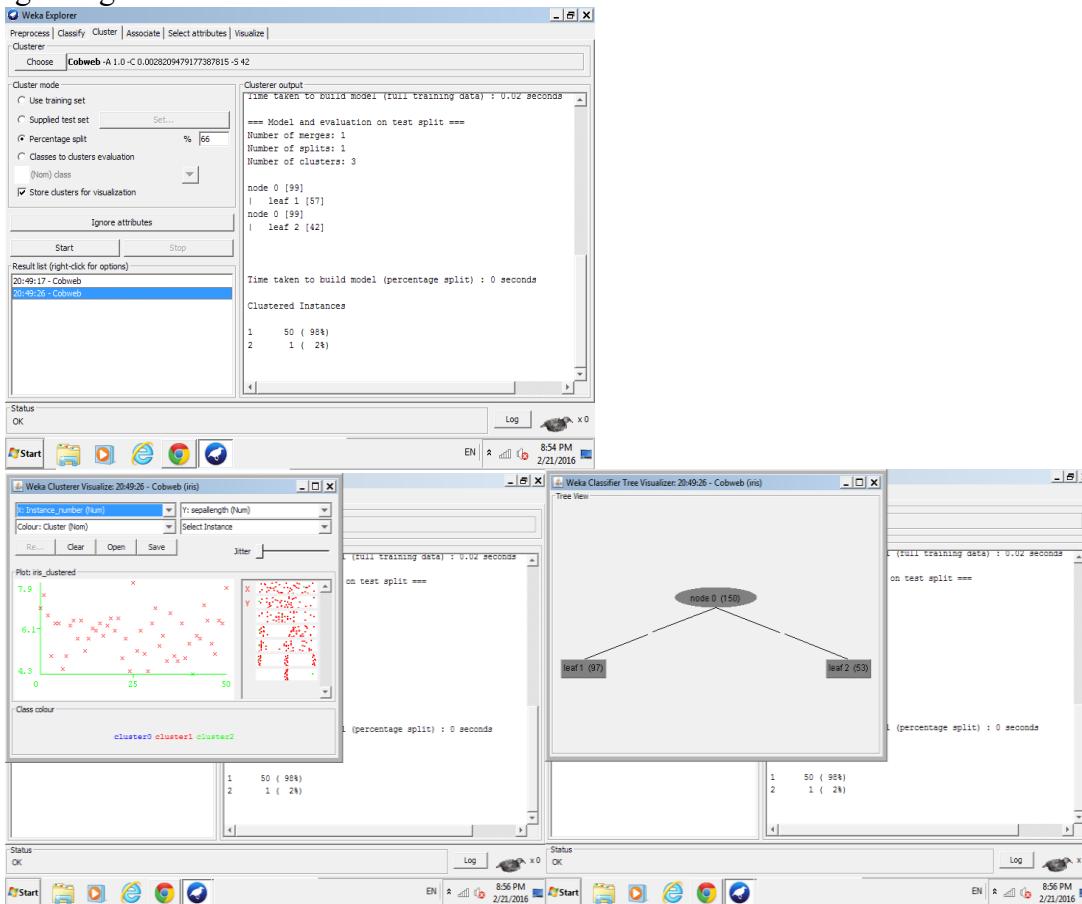


Orysya Stus



There are three leafs/clusters produced, which makes sense since the class attribute has 3 different values (Iris-setosa, Iris-versicolor, or Iris-virginica at 50:50:50).

Ignoring the “class” attribute:

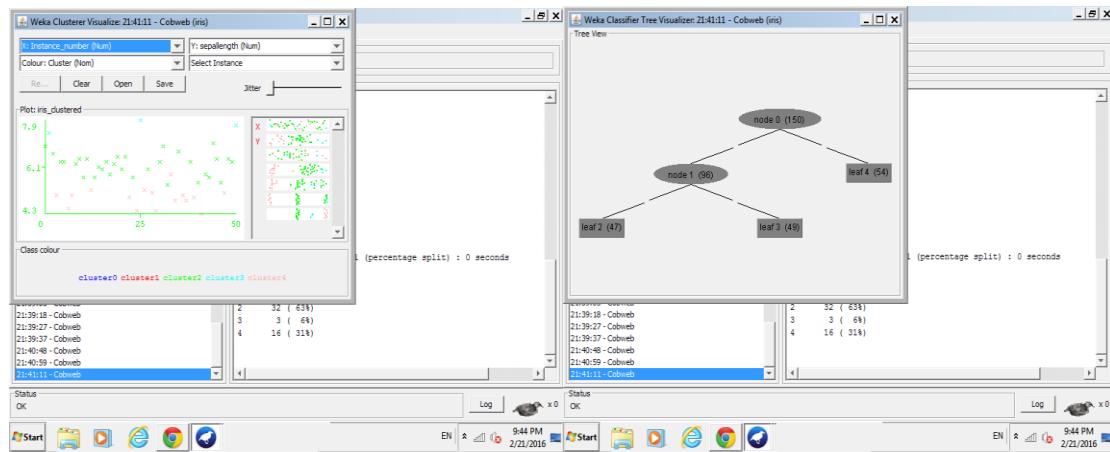
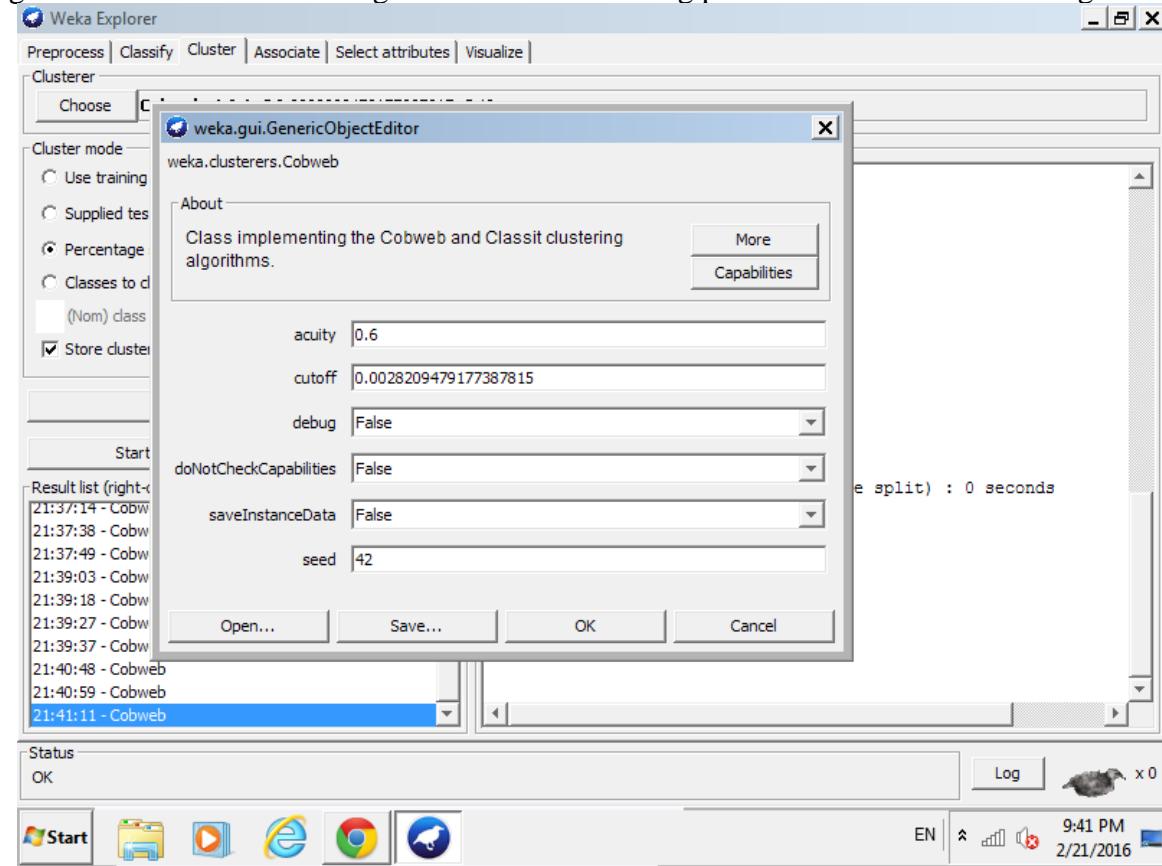


When ignoring the class attribute, only 2 cluster/leafs are generated with a 97:53 ratio which does not clearly cluster the data as well as the clustering above.

Ignoring the class, further changes to the acuity and cutoff parameters will be made to try to cluster the Iris types together and get a result as similar as possible to when the class is not ignored. Through systematic experimentation with acuity and cutoff parameters, it was found

Orysya Stus

that utilizing a acuity of 0.6 and maintaining the cutoff at the default 0.0028209479177387815, gives us the closest clustering to that of the clustering performed when class is not ignored.



The ratio of the clusters/leaves are 47:49:54, which is the closest to the 50:50:50 ration seen previously.