

## *Data Mining for Scientific Applications*

Course No. CSE-40770

### *Laboratory Assignment III:*

**To download additional .arff data sets go to:**

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

or

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

**zoo.arff, wine.arff, soybean.arff, zoo2\_x.arff,  
sunburn.arff, disease.arff**

or

**UCI data repository**

*You can find all of these files under the Resources section of the Blackboard as well!*

1. Use the following learning schemes to compare the training set and 10-fold stratified cross-validation scores of the disease data (in disease.arff):

Decision table - weka.classifiers.DecisionTable -R

C4.5 - weka.classifiers.j48.J48

Id3 - weka.clusterers.Id3

A. The disease.arff file was opened in Weka. “Visualize all” was used to determine the instance distribution for all of the attributes. Note: all attribute values are nominal with the “class” being manic or normal.



## B. Selecting weka.classifiers.DecisionTable -R. Training Set Evaluation:

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'DecisionTable' classifier is chosen, and the 'Test options' are set to 'Use training set'. The 'Classifier output' pane displays the following information:

Class  
CellWall  
Mucous  
Tails  
Colour

Test mode: reevaluate on training data

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 10  
Number of Rules : 2  
Non matches covered by Majority class.

Best first.  
Start set: no attributes  
Search direction: forward  
Stale search after 5 node expansions  
Total number of subsets evaluated: 14  
Merit of best subset found: 70  
Evaluation (for feature selection): CV (leave one out)  
Feature set: 2,1

Time taken to build model: 0 seconds

=== Evaluation on training set ===

--- Summary ---

Correctly Classified Instances	7	70	%		
Incorrectly Classified Instances	3	30	%		
Kappa statistic	0.4				
Mean absolute error	0.4333				
Root mean squared error	0.4595				
Relative absolute error	89.6552 %				
Root relative squared error	89.7343 %				
Total Number of Instances	10				

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.667	0.25	0.8	0.667	0.727	0.708	manic
	0.75	0.333	0.6	0.75	0.667	0.708	normal
Weighted Avg.	0.7	0.283	0.72	0.7	0.703	0.708	

--- Confusion Matrix ---

```

a b  <-- classified as
4 2 | a = manic
1 3 | b = normal
    
```

## Cross-validation 10 fold evaluation:

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'DecisionTable' classifier is chosen, and the 'Test options' are set to 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' pane displays the following information:

Class  
CellWall  
Mucous  
Tails  
Colour

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 10  
Number of Rules : 2  
Non matches covered by Majority class.

Best first.  
Start set: no attributes  
Search direction: forward  
Stale search after 5 node expansions  
Total number of subsets evaluated: 14  
Merit of best subset found: 70  
Evaluation (for feature selection): CV (leave one out)  
Feature set: 2,1

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

--- Summary ---

Correctly Classified Instances	5	50	%		
Incorrectly Classified Instances	5	50	%		
Kappa statistic	-0.087				
Mean absolute error	0.51				
Root mean squared error	0.5417				
Relative absolute error	96.7241 %				
Root relative squared error	105.0399 %				
Total Number of Instances	10				

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.667	0.75	0.371	0.667	0.615	0.438	manic
	0.25	0.333	0.333	0.25	0.286	0.438	normal
Weighted Avg.	0.5	0.583	0.476	0.5	0.484	0.438	

--- Confusion Matrix ---

```

a b  <-- classified as
4 2 | a = manic
1 3 | b = normal
    
```

## C. Selecting weka.classifiers.j48.J48. Training set evaluation:

# Orsysa Stus

**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  
☐ Cross-validation Folds 10  
☐ Percentage split % 66  
More options...

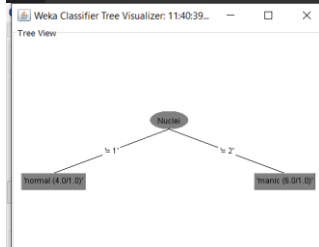
(Nom) Class: Start Stop

Result list (right-click for options):  
11:38:02 - rules.DecisionTable  
11:39:03 - rules.DecisionTable  
11:40:39 - trees.J48

Classifier output:  
Relation: Imaginary disease  
Instances: 10  
Attributes: 5  
Class  
CellWall  
Nuclei  
Tailx  
Colour

Test mode: evaluate on training data  
=== Classifier model (full training set) ===  
J48 pruned tree  
-----  
Nuclei = 1: normal (4.0/1.0)  
Nuclei = 2: manic (6.0/1.0)  
Number of Leaves : 2  
Size of the tree : 3  
Time taken to build model: 0.02 seconds  
=== Evaluation on training set ===  
=== Summary ===  
Correctly Classified Instances 8 80 %  
Incorrectly Classified Instances 2 20 %  
Kappa statistic 0.5533  
Mean absolute error 0.3167  
Root mean squared error 0.3979  
Relative absolute error 49.5172 %  
Root relative squared error 51.1763 %  
Total Number of Instances 10  
=== Detailed Accuracy By Class ===  
TP Rate FP Rate Precision Recall F-Measure ROC Area Class  
0.833 0.25 0.833 0.833 0.833 0.792 manic  
0.167 0.75 0.167 0.167 0.167 0.792 normal  
Weighted Avg. 0.8 0.217 0.8 0.8 0.8 0.792  
=== Confusion Matrix ===  
a b <-- Classified as  
5 1 a = manic  
1 3 b = normal

Status: OK



## Cross-validation 10 fold evaluation:

**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  
☒ Cross-validation Folds 10  
☐ Percentage split % 66  
More options...

(Nom) Class: Start Stop

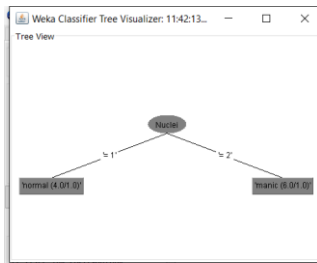
Result list (right-click for options):  
11:38:02 - rules.DecisionTable  
11:39:03 - rules.DecisionTable  
11:40:39 - trees.J48  
11:42:13 - trees.J48

Classifier output:  
Relation: Imaginary disease  
Instances: 10  
Attributes: 5  
Class  
CellWall  
Nuclei  
Tailx  
Colour

Test mode: 10-fold cross-validation  
=== Classifier model (full training set) ===  
J48 pruned tree  
-----  
Nuclei = 1: normal (4.0/1.0)  
Nuclei = 2: manic (6.0/1.0)  
Number of Leaves : 2  
Size of the tree : 3  
Time taken to build model: 0 seconds  
=== Stratified cross-validation ===  
=== Summary ===  
Correctly Classified Instances 3 30 %  
Incorrectly Classified Instances 7 70 %  
Kappa statistic -0.5217  
Mean absolute error 0.71  
Root mean squared error 0.7903  
Relative absolute error 134.6552 %  
Root relative squared error 147.7818 %  
Total Number of Instances 10  
=== Detailed Accuracy By Class ===  
TP Rate FP Rate Precision Recall F-Measure ROC Area Class  
0.5 1 0.429 0.5 0.462 0 manic  
0 0.5 0 0 0 0 normal  
Weighted Avg. 0.3 0.8 0.257 0.3 0.277 0  
=== Confusion Matrix ===  
a b <-- Classified as  
3 3 a = manic  
4 0 b = normal

Status: OK

## Orysya Stus



### D. Selecting weka.classifier.Id3. Training set evaluation:

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: **Id3**

Test options:

- ☒ Use training set
- ☐ Supplied test set
- ☐ Cross-validation
- ☐ Percentage split

(Nom) Class: **Normal**

Result list (right-click for options):

- 11:38:02 - rules.DecisionTable
- 11:39:03 - rules.DecisionTable
- 11:40:39 - trees.J48
- 11:42:13 - trees.J48
- 11:46:41 - trees.Id3**

Classifier output:

```
Class
CellWall
Nuclei
Tails
Colour

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Id3

Nuclei = 1
| CellWall = thick
| | Colour = light: normal
| | Colour = dark: manic
| CellWall = thin: normal
Nuclei = 2
| Tails = 1
| | CellWall = thick: manic
| | CellWall = thin: normal
| Tails = 2: manic

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances      10          100 %
Incorrectly Classified Instances    0           0 %
Kappa statistic                    1
Mean absolute error                0
Root mean squared error            0
Relative absolute error            0 %
Root relative squared error        0 %
Total Number of Instances         10

=== Detailed Accuracy By Class ===

TP Rate  PP Rate  Precision  Recall  F-Measure  ROC Area  Class
1       0       1       1       1       1       manic
1       0       1       1       1       1       normal
Weighted Avg.  1       0       1       1       1       1

=== Confusion Matrix ===

a b  <-- Classified as
0 0 | a = manic
0 4 | b = normal
```

### Cross-validation 10 fold evaluation:

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose Id3

Test options:  
☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds: 10  
☐ Percentage split %: 66  
 More options...

(Nom) Class: Id3

Result list (right-click for options):  
 11:38:52 - rules.DecisionTable  
 11:39:53 - rules.DecisionTable  
 11:40:39 - trees.J48  
 11:42:13 - trees.J48  
 11:48:41 - trees.J48  
 11:49:53 - trees.Id3

Classifier output:

```

Class
CellWall
Muciei
Tails
Colour

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===

Id3

Muciei = 1
| CellWall = thick
| | Colour = light: normal
| | Colour = dark: manic
Muciei = 2
| Tails = 1
| | CellWall = thick: manic
| | CellWall = thin: normal
| Tails = 2: manic

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      4      40 %
Incorrectly Classified Instances    6      60 %
Kappa statistic                    -0.3636
Mean absolute error                 0.6
Root mean squared error             0.7746
Relative absolute error             113.7931 %
Root relative squared error         144.8539 %
Total Number of Instances          10

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
0.667    1        0.5        0.667  0.571    0.333    manic
0        0.333    0         0      0         0        normal
Weighted Avg.  0.4    0.733    0.3    0.4    0.343    0.333

=== Confusion Matrix ===
a b  <-- classified as
4 2 | a = manic
4 0 | b = normal
  
```

Status: OK

Log

Search the web and Windows

11:49 AM 3/16/2016

### Analysis:

Learning Scheme	Instances correctly classified (#/10)	Mean absolute error	Information about classifier used	Means of validation information
Decision table full training set	7	0.4333	Precise yet compact way to model complex rule sets and their corresponding actions	Entire data set used since the data set is small
Decision table 10-fold cross validation	5	0.51	Precise yet compact way to model complex rule sets and their corresponding actions	Data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing
J48 full training set	8	0.3167	Decision tree for classification with min of 2 instances per node.	Entire data set used since the data set is small
J48 10-fold cross validation	3	0.71	Decision tree for classification with min of 2 instances per node.	Data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing
Id3 full training set	10	0	Recursive decision tree classifier which determines the attribute with the minimum entropy and	Entire data set used since the data set is small

			selects it.	
Id3 10-fold cross validation	4	0.6	Recursive decision tree classifier which determines the attribute with the minimum entropy and selects it.	Data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing

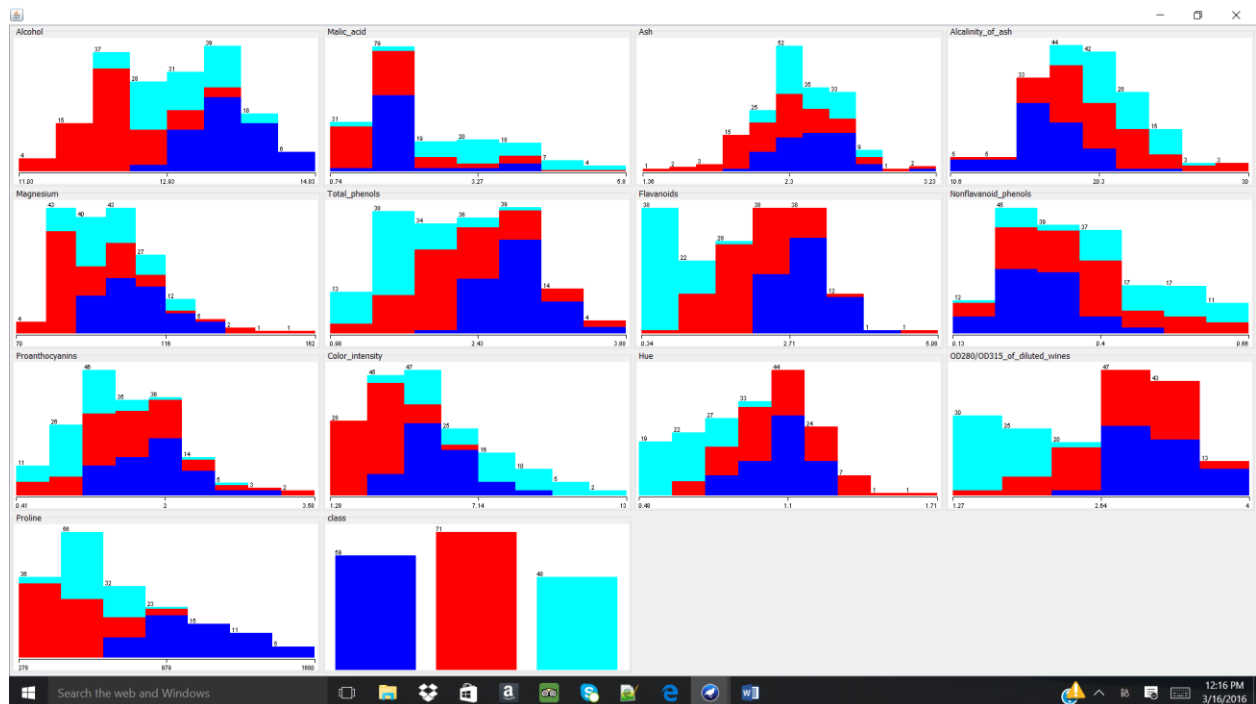
Using Id3 full training set evaluation on the disease.arff allows for all of the 10 instances to become correctly classified with a mean absolute error of zero. Using full training set evaluation is optimal since the dataset is very small, only 10 instances, and therefore each instance is representative and necessary for creating a model. Id3 allows for the “class” attribute to be determined from the attribute “nuclei”, showing that “nuclei” provides the most information gain or is the least entropic attribute in the data set.

2. Use the following learning schemes to analyze the wine data (in wine.arff).

C4.5 - weka.classifiers.j48.J48

Decision List - weka.classifiers.PART

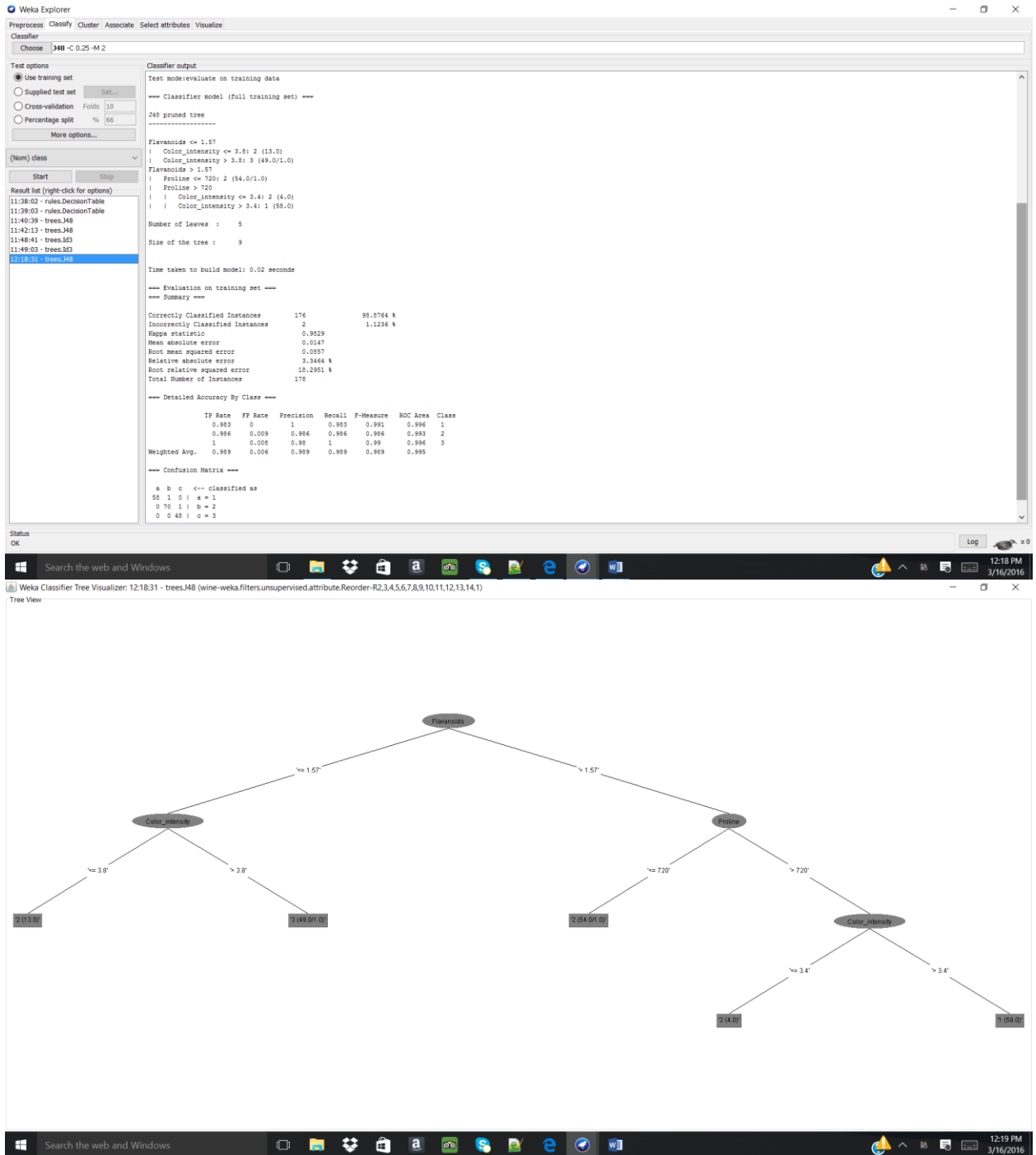
A. The wine.arff dataset was opened in Weka and “Visualize all” was used to see the instance distribution for each attribute. Note: all attribute values are numeric except the “class” attribute which is nominal.



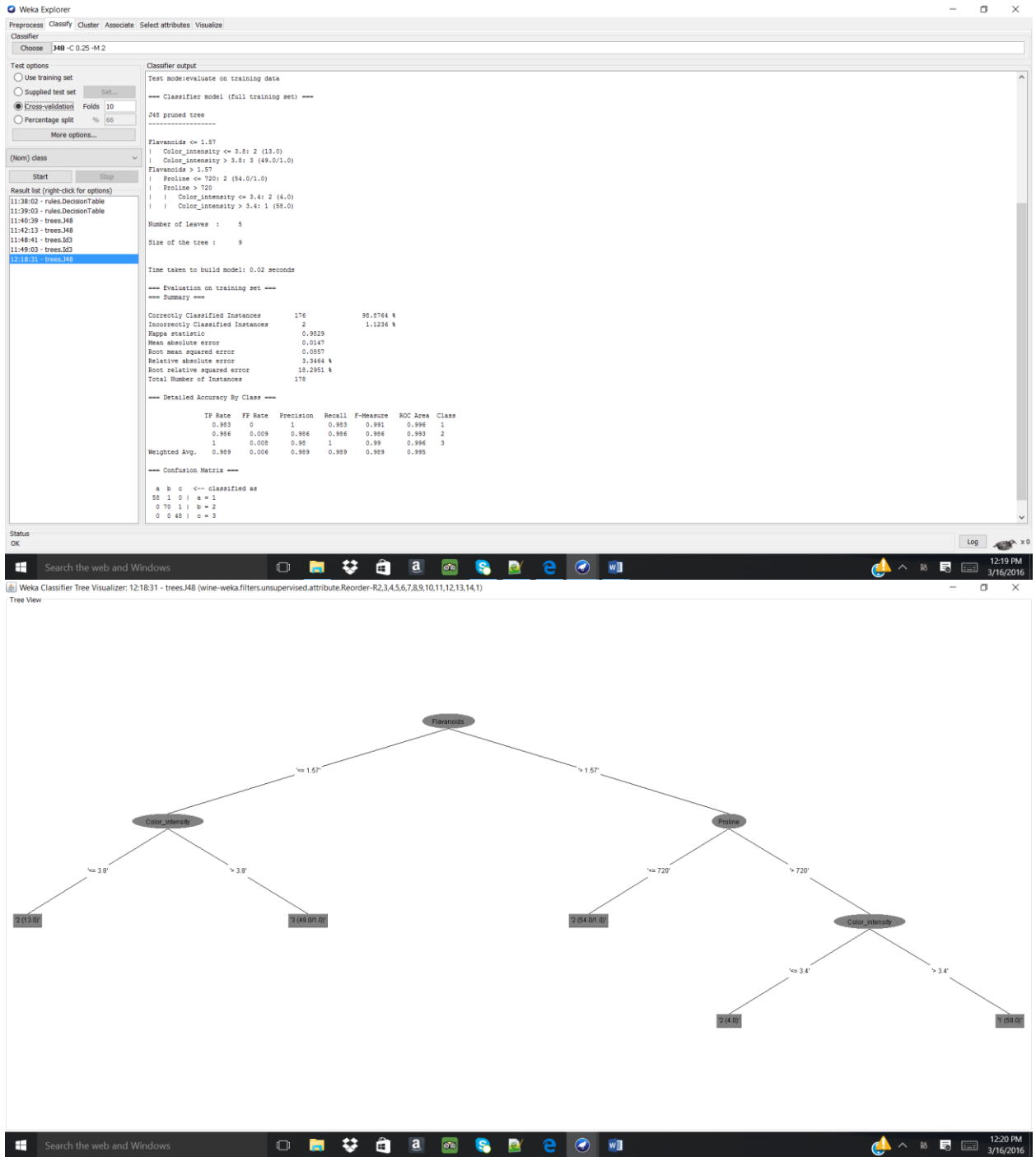
B. Selecting weka.classifiers.j48.J48.

Training set evaluation:

# Orysya Stus



Cross-validation 10 fold evaluation:



C. Selecting weka.classifiers.PART.  
Training set evaluation:



# Orysya Stus

**Weka Explorer**  
Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **PART-M2-C 0.25-Q 1**

Test options:  
☒ Use training set  
☐ Supplied test set  
☐ Cross-validation Folds: 10  
☐ Percentage split %: 66  
More options...

(Nom) class: Start Stop

Result list (right-click for options):  
11:38:02 - rules.DecisionTable  
11:39:03 - rules.DecisionTable  
11:40:39 - trees.M48  
11:42:13 - trees.M48  
11:48:41 - trees.M53  
11:49:03 - trees.M53  
12:18:31 - trees.M48  
12:21:58 - rules.PART

**Classifier output**

==== Classifier model (full training set) ====

PART decision list

Flavanoids <= 1.57 AND  
Color\_intensity > 3.8: 3 (49.0/1.0)

Proline <= 750 AND  
Alcohol <= 13.17: 2 (62.0)

Color\_intensity > 3.8: 1 (55.0)

Malic\_acid <= 1.73: 2 (9.0/1.0)

: 1 (3.0)

Number of Rules : 5

Time taken to build model: 0.01 seconds

==== Evaluation on training set ====

==== Summary ====

Correctly Classified Instances	176	98.8764 %
Incorrectly Classified Instances	2	1.1236 %
Kappa statistic	0.9929	
Mean absolute error	0.014	
Root mean squared error	0.037	
Relative absolute error	3.1894 %	
Root relative squared error	17.558 %	
Total Number of Instances	178	

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.983	0	1	0.983	0.991	0.999	1
	0.956	0.009	0.956	0.956	0.956	0.996	2
	1	0.008	0.98	1	0.99	0.996	3
Weighted Avg.	0.959	0.006	0.959	0.959	0.959	0.997	

==== Confusion Matrix ====

a	b	c	<-- classified as
58	1	0	a = 1
0	70	1	b = 2
0	0	48	c = 3

**Weka Explorer**  
Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **PART-M2-C 0.25-Q 1**

Test options:  
☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds: 10  
☐ Percentage split %: 66  
More options...

(Nom) class: Start Stop

Result list (right-click for options):  
11:38:02 - rules.DecisionTable  
11:39:03 - rules.DecisionTable  
11:40:39 - trees.M48  
11:42:13 - trees.M48  
11:48:41 - trees.M53  
11:49:03 - trees.M53  
12:18:31 - trees.M48  
12:21:58 - rules.PART  
12:22:46 - rules.PART

**Classifier output**

==== Classifier model (full training set) ====

PART decision list

Flavanoids <= 1.57 AND  
Color\_intensity > 3.8: 3 (49.0/1.0)

Proline <= 750 AND  
Alcohol <= 13.17: 2 (62.0)

Color\_intensity > 3.8: 1 (55.0)

Malic\_acid <= 1.73: 2 (9.0/1.0)

: 1 (3.0)

Number of Rules : 5

Time taken to build model: 0.01 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	166	93.2584 %
Incorrectly Classified Instances	12	6.7416 %
Kappa statistic	0.897	
Mean absolute error	0.05	
Root mean squared error	0.2137	
Relative absolute error	11.394 %	
Root relative squared error	45.6132 %	
Total Number of Instances	178	

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.949	0.017	0.966	0.949	0.957	0.971	1
	0.953	0.004	0.953	0.956	0.954	0.907	2
	0.875	0.008	0.977	0.875	0.923	0.931	3
Weighted Avg.	0.933	0.041	0.936	0.933	0.933	0.947	

==== Confusion Matrix ====

a	b	c	<-- classified as
56	3	0	a = 1
2	68	1	b = 2
0	6	42	c = 3

## Analysis:

Learning Scheme	Instances correctly classified (#/178)	Mean absolute error	Information about classifier used	Means of validation information
J48 full training set	176	0.0147	Decision tree for classification with min of 2 instances per	Entire data set used

			node.	
J48 10-fold cross validation	167	0.0486	Decision tree for classification with min of 2 instances per node.	Data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing
PART full training set	176	0.0486	Created IF/AND rules for classifying data sets.	Entire data set used
PART 10-fold cross validation	166	0.05	Created IF/AND rules for classifying data sets.	Data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing

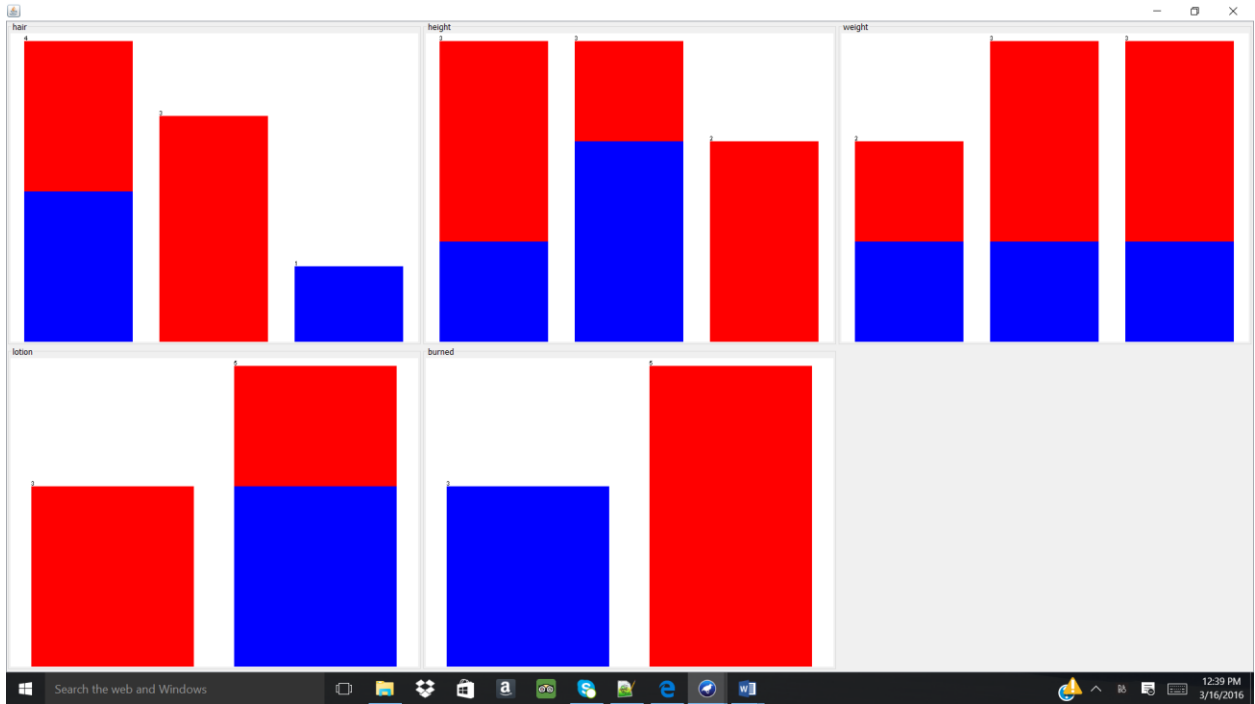
The attribute “Flavanoids” is the most important descriptor in the win.arff dataset as it provides the greatest information gain in classifying the types of “class”. Using the absolute mean error to quantify the success of the learning schemes, it was found that the J48 full training set performed the best with the lowest error at 0.0147 while classifying the majority of the instances. Therefore, the J48 full training set is trusted more over the other learning schemes and evaluations. Both learning schemes were able to classifying the instances into one of the three types of class.

3. Perform the same analysis of sunburn.arff as in Question #2. Instead of 10-fold cross-validations use 5-fold cross-validation.

Answer the same questions as in A)-E) in the question #2.

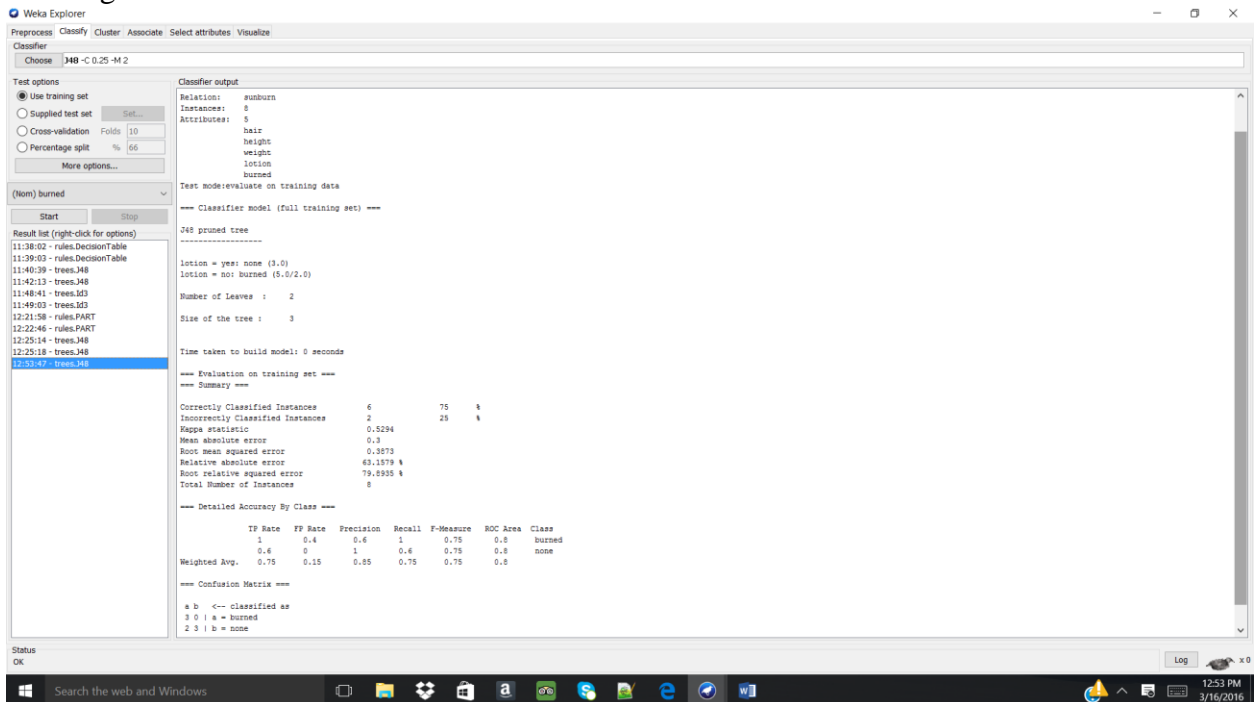
A. The sunburn.arff dataset was opened in Weka and “Visualize All” was used to see the instances distributions for each of the attributes. Note: all attribute values are nominal with attribute “burned” being the class attribute.

## Orysyta Stus

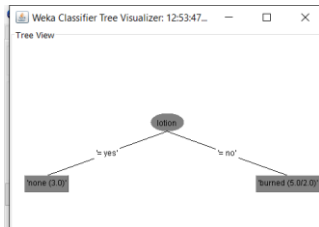


## B. Selecting weka.classifiers.j48.J48.

### Training set evaluation:



# Orysya Stus



Cross-validation 5 fold evaluation:

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
- ☒ Cross-validation Folds: 5
- ☐ Percentage split %: 66

More options...

(Name) burned

Start Stop

Result list (right-click for options)

- 11:38:02 - rules.DecisionTable
- 11:39:03 - rules.DecisionTable
- 11:40:39 - trees.J48
- 11:42:13 - trees.J48
- 11:48:41 - trees.J48
- 11:49:03 - trees.J48
- 12:21:58 - rules.PART
- 12:22:46 - rules.PART
- 12:25:14 - trees.J48
- 12:25:18 - trees.J48
- 12:53:47 - trees.J48
- 12:55:05 - trees.J48

Classifier output

Relation: sunburn  
Instances: 8  
Attributes: 5  
hair  
height  
weight  
lotion  
burned

Test mode: 5-fold cross-validation

== Classifier model (full training set) ==

J48 pruned tree

```
lotion = yes: none (3.0)
lotion = no: burned (5.0/2.0)
Number of Leaves : 2
Size of the tree : 3
```

Time taken to build model: 0 seconds

== Stratified cross-validation ==

== Summary ==

Correctly Classified Instances	3	37.5 %
Incorrectly Classified Instances	5	62.5 %
Kappa statistic	-0.25	
Mean absolute error	0.5521	
Root mean squared error	0.5721	
Relative absolute error	113.9714 %	
Root relative squared error	115.9451 %	
Total Number of Instances	8	

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.333	0.6	0.25	0.333	0.286	0.333	burned
	0.4	0.667	0.5	0.4	0.444	0.333	none
Weighted Avg.	0.375	0.642	0.406	0.375	0.385	0.333	

== Confusion Matrix ==

```
a b -- Classified as
1 2 | a = burned
3 2 | b = none
```

C. Selecting weka.classifiers.PART.

Training set evaluation:

# Orysya Stus

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose PART-M2-C 0.25-Q 1

Test options:  
☒ Use training set  
☐ Supplied test set  
☐ Cross-validation Folds 5  
☐ Percentage split % 66  
More options...

(Nom) burned

Start Stop

Result list (right-click for options):  
11:38:02 - rules.DecisionTable  
11:39:03 - rules.DecisionTable  
11:40:39 - trees.J48  
11:42:13 - trees.J48  
11:48:41 - trees.J48  
11:49:03 - trees.J48  
12:21:58 - rules.PART  
12:22:46 - rules.PART  
12:25:14 - trees.J48  
12:25:18 - trees.J48  
12:53:47 - trees.J48  
12:55:05 - trees.J48  
13:06:19 - rules.PART

Classifier output:  
Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1  
Relation: sunburn  
Instances: 8  
Attributes: 5  
hair  
height  
weight  
lotion  
burned  
Test mode: evaluate on training data

==== Classifier model (full training set) ====

PART decision list  
-----  
lotion = no: burned (5.0/2.0)  
: none (3.0)  
Number of Rules : 2

Time taken to build model: 0 seconds

==== Evaluation on training set ====

==== Summary ====

Correctly Classified Instances	6	75 %
Incorrectly Classified Instances	2	25 %
Kappa statistic	0.5294	
Mean absolute error	0.3	
Root mean squared error	0.3873	
Relative absolute error	63.1579 %	
Root relative squared error	79.8935 %	
Total Number of Instances	8	

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.4	0.6	1	0.75	0.8	burned
0.6	0	1	0.6	0.75	0.8	none
Weighted Avg.	0.75	0.15	0.85	0.75	0.75	

==== Confusion Matrix ====

```
a b <-- classified as
3 0 | a = burned
2 2 | b = none
```

Status OK

## Cross-validation 5 fold evaluation:

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose PART-M2-C 0.25-Q 1

Test options:  
☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds 5  
☐ Percentage split % 66  
More options...

(Nom) burned

Start Stop

Result list (right-click for options):  
11:38:02 - rules.DecisionTable  
11:39:03 - rules.DecisionTable  
11:40:39 - trees.J48  
11:42:13 - trees.J48  
11:48:41 - trees.J48  
11:49:03 - trees.J48  
12:21:58 - rules.PART  
12:22:46 - rules.PART  
12:25:14 - trees.J48  
12:25:18 - trees.J48  
12:53:47 - trees.J48  
12:55:05 - trees.J48  
13:06:19 - rules.PART  
13:06:14 - rules.PART

Classifier output:  
Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1  
Relation: sunburn  
Instances: 8  
Attributes: 5  
hair  
height  
weight  
lotion  
burned  
Test mode: 5-fold cross-validation

==== Classifier model (full training set) ====

PART decision list  
-----  
lotion = no: burned (5.0/2.0)  
: none (3.0)  
Number of Rules : 2

Time taken to build model: 0 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	3	37.5 %
Incorrectly Classified Instances	5	62.5 %
Kappa statistic	-0.25	
Mean absolute error	0.5521	
Root mean squared error	0.5761	
Relative absolute error	113.5714 %	
Root relative squared error	113.9451 %	
Total Number of Instances	8	

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.333	0.6	0.25	0.333	0.286	0.333	burned
0.4	0.667	0.5	0.4	0.444	0.333	none
Weighted Avg.	0.375	0.442	0.456	0.375	0.385	

==== Confusion Matrix ====

```
a b <-- classified as
1 2 | a = burned
3 2 | b = none
```

Status OK

## Analysis:

Learning Scheme	Instances correctly classified (#/8)	Mean absolute error	Information about classifier used	Means of validation information
J48 full training set	6	0.3	Decision tree for classification with min of 2 instances per	Entire data set used since this is a small data set

			node.	
J48 5-fold cross validation	3	0.5521	Decision tree for classification with min of 2 instances per node.	Data is split into training and test data sets 5 times with models being created, iterated, and polished between each training and testing
PART full training set	6	0.3	Created IF/AND rules for classifying data sets.	Entire data set used since this is a small data set
PART 5-fold cross validation	3	0.5521	Created IF/AND rules for classifying data sets.	Data is split into training and test data sets 5 times with models being created, iterated, and polished between each training and testing

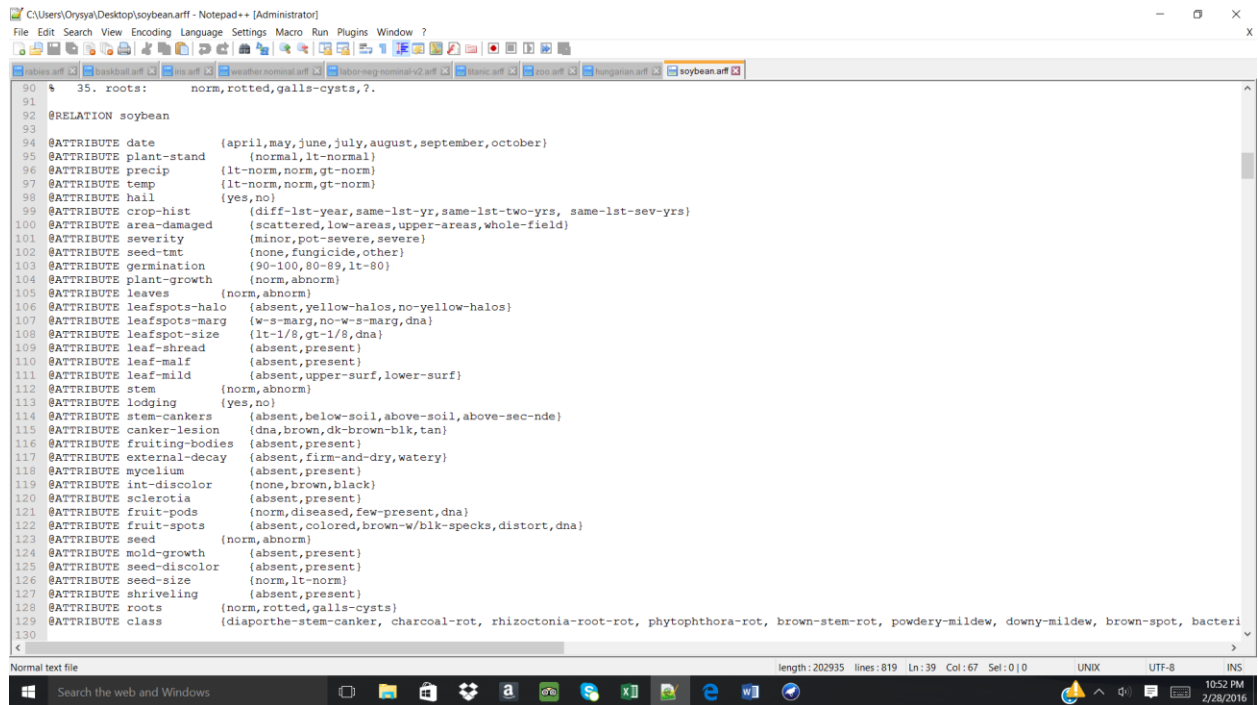
The attribute “lotion” is the most important descriptor in the sunburn.arff dataset as it provides the greatest information gain in classifying the types of “class”. Using the absolute mean error to quantify the success of the learning schemes, it was found that both the J48 and PART full training set performed the best, with the lowest error at 0.0147 while classifying the majority of the instances. Since the data set is so small, 8 instances, the full training set should be used for evaluation as each instance is representative of the data. Therefore, the J48 or PART full training set is trusted more over the other learning schemes and evaluations. Both learning schemes were able to classifying the instances into “burned” or “none”. 10-fold evaluation was not used in this example because you cannot have more folds than instances as an error in the system will occur, therefore 5-fold evaluation was used.

4. Choose one of the following three files: soybean.arff, zoo.arff or zoo2\_x.arff and use any two schemas of your choice to build and compare the models. Describe in details the process of building (data set, parameter settings/changes, etc) and evaluation of each individual model and comparison of the different models.

For this problem, file soybean.arff was selected.

A. The data set soybean.arff was opened in Notepad to understand information about the data set.

## Orysya Stus



```
35. roots: norm,rotted,galls-cysts,7.

@RELATION soybean

@ATTRIBUTE date (april,may,june,july,august,september,october)
@ATTRIBUTE plant-stand (normal,lt-normal)
@ATTRIBUTE precip (lt-norm,norm,gt-norm)
@ATTRIBUTE temp (lt-norm,norm,gt-norm)
@ATTRIBUTE hail (yes,no)
@ATTRIBUTE crop-hist (diff-1st-year,same-1st-yr,same-1st-two-yr, same-1st-sev-yr)
@ATTRIBUTE area-damaged (scattered,low-areas,upper-areas,whole-field)
@ATTRIBUTE severity (minor,pot-severe,severe)
@ATTRIBUTE seed-tmt (none,fungicide,other)
@ATTRIBUTE germination (90-100,80-89,lt-80)
@ATTRIBUTE plant-growth (norm,abnorm)
@ATTRIBUTE leaves (norm,abnorm)
@ATTRIBUTE leafspots-halo (absent,yellow-halos,no-yellow-halos)
@ATTRIBUTE leafspots-marg (w-s-marg,no-w-s-marg,dna)
@ATTRIBUTE leafspot-size (lt-1/8,gt-1/8,dna)
@ATTRIBUTE leaf-shread (absent,present)
@ATTRIBUTE leaf-malf (absent,present)
@ATTRIBUTE leaf-mild (absent,upper-surf,lower-surf)
@ATTRIBUTE stem (norm,abnorm)
@ATTRIBUTE lodging (yes,no)
@ATTRIBUTE stem-cankers (absent,below-soil,above-soil,above-sec-nde)
@ATTRIBUTE canker-lesion (dna,brown,dk-brown-blk,tan)
@ATTRIBUTE fruiting-bodies (absent,present)
@ATTRIBUTE external-decay (absent,firm-and-dry,watery)
@ATTRIBUTE mycelium (absent,present)
@ATTRIBUTE int-discolor (none,brown,black)
@ATTRIBUTE sclerotia (absent,present)
@ATTRIBUTE fruit-pods (norm,diseased,few-present,dna)
@ATTRIBUTE fruit-spots (absent,colored,brown-w/blk-specks,distort,dna)
@ATTRIBUTE seed (norm,abnorm)
@ATTRIBUTE mold-growth (absent,present)
@ATTRIBUTE seed-discolor (absent,present)
@ATTRIBUTE seed-size (norm,lt-norm)
@ATTRIBUTE shriveling (absent,present)
@ATTRIBUTE roots (norm,rotted,galls-cysts)
@ATTRIBUTE class (diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacteri
```

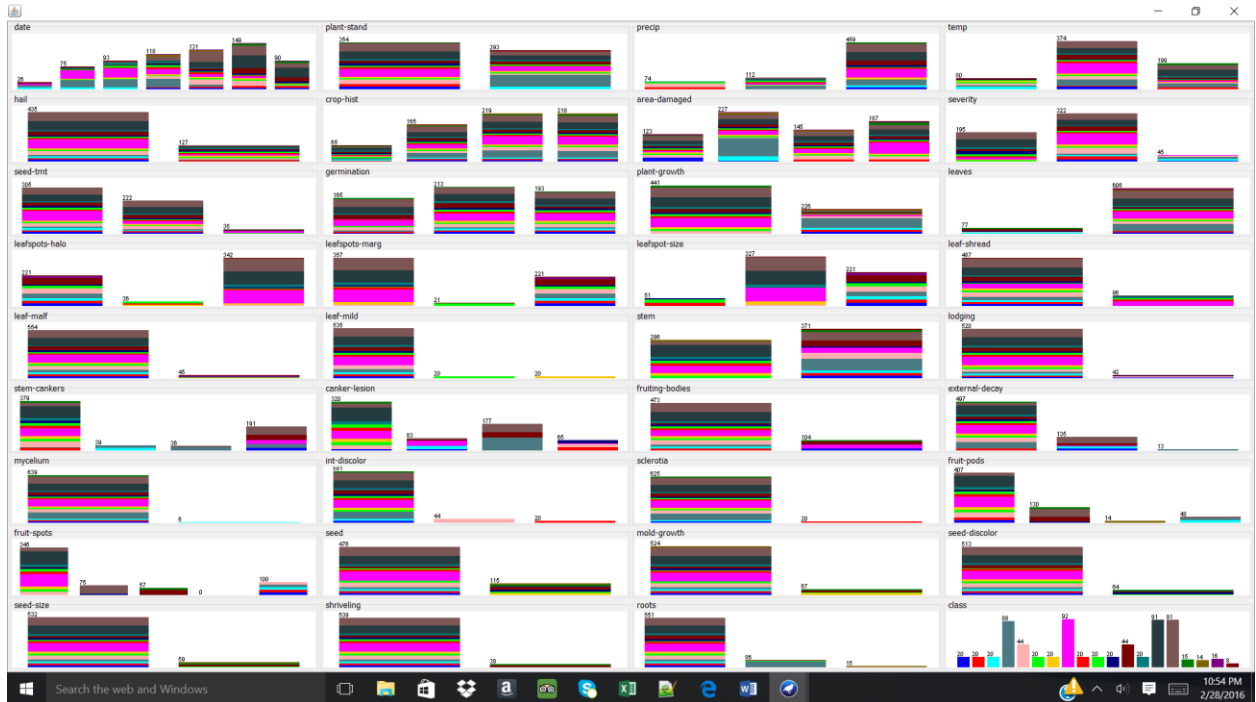
This dataset contains a multitude of attributes which will be used to classify the class of soybean.

B. The dataset soybean.arff was opened in WEKA and “Visualize all” was used to see the instance distribution for the attribute values. Note: all attributes values are nominal and the class attribute “class” had 19 different labels. The total number of instances in this dataset is 683. Since this is a fairly large dataset, the test option for modeling will be 10-fold cross validation vs. full training set. “Cross-validation” at “Fold 10” is 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 learning schemes will be utilized to understand which attributes contribute the most in understanding how the class attribute of soybean is classified:

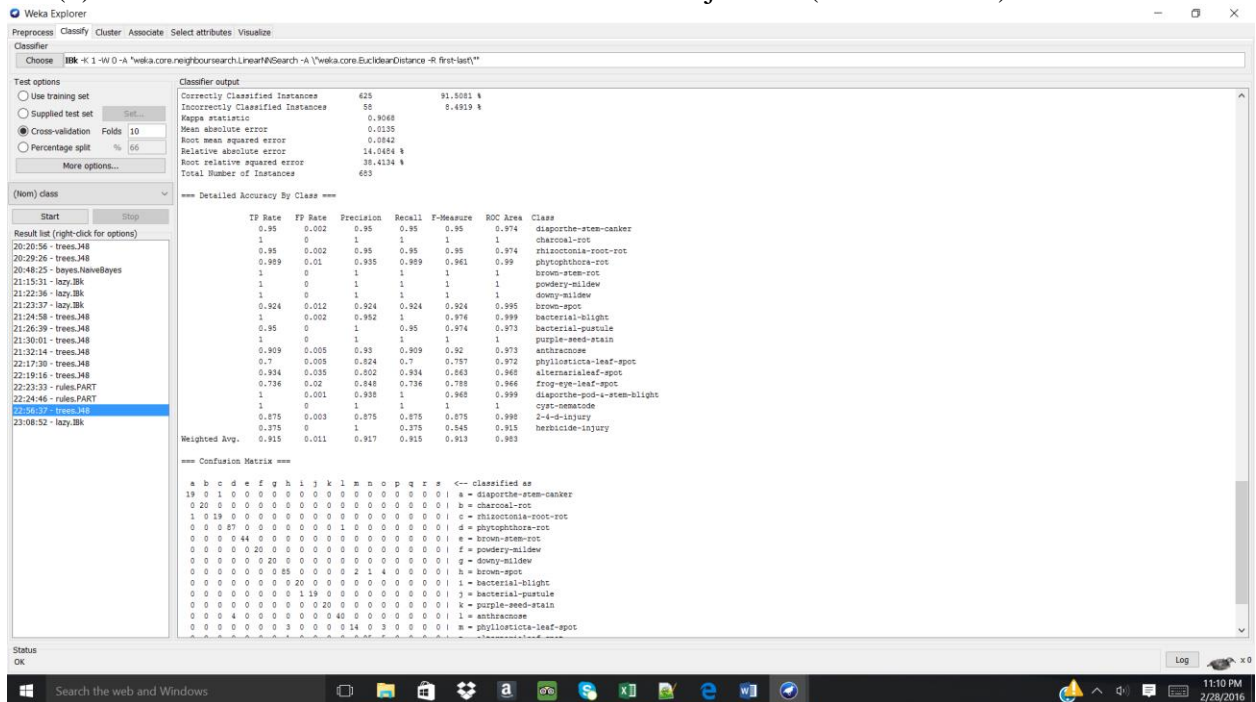
- (1) C4.5 10-fold cross validation -weka.classifiers.j48.J48 (-M2 was used)
- (2) IBK 10-fold cross validation with k=1

# Orysa Stus



C. The following learning schemes were applied:

(1) C4.5 10-fold cross validation -weka.classifiers.j48.J48 (-M2 was used)



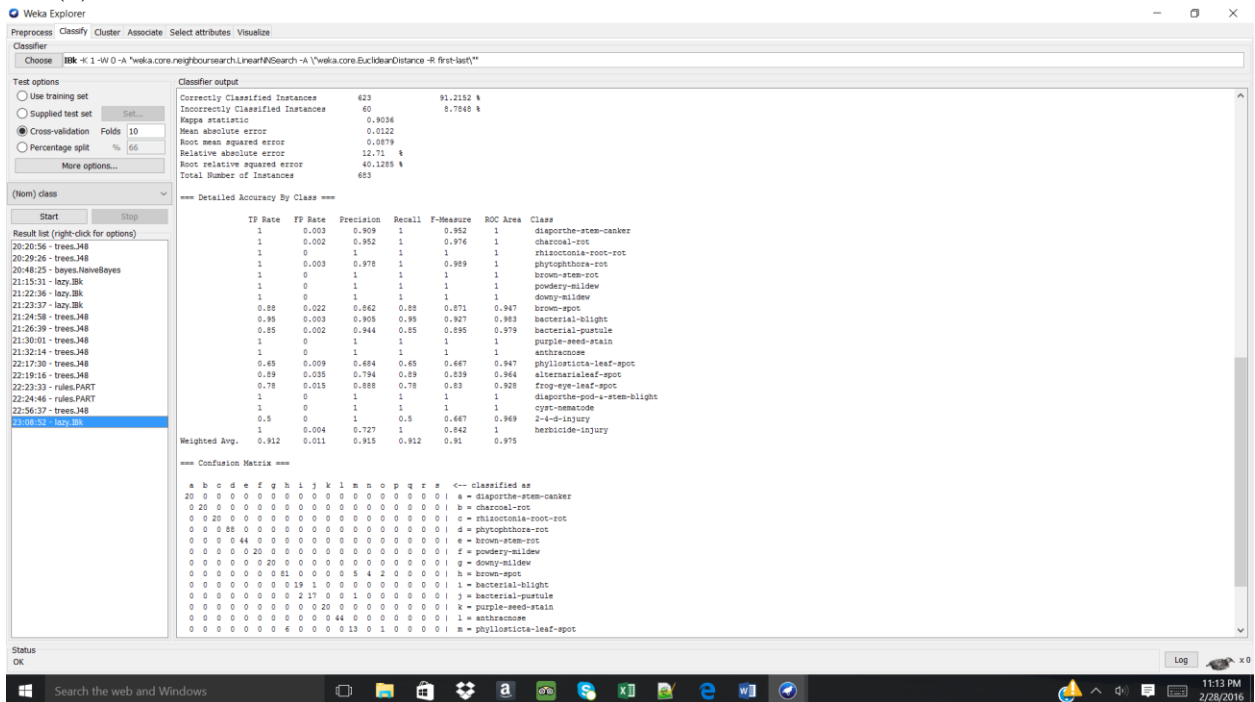


# Orysya Stus

Weka Classifier Tree Visualizer: 22:56:37 - trees.J48 (soybean)



## (2) IBK 10-fold cross validation with k=1



Understanding and analysing the above learning schemes:

Learning Scheme	Instances correctly classified (#/683)	Mean absolute error	Information about classifier used
C4.5 10-fold cross validation	625	0.0135	Decision tree for classification with min of 2 instances per node.
IBK 10-fold cross validation	623	0.0122	Simple instance-based learner that uses the class of the nearest k training instances for the class of the test instances

with k=1

Analysis of the table above, reveals that the IBk 10-fold cross validation with k=1 has a lower mean absolute error compared to the C4.5 10-fold cross validation. But in order to further understand what is occurring in the system, a closer look on the “Detailed Accuracy By Class” can be completed.

C4.5:

=== Detailed Accuracy By Class ===						
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.95	0.002	0.95	0.95	0.95	0.974	diaporthe-stem-canker
1	0	1	1	1	1	charcoal-rot
0.95	0.002	0.95	0.95	0.95	0.974	rhizoctonia-root-rot
0.989	0.01	0.935	0.989	0.961	0.99	phytophthora-rot
1	0	1	1	1	1	brown-stem-rot
1	0	1	1	1	1	powdery-mildew
1	0	1	1	1	1	downy-mildew
0.924	0.012	0.924	0.924	0.924	0.995	brown-spot
1	0.002	0.952	1	0.976	0.999	bacterial-blight
0.95	0	1	0.95	0.974	0.973	bacterial-pustule
1	0	1	1	1	1	purple-seed-stain
0.909	0.005	0.93	0.909	0.92	0.973	anthracnose
0.7	0.005	0.824	0.7	0.757	0.972	phyllosticta-leaf-spot
0.934	0.035	0.802	0.934	0.863	0.968	alternarialeaf-spot
0.736	0.02	0.848	0.736	0.788	0.966	frog-eye-leaf-spot
1	0.001	0.938	1	0.968	0.999	diaporthe-pod-&-stem-blight
1	0	1	1	1	1	cyst-nematode
0.875	0.003	0.875	0.875	0.875	0.998	2-4-d-injury
0.375	0	1	0.375	0.545	0.915	herbicide-injury
Weighted Avg.	0.915	0.011	0.917	0.915	0.913	0.983

IBk:

=== Detailed Accuracy By Class ===						
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.003	0.909	1	0.952	1	diaporthe-stem-canker
1	0.002	0.952	1	0.976	1	charcoal-rot
1	0	1	1	1	1	rhizoctonia-root-rot
1	0.003	0.978	1	0.989	1	phytophthora-rot
1	0	1	1	1	1	brown-stem-rot
1	0	1	1	1	1	powdery-mildew
1	0	1	1	1	1	downy-mildew
0.88	0.022	0.862	0.88	0.871	0.947	brown-spot
0.95	0.003	0.905	0.95	0.927	0.983	bacterial-blight
0.85	0.002	0.944	0.85	0.895	0.979	bacterial-pustule
1	0	1	1	1	1	purple-seed-stain
1	0	1	1	1	1	anthracnose
0.65	0.009	0.684	0.65	0.667	0.947	phyllosticta-leaf-spot
0.89	0.035	0.794	0.89	0.839	0.964	alternarialeaf-spot
0.78	0.015	0.888	0.78	0.83	0.928	frog-eye-leaf-spot
1	0	1	1	1	1	diaporthe-pod-&-stem-blight
1	0	1	1	1	1	cyst-nematode
0.5	0	1	0.5	0.667	0.969	2-4-d-injury
1	0.004	0.727	1	0.842	1	herbicide-injury
Weighted Avg.	0.912	0.011	0.915	0.912	0.91	0.975

As seen above, the 2 models perform fairly closer to one another with the C4.5 classifying slightly more instances correctly vs. the IBk model. Also comparing the weighted averages, it can be seen that the C4.5 model has a slightly higher true positive rate and higher precision. Therefore, I would keep the C4.5 model and use the decision tree for testing and learning schemes. From the C4.5 model, it can be learned that “leafspot-size” is the most important attribute used to classify the class attribute into 19 values while the IBk model can be used to apply different weights to the attributes for better modeling.