Orysya Stus

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!

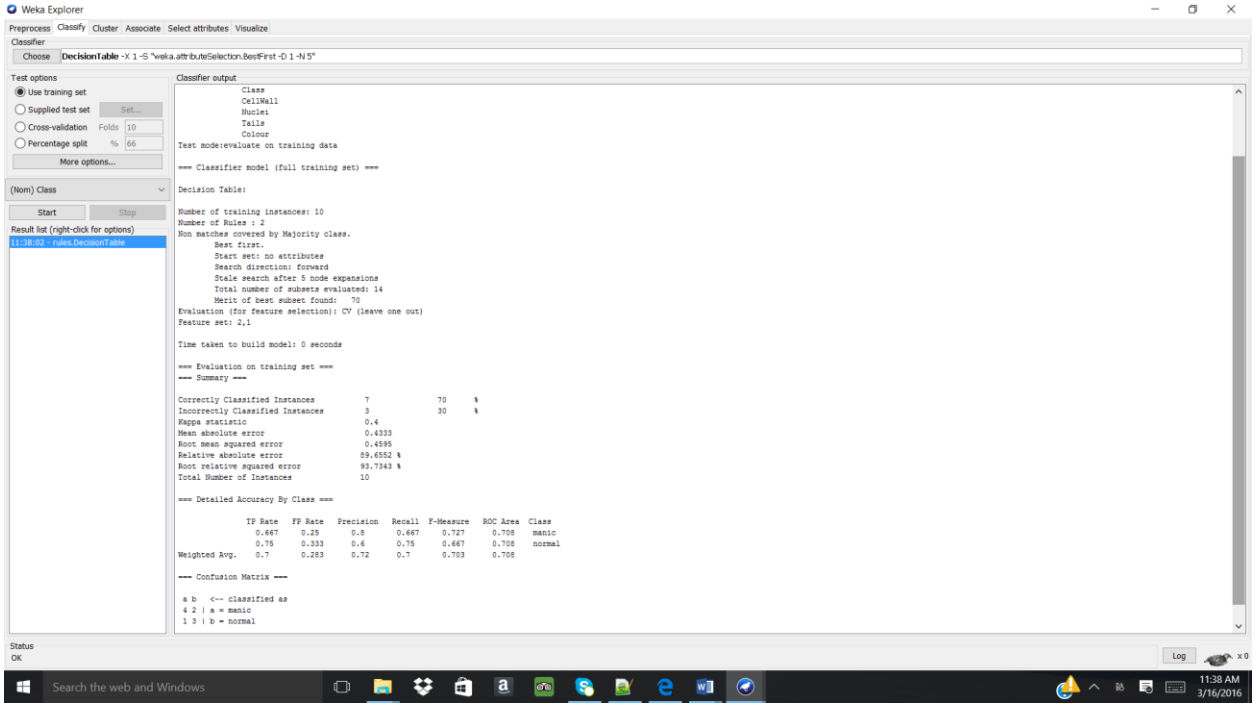
1. Use the following learning schemes to compare the training set and 10-fold stratified cross-validation scores of the disease data (in disease.arff):

   Decision table - weka.classifiers.DecisionTable -R
   C4.5 - weka.classifiers.j48.J48
   Id3 - weka.clusterers.Id3

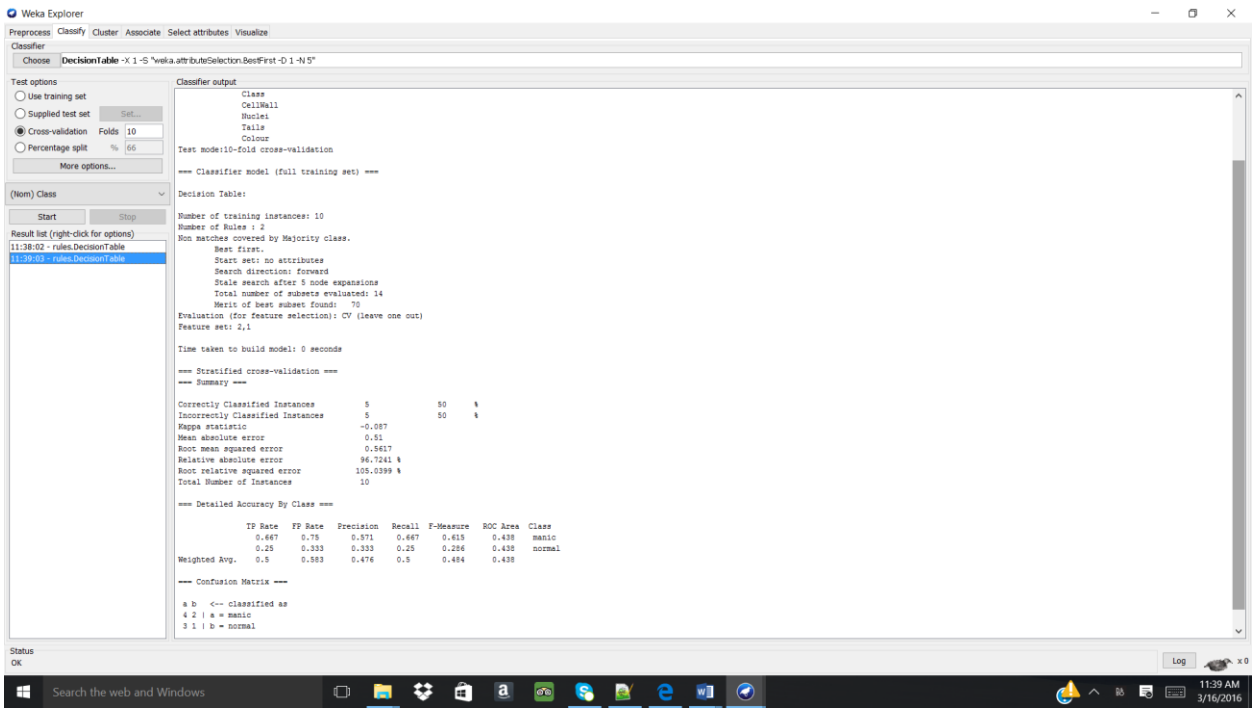
A. The disease.arff file was opened in Weka. “Visualize all” was used to determine the instance distribution for all of the attributes. Note: all attribute values are nominal with the “class” being manic or normal.

Orysya Stus

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

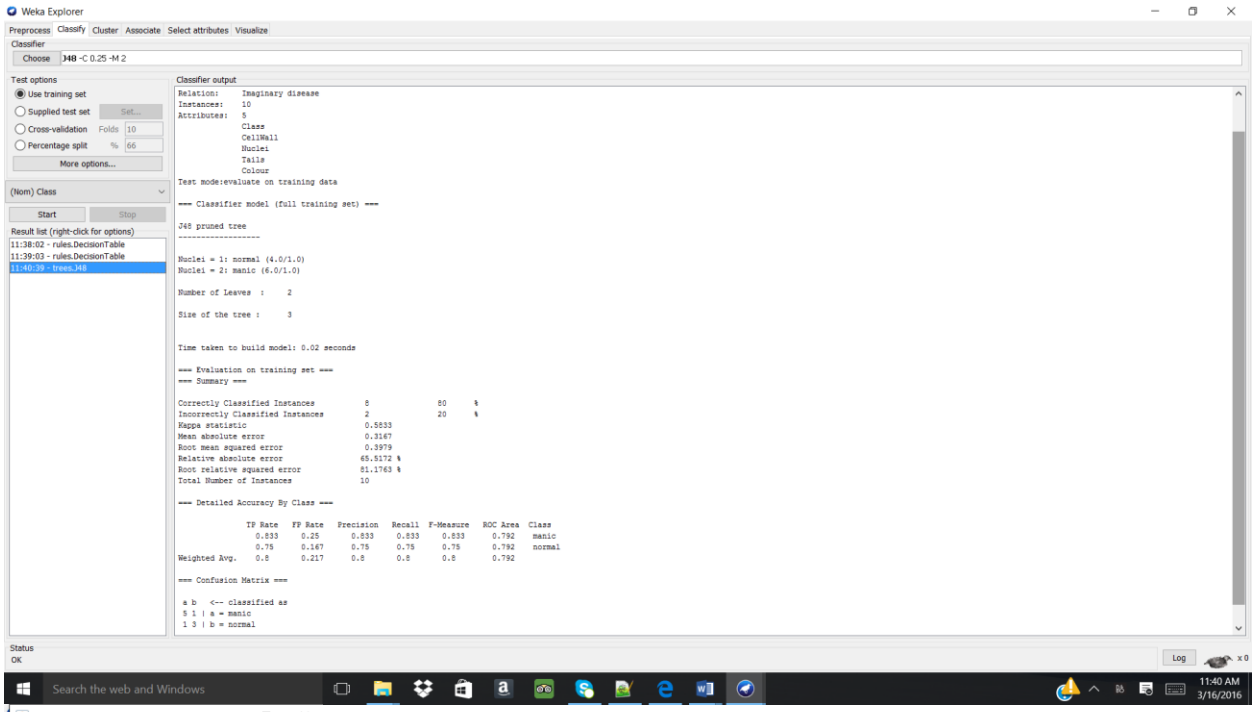


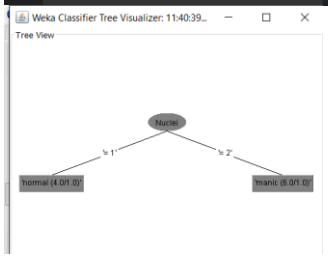
Cross-validation 10 fold evaluation:



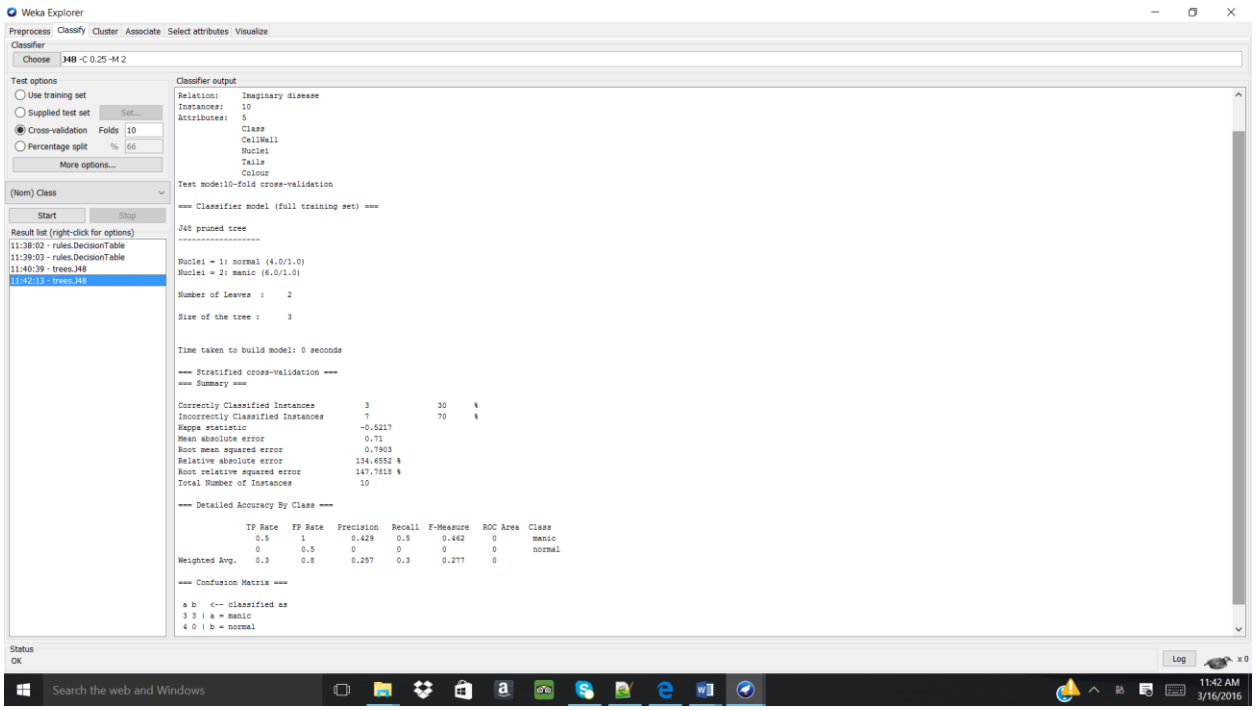
C. Selecting weka.classifiers.j48.J48.
Training set evaluation:

Orysya Stus

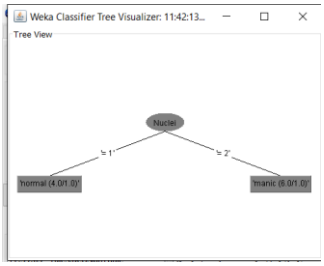




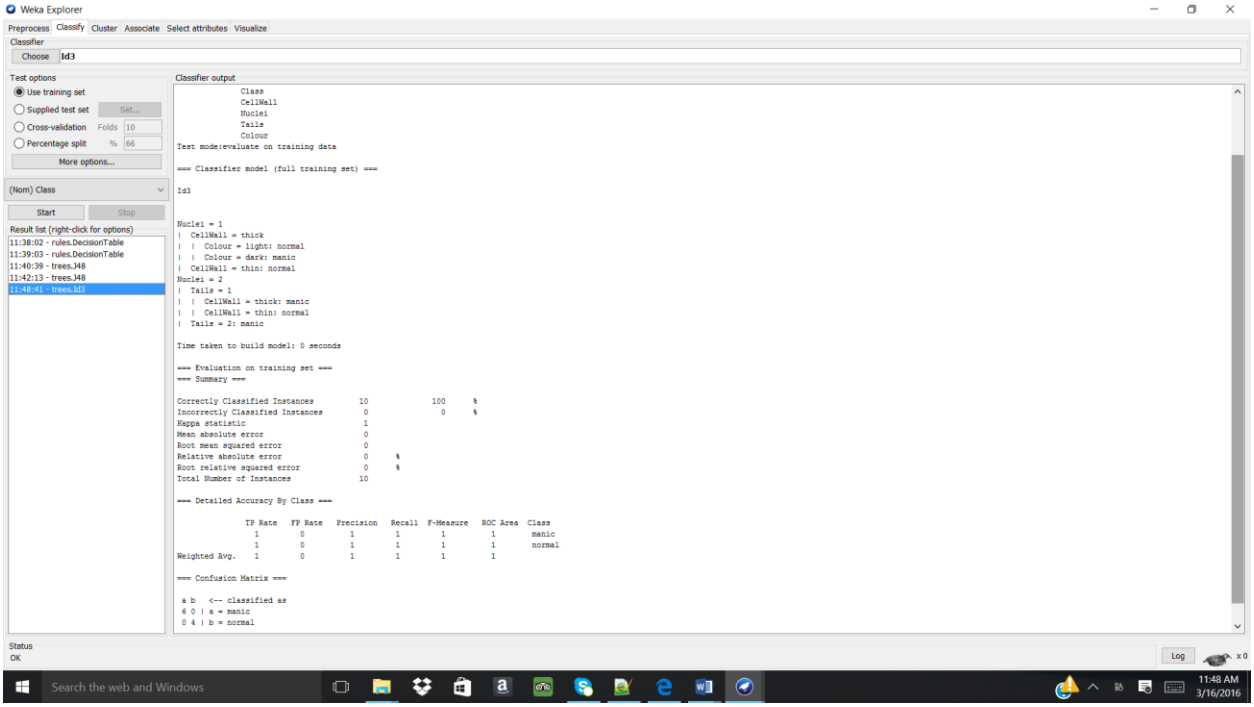
Cross-validation 10 fold evaluation:



Orysya Stus

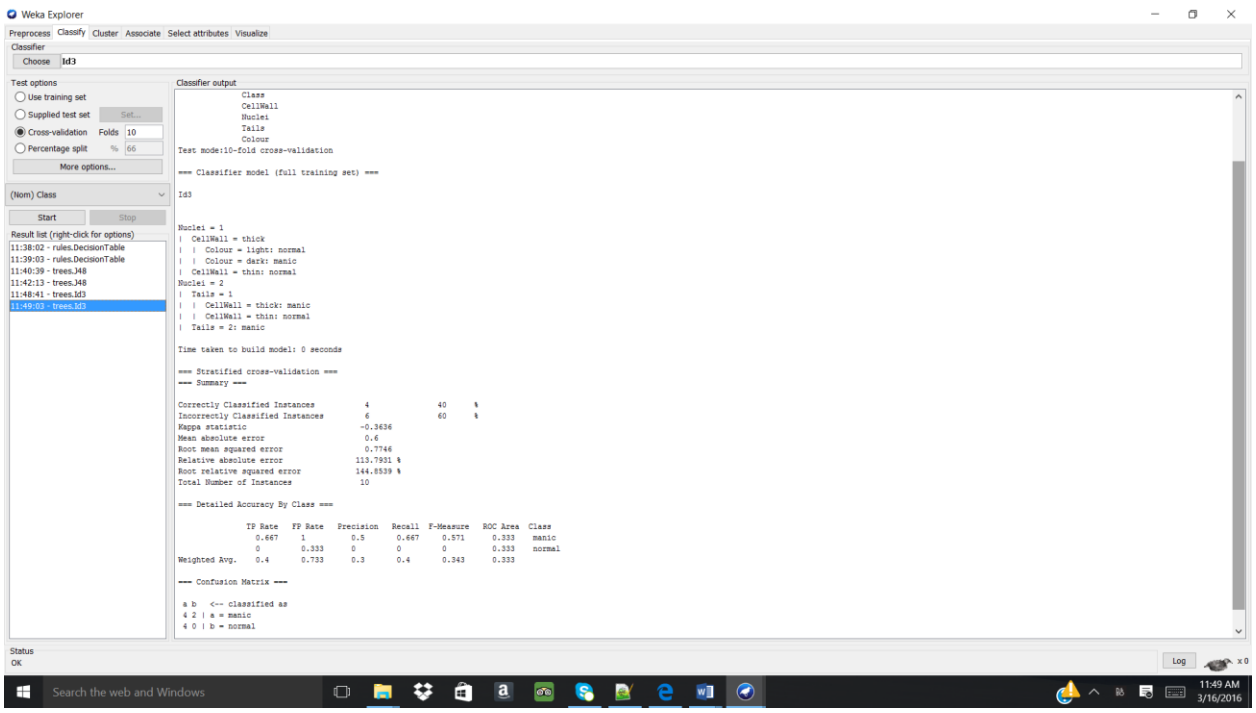


D. Selecting weka.classifier.Id3.
Training set evaluation:



Cross-validation 10 fold evaluation:

Orysya Stus



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

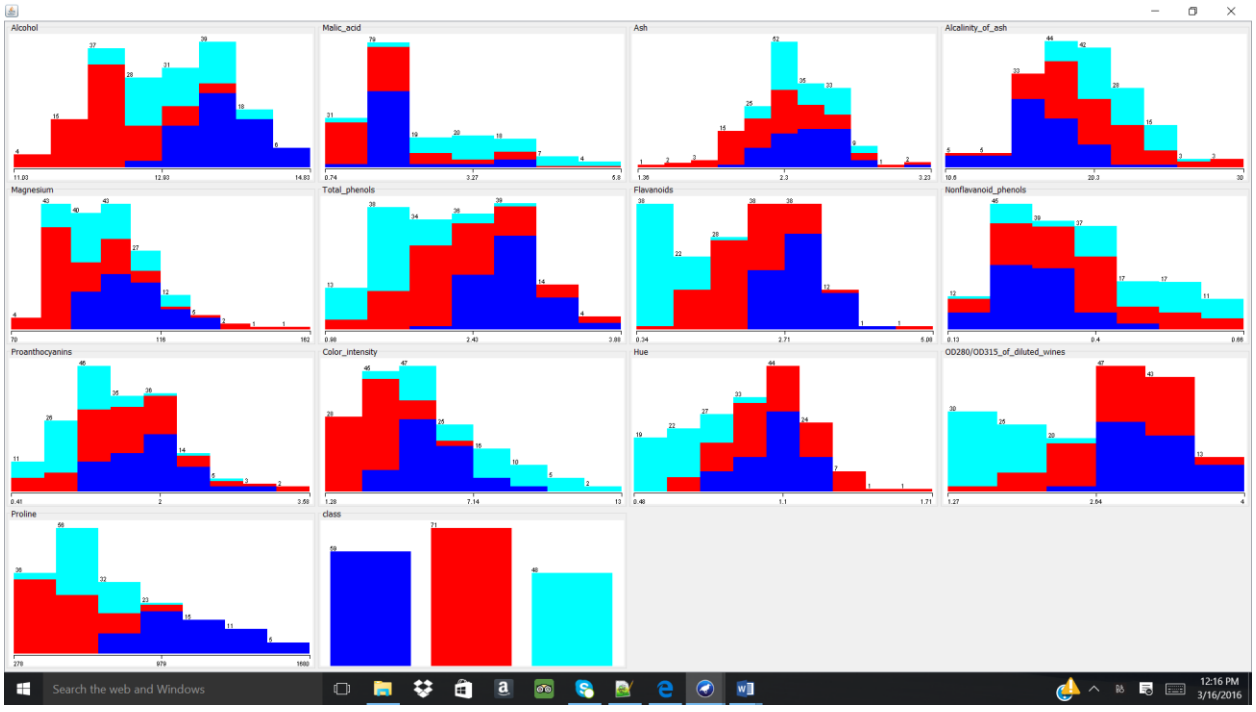
			selects it.	
Id3 10-fold cross validation	4	0.6	Recursive decision tree classifier which determines the attribute with the minimum entropy and selects it.	Data is split into training and test data sets 10 times with models being created, iterated, and polished between each training and testing

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 in very small, only 10 instances, and therefore each instances 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).

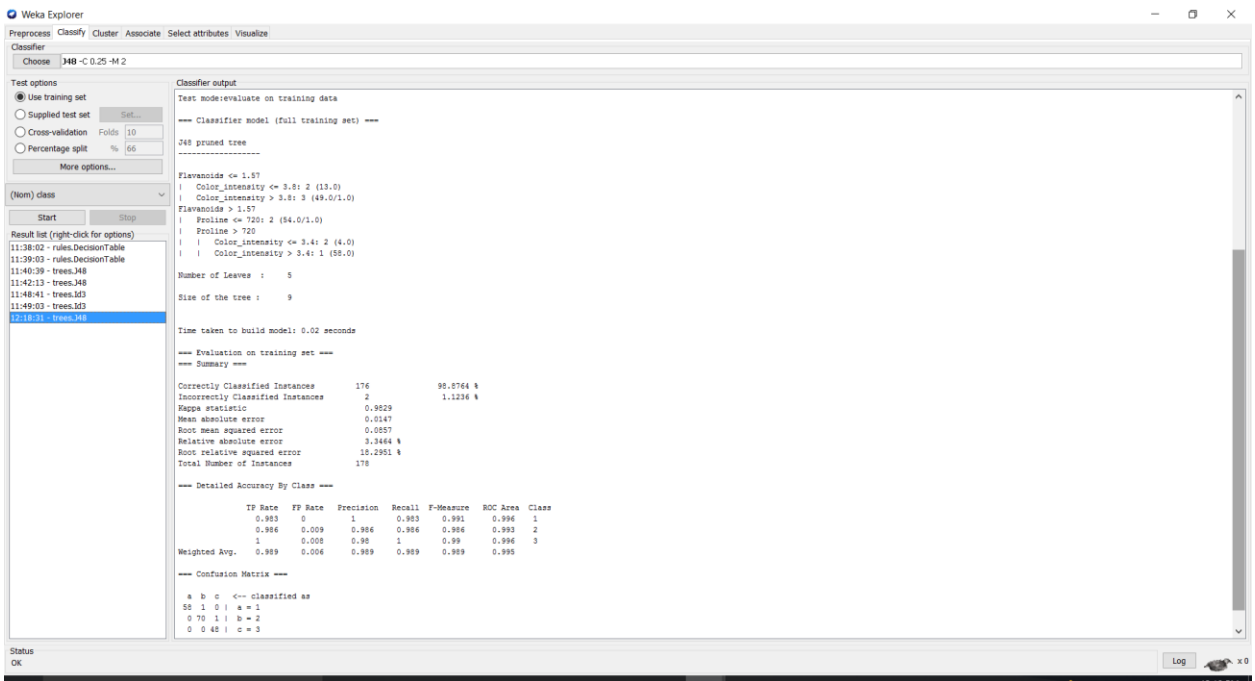
C4.5 - weka.classifiers.j48.J48
Decision List - weka. classifiers.PART

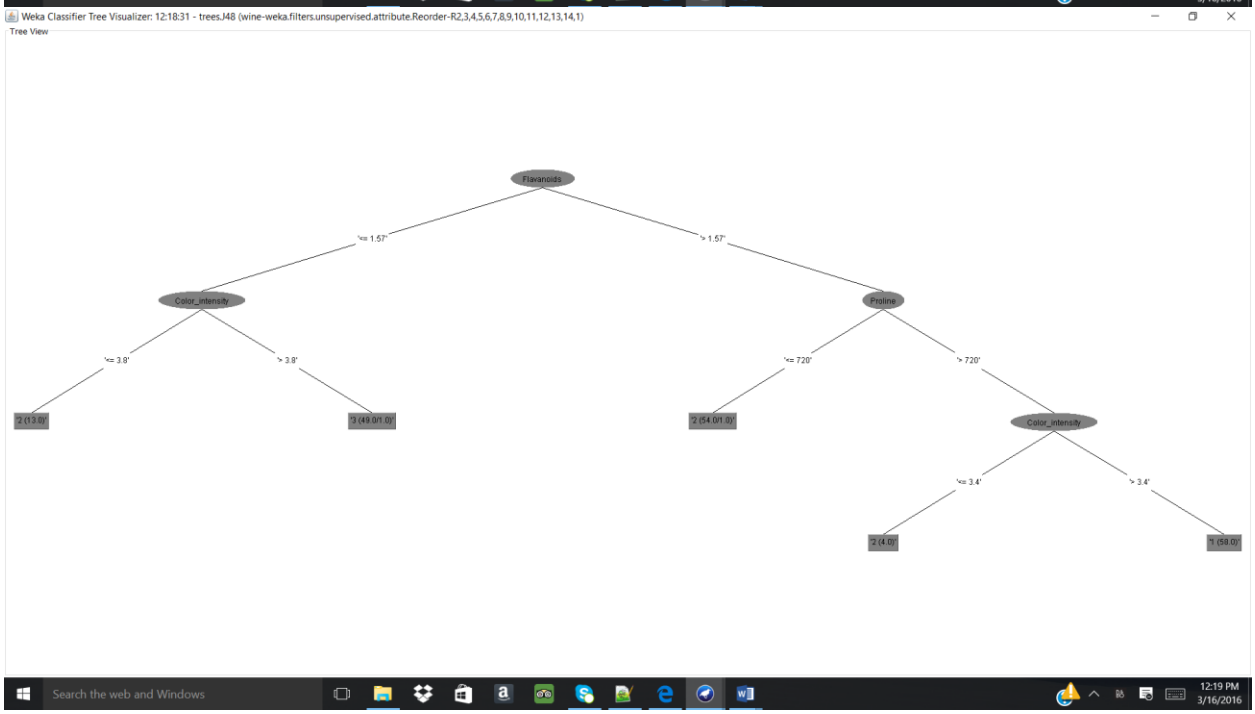
A. The wine.arff dataset was opened in Weka and "Visualize all" was used to see the instance distribution for each attribute. Note: all attribute values are numeric except the "class" attribute which is nominal.



B. Selecting weka.classifiers.j48.J48.
Training set evaluation:

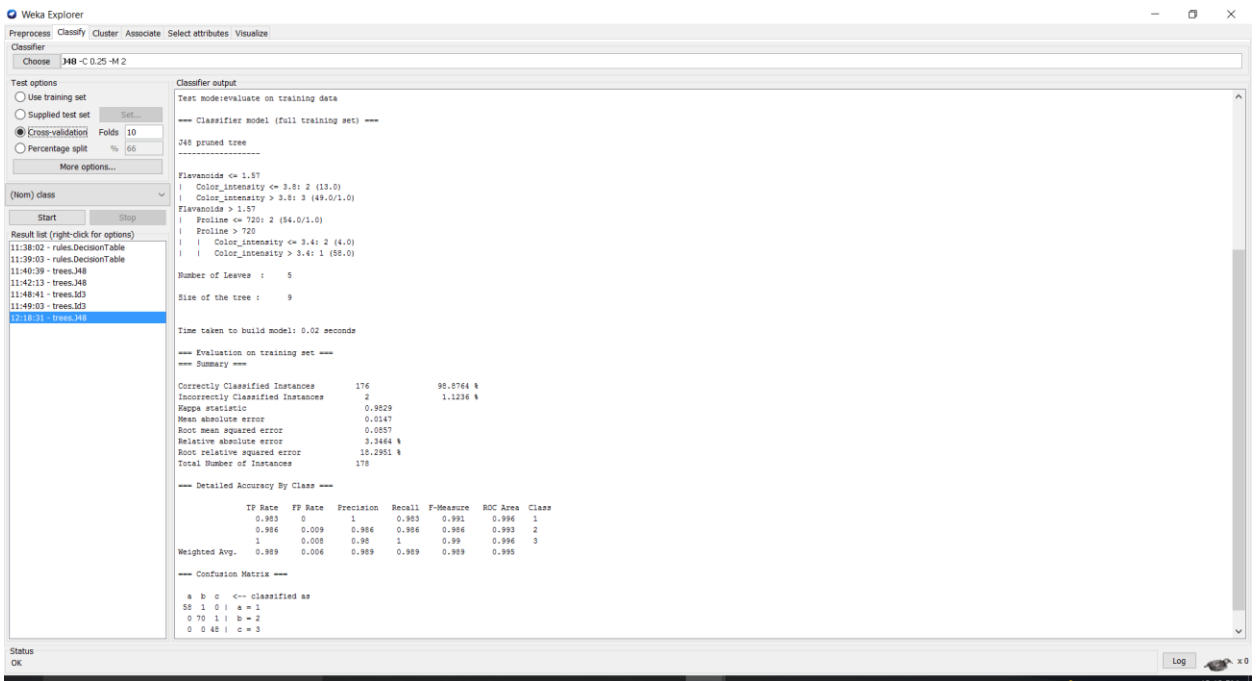
Orysya Stus

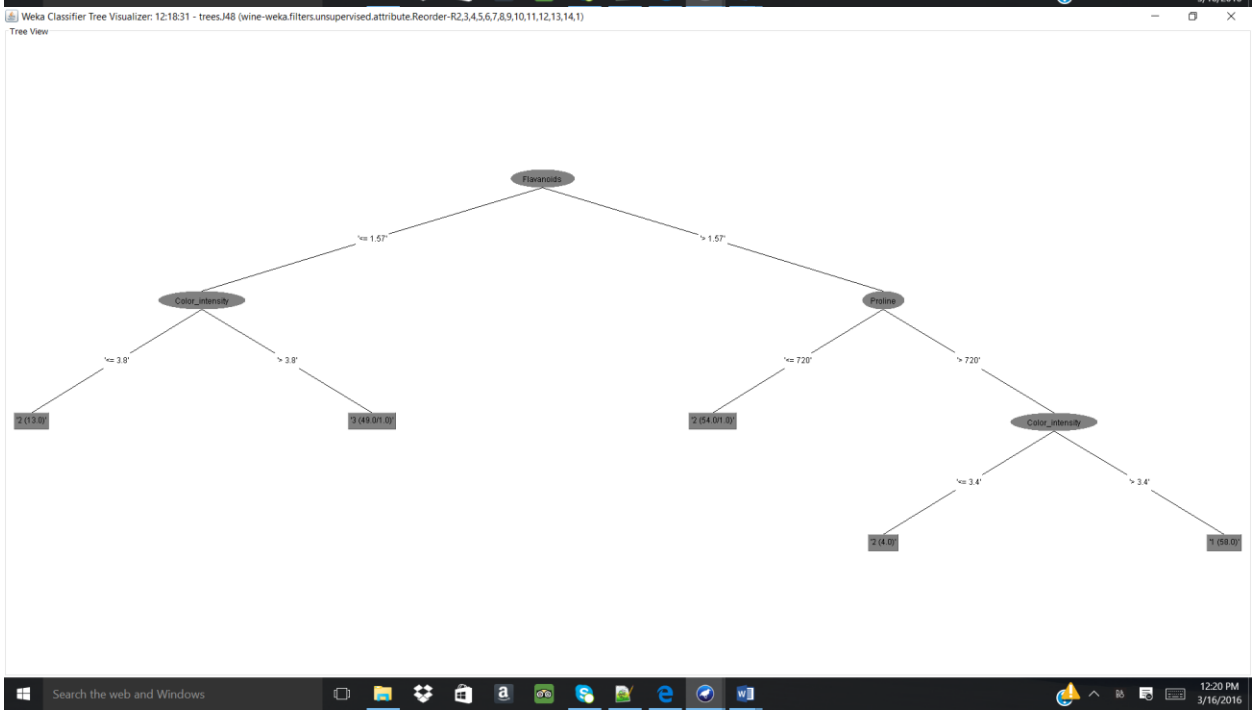




Cross-validation 10 fold evaluation:

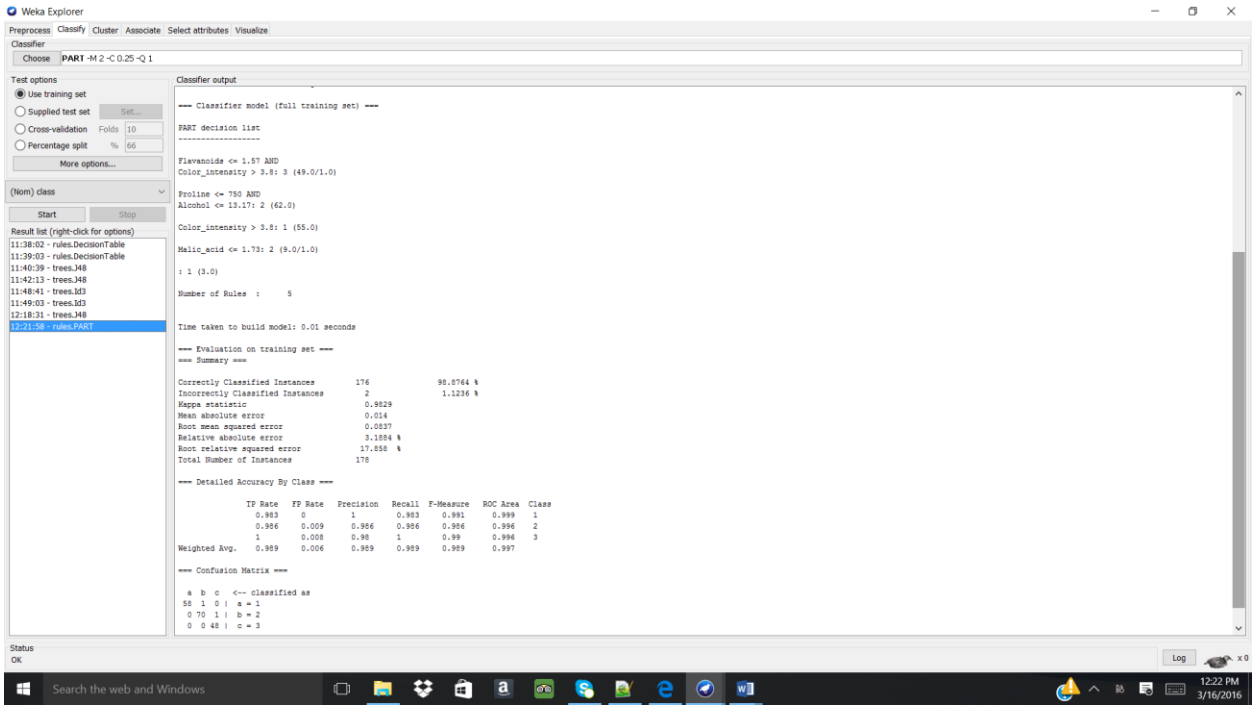
Orysya Stus



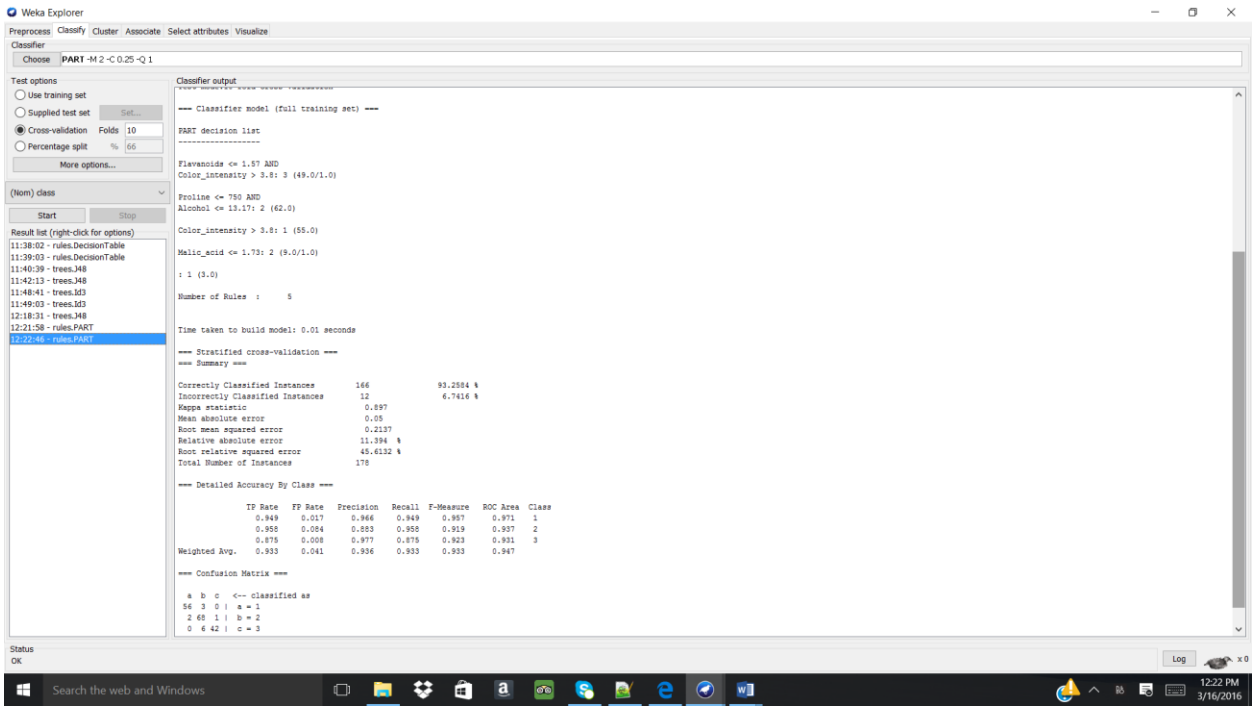


C. Selecting weka.classifiers.PART.
Training set evaluation:

Orysya Stus



+



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

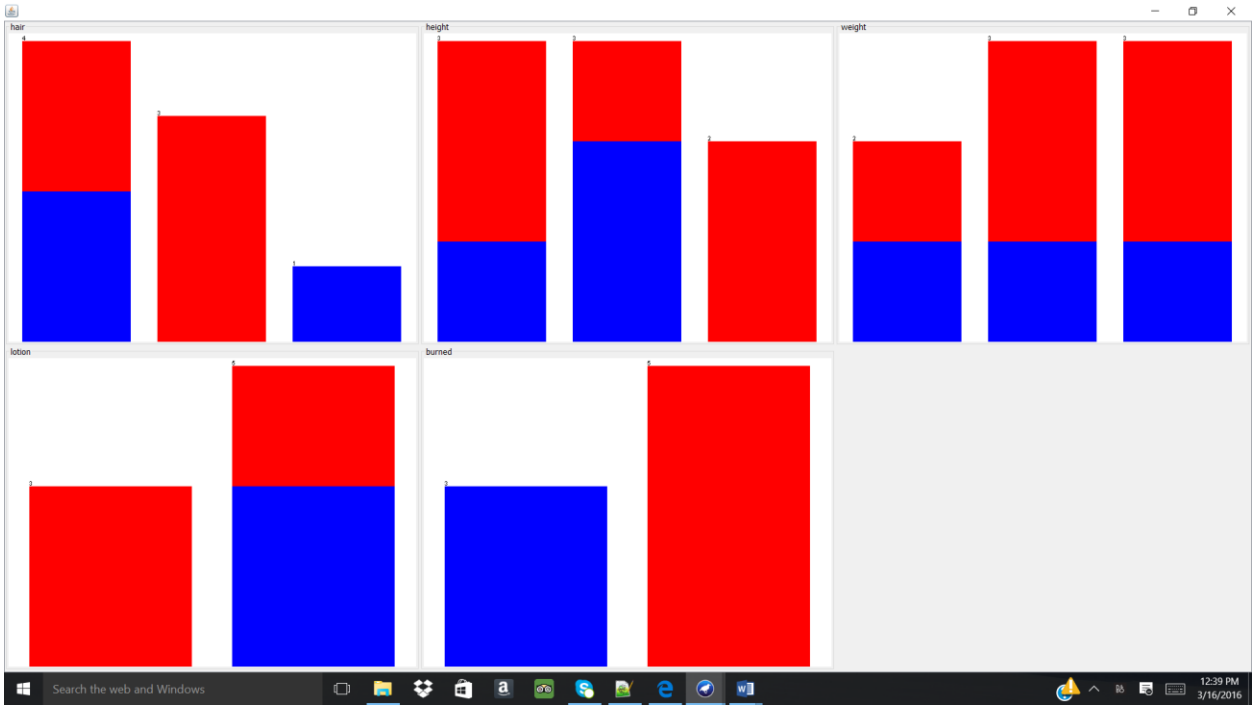
The attribute "Flavanoids" is the most important descriptor in the win.arff dataset as it provides the greatest information gain in classifying the types of "class". Using the absolute mean error to quantify the success of the learning schemes, it was found that the J48 full training set performed the 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.

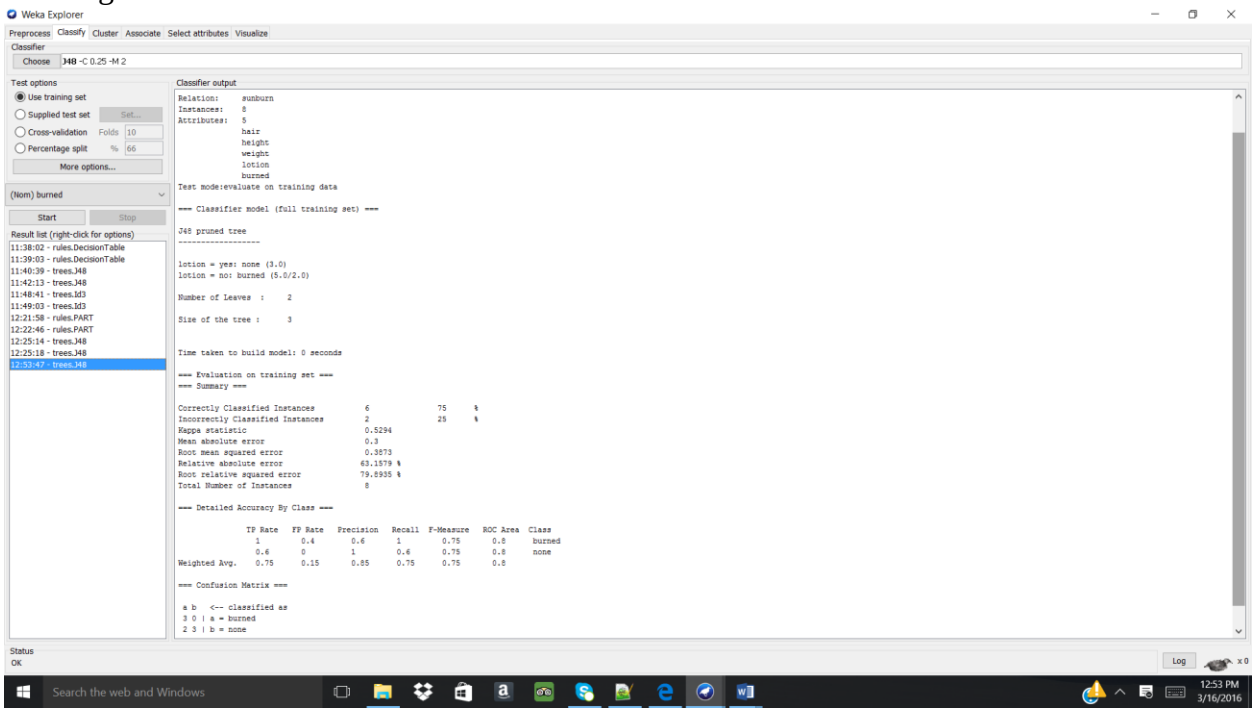
Answer the same questions as in A)-E) in the question #2.

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.

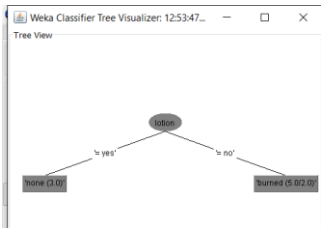
Orysya Stus



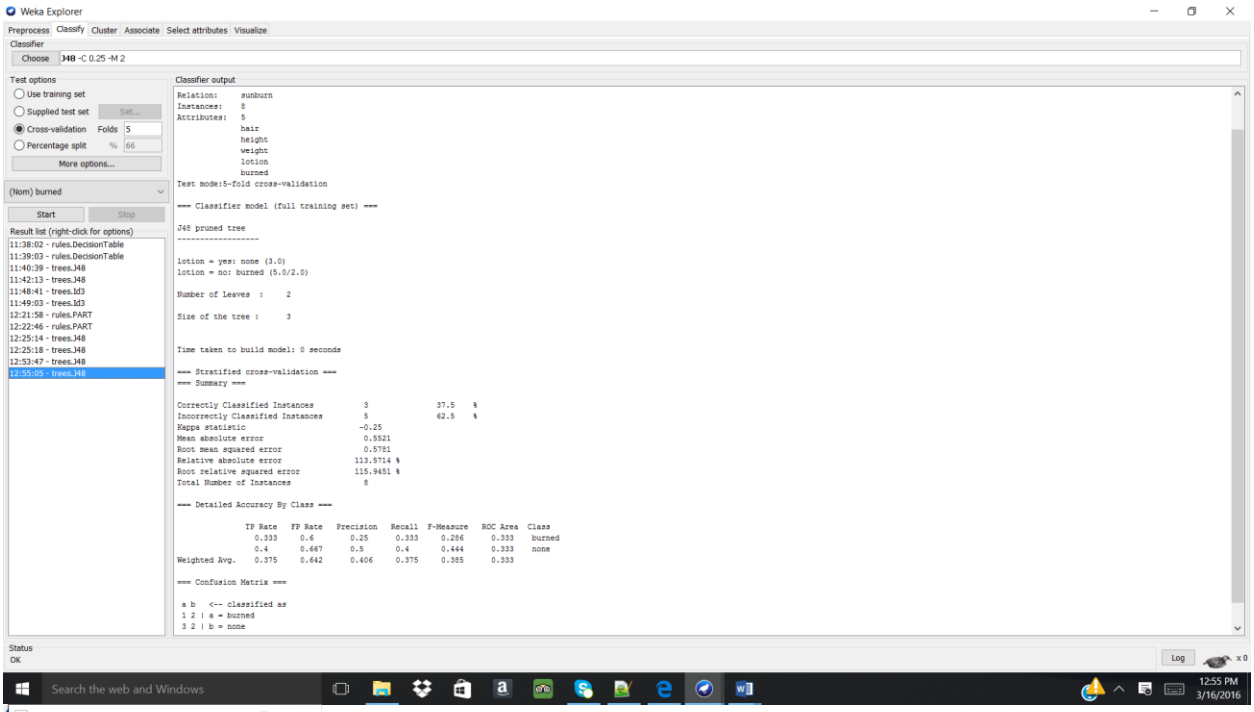
B. Selecting weka.classifiers.j48.J48.
Training set evaluation:

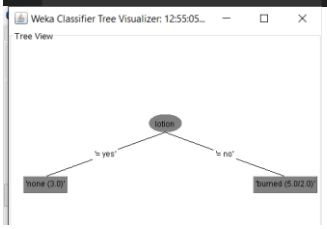


Orysya Stus



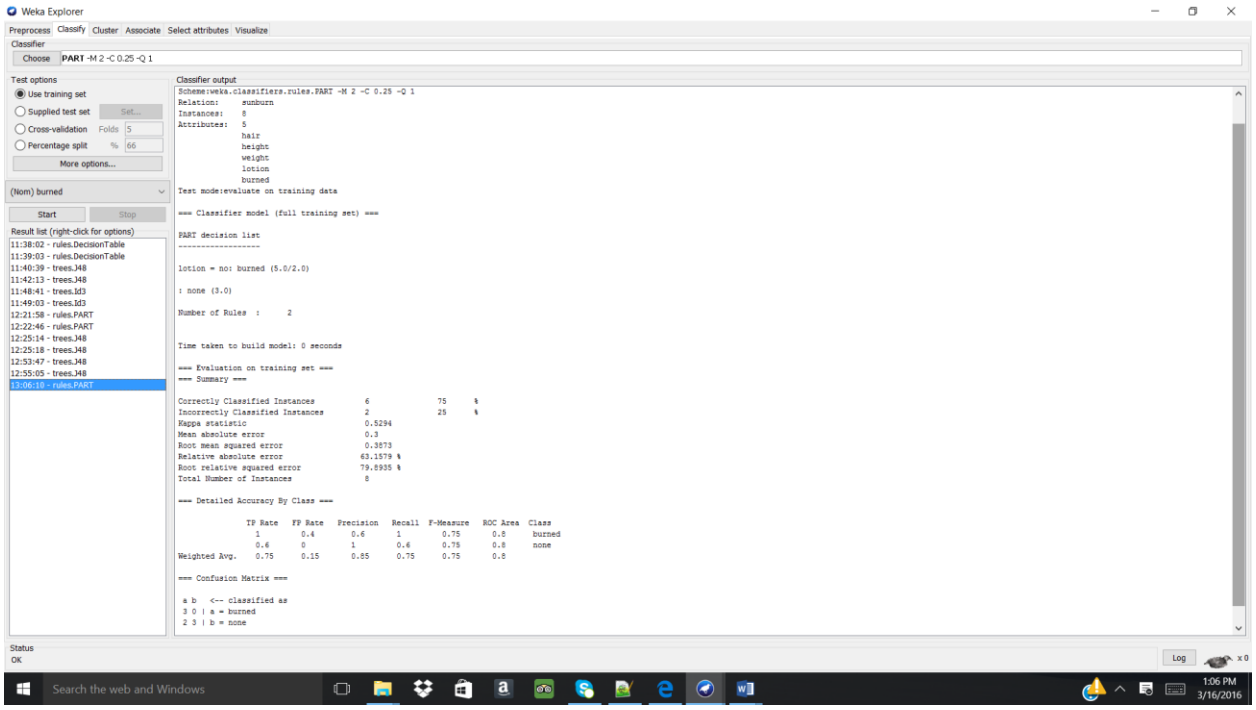
Cross-validation 5 fold evaluation:



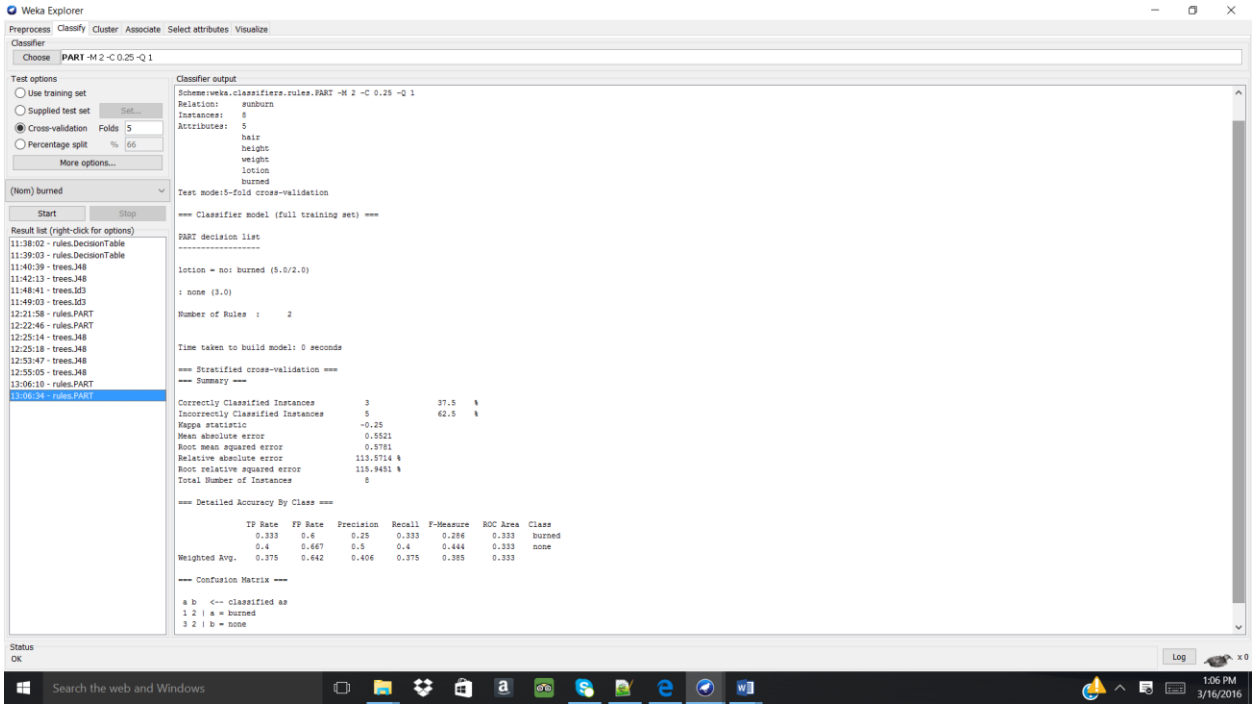


C. Selecting weka.classifiers.PART.
Training set evaluation:

Orysya Stus



Cross-validation 5 fold evaluation:



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

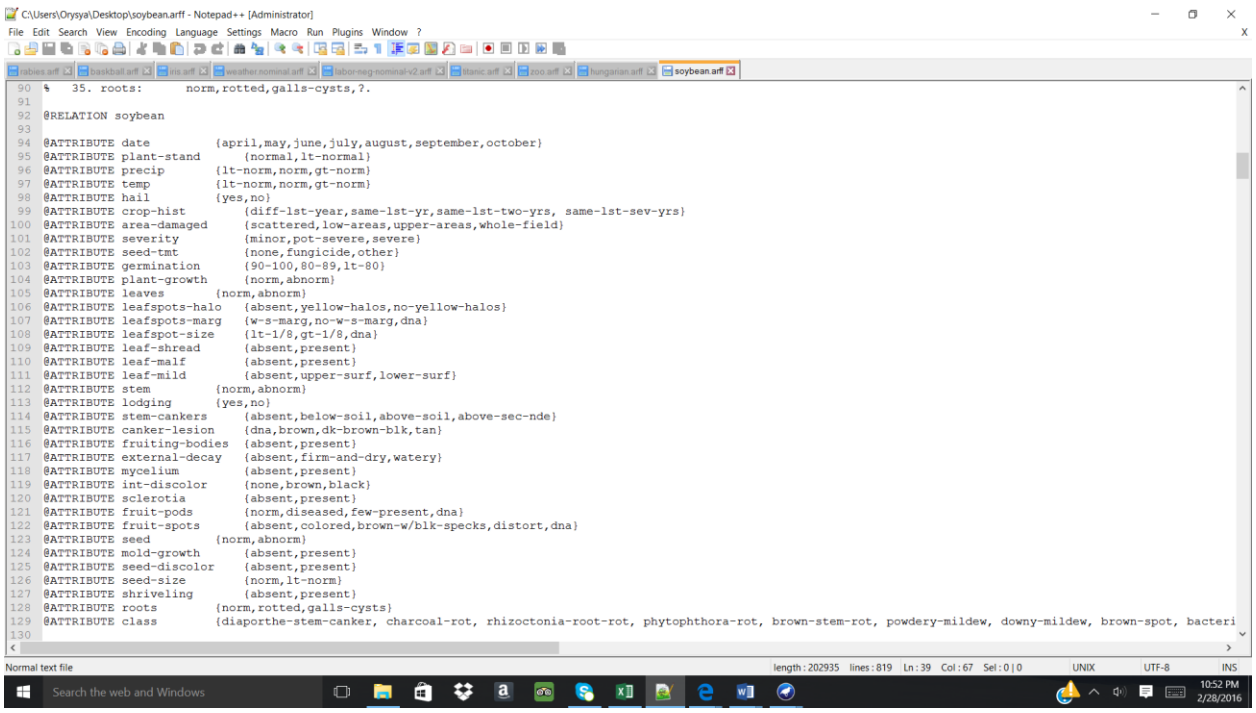
The attribute "lotion" is the most important descriptor in the sunburn.arff dataset as it provides the greatest information gain in classifying the types of "class". Using the absolute mean error to quantify the success of the learning schemes, it was found that both the J48 and PART full training set 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



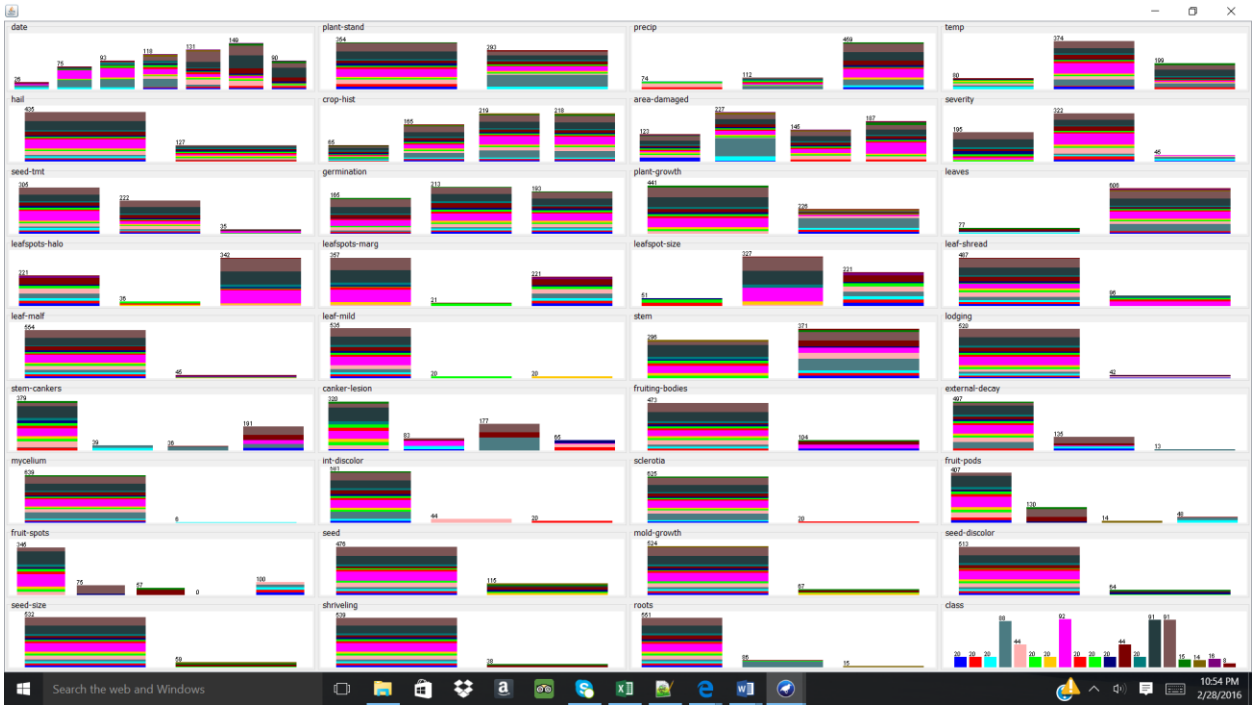
This dataset contains a multitude of attributes which will be used to classify the class of soybean.

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:

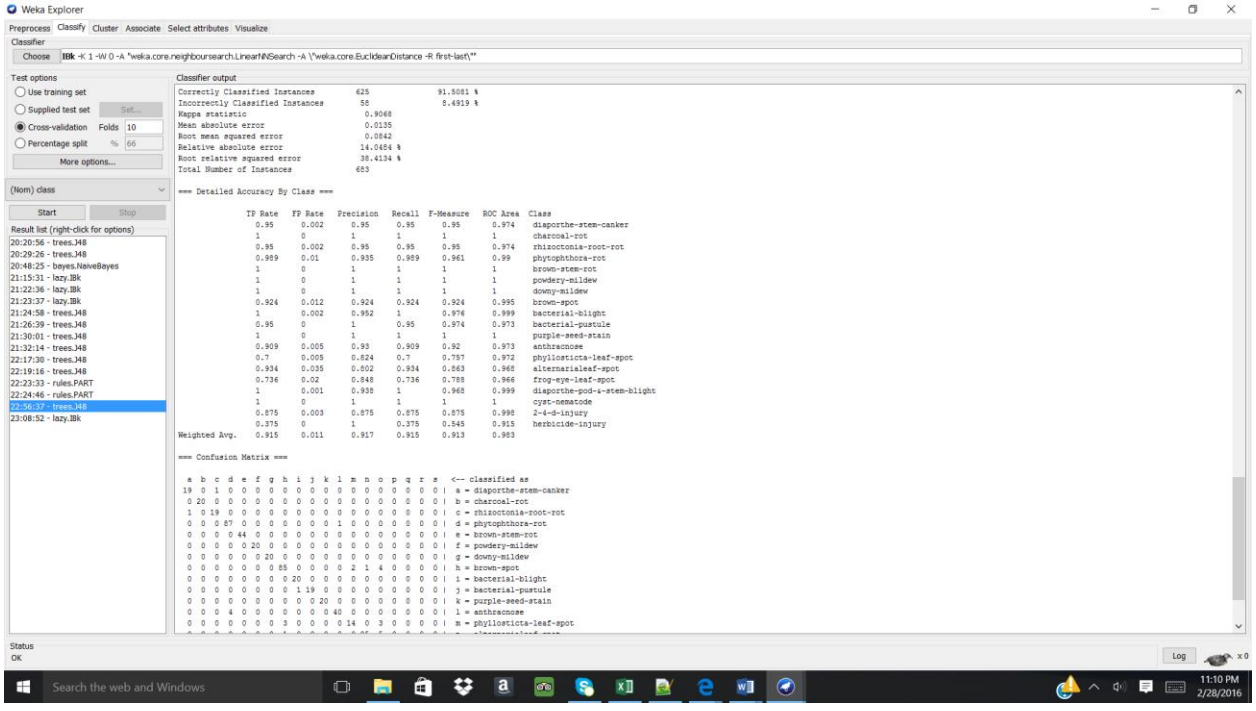
(1) C4.5 10-fold cross validation -weka.classifiers.j48.J48 (-M2 was used)
(2) IBK 10-fold cross validation with k=1

Orysya Stus

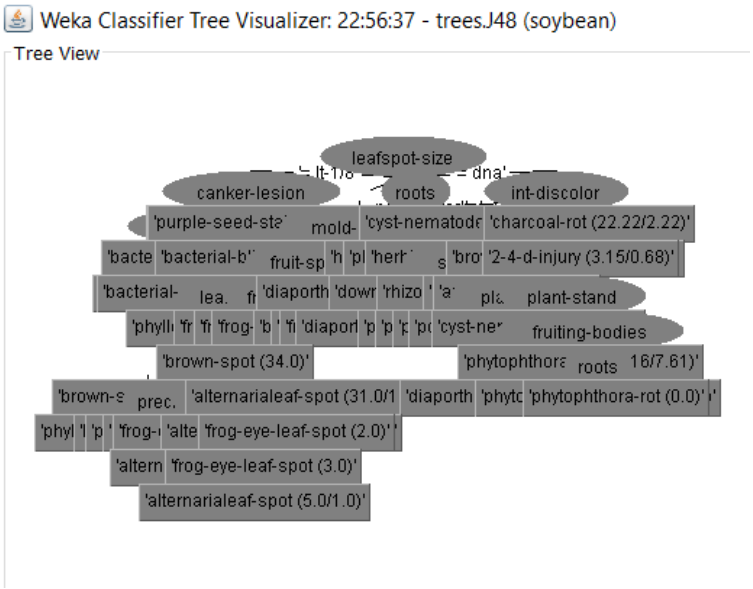


C. The following learning schemes were applied:

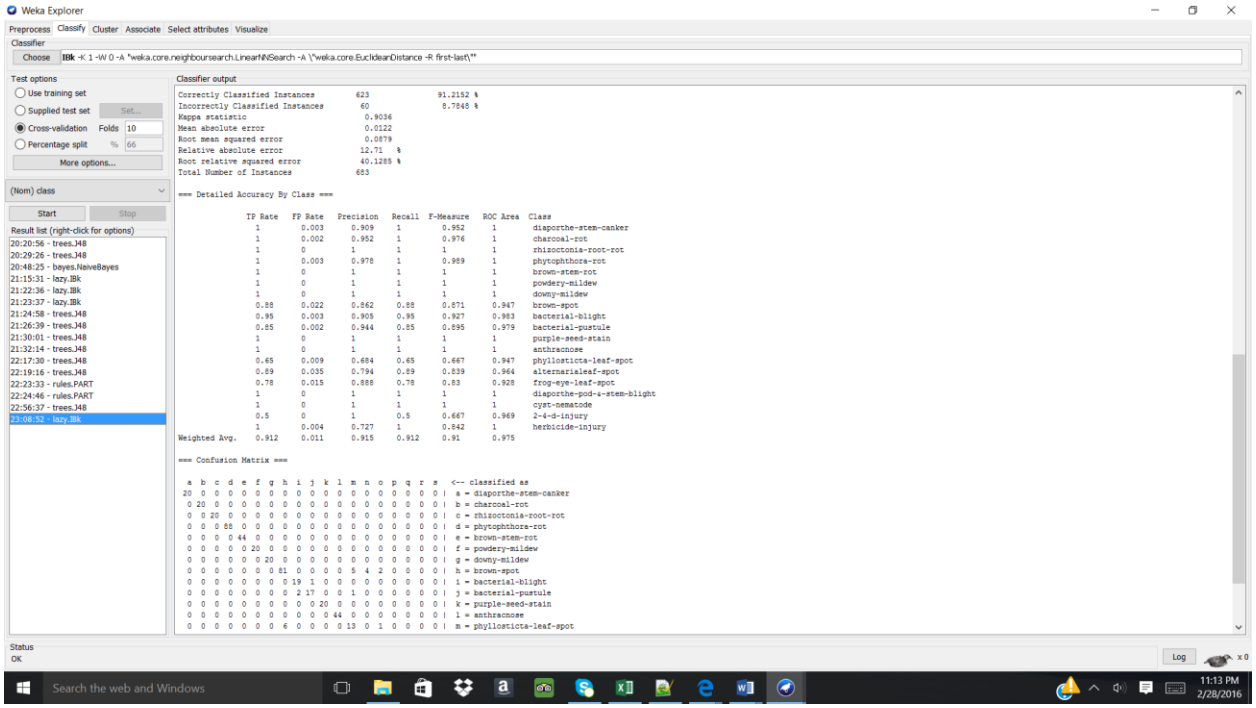
(1) C4.5 10-fold cross validation -weka.classifiers.j48.J48 (-M2 was used)



Orysya Stus



(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 aslightly 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.