EXERCISE 3: CLASSIFICATION PROCESS USING WEKATOOL

a) Classification on Iris.arff dataset using Naive Bayes Classifier Algorithm

The streamlined steps to classify the Iris dataset using the Naive Bayes Classifier Algorithm in Weka:

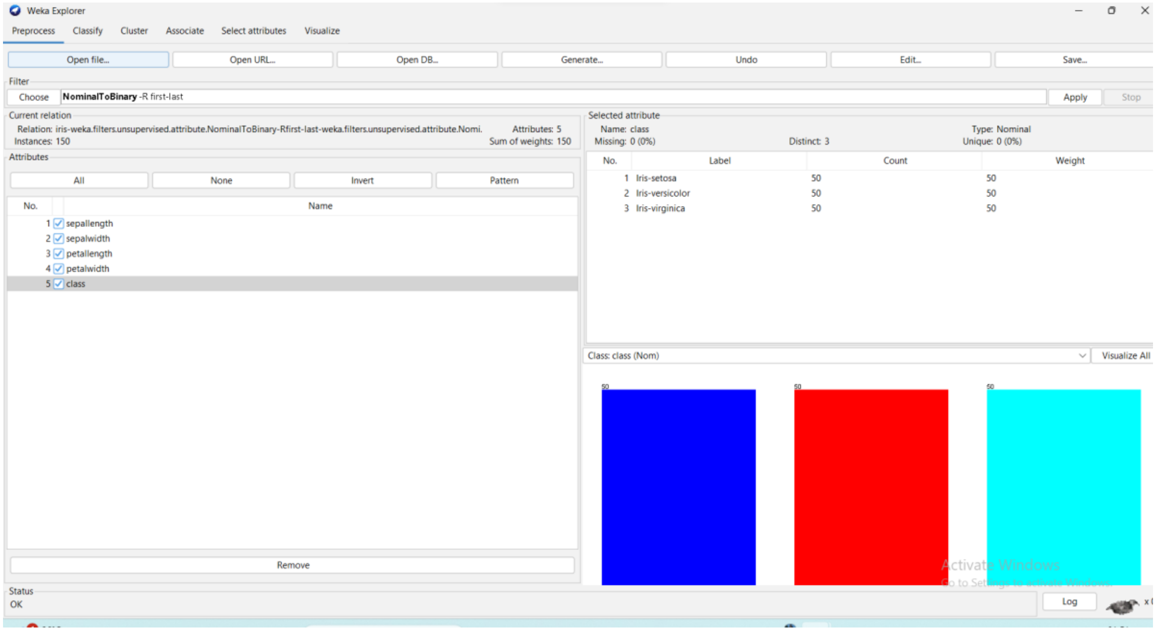
Setup:

• Launch Weka.

• Load Iris.arff via the "Explorer" > "Open file...".

• Review dataset in "Preprocess" tab.

• Apply preprocessing if needed (e.g., filters for missing values or nominal to numeric conversion).



Filter: **NominalToBinary** -R first-last

Current relation
Relation: iris-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last-weka.filters.unsupervised.attribute.Nomi.
Instances: 150
Attributes: 5
Sum of weights: 150

No.		Name
1	☑	sepallength
2	☑	sepalwidth
3	☑	petallength
4	☑	petalwidth
5	☑	class

Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Iris-setosa	50	50
2	Iris-versicolor	50	50
3	Iris-virginica	50	50

Class: class (Nom)

Status
OK

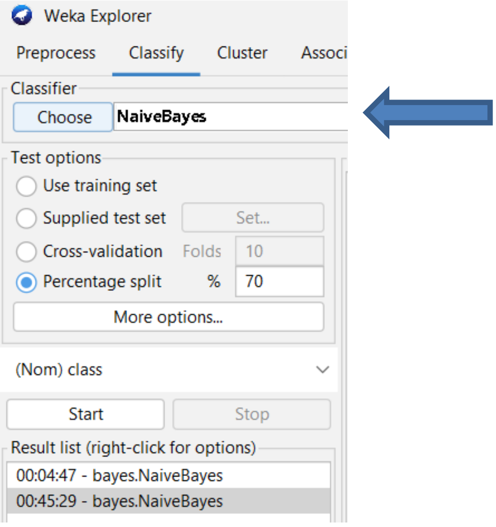
Classification:

• Go to "Classify" tab.

• Select Naive Bayes under "Choose".

• Configure classifier options as needed.

• Use cross-validation (10 folds) or percentage split (70% training, 30% testing) for dataset split.



NaiveBayes

Test options

- ○ Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ● Percentage split % 70

(Nom) class

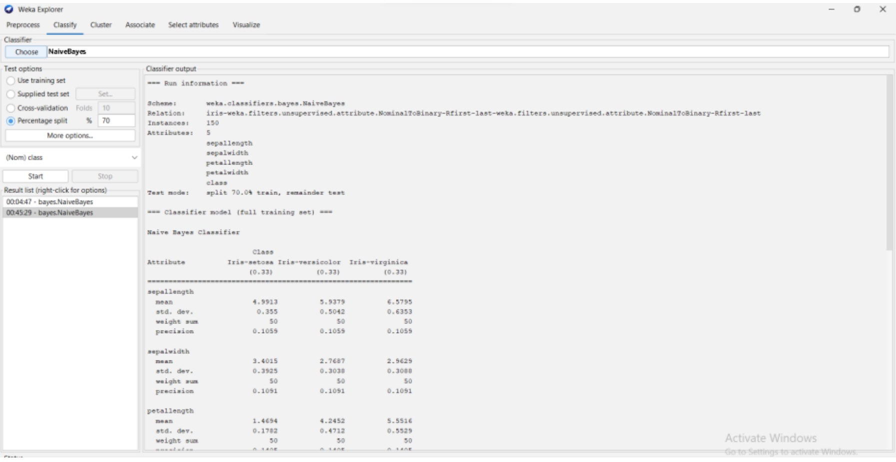
Result list (right-click for options)

- 00:04:47 - bayes.NaiveBayes
- 00:45:29 - bayes.NaiveBayes

• Start the classification process.

Evaluation:

• Observe output for accuracy, precision, recall, and F1-score.



NaiveBayes

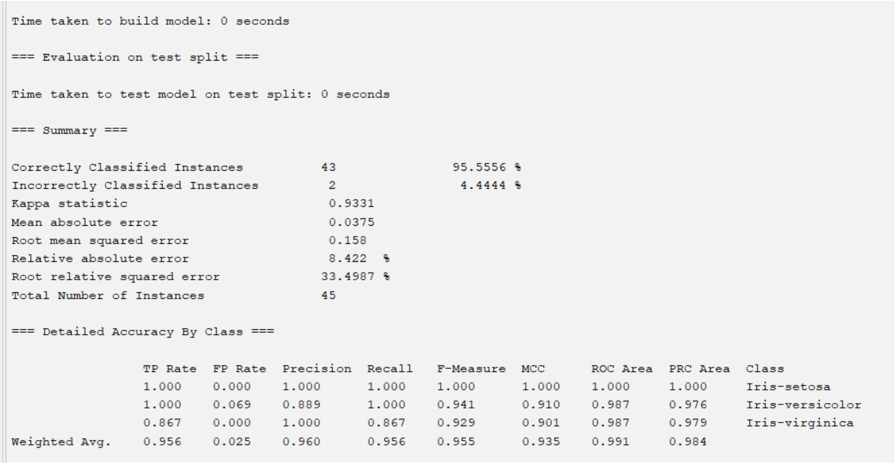
Test options: ● Percentage split % 70; (Nom) class

Result list (right-click for options)

- 00:04:47 - bayes.NaiveBayes
- 00:45:29 - bayes.NaiveBayes

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     iris-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 70.0% train, remainder test

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

Naive Bayes Classifier

                       Class
Attribute        Iris-setosa Iris-versicolor  Iris-virginica
                      (0.33)          (0.33)          (0.33)
=============================================================
sepallength
  mean                4.9913          5.9379          6.5795
  std. dev.            0.355          0.5042          0.6353
  weight sum              50              50              50
  precision           0.1059          0.1059          0.1059

sepalwidth
  mean                3.4015          2.7687          2.9629
  std. dev.           0.3925          0.3038          0.3088
  weight sum              50              50              50
  precision           0.1091          0.1091          0.1091

petallength
  mean                1.4694          4.2452          5.5516
  std. dev.           0.1782          0.4712          0.5529
  weight sum              50              50              50
```

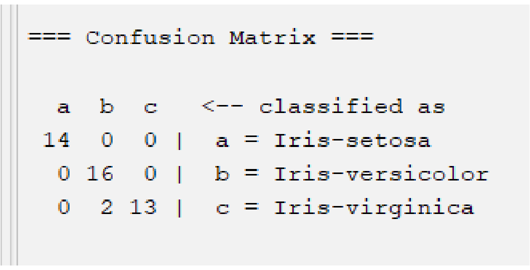


```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          43               95.5556 %
Incorrectly Classified Instances         2                4.4444 %
Kappa statistic                          0.9331
Mean absolute error                      0.0375
Root mean squared error                  0.158
Relative absolute error                  8.422  %
Root relative squared error             33.4987 %
Total Number of Instances               45

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 1.000    0.069    0.889      1.000    0.941      0.910    0.987     0.976     Iris-versicolor
                 0.867    0.000    1.000      0.867    0.929      0.901    0.987     0.979     Iris-virginica
Weighted Avg.    0.956    0.025    0.960      0.956    0.955      0.935    0.991     0.984
```

• Weka provides these metrics automatically post-classification.

Interpretation and Reporting:

- Confusion Matrix: Directly visible in the output.
- ROC Curve: Right-click on result > "Visualize" > "Threshold curve" > select "ROC".
- Precision-Recall Curve: Similarly, through "Visualize" > "Threshold curve" > select "Precision/Recall".
- Interpret the confusion matrix for accuracy per class and overall.
- Assess ROC and Precision-Recall curves for performance insights.



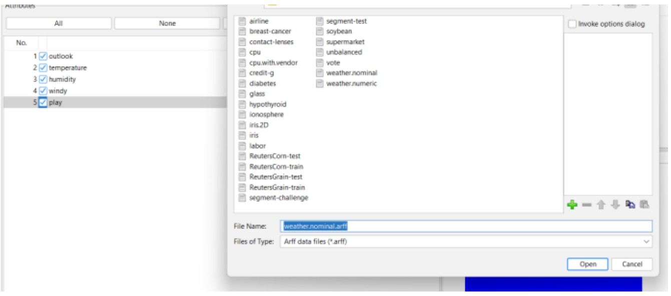
```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 14  0  0 |  a = Iris-setosa
  0 16  0 |  b = Iris-versicolor
  0  2 13 |  c = Iris-virginica
```

b) Classification on weather.nominal.arff dataset using J48 Tree Classifier Algorithm :

Setup:

Open Weka.

Load the weather.nominal.arff dataset through "Explorer" > "Open file...".

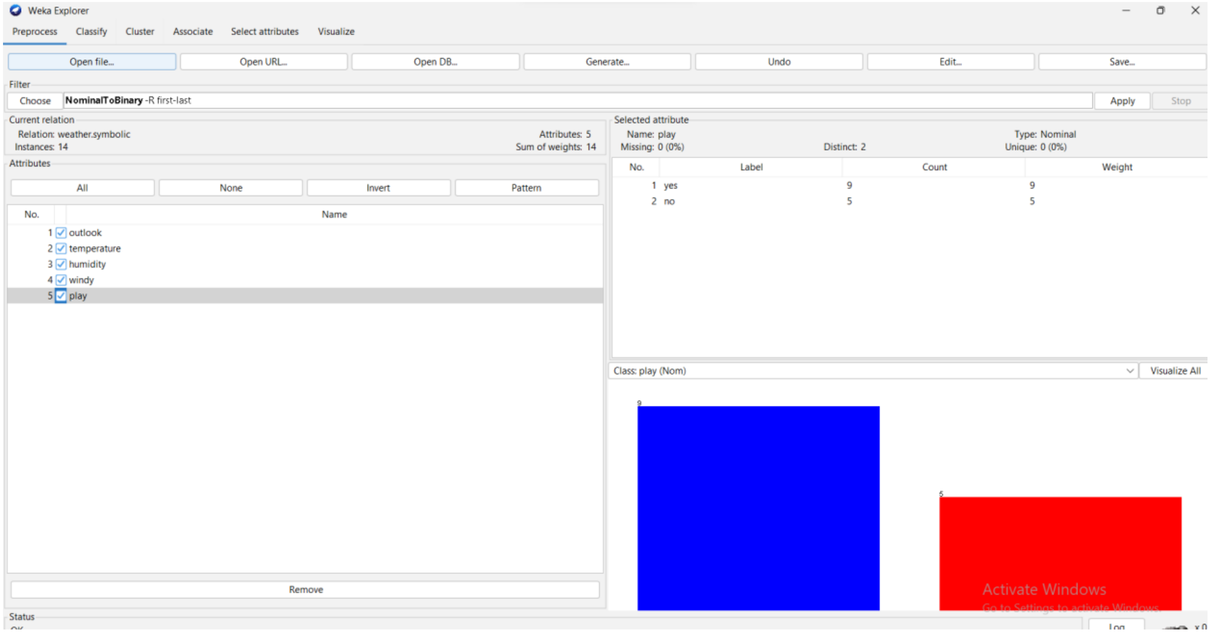


File Name: weather.nominal.arff

Files of Type: Arff data files (*.arff)

Inspect dataset attributes and structure under the "Preprocess" tab.

Apply any necessary preprocessing, such as handling missing values or converting nominal attributes to numeric,



Filter: Choose **NominalToBinary** -R first-last (Apply)

Current relation
Relation: weather.symbolic Attributes: 5
Instances: 14 Sum of weights: 14

Attributes

No.	Name
1	☑ outlook
2	☑ temperature
3	☑ humidity
4	☑ windy
5	☑ play

Selected attribute
Name: play Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	9	9
2	no	5	5

Class: play (Nom)

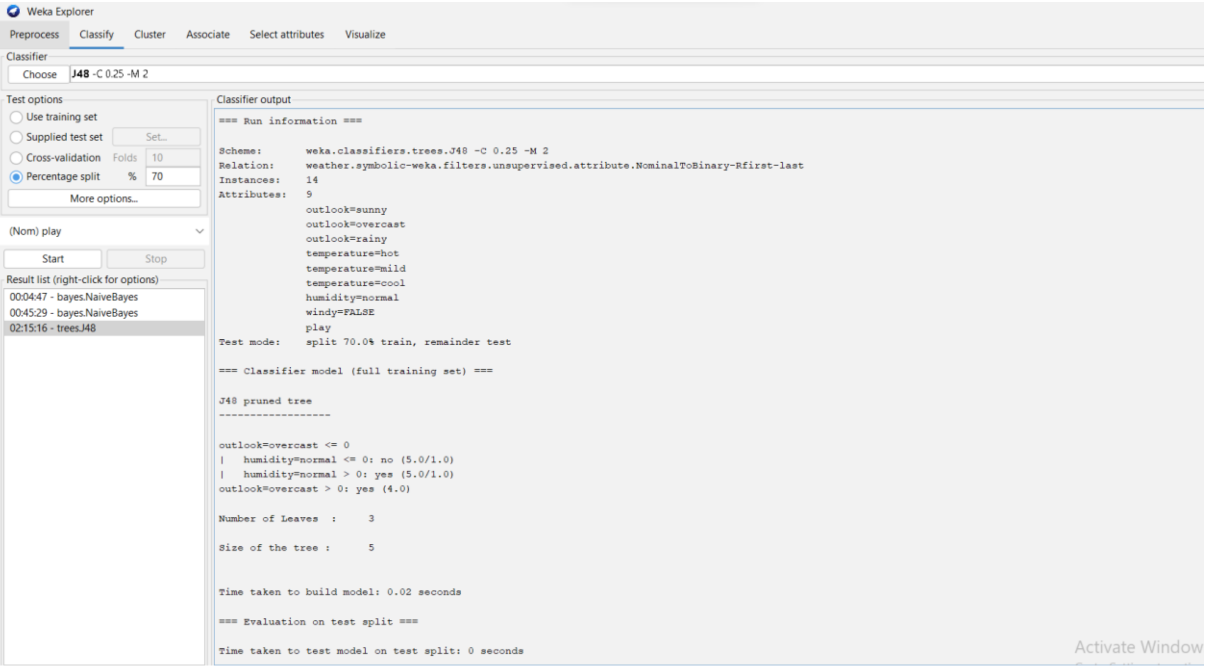
Though for J48, nominal to numeric conversion is not typically necessary.

Classification:

Navigate to the "Classify" tab.

Select the J48 classifier from the list under "trees".

(Optional) Adjust classifier options as required by clicking on the J48 name.



Classifier: Choose **J48** -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 70

(Nom) play

Result list (right-click for options)
00:04:47 - bayes.NaiveBayes
00:45:29 - bayes.NaiveBayes
02:15:16 - trees.J48

Classifier output

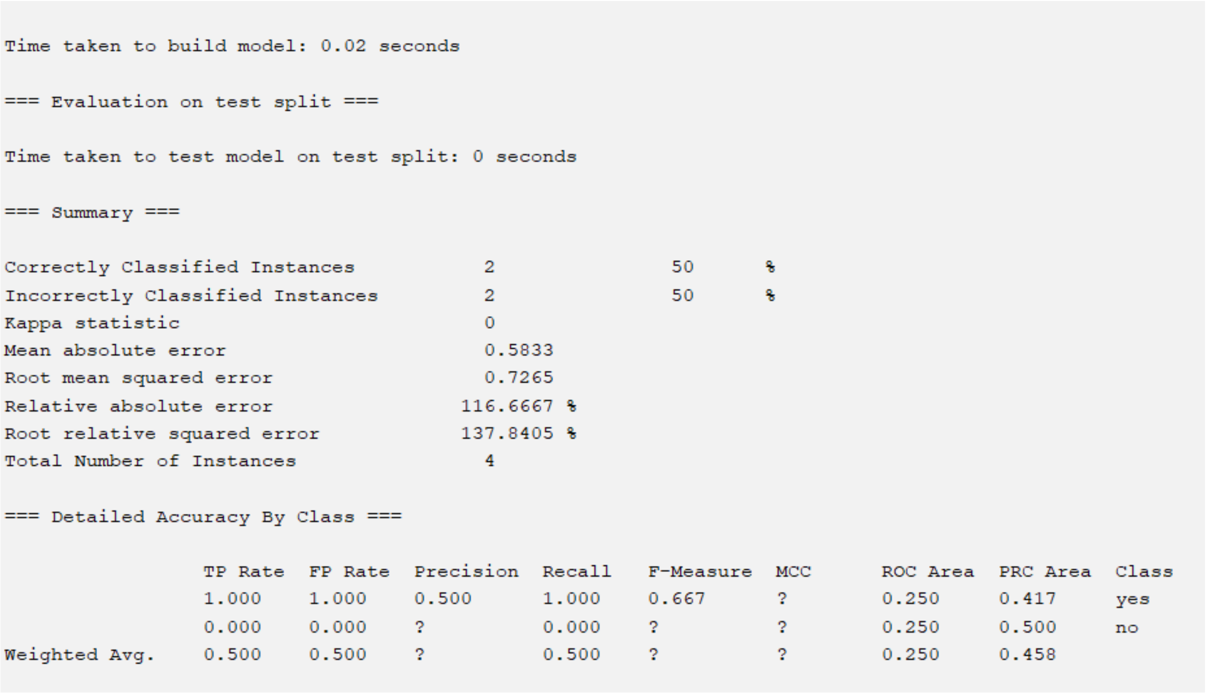
```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather.symbolic-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last
Instances:    14
Attributes:   9
              outlook=sunny
              outlook=overcast
              outlook=rainy
              temperature=hot
              temperature=mild
              temperature=cool
              humidity=normal
              windy=FALSE
              play
Test mode:    split 70.0% train, remainder test

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

J48 pruned tree
------------------

outlook=overcast <= 0
|   humidity=normal <= 0: no (5.0/1.0)
|   humidity=normal > 0: yes (5.0/1.0)
outlook=overcast > 0: yes (4.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds
```

For splitting the dataset, choose either "Cross-validation" with a standard 10 folds or "Percentage split" (e.g., 70% for training).

Click "Start" to run the J48 classifier on the dataset.

Evaluation:

After classification, review the output for accuracy, precision, recall, and F1-score, provided directly in the "Classifier output" section.



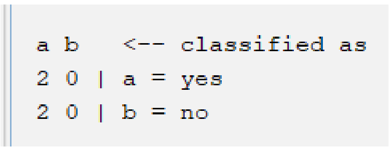
```
Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances           2               50      %
Incorrectly Classified Instances         2               50      %
Kappa statistic                          0
Mean absolute error                      0.5833
Root mean squared error                  0.7265
Relative absolute error                116.6667 %
Root relative squared error            137.8405 %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.500      1.000    0.667      ?        0.250     0.417     yes
                 0.000    0.000    ?          0.000    ?          ?        0.250     0.500     no
Weighted Avg.    0.500    0.500    ?          0.500    ?          ?        0.250     0.458
```

Weka automatically generates these metrics post-classification.

Interpretation and Reporting:

Confusion Matrix: Found in the classifier output, detailing true vs. predicted labels.

Decision Tree Visualization: Right-click on the result in the history list > "Visualize tree" to see the model's structure.



```
a b   <-- classified as
2 0 | a = yes
2 0 | b = no
```

Interpretation:

Examine the confusion matrix to evaluate accuracy and misclassifications among weather conditions.

Analyze the decision tree visualization for insights into the classifier's decision-making process.

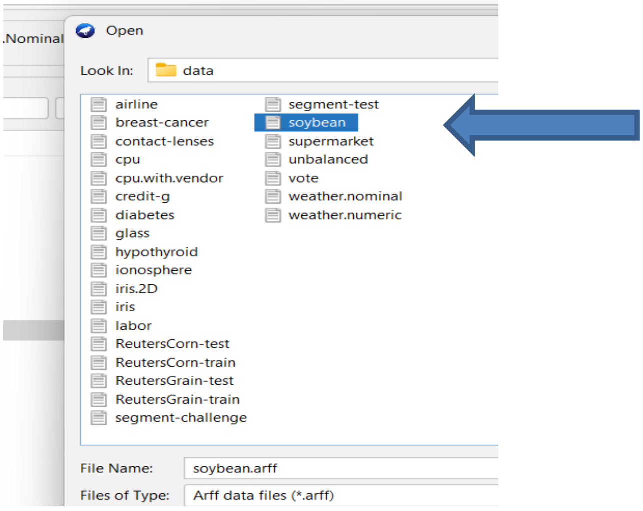
Highlight significant performance metrics in the classification summary.

c) Classification on soybean.arff dataset using RandomForest Tree Classifier:

Setup:

Open Weka on your computer.

Load the soybean.arff dataset through "Explorer" > "Open file...".

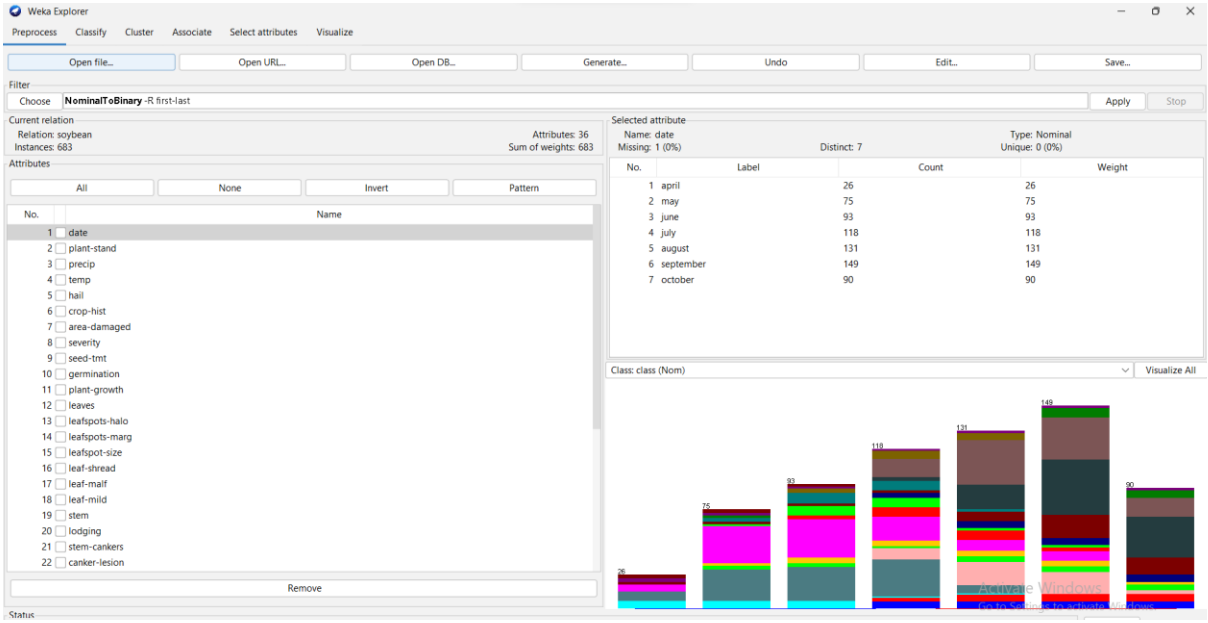


Look In: data

File Name: soybean.arff

Files of Type: Arff data files (*.arff)

Review the dataset's attributes and structure in the "Preprocess" tab.



Filter: NominalToBinary -R first-last

Current relation
Relation: soybean
Instances: 683
Attributes: 36
Sum of weights: 683

Selected attribute
Name: date
Missing: 1 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

Class: class (Nom)

Attributes

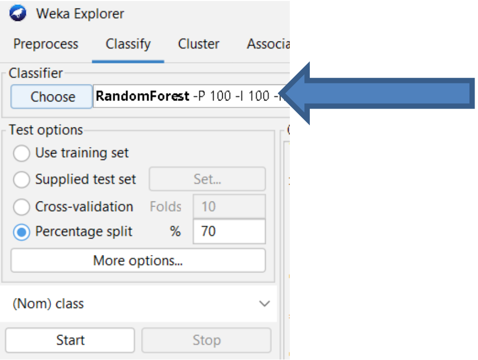
No.	Name
1	date
2	plant-stand
3	precip
4	temp
5	hail
6	crop-hist
7	area-damaged
8	severity
9	seed-tmt
10	germination
11	plant-growth
12	leaves
13	leafspots-halo
14	leafspots-marg
15	leafspot-size
16	leaf-shread
17	leaf-malf
18	leaf-mild
19	stem
20	lodging
21	stem-cankers
22	canker-lesion

Apply necessary preprocessing steps, such as using the ReplaceMissingValues filter if there are any missing values. RandomForest can handle nominal attributes directly, so conversion to numeric might not be required.

Classification:

Switch to the "Classify" tab.

Choose the RandomForest algorithm from the list under "trees" or "meta" depending on your Weka version.



RandomForest -P 100 -I 100 -

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 70

(Nom) class

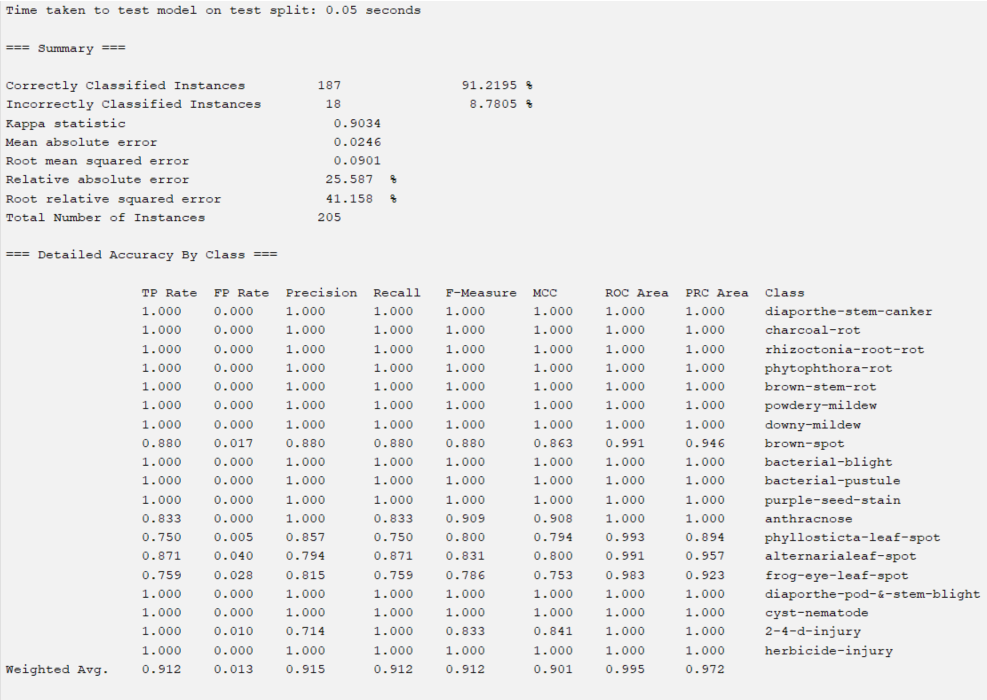
Configure any specific options for RandomForest by clicking on the algorithm name (though the default settings are often adequate).

Decide on the method for splitting the dataset: either use "Cross-validation" with a common choice of 10 folds or "Percentage split" (a typical split might be 70% for training and 30% for testing).

Initiate the classification by clicking "Start".

Evaluation:

Observe the output in the "Classifier output" section for accuracy, precision, recall, and F1-score.



```
Time taken to test model on test split: 0.05 seconds

=== Summary ===

Correctly Classified Instances         187               91.2195 %
Incorrectly Classified Instances        18                8.7805 %
Kappa statistic                          0.9034
Mean absolute error                      0.0246
Root mean squared error                  0.0901
Relative absolute error                 25.587  %
Root relative squared error             41.158  %
Total Number of Instances              205

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     diaporthe-stem-canker
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     charcoal-rot
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     rhizoctonia-root-rot
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     phytophthora-rot
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     brown-stem-rot
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     powdery-mildew
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     downy-mildew
                 0.880    0.017    0.880      0.880    0.880      0.863    0.991     0.946     brown-spot
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bacterial-blight
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bacterial-pustule
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     purple-seed-stain
                 0.833    0.000    1.000      0.833    0.909      0.908    1.000     1.000     anthracnose
                 0.750    0.005    0.857      0.750    0.800      0.794    0.993     0.894     phyllosticta-leaf-spot
                 0.871    0.040    0.794      0.871    0.831      0.800    0.991     0.957     alternarialeaf-spot
                 0.759    0.028    0.815      0.759    0.786      0.753    0.983     0.923     frog-eye-leaf-spot
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     diaporthe-pod-&-stem-blight
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     cyst-nematode
                 1.000    0.010    0.714      1.000    0.833      0.841    1.000     1.000     2-4-d-injury
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     herbicide-injury
Weighted Avg.    0.912    0.013    0.915      0.912    0.912      0.901    0.995     0.972
```

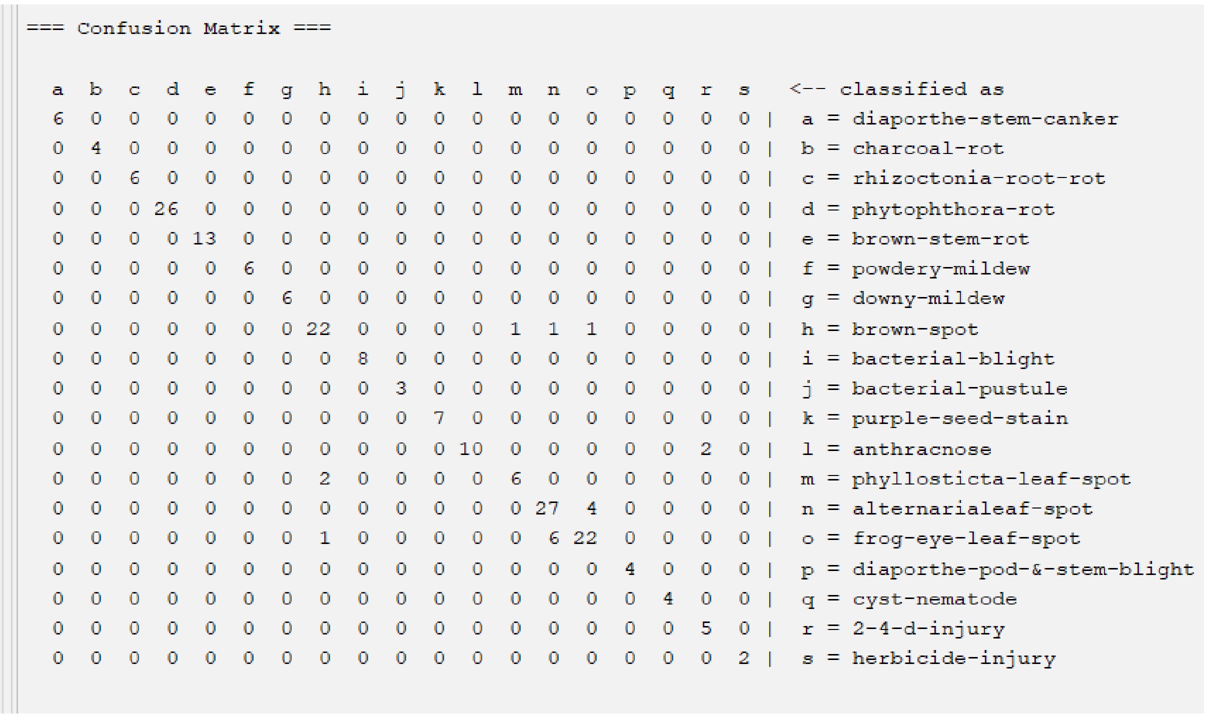
These metrics are provided by Weka following the classification.

Interpretation and Reporting:

Confusion Matrix: Directly available in the output, it shows the actual vs. predicted class allocations.

Feature Importance Plot: While Weka does not directly generate a feature importance plot in the GUI, you can interpret the textual output of feature importance provided by RandomForest or use additional tools or scripts to visualize this data.

Out-of-Bag Error Plot: Similar to feature importance, the out-of-bag error rate can be reviewed from the classifier's textual output. Visualization would require external tools.



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s   <-- classified as
  6  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = diaporthe-stem-canker
  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  b = charcoal-rot
  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = rhizoctonia-root-rot
  0  0  0 26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  d = phytophthora-rot
  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = brown-stem-rot
  0  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0  0 |  f = powdery-mildew
  0  0  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0 |  g = downy-mildew
  0  0  0  0  0  0  0 22  0  0  0  0  1  1  1  0  0  0  0 |  h = brown-spot
  0  0  0  0  0  0  0  0  8  0  0  0  0  0  0  0  0  0  0 |  i = bacterial-blight
  0  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0 |  j = bacterial-pustule
  0  0  0  0  0  0  0  0  0  0  7  0  0  0  0  0  0  0  0 |  k = purple-seed-stain
  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  2  0 |  l = anthracnose
  0  0  0  0  0  0  0  2  0  0  0  0  6  0  0  0  0  0  0 |  m = phyllosticta-leaf-spot
  0  0  0  0  0  0  0  0  0  0  0  0  0 27  4  0  0  0  0 |  n = alternarialeaf-spot
  0  0  0  0  0  0  0  1  0  0  0  0  0  6 22  0  0  0  0 |  o = frog-eye-leaf-spot
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0  0 |  p = diaporthe-pod-&-stem-blight
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0 |  q = cyst-nematode
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  0 |  r = 2-4-d-injury
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2 |  s = herbicide-injury
```

Interpretation:

Analyze the confusion matrix to assess classification accuracy and identify any notable misclassifications across the soybean classes.

Investigate the provided information on feature importance to determine which attributes most significantly impact soybean classification.

Consider the out-of-bag error as an estimate of the model's generalization error.

Result: