Name: Mareena Fernandes

TE IT	Roll number : 8669
Expt. number : 4	Date of implementation: 10/05/2021

Aim : To analyze and evaluate the performance of different classification algorithms using WEKA (data mining tool)

Related Course outcome : CO3

Upon completion of this course students will be able to evaluate the performance of different data mining algorithms using latest tools

Theory : WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools.

When 'WEKA GUI Chooser' window appears on the screen, we can select one of the four options. The options are 1. Simple CLI : provides a simple command line interface and allows direct execution of WEKA commands. 2. Explorer : is an environment for exploring data. 3. Experimenter : is an environment for performing experiments and conducting statistical tests between learning schemes. 4. Knowledge Flow : is a Java-Beans-based interface for setting up and running machine learning experiments.

Classifiers in WEKA are the models for predicting nominal or numeric quantities. The learning schemes available in WEKA include decision trees, Bayes net, neural network, support vector machine and so on.

In this experiment analysis and evaluation of three classification algorithm: Naive Bayesian algorithm, C4.5 algorithm, zero R is done. Before running the classification algorithm it is required to set test options. The selected test options were :

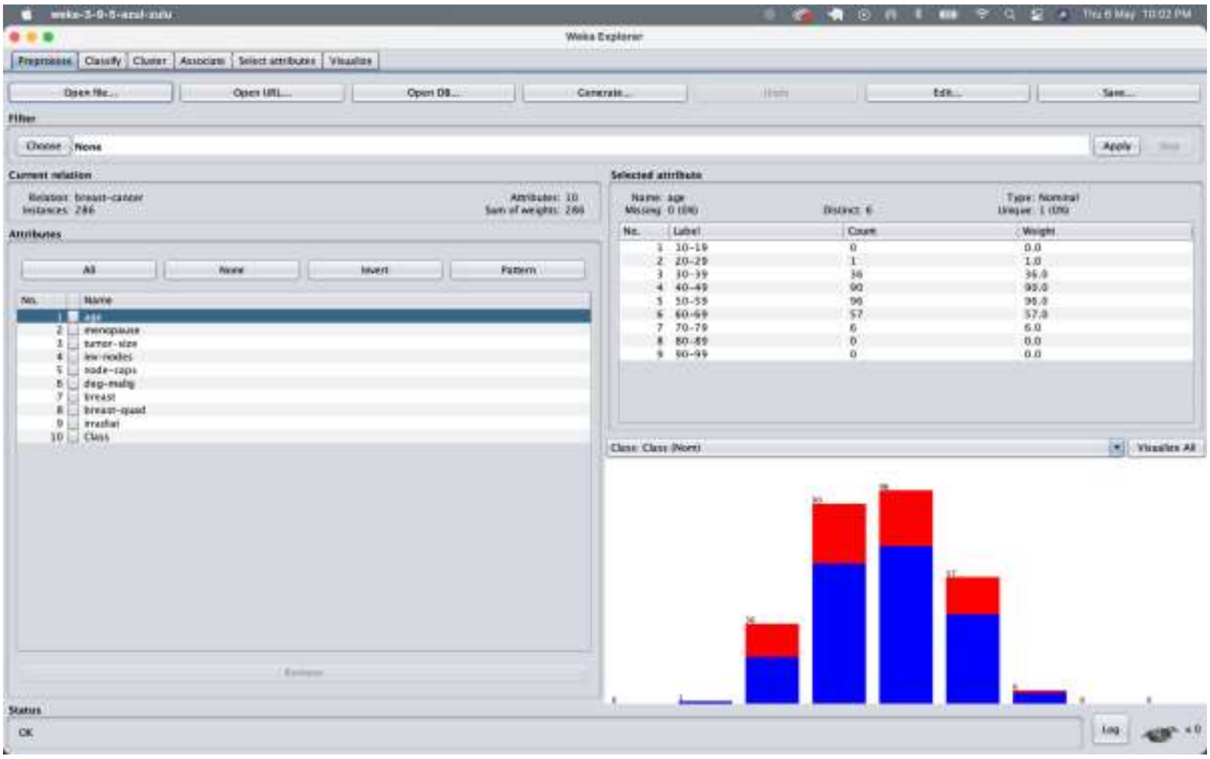
1. Use training set : Evaluates the classifier on how well it predicts the class of the instances it was trained on.
2. Cross-validation : Evaluates the classifier by cross-validation, using the number of folds (10)
3. Percentage split : Evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing.

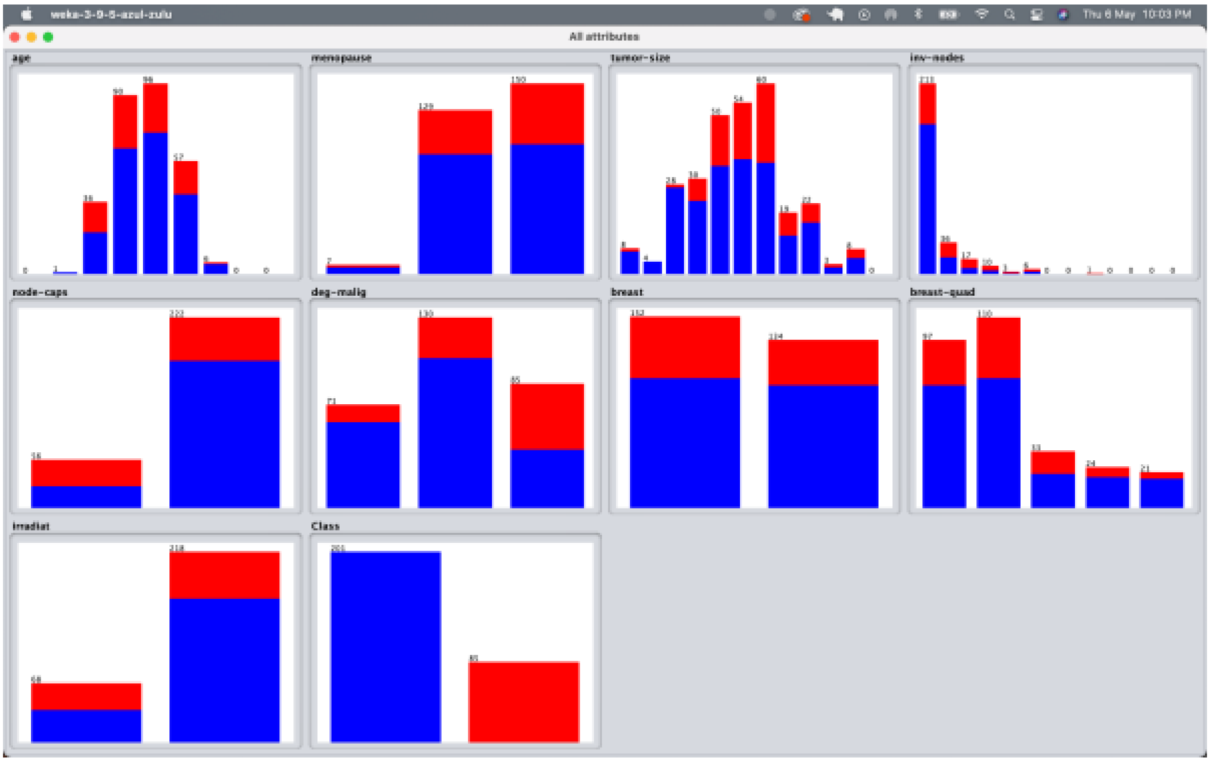
In the classifier evaluation options following options are checked :

1. Output model : The output is the classification model on the full training set.
2. Output per-class stats : The precision/recall and true/false statistics for each class output
3. Output confusion matrix : The confusion matrix of one classifier's prediction is included in the output

When training set is complete, the 'classifier' output area on the right panel of the classifier window is filled with text describing the results of training and testing.

Dataset:

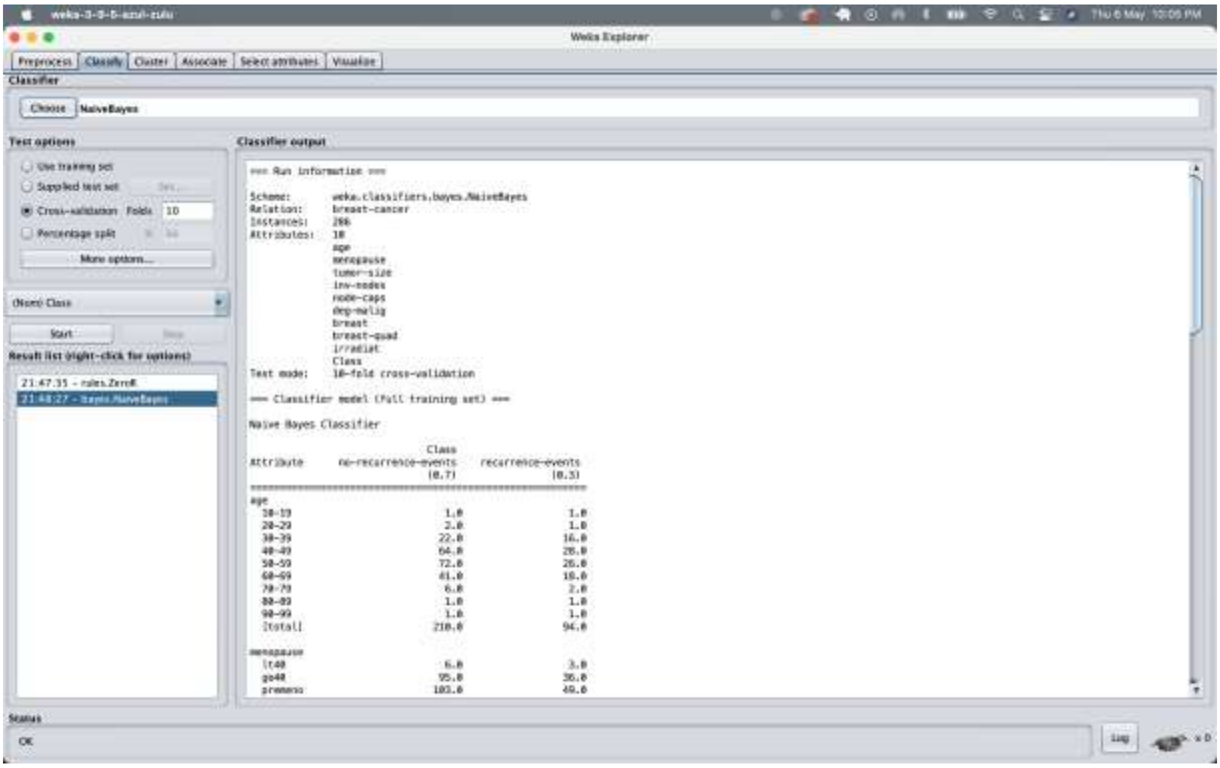

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: breast-cancer
Instances: 286
Attributes: 10
Sum of weights: 286
Attributes
All
None
Invert
Pattern
No.
Name
1 age
2 menopause
3 tumor-size
4 inv-nodes
5 node-caps
6 deg-malig
7 breast
8 breast-quad
9 irradiat
10 Class
Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)
No. Label Count Weight
1 10-19 0 0.0
2 20-29 1 1.0
3 30-39 36 36.0
4 40-49 90 90.0
5 50-59 96 96.0
6 60-69 57 57.0
7 70-79 6 6.0
8 80-89 0 0.0
9 90-99 0 0.0
Class: Class (Nom)
Visualize All
Status
OK
Log

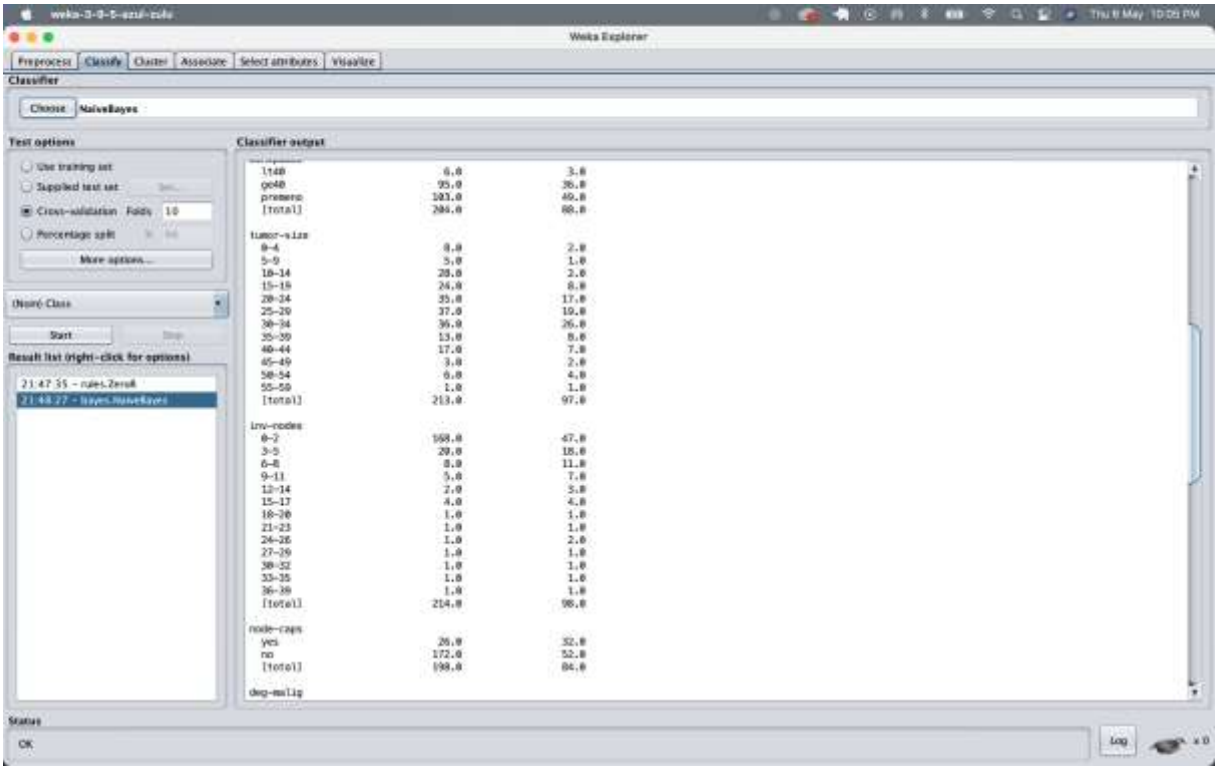

All attributes
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Class

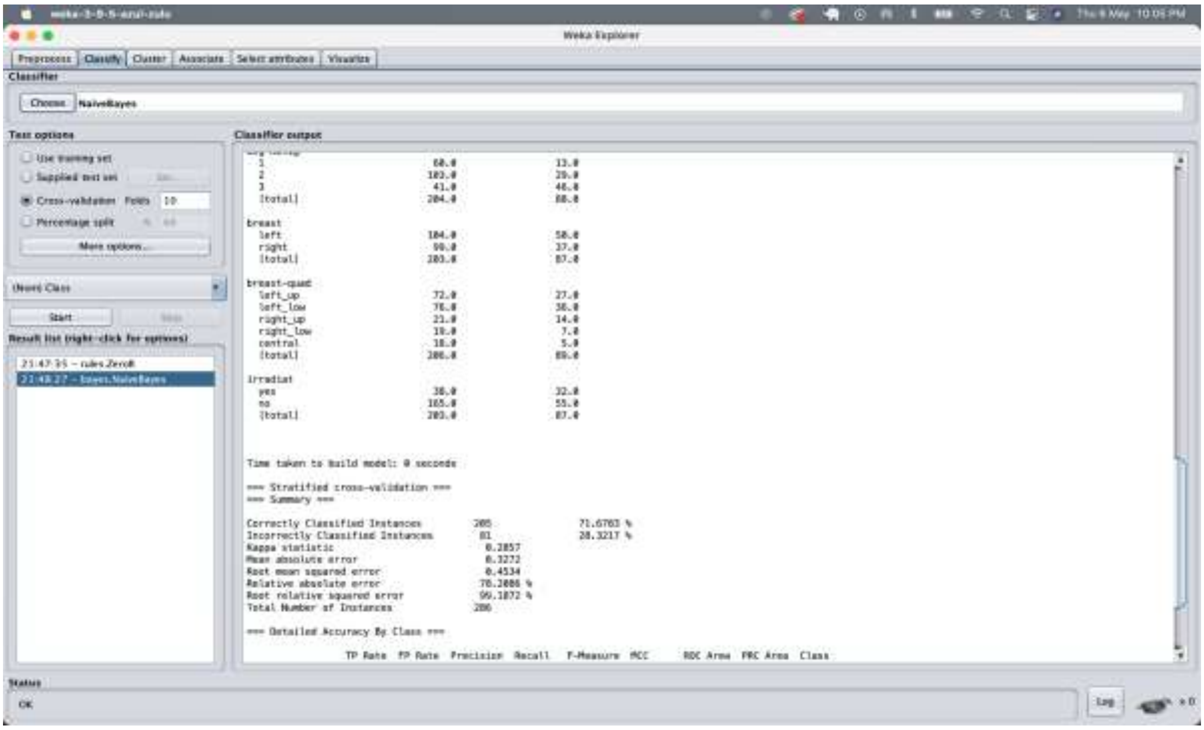
<u>Classifier Output:</u>

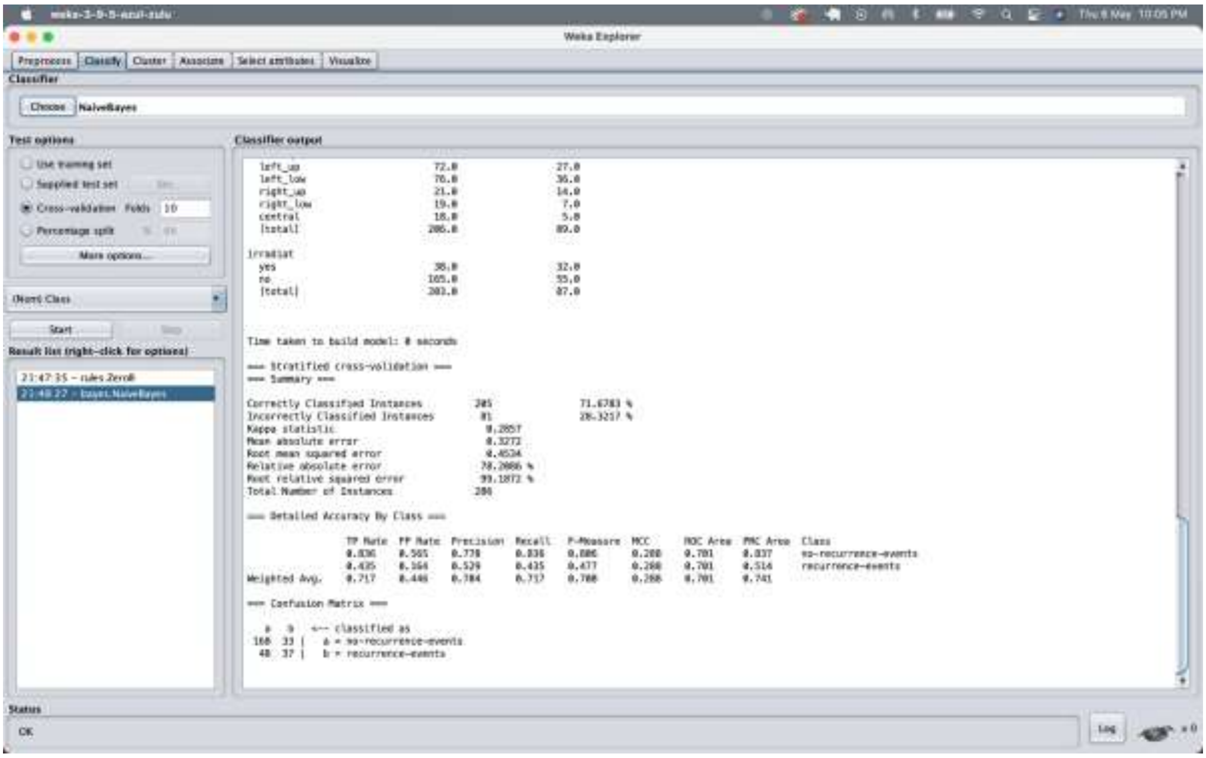
Naive Bayes

Cross Validation

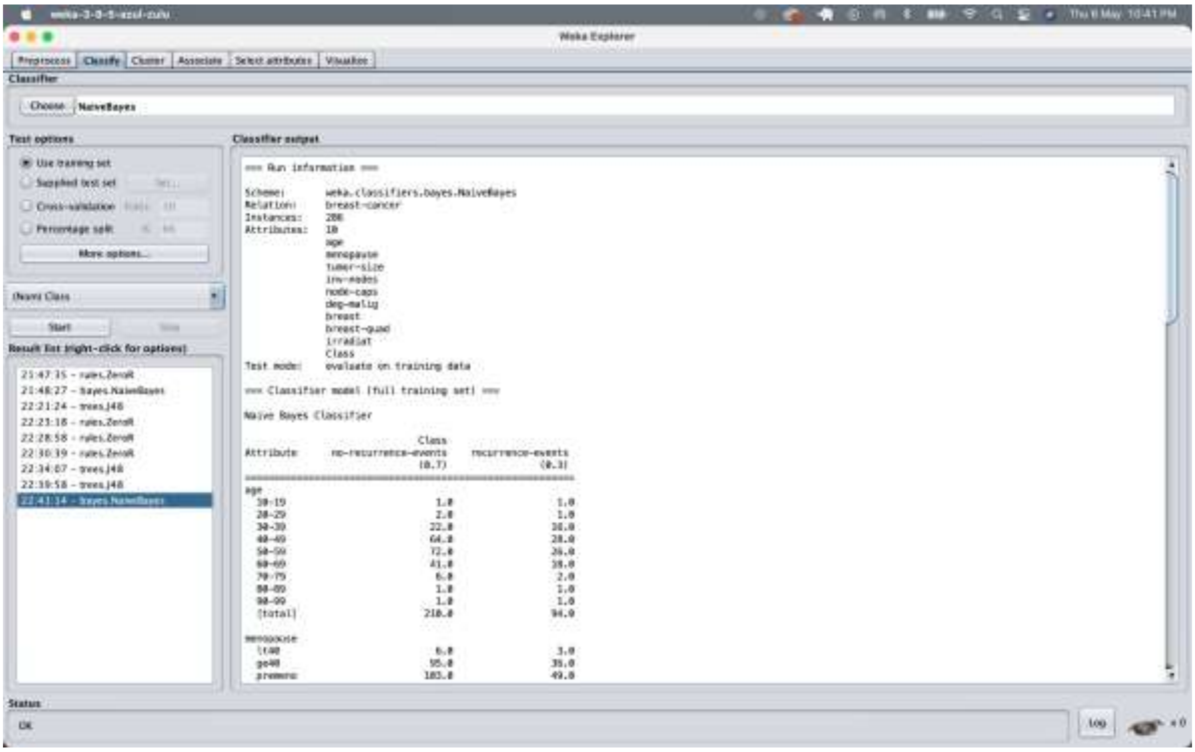


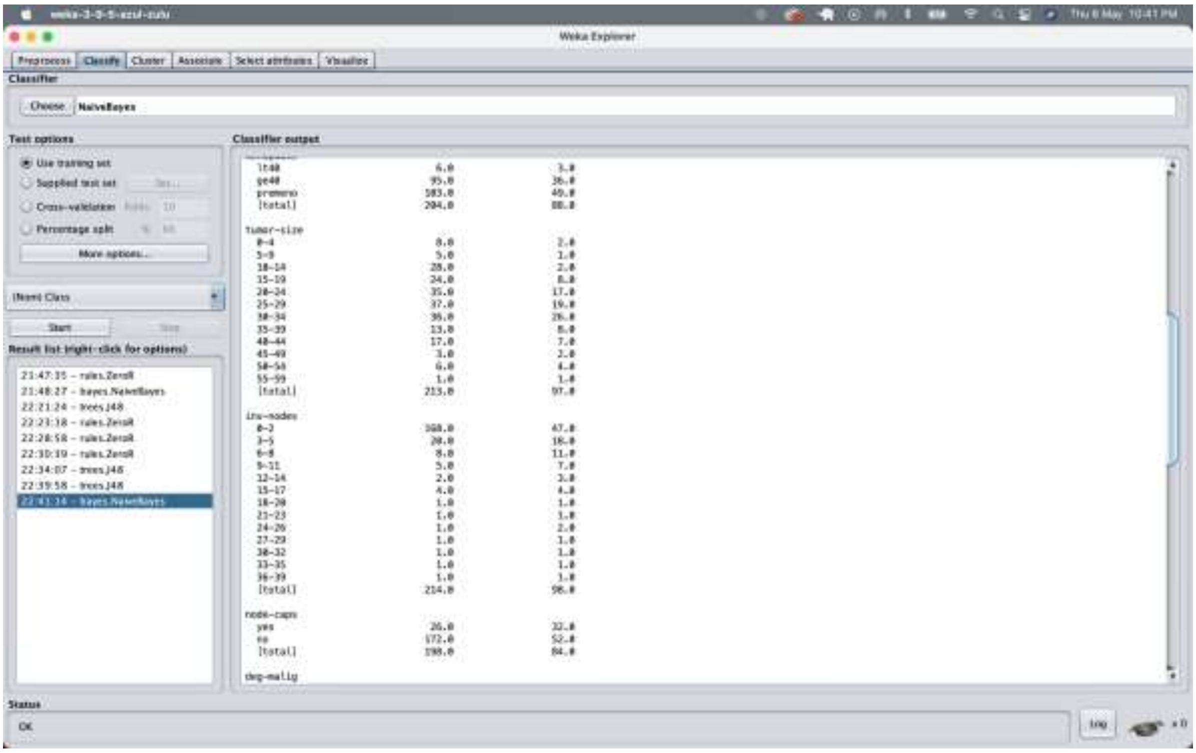


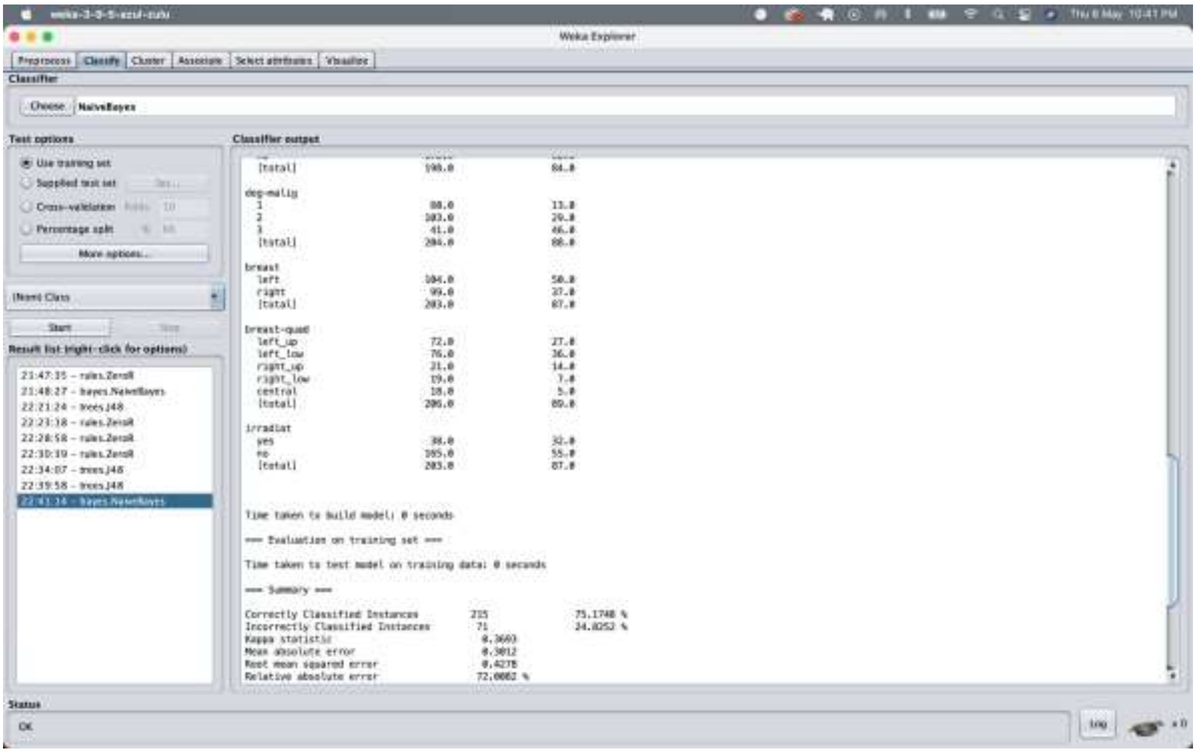

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation Folds 10
Percentage split
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
21:47:35 - rules.ZeroR
21:48:27 - bayes.NaiveBayes
Classifier output
breast
left
right
[total]
breast-quad
left_up
left_low
right_up
right_low
central
[total]
irradiat
yes
no
[total]
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 205 71.6783 %
Incorrectly Classified Instances 81 28.3217 %
Kappa statistic 0.2857
Mean absolute error 0.3272
Root mean squared error 0.4534
Relative absolute error 78.2086 %
Root relative squared error 99.1872 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation Folds 10
Percentage split
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
21:47:35 - rules.ZeroR
21:48:27 - bayes.NaiveBayes
Classifier output
left_up
left_low
right_up
right_low
central
[total]
irradiat
yes 38.0 32.0
no 165.0 55.0
[total] 203.0 87.0
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 205 71.6783 %
Incorrectly Classified Instances 81 28.3217 %
Kappa statistic 0.2857
Mean absolute error 0.3272
Root mean squared error 0.4534
Relative absolute error 78.2086 %
Root relative squared error 99.1872 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.836 0.565 0.778 0.836 0.806 0.288 0.701 0.837 no-recurrence-events
0.435 0.164 0.529 0.435 0.477 0.288 0.701 0.514 recurrence-events
Weighted Avg. 0.717 0.446 0.704 0.717 0.708 0.288 0.701 0.741
=== Confusion Matrix ===
a b <-- classified as
168 33 | a = no-recurrence-events
48 37 | b = recurrence-events
Status
OK
Log

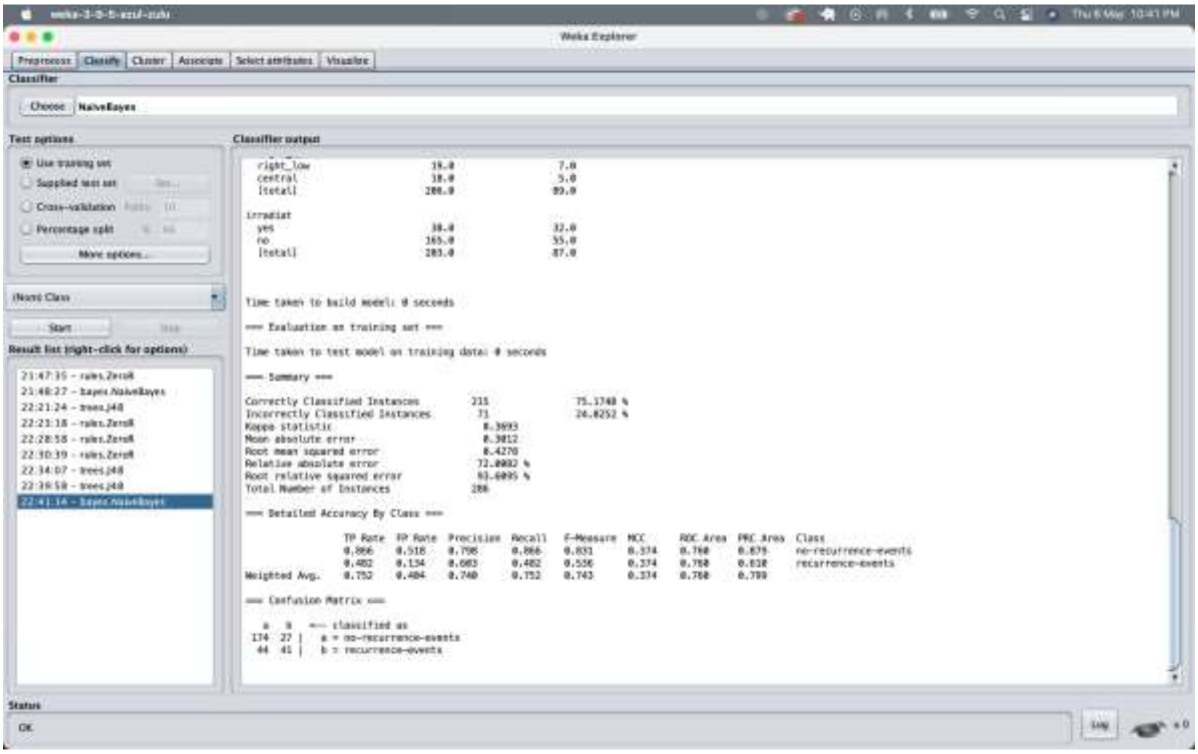
Use Training Sets:





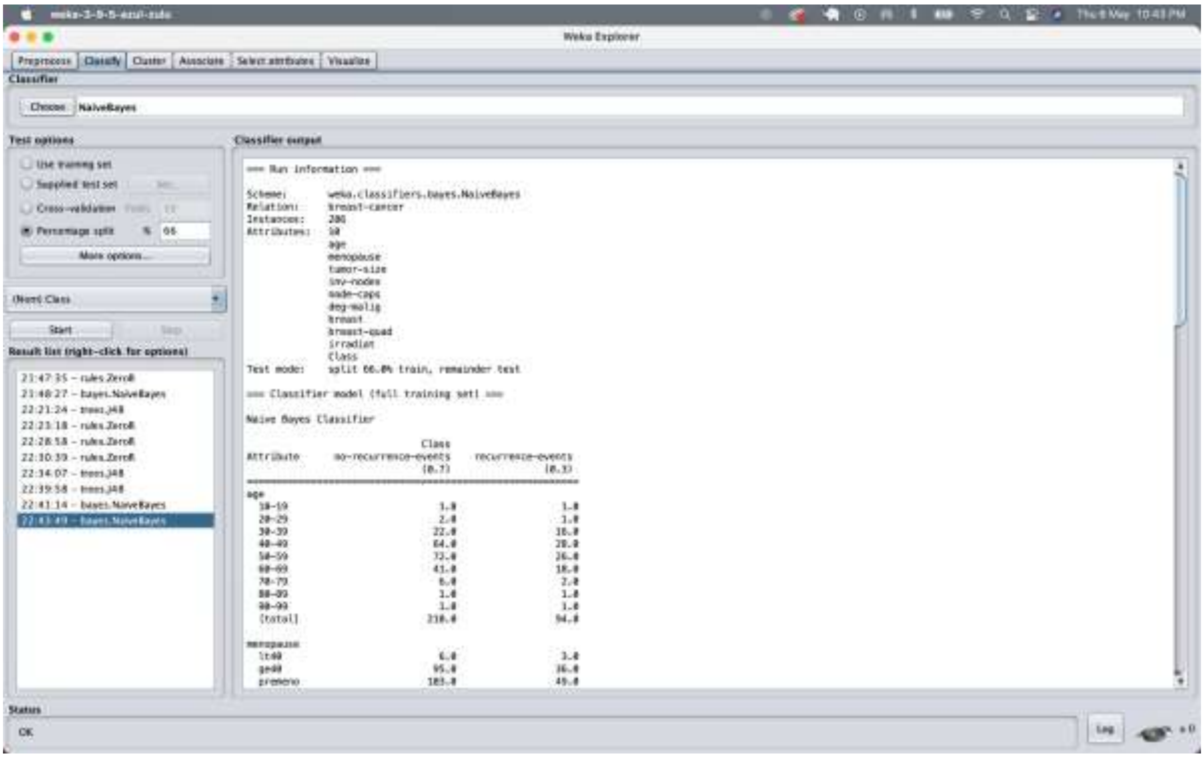


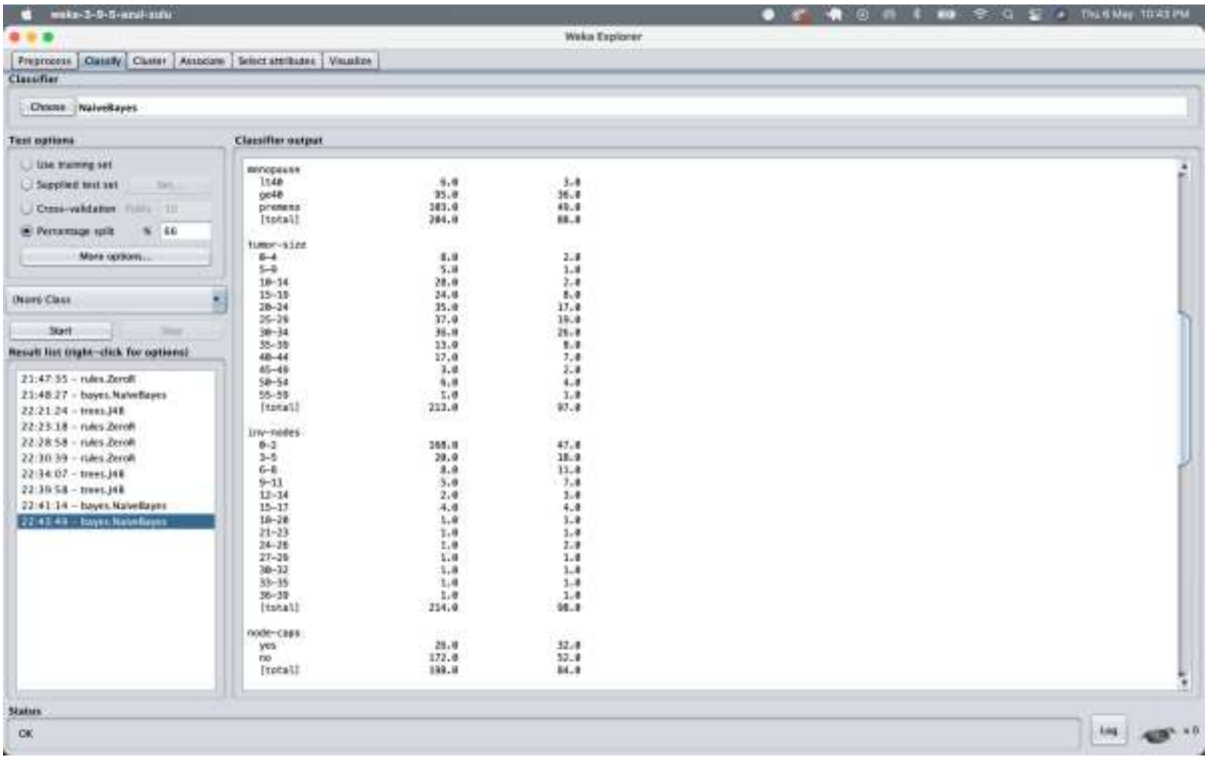
weka-3-8-5-azul-zulu
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
21:47:35 - rules.ZeroR
21:48:27 - bayes.NaiveBayes
22:21:24 - trees.J48
22:23:18 - rules.ZeroR
22:28:58 - rules.ZeroR
22:30:39 - rules.ZeroR
22:34:07 - trees.J48
22:39:58 - trees.J48
22:41:14 - bayes.NaiveBayes
Classifier output
[total] 198.0 84.0
deg-malig
1 88.0 13.0
2 103.0 29.0
3 41.0 46.0
[total] 204.0 88.0
breast
left 104.0 50.0
right 99.0 37.0
[total] 203.0 87.0
breast-quad
left_up 72.0 27.0
left_low 76.0 36.0
right_up 21.0 14.0
right_low 19.0 7.0
central 18.0 5.0
[total] 206.0 89.0
irradiat
yes 38.0 32.0
no 165.0 55.0
[total] 203.0 87.0
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 215 75.1748 %
Incorrectly Classified Instances 71 24.8252 %
Kappa statistic 0.3693
Mean absolute error 0.3012
Root mean squared error 0.4278
Relative absolute error 72.0082 %
Status
OK
Log

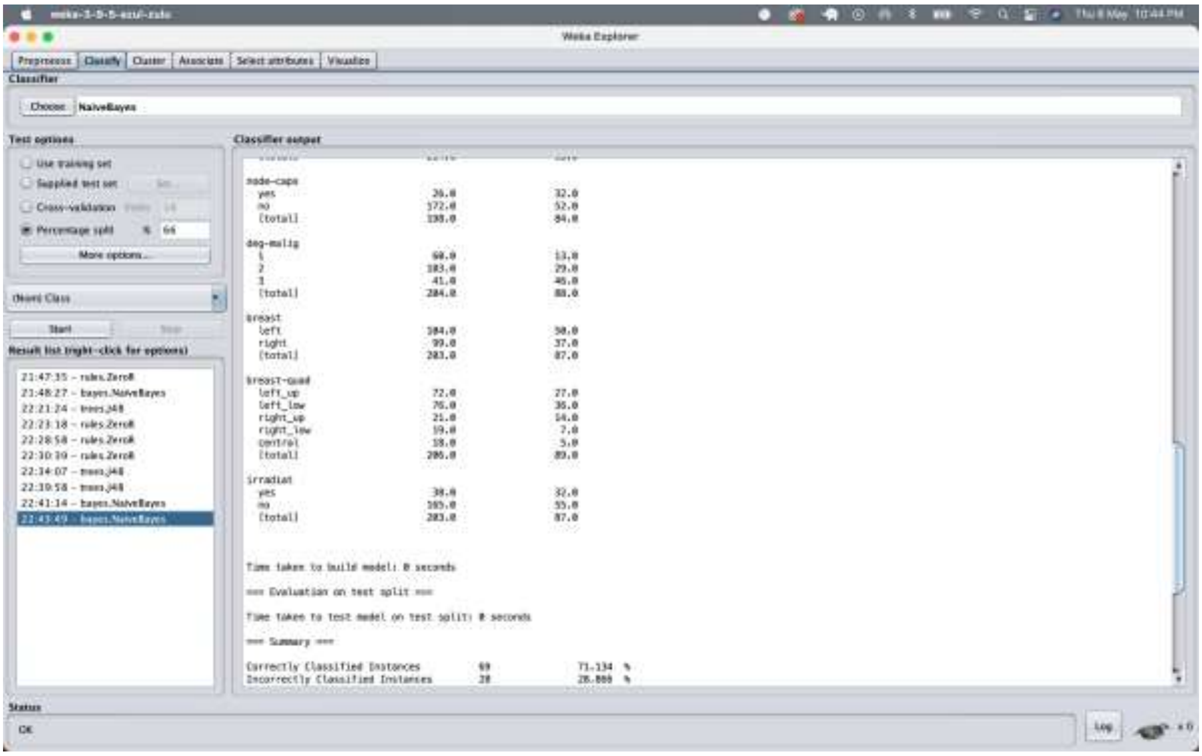


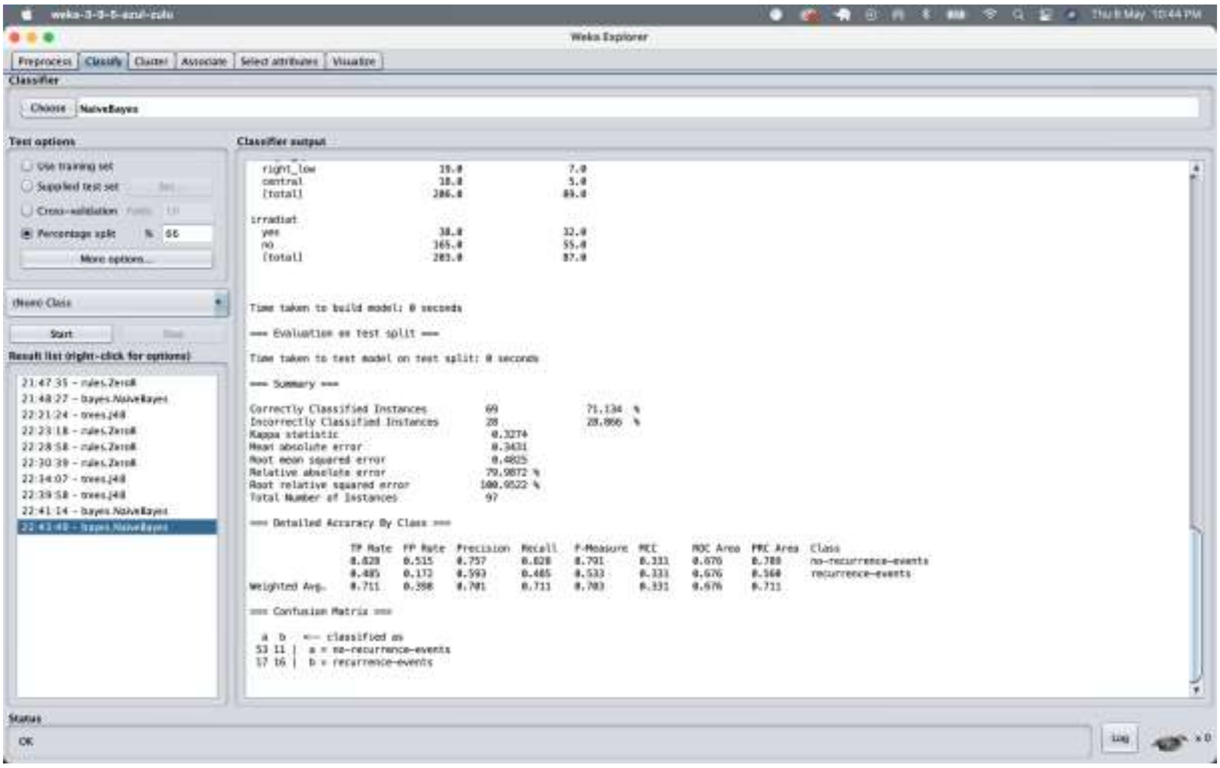
weka-3-8-5-azul-zulu
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
21:47:35 - rules.ZeroR
21:48:27 - bayes.NaiveBayes
22:21:24 - trees.J48
22:23:18 - rules.ZeroR
22:28:58 - rules.ZeroR
22:30:39 - rules.ZeroR
22:34:07 - trees.J48
22:39:58 - trees.J48
22:41:14 - bayes.NaiveBayes
Classifier output
right_low 19.0 7.0
central 18.0 5.0
[total] 206.0 89.0
irradiat
yes 38.0 32.0
no 165.0 55.0
[total] 203.0 87.0
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 215 75.1748 %
Incorrectly Classified Instances 71 24.8252 %
Kappa statistic 0.3693
Mean absolute error 0.3012
Root mean squared error 0.4278
Relative absolute error 72.0082 %
Root relative squared error 93.6095 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.866 0.518 0.798 0.866 0.831 0.374 0.760 0.875 no-recurrence-events
0.482 0.134 0.603 0.482 0.536 0.374 0.760 0.610 recurrence-events
Weighted Avg. 0.752 0.404 0.740 0.752 0.743 0.374 0.760 0.799
=== Confusion Matrix ===
a b <-- classified as
174 27 | a = no-recurrence-events
44 41 | b = recurrence-events
Status
OK
Log

Percentage Split:





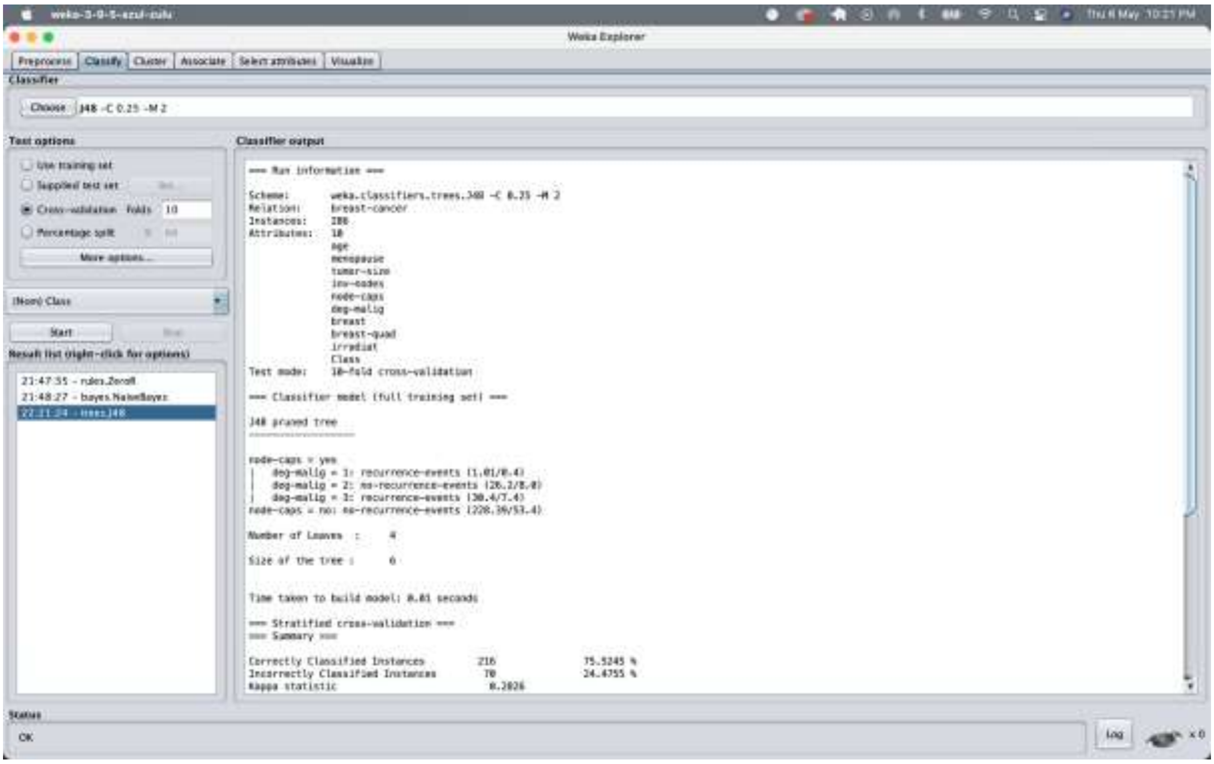

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
21:47:35 - rules.ZeroR
21:48:27 - bayes.NaiveBayes
22:21:24 - trees.J48
22:23:18 - rules.ZeroR
22:28:58 - rules.ZeroR
22:30:39 - rules.ZeroR
22:34:07 - trees.J48
22:39:58 - trees.J48
22:41:14 - bayes.NaiveBayes
22:43:49 - bayes.NaiveBayes
Classifier output
node-caps
yes 26.0 32.0
no 172.0 52.0
[total] 198.0 84.0
deg-malig
1 60.0 13.0
2 103.0 29.0
3 41.0 46.0
[total] 204.0 88.0
breast
left 104.0 50.0
right 99.0 37.0
[total] 203.0 87.0
breast-quad
left_up 72.0 27.0
left_low 76.0 36.0
right_up 21.0 14.0
right_low 19.0 7.0
central 18.0 5.0
[total] 206.0 89.0
irradiat
yes 38.0 32.0
no 165.0 55.0
[total] 203.0 87.0
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 69 71.134 %
Incorrectly Classified Instances 28 28.866 %
Status
OK
Log

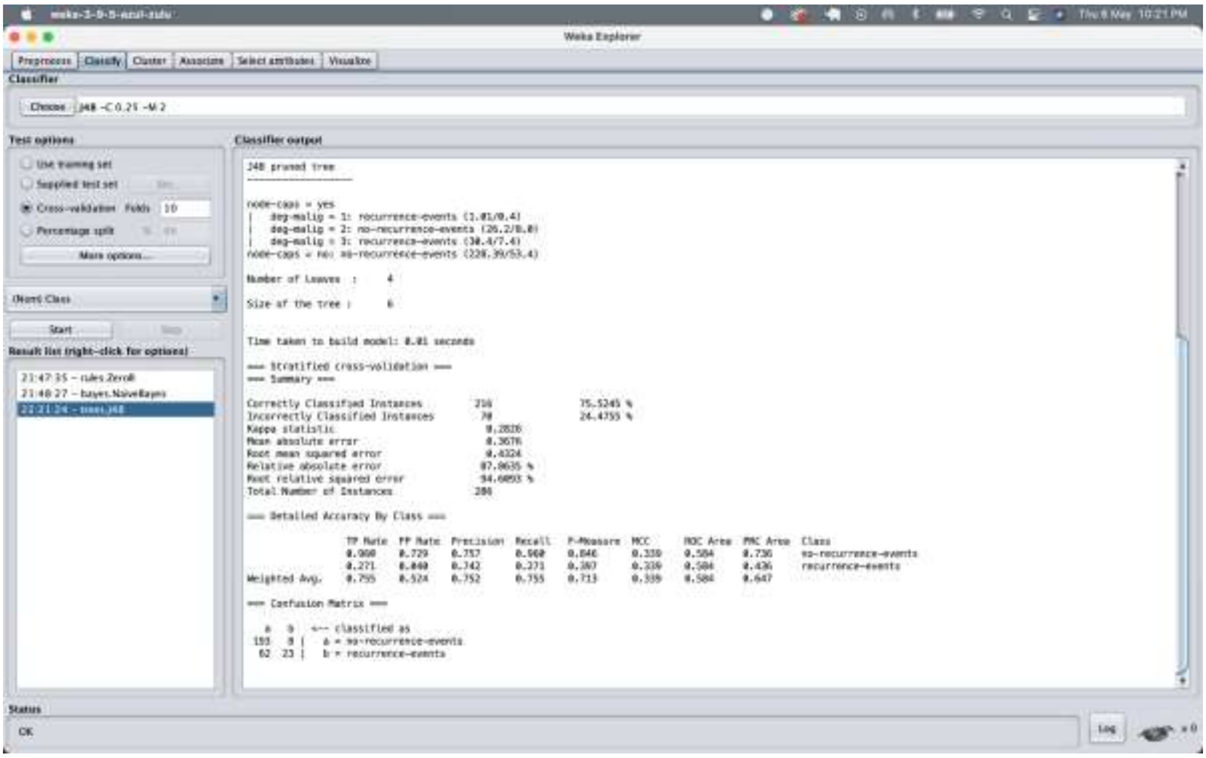

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
21:47:35 - rules.ZeroR
21:48:27 - bayes.NaiveBayes
22:21:24 - trees.J48
22:23:18 - rules.ZeroR
22:28:58 - rules.ZeroR
22:30:39 - rules.ZeroR
22:34:07 - trees.J48
22:39:58 - trees.J48
22:41:14 - bayes.NaiveBayes
22:43:49 - bayes.NaiveBayes
Classifier output
right_low 19.0 7.0
central 18.0 5.0
[total] 206.0 89.0
irradiat
yes 38.0 32.0
no 165.0 55.0
[total] 203.0 87.0
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 69 71.134 %
Incorrectly Classified Instances 28 28.866 %
Kappa statistic 0.3274
Mean absolute error 0.3431
Root mean squared error 0.4825
Relative absolute error 79.9872 %
Root relative squared error 100.9522 %
Total Number of Instances 97
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.828 0.515 0.757 0.828 0.791 0.331 0.676 0.788 no-recurrence-events
0.485 0.172 0.593 0.485 0.533 0.331 0.676 0.560 recurrence-events
Weighted Avg. 0.711 0.398 0.701 0.711 0.703 0.331 0.676 0.711
=== Confusion Matrix ===
a b <-- classified as
53 11 | a = no-recurrence-events
17 16 | b = recurrence-events
Status
OK
Log

Classifier Output:

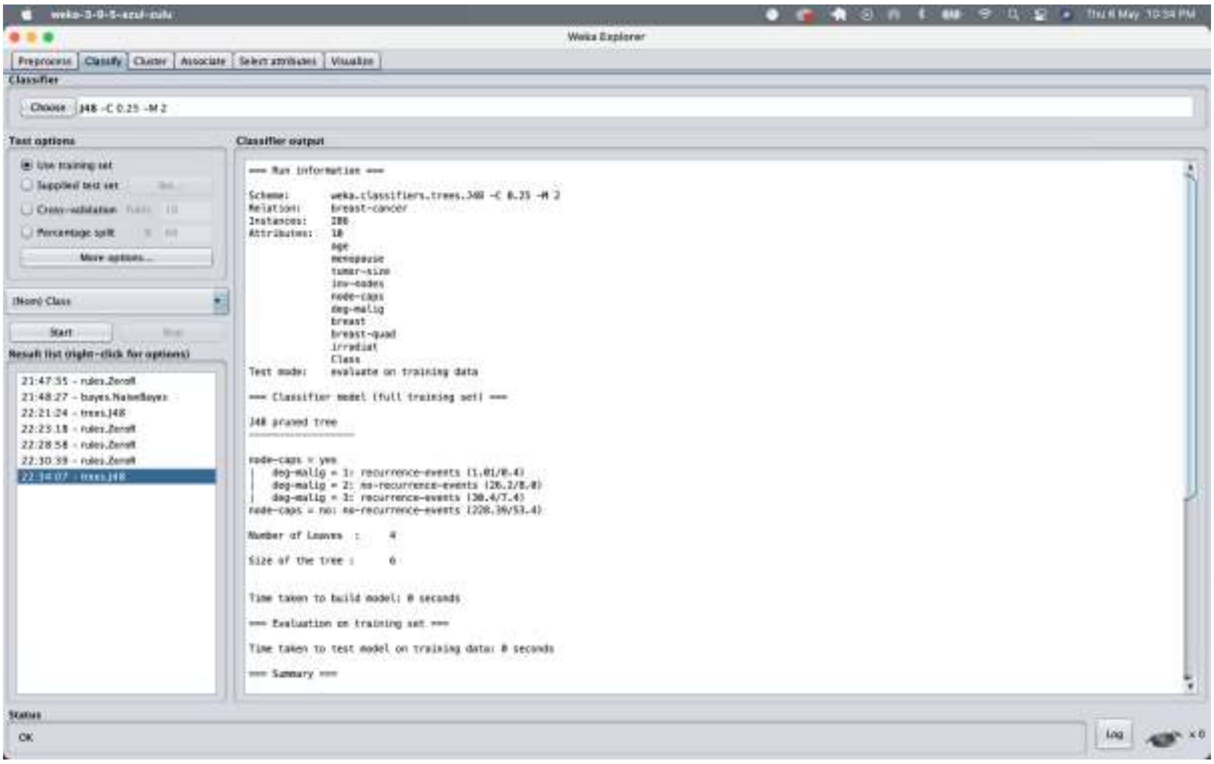
Decision Tree

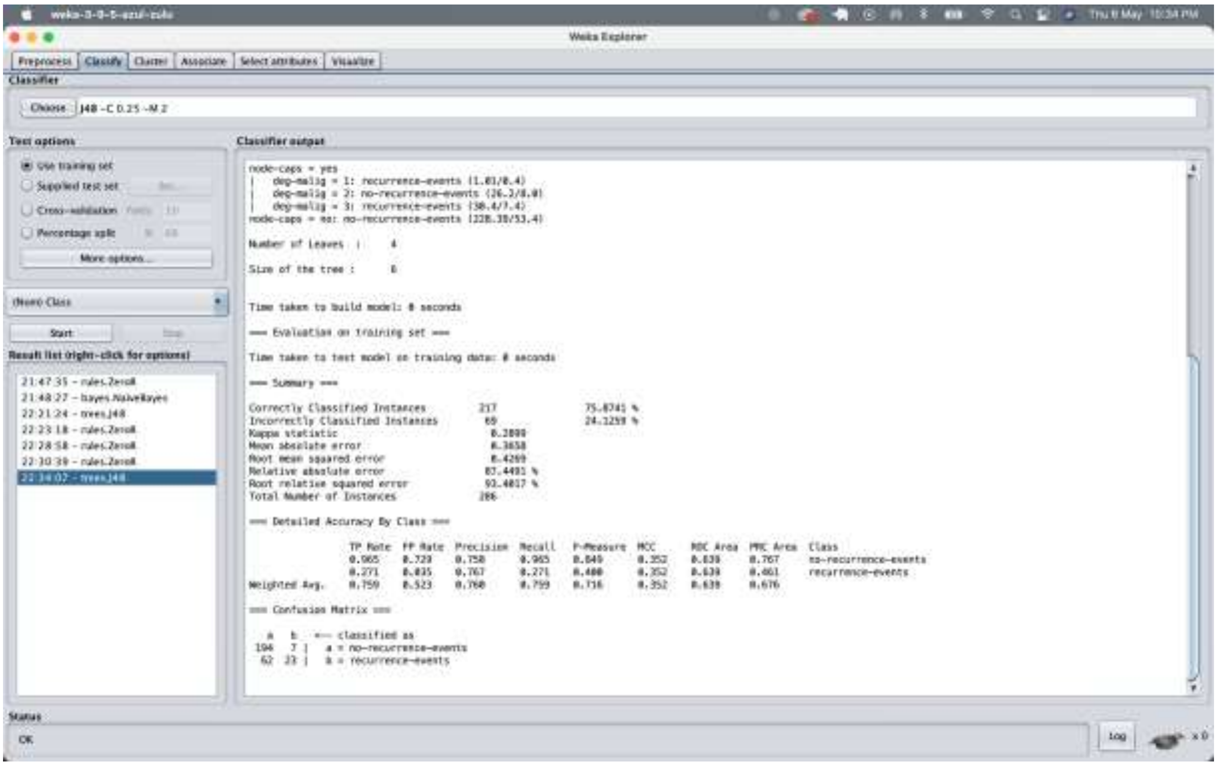
Cross Validation



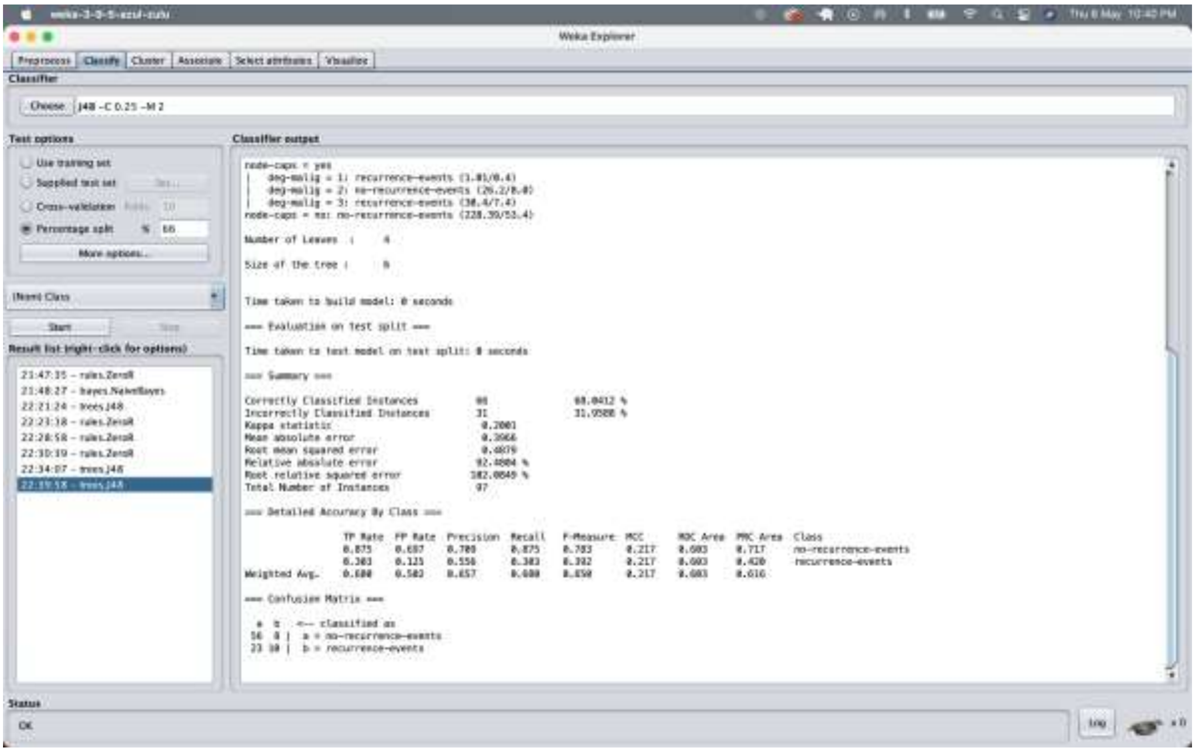


Use Training Set:





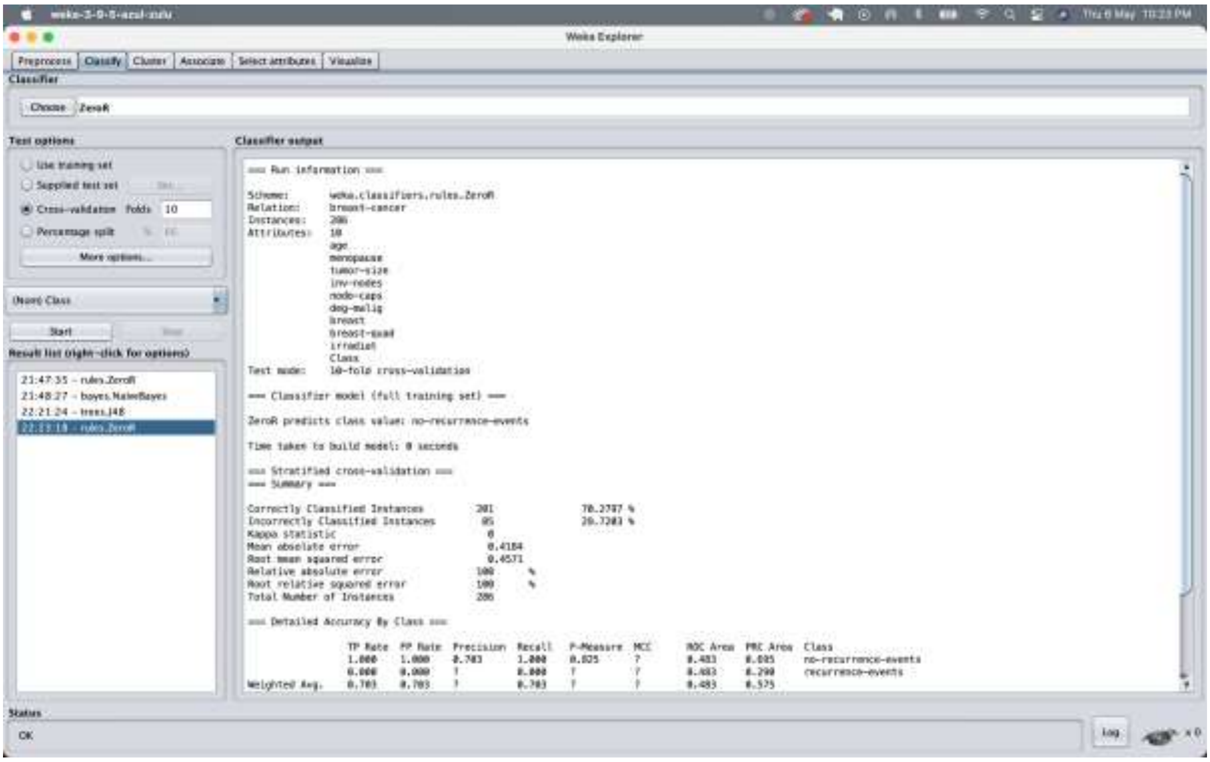
Percentage Split:

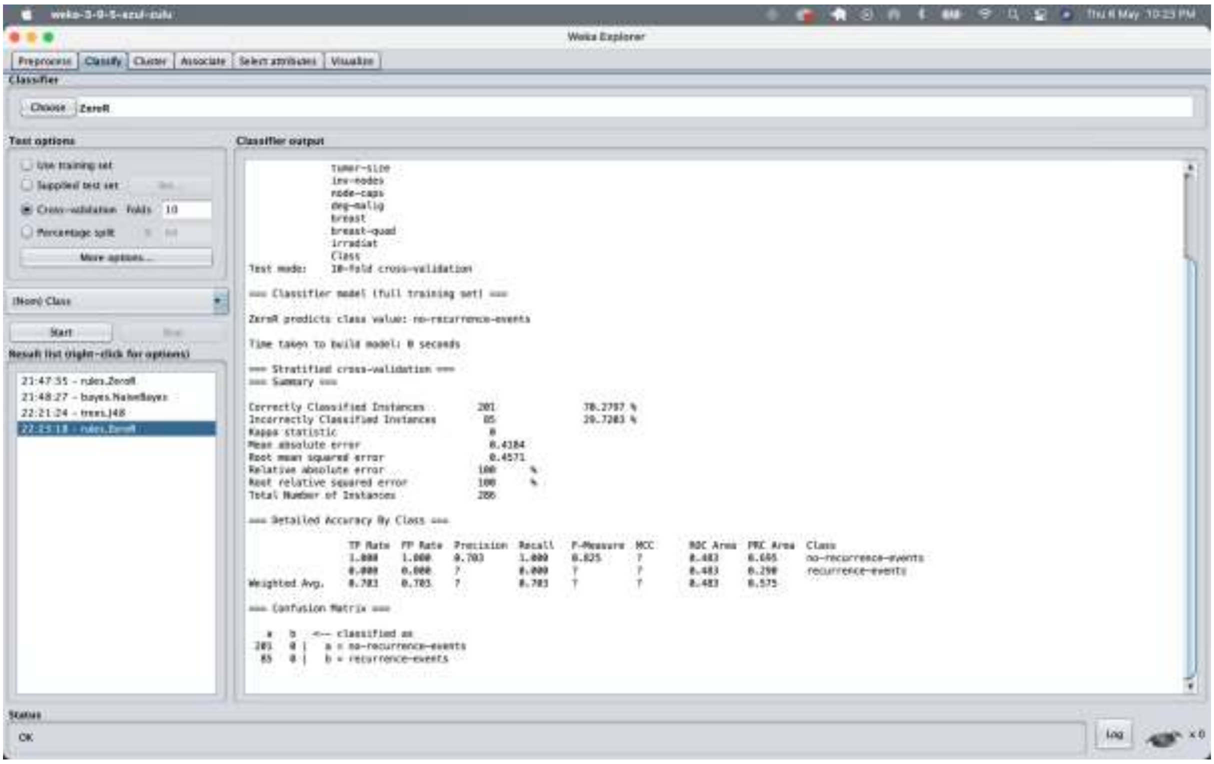


Classifier Output:

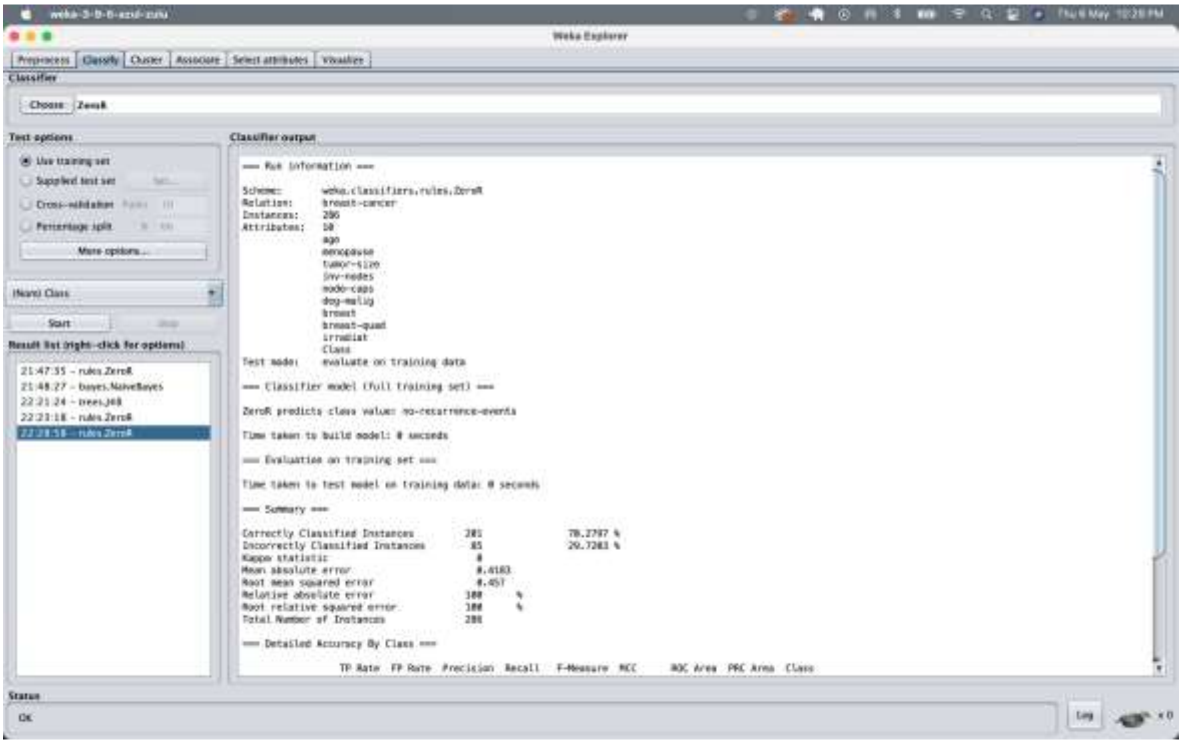
Zero R

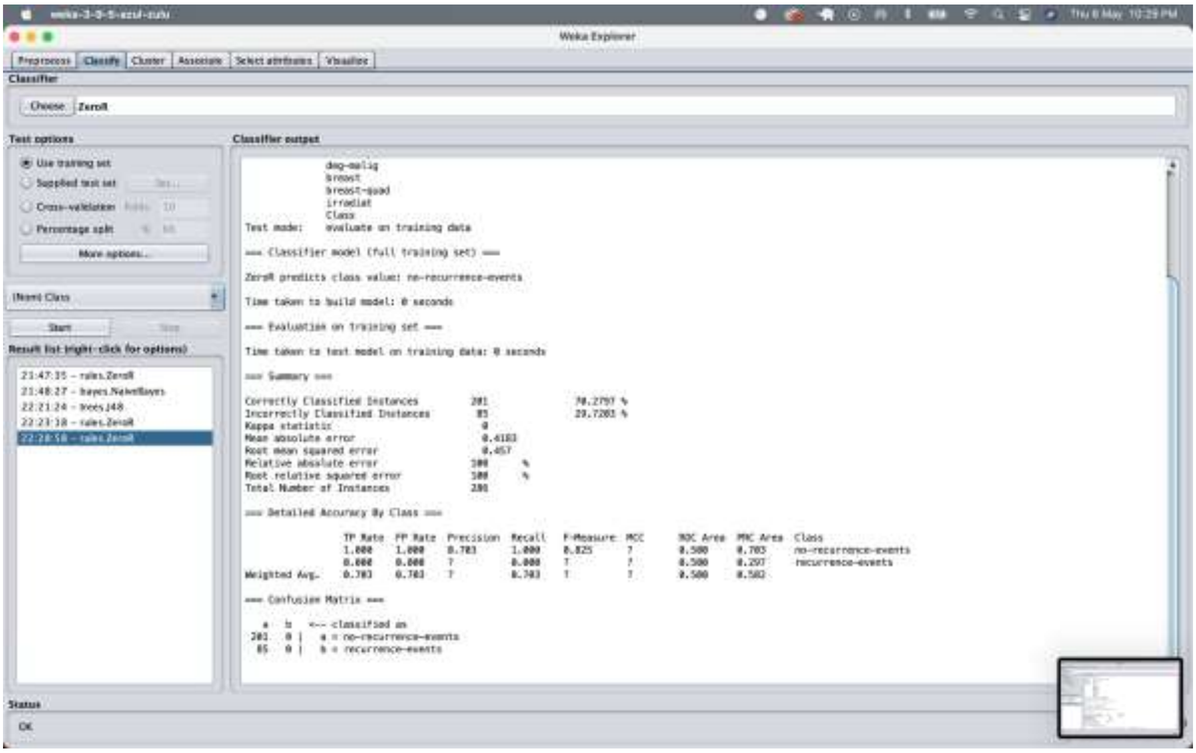
Cross Validation



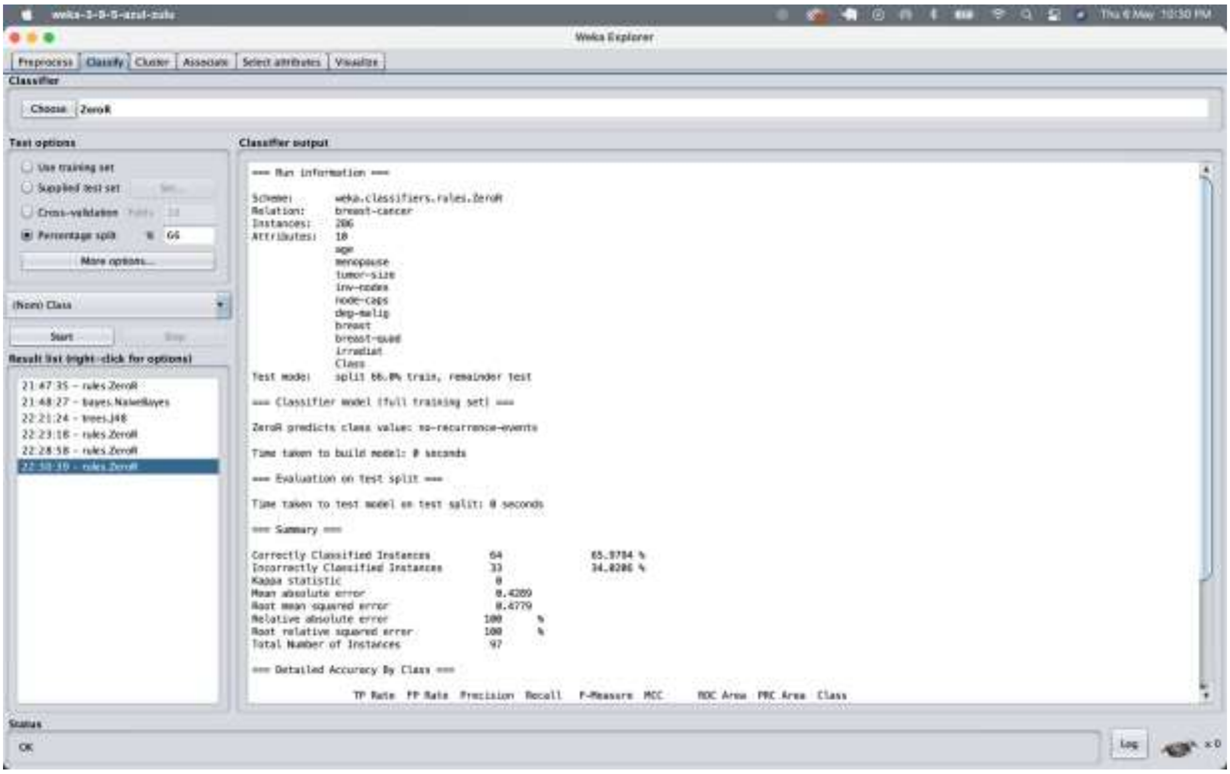


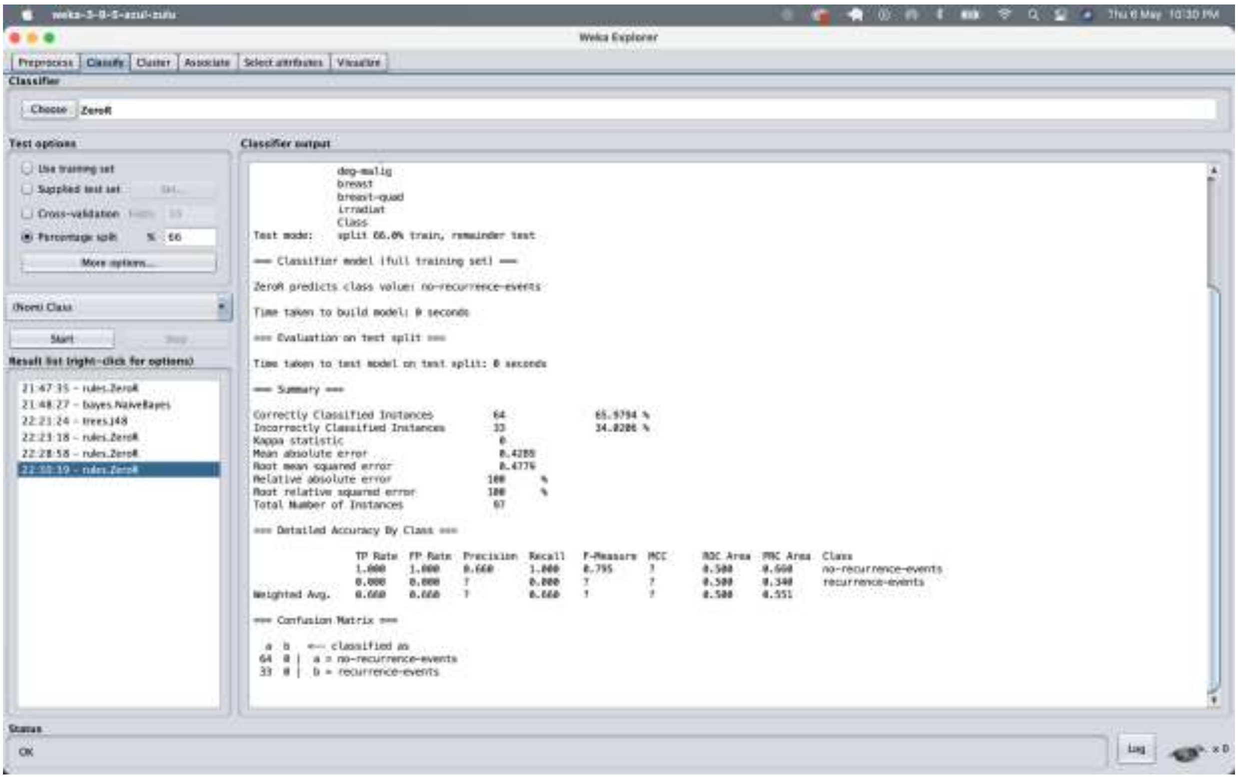
Use training Set:





Percentage Split:





Post Lab Questions

1. Explain ARFF file format with example

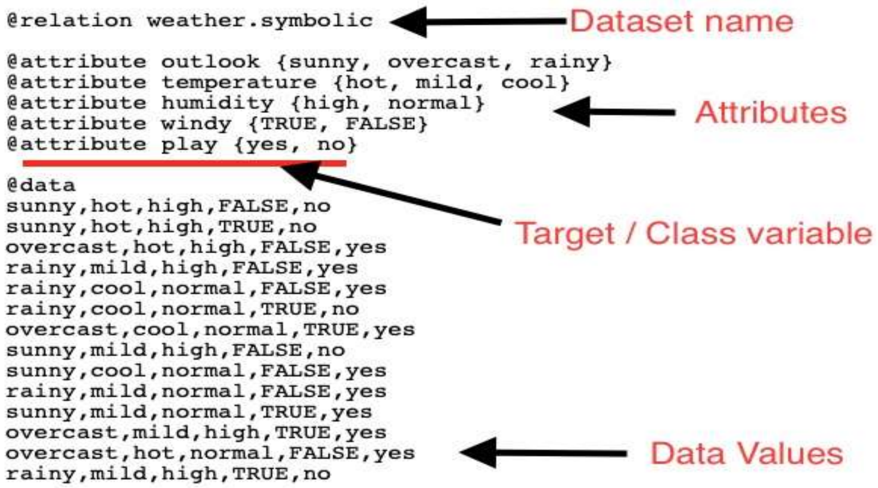
Ans:

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. An Arff file contains two sections - header and data.

The header describes the attribute types.

The data section contains a comma separated list of data.

As an example for Arff format, the Weather data file loaded from the WEKA sample databases is shown below



From the screenshot, you can infer the following points −

The @relation tag defines the name of the database.

The @attribute tag defines the attributes.

The @data tag starts the list of data rows each containing the comma separated fields.

2. What are WEKA filters? Explain any one

Ans:

a. Weka includes many filters that can be used before invoking a classifier to clean up the dataset, or alter it in some way. Filters help with data preparation. For example, you can easily remove an attribute. Or you can remove all instances that have a certain value for an attribute (e.g. instances for which humidity has the value high). Surprisingly, removing attributes sometimes leads to better classification! – and also simpler decision trees.
b. Some machine learning algorithms prefer or find it easier to work with discrete attributes.
c. For example, decision tree algorithms can choose split points in real valued attributes, but are much cleaner when split points are chosen between bins or predefined groups in the real valued attributes.
d. Discrete attributes are those that describe a category, called nominal attributes. Those attributes that describe a category where there is a meaning in the order for the categories are called ordinal attributes. The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization.
e. You can discretize your real valued attributes in Weka using the Discretize filter.