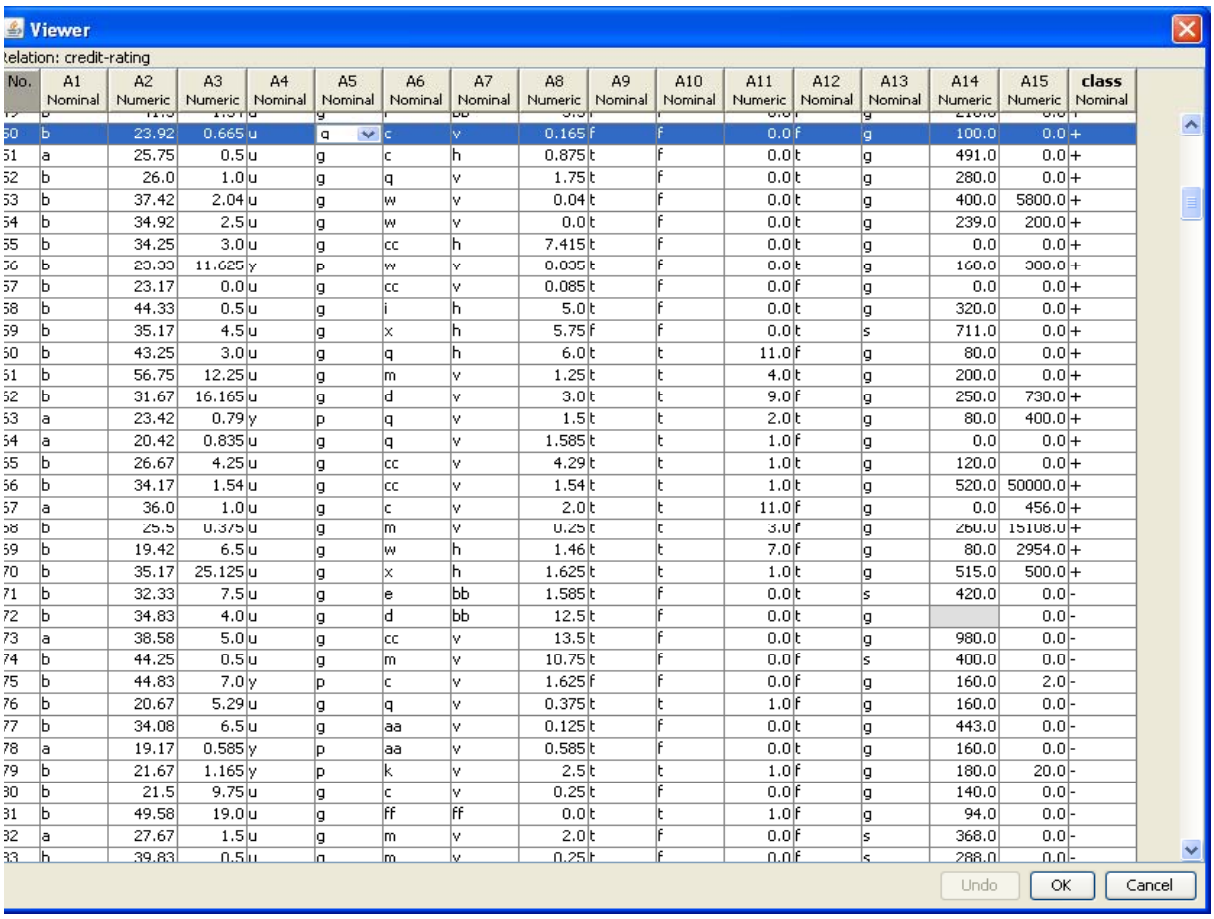
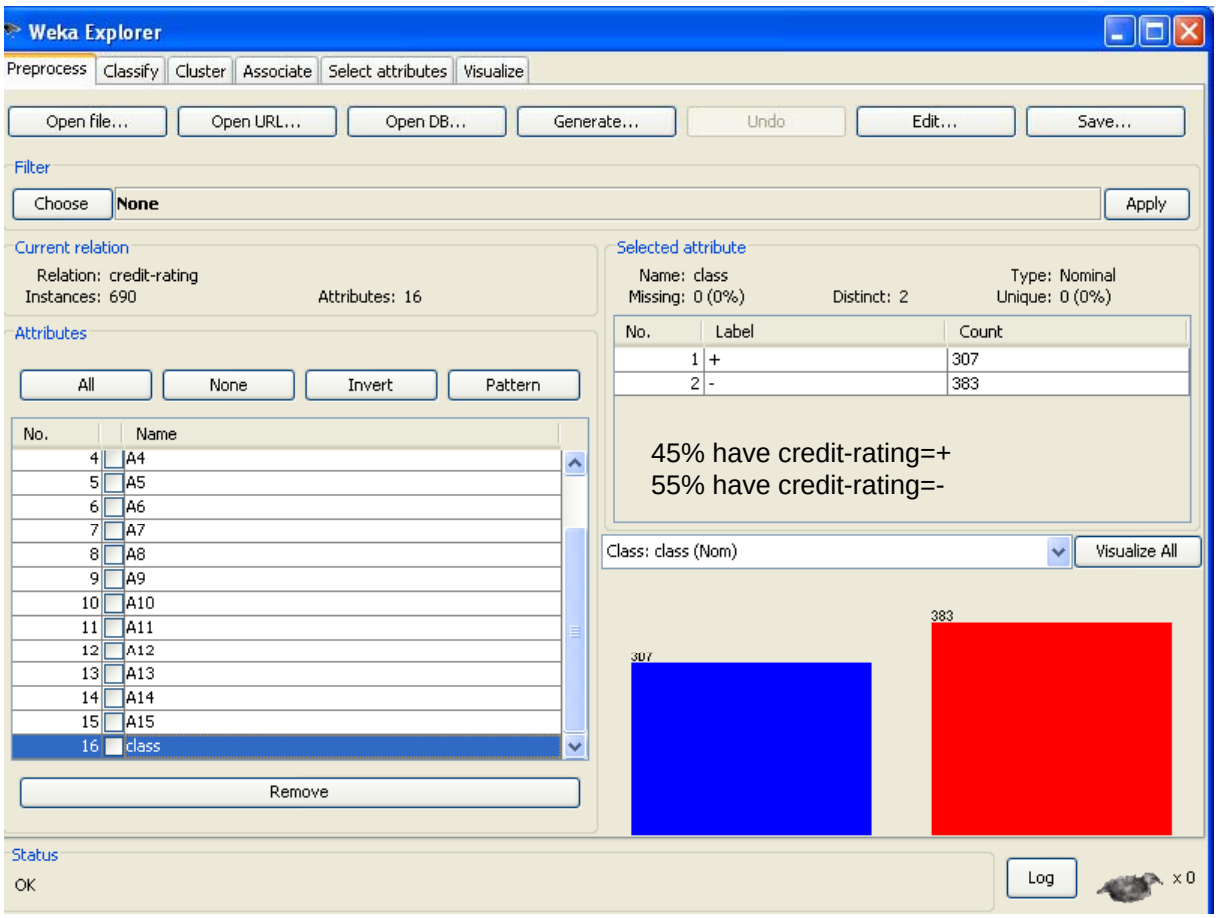
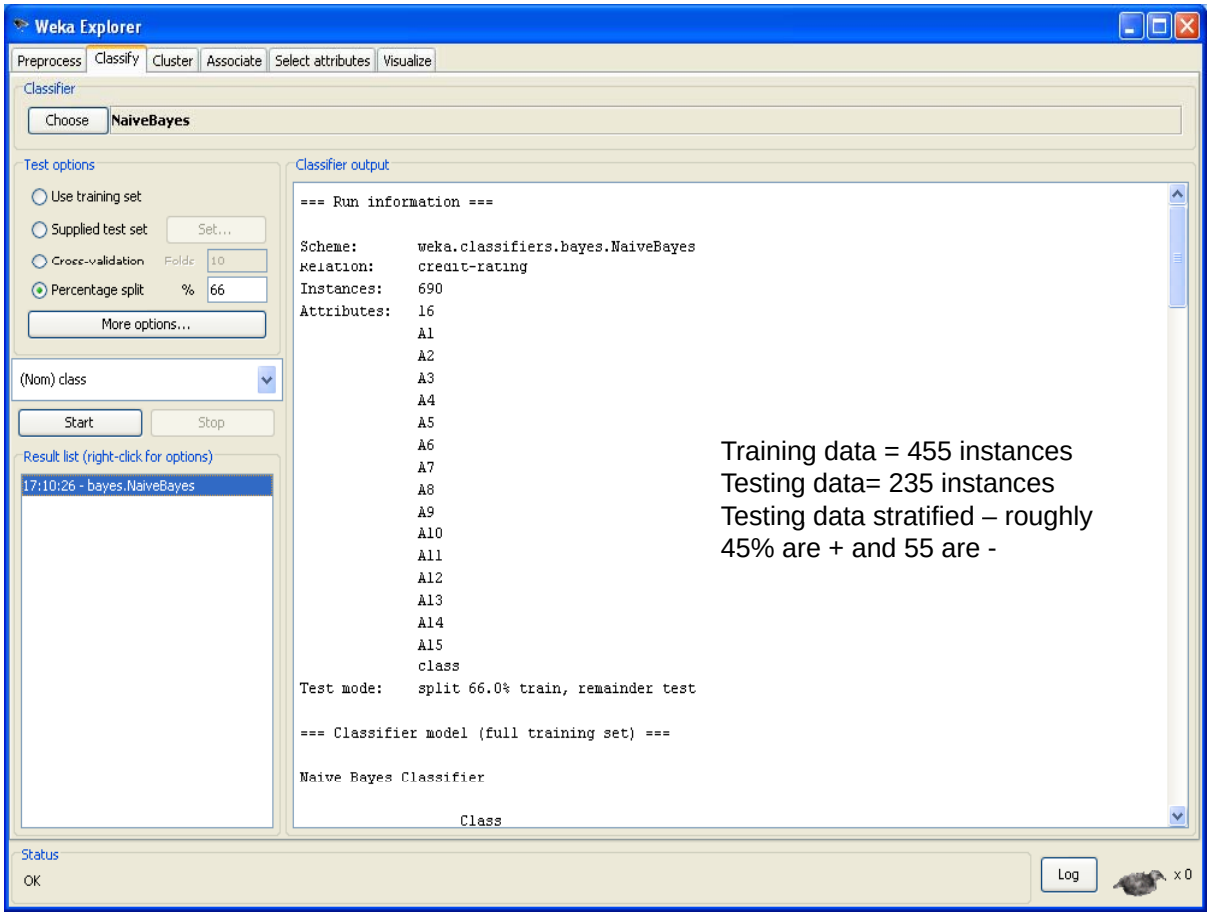
Data Mining

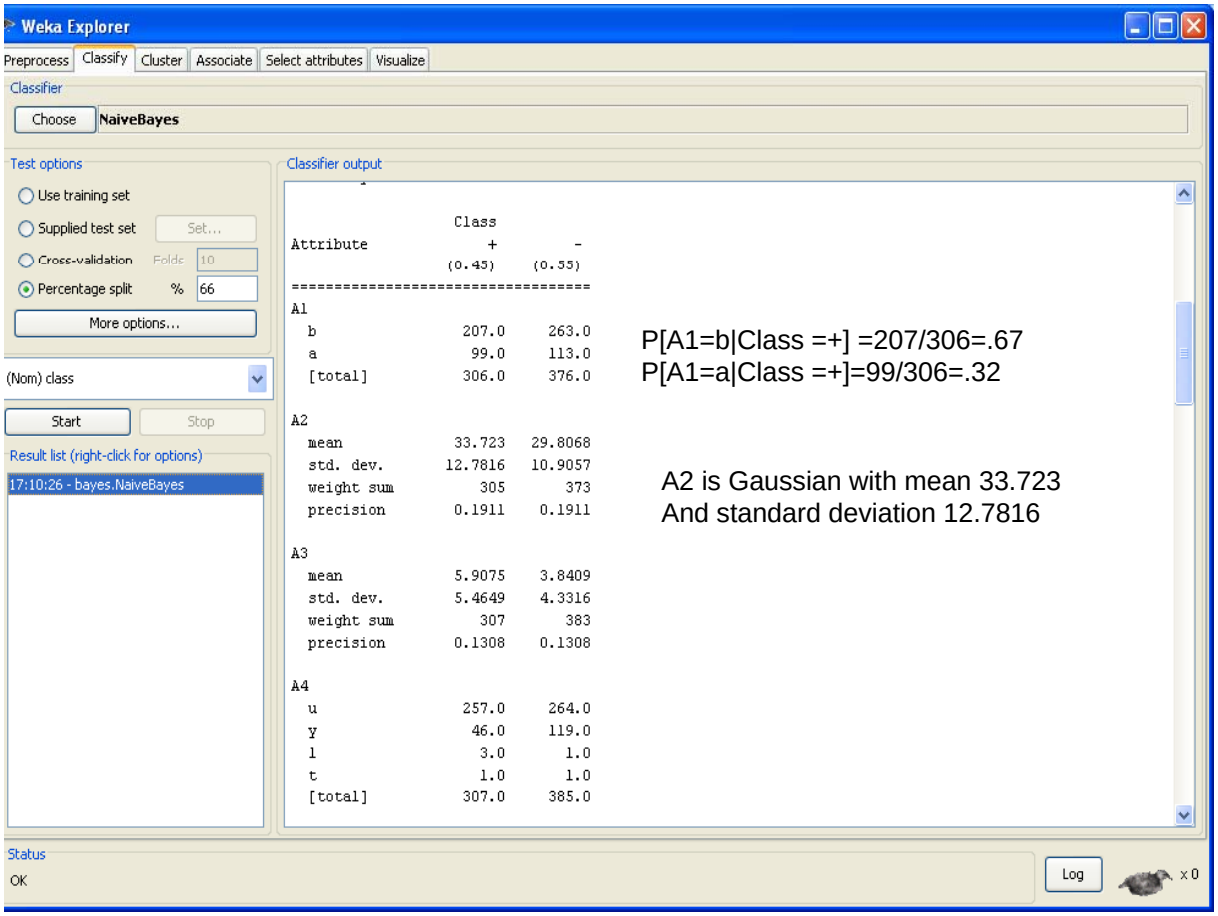
Practical Machine Learning Tools and Techniques

Examples Illustrating Testing and Evaluation Concepts (Sections 5.1-5.4, and 5.7)



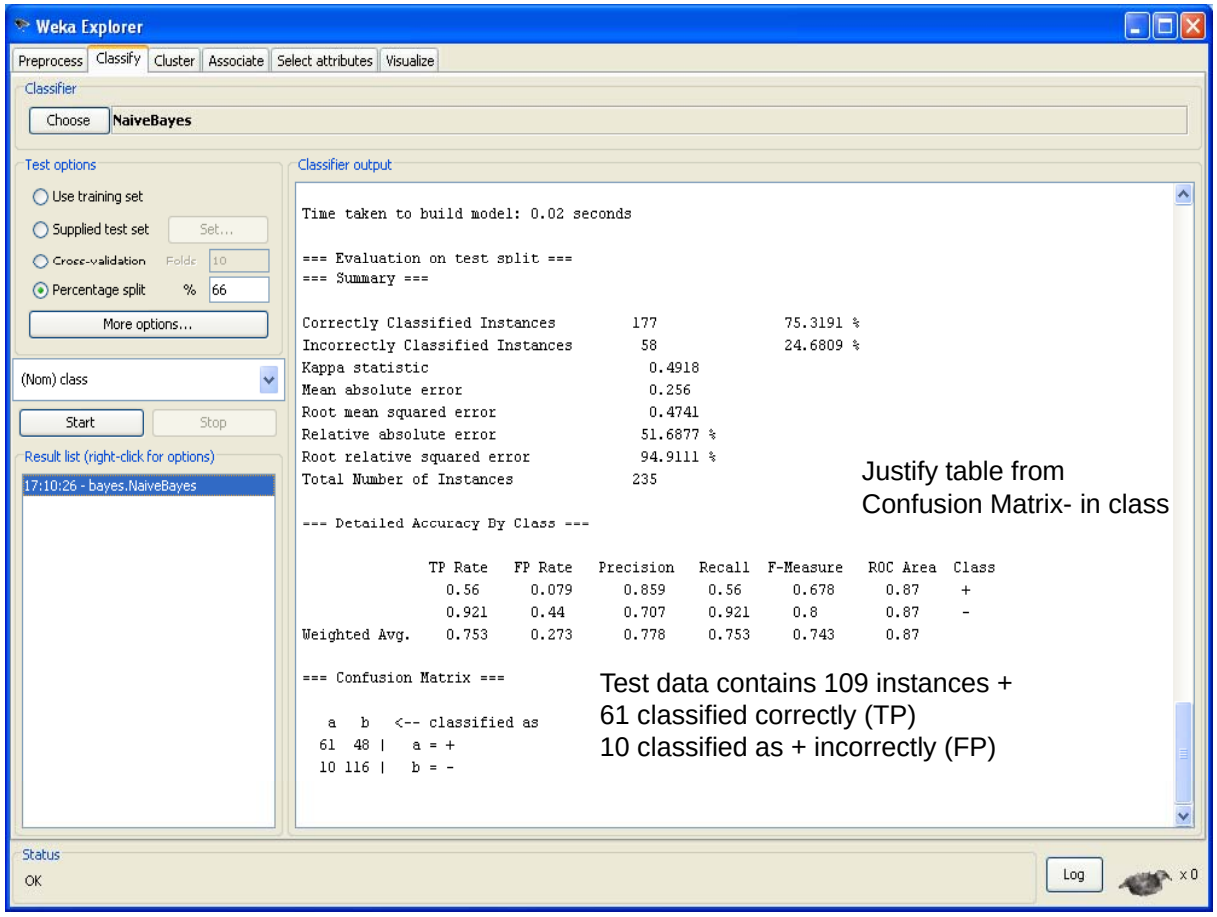

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: credit-rating
Instances: 690
Attributes: 16
Attributes
All
None
Invert
Pattern
No.
Name
4 A4
5 A5
6 A6
7 A7
8 A8
9 A9
10 A10
11 A11
12 A12
13 A13
14 A14
15 A15
16 class
Remove
Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No. Label Count
1 + 307
2 - 383
45% have credit-rating=+
55% have credit-rating=-
Class: class (Nom)
Visualize All
383
307
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
17:10:26 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: credit-rating
Instances: 690
Attributes: 16
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
class
Test mode: split 66.0% train, remainder test
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Training data = 455 instances
Testing data= 235 instances
Testing data stratified – roughly
45% are + and 55 are -
Status
OK
Log
x 0



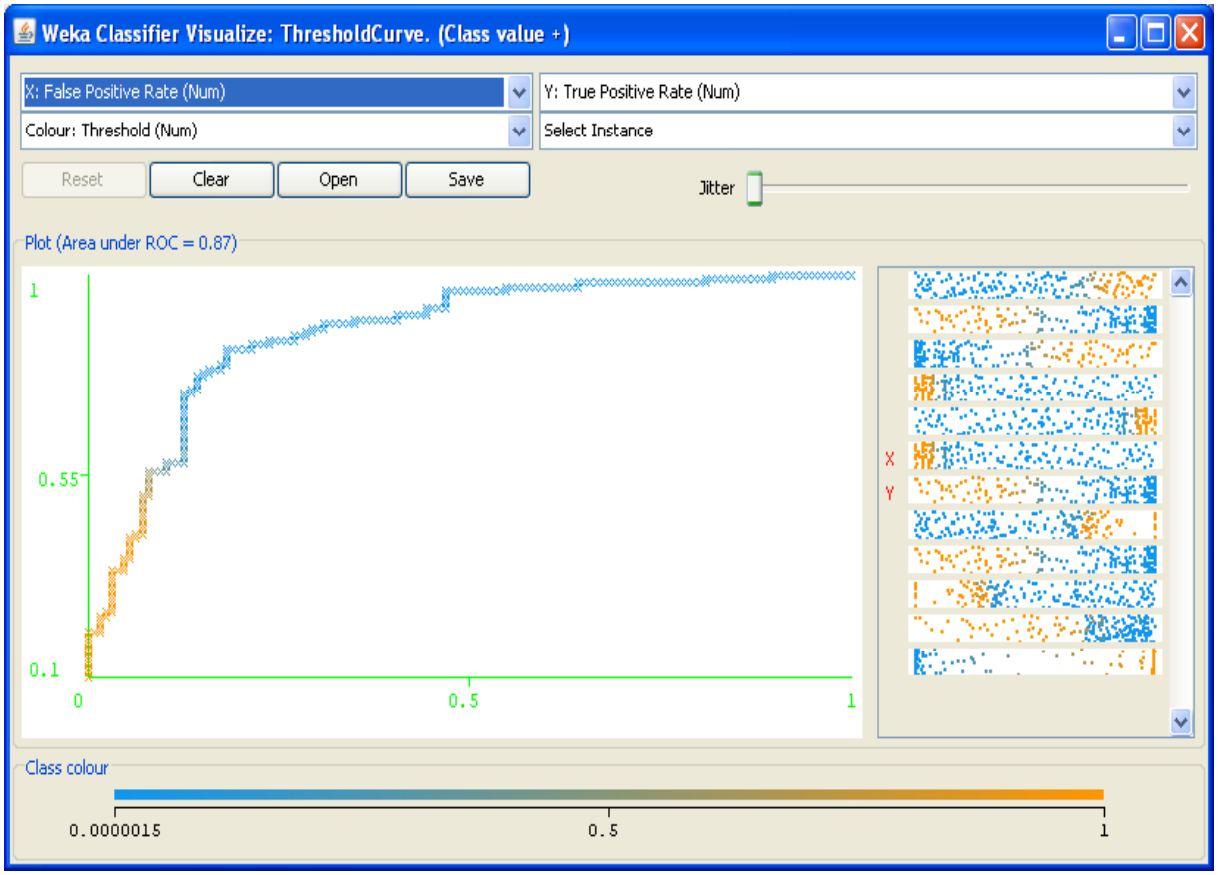
P[A1=b|Class =+] =207/306=.67
P[A1=a|Class =+]=99/306=.32

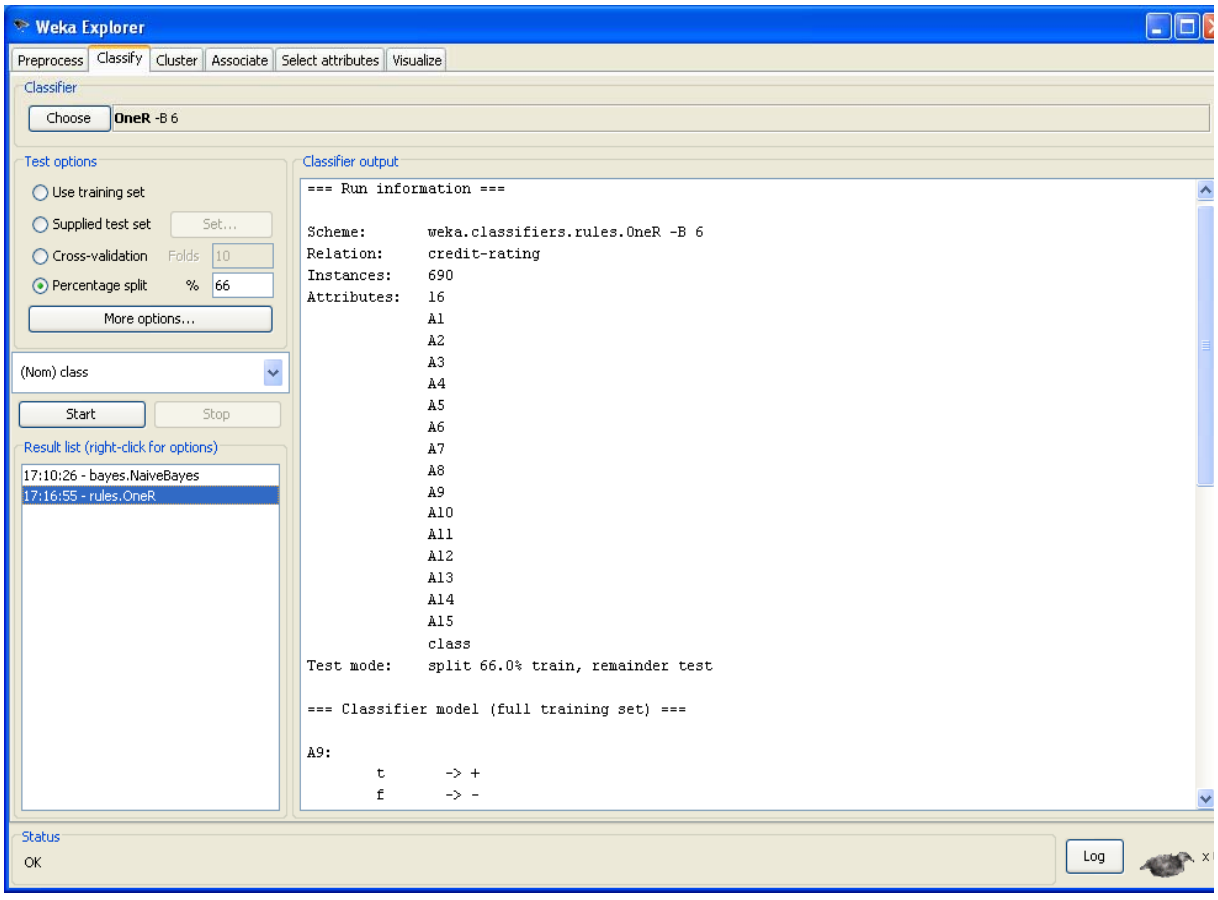
A2 is Gaussian with mean 33.723
And standard deviation 12.7816

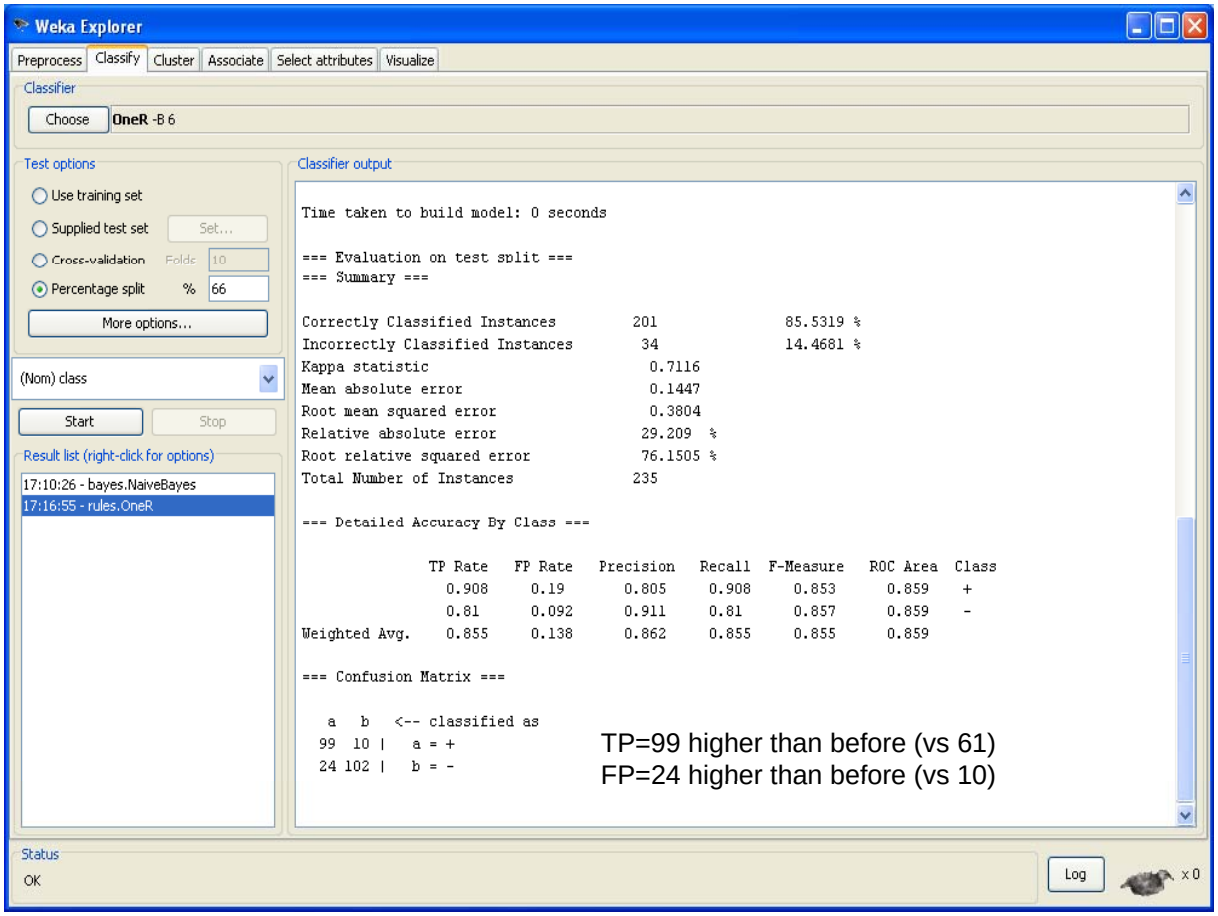


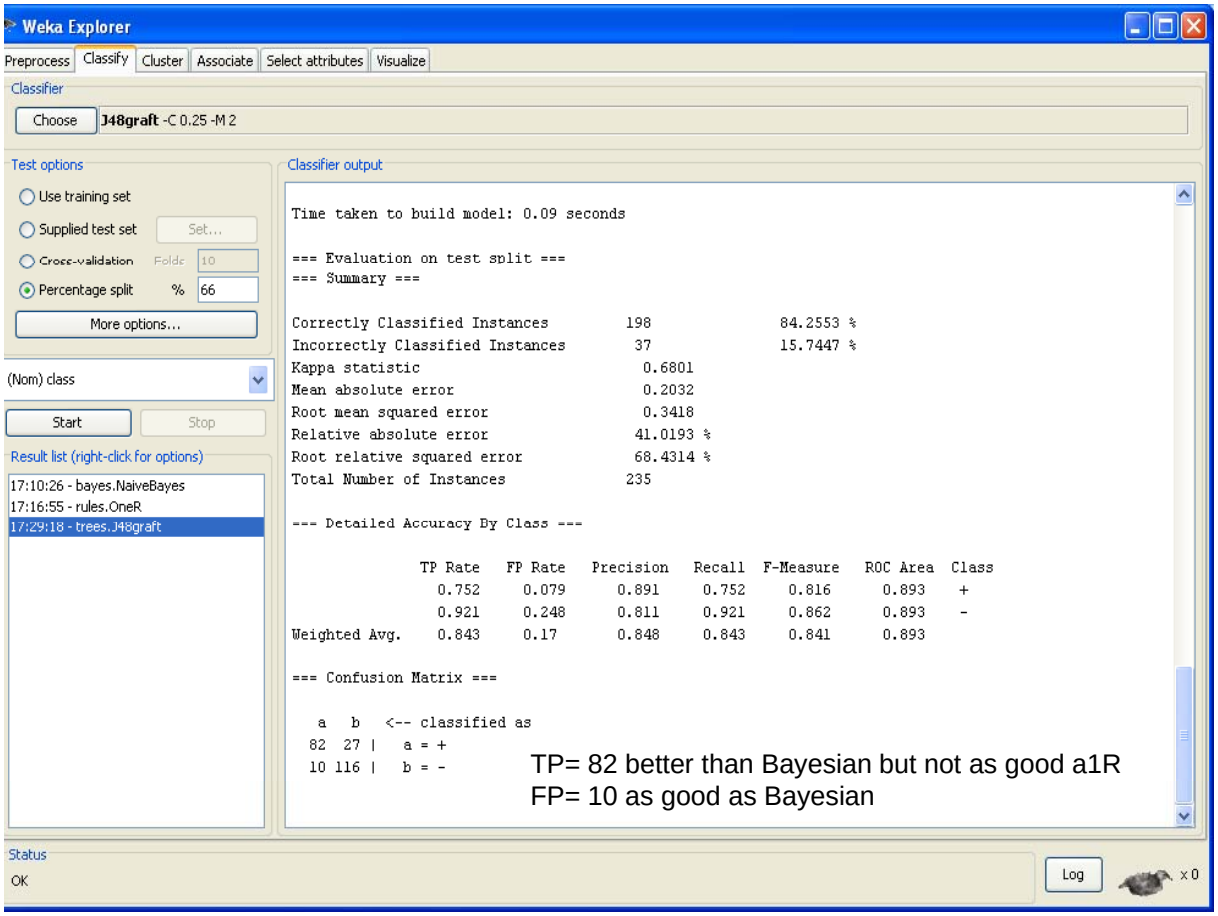
Justify table from Confusion Matrix- in class

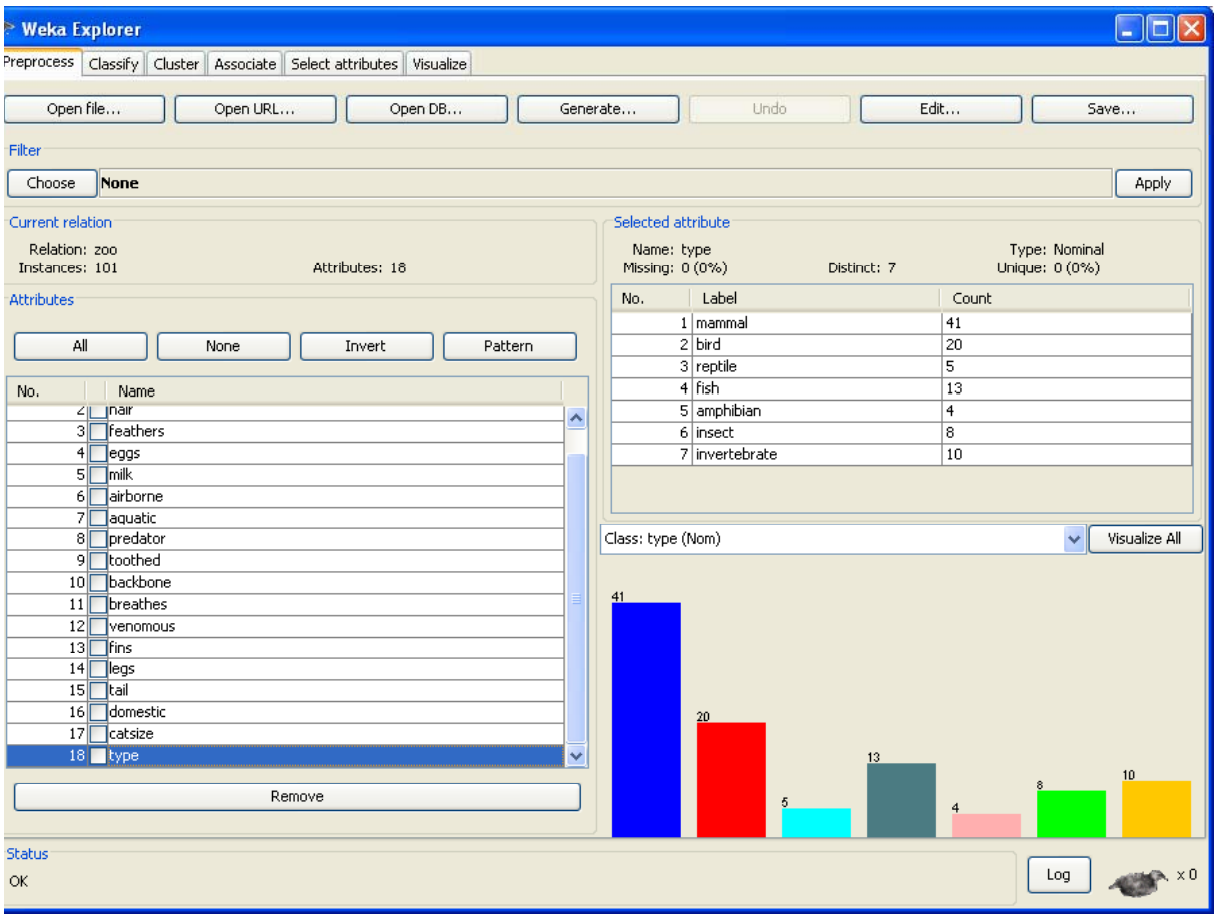
Test data contains 109 instances +
61 classified correctly (TP)
10 classified as + incorrectly (FP)

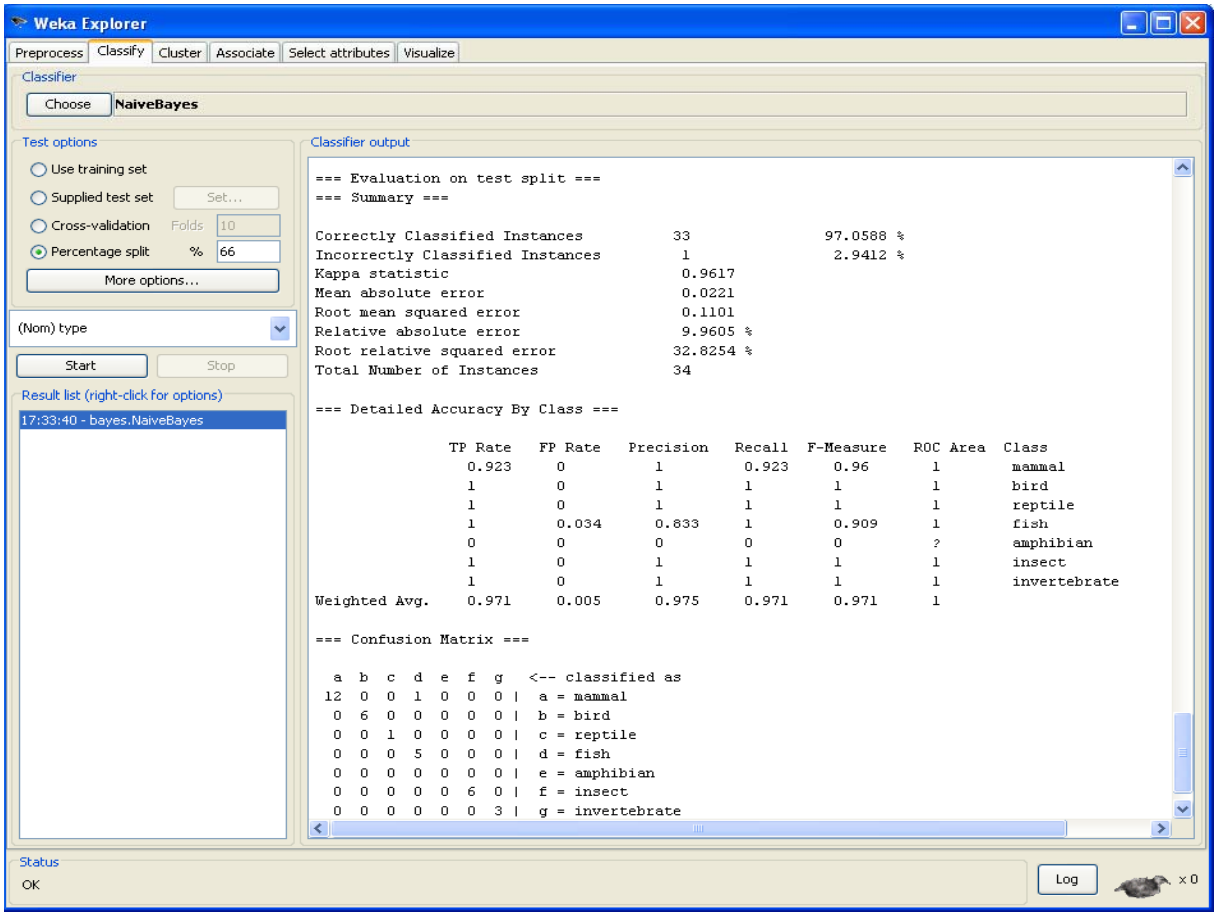

Weka Classifier Visualize: ThresholdCurve. (Class value +)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.87)
1
0.55
0.1
0
0.5
1
X
Y
Class colour
0.0000015
0.5
1

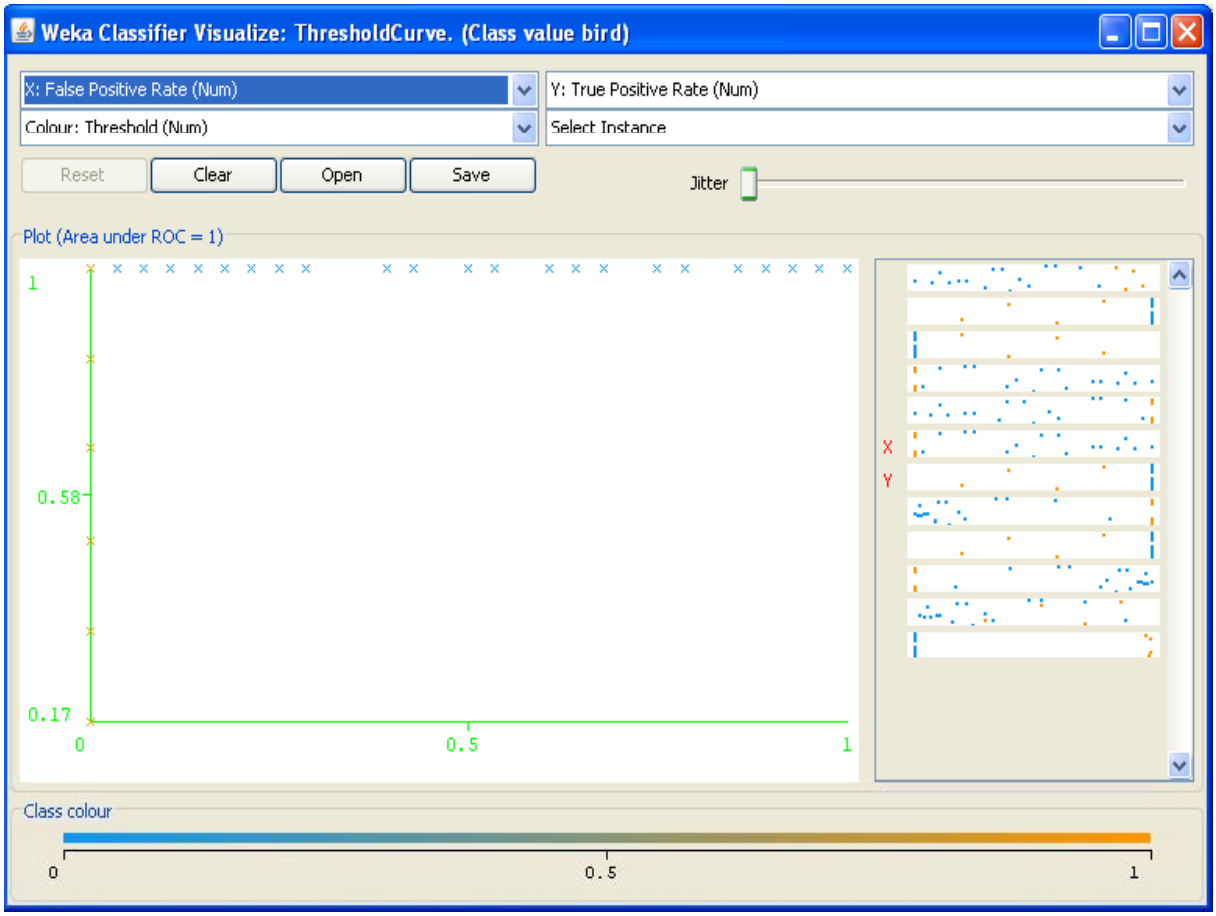

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
17:10:26 - bayes.NaiveBayes
17:16:55 - rules.OneR
Classifier output
=== Run information ===
Scheme: weka.classifiers.rules.OneR -B 6
Relation: credit-rating
Instances: 690
Attributes: 16
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
class
Test mode: split 66.0% train, remainder test
=== Classifier model (full training set) ===
A9:
t -> +
f -> -
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Time taken to build model: 0 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 201 85.5319 %
Incorrectly Classified Instances 34 14.4681 %
Kappa statistic 0.7116
Mean absolute error 0.1447
Root mean squared error 0.3804
Relative absolute error 29.209 %
Root relative squared error 76.1505 %
Total Number of Instances 235
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.908 0.19 0.805 0.908 0.853 0.859 +
0.81 0.092 0.911 0.81 0.857 0.859 -
Weighted Avg. 0.855 0.138 0.862 0.855 0.855 0.859
=== Confusion Matrix ===
a b <-- classified as
99 10 | a = +
24 102 | b = -
(Nom) class
Start
Stop
Result list (right-click for options)
17:10:26 - bayes.NaiveBayes
17:16:55 - rules.OneR
Status
OK
Log
x 0
TP=99 higher than before (vs 61)
FP=24 higher than before (vs 10)


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48graft -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Time taken to build model: 0.09 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 198 84.2553 %
Incorrectly Classified Instances 37 15.7447 %
Kappa statistic 0.6801
Mean absolute error 0.2032
Root mean squared error 0.3418
Relative absolute error 41.0193 %
Root relative squared error 68.4314 %
Total Number of Instances 235
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.752 0.079 0.891 0.752 0.816 0.893 +
0.921 0.248 0.811 0.921 0.862 0.893 -
Weighted Avg. 0.843 0.17 0.848 0.843 0.841 0.893
=== Confusion Matrix ===
a b <-- classified as
82 27 | a = +
10 116 | b = -
(Nom) class
Start
Stop
Result list (right-click for options)
17:10:26 - bayes.NaiveBayes
17:16:55 - rules.OneR
17:29:18 - trees.J48graft
Status
OK
Log
x 0
TP= 82 better than Bayesian but not as good a1R
FP= 10 as good as Bayesian


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: zoo
Instances: 101
Attributes: 18
Selected attribute
Name: type
Missing: 0 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
2 hair
3 feathers
4 eggs
5 milk
6 airborne
7 aquatic
8 predator
9 toothed
10 backbone
11 breathes
12 venomous
13 fins
14 legs
15 tail
16 domestic
17 catsize
18 type
Remove
No.
Label
Count
1 mammal 41
2 bird 20
3 reptile 5
4 fish 13
5 amphibian 4
6 insect 8
7 invertebrate 10
Class: type (Nom)
Visualize All
41
20
5
13
4
8
10
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) type
Start
Stop
Result list (right-click for options)
17:33:40 - bayes.NaiveBayes
Classifier output
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 33 97.0588 %
Incorrectly Classified Instances 1 2.9412 %
Kappa statistic 0.9617
Mean absolute error 0.0221
Root mean squared error 0.1101
Relative absolute error 9.9605 %
Root relative squared error 32.8254 %
Total Number of Instances 34
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.923 0 1 0.923 0.96 1 mammal
1 0 1 1 1 1 bird
1 0 1 1 1 1 reptile
1 0.034 0.833 1 0.909 1 fish
0 0 0 0 0 ? amphibian
1 0 1 1 1 1 insect
1 0 1 1 1 1 invertebrate
Weighted Avg. 0.971 0.005 0.975 0.971 0.971 1
=== Confusion Matrix ===
a b c d e f g <-- classified as
12 0 0 1 0 0 0 | a = mammal
0 6 0 0 0 0 0 | b = bird
0 0 1 0 0 0 0 | c = reptile
0 0 0 5 0 0 0 | d = fish
0 0 0 0 0 0 0 | e = amphibian
0 0 0 0 0 6 0 | f = insect
0 0 0 0 0 0 3 | g = invertebrate
Status
OK
Log
x 0


Weka Classifier Visualize: ThresholdCurve. (Class value bird)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 1)
1
0.58
0.17
0
0.5
1
X
Y
Class colour
0
0.5
1