CHAPTER 12

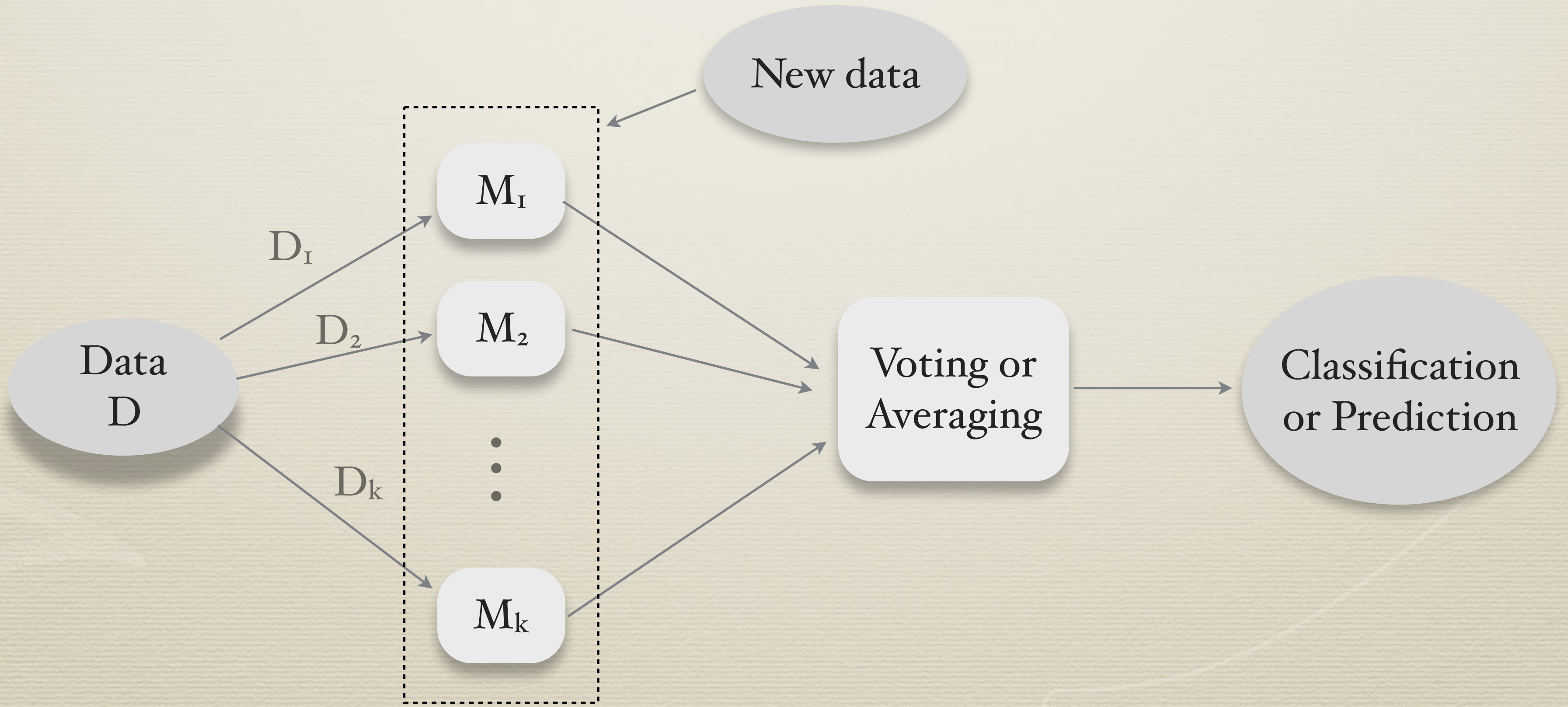
Ensemble Learning

Outline

- Introduction
- Bagging
- Boosting and AdaBoost
- Stacking
- Random Forests
- Randomization

Introduction (1/2)

- Group decision vs. Solitary intelligence
- Ensemble methods
  - Use a combination of models to increase accuracy
  - Combine a series of k learned models, M_1, M_2, …,M_k, with the aim of creating an improved model M*
  - Loss of interpretability



Introduction (2/2)

- Techniques
  - Bagging
    - Same data mining algorithm, different training datasets
    - Equal weight for each learned model
  - Boosting
    - Same data mining algorithm, different training datasets
    - Assign a weight to each learned model
  - Stacking
    - Different data mining algorithms
    - Two levels of learning
  - Random forests
    - A collection of CART-like trees
    - Randomness on training dataset and split selection of attributes

Bagging (1/9)

- 0.632 Bootstrap
  - A dataset of n instances is sampled n times, with replacement, to form the training set, and the remainder will be the test set

- Bootstrap AGGregation
- Supervised learning approach
- Analogy
  - Diagnosis based on multiple doctors' majority vote
- Training

63.2%

  - Given a set D of d tuples, at each iteration i, a training set D_i of d_i tuples is sampled with replacement from D
  - A classifier model M_i is learned for each training set D_i

Bagging (2/9)

- Classification: voting
  - Each classifier M_i returns its class prediction
  - The bagged classifier M* counts the votes and assigns the class with the most votes to X
- Prediction: averaging
  - Taking the average value of each prediction for a given test tuple
- Accuracy
  - Often significant better than a single classifier
    - Unstable classifiers: trees, neural nets
  - Proved improved accuracy in prediction

decision stump

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

round 1: x<=0.35 --> y=1
x>0.35 --> y=-1

x	0.1	0.2	0.3	0.4	0.5	0.8	0.9	1	1	1
y	1	1	1	-1	-1	1	1	1	1	1

round 2: x<=0.65 --> y=1
x>0.65 --> y=1

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

round 3: x<=0.35 --> y=1
x>0.35 --> y=-1

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

round 4: x<=0.3 --> y=1
x>0.3 --> y=-1

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

round 5: x<=0.35 --> y=1
x>0.35 --> y=-1

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

round 6: x<=0.75 --> y=-1
x>0.75 --> y=1

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

round 7: x<=0.75 --> y=-1
x>0.75 --> y=1

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

round 8: x<=0.75 --> y=-1
x>0.75 --> y=1

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

round 9: x<=0.75 --> y=-1
x>0.75 --> y=1

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

round 10: x<=0.05 --> y=-1
x>0.05 --> y=1

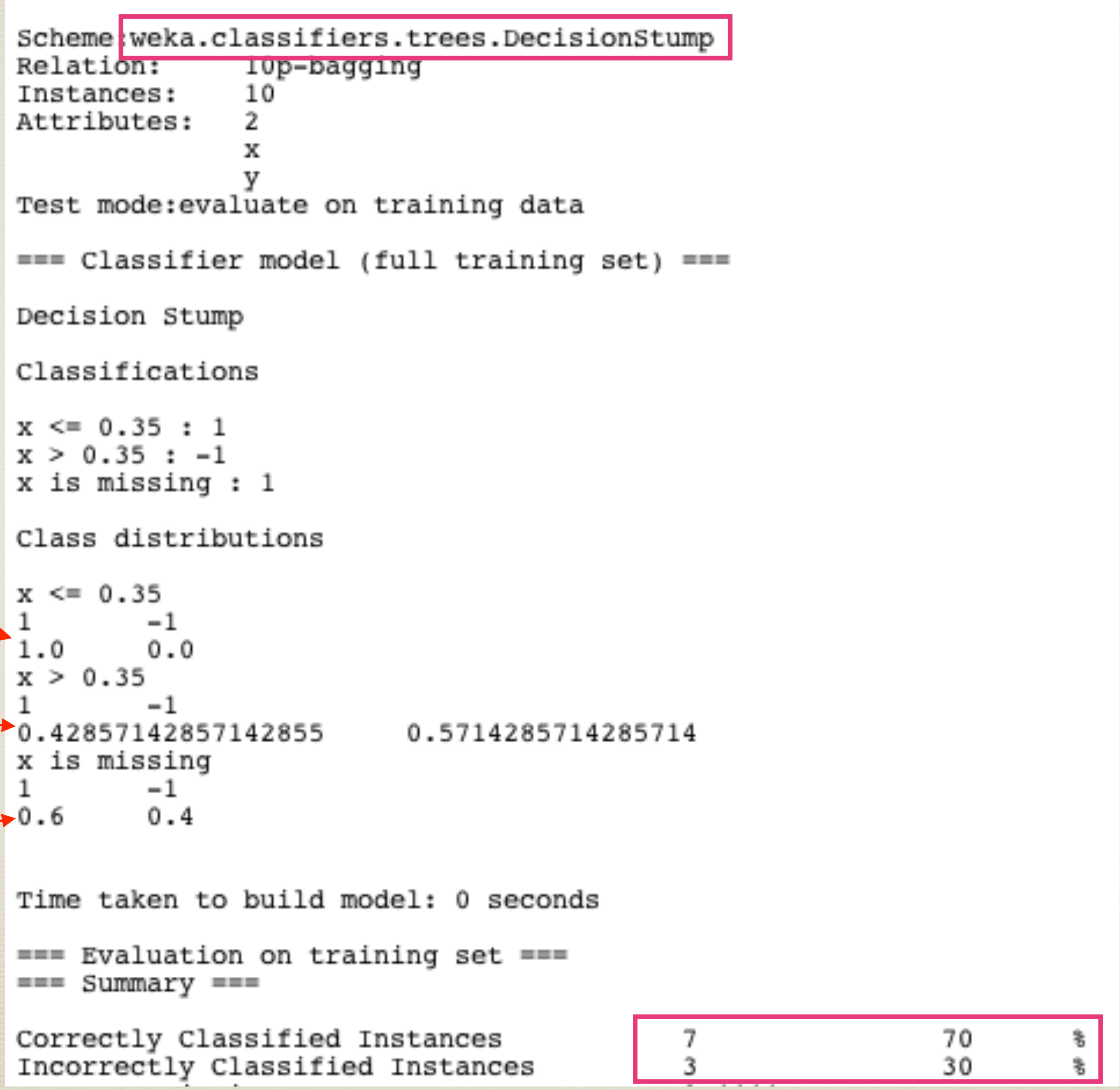
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

Without bagging: 70% precision rate

run	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	0.7
2	1	1	1	1	1	1	1	1	1	1	0.6
3	1	1	1	-1	-1	-1	-1	-1	-1	-1	0.7
4	1	1	1	-1	-1	-1	-1	-1	-1	-1	0.7
5	1	1	1	-1	-1	-1	-1	-1	-1	-1	0.7
6	-1	-1	-1	-1	-1	-1	-1	1	1	1	0.7
7	-1	-1	-1	-1	-1	-1	-1	1	1	1	0.7
8	-1	-1	-1	-1	-1	-1	-1	1	1	1	0.7
9	-1	-1	-1	-1	-1	-1	-1	1	1	1	0.7
10	1	1	1	1	1	1	1	1	1	1	0.6
sum	2	2	2	-6	-6	-6	-6	2	2	2	
y’	1	1	1	-1	-1	-1	-1	1	1	1	
y	1	1	1	-1	-1	-1	-1	1	1	1	

100% precision rate

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

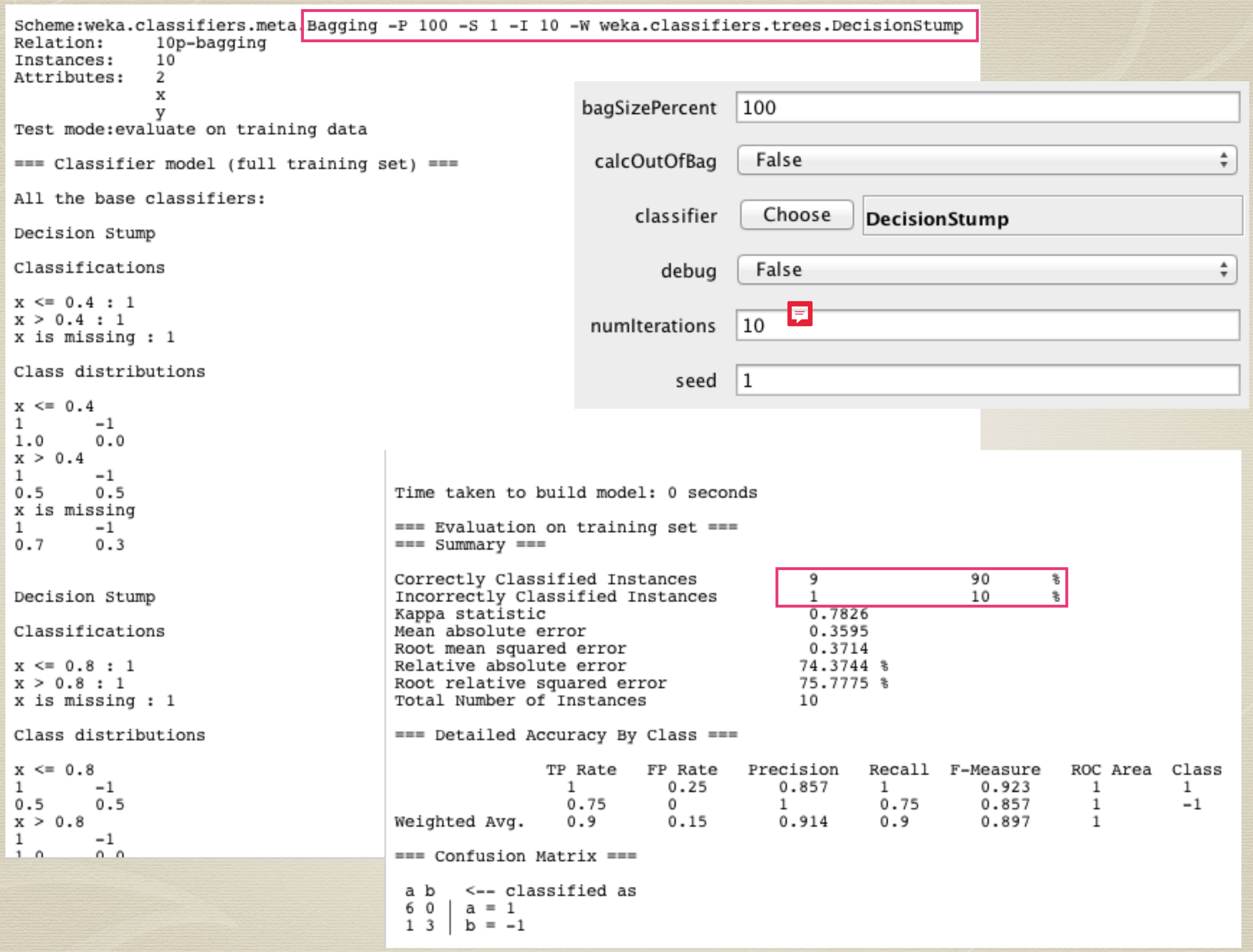


```
Scheme:weka.classifiers.trees.DecisionStump
Relation:     10p-bagging
Instances:    10
Attributes:   2
              x
              y
Test mode:evaluate on training data

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

Decision Stump

Classifications

x <= 0.35 : 1
x > 0.35 : -1
x is missing : 1

Class distributions

x <= 0.35
1       -1
1.0     0.0
x > 0.35
1       -1
0.42857142857142855     0.5714285714285714
x is missing
1       -1
0.6     0.4


Time taken to build model: 0 seconds

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

Correctly Classified Instances           7               70      %
Incorrectly Classified Instances         3               30      %
```

3/3

3/7

6/10



```
Scheme:weka.classifiers.meta.Bagging -P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump
Relation:     10p-bagging
Instances:    10
Attributes:   2
              x
              y
Test mode:evaluate on training data

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

All the base classifiers:

Decision Stump

Classifications

x <= 0.4 : 1
x > 0.4 : 1
x is missing : 1

Class distributions

x <= 0.4
1       -1
1.0     0.0
x > 0.4
1       -1
0.5     0.5
x is missing
1       -1
0.7     0.3


Decision Stump

Classifications

x <= 0.8 : 1
x > 0.8 : 1
x is missing : 1

Class distributions

x <= 0.8
1       -1
0.5     0.5
x > 0.8
1       -1
1.0     0.0
```

bagSizePercent: 100

calcOutOfBag: False

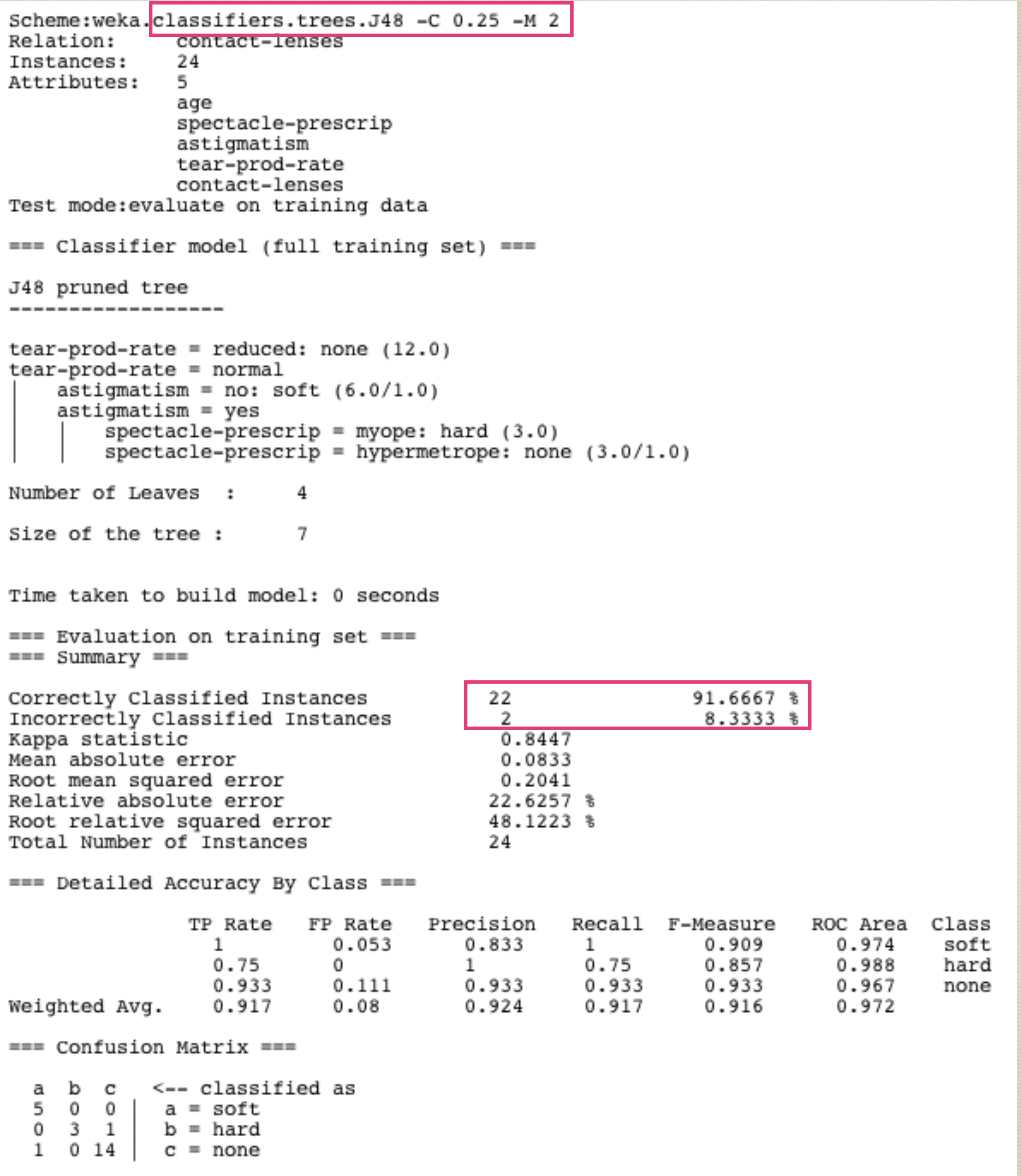
classifier: DecisionStump

debug: False

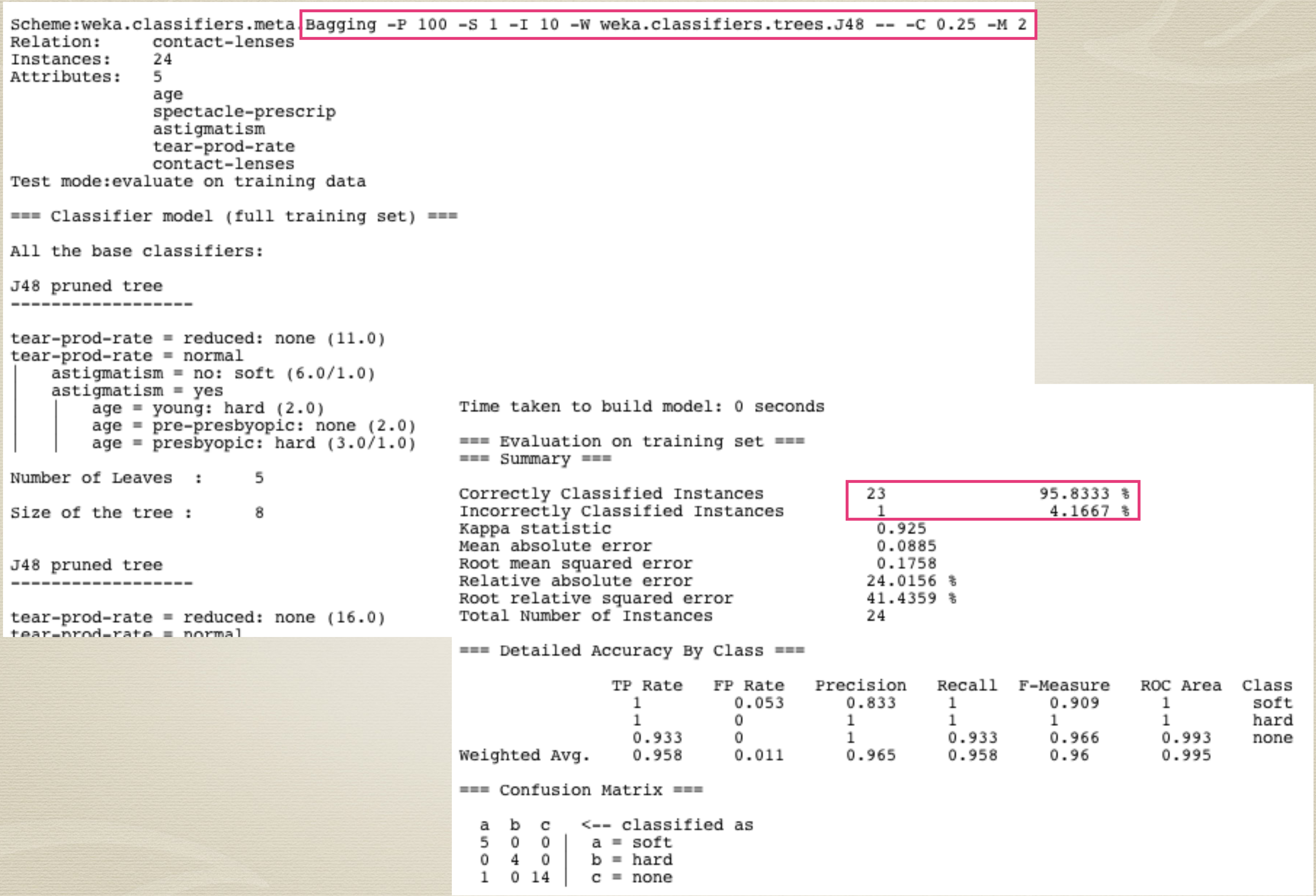
numIterations: 10

seed: 1

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances           9               90      %
Incorrectly Classified Instances         1               10      %
Kappa statistic                          0.7826
Mean absolute error                      0.3595
Root mean squared error                  0.3714
Relative absolute error                 74.3744 %
Root relative squared error             75.7775 %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.25      0.857     1         0.923       1        1
                 0.75      0         1         0.75      0.857       1        -1
Weighted Avg.    0.9       0.15      0.914     0.9       0.897       1

=== Confusion Matrix ===

 a b   <-- classified as
 6 0 | a = 1
 1 3 | b = -1
```

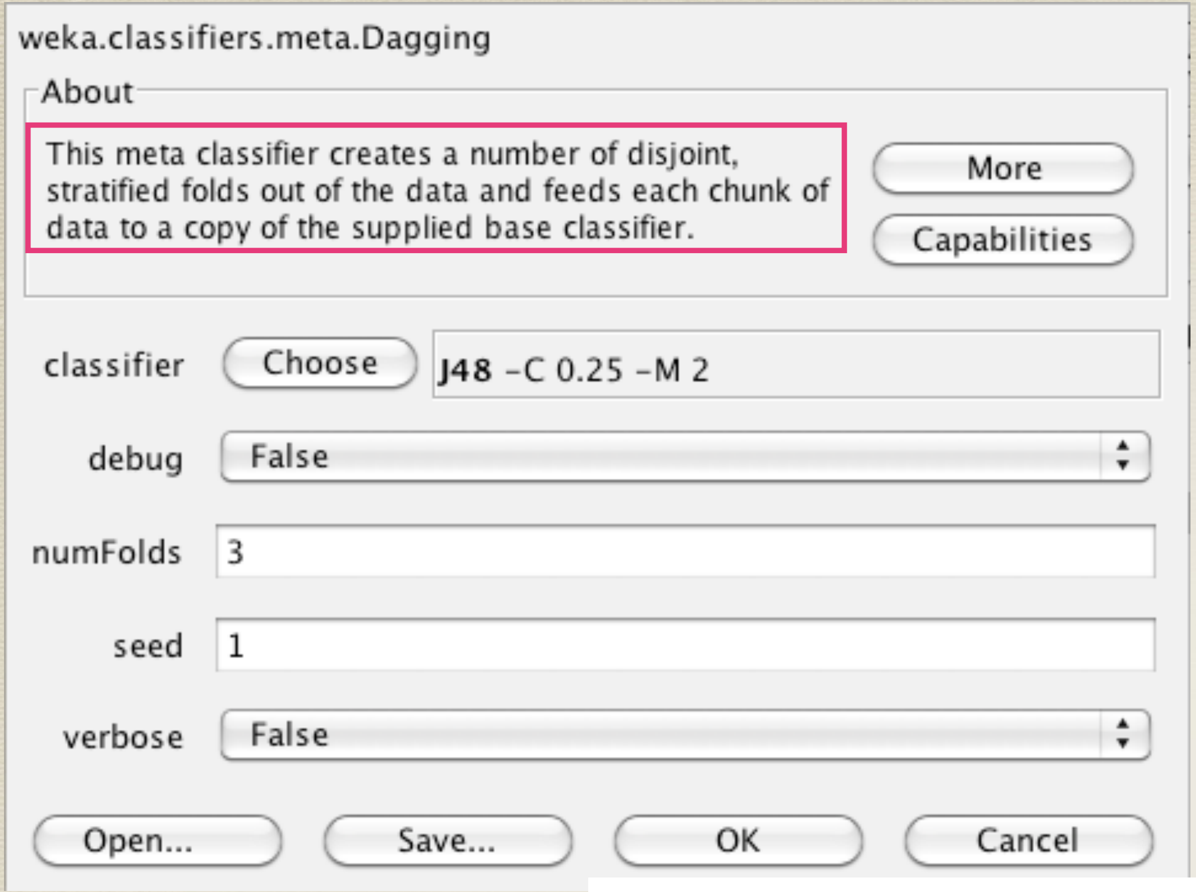


```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     contact-lenses
Instances:    24
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:evaluate on training data

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

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

tear-prod-rate = reduced: none (12.0)
tear-prod-rate = normal
|   astigmatism = no: soft (6.0/1.0)
|   astigmatism = yes
|   |   spectacle-prescrip = myope: hard (3.0)
|   |   spectacle-prescrip = hypermetrope: none (3.0/1.0)

Number of Leaves  :      4

Size of the tree :       7


Time taken to build model: 0 seconds

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

Correctly Classified Instances          22               91.6667 %
Incorrectly Classified Instances         2                8.3333 %
Kappa statistic                          0.8447
Mean absolute error                      0.0833
Root mean squared error                  0.2041
Relative absolute error                 22.6257 %
Root relative squared error             48.1223 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.053     0.833      1         0.909      0.974    soft
                 0.75      0         1          0.75      0.857      0.988    hard
                 0.933     0.111     0.933      0.933     0.933      0.967    none
Weighted Avg.    0.917     0.08      0.924      0.917     0.916      0.972

=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  0  0 |  a = soft
  0  3  1 |  b = hard
  1  0 14 |  c = none
```



```
Scheme:weka.classifiers.meta.Bagging -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Relation:     contact-lenses
Instances:    24
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:evaluate on training data

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

All the base classifiers:

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

tear-prod-rate = reduced: none (11.0)
tear-prod-rate = normal
|   astigmatism = no: soft (6.0/1.0)
|   astigmatism = yes
|   |   age = young: hard (2.0)
|   |   age = pre-presbyopic: none (2.0)
|   |   age = presbyopic: hard (3.0/1.0)

Number of Leaves  :     5

Size of the tree :      8


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

tear-prod-rate = reduced: none (16.0)
tear-prod-rate = normal
```

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          23               95.8333 %
Incorrectly Classified Instances         1                4.1667 %
Kappa statistic                          0.925
Mean absolute error                      0.0885
Root mean squared error                  0.1758
Relative absolute error                 24.0156 %
Root relative squared error             41.4359 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.053      0.833      1         0.909      1        soft
                 1         0          1          1         1          1        hard
                 0.933     0          1          0.933     0.966      0.993    none
Weighted Avg.    0.958     0.011      0.965      0.958     0.96       0.995

=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  0  0 |  a = soft
  0  4  0 |  b = hard
  1  0 14 |  c = none
```



weka.classifiers.meta.Dagging

About

This meta classifier creates a number of disjoint, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base classifier.

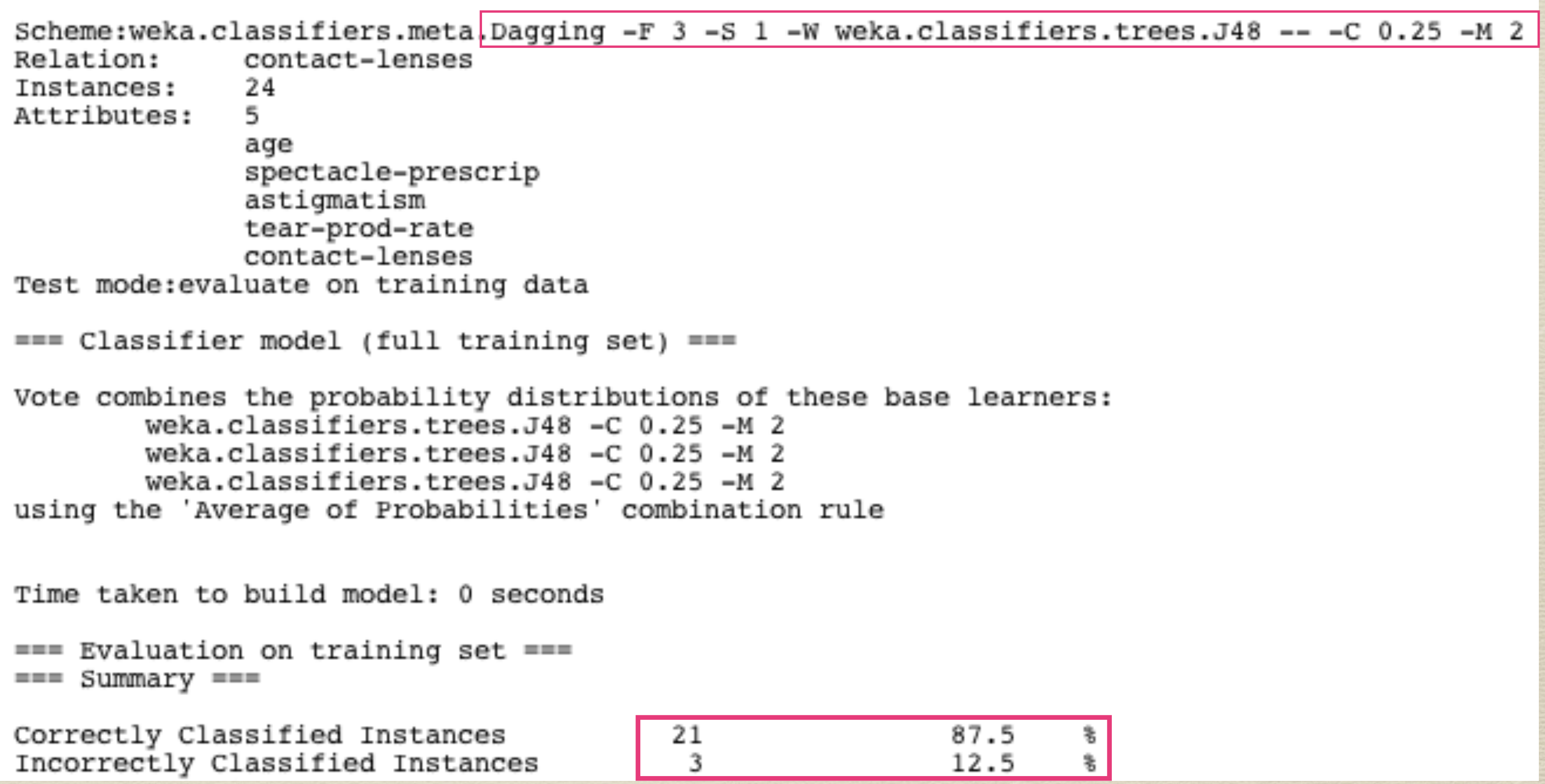
classifier: J48 -C 0.25 -M 2

debug: False

numFolds: 3

seed: 1

verbose: False



```
Scheme:weka.classifiers.meta.Dagging -F 3 -S 1 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Relation:     contact-lenses
Instances:    24
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:evaluate on training data

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

Vote combines the probability distributions of these base learners:
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
using the 'Average of Probabilities' combination rule


Time taken to build model: 0 seconds

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

Correctly Classified Instances          21               87.5    %
Incorrectly Classified Instances         3               12.5    %
```

Boosting (1/6)

- Analogy
  - Consult several doctors, based on a combination of weighted diagnoses-weight assigned based on the previous diagnosis accuracy
- How boosting works?
  - Weights are assigned to each training tuple
  - A series of k classifiers is iteratively learned
  - After a classified M_i is learned, the weights are updated to allow the subsequent classifier, M_{i+1}, to pay more attention to the training tuples that were misclassified by M_i
  - The final M* combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy

Boosting (2/6)

- The boosting algorithm can be extended for the prediction of continuous values
- Comparing with bagging
  - Boosting tends to achieve greater accuracy, but it also risks overfitting the model to misclassified data

Adaboost (Freund and Schapire, 1997)

Algorithm: AdaBoost

Input:

D: a set of d class-labeled training tuples;
k: the number of rounds (one classifier is generated per round);
a classification learning scheme;

Output: a composite model

Method:

```
(1)  initialize the weight of each tuple in D to 1/d;
(2)  for i=1 to k do
(3)      sample D with replacement according to the tuple weights to obtain D_i, of size d;
(4)      use training set D_i to derive a model M_i;
(5)      compute error(M_i), the error rate of M_i, over D_i;
(6)      if error(M_i) > 0.5 then
(7)          go back to step 3 and try again;
(8)      endif
(9)      for each tuple D_i in that was correctly classified do
(10)         multiply the weight of the tuple by error(M_i)/(1-error(M_i));
(11)     normalize the weight of each tuple;
(12) endfor
```

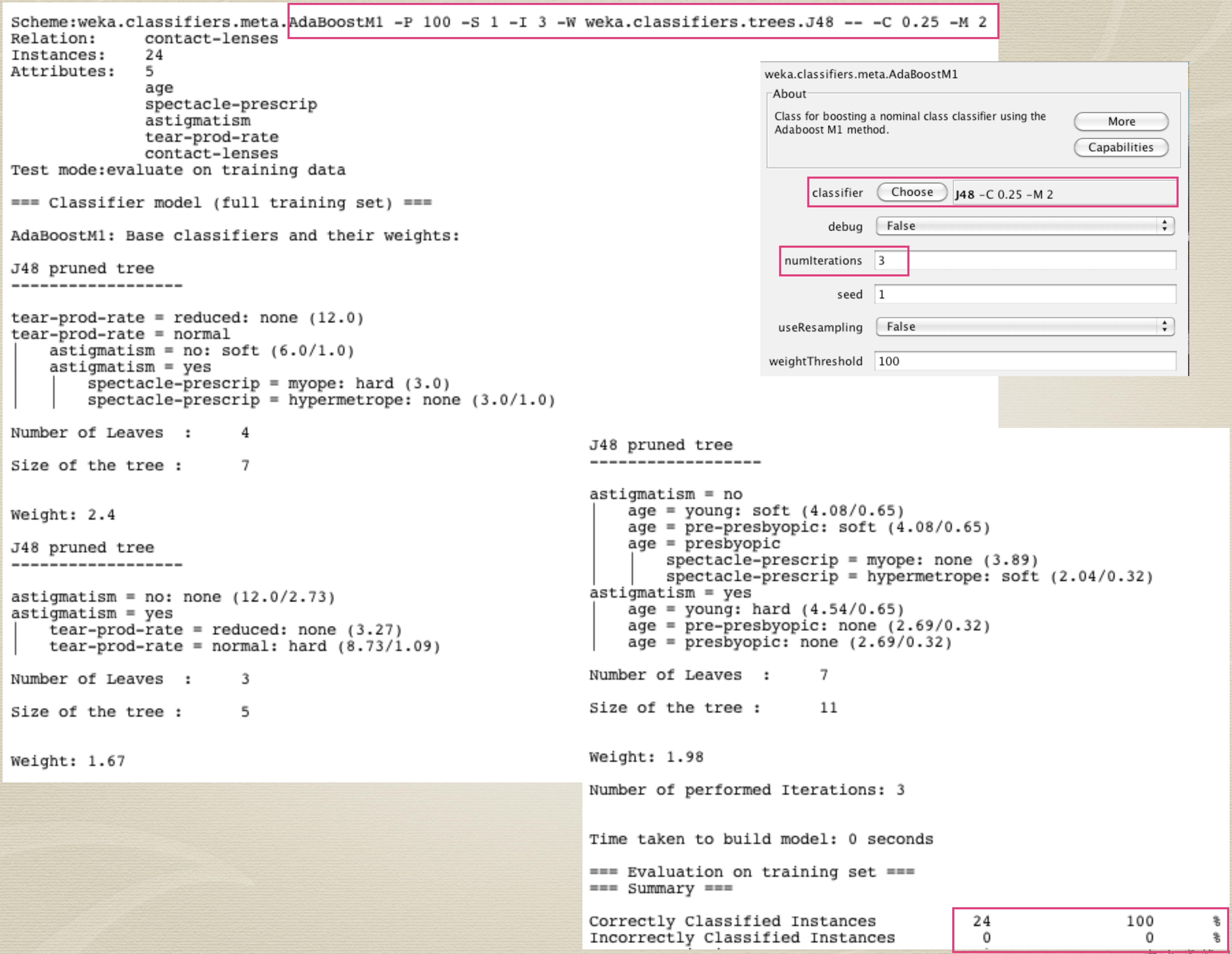
$$new\_weight * \frac{1}{\sum new\_weight}$$

To use the ensemble to classify tuple, X:

```
(1)  initialize weight of each class to 0;
(2)  for i=1 to k do
(3)      $w_i = \log \frac{1-error(M_i)}{error(M_i)}$
(4)      c = M_i(X);
(5)      add w_i to weight for class c;
(6)  endfor
(7)  return the class with the largest weight;
```

$$error(M_i) = \sum_{j}^{d} w_j \times err(\mathbf{X_j})$$

err(X_j)=1 if misclassified;
otherwise, err(X_j)=0



```
Scheme:weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 3 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Relation:     contact-lenses
Instances:    24
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:evaluate on training data

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

AdaBoostM1: Base classifiers and their weights:

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

tear-prod-rate = reduced: none (12.0)
tear-prod-rate = normal
|   astigmatism = no: soft (6.0/1.0)
|   astigmatism = yes
|   |   spectacle-prescrip = myope: hard (3.0)
|   |   spectacle-prescrip = hypermetrope: none (3.0/1.0)

Number of Leaves  :     4

Size of the tree :      7


Weight: 2.4

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

astigmatism = no: none (12.0/2.73)
astigmatism = yes
|   tear-prod-rate = reduced: none (3.27)
|   tear-prod-rate = normal: hard (8.73/1.09)

Number of Leaves  :     3

Size of the tree :      5


Weight: 1.67

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

astigmatism = no
|   age = young: soft (4.08/0.65)
|   age = pre-presbyopic: soft (4.08/0.65)
|   age = presbyopic
|   |   spectacle-prescrip = myope: none (3.89)
|   |   spectacle-prescrip = hypermetrope: soft (2.04/0.32)
astigmatism = yes
|   age = young: hard (4.54/0.65)
|   age = pre-presbyopic: none (2.69/0.32)
|   age = presbyopic: none (2.69/0.32)

Number of Leaves  :     7

Size of the tree :      11


Weight: 1.98

Number of performed Iterations: 3


Time taken to build model: 0 seconds

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

Correctly Classified Instances          24              100      %
Incorrectly Classified Instances         0                0      %
```

weka.classifiers.meta.AdaBoostM1

About

Class for boosting a nominal class classifier using the Adaboost M1 method.

classifier: J48 -C 0.25 -M 2

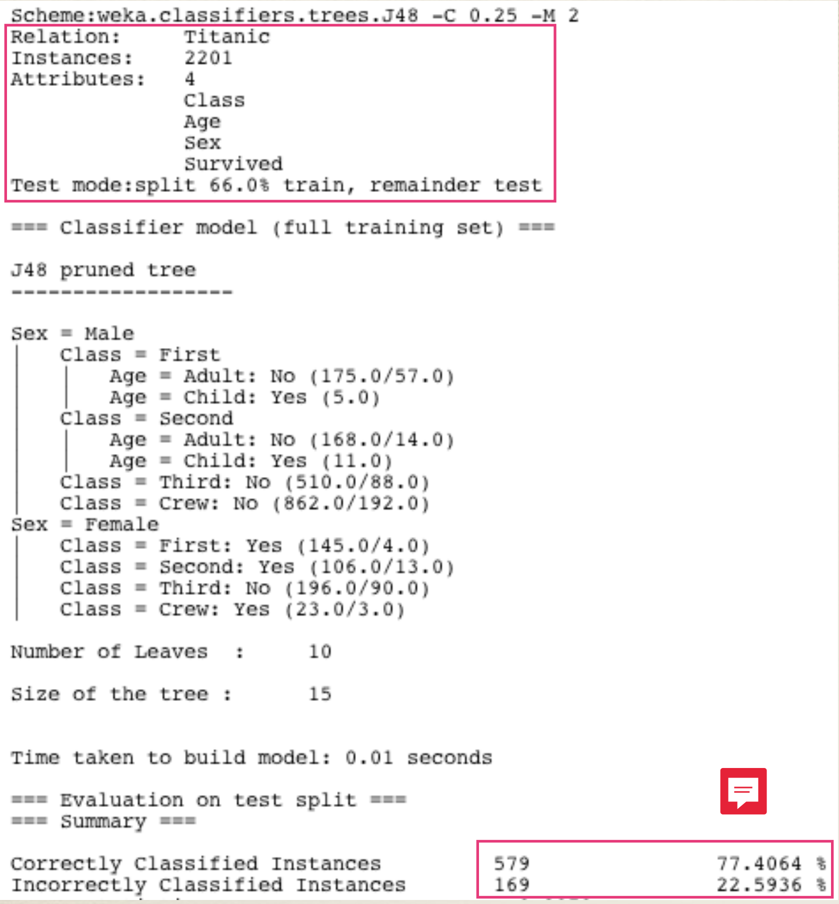
debug: False

numIterations: 3

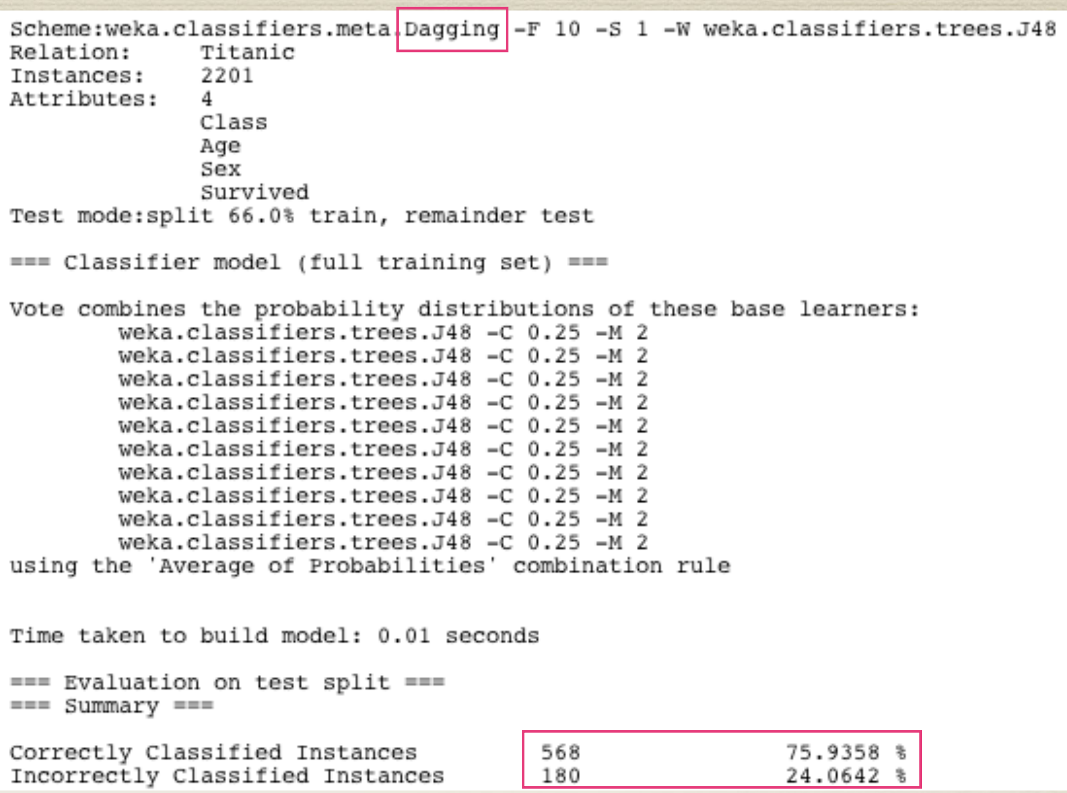
seed: 1

useResampling: False

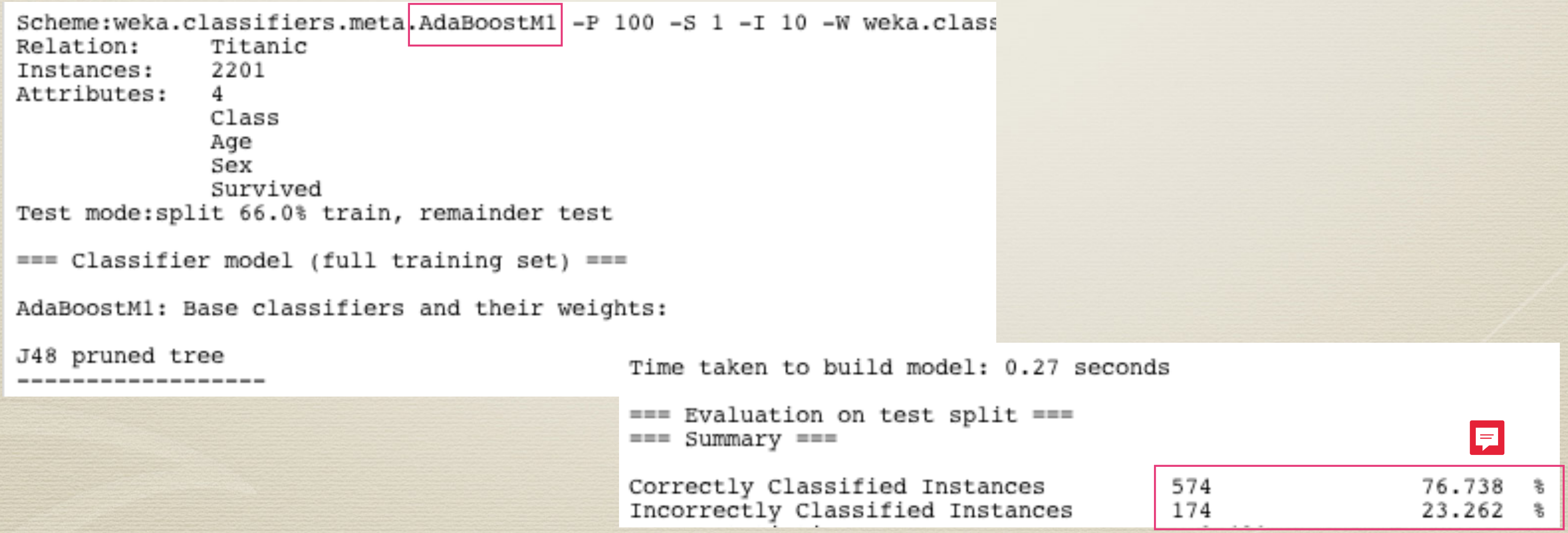
weightThreshold: 100



```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Titanic
Instances:    2201
Attributes:   4
              Class
              Age
              Sex
              Survived
Test mode:split 66.0% train, remainder test

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

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

Sex = Male
|   Class = First
|   |   Age = Adult: No (175.0/57.0)
|   |   Age = Child: Yes (5.0)
|   Class = Second
|   |   Age = Adult: No (168.0/14.0)
|   |   Age = Child: Yes (11.0)
|   Class = Third: No (510.0/88.0)
|   Class = Crew: No (862.0/192.0)
Sex = Female
|   Class = First: Yes (145.0/4.0)
|   Class = Second: Yes (106.0/13.0)
|   Class = Third: No (196.0/90.0)
|   Class = Crew: Yes (23.0/3.0)

Number of Leaves  :     10

Size of the tree :      15


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         579               77.4064 %
Incorrectly Classified Instances       169               22.5936 %
```



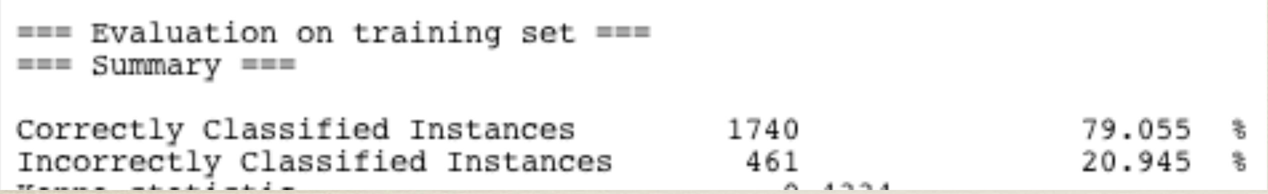
```
Scheme:weka.classifiers.meta.Dagging -F 10 -S 1 -W weka.classifiers.trees.J48
Relation:     Titanic
Instances:    2201
Attributes:   4
              Class
              Age
              Sex
              Survived
Test mode:split 66.0% train, remainder test

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

Vote combines the probability distributions of these base learners:
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
        weka.classifiers.trees.J48 -C 0.25 -M 2
using the 'Average of Probabilities' combination rule


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         568               75.9358 %
Incorrectly Classified Instances       180               24.0642 %
```



```
Scheme:weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.class
Relation:     Titanic
Instances:    2201
Attributes:   4
              Class
              Age
              Sex
              Survived
Test mode:split 66.0% train, remainder test

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

AdaBoostM1: Base classifiers and their weights:

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

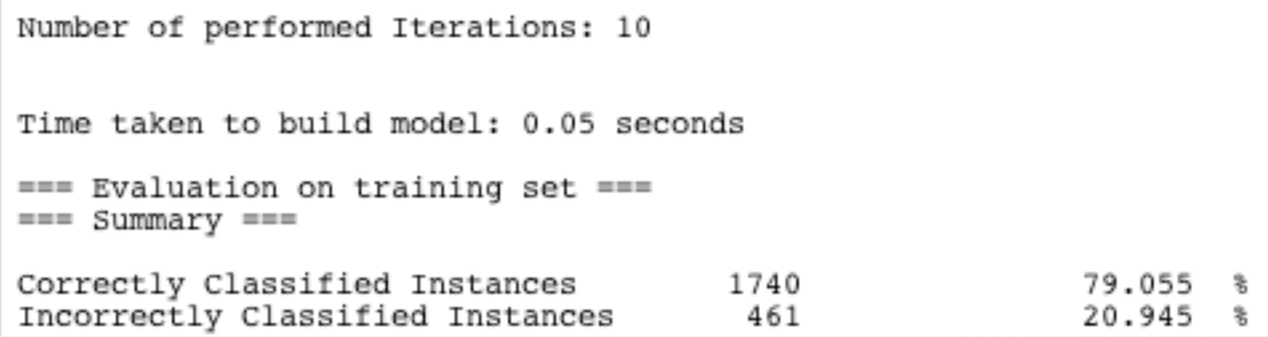
```
Time taken to build model: 0.27 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         574               76.738  %
Incorrectly Classified Instances       174               23.262  %
```

J48



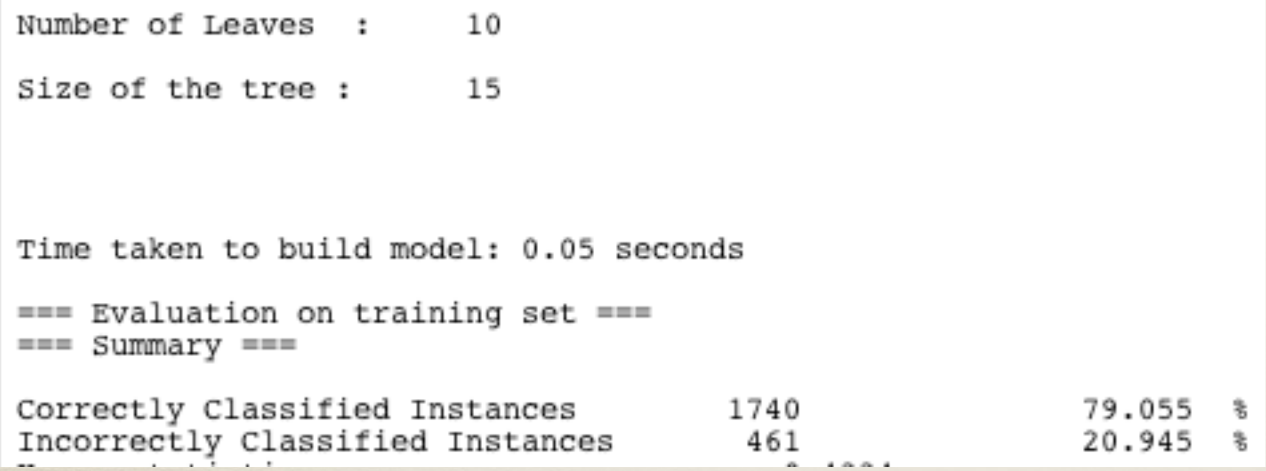
```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
```

AdaBoostM1



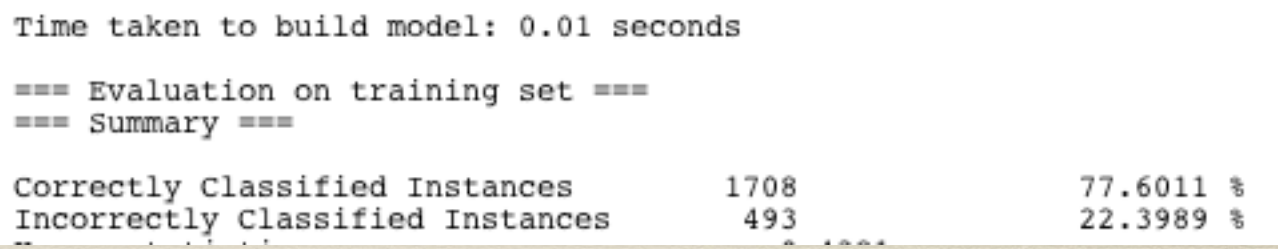
```
Number of performed Iterations: 10


Time taken to build model: 0.05 seconds

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

Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
```

Bagging



```
Number of Leaves  :     10

Size of the tree :      15



Time taken to build model: 0.05 seconds

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

Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
```

Dagging



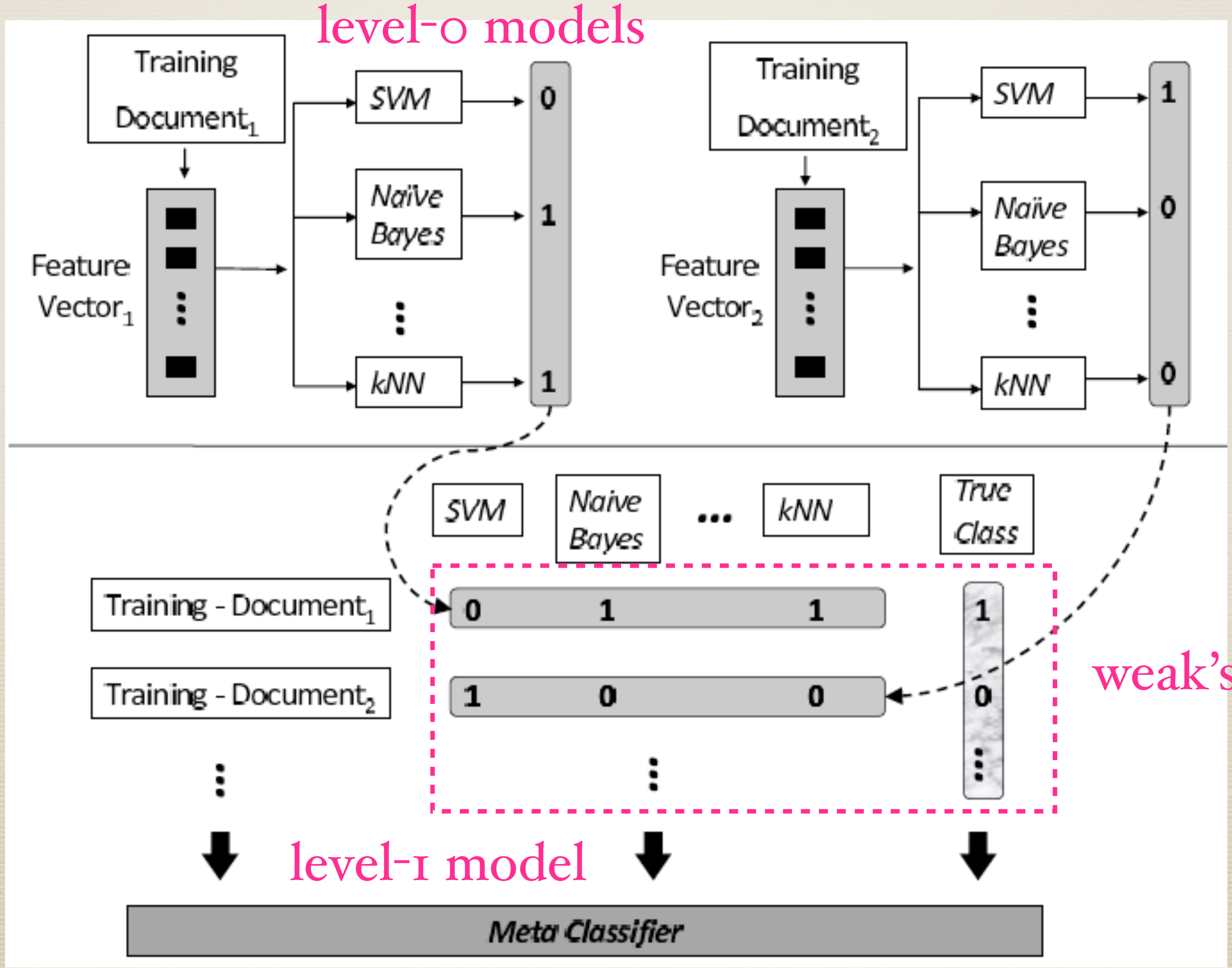
```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances        1708               77.6011 %
Incorrectly Classified Instances       493               22.3989 %
```

Stacking (1/3)

- Stacked generalization
- Less widely used than bagging and boosting
- Built by different learning algorithms

example

level-0 models

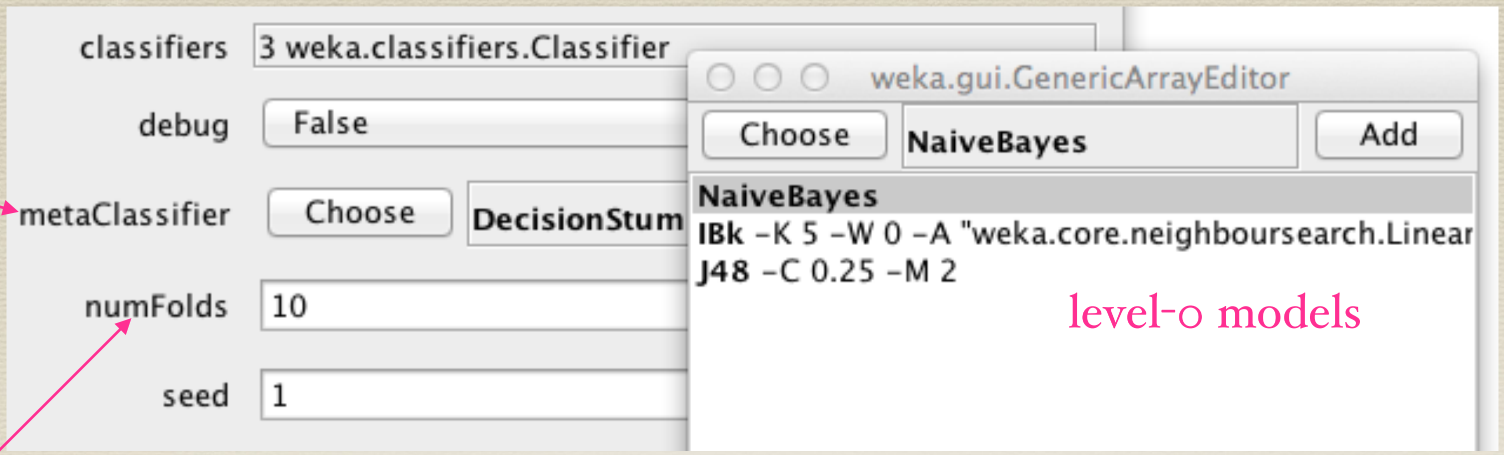


weak's stacking

level-1 model

Stacking (2/3)

- Training using "holdout"
  - Reserve some instances for training level-1 model and build level-0 models from the remaining data



```
classifiers     3 weka.classifiers.Classifier
debug           False
metaClassifier  Choose  DecisionStum
numFolds        10
seed            1
```

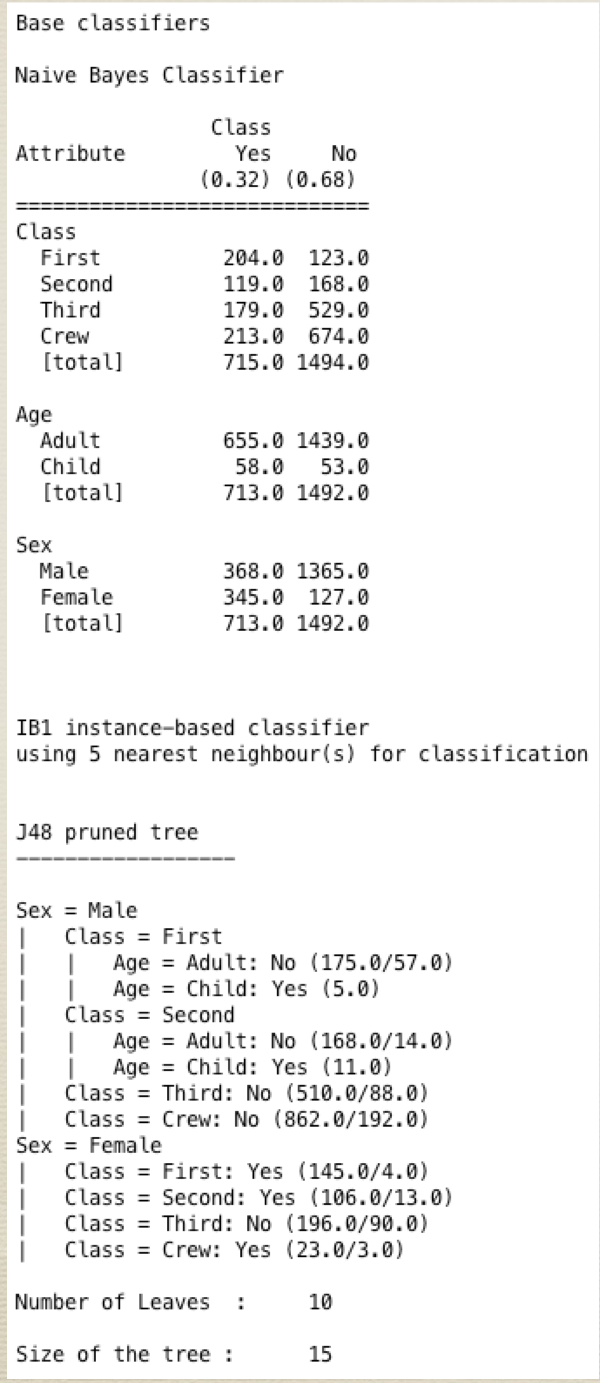
```
NaiveBayes
NaiveBayes
IBk -K 5 -W 0 -A "weka.core.neighboursearch.Linear
J48 -C 0.25 -M 2
```

level-1 models

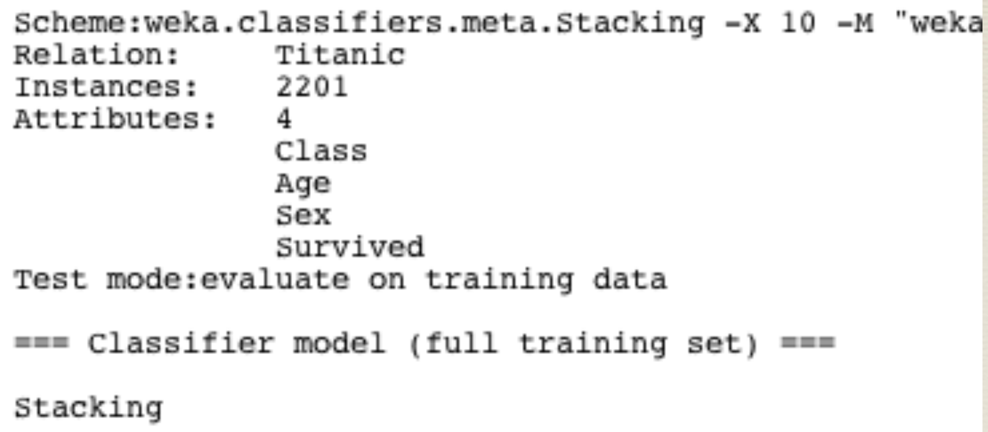
level-0 models

The number of folds used for cross-validation

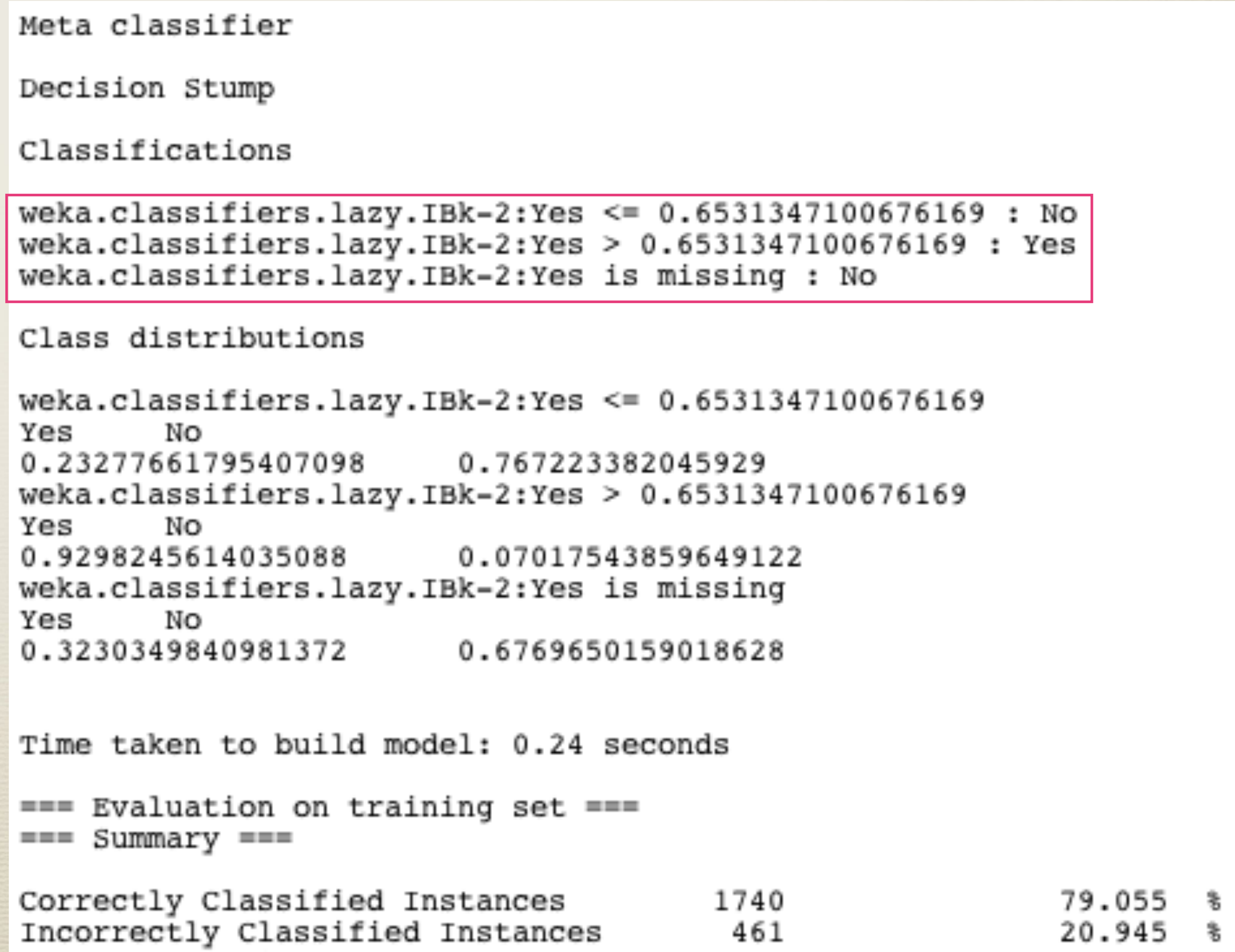
StackingC: efficient version of Stacking



```
Base classifiers

Naive Bayes Classifier

                 Class
Attribute          Yes      No
                (0.32)  (0.68)
=============================
Class
  First          204.0   123.0
  Second         119.0   168.0
  Third          179.0   529.0
  Crew           213.0   674.0
  [total]        715.0  1494.0

Age
  Adult          655.0  1439.0
  Child           58.0    53.0
  [total]        713.0  1492.0

Sex
  Male           368.0  1365.0
  Female         345.0   127.0
  [total]        713.0  1492.0



IB1 instance-based classifier
using 5 nearest neighbour(s) for classification


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

Sex = Male
|   Class = First
|   |   Age = Adult: No (175.0/57.0)
|   |   Age = Child: Yes (5.0)
|   Class = Second
|   |   Age = Adult: No (168.0/14.0)
|   |   Age = Child: Yes (11.0)
|   Class = Third: No (510.0/88.0)
|   Class = Crew: No (862.0/192.0)
Sex = Female
|   Class = First: Yes (145.0/4.0)
|   Class = Second: Yes (106.0/13.0)
|   Class = Third: No (196.0/90.0)
|   Class = Crew: Yes (23.0/3.0)

Number of Leaves  :     10

Size of the tree :      15
```



```
Scheme:weka.classifiers.meta.Stacking -X 10 -M "weka
Relation:     Titanic
Instances:    2201
Attributes:   4
              Class
              Age
              Sex
              Survived
Test mode:evaluate on training data

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

Stacking
```



```
Meta classifier

Decision Stump

Classifications

weka.classifiers.lazy.IBk-2:Yes <= 0.6531347100676169 : No
weka.classifiers.lazy.IBk-2:Yes > 0.6531347100676169 : Yes
weka.classifiers.lazy.IBk-2:Yes is missing : No

Class distributions

weka.classifiers.lazy.IBk-2:Yes <= 0.6531347100676169
Yes      No
0.23277661795407098     0.767223382045929
weka.classifiers.lazy.IBk-2:Yes > 0.6531347100676169
Yes      No
0.9298245614035088      0.07017543859649122
weka.classifiers.lazy.IBk-2:Yes is missing
Yes      No
0.3230349840981372      0.6769650159018628


Time taken to build model: 0.24 seconds

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

Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
```

Random Forests (1/11)

- A powerful new approach to data exploration, data analysis, and predictive modeling
- Developed by Leo Breiman (father of CART) at University of California, Berkeley
- A random forest is a collection of CART-like trees following specific rules for
  - Tree growing
  - Self-testing
  - Tree combination

Random Forests (2/11)

Algorithm: Random Forests

Input:

D: a set of d class-labeled training tuples;

k: the number of rounds (one classifier is generated per round);

a CART-like tree classifier;

Output: a composite model

Method:

```
for i=1 to k do
    sample D with replacement to obtain $D_i$;  //Bootstrap
    repeat
       randomly select a attribute subset F;
       split out node with the best suitable attribute of F
    until all terminal nodes of $M_i$ contain only one data record
endfor
```

Random Forests: Tree Growing (3/11)

- Trees are grown using “binary” partitioning (each parent node is split into no more than two children)
- Each tree is grown at least partially at random
  - Randomness is injected by growing each tree on a different random subsample of the training data
  - Randomness is injected into the split selection process so that the splitter at any node is determined partly at random

Random Forests: Tree Growing (4/11)

- Split selection
  - First select a small subset of available attributes at random
    - Typically we select about (1/2)**sqrt(K), sqrt(K), or 2*sqrt(K),* where K is the total number of attributes available
  - We split out node with the best attribute among the random subset
    - Gini index
- Split selection is applied on each child node until all terminal nodes contain only one data record
  - Trains rapidly even with thousands of potential attributes

Gini Index

- If a data set D contains examples from n classes, gini index, gini(D) is defined as

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the relative frequency of class j in D

- If a data set D is split on A into two subsets D_1 and D_2, the gini index gini(D) is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- The attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node (need to enumerate all the possible splitting points for each attribute)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- Ex. D has 9 tuples in buys_computer = "yes" and 5 in "no"

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Suppose the attribute income partitions D into 10 in D$_1$: {low, medium} and 4 in D$_2$

$$\begin{aligned} gini_{income \in \{low, medium\}}(D) &= \left(\frac{10}{14}\right) gini(D_1) + \left(\frac{4}{14}\right) gini(D_2) \\ &= \frac{10}{14} * \left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14} * \left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) \\ &= 0.443 \\ &= Gini_{income \in \{high\}}(D) \end{aligned}$$

but gini$_{\{medium,high\}}$ is 0.30 and thus the best since it is the lowest

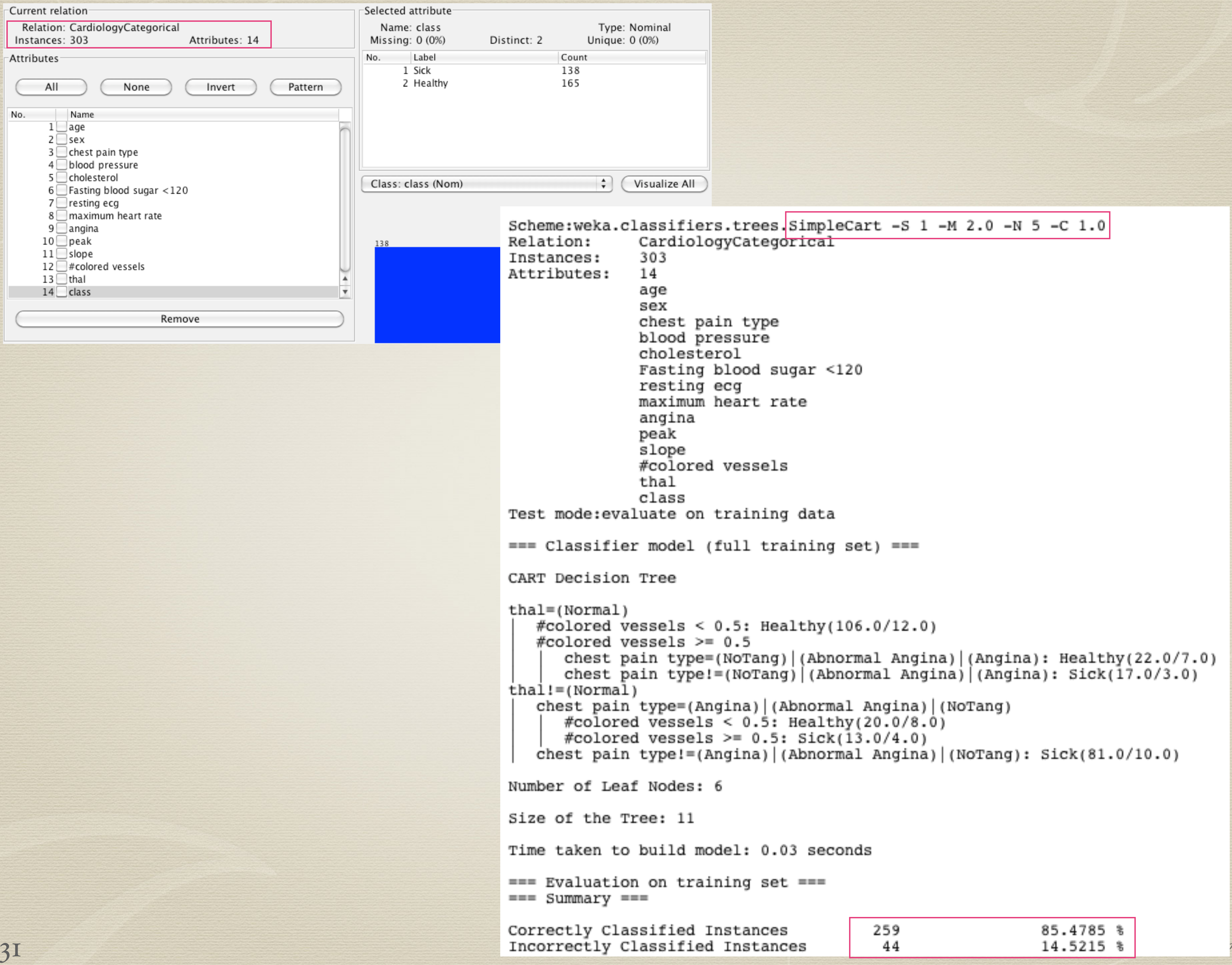
-

Random Forests: Self Testing (7/11)

- Each tree is grown on about 63.2% of the original training data (due to the bootstrap sampling process)
- Remaining, 36.8% of the data (OOB, Out of Bag), is available to test the single tree
- All performance statistics reported by random forests are based on OOB calculations

Random Forests: Combining Trees (8/11)

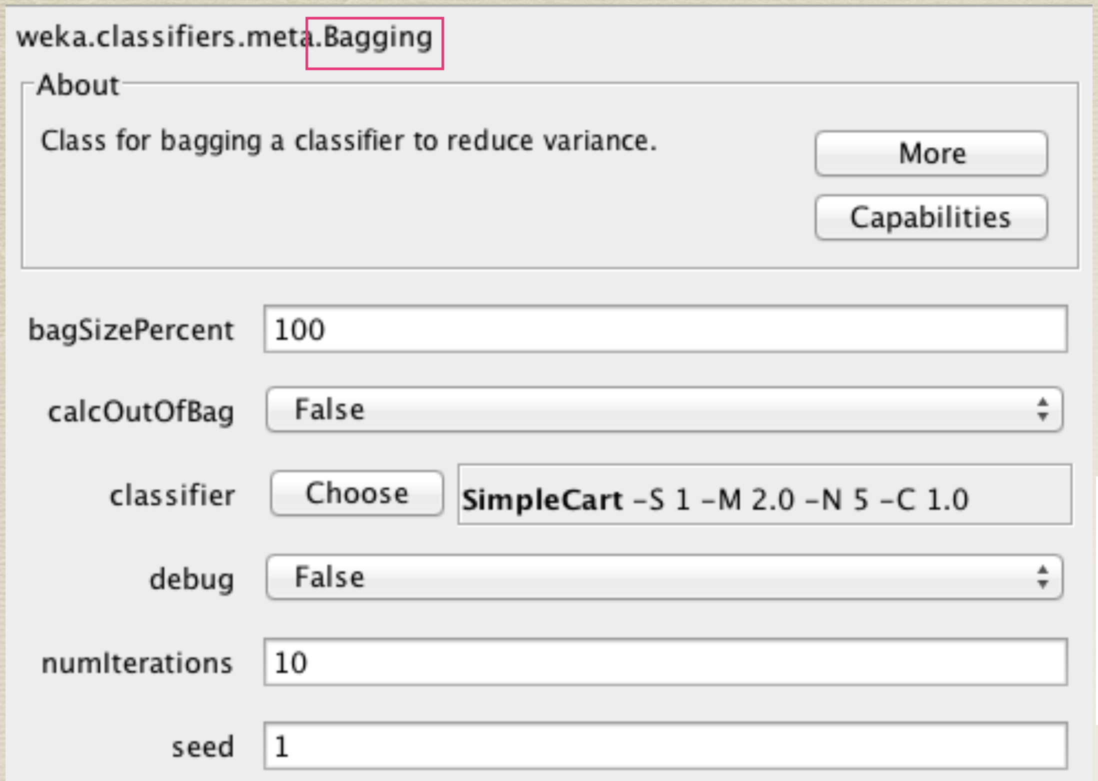
- Grow many trees
  - Recommend 500 but for large data sets 150 or so may be sufficient
- When multiple models are generated they are normally combined by
  - Voting a classification problems, perhaps weighted
  - Averaging in regression problems, perhaps weighted



```
Current relation
  Relation: CardiologyCategorical
 Instances: 303                    Attributes: 14

Attributes
No.  Name
  1  age
  2  sex
  3  chest pain type
  4  blood pressure
  5  cholesterol
  6  Fasting blood sugar <120
  7  resting ecg
  8  maximum heart rate
  9  angina
 10  peak
 11  slope
 12  #colored vessels
 13  thal
 14  class
```

```
Selected attribute
  Name: class                              Type: Nominal
 Missing: 0 (0%)        Distinct: 2        Unique: 0 (0%)
No.   Label           Count
  1   Sick            138
  2   Healthy         165

Class: class (Nom)
```

```
Scheme:weka.classifiers.trees.SimpleCart -S 1 -M 2.0 -N 5 -C 1.0
Relation:     CardiologyCategorical
Instances:    303
Attributes:   14
              age
              sex
              chest pain type
              blood pressure
              cholesterol
              Fasting blood sugar <120
              resting ecg
              maximum heart rate
              angina
              peak
              slope
              #colored vessels
              thal
              class
Test mode:evaluate on training data

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

CART Decision Tree

thal=(Normal)
|  #colored vessels < 0.5: Healthy(106.0/12.0)
|  #colored vessels >= 0.5
|  |  chest pain type=(NoTang)|(Abnormal Angina)|(Angina): Healthy(22.0/7.0)
|  |  chest pain type!=(NoTang)|(Abnormal Angina)|(Angina): Sick(17.0/3.0)
thal!=(Normal)
|  chest pain type=(Angina)|(Abnormal Angina)|(NoTang)
|  |  #colored vessels < 0.5: Healthy(20.0/8.0)
|  |  #colored vessels >= 0.5: Sick(13.0/4.0)
|  chest pain type!=(Angina)|(Abnormal Angina)|(NoTang): Sick(81.0/10.0)

Number of Leaf Nodes: 6

Size of the Tree: 11

Time taken to build model: 0.03 seconds

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

Correctly Classified Instances         259               85.4785 %
Incorrectly Classified Instances        44               14.5215 %
```



weka.classifiers.meta.Bagging

About

Class for bagging a classifier to reduce variance.

bagSizePercent: 100

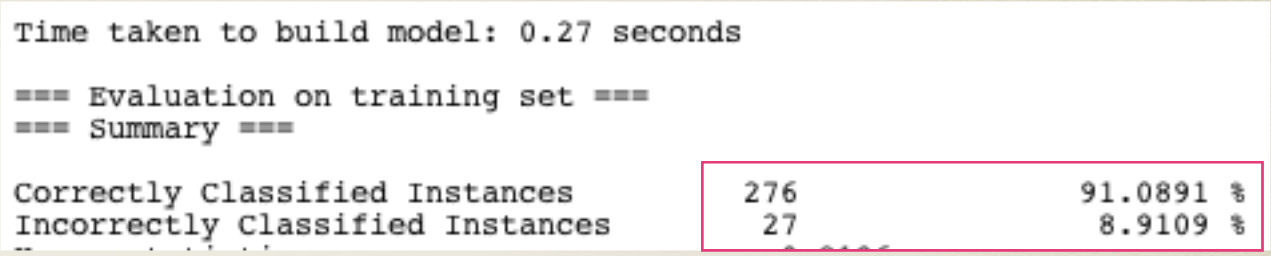
calcOutOfBag: False

classifier: **SimpleCart** -S 1 -M 2.0 -N 5 -C 1.0

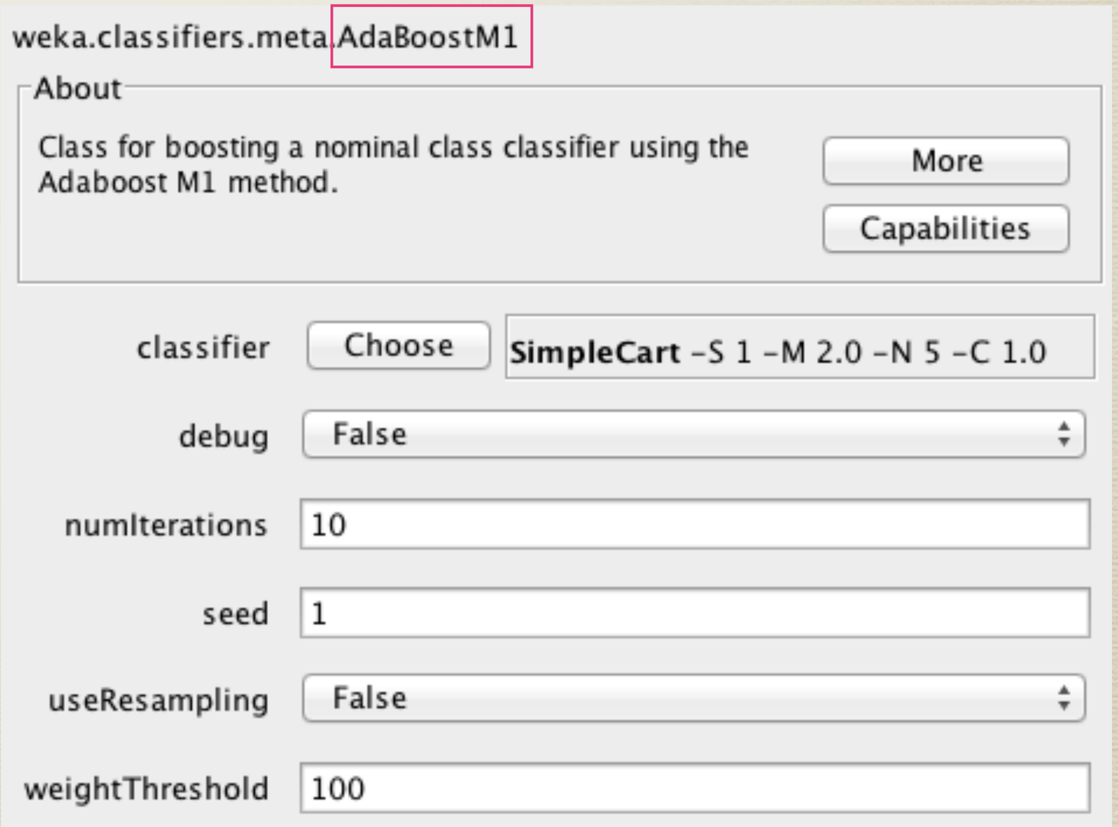
debug: False

numIterations: 10

seed: 1



```
Time taken to build model: 0.27 seconds

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

Correctly Classified Instances         276               91.0891 %
Incorrectly Classified Instances        27                8.9109 %
```



weka.classifiers.meta.AdaBoostM1

About

Class for boosting a nominal class classifier using the Adaboost M1 method.

classifier: **SimpleCart** -S 1 -M 2.0 -N 5 -C 1.0

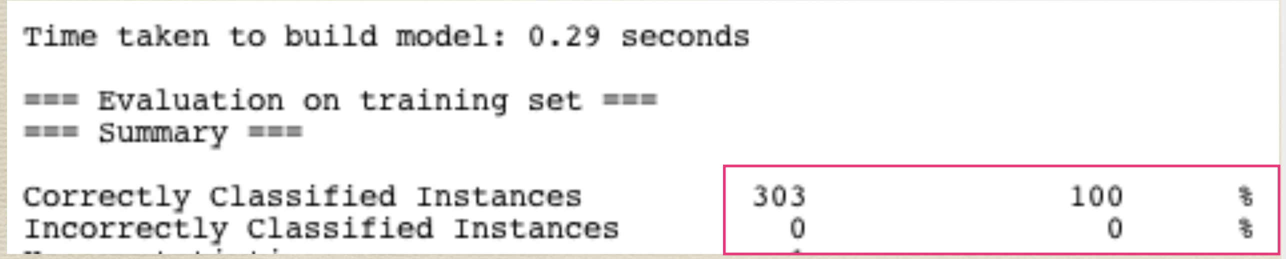
debug: False

numIterations: 10

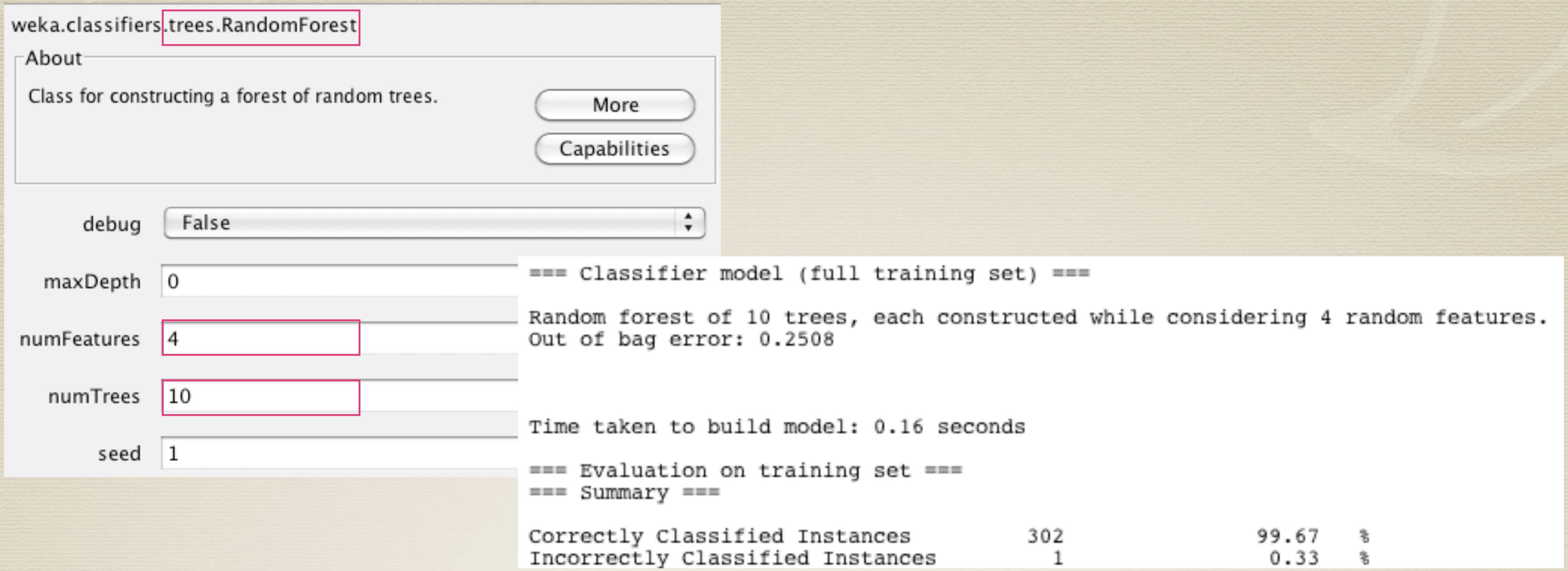
seed: 1

useResampling: False

weightThreshold: 100



```
Time taken to build model: 0.29 seconds

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

Correctly Classified Instances         303              100      %
Incorrectly Classified Instances         0                0      %
```



weka.classifiers.trees.RandomForest

About

Class for constructing a forest of random trees.

debug: False

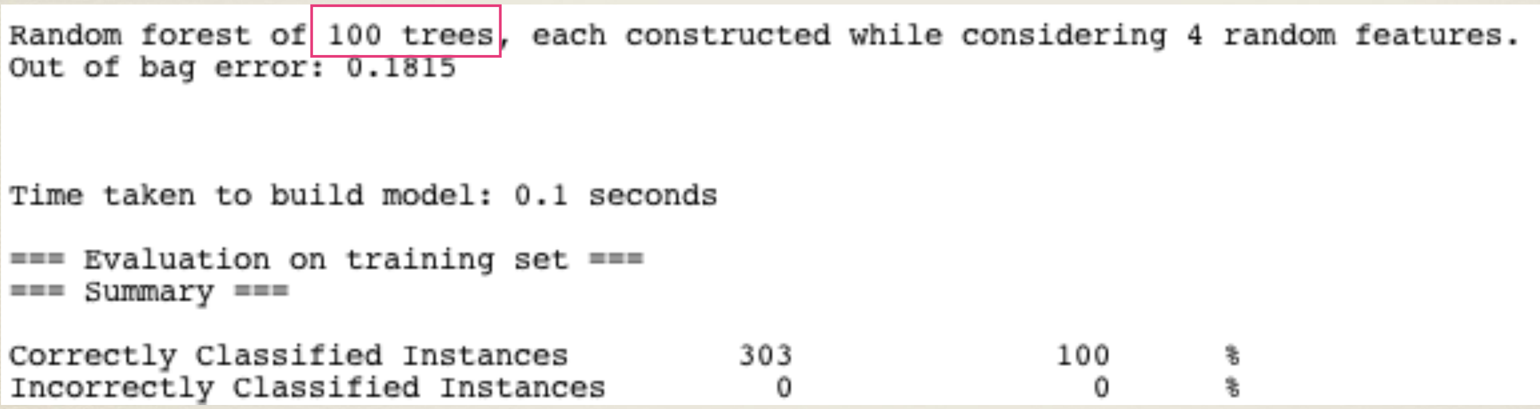
maxDepth: 0

numFeatures: 4

numTrees: 10

seed: 1

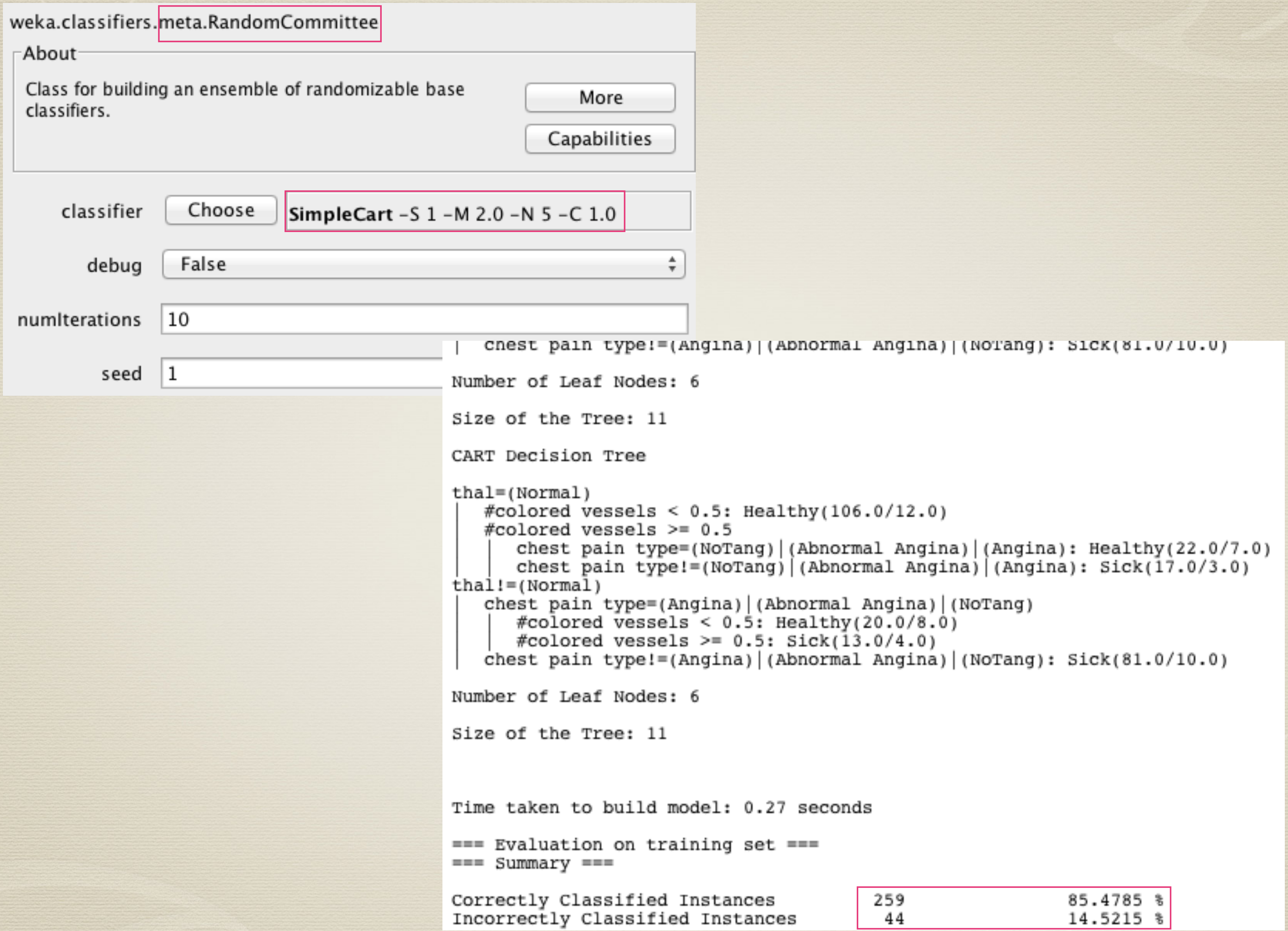
```
=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 4 random features.
Out of bag error: 0.2508



Time taken to build model: 0.16 seconds

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

Correctly Classified Instances         302               99.67   %
Incorrectly Classified Instances         1                0.33   %
```



```
Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.1815



Time taken to build model: 0.1 seconds

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

Correctly Classified Instances         303                100     %
Incorrectly Classified Instances         0                  0     %
```

Randomization (1/5)

- Random number seeds
  - RandomCommittee
- Random sampling training data
  - Bagging
  - RandomForest
- Random subsets of attributes
  - RandomSubSpace
    - Randomly select at the beginning to build tree
  - RandomTree
    - Randomly select at each node
    - One of a random forest
  - RandomForest



weka.classifiers.meta.RandomCommittee

About

Class for building an ensemble of randomizable base classifiers.

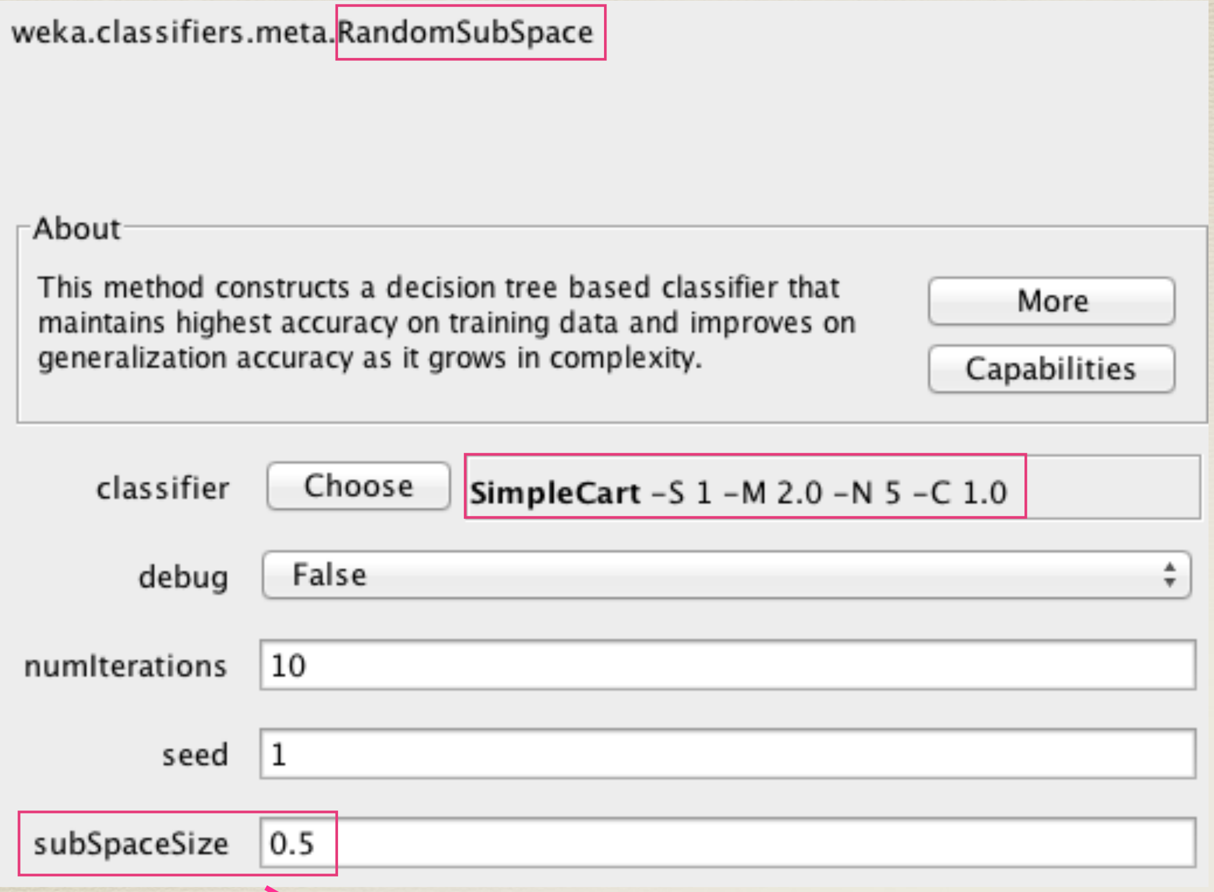
classifier: SimpleCart -S 1 -M 2.0 -N 5 -C 1.0

debug: False

numIterations: 10

seed: 1

```
|   chest pain type!=(Angina)|(Abnormal Angina)|(NoTang): Sick(81.0/10.0)

Number of Leaf Nodes: 6

Size of the Tree: 11

CART Decision Tree

thal=(Normal)
|  #colored vessels < 0.5: Healthy(106.0/12.0)
|  #colored vessels >= 0.5
|  |  chest pain type=(NoTang)|(Abnormal Angina)|(Angina): Healthy(22.0/7.0)
|  |  chest pain type!=(NoTang)|(Abnormal Angina)|(Angina): Sick(17.0/3.0)
thal!=(Normal)
|  chest pain type=(Angina)|(Abnormal Angina)|(NoTang)
|  |  #colored vessels < 0.5: Healthy(20.0/8.0)
|  |  #colored vessels >= 0.5: Sick(13.0/4.0)
|  chest pain type!=(Angina)|(Abnormal Angina)|(NoTang): Sick(81.0/10.0)

Number of Leaf Nodes: 6

Size of the Tree: 11



Time taken to build model: 0.27 seconds

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

Correctly Classified Instances         259               85.4785 %
Incorrectly Classified Instances        44               14.5215 %
```



weka.classifiers.meta.RandomSubSpace

About

This method constructs a decision tree based classifier that maintains highest accuracy on training data and improves on generalization accuracy as it grows in complexity.

classifier: SimpleCart -S 1 -M 2.0 -N 5 -C 1.0

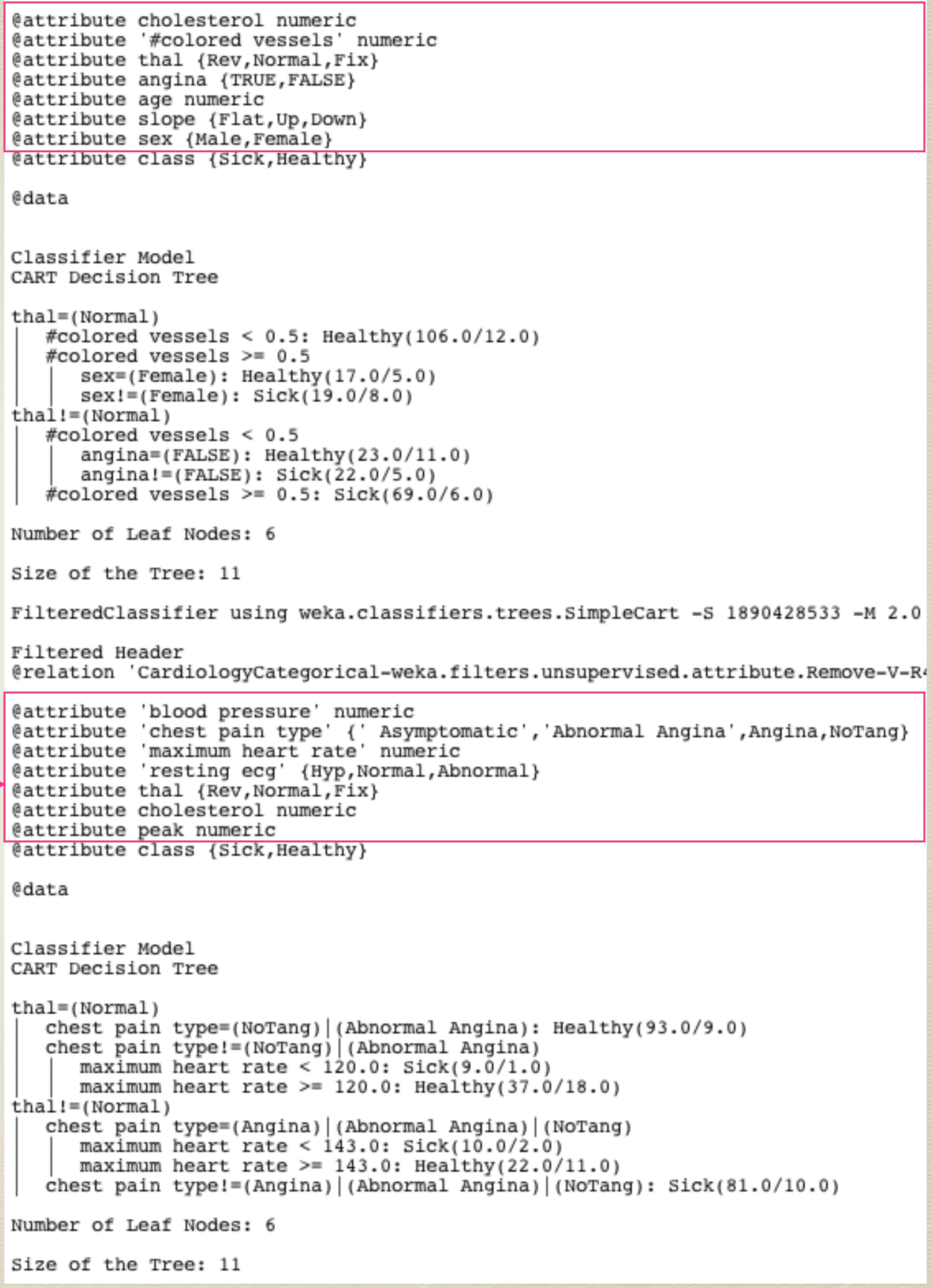
debug: False

numIterations: 10

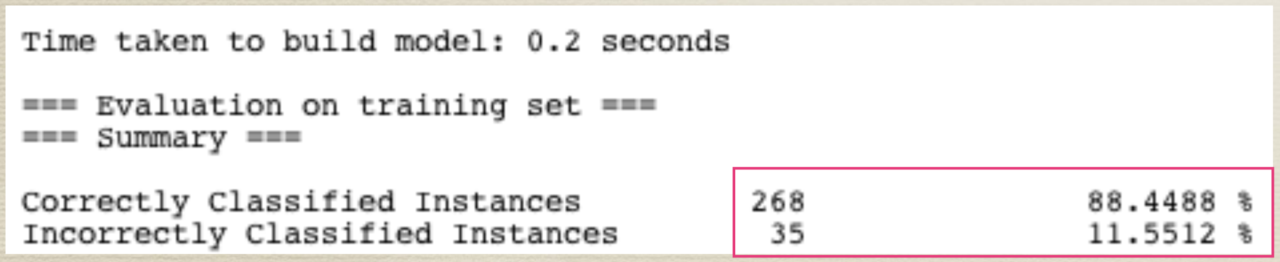
seed: 1

subSpaceSize: 0.5

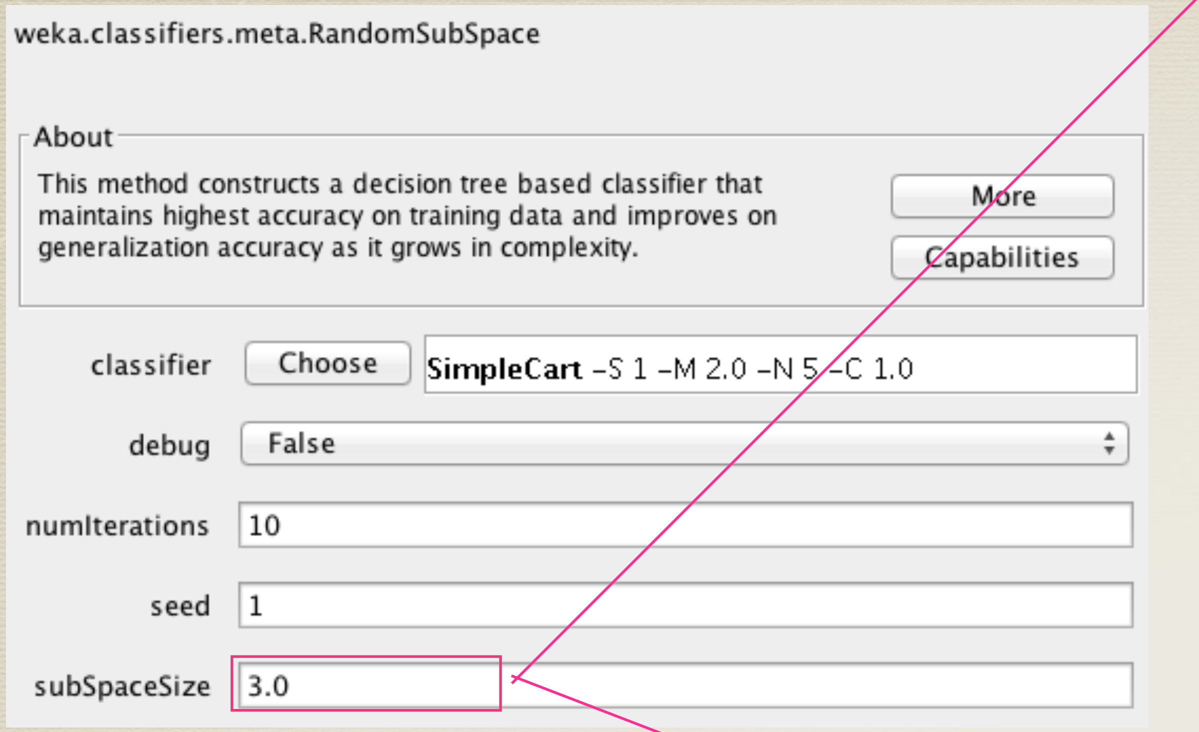
13*0.5



```
@attribute cholesterol numeric
@attribute '#colored vessels' numeric
@attribute thal {Rev,Normal,Fix}
@attribute angina {TRUE,FALSE}
@attribute age numeric
@attribute slope {Flat,Up,Down}
@attribute sex {Male,Female}
@attribute class {Sick,Healthy}

@data


Classifier Model
CART Decision Tree

thal=(Normal)
|  #colored vessels < 0.5: Healthy(106.0/12.0)
|  #colored vessels >= 0.5
|  |  sex=(Female): Healthy(17.0/5.0)
|  |  sex!=(Female): Sick(19.0/8.0)
thal!=(Normal)
|  #colored vessels < 0.5
|  |  angina=(FALSE): Healthy(23.0/11.0)
|  |  angina!=(FALSE): Sick(22.0/5.0)
|  #colored vessels >= 0.5: Sick(69.0/6.0)

Number of Leaf Nodes: 6

Size of the Tree: 11

FilteredClassifier using weka.classifiers.trees.SimpleCart -S 1890428533 -M 2.0

Filtered Header
@relation 'CardiologyCategorical-weka.filters.unsupervised.attribute.Remove-V-R

@attribute 'blood pressure' numeric
@attribute 'chest pain type' {' Asymptomatic','Abnormal Angina',Angina,NoTang}
@attribute 'maximum heart rate' numeric
@attribute 'resting ecg' {Hyp,Normal,Abnormal}
@attribute thal {Rev,Normal,Fix}
@attribute cholesterol numeric
@attribute peak numeric
@attribute class {Sick,Healthy}

@data


Classifier Model
CART Decision Tree

thal=(Normal)
|  chest pain type=(NoTang)|(Abnormal Angina): Healthy(93.0/9.0)
|  chest pain type!=(NoTang)|(Abnormal Angina)
|  |  maximum heart rate < 120.0: Sick(9.0/1.0)
|  |  maximum heart rate >= 120.0: Healthy(37.0/18.0)
thal!=(Normal)
|  chest pain type=(Angina)|(Abnormal Angina)|(NoTang)
|  |  maximum heart rate < 143.0: Sick(10.0/2.0)
|  |  maximum heart rate >= 143.0: Healthy(22.0/11.0)
|  chest pain type!=(Angina)|(Abnormal Angina)|(NoTang): Sick(81.0/10.0)

Number of Leaf Nodes: 6

Size of the Tree: 11
```



```
Time taken to build model: 0.2 seconds

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

Correctly Classified Instances         268               88.4488 %
Incorrectly Classified Instances        35               11.5512 %
```



weka.classifiers.meta.RandomSubSpace

About

This method constructs a decision tree based classifier that maintains highest accuracy on training data and improves on generalization accuracy as it grows in complexity.

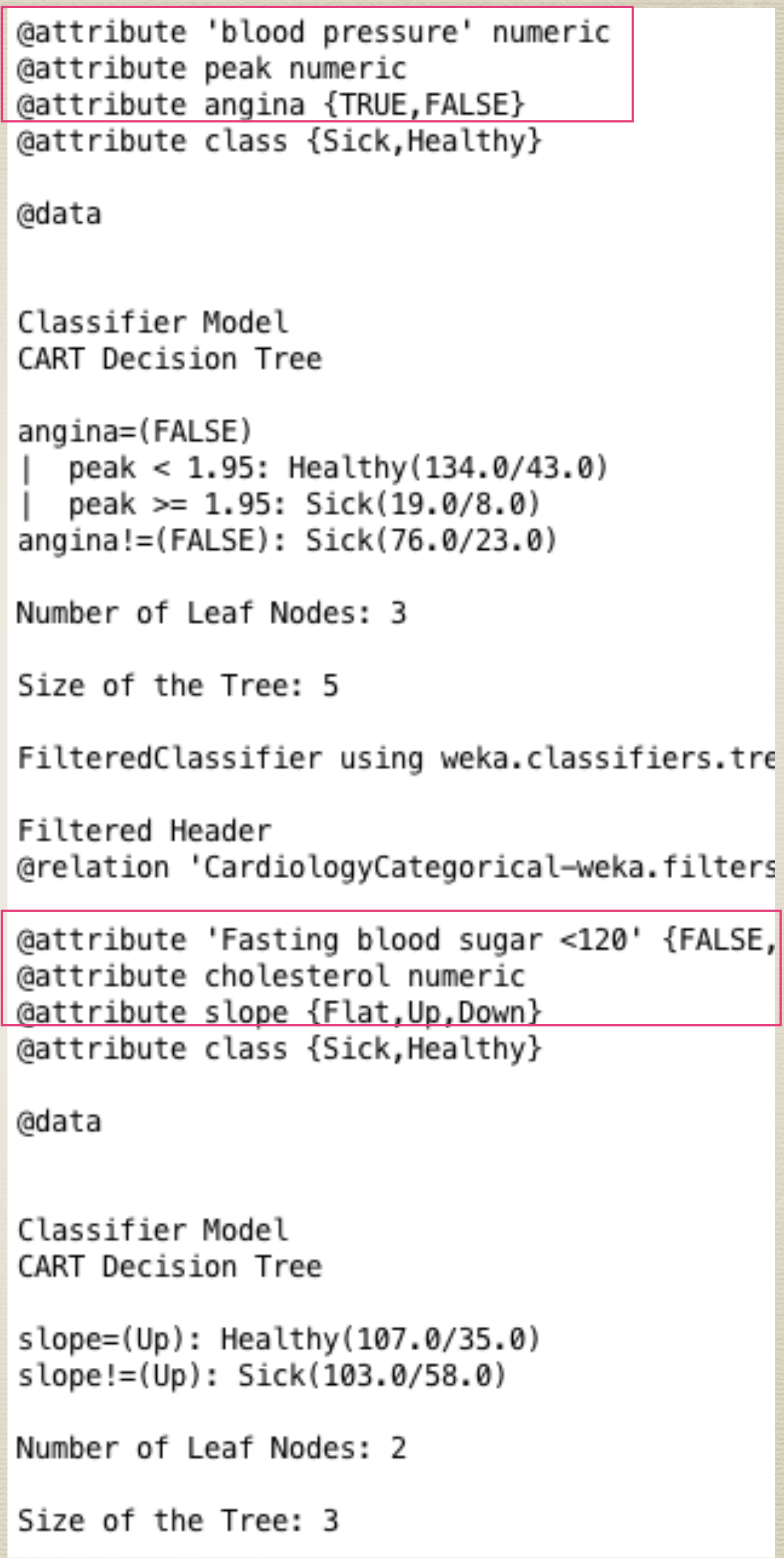
classifier: SimpleCart -S 1 -M 2.0 -N 5 -C 1.0

debug: False

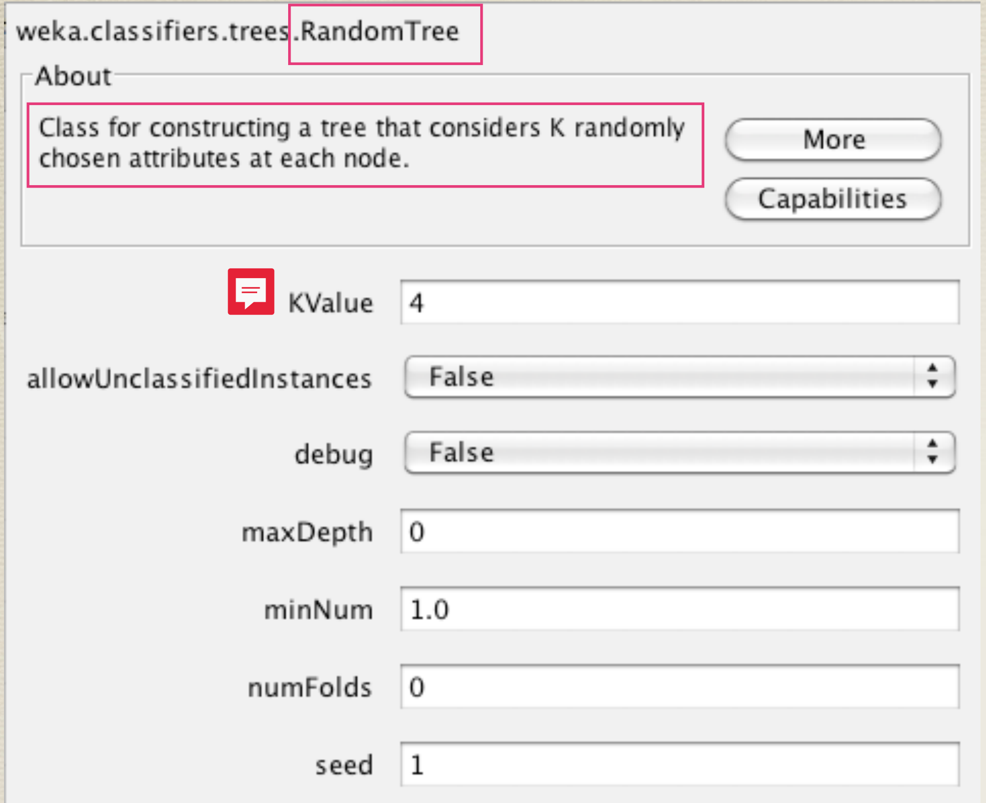
numIterations: 10

seed: 1

subSpaceSize: 3.0



```
@attribute 'blood pressure' numeric
@attribute peak numeric
@attribute angina {TRUE,FALSE}
@attribute class {Sick,Healthy}

@data


Classifier Model
CART Decision Tree

angina=(FALSE)
|  peak < 1.95: Healthy(134.0/43.0)
|  peak >= 1.95: Sick(19.0/8.0)
angina!=(FALSE): Sick(76.0/23.0)

Number of Leaf Nodes: 3

Size of the Tree: 5

FilteredClassifier using weka.classifiers.tre

Filtered Header
@relation 'CardiologyCategorical-weka.filters

@attribute 'Fasting blood sugar <120' {FALSE,
@attribute cholesterol numeric
@attribute slope {Flat,Up,Down}
@attribute class {Sick,Healthy}

@data


Classifier Model
CART Decision Tree

slope=(Up): Healthy(107.0/35.0)
slope!=(Up): Sick(103.0/58.0)

Number of Leaf Nodes: 2

Size of the Tree: 3
```

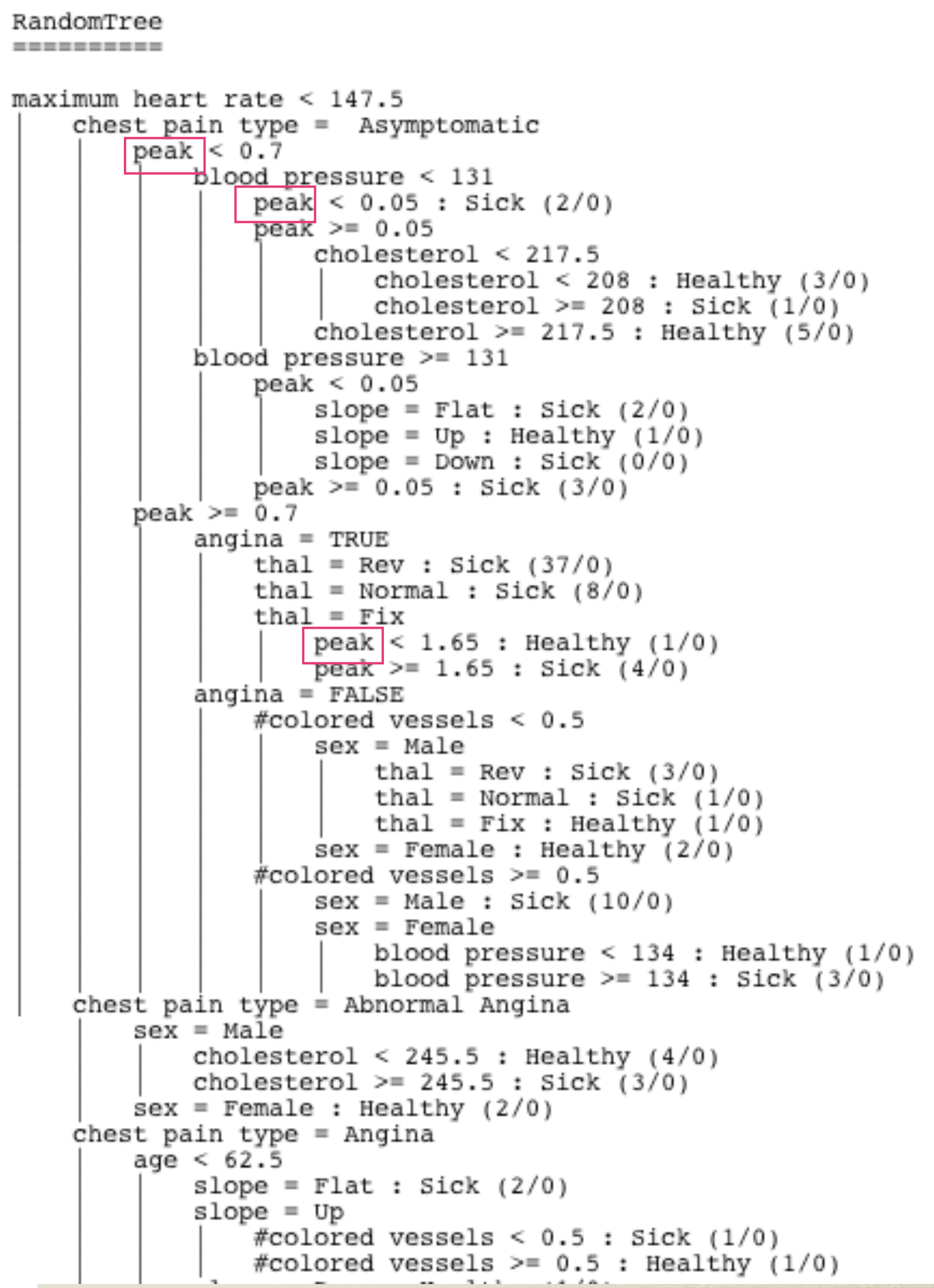


weka.classifiers.trees.RandomTree

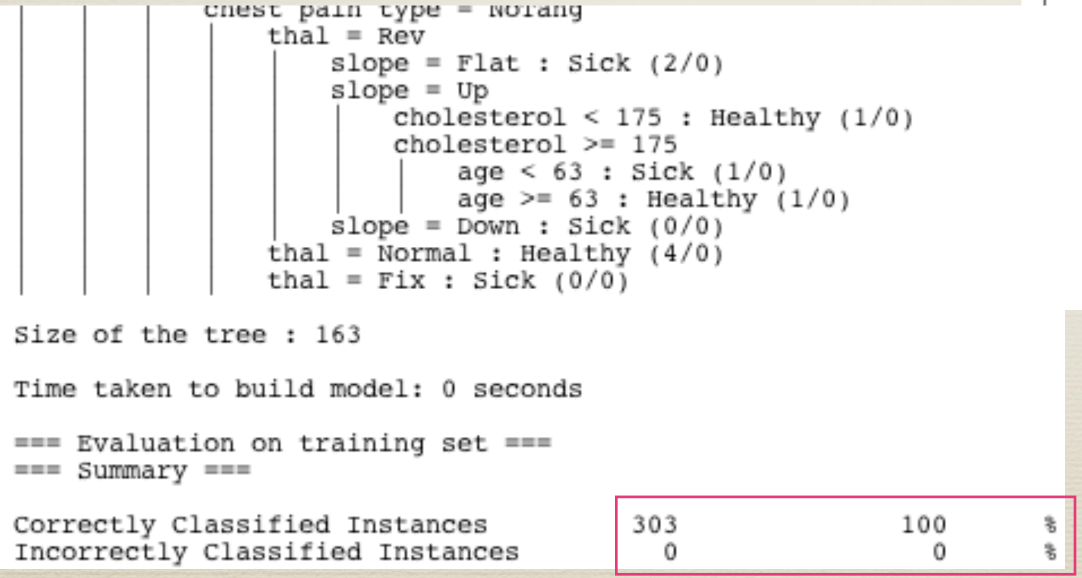
About

Class for constructing a tree that considers K randomly chosen attributes at each node.

- KValue: 4
- allowUnclassifiedInstances: False
- debug: False
- maxDepth: 0
- minNum: 1.0
- numFolds: 0
- seed: 1



```
RandomTree
==========

maximum heart rate < 147.5
|   chest pain type =  Asymptomatic
|   |   peak < 0.7
|   |   |   blood pressure < 131
|   |   |   |   peak < 0.05 : Sick (2/0)
|   |   |   |   peak >= 0.05
|   |   |   |   |   cholesterol < 217.5
|   |   |   |   |   |   cholesterol < 208 : Healthy (3/0)
|   |   |   |   |   |   cholesterol >= 208 : Sick (1/0)
|   |   |   |   |   cholesterol >= 217.5 : Healthy (5/0)
|   |   |   blood pressure >= 131
|   |   |   |   peak < 0.05
|   |   |   |   |   slope = Flat : Sick (2/0)
|   |   |   |   |   slope = Up : Healthy (1/0)
|   |   |   |   |   slope = Down : Sick (0/0)
|   |   |   |   peak >= 0.05 : Sick (3/0)
|   |   peak >= 0.7
|   |   |   angina = TRUE
|   |   |   |   thal = Rev : Sick (37/0)
|   |   |   |   thal = Normal : Sick (8/0)
|   |   |   |   thal = Fix
|   |   |   |   |   peak < 1.65 : Healthy (1/0)
|   |   |   |   |   peak >= 1.65 : Sick (4/0)
|   |   |   angina = FALSE
|   |   |   |   #colored vessels < 0.5
|   |   |   |   |   sex = Male
|   |   |   |   |   |   thal = Rev : Sick (3/0)
|   |   |   |   |   |   thal = Normal : Sick (1/0)
|   |   |   |   |   |   thal = Fix : Healthy (1/0)
|   |   |   |   |   sex = Female : Healthy (2/0)
|   |   |   |   #colored vessels >= 0.5
|   |   |   |   |   sex = Male : Sick (10/0)
|   |   |   |   |   sex = Female
|   |   |   |   |   |   blood pressure < 134 : Healthy (1/0)
|   |   |   |   |   |   blood pressure >= 134 : Sick (3/0)
|   chest pain type = Abnormal Angina
|   |   sex = Male
|   |   |   cholesterol < 245.5 : Healthy (4/0)
|   |   |   cholesterol >= 245.5 : Sick (3/0)
|   |   sex = Female : Healthy (2/0)
|   chest pain type = Angina
|   |   age < 62.5
|   |   |   slope = Flat : Sick (2/0)
|   |   |   slope = Up
|   |   |   |   #colored vessels < 0.5 : Sick (1/0)
|   |   |   |   #colored vessels >= 0.5 : Healthy (1/0)
```



```
|   |   |   chest pain type = NoTang
|   |   |   |   thal = Rev
|   |   |   |   |   slope = Flat : Sick (2/0)
|   |   |   |   |   slope = Up
|   |   |   |   |   |   cholesterol < 175 : Healthy (1/0)
|   |   |   |   |   |   cholesterol >= 175
|   |   |   |   |   |   |   age < 63 : Sick (1/0)
|   |   |   |   |   |   |   age >= 63 : Healthy (1/0)
|   |   |   |   |   slope = Down : Sick (0/0)
|   |   |   |   thal = Normal : Healthy (4/0)
|   |   |   |   thal = Fix : Sick (0/0)

Size of the tree : 163

Time taken to build model: 0 seconds

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

Correctly Classified Instances         303              100      %
Incorrectly Classified Instances         0                0      %
```