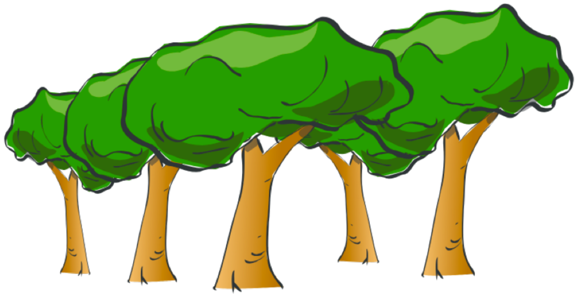
What’s better than a tree?

Random Forest



- Can often improve performance of a decision tree classifier using a set of decision trees (a forest)
- Each tree trained on a random subset of training data
- Classify a data instance using all trees
- Combine answers to make classification
  - E.g., vote for most common class

Bagging

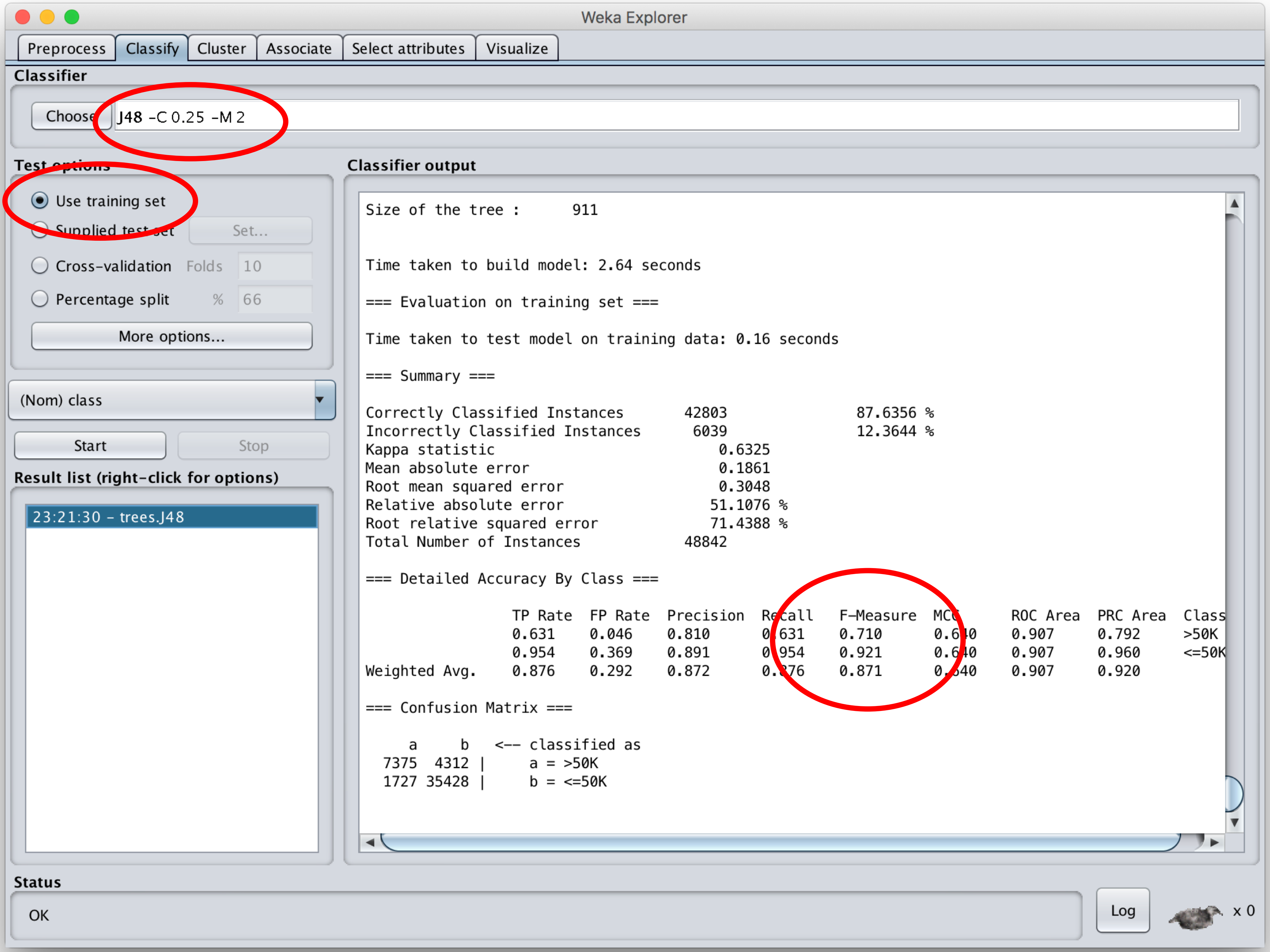
- Idea can be used on any classifier!
- Bagging = ***B****ootstrap* ***agg****regat****ing***

  An ***ensemble*** *meta-algorithm that can improve stability & accuracy of algorithms for statistical classification and regression*

- Helps avoid overfitting

Choosing subsets of training data

- Classic bagging: select random set of training instances **with replacement**
- Pasting: select random subset of training instances
- Random Subspaces: use all training instances, but with a random subset of features
- Random Patches: random subset of instances and random subset of features
- What’s best? YMMV: depends on problem, training data, algorithm



Classifier: J48 -C 0.25 -M 2

Test options:

● Use training set

○ Supplied test set

○ Cross-validation Folds 10

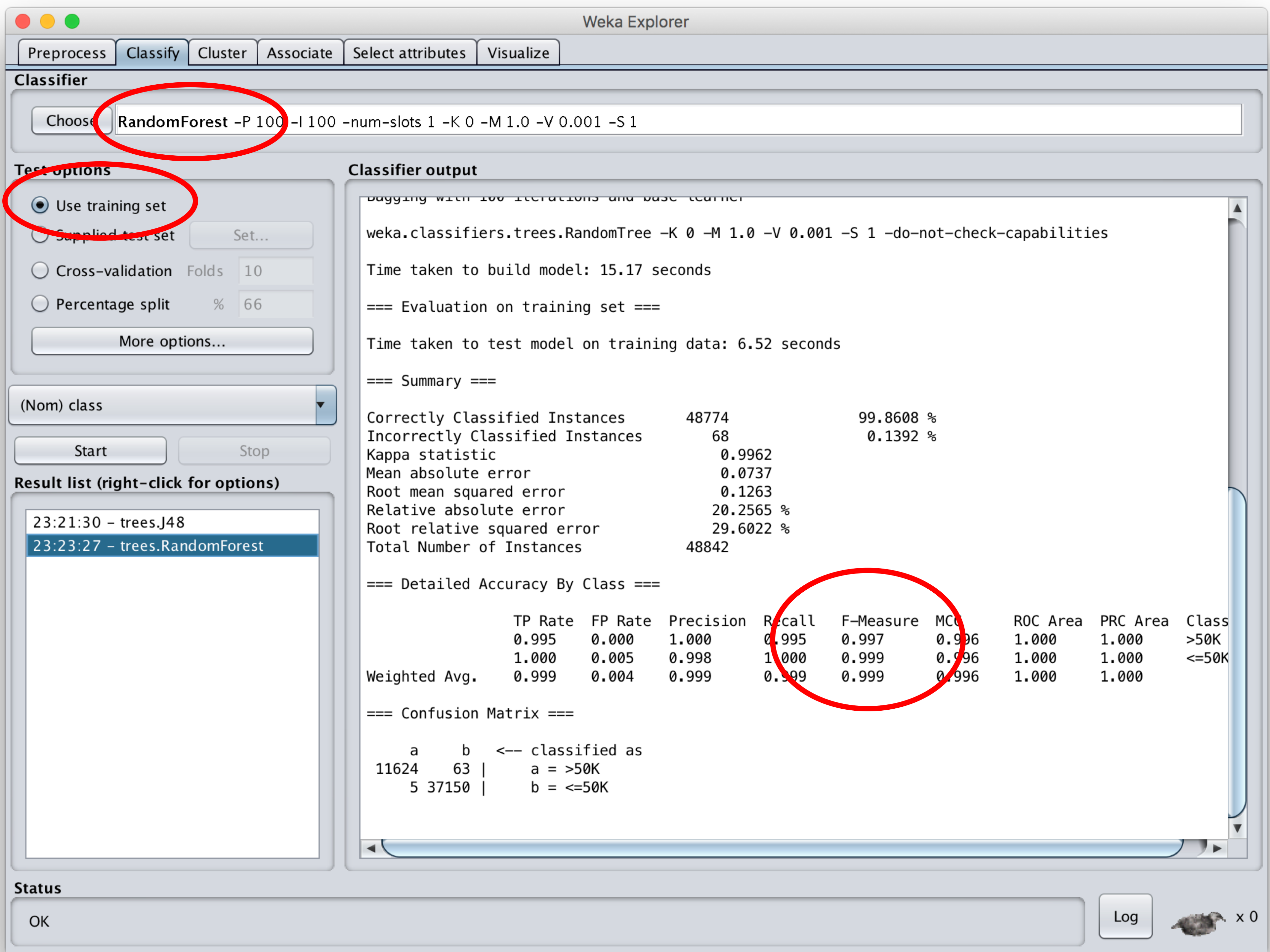
○ Percentage split % 66

(Nom) class

23:21:30 - trees.J48

```
Size of the tree :      911


Time taken to build model: 2.64 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.16 seconds

=== Summary ===

Correctly Classified Instances       42803               87.6356 %
Incorrectly Classified Instances      6039               12.3644 %
Kappa statistic                          0.6325
Mean absolute error                      0.1861
Root mean squared error                  0.3048
Relative absolute error                 51.1076 %
Root relative squared error             71.4388 %
Total Number of Instances            48842

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.631    0.046    0.810      0.631    0.710      0.640    0.907     0.792     >50K
                 0.954    0.369    0.891      0.954    0.921      0.640    0.907     0.960     <=50K
Weighted Avg.    0.876    0.292    0.872      0.876    0.871      0.640    0.907     0.920

=== Confusion Matrix ===

     a     b   <-- classified as
  7375  4312 |     a = >50K
  1727 35428 |     b = <=50K
```

Status: OK



Classifier: RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options:
- ● Use training set
- ○ Supplied test set Set...
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) class

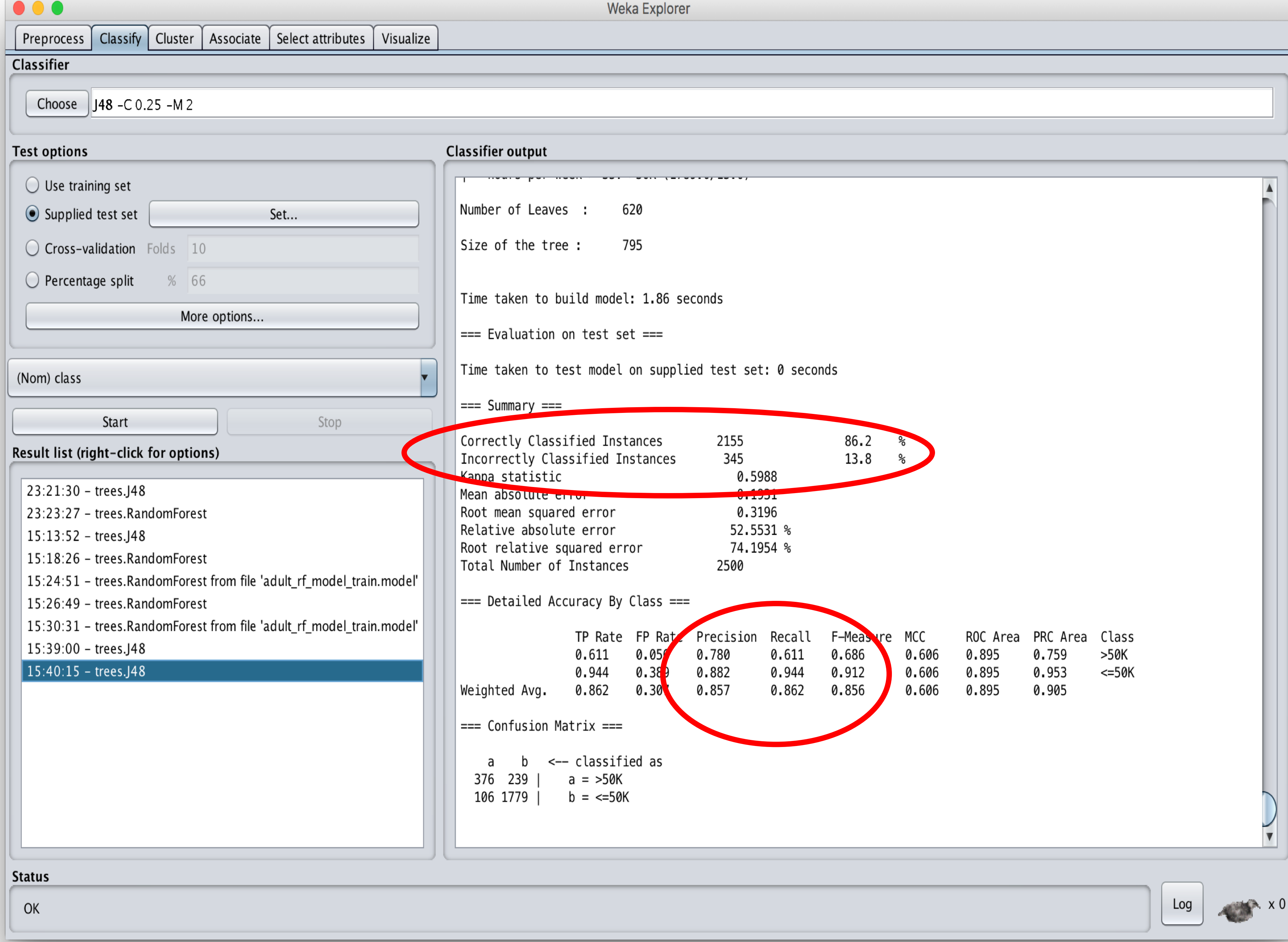
Result list:
- 23:21:30 - trees.J48
- 23:23:27 - trees.RandomForest

Status: OK

```
Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 15.17 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 6.52 seconds

=== Summary ===

Correctly Classified Instances       48774               99.8608 %
Incorrectly Classified Instances        68                0.1392 %
Kappa statistic                          0.9962
Mean absolute error                      0.0737
Root mean squared error                  0.1263
Relative absolute error                 20.2565 %
Root relative squared error             29.6022 %
Total Number of Instances            48842

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.000    1.000      0.995    0.997      0.996    1.000     1.000     >50K
                 1.000    0.005    0.998      1.000    0.999      0.996    1.000     1.000     <=50K
Weighted Avg.    0.999    0.004    0.999      0.999    0.999      0.996    1.000     1.000

=== Confusion Matrix ===

     a     b   <-- classified as
 11624    63 |     a = >50K
     5 37150 |     b = <=50K
```

Created a train and test collection

- Train has ~95% of data, test 5%
- Trained models for J48 and random forest using train dataset
- Tested on test data set
- Results were that random forest was (at best) about the same as J48



J48 -C 0.25 -M 2

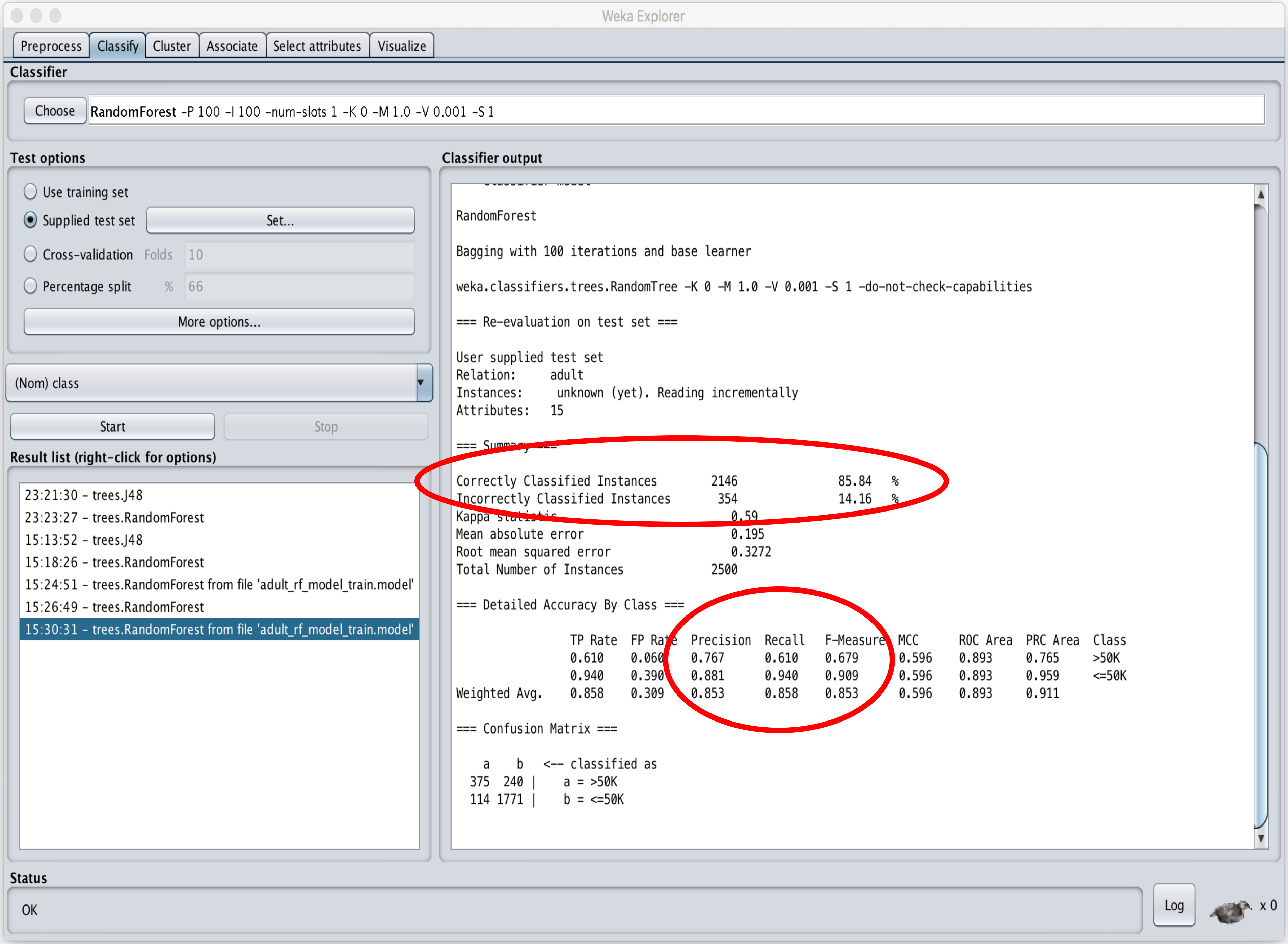
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

23:21:30 - trees.J48
23:23:27 - trees.RandomForest
15:13:52 - trees.J48
15:18:26 - trees.RandomForest
15:24:51 - trees.RandomForest from file 'adult_rf_model_train.model'
15:26:49 - trees.RandomForest
15:30:31 - trees.RandomForest from file 'adult_rf_model_train.model'
15:39:00 - trees.J48
15:40:15 - trees.J48

```
Number of Leaves  :     620

Size of the tree :      795


Time taken to build model: 1.86 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances        2155               86.2    %
Incorrectly Classified Instances       345               13.8    %
Kappa statistic                          0.5988
Mean absolute error                      0.1931
Root mean squared error                  0.3196
Relative absolute error                 52.5531 %
Root relative squared error             74.1954 %
Total Number of Instances             2500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.611    0.056    0.780      0.611    0.686      0.606    0.895     0.759     >50K
                 0.944    0.389    0.882      0.944    0.912      0.606    0.895     0.953     <=50K
Weighted Avg.    0.862    0.307    0.857      0.862    0.856      0.606    0.895     0.905

=== Confusion Matrix ===

    a    b   <-- classified as
  376  239 |    a = >50K
  106 1779 |    b = <=50K
```

OK



Classifier

RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

○ Use training set

● Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

23:21:30 - trees.J48
23:23:27 - trees.RandomForest
15:13:52 - trees.J48
15:18:26 - trees.RandomForest
15:24:51 - trees.RandomForest from file 'adult_rf_model_train.model'
15:26:49 - trees.RandomForest
15:30:31 - trees.RandomForest from file 'adult_rf_model_train.model'

Classifier output

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

=== Re-evaluation on test set ===

User supplied test set
Relation:     adult
Instances:     unknown (yet). Reading incrementally
Attributes:   15

=== Summary ===

Correctly Classified Instances        2146               85.84   %
Incorrectly Classified Instances       354               14.16   %
Kappa statistic                          0.59
Mean absolute error                      0.195
Root mean squared error                  0.3272
Total Number of Instances             2500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.610    0.060    0.767      0.610    0.679      0.596    0.893     0.765     >50K
                 0.940    0.390    0.881      0.940    0.909      0.596    0.893     0.959     <=50K
Weighted Avg.    0.858    0.309    0.853      0.858    0.853      0.596    0.893     0.911

=== Confusion Matrix ===

    a    b   <-- classified as
  375  240 |    a = >50K
  114 1771 |    b = <=50K
```

Status

OK