CHEMICAL TOXICITY PREDICTION

By

Naga Anusha Bommidi

ID: 2768523

Mahitha Doddala

ID: 2787070

CONTENTS

1.TOXICITY PREDICTION OF CHEMICALS

1.1.Background

Computational prediction of toxicity based on a chemical's structure is a challenging problem. The major reason for this is the fact that the process between the phase where the body is exposed to the chemical to the phase where physical symptoms start appearing is highly complex and may involve multiple biological networks. Various researchers have been trying to seek a practically feasible solution for this problem for over a decade now. To better understand the toxicity of chemicals and to better preserve our environment, the US Environmental Protect Agency (EPA) organized the tox21 competition.

For this class project, we use machine learning techniques to predict how protein interacts with chemicals, a critical step towards the better modeling of toxicity of chemicals. For this purpose, scientist developed in vitro assays (assay for simplicity), which are lab experiments that are performed outside of the live animals. Our specific task is to predict the result of some assays.

1.2.Data Source

There are a total of 12 assays and a test set provided by the instructor, named as follows:

- ECCS_738_sr_hse_train.arff
- ECCS_738_sr_mmp_train.arff
- ECCS_738_TestSet.arff
- EECS_738_nr_ahr_train.arff
- EECS_738_nr_ar_lbd_train.arff
- EECS_738_nr_ar_train.arff
- EECS_738_nr_aromatase_train.arff
- EECS_738_nr_er_lbd_train.arff
- EECS_738_nr_er_train.arff
- EECS_738_nr_ppar_gamma_train.arff
- EECS_738_sr_are_train.arff
- EECS_738_sr_atad5_train.arff
- EECS_738_sr_p53_train.arff

2.PROBLEM STATEMENT

2.1.Goal

In the given data, there are output results from twelve different assays. You can pick any three assays and build models to predict the result of these assays.

2.2.Instructions

1) Download the data from the source that is specified by the instructor.
2) Build a model to predict the labels of at least three of the assays provided in the dataset.
3) For each assay, use at least three different machine learning algorithms such as SVM, KNN, Neural Networks or any other algorithms. Optimize the algorithm by optimizing the parameters of the algorithms.
4) Feature representation is a big aspect of modeling. Use at least one kind of feature selection machine algorithm and see how much the results improve.
5) For model evaluation, report the performance of the above algorithms on MCC (Matthew's Correlation Coefficient).
   MCC = (TP*TN-FP*FN)/sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)) , where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively. When computing MCC, you need to run the experiments at least 50 times.

3.EXPERIMENTAL DESIGN AND STUDY RESULTS

3.1.BUILDING MODELS

Out of the 12 assays, the 3 assays that we have chosen to built the models are as follows:

- EECS_738_nr_ar_train.arff
- EECS_738_nr_aromatase_train.arff
- EECS_sr_are_train.arff

And the classifiers we have chosen are as follows:

- RandomForest
- RandomTree
- REPTree

For each assay, we used above mentioned three classifiers to build 9 different models.
We used cross-validation with 10 folds which is also a default option. But before finalizing the models we did parameter optimization which is discussed in the next section.

3.2.OPTIMIZING MODELS

In order to build an optimized model, we first checked the ROC values by changing the parameter values for each classifier. And optimized parameters are shown in the Table.1. The screenshots of the models using the optimized parameters are in Appendix A.

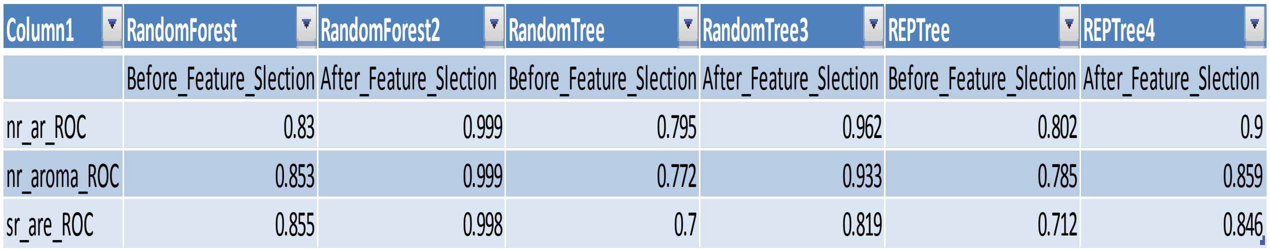
	RandomForest	RandomTree	REPTree
EECS_738_nr_ar_train.arff	-I 60 -K 30	-K 17 -M 35	-M 0
EECS_738_nr_aromatase_train.arff	-I 40 -K 25	-K 20 -M 40	-M 0
EECS_sr_are_train.arff	-I 60 -K 30	-K 50 -M 200 -S 5	-M 0

Table.1

3.3.FEATURE EXTRACTION

We used Ranker Search algorithm with threshold 0.0 and InfoGainAttributeEval evaluator algorithm in the Weka 'Select Attributes' tab to perform feature selection which reduced the total number of instances to around one-third compared to the initial given instances.

We saved this set in the Weka 'Preprocess' tab in the arff format and supplied it to the optimized models constructed above. And the ROC values before and after feature selection are shown in the table below. The screenshots of the models after feature extraction are shown in Appendix B.



Column1	RandomForest	RandomForest2	RandomTree	RandomTree3	REPTree	REPTree4
	Before_Feature_Slection	After_Feature_Slection	Before_Feature_Slection	After_Feature_Slection	Before_Feature_Slection	After_Feature_Slection
nr_ar_ROC	0.83	0.999	0.795	0.962	0.802	0.9
nr_aroma_ROC	0.853	0.999	0.772	0.933	0.785	0.859
sr_are_ROC	0.855	0.998	0.7	0.819	0.712	0.846

Table.2

3.4.MODELS EVALUATION

For Model Evaluation, we noted the MCC values of 9 models by changing the seed for 50 times. All these 50 MCC values for 9 models are plotted and shown in the graph below.

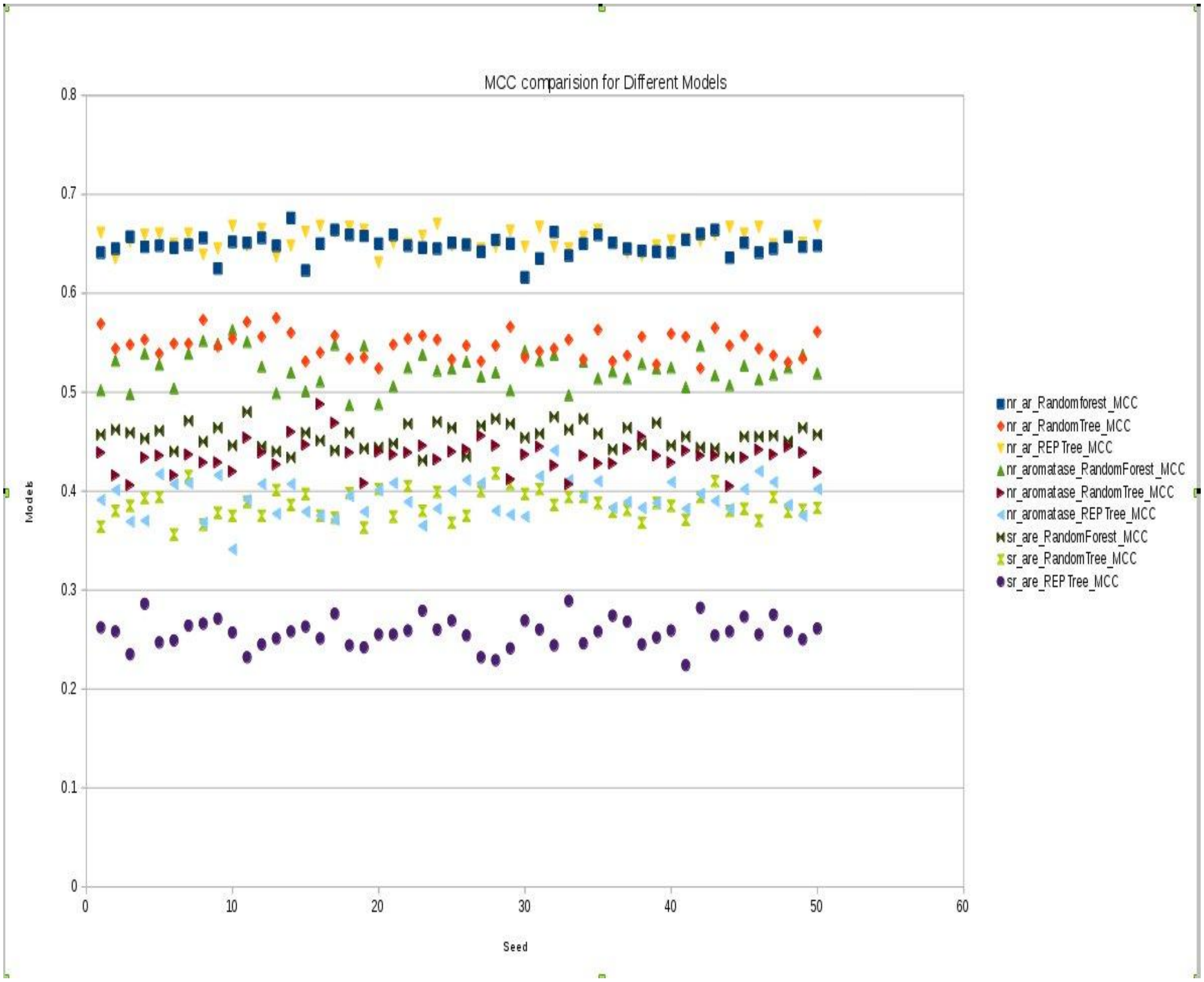


Figure.1

4. RESULTS ANALYSIS AND DISCUSSION

After the step-wise experiments and the results obtained above, the best values from each case are chosen.

- In case of optimizing parameters, the parameters for which highest ROC value is obtained is chosen in each model.
- Next, feature selection algorithms mentioned in section 3.3. are used along with the machine learning algorithms mentioned in section 3.1 to select attributes that can contribute best in the prediction of results.
- For model evaluation, MCC values are noted 50 times for each model, and best MCC value is observed. The corresponding seed value is noted to use in the final model construction.
- From, the figure.1, it is clear that, for each Assay, RandomForest gave best results.
- Finally, based on optimizing, feature selection and model evaluation, we selected Assay EECS_738_nr_ar_train.arff with RandomForest classifier as best for final model construction.

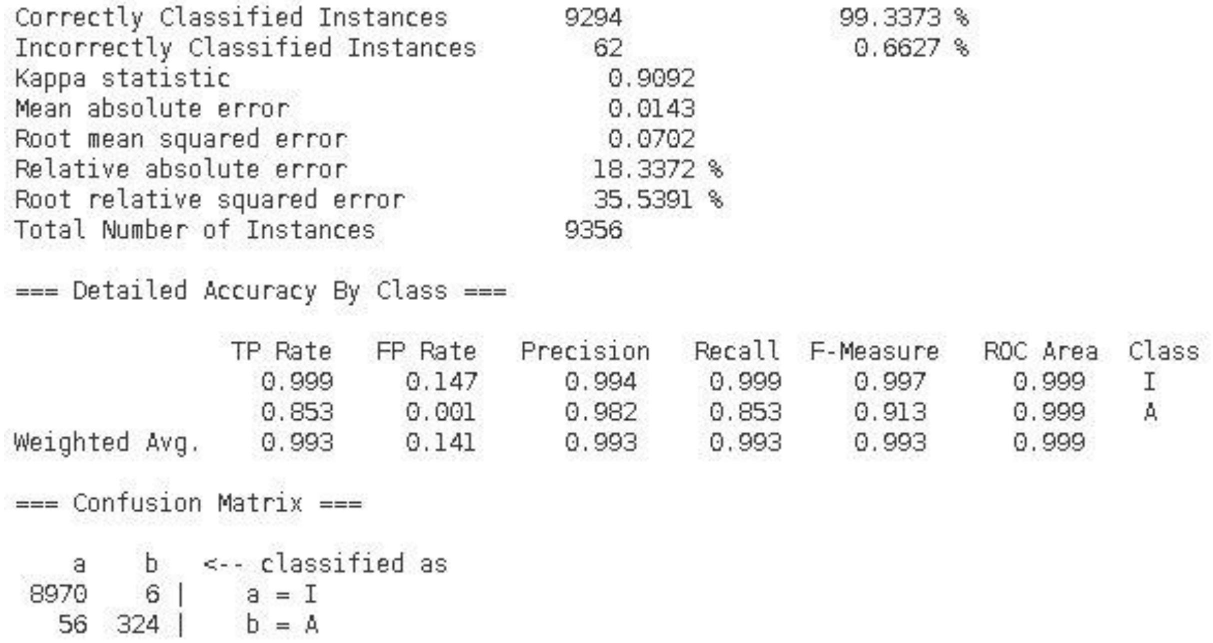
5.FINAL MODEL CONSTRUCTION

Based on the results analysis, the final model is constructed with the Assay, classifier, feature selection algorithms and parameter values shown in Table.3.

Assay Name	EECS_738_nr_ar_train.arff
Classifier Name	RandomForest
Evaluator Name	InfoGainAttributeEval
Search Algorithm Name	Ranker (Threshold: 0.0)
Parameter Values	-I 60, -K 30, -S 14

Table.3

The screenshot of the final model constructed with the above mentioned values is shown in figure.2 below.



```
Correctly Classified Instances        9294               99.3373 %
Incorrectly Classified Instances        62                0.6627 %
Kappa statistic                          0.9092
Mean absolute error                      0.0143
Root mean squared error                  0.0702
Relative absolute error                 18.3372 %
Root relative squared error             35.5391 %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.147      0.994     0.999     0.997      0.999    I
                 0.853     0.001      0.982     0.853     0.913      0.999    A
Weighted Avg.    0.993     0.141      0.993     0.993     0.993      0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 8970    6 |    a = I
   56  324 |    b = A
```

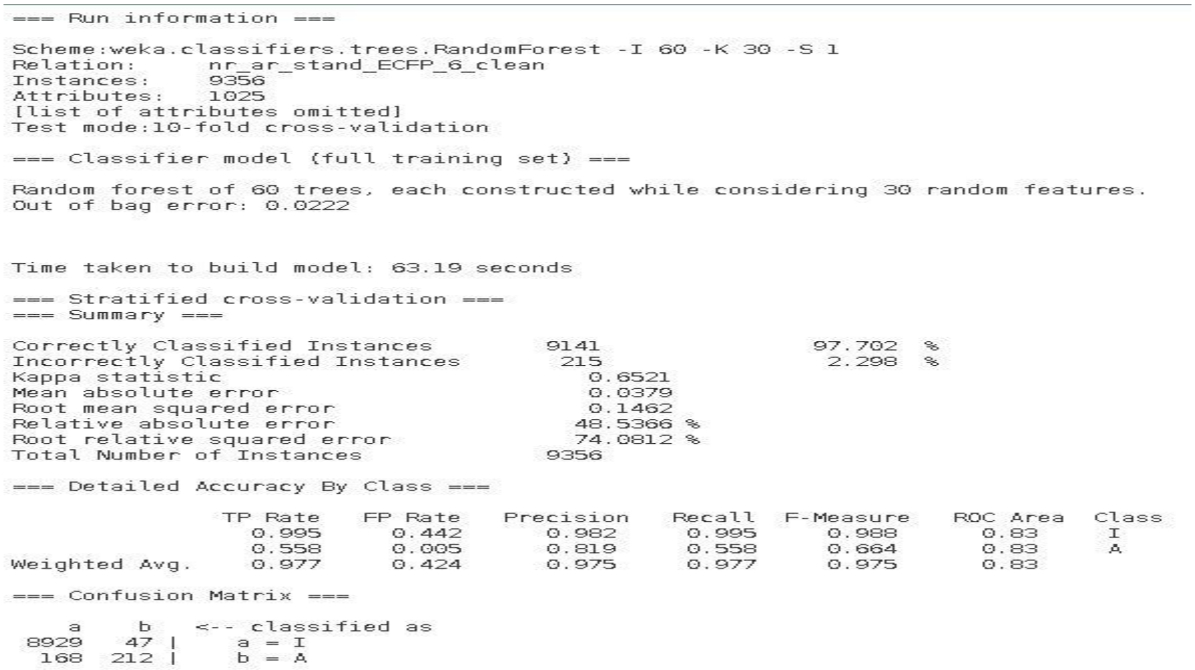
figure.2

6.PREDICTION RESULTS

The prediction results for final model are attached as a separate file with the name EECS738_Bommidi_Doddala_(EECS_738_nr_ar_train.arff)_testing.txt

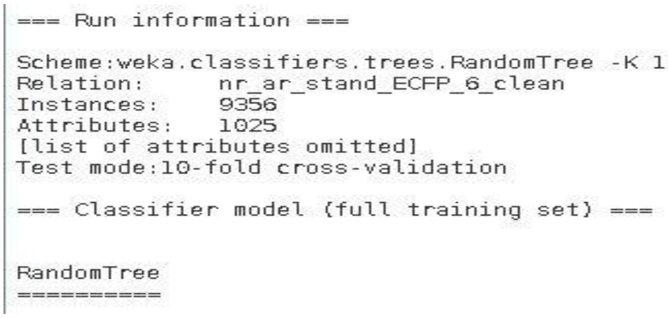
7. APPENDIX A

nr_ar_train_random_forest_optimized.model

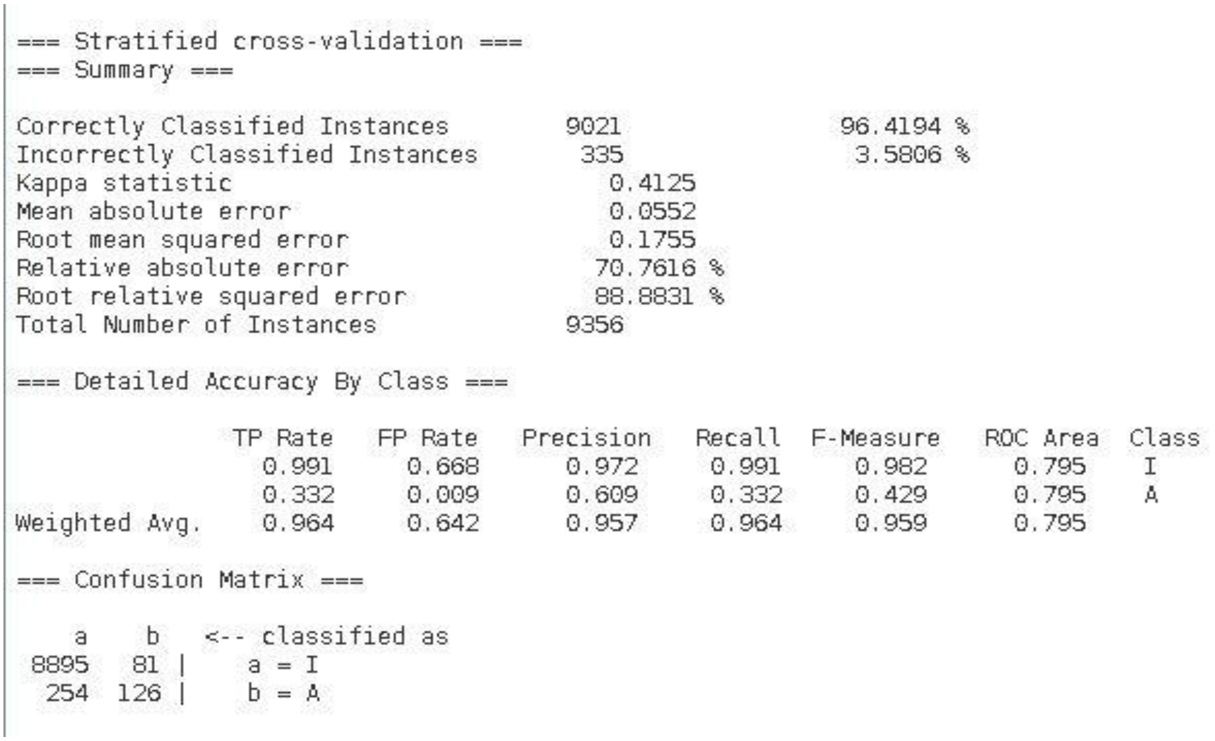


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 60 -K 30 -S 1
Relation:     nr_ar_stand_ECFP_6_clean
Instances:    9356
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation

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

Random forest of 60 trees, each constructed while considering 30 random features.
Out of bag error: 0.0222


Time taken to build model: 63.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9141               97.702  %
Incorrectly Classified Instances       215                2.298  %
Kappa statistic                          0.6521
Mean absolute error                      0.0379
Root mean squared error                  0.1462
Relative absolute error                 48.5366 %
Root relative squared error             74.0812 %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995     0.442      0.982     0.995     0.988      0.83     I
                 0.558     0.005      0.819     0.558     0.664      0.83     A
Weighted Avg.    0.977     0.424      0.975     0.977     0.975      0.83

=== Confusion Matrix ===

    a    b   <-- classified as
 8929   47 |    a = I
  168  212 |    b = A
```

nr_ar_train_random_tree_optimized.model

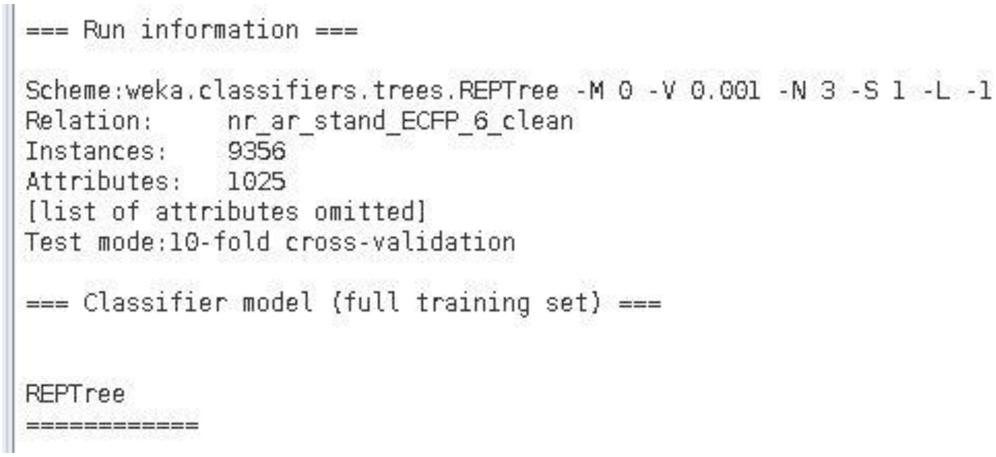


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 17 -M 35.0 -S 1
Relation:     nr_ar_stand_ECFP_6_clean
Instances:    9356
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation

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


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

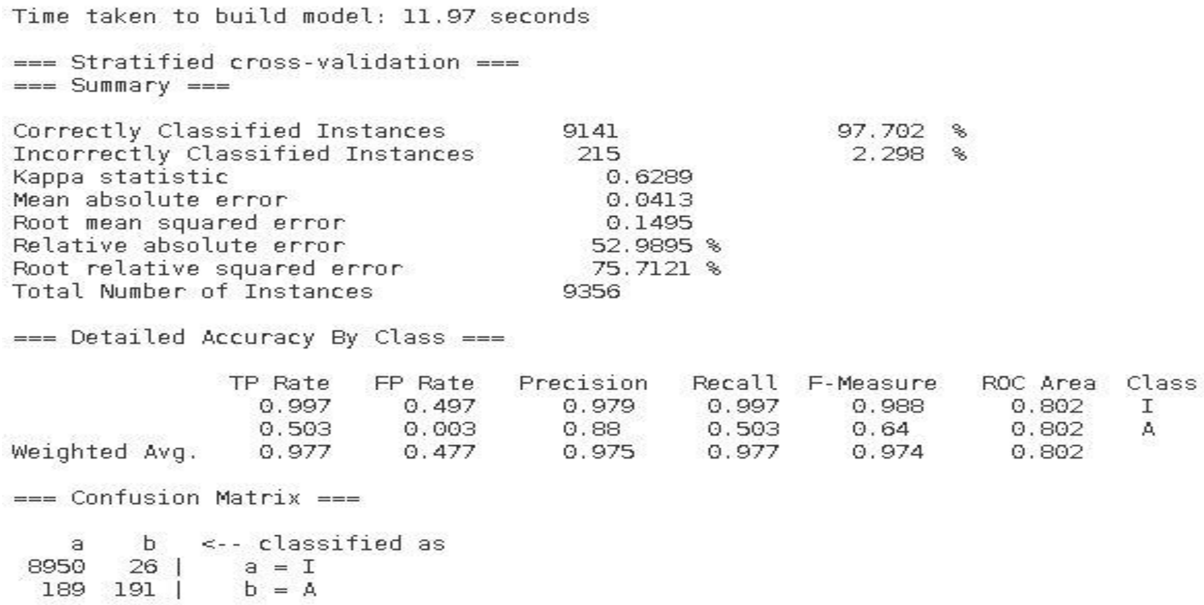


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9021               96.4194 %
Incorrectly Classified Instances       335                3.5806 %
Kappa statistic                          0.4125
Mean absolute error                      0.0552
Root mean squared error                  0.1755
Relative absolute error                 70.7616 %
Root relative squared error             88.8831 %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.991     0.668      0.972     0.991     0.982      0.795    I
                 0.332     0.009      0.609     0.332     0.429      0.795    A
Weighted Avg.    0.964     0.642      0.957     0.964     0.959      0.795

=== Confusion Matrix ===

    a    b   <-- classified as
 8895   81 |    a = I
  254  126 |    b = A
```

nr_ar_train_rep_tree_optimized.model

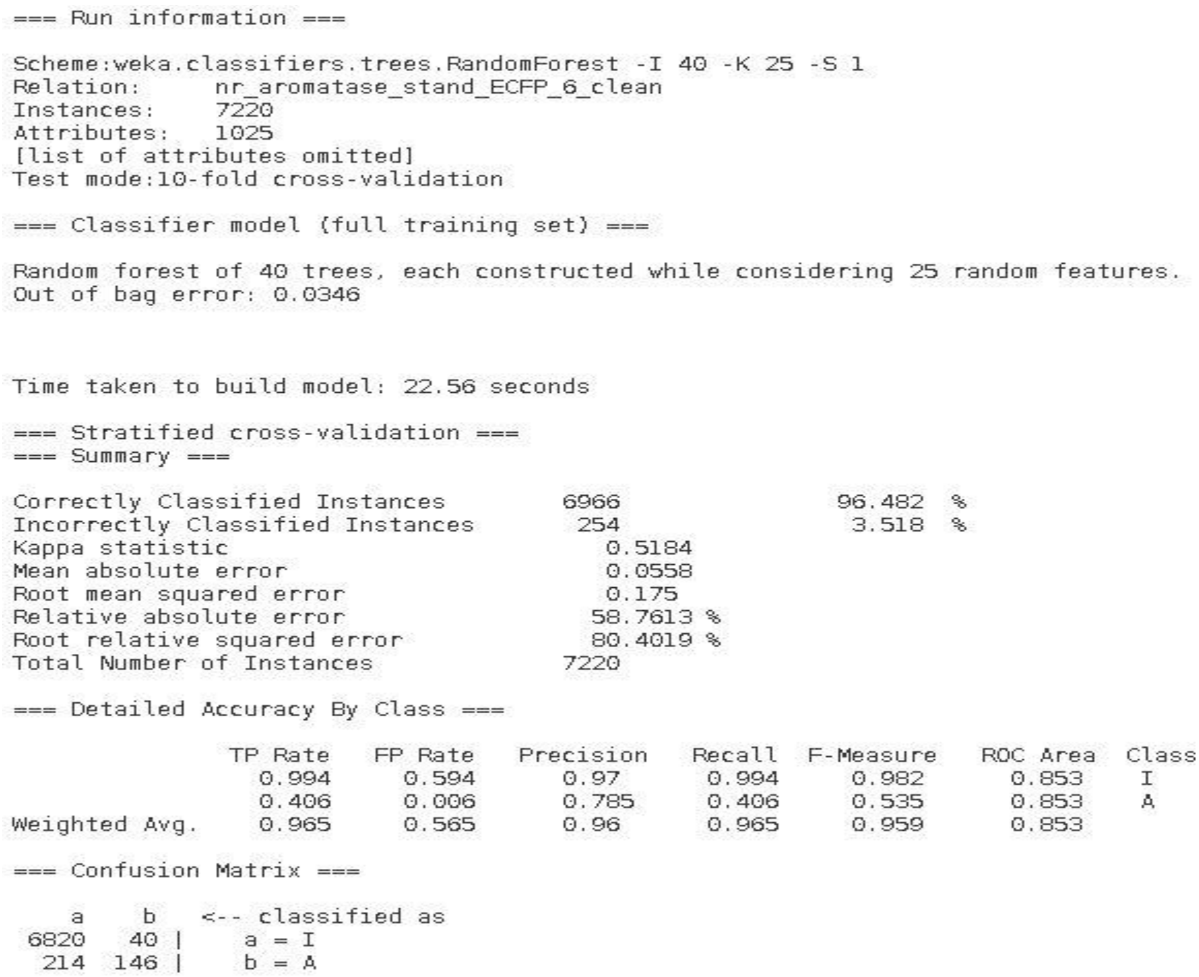


```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 0 -V 0.001 -N 3 -S 1 -L -1
Relation:     nr_ar_stand_ECFP_6_clean
Instances:    9356
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation

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


REPTree
============
```



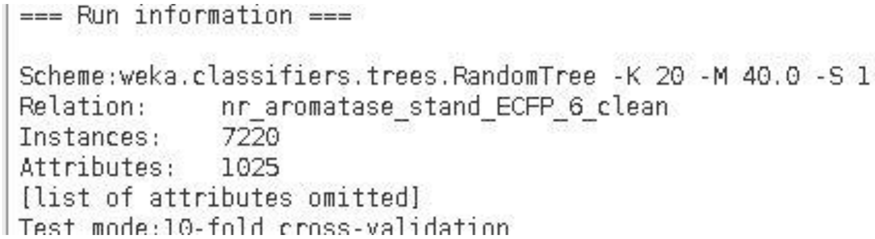
```
Time taken to build model: 11.97 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9141               97.702  %
Incorrectly Classified Instances       215                2.298  %
Kappa statistic                          0.6289
Mean absolute error                      0.0413
Root mean squared error                  0.1495
Relative absolute error                 52.9895 %
Root relative squared error             75.7121 %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.997     0.497      0.979     0.997     0.988      0.802    I
                 0.503     0.003      0.88      0.503     0.64       0.802    A
Weighted Avg.    0.977     0.477      0.975     0.977     0.974      0.802

=== Confusion Matrix ===

    a    b   <-- classified as
 8950   26 |    a = I
  189  191 |    b = A
```

nr_aromatase_random_forest_optimized.model

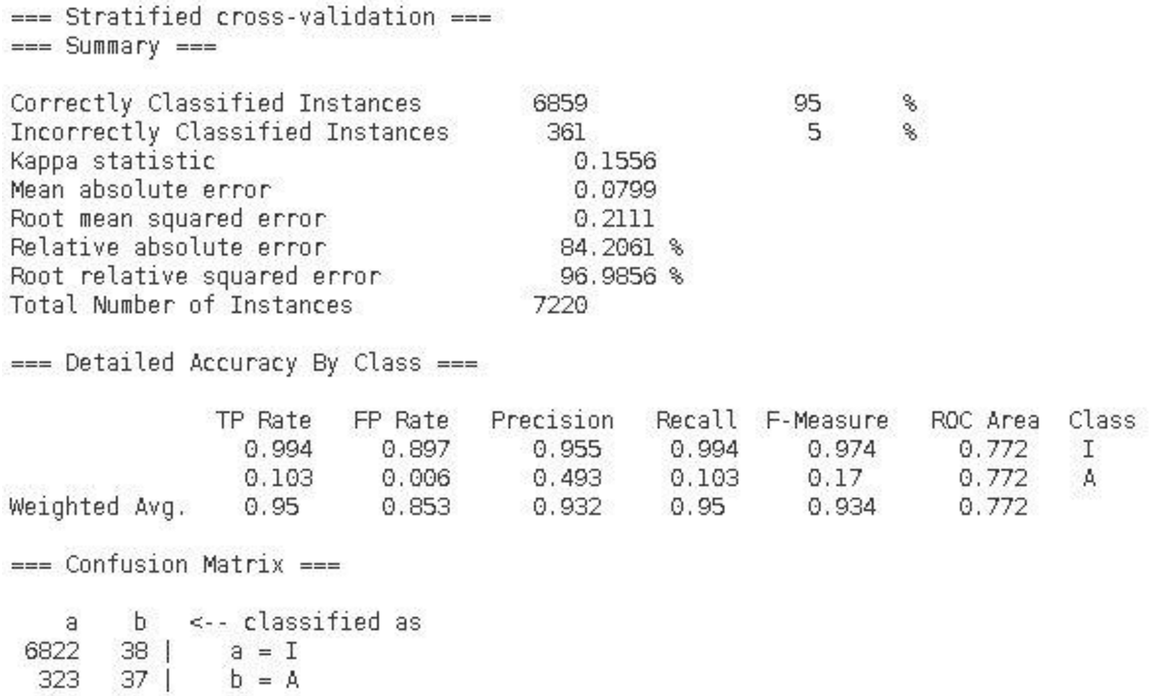


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 40 -K 25 -S 1
Relation:     nr_aromatase_stand_ECFP_6_clean
Instances:    7220
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation

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

Random forest of 40 trees, each constructed while considering 25 random features.
Out of bag error: 0.0346



Time taken to build model: 22.56 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6966               96.482  %
Incorrectly Classified Instances       254                3.518  %
Kappa statistic                          0.5184
Mean absolute error                      0.0558
Root mean squared error                  0.175
Relative absolute error                 58.7613 %
Root relative squared error             80.4019 %
Total Number of Instances             7220

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.994     0.594      0.97      0.994     0.982      0.853    I
                 0.406     0.006      0.785     0.406     0.535      0.853    A
Weighted Avg.    0.965     0.565      0.96      0.965     0.959      0.853

=== Confusion Matrix ===

    a    b   <-- classified as
 6820   40 |    a = I
  214  146 |    b = A
```

nr_aromatase_random_tree_optimized.model

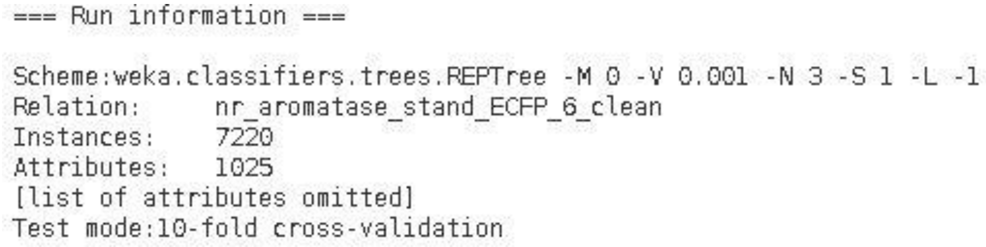


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 20 -M 40.0 -S 1
Relation:     nr_aromatase_stand_ECFP_6_clean
Instances:    7220
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation
```

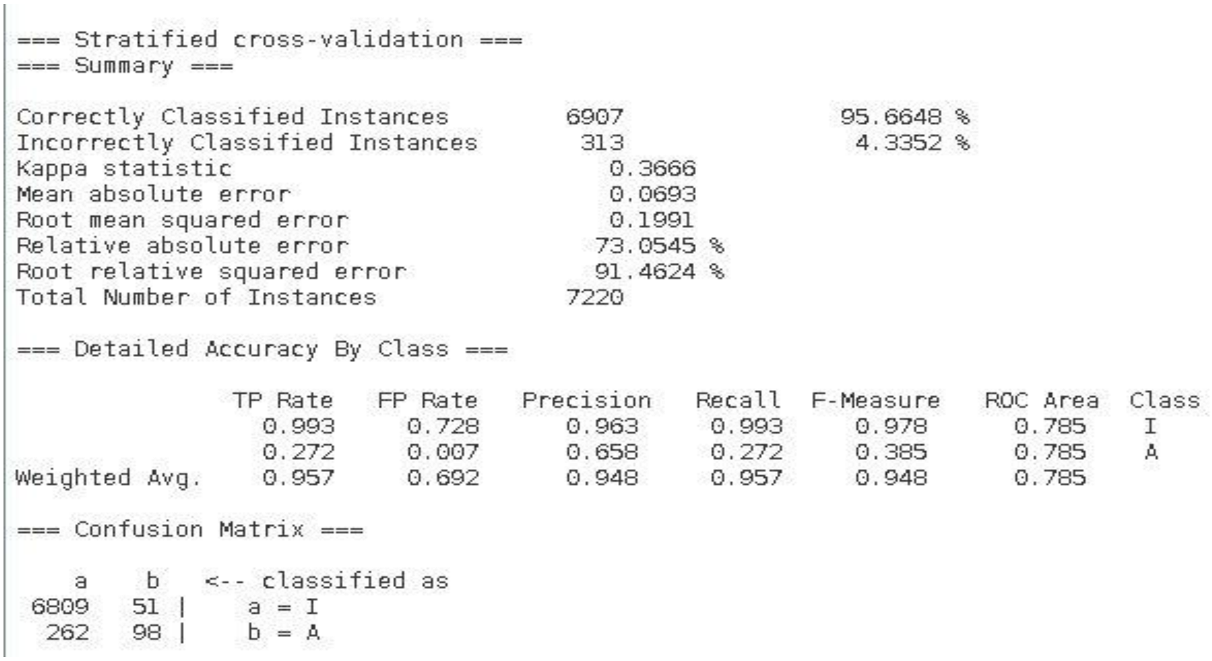


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6859               95      %
Incorrectly Classified Instances       361                5      %
Kappa statistic                          0.1556
Mean absolute error                      0.0799
Root mean squared error                  0.2111
Relative absolute error                 84.2061 %
Root relative squared error             96.9856 %
Total Number of Instances             7220

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.994     0.897      0.955     0.994     0.974      0.772    I
                 0.103     0.006      0.493     0.103     0.17       0.772    A
Weighted Avg.    0.95      0.853      0.932     0.95      0.934      0.772

=== Confusion Matrix ===

    a    b   <-- classified as
 6822   38 |    a = I
  323   37 |    b = A
```

nr_aromatase_rep_tree_optimized.model

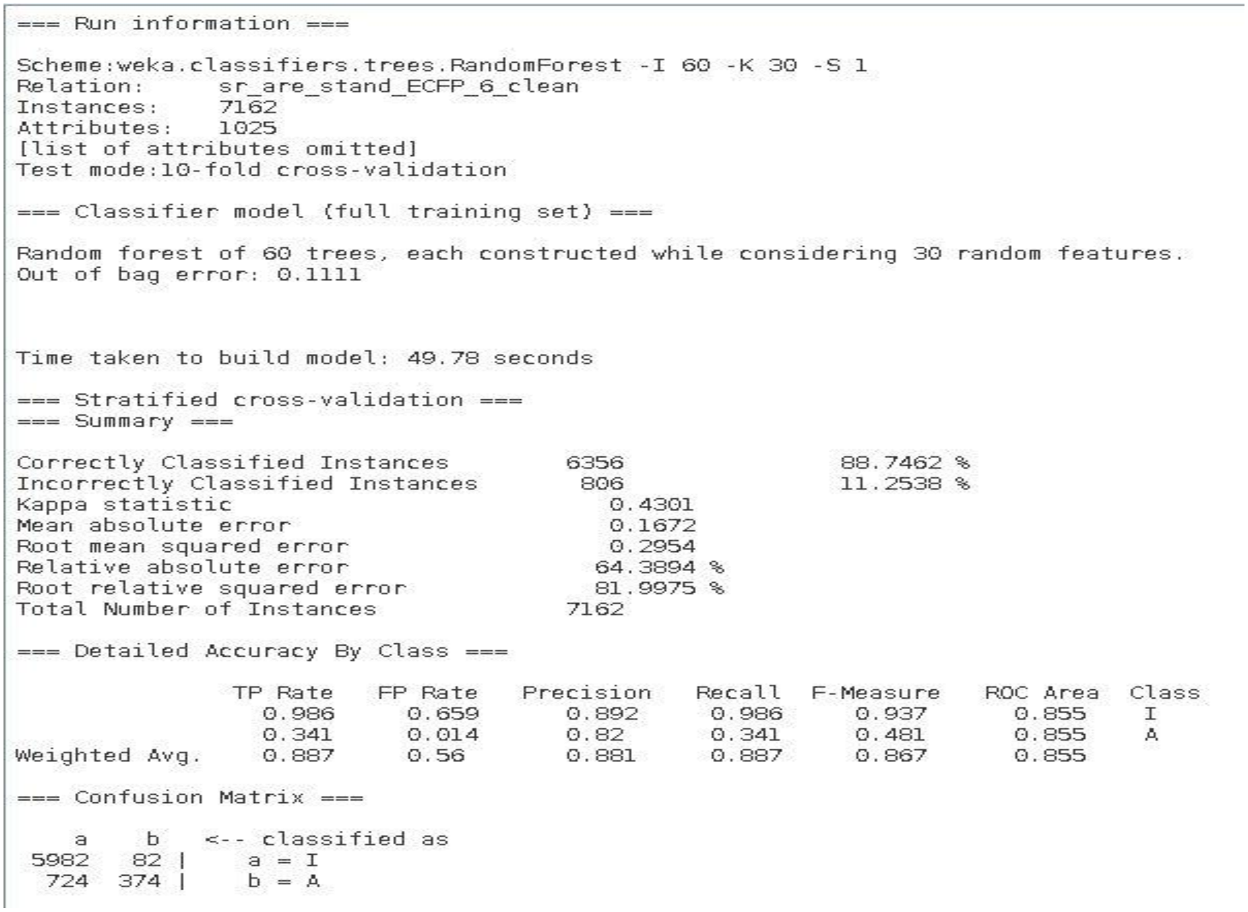


```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 0 -V 0.001 -N 3 -S 1 -L -1
Relation:     nr_aromatase_stand_ECFP_6_clean
Instances:    7220
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation
```



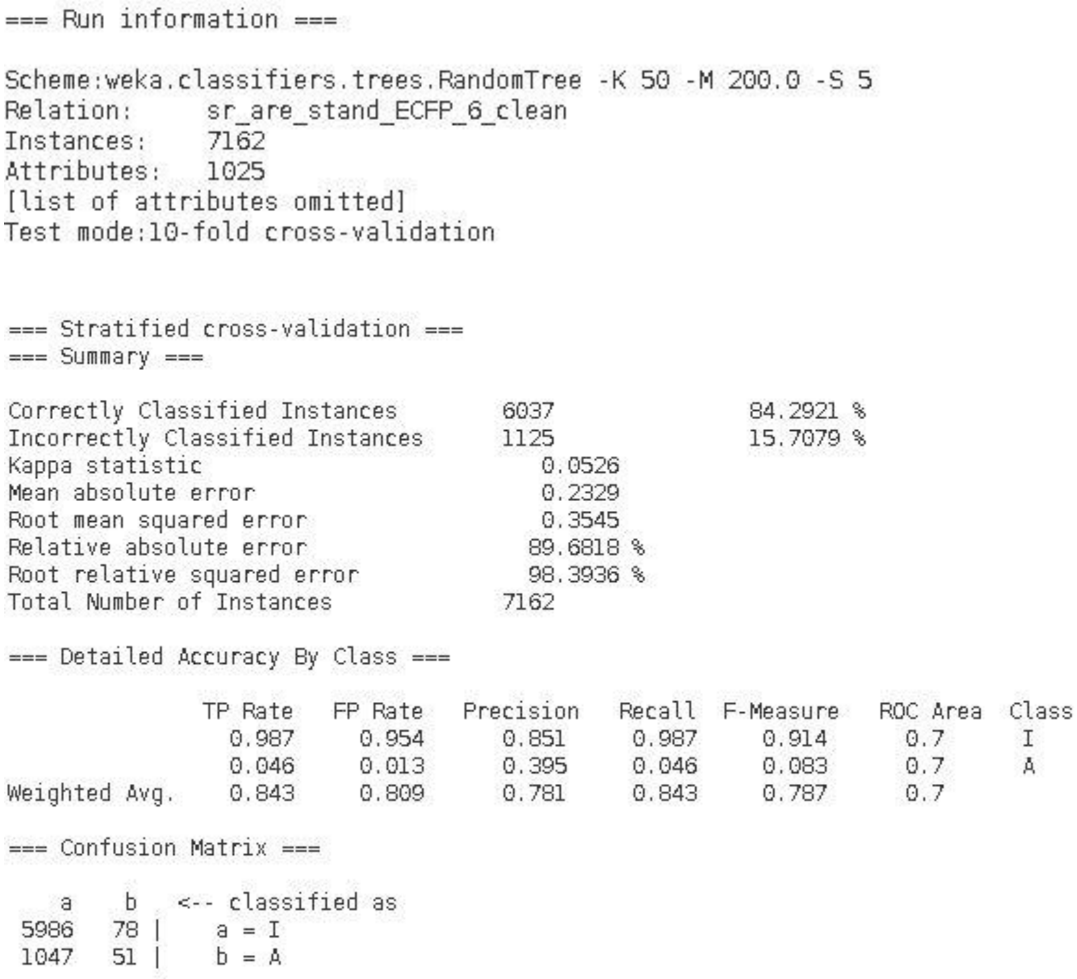
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6907               95.6648 %
Incorrectly Classified Instances       313                4.3352 %
Kappa statistic                          0.3666
Mean absolute error                      0.0693
Root mean squared error                  0.1991
Relative absolute error                 73.0545 %
Root relative squared error             91.4624 %
Total Number of Instances             7220

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.993     0.728      0.963     0.993     0.978      0.785    I
                 0.272     0.007      0.658     0.272     0.385      0.785    A
Weighted Avg.    0.957     0.692      0.948     0.957     0.948      0.785

=== Confusion Matrix ===

    a    b   <-- classified as
 6809   51 |    a = I
  262   98 |    b = A
```

sr_are_random_forest_optimized.model



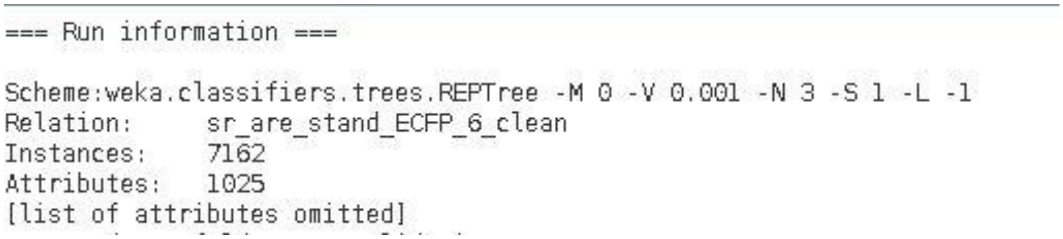
```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 60 -K 30 -S 1
Relation:     sr_are_stand_ECFP_6_clean
Instances:    7162
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation

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

Random forest of 60 trees, each constructed while considering 30 random features.
Out of bag error: 0.1111



Time taken to build model: 49.78 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6356               88.7462 %
Incorrectly Classified Instances       806               11.2538 %
Kappa statistic                          0.4301
Mean absolute error                      0.1672
Root mean squared error                  0.2954
Relative absolute error                 64.3894 %
Root relative squared error             81.9975 %
Total Number of Instances             7162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.986     0.659      0.892     0.986     0.937      0.855    I
                 0.341     0.014      0.82      0.341     0.481      0.855    A
Weighted Avg.    0.887     0.56       0.881     0.887     0.867      0.855

=== Confusion Matrix ===

    a    b   <-- classified as
 5982   82 |    a = I
  724  374 |    b = A
```

sr_are_random_tree_optimized.model

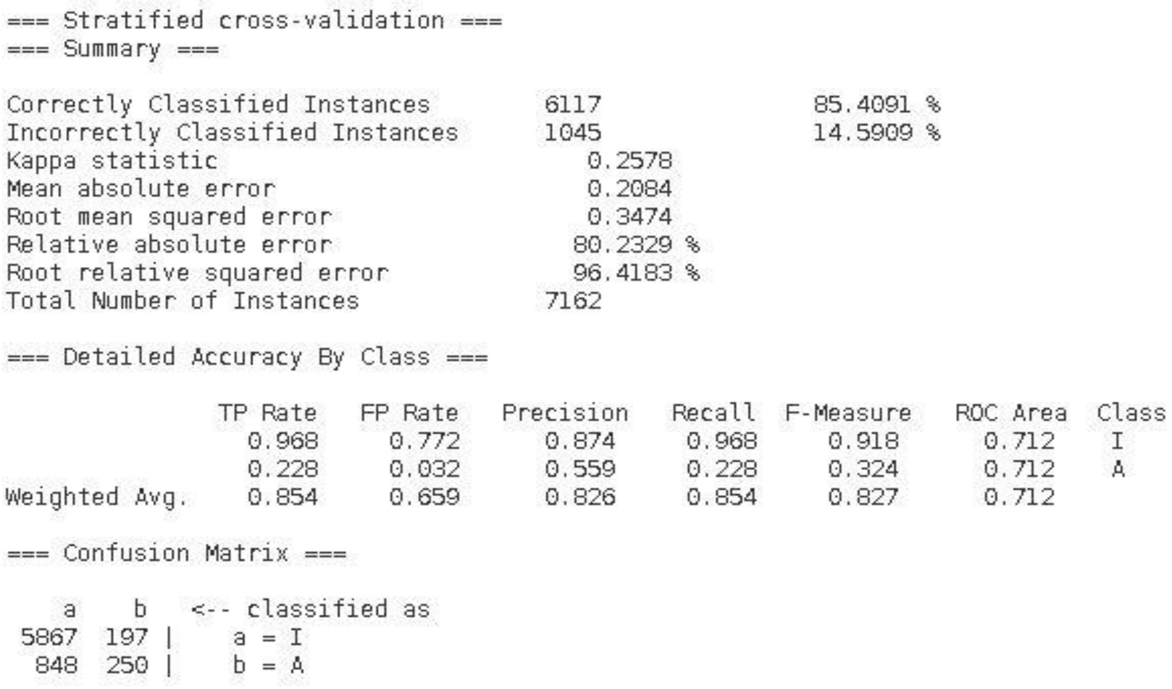


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 50 -M 200.0 -S 5
Relation:     sr_are_stand_ECFP_6_clean
Instances:    7162
Attributes:   1025
[list of attributes omitted]
Test mode:10-fold cross-validation


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6037               84.2921 %
Incorrectly Classified Instances      1125               15.7079 %
Kappa statistic                          0.0526
Mean absolute error                      0.2329
Root mean squared error                  0.3545
Relative absolute error                 89.6818 %
Root relative squared error             98.3936 %
Total Number of Instances             7162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.987     0.954      0.851     0.987     0.914      0.7      I
                 0.046     0.013      0.395     0.046     0.083      0.7      A
Weighted Avg.    0.843     0.809      0.781     0.843     0.787      0.7

=== Confusion Matrix ===

    a    b   <-- classified as
 5986   78 |    a = I
 1047   51 |    b = A
```

sr_are_rep_tree_optimized.model



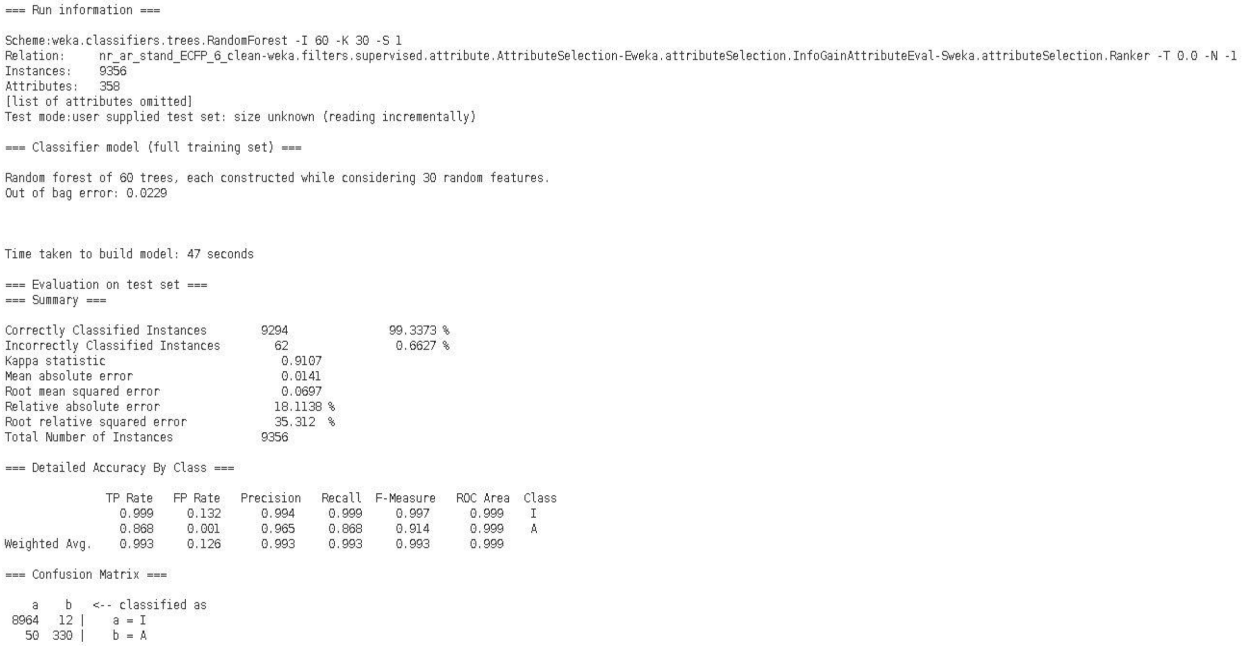
```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 0 -V 0.001 -N 3 -S 1 -L -1
Relation:     sr_are_stand_ECFP_6_clean
Instances:    7162
Attributes:   1025
[list of attributes omitted]
```



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6117               85.4091 %
Incorrectly Classified Instances      1045               14.5909 %
Kappa statistic                          0.2578
Mean absolute error                      0.2084
Root mean squared error                  0.3474
Relative absolute error                 80.2329 %
Root relative squared error             96.4183 %
Total Number of Instances             7162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.968     0.772      0.874     0.968     0.918      0.712    I
                 0.228     0.032      0.559     0.228     0.324      0.712    A
Weighted Avg.    0.854     0.659      0.826     0.854     0.827      0.712

=== Confusion Matrix ===

    a    b   <-- classified as
 5867  197 |    a = I
  848  250 |    b = A
```

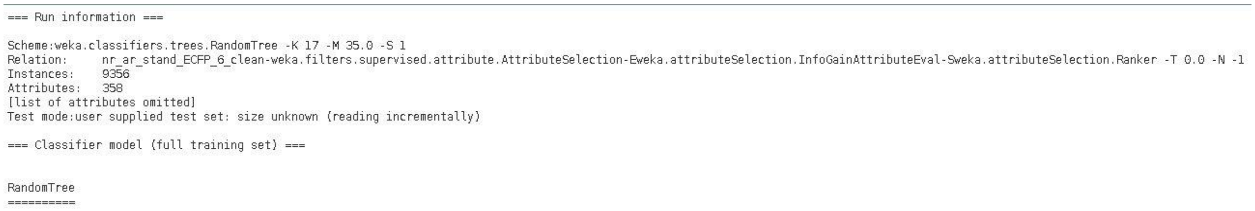
8.APPENDIX B

nr_ar_train_random_forest_feature_selection.model

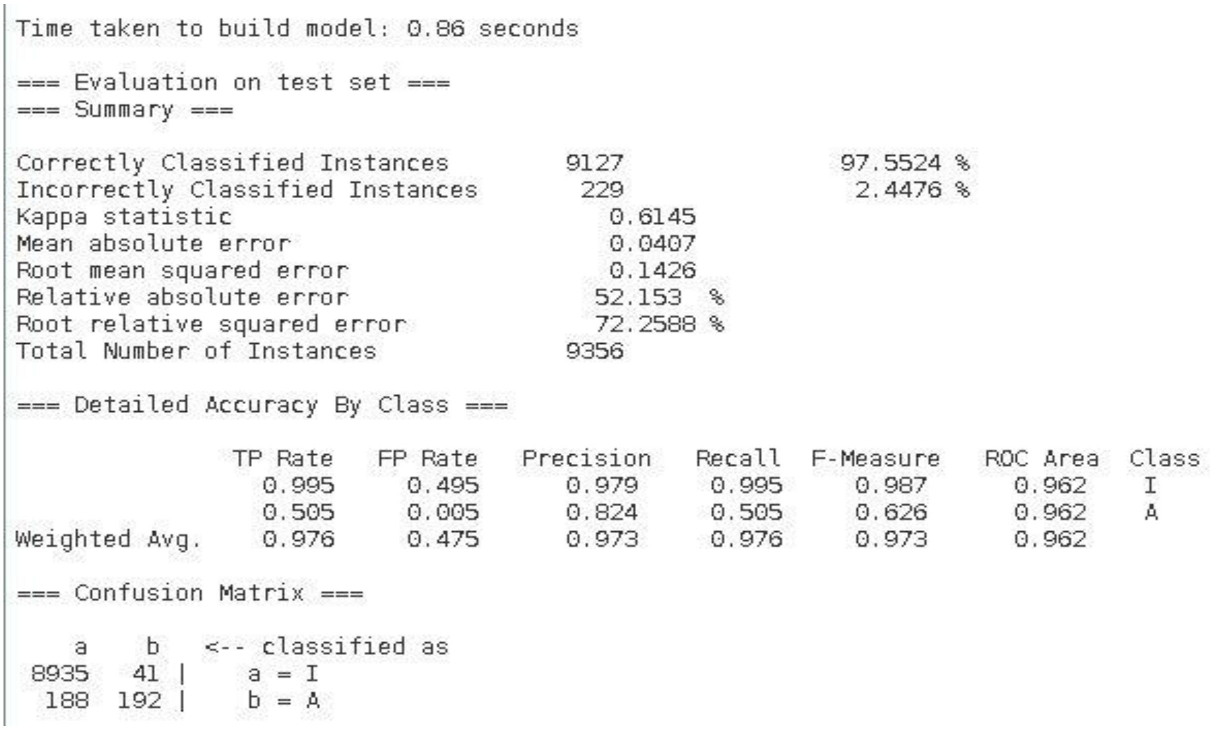


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 60 -K 30 -S 1
Relation:     nr_ar_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    9356
Attributes:   358
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)

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

Random forest of 60 trees, each constructed while considering 30 random features.
Out of bag error: 0.0229



Time taken to build model: 47 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        9294               99.3373 %
Incorrectly Classified Instances        62                0.6627 %
Kappa statistic                          0.9107
Mean absolute error                      0.0141
Root mean squared error                  0.0697
Relative absolute error                 18.1138 %
Root relative squared error             35.312  %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.132      0.994     0.999     0.997      0.999    I
                 0.868     0.001      0.965     0.868     0.914      0.999    A
Weighted Avg.    0.993     0.126      0.993     0.993     0.993      0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 8964   12 |    a = I
   50  330 |    b = A
```

nr_ar_random_tree_feature_selection.model

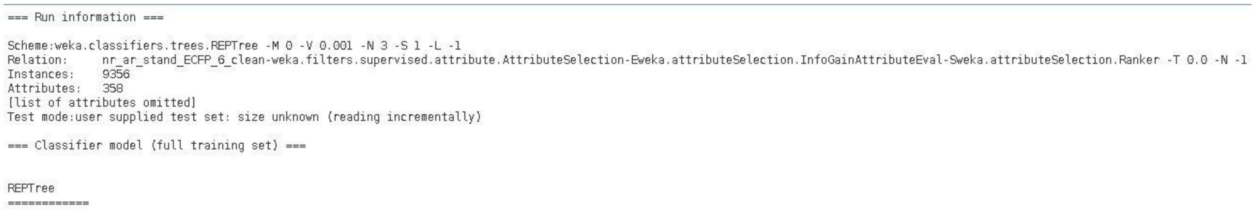


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 17 -M 35.0 -S 1
Relation:     nr_ar_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    9356
Attributes:   358
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)

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


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

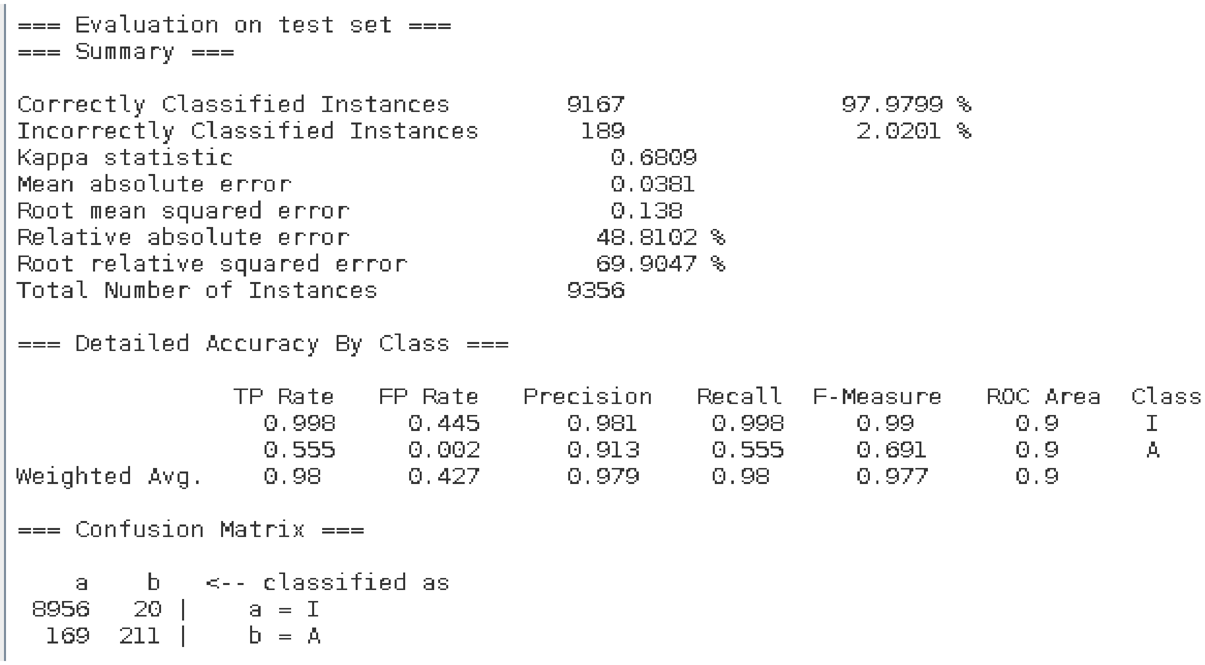


```
Time taken to build model: 0.86 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        9127               97.5524 %
Incorrectly Classified Instances       229                2.4476 %
Kappa statistic                          0.6145
Mean absolute error                      0.0407
Root mean squared error                  0.1426
Relative absolute error                 52.153  %
Root relative squared error             72.2588 %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995     0.495      0.979     0.995     0.987      0.962    I
                 0.505     0.005      0.824     0.505     0.626      0.962    A
Weighted Avg.    0.976     0.475      0.973     0.976     0.973      0.962

=== Confusion Matrix ===

    a    b   <-- classified as
 8935   41 |    a = I
  188  192 |    b = A
```

nr_ar_rep_tree_feature_selection.model

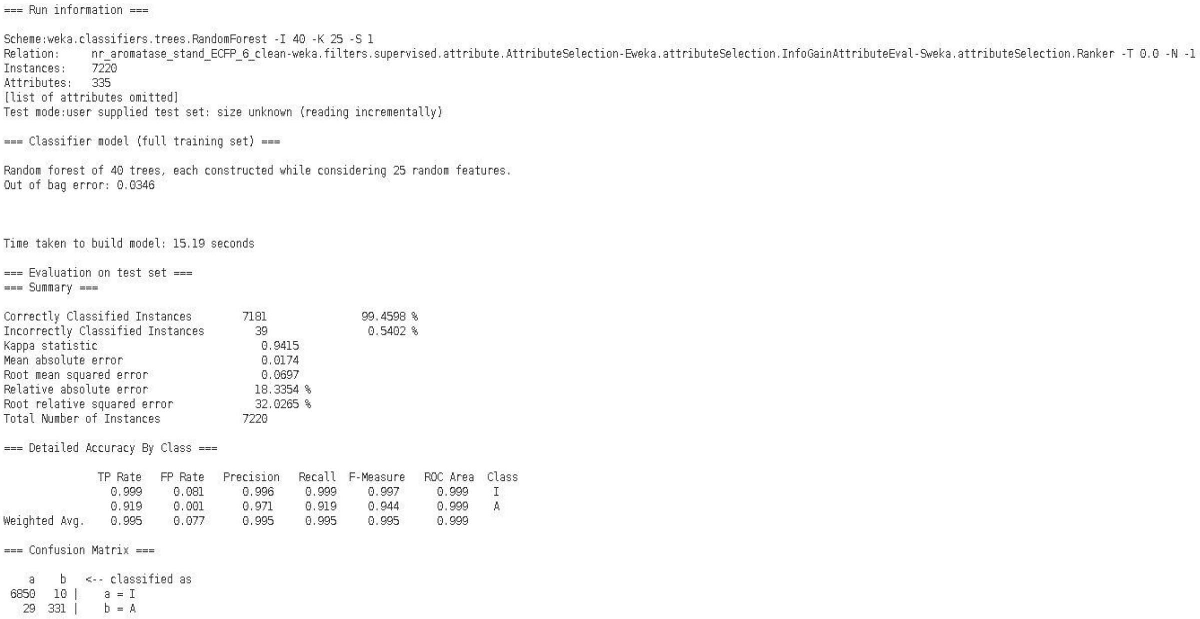


```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 0 -V 0.001 -N 3 -S 1 -L -1
Relation:     nr_ar_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    9356
Attributes:   358
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)

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


REPTree
============
```



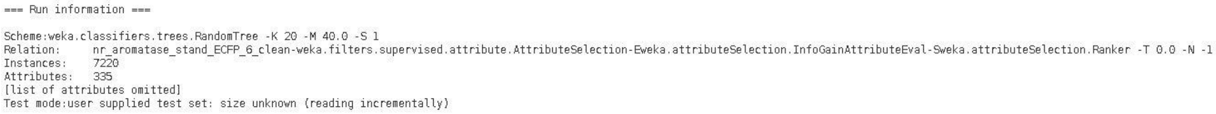
```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        9167               97.9799 %
Incorrectly Classified Instances       189                2.0201 %
Kappa statistic                          0.6809
Mean absolute error                      0.0381
Root mean squared error                  0.138
Relative absolute error                 48.8102 %
Root relative squared error             69.9047 %
Total Number of Instances             9356

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.445      0.981     0.998     0.99       0.9      I
                 0.555     0.002      0.913     0.555     0.691      0.9      A
Weighted Avg.    0.98      0.427      0.979     0.98      0.977      0.9

=== Confusion Matrix ===

    a    b   <-- classified as
 8956   20 |    a = I
  169  211 |    b = A
```

nr_aromatase_random_forest_feature_selection.model

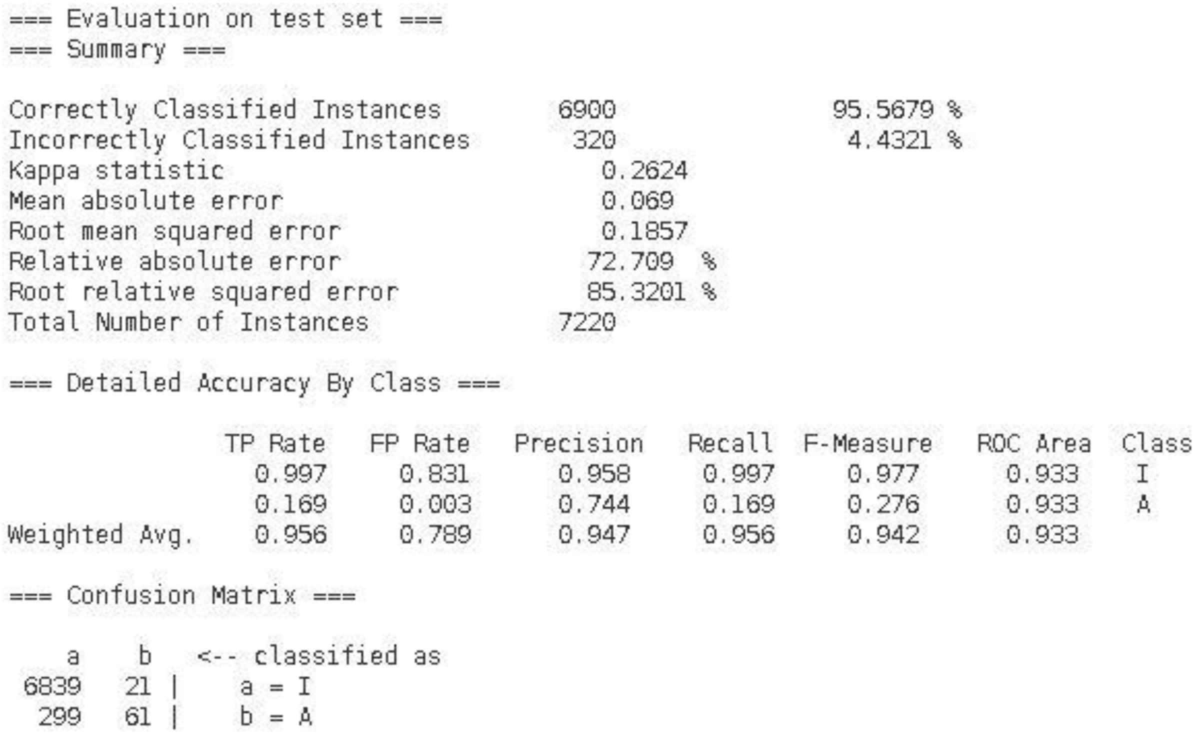


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 40 -K 25 -S 1
Relation:     nr_aromatase_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    7220
Attributes:   335
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)

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

Random forest of 40 trees, each constructed while considering 25 random features.
Out of bag error: 0.0346



Time taken to build model: 15.19 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        7181               99.4598 %
Incorrectly Classified Instances        39                0.5402 %
Kappa statistic                          0.9415
Mean absolute error                      0.0174
Root mean squared error                  0.0697
Relative absolute error                 18.3354 %
Root relative squared error             32.0265 %
Total Number of Instances             7220

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.081      0.996     0.999     0.997      0.999    I
                 0.919     0.001      0.971     0.919     0.944      0.999    A
Weighted Avg.    0.995     0.077      0.995     0.995     0.995      0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 6850   10 |    a = I
   29  331 |    b = A
```

nr_aromatase_random_tree_feature_selection.model

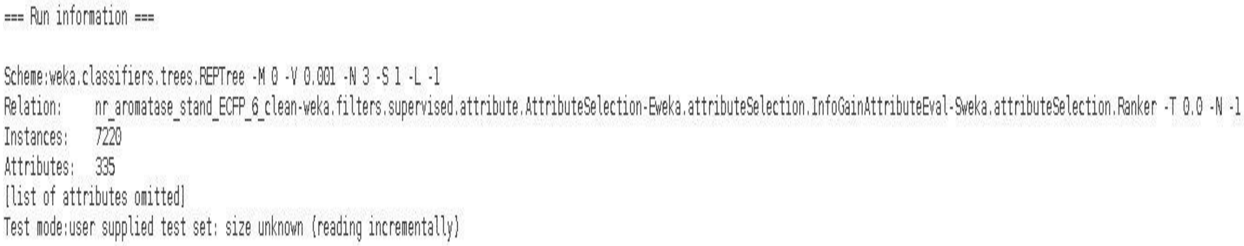


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 20 -M 40.0 -S 1
Relation:     nr_aromatase_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    7220
Attributes:   335
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)
```

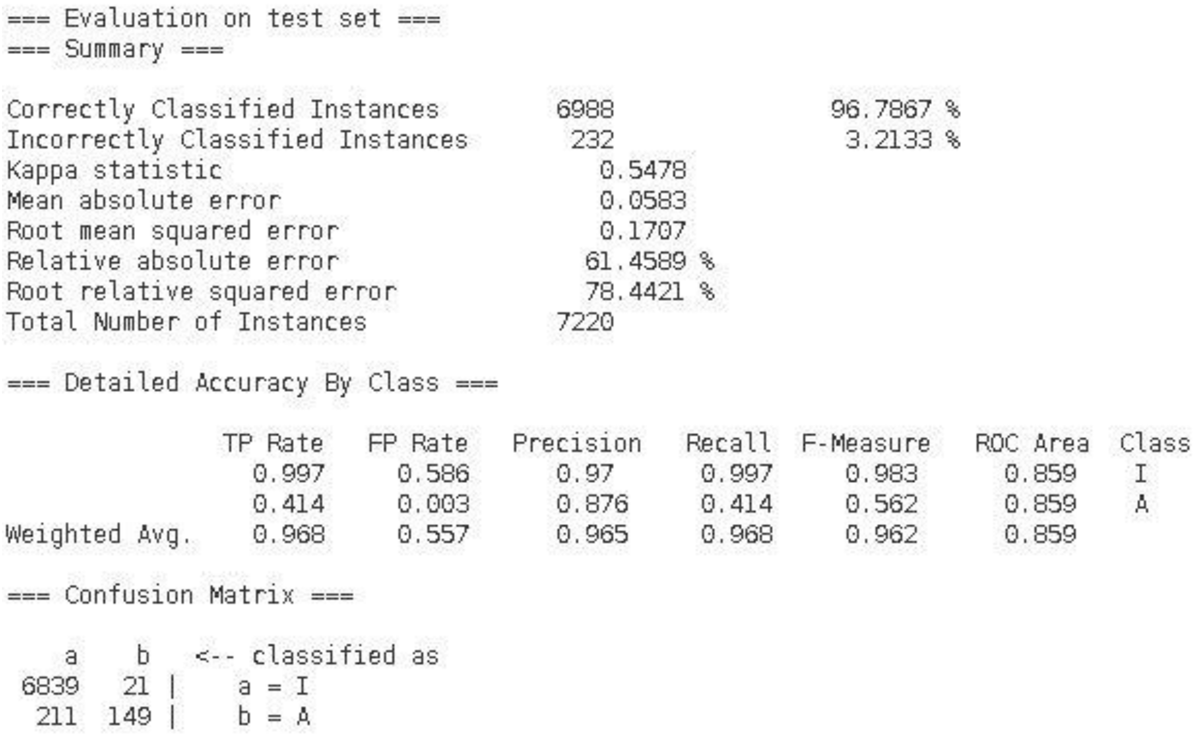


```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        6900               95.5679 %
Incorrectly Classified Instances       320                4.4321 %
Kappa statistic                          0.2624
Mean absolute error                      0.069
Root mean squared error                  0.1857
Relative absolute error                 72.709  %
Root relative squared error             85.3201 %
Total Number of Instances             7220

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.997     0.831      0.958     0.997     0.977      0.933    I
                 0.169     0.003      0.744     0.169     0.276      0.933    A
Weighted Avg.    0.956     0.789      0.947     0.956     0.942      0.933

=== Confusion Matrix ===

    a    b   <-- classified as
 6839   21 |    a = I
  299   61 |    b = A
```

nr_aromatase_rep_tree_feature_selection.model

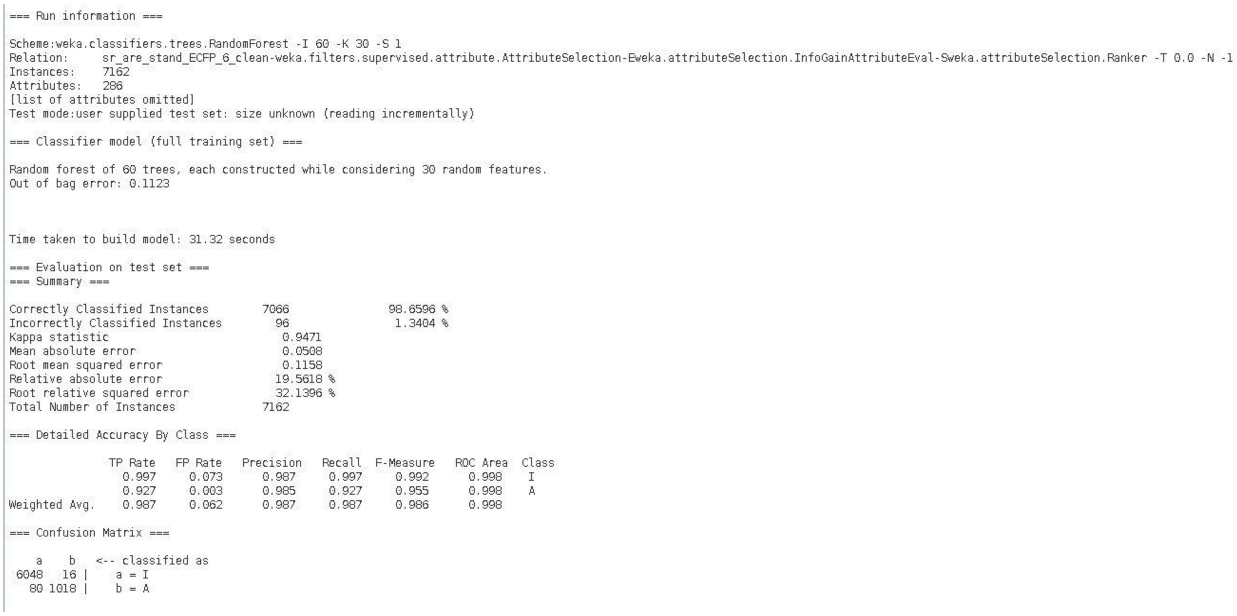


```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 0 -V 0.001 -N 3 -S 1 -L -1
Relation:     nr_aromatase_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    7220
Attributes:   335
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)
```



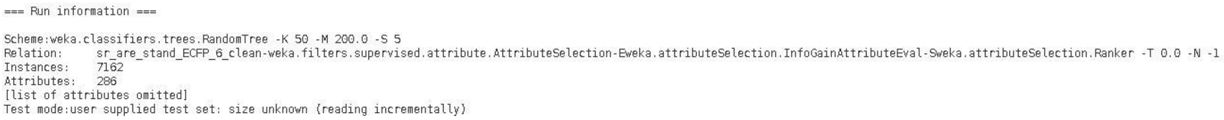
```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        6988               96.7867 %
Incorrectly Classified Instances       232                3.2133 %
Kappa statistic                          0.5478
Mean absolute error                      0.0583
Root mean squared error                  0.1707
Relative absolute error                 61.4589 %
Root relative squared error             78.4421 %
Total Number of Instances             7220

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.997     0.586      0.97      0.997     0.983      0.859    I
                 0.414     0.003      0.876     0.414     0.562      0.859    A
Weighted Avg.    0.968     0.557      0.965     0.968     0.962      0.859

=== Confusion Matrix ===

    a    b   <-- classified as
 6839   21 |    a = I
  211  149 |    b = A
```

sr_are_random_forest_feature_selection.model

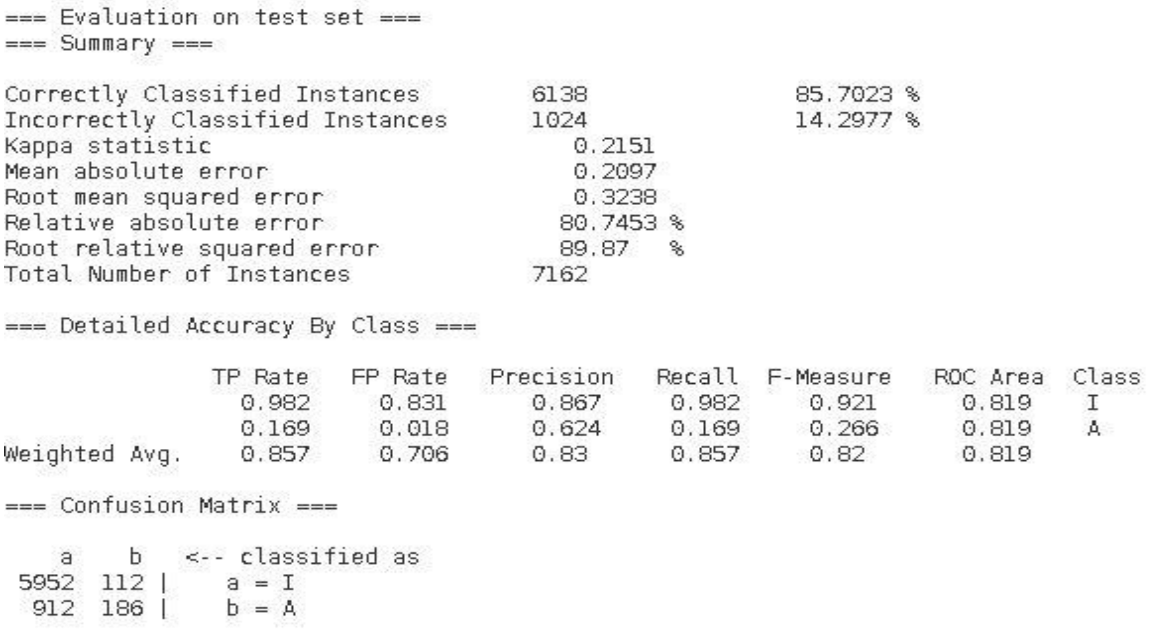


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 60 -K 30 -S 1
Relation:     sr_are_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    7162
Attributes:   286
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)

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

Random forest of 60 trees, each constructed while considering 30 random features.
Out of bag error: 0.1123



Time taken to build model: 31.32 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        7066               98.6596 %
Incorrectly Classified Instances        96                1.3404 %
Kappa statistic                          0.9471
Mean absolute error                      0.0508
Root mean squared error                  0.1158
Relative absolute error                 19.5618 %
Root relative squared error             32.1396 %
Total Number of Instances             7162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.997     0.073      0.987     0.997     0.992      0.998    I
                 0.927     0.003      0.985     0.927     0.955      0.998    A
Weighted Avg.    0.987     0.062      0.987     0.987     0.986      0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 6048   16 |    a = I
   80 1018 |    b = A
```

sr_are_random_tree_feature_selection.model

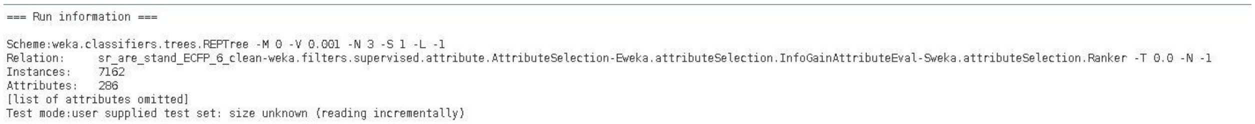


```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 50 -M 200.0 -S 5
Relation:     sr_are_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    7162
Attributes:   286
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)
```

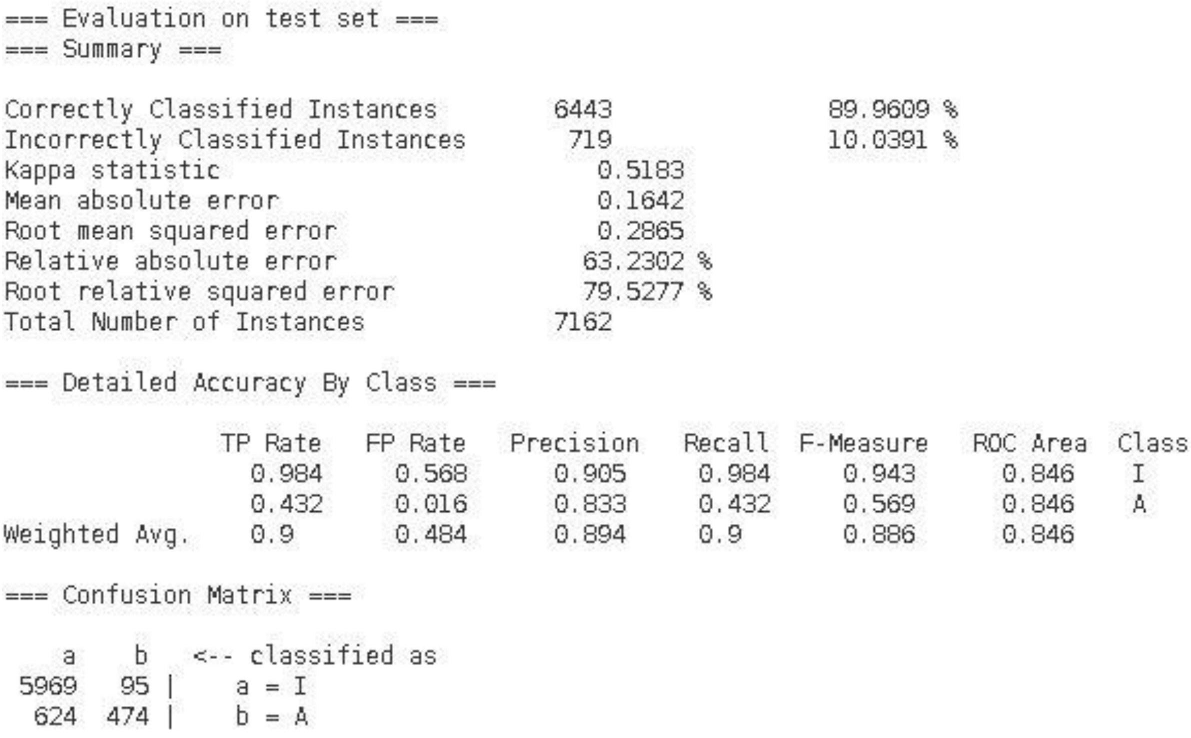


```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        6138               85.7023 %
Incorrectly Classified Instances      1024               14.2977 %
Kappa statistic                          0.2151
Mean absolute error                      0.2097
Root mean squared error                  0.3238
Relative absolute error                 80.7453 %
Root relative squared error             89.87   %
Total Number of Instances             7162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.982     0.831      0.867     0.982     0.921      0.819    I
                 0.169     0.018      0.624     0.169     0.266      0.819    A
Weighted Avg.    0.857     0.706      0.83      0.857     0.82       0.819

=== Confusion Matrix ===

    a    b   <-- classified as
 5952  112 |    a = I
  912  186 |    b = A
```

sr_are_rep_tree_feature_selection.model

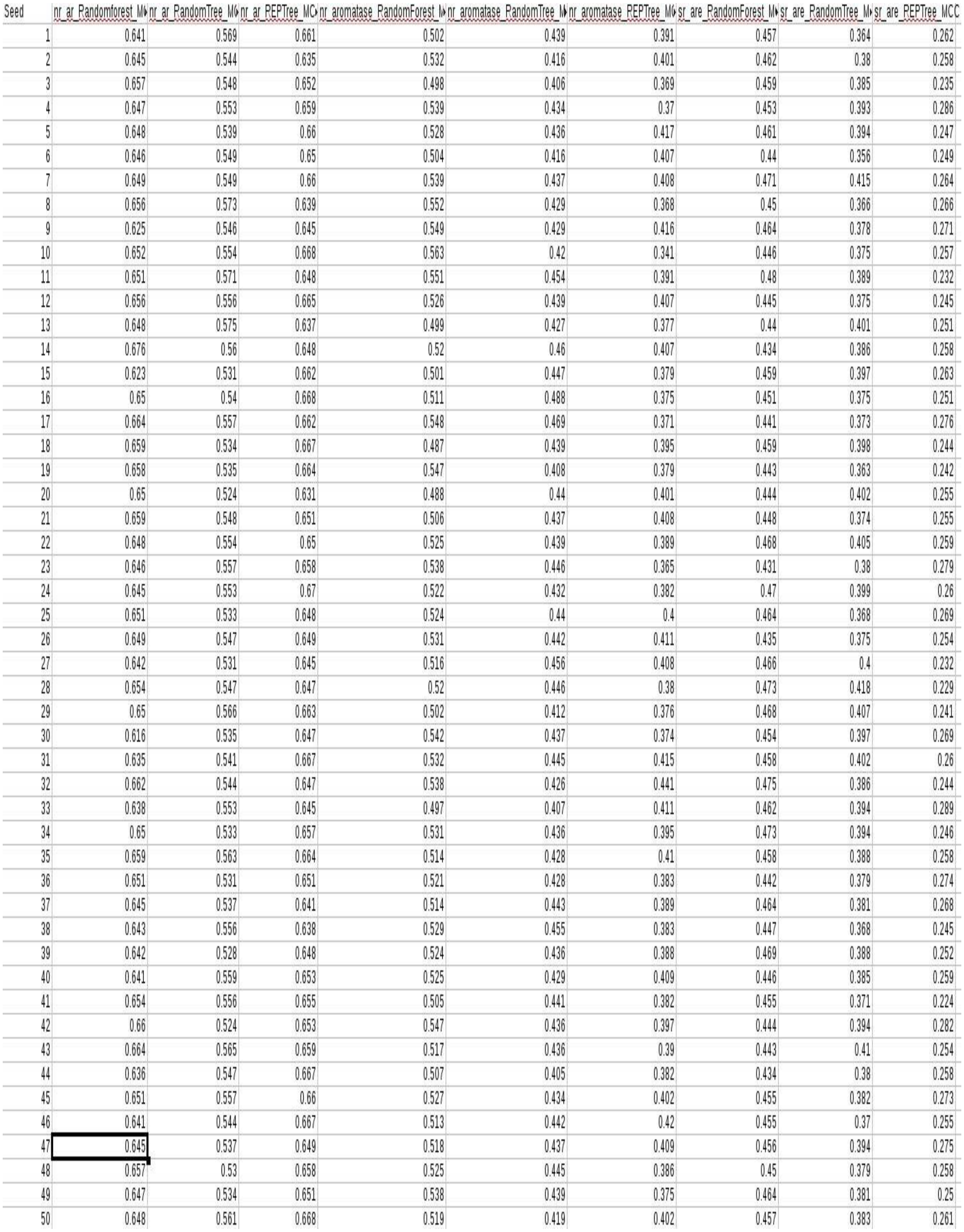


```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 0 -V 0.001 -N 3 -S 1 -L -1
Relation:     sr_are_stand_ECFP_6_clean-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T 0.0 -N -1
Instances:    7162
Attributes:   286
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)
```



```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        6443               89.9609 %
Incorrectly Classified Instances       719               10.0391 %
Kappa statistic                          0.5183
Mean absolute error                      0.1642
Root mean squared error                  0.2865
Relative absolute error                 63.2302 %
Root relative squared error             79.5277 %
Total Number of Instances             7162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.984     0.568      0.905     0.984     0.943      0.846    I
                 0.432     0.016      0.833     0.432     0.569      0.846    A
Weighted Avg.    0.9       0.484      0.894     0.9       0.886      0.846

=== Confusion Matrix ===

    a    b   <-- classified as
 5969   95 |    a = I
  624  474 |    b = A
```

9. APPENDIX C



Seed	nr_ar_Randomforest_M	nr_ar_RandomTree_M	nr_ar_PEPTree_MC	nr_aromatase_RandomForest_M	nr_aromatase_RandomTree_M	nr_aromatase_PEPTree_M	sr_are_RandomForest_M	sr_are_RandomTree_M	sr_are_PEPTree_MCC
1	0.641	0.569	0.661	0.502	0.439	0.391	0.457	0.364	0.262
2	0.645	0.544	0.635	0.532	0.416	0.401	0.462	0.38	0.258
3	0.657	0.548	0.652	0.498	0.406	0.369	0.459	0.385	0.235
4	0.647	0.553	0.659	0.539	0.434	0.37	0.453	0.393	0.286
5	0.648	0.539	0.66	0.528	0.436	0.417	0.461	0.394	0.247
6	0.646	0.549	0.65	0.504	0.416	0.407	0.44	0.356	0.249
7	0.649	0.549	0.66	0.539	0.437	0.408	0.471	0.415	0.264
8	0.656	0.573	0.639	0.552	0.429	0.368	0.45	0.366	0.266
9	0.625	0.546	0.645	0.549	0.429	0.416	0.464	0.378	0.271
10	0.652	0.554	0.668	0.563	0.42	0.341	0.446	0.375	0.257
11	0.651	0.571	0.648	0.551	0.454	0.391	0.48	0.389	0.232
12	0.656	0.556	0.665	0.526	0.439	0.407	0.445	0.375	0.245
13	0.648	0.575	0.637	0.499	0.427	0.377	0.44	0.401	0.251
14	0.676	0.56	0.648	0.52	0.46	0.407	0.434	0.386	0.258
15	0.623	0.531	0.662	0.501	0.447	0.379	0.459	0.397	0.263
16	0.65	0.54	0.668	0.511	0.488	0.375	0.451	0.375	0.251
17	0.664	0.557	0.662	0.548	0.469	0.371	0.441	0.373	0.276
18	0.659	0.534	0.667	0.487	0.439	0.395	0.459	0.398	0.244
19	0.658	0.535	0.664	0.547	0.408	0.379	0.443	0.363	0.242
20	0.65	0.524	0.631	0.488	0.44	0.401	0.444	0.402	0.255
21	0.659	0.548	0.651	0.506	0.437	0.408	0.448	0.374	0.255
22	0.648	0.554	0.65	0.525	0.439	0.389	0.468	0.405	0.259
23	0.646	0.557	0.658	0.538	0.446	0.365	0.431	0.38	0.279
24	0.645	0.553	0.67	0.522	0.432	0.382	0.47	0.399	0.26
25	0.651	0.533	0.648	0.524	0.44	0.4	0.464	0.368	0.269
26	0.649	0.547	0.649	0.531	0.442	0.411	0.435	0.375	0.254
27	0.642	0.531	0.645	0.516	0.456	0.408	0.466	0.4	0.232
28	0.654	0.547	0.647	0.52	0.446	0.38	0.473	0.418	0.229
29	0.65	0.566	0.663	0.502	0.412	0.376	0.468	0.407	0.241
30	0.616	0.535	0.647	0.542	0.437	0.374	0.454	0.397	0.269
31	0.635	0.541	0.667	0.532	0.445	0.415	0.458	0.402	0.26
32	0.662	0.544	0.647	0.538	0.426	0.441	0.475	0.386	0.244
33	0.638	0.553	0.645	0.497	0.407	0.411	0.462	0.394	0.289
34	0.65	0.533	0.657	0.531	0.436	0.395	0.473	0.394	0.246
35	0.659	0.563	0.664	0.514	0.428	0.41	0.458	0.388	0.258
36	0.651	0.531	0.651	0.521	0.428	0.383	0.442	0.379	0.274
37	0.645	0.537	0.641	0.514	0.443	0.389	0.464	0.381	0.268
38	0.643	0.556	0.638	0.529	0.455	0.383	0.447	0.368	0.245
39	0.642	0.528	0.648	0.524	0.436	0.388	0.469	0.388	0.252
40	0.641	0.559	0.653	0.525	0.429	0.409	0.446	0.385	0.259
41	0.654	0.556	0.655	0.505	0.441	0.382	0.455	0.371	0.224
42	0.66	0.524	0.653	0.547	0.436	0.397	0.444	0.394	0.282
43	0.664	0.565	0.659	0.517	0.436	0.39	0.443	0.41	0.254
44	0.636	0.547	0.667	0.507	0.405	0.382	0.434	0.38	0.258
45	0.651	0.557	0.66	0.527	0.434	0.402	0.455	0.382	0.273
46	0.641	0.544	0.667	0.513	0.442	0.42	0.455	0.37	0.255
47	0.645	0.537	0.649	0.518	0.437	0.409	0.456	0.394	0.275
48	0.657	0.53	0.658	0.525	0.445	0.386	0.45	0.379	0.258
49	0.647	0.534	0.651	0.538	0.439	0.375	0.464	0.381	0.25
50	0.648	0.561	0.668	0.519	0.419	0.402	0.457	0.383	0.261