Artificial Intelligence ENCS3340

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Homework (1)

Prepared By:

Student Name and ID: Ibrahim Nobani 1190278

Student Name and ID: Mahmoud Nobani 1180729

Instructor: Adnan Yahya

Section: 2

Date: 11/6/2022

## Abstract:

The aim for this project is study how various machine learning techniques can affect the level of accurate results we can get, for experimenting we will use weka IDE.

## Data Set Information:

The data set we are using obtained with CVS, studied various grown in turkey **Kecimen and Besni raisin Images**, the number of raisin grains used is **900**, **450 pieces from both varieties.** These images were subjected to various stages of pre-processing and 7 morphological features were extracted. These features have been classified using three different artificial intelligence techniques.

## Attribute Information:

1.) **Area:** Grants the number of pixels within the raisin's boundaries.
2.) **Perimeter:** Gives the environment measurements, by computing the distance between the raisin's boundaries and the pixels around it.
3.) **MajorAxisLength:** Grants main axis length, this is the longest line on the raisin that can be drawn.
4.) **MinorAxisLength**: Grants small axis length, this is the shortest line on the raisin that can be drawn.
5.) **Eccentricity:** Grants the eccentricity of the ellipse measure, which has the same moments as raisins.
6.) **ConvexArea:** Grants the number of pixels of the smallest convex shell of the region formed by the raisin.
7.) **Extent:** Gives the region ratio formed by the raisin to the total pixels in the bounding box.
8.) **Class:** Kecimen and Besni raisin.

For this project we are preforming different machine learning techniques on the given dataset, in hope of introducing a machine capable of differentiating the two types of raisin given above, supervised learning is the type of feedback used, which has the correct answer within the dataset we will train.

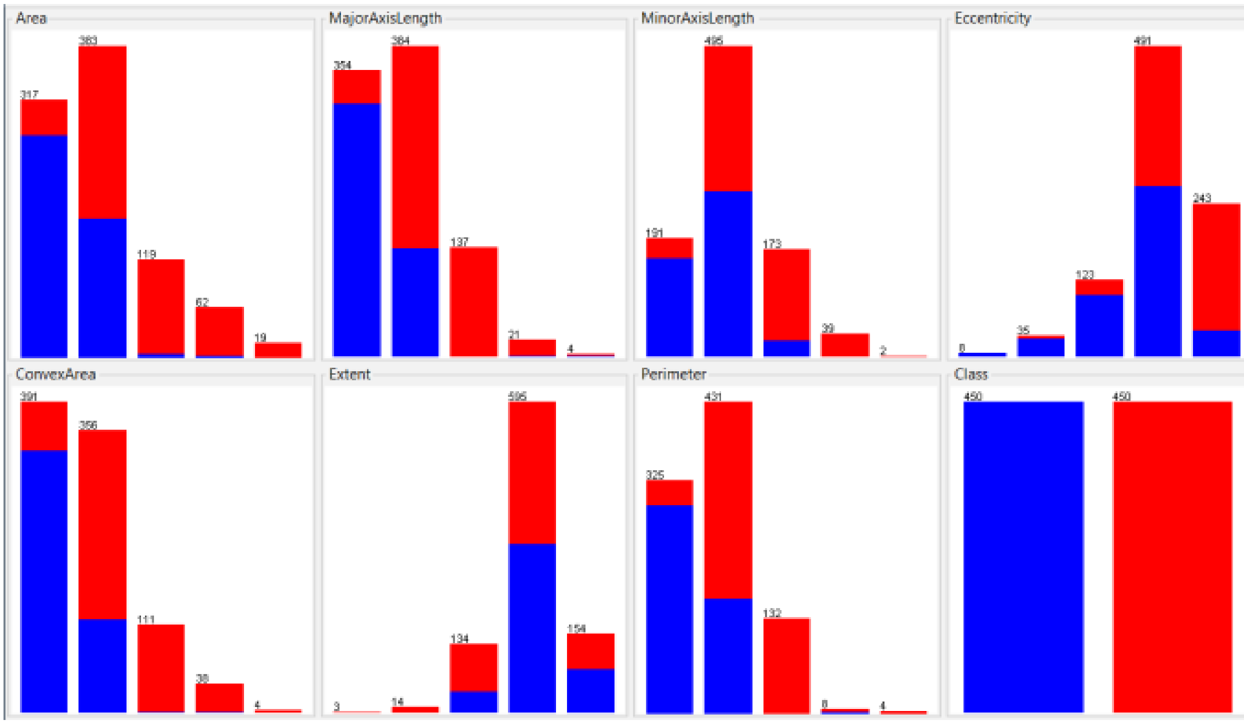
Three machine learning models were used:

- **Decision Tree:** is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
- **Naïve bayes:** is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems, it is mainly used in *text classification*
- **Random Forest:** based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

To evaluate the results obtained, we calculated different type of parameters, the parameters we have are:

1- **Accuracy**: percentage of correct predictions made by our classification model
2- **Inaccuracy (Error):** percentage of incorrect predictions made by our classification model
3- **True Positive Rate (TP rate):** percentage of the actual positive from the predicted values (also positive)
4- **False Positive Rate (FP rate):** percentage of the actual negative from the predicted values (positive)
5- **Precision:** out of all positive predictions, how many are actually positive
6- **Recall:** indicates out of all actually positive values, how many are predicted positive
7- **F-Measure**: the harmonic means of precision and recall, which allows the model to be evaluated taking both the precision and recall into account
8- **Confusion Matrix**: is a technique for summarizing the performance of a classification algorithm/ model.

We took **“area”** as an attribute and aimed to study it using machine learning, we started **discretization of all attribute to 5** bins as seen in the figure that shows the visualization of all attributes including area attribute (Top Left):




We started with the first which is the **decision tree**, we applied it to the **area attribute, for the test 5-fold cross validation were used**, the following results were generated:

| | | |
|---|---|---|
| Correctly Classified Instances | 777 | 86.3333 % |
| Incorrectly Classified Instances | 123 | 13.6667 % |

| | TP rate | FP rate | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| (-inf-67319] | 0.975 | 0.05 | 0.914 | 0.975 | 0.944 |
| (67319-109251] | 0.919 | 0.093 | 0.880 | 0.919 | 0.899 |
| (109251-151183] | 0.580 | 0.026 | 0.775 | 0.580 | 0.663 |
| (151183-193115] | 0.645 | 0.024 | 0.667 | 0.645 | 0.656 |
| (193115-inf) | 0.368 | 0.007 | 0.538 | 0.368 | 0.438 |
| total | 0.863 | 0.062 | 0.856 | 0.863 | 0.857 |

**Confusion matrix:**

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a = '(-inf-67319]' | 309 | 8 | 0 | 0 | 0 |
| b = '(67319-109251]' | 29 | 352 | 2 | 0 | 0 |
| c = '(109251-151183]' | 0 | 40 | 69 | 9 | 1 |
| d = '(151183-193115]' | 0 | 0 | 17 | 40 | 5 |
| e = '(193115-inf)' | 0 | 0 | 1 | 11 | 7 |

For the snapshoot refer to Decision tree, with no change in hyper parameters

A change to the **hyper parameters** was applied to the model, specifically we removed **pruning**, the output changed to the following:

| | | |
|---|---|---|
| Correctly Classified Instances | 768 | 85.3333 % |
| Incorrectly Classified Instances | 132 | 14.6667 % |

| | TP rate | FP rate | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| (-inf-67319] | 0.968 | 0.053 | 0.908 | 0.968 | 0.937 |
| (67319-109251] | 0.854 | 0.060 | 0.913 | 0.854 | 0.883 |
| (109251-151183] | 0.706 | 0.056 | 0.656 | 0.706 | 0.680 |
| (151183-193115] | 0.597 | 0.021 | 0.673 | 0.597 | 0.632 |
| (193115-inf) | 0.684 | 0.009 | 0.619 | 0.684 | 0.650 |
| total | 0.853 | 0.053 | 0.855 | 0.853 | 0.853 |

**Confusion Matrix:**

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a = '(-inf-67319]' | 307 | 8 | 0 | 0 | 0 |
| b = '(67319-109251]' | 31 | 327 | 25 | 0 | 0 |
| c = '(109251-151183]' | 0 | 21 | 84 | 13 | 1 |
| d = '(151183-193115]' | 0 | 0 | 18 | 37 | 7 |
| e = '(193115-inf)' | 0 | 0 | 1 | 5 | 13 |

For the snapshoot refer to Decision tree with change to the hyper parameters

**Naïve Bayes:**

| | | |
|---|---|---|
| Correctly Classified Instances | 766 | 85.1111 % |
| Incorrectly Classified Instances | 134 | 14.8889 % |

| | TP rate | FP rate | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| (-inf-67319] | 0.956 | 0.069 | 0.883 | 0.956 | 0.918 |
| (67319-109251] | 0.872 | 0.085 | 0.884 | 0.872 | 0.878 |
| (109251-151183] | 0.647 | 0.032 | 0.755 | 0.647 | 0.697 |
| (151183-193115] | 0.694 | 0.026 | 0.662 | 0.694 | 0.677 |
| (193115-inf) | 0.474 | 0.003 | 0.750 | 0.474 | 0.581 |
| total | 0.851 | 0.066 | 0.848 | 0.851 | 0.848 |

**Confusion Matrix:**

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a = '(-inf-67319]' | 303 | 14 | 0 | 0 | 0 |
| b = '(67319-109251]' | 40 | 334 | 9 | 0 | 0 |
| c = '(109251-151183]' | 0 | 30 | 77 | 12 | 0 |
| d = '(151183-193115]' | 0 | 0 | 16 | 43 | 3 |
| e = '(193115-inf)' | 0 | 0 | 0 | 10 | 9 |

Many changes were applied to **the hyper parameters**, **none** of them affected the **result**.

For the snapshoot refer to Naïve bayes

**Random Forest:**

Random forest, 5-fold, 5 discrete, no hyper

| | | |
|---|---|---|
| Correctly Classified Instances | 774 | 86 % |
| Incorrectly Classified Instances | 126 | 14 % |

| | TP rate | FP rate | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| (-inf-67319] | 0.968 | 0.048 | 0.916 | 0.942 | 0.942 |
| (67319-109251] | 0.584 | 0.050 | 0.926 | 0.889 | 0.889 |
| (109251-151183] | 0.731 | 0.058 | 0.659 | 0.693 | 0.693 |
| (151183-193115] | 0.629 | 0.025 | 0.650 | 0.629 | 0.639 |
| (193115-inf) | 0.737 | 0.007 | 0.700 | 0.737 | 0.718 |
| total | 0.860 | 0.048 | 0.864 | 0.860 | 0.809 |

**Confusion Matrix**:

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a = '(-inf-67319]' | 307 | 10 | 0 | 0 | 0 |
| b = '(67319-109251]' | 28 | 327 | 28 | 0 | 0 |
| c = '(109251-151183]' | 0 | 16 | 87 | 16 | 0 |
| d = '(151183-193115]' | 0 | 0 | 17 | 39 | 6 |
| e = '(193115-inf)' | 0 | 0 | 0 | 5 | 14 |

For the snapshoot refer to Random forest without a change to hyper parameters

After applying a **change** to the **max depth parameter**, the following result was obtained:

Random forest, 5-fold, 5 discrete, max depth 1

Correctly Classified Instances 741 82.3333 %

Incorrectly Classified Instances 159 17.6667 %

| | TP rate | FP rate | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| (-inf-67319] | 0.968 | 0.062 | 0.895 | 0.968 | 0.930 |
| (67319-109251] | 0.906 | 0.128 | 0.840 | 0.906 | 0.872 |
| (109251-151183] | 0.521 | 0.050 | 0.614 | 0.521 | 0.564 |
| (151183-193115] | 0.339 | 0.018 | 0.583 | 0.339 | 0.429 |
| (193115-inf) | 0.211 | 0.003 | 0.571 | 0.211 | 0.308 |
| total | 0.823 | 0.084 | 0.806 | 0.823 | 0.809 |

**Confusion Matrix:**

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a = '(-inf-67319]' | 307 | 10 | 0 | 0 | 0 |
| b = '(67319-109251]' | 36 | 347 | 0 | 0 | 0 |
| c = '(109251-151183]' | 0 | 56 | 62 | 0 | 1 |
| d = '(151183-193115]' | 0 | 0 | 39 | 21 | 2 |
| e = '(193115-inf)' | 0 | 0 | 0 | 51 | 4 |

For the snapshoot refer to Random forest with a change to the hyper parameters

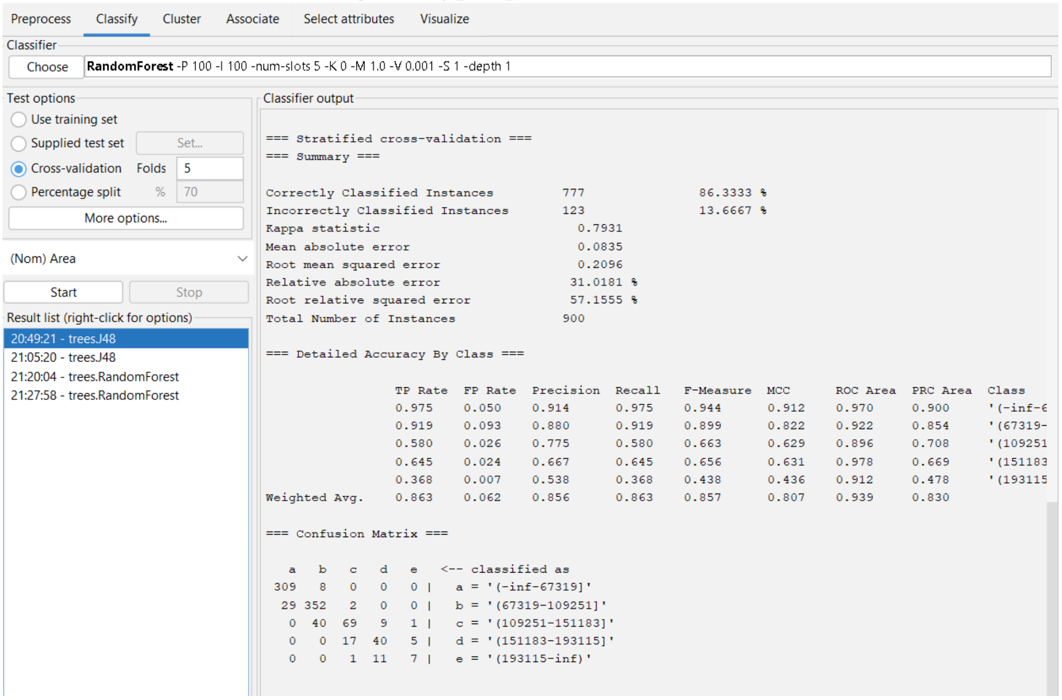
We then compare the results of each model giving us the following result:

1) Decision tree (no change on hyper parameters): 86.33%
2) Random forest (no change on hyper parameters): 86%
3) Decision tree (change on hyper parameters): 85.33%
4) Naïve bayes: 85.11%
5) Random forest (change on hyper parameters): 82.33%

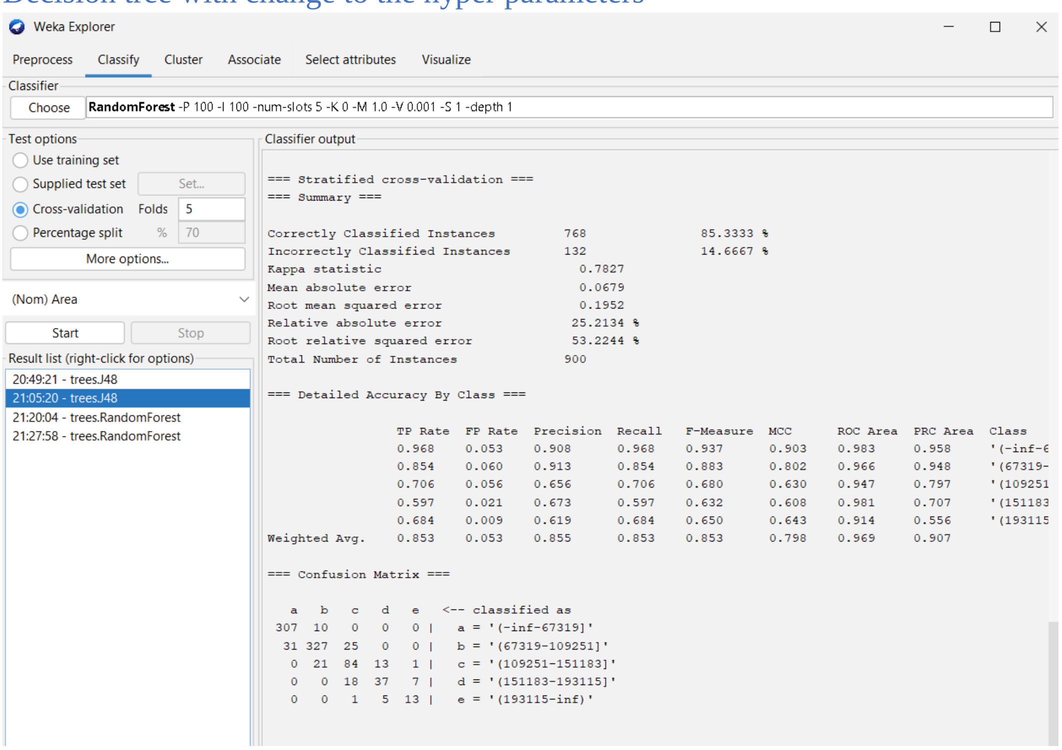
Thus, we can conclude that decision tree is the best model.

# Snapshoots from the tool:

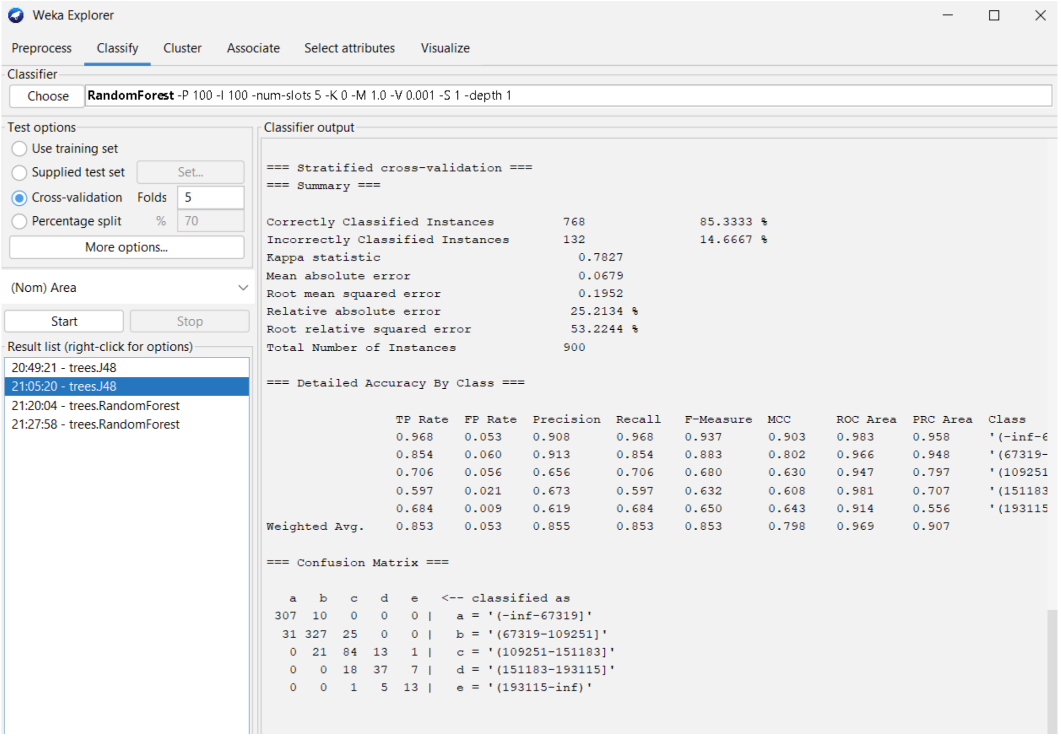
## Decision tree, with no change in hyper parameters



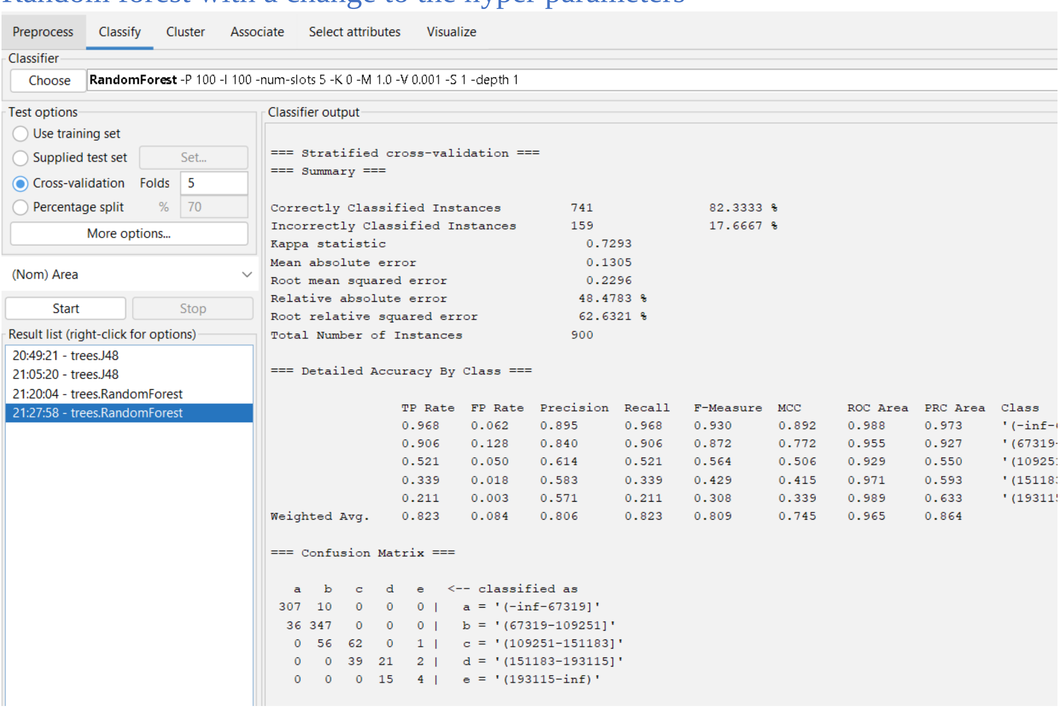

## Decision tree with change to the hyper parameters



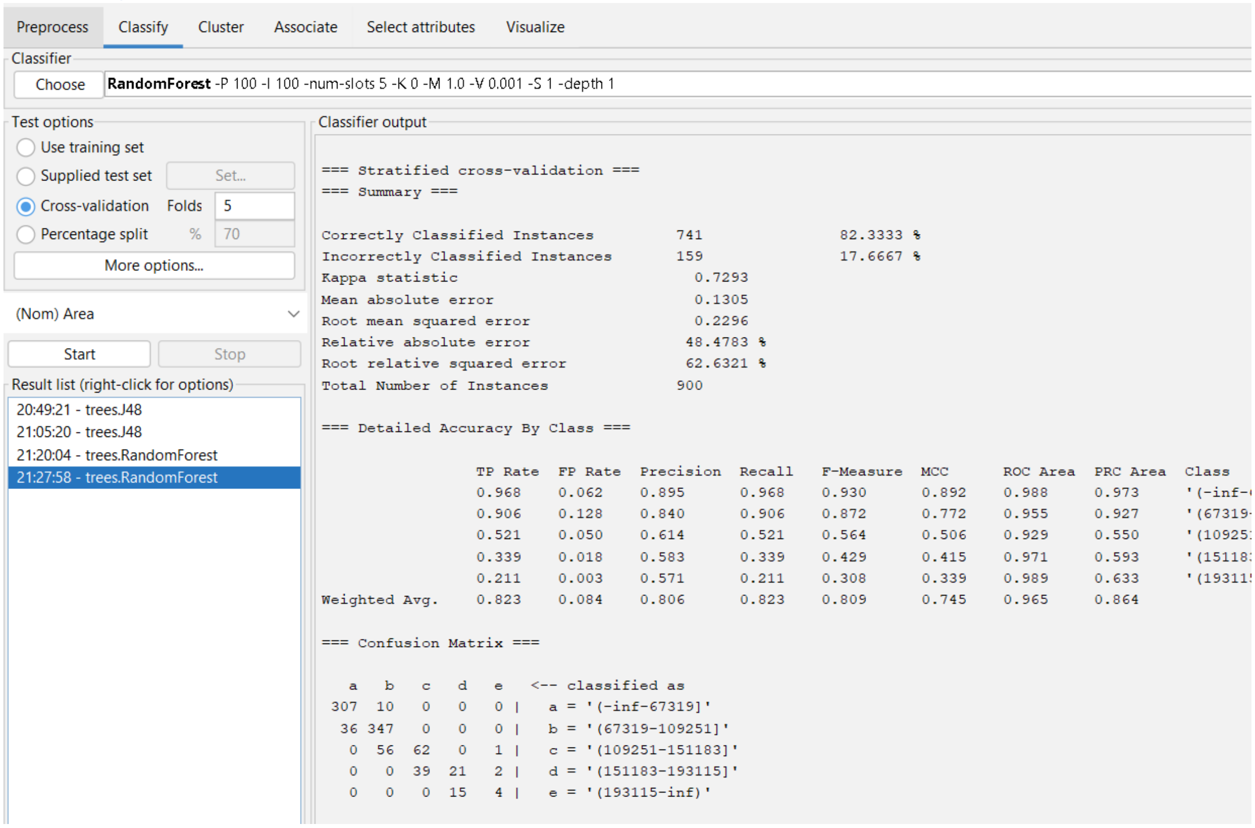
# Random forest without a change to hyper parameters




# Random forest with a change to the hyper parameters



# Naïve bayes



Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 5 -K 0 -M 1.0 -V 0.001 -S 1 -depth 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
5
Percentage split
%
70
More options...
(Nom) Area
Start
Stop
Result list (right-click for options)
20:49:21 - trees.J48
21:05:20 - trees.J48
21:20:04 - trees.RandomForest
21:27:58 - trees.RandomForest
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 741 82.3333 %
Incorrectly Classified Instances 159 17.6667 %
Kappa statistic 0.7293
Mean absolute error 0.1305
Root mean squared error 0.2296
Relative absolute error 48.4783 %
Root relative squared error 62.6321 %
Total Number of Instances 900
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.968 0.062 0.895 0.968 0.930 0.892 0.988 0.973 '(-inf-
0.906 0.128 0.840 0.906 0.872 0.772 0.955 0.927 '(67319-
0.521 0.050 0.614 0.521 0.564 0.506 0.929 0.550 '(10925
0.339 0.018 0.583 0.339 0.429 0.415 0.971 0.593 '(15118
0.211 0.003 0.571 0.211 0.308 0.339 0.989 0.633 '(19311
Weighted Avg. 0.823 0.084 0.806 0.823 0.809 0.745 0.965 0.864
=== Confusion Matrix ===
a b c d e <-- classified as
307 10 0 0 0 | a = '(-inf-67319]'
36 347 0 0 0 | b = '(67319-109251]'
0 56 62 0 1 | c = '(109251-151183]'
0 0 39 21 2 | d = '(151183-193115]'
0 0 0 15 4 | e = '(193115-inf)'