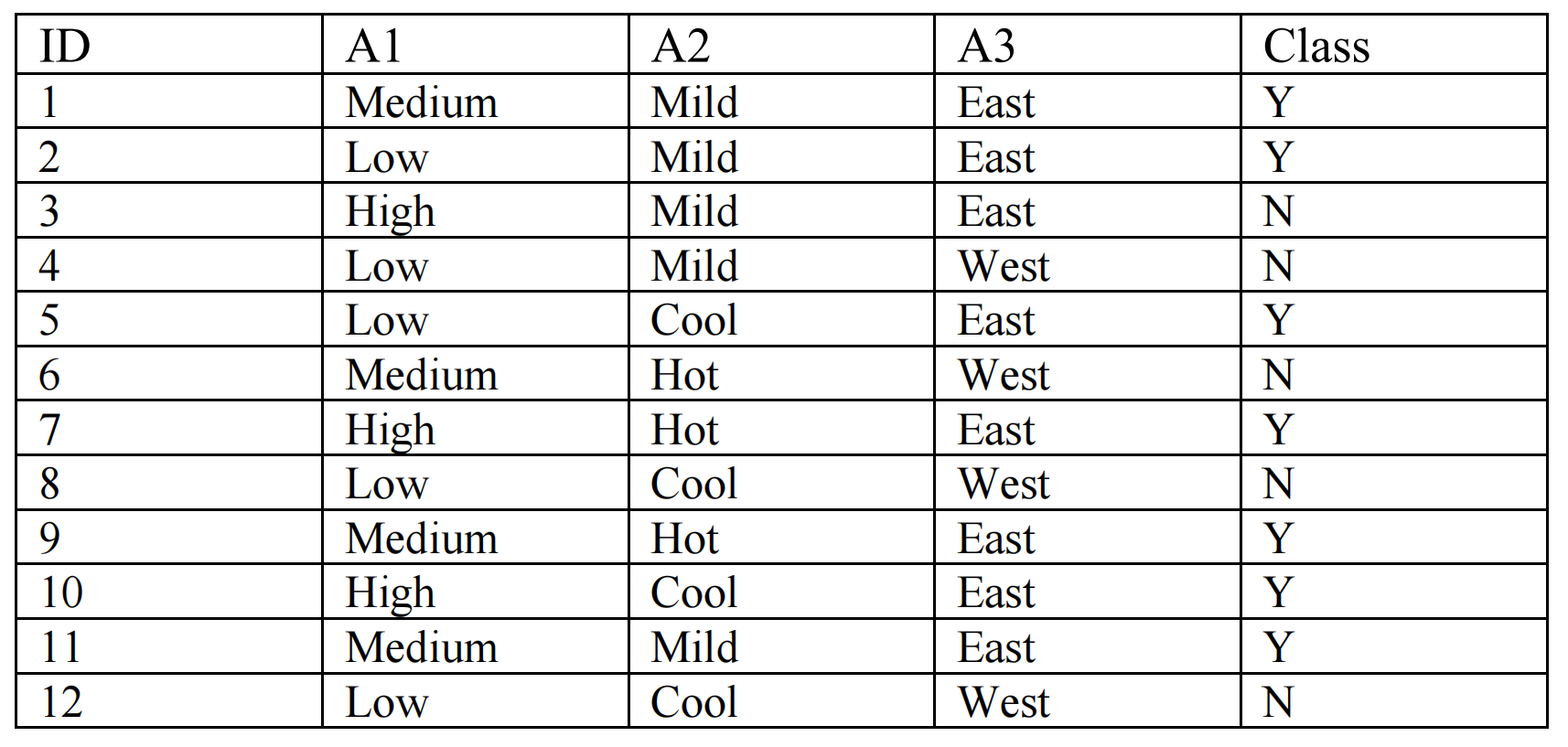
Yiduo Feng

CS 699

Assignment 3

Date: 5/27/2022

**Problem 1 (20 points)**. This problem is about the decision tree algorithm we discussed.Calculate the information gain of A2 and A3 and determine which is better as the test attribute at the root. You must show all calculations, including the calculation of *info* and *information gain*.



 There are 12 tuples in D

 7 Yes’s and 5 No’s

Info(D) = I(7,5) = -7/12log2(7/12) - 5/12log2(5/12) = 0.97986

* In A2

4 Cool: 2yes 2no -> I(2,2) = -2/4log2(2/4) - 2/4log2(2/4) = 1

5 Mild: 3yes 2no -> I(3,2) = -3/5log2(3/5) - 2/5log2(2/5) = 0.97095

3 Hot: 2yes 1no -> I(2,1) = -2/3log2(2/3) - 1/3log2(1/3) = 0.9182958

Info\_A2(D) = 4/12\*I(2,2)+5/12\*I(3,2)+3/12\*I(2,1) = 0.96747

Gain(A2) = Info(D) - Info\_A2(D) = 0.97986 - 0.96747 = 0.01239

* In A3

8 East: 7yes 1no -> I(7,1) = -7/8log2(7/8) - 1/8log2(1/8) = 0.54356

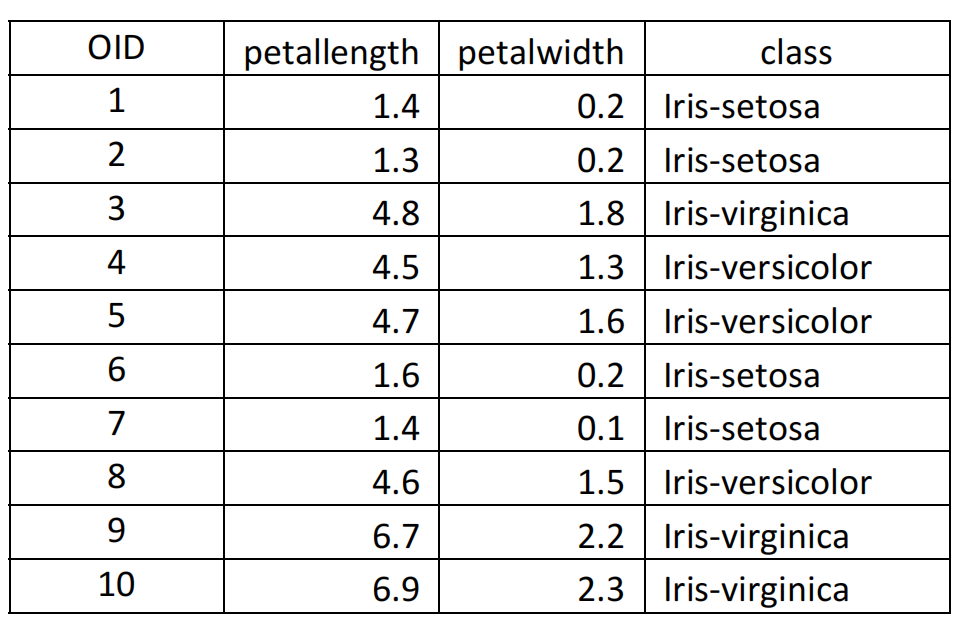
4 West: 0yes 4no -> I(0,4) = - log2(1) = 0

Info\_A3(D) = 8/12\*I(7,1)+4/12\*I(0,4) = 8/12 \* 0.54356 = 0.36237

Gain(A3) = Info(D) - Info\_A3(D) = 0.97986 - 1.14553 = 0.97986 - 0.36237 =0.61749

A3 is better as test attribute at the root because it has higher information gain.

**Problem 2 (20 points)**. Consider the following dataset, which is a part of the *iris* dataset:



Suppose you want to classify an unseen object X: <petallength = 4.2, petalwidth = 1.3> using the KNN method we discussed in the class.

1. . Calculate the distance between X and all 10 objects. Use the Euclidean distance.

1: sqrt((4.2-1.4)^2+(1.3-0.2)^2) = 3.008321791

2: sqrt((4.2-1.3)^2+(1.3-0.2)^2) = 3.101612484

3: sqrt((4.2-4.8)^2+(1.3-1.8)^2) = 0.781024968

4: sqrt((4.2-4.5)^2+(1.3-1.3)^2) = 0.3

5: sqrt((4.2-4.7)^2+(1.3-1.6)^2) = 0.583095189

6: sqrt((4.2-1.6)^2+(1.3-0.2)^2) = 2.823118843

7: sqrt((4.2-1.4)^2+(1.3-0.1)^2) = 3.046309242

8: sqrt((4.2-4.6)^2+(1.3-1.5)^2) = 0.447213595

9: sqrt((4.2-6.7)^2+(1.3-2.2)^2) = 2.657066051

10: sqrt((4.2-6.9)^2+(1.3-2.3)^2) = 2.87923601

1. . Classify X using five nearest neighbors.

According to the results in (1), 3,4,5,8,9 are nearest neighbors, and there are 3 Iris‐versicolor and 2 Iris‐virginica, so X is classified as Iris‐versicolor.

**Problem 3 (20 points)**. This problem is about the logistic regression we discussed in the class. Consider a dataset that has two independent variables A1 and A2 and a class

attribute, which takes on either *yes* or *no*. Suppose you ran a logistic regression algorithm on the dataset and obtained the following coefficients for class *yes*:

Coefficient of A1 = 0.045

Coefficient of A2 = 0.003

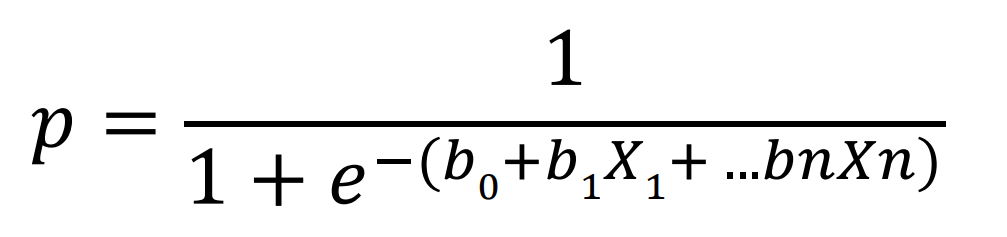
Intercept = -3.485

Classify the following two unseen objects using the above model:

O1: <A1 = 47, A2 = 213>

O2: <A1 = 65, A2 = 276>

Assume that the classification threshold is 0.5. You must not use any software except for calculation and you must show all calculations.



Q1:

P(class=yes|A1=47|A2=213) = 1/(1+e^(-(-3.485 + 0.045\*47+0.003\*213))) = 0.32497

Threshold is 0.5, so Q1 is classified as no.

O2:

P(class=yes|A1=65|A2=276) = 1/(1+e^(-(-3.485 + 0.045\*65+0.003\*276))) = 0.56660

Threshold is 0.5, so Q2 is classified as yes.

**Problem 4 (20 points).** Study *discriminant analysis* classification method, and

1. . Write a brief, one-page description of the method.

Linear discriminant analysis is a supervised learning, so it is based on label analysis. Its most prominent feature is that it can find a linear transformation method to reduce the dimension to achieve the effect of more concise classification.

It can be used for dimensionality reduction or classification, and the representation of LDA is very simple. It contains statistical properties of the data, computed for each class. For a single input variable, this includes: the mean for each category. Variance calculated across all categories. Predictions are made by computing the discriminant value for each class and predicting the class with the largest value. This technique assumes that the data has a Gaussian distribution, so it is best to remove outliers from the data beforehand. It is a simple and powerful method for classifying predictive modeling problems.

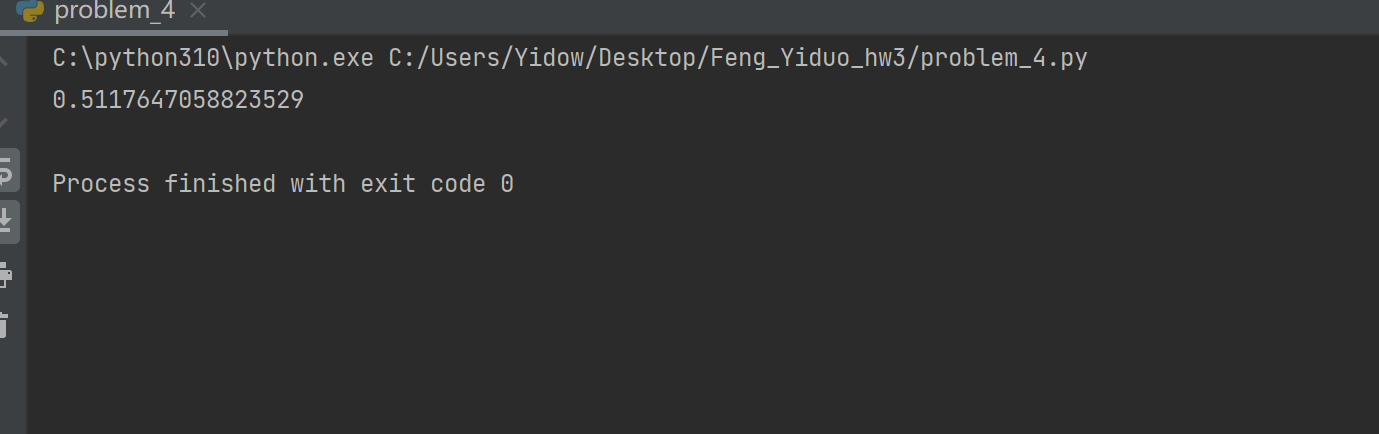
(2). Run a *linear discriminant analysis* method on Accidents1000 dataset:

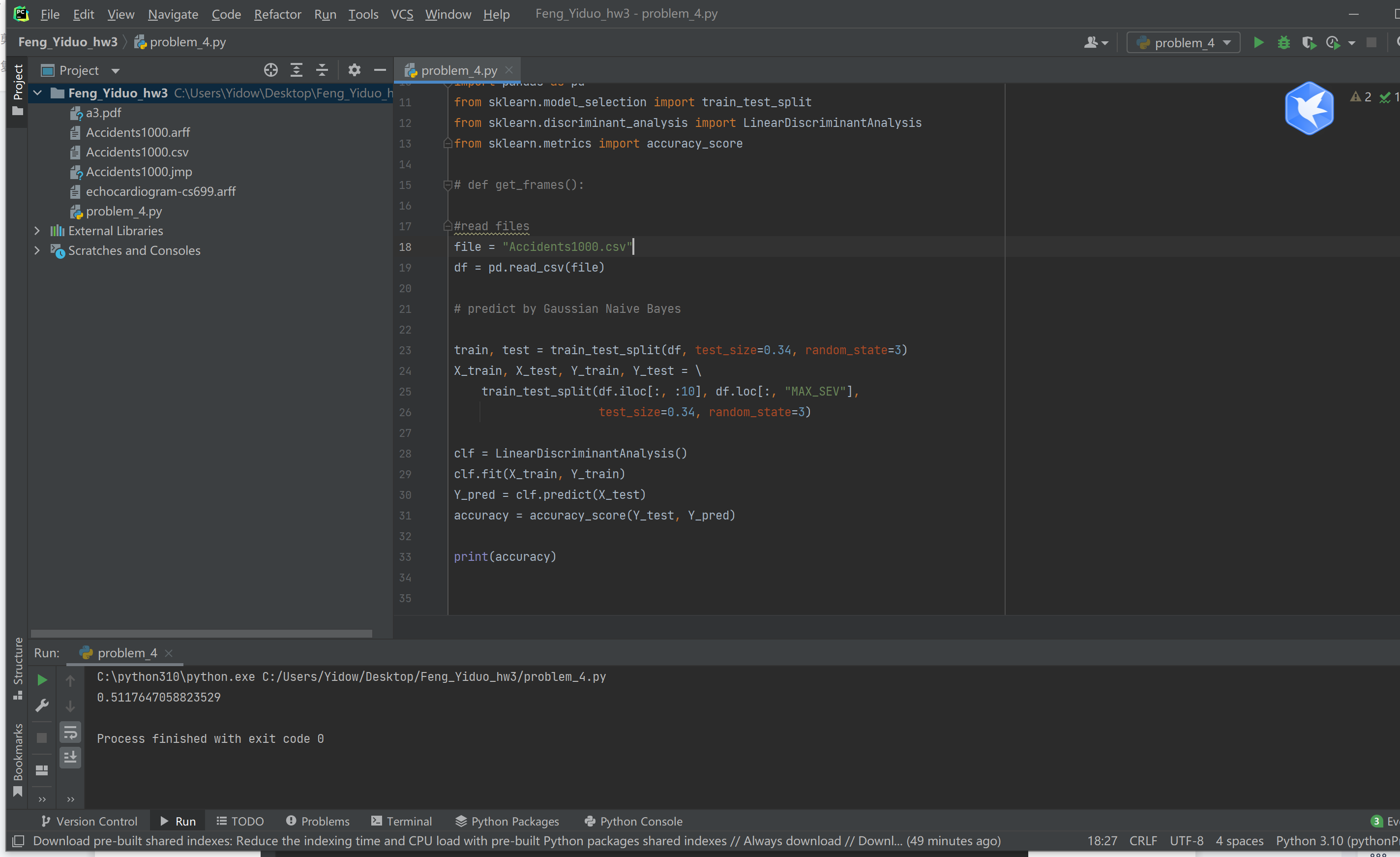
 Weka: Use *Accidents1000.arff*. Choose *Percentage split* and set 66%. Submit the screenshot of output window.

 JMP Pro: use *Accidents1000.jmp*. Validation column is already created. Submit the screenshot of output window.

 Other tools: Use *Accidents1000.csv*. Split the dataset with 66-34 ratio. Submit the script file and the screenshot of your output.

Include the prediction accuracy on the test (or validation) dataset in your submission.





**Problem 5 (20 points).** Using Weka**,** run six classifier algorithms on the *echocardiogram-cs699.arff* dataset, which was downloaded from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/echocardiogram) and modified for this assignment.

For each classifier algorithm, perform:

(1) Run the classifier algorithm alone

(2) Run Bagging with the classifier

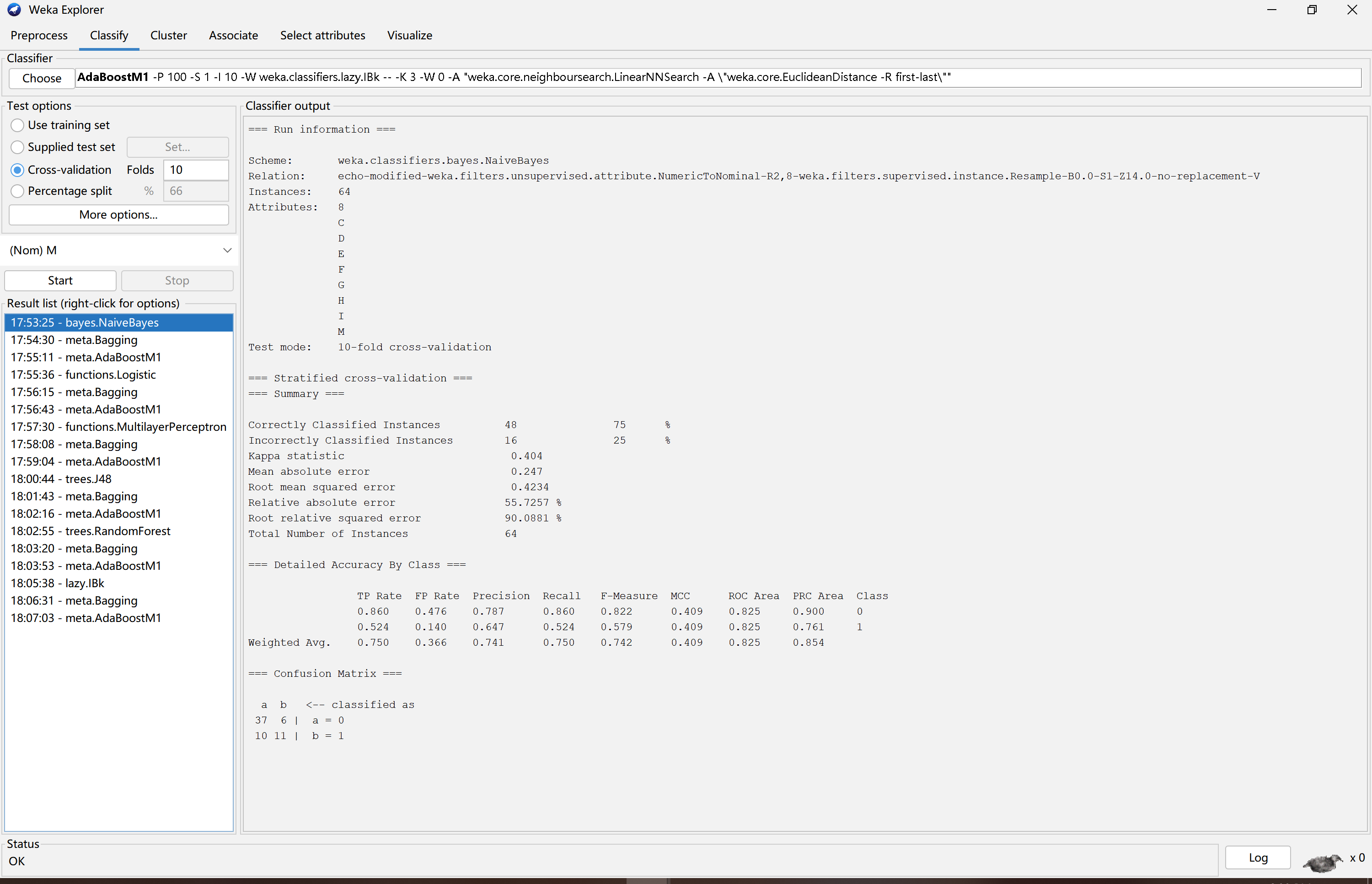
(3) Run AdaBoostM1 with the classifier

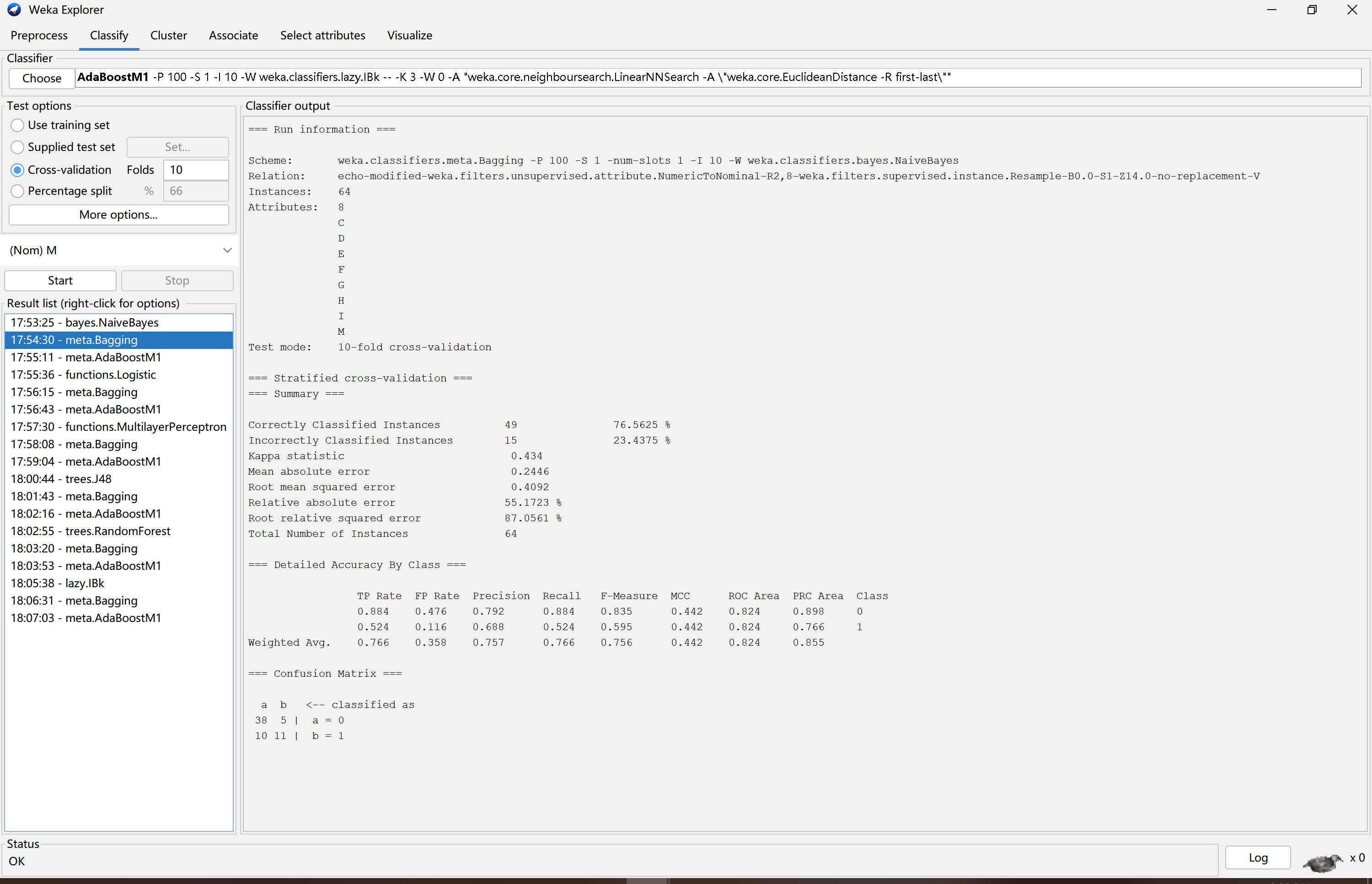
(1). Collect all accuracies (% correctly classified instances) and enter these accuracies in the following table:

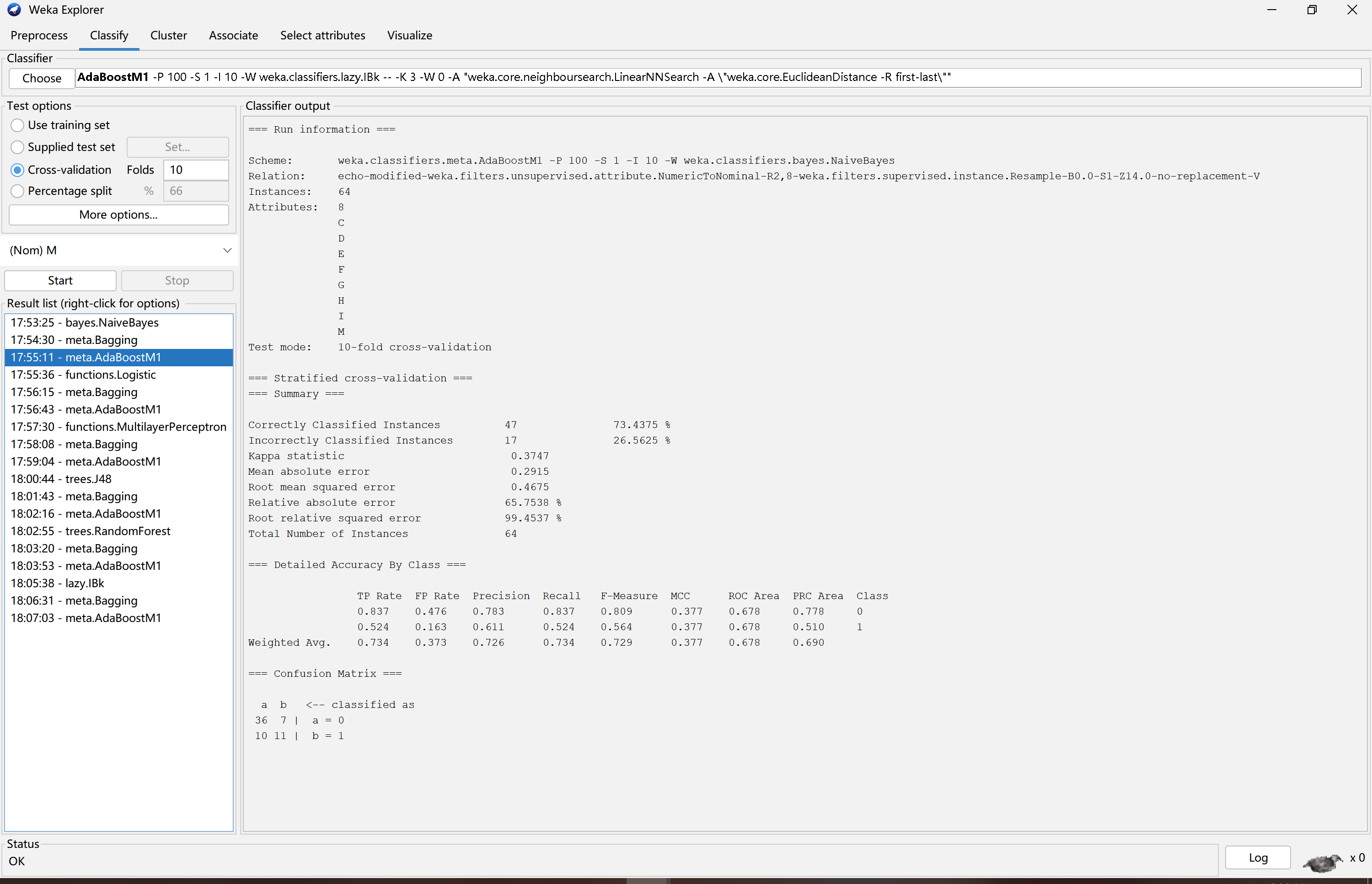
|  |  |  |  |
| --- | --- | --- | --- |
|  | Classifier alone | Bagging with classifier | AdaBoostM1 with classifier |
| Naïve Bayes | 75% | 76.5625% | 73.4375 % |
| Logistic | 75% | 73.4375 % | 75% |
| MultilayerPerceptron | 81.25% | 85.9375 % | 76.5625 % |
| J48 | 71.875% | 76.5625 % | 78.125 % |
| RandomForest | 76.5625 % | 79.6875 % | 78.125 % |
| IBk (with k = 3) | 79.6875 % | 75% | 76.5625 % |

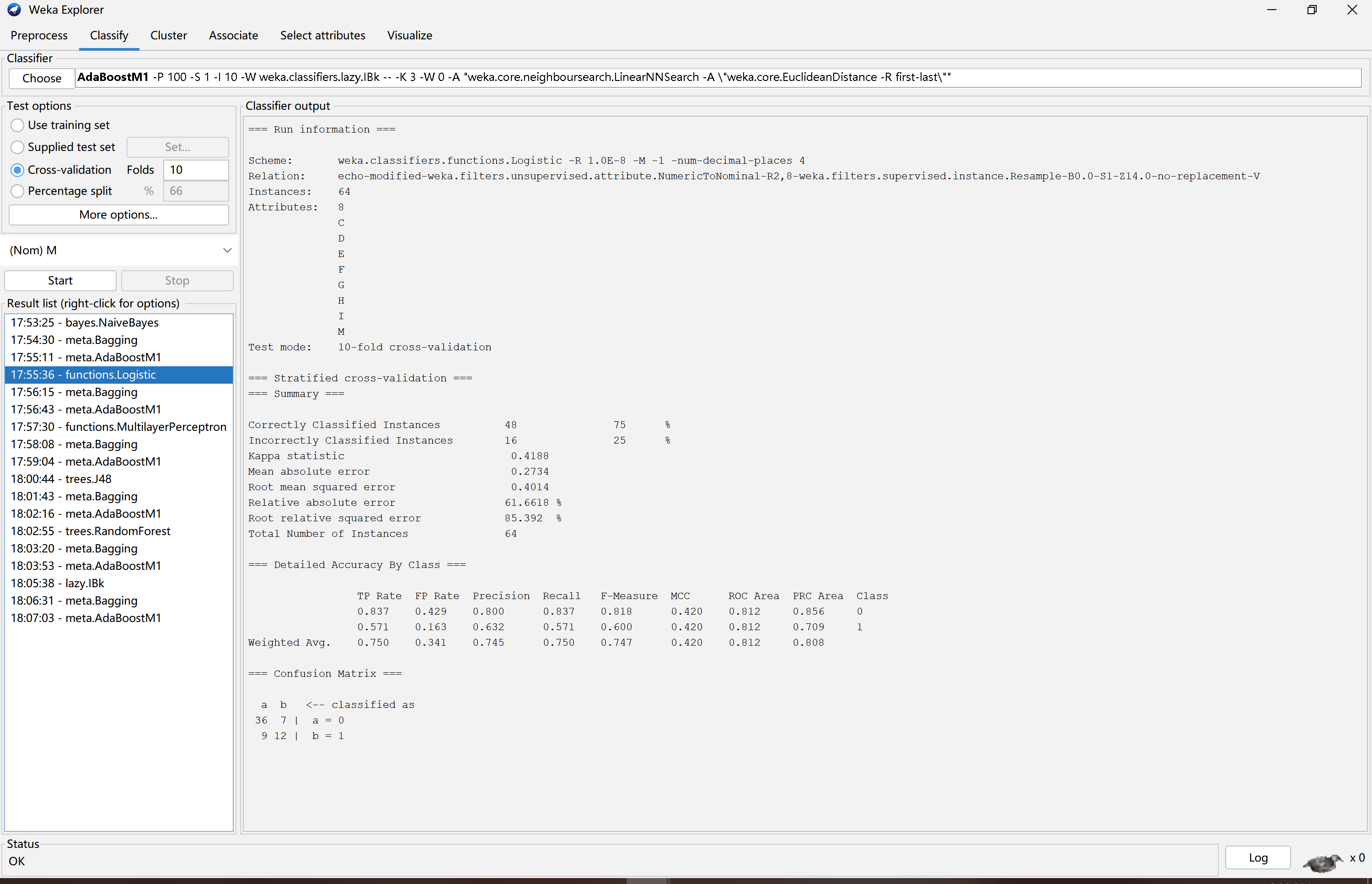
(2). Include in your submission screenshots of all Weka’s classification output, which

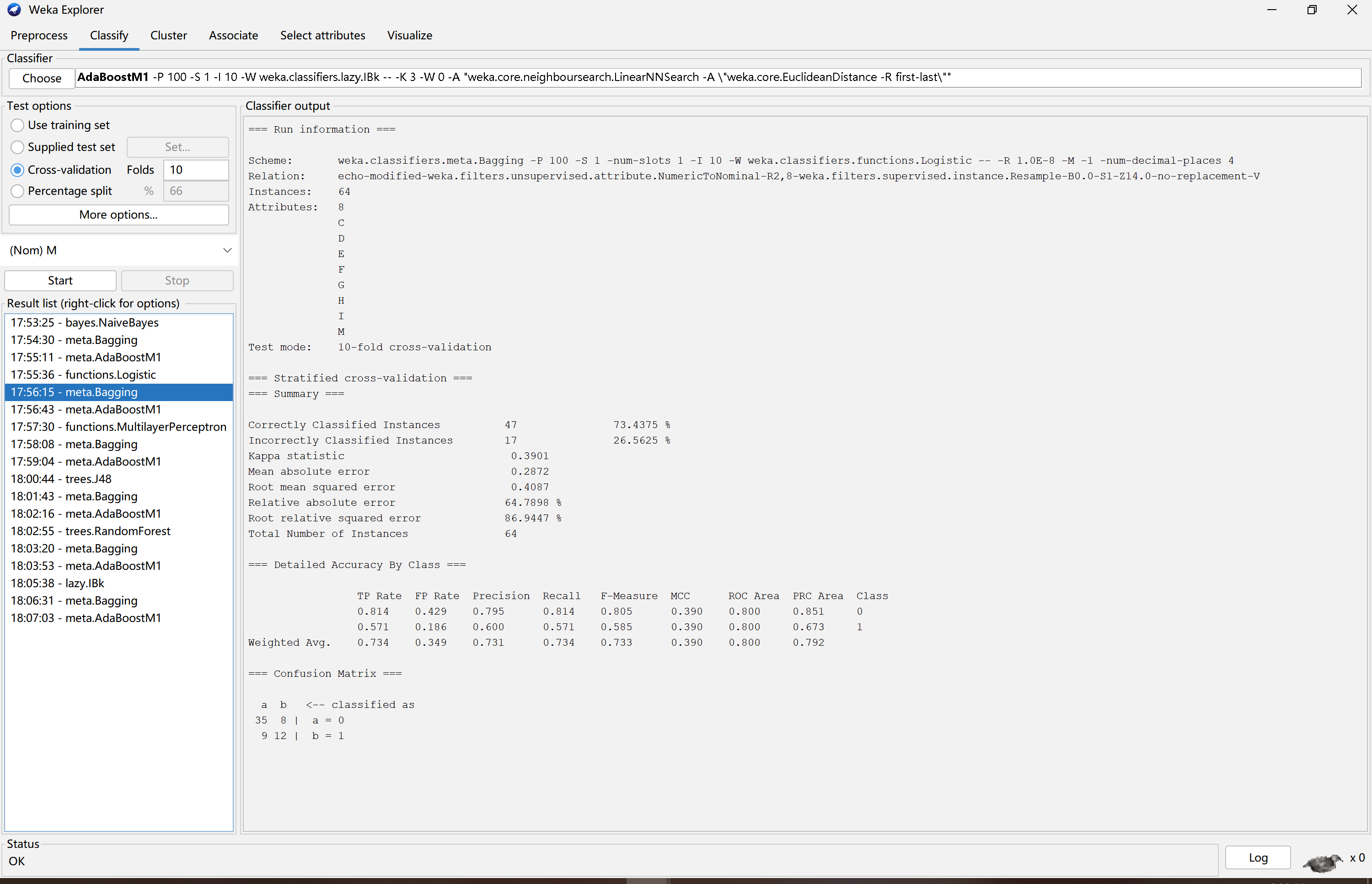
includes all performance measures and a confusion matrix.

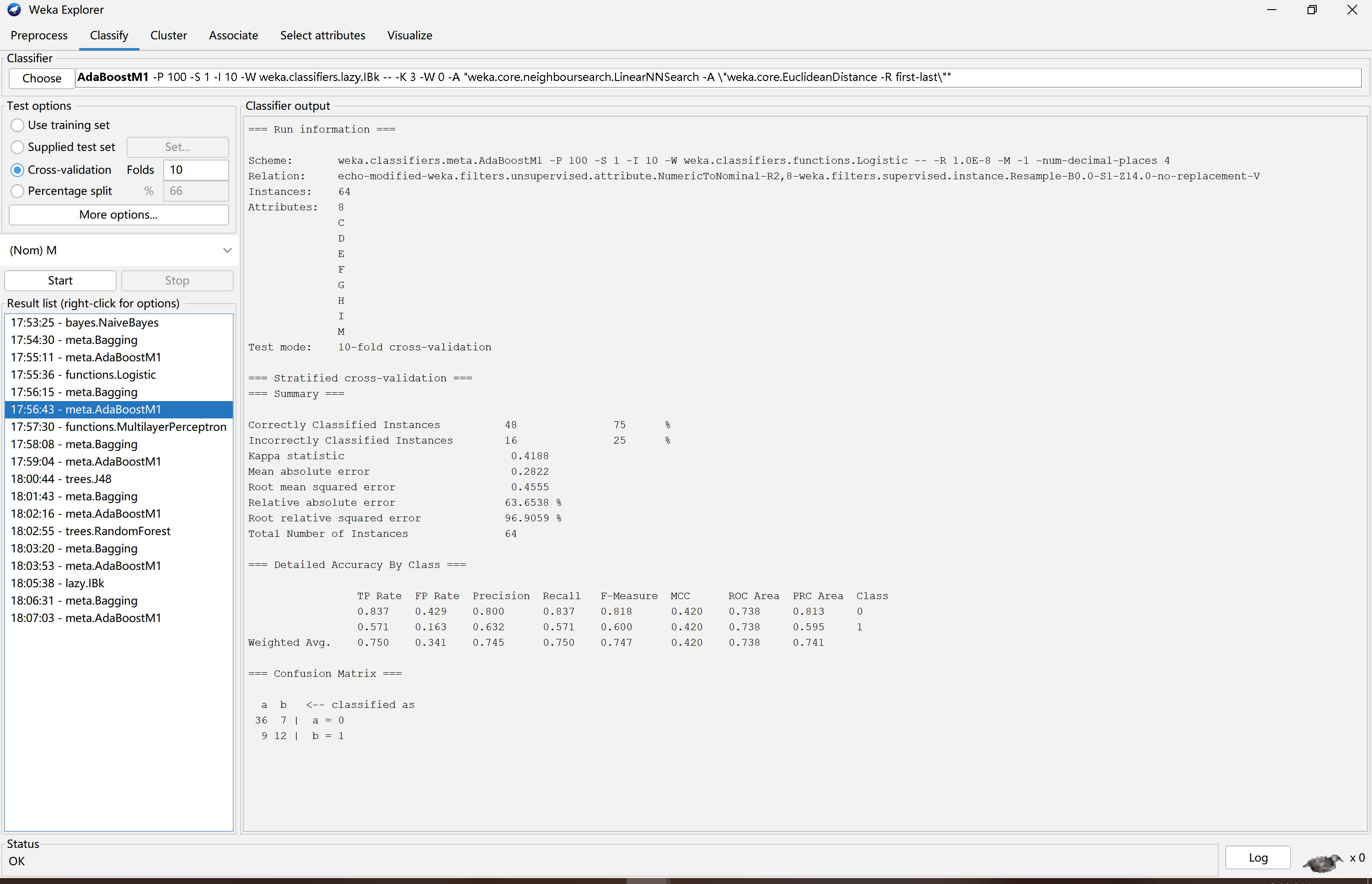


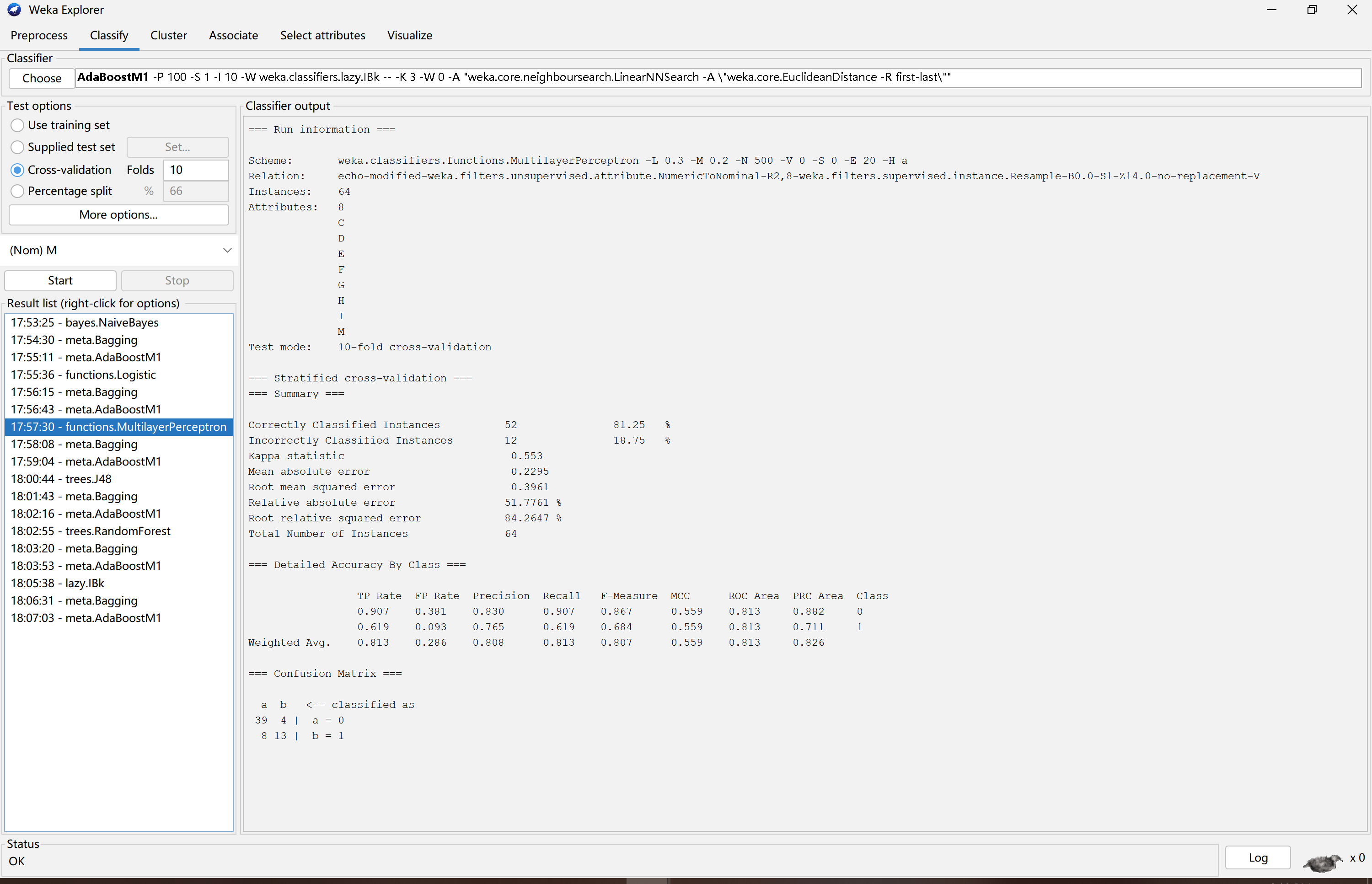


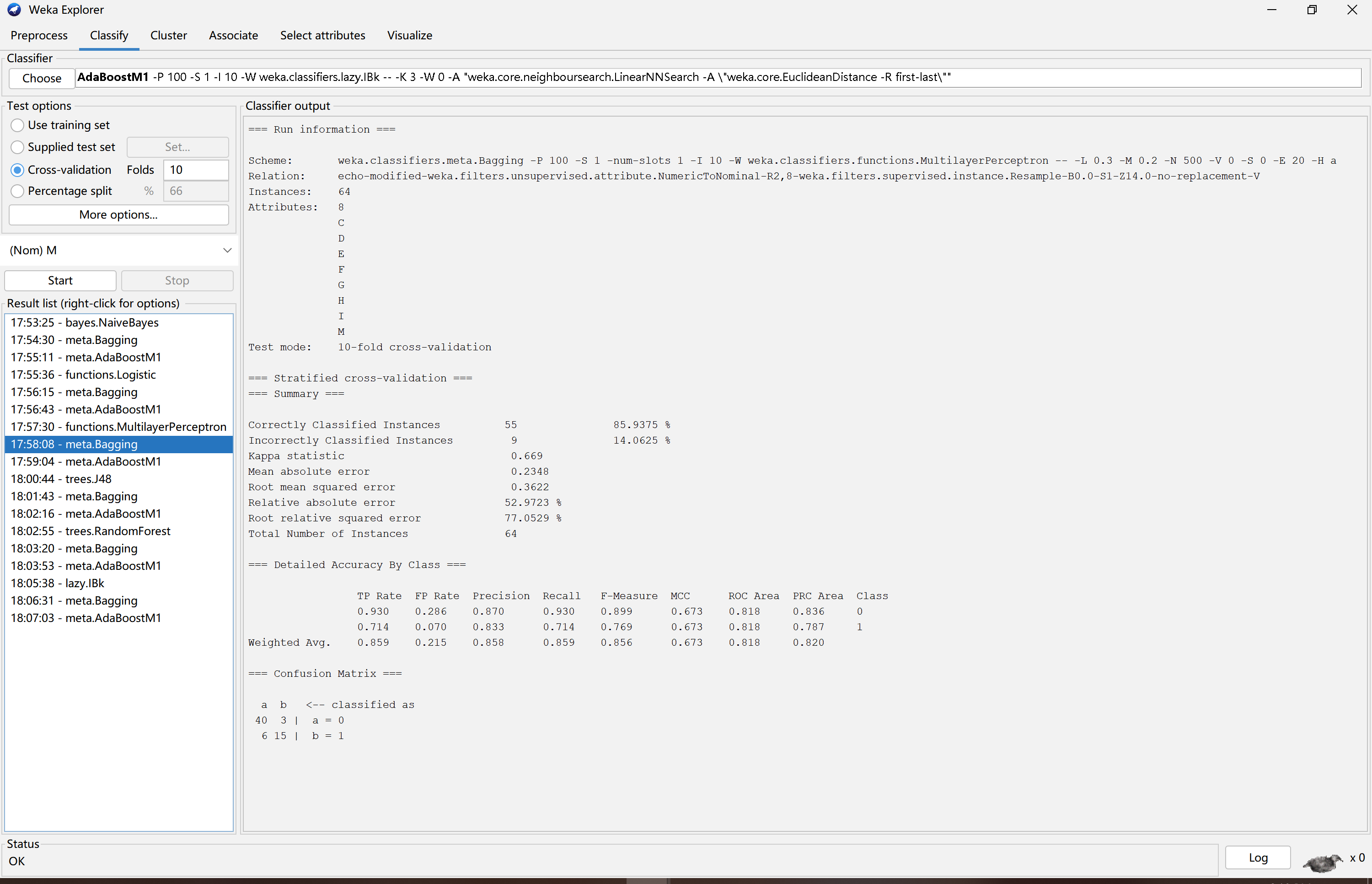


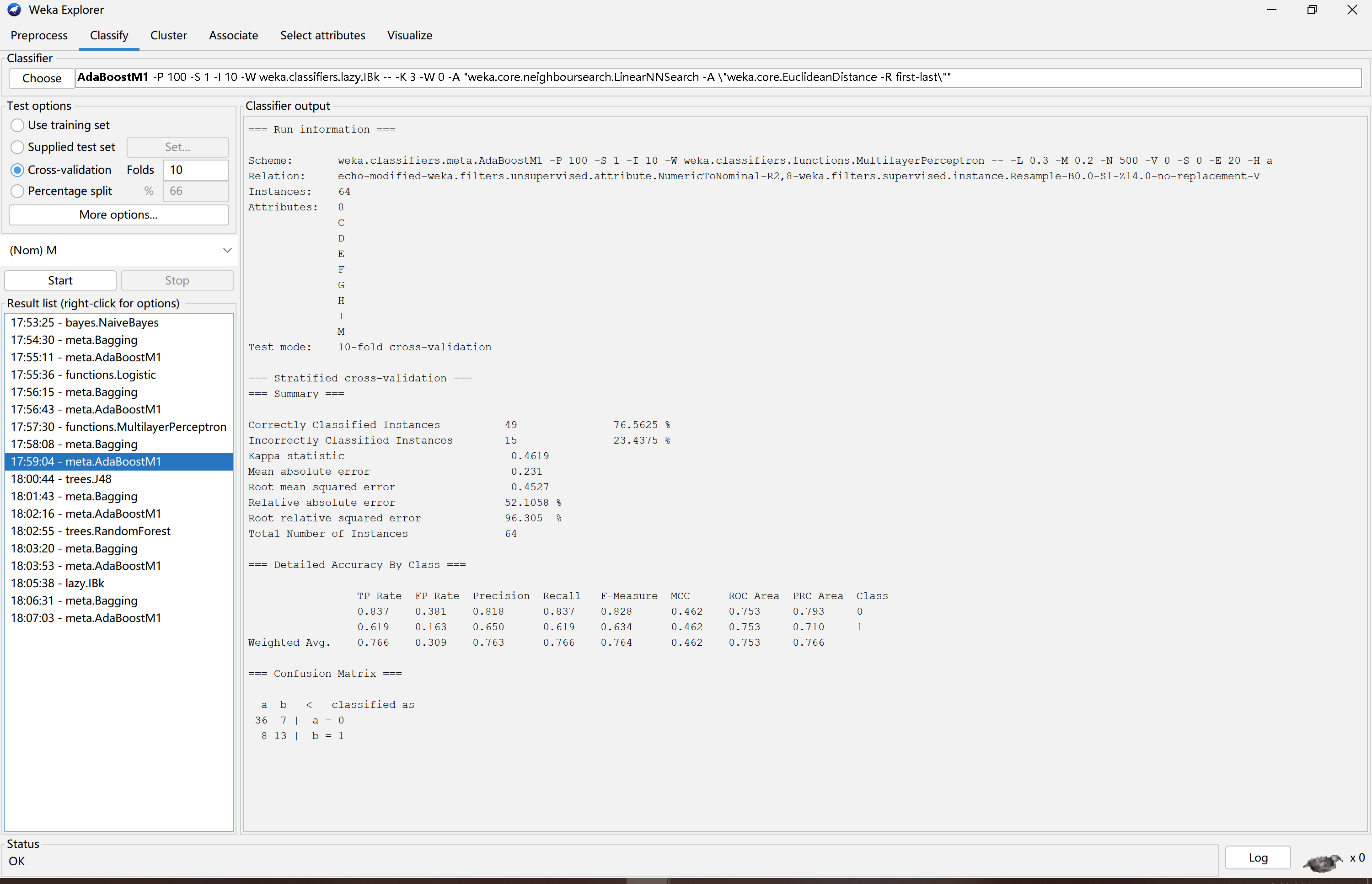


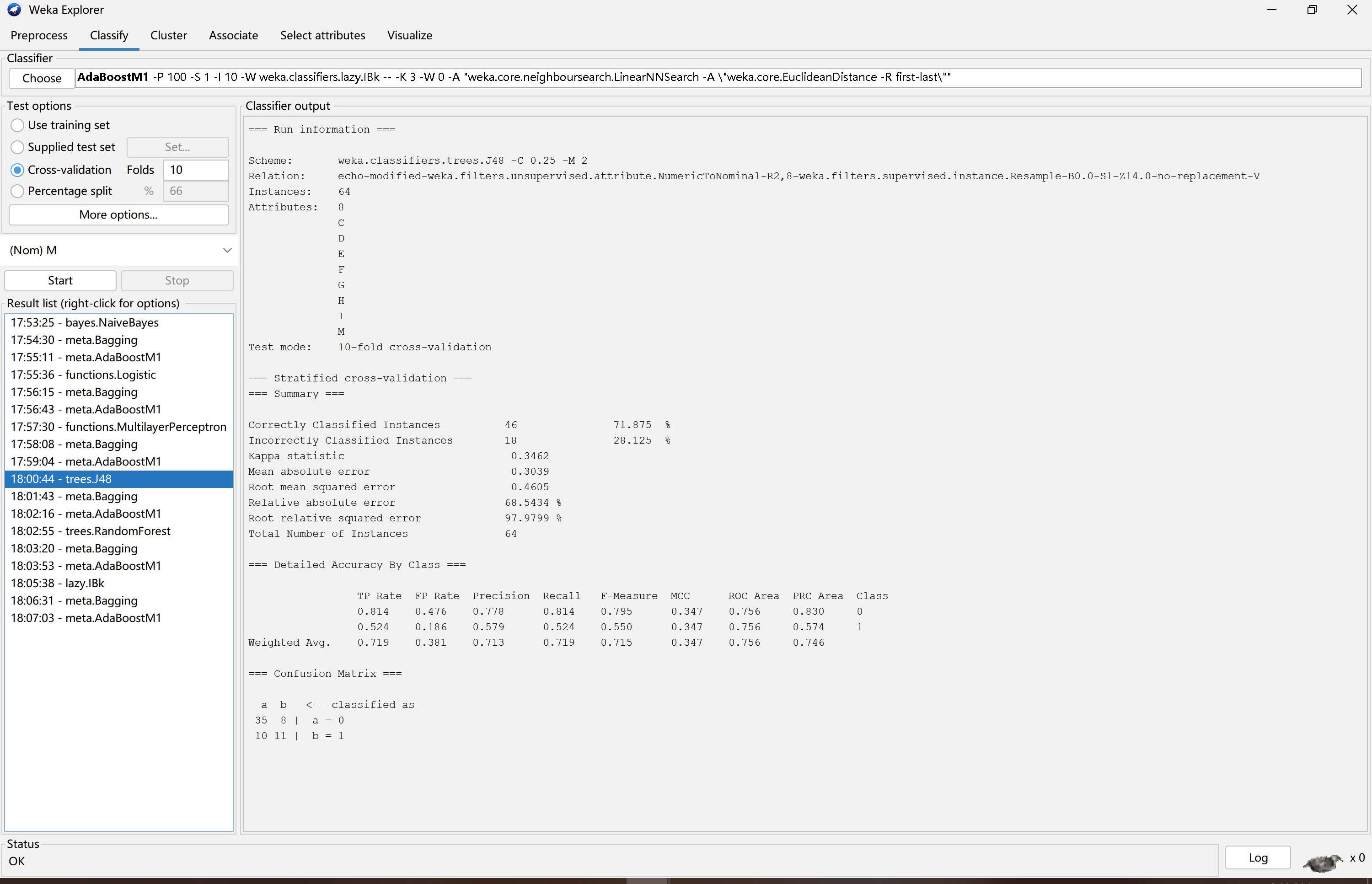


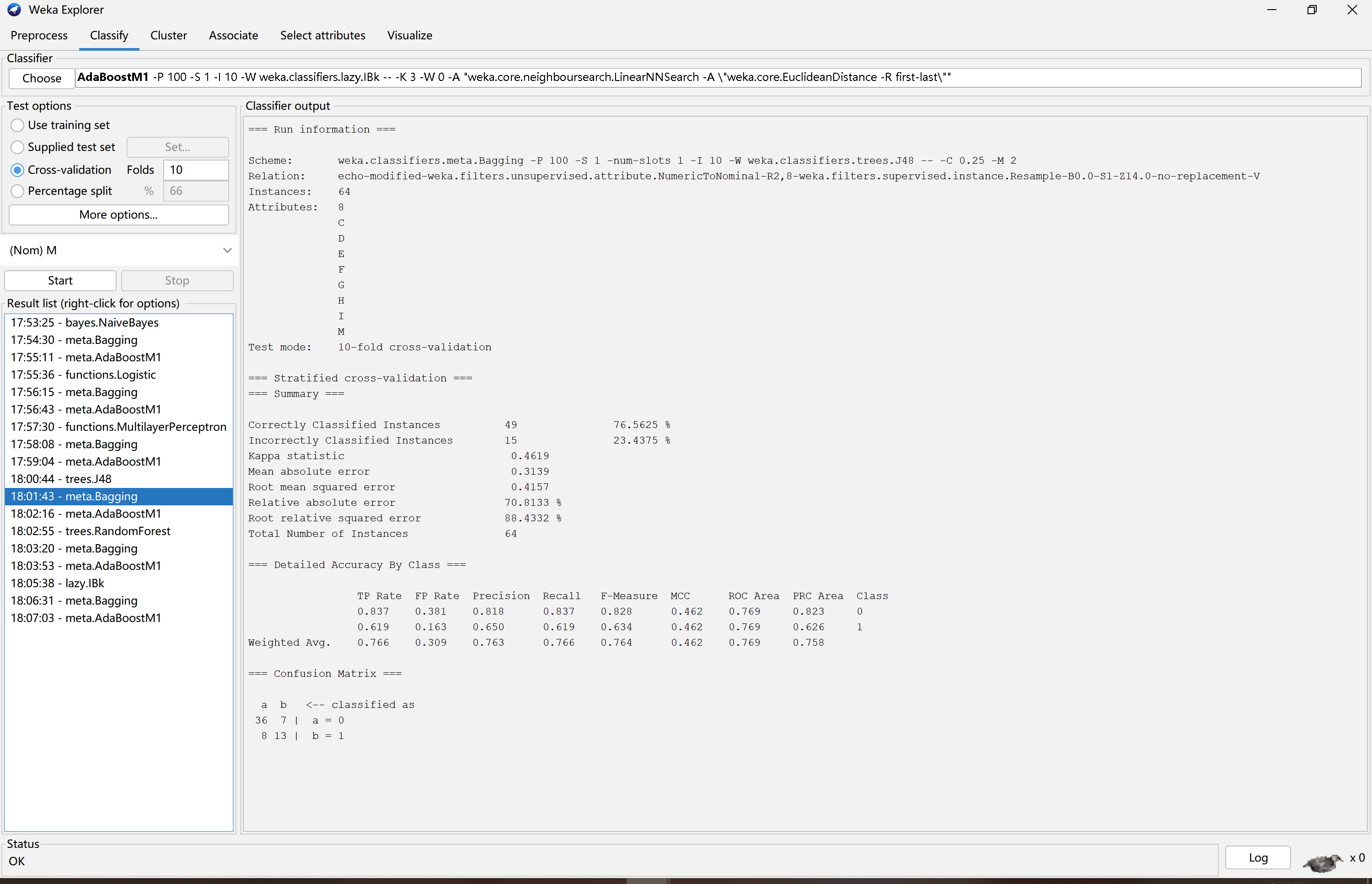


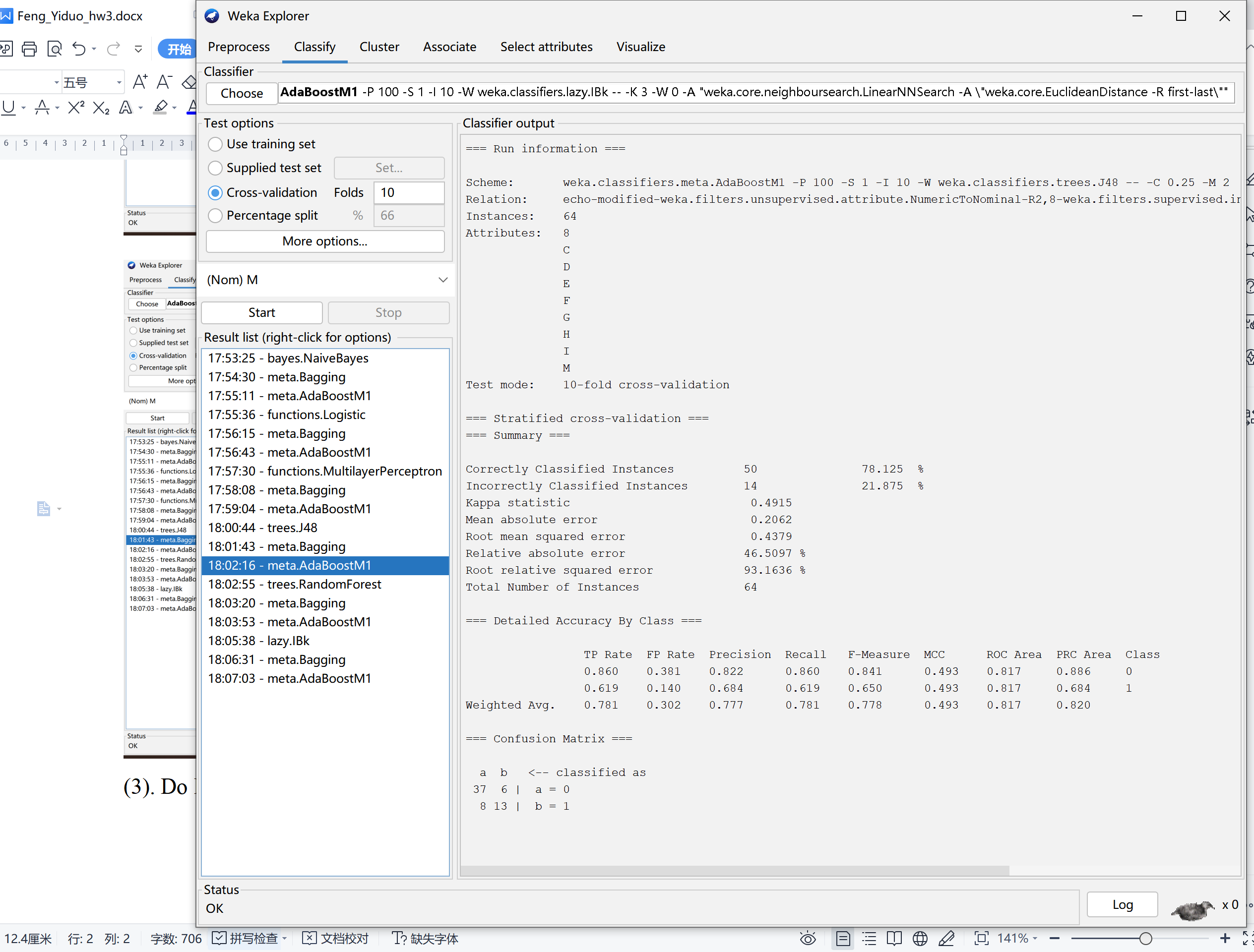


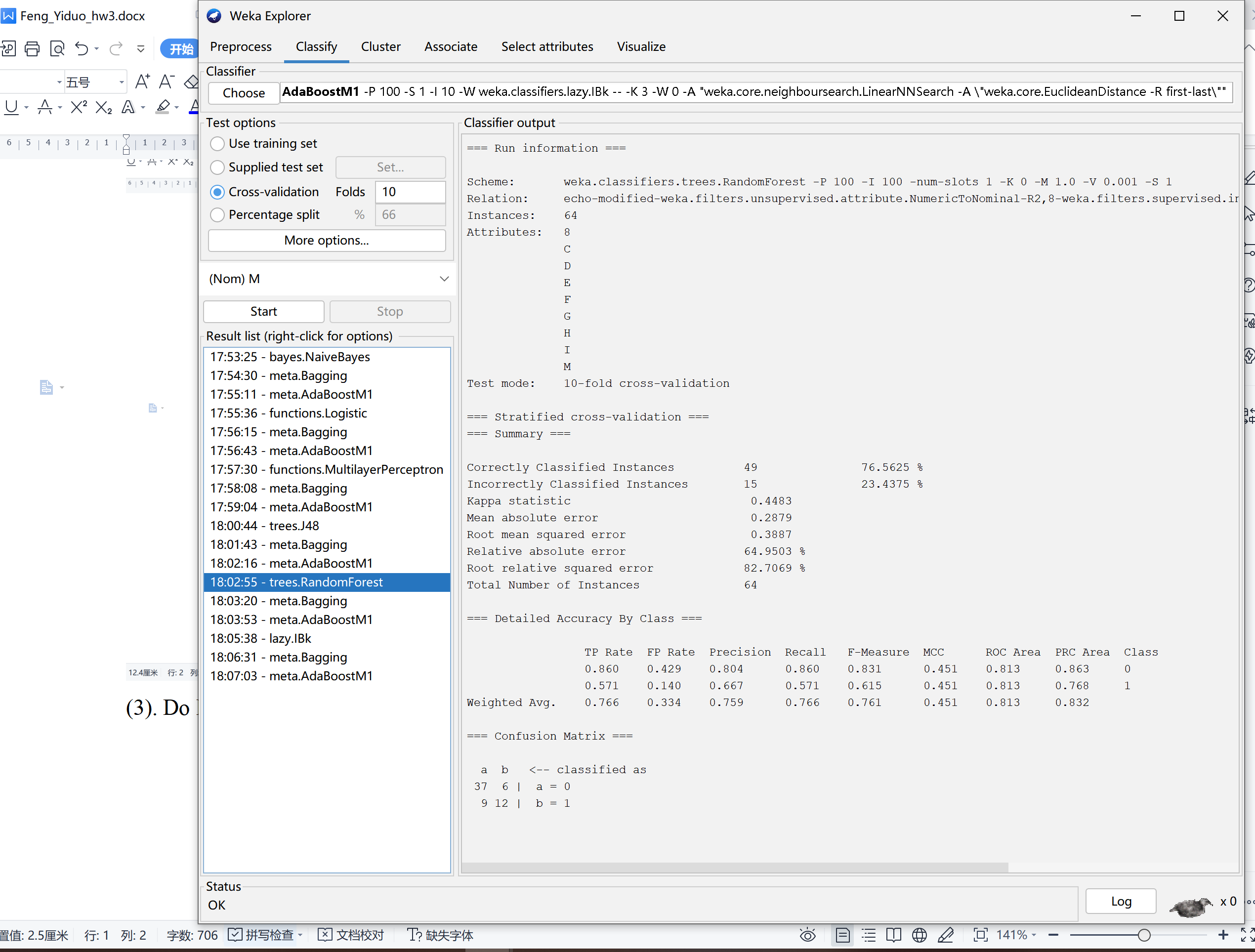


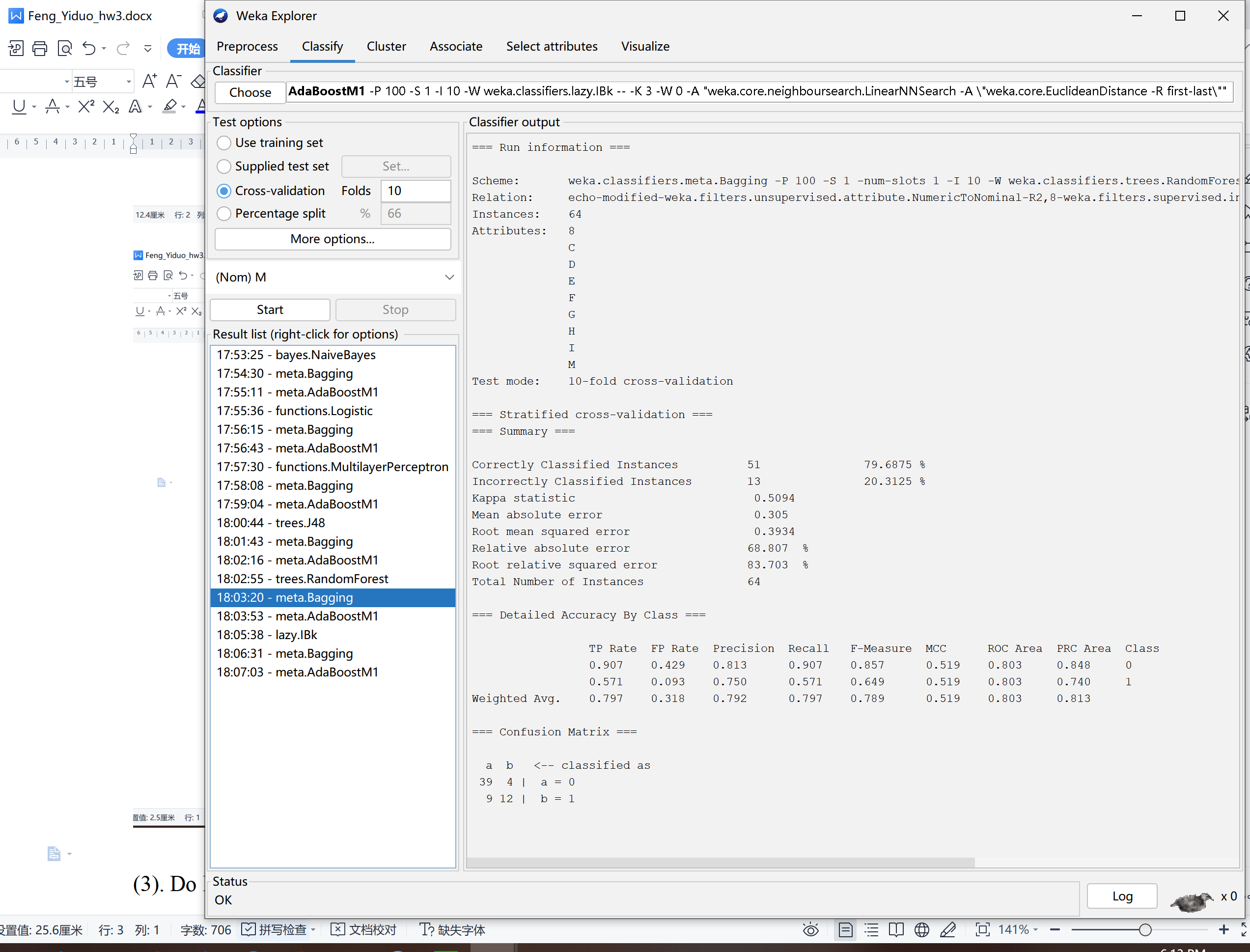


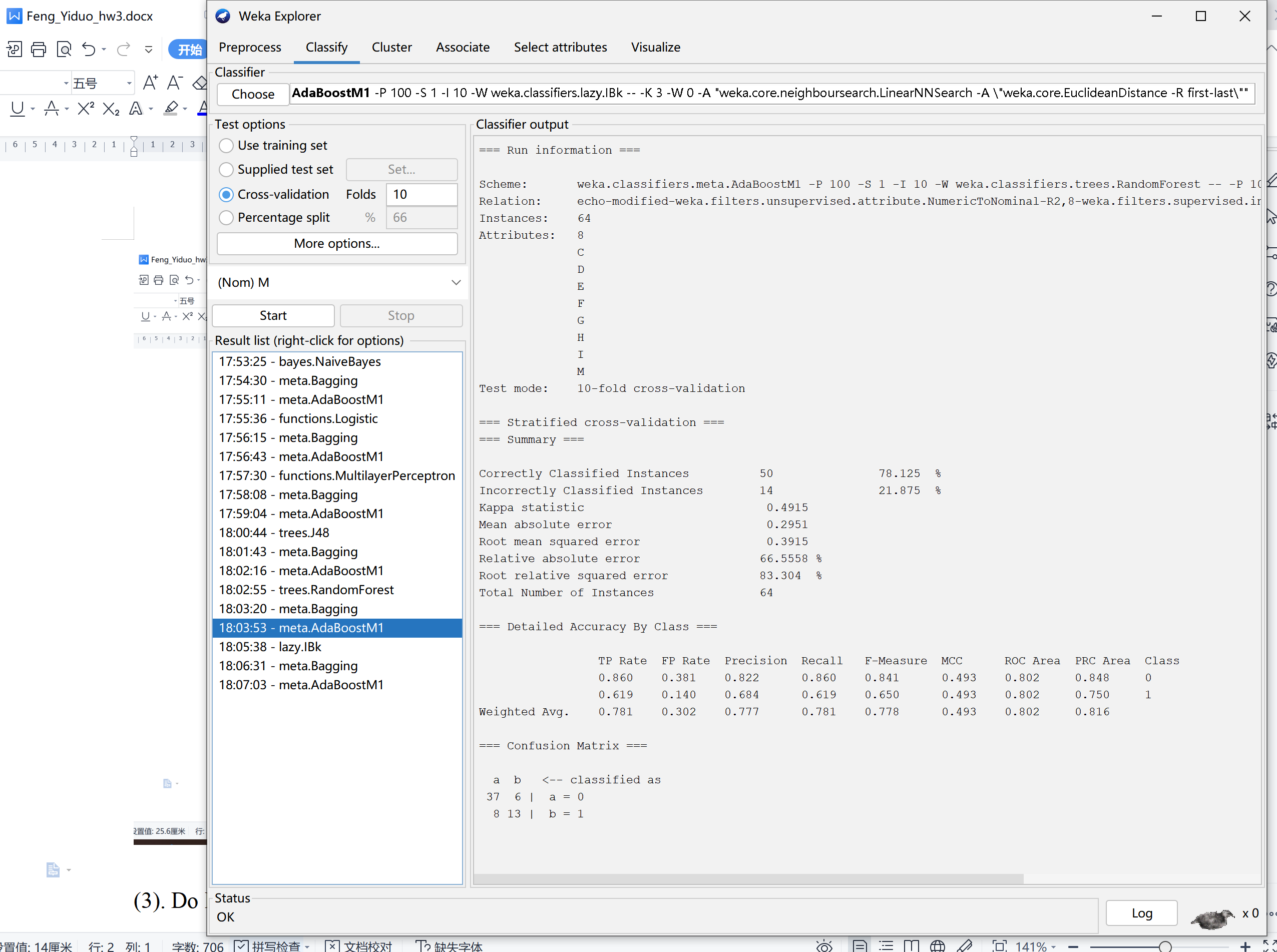


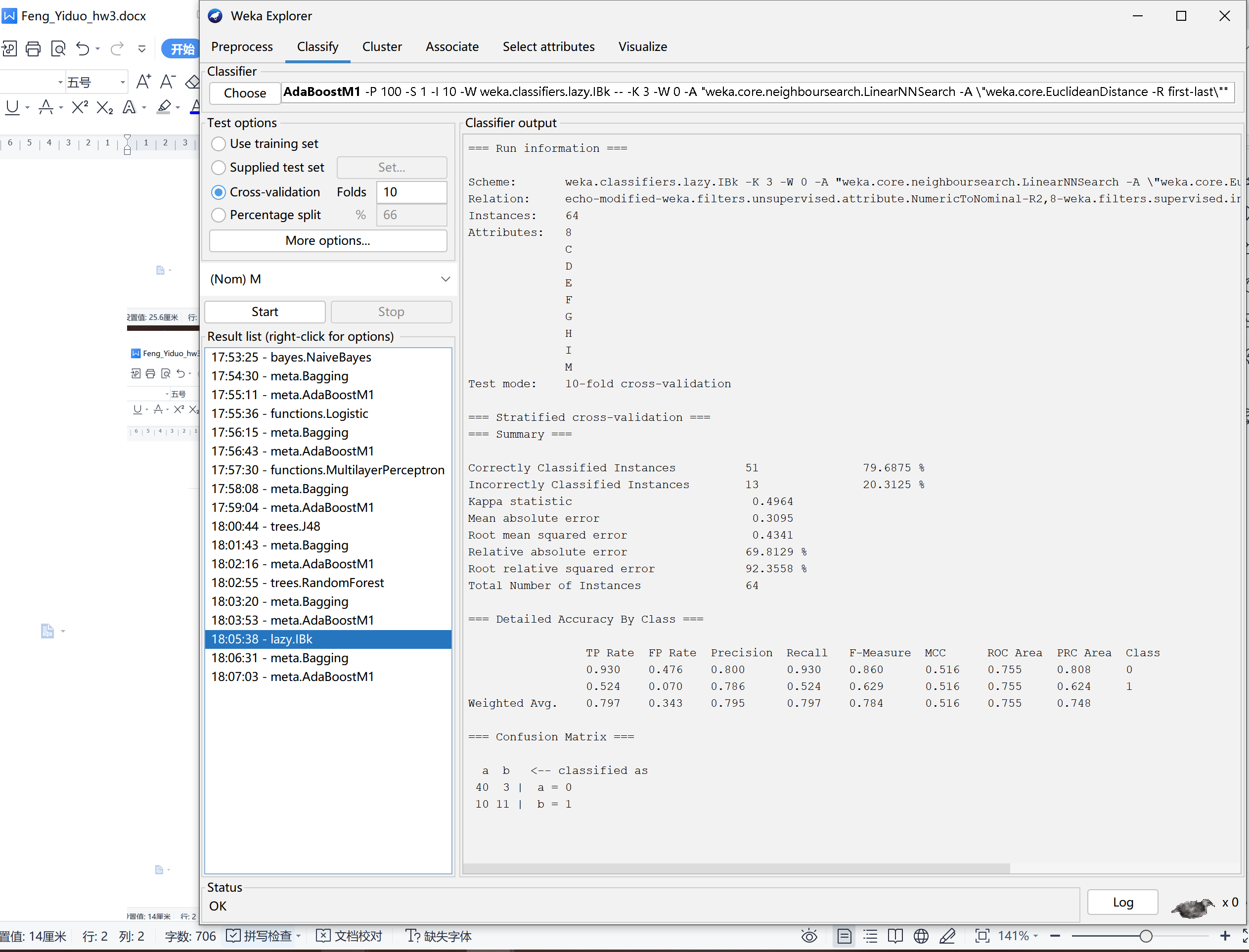


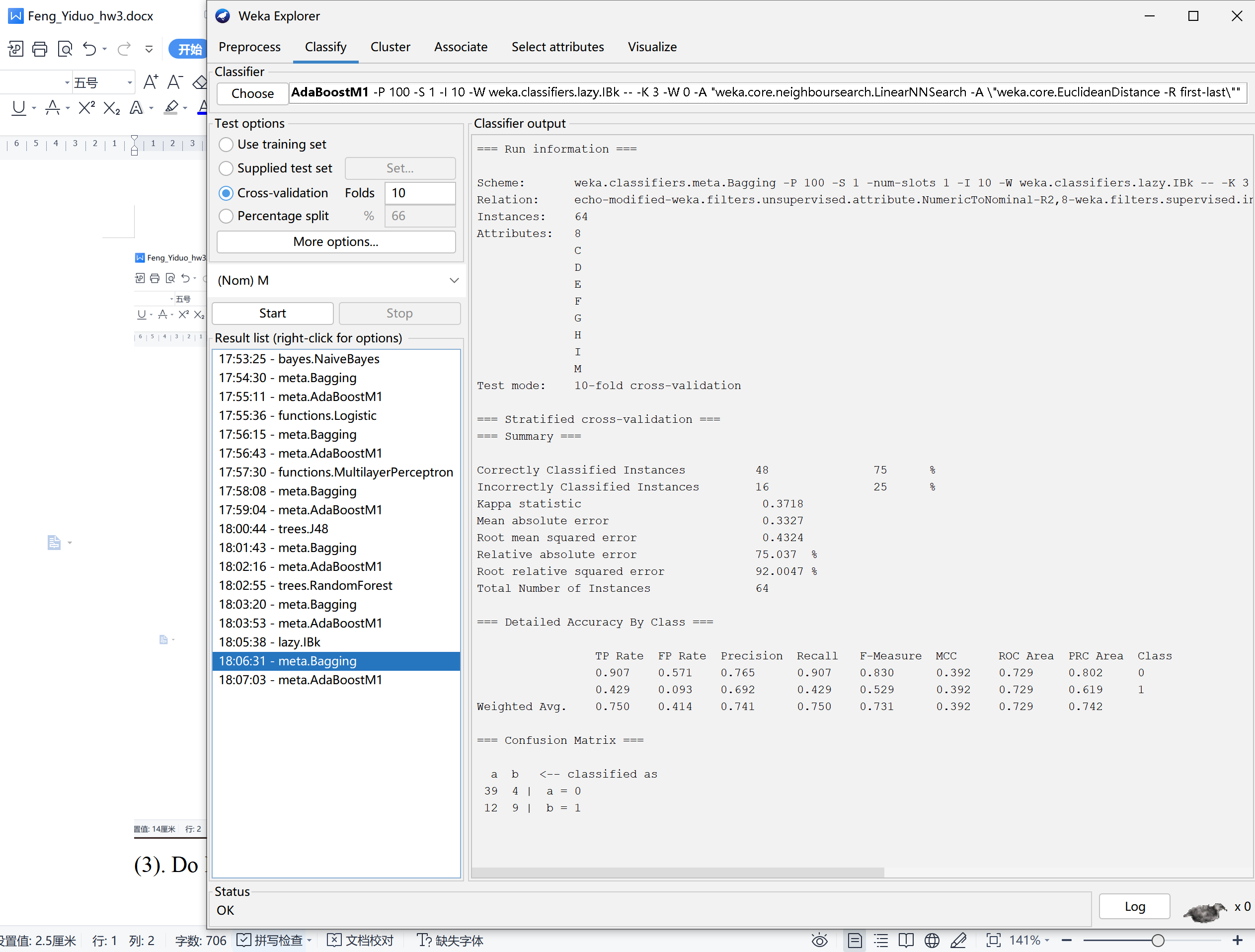


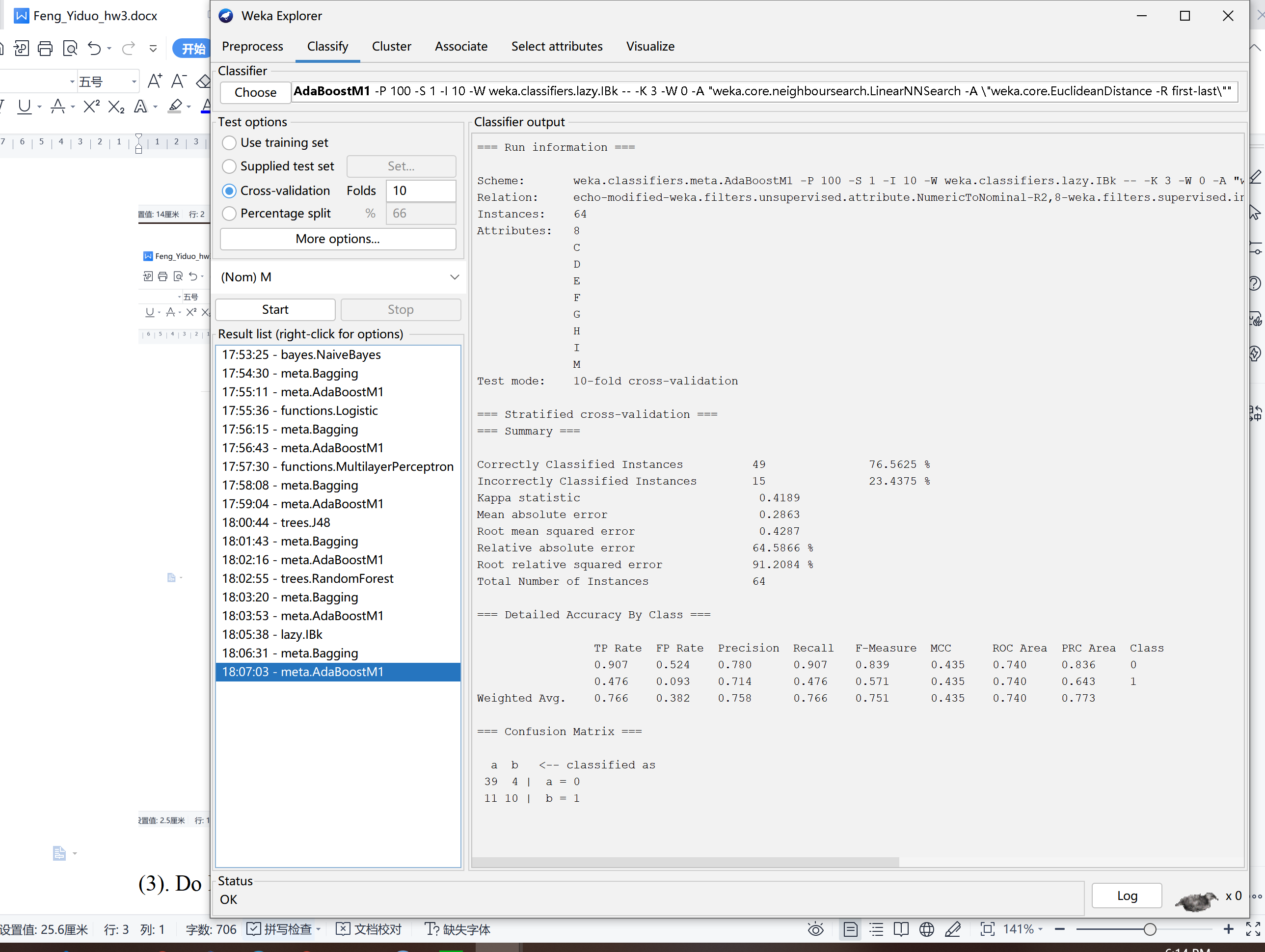












(3). Do Bagging and AdaBoostM1 increase accuracies?

Bagging increase accuracy for Naive Bayes, MultilayerPerceptron, J48, RandomForest.

AdaBoostM1 increase accuracy for J48 and RandomForest.

They are not always increasing the accuracy, but in this case, both of them increase in J48 and RandomForest.