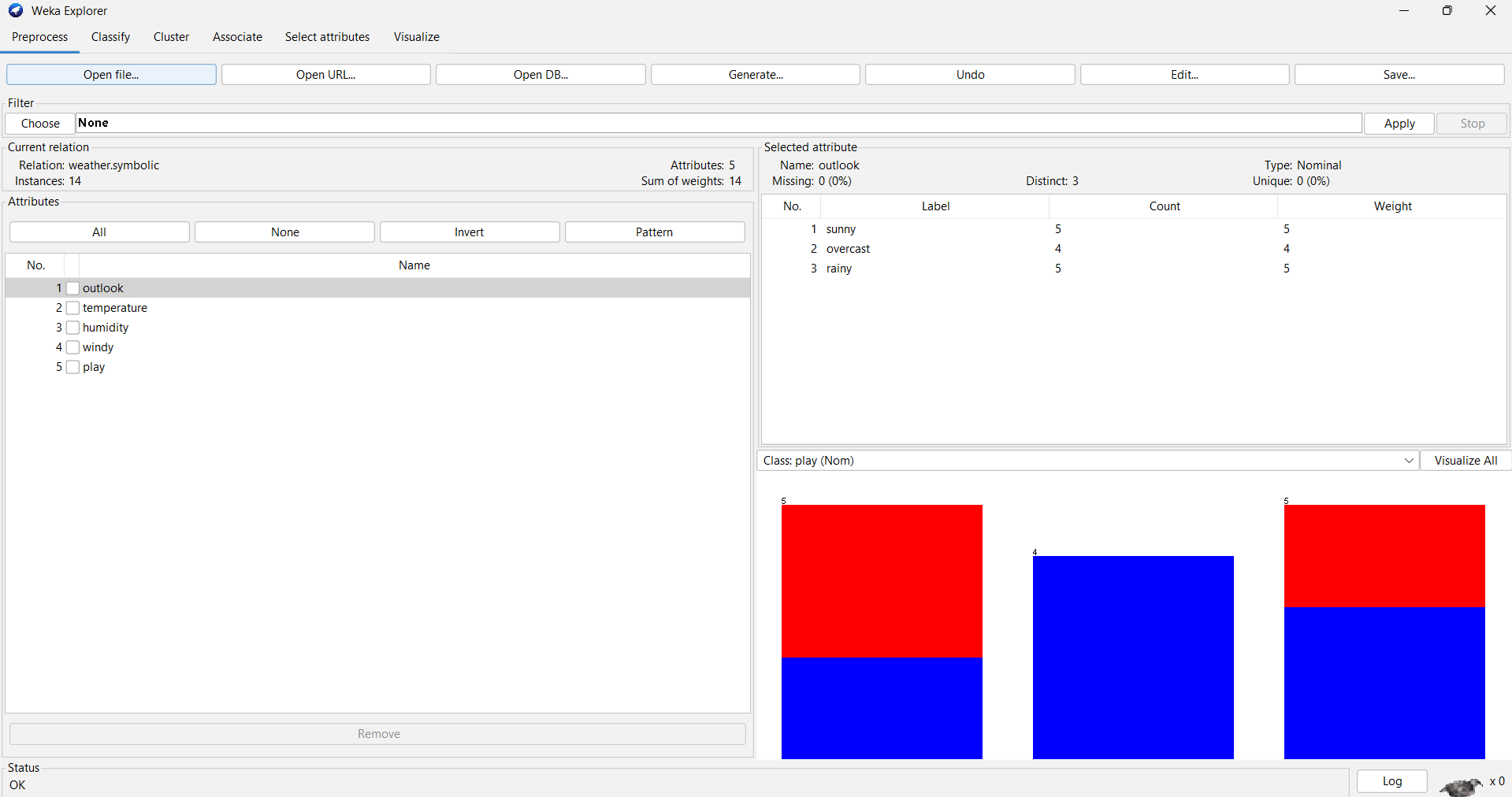
**Lab 6-7: Classification**

**Exercise 1: Using Weka Explorer to construct decision tree for the weather.nominal.arff with the J48 algorithm (or the other classification model) and classify it.**1. Launch the Weka tool and activate the **Explorer** environment.  
2. Open the “**weather.nominal**” dataset

  
3. Go to the **Classify** tab. Select the **J48** classifier. Choose the “Cross-validation” (10 folds) test mode.  
Run the classifier and observe the results shown in the “Classifier output” window.

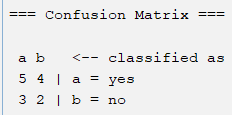
  
- How many instances are incorrectly classified?



+ Incorrectly Classified Instances 7 50 %  
- What is the MAE (mean absolute error) made by the classifier?



+ Mean absolute error 0.4167  
- What can you infer from the information shown in the Confusion Matrix?

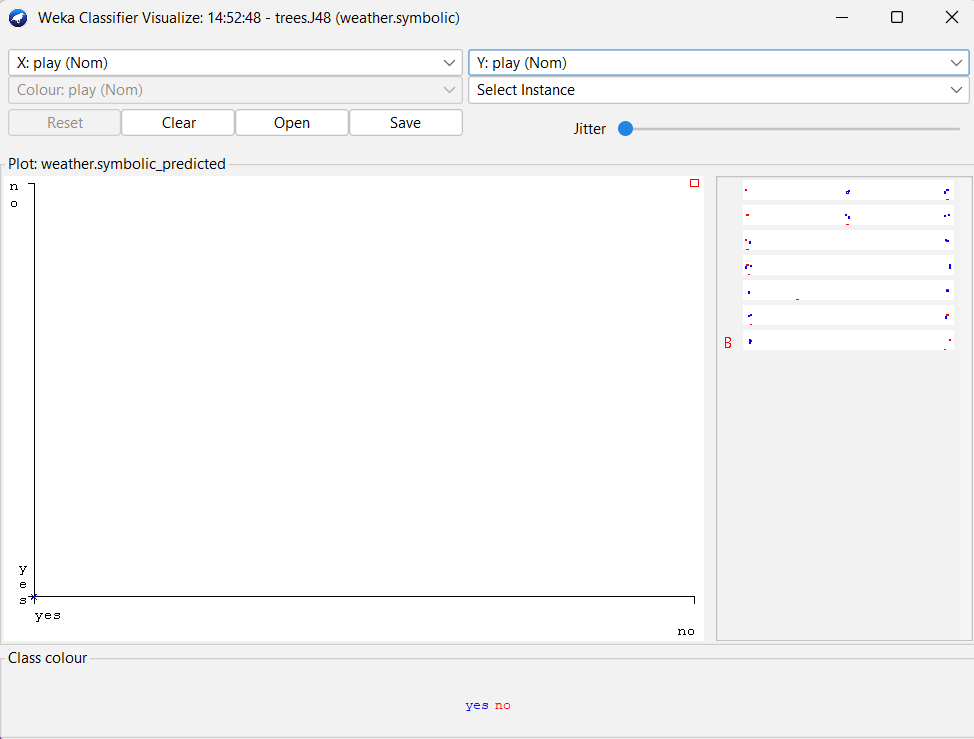


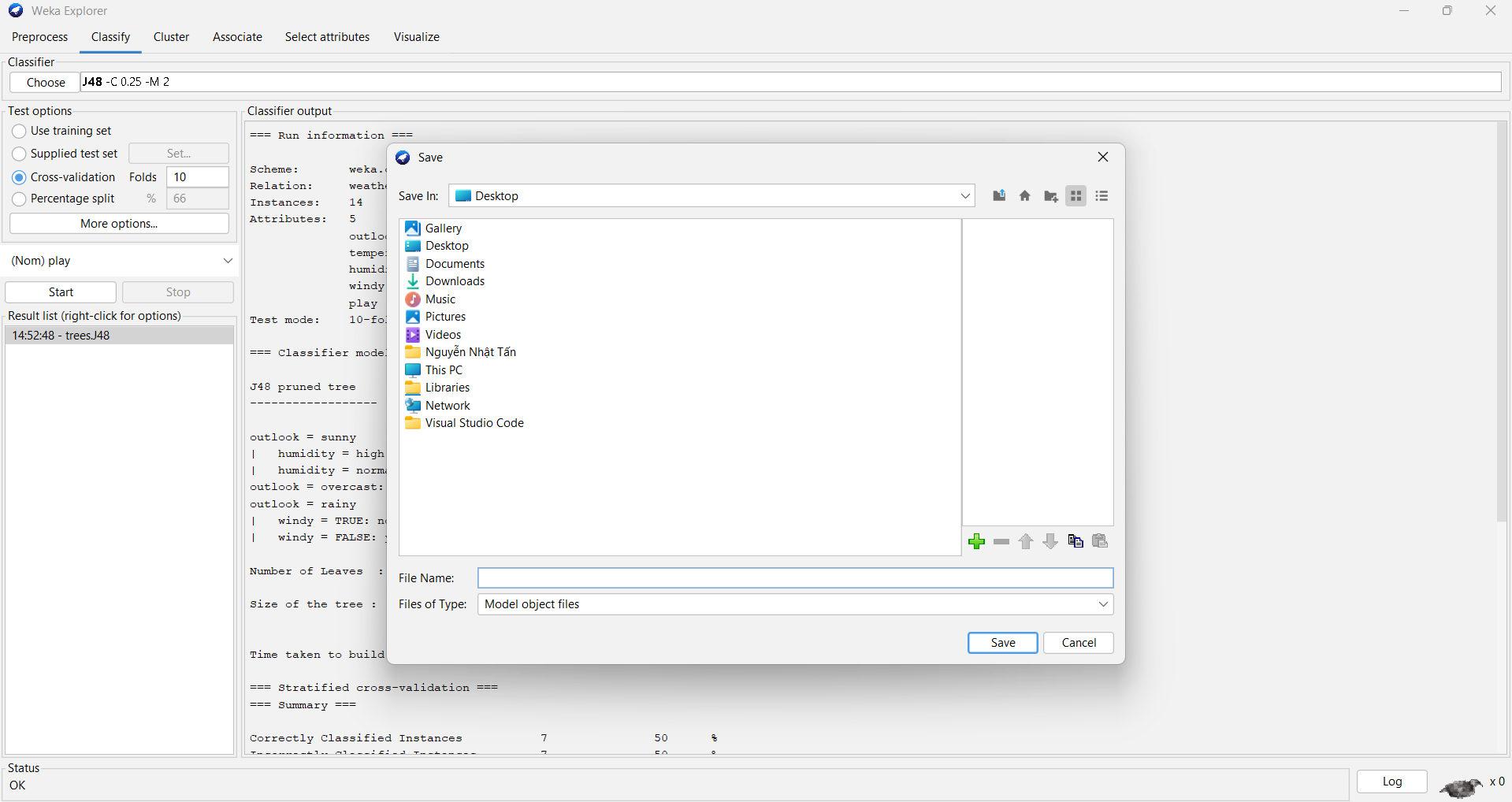
=== Confusion Matrix ===

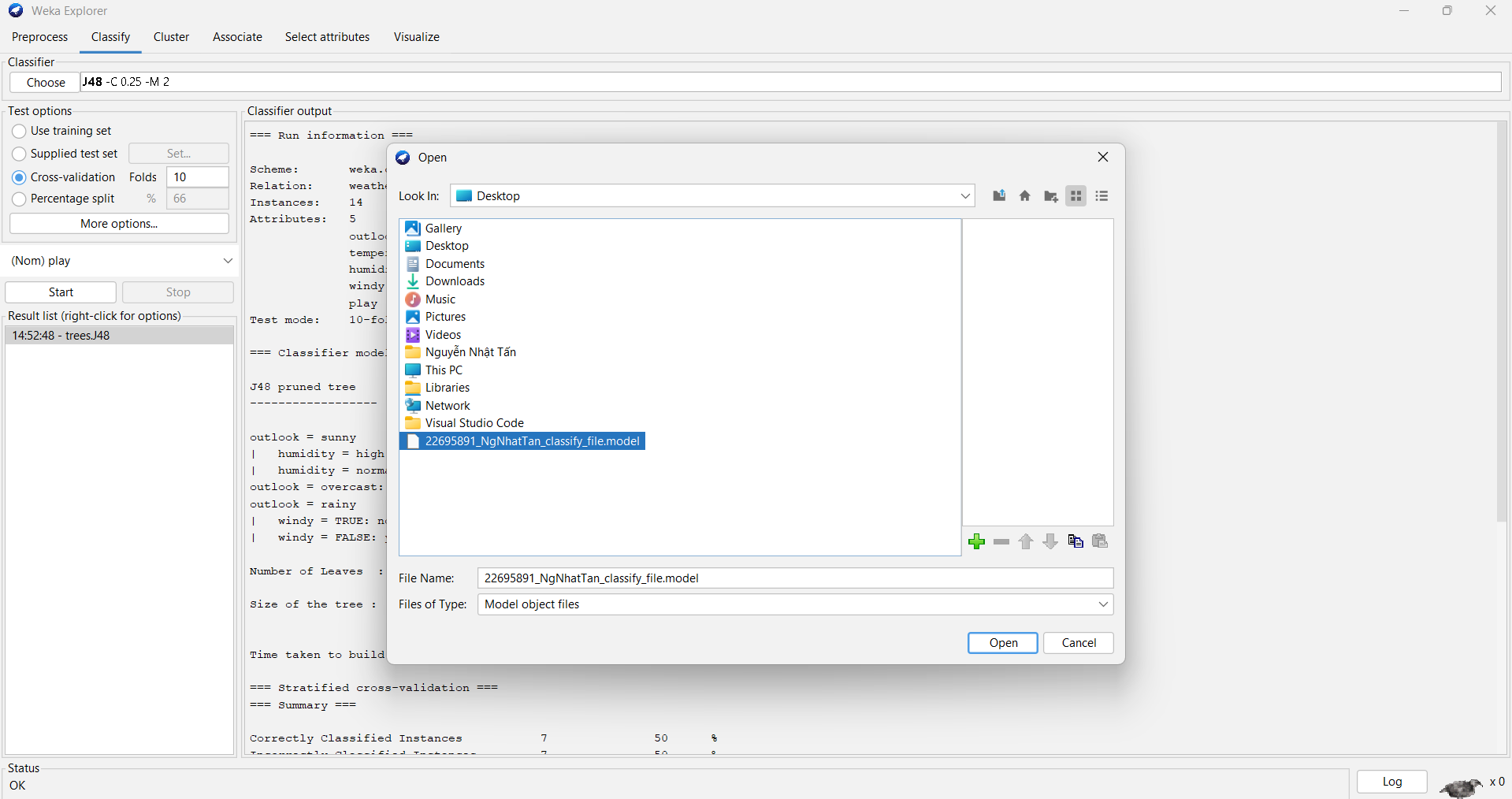
a b <-- classified as

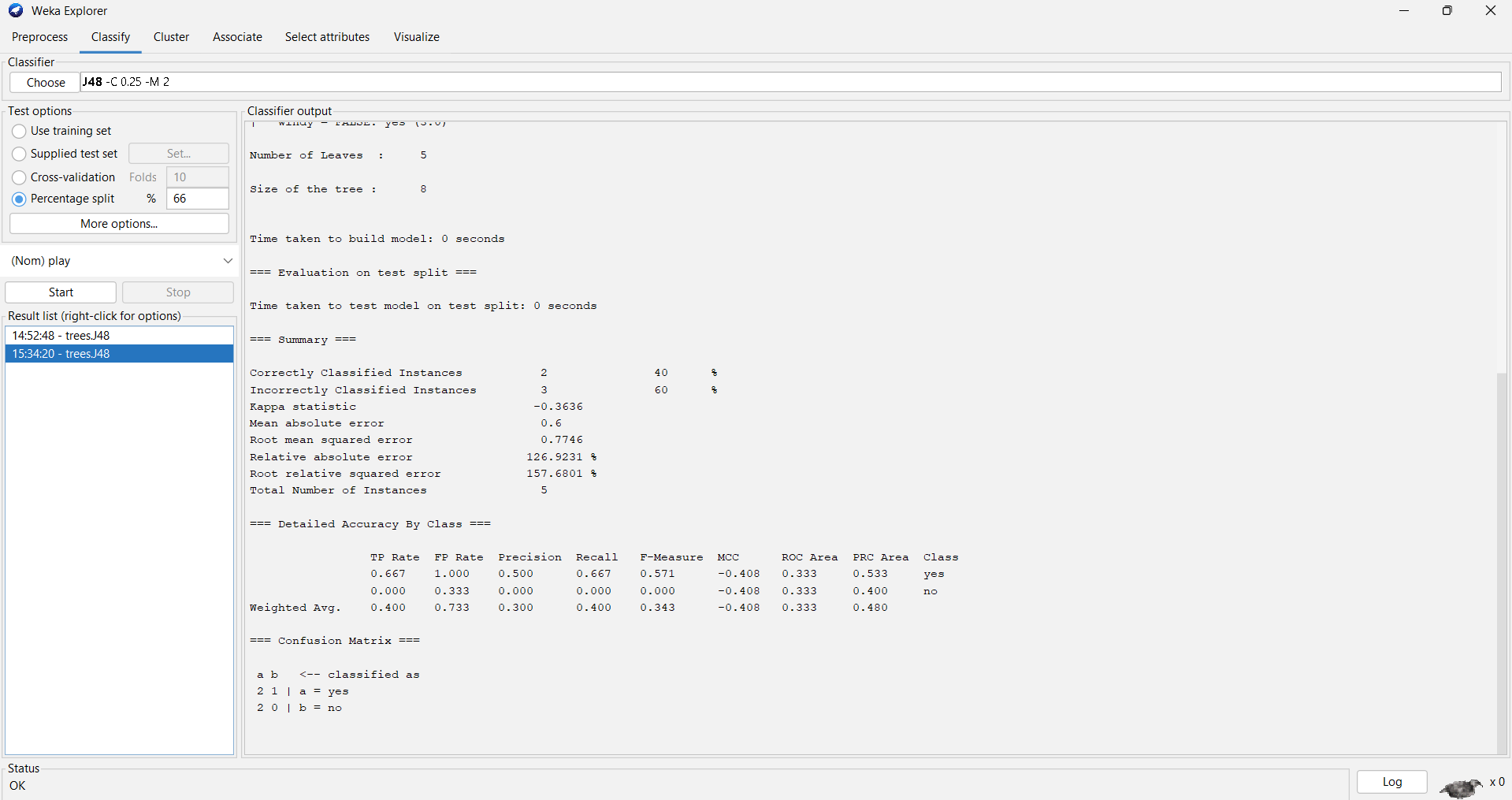
5 4 | a = yes

3 2 | b = no  
- Visualize the classifier errors. In the plot, how can you differentiate between the correctly and incorrectly classified instances?

  
- How can you save the learned classifier to a file?

  
- How can you load a learned classifier from a file?

  
4. Choose the “Percentage split” (66% for training) test mode. Run the **J48** classifier and observe the results shown in the “Classifier output” window.

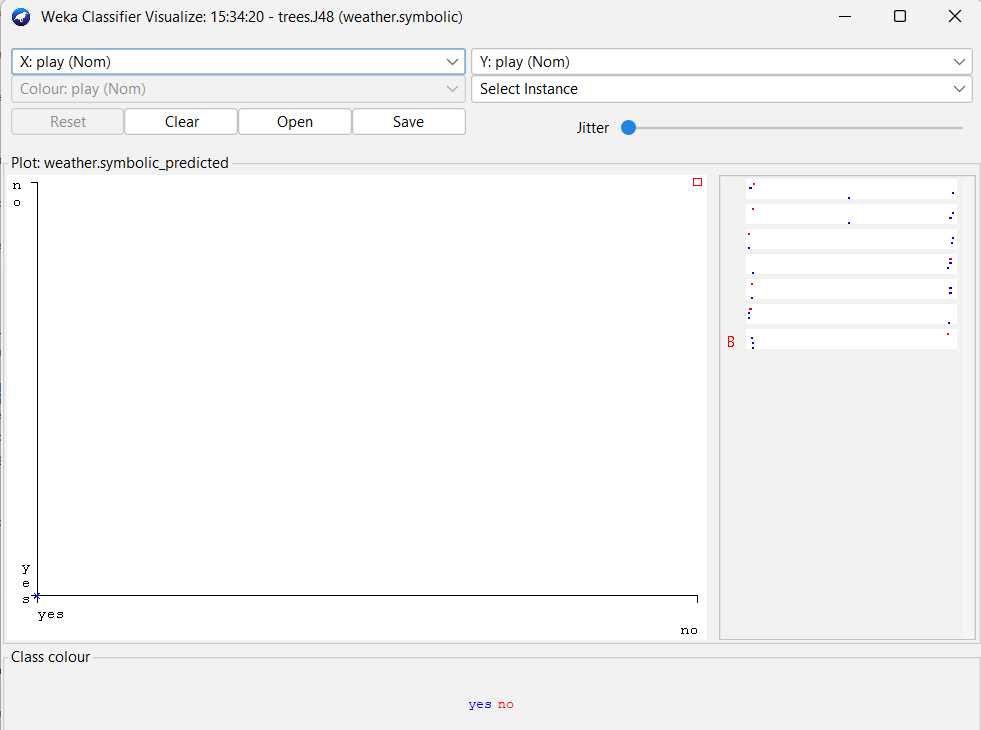
  
- How many instances are incorrectly classified? Why this number is smaller than that observed in the previous experiment (i.e., using the cross-validation test mode)?



Incorrectly Classified Instances 3 60 %  
- What is the MAE made by the classifier?



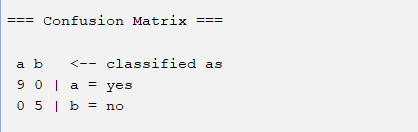
Mean absolute error 0.6   
- Visualize the classifier errors to see the detailed information.

  
5. Choose the “Use training set” (66% for training) test mode. Run the **J48** classifier and observe the results shown in the “Classifier output” window.

  
- How many instances are incorrectly classified? Why this number is smaller than that observed in the previous experiment (i.e., using the cross-validation test mode)?



Incorrectly Classified Instances 0 0 %  
- What is the MAE made by the classifier?

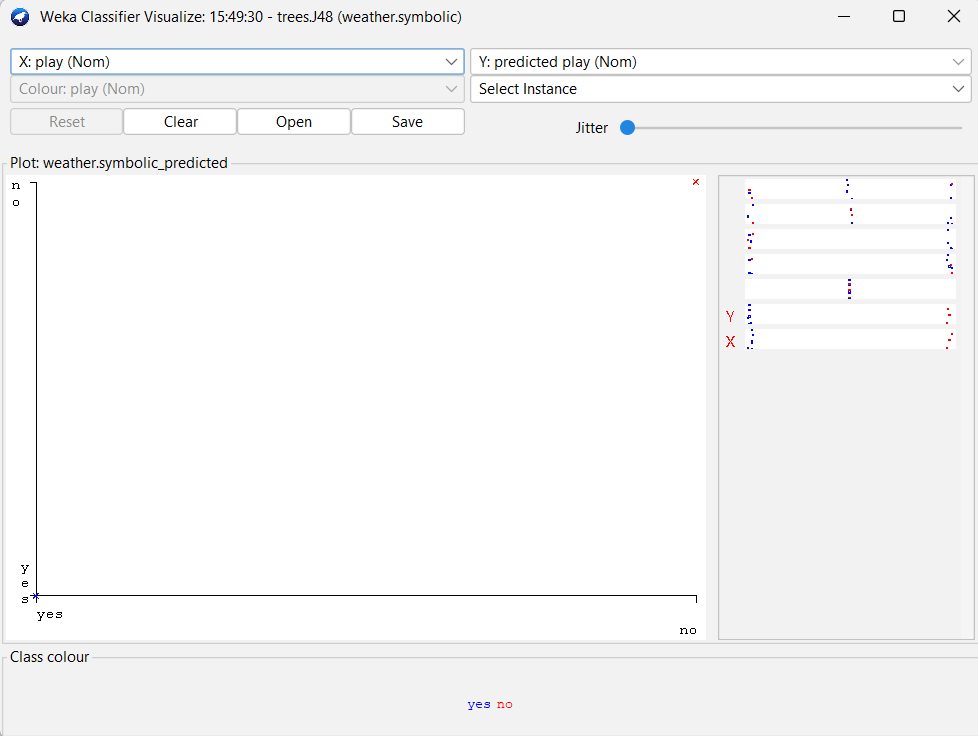


=== Confusion Matrix ===

a b <-- classified as

9 0 | a = yes

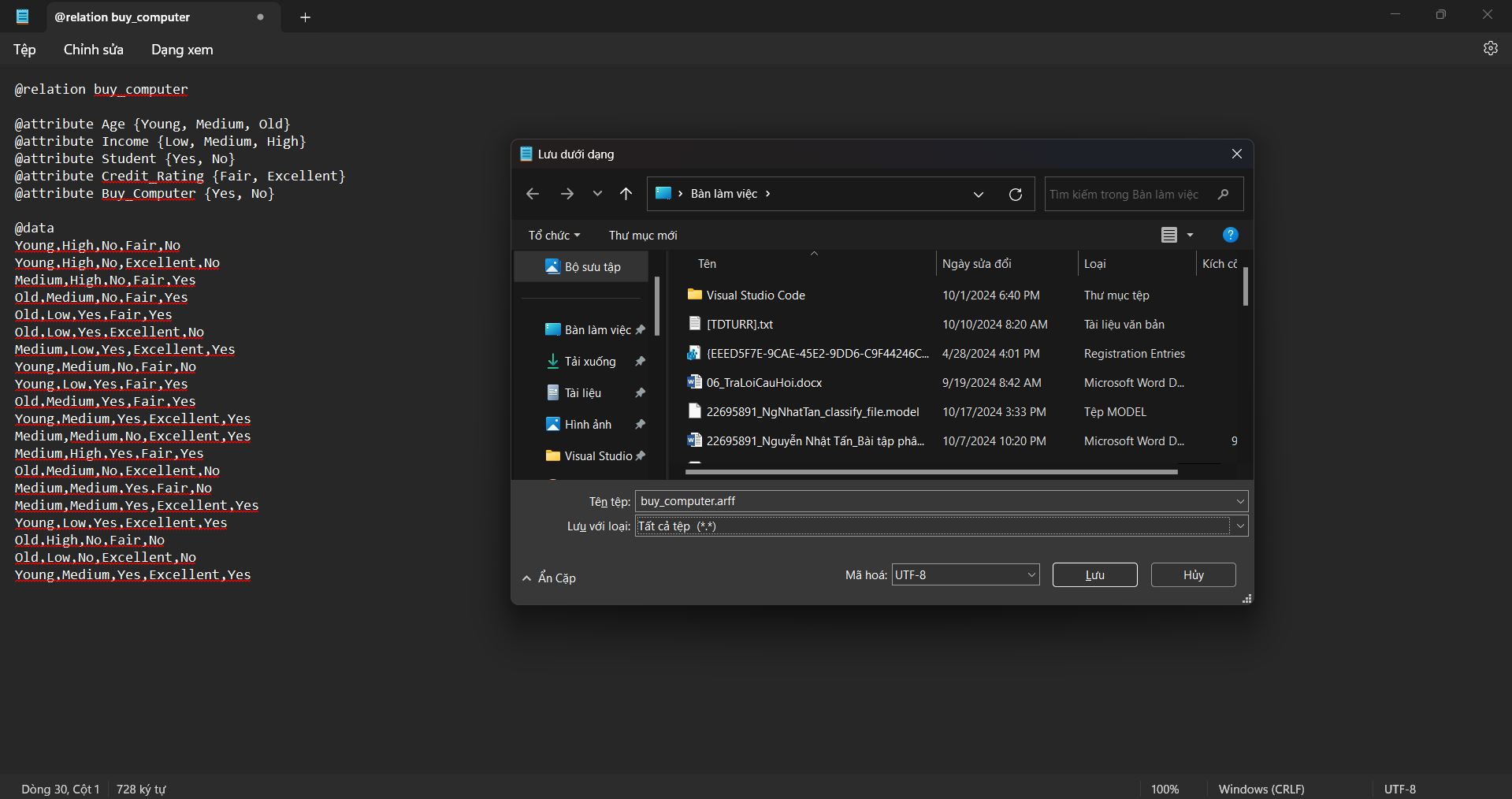
0 5 | b = no  
- Visualize the classifier errors to see the detailed information.

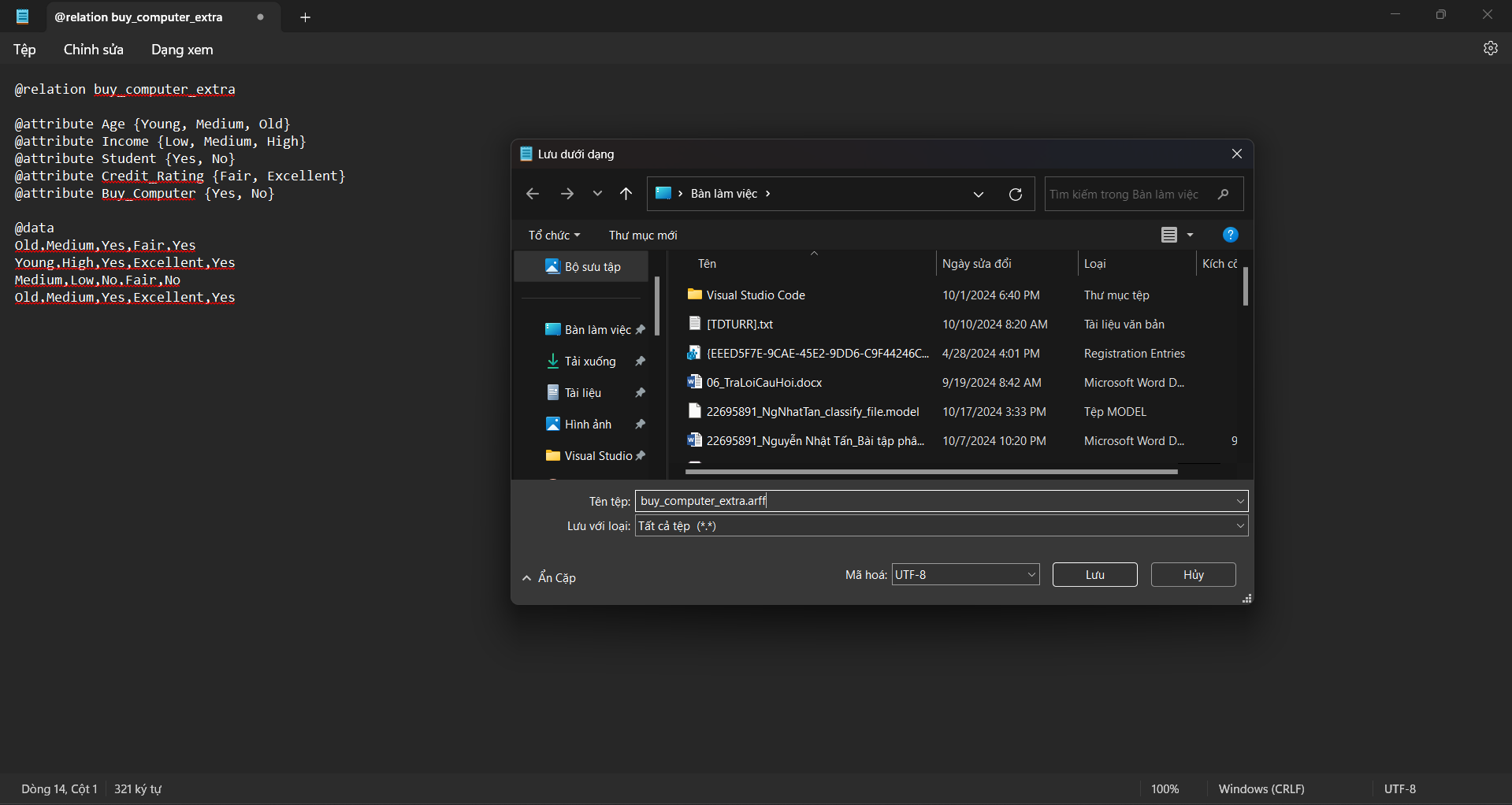
  
6. Compare test modes provided in the J48 classifiers and fill-in the following table:

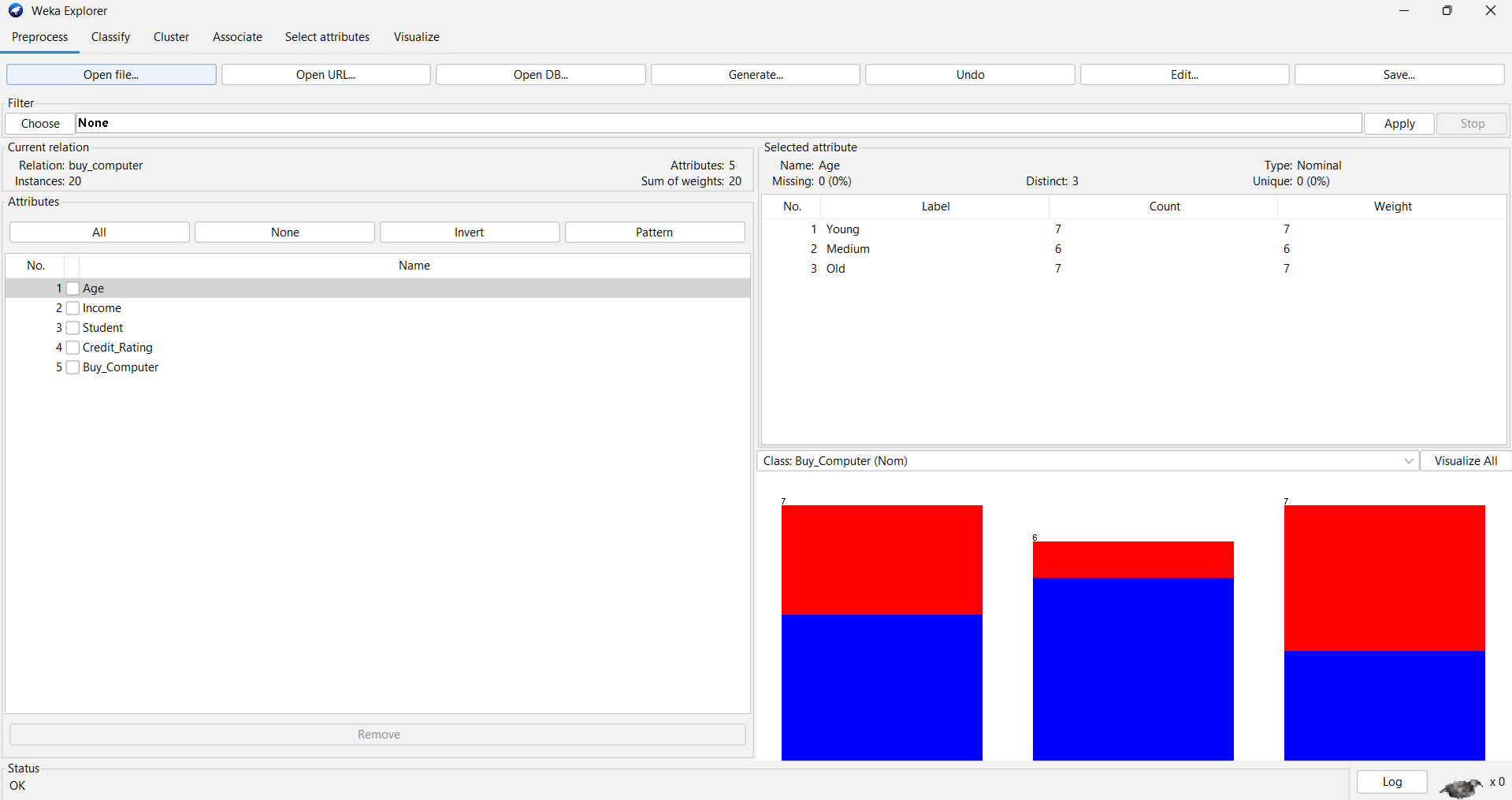
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test option** | **Kappa statistic** | **Mean absolute error** | **Root mean squared error** | **Relative absolute error** | **Root relative squared error** |
| ***Use training set*** |  |  |  |  |  |
| ***Cross-validation*** |  |  |  |  |  |
| ***Percenttagesplit*** |  |  |  |  |  |

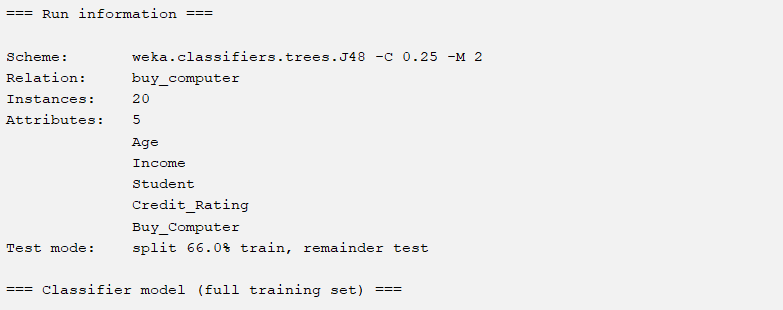
7. Compare the performance of other classification approaches in this dataset:  
- J48  
- OneR  
- Naive Bayes  
**Exercise 2:**

Let’s assume that we have collected the following data set of users who decided to buy a computer  
and others who decided not. Each user record (i.e., example) is represented by the 5 attributes.  
- Age, with the possible values {Young, Medium, Old}.  
- Income, with the possible values {Low, Medium, High}.  
- Student, with the possible values {Yes, No}.  
- Credit\_Rating, with the possible values {Fair, Excellent}.  
- Buy\_Computer – the classification attribute, with the possible values {Yes, No}.  
***Data Mining and Applications Department of Computer Science***25  
**UserID Age Income Student Credit\_Rating Buy\_Computer**1 Young High No Fair No  
2 Young High No Excellent No  
3 Medium High No Fair Yes  
4 Old Medium No Fair Yes  
5 Old Low Yes Fair Yes  
6 Old Low Yes Excellent No  
7 Medium Low Yes Excellent Yes  
8 Young Medium No Fair No  
9 Young Low Yes Fair Yes  
10 Old Medium Yes Fair Yes  
11 Young Medium Yes Excellent Yes  
12 Medium Medium No Excellent Yes  
13 Medium High Yes Fair Yes  
14 Old Medium No Excellent No  
15 Medium Medium Yes Fair No  
16 Medium Medium Yes Excellent Yes  
17 Young Low Yes Excellent Yes  
18 Old High No Fair No  
19 Old Low No Excellent No  
20 Young Medium Yes Excellent Yes  
We want to predict, for each of the following users, if s/he will buy a computer or not.  
- User #21. A young student with medium income and fair credit rating.  
- User #22. A young non-student with low income and fair credit rating.  
- User #23. A medium student with high income and excellent credit rating.  
- User #24. An old non-student with high income and excellent credit rating.  
**Requirement:**- Create the dataset containing 20 examples (i.e., Users #1-20) into the ARFF format and save it in the “buy\_comp.arff” file.

  
- For each user in the set of Users #21-24, set the values of the Buy\_Computer attribute by the predictions computed manually. Create the data of these four users into the ARFF format, and save it in the “buy\_comp\_extra.arff” file.

  
- Use the Weka Explorer tool Open the “buy\_comp” dataset (i.e., saved in the “buy\_comp.arff”file).

  
- Go to the “Classify” tab. Select the **J48** classifier **(or the other classification model)**. Choose “Percentage split” (66% for training) test mode. Run the classifier and observe the results shown in the “Classifier output” window.  
- How many instances used for the training? How many for the test?



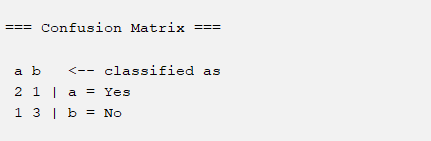
Test mode: split 66.0% train, remainder test  
- Does the test set currently used include the four instances of Users #21-24? None  
- How many instances are incorrectly classified?



Incorrectly Classified Instances 2 28.5714 %  
- What is the MAE (mean absolute error) made by the learned DT?



Mean absolute error 0.3036  
- What can you infer from the information shown in the Confusion Matrix?

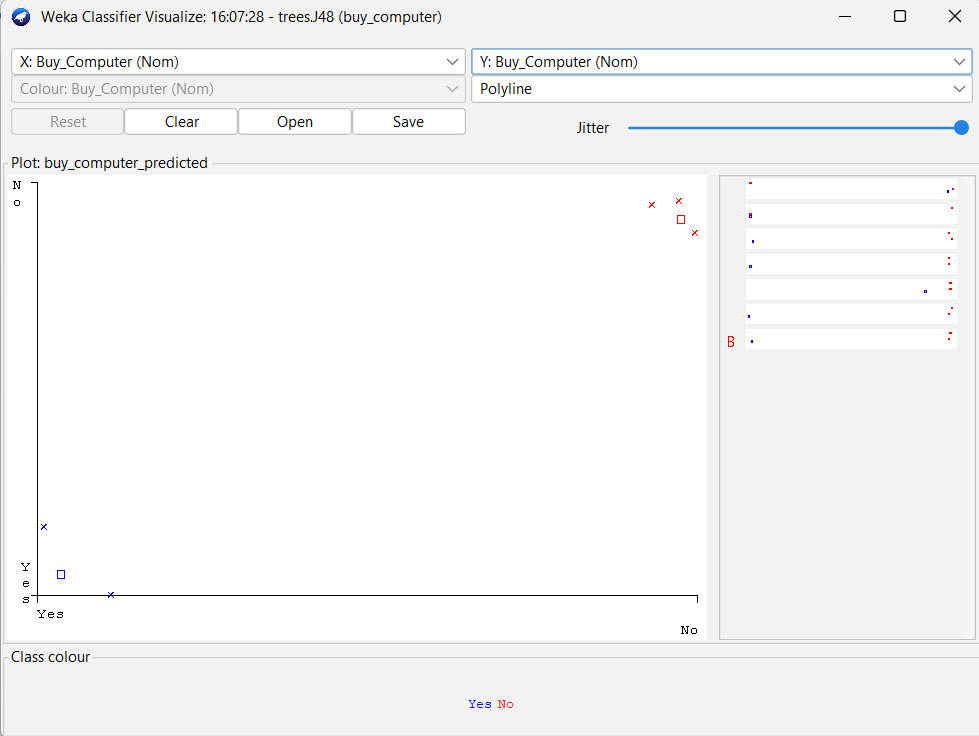


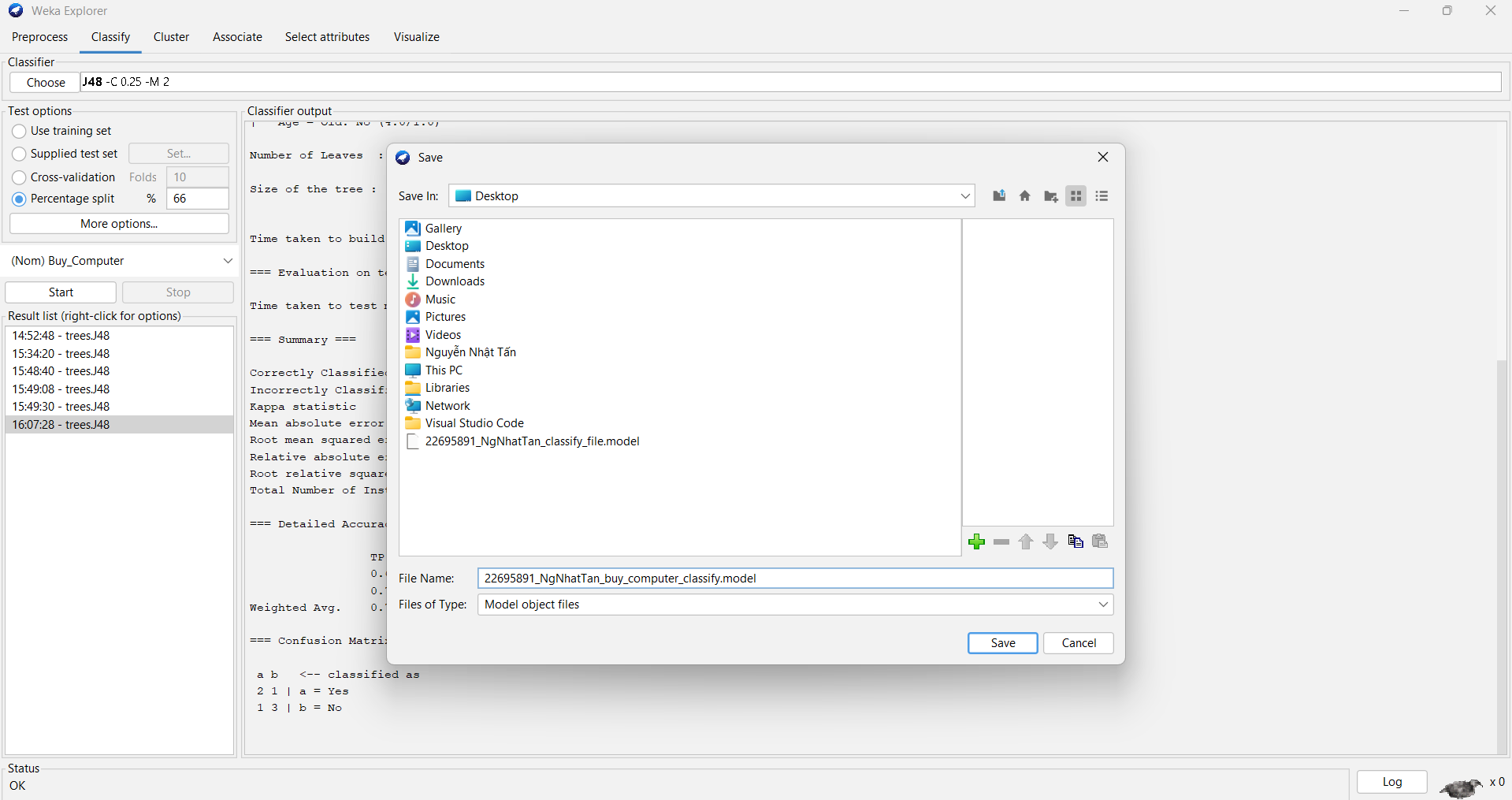
=== Confusion Matrix ===

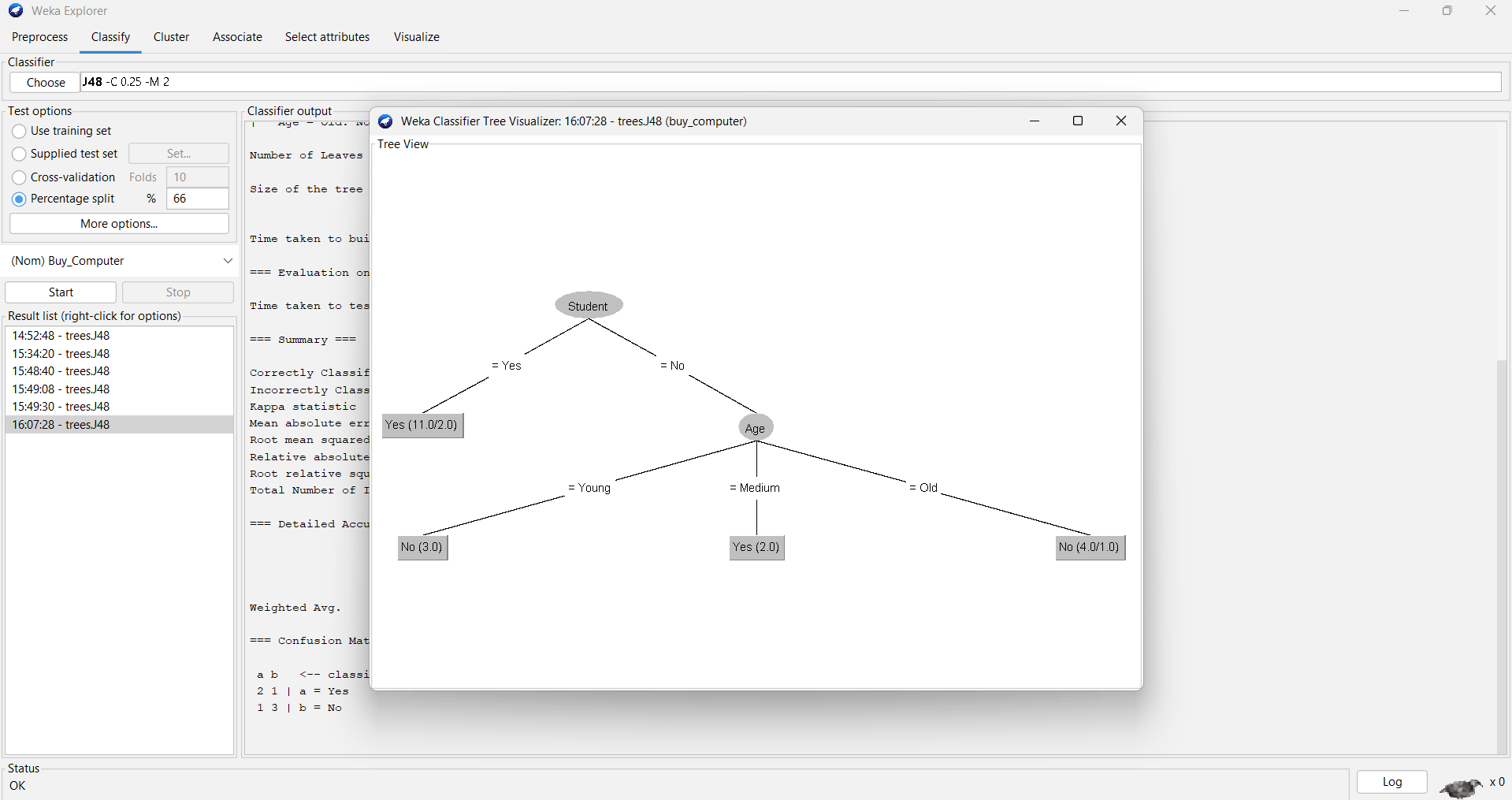
a b <-- classified as

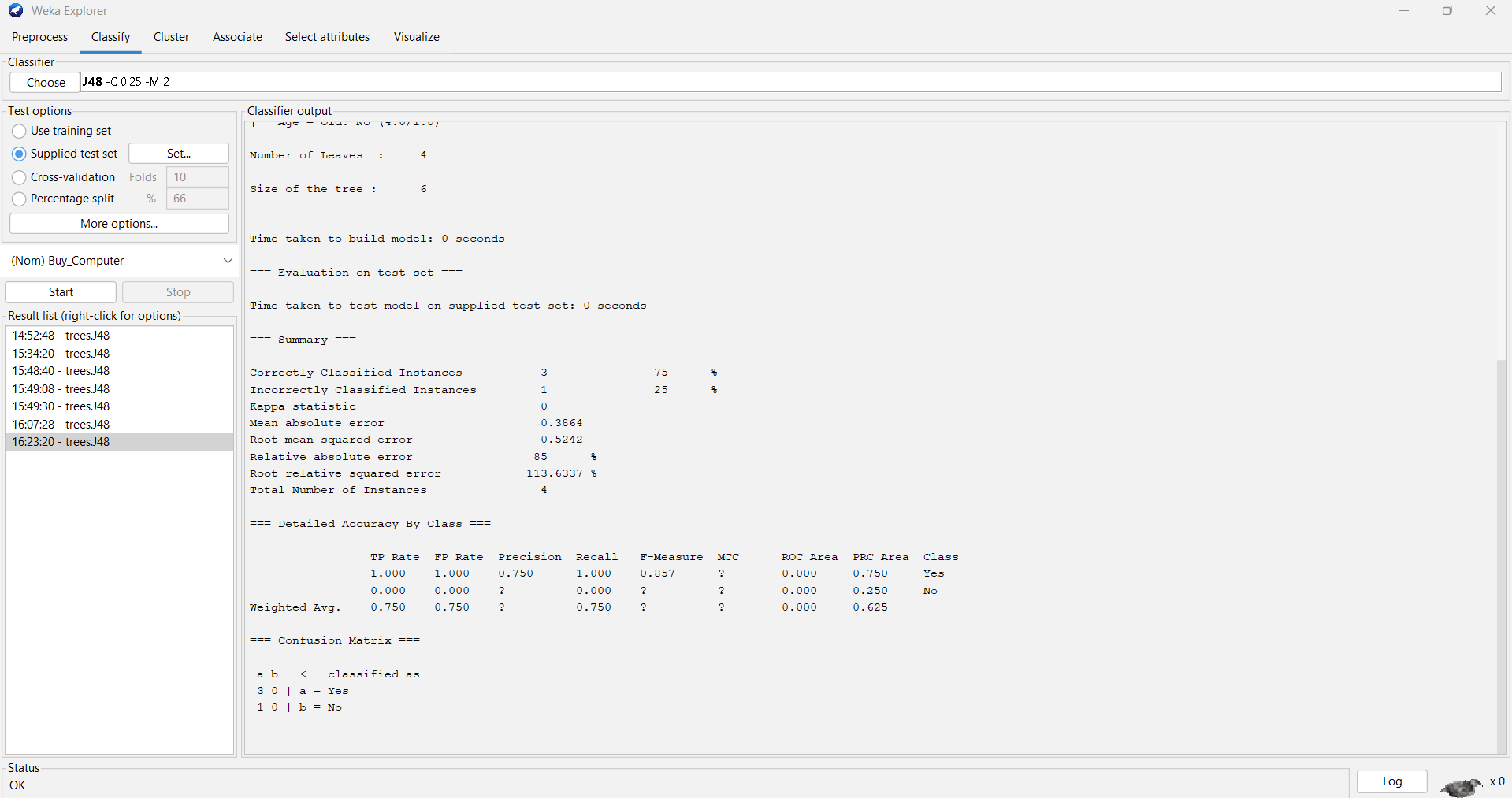
2 1 | a = Yes

1 3 | b = No  
- Visualize the errors made by the learned DT. In the plot, how can you differentiate between the correctly and incorrectly classified instances? In the plot, how can you see the detailed information of an incorrectly classified instance?

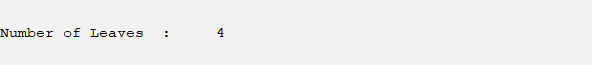
  
- How can you save the learned DT to a file?

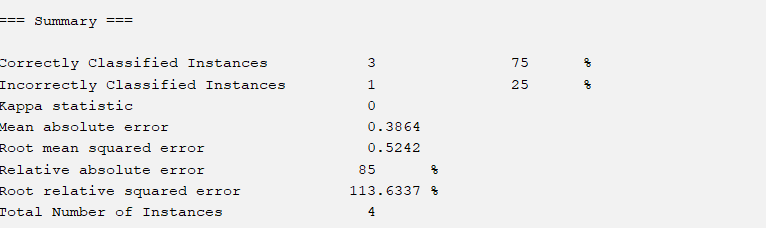
  
- How can you visualize the structure of the learned DT?

  
- In the “Test options” panel select the “Supplied test set” option. Activate the nearby “Set...” button and locate the “buy\_comp\_extra.arff” file. Run the classifier and observe the results shown in the “Classifier output” window.

  
- How many instances used for the training? How many for the test?

  
- Does the test set currently used include the four examples (i.e., Users #21-24)?

  
- In the “Classifier output” window, where you can find the information that says for which of the four users (i.e., Users #21-24) the learned DT predicts correctly and for which others it predicts incorrectly?



Correctly Classified Instances 3 75 %

Incorrectly Classified Instances 1 25 %  
- What is the MAE (mean absolute error) made by the learned DT?



Mean absolute error 0.3864  
**Exercise 3: Using Weka Knowledge Flow to construct a decision tree for the  
weather.nominal.arff with the J-48 algorithm (or the other classification model) for the above exercises**

**Exercise 4: Using Java programming language to construct a decision tree with the J-48 algorithm (or the other classification model) for the above exercises**