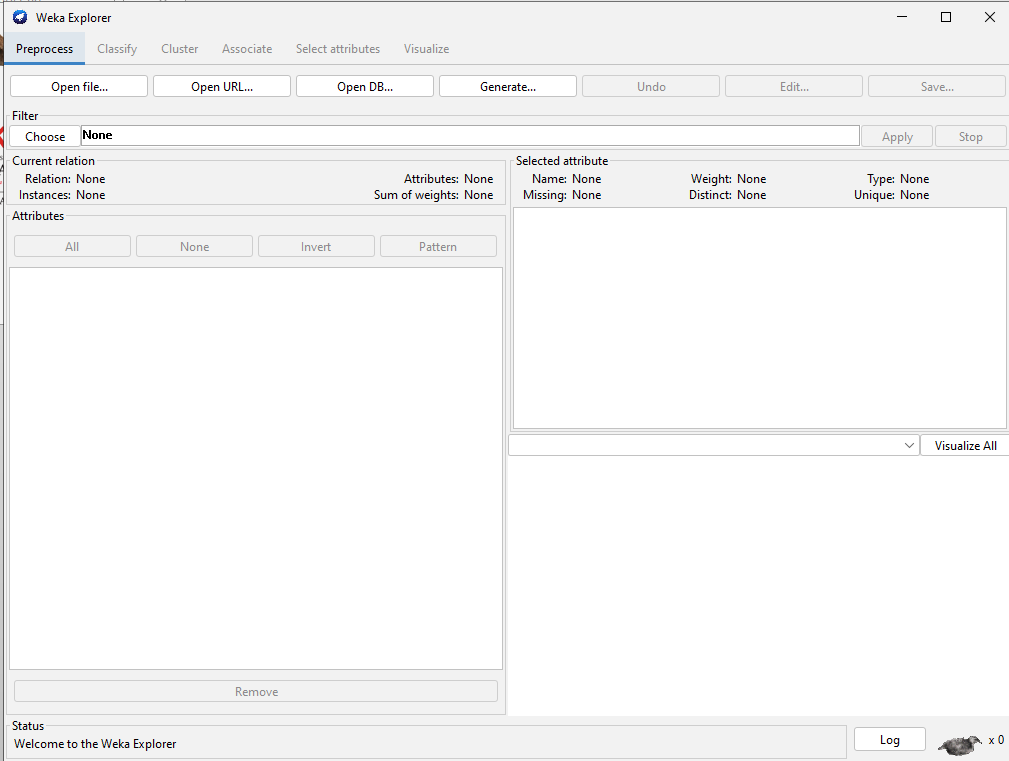
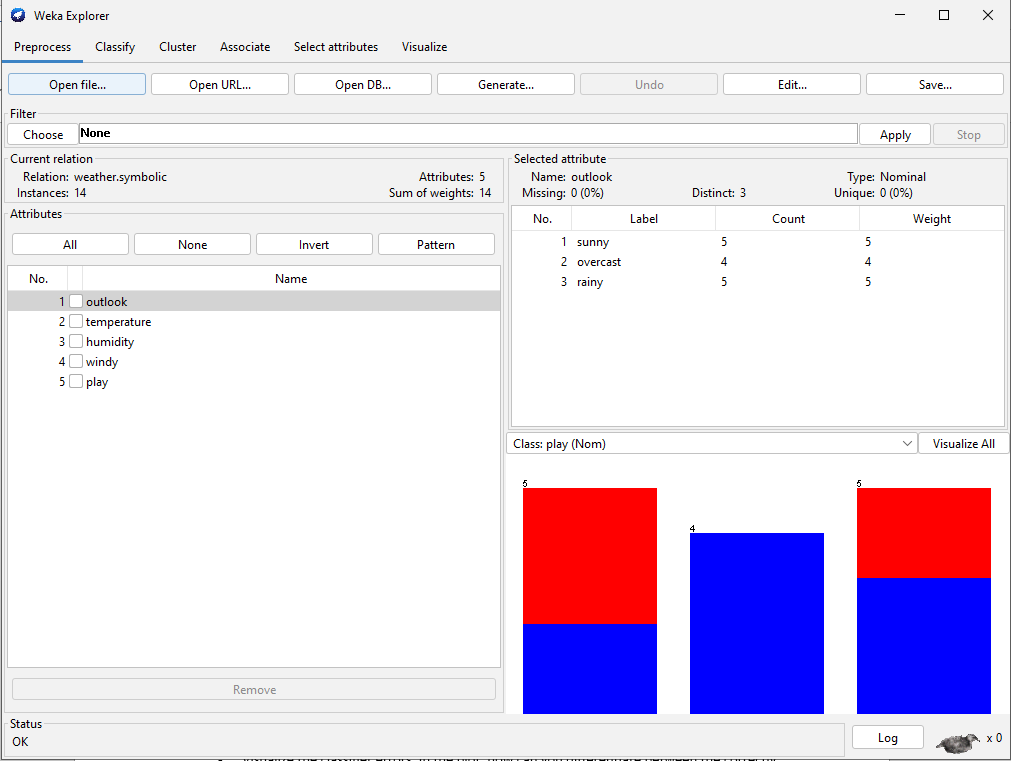
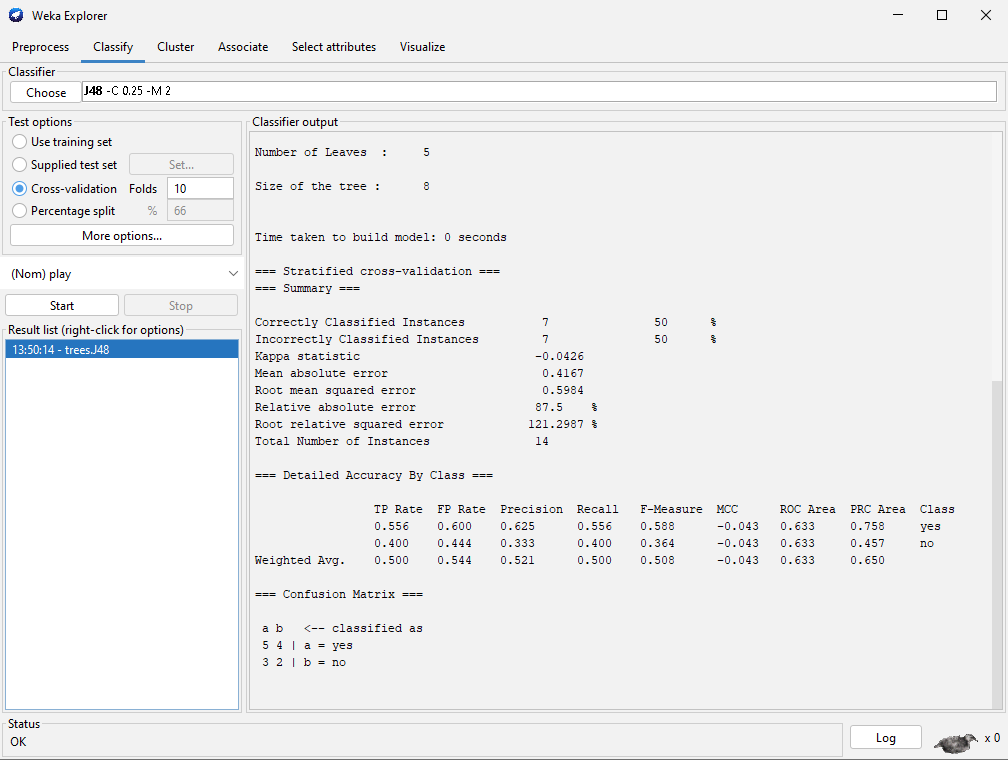
LAB 6 -7: CLASSIFICATION DATA

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. 
   1. Run the classifier and observe the results shown in the “Classifier output” window.
      * How many instances are incorrectly classified?

Các trường hợp được phân loại không chính xác: 7 trong số 14 trường hợp, chiếm 50% số trường hợp.

* + - What is the MAE (mean absolute error) made by the classifier?

Sai số tuyệt đối trung bình (MAE): 0,4167, biểu thị sai số trung bình do bộ phân loại gây ra.

* + - What can you infer from the information shown in the Confusion Matrix?

a b <-- classified as

5 4 | a = yes

3 2 | b = no

True Positives (TP) cho 'có': 5 (phân loại đúng là 'có')

False Positives (FP) cho 'có': 3 (phân loại sai là 'có')

True Negatives (TN) cho 'không': 2 (phân loại đúng là 'không')

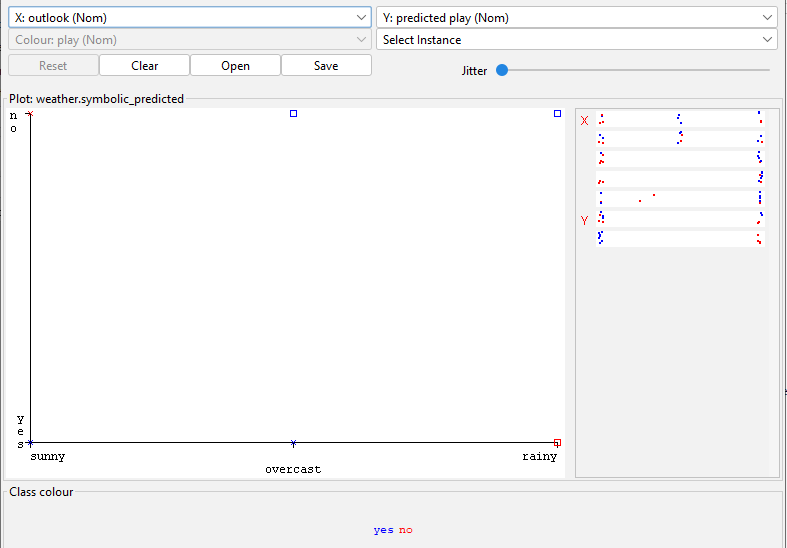
False Negatives (FN) cho 'không': 4 (phân loại sai là 'không') Suy luận:

Bộ phân loại không thực sự chính xác vì nó chỉ phân loại đúng 50% các trường hợp.

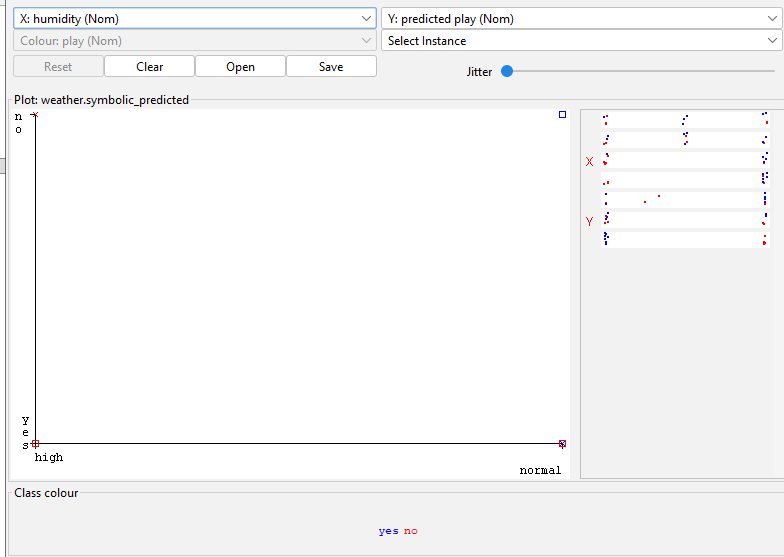
Nó gặp khó khăn trong việc phân biệt giữa các lớp "có" và "không".

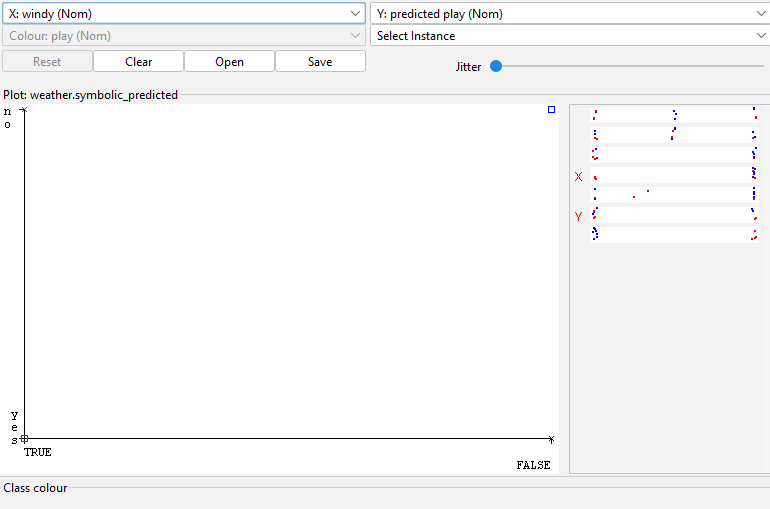
Tỷ lệ dương tính thực (TPR) là 55,6% đối với lớp "có" và chỉ 40% đối với lớp "không", cho thấy hiệu suất kém.

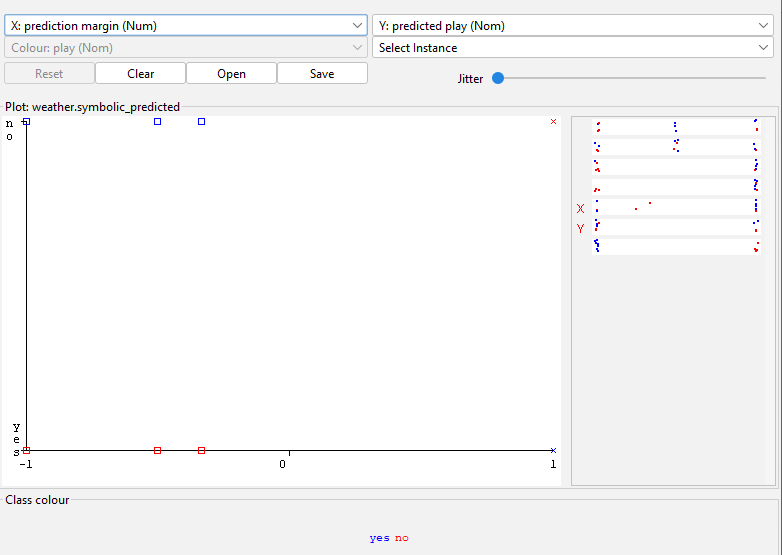
* + - Visualize the classifier errors. In the plot, how can you differentiate between the correctly and incorrectly classified instances?

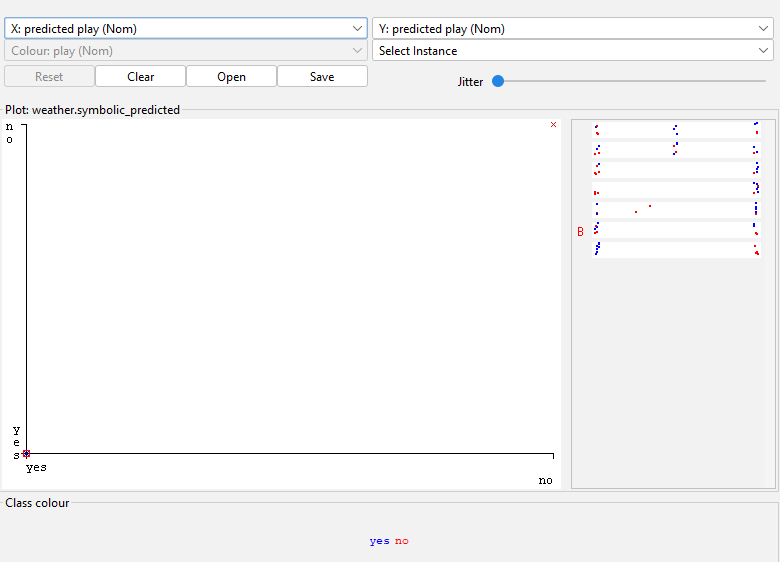
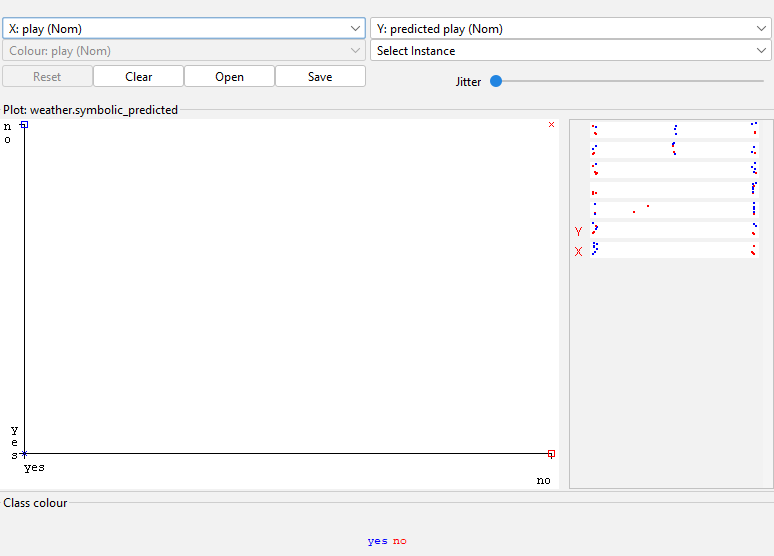










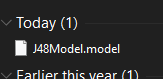
* + - How can you save the learned classifier to a file?

B1. Sau khi xây dựng mô hình, hãy vào "Result list" trong Weka.

B2. Nhấp chuột phải vào mô hình.

B3. Chọn "Save model."

B4. Chọn một vị trí và lưu dưới dạng tệp .model.

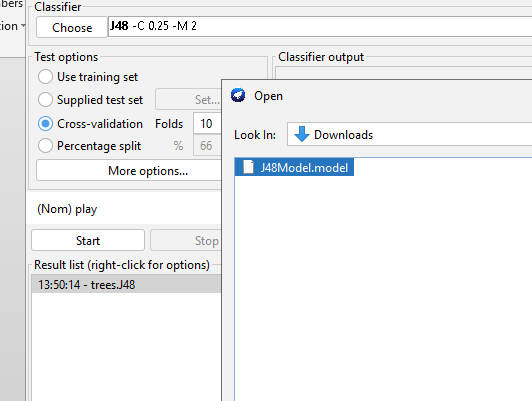


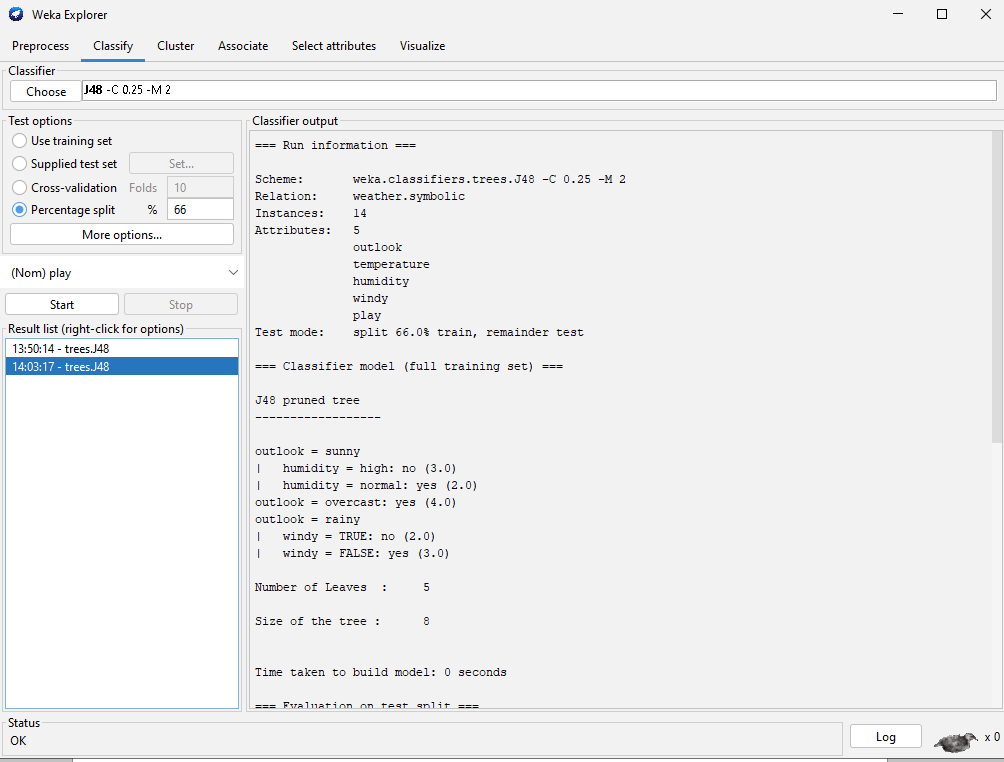
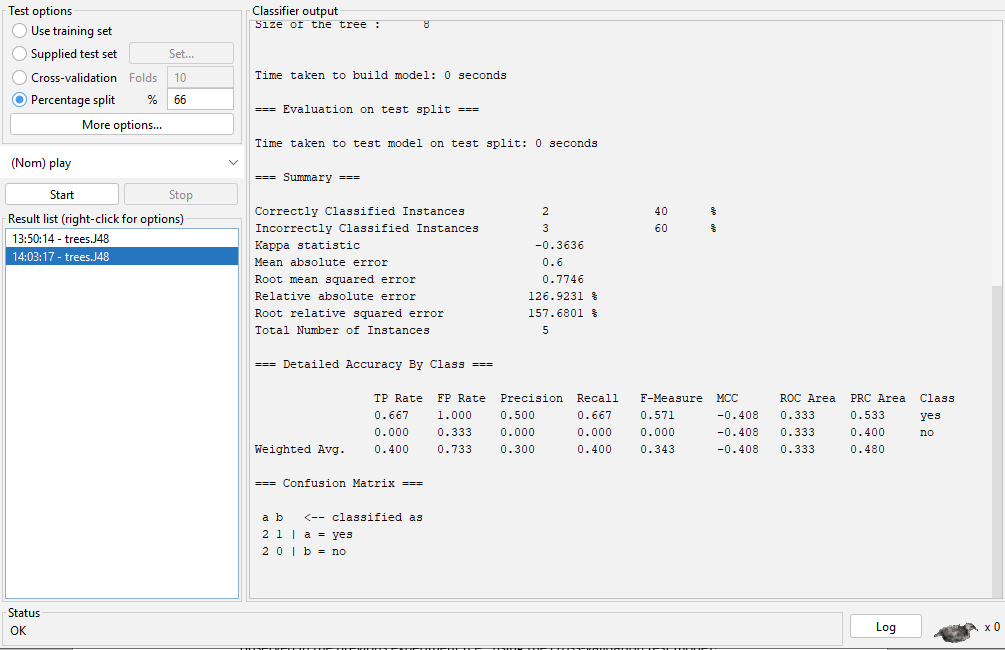
* + - How can you load a learned classifier from a file?

B1. Mở tab "Classify".

B2. Nhấp vào "Load model"

B3. Chọn tệp .model đã lưu để tải tệp đó.

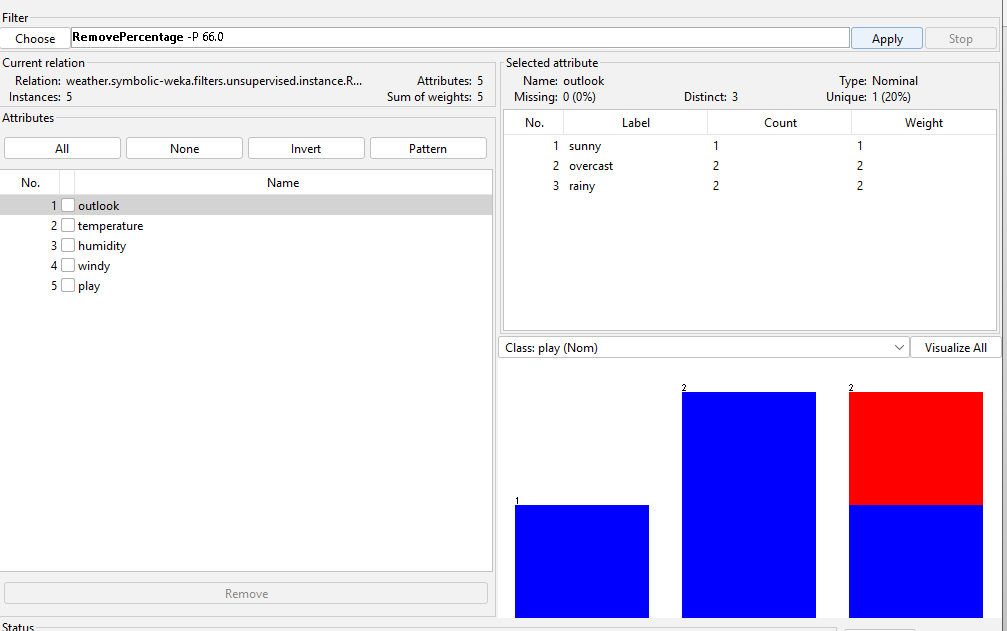


1. Choose the “Percentage split” (66% for training) test mode. Run the J48 classifier and observe the results shown in the “Classifier output” window  
2. Choose the “Use training set” (66% for training) test mode. Run the J48 classifier and observe the results shown in the “Classifier output” window.

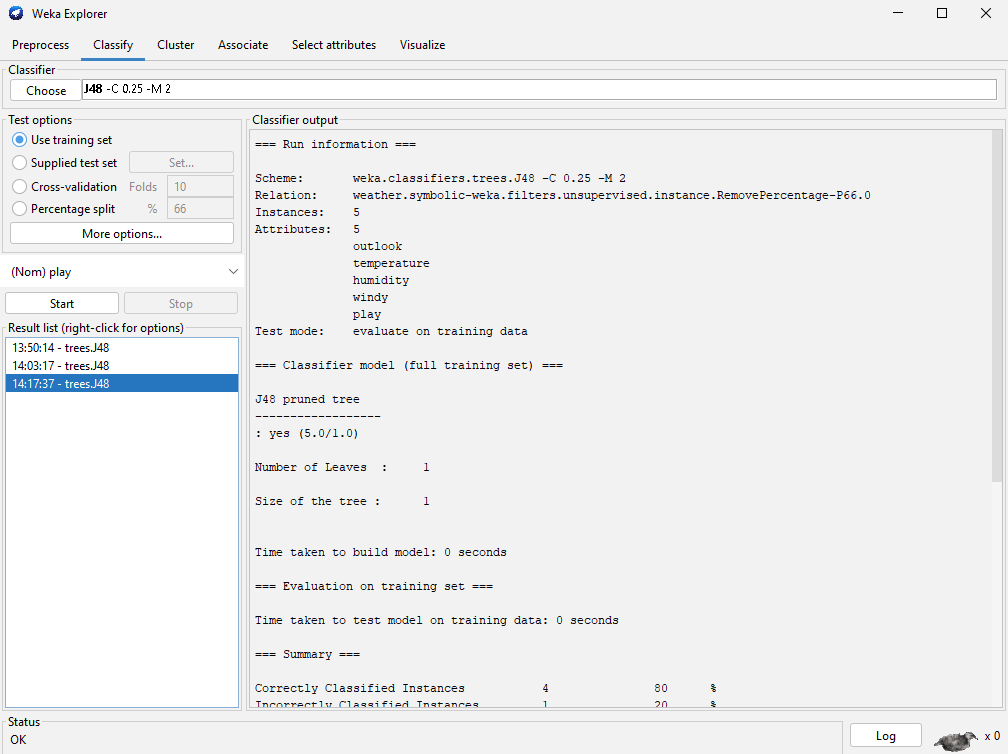
B1: Quay lại Tab ProPressing

B2: Chọn Filter > Unsupervised > Instance

B3: Chọn RemovePercentage cài đặt percentage là 66%



Thực hiện use training set



* + - 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)?

Các trường hợp phân loại không chính xác: 1 trường hợp trong số 5 (20%).

Tại sao con số này nhỏ hơn so với trong xác thực chéo:

Bộ dữ liệu nhỏ hơn và trong trường hợp này, mô hình được đào tạo và thử nghiệm trên cùng một bộ đào tạo (chỉ bao gồm 5 trường hợp sau khi chia tách).

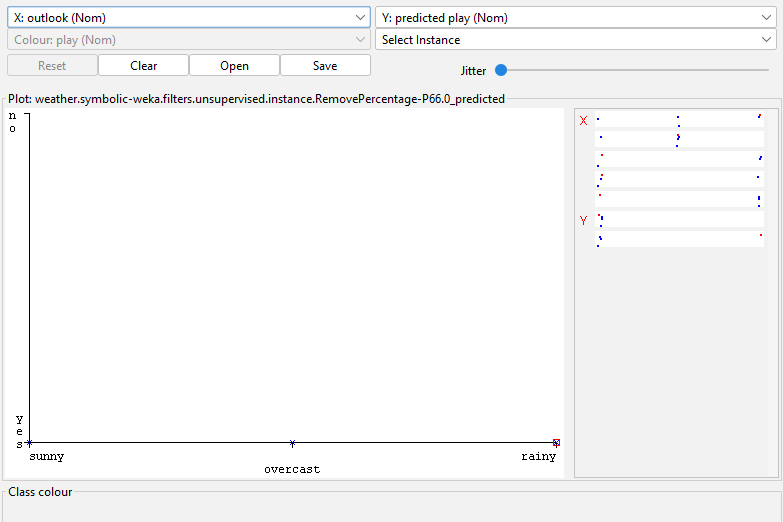
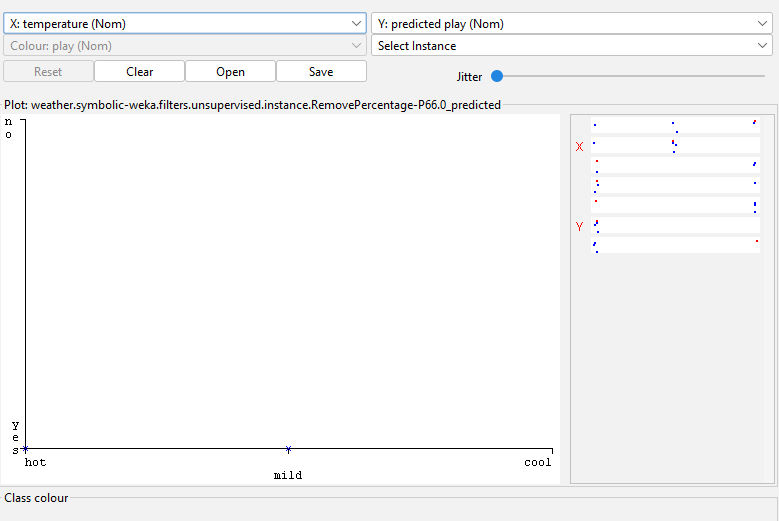
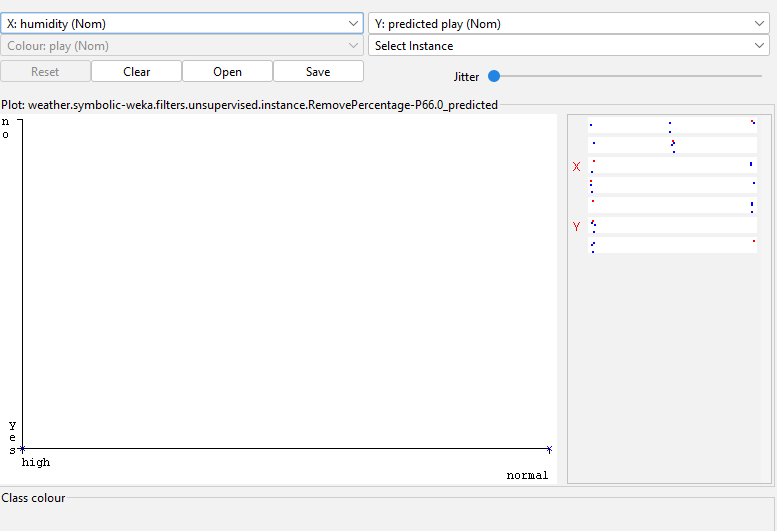
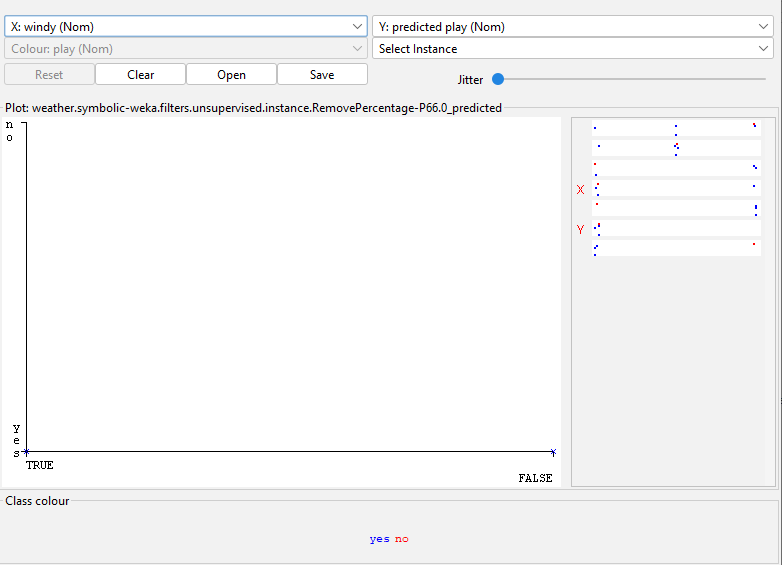
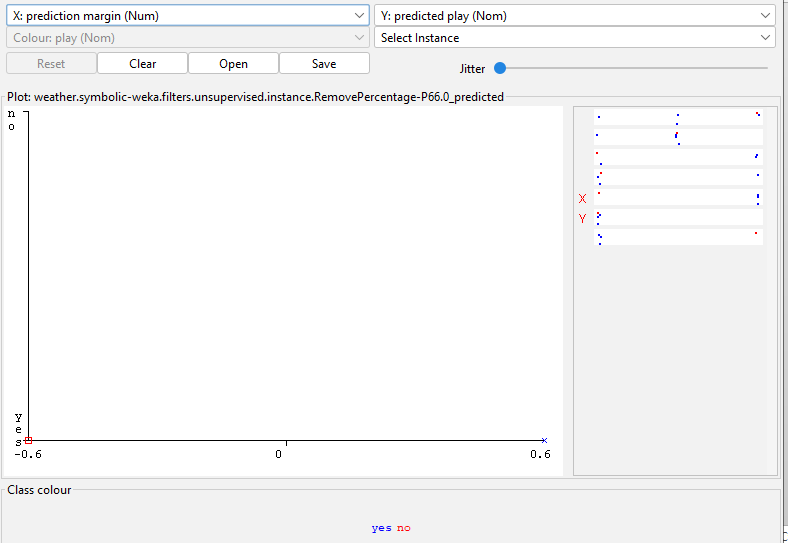
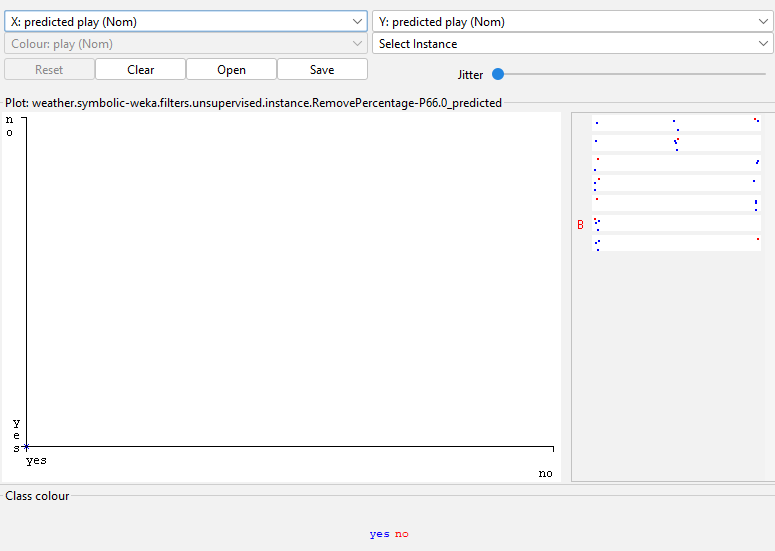
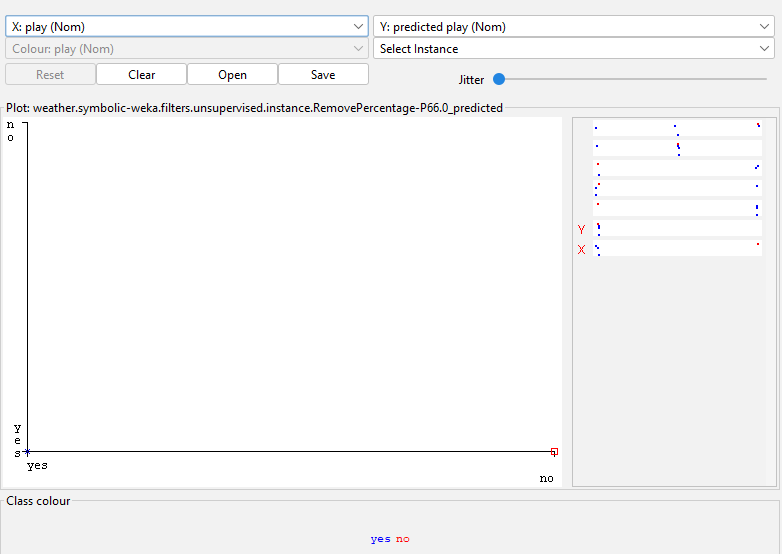
Điều này có xu hướng dẫn đến hiệu suất tốt hơn trên dữ liệu đào tạo, thường là do quá khớp. Mô hình được đào tạo cụ thể trên dữ liệu này và có thể đã ghi nhớ các mẫu thay vì khái quát hóa chúng.

Trong xác thực chéo, mô hình được thử nghiệm trên dữ liệu chưa biết, thường phản ánh khả năng khái quát hóa của mô hình đối với các trường hợp mới, thường dẫn đến nhiều lỗi phân loại hơn.

* + - What is the MAE made by the classifier?

Điều này có nghĩa là sự khác biệt trung bình giữa giá trị dự đoán và giá trị thực tế là 0,32 trên thang điểm từ 0 đến 1.

* + - Visualize the classifier errors to see the detailed information

1. 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 | 0 | 0.32 | 0.4 | |  | | --- |   86.1538% | 97.7802% |
| Cross-validation | -0.0426 | 0.4167 | 0.5984 | 87.5% | 121.2987% |
| Percenttagesplit | 0.363 | 0.6 | 0.7746 | 126.9231% | 157.6801% |

1. Compare the performance of other classification approaches in this dataset:

| **Classifier** | **Kappa Statistic** | **Mean Absolute Error** | **Root Mean Squared Error** | **Relative Absolute Error (%)** | **Root Relative Squared Error (%)** | **Accuracy (% Correct)** |
| --- | --- | --- | --- | --- | --- | --- |

| **J48** | 0 | 0.32 | 0.4 | 86.1538 | 97.7802 | 80% |
| --- | --- | --- | --- | --- | --- | --- |

| **OneR** | 0 | 0.2 | 0.4472 | 53.8462 | 109.3216 | 80% |
| --- | --- | --- | --- | --- | --- | --- |

| **Naive Bayes** | 1 | 0.227 | 0.2569 | 61.1199 | 62.8035 | 100% |
| --- | --- | --- | --- | --- | --- | --- |

Naive Bayes vượt trội hơn cả J48 và OneR trên tập dữ liệu nhỏ này, đạt được phân loại hoàn hảo (độ chính xác 100%) với lỗi thấp nhất (RMSE và RAE).

OneR có lợi thế nhỏ về MAE và RAE nhưng lại tụt hậu về RMSE và RRSE.

J48, mặc dù phân loại chính xác 80% các trường hợp, nhưng lại có lỗi cao hơn (MAE, RAE) so với OneR và Naive Bayes, cho thấy dự đoán của nó kém tin cậy hơn đối với tập dữ liệu này.

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}.

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

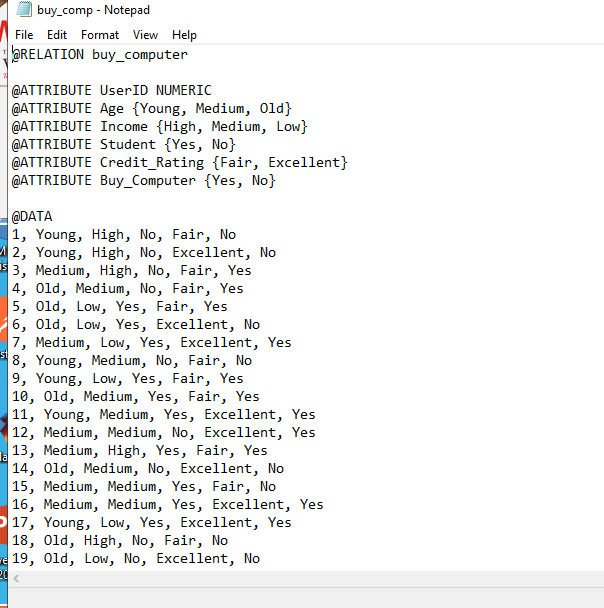
We want to predict, for each of the following users, if s/he will buy a computer or not.

- User #21. A young student with medium income and fair credit rating.

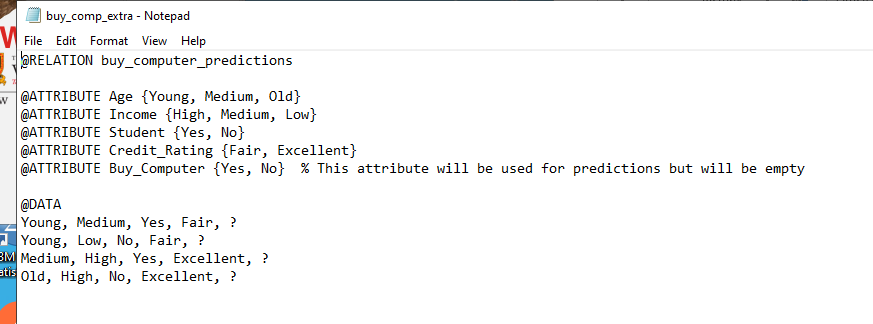
- User #22. A young non-student with low income and fair credit rating.

- User #23. A medium student with high income and excellent credit rating.

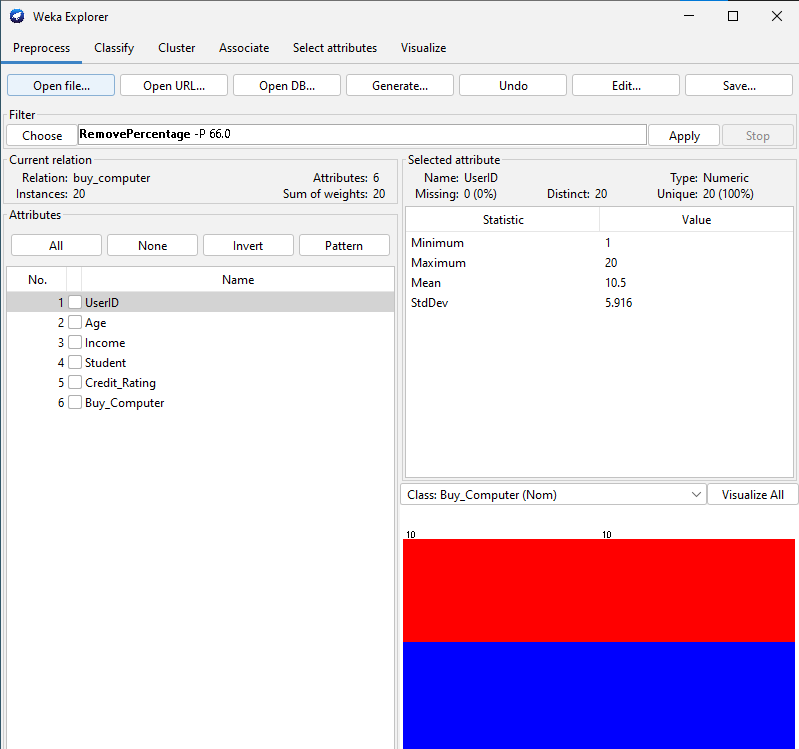
- User #24. An old non-student with high income and excellent credit rating.

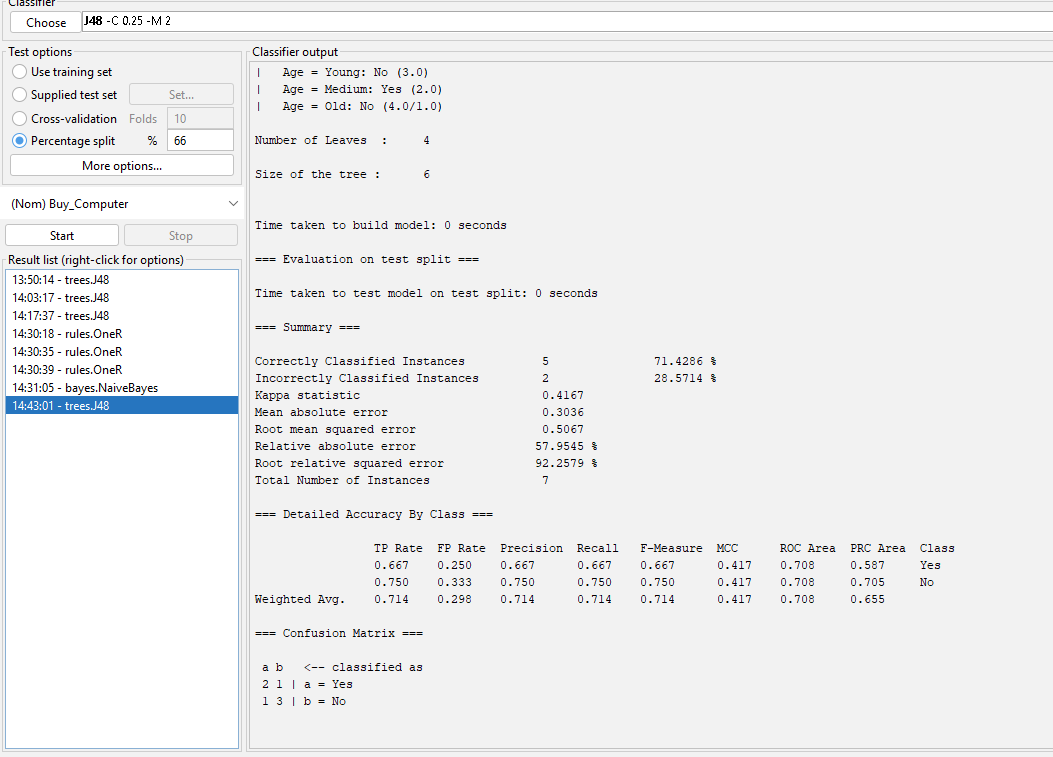
- Create the dataset containing 20 examples (i.e., Users #1-20) into the ARFF format and save it in the “buy\_comp.arff” file. 

* 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?

Training Instances: 20

Testing Instances: 7

* Does the test set currently used include the four instances of Users #21-24?

Các trường hợp cho Người dùng #21-24 không phải là một phần của tập kiểm tra hiện tại mà bạn đã sử dụng với bộ phân loại J48. Bạn sẽ cần tải chúng vào tập dữ liệu mới hoặc tệp ARFF được thiết kế riêng để đưa ra dự đoán bằng mô hình đã đào tạo của bạn.

* How many instances are incorrectly classified?

Total Test Instances: 7

Incorrectly Classified Instances: 2

Accuracy: 71.43% (which means 5 instances were correctly classified).

- What is the MAE (mean absolute error) made by the learned DT?

Sai số tuyệt đối trung bình (MAE) do cây quyết định J48 đã học tạo ra là 0,3036.

- What can you infer from the information shown in the Confusion Matrix?

Ma trận nhầm lẫn cung cấp những hiểu biết sau:

Dự đoán Có Không

Có 2 1

Không 1 3

True Positives (TP): 2 trường hợp được phân loại đúng là "Có".

False Positives (FP): 1 trường hợp được phân loại sai là "Có" (thực tế là "Không").

True Negatives (TN): 3 trường hợp được phân loại đúng là "Không".

False Negatives (FN): 1 trường hợp được phân loại sai là "Không" (thực tế là "Có").

Suy luận:

Mô hình thực hiện tốt hơn trong việc dự đoán "Không" so với "Có", điều này được phản ánh trong số lượng true negatives cao hơn so với true positives.

Thống kê Kappa (0,4167) cho thấy sự đồng thuận vừa phải giữa phân loại dự đoán và phân loại thực tế, cho thấy mô hình có chỗ để cải thiện.

- 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?

Để lưu mô hình cây quyết định đã học vào một tệp trong Weka:

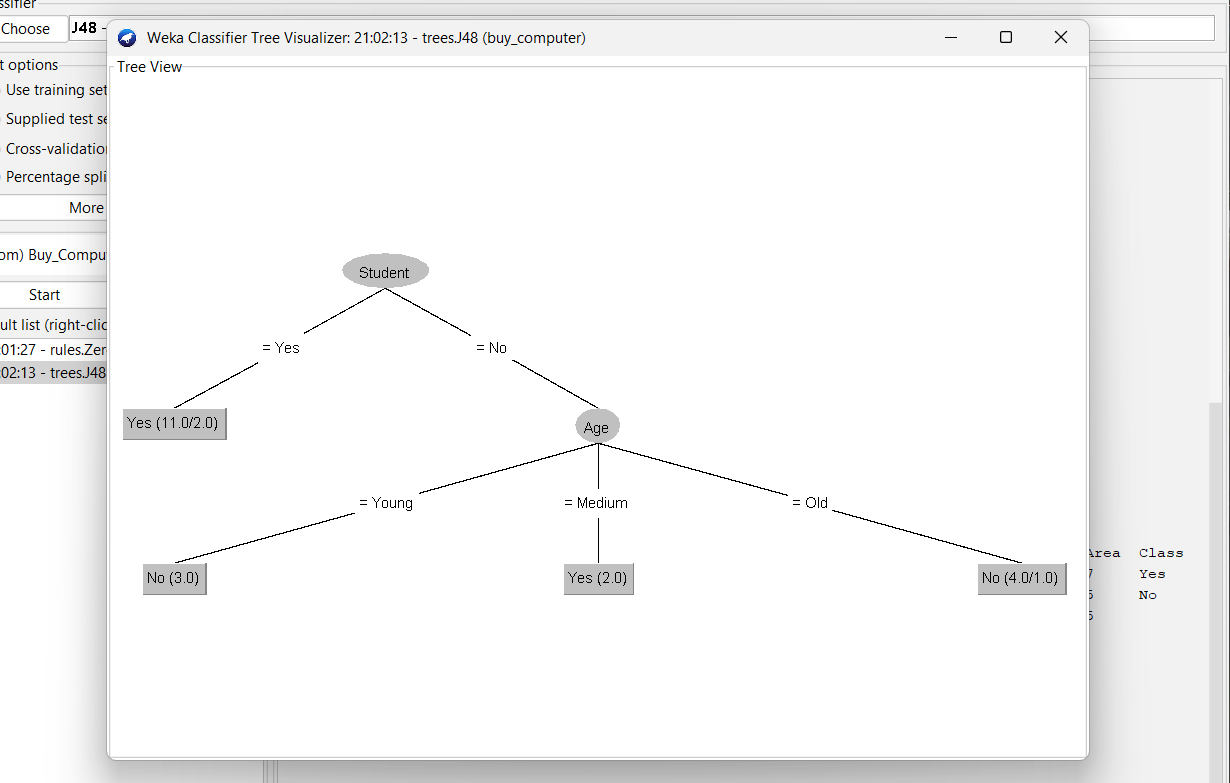
B1. Sau khi xây dựng mô hình, hãy chuyển đến tab Phân loại.

B2. Nhấp vào nút Lưu mô hình (thường trông giống như biểu tượng đĩa).

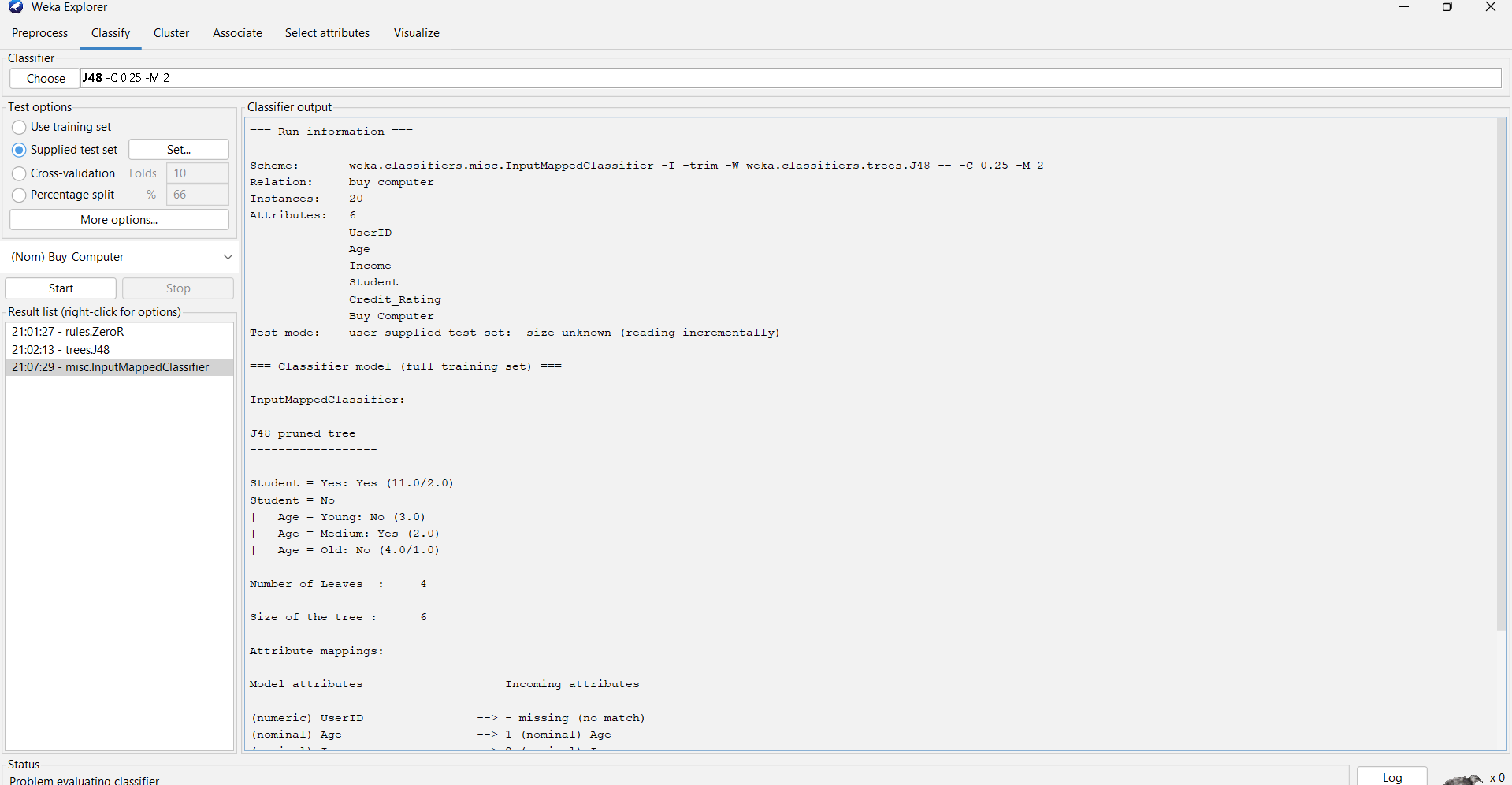
B3. Chọn tên tệp và vị trí, sau đó lưu tệp với phần mở rộng .model hoặc .bin.

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

Sau khi đào tạo mô hình của bạn, hãy nhấp vào nút Visualize tree trong cửa sổ đầu ra của Classifier.

Thao tác này sẽ mở một cửa sổ hiển thị cấu trúc cây, nơi bạn có thể thấy các phần chia, nút và lá ở định dạng đồ họa.

* 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?

Với tập huấn luyện ban đầu có 20 mẫu, tất cả các mẫu này đã được sử dụng cho việc huấn luyện. Bộ kiểm tra (buy\_comp\_extra.arff) chứa các mẫu kiểm tra, bao gồm 4 người dùng (từ #21 đến #24), cùng với các mẫu khác nếu có.

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

Có, nếu tập kiểm tra là buy\_comp\_extra.arff, nó sẽ bao gồm 4 người dùng (#21-24). Bạn có thể xác nhận bằng cách mở file buy\_comp\_extra.arff để kiểm tra xem người dùng #21-24 có nằm trong đó không.

* 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?

Để biết mô hình dự đoán đúng hay sai đối với các người dùng #21-24:

Sau khi chạy mô hình, cuộn xuống trong cửa sổ **"Classifier output"**.

Tìm phần kết quả dự đoán, thường có tiêu đề như **Predicted Results** hoặc **Instances**.

Phần này sẽ hiển thị lớp dự đoán, lớp thực tế và cho biết dự đoán có đúng (+) hay sai (-).

Bạn có thể dựa vào giá trị các thuộc tính như **Age** (Tuổi), **Income** (Thu nhập), **Student** (Sinh viên), v.v. để nhận diện các người dùng #21-24 và xem mô hình dự đoán đúng hay sai.

* What is the MAE (mean absolute error) made by the learned DT?

MAE (Mean Absolute Error) được hiển thị trong phần Summary của cửa sổ Classifier output. Sau khi bạn chạy mô hình với tập kiểm tra (buy\_comp\_extra.arff), MAE sẽ được tính toán và hiển thị trong kết quả tóm tắt.

Ví dụ, trong các lần chạy trước, MAE đã được hiển thị như sau:

Mean absolute error: 0.3036 cho một lần kiểm tra trước đó.

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

