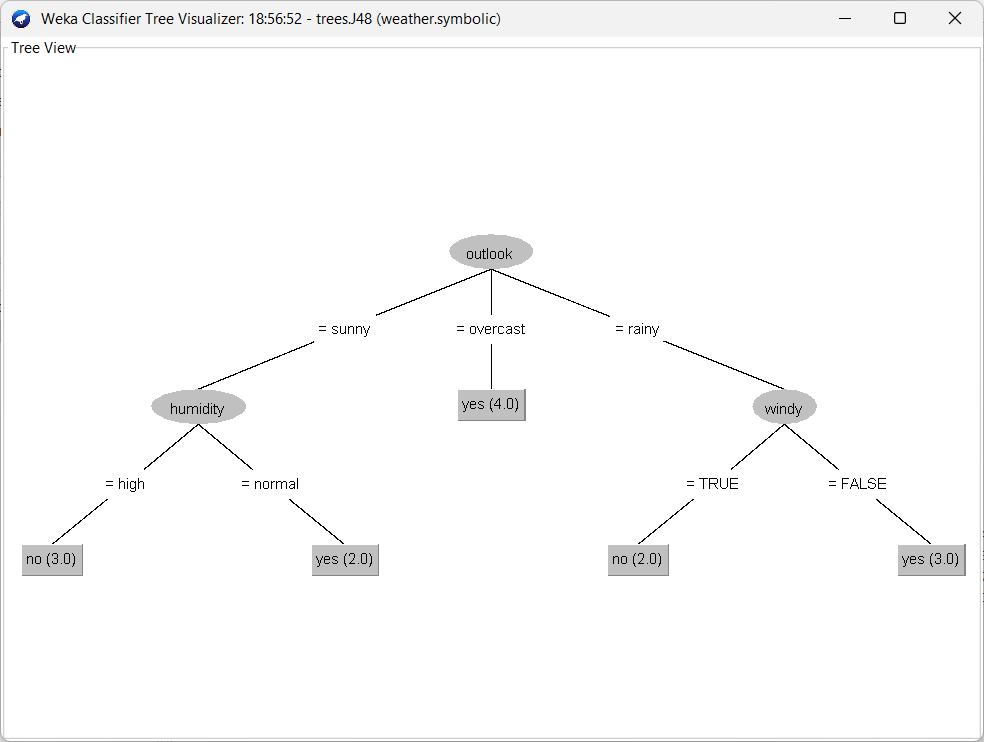
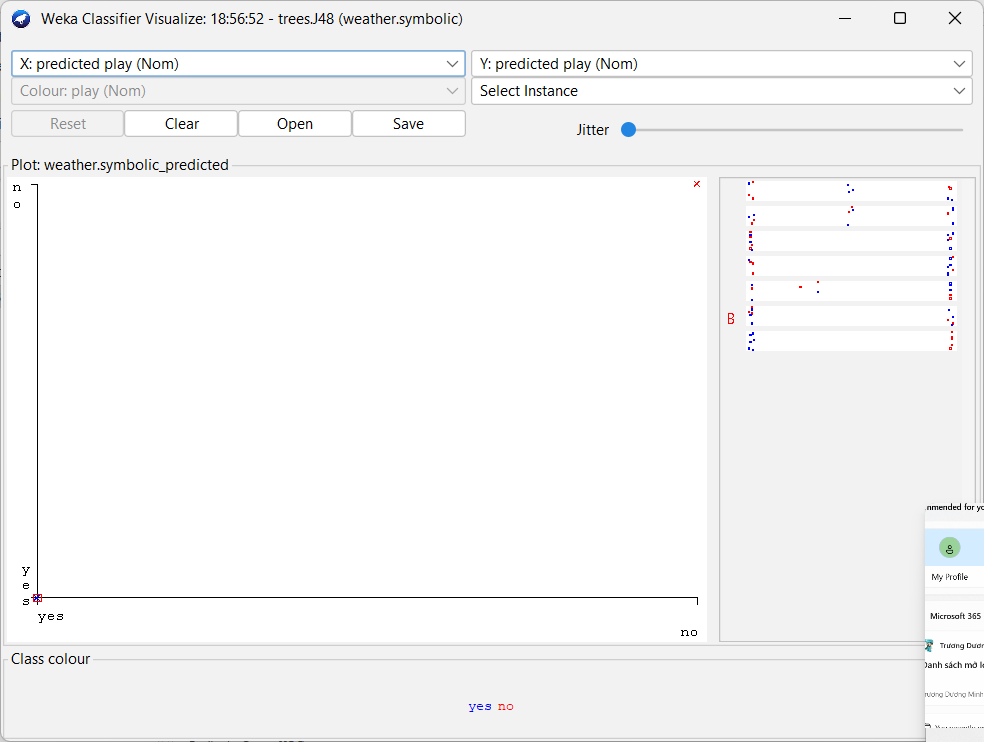
**21083791 – Lê Hoàng Khang – Lab 6-7: : CLASSIFICATION DATA**

**Exercise 1:**



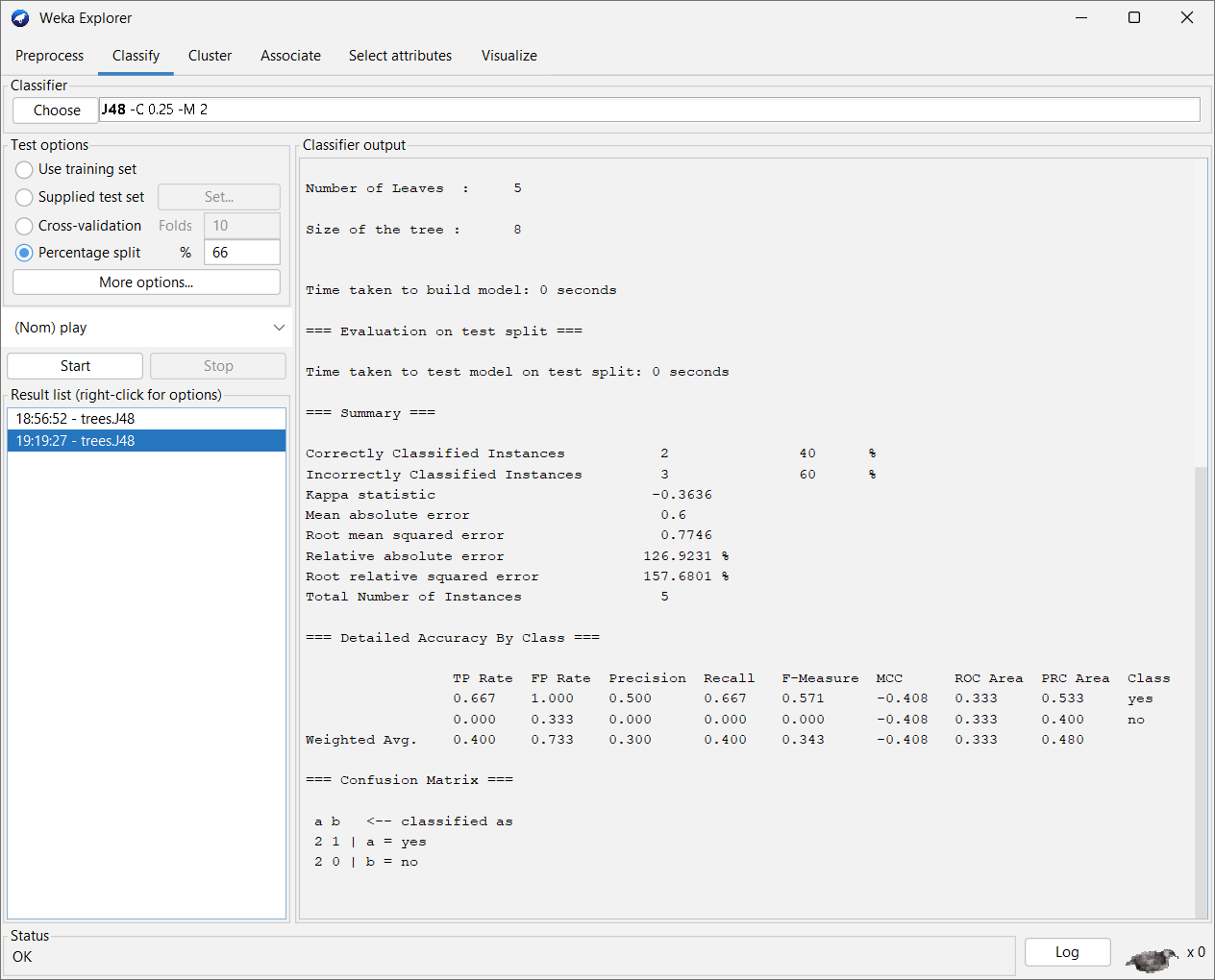
* Có 7 trường hợp được phân loại sai (3 yes, 2 no)
* Mean absolute error(Trung bình lỗi): 0.4167
* Từ thông tin trong Ma trận nhầm lẫn (Confusion Matrix), chúng ta có thể rút ra các kết luận sau: Số lượng dữ liệu được phân loại đúng và sai: Ma trận cho biết số lượng các điểm dữ liệu được phân loại đúng và sai cho mỗi lớp. Trong trường hợp này, có 5 điểm dữ liệu nhãn "yes" được phân loại đúng, 4 điểm nhãn "no" được phân loại đúng. Tuy nhiên, có 3 điểm nhãn "yes" và 2 điểm nhãn "no" bị phân loại sai.
* Visualize the classifier errors. In the plot, how can you differentiate between the correctly and incorrectly classified instances?



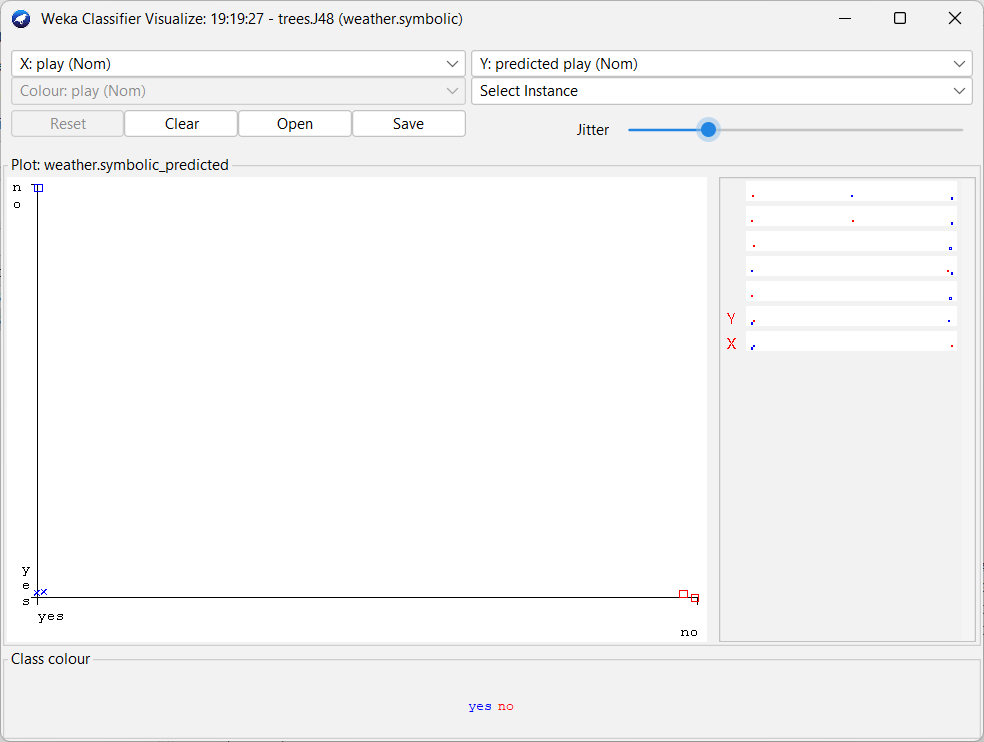
* Để lưu mô hình: Nhấp chuột phải vào Result muốn lưu trong Result List 🡪 Save model

4. Choose the “Percentage split” (66% for training)

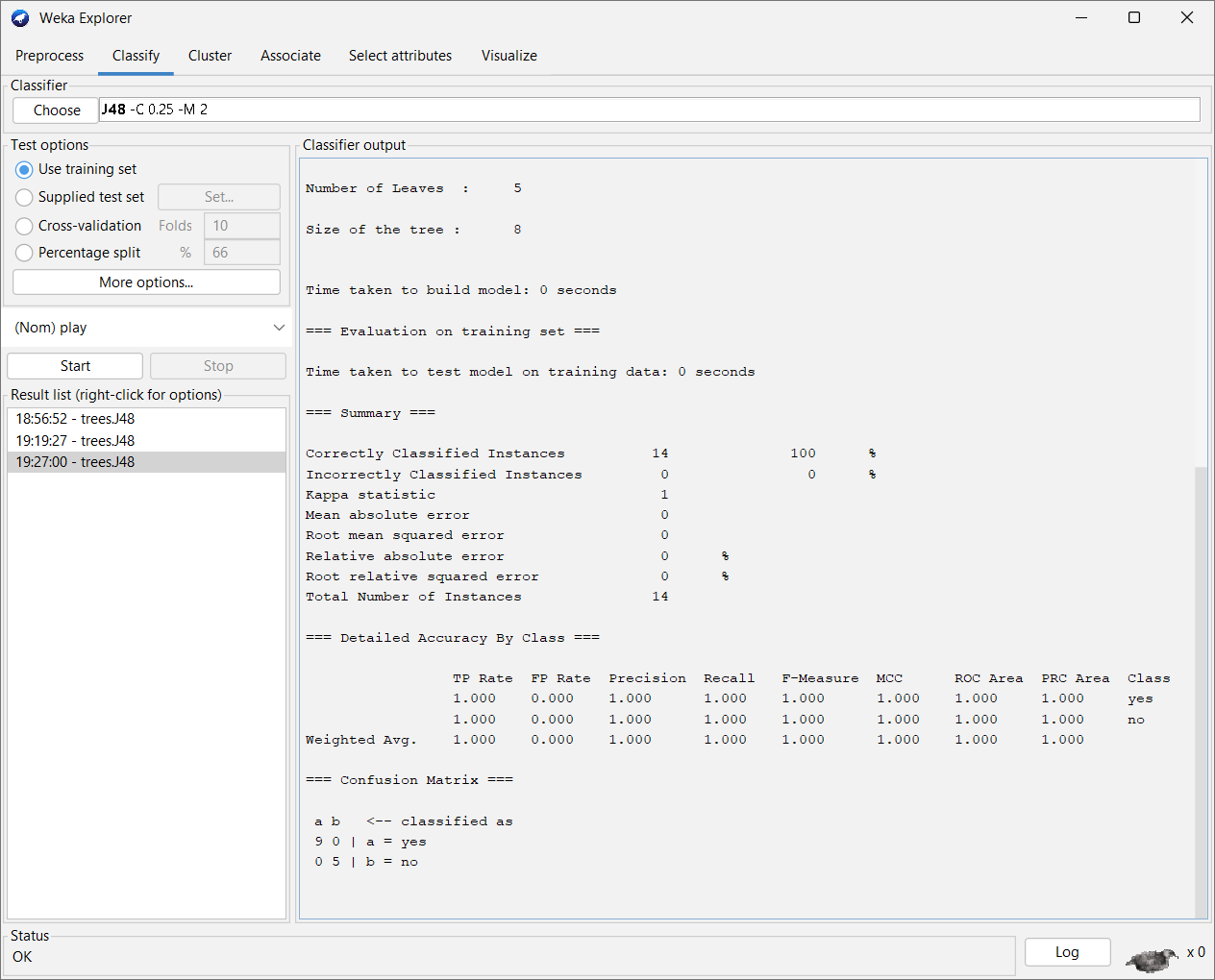
- Kết quả



* Có 2 điểm dữ liệu được phân loại đúng vào lớp "yes", và 0 điểm dữ liệu được phân loại đúng vào lớp "no". Tuy nhiên, có 1 điểm dữ liệu lớp "yes" bị phân loại sai thành lớp "no", và 2 điểm dữ liệu lớp "no" bị phân loại sai.
* Con số nhỏ hơn trong thử nghiệm này có thể là do sự biến động trong dữ liệu, hoặc cũng có thể là do phân chia dữ liệu huấn luyện và kiểm tra dẫn đến một phân phối không đồng đều giữa các lớp hoặc các mẫu không đại diện cho dữ liệu tổng thể.
* MAE = 0.6
* Visualize the classifier errors to see the detailed information.



5. Choose the “Use training set”



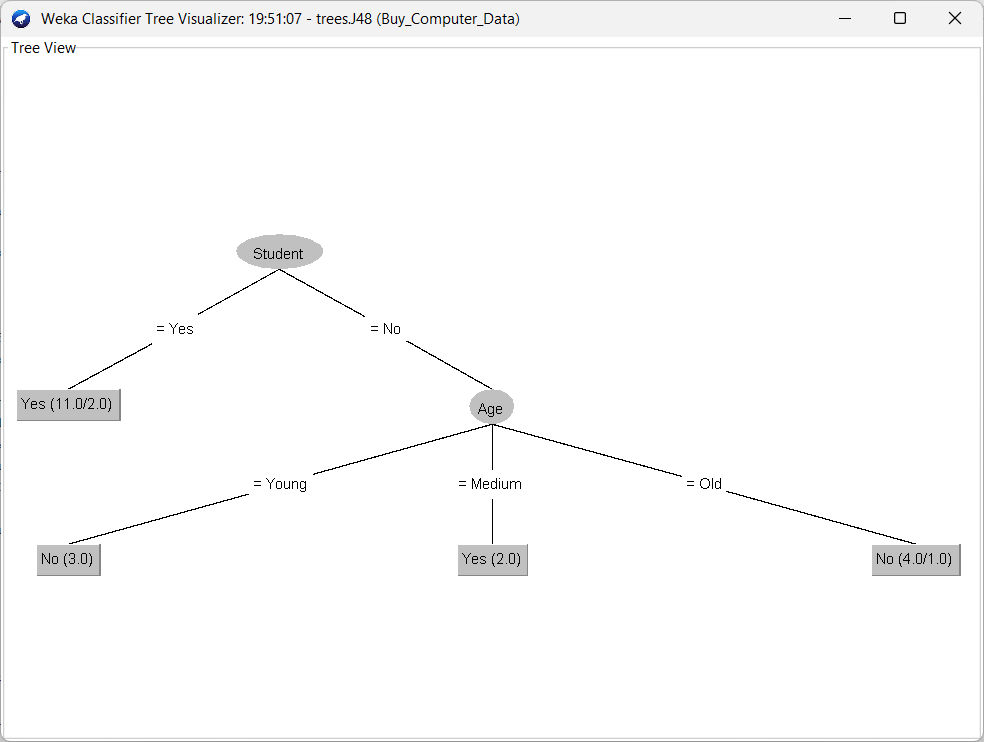
* Tất cả 14 điểm dữ liệu đều được phân loại đúng, đạt tỷ lệ phân loại chính xác là 100%. Không có dữ liệu nào được phân loại sai 0%.
* Lý do chính là trong trường hợp "Use training set", mô hình được đánh giá trên cùng tập dữ liệu được sử dụng để huấn luyện. Do đó, mô hình có khả năng học được từ dữ liệu đó và có thể dự đoán chính xác các nhãn mà không gặp phải dữ liệu mới nào. Tuy nhiên, điều này không phản ánh được khả năng tổng quát hóa của mô hình khi áp dụng cho dữ liệu mới.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Option | Kappa statistic | Mean absolute  error | Root mean squared error | Relative absolute error | Root retaltive squared error |
| User trainning set | 1 | 0 | 0 | 0 | 0 |
| Cross-validation | -0.0426 | 0.4167 | 0.5984 | 87.5% | 121.2987% |
| Percenttagesplit | -0.3636 | 0.6 | 0.7746 | 126.9231% | 157.6801% |

**Exercise 2:**

Load data

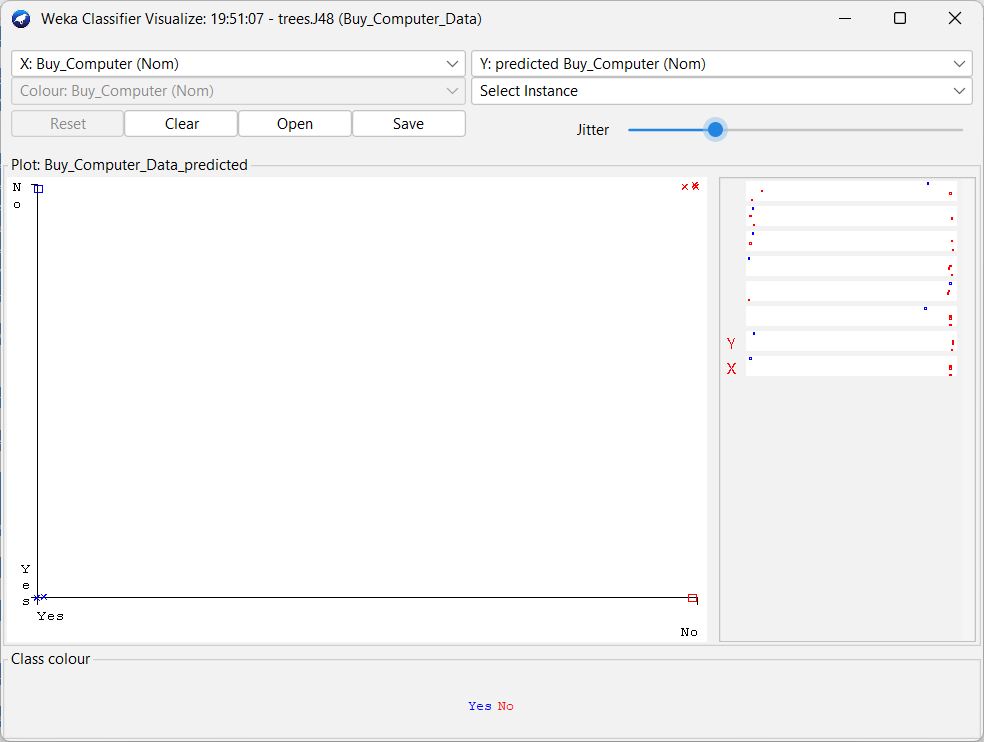
Kết quả khi áp dụng:



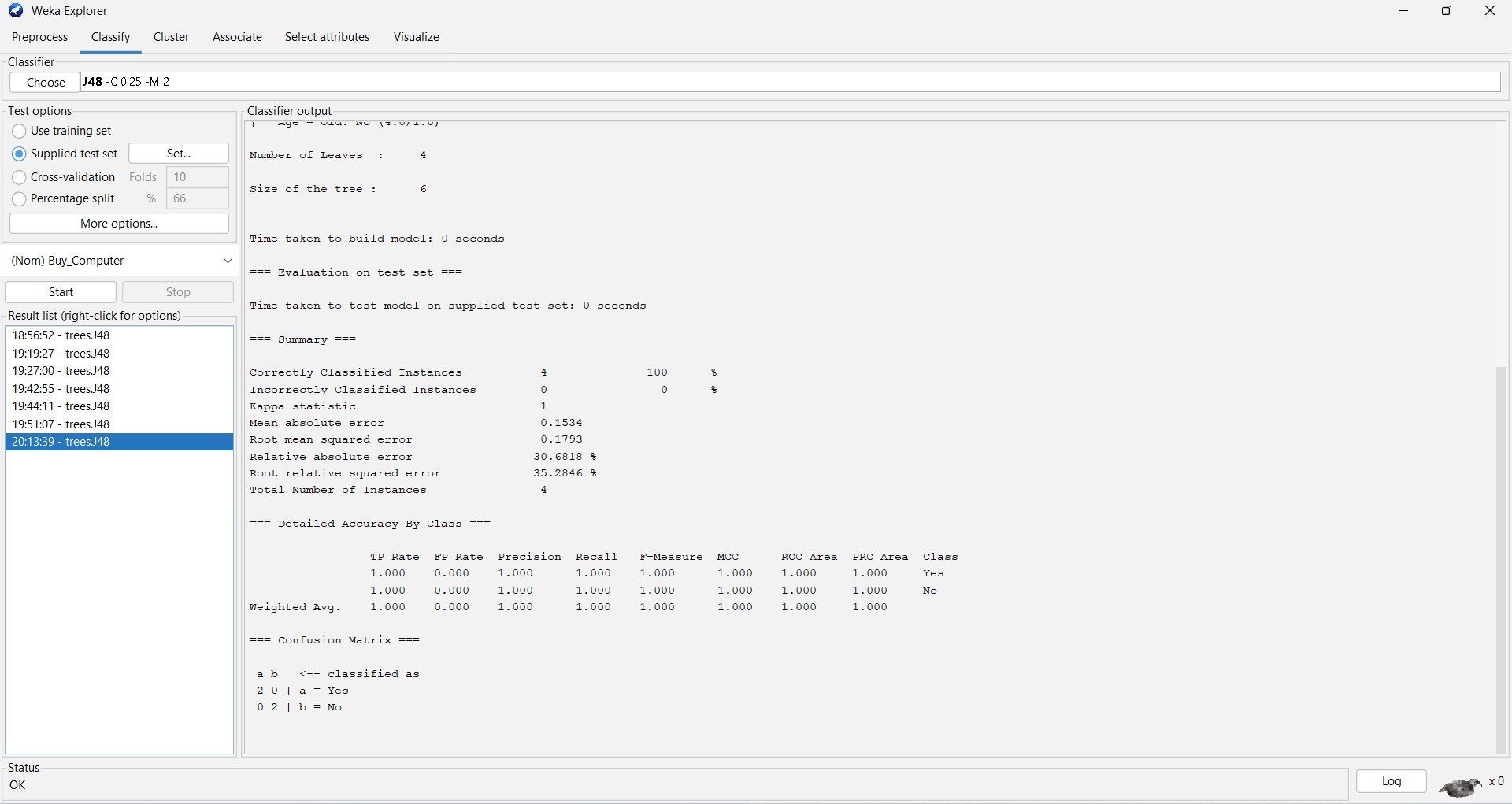
* Dự đoán cho tập dữ liệu mới
  + 21, Young, Medium, Yes, Fair, Yes 🡪 Học sinh 🡪 Yes
  + 22, Young, Low, No, Fair, No 🡪 Không phải học sinh + Trẻ 🡪 No
  + 23, Medium, High, Yes, Excellent, Yes 🡪 Không phải học sinh + Trung bình 🡪 Yes
  + 24, Old, High, No, Excellent, No 🡪 Không phải học sinh + Già 🡪 No

Dựa vào kết quả trên:

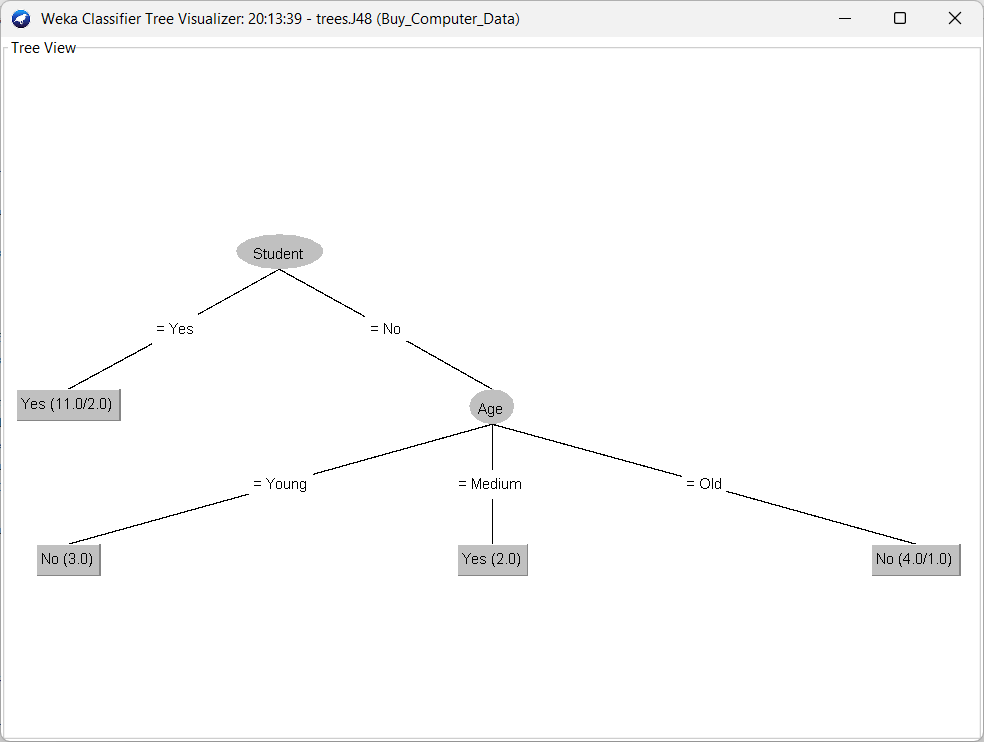
* Có 13 dữ liệu để huấn luyện và 7 dữ liệu để kiểm tra
* Có tổng cộng 2 trường hợp được phân loại sai.
* MAE = 0.3036.
* Ma trận nhầm lẫn cung cấp số lượng các trường hợp được phân loại đúng và sai cho mỗi lớp. Trong trường hợp này, có 2 trường hợp nhãn "Yes" được phân loại sai thành "No", và 1 trường hợp nhãn "No" được phân loại sai thành "Yes".



* Sử dụng “Test options” với file “buy\_comp\_extra.arff”



Cây quyết định



* Có 20 dữ liệu cho trainning và 4 dữ liệu cho test
* Bộ dữ liệu test hiện tại có 4 người dùng từ 21 đến 24
* Trong cửa sổ "Classifier output", thông tin về các trường hợp dự đoán đúng và sai được hiển thị trong phần "Confusion Matrix". Trong trường hợp này, có 2 trường hợp được dự đoán đúng cho mỗi lớp, không có trường hợp nào được dự đoán sai.
* MAE = 0.1534.

DEMO CODE JAVA:

