**Buổi 1. WEKA**

File dữ liệu trong weka

Định dạng ARFF

|  |
| --- |
| % 5. Number of Instances: 150 (50 in each of three classes)  %  % 6. Number of Attributes: 4 numeric, predictive attributes and the class % % 7. Attribute Information:  % 1. sepal length in cm  % 2. sepal width in cm  % 3. petal length in cm  % 4. petal width in cm % 5. class:  % -- Iris Setosa  % -- Iris Versicolour  % -- Iris Virginica  %  % 8. Missing Attribute Values: None  % % Summary Statistics:  % Min Max Mean SD Class  Correlation  % sepal length: 4.3 7.9 5.84 0.83 0.7826  % sepal width: 2.0 4.4 3.05 0.43 -0.4194  % petal length: 1.0 6.9 3.76 1.76 0.9490  (high!)  % petal width: 0.1 2.5 1.20 0.76 0.9565  (high!)  % % 9. Class Distribution: 33.3% for each of 3 classes.  @RELATION iris  @ATTRIBUTE sepallength REAL  @ATTRIBUTE sepalwidth REAL  @ATTRIBUTE petallength REAL  @ATTRIBUTE petalwidth REAL  @ATTRIBUTE class {Iris-setosa,Iris-versicolor,Irisvirginica}  @DATA  5.1,3.5,1.4,0.2,Iris-setosa  4.9,3.0,1.4,0.2,Iris-setosa |

# Explorer

* Preprocess: tải tập dữ liệu, chuyển đổi dữ liệu về dạng mong muốn
* Classify: lựa chọn và chạy các thuật toán phân loại và hồi quy trên tập dữ liệu đã chọn
* Cluster: chạy thuật toán phân cụmh
* Associate: chạy thuật toán kết hợp
* Select Attributes: các thuật toán lựa chọn thuộc tính
* Visualize: trực quan mối quan hệ giữa các thuộc tính

# Experiment

* Setup: thiết kế thực nghiệm
* Run: thực thi thí nghiệm
* Analyse: phân tích kết quả thí nghiệm

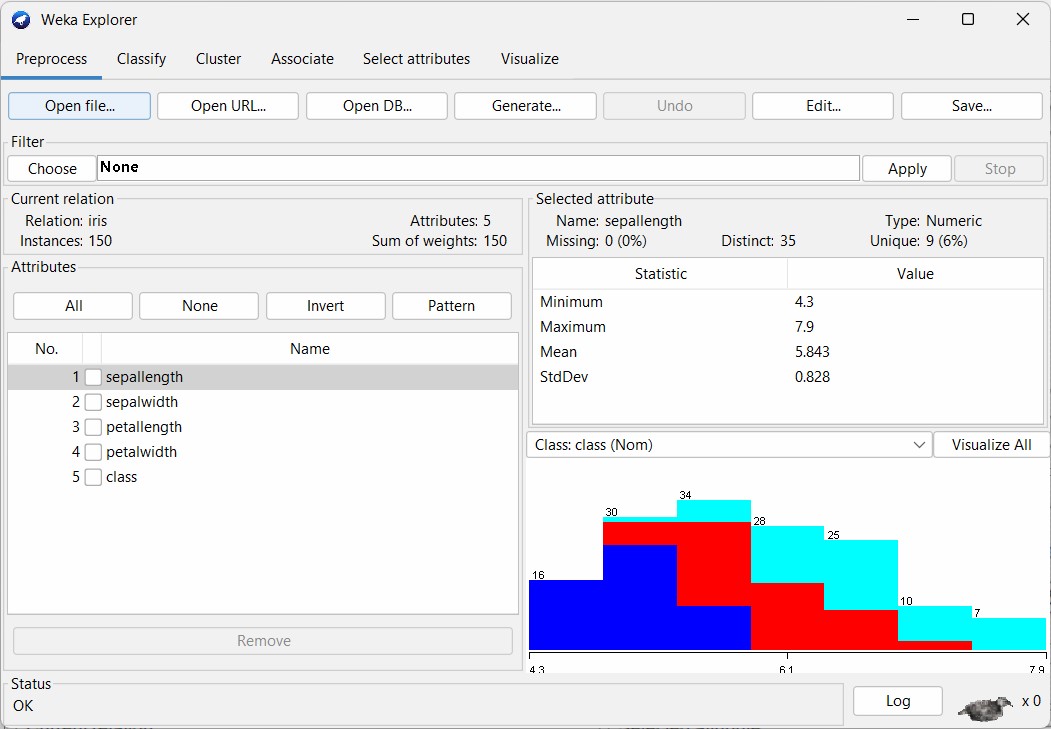
Trong học phần sử dụng Explorer

**1. Preprocess**

# Load dữ liệu

Dữ liệu sử dụng có thể từ các file định dạng arff, csv, c4.5 hay binary.

Weka có thể đọc dữ liệu từ một url hoặc từ một CSDL SQL



**Câu 1:**

* Tải tập dữ liệu *weather.nominal.arff*
* Cho biết tập dữ liệu có bao nhiêu thuộc tính (không thuộc tính lớp)? Các thuộc tính kiểu gì

\*Tập dữ liệu có 5 thuộc tính không phải là thuộc tính lớp: outlook, temperature, humidity, windy và play.

\* Các thuộc tính kiểu norminal.

* Các giá trị có thể có của temperature là gì?

\*Các giá trị hot, mild, cool.

**Câu 2:**

* Tải tập dữ liệu *glass.arff*
* Cho biết tập dữ liệu có bao nhiêu thuộc tính? Kiểu dữ liệu của từng loại thuộc tính

\* Tập dữ liệu có 10 thuộc tính,

\*9 thuộc tính đầu kiểu numberic, thuộc tính cuối kiểu nominal.

* Có tất cả bao nhiêu nhãn? (Mỗi đối tượng có thể có bao nhiêu nhãn)

+Có 7 nhãn.

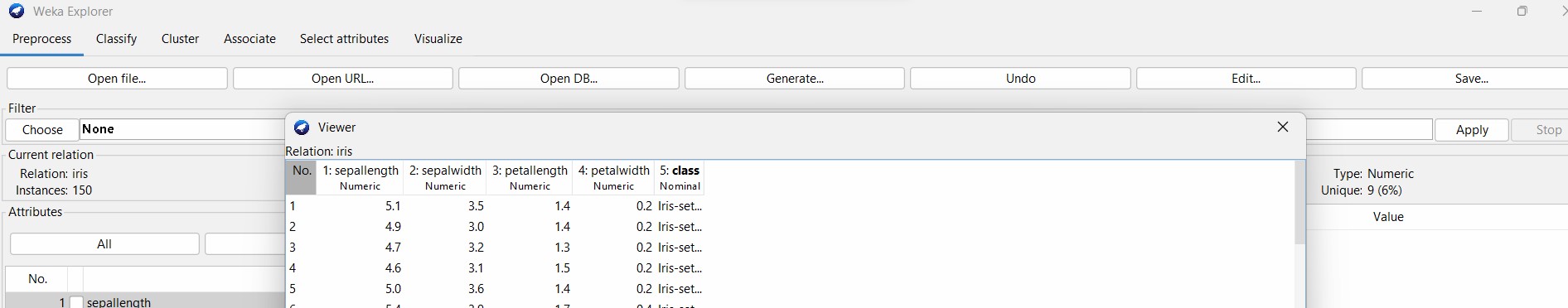
+Mỗi đối tượng có một nhãn

* Các giá trị có thể có của thuộc tính K nằm trong khoảng nào?

\*Từ 0 tới 6.21

# Edit dữ liệu

Chọn Edit để xem chi tiết dữ liệu



**Câu 3:**

* Xem chi tiết tập dữ liệu *weather.nominal.arff*
* Cột đầu tiên có ý nghĩa gì?

\*Số thứ tự.

* Bộ dữ liệu có số thứ 8 trong tập dữ liệu được phân lớp gì?

\*Phân lớp “NO”.

**Câu 4:**

* Xem chi tiết tập dữ liệu iris
* Các thuộc tính kiểu số và thuộc tính rời rạc là gì?

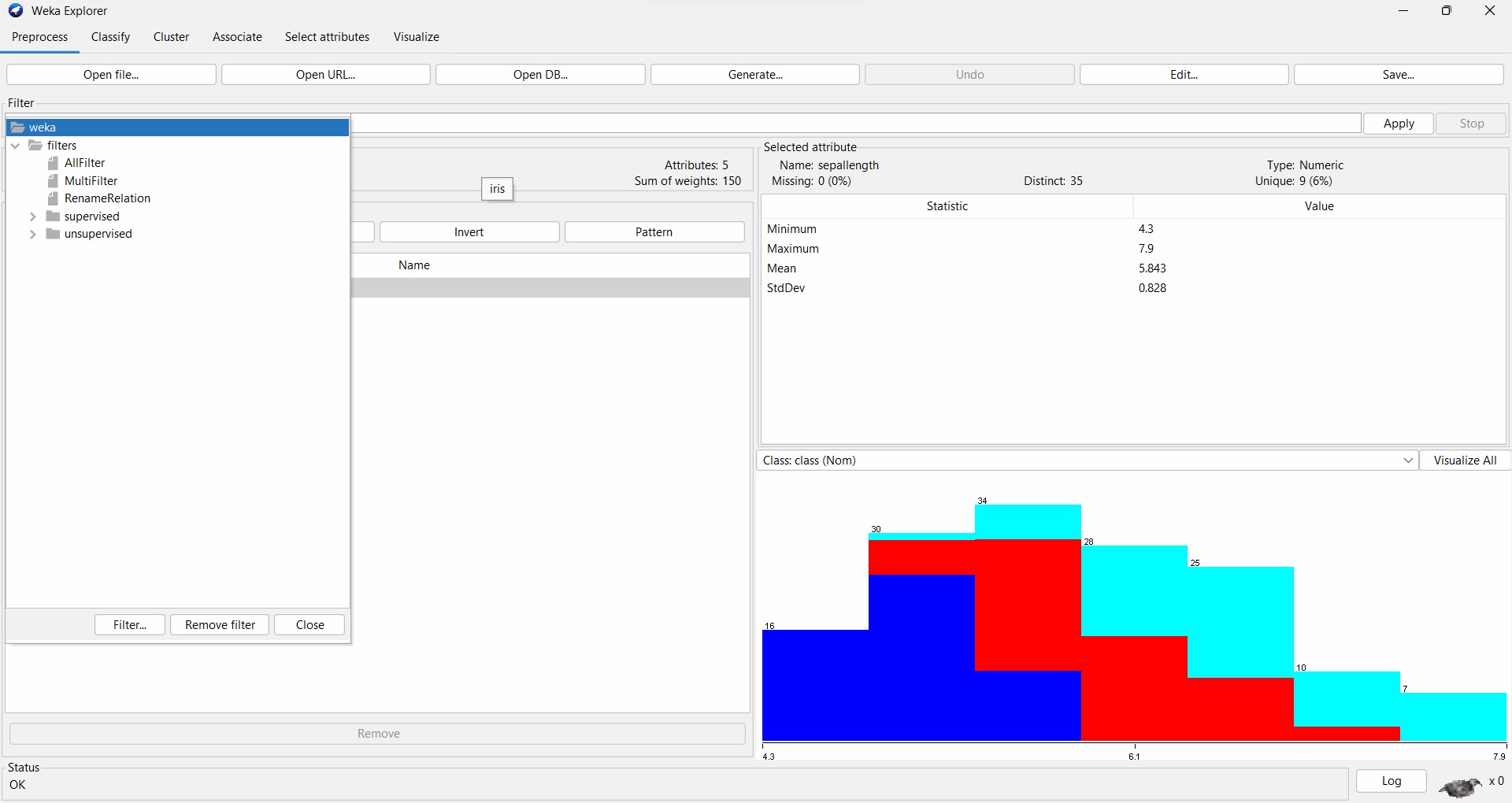
\*4 thuộc kiểu số và 1 thuộc tính rời rạc.

# Áp dụng Filter

Các công cụ trong Weka giúp điều chỉnh tập dữ liệu gọi là các **filter**.

Các filter hỗ trợ chính bởi Weka: discretization, normalization, resampling, attribute selection, transforming and combining attributes…

Chọn filter



Các loại filter:

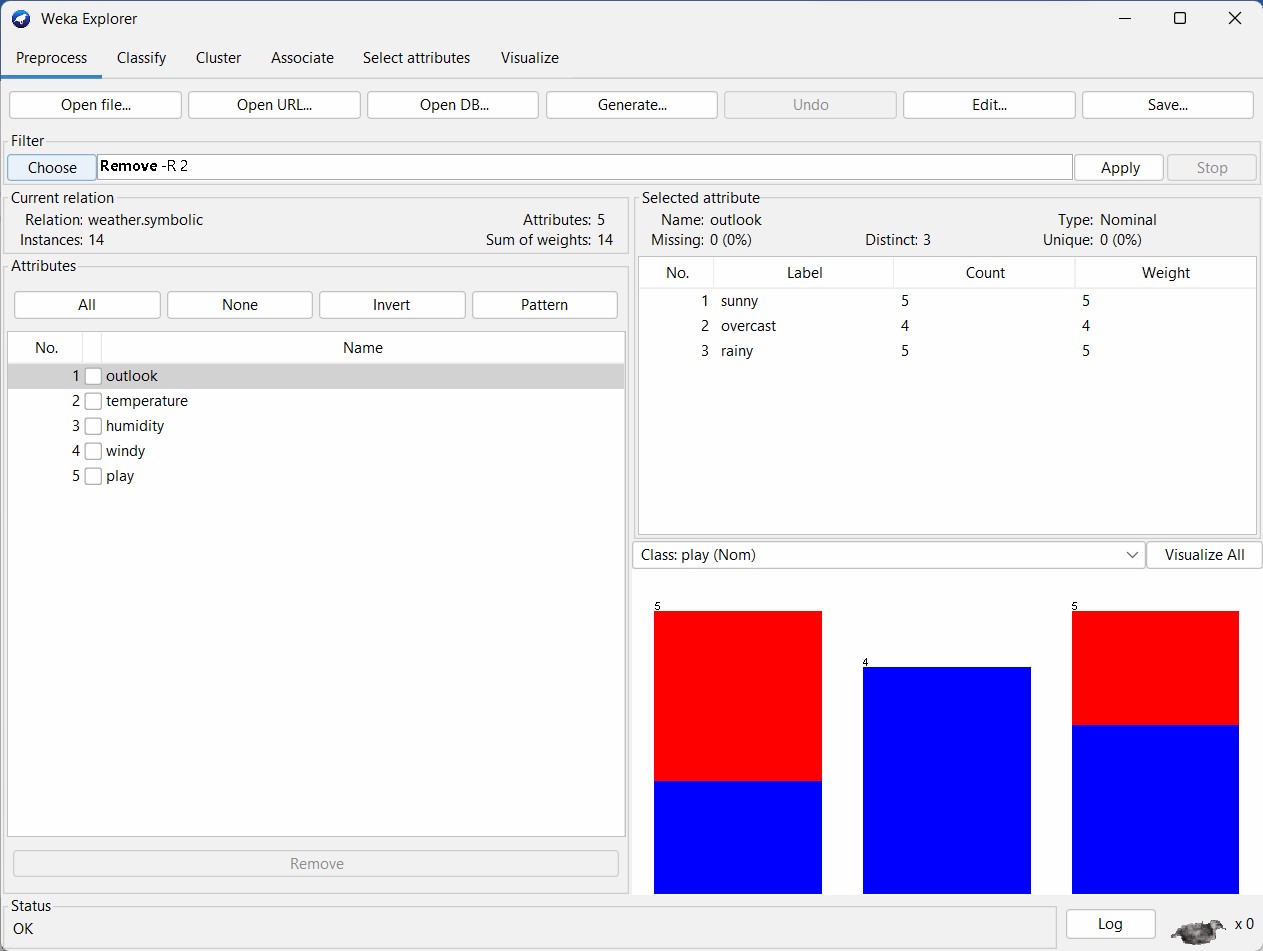
* Supervised filters: yêu cầu sự kiểm soát của người dùng- Unsupervised filter: ví dụ như chuẩn hóa dữ liệu về khoảng 0-1

Mỗi nhóm filter có 02 nhóm con:

* Attribute filters: áp dụng trên thuộc tính- Instance filters: áp dụng trên các bộ dữ liệu

*Ví dụ*:

Bỏ cột temperature trong tập dữ liệu weather.nominal.arff



**Câu 5:**

Tập dữ liệu sử dụng *weather.nominal*, sử dụng filter RemoveWithValues để bỏ các dòng dữ liệu có *humidity* là *high*.

A screenshot of a computer

Description automatically generated

(Chụp lại màn hình filter và dữ liệu chi tiết sau filter)

**Câu 6:**

Tập dữ liệu sử dụng *iris*, sử dụng filter AttributeSelection để loại bỏ các thuộc tính theo giải thuật lựa chọn thuộc tính.

A screenshot of a computer

Description automatically generated

Thuộc tính giữ lại là thuộc tính nào?

+petallength (độ dài cánh hoa)

+petalwidth (độ rộng cánh hoa)

+class (loại hoa)

(Chụp lại màn hình filter)

**Câu 7**

Tải tập dữ liệu *bank-data.csv*.

* Thực hiện bỏ thuộc tính id.
* Cho biết tập dữ liệu còn lại bao nhiêu thuộc tính (kể cả thuộc tính lớp)

A screenshot of a computer

Description automatically generated

+Tập dữ liệu còn lại 11 thuộc tính

Lưu lại tập dữ liệu dưới dạng file arff với tên *bank-data2*

A screenshot of a computer

Description automatically generated

* Mở file vừa lưu và xem lại trình bày có gì khác so với dữ liệu gốc không?
* Không khác gì với file ban đầu.
* Tải tập *bank-data2*.
* Áp dụng Discrete Filter cho các thuộc tính dạng số. Chụp lại các màn hình với các thuộc tính chuyển đổi.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Undo và điều chỉ các tham số sau: áp dụng các khoảng chia là 6.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

-Undo lại bước trước áp dụng Discrete Filter, áp dụng NumericToNominal cho các thuộc tính số.

A screenshot of a computer

Description automatically generated

Kết quả filter này so với Discrete Filter có gì khác biệt?

* Undo lại bước trước, áp dụng Normalize. Những thuộc tính nào có giá trị thay đổi?

A screenshot of a computer

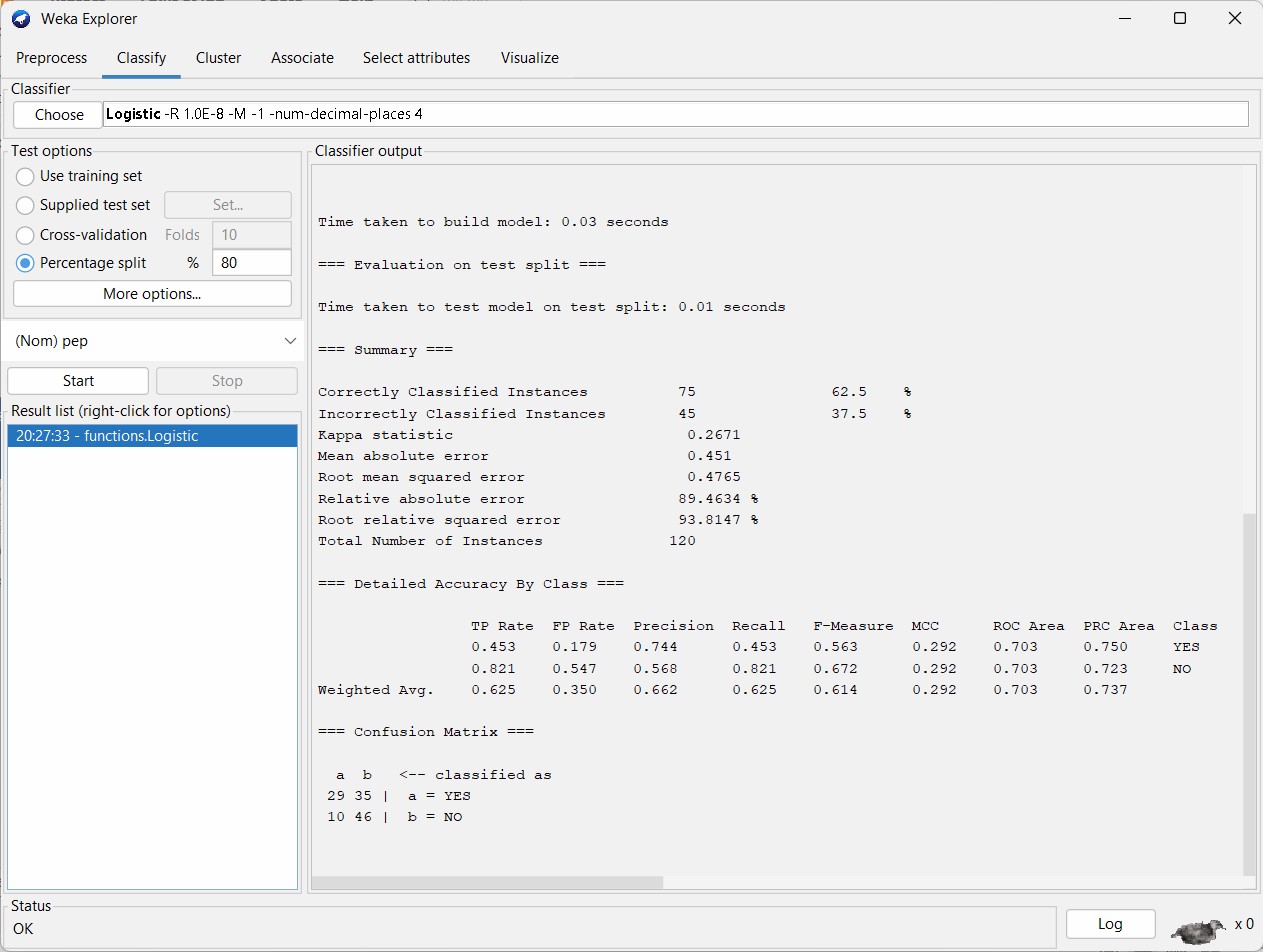
Description automatically generated

* Tiếp tục áp dụng Attribute Selection thì những thuộc tính nào còn lại?

A screenshot of a computer

Description automatically generated

**2. Classify**



Cho phép chọn các mô hình học cho bài toán phân loại và hồi quy.

**Câu 8**

Sử dụng tập dữ liệu *weather.nominal*. Mô hình học sử dụng J48.

* Sử dụng tùy chọn training set để thực hiện việc học.
* Độ chính xác là bao nhiêu?

A screenshot of a computer

Description automatically generated

* Cây học được có dạng thế nào? (Chụp hình)

A screenshot of a computer program

Description automatically generated

* Nếu một trường hợp có các giá trị ứng với các thuộc tính như sau:

Outlook = sunny, temperature = cool, humidity = high, windy = TRUE

Thì trường hợp này được phân loại là gì với cây vừa học.

+no

**Câu 9**

Sử dụng tập dữ liệu *bank.csv*

* Mô tả về tập dữ liệu (số lượng các trường hợp, số lượng thuộc tính, các kiểu dữ liệu, bài toán phân loại hay hồi quy...)
* Có 600 trường hợp (instance)
* Có tất cả 12 thuộc tính, trong đó age, income, children thuộc kiểu numeric, các trường còn lại có kiểu là Nominal.

\*Đây là một bài toán phân loại vì ta đang phân loại dựa theo thuộc tính Pep.

A screenshot of a computer

Description automatically generated

* Thực hiện chuẩn hóa dữ liệu (lựa chọn normalize hoặc standardize..)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Áp dụng lựa chọn các thuộc tính (bằng attributeselection và bằng cách chọn 50% thuộc tính có rank cao nhất, lưu lại dưới 2 tập dữ liệu với tên khác nhau).

A screenshot of a computer

Description automatically generated

* Thực hiện việc chia train và test và lưu thành các file tương ứng

A screenshot of a computer

Description automatically generated

* Xây dựng các cây quyết định tương ứng với các tập dữ liệu vừa tạo.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated