GIỚI THIỆU WEKA – PHÂN LỚP DỮ LIỆU VỚI WEKA

Giới thiệu về WEKA

- **WEKA** – **W**aikato **E**nvironment for **K**nowledge **A**nalysis.
- Là phần mềm khai thác dữ liệu, thuộc dự án nghiên cứu của đại học Waikato, New Zealand.
- Mục tiêu: xây dựng một công cụ hiện đại nhằm phát triển các kỹ thuật máy học và áp dụng chúng vào bài toán khai thác dữ liệu trong thực tế.

Lịch sử phát triển

- **1993** – Đại học Waikato, New Zealand, khởi động dự án, xây dựng phiên bản đầu tiên của Weka.
- **1997** – Quyết định xây dựng lại Weka từ đầu bằng Java, có cài đặt các thuật toán mô hình hóa.
- **2005** – Weka nhận giải thưởng SIGKDD Data Mining and Knowledge Discovery Service Award.
- **Xếp hạng** trên Sourceforge.net từ 25-06-2007: **241** (907,318 lượt).

Cấu trúc phần mềm

- WEKA được xây dựng bằng ngôn ngữ Java, cấu trúc gồm hơn 600 lớp, tổ chức thành 10 packages.
- Các chức năng chính của phần mềm:
  - Khảo sát dữ liệu: tiền xử lí dữ liệu, phân lớp, gom nhóm dữ liệu, và khai thác luật kết hợp.
  - Thực nghiệm mô hình: cung cấp phương tiện để kiểm chứng, đánh giá các mô hình học.
  - Biểu diễn trực quan dữ liệu bằng nhiều dạng đồ thị khác nhau.

Các phiên bản WEKA

- *Snapshots* là các bản vá lỗi mới nhất, thường là được cập nhật hàng đêm.
- *Book versions* là các phiên bản thể hiện những chức năng được mô tả trong quyển sách *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)* của Ian.H.Witten và Eibe Frank.
- *Developer versions* là các phiên bản thử nghiệm, hỗ trợ nhiều tính năng mới nhưng còn chưa ổn định.

Download: Trang chủ: http://www.cs.waikato.ac.nz/ml/weka/

Các chức năng của WEKA explorer

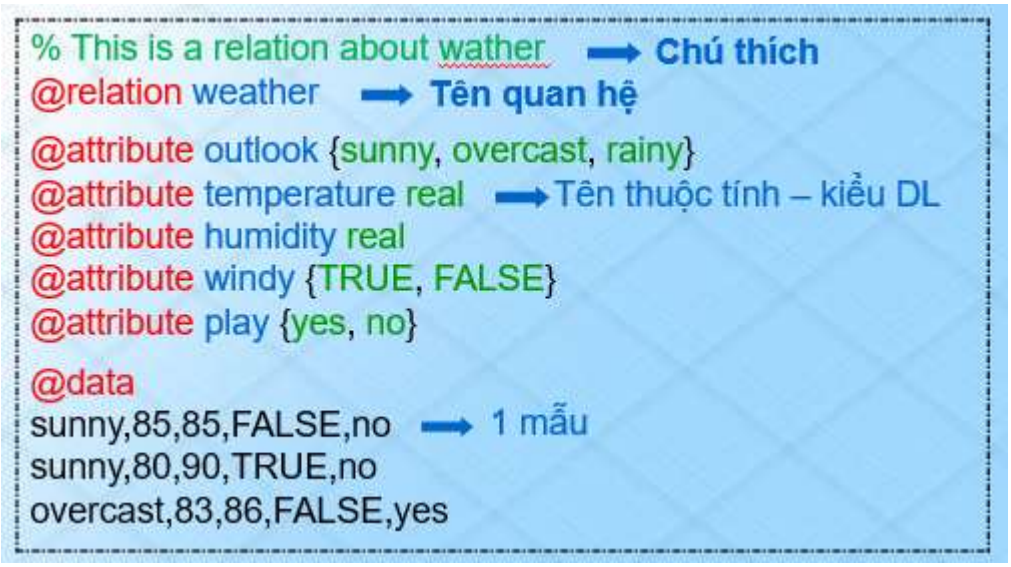
Explorer: là ứng dụng con cho phép thực nghiệm các nhiệm vụ khai thác dữ liệu thường gặp như:

- Tiền xử lý dữ liệu
- Khai thác luật kết hợp
- Phân lớp
- Gom nhóm

XỬ LÝ DỮ LIỆU VỚI WEKA

Cấu trúc tập tin ARFF (Attribute-Relation File Format)

- *ARFF* là định dạng dữ liệu chuyên biệt của Weka, tổ chức dữ liệu theo cấu trúc được qui định trước.
- Cấu trúc tập tin *.ARFF bao gồm các thành phần:
  - Header: chứa khai báo quan hệ, danh sách các thuộc tính (tên, kiểu dữ liệu).
  - Data: gồm nhiều dòng, mỗi dòng thể hiện giá trị của các thuộc tính cho một mẫu.



- Phần khai báo:
  @relation <tên dữ liệu>
  @attribute <tên thuộc tính 1> <Kiểu dữ liệu>
  @attribute <tên thuộc tính 2> <Kiểu dữ liệu>
  …
  @attribute <tên thuộc tính n> <Kiểu dữ liệu>
  - Các kiểu dữ liệu
    - numeric: là kiểu dữ liệu số, gồm real và integer
    - nominal: là kiểu dữ liệu danh sách.
    - string: là kiểu dữ liệu dạng chuỗi
    - date: kiểu dữ liệu thời gian (ngày tháng năm, giờ phút giây…)

```
@relation nhanvien
@attribute hoten string
@attribute ngaysinh date "dd/MM/yy"
@attribute gioitinh {nam, nu}
@attribute hesoluong real

@data
'Nguyen Van A', 10/12/1957, nam, 1.34
'Tran Thi B', ?, nu, 1.5
```

- Phần dữ liệu:

  Mỗi mẫu dữ liệu được đặt trên một dòng, giá trị của các thuộc tính được liệt kê theo thứ tự từ trái qua phải và ngăn cách bởi dấu phẩy “,”

  - **Chú ý**

- Dòng ghi chú được bắt đầu bằng dấu %.
- Dữ liệu thiếu được biểu diễn bằng dấu ?.
- Chuỗi nếu có khoảng trắng phải đặt trong dấu nháy đơn.
- Các giá trị trong phần data phải tuyệt đối theo đúng thông tin đã khai báo trong header.

Comma Separated Values (*.csv)

- Là tập tin văn bản
- Cấu trúc tương tự phần dữ liệu của tập tin arff: Các mẫu được lưu trên một dòng, các thuộc tính được ngăn cách bằng dấu phẩy.
- Dòng đầu tiên chứa tên các thuộc tính.

Ví dụ tập tin csv:

```
outlook,temperature,humidity,windy,play
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Có nghĩa là dữ liệu này gồm có 14 mẫu và 5 thuộc tính

(outlook, temperature, humidity, windy, play).

Hiển thị tập tin này bằng arffViewer:

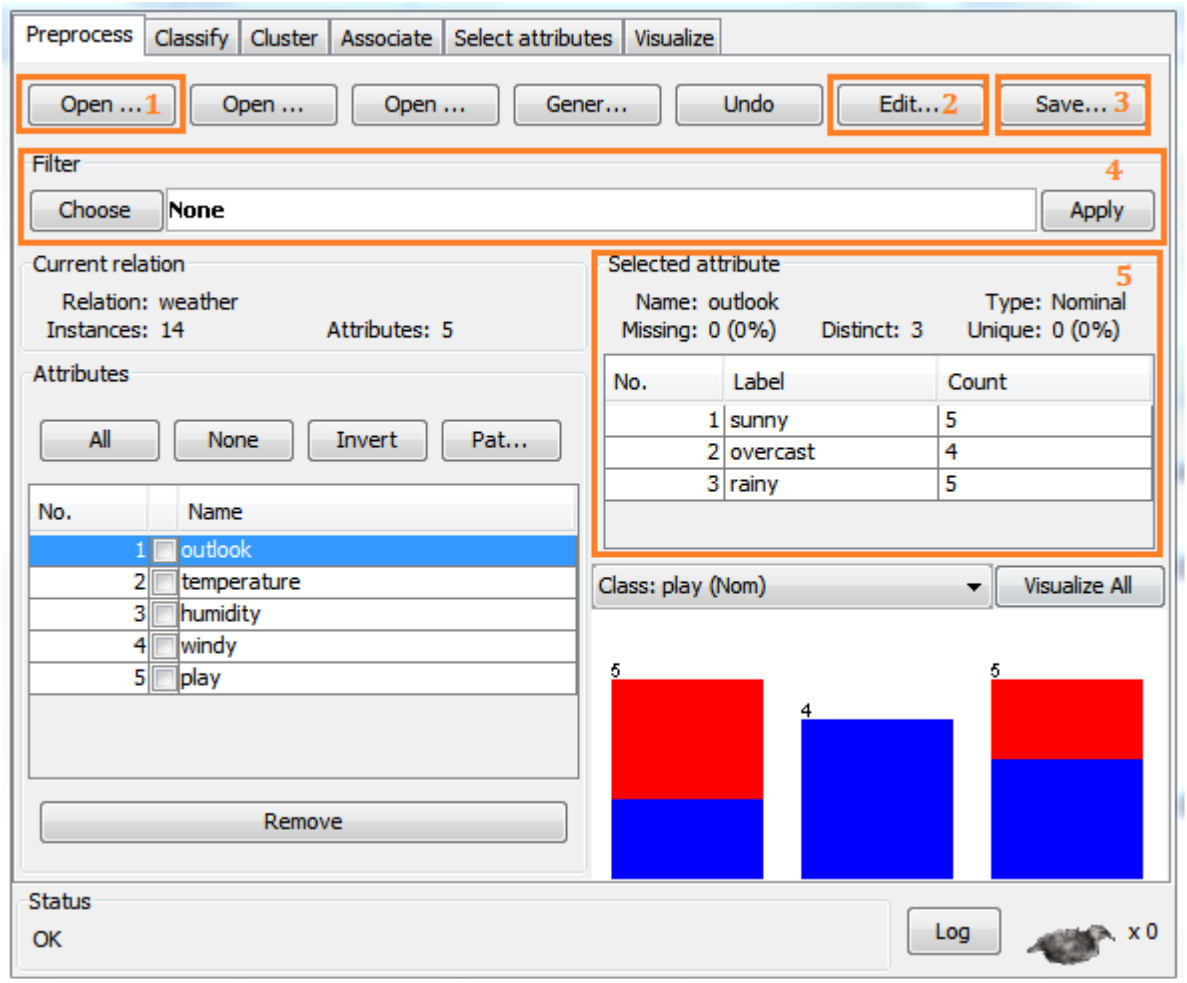
weather.nominal.arff.csv

Relation: weather.nominal.arff

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Khảo sát dữ liệu: sử dụng thẻ Preprocess

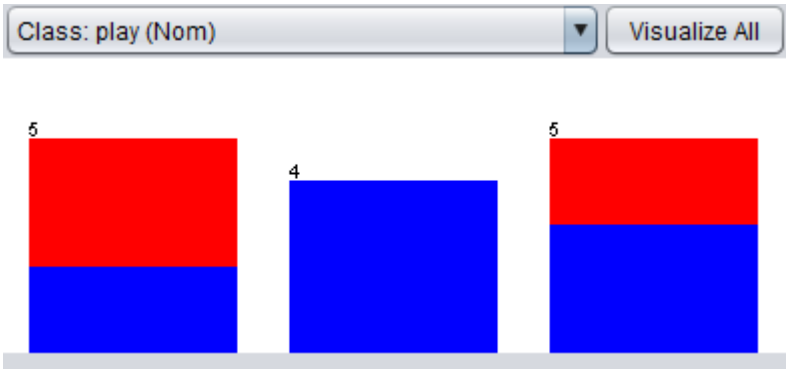
- (1) **Open file…**: Mở một tập tin dữ liệu.
- (2) **Edit…:** Hiển thị và chỉnh sửa dữ liệu bằng tay nếu cần thiết.
- (3) **Save…:** Lưu dữ liệu hiện tại ra tập tin. Weka Explorer hỗ trợ một số định dạng trong đó có 2 định dạng chính cần quan tâm là *.arff và *.csv
- (4) **Filter:** Các tác vụ tiền xử lý được gọi là các bộ lọc.

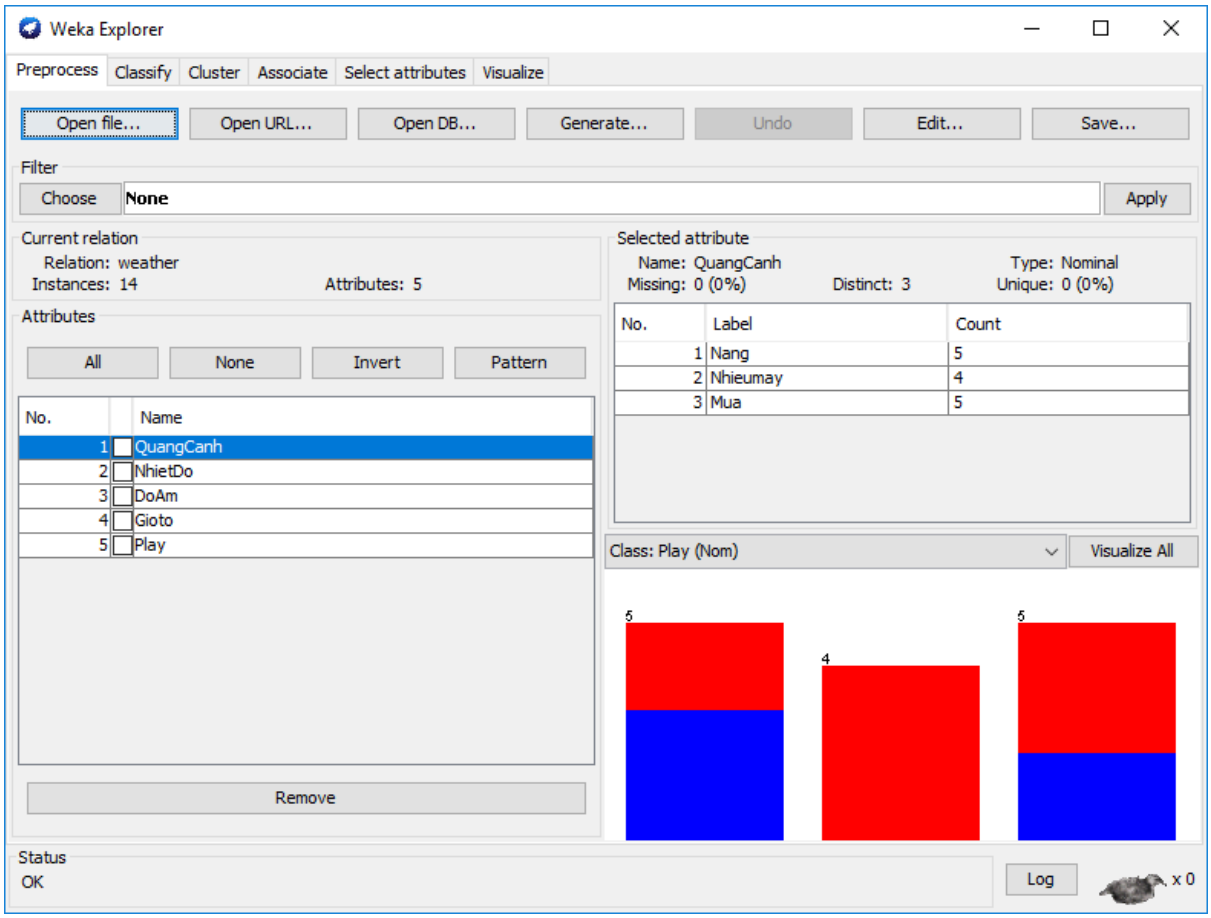


- (5) **Selected attribute:** Thông tin về thuộc tính đang được chọn:
  - **Type**: Kiểu dữ liệu của thuộc tính (**Numeric**: Dạng số, **Nominal**: Dạng rời rạc/phi số).
  - **Missing:** Số mẫu thiếu giá trị trên thuộc tính đang xét
  - **Distinct:** Số giá trị phân biệt
  - **Unique:** Số mẫu không có giá trị trùng với mẫu khác
  - **Bảng thống kê:**
    - **Dạng phi số:** Thể hiện các giá trị và tần suất của mỗi giá trị

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

    - **Biểu diễn trực quan:**





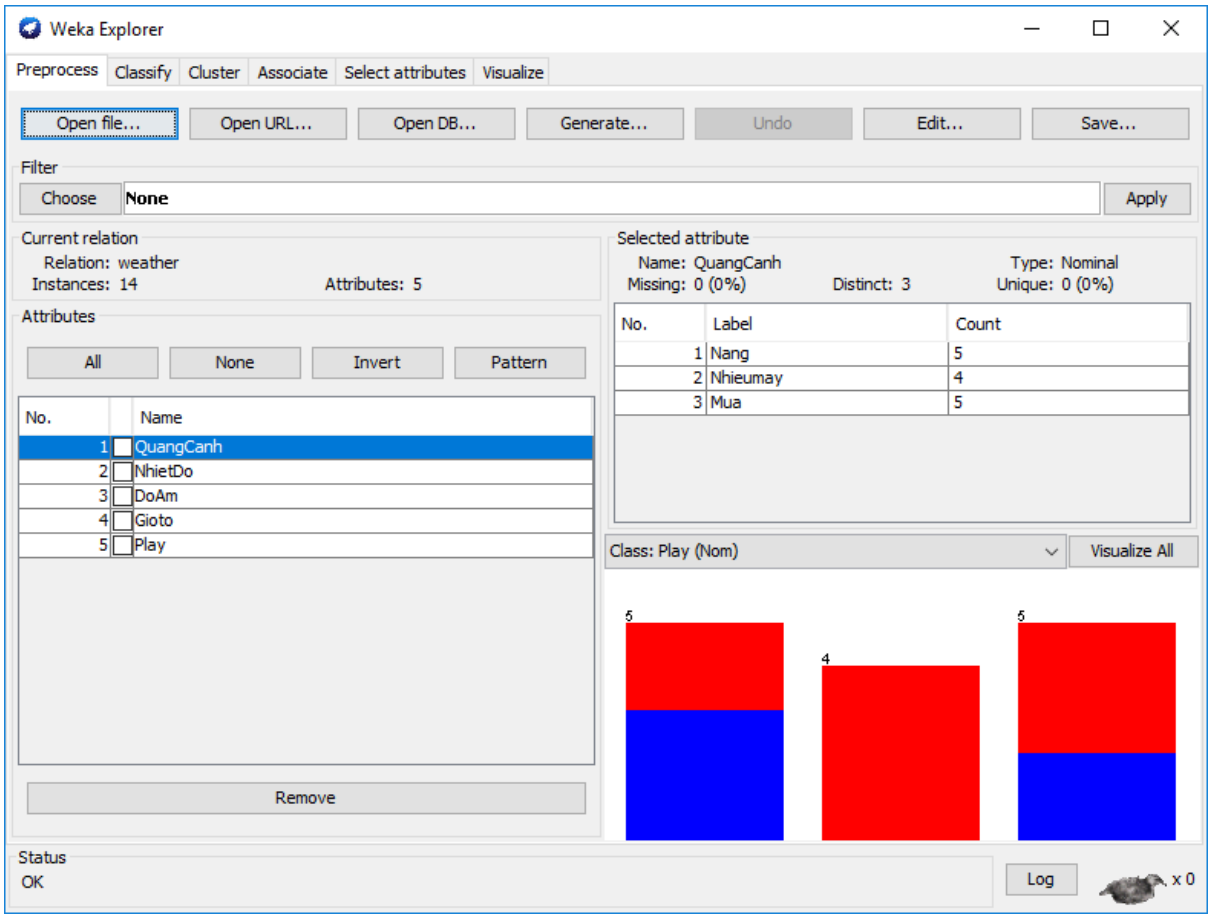
Các chức năng chính của Weka Explorer thể hiện trong các thẻ (tab) của màn hình chính, bao gồm:

- **Preprocess**: Cho phép mở, điều chỉnh, lưu một tập tin dữ liệu, thẻ này chứa các thuậtt toán áp dụng trong tiền xử lý dữ liệu.
- **Classify**: Cung cấp các mô hình phân loại dữ liệu hoặc hồi quy.
- **Cluster**: Cung cấp các mô hình gom cụm.
- **Associate**: Khai thác tập phổ biến và luật kết hợp.
- **Select Attributes**: Lựa chọn các thuộc tính thích hợp nhất trong tập dữ liệu
- **Visualize**: Thể hiện dữ liệu dưới dạng biểu đồ

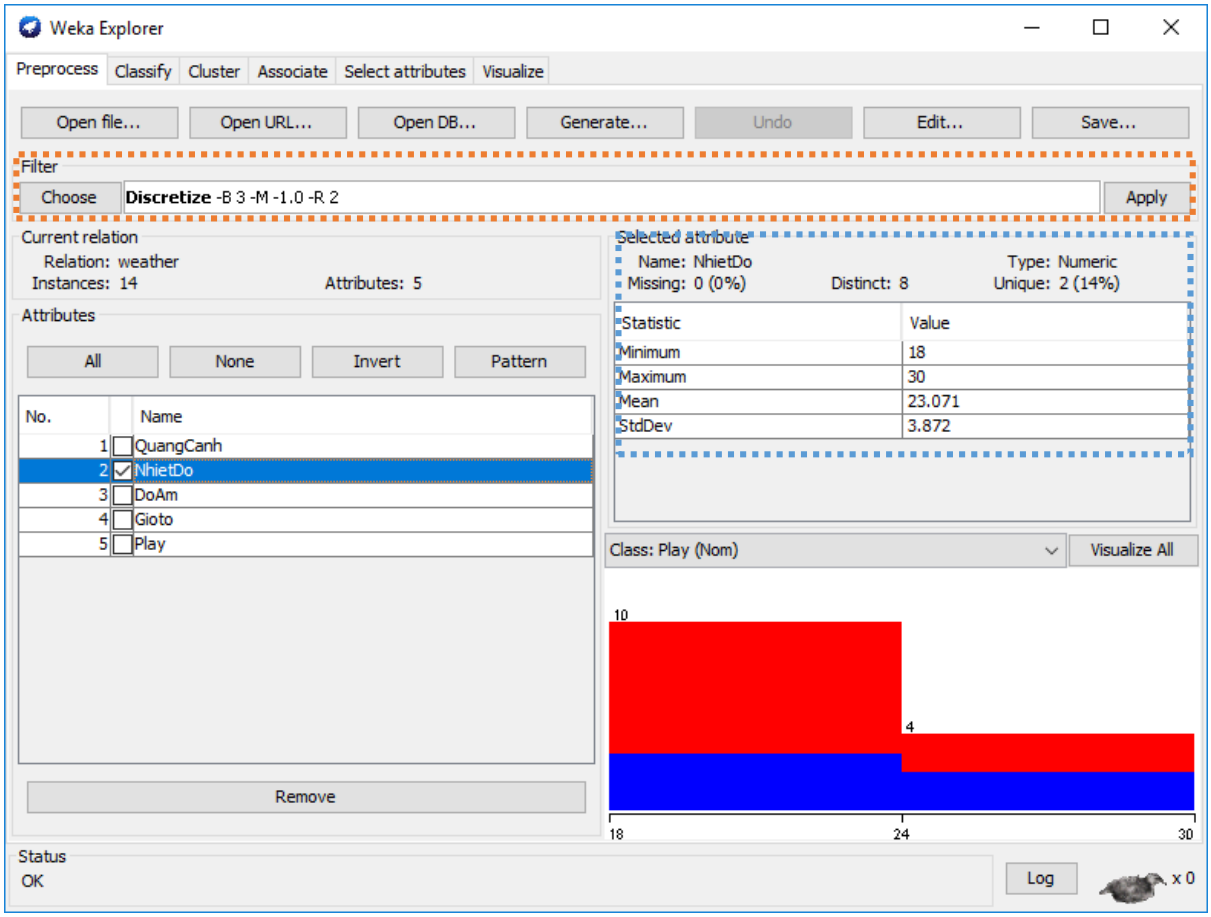
Tiền xử lý dữ liệu: xử lý dữ liệu liên tục

Trong Data Mining, một số kỹ thuật như khai phá luật kết hợp (association rule mining) chỉ có thể thực hiện trên các dữ liệu phân loại (categorical/ nominal data). Điều này yêu cầu phải thực hiện việc rời rạc hóa trên các thuộc tính có kiểu dữ liên tục (như kiểu numeric chẳn hạn)

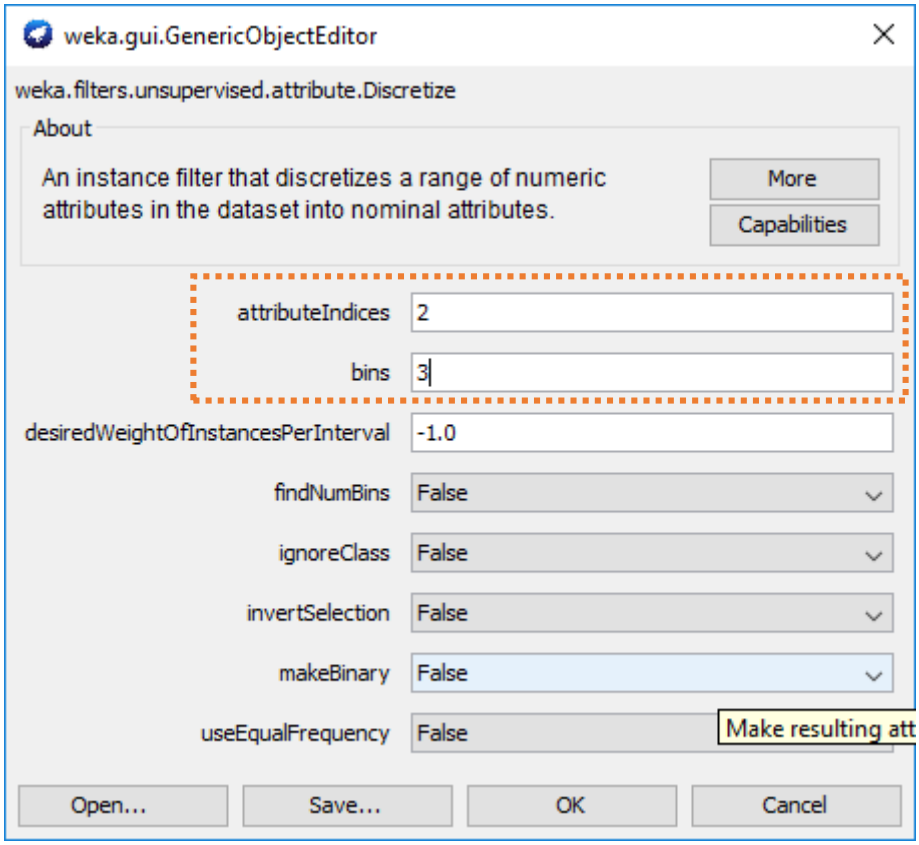
Bước 1: nạp file dữ liệu



Bước 2: *mở hộp thoại Filter và chọn: filters.unsupervised.attribute.Discretize*

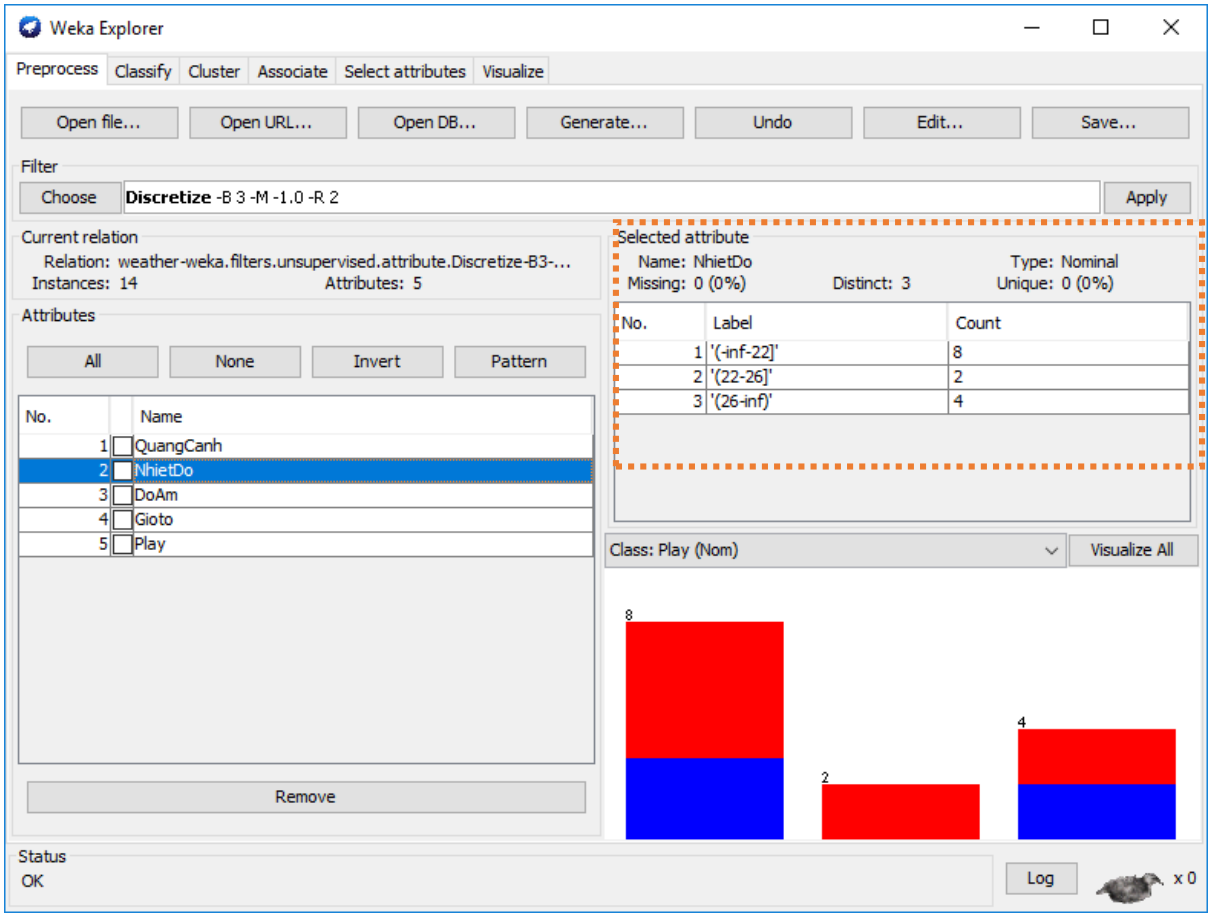


Bước 3: bấm chuột vào text box ngay bên phải nút "Choose" và thiết lập các tham số để thực hiện việc rời rạc hóa.

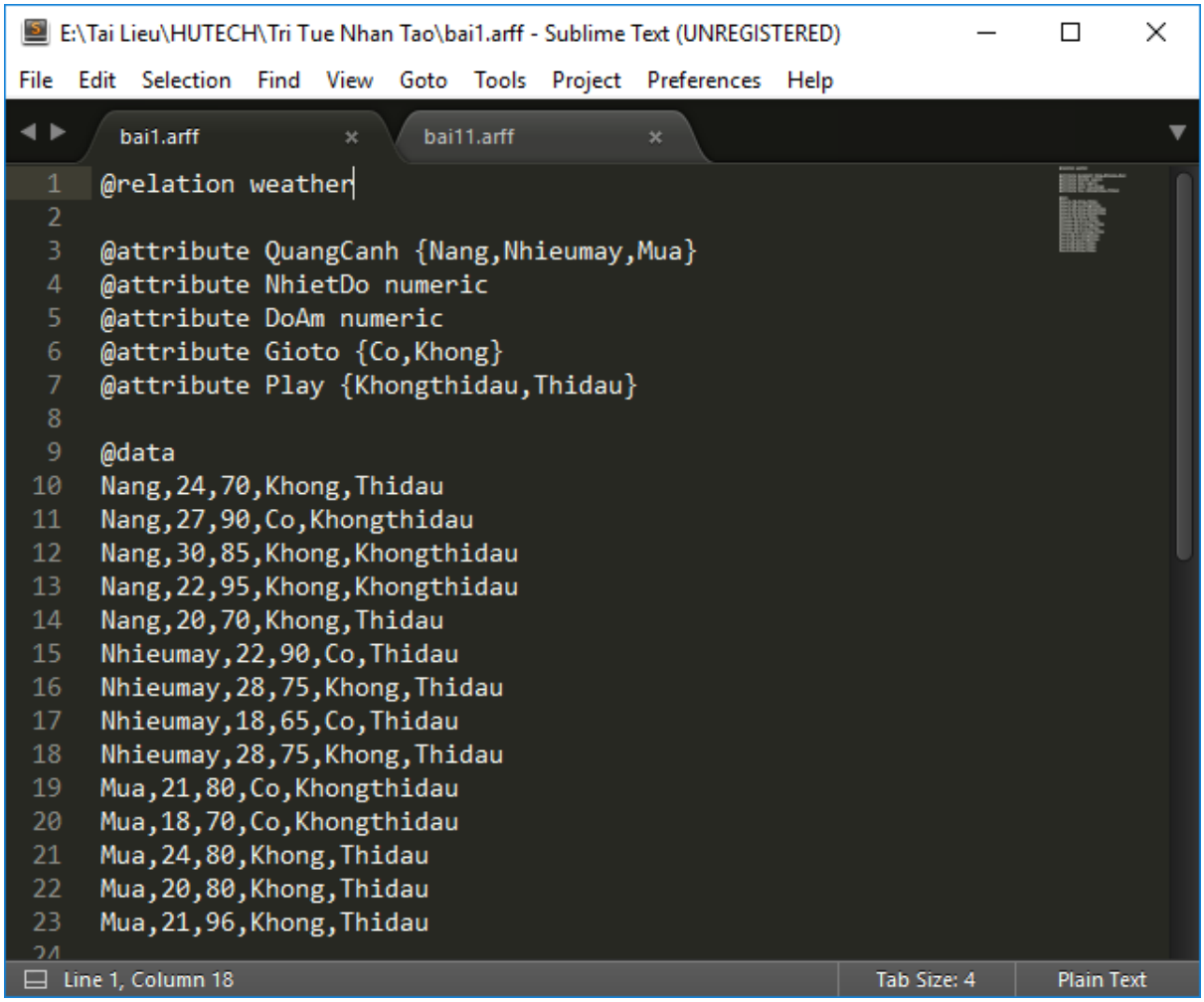


- *attributeIndices* nhập số tương ứng với index của thuộc tính liên tục mà ta muốn rời rạc.
- *bins* nhập số khoảng muốn chia

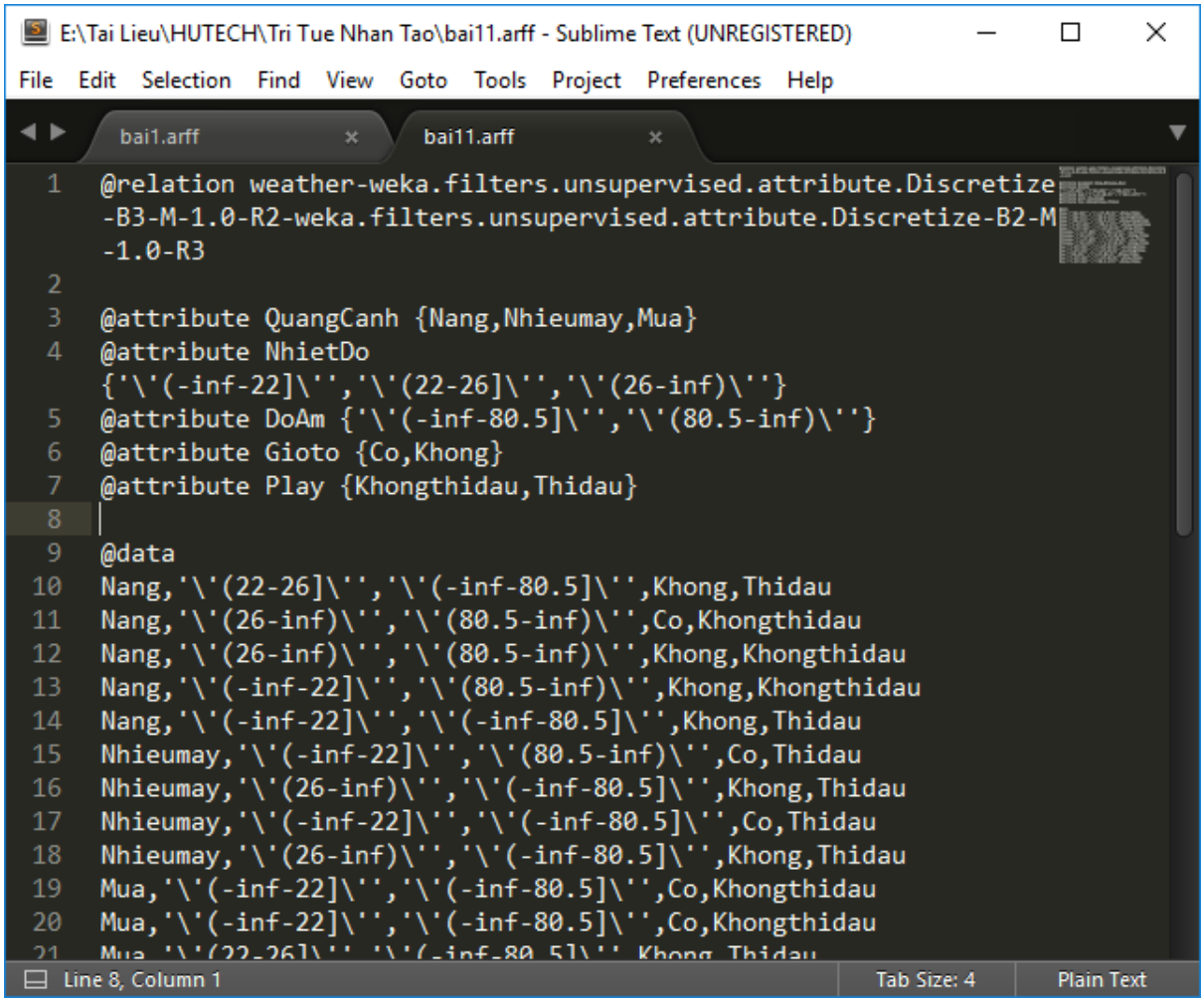
Bước 4: Click "Apply" để thực hiện, click “Save” để lưu lại dữ liệu đã rời rạc hóa



File ARFF trước khi chuẩn hóa

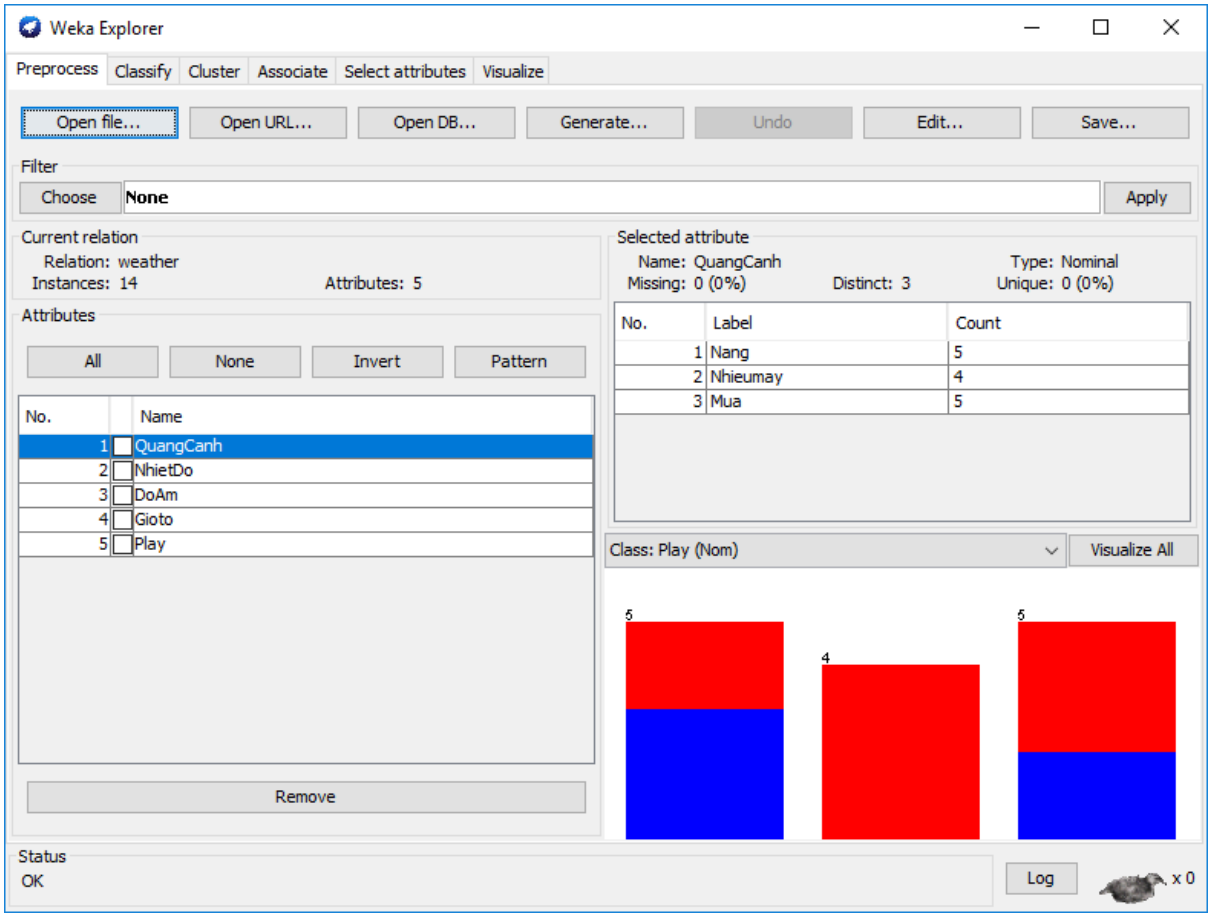


File arff sau khi chuẩn hóa:

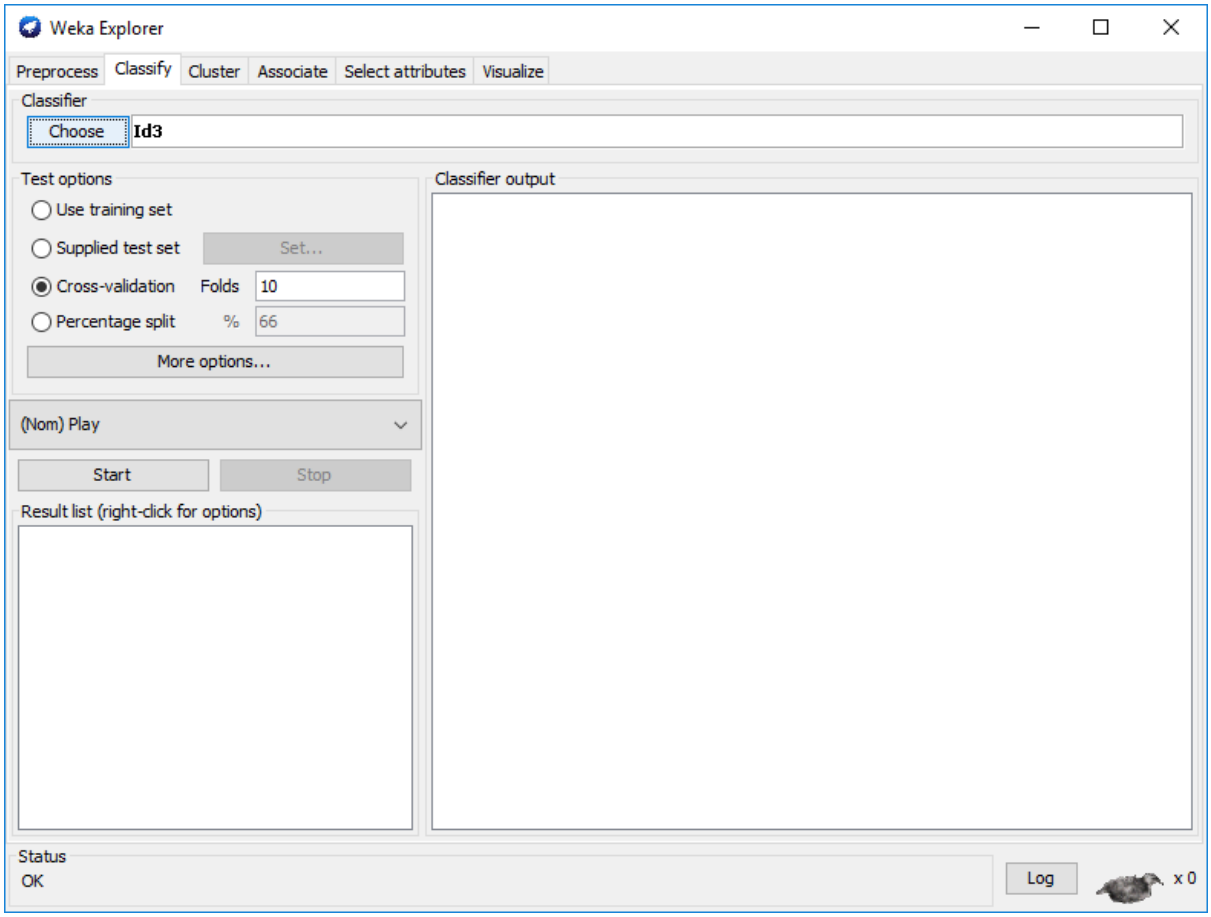


Phân lớp sử dụng ID3 với Weka

- Bước 0: tại tab Preprocess, chọn tập dữ liệu và thực hiện tiền xử lí dữ liệu nếu cần thiết.



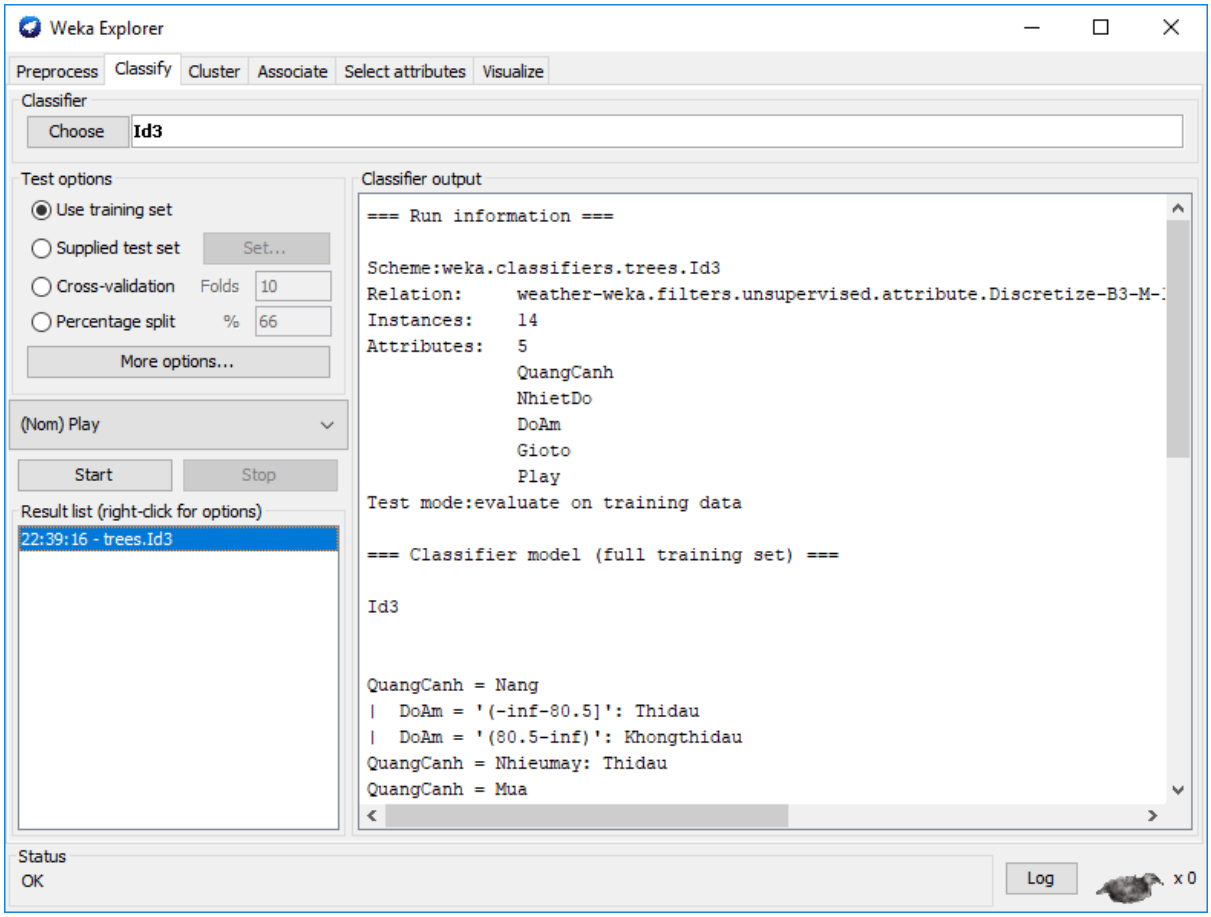
- Bước 1: chọn thuật toán ID3 tại tab Classify



- Bước 2: chọn kiểu test và nhập dữ liệu test nếu cần.

Có rất nhiều kiểu test, nhưng có 4 kiểu chính:

  - Use training set: sử dụng chính tập huấn luyện là tập test
  - Supplied test set: chỉ định tập test mới
  - Cross-validation: lấy bao nhiêu dòng dữ liệu làm dữ liệu test
  - Percentage split: chia tỷ lệ phần trăm.
- Bước 3: tiến hành phân lớp, bấm Start
- Bước 4: ghi nhận kết quả



- Run information: thông tin về mô hình học, tên quan hệ, số mẫu, thuộc tính và kiểu test.

```
=== Run information ===

Scheme:weka.classifiers.trees.Id3
Relation:     weather-weka.filters.unsupervis
Instances:    14
Attributes:   5
              QuangCanh
              NhietDo
              DoAm
              Gioto
              Play
Test mode:evaluate on training data
```

- Classifier model (full training set): cho biết mô hình phân lớp được xây dựng dựa trên cả tập huấn luyện.

```
=== Classifier model (full training set) ===

Id3


QuangCanh = Nang
|   DoAm = '(-inf-80.5]': Thidau
|   DoAm = '(80.5-inf)': Khongthidau
QuangCanh = Nhieumay: Thidau
QuangCanh = Mua
|   Gioto = Co: Khongthidau
|   Gioto = Khong: Thidau

Time taken to build model: 0 seconds
```

- Tổng kết: số liệu thống kê cho biết độ chính xác của bộ phân lớp theo một kiểu test cụ thể.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          14               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               14
```

- Độ chính xác chi tiết từng phân lớp

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        Khongthidau
                 1         0          1         1         1          1        Thidau
Weighted Avg.    1         0          1         1         1          1
```

- Confusion matrix: cho biết bao nhiêu mẫu được gán vào từng lớp. Các phần tử của ma trận thể hiện số mẫu test có lớp thật sự là dòng và lớp dự đoán là cột.

```
=== Confusion Matrix ===

 a b   <-- classified as
 5 0 | a = Khongthidau
 0 9 | b = Thidau
```

Phân lớp sử dụng Navie bayes

(xem clip)

-Huớng dẫn cài đặt phần mềm Weka

https://www.youtube.com/watch?v=C9YL8kQE7Ns

-Các chức năng chính của phần mềm WEKA

https://www.youtube.com/watch?v=7hLXzifK7r8

- Chức năng tiền xử lí dữ liệu

https://www.youtube.com/watch?v=7H7PgfvmJY8

- Ví dụ minh họa về tiền xử lý dữ liệu

https://www.youtube.com/watch?v=9Psn1wKGcYA

- Chức năng phân lớp Classify trên Weka

https://www.youtube.com/watch?v=6_jcJVFGymk

-Minh họa thuật toán ID3 trên Weka

https://www.youtube.com/watch?v=docE3QtP6AI

-Minh họa Navie Bayes trên Weka

https://www.youtube.com/watch?v=9z7FBV5BUrA&t=68s

Bài tập

1. Tạo tập tin ARFF cho tất cả các bài tập lab01, lab02
2. Chuẩn hóa dữ liệu nếu cần
3. Sử dụng weka để chạy thuật toán ID3 (lab01), Navie Bayes (lab02), ghi nhận lại kết quả và so sánh với kết quả chạy tay ở những tuần trước