

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

BÀI TẬP THỰC HÀNH 1

PREPROCESSING

Nhóm 24

1712607 - Nguyễn Văn Hoài Nam

1712845 - Nguyễn Ngọc Trung

1712922 - Phạm Hoàng Vũ



Khoa Công nghệ Thông tin
Đại học Khoa học Tự nhiên TP HCM
Tháng 09/2019

MỤC LỤC

1	Tổng quan.....	3
	Thông tin nhóm.....	3
	Thông tin bài tập.....	3
2	Chi tiết bài tập.....	5
2.1	Báo cáo viết	5
2.1.1	Tạo tập tin arff từ tập dữ liệu Course Ratings	5
2.1.2	Khảo sát tập dữ liệu Weather (thuộc tính rời rạc)	10
2.1.3	Khảo sát tập dữ liệu Iris (thuộc tính số liên tục).....	14
2.1.4	Khảo sát tập dữ liệu Weather (thuộc tính số liên tục).....	25
2.2	Nội dung hướng dẫn sử dụng cài đặt	33
2.2.1	Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản.....	33
2.2.1	Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước.....	34
2.3	Tài liệu tham khảo.....	35

1

Tổng quan

Thông tin nhóm

MSSV	Họ tên	Email	Tỷ lệ thực hiện
1712607	Nguyễn Văn Hoài Nam	1712607@student.hcmus.edu.vn	
1712845	Nguyễn Ngọc Trung	1712845@student.hcmus.edu.vn	
1712922	Phạm Hoàng Vũ	1712922@student.hcmus.edu.vn	

Thông tin bài tập

Nội dung	Yêu cầu	Hoàn thành	Ghi chú
A1-Tạo tập tin arff từ tập dữ liệu Course Ratings	Tạo tập tin từ tập dữ liệu Course Ratings, đặt tên là "course_rating.arff" và nộp lại cho giáo viên	100%	
	Đọc tập dữ liệu vào Weka	100%	
	Quan sát, nhận xét và trả lời câu hỏi	100%	
	So sánh nội dung các đồ thị trình diễn do chức năng Visualize All cung cấp	100%	
	Minh họa và chụp màn hình làm minh chứng	100%	
A2-Khảo sát tập dữ liệu Weather (thuộc tính rời rạc)	Quan sát, mô tả và trả lời câu hỏi	100%	
	Sử dụng các bộ lọc của Weka. Đặt tên tập tin là "weather_remove3.arff" và nộp lại cho giáo viên	100%	
	Sử dụng bộ lọc weka.unsupervised.instance.RemoveWithValues để loại bỏ mọi mẫu có giá trị thuộc tính humidity là high	100%	
	Minh họa và chụp màn hình làm minh chứng	100%	
A3-Khảo sát tập dữ liệu Iris	Quan sát, mô tả và trả lời câu hỏi	100%	
	Sử dụng tab Visualize	100%	

(thuộc tính số liên tục)	Sử dụng Jitter slider, Select Instance, Reset, Clear và Save	100%	
	Minh họa và chụp màn hình làm minh chứng	100%	
A4-Khảo sát tập dữ liệu Weather (thuộc tính số liên tục)	Quan sát, mô tả và trả lời câu hỏi theo phần 3	100%	
	Minh họa và chụp màn hình làm minh chứng	100%	
B1-Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản	Chuẩn hóa min-max trên danh sách thuộc tính chỉ định	100%	
	Chuẩn hóa Z-scores trên danh sách thuộc tính chỉ định	100%	
	Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ rộng trên danh sách thuộc tính chỉ định	100%	
	Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ sâu trên danh sách thuộc tính chỉ định	100%	
	Xóa các mẫu dữ liệu thiếu giá trị trên danh sách thuộc tính chỉ định	100%	
	Điền giá trị bị thiếu trên danh sách thuộc tính chỉ định, giá trị được điền là giá trị trung bình (mean) của thuộc tính nếu đó là thuộc tính số hoặc điền giá trị có tần số xuất hiện cao nhất (mode) nếu là thuộc tính rời rạc	100%	
	Nhận đầu vào là một tập tin CSV (.csv) và tạo đầu ra cũng là một tập tin CSV	100%	
	Hoạt động theo cơ chế console, yêu cầu người dùng được đặc tả thông qua tham số dòng lệnh	100%	
B2-Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước	Cài đặt chương trình chuyển tập tin countries.txt trên thành tập tin CSV (.csv)	100%	
	Xóa các mẫu rỗng	100%	
	Xóa các mẫu bị trùng lặp	100%	
	Chuyển diện tích về km2	100%	
	Sử dụng chương trình đã cài đặt ở phần B-1. để xóa các mẫu bị thiếu diện tích	100%	
	Hoạt động theo cơ chế console, yêu cầu người dùng được đặc tả thông qua tham số dòng lệnh	100%	

2

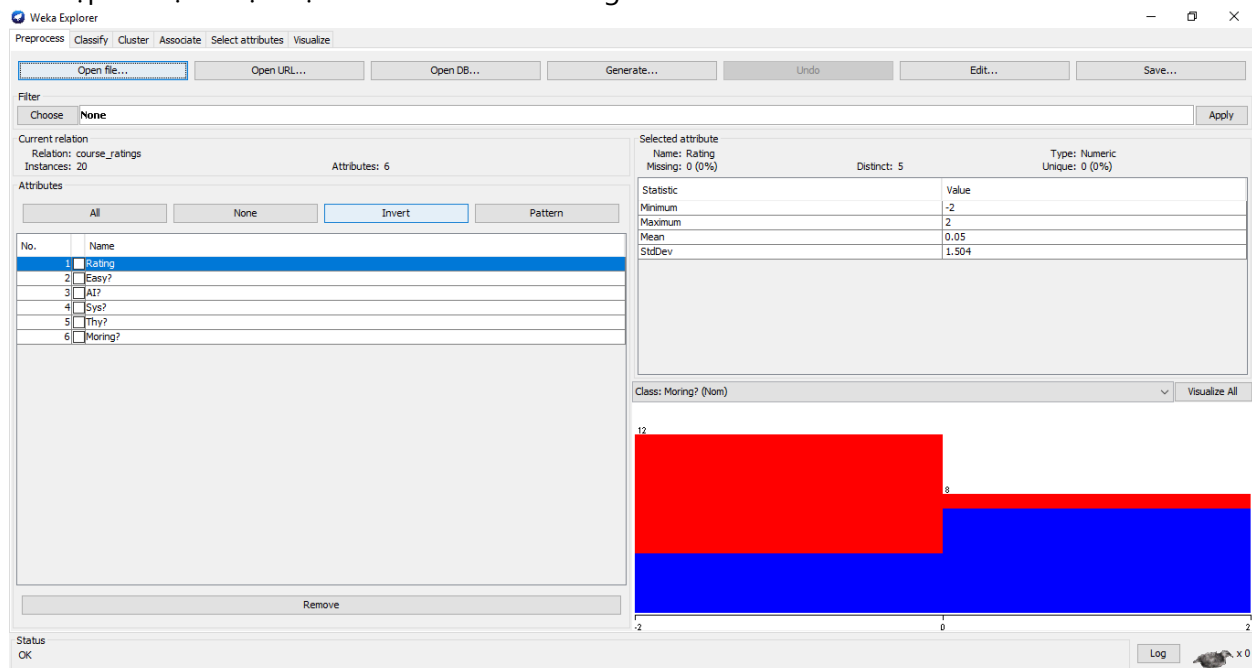
Chi tiết bài tập

2.1

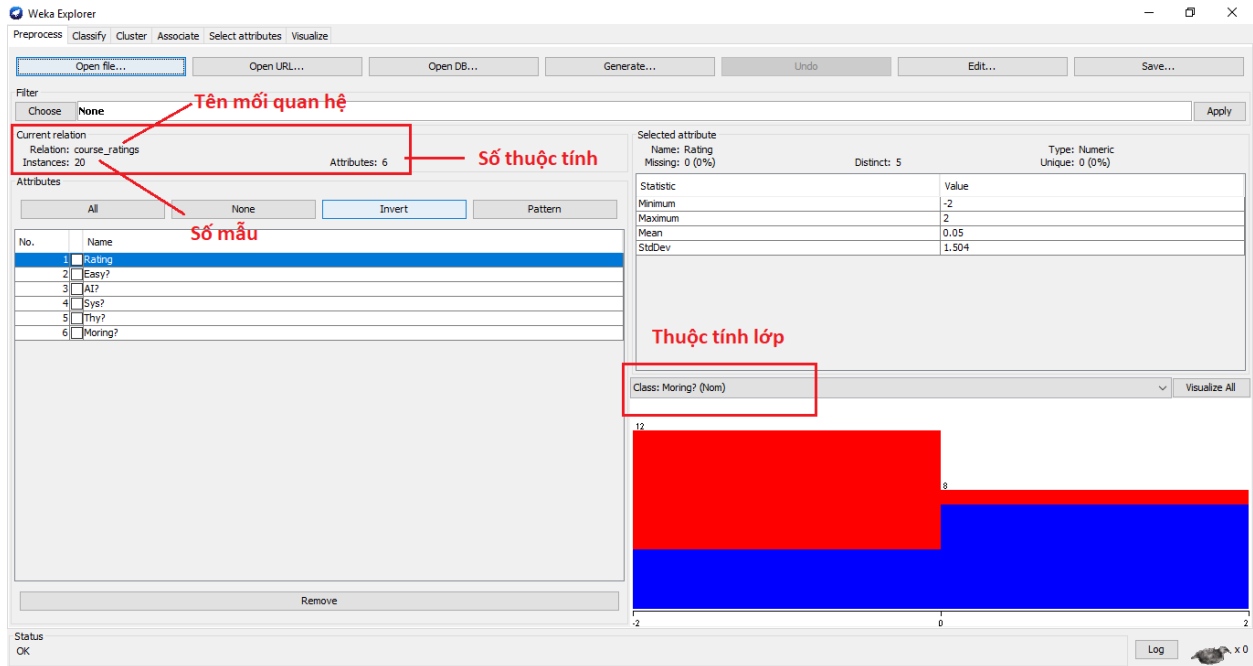
Báo cáo viết

2.1.1 Tạo tập tin arff từ tập dữ liệu Course Ratings

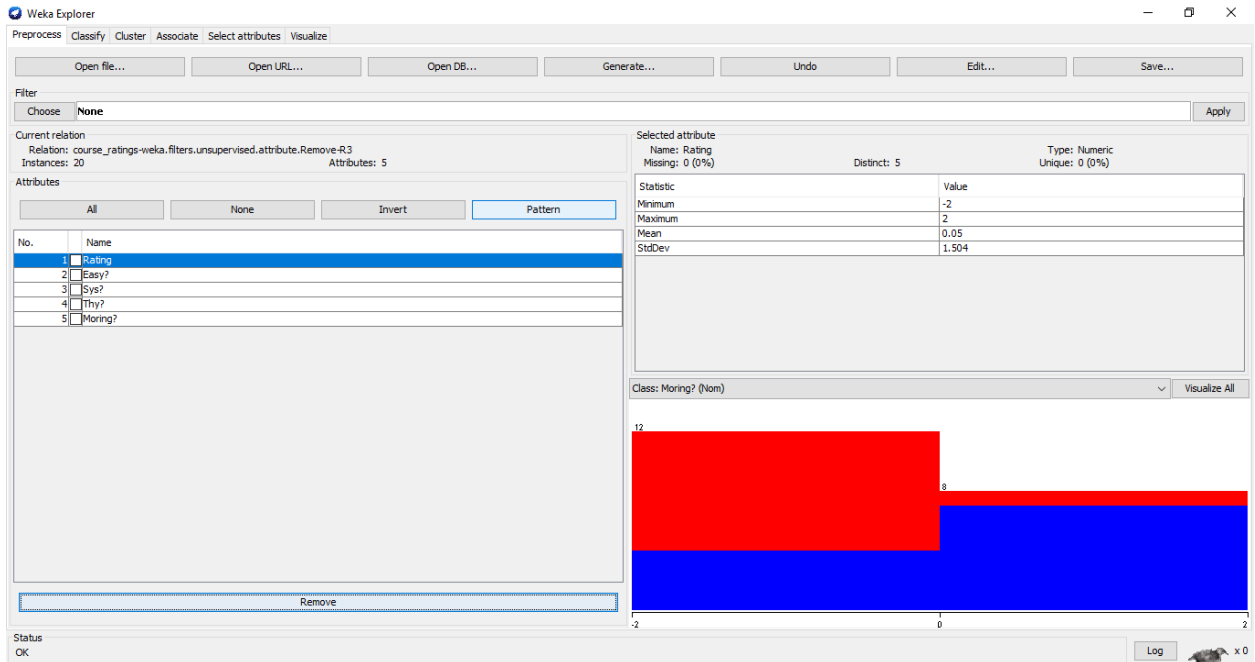
1. Tập dữ liệu được đọc vào Weka thành công



2. - Tên của mối quan hệ (relation) trong dữ liệu là: course_ratings
 - Tập dữ liệu có 20 mẫu (instances)
 - Tập dữ liệu có 6 thuộc tính (attributes)
 - Thuộc tính Morning? là thuộc tính lớp (class)

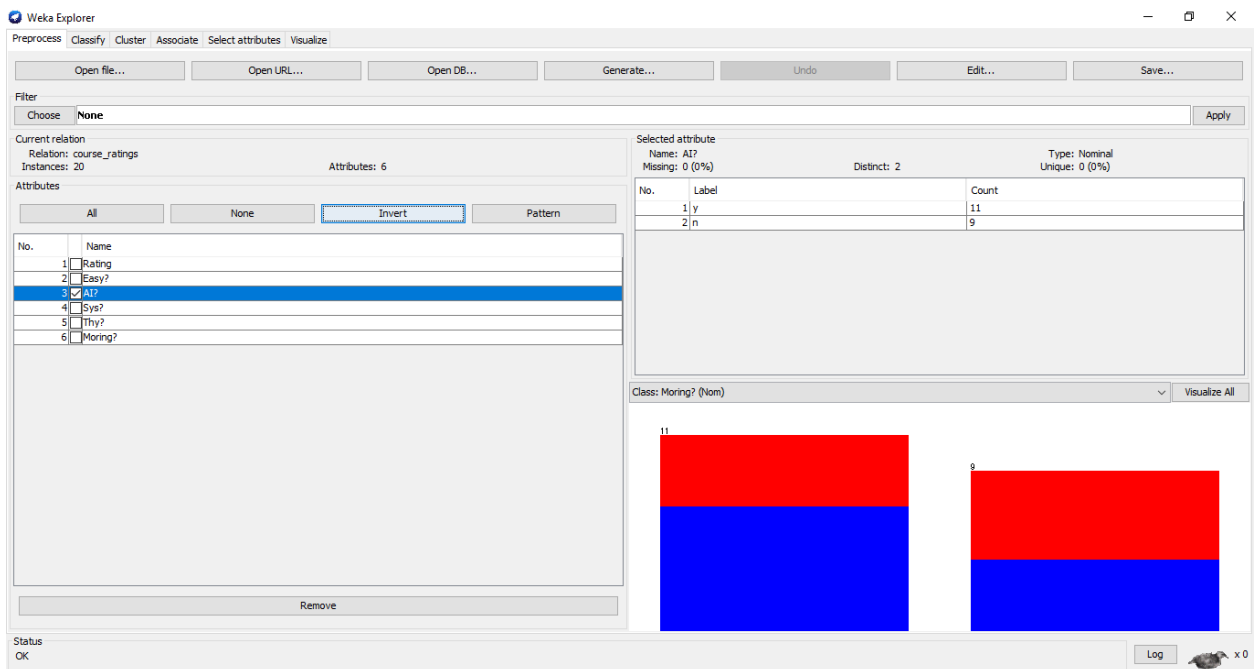


3. Nhận thấy điều gì đáng chú ý khi quan sát các thông tin thống kê và đồ thị trình diễn?
Đồ thị thống kê và các thông tin thống kê được thay đổi theo thuộc tính được chọn. Đồ thị trình diễn biểu thị các thông số đã được đề cập ở thông tin thống kê
4. Khi thuộc tính lớp là thuộc tính đầu tiên bên trái thì đồ thị trình diễn chức năng Visualize All không cho ra kết quả như khi ta chọn thuộc tính lớp là thuộc tính cuối cùng như trước.
5. Thao tác xóa:
Giả sử ta muốn xóa thuộc tính AI?
Kết quả:



Thao tác đảo ngược:

Giả sử ban đầu ta chọn thuộc tính AI? như hình



Sau khi thao tác đảo ngược với Invert, ta được như sau:

Weka Explorer - Preprocess tab

Current relation: course_ratings
Instances: 20
Attributes: 6

Selected attribute: AI?
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	y	11
2	n	9

Class: Moring? (Nom) Visualize All

Hủy thao tác trở về trước đó:

Giả sử sau khi thao tác xóa đi thuộc tính AI? như thao tác xóa

Weka Explorer - Preprocess tab

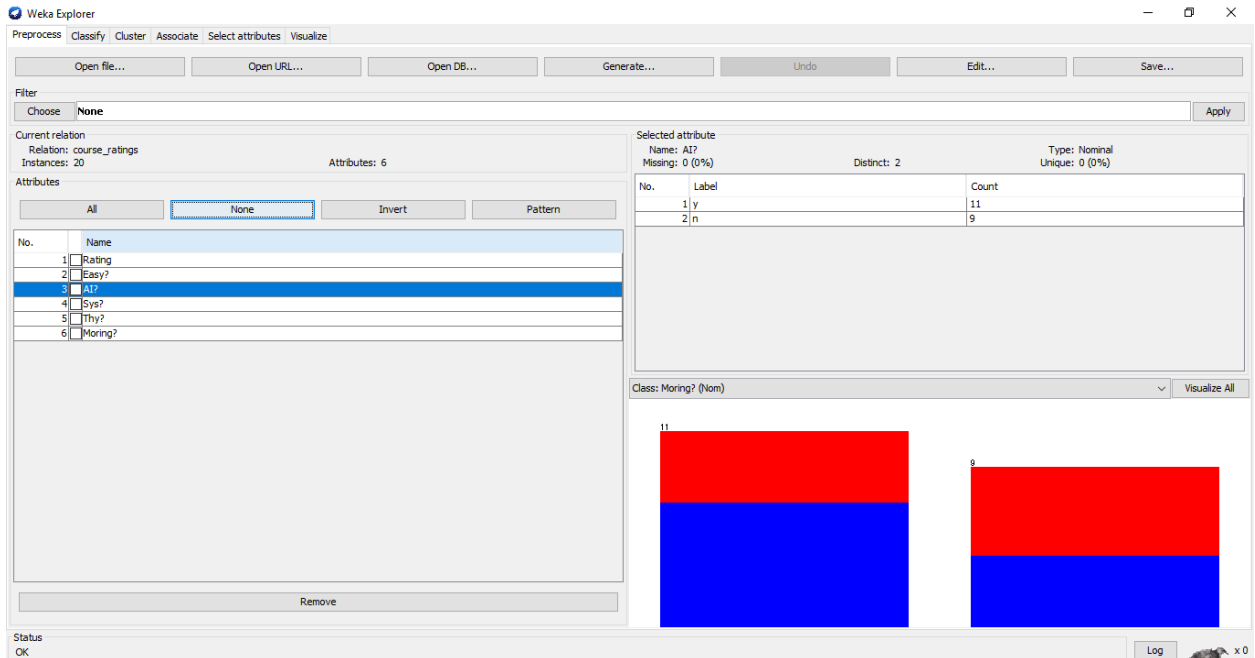
Current relation: course_ratings-weka.filters.unsupervised.attribute.Remove-R3
Instances: 20
Attributes: 5

Selected attribute: Rating
Missing: 0 (0%)
Distinct: 5
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	-2
Maximum	2
Mean	0.05
StdDev	1.504

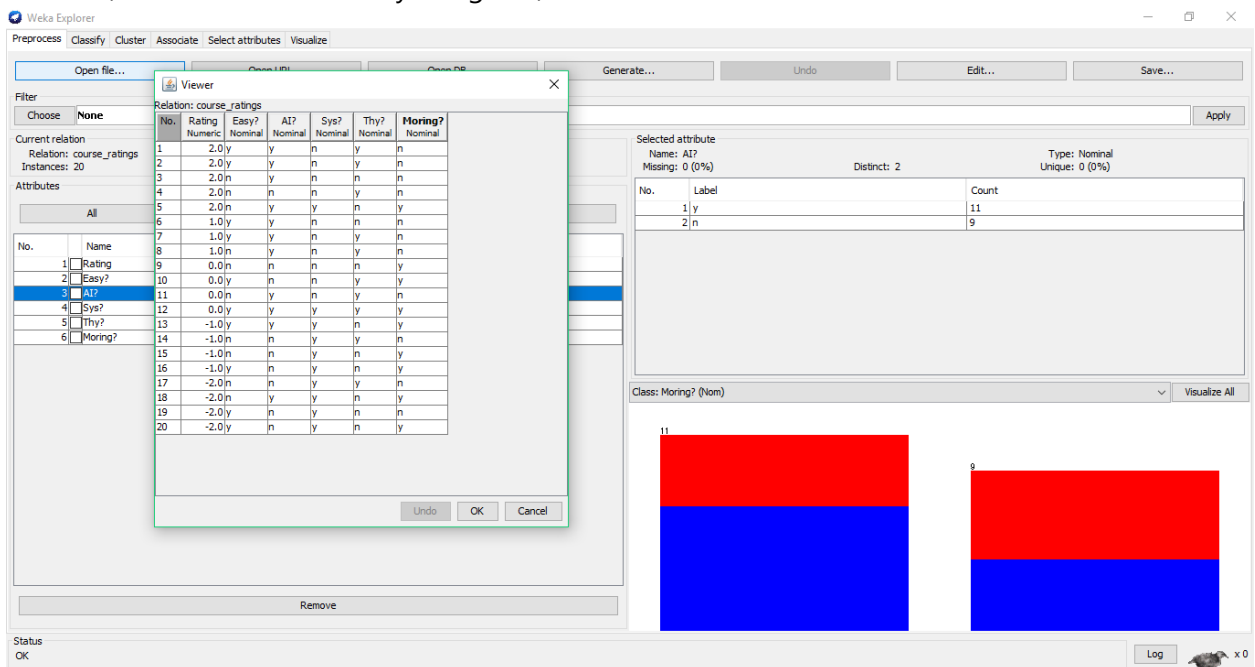
Class: Moring? (Nom) Visualize All

nhấn undo ta được trở về trạng thái trước đó



Chỉnh sửa:

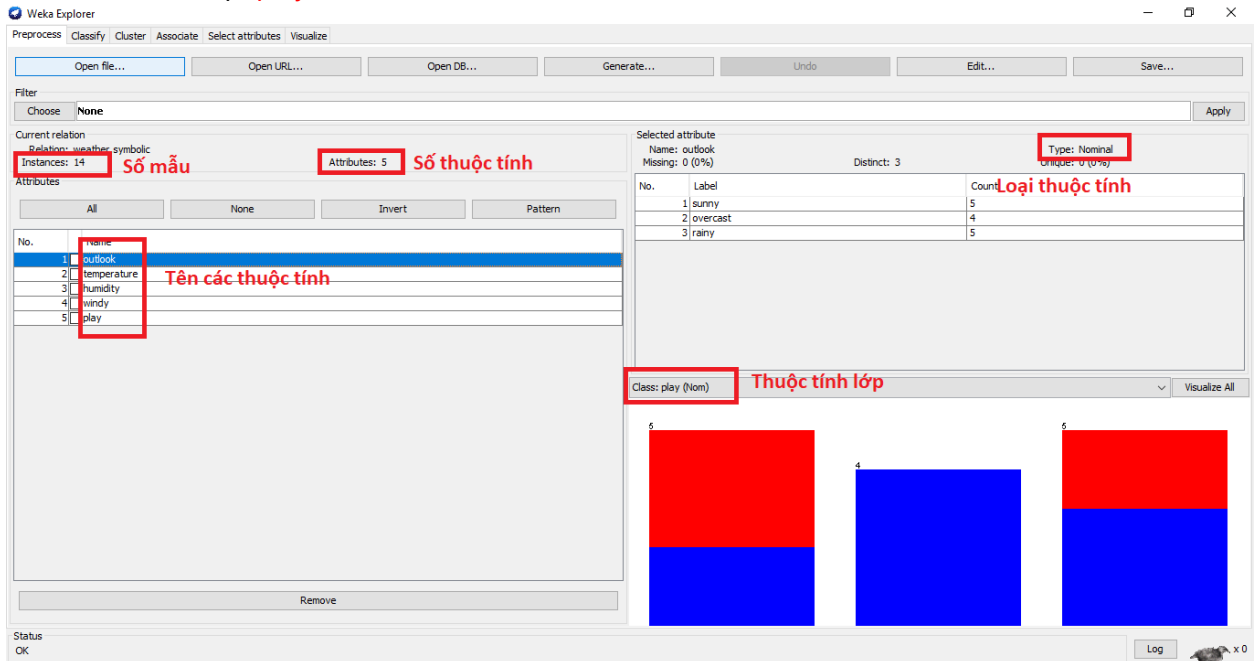
Ta nhấn chọn Edit, sau đó xuất hiện cửa sổ View như hình, thay đổi các giá trị của từng mẫu bất kì và chọn OK nhấn lưu để thay đổi giá trị



2.1.2 Khảo sát tập dữ liệu Weather (thuộc tính rời rạc)

6.

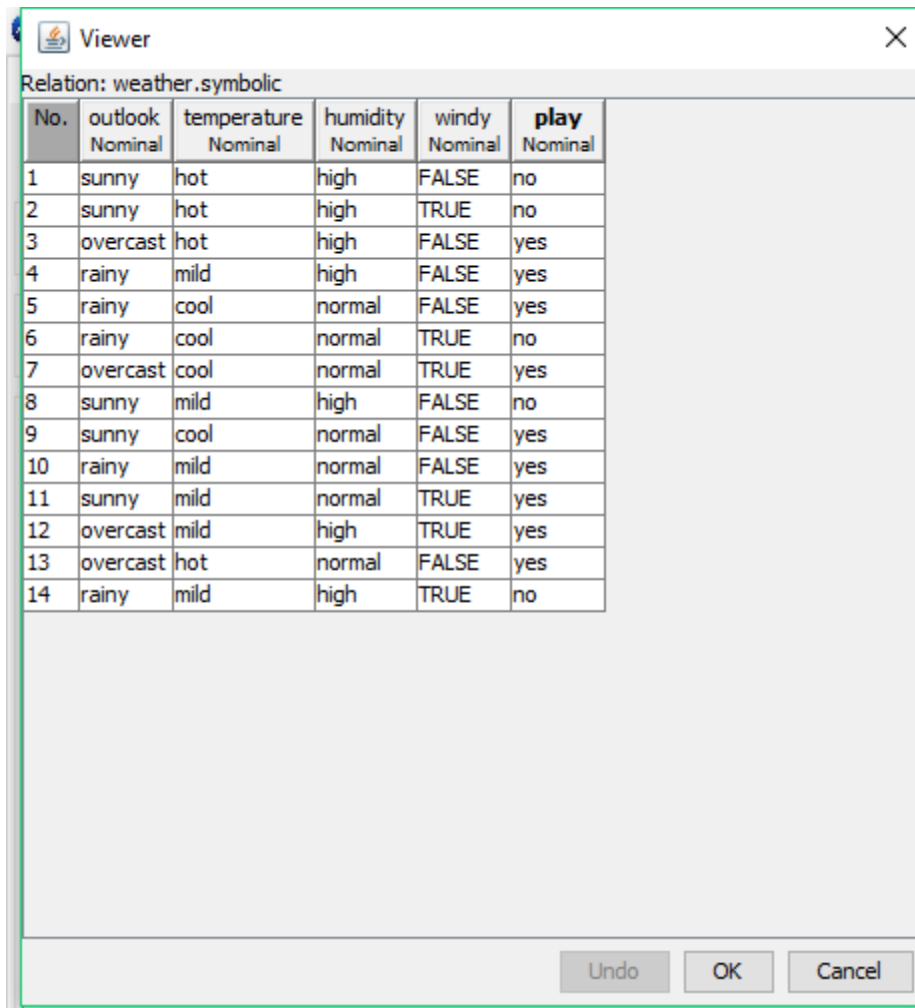
- Tập dữ liệu có **14** mẫu
- **5** thuộc tính
- Các thuộc tính gồm: **outlook** (kiểu nominal), **temperature** (kiểu nominal), **humidity** (kiểu nominal), **windy** (kiểu nominal), **play** (kiểu nominal)
- Thuộc tính lớp: **play**



7. Cột viewer

- Chức năng cột đầu tiên của cửa sổ Viewer là đánh số các mẫu
- Lớp của mẫu thứ 8 là **play**

8. Cửa sổ Viewer trước khi thực hiện:

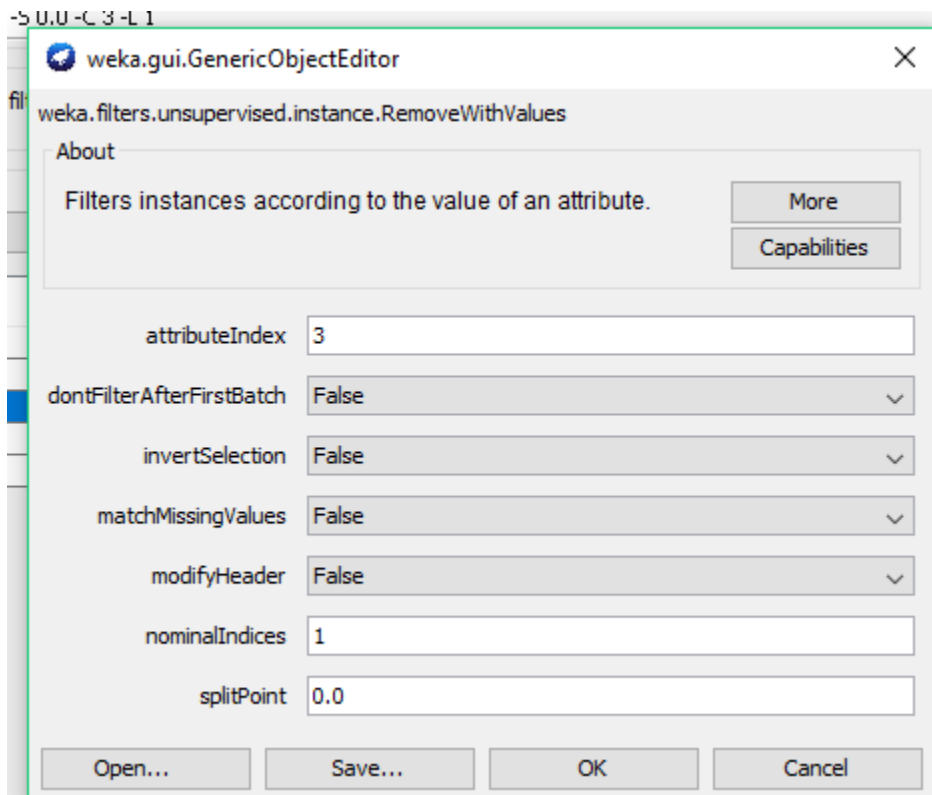


Relation: weather.symbolic

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

Undo OK Cancel

Tham số đã lập:



Kết quả của sổ Viewer sau khi đã lọc:

Viewer

Relation: weather.symbolic-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C...

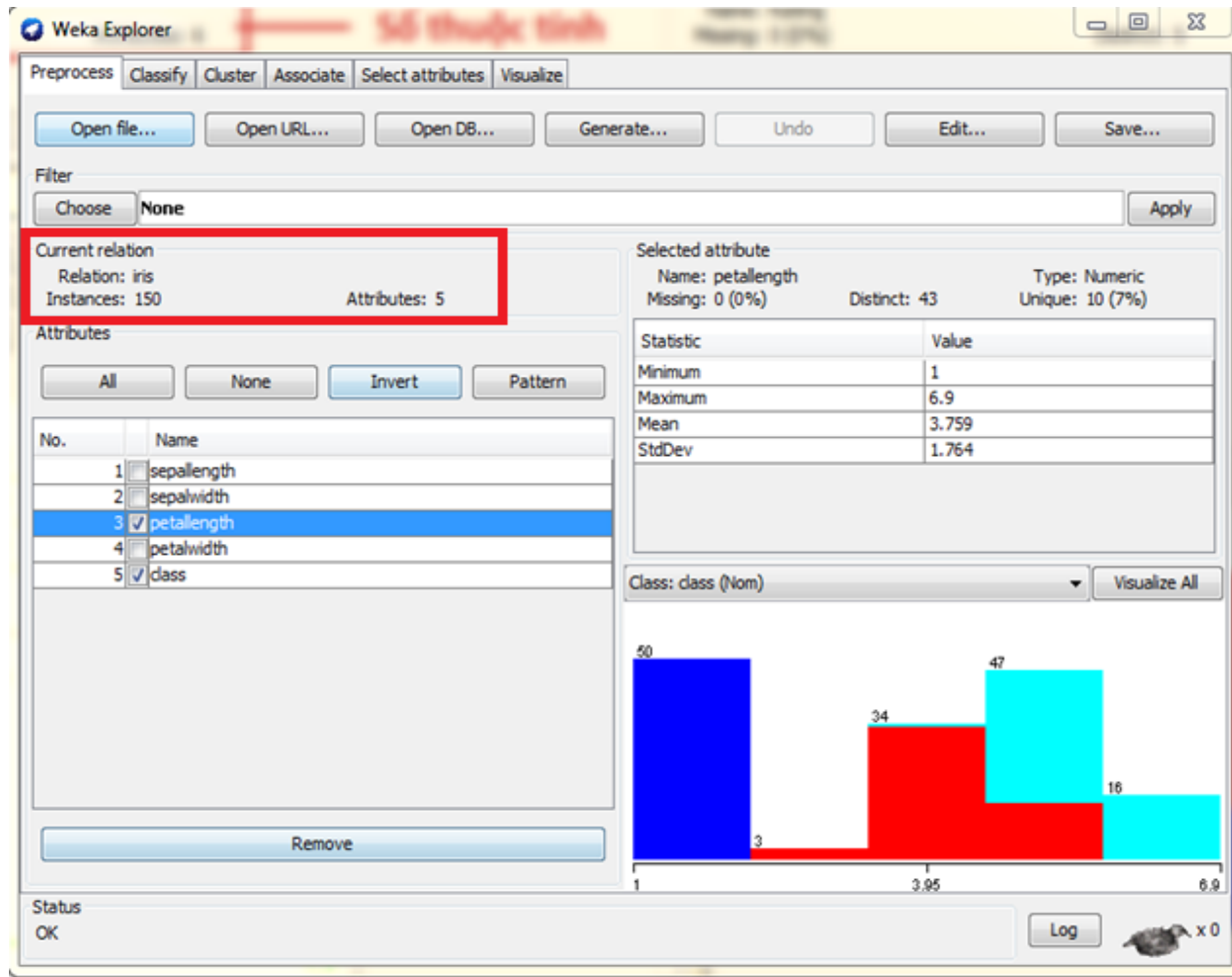
No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	rainy	cool	normal	FALSE	yes
2	rainy	cool	normal	TRUE	no
3	overcast	cool	normal	TRUE	yes
4	sunny	cool	normal	FALSE	yes
5	rainy	mild	normal	FALSE	yes
6	sunny	mild	normal	TRUE	yes
7	overcast	hot	normal	FALSE	yes

Undo OK Cancel

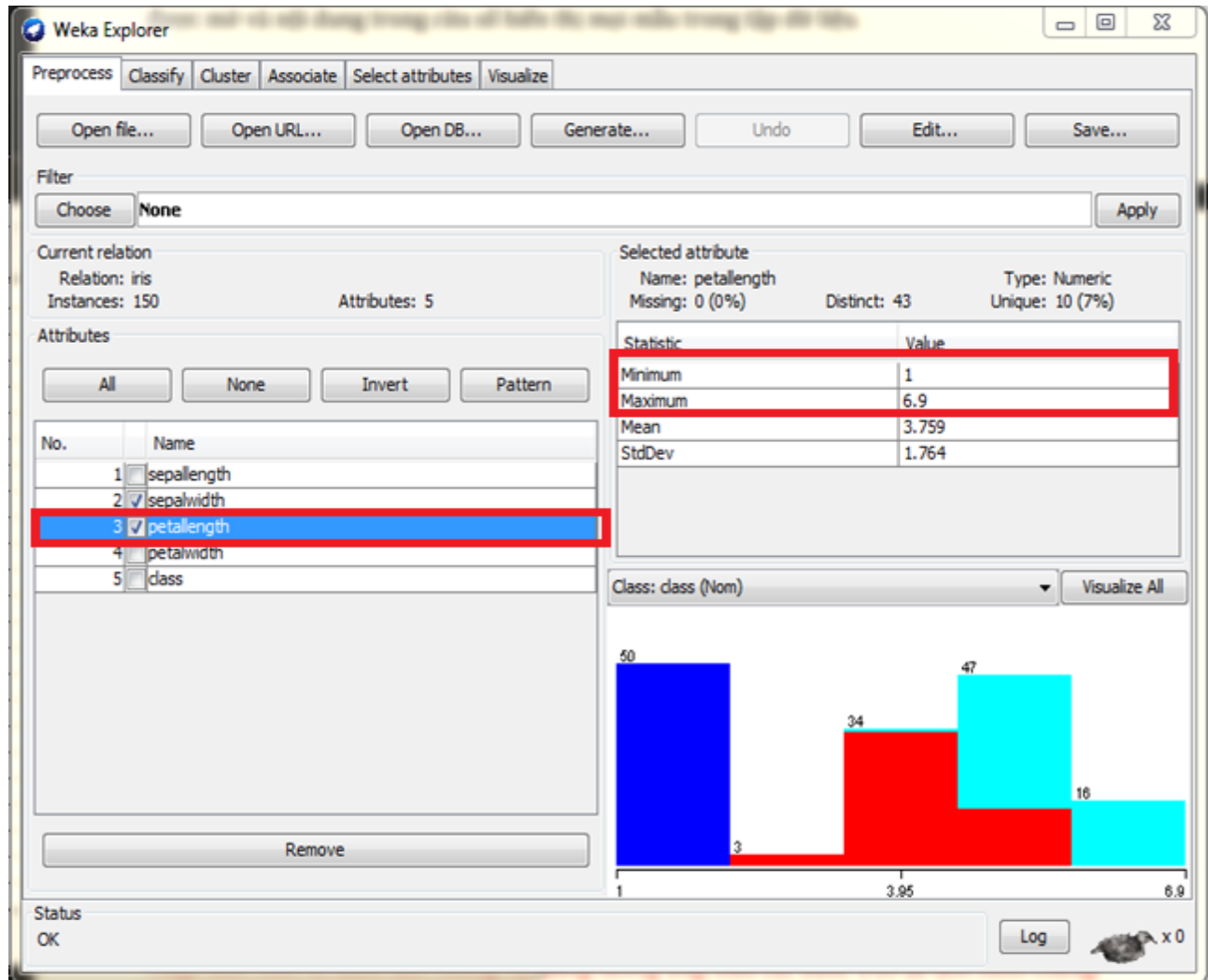
2.1.3 Khảo sát tập dữ liệu Iris (thuộc tính số liên tục)

Đã mở file dữ liệu thành công

- Tập dữ liệu trên có **150** mẫu (Instances)
- Có **5** thuộc tính(Attributes) : **sepalength, sepalwidth, petallength, petalwidth, class**



-Miền giá trị của thuộc tính petallength là từ : **1 đến 6.9**



- Mẫu dữ liệu trên có **4** thuộc tính số (Num): **sepalength, sepalwidth, petallength, petalwidth** và **1** thuộc tính rời rạc (Nom): **Class**
- Tên của thuộc tính lớp là : **Class**

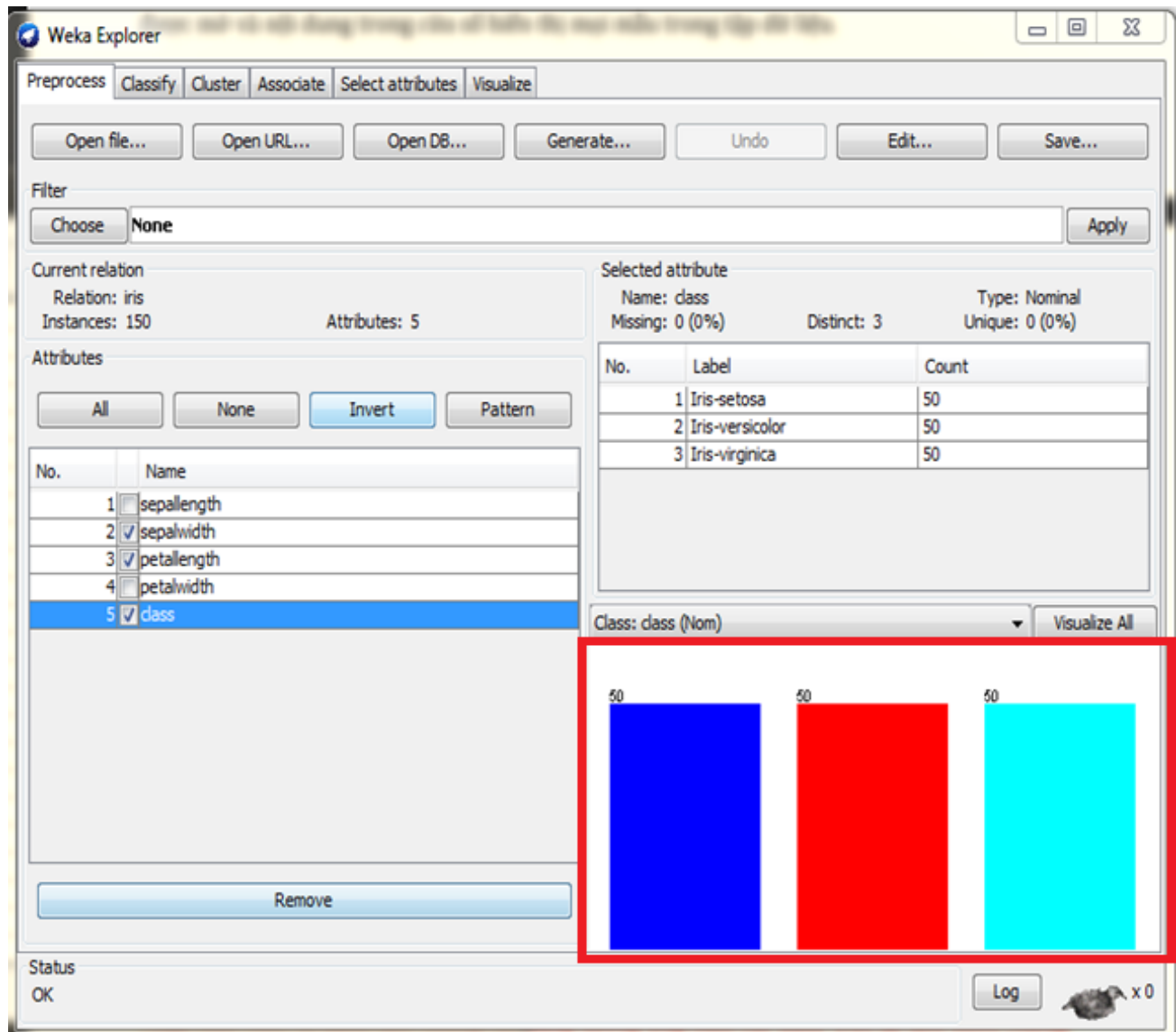
Viewer

Relation: iris

No.	sepalength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa
23	4.6	3.6	1.0	0.2	Iris-setosa
24	5.1	3.3	1.7	0.5	Iris-setosa

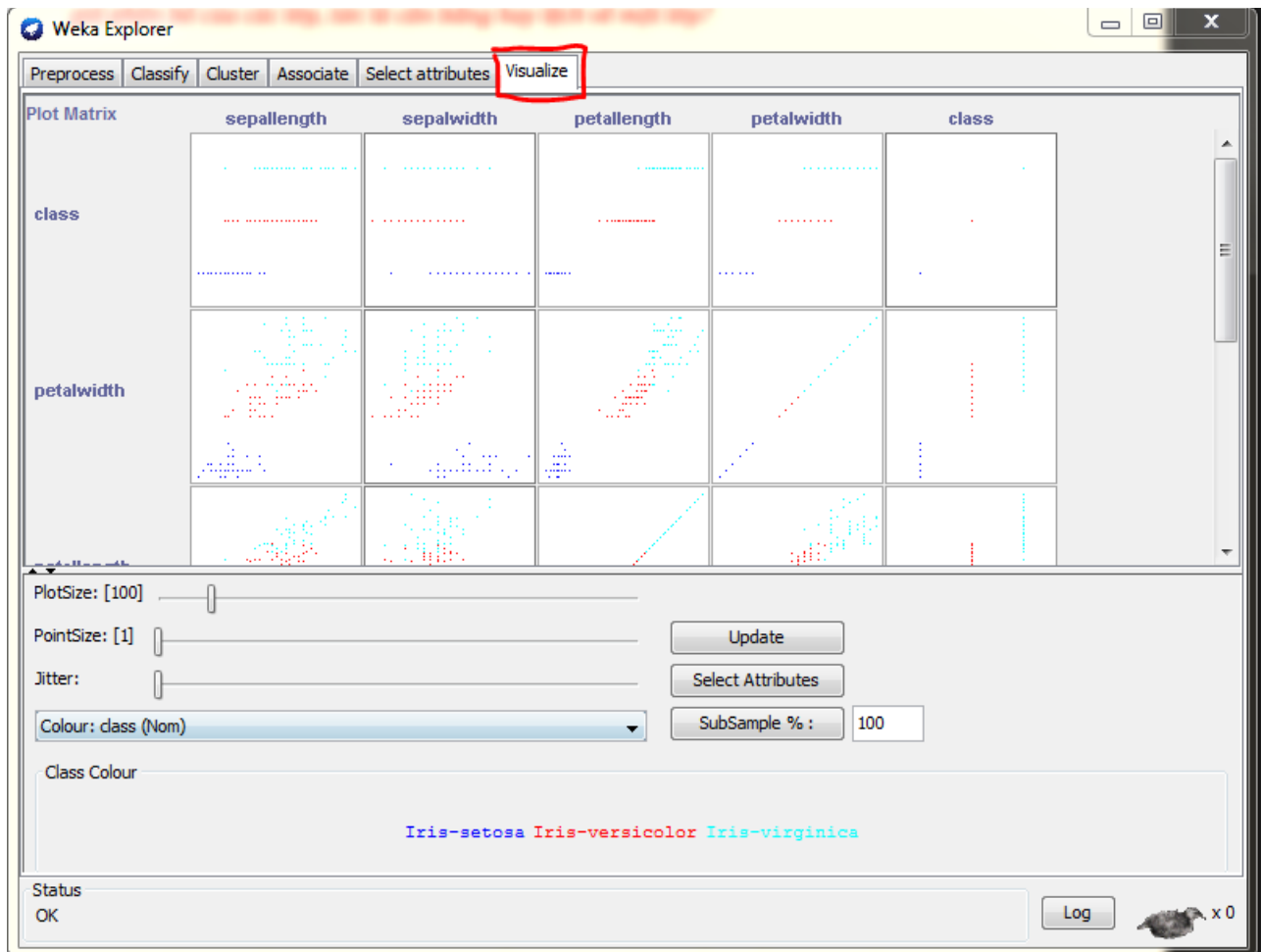
Undo OK Cancel

- **Đánh giá phân bố các lớp là cân bằng**

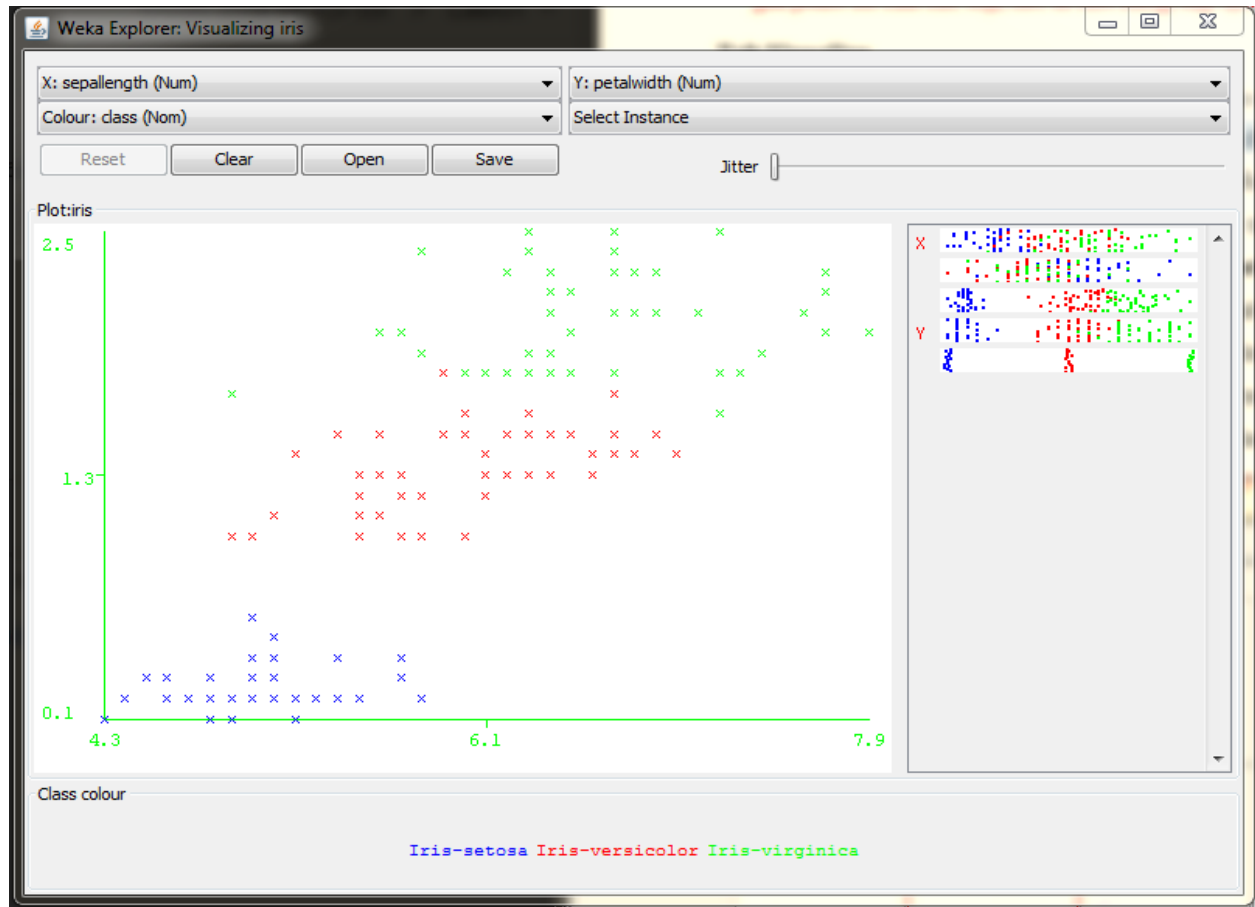


Tab Visualize

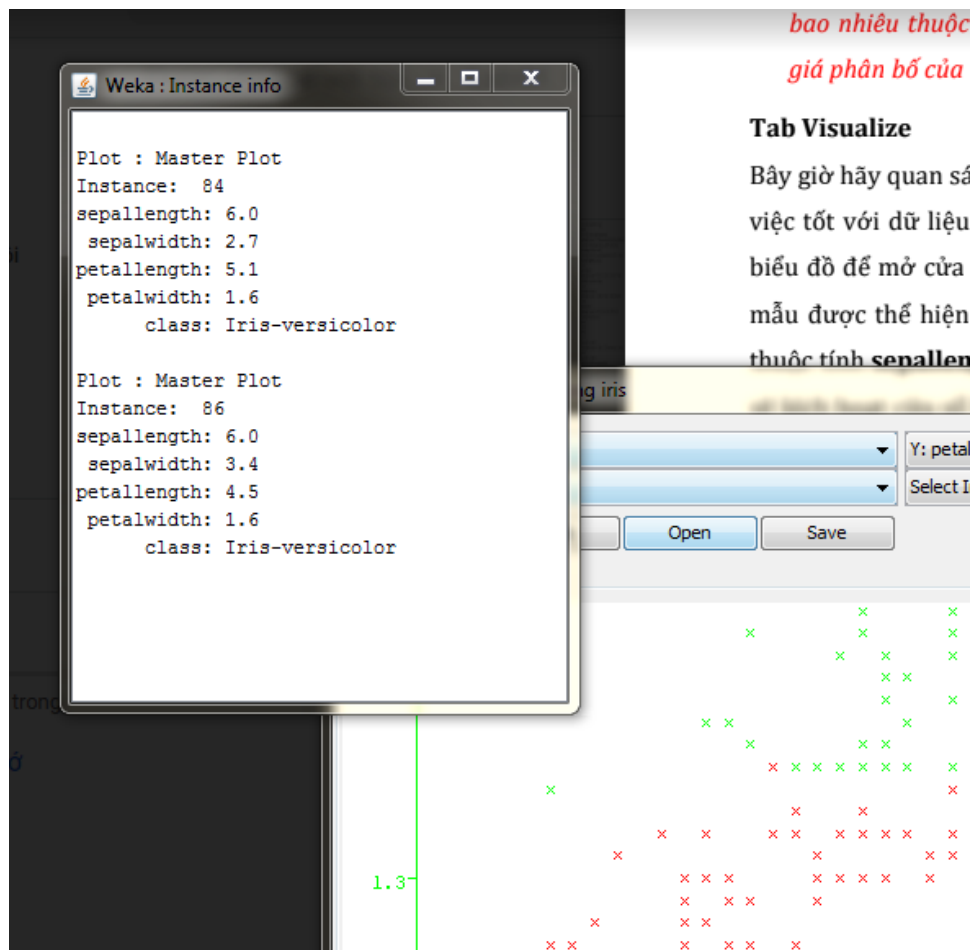
Mở tab Visualize



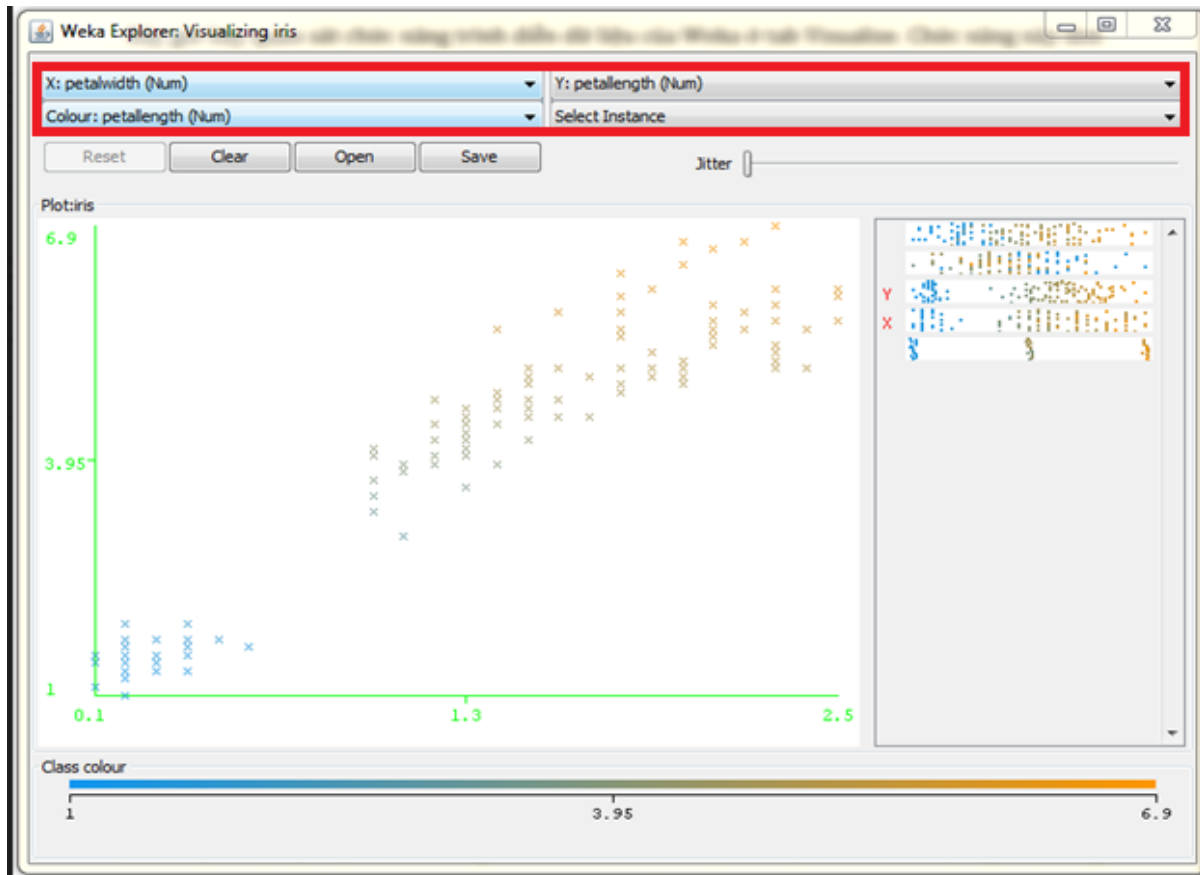
Đã chọn biểu đồ theo yêu cầu



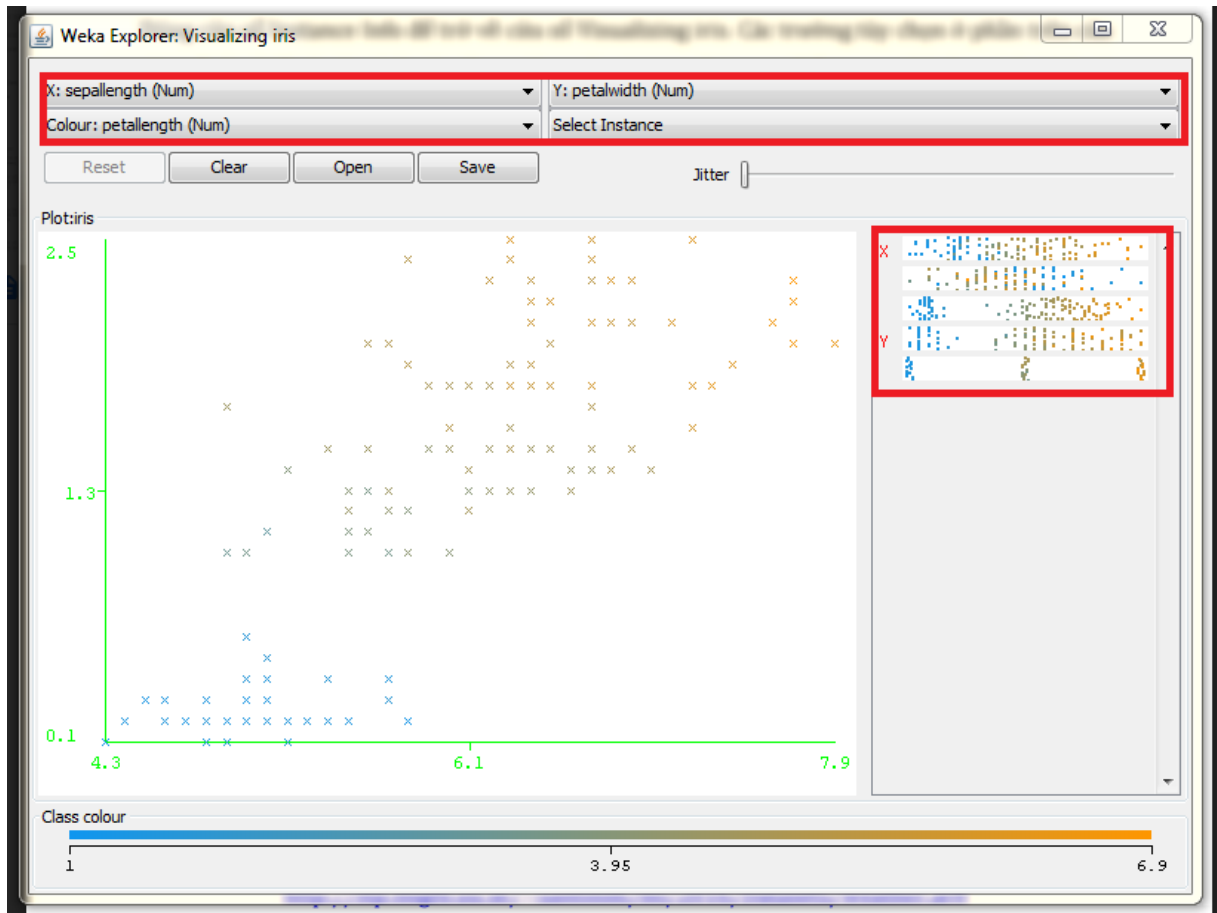
Nhấp chuột đôi vào dấu x:



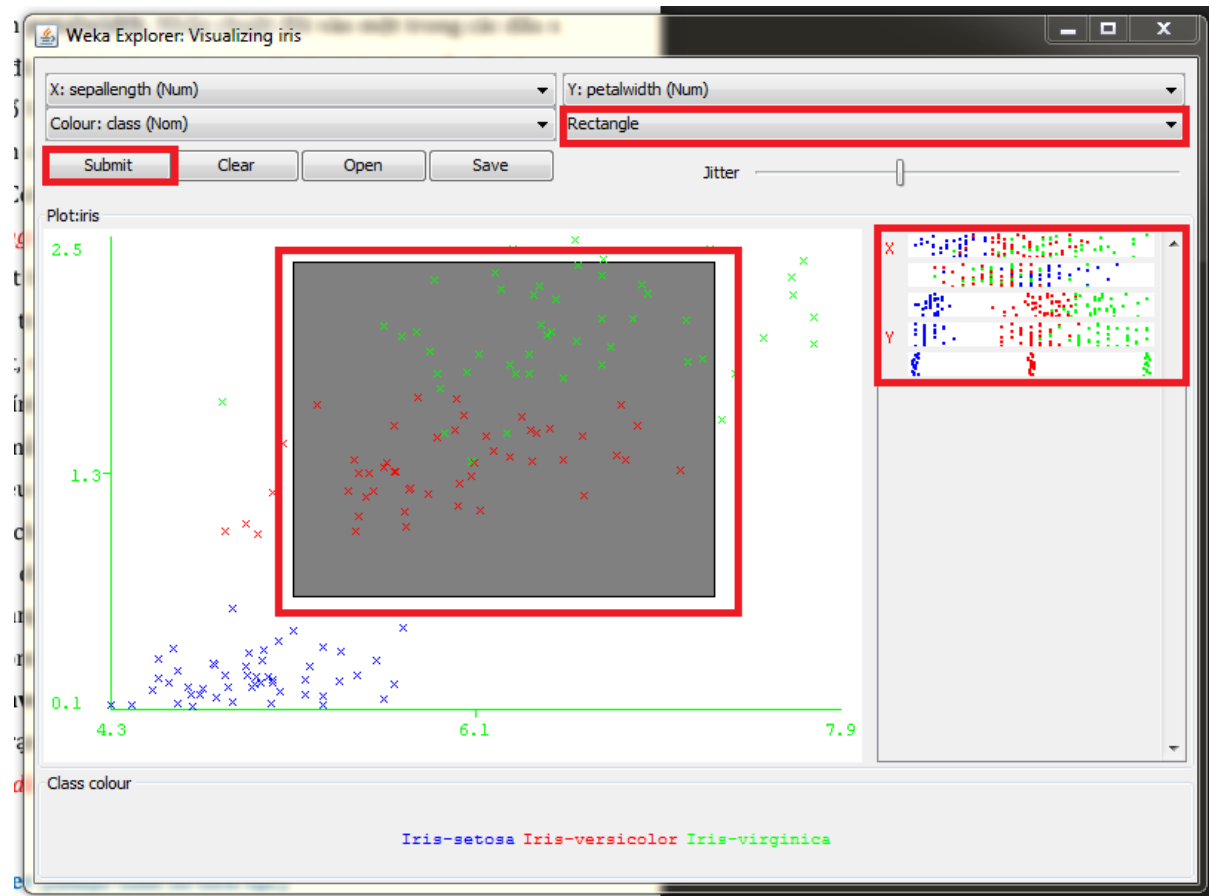
Đổi trục x thành petalwidth (Num) và trục y thành petallength (Num) và thay đổi trường class để đổi kí hiệu màu



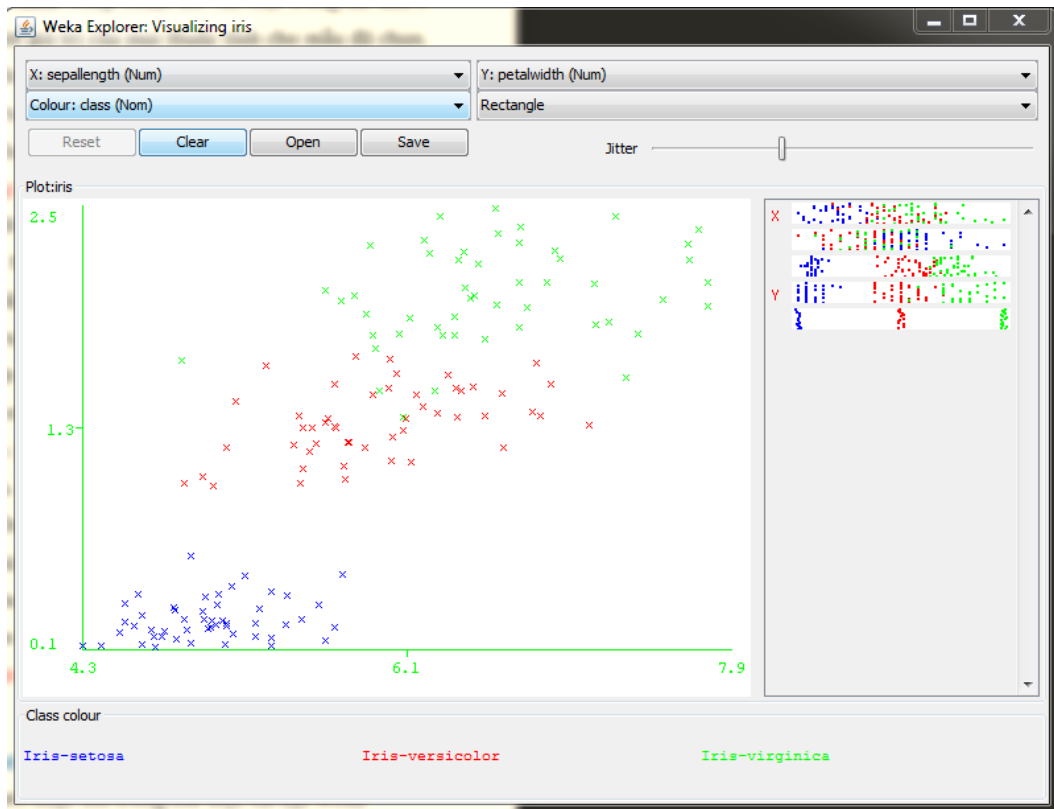
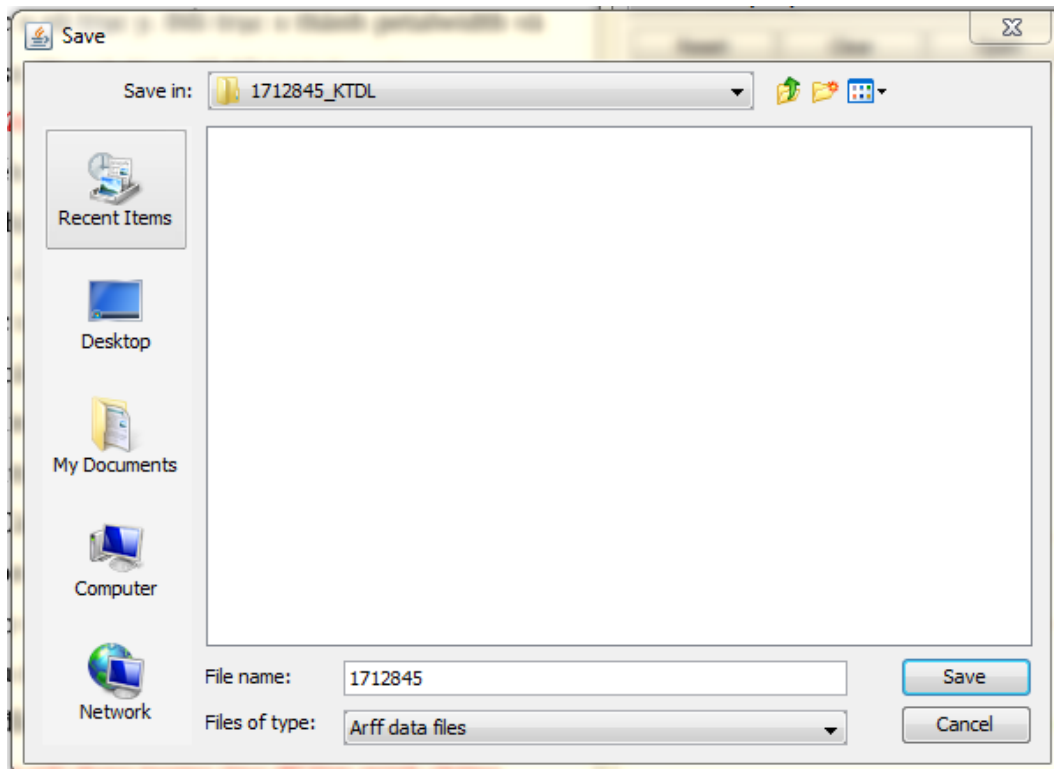
Sử dụng scatter plot set lại thuộc tính



Trải nghiệm jitter và set Rectangle

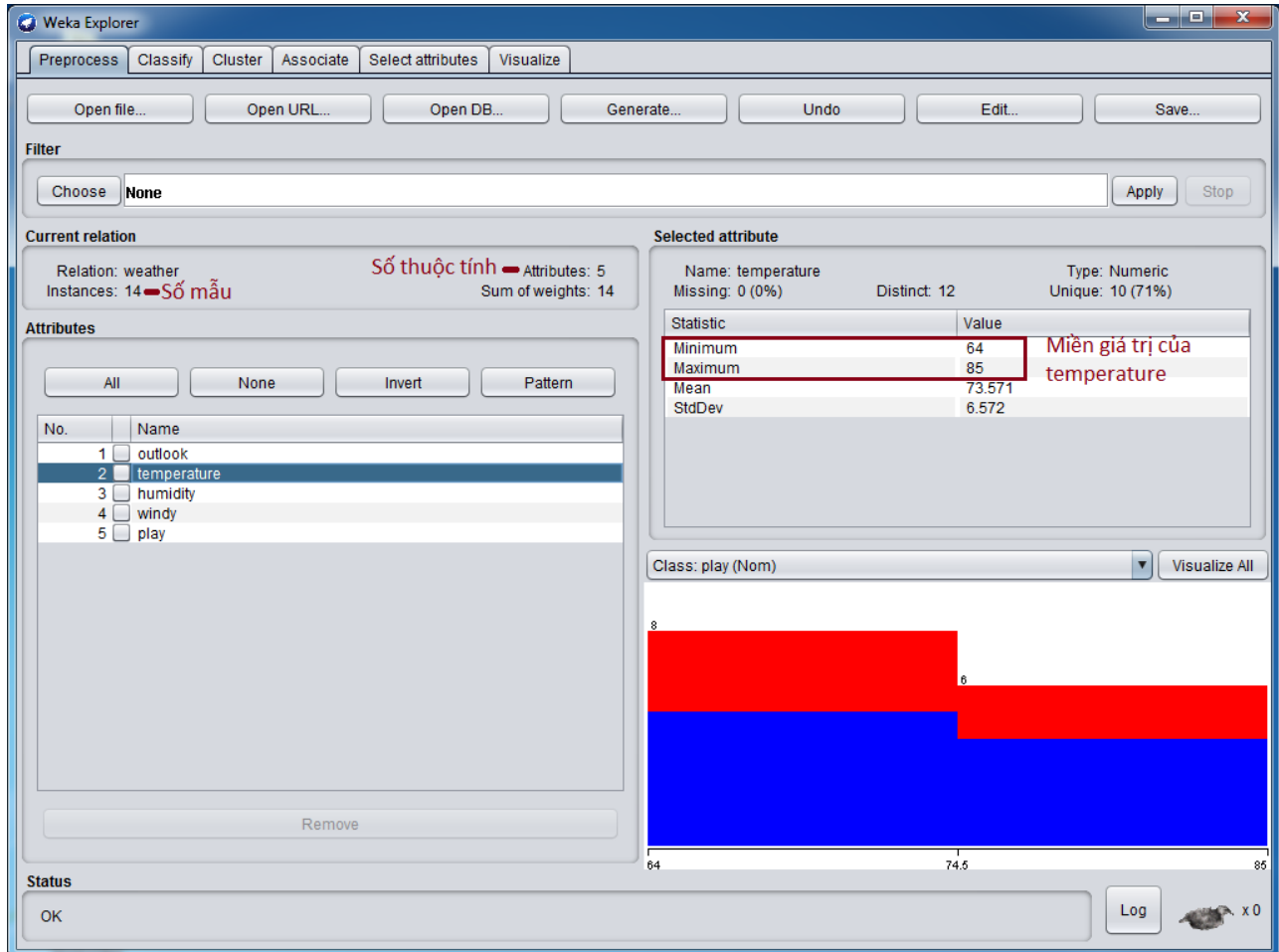


Sao lưu và reset lại



2.1.4 Khảo sát tập dữ liệu Weather (thuộc tính số liên tục)

Mô tả tập dữ liệu.



Tập dữ liệu có bao nhiêu mẫu?

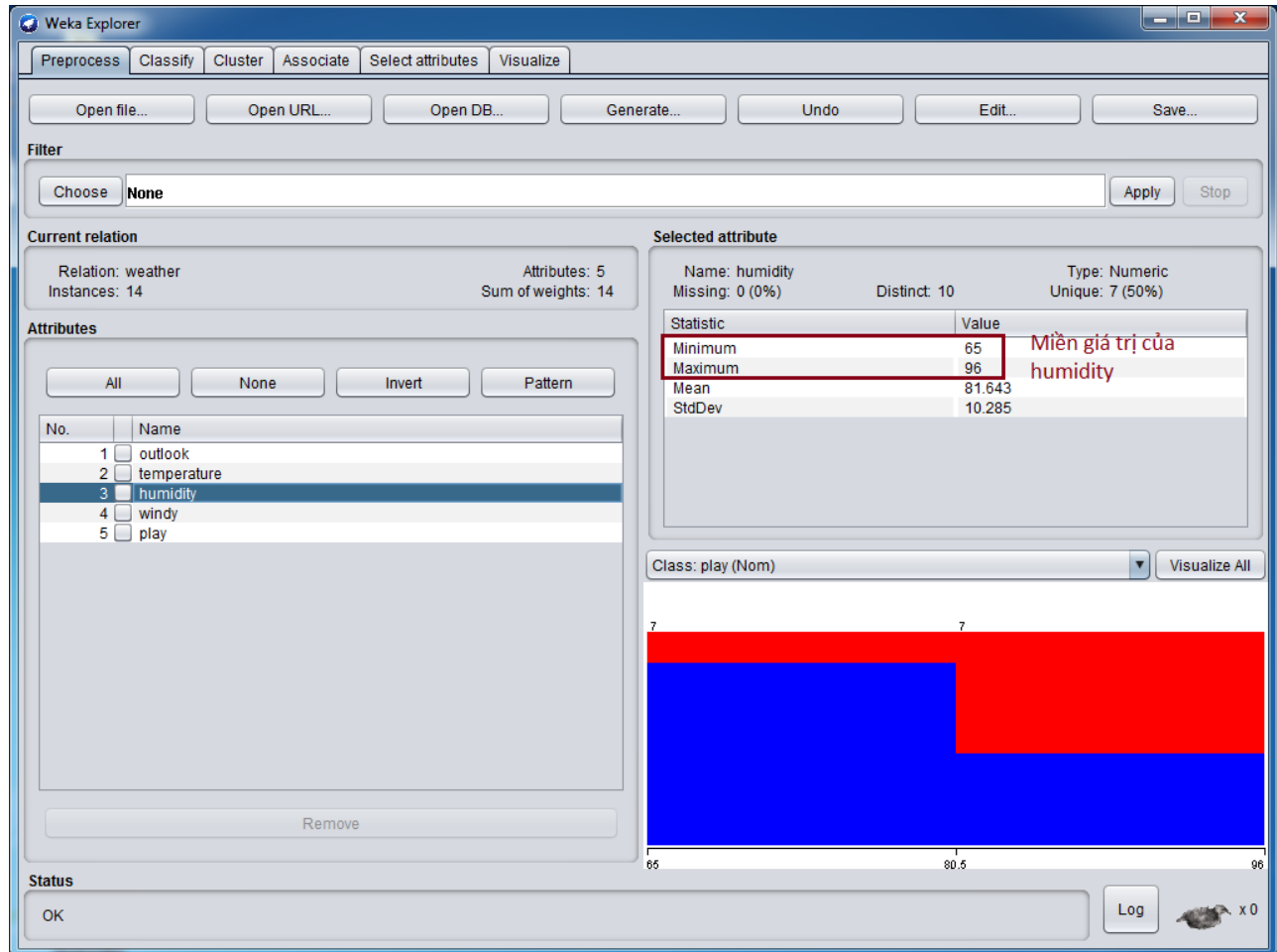
- Tập dữ liệu có **14** mẫu

Bao nhiêu thuộc tính?

- **5** thuộc tính

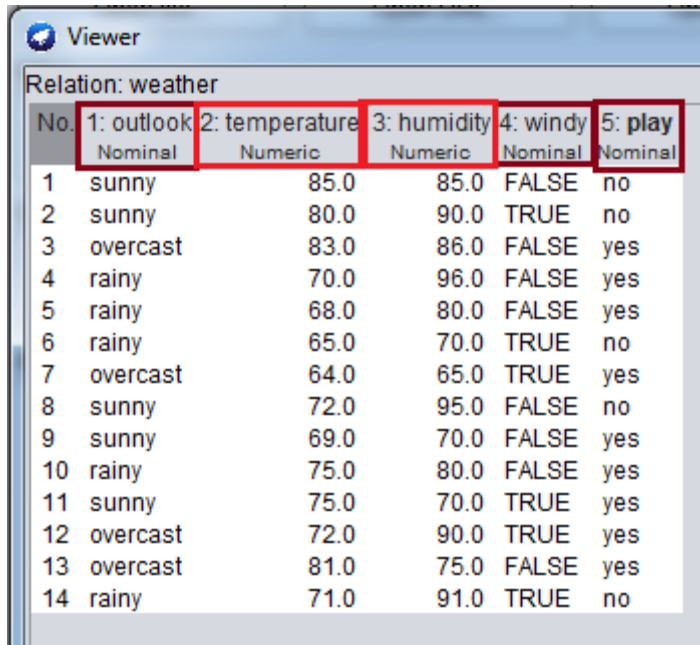
Miền giá trị của thuộc tính temperature là gì?

- Miền giá trị của thuộc tính **temperature** là **[64 - 85]** độ F



Miền giá trị của thuộc tính humidity là gì?

- Miền giá trị của thuộc tính **humidity** là **[65 - 96] %**



No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

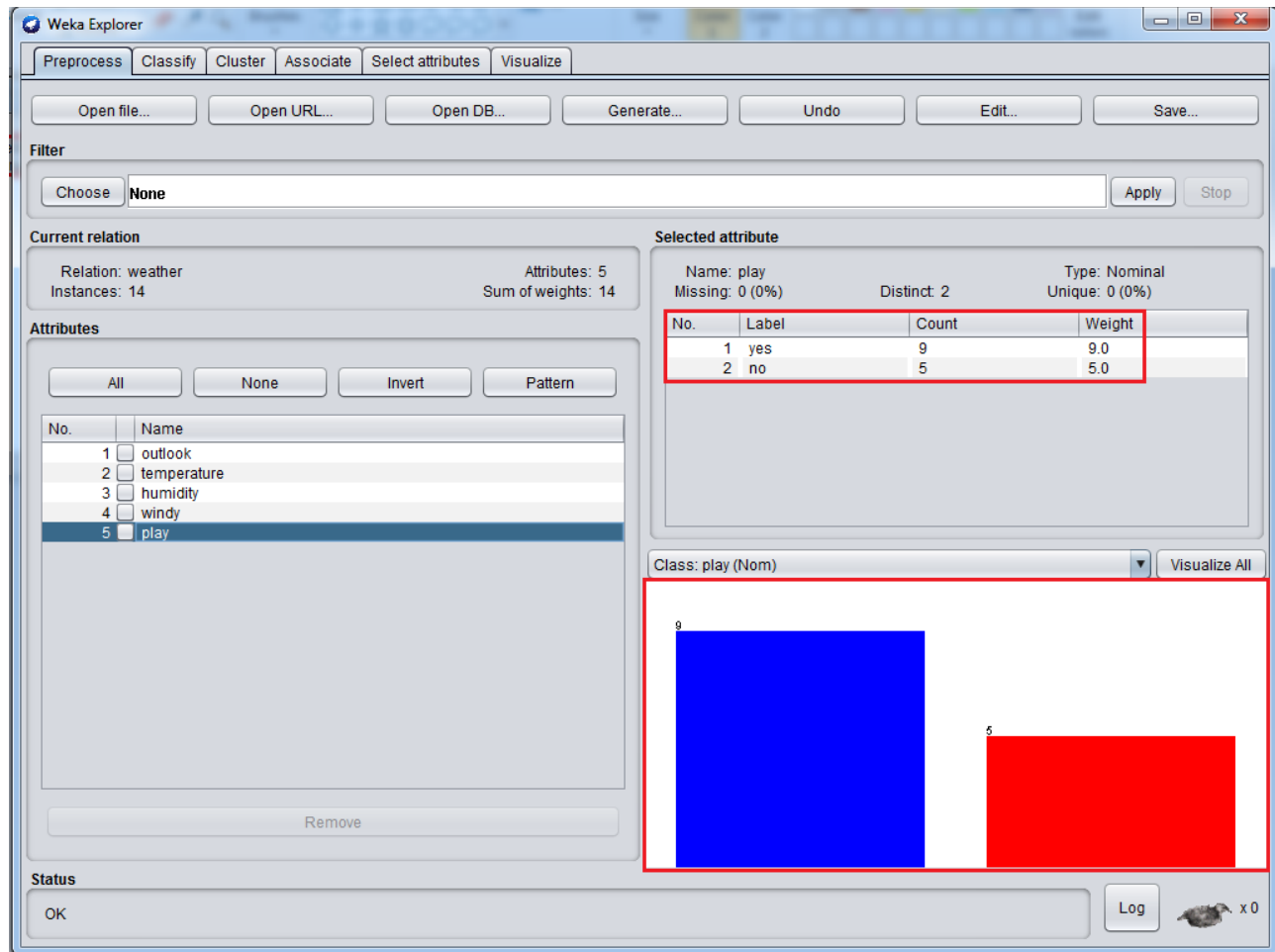
Tập dữ liệu có bao nhiêu thuộc tính số và bao nhiêu thuộc tính rời rạc?

- Tập dữ liệu có 2 thuộc tính số: temperature, humidity và 3 thuộc tính rời rạc: outlook, windy, play

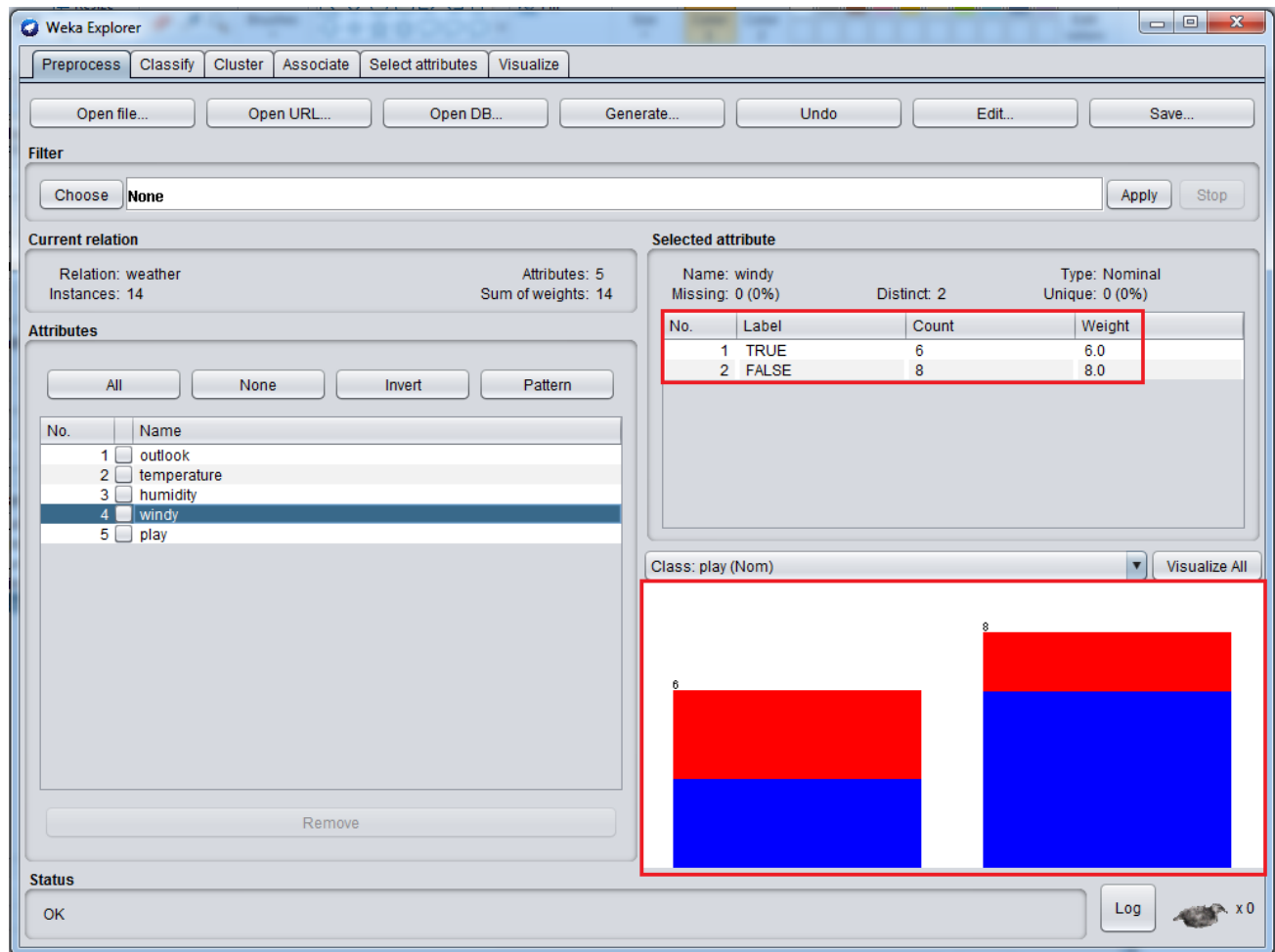
Tên của thuộc tính lớp là gì?

- Tên của thuộc tính lớp là play

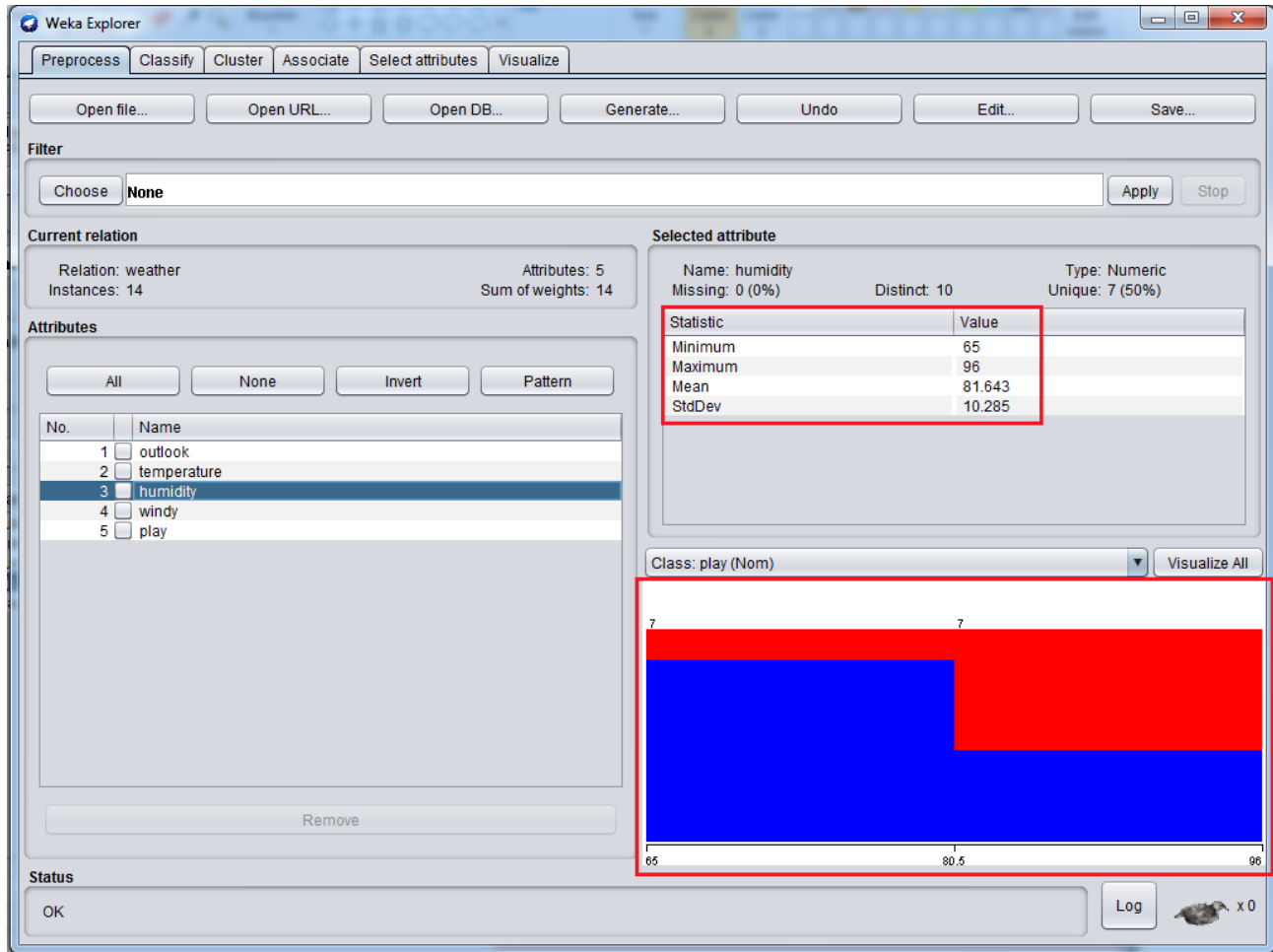
Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?



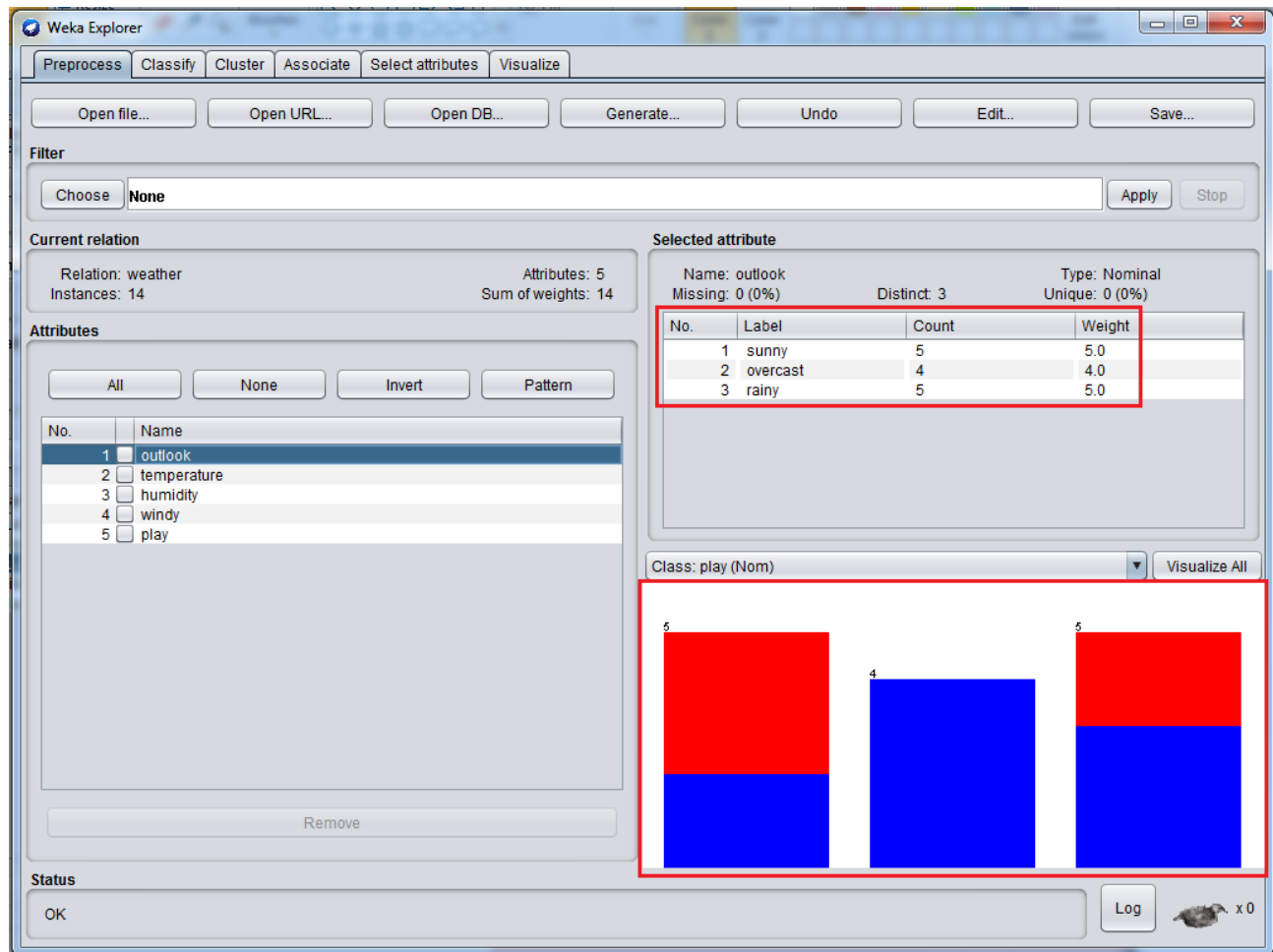
Giá trị ở lớp play này lệch nhiều về giá trị yes



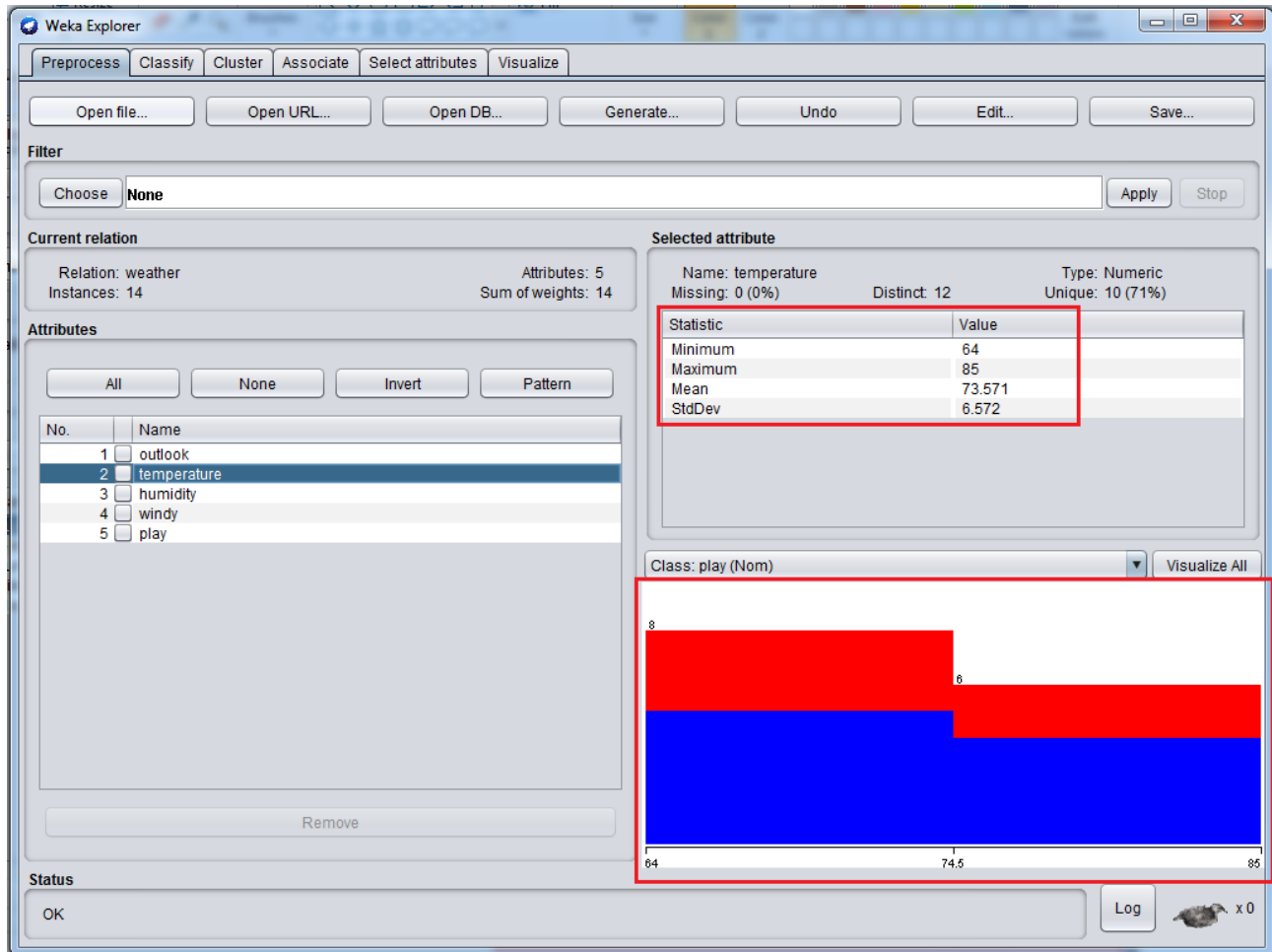
Ở thuộc tính windy khi giá trị là 6 thì cân bằng, khi giá trị là 8 thì lệch về giá trị yes của lớp play.



Ở Thuộc tính humidity khi giá trị thấp hơn 80.5 thì lệch về thì giá trị yes của lớp play, khi giá trị cao hơn 80.5 thì thì lệch về thì giá trị no của lớp play.



Ở thuộc tính outlook, khi giá trị là sunny thì lệch về giá trị no của lớp play, khi giá trị là overcast thì lệch hoàn toàn về giá trị yes của lớp play, khi giá trị là rainy thì lệch nhẹ về giá trị yes của lớp play



Ở thuộc tính temperature khi giá trị thấp hơn 74.5 thì giá trị no của lớp play tăng rõ ràng, nhưng tổng thể vẫn lệch về phía giá trị yes của lớp play.

2.2

Nội dung hướng dẫn sử dụng cài đặt

Chức năng, tham số vào hàm, giá trị trả về của từng hàm đã comment trong file cài đặt. Phần này chỉ trình bày hướng dẫn sử dụng tham số dòng lệnh.

Bài nộp có bao gồm một số file .csv để test cho phần 2.2.1: nz_weather.csv, school_earnings.csv, styled-line.csv.

2.2.1 Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản

a. Chuẩn hóa min-max trên danh sách thuộc tính chỉ định.

Cú pháp tham số dòng lệnh: 24_B1.py input.csv output.csv min-max propList{id,age}

Với id, age là ví dụ về 2 thuộc tính trong tập dữ liệu

b. Chuẩn hóa Z-scores trên danh sách thuộc tính chỉ định.

Cú pháp tham số dòng lệnh: 24_B1.py input.csv output.csv z-score propList{id,age}

c. Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ rộng trên danh sách thuộc tính chỉ định.

Cú pháp tham số dòng lệnh: 24_B1.py input.csv output.csv equal-width bin-N propList{id,age}

Với bin-N và N là số giỏ sẽ chia

d. Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ sâu trên danh sách thuộc tính chỉ định.

Cú pháp tham số dòng lệnh: 24_B1.py input.csv output.csv equal-depth bin-N propList{id,age}

e. Xóa các mẫu dữ liệu thiếu giá trị trên danh sách thuộc tính chỉ định.

Cú pháp tham số dòng lệnh: 24_B1.py input.csv output.csv remove propList{id,age}

f. Điền giá trị bị thiếu trên danh sách thuộc tính chỉ định, giá trị được điền là giá trị trung bình (mean) của thuộc tính nếu đó là thuộc tính số hoặc điền giá trị có tần số xuất hiện cao nhất (mode) nếu là thuộc tính rời rạc.

Cú pháp tham số dòng lệnh: 24_B1.py input.csv output.csv fill-missingvalue propList{id,age}

2.2.1 Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước

Cú pháp tham số dòng lệnh: 24_B2.py countries.txt output.csv

Cú pháp thực hiện cả 4 chức năng:

- a. Xóa các mẫu rỗng.
- b. Xóa các mẫu bị trùng lặp
- c. Chuyển diện tích về km^2
- d. Sử dụng chương trình đã cài đặt ở phần B-1. để xóa các mẫu bị thiếu diện tích.

2.3

Tài liệu tham khảo

<https://www.w3schools.com/python/>

<https://www.geeksforgeeks.org/finding-mean-median-mode-in-python-without-libraries/>

<https://machinelearningmastery.com>

<https://stackoverflow.com/a/43136765>

<https://stackoverflow.com/a/39495168>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html>

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_csv.html

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.isna.html>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.index.html>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.iterrows.html>

<https://docs.scipy.org/doc/numpy/reference/generated/numpy.isnan.html>