TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP. HCM

KHOA CÔNG NGHỆ THÔNG TIN

DATA MINING

BÁO CÁO LAB1:

SINH VIÊN THỰC HIỆN:

1712919 Lê Văn Vũ

1712502 Trần Quang Huy

GV LÝ THUYẾT/ HD THỰC HÀNH:

Thầy **Lê Hoài Bắc**

Thầy **Dương Nguyễn Thái Bảo**

Thầy **Hoàng Xuân Trường**

Thầy **Nguyễn Ngọc Đức**

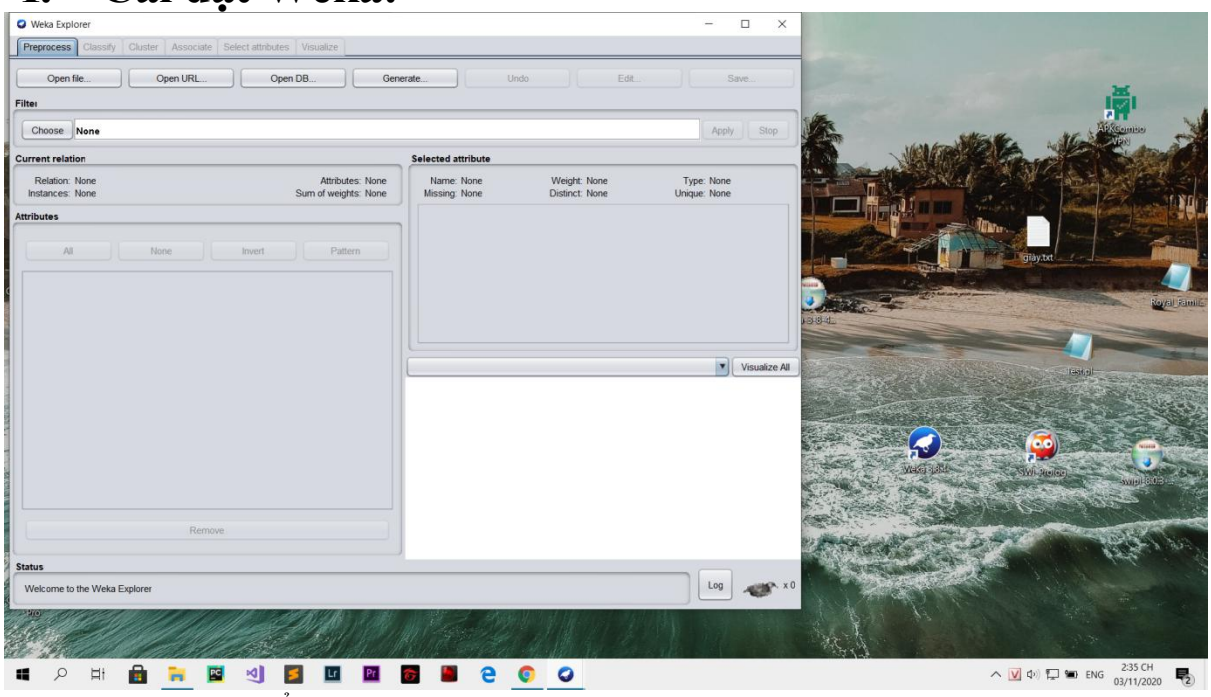
MỤC LỤC:

A. ĐÁNH GIÁ:

Yêu cầu:		Người thực hiện	Ghi chú	Mức độ hoàn thành (%)
Cài đặt Weka		Cả 2		100
Làm quen với Weka	2.1	1712502_Huy		100
	2.2	1712919_Vũ		100
	2.3	Cả 2	- Huy làm phần 1, 2 - Vũ làm phần 3 và 4	100
Cài đặt tiền xử lý dữ liệu	1	1712919_Vũ		100
	2	1712919_Vũ		100
	3	1712502_Huy		100
	4	1712502_Huy		0
	5	1712919_Vũ		100
	6	1712502_Huy		100
	7.1	1712919_Vũ		100
	7.2	1712502_Huy		100
	8	1712502_Huy		0

B. Tìm hiểu Weka:

I. Cài đặt Weka:



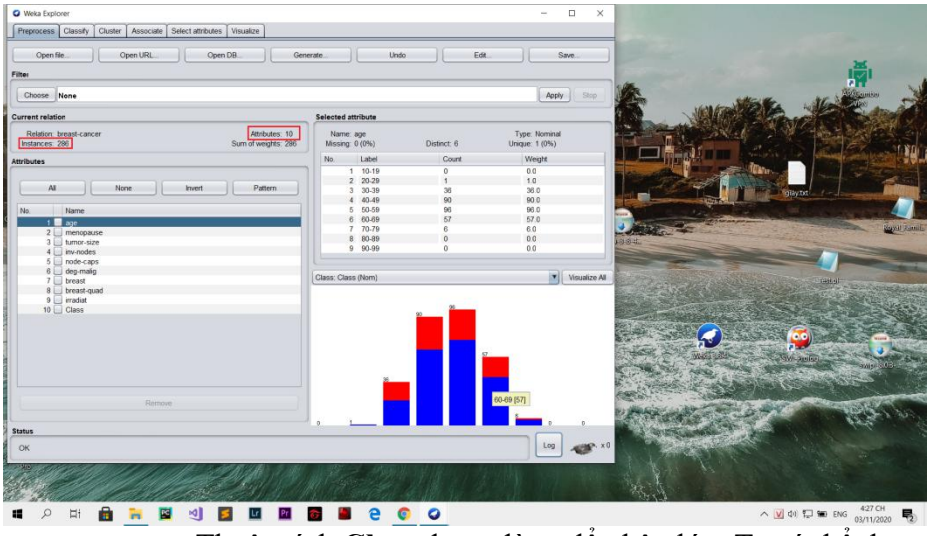
- Preprocess: Để chọn và xử lý dữ liệu làm việc
  - Current relation: cung cấp thông tin về tập dữ liệu như tên quan hệ, số mẫu, số thuộc tính.
  - Attributes: danh sách các thuộc tính (có thứ tự).
  - Selected attribute: thông tin chi tiết của một thuộc tính như tên thuộc tính, tỉ lệ thiếu, loại dữ liệu, các giá trị và số lần xuất hiện.

- Classify: Để huấn luyện và kiểm tra các mô hình học máy (phân loại, hồi quy/dự đoán).
- Cluster: Gom nhóm
- Associate: Để khám phá các luật kết hợp từ dữ liệu
- Select attributes: Để xác định và lựa chọn thuộc tính liên quan (quan trọng) nhất của dữ liệu.
- Visualize: Để xem (hiển thị) biểu đồ tương tác 2 chiều đối với dữ liệu

II. Làm quen với Weka:

1. Đọc dữ liệu vào Weka

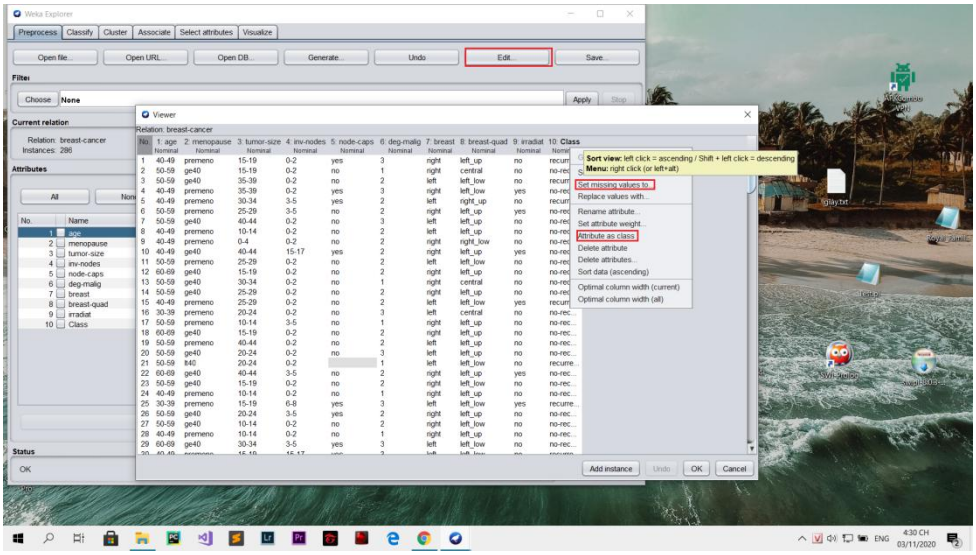
a. Instances: 286
b. Attributes: 10



c. Thuộc tính **Class** được dùng để phân lớp. Ta có thể thay đổi thuộc tính làm lớp bằng cách nhấp vào **Edit** → nhấp chuột phải vào thuộc tính cần chọn làm lớp → chọn **Attribute as class.**

d. Có 2 thuộc tính bị thiếu dữ liệu: node-caps và breast-quad. Thuộc tính thiếu dữ liệu nhiều nhất là node-caps và ít nhất là breast-quad.

- Để giải quyết vấn đề missing values, ta có thể nhấp vào **Edit** → nhấp chuột phải vào cột thuộc tính bị thiếu dữ liệu → chọn **Set missing values to** hoặc nhấp vào ô dữ liệu bị thiếu để bổ sung dữ liệu.



e. Đồ thị biểu diễn trực quan cho từng thuộc tính. Màu xanh biểu thị cho lớp **Yes**, còn đỏ biểu thị cho lớp **No.**

- Visualize all: dùng để biểu diễn đồ thị cho tất cả các thuộc tính.

2. Khám phá tập dữ liệu Weather:

Tập dữ liệu **weather.numeric.arff:**

1. Tập dữ liệu có bao nhiêu thuộc tính? Bao nhiêu mẫu? Phân loại các thuộc tính theo kiểu dữ liệu (categorical/numeric). Thuộc tính nào là lớp?

Tập dữ liệu có:

- 5 thuộc tính:

+ outlook {sunny, overcast, rainy}
+ temperature numeric
+ humidity numeric
+ windy {TRUE, FALSE}
+ play {yes, no}

- 14 mẫu:

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no

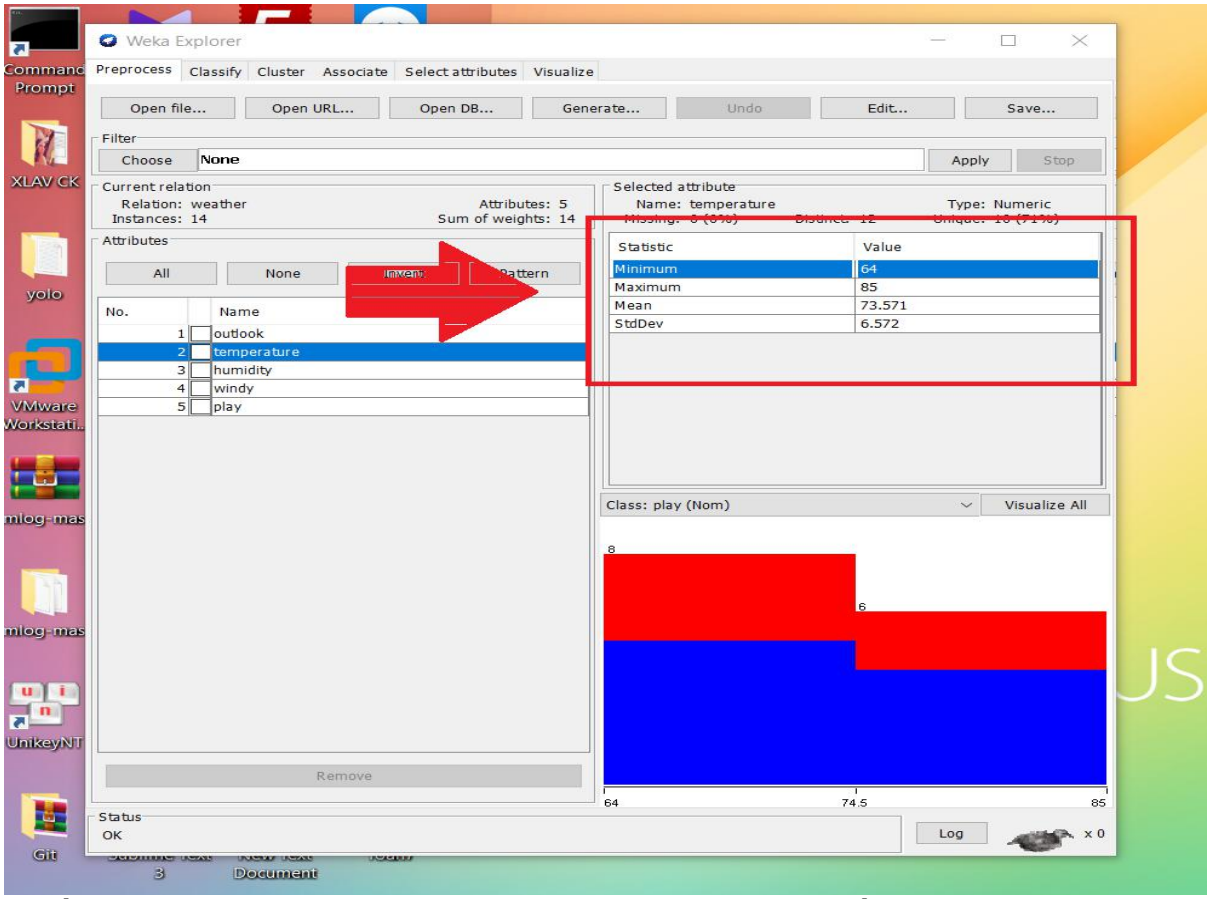
- Phân loại:
  + Categorical: outlook, windy, play.
  + Numeric: temperature, numeric.
- Thuộc tính là lớp: play

2. Liệt kê five-number summary của thuộc tính temperature và humidity. Weka có cung cấp những giá trị này không?

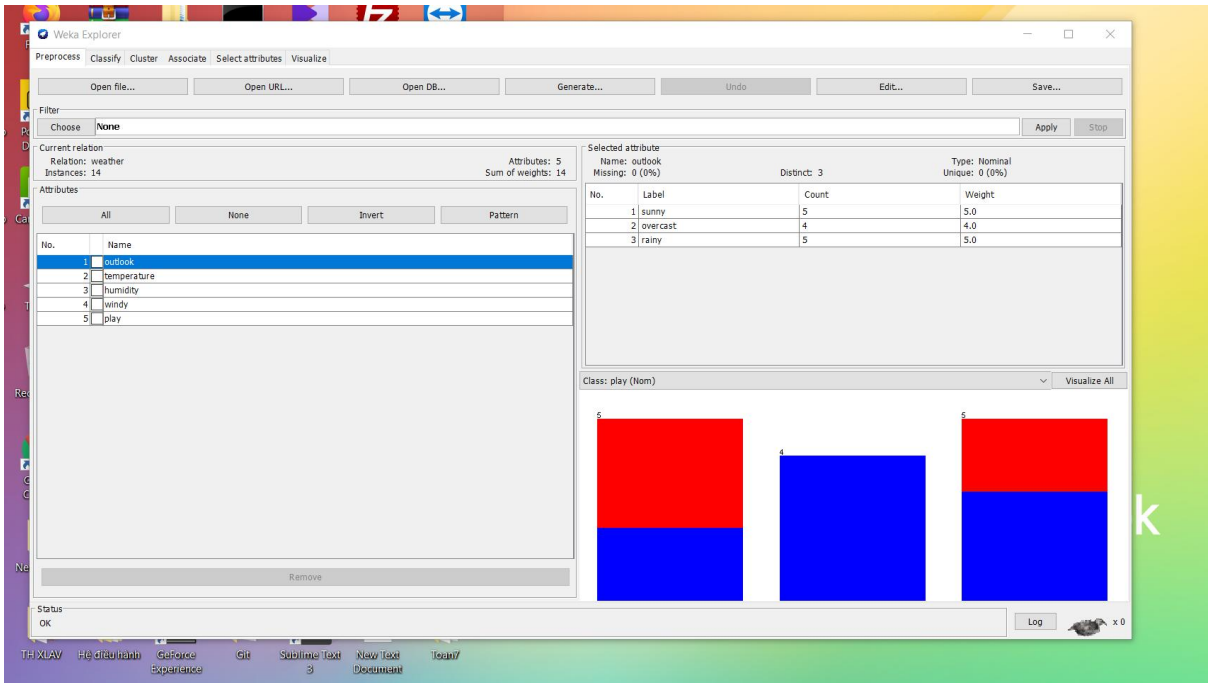
****five-number summary:***

temperature	**humidity**
Min: 64.000	Min: 65.000
Q1: 69.250	Q1: 71.250
Median: 72.000	Median: 82.500
Q3: 78.750	Q3: 90.000
Max: 85.000	Max: 96.000

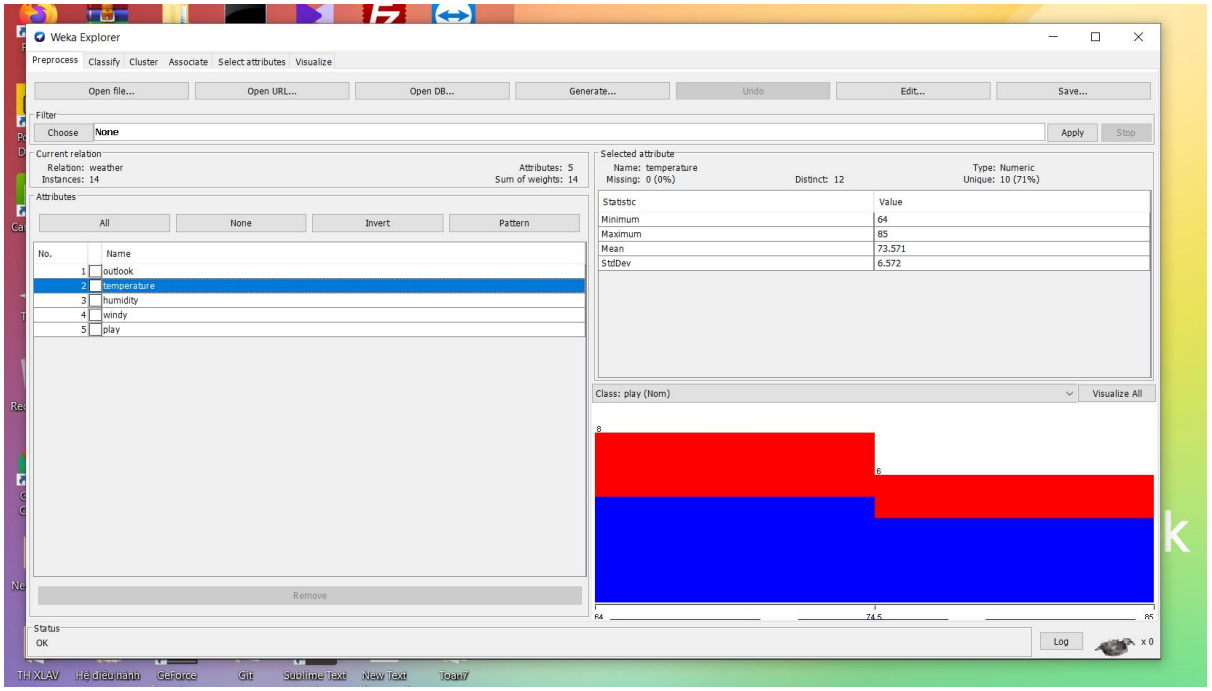
➔ Weka chỉ cung cấp giá trị Min và Max:



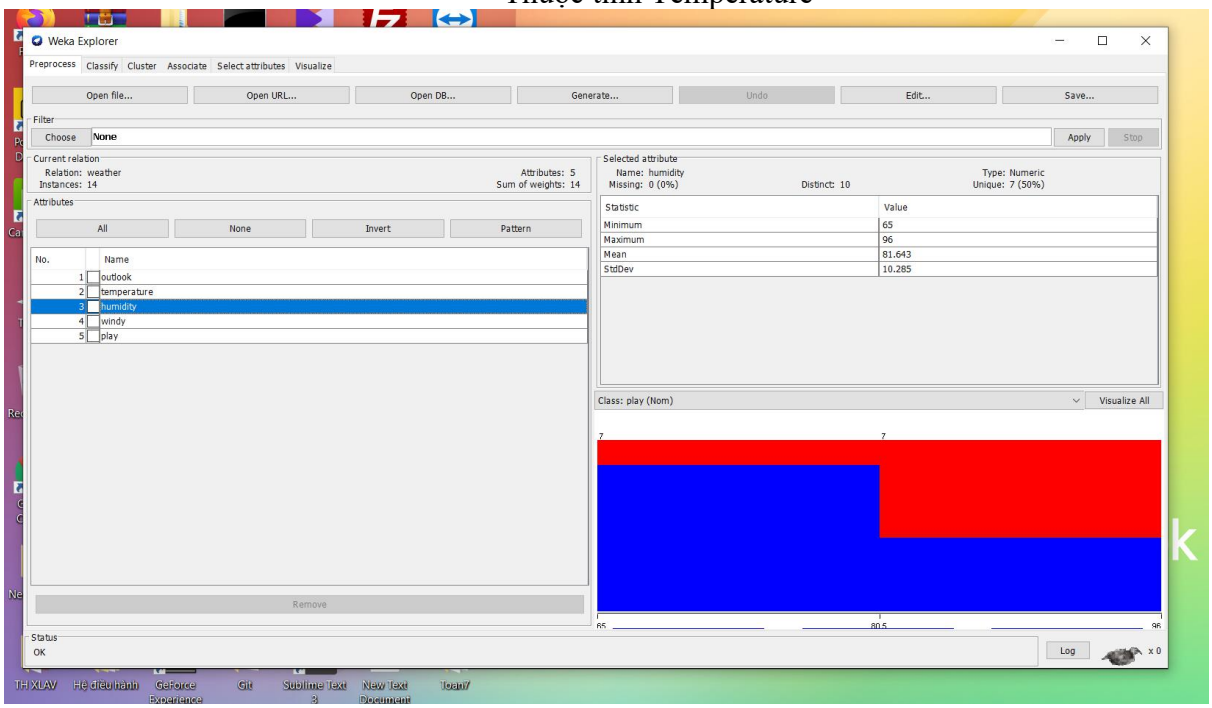
3. Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị. Dán các ảnh chụp màn hình vào bài làm.



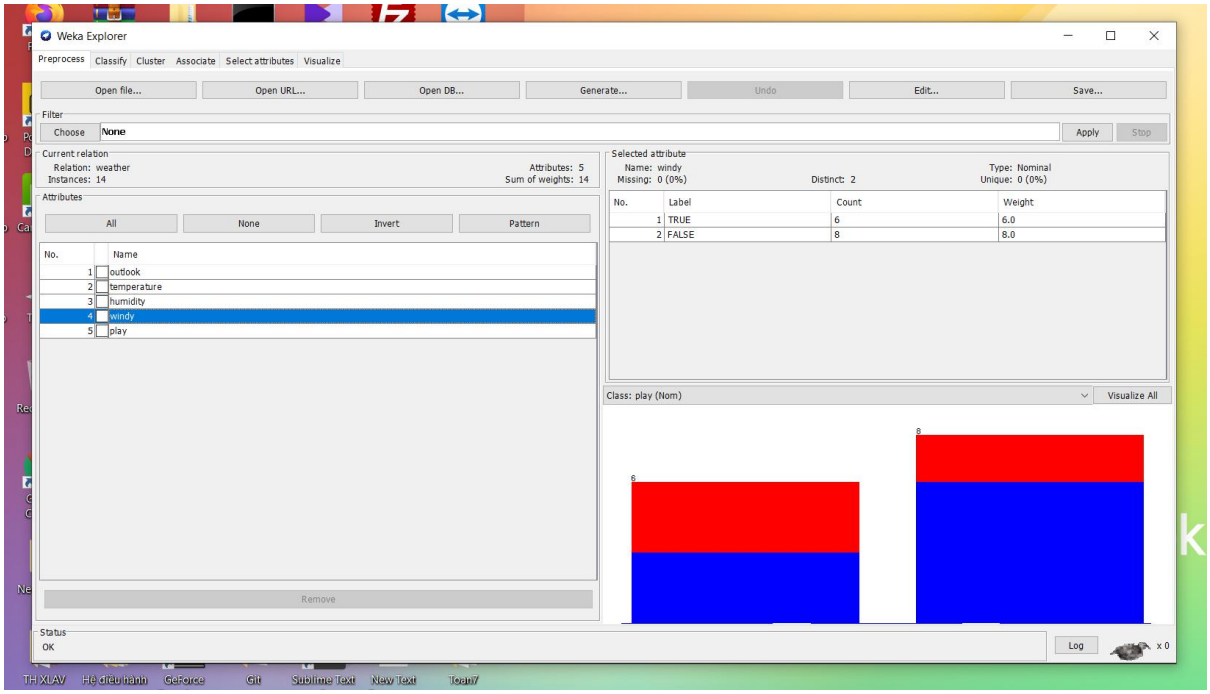
Thuộc tính Outlook



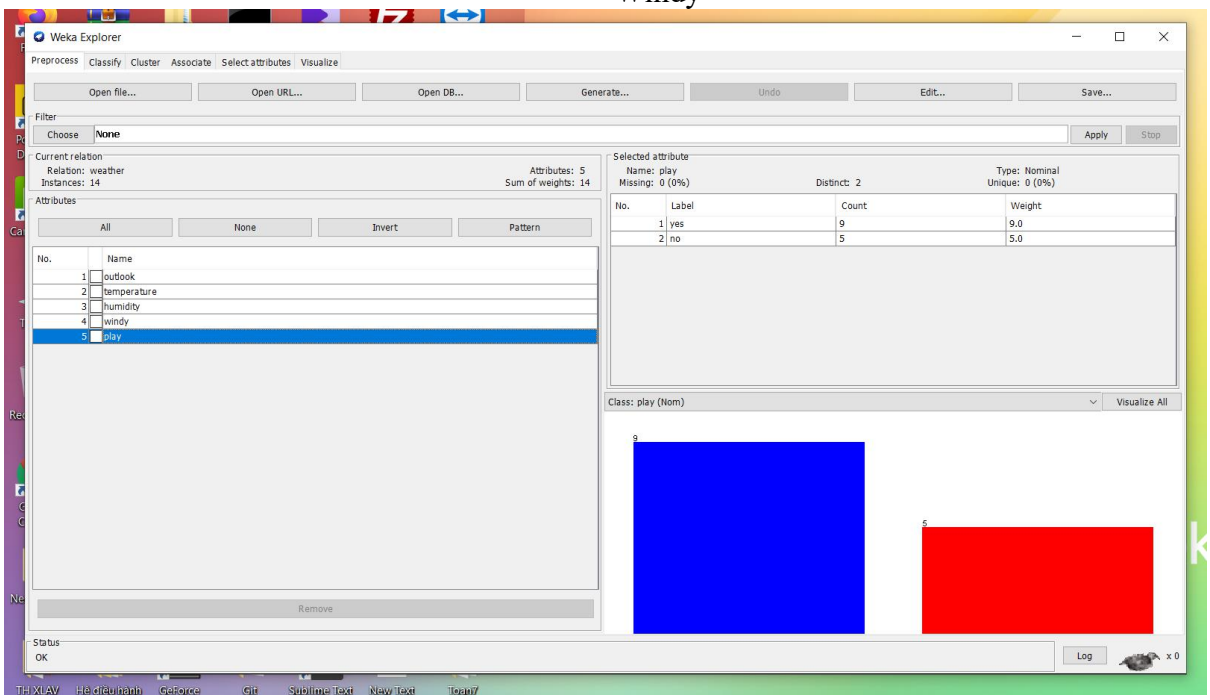
Thuộc tính Temperature



Humidity


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: windy
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 TRUE 6 6.0
2 FALSE 8 8.0
Class: play (Nom)
Visualize All
8
6
Status
OK
Log
x 0

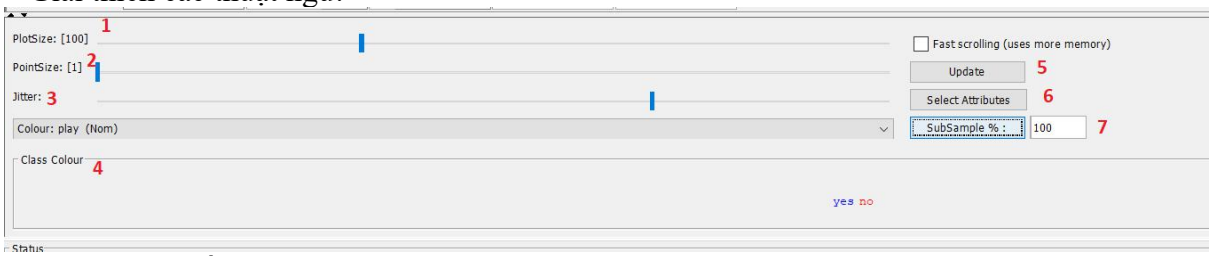
Windy


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: play
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 yes 9 9.0
2 no 5 5.0
Class: play (Nom)
Visualize All
9
5
Status
OK
Log
x 0

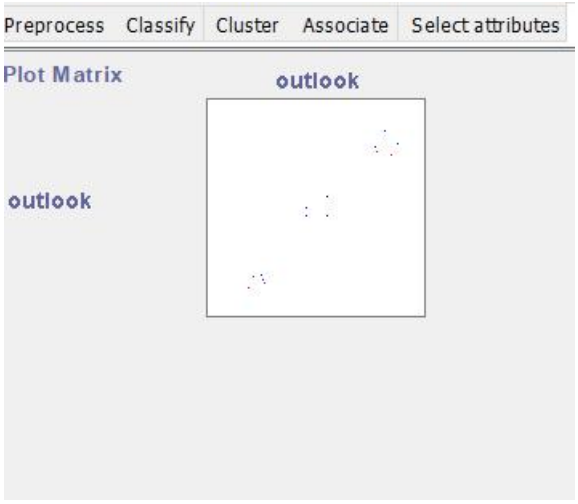
Play

4.Chuyển sang tab Visualize. Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị ở đây là gì? Chọn jitter tối đa để thấy tổng quan hơn về phân bố dữ liệu. Theo bạn có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?

* Giải thích các thuật ngữ:



1: Kích thước đồ thị.
2: kích thước điểm.
3: hỗ trợ việc hiển thị rõà h r ngơn, khi có quá nhiều ví dụ (điểm) tập trung xung quanh một vị trí trên biểu đồ
4: Màu của class (ở đây mà blue tương ứng với yes, red tương ứng với no).
5: update các vị trí point trên biểu đồ.
6: lựa chọn thuộc tính để hiển thị:
Ví dụ nếu chọn Outlook thì biểu đồ hiển thị:



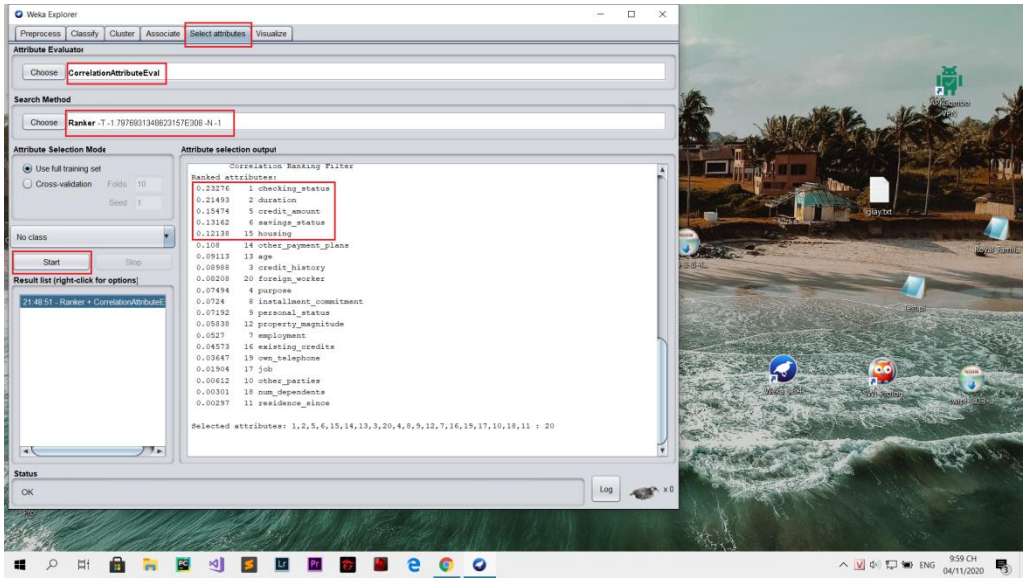
7: Tùy chỉnh Random seed và Tỷ lệ Subsample của tập input.
* Những cặp thuộc tính khác nhau có vẻ như tương quan với nhau: (theo thứ tự (hoành-tung) trên đồ thị):
+ Outlook-Play và Outlook-windy
+ temperature-Play, humidity-Play, temperature-windy
+ temperature-outlook và humidity-outlook
+ windy-outlook và play-outlook

3. Khám phá tập dữ liệu Tín dụng Đức

a. Tập dữ liệu có 1000 mẫu và 21 thuộc tính
b. Tên của thuộc tính lớp tên là **class.** Đa số các lớp đều bị phân bố lệch về một lớp
c. Các phương pháp để lựa chọn thuộc tính:
   - *CfsSubsetEval:* Đánh giá giá trị của một tập hợp con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng tính năng cùng với mức độ dư thừa giữa chúng.
   - *ClassifierAttributeEval:* Đánh giá giá trị của một thuộc tính bằng cách sử dụng trình phân loại do người dùng chỉ định.

- *ClassifierSubsetEval:* Đánh giá tập hợp con thuộc tính trên dữ liệu đào tạo hoặc tập thử nghiệm tạm dừng riêng biệt. Sử dụng bộ phân loại để ước tính "giá trị" của một tập hợp các thuộc tính.
- *CorrelationAttributeEval:* Lựa chọn tính năng dựa trên mối tương quan.
- *GainRatioAttributeEval:* Đánh giá giá trị của một thuộc tính bằng cách đo tỷ lệ khuếch đại liên quan đến lớp.
- *InfogainAttributeEval:* Đánh giá giá trị của một thuộc tính bằng cách đo lường mức tăng thông tin liên quan đến lớp.
- *OneRAttributeEval:* Đánh giá giá trị của một thuộc tính bằng cách sử dụng bộ phân loại OneR.
- *PrincipalComponents:*
- *ReliefFAttributeEval:*
- *SymmetricalUncertAttributeEval*
- *WrapperSubsetEval:* Đánh giá các tập thuộc tính bằng cách sử dụng lược đồ học tập. Xác thực chéo được sử dụng để ước tính độ chính xác của lược đồ học tập cho một tập hợp các thuộc tính.

d. Ở phần **Attribute Evaluator** chúng ta sử dụng bộ lọc **CorrelationAttributeEva**l, mặc định ở phần **Search Method** sẽ yêu cầu chúng ta chọn **Ranker** để hiển thị ra mối tương quan của các thuộc tính theo thứ tự giảm dần như trong hình dưới.



C. THỰC HÀNH:

Các chức năng được đánh số như ssau:

1. Liệt kê các cột bị thiếu dữ liệu.

2. Đếm số dòng bị thiếu dữ liệu.

3. Điền giá trị bị thiếu bằng phương pháp mean, median (cho thuộc tính numeric) và mode (cho thuộc tính categorical). Lưu ý: khi tính mean, median hay mode các bạn bỏ qua giá trị bị thiếu.

4. Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các dòng bị thiếu hơn 50% giá trị các thuộc tính).

5. Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các cột bị thiếu giá trị thuộc tính ở hơn 50% số mẫu).

6. Xóa các mẫu bị trùng lặp.
7. Chuẩn hóa một thuộc tính numeric bằng phương pháp min-max**(7.1)** và Z-score**(7.2)**.
8. Tính giá trị biểu thức thuộc tính: ví dụ đối với một tập dữ liệu có chứa 2 thuộc tính *width* và *height* thì biểu thức *width* * *height* sẽ trả về tập dữ liệu cũ với một thuộc tính mới có giá trị ở mỗi mẫu là tích của thuộc tính *width* và *height* trong mẫu tương ứng, với điều kiện cả 2 giá trị *width* và *height* đều không bị thiếu, trong trường hợp bị thiếu thì giá trị biểu thức coi như bị thiếu. Lưu ý: biểu thức có thể có nhiều thuộc tính và nhiều phép toán bao gồm cộng, trừ, nhân, chia.

Các chức năng đã hoàn thành:

- Cách sử dụng: chạy bằng tham số dòng lệnh:

source.py Mã_chức_năng file_input ...

Chức năng:	Ví dụ Cách sử dụng	Kết quả	Ghi chú
1.	source.py **1** house-prices.csv	Xuất ra file output có tên: source.py 1 house-prices.csv	
2.	source.py **2** house-prices.csv	In ra màn hình console: So dong bi thieu du lieu la:", count_miss_data_rows(data), "(dong)	
3.	source.py **3** house-prices.csv	Xuất ra file output có tên: FilledMissingData.csv	
5.	source.py **5** house-prices.csv **50**	Xuất ra file output có tên: removed_miss_data_column.csv	**50** là tỷ lệ phần trăm dữ liệu bị thiếu
6.	source.py **6** house-prices.csv	Xuất ra file output có tên: house-prices_deduped.csv	
7.1	source.py **7.1** house-prices.csv **2 8**	Xuất ra file output có tên: MinMaxDataNormalization.csv	**2 và 8** là giá trị min và max
7.2	source.py **7.2** house-prices.csv	Xuất ra file output có tên: Z_ScoresDataNormalization.csv	

Nguồn tham khảo:

https://www.tutorialspoint.com/weka/what_is_weka.htm
https://www.cs.waikato.ac.nz/ml/weka/
https://www.springboard.com/blog/data-mining-python-tutorial/
https://github.com/ChangyuYan/Data-Mining
https://github.com/mrsan22/Data-Mining-Project

Và nhiều nguồn khác