

BÁO CÁO

LAB 01

TIỀN XỬ LÝ VÀ KHÁM PHÁ DỮ LIỆU

SINH VIÊN THỰC HIỆN

20120130 - ĐINH THỊ HOÀNG LINH

20120146 - NGUYỄN THỊ CHÂU NGỌC

LỜI CẢM ƠN

Sau một khoảng thời gian học tập và làm việc ở trường Đại học Khoa học Tự nhiên, bằng lòng biết ơn và sự kính trọng sâu sắc, chúng xin gửi lời cảm ơn chân thành đến Ban giám hiệu nhà trường, các phòng ban,... đã tạo điều kiện tốt nhất để chúng bước đầu làm quen với môi trường học tập đại học. Hơn thế nữa, em muốn cảm ơn Khoa Công nghệ Thông tin đã tạo ra môi trường học tập thân thiện với những kiến thức bổ ích là nền tảng ban đầu để chúng em có thể tiếp cận với chuyên ngành. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – Thầy Lê Hoài Bắc, cô Nguyễn Thị Thu Hằng đã dạy dỗ, truyền đạt kinh nghiệm và kiến thức quý báu cho chúng em, tạo động lực cho chúng em làm quen với những kiến thức mới lạ. Đây chắc chắn sẽ là động lực, là hành trang cho chúng em có thể vững bước sau này.

Khai thác dữ liệu và ứng dụng là môn học thú vị, vô cùng bổ ích và có tính thực tế cao, đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn khả năng còn nhiều hạn chế nên khó tránh khỏi những sai sót. Mặc dù em đã cố gắng hết sức nhưng chắc chắn bài báo cáo khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong thầy, cô xem xét và góp ý để bài báo cáo của em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

MỤC LỤC

LỜI CẢM ƠN.....	1
MỤC LỤC.....	2
PHẦN NỘI DUNG BÁO CÁO	3
Sơ lược nội dung báo cáo:	3
1. Thông tin chung.....	3
2. Cài đặt Weka.	3
3. Làm quen với Weka.....	3
4. Tiền xử lý dữ liệu trong python.....	3
1. Thông tin chung:	4
1.1 Thông tin thành viên nhóm:.....	4
1.2 Tỷ lệ hoàn thành công việc:.....	4
2. Cài đặt Weka:.....	4
2.1 Yêu cầu 1:.....	4
2.2 Yêu cầu 2:.....	5
3. Làm quen với Weka:	9
3.1 Khám phá dữ liệu Breast Cancer.	9
3.2 Khám phá dữ liệu Weather.	17
3.3 Khám phá dữ liệu tín dụng ở Germany.	25
4. Tiền xử lý dữ liệu trong python:.....	32
4.1 Xuất ra các cột có giá trị thiếu:.....	32
4.2 Đếm số hàng có giá trị thiếu:.....	33
4.3 Điền các giá trị thiếu sử dụng mean, median và mode:.....	33
4.4 Xóa các hàng có tỉ lệ giá trị thiếu cao hơn 1 ngưỡng nhất định:	34
4.5 Xóa các cột có tỉ lệ giá trị thiếu cao hơn 1 ngưỡng nhất định:	35
4.6 Xóa các hàng trùng lặp:	35
4.7 Chuẩn hóa dữ liệu sử dụng min-max hoặc z-score:	36
4.8 Thực hiện tính toán trên 2 cột dữ liệu dạng số:	36
TÀI LIỆU THAM KHẢO	38

PHẦN NỘI DUNG BÁO CÁO

Sơ lược nội dung báo cáo:

1. Thông tin chung.
 - 1.1. Thông tin thành viên nhóm.
 - 1.2. Tỷ lệ hoàn thành công việc.
2. Cài đặt Weka.
 - 2.1. Yêu cầu 1.
 - 2.2. Yêu cầu 2.
3. Làm quen với Weka.
 - 3.1. Khám phá dữ liệu Breast Cancer.
 - 3.2. Khám phá dữ liệu Weather.
 - 3.3. Khám phá dữ liệu tín dụng ở Germany.
4. Tiền xử lý dữ liệu trong python.

1. Thông tin chung:

1.1 Thông tin thành viên nhóm:

MSSV	HỌ VÀ TÊN	Email
20120130	Đinh Thị Hoàng Linh	20120130@student.hcmus.edu.vn
20120146	Nguyễn Thị Châu Ngọc	20120146@student.hcmus.edu.vn

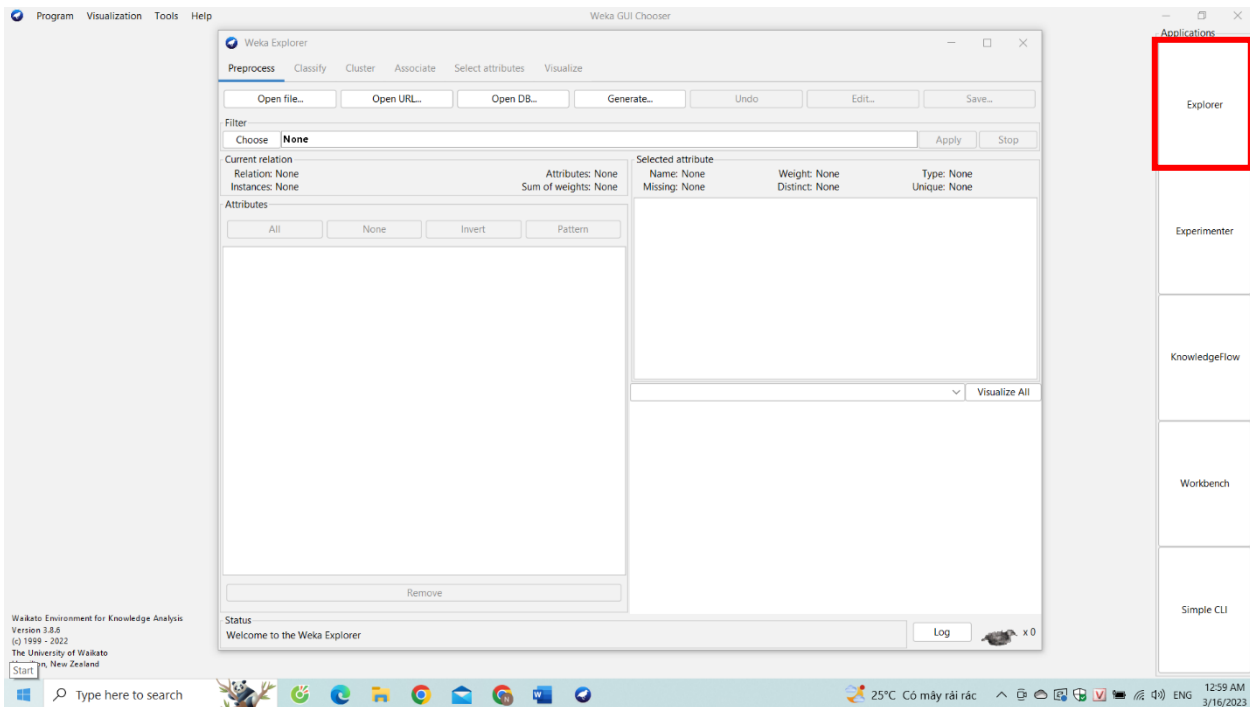
1.2 Tỷ lệ hoàn thành công việc:

STT	MSSV	Công việc	Tỷ lệ hoàn thành
1	20120146	Cài đặt Weka	100%
2	20120146	Khám phá dữ liệu Breast Cancer	100%
3	20120146 20120130	Khám phá dữ liệu Weather	100%
4	20120130	Khám phá dữ liệu tín dụng ở Germany	100%
5	20120130	Tiền xử lý dữ liệu trong python	100%
6	20120146	Trình bày báo cáo	100%

2. Cài đặt Weka:

2.1 Yêu cầu 1:

After installing, you capture a screen that contains the "Explorer" function in your desktop background.



2.2 Yêu cầu 2:

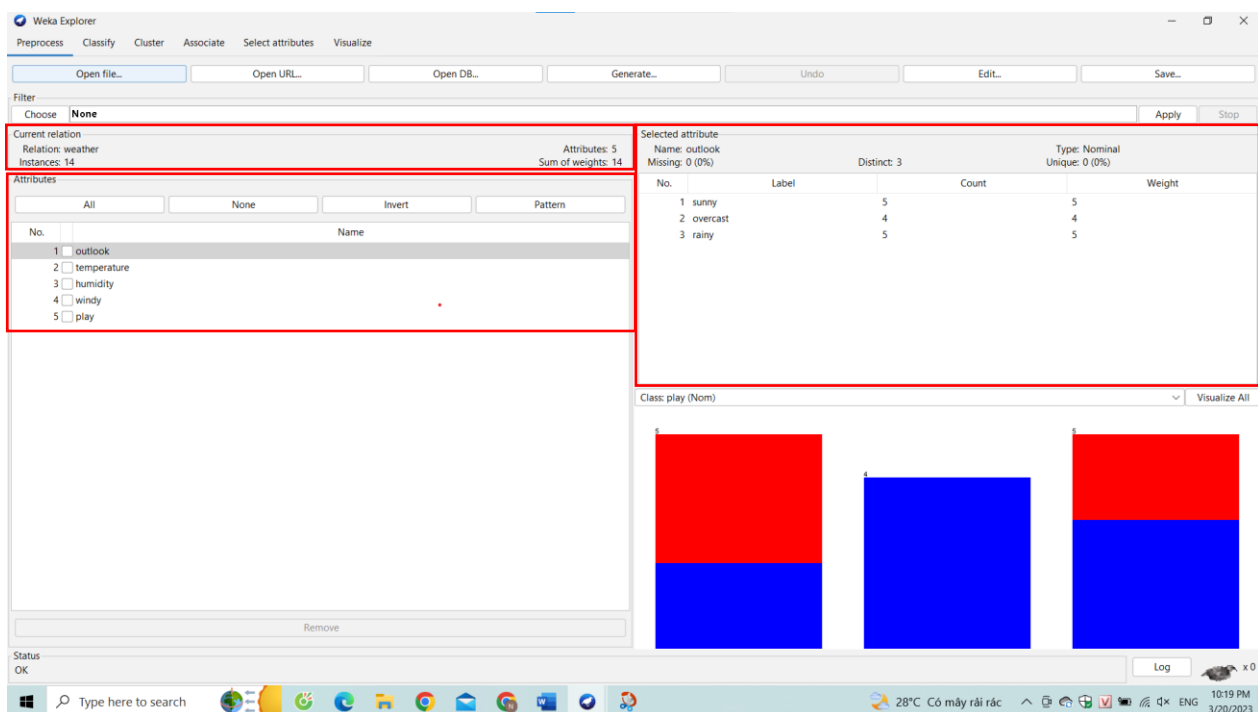
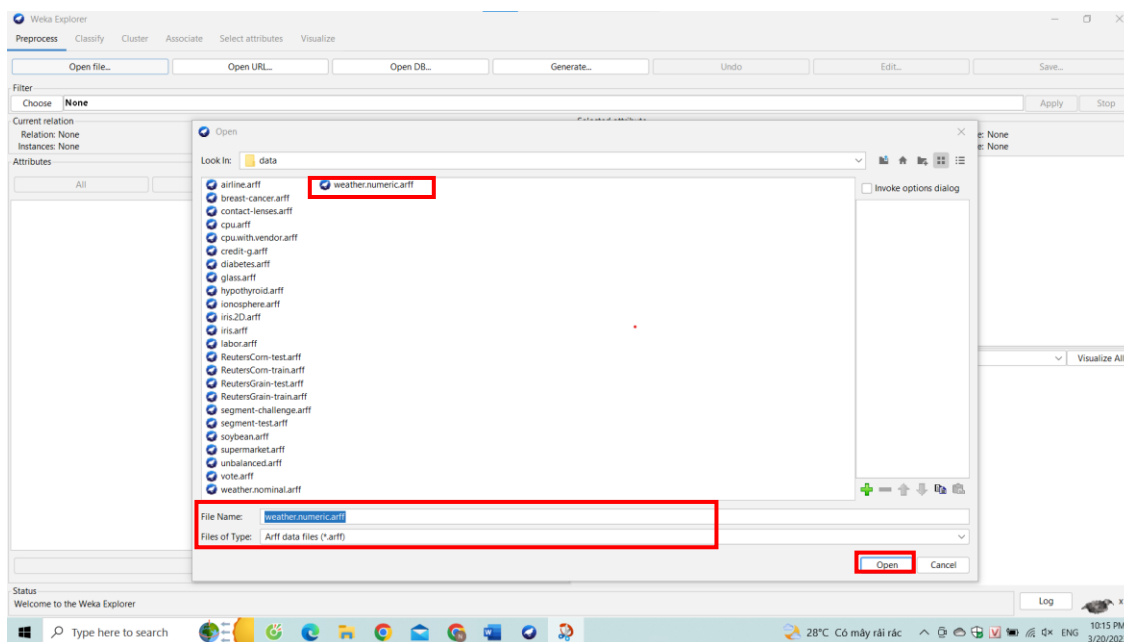
Students open any data set (with extended part .arff). Explain the meaning of Current Relation, Attributes, and Selected Attribute in Preprocess tag. Briefly explain the meaning of the other tags in WEKA Explorer.

1. Ý nghĩa nhóm lệnh điều khiển trong tab Preprocess:

- Current relation (tạm dịch ‘Quan hệ hiện tại’): Cho biết các thông tin chung về tập dữ liệu hiện tại như tên tập dữ liệu, số mẫu, số thuộc tính.
- Attributes (tạm dịch ‘Thuộc tính’): Cho biết danh sách các thuộc tính hiện tại trong tập dữ liệu.
- Selected attribute (tạm dịch ‘Thuộc tính được chọn’): Cho biết thông tin chung liên quan đến thống kê như trung bình cộng, giá trị min, giá trị max, độ lệch chuẩn của thuộc tính đã được chọn trước trong phần Attributes nếu dữ liệu thuộc tính ở dạng định lượng. Đối với dữ liệu dạng định danh, Weka sẽ cung cấp danh sách các định danh và số lượng mỗi định danh.

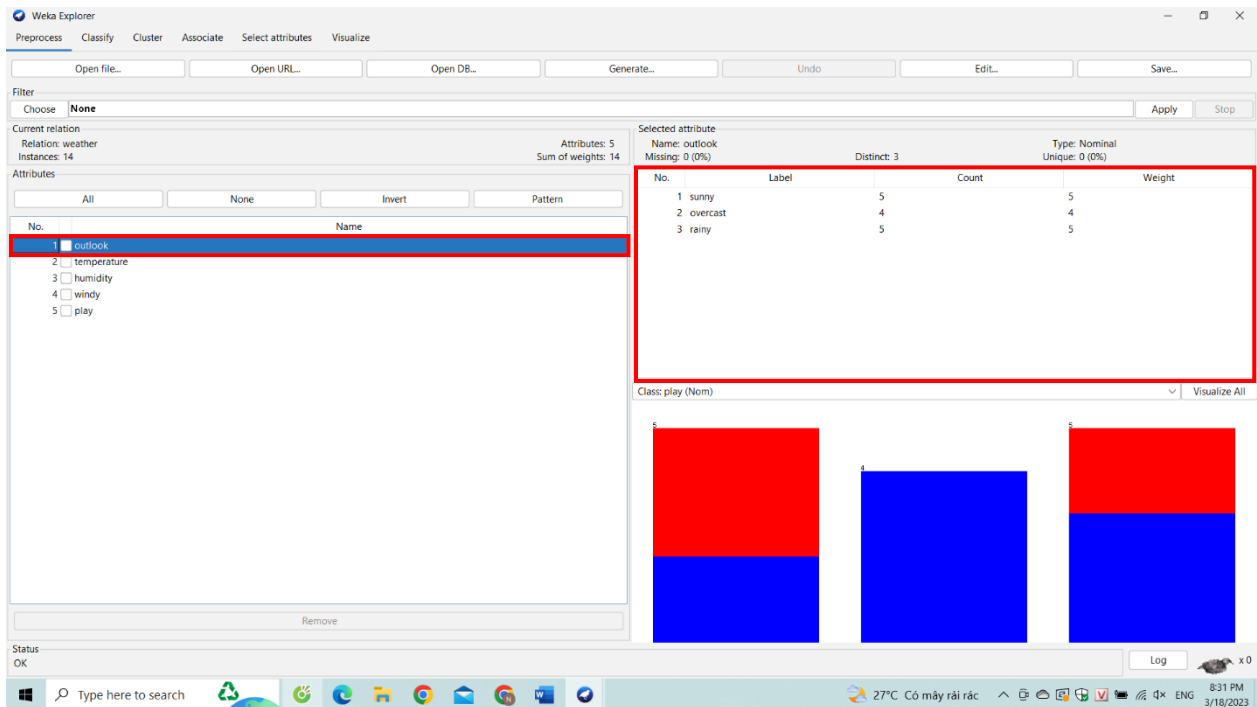
2. Giải thích về ý nghĩa của Current Relation, Attributes, and Selected Attribute trong Preprocess tag với dataset “weather.numeric.arff”.

- Mở dataset “weather.numeric.arff”.

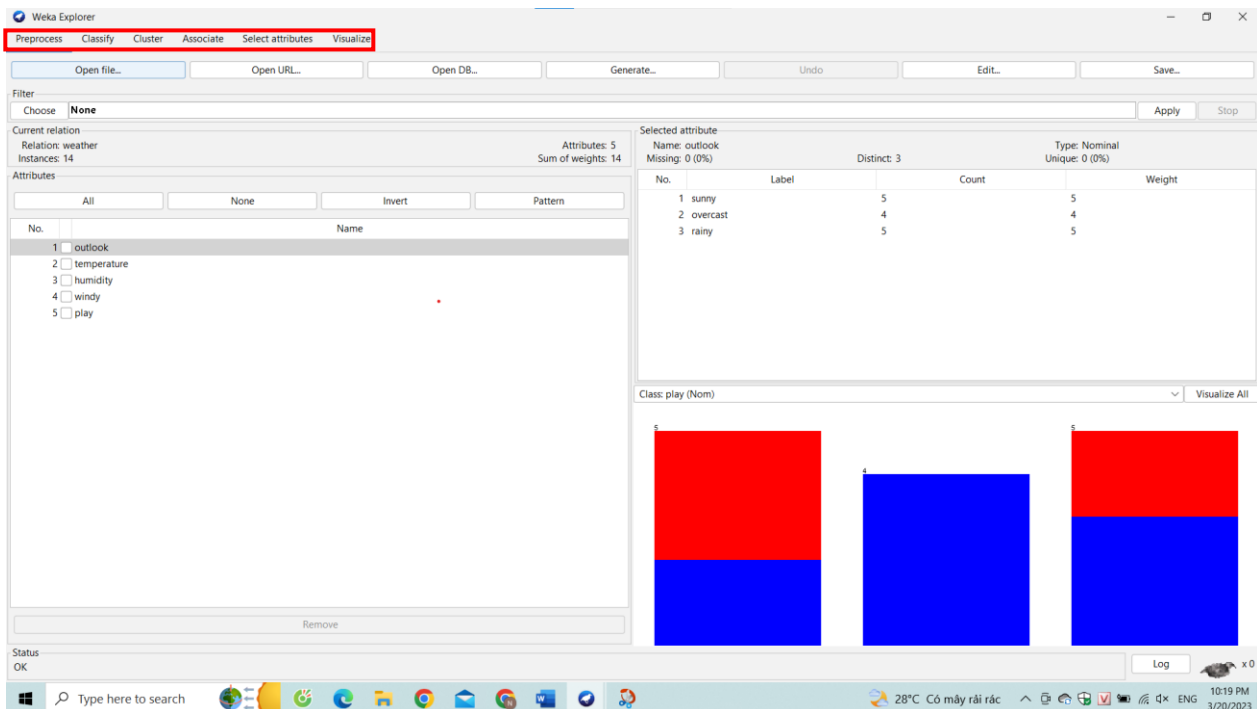


- **Current Relation:** Thông tin về quan hệ hiện tại
 - Relation: Tên của quan hệ được đưa ra trong file đã được nạp là weather.
 - Instance: Số lượng các bản ghi là 14.

- Attributes: Số các thuộc tính trong dữ liệu là 5(outlook, temperature, humidity, windy, play).
- Sum of weights: Tổng trọng lượng của mỗi bản ghi trong quan hệ là 14.
- **Attributes:** Có 4 nút(All: Tất cả các hộp Selection tick đều được tích, None: Bỏ dấu tích ở tất cả các hộp, Invert: Chuyển đổi trạng thái các hộp đã tích thành chưa tích và ngược lại, Pattern: Cho phép người dùng lựa chọn các thuộc tính dựa trên công thức) và ở dưới chúng là một danh sách các thuộc tính trong quan hệ hiện tại. Danh sách này có ba cột:
 - No: Một số xác định thuộc tính theo thứ tự chúng được quy định cụ thể trong file dữ liệu.
 - Selection tick boxes: Cho phép lựa chọn với các thuộc tính được mô tả trong quan hệ.
 - Name: Tên của thuộc tính. Khi ta nhấp chuột vào các dòng khác nhau trong danh sách thuộc tính, các trường sẽ thay đổi trong hộp bên phải có tên là “Selected attribute”. Trong tập dữ liệu được quan sát dưới đây thì có 6 thuộc tính bao gồm: outlook, temperature, humidity, windy, play.
- **Selected Attribute:** Thông tin về thuộc tính đang được chọn:
 - Name: Tên của thuộc tính là outlook.
 - Type: Kiểu của thuộc tính (kiểu Nominal: dạng rời rạc/phi số hoặc Numeric: Dạng số). Hình ảnh bên dưới hiển thị thuộc tính outlook có kiểu thuộc tính là Nominal.
 - Missing: số mẫu và phần trăm tương ứng thiếu giá trị trên thuộc tính outlook là 0(0%).
 - Distinct: Số giá trị phân biệt là 3.
 - Unique: Số mẫu và phần trăm tương ứng không có giá trị trùng với mẫu khác là 0(0%).
 - Phía dưới là bảng thống kê một danh sách các thông tin bổ sung về những giá trị được chứa trong thuộc tính này. Nếu các thuộc tính là “nominal” thì danh sách bao gồm các giá trị có thể đối với thuộc tính cùng với số các trường hợp có của giá trị đó. Nếu thuộc tính là “numeric”, danh sách đưa ra bốn số liệu thống kê mô tả sự phân bố của các giá trị gồm: giá trị tối thiểu, tối đa, trung bình và độ lệch chuẩn. Dưới đây là thuộc tính outlook, vì có kiểu dữ liệu thuộc tính là nominal nên bảng ta có bảng thống kê như sau:



3. Giải thích về ý nghĩa của các thẻ trong Weka Explorer:



- Preprocess: Cho phép mở, điều chỉnh, lưu một tập tin dữ liệu, thẻ này chứa các thuật toán áp dụng trong tiền xử lý dữ liệu.

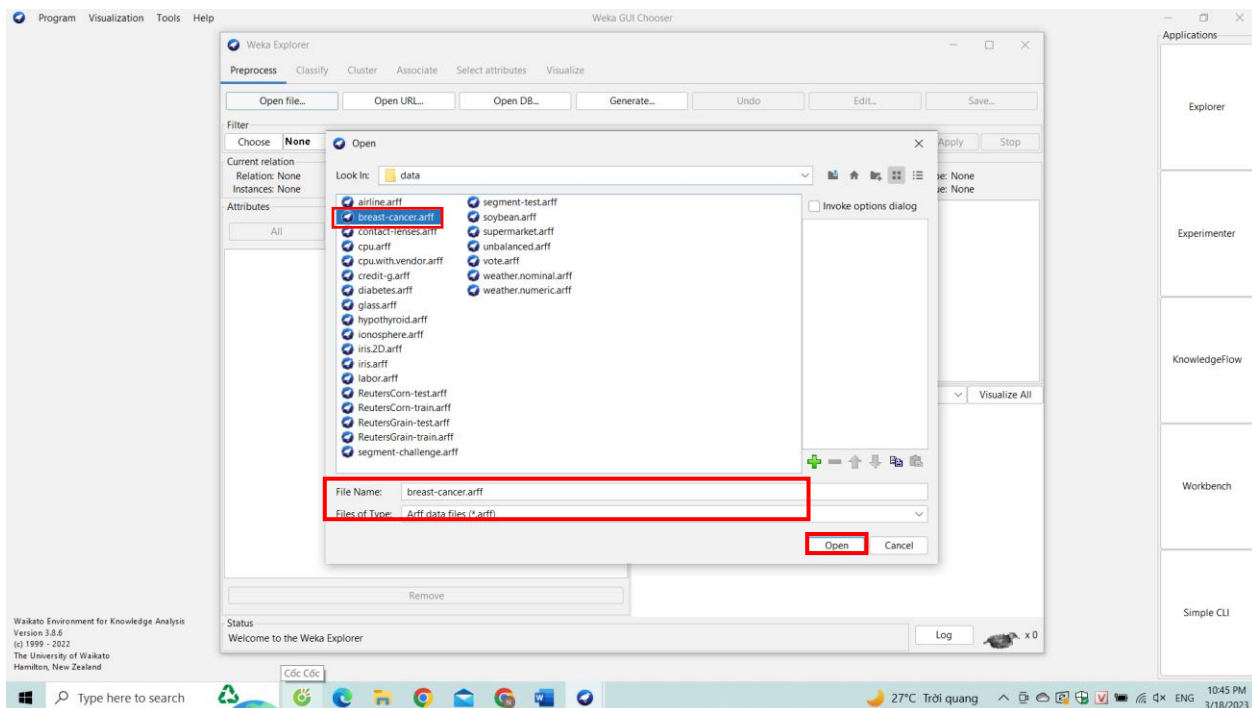
- Classify: Huấn luyện các chương trình học thực hiện phân loại hoặc hồi quy và đánh giá chúng .

- Cluster: Cung cấp các mô hình gom cụm.
- Associate: Khai thác tập phổ biến, luật kết hợp cho dữ liệu và đánh giá chúng.
- Select attributes: Chọn các thuộc tính phù hợp nhất của tập dữ liệu.
- Visualize: Thể hiện dữ liệu dưới dạng biểu đồ.

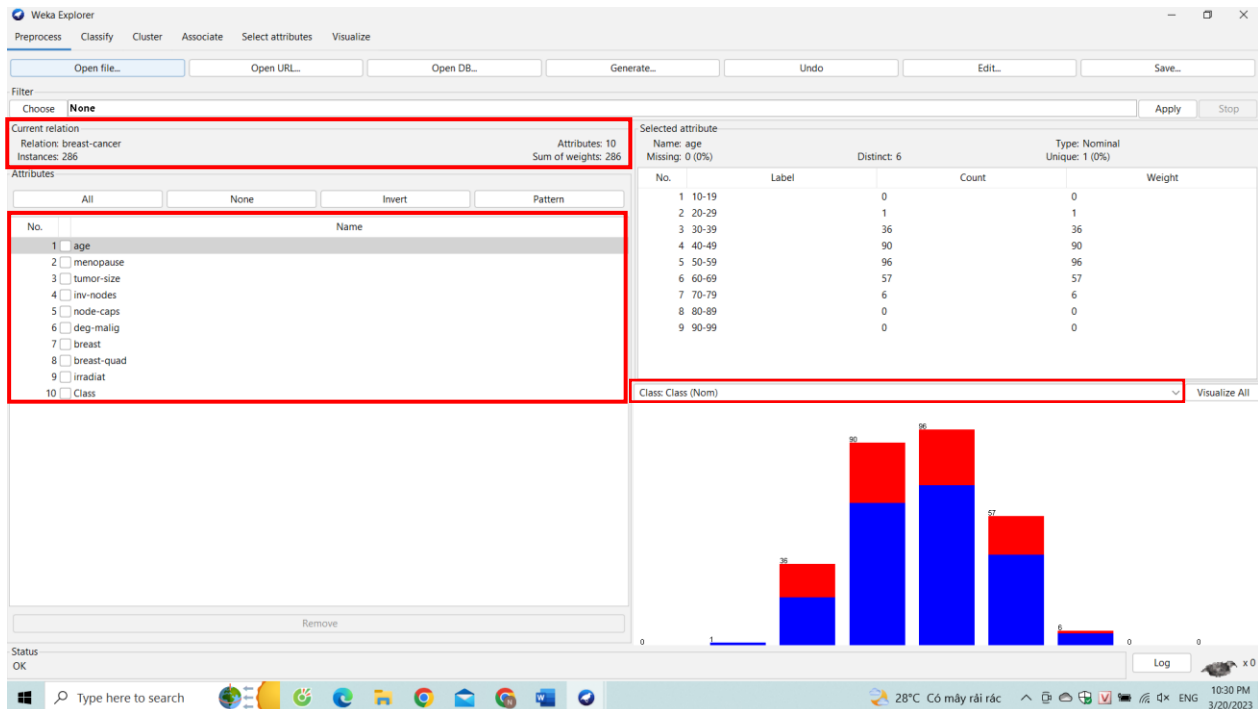
3. Làm quen với Weka:

3.1 Khám phá dữ liệu Breast Cancer.

- Mở tập dữ liệu “breast_cancer.arff” trong giao diện Weka Explorer.



- Trả lời các câu hỏi:



1. How many instances does this data set have?

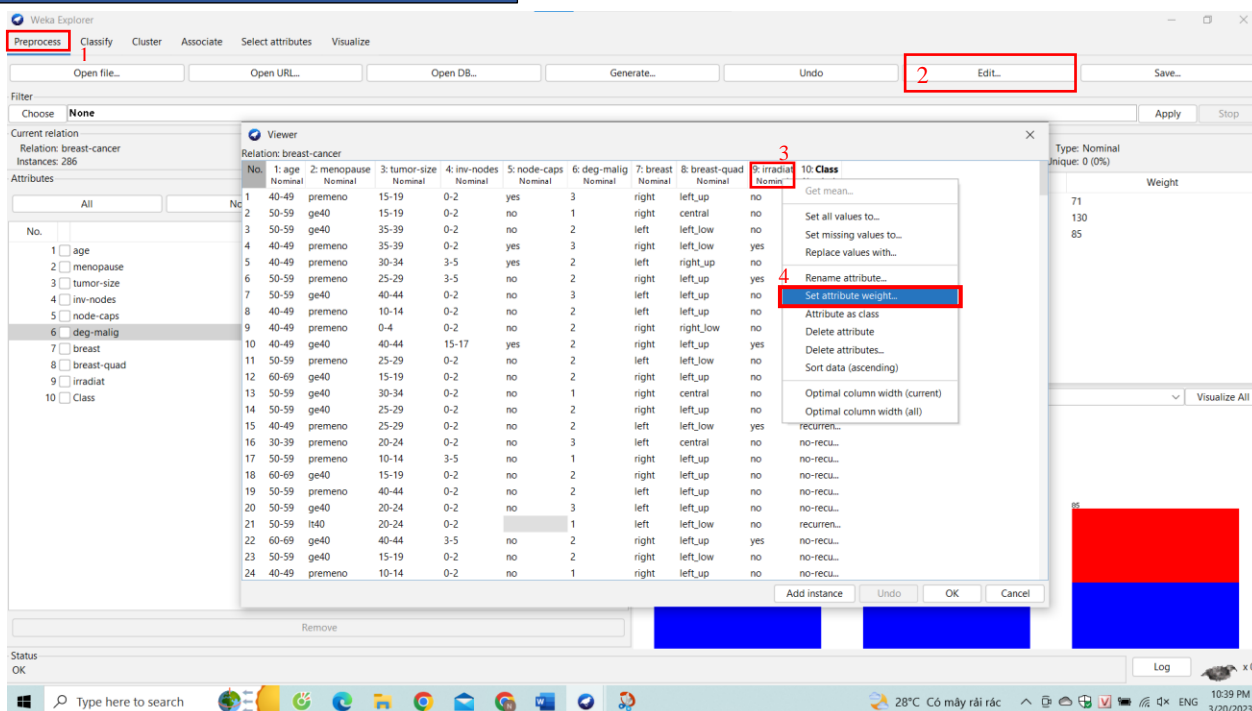
Tập dữ liệu có 286 instances.

2. How many attributes does this data set have?

Tập dữ liệu có 10 attributes.

3. Which attribute is used for the label? Can it be changed? How?

Thuộc tính được sử dụng làm nhãn là “Class”. Thuộc tính này gồm hai giá trị là: no-recurrence-events và recurrence-events. Ta có thể đổi thuộc tính dùng làm nhãn. Cách thực hiện: Weka Explorer → Preprocess → Edit → nhấn chuột phải vào thuộc tính muốn dùng làm nhãn → Chọn Attribute as class.



4. What is the meaning of each attribute?

Name	Description	Type	Limits
age	Độ tuổi	Real	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	Thời điểm mãn kinh	Discrete	lt40, ge40, premeno
tumor-size	Kích thước khối u cắt theo mm	Real	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
inv-nodes	Thước đo về sự hiện diện	Real	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
node-caps	Bằng chứng tế bào ung thư	Discrete	yes, no
deg-malig	cấp độ mô học khối u	Real	1, 2, 3

breast	Vú bị ảnh hưởng	Discrete	left, right
breast-quad	Góc phần tư vú	Discrete	left-up, left-low, right-up, right-low, central
irradiat	Xạ trị	Discrete	yes, no
Class	Phân lớp	Discrete	no-recurrence- events, recurrence- events

5. Let's investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.

- Thuộc tính không có giá trị thiếu: age, menopause, tumor-size, inv-nodes, deg-malig, breast, irradiat, Class.

Selected attribute Name: age Missing: 0 (0%)	Distinct: 6	Type: Nominal Unique: 1 (0%)
Selected attribute Name: menopause Missing: 0 (0%)	Distinct: 3	Type: Nominal Unique: 0 (0%)
Selected attribute Name: tumor-size Missing: 0 (0%)	Distinct: 11	Type: Nominal Unique: 0 (0%)
Selected attribute Name: inv-nodes Missing: 0 (0%)	Distinct: 7	Type: Nominal Unique: 1 (0%)
Selected attribute Name: deg-malig Missing: 0 (0%)	Distinct: 3	Type: Nominal Unique: 0 (0%)
Selected attribute Name: breast Missing: 0 (0%)	Distinct: 2	Type: Nominal Unique: 0 (0%)
Selected attribute Name: irradiat Missing: 0 (0%)	Distinct: 2	Type: Nominal Unique: 0 (0%)
Selected attribute Name: Class Missing: 0 (0%)	Distinct: 2	Type: Nominal Unique: 0 (0%)

- Thuộc tính có giá trị thiếu: node-caps(8 giá trị _ 3% trên toàn thuộc tính), breast-quad(1 giá trị _ 0% trên toàn thuộc tính).

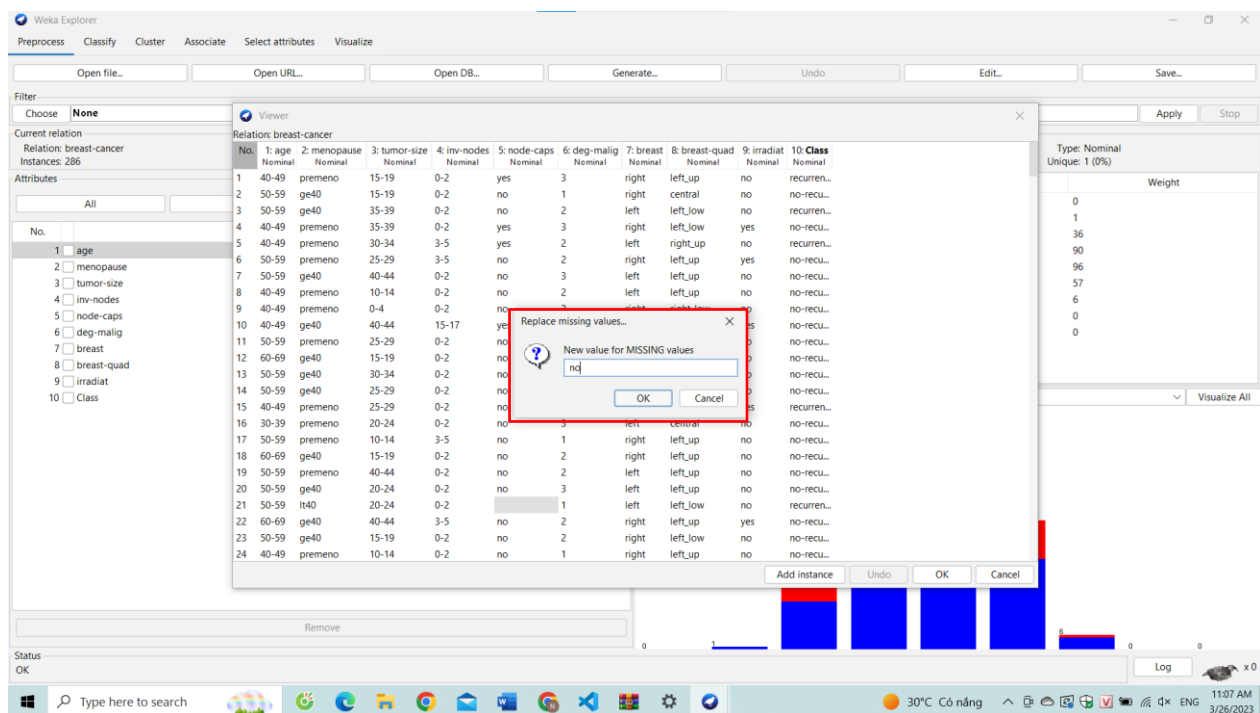
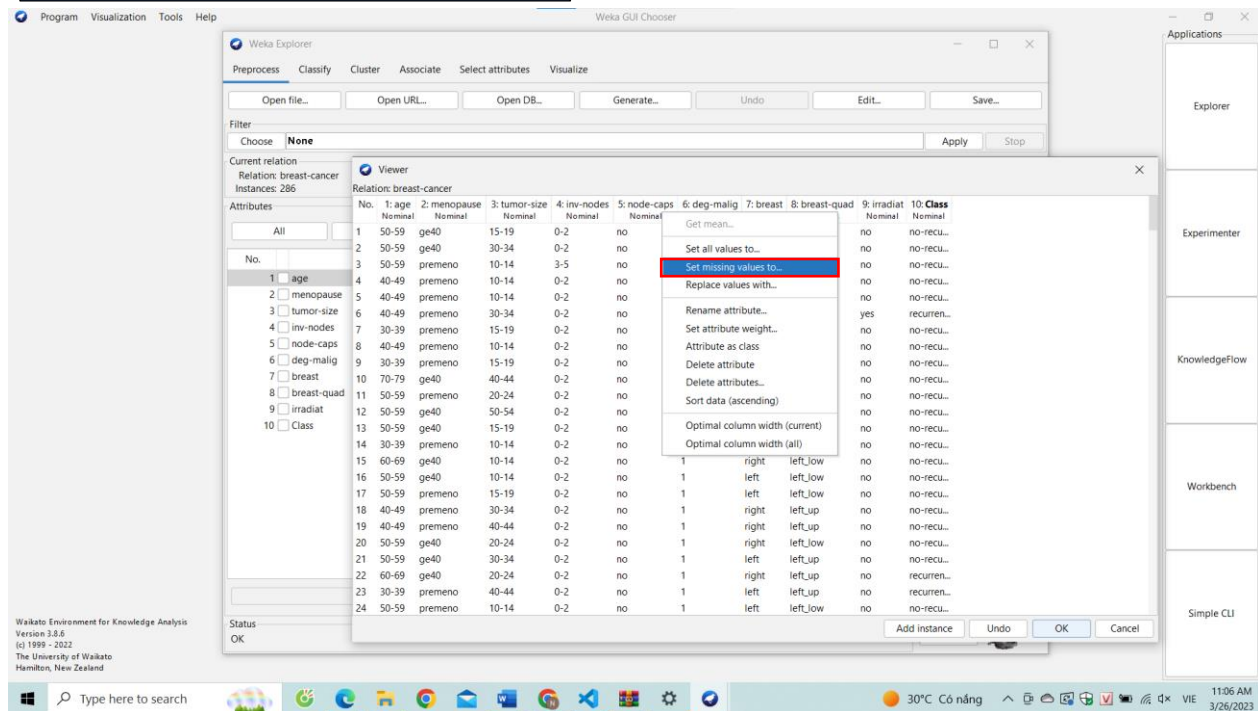
Selected attribute Name: node-caps Missing: 8 (3%)	Distinct: 2	Type: Nominal Unique: 0 (0%)
Selected attribute Name: breast-quad Missing: 1 (0%)	Distinct: 5	Type: Nominal Unique: 0 (0%)

- Khi gặp tình trạng thiếu dữ liệu, ta có thể bỏ dữ liệu thiếu hoặc bổ sung dữ liệu thiếu.

6. Let's propose solutions to the problem of missing values in the specific attribute.

- Tự điền vào tất cả giá trị thiếu của một thuộc tính bằng giá trị cụ thể nằm trong thuộc tính đó:

Explore → Preprocess → Edit → nhấp chuột phải vào cột thuộc tính có giá trị thiếu → chọn Set missing value to... → nhập giá trị muốn điền vào chỗ thiếu giá trị (giá trị này phải thuộc thuộc tính đó) → nhấn OK.



- Sử dụng Filter ReplaceMissingValues/ ReplaceMissingwithUserConstant:

Explore → Preprocess → trong Filter chọn Choose → filters → unsupervised → attribute → ReplaceMissingValues → Apply.

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Choose' button is highlighted in red, and the 'ReplaceMissingValues' filter is selected. The 'Selected attribute' section shows 'Name: node-caps' and 'Missing: 0 (0%)'. The 'Attributes' list on the left shows 'node-caps' selected. The 'Class: Class (Nom)' dropdown is set to 'Class (Nom)'. The 'Visualize All' button is visible. The status bar at the bottom shows 'Status OK' and the system clock.

No.	Label	Count	Weight
1	yes	56	56
2	no	230	230

Explore → Preprocess → trong Filter chọn Choose → filters → unsupervised → attribute → ReplaceMissingWithUserConstant → Lựa chọn thông số phù hợp → Apply.

Thứ tự thuộc tính có giá trị thiếu.

Giá trị thay vào các vị trí bị thiếu

Count	Weight
0	0
1	1
36	36
90	90
96	96
57	57
6	6
0	0
0	0

- Thực hiện xóa các giá trị bị thiếu:

Với node-caps và breast-quad đều có type là nominal ta dùng ReplaceMissingWithUserConstant để thay thế các giá trị thiếu bằng “Unknown”. Sau đó trong Filter chọn Choose → filters → unsupervised → instance → RemoveWithValues → Lựa chọn thông số phù hợp → Apply.

Thư tự thuộc tính có giá trị thiếu

matchMissingValues thành True

Vị trí có giá trị thiếu đã được đánh dấu bằng "Unknown"

No.	Label	Count	Weight
1	Unknown	8	8
2	yes	36	36
3	no	222	222

7. Let's explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend.

- Ý nghĩa của biểu đồ trong WEKA Explorer: Đồ thị gồm hai màu là màu xanh và màu đỏ. Màu đỏ biểu thị tại mỗi khoảng thuộc tính được chọn, có bao nhiêu mẫu cho kết quả là recurrence-events. Còn tương tự với màu xanh là số mẫu cho kết quả no-recurrence-events. Có thể đặt cho đồ thị này là đồ thị phân lớp.

3.2 Khám phá dữ liệu Weather.

1. How many attributes does this data set have? How many samples? Which attributes have data type categorical? Which attributes have a data type that is numerical? Which attribute is used for the label?

- Dataset có 5 thuộc tính.
- Có 14 mẫu.

Current relation

Relation: weather	Attributes: 5
Instances: 14	Sum of weights: 14

Attributes

All None Invert Pattern

No.	Name
1 <input checked="" type="checkbox"/>	outlook
2 <input type="checkbox"/>	temperature
3 <input type="checkbox"/>	humidity
4 <input type="checkbox"/>	windy
5 <input type="checkbox"/>	play

- Thuộc tính có loại là categorical: outlook, windy, play.

Selected attribute

Name: play
Missing: 0 (0%)

Distinct: 2

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	9	9
2	no	5	5

Selected attribute

Name: windy
Missing: 0 (0%)

Distinct: 2

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	TRUE	6	6
2	FALSE	8	8

Selected attribute

Name: outlook
Missing: 0 (0%)

Distinct: 3

Type: Nominal
Unique: 0 (0%)

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

- Thuộc tính có loại là numerical: temperature, humidity.

Selected attribute		
Name: temperature		Type: Numeric
Missing: 0 (0%)	Distinct: 12	Unique: 10 (71%)
Statistic	Value	
Minimum	64	
Maximum	85	
Mean	73.571	
StdDev	6.572	

Selected attribute		
Name: humidity		Type: Numeric
Missing: 0 (0%)	Distinct: 10	Unique: 7 (50%)
Statistic	Value	
Minimum	65	
Maximum	96	
Mean	81.643	
StdDev	10.285	

- Thuộc tính được dùng làm label: play.

Selected attribute				
Name: play			Type: Nominal	
Missing: 0 (0%)		Distinct: 2	Unique: 0 (0%)	
No.	Label	Count	Weight	
1	yes	9	9	
2	no	5	5	

Class: play (Nom) ▼

2. Let's list five-number summary of two attributes temperature and humidity. Does WEKA provide these values?

- Danh sách five-number summary của hai thuộc tính “temperature” và humidity.

	temperature	humidity
Minimum	64.0	65.0
Lower quartile(Q1)	69.0	70.0
Median(Q2)	72.0	82.5
Upper quartile(Q3)	80.0	90.0
Maximum	85.0	96.0

- WEKA chỉ cung cấp hai trong năm giá trị trên đó là minimum và maximum.

Selected attribute		
Name: humidity		Type: Numeric
Missing: 0 (0%)	Distinct: 10	Unique: 7 (50%)
Statistic	Value	
Minimum	65	
Maximum	96	
Mean	81.643	
StdDev	10.285	

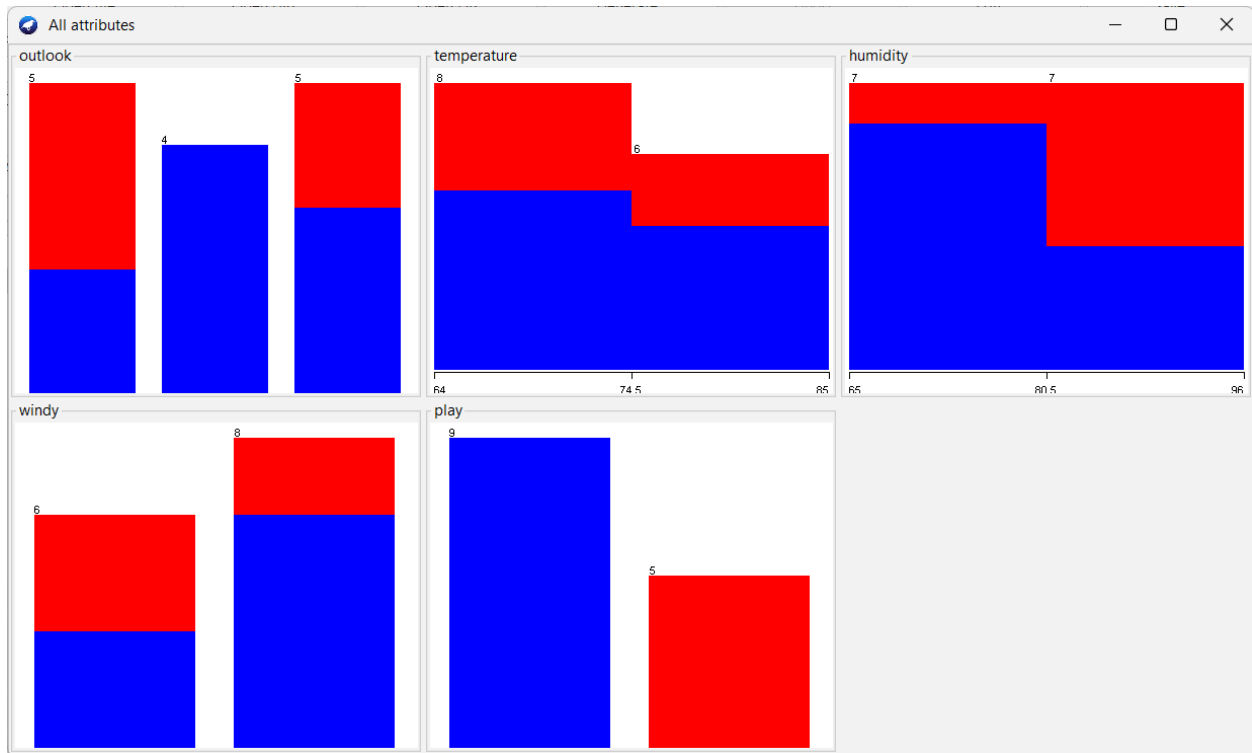
Selected attribute		
Name: temperature		Type: Numeric
Missing: 0 (0%)	Distinct: 12	Unique: 10 (71%)
Statistic	Value	
Minimum	64	
Maximum	85	
Mean	73.571	
StdDev	6.572	

3. Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.

Các biểu đồ đều biểu diễn phân bố các giá trị của thuộc tính cùng tên (ví dụ: biểu đồ đầu tiên có tên là “outlook” và biểu diễn phân bố giá trị của thuộc tính “outlook”).

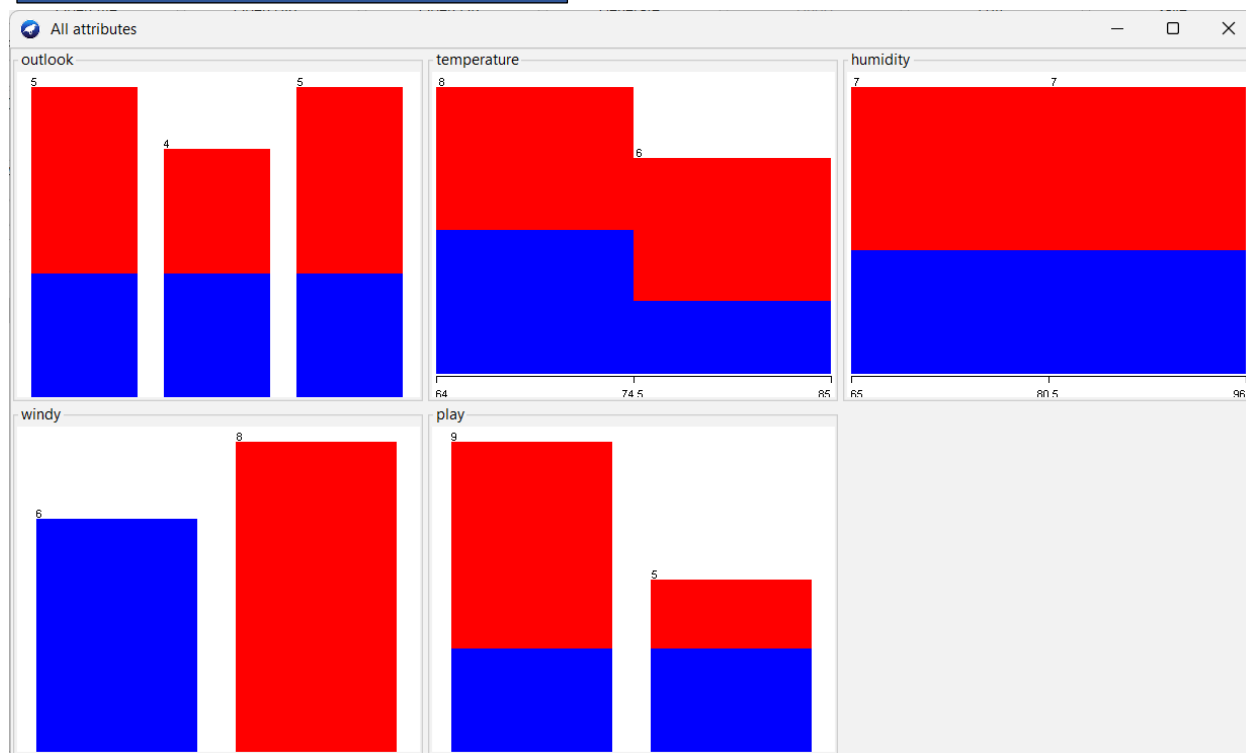
Đối với các cột có dạng numeric như “temperature” và “humidity” thì biểu đồ biểu diễn phân bố các giá trị theo một khoảng nhất định.

Các biểu đồ trong tab Explorer khi chọn class “play”:



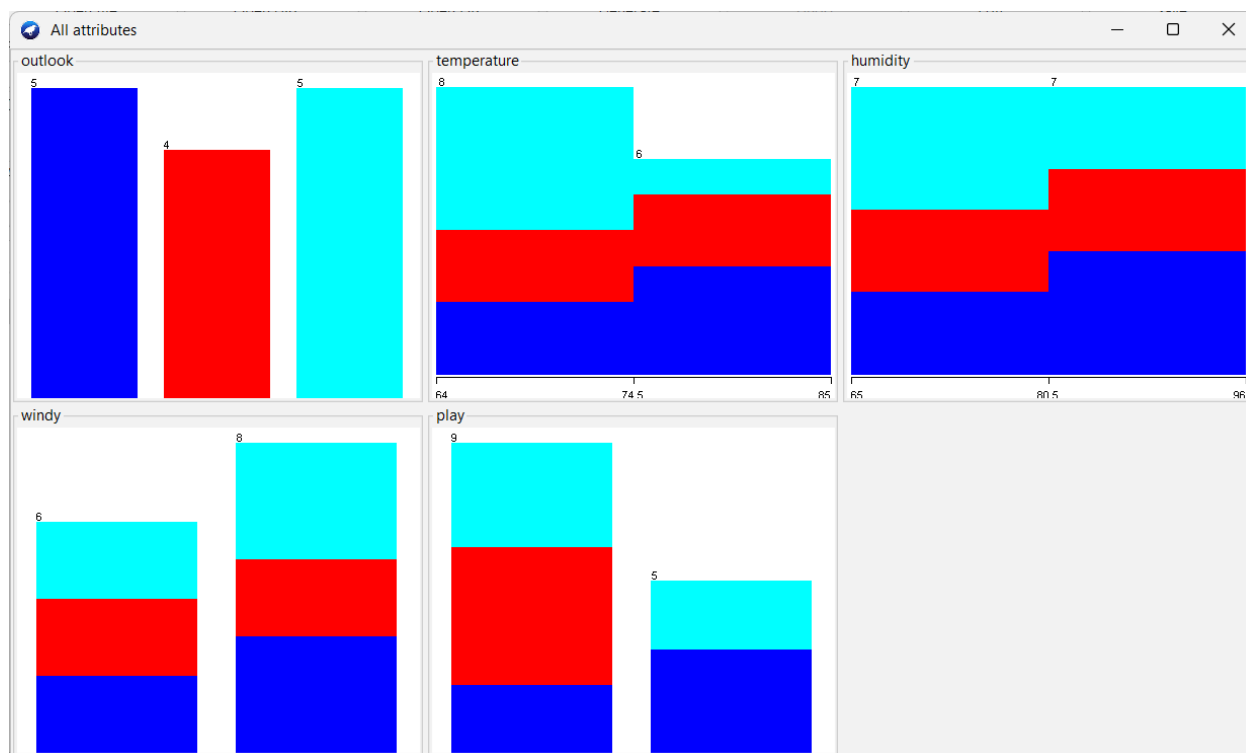
Màu của các cột biểu diễn giá trị của cột “play”, với màu đỏ tượng trưng cho các hàng có play = “no” và màu xanh tượng trưng cho các hàng có play = “yes”.

Các biểu đồ trong tab Explorer khi chọn class “windy”:



Ở đây màu xanh tượng trưng cho các hàng có windy = "TRUE" và màu đỏ tượng trưng cho các hàng có windy = "FALSE".

Các biểu đồ trong tab "Explorer" khi chọn class "outlook":



Phần màu xanh nhạt tượng trưng cho các hàng có outlook = “rainy”, phần màu xanh đậm tượng trưng cho các hàng có outlook = “sunny” và màu đỏ tượng trưng cho các hàng có outlook = “overcast”

Các biểu đồ trong tab “Explorer” khi chọn class “temperature”:

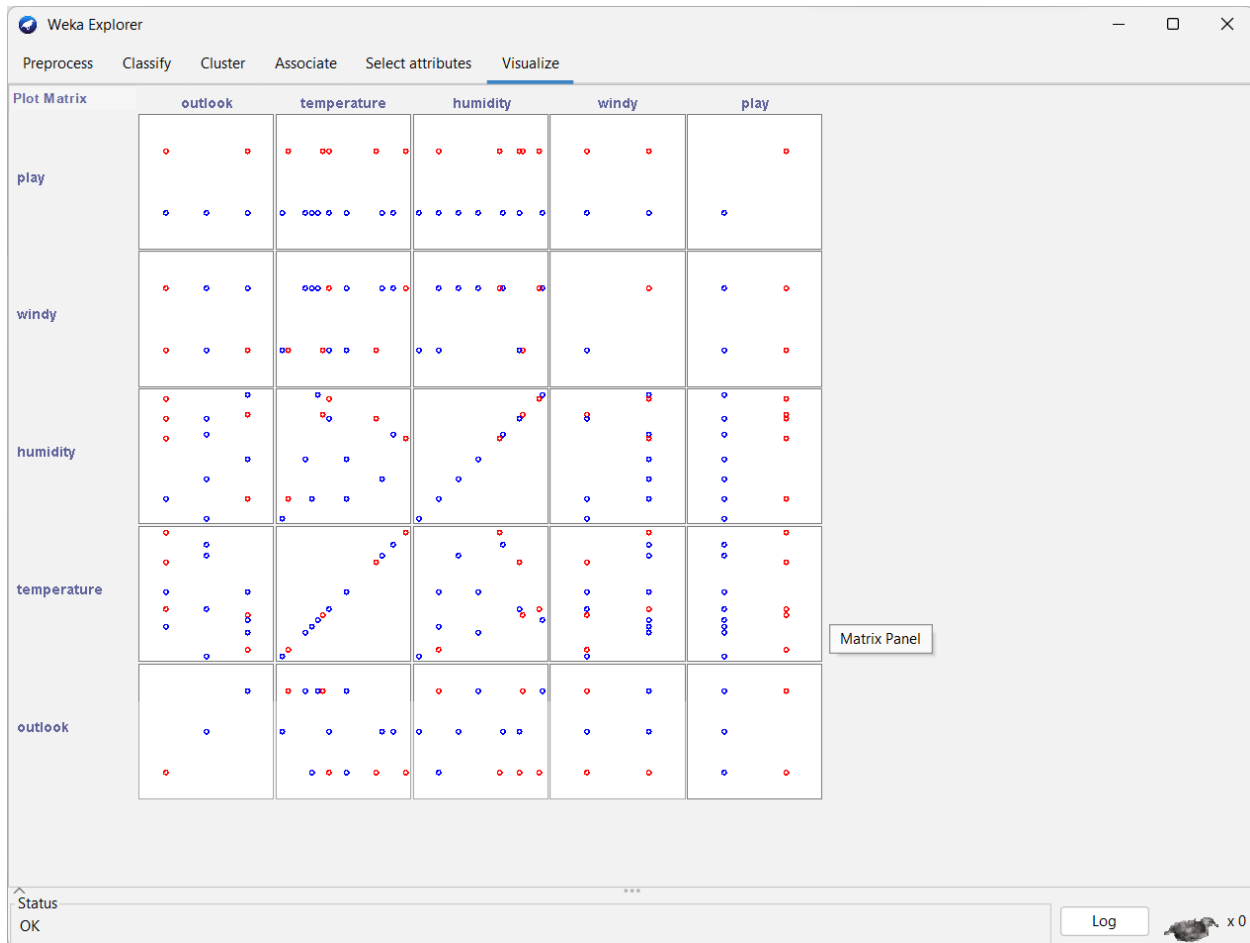


Các biểu đồ trong tab Explorer khi chọn class “humidity”:



Vì cả hai cột “temperature” và “humidity” đều có kiểu dữ liệu numeric nên Weka không phân chia theo màu trong từng cột nữa. Thay vào đó các biểu đồ phân bố giá trị của các thuộc tính numeric đều có các cột màu trắng và các biểu đồ phân bố giá trị của các thuộc tính nominal có các cột màu đen.

4. Let’s move to the Visualize tag. What’s the name of this chart? Do you think there are any pairs of different attributes that have correlated?



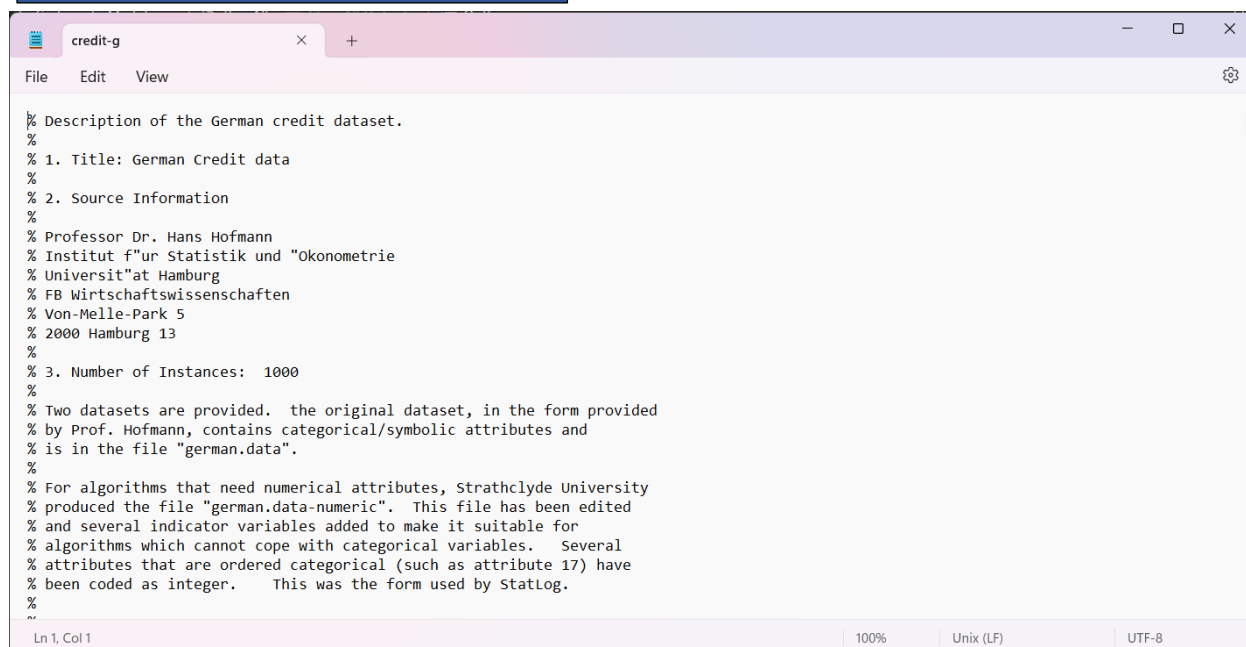
Biểu đồ được sử dụng là biểu đồ scatterplot.

Một số tương quan giữa các thuộc tính:

- “temperature” và “humidity”: 2 thuộc tính này tỉ lệ nghịch với nhau.
- “temperature” và “outlook”: outlook = “sunny” thì temperature cao, outlook = “overcast” thì nhiệt độ trung bình, outlook = “rainy” thì nhiệt độ thấp => thời tiết càng nắng thì nhiệt độ càng cao
- “windy” và “humidity”: trời có gió thì độ ẩm thấp và ngược lại.
- “windy” và “temperature”: trời có gió thì nhiệt độ thấp và ngược lại.

3.3 Khám phá dữ liệu tín dụng ở Germany.

1. What is the content of the comments section in credit-g.arff (when opened with any text editor) about? How many samples does the data set have? How many attributes? Describe any five attributes (must have both discrete and continuous attributes).



```
% Description of the German credit dataset.
%
% 1. Title: German Credit data
%
% 2. Source Information
%
% Professor Dr. Hans Hofmann
% Institut f"ur Statistik und "Okonometrie
% Universit"at Hamburg
% FB Wirtschaftswissenschaften
% Von-Melle-Park 5
% 2000 Hamburg 13
%
% 3. Number of Instances: 1000
%
% Two datasets are provided. the original dataset, in the form provided
% by Prof. Hofmann, contains categorical/symbolic attributes and
% is in the file "german.data".
%
% For algorithms that need numerical attributes, Strathclyde University
% produced the file "german.data-numeric". This file has been edited
% and several indicator variables added to make it suitable for
% algorithms which cannot cope with categorical variables. Several
% attributes that are ordered categorical (such as attribute 17) have
% been coded as integer. This was the form used by Statlog.
%
```

Mục comments chứa các thông tin về dataset Credits in Germany. Có 1000 samples và 20 attributes.

Miêu tả 5 thuộc tính bất kì:

- Thuộc tính 1: “checking status”
Loại: nominal
Số giá trị thiếu: 0 (0%)
Số giá trị khác nhau: 4
Số giá trị chỉ xuất hiện 1 lần: 0 (0%)
- Thuộc tính 2: “duration”
Loại: numeric, quantitative, discrete
Số giá trị thiếu: 0 (0%)
Số giá trị khác nhau: 33
Số giá trị chỉ xuất hiện 1 lần: 5 (1%)
- Thuộc tính 3: “credit history”
Loại: nominal
Số giá trị thiếu: 0 (0%)
Số giá trị khác nhau: 5
Số giá trị chỉ xuất hiện 1 lần: 0 (0%)
- Thuộc tính 4: “duration”
Loại: nominal
Số giá trị thiếu: 0 (0%)

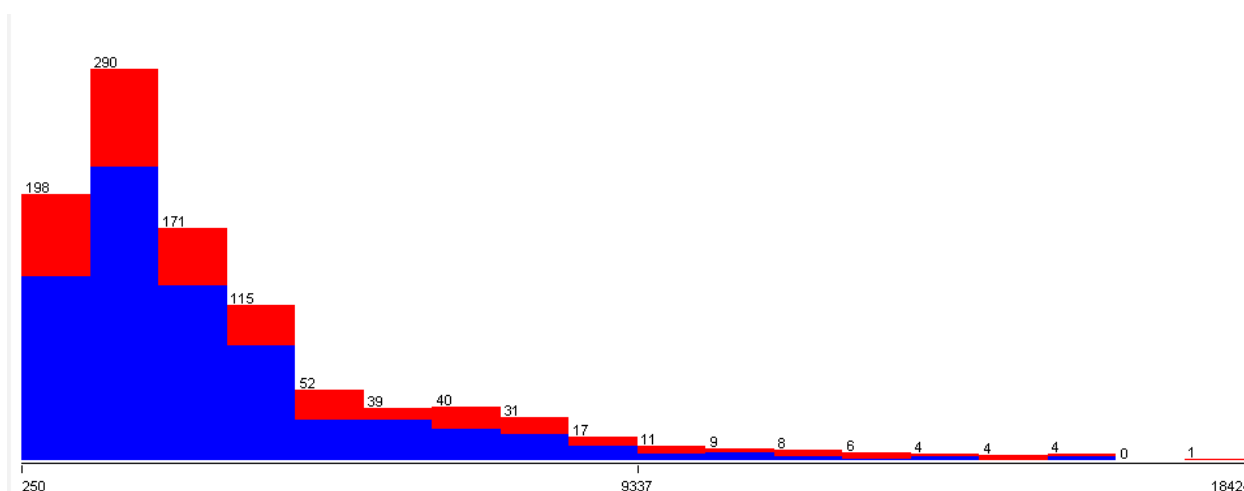
- Số giá trị khác nhau: 10
- Số giá trị chỉ xuất hiện 1 lần: 0 (0%)
- Thuộc tính 5: “credit amount”
 - Loại: numeric, quantitative, continuous
 - Số giá trị thiếu: 0 (0%)
 - Số giá trị khác nhau: 921
 - Số giá trị chỉ xuất hiện 1 lần: 847 (85%)

2. Which attribute is used for the label?

Thuộc tính “class” được dùng làm nhãn.

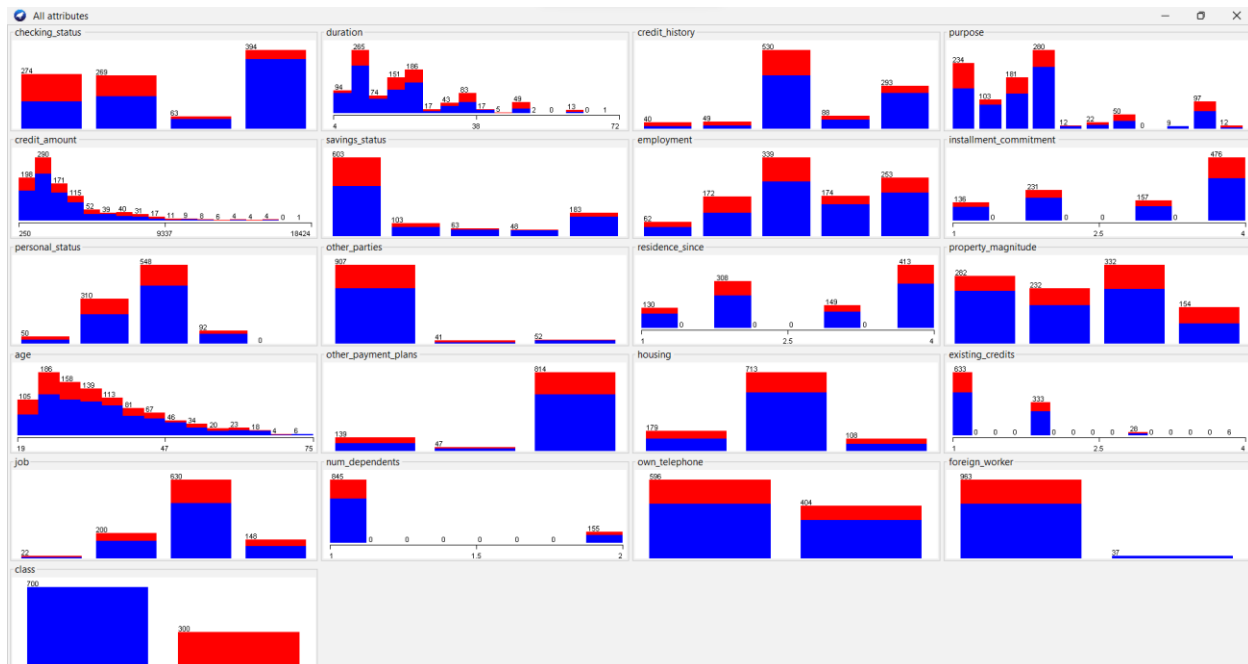
3. Let’s describe the distribution of continuous attributes (left or right-skewed)?

Ta sẽ miêu tả phân bố dữ liệu của cột “credit_amount”



Có thể thấy phân bố dữ liệu của cột này lệch về bên phải.

4. Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.



Các biểu đồ này biểu diễn phân bố giá trị theo từng cột. Phần màu đỏ tượng trưng cho các hàng có nhãn class = “bad”, trong khi phần màu xanh tượng trưng cho các hàng có “class” = “good”.

5. Let’s move to the Select attributes tag. Describe all of the options for attribute selection.

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator
Choose **CorrelationAttributeEval**

Search Method
Choose **Ranker -T -1.7976931348623157E308 -N 5**

Attribute Selection Mode
☒ Use full training set
☐ Cross-validation Folds: 10 Seed: 1

(Nom) class
 Start Stop

Result list (right-click for options)
 18:06:04 - Ranker - CorrelationAttributeEval

Attribute selection output

```

=====
personal_status
other_parties
residence_since
property_magnitude
age
other_payment_plans
housing
existing_credits
job
num_dependents
own_telephone
foreign_worker
class

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
Correlation Ranking Filter

Ranked attributes:
0.233 1 checking_status
0.215 2 duration
0.155 5 credit_amount
0.132 6 savings_status
0.121 15 housing

Selected attributes: 1,2,5,6,15 : 5
  
```

Status
OK

Log

Có 4 mục cần chú ý: Attribute Evaluator, Search Method, Attribute Selection Mode và class (ngay dưới mục Attribute Selection Mode).

Mục Attribute Evaluator gồm nhiều lựa chọn, có thể kể tới:

- CfsSubsetEval: Đánh giá mức độ quan trọng của một tập hợp con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng đặc điểm cùng với mức độ dư thừa giữa chúng (tập các thuộc tính có tương quan cao với class attribute và ít bị tương quan với nhau (intercorrelate) được ưu tiên).

- ClassifierAttributeEval: đánh giá mức độ quan trọng của một thuộc tính bằng cách sử dụng bộ phân lớp do người dùng chỉ định.

- CorrelationAttributeEval: đánh giá mức độ quan trọng của một thuộc tính bằng cách đo lường mối tương quan giữa nó và lớp bằng hệ số tương quan Pearson.

- ClassifierSubsetEvaluator: đánh giá giá trị thuộc tính trên dữ liệu training hoặc một tập cross validation (hold-out) riêng biệt.

- PrincipleComponents: sử dụng giải thuật PCA (phân tích các thành phần chính) để giảm số chiều (số cột) của dữ liệu.

Mục Search Method có 3 lựa chọn:

- GreedyStepwise: Thực hiện tìm kiếm tiến hoặc lùi sử dụng giải thuật tham lam trong không gian thuộc tính.
- BestFirst: Tìm kiếm leo đồi sử dụng giải thuật tham lam trong không gian thuộc tính, cho phép lần ngược (backtracking).
- Ranker: Xếp hạng các thuộc tính theo giá trị tìm được của chúng (khi sử dụng evaluator nhất định như CorrelationAttributeEval, InfoGainEval...)

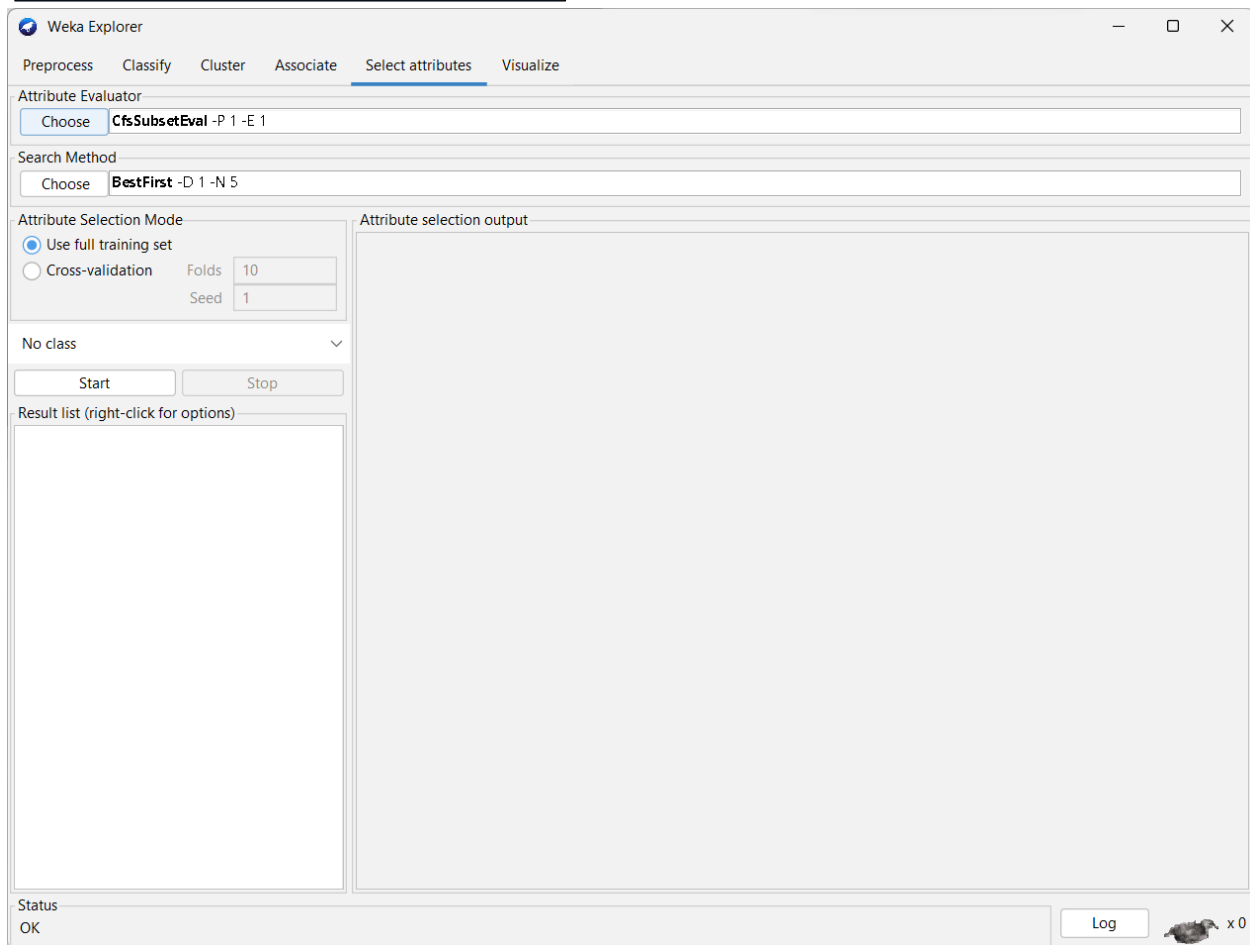
Mục Attribute Selection Mode cho phép chọn giữa 2 lựa chọn: use full training set (sử dụng toàn bộ dữ liệu) và sử dụng cross-validation (cho phép chọn số folds và seed tùy ý).

Mục class cho phép chọn thuộc tính bất kỳ để làm class.

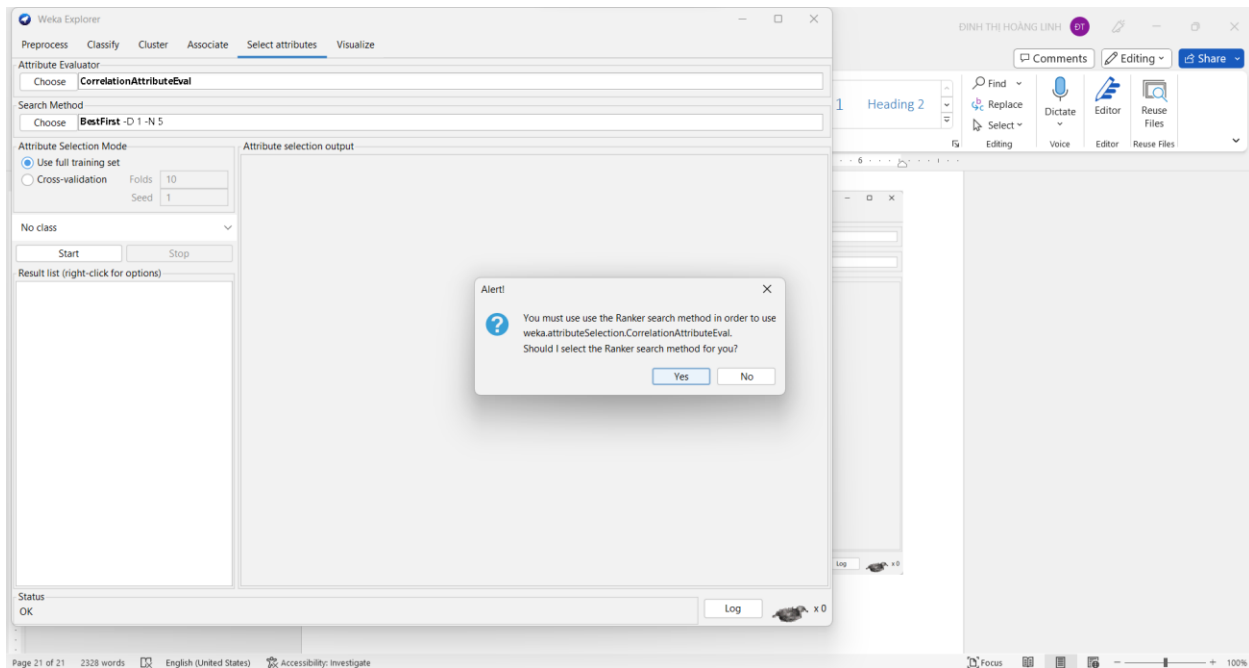
5. Which options should be used to select the 5 attributes with the highest correlation? (Step-by-step description, with step-by-step photos and final results)

Để chọn 5 thuộc tính có tương quan cao nhất ta cần sử dụng bộ lọc CorrelationAttributeEval.

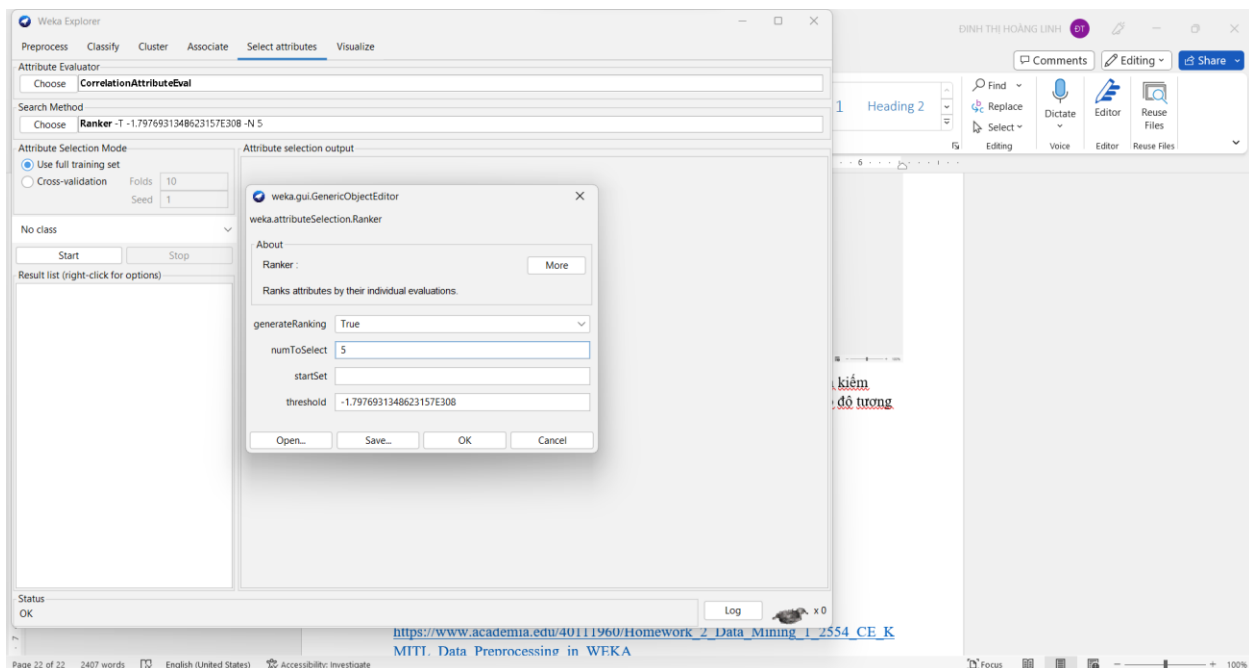
Bước 1: Từ menu Explorer, chọn tab Select Attributes.



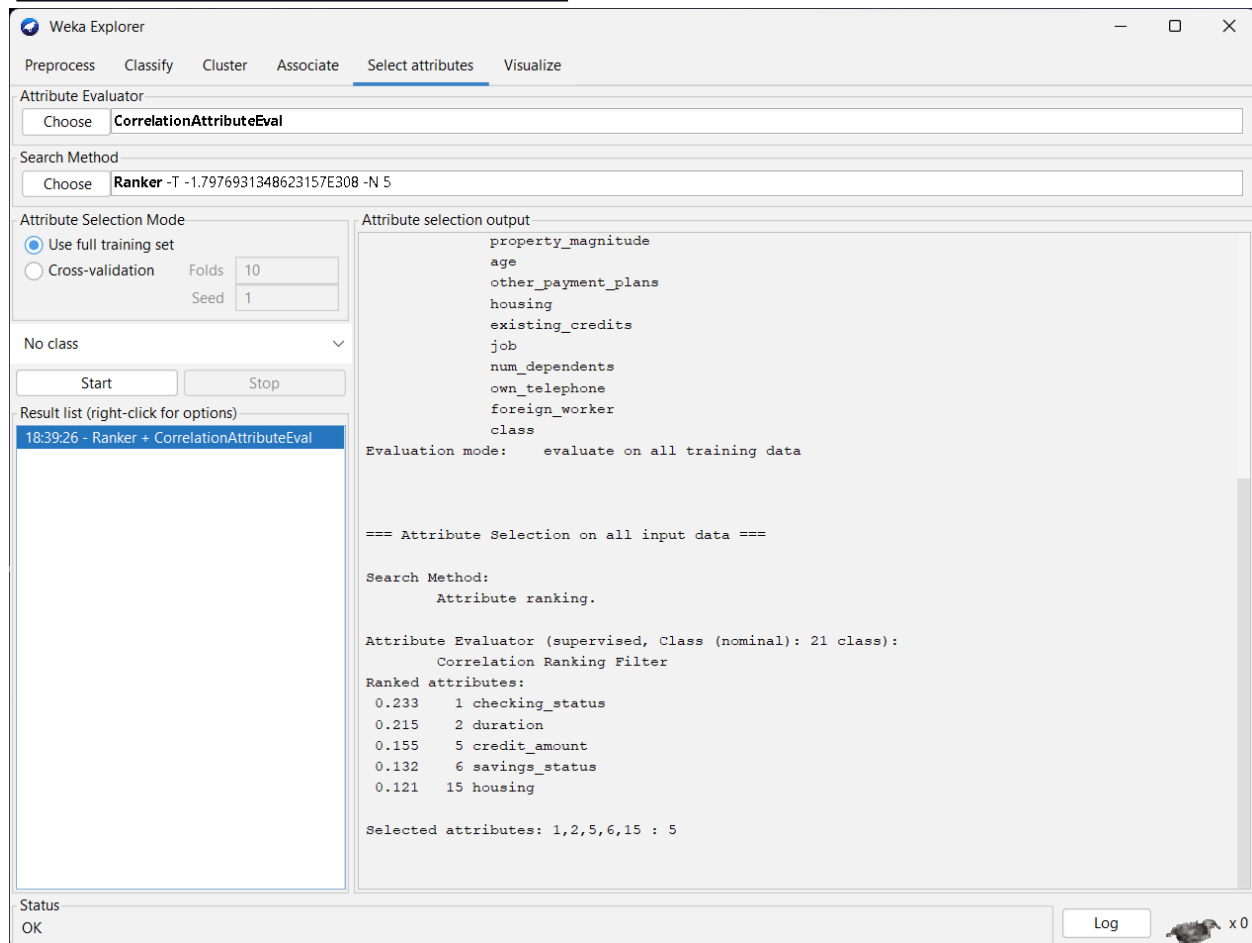
Bước 2: Ở mục AttributeEvaluator chọn CorrelationAttributeEval, ở đây Weka hiện hộp thoại bắt ta sử dụng Search Method Ranker, chọn Yes để Weka tự động đổi Search Method luôn (đương nhiên vẫn có thể tự đổi nếu biết evaluator nào đi với search method nào).



Bước 3: Click vào text box có chữ “Ranker” để tinh chỉnh thuật toán tìm kiếm thuộc tính. Nhập “5” vào trường “numToSelect” để chọn 5 thuộc tính có độ tương quan cao nhất với thuộc tính class. Nhấn “OK”.



Bước 4: Chọn “Start”. Sau khi thuật toán chạy xong, kết quả sẽ hiện ở textbox “Attribute selection output”.



Có thể thấy 5 thuộc tính được chọn lần lượt là “checking_status”, “duration”, “credit_amount”, “savings_status” và “housing”.

4. Tiền xử lý dữ liệu trong python:

Về căn bản thì nhóm đã làm xong đủ 8 chức năng.

4.1 Xuất ra các cột có giá trị thiếu:

Output của terminal sau khi gọi lệnh help:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./extract.py -h
usage: extract.py [-h] -f FILENAME

Extract column with missing values.

options:
  -h, --help            show this help message and exit
  -f FILENAME, --filename FILENAME
                        Input filename.
```

Kết quả test trên tập tin “house-price.csv”:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./extract.py -f house-prices.csv
Column 3
Column 6
Column 25
Column 26
Column 30
Column 31
Column 32
Column 33
Column 35
Column 57
Column 58
Column 59
Column 60
Column 63
Column 64
Column 72
Column 73
Column 74
```

4.2 Đếm số hàng có giá trị thiếu:

Output của terminal sau khi gọi lệnh help:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./count.py -h
usage: count.py [-h] -f FILENAME

Count the number of lines with missing data.

options:
  -h, --help            show this help message and exit
  -f FILENAME, --filename FILENAME
                        Input filename.
```

Kết quả test trên tập tin “house-prices.csv”:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./count.py -f house-prices.csv
Number of rows with missing values: 1000
```

4.3 Điền các giá trị thiếu sử dụng mean, median và mode:

Output của terminal sau khi gọi lệnh help:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./fill.py -h
usage: fill.py [-h] -i INPUT -o {fill,find} [-c COLUMN] [-m {mean,median,mode}] [-of OUTPUT]

Fill missing values in column using mean, median or mode.

options:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input filename.
  -o {fill,find}, --option {fill,find}
                        find: find name and datatype (numeric/categorical) of columns, fill: fill missing value
  -c COLUMN, --column COLUMN
                        Name of column to be filled.
  -m {mean,median,mode}, --method {mean,median,mode}
                        Method to fill column with.
  -of OUTPUT, --output OUTPUT
                        Output filename.
```

Kết quả test trên tập tin “house-prices.csv” được lưu trong thư mục output, gồm 3 file:

- output3_mean.csv:
Cột được chọn để test là cột “LotFrontage” (để dễ kiểm tra output).
Command line argument:
`py -3 ./fill.py -i house-prices.csv -o fill -c LotFrontage -m mean -of output3`
- output3_median.csv:
Cột được chọn để test là cột “LotFrontage”.
Command line argument:
`py -3 ./fill.py -i house-prices.csv -o fill -c LotFrontage -m median -of output3`
- output3_mode.csv:
Cột được chọn để test là cột “Alley”.
Command line argument:
`py -3 ./fill.py -i house-prices.csv -o fill -c Alley -m mode -of output3`

4.4 Xóa các hàng có tỉ lệ giá trị thiếu cao hơn 1 ngưỡng nhất định:

Output của terminal sau khi gọi lệnh help:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./delete_rows.py -h
usage: delete_rows.py [-h] -i INPUT -t THRESHOLD -o OUTPUT

Delete rows with percentage of missing values over a certain threshold.

options:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input filename.
  -t THRESHOLD, --threshold THRESHOLD
                        Define a threshold.
  -o OUTPUT, --output OUTPUT
                        Output filename.
```

Kết quả test trên tập tin “house-prices.csv” là tập tin output4.csv được lưu trong thư mục output, ngưỡng được chọn là 0.1.

Command line argument:

```
py -3 ./delete_rows.py -if house-prices.csv -t 0.1 -of output4.csv
```

4.5 Xóa các cột có tỉ lệ giá trị thiếu cao hơn 1 ngưỡng nhất định:

Output của terminal sau khi gọi lệnh help:

```
usage: delete_columns.py [-h] -i INPUT -t THRESHOLD -o OUTPUT

Delete columns with percentage of missing values over a certain threshold.

options:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input filename.
  -t THRESHOLD, --threshold THRESHOLD
                        Define a threshold.
  -o OUTPUT, --output OUTPUT
                        Output filename.
```

Kết quả test trên tập tin “house-prices.csv” là tập tin output5.csv được lưu trong thư mục output, ngưỡng được chọn là 0.5.

Command line argument:

```
py -3 ./delete_columns.py -if house-prices.csv -t 0.5 -of output5.csv
```

4.6 Xóa các hàng trùng lặp:

Output của terminal sau khi gọi lệnh help:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./drop_duplicates.py -h
usage: drop_duplicates.py [-h] -i INPUT -o OUTPUT

Drop duplicates.

options:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input filename.
  -o OUTPUT, --output OUTPUT
                        Output filename.
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> |
```

Kết quả test trên tập tin “house-prices.csv” là tập tin output6.csv được lưu trong thư mục output.

Command line argument:

```
py -3 ./drop_duplicates.py -if house-prices.csv -of output6.csv
```

4.7 Chuẩn hóa dữ liệu sử dụng min-max hoặc z-score:

Output của terminal sau khi gọi lệnh help:

```
PS C:\Users\brocc\Dropbox\PC\Downloads\20120130_20120146\Source> py -3 ./normalization.py -h
usage: normalization.py [-h] -if INPUT -o {norm,find} [-c COLUMN] [-m {min-max,z-score}] [-of OUTPUT]

Perform normalization on a column.

options:
  -h, --help            show this help message and exit
  -if INPUT, --input INPUT
                        Input filename.
  -o {norm,find}, --option {norm,find}
                        find: find name and datatype (numeric/categorical) of columns, norm: perform normalization
  -c COLUMN, --column COLUMN
                        Name of column to be filled.
  -m {min-max,z-score}, --method {min-max,z-score}
                        Normalization.
  -of OUTPUT, --output OUTPUT
                        Output filename.
```

Kết quả test trên tập tin “house-prices.csv” được lưu trong thư mục output, gồm 2 file:

- output7_min-max.csv:
Cột được chọn để test là cột “LotArea” (để dễ kiểm tra output).
Command line argument:

```
py -3 ./fill.py -if house-prices.csv -o norm -c LotArea -m min-max -of output7
```
- output7_z-score.csv:
Cột được chọn để test là cột “LotArea”.
Command line argument:

```
py -3 ./fill.py -if house-prices.csv -o fill -c LotArea -m z-score -of output3
```

4.8 Thực hiện tính toán trên 2 cột dữ liệu dạng số:

Output của terminal sau khi gọi lệnh help:

```
usage: calculate.py [-h] -if INPUT -o {find,calc} [-c1 COLUMN1] [-c2 COLUMN2] [-m {add,sub,mul,div}] [-of OUTPUT]

Perform calculation (add/subtract/multiply/divide) on two columns.

options:
  -h, --help            show this help message and exit
  -if INPUT, --input INPUT
                        Input filename.
  -o {find,calc}, --option {find,calc}
                        find: find name and datatype (numeric/categorical) of columns, norm: perform calculation
  -c1 COLUMN1, --column1 COLUMN1
                        Name of column 1.
  -c2 COLUMN2, --column2 COLUMN2
                        Name of column 2.
  -m {add,sub,mul,div}, --method {add,sub,mul,div}
                        Calculation method.
  -of OUTPUT, --output OUTPUT
                        Output filename.
```

Kết quả test trên tập tin “house-prices.csv” được lưu trong thư mục output, gồm 4 file:

- output8_add.csv:
2 cột được chọn để test lần lượt là cột “LotArea” và “LotFrontage” (để dễ kiểm tra output).
Command line argument:
py -3 ./calculate.py -if house-prices.csv -o calc -c1 LotArea -c2 LotFrontage -m add -of output7
- output8_sub.csv:
2 cột được chọn để test lần lượt là cột “LotArea” và “LotFrontage” (để dễ kiểm tra output).
Command line argument:
py -3 ./calculate.py -if house-prices.csv -o calc -c1 LotArea -c2 LotFrontage -m sub -of output7
- output8_mul.csv:
2 cột được chọn để test lần lượt là cột “LotArea” và “LotFrontage” (để dễ kiểm tra output).
Command line argument:
py -3 ./calculate.py -if house-prices.csv -o calc -c1 LotArea -c2 LotFrontage -m mul -of output7
- output8_div.csv:
2 cột được chọn để test lần lượt là cột “LotArea” và “LotFrontage” (để dễ kiểm tra output).
Command line argument:
py -3 ./calculate.py -if house-prices.csv -o calc -c1 LotArea -c2 LotFrontage -m div -of output7

TÀI LIỆU THAM KHẢO

1. *Lecture slides*
2. <https://www.cs.waikato.ac.nz/ml/weka/>
3. Textbook: J. Han and M. Kamber: *Data Mining, Concepts, and Techniques, Second Edition - Chapter 2: Data Preprocessing.*
4. Textbook: I. H. Witten and E. Frank: *Data mining, Practical Machine Learning Tools and Techniques.*
5. WEKA Datasets, Classifier And J48 Algorithm For Decision Tree (softwaretestinghelp.com)
6. Argparse Tutorial — Python 3.11.2 documentation