

KHOA CÔNG NGHỆ THÔNG TIN  
ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA  
TPHCM

# Khai thác dữ liệu và ứng dụng Lab 1: Tiền xử lý và khám phá dữ liệu



Nhóm sinh viên thực hiện: 20120099 – Trần Huỳnh Hương

20120547 – Võ Thành Phong

ĐỒ ÁN/BÀI TẬP MÔN HỌC - KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG  
HỌC KỲ II – NĂM HỌC 2022-2023



## BẢNG THÔNG TIN CHI TIẾT NHÓM

MSSV	Họ tên	Email
20120099	Trần Huỳnh Hương	20120099@student.hcmus.edu.vn
20120547	Võ Thành Phong	20120547@student.hcmus.edu.vn

Bảng phân công & đánh giá hoàn thành công việc			
Người thực hiện	Công việc thực hiện	Tỉ lệ đóng góp	Mức độ hoàn thành
Trần Huỳnh Hương	<ul style="list-style-type: none"> <li>- Weka:</li> <li>+ Khám phá bộ dữ liệu ‘Weather’.</li> <li>+ Khám phá bộ dữ liệu ‘Credit in Germany’.</li> <li>- Tiền xử lý dữ liệu: Cài đặt các hàm sau: <ul style="list-style-type: none"> <li>+ 1. Liệt kê các cột bị thiếu dữ liệu.</li> <li>+ 2. Đếm số dòng bị thiếu dữ liệu.</li> <li>+ 3. Điền các giá trị thiếu bằng mean/median/mode.</li> <li>+ 8. Cộng/Trừ/Nhân/Chia hai cột dữ liệu dạng số.</li> </ul> </li> <li>- Viết báo cáo.</li> </ul>	50%	100%
Võ Thành Phong	<ul style="list-style-type: none"> <li>- Weka:</li> <li>+ Cài đặt Weka.</li> <li>+ Khám phá bộ dữ liệu ‘Breast Cancer’.</li> <li>- Tiền xử lý dữ liệu: Cài đặt các hàm sau: <ul style="list-style-type: none"> <li>+ 4. Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.</li> <li>+ 5. Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.</li> <li>+ 6. Xóa các mẫu trùng lặp.</li> <li>+ 7. Chuẩn hóa một thuộc tính dạng số bằng phương pháp min-max hoặc Z-score.</li> </ul> </li> <li>- Viết báo cáo.</li> </ul>	50%	100%



## MỤC LỤC

A. Yêu cầu của bài .....	4
B. Kết quả.....	4
1. CÀI ĐẶT WEKA .....	4
1.1. Yêu cầu 1: Chụp ảnh màn hình có chứa hàm “Explorer” của Weka.....	4
1.2. Yêu cầu 2: Giải thích ý nghĩa một số vùng quan trọng trong thẻ ‘Preprocess’ của Weka Explorer.....	5
2. LÀM QUEN VỚI WEKA.....	10
2.1. Khám phá tập dữ liệu ‘Breast Cancer’ .....	12
2.1.1. Tập dữ liệu có bao nhiêu mẫu?.....	12
2.1.2. Tập dữ liệu có bao nhiêu thuộc tính?.....	12
2.1.3. Thuộc tính nào được dùng để làm nhãn? Có thay đổi được không? Bằng cách nào? .....	12
2.1.4. Ý nghĩa của mỗi thuộc tính .....	15
2.1.5. Tìm hiểu tình trạng thiếu giá trị (missing values) trong mỗi thuộc tính và mô tả các cách giải quyết tổng quát cho vấn đề thiếu giá trị.....	15
2.1.6. Đề xuất giải pháp cho vấn đề thiếu giá trị trong thuộc tính cụ thể .....	18
2.1.7. Giải thích và thực hiện một số cài đặt cho biểu đồ trong Weka Explorer.....	19
2.2. Khám phá tập dữ liệu Weather (weather.numeric.arff).....	20
2.2.1. Tổng quan về bộ dữ liệu .....	20
2.2.2. Liệt kê 5 số liệu thống kê (five-number summary) của 2 thuộc tính ‘temperature’ và ‘humidity’. Weka có cung cấp những số liệu này không?.....	22
2.2.3. Giải thích ý nghĩa tất cả biểu đồ trong Weka Explorer. Đặt tên cho biểu đồ và mô tả chú thích .....	22
2.2.4. Phân tích thẻ “Visualize” .....	26
2.3. Khám phá tập dữ liệu Germany (credit-g.arff) .....	27
2.3.1. Tổng quan về bộ dữ liệu .....	27
2.3.2. Xác định thuộc tính được chọn làm nhãn.....	34
2.3.3. Mô tả sự phân bố của các thuộc tính có giá trị liên tục (Lệch trái hay lệch phải) .....	35
2.3.4. Giải thích ý nghĩa tất cả biểu đồ trong Weka Explorer. Đặt tên cho biểu đồ và mô tả chú thích .....	37



2.3.5. Giải thích và miêu tả các lựa chọn trong tag Select attributes .....	40
2.3.6. Lựa chọn để lấy được 5 thuộc tính có độ tương quan cao nhất.....	43
<b>3. TIỀN XỬ LÝ DỮ LIỆU BẰNG PYTHON.....</b>	<b>44</b>
3.1. Liệt kê các cột bị thiếu dữ liệu.....	45
3.2 Đếm số dòng bị thiếu dữ liệu .....	46
3.3. Điền các giá trị còn thiếu trong cột với các phương thức: mean và median ( <i>đối với cột có kiểu dữ liệu là numeric và mode đối với cột có kiểu dữ liệu là mode</i> ) .....	46
3.4. Xóa các dòng bị thiếu dữ liệu với ngưỡng cho trước.....	50
3.5. Xóa các cột bị thiếu dữ liệu với ngưỡng cho trước .....	51
3.6. Xóa các mẫu bị trùng lặp .....	52
3.7. Chuẩn hóa một thuộc tính dạng số bằng phương pháp min-max hoặc Z-score.....	53
3.8. Thực hiện cộng, trừ, nhân, chia giữa các thuộc tính có kiểu dữ liệu là numeric... ..	56
C. Tài liệu tham khảo.....	61



## YÊU CẦU ĐỒ ÁN- BÀI TẬP

Loại bài tập	<input type="checkbox"/> Lý thuyết <input checked="" type="checkbox"/> Thực hành <input type="checkbox"/> Đồ án <input type="checkbox"/> Bài tập
Ngày bắt đầu	13/03/2023
Ngày kết thúc	26/03/2023

### A. Yêu cầu của bài

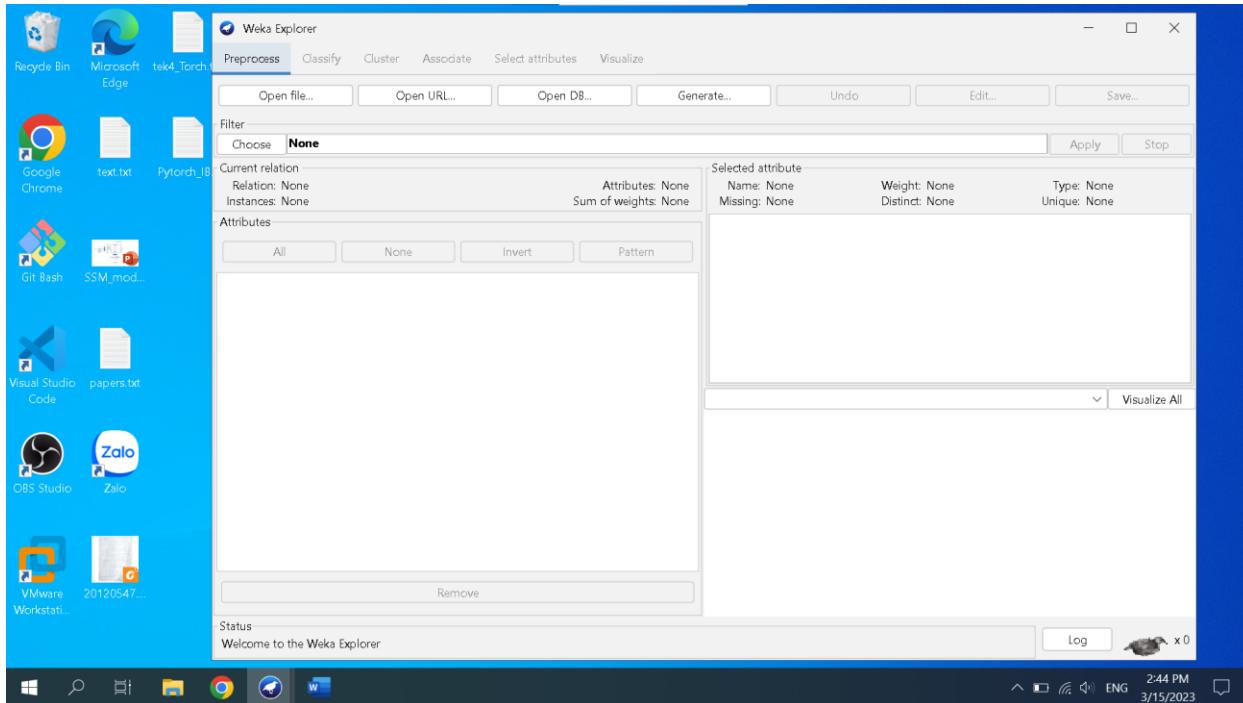
Gồm 3 phần chính như sau:

1. Cài đặt WEKA
2. Làm quen với WEKA với 3 tập dữ liệu: breast cancer, weather và credit in Germany
3. Tiền xử lý dữ liệu bằng Python

### B. Kết quả

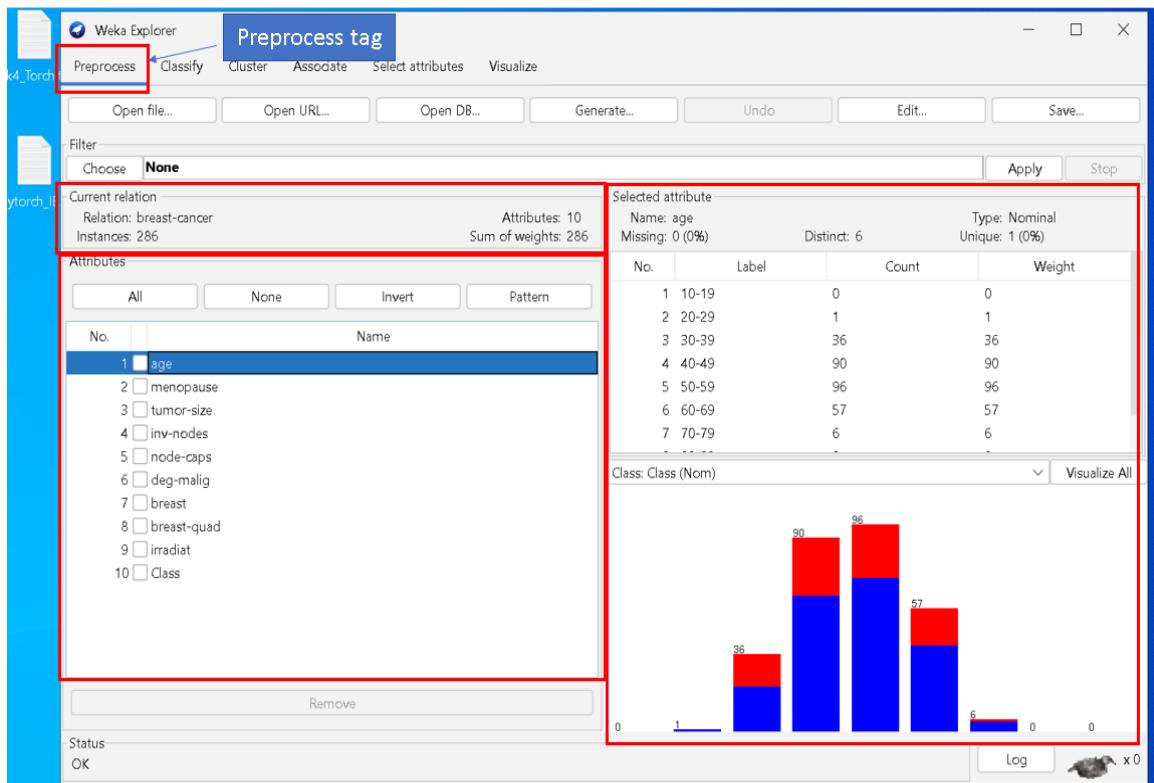
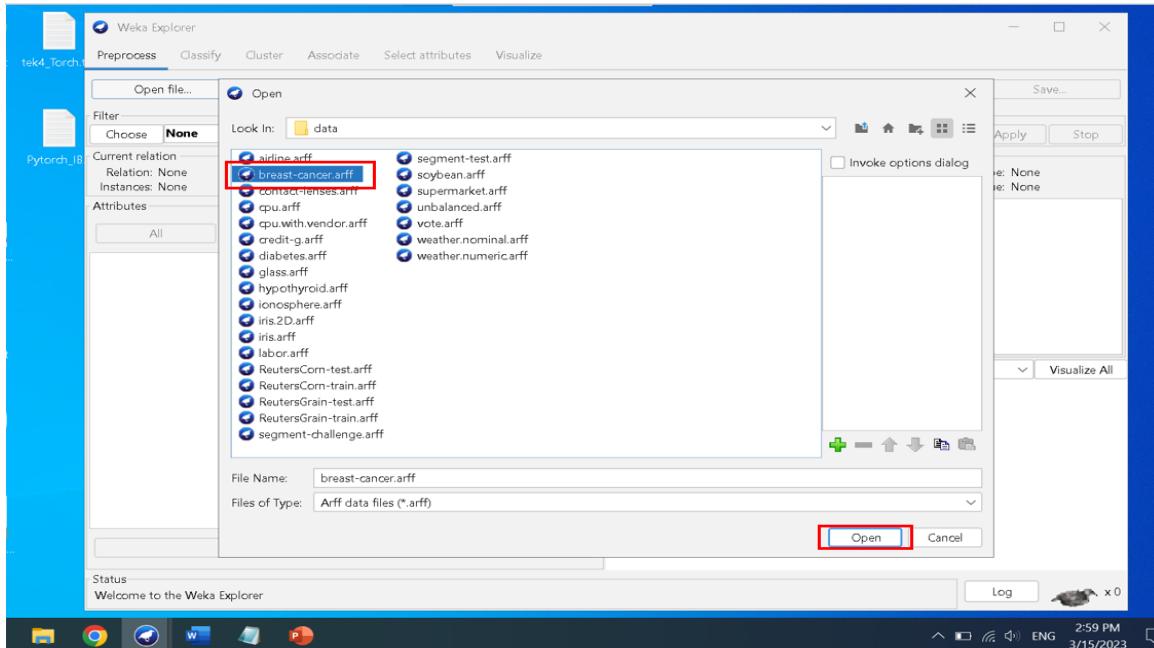
#### 1. CÀI ĐẶT WEKA

##### 1.1. Yêu cầu 1: Chụp ảnh màn hình có chứa hàm “Explorer” của Weka.



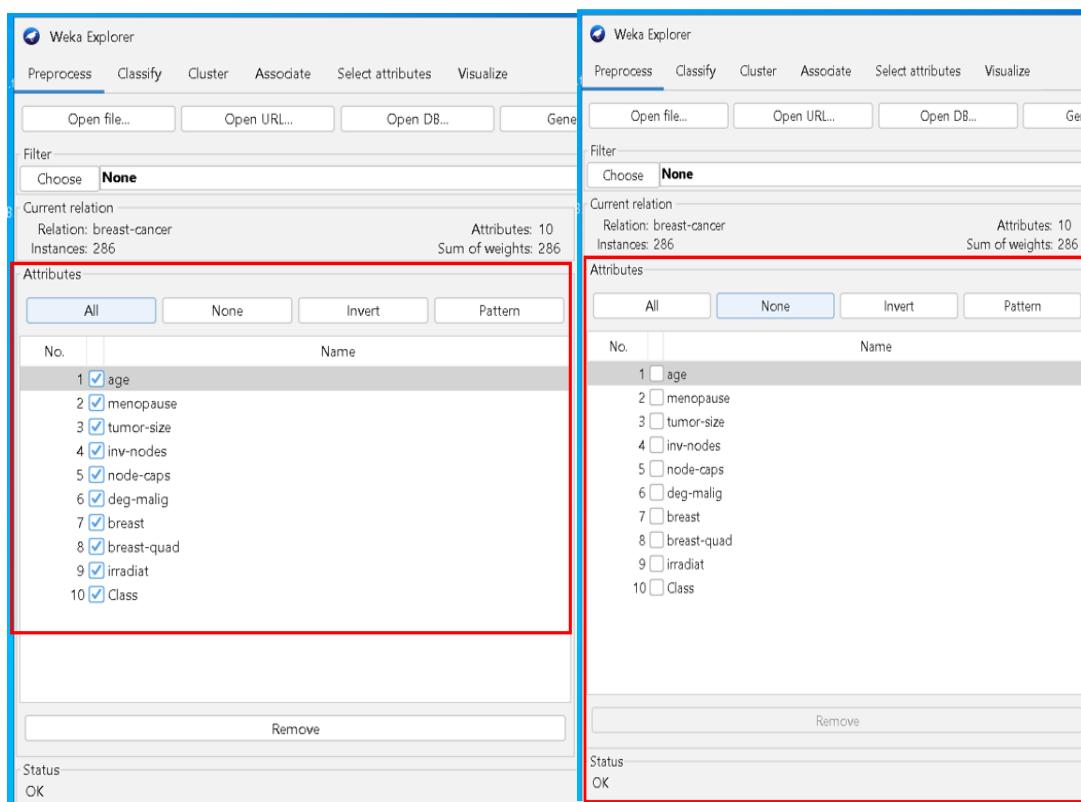
## 1.2. Yêu cầu 2: Giải thích ý nghĩa một số vùng quan trọng trong thẻ ‘Preprocess’ của Weka Explorer.

Trong phần này, nhóm chúng em sẽ thực hiện giải thích sau khi mở bộ dữ liệu có sẵn trong folder ‘data’ có tên là ‘breast-cancer’:



- **Bảng ‘Current relation’:** Cho biết những thông tin liên quan đến toàn bộ tập dữ liệu.

- + ‘Relation’: Tên bộ dữ liệu.
- + ‘Instances’: Số hàng (số mẫu) trong bộ dữ liệu.
- + ‘Attributes’: Số thuộc tính của bộ dữ liệu (số thuộc tính mà mỗi mẫu có) hay nói cách khác là số cột của bộ dữ liệu.
- + ‘Sum of weights’: Cho biết tổng trọng số của các mẫu trong bộ dữ liệu. Trọng số (hay cách hiểu đơn giản hơn là mức độ quan trọng) của mỗi mẫu trong bộ dữ liệu có thể là khác nhau, nếu trọng số của các mẫu là bằng nhau và bằng 1 (hoặc mặc định không gán trọng số) thì giá trị của ‘Sum of weights’ sẽ bằng với giá trị của ‘Instance’.
- **Bảng ‘Attributes’:** Hiển thị tất cả các thuộc tính có trong bộ dữ liệu, và khi chọn vào một thuộc tính nhất định thì thông tin chi tiết về thuộc tính đó sẽ được hiển thị ở bảng ‘Selected attribute’ (sẽ được trình bày sau). Đối với các bộ dữ liệu có quá nhiều thuộc tính thì có lựa chọn ‘All’ (Chọn tất cả thuộc tính), ‘None’ (Bỏ chọn tất cả thuộc tính), ‘Invert’ (Nếu 1 thuộc tính đang được chọn và nhấn invert thì thuộc tính đó sẽ được bỏ chọn và các thuộc tính khác sẽ được chọn), ‘Pattern’ (Lựa chọn thuộc tính có tên theo một mẫu nhất định).





Attributes

No.	Name
1	age
2	menopause
3	<input checked="" type="checkbox"/> tumor-size
4	inv-nodes
5	node-caps
6	deg-malig
7	breast
8	breast-quad
9	irradiat
10	Class

All None Invert Pattern

Chưa Invert

Remove

Status OK

Attributes

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input checked="" type="checkbox"/> menopause
3	<input type="checkbox"/> tumor-size
4	<input checked="" type="checkbox"/> inv-nodes
5	<input checked="" type="checkbox"/> node-caps
6	<input checked="" type="checkbox"/> deg-malig
7	<input checked="" type="checkbox"/> breast
8	<input checked="" type="checkbox"/> breast-quad
9	<input checked="" type="checkbox"/> irradiat
10	<input checked="" type="checkbox"/> Class

All None Invert Pattern

Invert

Remove

Status OK

Attributes

No.	Name
1	age
2	menopause
3	<input type="checkbox"/> tumor-size
4	inv-nodes
5	<input checked="" type="checkbox"/> node-caps
6	deg-malig
7	breast
8	breast-quad
9	irradiat
10	Class

All None Invert Pattern

Input

Enter a Perl regular expression

node-caps

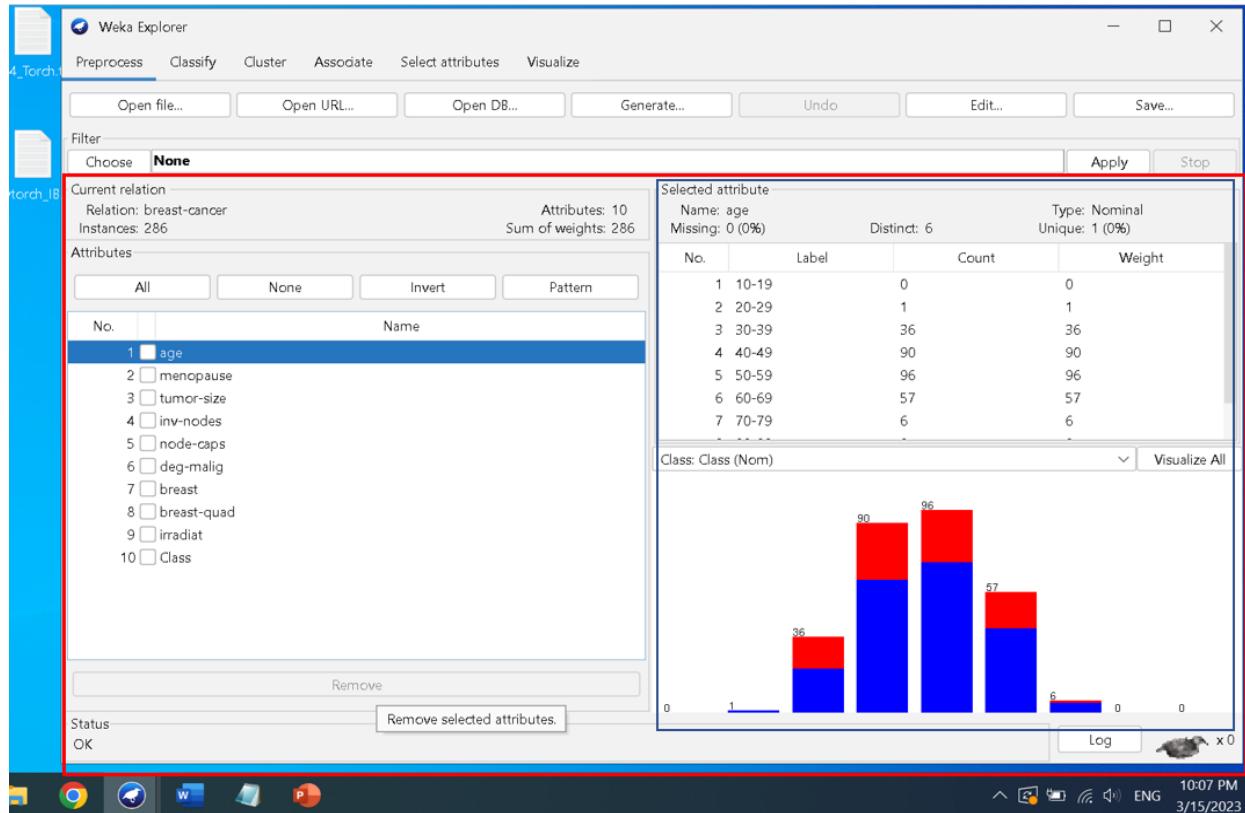
OK Cancel

Remove

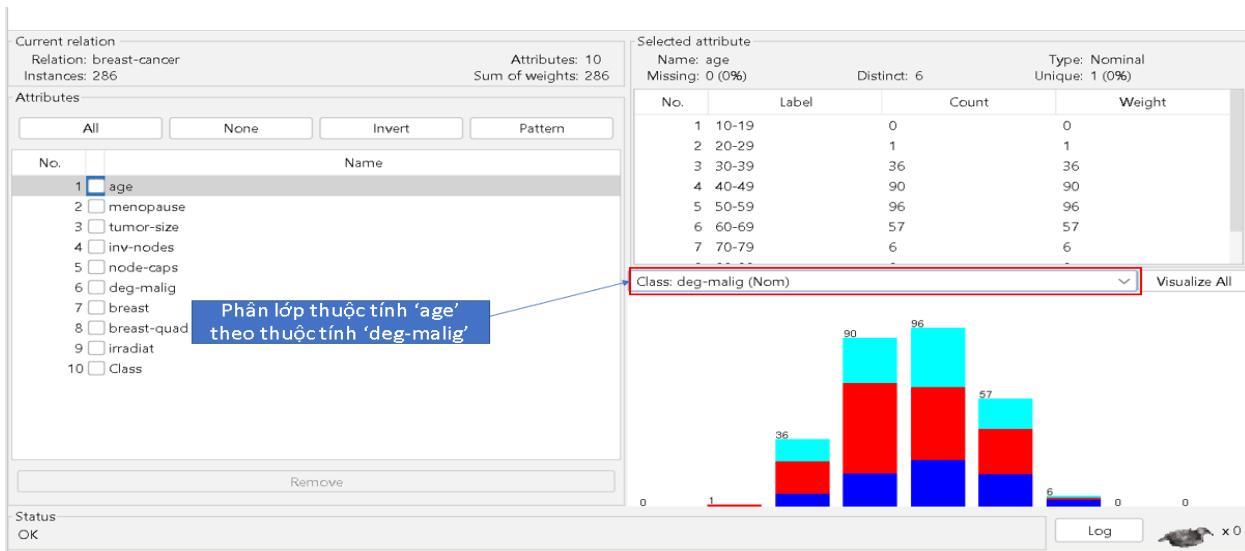
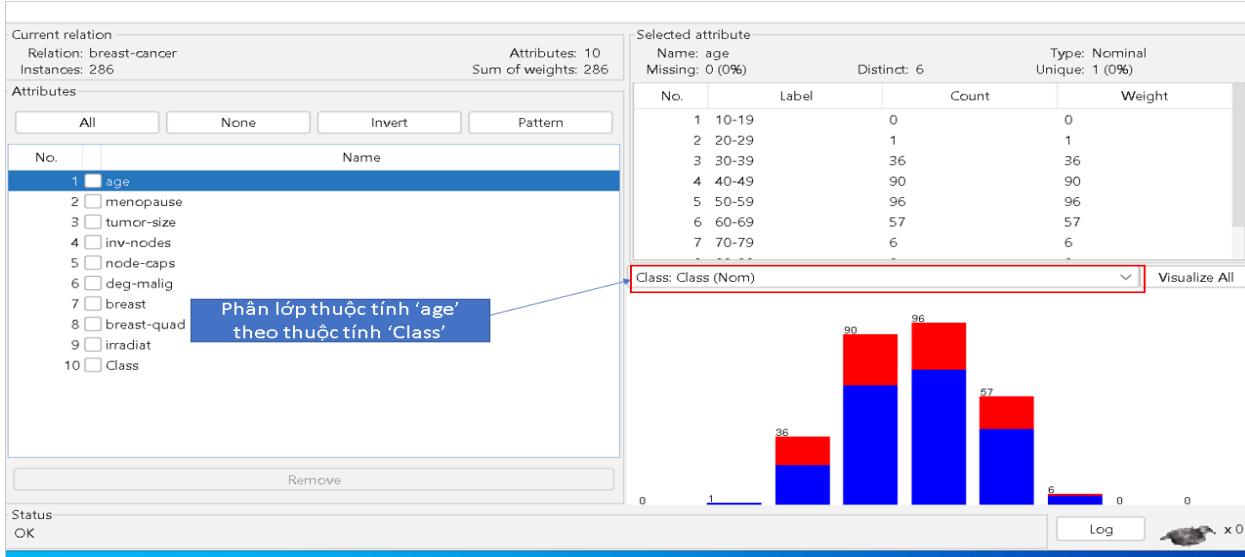
Status OK

Search 'node-caps'

- **Bảng ‘Selected attribute’:** Như đã đề cập ở trên, sau khi chọn một thuộc tính nhất định thì thông tin chi tiết cho thuộc tính được hiển thị ở bảng ‘Selected attribute’. Ví dụ chọn thuộc tính ‘age’



- + ‘**Name**’: tên thuộc tính.
- + ‘**Missing**’: số mẫu bị thiếu trên thuộc tính được chọn.
- + ‘**Type**’: loại thuộc tính (Nominal/Nominal).
- + ‘**Distinct**’: số giá trị riêng biệt trong thuộc tính.
- + ‘**Unique**’: tỉ lệ mẫu trong tập dữ liệu có giá trị thuộc tính này mà tập mẫu khác không có.
- + Bảng tiếp theo phía dưới sẽ hiển thị thông tin chi tiết cho từng giá trị (label) của thuộc tính như có bao nhiêu mẫu có giá trị tương ứng?, trọng số là bao nhiêu?
- + Biểu đồ cuối cùng thể hiện Histogram của thuộc tính, và sẽ được phân lớp tùy theo thuộc tính được chọn để phân lớp là gì.



### \* Phân tích ngắn gọn một số tag khác trong Weka Explorer:

- '**Classify**' tag: Áp dụng các mô hình phân lớp. Classify sẽ hỗ trợ nhiều loại thuật toán (Naïve Bayes, Id3, ...) để đánh giá và dự đoán class dựa trên dữ liệu hiện có.
- '**Cluster**' tag: Tag này cung cấp các thuật toán cho phân cụm. 'Cluster' tag trong Weka Explorer cung cấp một loạt các thuật toán phân cụm, bao gồm k-means, hierarchical, density-based và nhiều thuật toán khác nữa. Dùng để nhóm một tập hợp các đối tượng sao cho các đối tượng trong cùng một nhóm (gọi là một cụm) tương đồng hơn với nhau hơn so với những đối tượng ở các nhóm khác (các cụm).

- '**Associate**' tag: Là thẻ được sử dụng cho các thuật toán khai thác quy luật kết hợp (association rule mining), tìm kiếm các mối quan hệ giữa các biến trong tập dữ liệu.
- '**Select attributes**' tag: Cung cấp một số lựa chọn cho việc lựa chọn thuộc tính như: PCA giảm chiều dữ liệu, đánh giá thuộc tính, đánh giá tập con, các phương pháp tìm kiếm, ....
- '**Visualize**' tag: Dùng để trực quan và quan sát sự trực quan đó của bộ dữ liệu được chọn.



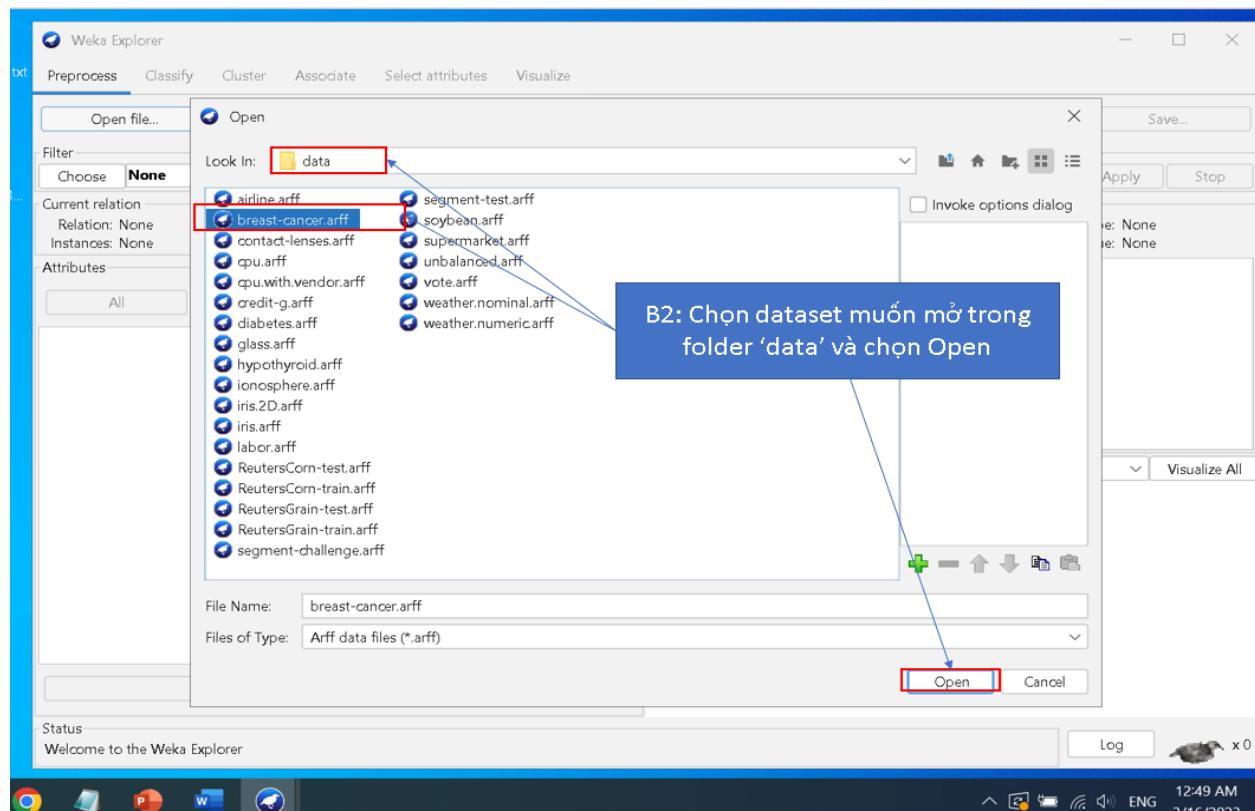
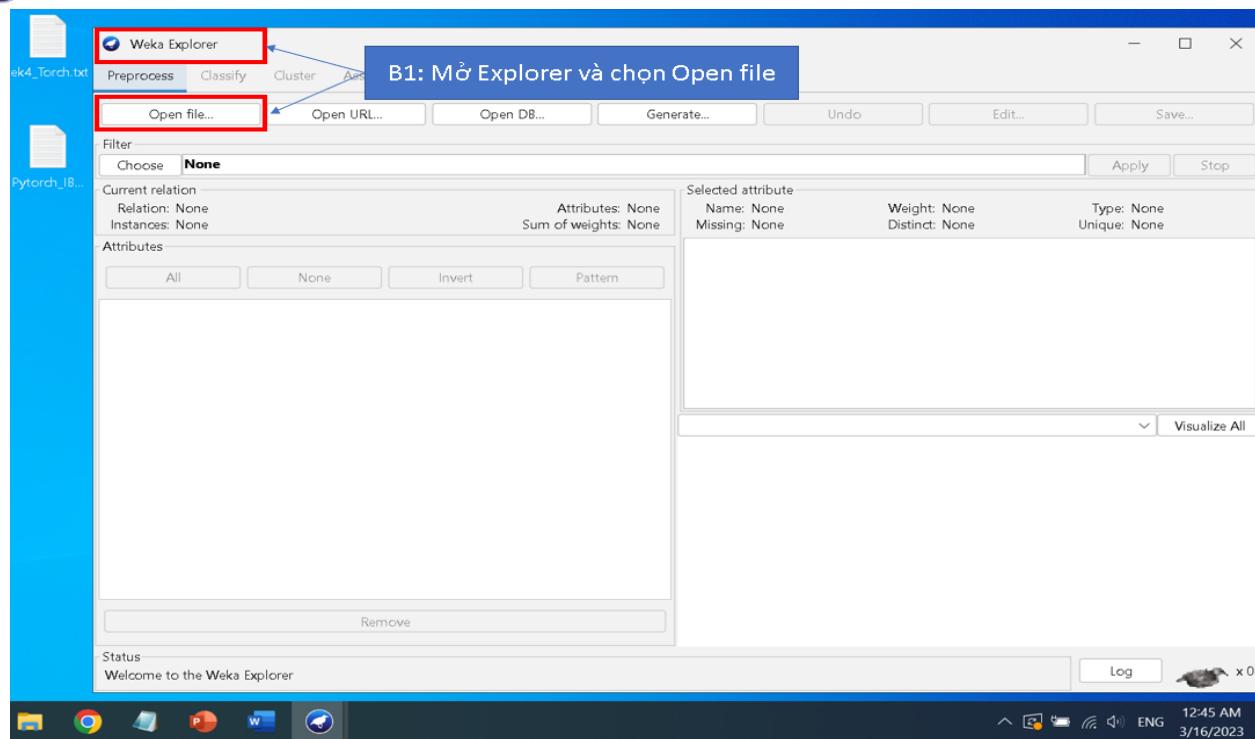
## 2. LÀM QUEN VỚI WEKA

Trong phần này, nhóm được yêu cầu khám phá các bộ dữ liệu khác nhau để làm quen với Weka, việc đầu tiên là đọc các tập dữ liệu có sẵn được cung cấp bởi Weka, cách đọc như sau:

**Đọc file có tên là ‘breast cancer.arf’.**

Bước 1: Mở cửa sổ Explorer, sau đó chọn Open file.

Bước 2: Chọn vào folder data trong đường dẫn cài đặt Weka ban đầu vào máy, sau đó chọn file dataset mong muốn và nhấn Open.



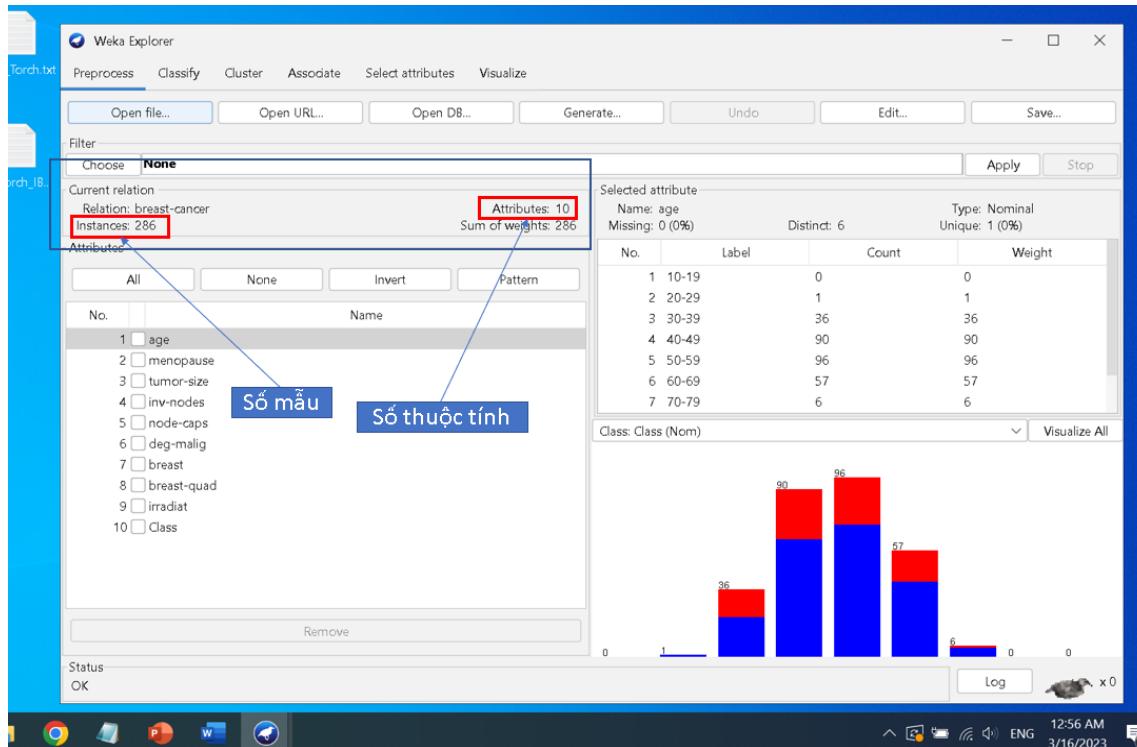
## 2.1. Khám phá tập dữ liệu ‘Breast Cancer’

### 2.1.1. Tập dữ liệu có bao nhiêu mẫu?

- Tập dữ liệu có 286 mẫu (minh họa hình 1.1).

### 2.1.2. Tập dữ liệu có bao nhiêu thuộc tính?

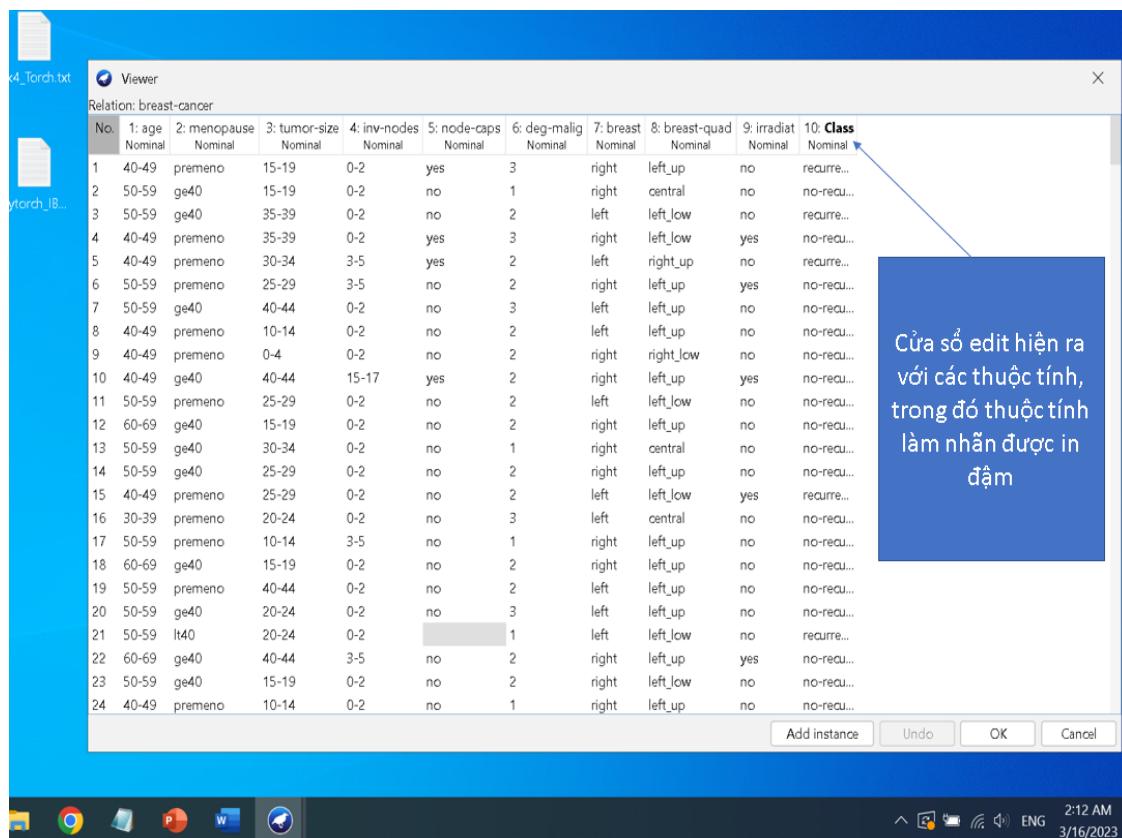
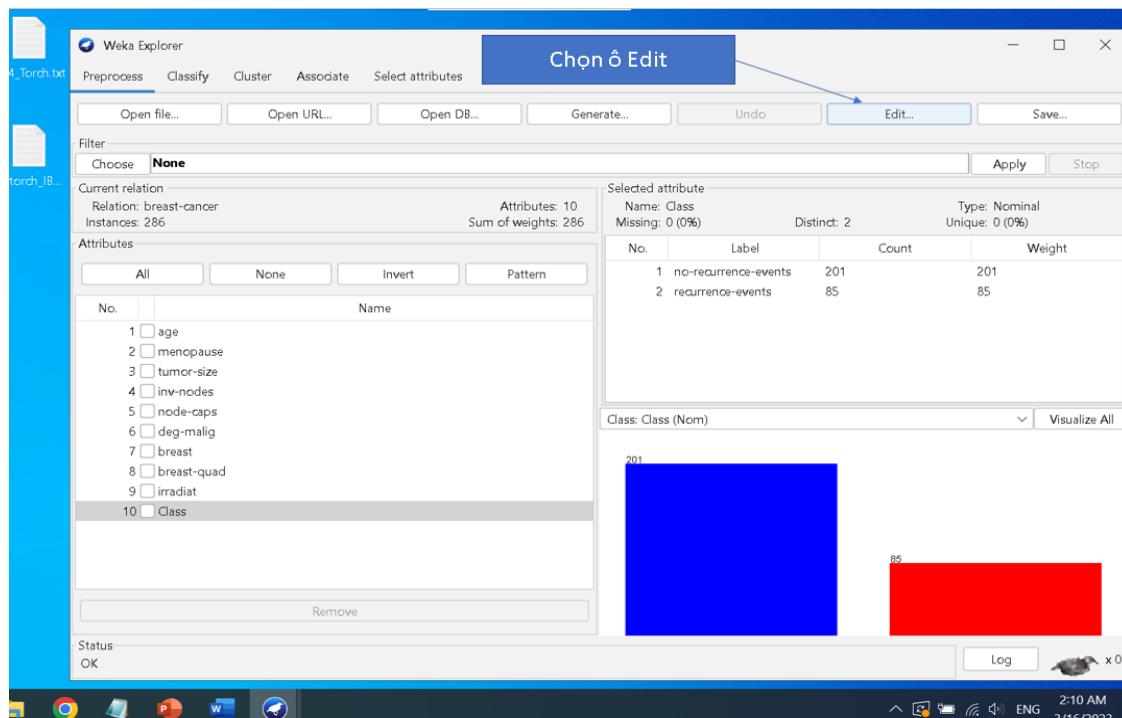
- Tập dữ liệu có 10 thuộc tính (minh họa hình 1.1).



Hình 1.1. Minh họa trả lời phần 3.2.1.1 và 3.2.1.2

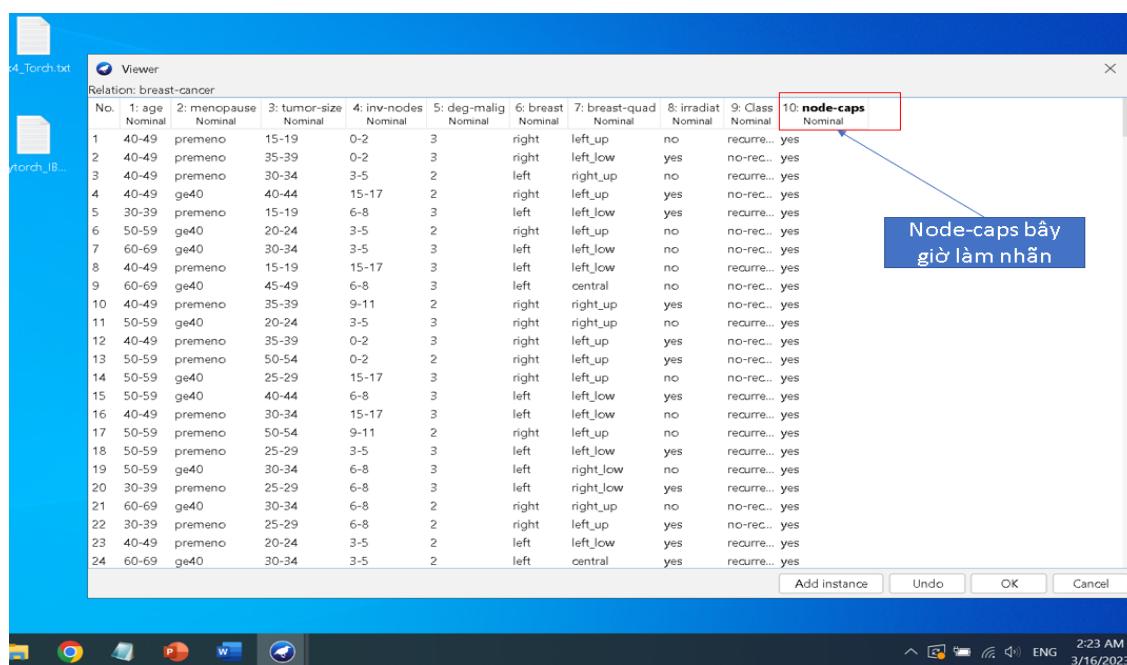
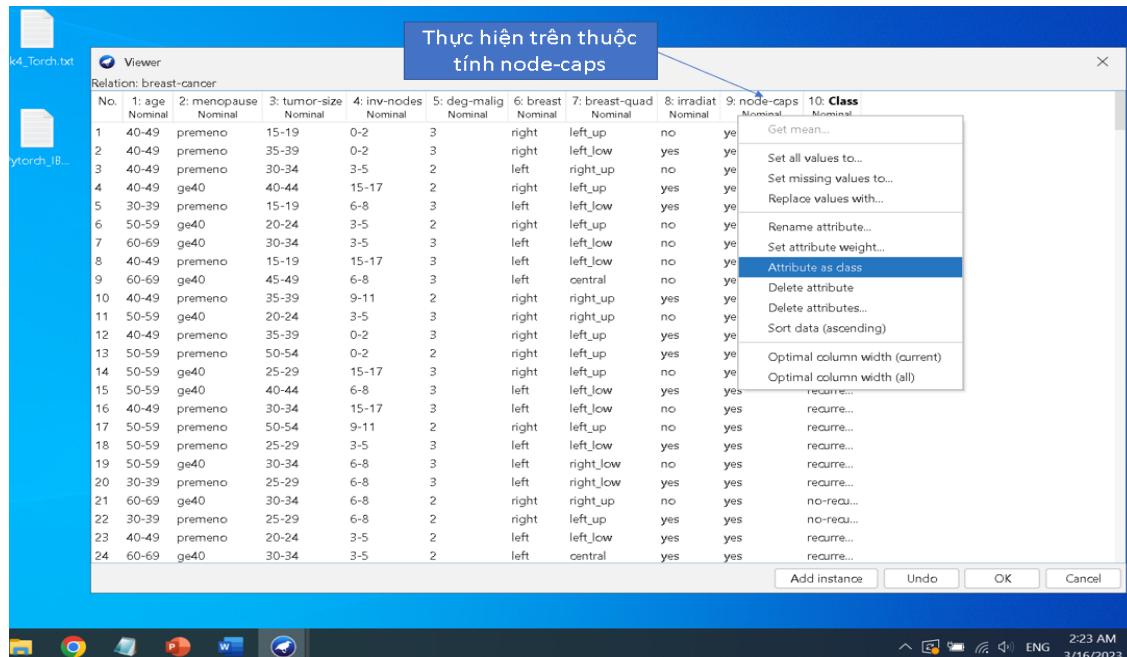
### 2.1.3. Thuộc tính nào được dùng để làm nhãn? Có thay đổi được không? Bằng cách nào?

- Thuộc tính ‘**Class**’ được dùng để làm nhãn. Xem bằng cách chọn vào ô ‘**Edit**’, thuộc tính nào làm nhãn sẽ được in đậm.



- Thuộc tính làm nhãn có thể thay đổi được. Bằng cách:

Vẫn trong cửa sổ ‘Viewer’ vừa được trình bày cách mở ở trên, chọn thuộc tính mới muốn làm nhãn và click chuột phải vào lựa chọn ‘**attribute as class**’ để đặt thuộc tính này làm nhãn mới.



- Sau đó nhấn ‘Ok’ để lưu thay đổi.



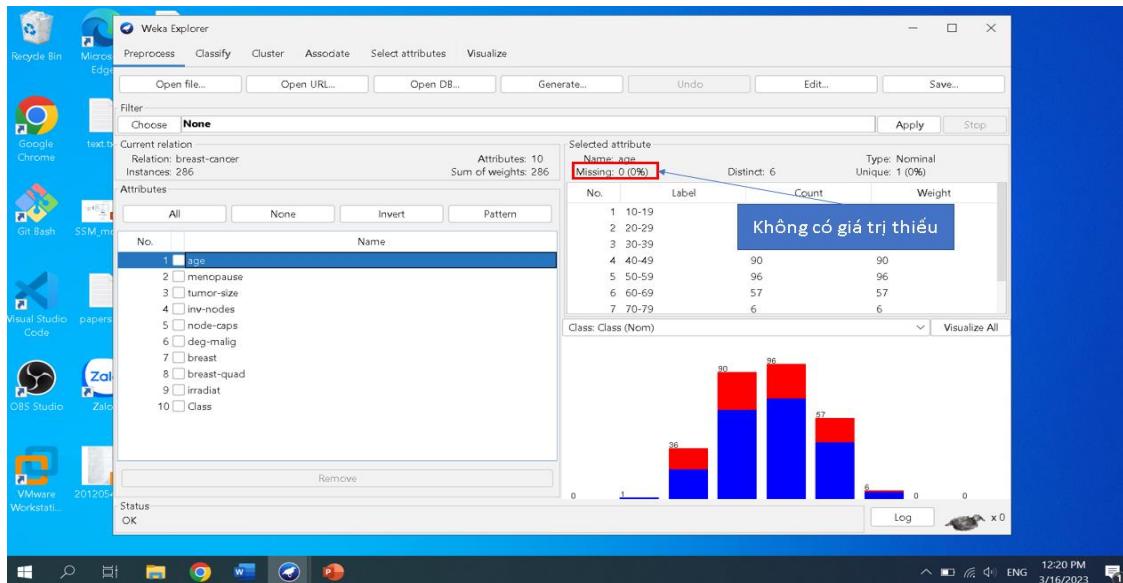
#### 2.1.4. Ý nghĩa của mỗi thuộc tính

Tên (loại) thuộc tính	Ý nghĩa
age (nominal)	Khoảng độ tuổi của bệnh nhân mắc ung thư vú.
Menopause (nominal)	Thể hiện tình trạng mãn kinh của bệnh nhân: ‘premeno’: Tiền mãn kinh. ‘lt40’: Tuổi mãn kinh nhỏ hơn 40 tuổi. ‘ge40’: Tuổi mãn kinh lớn hơn hoặc bằng 40 tuổi.
tumor-size (nominal)	Thể hiện kích thước của khối u ở vú (milimet).
inv-nodes (nominal)	Thể hiện số lượng các nút bạch huyết trong cơ thể đã bị ảnh hưởng bởi khối ung thư.
node-caps (nominal)	Cho biết các u nang có bám vào mạch máu không: yes(có), no(không).
deg-malig (nominal)	Độ xấu của các tế bào u nang: 1: thấp. 2: vừa phải. 3: cao.
breast (nominal)	Chỉ ra vị trí của u nang trong ngực khi được chia làm 2 phần: left(trái), right(phải).
breast-quad (nominal)	Chỉ ra vị trí của u nang trong ngực khi được chia làm 4 phần (có 5 giá trị): left_up: trái trên. left_down: trái dưới. right_up: phải trên. right_down: phải dưới central: trung tâm
irradiat (nominal)	Bệnh nhân có được xạ trị hay chưa: yes(có), no(không).
Class (nominal)	Phân loại các trường hợp bệnh nhân ung thư vú thành 2 lớp: no-recurrence-events: không tái phát. recurrence-events: tái phát.

#### 2.1.5. Tìm hiểu tình trạng thiếu giá trị (missing values) trong mỗi thuộc tính và mô tả các cách giải quyết tổng quát cho vấn đề thiếu giá trị.

Để xem mô tả về tình trạng thiếu giá trị của mỗi thuộc tính: Chọn vào thuộc tính ở bảng Attributes, sau đó quan sát thông tin ở bảng Selected attribute.

- Thuộc tính ‘age’ không xảy ra vấn đề thiếu giá trị:



(Để phần trình bày được rõ ràng thì đối với các thuộc tính sau chỉ chụp ảnh phần bảng Selected attribute).

- Một loạt các thuộc tính gồm: '*menopause*', '*tumor-size*', '*inv-nodes*', '*deg-malig*', '*breast*', '*irradiat*', '*Class*' không xảy ra vấn đề thiếu giá trị:

#### Selected attribute

Name: menopause	Type: Nominal
Missing: 0 (0%)	Distinct: 3

#### Selected attribute

Name: tumor-size	Type: Nominal
Missing: 0 (0%)	Distinct: 11

#### Selected attribute

Name: inv-nodes	Type: Nominal
Missing: 0 (0%)	Distinct: 7

#### Selected attribute

Name: deg-malig	Type: Nominal
Missing: 0 (0%)	Distinct: 3

#### Selected attribute

Name: breast	Type: Nominal
Missing: 0 (0%)	Distinct: 2

Selected attribute

Name: irradiat

Missing: 0 (0%)

Type: Nominal

Distinct: 2 Unique: 0 (0%)

Selected attribute

Name: Class

Missing: 0 (0%)

Type: Nominal

Distinct: 2 Unique: 0 (0%)

- Thuộc tính ‘node-caps’ thiếu giá trị nhiều nhất với 8 giá trị thiếu (chiếm xấp xỉ 3%):

Selected attribute

Name: node-caps

Missing: 8 (3%)

Type: Nominal

Distinct: 2 Unique: 0 (0%)

- Thuộc tính ‘breast-quad’ chỉ chứa 1 giá trị thiếu (tỉ lệ coi như gần bằng 0%):

Selected attribute

Name: breast-quad

Missing: 1 (0%)

Type: Nominal

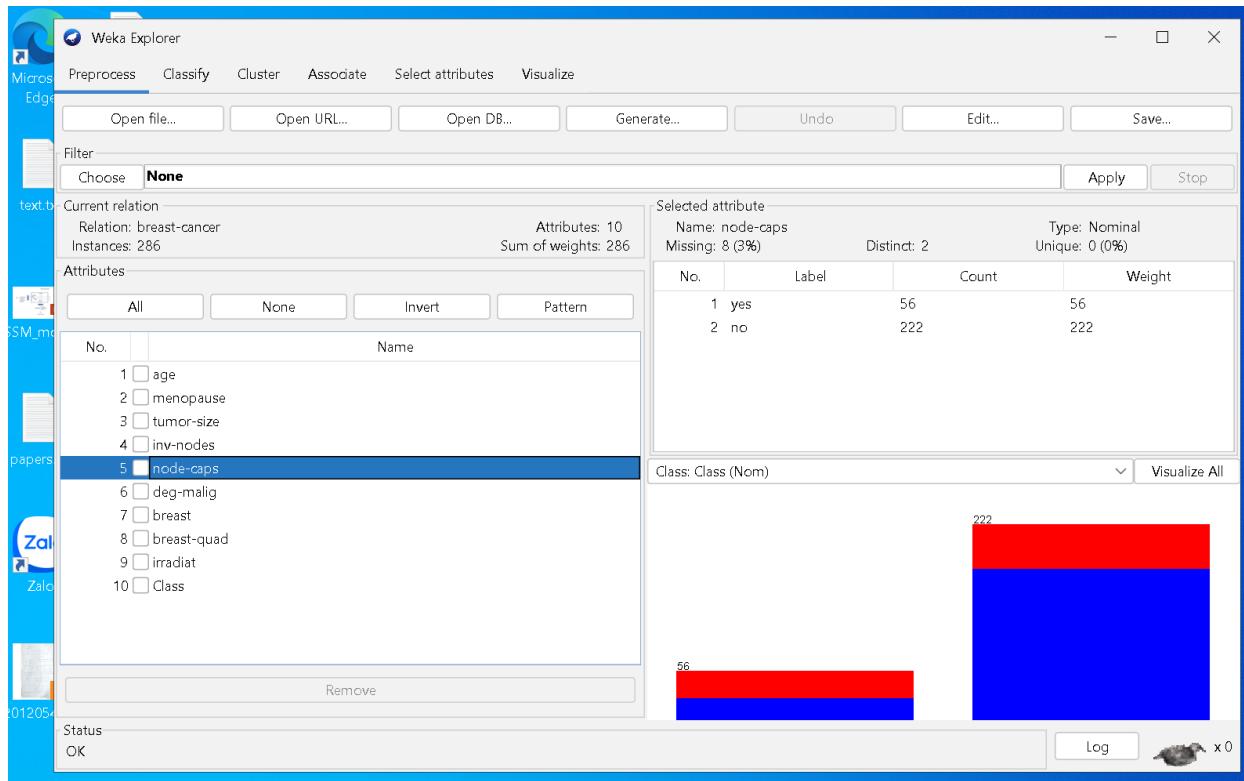
Distinct: 5 Unique: 0 (0%)

- Các cách giải quyết tổng quát cho vấn đề thiếu giá trị (missing values):

Kiểu dữ liệu	Cách xử lý
Numeric (dữ liệu số)	<ul style="list-style-type: none"><li>- Mean: thay thế các giá trị bị thiếu bằng trung bình của các giá trị không thiếu.</li><li>- Median: thay thế các giá trị bị thiếu bằng giá trị ở giữa của tập dữ liệu sau khi sắp xếp.</li><li>- Mode: thay thế các giá trị bị thiếu bằng giá trị xuất hiện nhiều nhất.</li><li>- Nếu số lượng mẫu chứa giá trị thiếu chiếm một tỉ lệ cực kì nhỏ, có thể xóa mẫu chứa giá trị thiếu ra khỏi dữ liệu.</li></ul>
Nominal (dữ liệu định danh)	<ul style="list-style-type: none"><li>- Mode: thay thế các giá trị bị thiếu bằng giá trị xuất hiện nhiều nhất.</li><li>- Nếu số lượng mẫu chứa giá trị thiếu chiếm một tỉ lệ cực kì nhỏ, có thể xóa mẫu chứa giá trị thiếu ra khỏi dữ liệu.</li></ul>

## 2.1.6. Đề xuất giải pháp cho vấn đề thiếu giá trị trong thuộc tính cụ thể

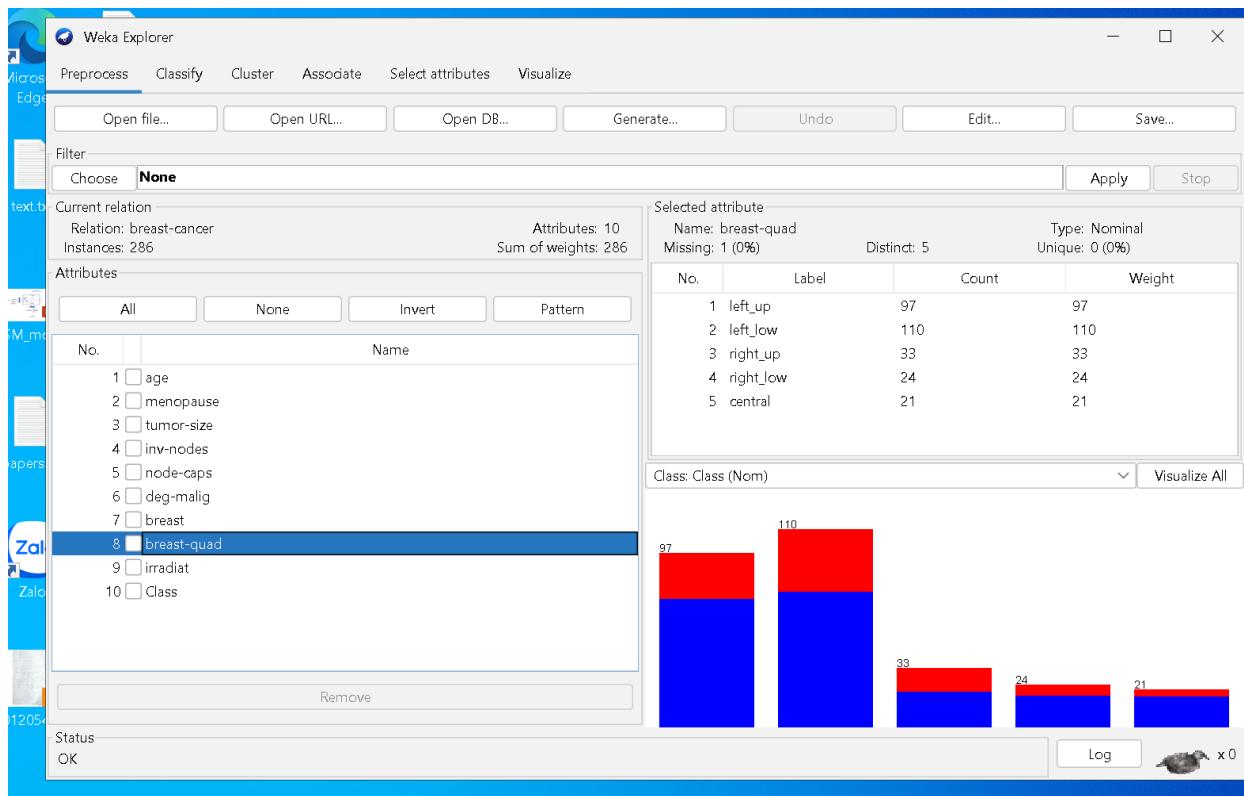
- Đối với thuộc tính ‘node-caps’:



+ Thuộc là kiểu dữ liệu nominal và chỉ chứa 2 loại giá trị là yes và no, ta có thể thấy được số mẫu mang giá trị ‘no’ nhiều hơn rất nhiều so với số mẫu mang giá trị ‘yes’.

+ Do đó có thể thay thế các giá trị thiếu bằng giá trị mode của thuộc tính (thay bằng giá trị ‘no’).

- Đối với thuộc tính ‘breast-quad’:



+ Tương tự như thuộc tính ‘node-caps’, ‘breast-quad’ cũng là dữ liệu kiểu nominal do đó có thể thay thế giá trị thiếu bằng giá trị mode của thuộc tính (giá trị left\_low).

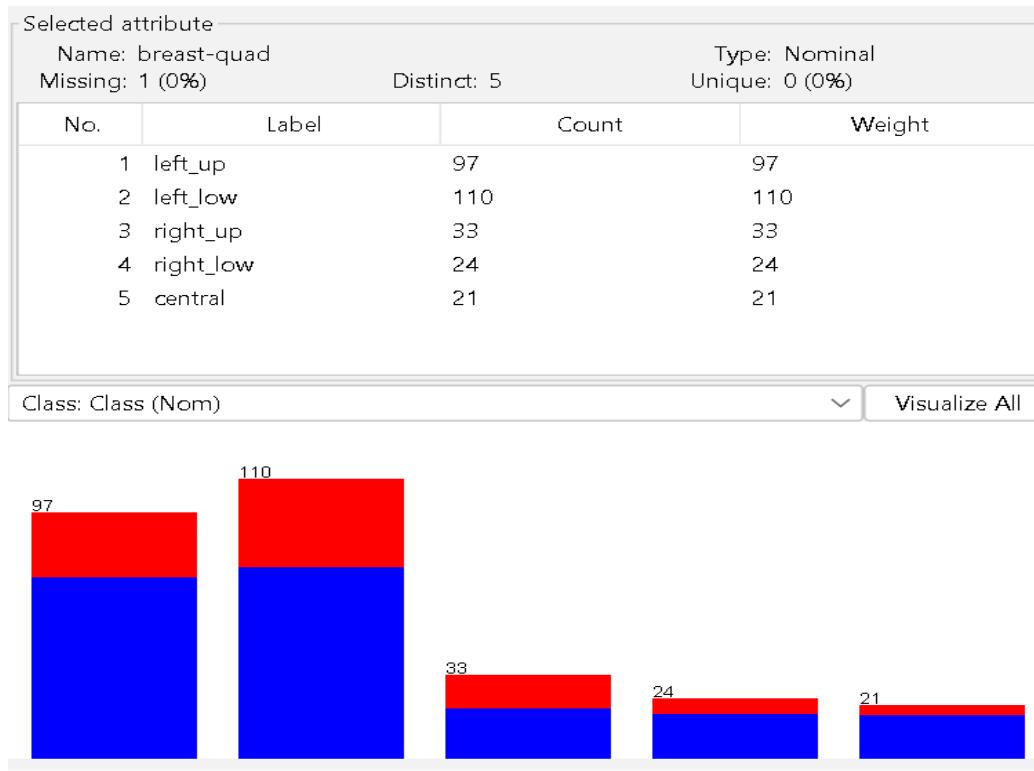
+ Ngoài ra, đối với thuộc tính ‘breast-quad’ chỉ tồn tại 1 giá trị thiếu, là cực kỳ nhỏ so với tổng số mẫu của bộ dữ liệu, do đó có thể chọn cách xóa mẫu chứa giá trị thiếu ra khỏi dữ liệu.

### **2.1.7. Giải thích và thực hiện một số cài đặt cho biểu đồ trong Weka Explorer**

- Biểu đồ được hiển thị khi chọn vào một thuộc tính nhất định là một biểu đồ cột **dùng để thể hiện sự phân bố các giá trị của thuộc tính và được mã hóa theo màu**. Số màu được chọn để mã hóa sẽ phụ thuộc vào số lớp trong thuộc tính được chọn làm nhãn

**- Đặt tên đồ thị: Đồ thị phân bố giá trị của thuộc tính theo nhãn.**

- Ví dụ với thuộc tính ‘breast-quad’:



- + Thuộc tính làm nhãn đang là thuộc tính ‘Class’ chỉ có 2 lớp ‘no-recurrence-events’ và ‘recurrence-events’, do đó số màu để mã hóa theo màu là 2.
- + Màu xanh và màu đỏ lần lượt đại diện cho lớp no\_recurrence\_events và recurrence\_events.
- + Thuộc tính ‘breast-quad’ có 5 loại giá trị với số lượng từng loại như sau: left\_up có 97 mẫu, left\_low có 110 mẫu, right\_up có 33 mẫu, right\_low có 24 mẫu, central có 21 mẫu.
- + Và trong mỗi loại giá trị thì số lượng mẫu thuộc về lớp ‘no-recurrence-events’ (màu xanh) luôn nhiều hơn số lượng mẫu thuộc về lớp ‘recurrence-events’ (màu đỏ).

## 2.2. Khám phá tập dữ liệu Weather (weather.numeric.arff)

### 2.2.1. Tổng quan về bộ dữ liệu

- Tập dữ liệu có 5 thuộc tính, 14 mẫu.



- Những thuộc tính có loại là categorical (nominal) là: outlook, windy, play.

Selected attribute		
Name: outlook		Type: Nominal
Missing: 0 (0%)	Distinct: 3	Unique: 0 (0%)

Selected attribute		
Name: windy		Type: Nominal
Missing: 0 (0%)	Distinct: 2	Unique: 0 (0%)

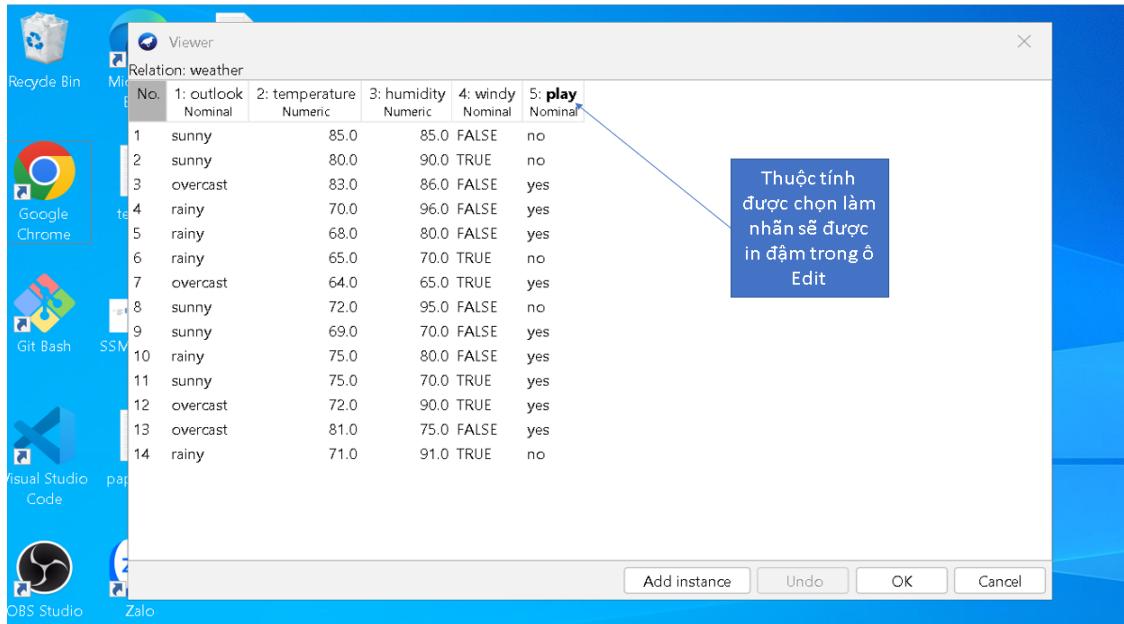
Selected attribute		
Name: play		Type: Nominal
Missing: 0 (0%)	Distinct: 2	Unique: 0 (0%)

- Những thuộc tính có loại là numerical là: temperature, humidity.

Selected attribute		
Name: temperature		Type: Numeric
Missing: 0 (0%)	Distinct: 12	Unique: 10 (71%)

Selected attribute		
Name: humidity		Type: Numeric
Missing: 0 (0%)	Distinct: 10	Unique: 7 (50%)

- Thuộc tính được chọn làm nhãn là thuộc tính ‘play’.



**2.2.2. Liệt kê 5 số liệu thống kê (five-number summary) của 2 thuộc tính ‘temperature’ và ‘humidity’. Weka có cung cấp những số liệu này không?**

Thuộc tính	Five-number summary
temperature	min: 64 max: 85 Q1: 69 median: 72 Q3: 80
humidity	min: 65 max: 96 Q1: 70 median: 82.5 Q3: 90

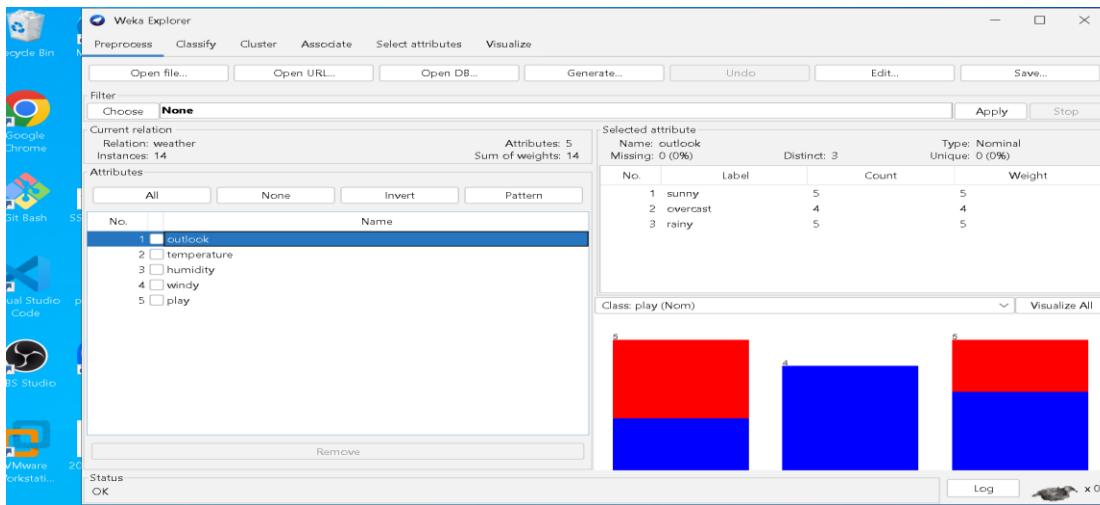
- Trong Weka, đối với thuộc tính dạng numerical thì không cung cấp thông tin thống kê về 5 con số này, chỉ có 2 giá trị là min và max.

**2.2.3. Giải thích ý nghĩa tất cả biểu đồ trong Weka Explorer. Đặt tên cho biểu đồ và mô tả chú thích**

- Tất cả các biểu đồ đều **dùng để thể hiện sự phân bố các giá trị của thuộc tính và được mã hóa theo màu**. Số màu được chọn để mã hóa sẽ phụ thuộc vào số lớp trong thuộc tính được chọn làm nhãn.

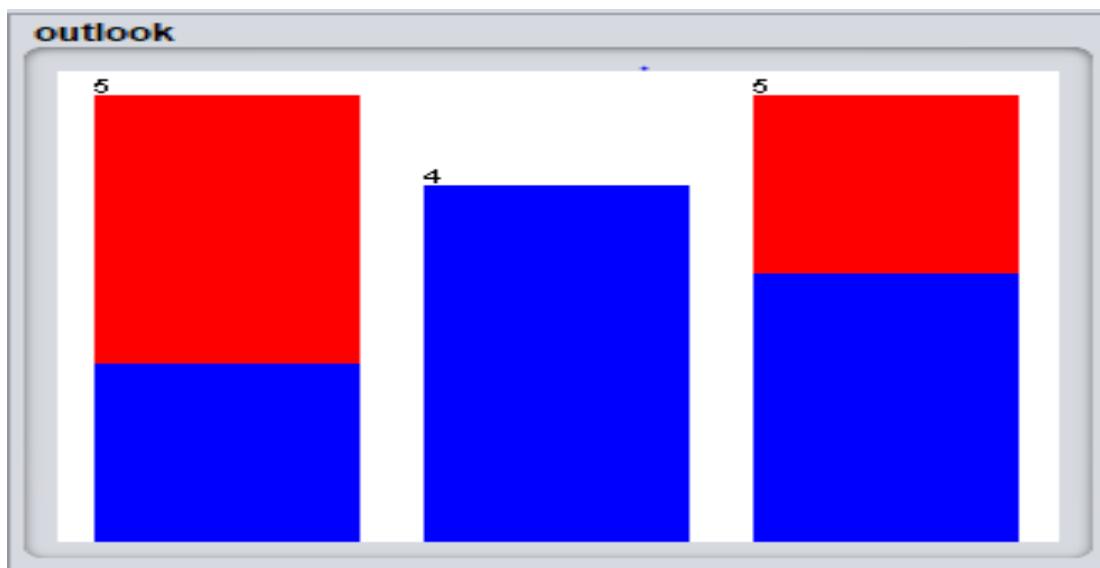
- Trong tập dữ liệu này thì thuộc tính ‘play’ được chọn làm nhãn với 2 lớp ‘yes’, ‘no’.
- Xét nhóm thuộc tính dạng nominal/categorical:

- **Thuộc tính ‘outlook’:**



Có 3 loại giá trị với số lượng mẫu ở mỗi loại giá trị như sau: ‘sunny’ 5 mẫu, ‘overcast’ 4 mẫu, ‘rainy’ 5 mẫu. Trong mỗi loại giá trị thì các mẫu sẽ được chia thành 2 lớp lần lượt là màu đỏ cho lớp ‘no’ (ý nghĩa là không đi chơi) và màu xanh cho lớp ‘yes’ (ý nghĩa là đi chơi)

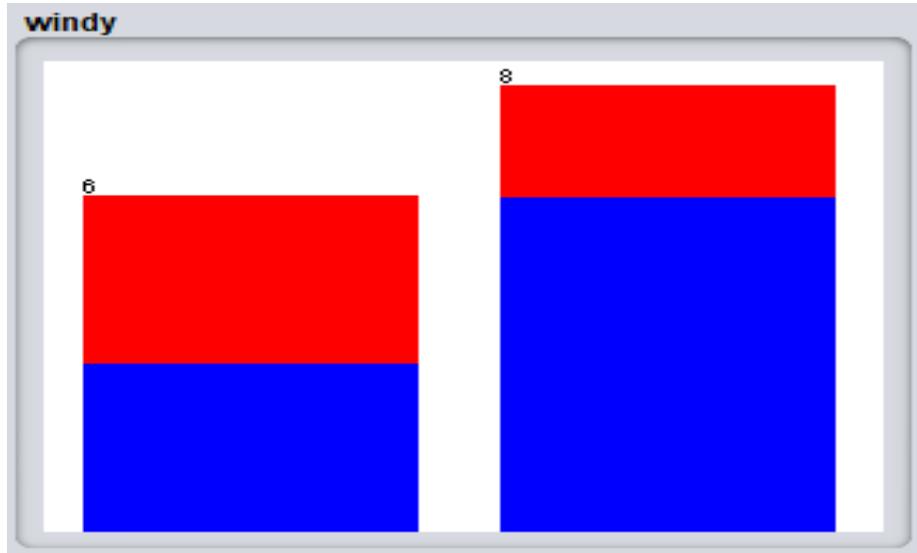
Đặt tên cho đồ thị:



*Đồ thị phân bố giá trị của thuộc tính outlook theo nhãn.*

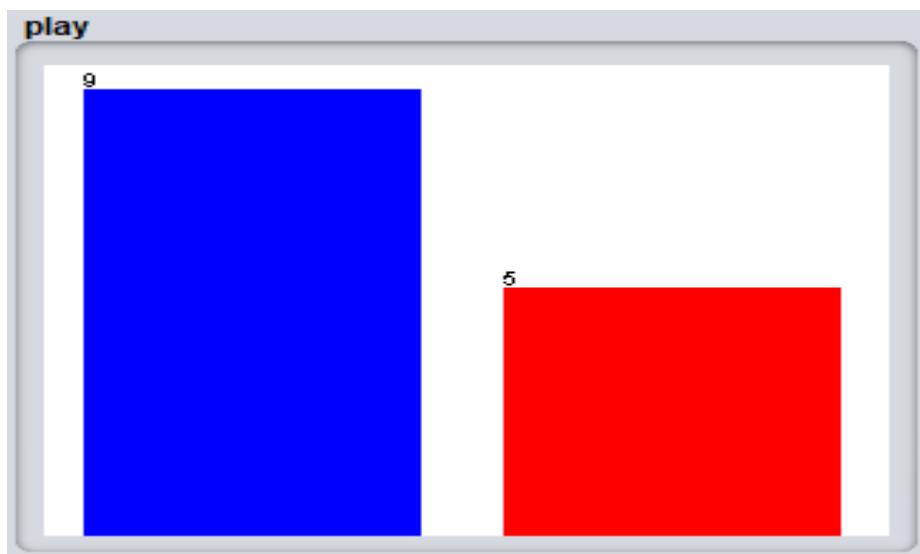
**Cách giải thích tương tự cho các thuộc tính khác**, phần sau sẽ chụp ảnh các đồ thị của các thuộc tính còn lại và đặt tên cho chúng.

- **Thuộc tính ‘windy’:**



Đồ thị phân bố giá trị của thuộc tính windy theo nhãn.

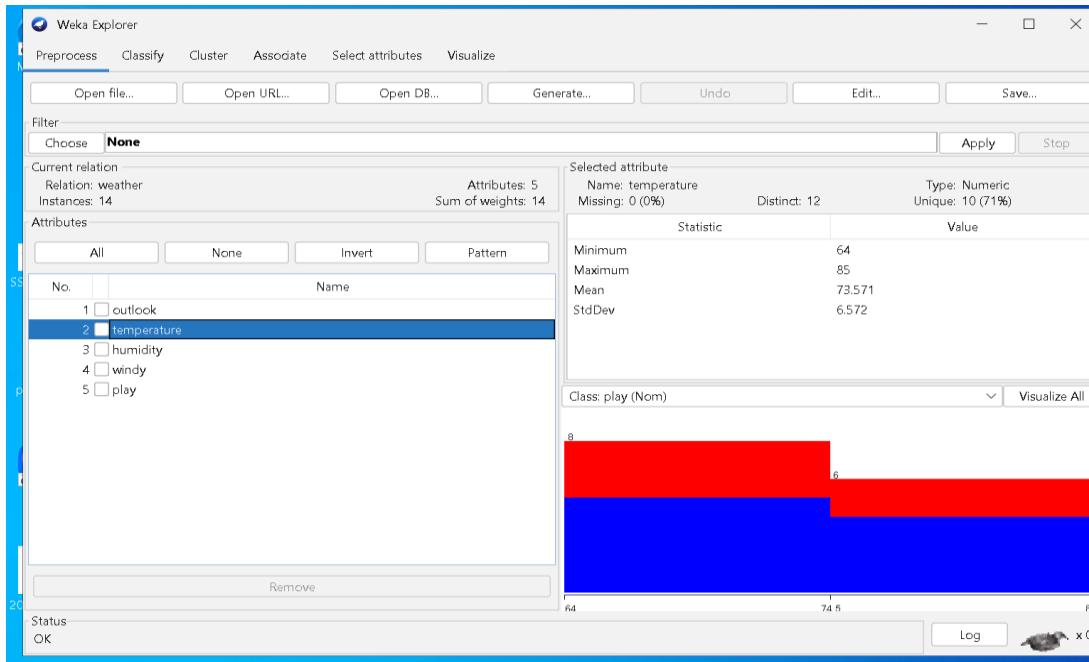
- **Thuộc tính ‘play’:**



Đồ thị phân bố giá trị của thuộc tính play theo nhãn.

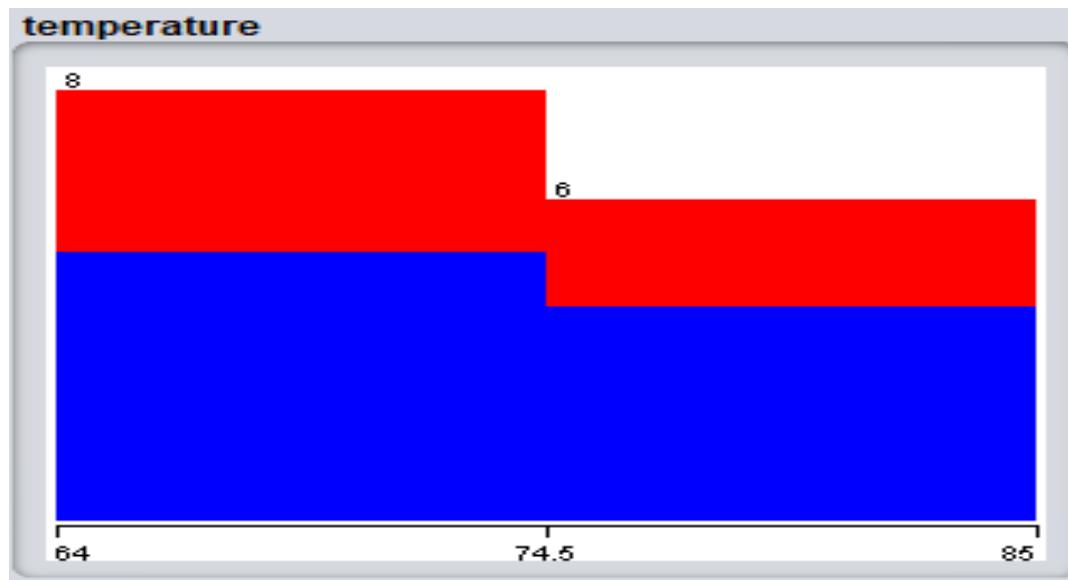
- Xét nhóm thuộc tính dạng numerical:

- **Thuộc tính ‘temperature’:**



Do thuộc tính dạng numerical nên đồ thị sẽ là một khối liên tục, vẫn thể hiện sự phân bố giá trị của thuộc tính và các giá trị cũng được phân theo 2 lớp là ‘yes’ (màu xanh) và ‘no’ (màu đỏ).

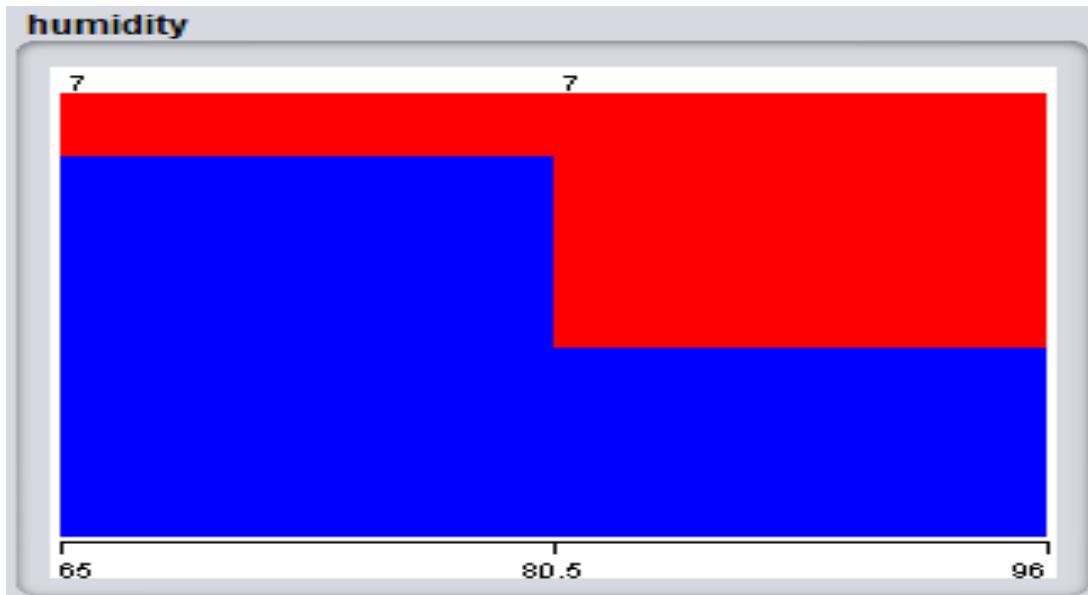
Đặt tên cho đồ thị:



*Đồ thị phân bố giá trị của thuộc tính temperature theo nhãn.*

- **Thuộc tính ‘humidity’:**

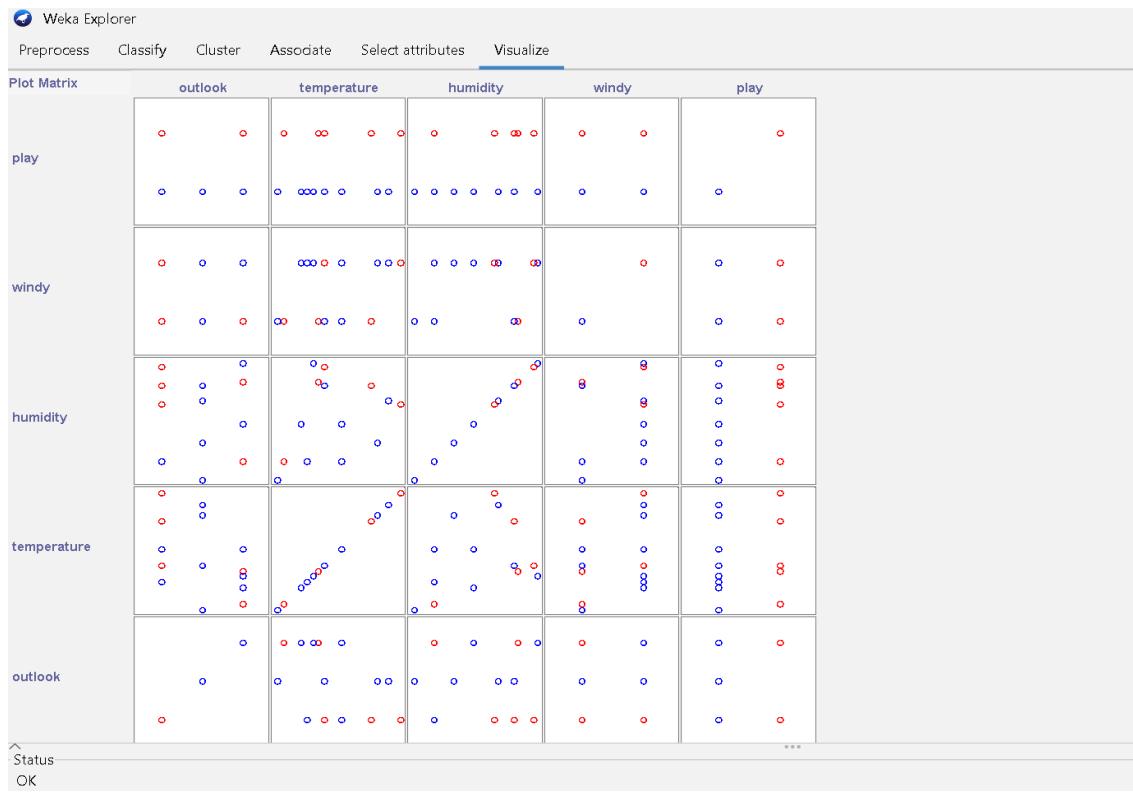
Tiếp theo về thuộc tính ‘humidity’ cũng sẽ được hiểu theo cách tương tự, đặt tên cho đồ thị của thuộc tính ‘humidity’:



Đồ thị phân bố giá trị của thuộc tính humidity theo nhän.

#### 2.2.4. Phân tích thẻ “Visualize”

- Các đồ thị này được gọi là biểu đồ phân tán (scatter plot).
- Không có cặp thuộc tính nào tương quan với nhau, do các điểm trong scatter plot đều được phân tán rộng, một cách rời rạc, không thể hiện xu hướng chung hay mối quan hệ tuyến tính nào.



## 2.3. Khám phá tập dữ liệu Germany (credit-g.arff)

### 2.3.1. Tổng quan về bộ dữ liệu

- Dựa vào đường dẫn khi cài đặt phần mềm, ta tìm thư mục dataset, mở tập dữ liệu credit-g.arff bằng công cụ Notepad, ta nhận được 278 dòng chú thích mô tả (tất cả các dòng bắt đầu bằng kí tự %) của tập dữ liệu ở phần đầu như sau:

```
credit-g.arff      +  
  
File Edit View  
  
% 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.  
%  
%  
% 6. Number of Attributes german: 20 (7 numerical, 13 categorical)  
%     Number of Attributes german.numer: 24 (24 numerical)
```

Sơ lược về toàn bộ phần chú thích mô tả:

- **Title:** Tiêu đề/ tên của tập dữ liệu (German Credit)

```
% Description of the German credit dataset.  
%  
% 1. Title: German Credit data
```

- **Source Information:** Thông tin về nguồn gốc của tập dữ liệu (Tên tác giả, nơi công tác, thời điểm thực hiện)

```
% 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  
%
```

- **Number of instances:** số lượng mẫu và một vài chú thích cho tập dữ liệu có nội dung tổng quan như sau:

Có hai bộ dữ liệu được cung cấp. Tập dữ liệu gốc được cung cấp bởi Giáo sư Hofmann, chứa các thuộc tính categorical/symbolic và nằm trong tệp "german.data".

Đại học Strathclyde đã tạo thêm tệp "german.data-numeric" dành cho các thuật toán cần thuộc tính numeric. Tập tin này đã được chỉnh sửa và thêm một số biến chỉ báo vào để làm cho nó phù hợp với các thuật toán không thẻ đối phó với các biến categorical. Một số các thuộc tính được sắp xếp theo thứ tự phân loại (chẳng hạn như thuộc tính 17) được mã hóa thành số nguyên.

```
% 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.
```

- **Number of Attributes german/ Number of Attributes german.numer:** Số lượng thuộc tính

```
% 6. Number of Attributes german: 20 (7 numerical, 13 categorical)
%     Number of Attributes german.numer: 24 (24 numerical)
```

- **Attribute description for german:** Mô tả sơ lược các thuộc tính cho tệp german (Hình sau là ví dụ cho mô tả 3 thuộc tính đầu trong tệp)

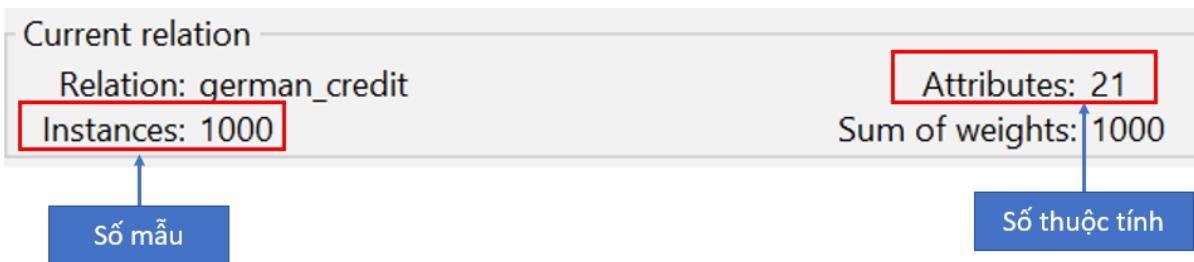
```
% 7. Attribute description for german
%
% Attribute 1: (qualitative)
%             Status of existing checking account
%             A11 : ... < 0 DM
%             A12 : 0 <= ... < 200 DM
%             A13 : ... >= 200 DM /
%                     salary assignments for at least 1 year
%             A14 : no checking account
%
% Attribute 2: (numerical)
%             Duration in month
%
% Attribute 3: (qualitative)
%             Credit history
%             A30 : no credits taken/
%                     all credits paid back duly
%             A31 : all credits at this bank paid back duly
%             A32 : existing credits paid back duly till now
%             A33 : delay in paying off in the past
%             A34 : critical account/
%                     other credits existing (not at this bank)
```

- **Cost matrix:**

```
% 8. Cost Matrix
%
% This dataset requires use of a cost matrix (see below)
%
%
%      1      2
% -----
% 1  0      1
% -----
% 2  5      0
%
% (1 = Good,  2 = Bad)
%
% the rows represent the actual classification and the columns
% the predicted classification.
%
% It is worse to class a customer as good when they are bad (5),
% than it is to class a customer as bad when they are good (1).
```

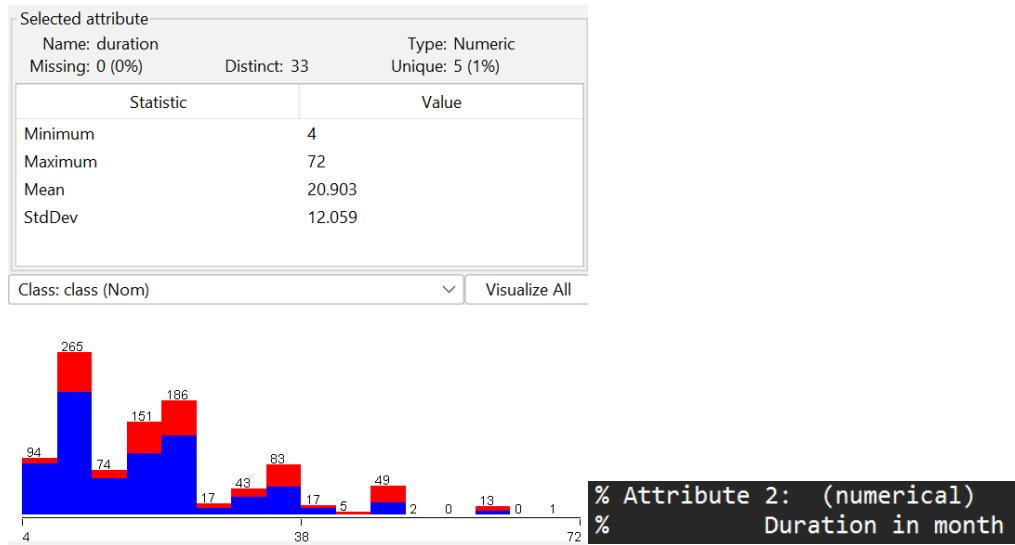
- Ma trận chi phí của bộ dữ liệu này xác định chi phí cho mỗi lớp được phân loại sai, bộ dữ liệu này có hai lớp là "tốt" và "xấu".
- Ma trận chi phí là một ma trận  $2 \times 2$  trong đó các hàng đại diện cho lớp thực tế và các cột đại diện cho lớp được dự đoán. Các phần tử đường chéo chính của ma trận đại diện cho chi phí của việc phân loại đúng lớp, các phần tử đường chéo phụ đại diện cho chi phí phân loại sai lớp.
- Ma trận chi phí này cho thấy chi phí phân loại sai lớp xấu (false negative) là 5 lần lớn hơn so với chi phí phân loại sai lớp tốt (false positive).

**Ta thấy: Tập dữ liệu có 21 thuộc tính, 1000 mẫu**



- Mô tả 5 thuộc tính bắt kì: duration (thuộc tính 2), credit amount (thuộc tính 5), personal status (thuộc tính 9), housing (thuộc tính 15), foreign worker (thuộc tính 20)
- Xét nhóm thuộc tính dạng numerical:

- **Thuộc tính ‘duration’ (Thời gian sử dụng tài khoản)**

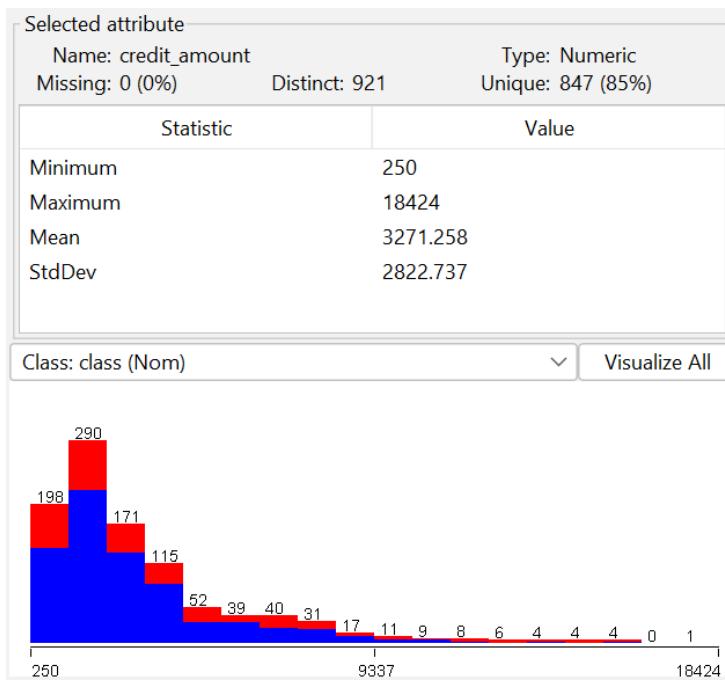


Thuộc tính có giá trị Minimum=4, Maximum=72, Mean (giá trị trung bình)=20.903, StdDev (độ lệch chuẩn)=12.059

Đồ thị là một khối liên tục, vẫn thể hiện sự phân bố giá trị của thuộc tính và các giá trị cũng được phân theo 2 lớp là '**good**' (màu xanh) và '**bad**' (màu đỏ).

- **Thuộc tính ‘credit amount’ (Số dư thẻ)**

% Attribute 5: (numerical)  
 % Credit amount



Thuộc tính có giá trị Minimum=250, Maximum=18424, Mean (giá trị trung bình)=3271.258, StdDev (độ lệch chuẩn)=2822.737

Đồ thị là một khối liên tục, vẫn thể hiện sự phân bố giá trị của thuộc tính và các giá trị cũng được phân theo 2 lớp là '**good**' (màu xanh) và '**bad**' (màu đỏ).

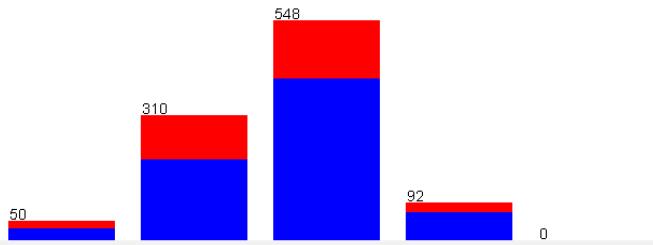
- Xét nhóm thuộc tính dạng nominal/categorical:

- **Thuộc tính 'personal status':**

```
% Attribute 9: (qualitative)
%          Personal status and sex
%          A91 : male   : divorced/separated
%          A92 : female : divorced/separated/married
%          A93 : male   : single
%          A94 : male   : married/widowed
%          A95 : female : single
```

Selected attribute			
Name: personal_status		Type: Nominal	
Missing: 0 (0%)		Distinct: 4	Unique: 0 (0%)
No.	Label	Count	Weight
1	male div/sep	50	50
2	female div/dep/mar	310	310
3	male single	548	548
4	male mar/wid	92	92
5	female single	0	0

Class: class (Nom) Visualize All



Thuộc tính này có 5 loại giá trị với số lượng mẫu tương ứng như: **male div/sep** (đàn ông đã li dị hoặc li thân) có 50 mẫu, **female div/dep/mar** (phụ nữ đã li dị hoặc li thân hoặc đã kết hôn) có 310 mẫu, **male single** (đàn ông độc thân) có 548 mẫu, **male mar/wid** (đàn ông đã kết hôn hoặc goá vợ) có 92 mẫu, **female single** (phụ nữ độc thân) có 0 mẫu.

Trong mỗi loại giá trị thì số lượng mẫu thuộc về lớp '**good**' (màu xanh) luôn nhiều hơn số lượng mẫu thuộc về lớp '**bad**' (màu đỏ).

- **Thuộc tính ‘housing’:**

```
% Attribute 15: (qualitative)
%           Housing
%           A151 : rent
%           A152 : own
%           A153 : for free
```

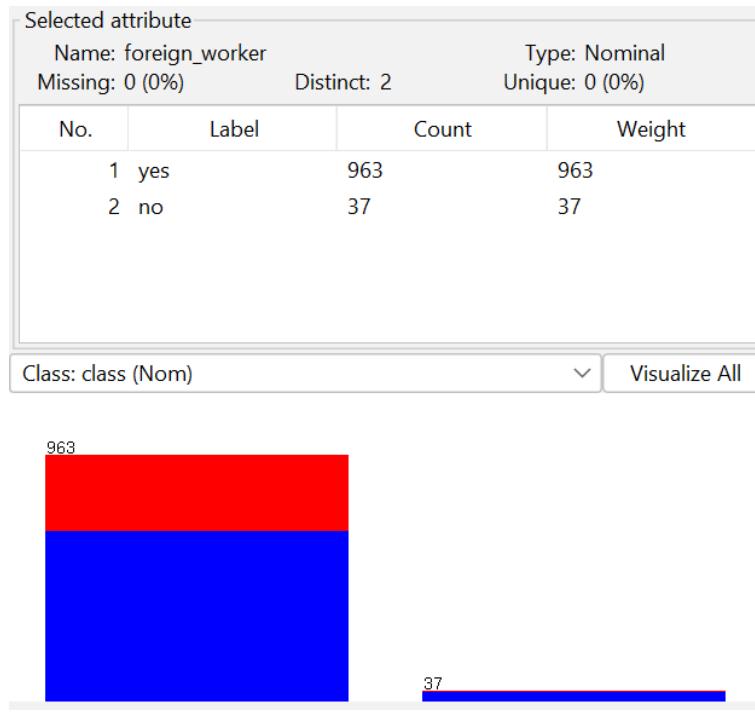


Thuộc tính này có 3 loại giá trị tương ứng với số lượng mẫu như sau: rent (nhà thuê) có 179 mẫu, own (nhà riêng) có 713 mẫu, for free (nhà trợ cấp) có 108 mẫu.

Trong mỗi loại giá trị thì số lượng mẫu thuộc về lớp ‘**good**’ (màu xanh) luôn nhiều hơn số lượng mẫu thuộc về lớp ‘**bad**’ (màu đỏ).

- **Thuộc tính ‘foreign worker’**

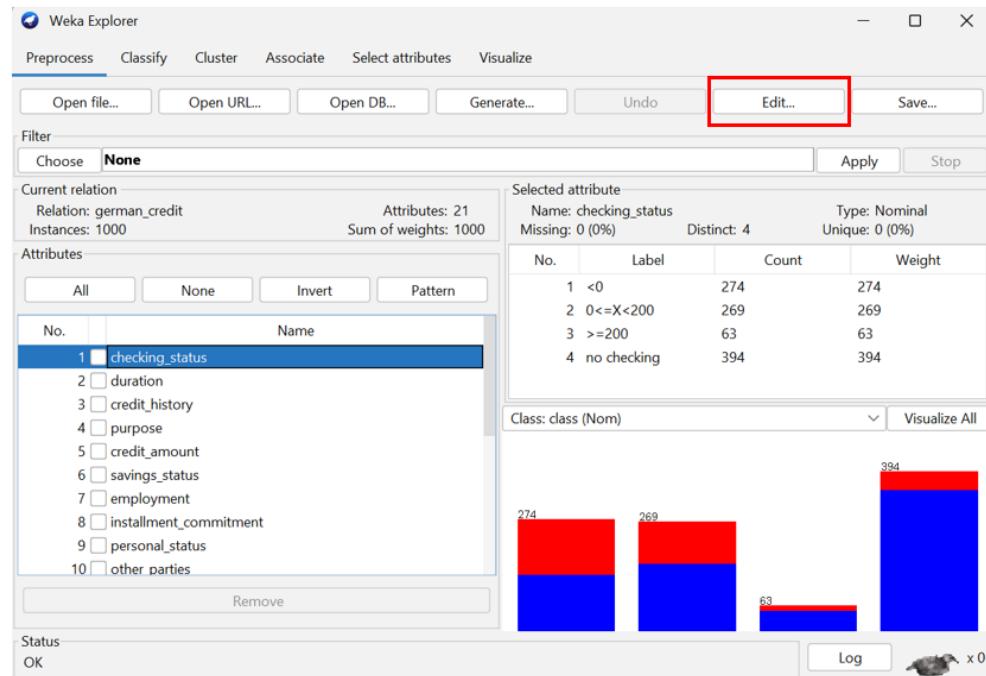
```
% Attribute 20: (qualitative)
%           foreign worker
%           A201 : yes
%           A202 : no
```



Thuộc tính này có 2 loại giá trị có số lượng mẫu tương ứng như sau: yes (là người lao động ngoại quốc) có 963 mẫu, no (không phải là người lao động ngoại quốc) có 37 mẫu.

### **2.3.2. Xác định thuộc tính được chọn làm nhãn**

- Thuộc tính được chọn làm nhãn là thuộc tính ‘**class**’. Ta xem bằng cách chọn vào ô ‘**Edit**’, thuộc tính nào làm nhãn sẽ được in đậm.

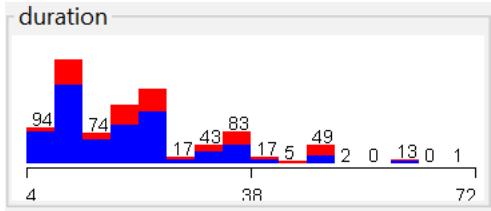


The screenshot shows the Weka Viewer interface. The title is 'Viewer' and the relation is 'german\_credit'. The table has columns: 15: housing (Nominal), 16: existing\_credits (Numeric), 17: job (Nominal), 18: num\_dependents (Numeric), 19: own\_telephone (Nominal), 20: foreign\_worker (Nominal), and 21: class (Nominal). The 'class' column is highlighted with a red box. The data rows show various combinations of attributes leading to 'good' or 'bad' outcomes. Buttons at the bottom include Add instance, Undo, OK, and Cancel.

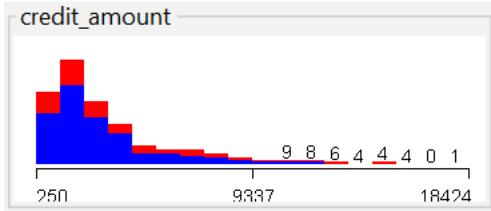
### 2.3.3. Mô tả sự phân bố của các thuộc tính có giá trị liên tục (Lệch trái hay lệch phải)

Xét nhóm thuộc tính dạng numerical gồm: *duration*, *credit\_amount*, *installment\_commitment*, *residence\_since*, *age*, *existing\_credits*, *num\_dependents*.

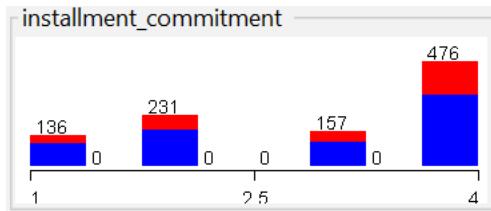
- Duration:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch phải



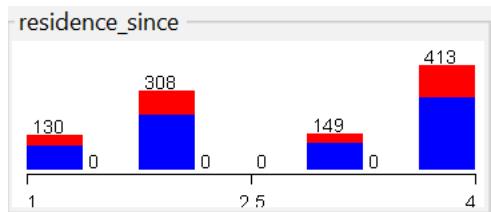
- **Credit\_amount:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch phải



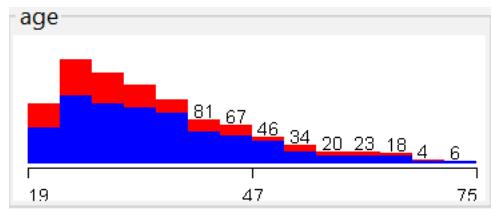
- **Installment\_commitment:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch trái



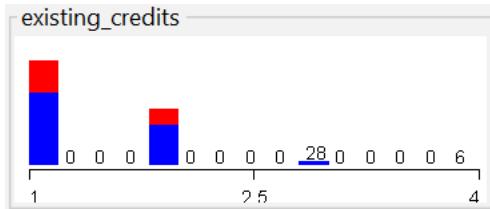
- **Residence\_since:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch trái



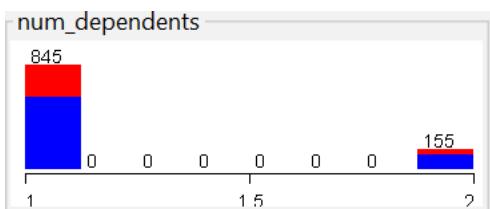
- **Age:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch phải



- **Existing\_credits:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch phải

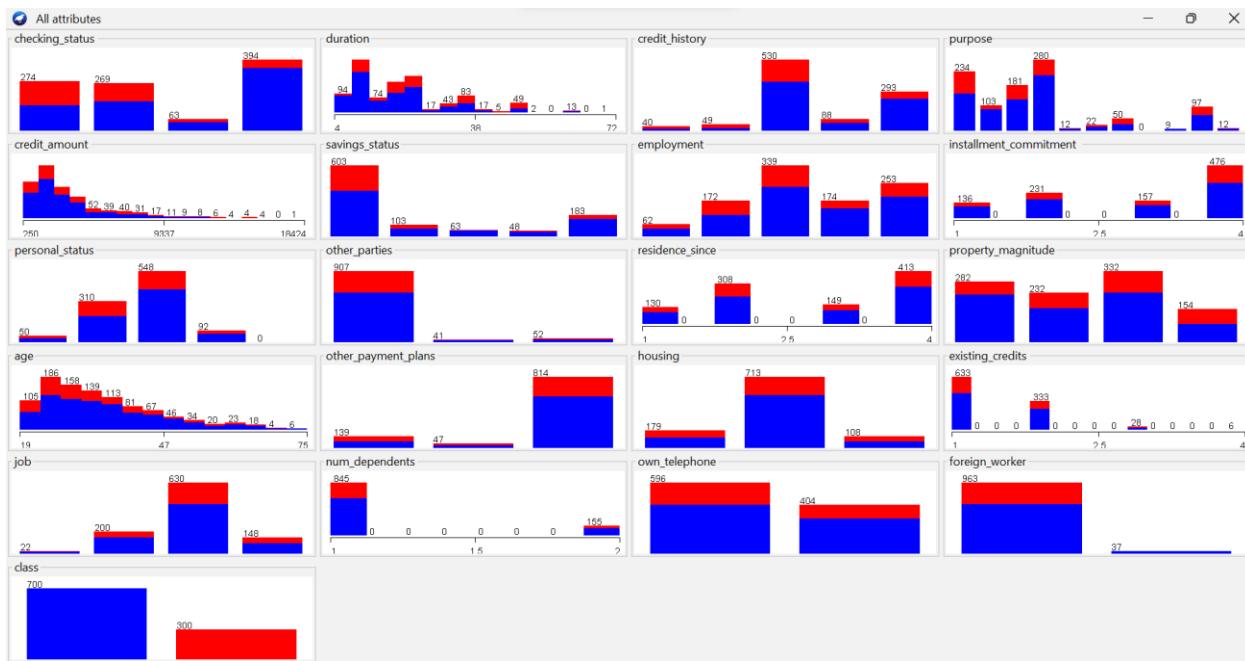


- **Num\_dependents:** Hình dáng phân phối của thuộc tính này trong dataset trên lệch phải



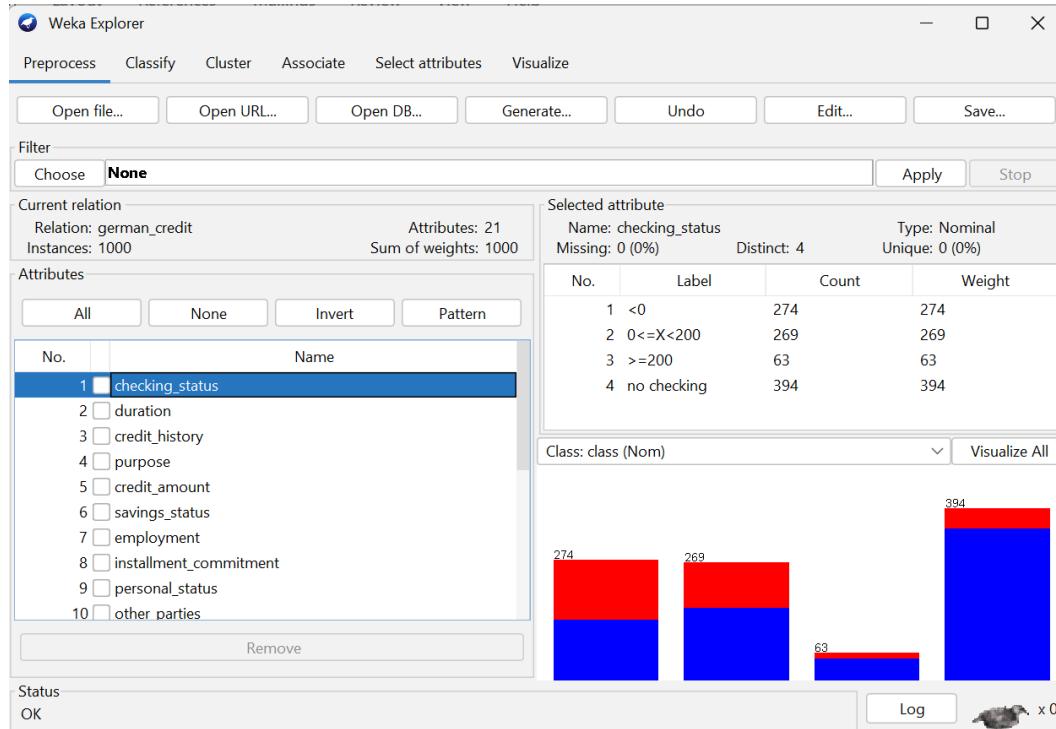
#### 2.3.4. Giải thích ý nghĩa tất cả biểu đồ trong Weka Explorer. Đặt tên cho biểu đồ và mô tả chú thích

- Tất cả biểu đồ đều **dùng để thể hiện sự phân bố các giá trị của thuộc tính và được mã hóa theo màu**. Số màu được chọn để mã hóa sẽ phụ thuộc vào số lớp trong thuộc tính được chọn làm nhãn.
- Trong tập dữ liệu này thì thuộc tính ‘class’ được chọn làm nhãn với 2 lớp ‘**good**’, ‘**bad**’.



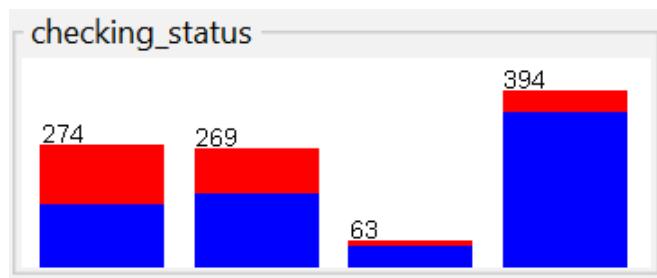
- Xét nhóm thuộc tính dạng nominal/categorical gồm: *checking\_status*, *credit\_history*, *purpose*, *savings\_status*, *employment*, *personal\_status*, *other\_parties*, *property\_magnitude*, *other\_payment\_plans*, *housing*, *job*, *own\_telephone*, *foreign\_worker*, *class*.

- **Thuộc tính: checking\_status**



Thuộc tính có 3 loại giá trị với số lượng mẫu tương ứng như sau: ‘<0’ có 274 mẫu, ‘0<=X<200’ có 269 mẫu, ‘>=200’ có 63 mẫu, ‘no checking’ có 394 mẫu. Trong mỗi loại giá trị thì các mẫu sẽ được chia thành 2 lớp lần lượt là màu xanh cho lớp ‘good’ (nghĩa là rủi ro tốt, tức là có khả năng hoàn trả) và màu đỏ cho lớp ‘bad’ (nghĩa là rủi ro xấu, không có khả năng hoàn trả).

Đặt tên cho đồ thị:

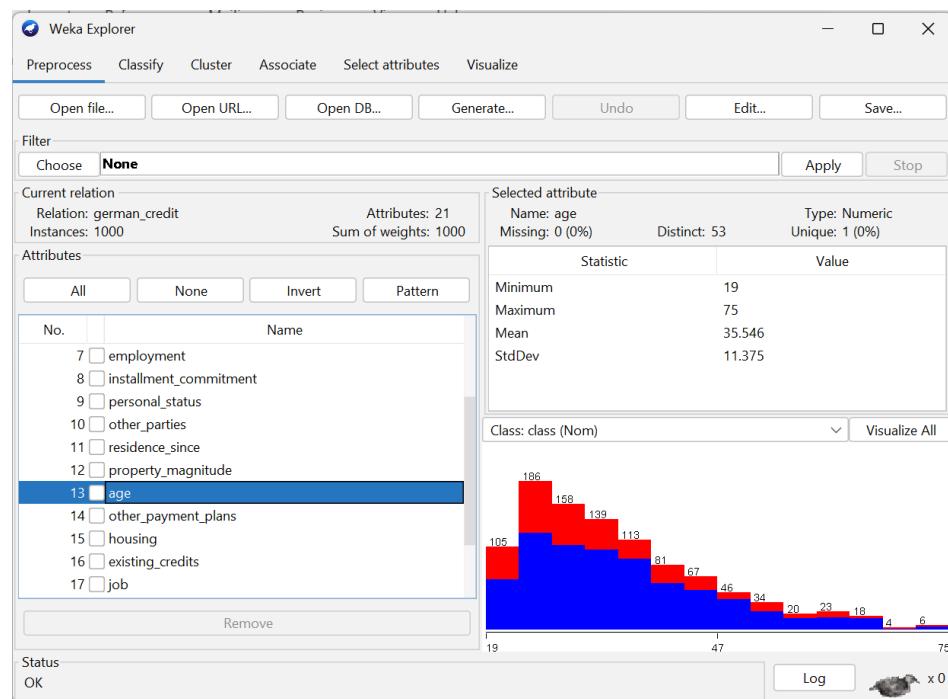


Đồ thị phân bố giá trị của thuộc tính checking\_status theo nhãn.

**Cách giải thích và đặt tên đồ thị tương tự cho các thuộc tính còn lại trong cùng nhóm.**

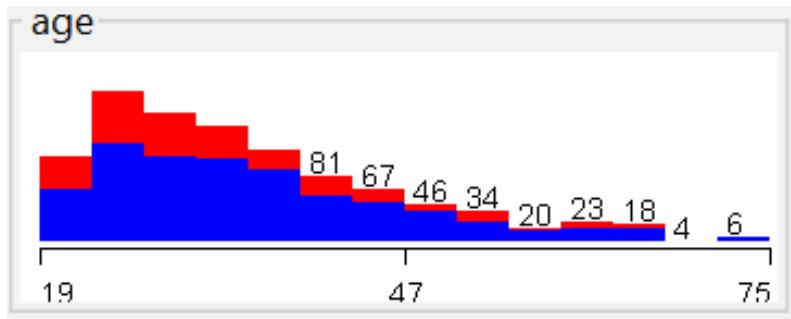
- Xét nhóm thuộc tính dạng numerical gồm: *duration*, *credit\_amount*, *installment\_commitment*, *residence\_since*, *age*, *existing\_credits*, *num\_dependents*.

#### • Thuộc tính: age



Do thuộc tính dạng numerical nên đồ thị sẽ là một khối liên tục, vẫn thể hiện sự phân bố giá trị của thuộc tính và các giá trị cũng được phân theo 2 lớp là ‘**good**’ (màu xanh) và ‘**bad**’ (màu đỏ).

Đặt tên cho đồ thị:

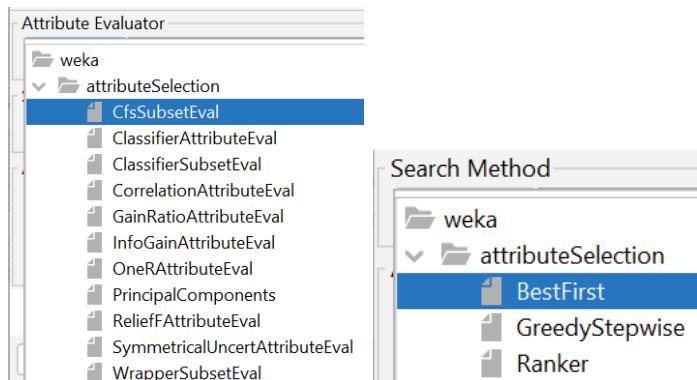
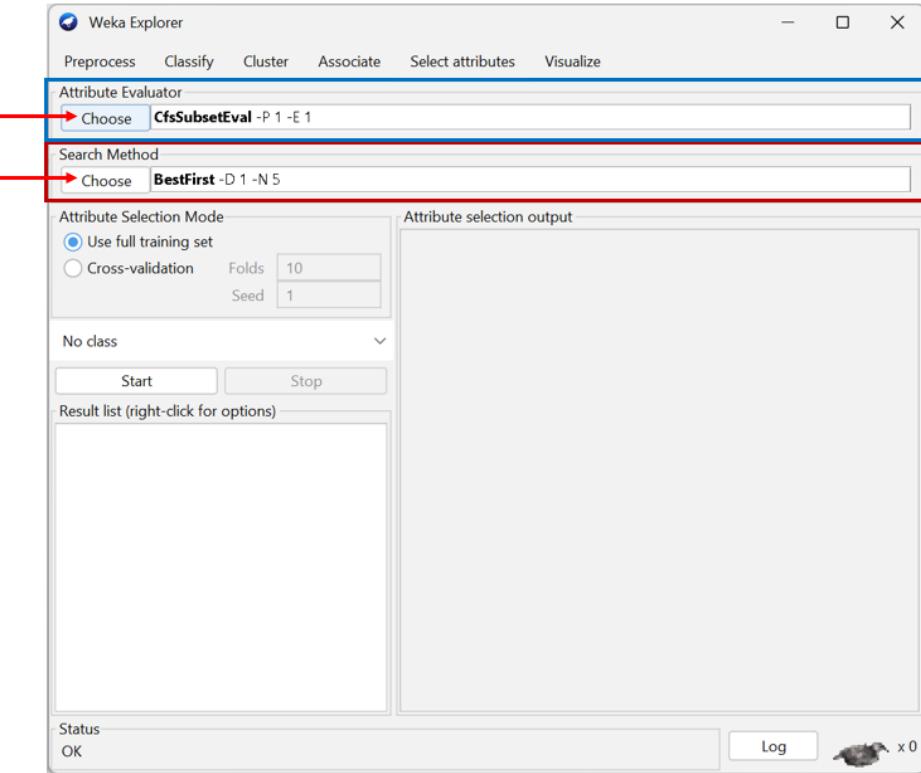


Đồ thị phân bố giá trị của thuộc tính age theo nhãn.

**Cách giải thích và đặt tên đồ thị tương tự cho các thuộc tính còn lại trong cùng nhóm.**

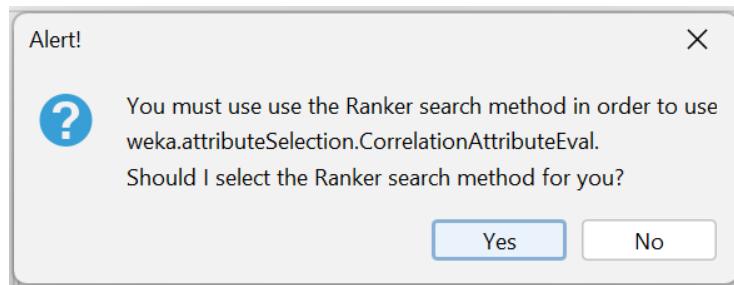
#### 2.3.5. Giải thích và miêu tả các lựa chọn trong tag *Select attributes*

Thẻ này gồm 2 phần chính là: Attribute Evaluator (Đánh giá thuộc tính) và Search Method (Phương pháp tìm kiếm). Mỗi phần có nhiều kỹ thuật để lựa chọn bằng cách ấn vào nút Choose, khi đó danh sách các kỹ thuật sẽ hiện ra.

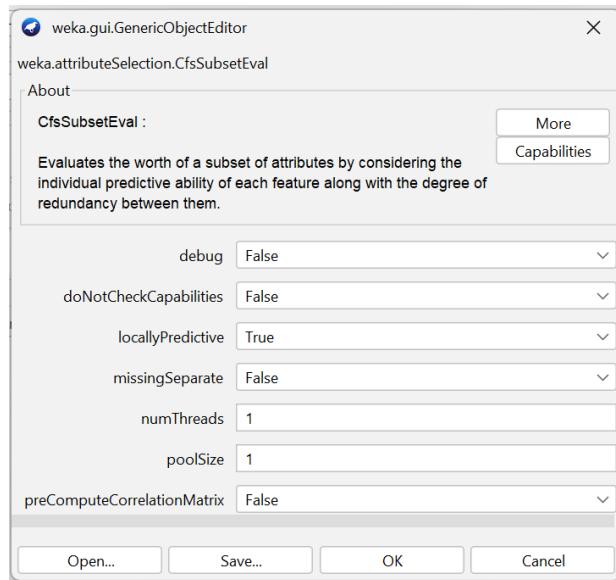


Attribute Evaluator là kỹ thuật mà mỗi thuộc tính trong tập dữ liệu của bạn (còn được gọi là cột hoặc tính năng) được đánh giá trong ngữ cảnh của biến đầu ra (ví dụ: lớp). Search Method là kỹ thuật để thử hoặc điều hướng các kết hợp thuộc tính khác nhau trong tập dữ liệu để đi đến một danh sách ngắn các tính năng đã chọn.

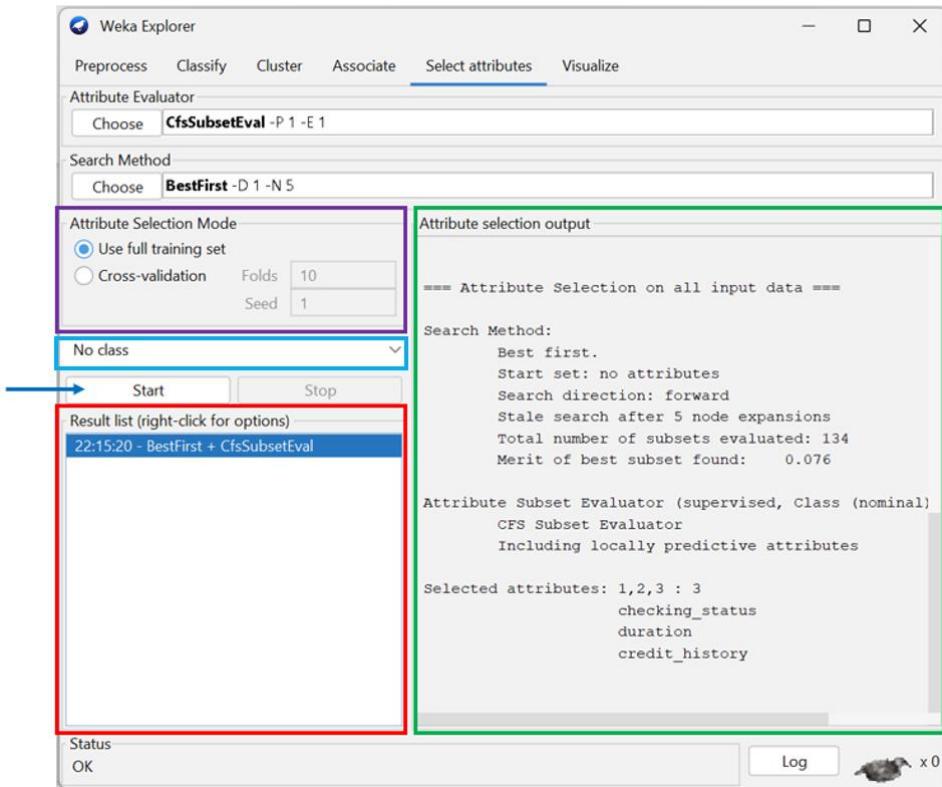
Một vài kỹ thuật Attribute Evaluator yêu cầu sử dụng các kỹ thuật Search Method cụ thể. Điển hình như: kỹ thuật CorrelationAttributeEval chỉ có thể được sử dụng với kỹ thuật Ranker Search Method, đánh giá từng thuộc tính và liệt kê kết quả theo thứ tự xếp hạng. Khi chọn các kỹ thuật Attribute Evaluator khác nhau, giao diện có thể yêu cầu thay đổi Search Method cho phù hợp với kỹ thuật đã chọn.



Ta có thể tự cấu hình các kỹ thuật ở cả 2 phần Attribute Evaluator và Search Method bằng cách nhấp vào tên của chính kỹ thuật đó để xuất hiện cửa sổ mới bao gồm các thông tin chi tiết cấu hình của kỹ thuật này.



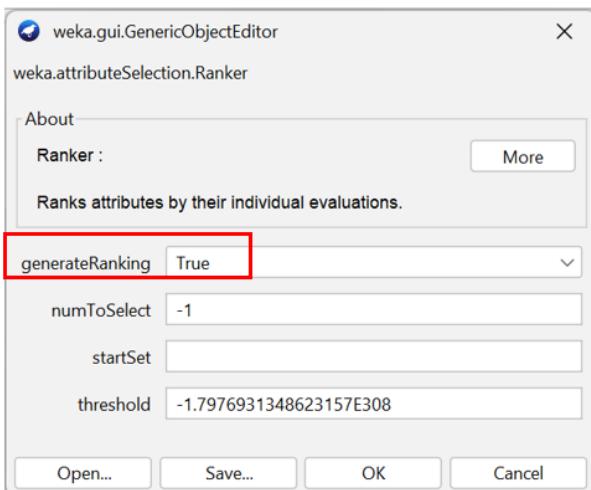
Ở mục Attribute Selection Mode có 2 lựa chọn là: Use full training set và Cross-validation. Sau khi chọn các kỹ thuật Attribute Evaluator, Search Method và Attribute Selection Mode và lớp, ta ấn nút Start để thực hiện. Kết quả sẽ thể hiện ở khung Result list và kết quả chi tiết ở khung Attribute selection output.



### 2.3.6. Lựa chọn để lấy được 5 thuộc tính có độ tương quan cao nhất

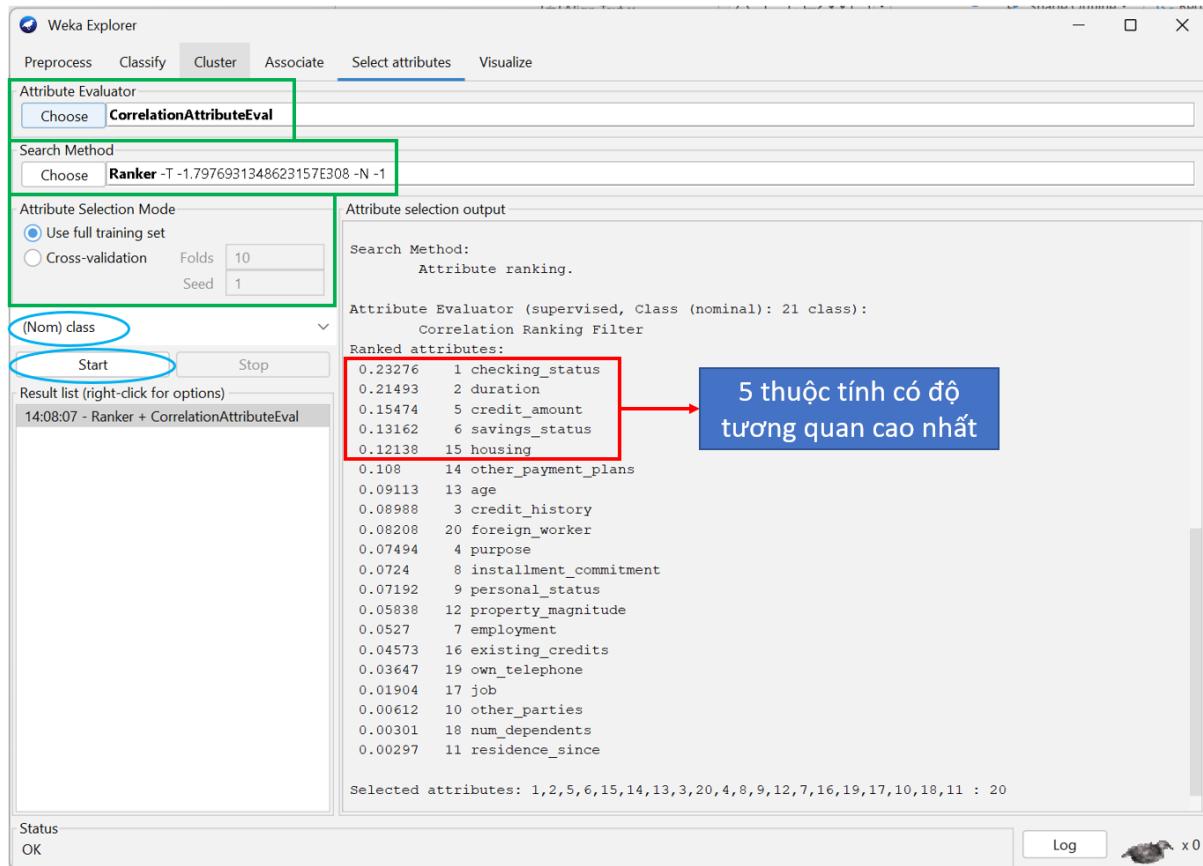
Ở thẻ Select attributes:

- Mục Attribute Evaluator: ta ấn Choose và chọn kỹ thuật CorrelationAttributeEval
- Mục Search Method: ta chọn Ranker và thiết lập Generate Ranking=True



- Mục Attribute Selection Mode: ta chọn chế độ Use full training set
- Cuối cùng, ta chọn ‘*class*’ là lớp

Sau đó, ta ấn Start để bắt đầu thực hiện, ta nhận được 5 thuộc tính có độ tương quan cao nhất là: checking\_status, duration, credit\_amount, saving\_status và housing.



### 3. TIỀN XỬ LÝ DỮ LIỆU BẰNG PYTHON

\*Lưu ý: Cài đặt các thư viện sau:

- Sử dụng thư viện Pandas cho việc đọc dữ liệu do đó cần cài thư viện pandas bằng dòng lệnh pip install pandas.
- Sử dụng thư viện Numpy cho việc so sánh các kiểu dữ liệu do đó cần cài thư viện numpy bằng dòng lệnh pip install numpy.

\*Cách chạy chương trình:

- Mở terminal của máy hoặc app bất kỳ hỗ trợ chạy terminal, chuyển đến đường dẫn chứa file .py (file chương trình) bằng cách gõ cd 'path'.
- Chạy chương trình, ví dụ code như sau: python list\_missing\_columns.py -i house-prices.csv.



### 3.1. Liệt kê các cột bị thiếu dữ liệu

- **Tránh trường hợp người dùng chạy chương trình nhưng không truyền vào tham số dòng lệnh nào:** Nhóm đã thiết kế để hiển thị thông báo người dùng xem hướng dẫn của chương trình để biết các đối số của chương trình.

A screenshot of a Windows PowerShell window titled "Windows PowerShell". The command entered is "python list\_missing\_columns.py". A red box highlights the command line. A blue box contains the text "Gọi chương trình nhưng không truyền vào tham số dòng lệnh nào" (Call the program but do not pass a command-line parameter). Another blue box at the bottom left contains the text "Thông báo xem hướng dẫn" (Information: See the guide). Red arrows point from the highlighted command line to the explanatory text boxes.

- Thêm tham số dòng lệnh -h/--help để xem hướng dẫn của chương trình:

A screenshot of a Windows PowerShell window titled "Windows PowerShell". The command entered is "python list\_missing\_columns.py -h". A red box highlights the command line. The output shows the program's usage information, including options for "-h/--help to show instructions" and "-i/--input=... to input path to csv file".

- Hình ảnh minh họa chương trình khi truyền vào tên file (nếu tham số dòng lệnh ở dạng short options thì giá trị truyền vào cách khoảng trắng so với tham số, nếu ở dạng long options thì truyền ngay sau dấu bằng).

```
python list_missing_columns.py --input=house-prices.csv
```



Windows PowerShell

```
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python list_missing_columns.py --input=house-prices.csv
List of missing columns:
LotFrontage
Alley
MasVnrType
MasVnrArea
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
FireplaceQu
GarageType
GarageYrBlt
GarageFinish
GarageQual
GarageCond
PoolQC
Fence
MiscFeature
Sum of missing columns: 18
```

Kết quả của chương trình

### 3.2 Đếm số dòng bị thiếu dữ liệu

- Thêm tham số dòng lệnh -h/--help để xem hướng dẫn của chương trình:

Windows PowerShell

```
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python count_missing_rows.py --help
=====
|| DATA PREPROCESSING: ||
|| -h/--help to show instructions
|| -i/--input=... to input path to csv file ||
=====
```

- Hình ảnh minh họa chương trình:

Windows PowerShell

```
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python count_missing_rows.py -i house-prices.csv
The number of rows that have missing values: 1000
```

Tham số dòng lệnh

### 3.3. Điền các giá trị còn thiếu trong cột với các phương thức: mean và median (*đối với cột có kiểu dữ liệu là numeric và mode đối với cột có kiểu dữ liệu là mode*)

**Lưu ý:** Nếu cột là rỗng (không có dữ liệu gì thì sẽ không điền dữ liệu). Ví dụ: Cột PoolQC khi điền giá trị sẽ chỉ hiện các dòng giá trị là nan (Xem hình ảnh cụ thể ở ví dụ trong trường hợp 3).



- Thêm tham số dòng lệnh `-h`/`--help` để xem hướng dẫn của chương trình:

```
D:\CODE\KTDLUD\LAB01\SourceCode> python fill_missing_values.py -h
=====
|| DATA PREPROCESSING:
|| 
|| -h/--help to show instructions
|| -i/--input=... to input path to csv file
|| -c/--column=... to input column you want to fill in
|| -m/--method=...(mean/median/mode) to input method you want to use to fill in
|| -a/--all=(true/false) to input if you want to fill all missing columns or not
|| -o/--output=... to output path to output file
|| DEFAULT METHOD IS MEAN, DEFAULT OUTPUT FILE PATH: "filled_file.csv"
|| DEFAULT ALL: FALSE
=====
```

- Hình ảnh minh họa chương trình:

- **Trường hợp 1:** Nếu người dùng chọn phương pháp không hợp lý cho cột cần điền giá trị thì chương trình sẽ báo dòng lệnh yêu cầu người dùng kiểm tra lại phương pháp và kết quả file xuất sẽ rỗng

```
D:\CODE\KTDLUD\LAB01\SourceCode> python fill_missing_values.py -i house-prices.csv -c Fence -m mean -a false -o filled.csv
This method is unreasonable, please check again
filled.csv  is empty
```

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd
1	1	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
2	2	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
3	3	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
4	4	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
5	5	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
6	6	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
7	7	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
8	8	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
9	9	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
10	10	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
11	11	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
12	12	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
13	13	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
14	14	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
15	15	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
16	16	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
17	17	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
18	18	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
19	19	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
20	20	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960
21	21	R	65.0	21800.0	Pave	N	Reg	Reg	AllPub	Corner	Gtl	Blmngtn	Normal	1Fam	Simp	7.0	7.0	1960	1960

- **Trường hợp 2:** Người dùng nhập phương pháp hợp lý và chỉ yêu cầu thay trên một cột (-a false/ --all=false)

Command Prompt

```
D:\CODE\KTDLUD\LAB01\SourceCode> python fill_missing_values.py -i house-prices.csv -c Fence -m mode -a false -o filled.csv
Filling missing values done, checking in filled.csv
```

	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
1	GarageArea	GarageQual	GarageCond	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence
2	954 TA	TA	Y		0	56	0	0	0	0	0	
3	462 TA	TA	Y		0	0	0	0	0	0	0	
4	208 TA	TA	Y		0	0	112	0	0	0	0	GdWo
5	160 Fa	TA	Y		0	141	0	0	0	0	0	MnPrv
6	312 TA	TA	Y		355	0	0	0	0	0	0	
7	792 TA	TA	Y		0	152	0	0	0	0	0	
8	480 TA	TA	Y		0	80	0	0	0	0	0	
9	402 TA	TA	Y		0	125	0	0	0	0	0	
10	560 TA	TA	Y		125	192	0	0	0	0	0	
11	539 TA	TA	Y		0	23	112	0	0	0	0	
12	294 TA	TA	Y		250	0	0	0	0	0	0	
13	884 TA	TA	Y		0	64	0	0	0	0	0	
14	451 TA	TA	Y		252	64	0	0	0	0	0	
15	480 TA	TA	Y		0	0	0	0	0	0	0	GdWo
16	665 TA	TA	Y		0	72	174	0	0	0	0	
17	338 TA	TA	Y		0	0	0	0	0	0	0	
18	513 Fa	Fa	Y		0	0	96	0	0	0	0	
19	506 TA	TA	Y		0	34	0	0	0	0	0	
20	576 TA	TA	Y		112	0	0	0	0	0	0	MnPrv
21	572 TA	TA	Y		100	110	0	0	0	0	0	

Cột Fence trước khi điền các giá trị thiếu



	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
1	GarageArea	GarageType	GarageCondition	PavedDrive	WoodDeckSqr	OpenPorchSqr	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence
2	954	TA	Y		0	56	0	0	0	0	0	MnWw
3	462	TA	Y		0	0	0	0	0	0	0	MnWw
4	208	TA	Y		0	0	112	0	0	0	0	MnWw
5	160	Fa	Y		0	141	0	0	0	0	0	MnWw
6	312	TA	Y		355	0	0	0	0	0	0	GdWo
7	792	TA	Y		0	152	0	0	0	0	0	MnWw
8	480	TA	Y		0	80	0	0	0	0	0	MnPrv
9	402	TA	Y		0	125	0	0	0	0	0	MnWw
10	560	TA	Y		125	192	0	0	0	0	0	MnWw
11	539	TA	Y		0	23	112	0	0	0	0	MnWw
12	294	TA	Y		250	0	0	0	0	0	0	MnWw
13	884	TA	Y		0	64	0	0	0	0	0	MnPrv
14	451	TA	Y		252	64	0	0	0	0	0	MnWw
15	480	TA	Y		0	0	0	0	0	0	0	GdWo
16	665	TA	Y		0	72	174	0	0	0	0	MnWw
17	338	TA	Y		0	0	0	0	0	0	0	MnWw
18	513	Fa	Y		0	0	96	0	0	0	0	MnWw
19	506	TA	Y		0	34	0	0	0	0	0	MnWw
20	576	TA	Y		112	0	0	0	0	0	0	MnPrv
21	572	TA	Y		100	110	0	0	0	0	0	MnWw

Cột Fence sau khi điền các giá trị thiếu bằng phương pháp mode

- Trường hợp 3:** Người dùng nhập phương pháp hợp lý và yêu cầu điền các giá trị còn thiếu vào tất cả các cột bị thiếu dữ liệu (-a true/ --all=true). Ví dụ ta xét 3 cột như sau:

```
C:\ Command Prompt
D:\CODE\KTDLUD\LAB01\SourceCode> python fill_missing_values.py -i house-prices.csv -a true -o filled.csv
Filling missing values done, checking in filled.csv
```

	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC
1	GarageArea	GarageType	GarageCondition	PavedDrive	WoodDeckSqr	OpenPorchSqr	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MosSold	YrsSold	SaleType	SaleCondi	SalePrice
2	954	TA	Y		0	56	0	0	0	0	0				0	6	2007 New	Partial	248328
3	462	TA	Y		0	0	0	0	0	0	0				0	3	2007 WD	Normal	101800
4	208	TA	Y		0	0	112	0	0	0	0				0	7	2008 WD	Normal	120000
5	160	Fa	Y		0	141	0	0	0	0	0				0	4	2008 WD	Normal	91000
6	312	TA	Y		355	0	0	0	0	0	0				0	4	2008 WD	Normal	141000
7	792	TA	Y		0	152	0	0	0	0	0				0	4	2009 WD	Normal	124000
8	480	TA	Y		0	80	0	0	0	0	0				0	6	2009 WD	Normal	139000
9	402	TA	Y		0	125	0	0	0	0	0				0	5	2006 WD	Normal	164000
10	560	TA	Y		125	192	0	0	0	0	0				0	6	2009 WD	Normal	215000
11	539	TA	Y		0	23	112	0	0	0	0				0	1	2009 WD	Normal	103000
12	294	TA	Y		250	0	0	0	0	0	0				0	6	2010 WD	Normal	145000
13	884	TA	Y		0	64	0	0	0	0	0				0	10	2006 WD	Normal	146000
14	451	TA	Y		252	64	0	0	0	0	0				0	6	2008 WD	Normal	176000
15	480	TA	Y		0	0	0	0	0	0	0				0	6	2007 WD	Normal	123000
16	665	TA	Y		0	72	174	0	0	0	0				0	5	2008 COD	Abnorml	287000
17	338	TA	Y		0	0	0	0	0	0	0				0	8	2009 WD	Normal	133500
18	513	Fa	Y		0	0	96	0	0	0	0				0	5	2008 COD	Abnorml	98000
19	506	TA	Y		0	34	0	0	0	0	0				0	3	2006 WD	Normal	183900
20	576	TA	Y		112	0	0	0	0	0	0				0	4	2009 WD	Normal	141500
21	572	TA	Y		100	110	0	0	0	0	0				0	6	2007 WD	Normal	129900

Dữ liệu trước khi điền các giá trị còn thiếu

A1	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC
1	GarageAr	GarageQu	GarageCo	PavedDriv	WoodDec	OpenPorc	EnclosedP	35SsnPorch	ScreenPor	PoolArea	PoolQC	Fence	MiscFeat	MiscVal	MoSold	YrSold	SaleType	SaleCondi	SalePrice
2	954	TA	TA	Y	0	56	0	0	0	0	nan	MnWw	Shed	0	6	2007	New	Partial	248328
3	462	TA	TA	Y	0	0	0	0	0	0	nan	MnWw	Shed	0	3	2007	WD	Normal	101800
4	208	TA	TA	Y	0	0	112	0	0	0	nan	MnWw	Shed	0	7	2008	WD	Normal	120000
5	160	Fa	TA	Y	0	141	0	0	0	0	nan	MnWw	Shed	0	4	2008	WD	Normal	91000
6	312	TA	TA	Y	355	0	0	0	0	0	nan	GdWo	Shed	0	4	2008	WD	Normal	141000
7	792	TA	TA	Y	0	152	0	0	0	0	nan	MnWw	Shed	0	4	2009	WD	Normal	124000
8	480	TA	TA	Y	0	80	0	0	0	0	nan	MnPrv	Shed	0	6	2009	WD	Normal	139000
9	402	TA	TA	Y	0	125	0	0	0	0	nan	MnWw	Shed	0	5	2006	WD	Normal	164000
10	560	TA	TA	Y	125	192	0	0	0	0	nan	MnWw	Shed	0	6	2009	WD	Normal	215000
11	539	TA	TA	Y	0	23	112	0	0	0	nan	MnWw	Shed	0	1	2009	WD	Normal	103000
12	294	TA	TA	Y	250	0	0	0	0	0	nan	MnWw	Shed	0	6	2010	WD	Normal	145000
13	884	TA	TA	Y	0	64	0	0	0	0	nan	MnPrv	Shed	0	10	2006	WD	Normal	146000
14	451	TA	TA	Y	252	64	0	0	0	0	nan	MnWw	Shed	0	6	2008	WD	Normal	176000
15	480	TA	TA	Y	0	0	0	0	0	0	nan	GdWo	Shed	0	6	2007	WD	Normal	123000
16	665	TA	TA	Y	0	72	174	0	0	0	nan	MnWw	Shed	0	5	2008	COD	Abnorml	287000
17	338	TA	TA	Y	0	0	0	0	0	0	nan	MnWw	Shed	0	8	2009	WD	Normal	133500
18	513	Fa	Y	0	0	96	0	0	0	0	nan	MnWw	Shed	0	5	2008	COD	Abnorml	98000
19	506	TA	TA	Y	0	34	0	0	0	0	nan	MnWw	Shed	0	3	2006	WD	Normal	183900
20	576	TA	TA	Y	112	0	0	0	0	0	nan	MnPrv	Shed	0	4	2009	WD	Normal	141500
21	572	TA	TA	Y	100	110	0	0	0	0	nan	MnWw	Shed	0	6	2007	WD	Normal	129900

Dữ liệu sau khi đã điền các giá trị thiếu

### 3.4. Xóa các dòng bị thiếu dữ liệu với ngưỡng cho trước

- Thêm tham số dòng lệnh -h/--help để xem hướng dẫn của chương trình:

```

PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python remove_rows.py -h
=====
|| DATA PREPROCESSING:
|| 
|| -h/--help to show instructions
|| -i/--input=... to input path to csv file
|| -t/--threshold=... to input allowed threshold percentage of missing value
|| -o/--output=... to output path to output file
|| DEFAULT THRESHOLD IS 0%, DEFAULT OUTPUT FILE PATH: "output.csv"
=====
```

- Hình ảnh minh họa chương trình:

**Lưu ý:** tham số threshold được thiết kế truyền vào giá trị ở dạng số nguyên (ví dụ 10% thì nhập vào là 10).

```

Tham số dòng lệnh
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python remove_rows.py -i house-prices.csv -t 10 -o result.csv
Removing done, checking in result.csv
The number of removed rows: 80
=====
```

Kiểm tra trong file result.csv:



File Explorer showing the directory structure:

```
This PC > Downloads > KTDLUD > Lab01 > SoureCode >
```

	Name	Date modified	Type	Size
Quick access	_pycache_	3/21/2023 8:51 PM	File folder	
Desktop	count_missing_rows.py	3/18/2023 7:31 PM	PY File	1 KB
Downloads	draft.py	3/19/2023 6:06 PM	PY File	1 KB
Documents	fill_missing_columns.py	3/21/2023 7:54 PM	PY File	0 KB
Pictures	house-prices.csv	3/11/2023 9:37 PM	Microsoft Excel Com...	302 KB
repos	list_missing_columns.py	3/19/2023 10:31 AM	PY File	1 KB
ML	remove_columns.py	3/18/2023 9:58 PM	PY File	2 KB
NLP with Pyth	remove_duplicate_samples.py	3/18/2023 10:21 PM	PY File	3 KB
NLP Coursera	remove_rows.py	3/18/2023 9:53 PM	PY File	2 KB
SourceCode	result.csv	3/21/2023 9:22 PM	Microsoft Excel Com...	294 KB
SourceCode	utils.py	3/21/2023 8:35 PM	PY File	5 KB

Excel spreadsheet showing house-prices.csv data:

A1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
911	1255	ZU RL		9550 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Wtches	Norm	Norm	1Fam	1Story	6	6		
912	1053	60 RL		100	9500 Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	mes	Artery	Norm	1Fam	2Story	8	5	
913	582	20 RL		98	12704 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	6	5	
914	1420	20 RL	nan	16381	Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	1Story	6	5	
915	1417	190 RM		60	11340 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	2fmCon	2Story	4	6	
916	668	20 RL		65	8125 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Norm	Norm	1Fam	1Story	6	5	
917	1190	60 RL		60	7500 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	
918	192	60 RL	nan	7472	Pave	nan	IR1	Lvl	AllPub	CulSac	Gtl	mes	Norm	Norm	1Fam	2Story	7	9	
919	990	60 FV		65	8125 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story	7	5	
920	982	60 RL		98	12203 Pave	nan	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	
921	862	190 RL		75	11625 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	2fmCon	1Story	5	4	
922																			
923																			
924																			
925	Số dòng chỉ còn 920 dòng																		
926																			

### 3.5. Xóa các cột bị thiếu dữ liệu với ngưỡng cho trước

- Thêm tham số dòng lệnh `-h`--help để xem hướng dẫn của chương trình:

Windows PowerShell window showing the command:

```
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python remove_columns.py -h
```

Output:

```
=====
|| DATA PREPROCESSING:
|| -h/--help to show instructions
|| -i/--input=... to input path to csv file
|| -t/--threshold=... to input allowed threshold percentage of missing value
|| -o/--output=... to output path to output file
|| DEFAULT THRESHOLD IS 0%, DEFAULT OUTPUT FILE PATH: "output.csv"
=====
```

- Hình ảnh minh họa chương trình:

**Lưu ý:** tham số threshold được thiết kế truyền vào giá trị ở dạng số nguyên (ví dụ 10% thì nhập vào là 10).



```
Windows PowerShell Tham số dòng lệnh
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python remove_columns.py -i house-prices.csv -t 10 -o result2.csv
Removing done, checking in result2.csv
Columns are removed: ['LotFrontage', 'Alley', 'MasVnrType', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature']
```

Kết quả

Kiểm tra lại bằng cách liệt kê lại số cột bị thiếu dữ liệu:

```
Windows PowerShell
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python list_missing_columns.py -i result2.csv
List of missing columns:
MasVnrArea
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
GarageType
GarageYrBlt
GarageFinish
GarageQual
GarageCond
Sum of missing columns: 11
```

### 3.6. Xóa các mẫu bị trùng lặp

- Thêm tham số dòng lệnh -h/--help để xem hướng dẫn của chương trình:

```
Windows PowerShell
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python remove_duplicate_samples.py --help
=====
|| DATA PREPROCESSING:
|| 
|| -h/--help to show instructions
|| -i/--input=... to input path to csv file
|| -o/--output=... to output path to output file
|| DEFAULT OUTPUT FILE PATH: "output.csv"
```

- Hình ảnh minh họa chương trình:



Tham số dòng lệnh

```
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python remove_duplicate_samples.py -i house-prices.csv
Removing done, checking in output.csv
The number of removed samples: 284
```

Kết quả

Kiểm tra trong file output.csv:

	Name	Date modified	Type	Size
Quick access				
Desktop	_pycache_	3/21/2023 8:51 PM	File folder	
Downloads	count_missing_rows.py	3/18/2023 7:31 PM	PY File	1 KB
Documents	draft.py	3/19/2023 6:06 PM	PY File	1 KB
Pictures	fill_missing_columns.py	3/21/2023 9:42 PM	PY File	1 KB
repos	house-prices.csv	3/11/2023 9:37 PM	Microsoft Excel Com...	302 KB
ML	list_missing_columns.py	3/19/2023 10:31 AM	PY File	1 KB
NLP with Pyth	remove_columns.py	3/21/2023 9:46 PM	Microsoft Excel Com...	229 KB
NLP Couresa	remove_duplicate_samples.py	3/18/2023 9:58 PM	PY File	2 KB
SourceCode	remove_rows.py	3/21/2023 9:44 PM	PY File	3 KB
SourcCode	result.csv	3/18/2023 9:53 PM	PY File	2 KB
SSMs	result2.csv	3/21/2023 9:22 PM	Microsoft Excel Com...	294 KB
	utils.py	3/21/2023 9:32 PM	Microsoft Excel Com...	289 KB
		3/21/2023 8:35 PM	PY File	5 KB

A1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
704	885	20 RL	90	11248 Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	mes	norm	norm	1Fam	1Story	9	5	
705	684	20 RL	85	9350 Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	ColgCr	Norm	Norm	1Fam	1Story	6	7	
706	254	80 RL	150	215245 Pave	nan	IR3	Low	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	SLvl	7	5	
707	314	20 RL	80	10197 Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Sev	Timber	Norm	Norm	1Fam	1Story	7	5
708	174	20 RL	72	8640 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	6	5	
709	213	60 FV	53227 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story	7	5		
710	458	20 RL	60	7200 Pave	nan	IR1	Low	AllPub	CulDSac	Mod	ClearCr	Norm	Norm	1Fam	1Story	4	6	
711	62	75 RM	114	14803 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	2.5Unf	5	7	
712	826	20 RL	75	10125 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	PosN	PosN	1Fam	1Story	10	5	
713	985	90 RL	98	12704 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	Duplex	1.5Fin	5	5	
714	582	20 RL	65	8125 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	
715	668	20 RL	60	7500 Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	SavvyerW	Norm	Norm	1Fam	1Story	6	5	
716	1190	60 RL	60	7472 Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	
717	192	nan	nan	nan	nan	Reg	Lvl	AllPub	CulDSac	Gtl	mes	Norm	Norm	1Fam	2Story	7	9	
718																		
719																		
720																		
721																		
722																		
723																		

Dữ liệu chỉ còn 716 dòng

### 3.7. Chuẩn hóa một thuộc tính dạng số bằng phương pháp min-max hoặc Z-score

- Thêm tham số dòng lệnh -h/--help để xem hướng dẫn của chương trình:

```

PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python normalize.py --help
=====
|| DATA PREPROCESSING:
|| -h/--help to show instructions
|| -i/--input=... to input path to csv file
|| -c/--column=["column1", "column2",...] to input columns you want to normalize
|| -m/--method=["method1", "method2",...] to input methods you want to use to normalize
|| -a/--all=... (true/false) to input if you want to normalize all columns or not
|| -M/--METHOD=... all selected columns are normalized by this method
|| -o/--output=... to output path to output file
|| -mi/--min=... new min to normalize data (default: 0)
|| -ma/--max=... new max to normalize data (default: 1)
|| DEFAULT OUTPUT FILE PATH: "normalizing_result.csv"
|| DEFAULT ALL: FALSE
|| DEFAULT METHOD FOR NUMERICAL ATTRIBUTE: "z_score" (ALL methods: "z_score"/"min_max")
|| DEFAULT -M is "" (empty string)
|| COLUMNS NAME MUST BE CONTAINED IN A LIST, SEPERATED BY ","
|| METHODS MUST BE CONTAINED IN A LIST, SEPERATED BY ","
|| IF A CATEGORICAL COLUMN IS PASSED, IT WILL BE IGNORE WHEN THE PROGRAM IS RAN
=====
```

- Giải thích một số tham số dòng lệnh của chương trình:

**-c/--column** sẽ nhận vào một list chứa tên các cột của bộ dữ liệu, list này có thể chứa 1 tên nếu như chỉ muốn chuẩn hóa một cột, hoặc nhiều tên nếu như muốn chuẩn hóa nhiều cột.

**-m/--method** sẽ nhận vào một list chứa tên các phương pháp dùng để chuẩn hóa, thứ tự phương pháp chứa trong list này sẽ là phương pháp tương ứng cho tên cột chứa trong list ở tham số -c. Nếu như tên trong list method chưa đủ số lượng tương ứng với tên trong list column thì phương pháp mặc định sẽ được sử dụng cho các column còn lại.

**-M/--METHOD** sẽ được dùng để chọn phương pháp cho việc chuẩn hóa tất cả các cột dạng số. Nếu không truyền giá trị cho -M thì phương pháp mặc định sẽ được dùng.

- Hình ảnh minh họa chương trình:

```

PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode> python normalize.py -i house-prices.csv -c [LotFrontage, LotArea]
Normalizing done, checking in normalizing_result.csv
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SoureCode>
```

Tham số dòng lệnh

Các cột trong dữ liệu

Hai cột LotFrontage và LotArea khi chưa được chuẩn hóa:

	A	B	C	D	E	F	G
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley
2	1242	20	RL	83	9849	Pave	
3	1233	90	RL	70	9842	Pave	
4	1401	50	RM	50	6000	Pave	
5	1377	30	RL	52	6292	Pave	
6	208	20	RL		12493	Pave	
7	1392	90	RL	65	8944	Pave	
8	980	20	RL	80	8816	Pave	
9	484	120	RM	32	4500	Pave	
10	392	60	RL	71	12209	Pave	
11	730	30	RM	52	6240	Pave	Grvl
12	255	20	RL	70	8400	Pave	
13	1094	20	RL	71	9230	Pave	
14	1021	20	RL	60	7024	Pave	
15	1341	20	RL	70	8294	Pave	
16	1025	20	RL		15498	Pave	
17	848	20	RL	36	15523	Pave	
18	457	70	RM	34	4571	Pave	Grvl
19	1266	160	FV	35	3735	Pave	
20				--	--	--	--

Hai cột LotFrontage và LotArea sau khi chuẩn hóa bằng Z-score:

	A	B	C	D	E	F	G
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley
2	1242	20	RL	0.6442	-0.0395	Pave	nan
3	1233	90	RL	0.0328	-0.0403	Pave	nan
4	1401	50	RM	-0.908	-0.4588	Pave	nan
5	1377	30	RL	-0.8139	-0.427	Pave	nan
6	208	20	RL	nan	0.2485	Pave	nan
7	1392	90	RL	-0.2024	-0.1381	Pave	nan
8	980	20	RL	0.5031	-0.152	Pave	nan
9	484	120	RM	-1.7546	-0.6222	Pave	nan
10	392	60	RL	0.0798	0.2175	Pave	nan
11	730	30	RM	-0.8139	-0.4326	Pave	Grvl
12	255	20	RL	0.0328	-0.1974	Pave	nan
13	1094	20	RL	0.0798	-0.107	Pave	nan
14	1021	20	RL	-0.4376	-0.3472	Pave	nan
15	1341	20	RL	0.0328	-0.2089	Pave	nan
16	1025	20	RL	nan	0.5758	Pave	nan
17	848	20	RL	-1.5665	0.5785	Pave	nan
18	457	70	RM	-1.6606	-0.6144	Pave	Grvl
19	1266	160	FV	-1.6135	-0.7055	Pave	nan
20	605	50	RM	-0.8600	-0.4457	Pave	nan

- Chuẩn hóa tất cả các cột dạng số bằng min-max:



```
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SourceCode> python normalize.py -i house-prices.csv -a true -M min_max -o result.csv
C:/Users/t480/Downloads/KTDLUD/Lab01/SourceCode/utils.py:333: RuntimeWarning: invalid value encountered in double_scalars
    result.append(round((float(data[col][i])-min_)/(max_-min_)*(newMax-newMin) + newMin),4))
Normalizing done, checking in result.csv
PS C:\Users\t480\Downloads\KTDLUD\Lab01\SourceCode>
```

Cảnh báo của python là do chia cột có toàn giá trị 0.

A1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	LotConfig	LandSlope	Neighborhood	Condition	BldgType	HouseStyle	OverallQual	OverallCond	Year
2	0.8505	0 RL	0.4697	0.0392	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Somers	Norm	Norm	1Fam	1Story	0.6667	0.625
3	0.8443	0.4118 RL	0.3712	0.0391	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	0.3333	0.5
4	0.9595	0.1765 RM	0.2197	0.0212	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	0.5556	0.75
5	0.9431	0.0588 RL	0.2348	0.0225	Pave	nan	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	0.5556	0.5
6	0.1413	0 RL	nan	0.0515	Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	0.3333	0.5
7	0.9534	0.4118 RL	0.3333	0.0349	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	0.4444	0.5
8	0.6708	0 RL	0.447	0.0343	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	0.4444	0.625
9	0.3306	0.5882 RM	0.0833	0.0141	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	0.5556	0.5
10	0.2675	0.2353 RL	0.3788	0.0502	Pave	nan	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	0.5556	0.5
11	0.4993	0.0588 RM	0.2348	0.0223	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	0.3333	0.5
12	0.1735	0 RL	0.3712	0.0324	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	0.4444	0.625
13	0.749	0 RL	0.3788	0.0363	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	0.4444	0.875
14	0.6989	0 RL	0.2955	0.0259	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	0.3333	0.5
15	0.9184	0 RL	0.3712	0.0319	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	0.3333	0.5
16	0.7016	0 RL	nan	0.0656	Pave	nan	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	0.7778	0.625
17	0.5802	0 RL	0.1136	0.0657	Pave	nan	IR1	Lvl	AllPub	CulDSac	Gtl	CollCr	Norm	Norm	1Fam	1Story	0.4444	0.625
18	0.3121	0.2941 RM	0.0985	0.0145	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	0.4444	0.5
19	0.8669	0.8235 FV	0.1061	0.0106	Pave	nan	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	0.6667	0.5
20	0.4752	0.1765 RM	0.2272	0.0217	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1.5Fin	1.5Fin	0.4444	0.625

Tất cả các cột dạng số đều có giá trị ở khoảng 0 đến 1.

### 3.8. Thực hiện cộng, trừ, nhân, chia giữa các thuộc tính có kiểu dữ liệu là numeric

- Thêm tham số dòng lệnh `-h`/`--help` để xem hướng dẫn của chương trình:

```
Command Prompt
D:\CODE\KTDLUD\LAB01\SourceCode> python calculate_2numeric_columns.py -h
=====
| DATA PREPROCESSING:
|
|-h--help to show instructions
|-i--input=... to input path to csv file
|-f--first=... to input first column you want to calculate
|-s--second=... to input second column you want to calculate
|-o--output=... to output path to output file
|-m--method=... (add/sub/mul/div) to calculate first and second column
| DEFAULT METHOD IS ADD, DEFAULT OUTPUT FILE PATH: "caculation_result.csv"
=====
```

- Hình ảnh minh họa chương trình:

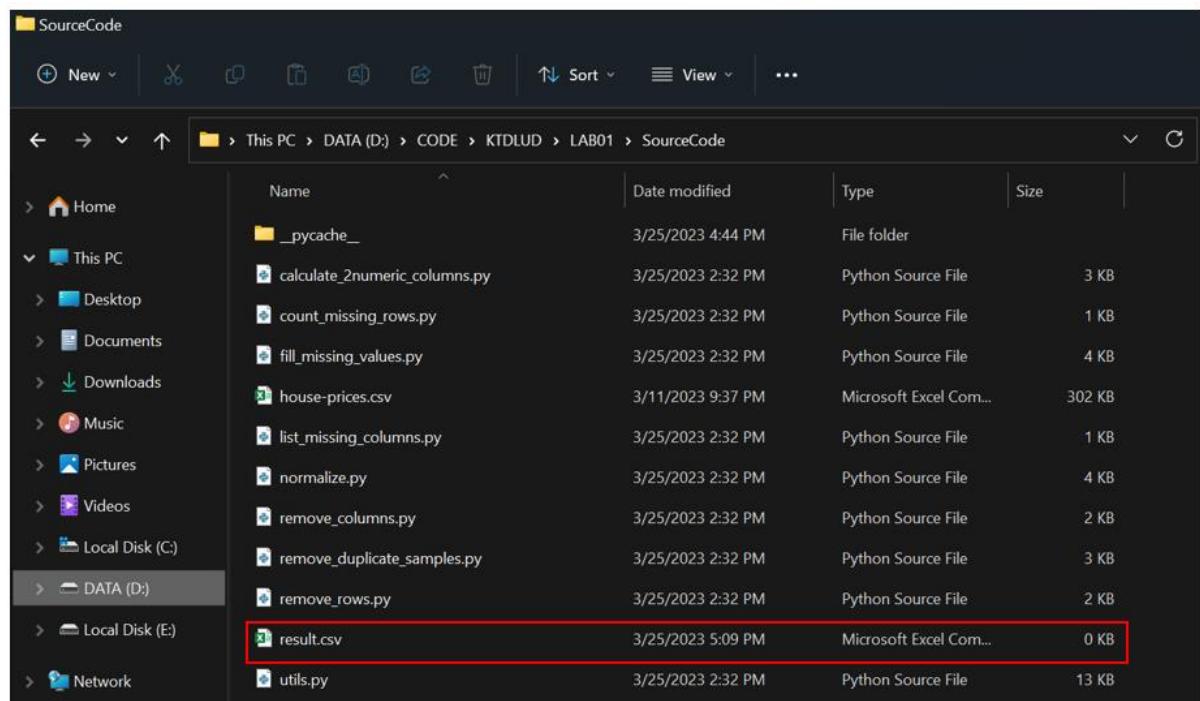
**Lưu ý:** Tham số `first` được thiết kế để truyền vào tên cột đầu tiên. Tham số `second` được thiết kế để truyền vào tên cột thứ hai. Khi thực hiện phép cộng, trừ, nhân hay chia thì chương trình được thiết kế lấy các giá trị cột đầu tiên thực hiện với các giá trị của cột thứ hai (**add**:  $f+s$ , **sub**:  $f-s$ , **mul**:  $f*s$ , **div**:  $f/s$ ).

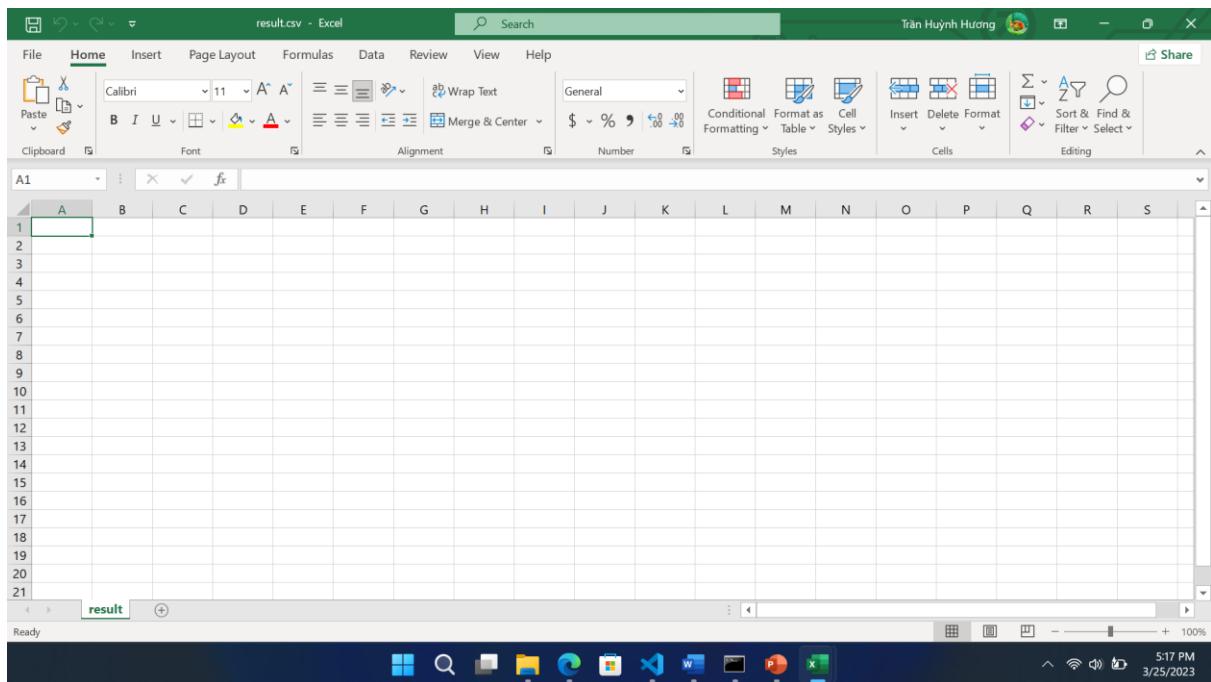
- **Trường hợp 1:** Người dùng yêu cầu thực hiện phép tính đối với các cột có dữ liệu là categorical. Chương trình sẽ yêu cầu kiểm tra lại kiểu thuộc tính của cột người dùng nhập và thông báo file kết quả sẽ không có dữ liệu.

```
D:\CODE\KTDLUD\LAB01\SourceCode> python calculate_2numeric_columns.py -i house-prices.csv -f Id -s Street -o result.csv
Please check type of these two column again
result.csv is empty cause type of these column are unresonable
```

Numeric      categorical

Kiểm tra file csv:





- **Trường hợp 2:** Thực hiện công (-m add/ --method=add)

```
C:\ Command Prompt
D:\CODE\KTDLUD\LAB01\SourceCode> python calculate_2numeric_columns.py -i house-prices.csv -f Id -s MSSubClass -m add -o result.csv
Calculating done, checking in  result.csv
```

A	B
1	Result
2	1262
3	1323
4	1451
5	1407
6	228
7	1482
8	1000
9	604
10	452
11	760
12	275
13	1114
14	1041
15	1361
16	1045
17	868
18	527
19	1426
20	745
21	144



- **Trường hợp 3:** Thực hiện trừ (-m sub/ --method=sub)

```
Command Prompt

D:\CODE\KTDLUD\LAB01\SourceCode> python calculate_2numeric_columns.py -i house-prices.csv -f Id -s MSSubClass -m sub -o result.csv
Calculating done, checking in result.csv
```

	A	B
1	Result	
2	1222	
3	1143	
4	1351	
5	1347	
6	188	
7	1302	
8	960	
9	364	
10	332	
11	700	
12	235	
13	1074	
14	1001	
15	1321	
16	1005	
17	828	
18	387	
19	1106	
20	645	
21	-96	

- **Trường hợp 4:** Thực hiện nhân (-m mul/ --method=mul)

```
Command Prompt

D:\CODE\KTDLUD\LAB01\SourceCode> python calculate_2numeric_columns.py -i house-prices.csv -f Id -s MSSubClass -m mul -o result.csv
Calculating done, checking in result.csv
```



A	B
1	Result
2	24840
3	110970
4	70050
5	41310
6	4160
7	125280
8	19600
9	58080
10	23520
11	21900
12	5100
13	21880
14	20420
15	26820
16	20500
17	16960
18	31990
19	202560
20	34750
21	2880

- **Trường hợp 5:** Thực hiện chia (-m div/ --method=div)

```
Command Prompt
D:\CODE\KTDLUD\LAB01\SourceCode> python calculate_2numeric_columns.py -i house-prices.csv -f Id -s MSSubClass -m div -o result.csv
Calculating done, checking in result.csv
```

A	B
1	Result
2	62.1
3	13.7
4	28.02
5	45.9
6	10.4
7	15.46667
8	49
9	4.033333
10	6.533333
11	24.33333
12	12.75
13	54.7
14	51.05
15	67.05
16	51.25
17	42.4
18	6.528571
19	7.9125
20	13.9
21	0.2



## C. Tài liệu tham khảo

- Geeksforgeeks: [Command Line Arguments in Python - GeeksforGeeks](#)
- Slides lý thuyết
- Trang chủ WEKA