

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
BÁO CÁO LAB01
TIỀN XỬ LÝ VÀ KHÁM PHÁ DỮ LIỆU

ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQGTPHCM

SVTH: HÀ NGHUYỄN THẢO VY

GVHD: NGUYỄN THỊ THU HÀNG

NH: 2022 - 2023

MỤC LỤC

I. Thông tin sinh viên.....	2
II. Bảng báo cáo công việc	2
III. Nội dung bài làm	4
1. Cài đặt WEKA	4
2. Làm quen với WEKA	7
2.1. Khám phá tập dữ liệu Breast Cancer	7
2.2. Khám phá tập dữ liệu Weather	19
2.3. Khám phá tập dữ liệu Credit in Germany	23
3. Tiền xử lý dữ liệu với Python.....	29

I. Thông tin sinh viên

Họ và tên: Hà Nguyễn Thảo Vy

MSSV: 20120237

Email: 20120237@student.hcmus.edu.vn

II. Bảng báo cáo công việc

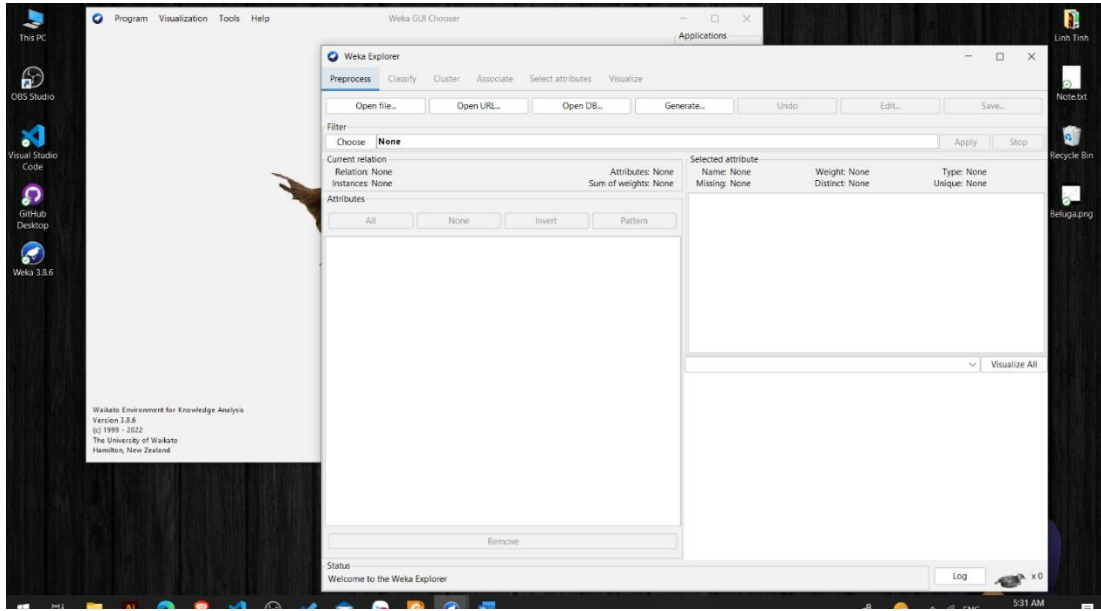
STT	Công việc		Mức độ hoàn thành
1	Install WEKA	After installing, you capture a screen that contains the “Explorer” function in your desktop background.	100%
2		Students open any data set (with extended part .arff). Explain the meaning of Current Relation, Attributes, and Selected Attribute in Preprocess tag. Briefly explain the meaning of the other tags in WEKA Explorer.	100%
3	Exploring Breast Cancer data set	How many instances does this data set have?	100%
4		How many attributes does this data set have?	100%
5		Which attribute is used for the label? Can it be changed? How?	100%
6		What is the meaning of each attribute?	100%
7		Let’s investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.	100%
8		Let’s propose solutions to the problem of missing values in the specific attribute.	100%
9		Let’s explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend.	100%
10	Exploring Weather data set	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?	100%
11		Let’s list five-number summary of two attributes temperature and humidity. Does WEKA provide these values?	100%
12		Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.	100%
13		Let’s move to the Visualize tag. What’s the name	100%

		of this chart? Do you think there are any pairs of different attributes that have correlated?	
14	Exploring Credit in Germany data set	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).	100%
15		Which attribute is used for the label?	100%
16		Let's describe the distribution of continuous attributes. (Left skewed or right skewed ?)	100%
17		Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.	100%
18		Let's move to the Select attributes tag. Describe all of the options for attribute selection.	100%
19		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)	100%
20	Preprocessing Data in Python	Extract columns with missing values	100%
21		Count the number of lines with missing data.	100%
22		Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute).	100%
23		Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).	100%
24		Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples).	100%
25		Delete duplicate samples.	100%
26		Normalize a numeric attribute using min-max and Z-score methods	100%
27		Performing addition, subtraction, multiplication, and division between two numerical attributes.	100%
28	Report	A lab report with format PDF	100%

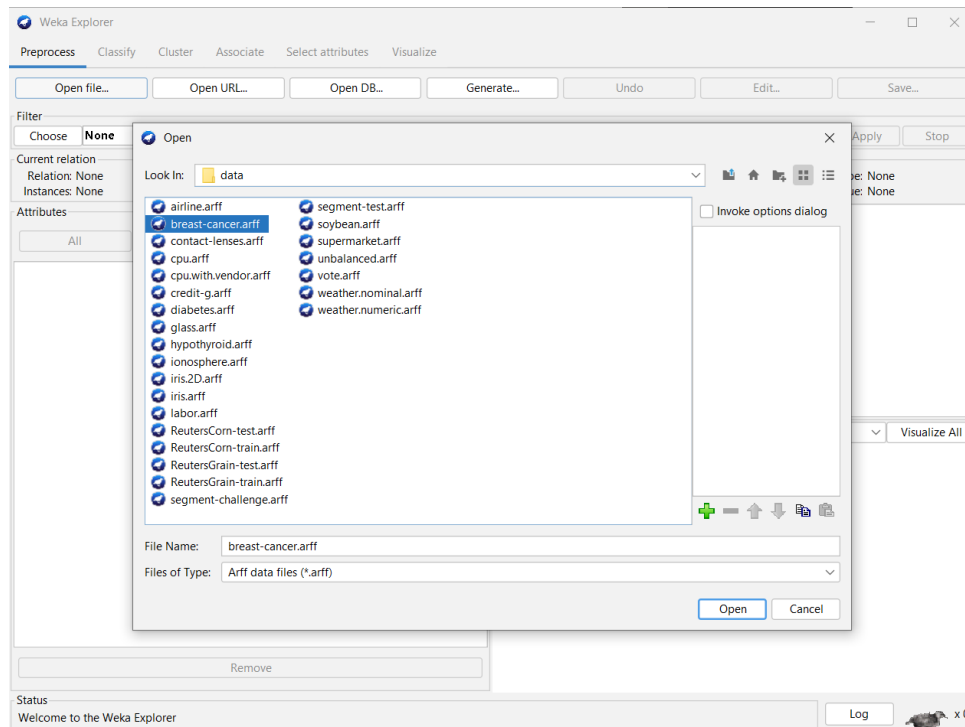
III. Nội dung bài làm

1. Cài đặt WEKA

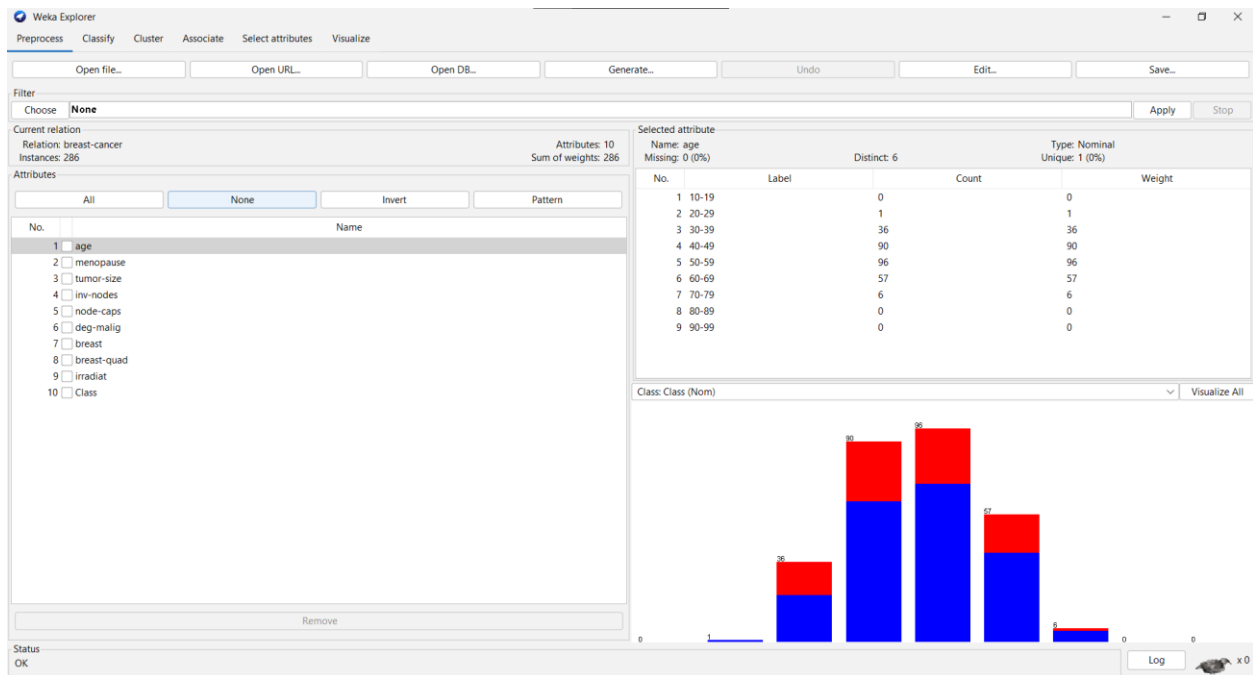
- Sau khi cài đặt WEKA, vào môi trường làm việc Explorer – đây là môi trường cho phép sử dụng tất cả khả năng của WEKA để khám phá dữ liệu.



- Mở 1 tập dữ liệu mẫu của WEKA (ở đây dùng tập dữ liệu breast-cancer.arff).



- Sau khi dữ liệu được load, ta được các thông tin như sau:



Ý nghĩa của các khu vực chức năng trong WEKA Explorer:

- Current Relation:**
 - Relation: tên của tập dữ liệu vừa được mở.
 - Instance và Sum of weights: Số các trường hợp/số mẫu trong tập dữ liệu.
 - Attributes: Số các thuộc tính trong tập dữ liệu.
- Attributes:** Ở đây có bốn nút và dưới chúng là danh sách các thuộc tính trong tập dữ liệu. Bốn nút được sử dụng để thay đổi lựa chọn các thuộc tính:
 - All: Tất cả các thuộc tính sẽ được chọn.
 - None: Bỏ chọn tất cả các thuộc tính.
 - Invert: Chuyển đổi trạng thái các thuộc tính đã chọn thành chưa chọn và ngược lại.
 - Pattern: Cho phép chọn các thuộc tính dựa trên công thức.
 - Dưới 4 nút chọn là bảng danh sách gồm 2 cột là số thứ tự và tên của từng thuộc tính.
- Selected attribute:** Khi nhấp chuột chọn vào các thuộc tính khác nhau trong danh sách thuộc tính của mục Attributes, các trường trong Selected attribute sẽ thay đổi.

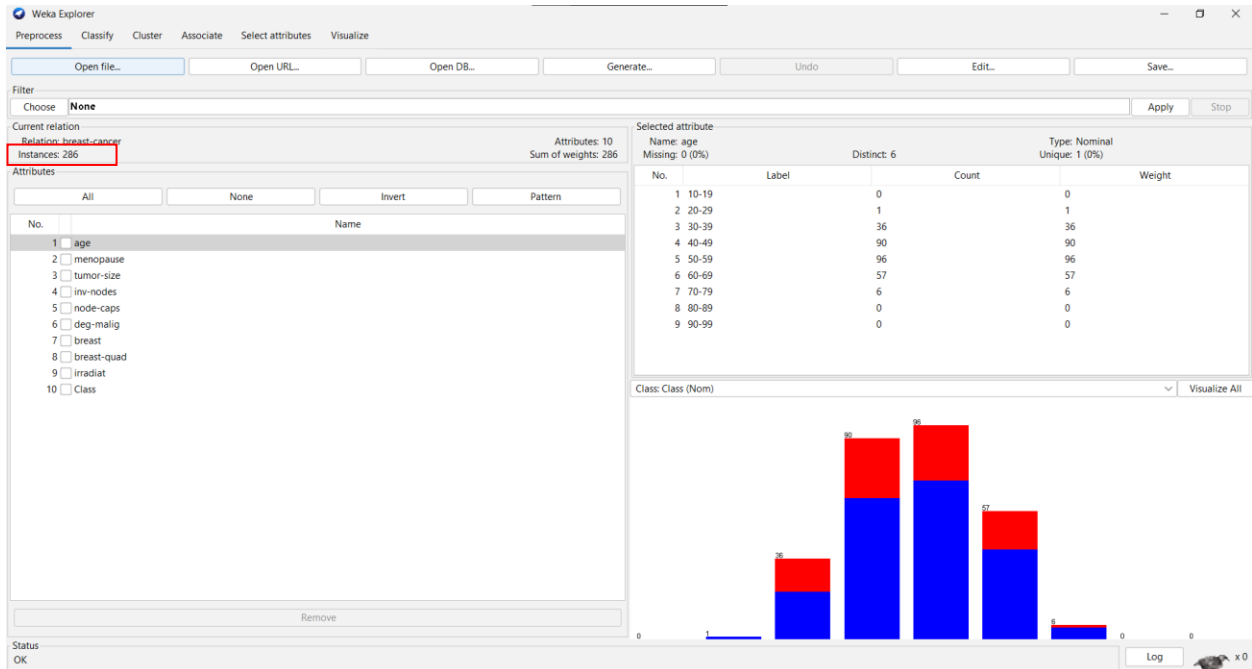
- Name: Tên của thuộc tính được chọn.
- Type: Kiểu của thuộc tính, thường là Nominal hoặc Numeric.
- Missing: Số và tỷ lệ phần trăm dữ liệu của thuộc tính này bị mất/thiếu.
- Distinct: Số lượng những giá trị ngoại lai chứa trong thuộc tính đã chọn.
- Unique: Số và tỷ lệ phần trăm các phiên bản có giá trị cho thuộc tính này mà không có phiên bản nào khác có trong dữ liệu.
- Dưới các trường là bảng số liệu với các cột: No (số thứ tự), Label (khoảng giá trị số và các giá trị khác nhau đối với dữ liệu chữ), Count và Weight (số lượng giá trị).
- Các mục khác trong tag Preprocess:
 - Open file: Mở file dữ liệu với định dạng .arff.
 - Open URL: Yêu cầu một địa chỉ URL dẫn đến vị trí đã lưu trữ dữ liệu.
 - Open DB...: Đọc dữ liệu từ một cơ sở dữ liệu (phải sử dụng file trong WEKA/experiment/DatabaseUtils.props trước khi thực hiện).
 - Generate...: Cho phép tạo dữ liệu ảo từ DataGenerators.
 - Undo: Hoàn tác thay đổi gần nhất với tập dữ liệu.
 - Edit: Mở tập dữ liệu hiện tại trong chế độ Viewing để chỉnh sửa.
 - Filter: chức năng lọc của WEKA. Click chuột vào nút Choose để chọn một bộ lọc thích hợp trong WEKA. Sau đó nhấn Apply và kiểm tra sự thay đổi của thuộc tính.
 - Class: Nó cũng được sử dụng như một lớp thuộc tính khi áp dụng bộ lọc để biểu diễn trực quan các giá trị.
 - Visualize All: Hiển thị tất cả biểu đồ của các thuộc tính trong tập dữ liệu trong một cửa sổ riêng.
 - Status: Hộp hiển thị các thông báo cho biết những gì đang diễn ra.
 - Log: Đây là nút để mở các bản ghi nhật ký ghi lại mọi hành động trong WEKA bao gồm cả các sự cố.
 - Hình ảnh một con chim: Là biểu tượng của WEKA. Biểu tượng 'x' bên cạnh biểu thị cho số quy trình đang chạy đồng thời. Con chim ngồi xuống nghĩa là không có tiến trình nào đang chạy và ngược lại.

2. Làm quen với WEKA

2.1. Khám phá tập dữ liệu Breast Cancer

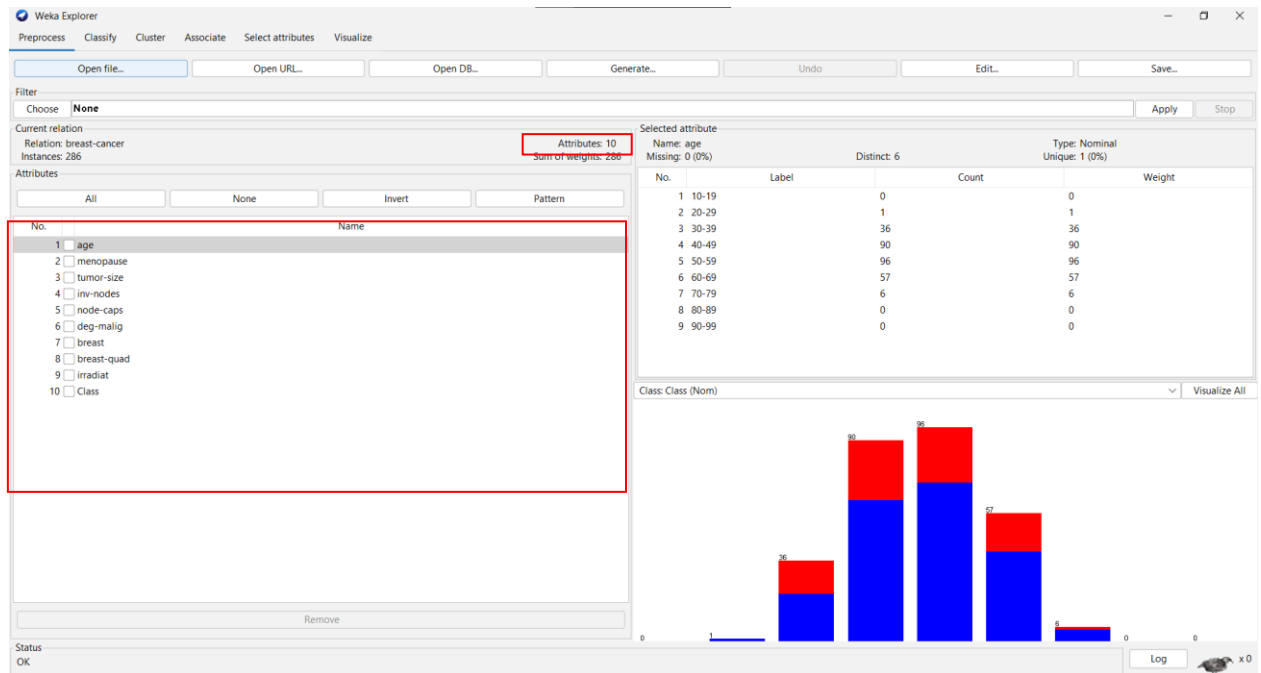
a. How many instances does this data set have?

Có 286 trường hợp trong tập dữ liệu.



b. How many attributes does this data set have?

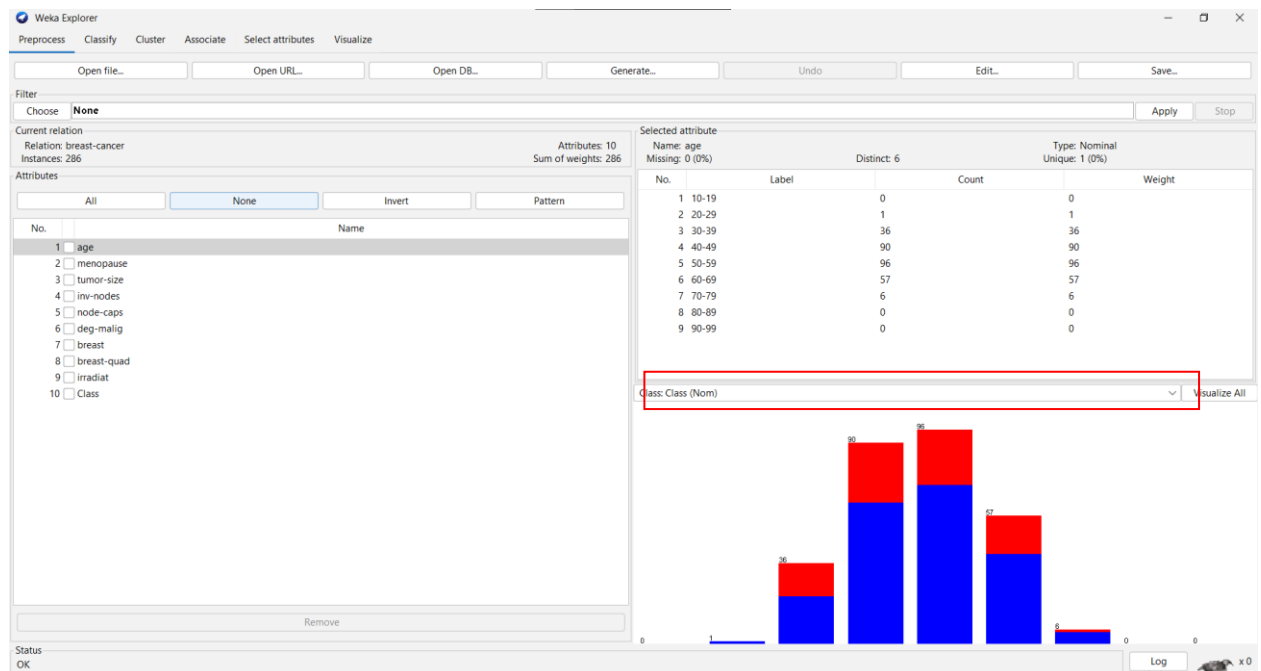
Có 10 thuộc tính trong tập dữ liệu.



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

Trong tập dữ liệu gốc thì label được định sẵn là thuộc tính Class.

Ta có thể xác định bằng cách xem thông qua bảng chọn cạnh nút Visualize All như sau (Mặc định là **Class: Class (Nom)**):



Hoặc nhấn vào Edit sẽ hiện ra cửa sổ Viewer để xác định label (cột có thuộc tính in đậm chính là label):

Viewer										
Relation: breast-cancer										
No.	1: age Nominal	2: menopause Nominal	3: tumor-size Nominal	4: inv-nodes Nominal	5: node-caps Nominal	6: deg-malig Nominal	7: breast Nominal	8: breast-quad Nominal	9: irradiat Nominal	10: Class Nominal
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurren...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recu...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurren...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recu...
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	recurren...
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	no-recu...
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recu...
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	no-recu...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recu...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recu...
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recu...
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
13	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recu...
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recu...
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurren...
16	30-39	premeno	20-24	0-2	no	3	left	central	no	no-recu...
17	50-59	premeno	10-14	3-5	no	1	right	left_up	no	no-recu...
18	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
19	50-59	premeno	40-44	0-2	no	2	left	left_up	no	no-recu...
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-recu...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurren...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-recu...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-recu...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-recu...

Có thể thay đổi label bằng cách thay đổi thuộc tính trong bảng chọn cạnh nút **Visualize All**. Khi đó label trong cửa sổ Viewer cũng thay đổi tương ứng:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Filter

Choose **None**

Apply Stop

Current relation

Relation: breast-cancer

Instances: 286

Attributes: 10

Sum of weights: 286

Attributes

All None Invert Pattern

No.

Name

1 ☐ age

2 ☐ menopause

3 ☐ tumor-size

4 ☐ inv-nodes

5 ☐ node-caps

6 ☐ deg-malig

7 ☒ breast

8 ☐ breast-quad

9 ☐ irradiat

10 ☐ Class

Remove

Selected attribute

Name: breast

Missing: 0 (0%)

Distinct: 2

Type: Nominal

Unique: 0 (0%)

No.	Label	Count	Weight
1	left	152	152
2	right	134	134

Class: breast (Nom)

No class

Class: age (Nom)

Class: menopause (Nom)

Class: tumor-size (Nom)

Class: inv-nodes (Nom)

Class: node-caps (Nom)

Class: deg-malig (Nom)

Class: breast (Nom)

Class: breast-quad (Nom)

Class: irradiat (Nom)

Class: Class (Nom)

Visualize All

Status

OK

Log

x 0

Viewer

Relation: breast-cancer

No.	1: age Nominal	2: menopause Nominal	3: tumor-size Nominal	4: inv-nodes Nominal	5: node-caps Nominal	6: deg-malign Nominal	7: breast Nominal	8: breast-quad Nominal	9: irradiat Nominal	10: Class Nominal
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurren...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recu...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurren...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recu...
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	recurren...
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	no-recu...
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recu...
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	no-recu...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recu...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recu...
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recu...
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
13	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recu...
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recu...
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurren...
16	30-39	premeno	20-24	0-2	no	3	left	central	no	no-recu...
17	50-59	premeno	10-14	3-5	no	1	right	left_up	no	no-recu...
18	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
19	50-59	premeno	40-44	0-2	no	2	left	left_up	no	no-recu...
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-recu...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurren...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-recu...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-recu...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-recu...

Add instance Undo OK Cancel

Nên chọn các lớp label là các thuộc tính thuộc dạng Nominal, có biến categorical vì nó sẽ cho thấy được mối quan hệ với các thuộc tính khác. Nó sẽ chứa các giá trị riêng biệt mà bạn muốn dự đoán dựa trên giá trị của những thuộc tính khác.

d. What is the meaning of each attribute?

- Age: Tuổi của bệnh nhân tại thời điểm chẩn đoán.
- Menopause (Mãn kinh): Bệnh nhân đang ở giai đoạn trước, trong hay sau mãn kinh tại thời điểm chẩn đoán.
- Tumor-size (Kích thước khối u): Đường kính lớn nhất (được tính bằng mm) của khối u được cắt bỏ.
- Inv-nodes: Số lượng hạch bạch huyết ở nách chứa tế bào ung thư vú di căn có thể nhìn thấy khi tiến hành kiểm tra mô học (số lượng trong khoảng 0 – 39).
- Node-caps (Mũ hạch): Nếu ung thư di căn đến hạch bạch huyết, mặc dù hạch nằm bên ngoài vị trí ban đầu của khối u, nhưng nó vẫn có thể bao bên ngoài vỏ hạch. Tuy nhiên, theo thời gian và tình trạng bệnh nặng hơn, hạch bạch huyết có thể bị thay thế bởi khối

u và sau đó xâm nhập vào viêm nang, cho phép nó xâm lấn đến các mô xung quang. Thuộc tính này cho thấy có hay không việc các khối u xâm nhập vào viêm nang.

- Deg-malig (Degree of malignancy): Mức độ ác tính của khối u (trong khoảng 1 – 3).
- Breast: Ung thư có thể xảy ra bên trái hay bên phải.
- Breast-quad (Breast quadrant): Vú có thể chia thành 4 phần tư và 1 vùng trung tâm tại núm vú. Thuộc tính cho thấy số lượng mắc phải thuộc các vùng nào.
- Irradiat (Irradiation): Số lượng bệnh nhân có hoặc không áp dụng phương pháp xạ trị.
- Class: Số lượng bệnh nhân có hay không có tái phát bệnh.

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

- Có 2 thuộc tính có dữ liệu bị thiếu:
 - Thuộc tính **node-caps** bị mất 8 dòng dữ liệu, chiếm 3%. Các dòng dữ liệu bị thiếu sẽ được bôi đen (ảnh chụp chỉ thể hiện 1 phần dữ liệu bị mất).

Selected attribute				
Name: node-caps				
Missing: 8 (3%)				
Distinct: 2				
Type: Nominal				
Unique: 0 (0%)				
No.	Label	Count	Weight	
1	yes	56	56	
2	no	222	222	

Relation: breast-cancer										
No.	1: age	2: menopause	3: tumor-size	4: inv-nodes	5: node-caps	6: deg-malig	7: breast	8: breast-quad	9: irradiat	10: Class
31	60-69	ge40	30-34	0-2	no	3	right	central	no	recurren...
32	60-69	ge40	25-29	3-5		1	right	left_low	yes	no-recu...
33	50-59	ge40	25-29	0-2	no	3	left	right_up	no	no-recu...
34	50-59	ge40	20-24	0-2	no	3	right	left_up	no	no-recu...
35	40-49	premeno	30-34	0-2	no	1	left	left_low	yes	recurren...
36	30-39	premeno	15-19	0-2	no	1	left	left_low	no	no-recu...
37	40-49	premeno	10-14	0-2	no	2	right	left_low	no	no-recu...
38	60-69	ge40	45-49	6-8	yes	3	left	central	no	no-recu...
39	40-49	ge40	20-24	0-2	no	3	left	left_low	no	no-recu...
40	40-49	premeno	10-14	0-2	no	1	right	right_low	no	no-recu...
41	30-39	premeno	35-39	0-2	no	3	left	left_low	no	recurren...
42	40-49	premeno	35-39	9-11	yes	2	right	right_up	yes	no-recu...
43	60-69	ge40	25-29	0-2	no	2	right	left_low	no	no-recu...
44	50-59	ge40	20-24	3-5	yes	3	right	right_up	no	recurren...
45	30-39	premeno	15-19	0-2	no	1	left	left_low	no	no-recu...
46	50-59	premeno	30-34	0-2	no	3	left	right_low	no	recurren...
47	60-69	ge40	10-14	0-2	no	2	right	left_low	yes	no-recu...
48	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recu...
49	50-59	premeno	50-54	0-2	yes	2	right	left_low	yes	no-recu...
50	50-59	ge40	40-44	0-2	no	3	right	left_low	no	no-recu...
51	70-79	ge40	15-19	9-11		1	left	left_low	yes	recurren...
52	50-59	lt40	30-34	0-2	no	3	right	left_low	no	no-recu...
53	40-49	premeno	0-4	0-2	no	3	left	central	no	no-recu...
54	70-79	ge40	40-44	0-2	no	1	right	right_low	no	no-recu...
55	40-49	premeno	25-29	0-2		2	left	right_low	yes	no-recu...
56	50-59	ge40	25-29	15-17	yes	3	right	left_low	no	no-recu...
57	50-59	premeno	20-24	0-2	no	1	left	left_low	no	no-recu...
58	50-59	ge40	35-39	15-17	no	3	left	left_low	no	no-recu...

- Thuộc tính breast-quad có 1 dòng dữ liệu bị thiếu và chỉ chiếm chưa đến 1%.

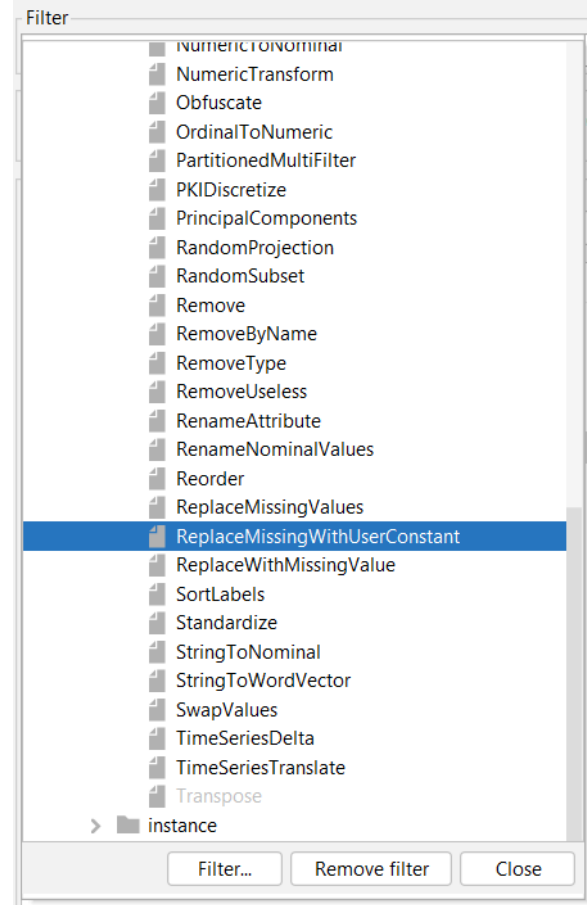
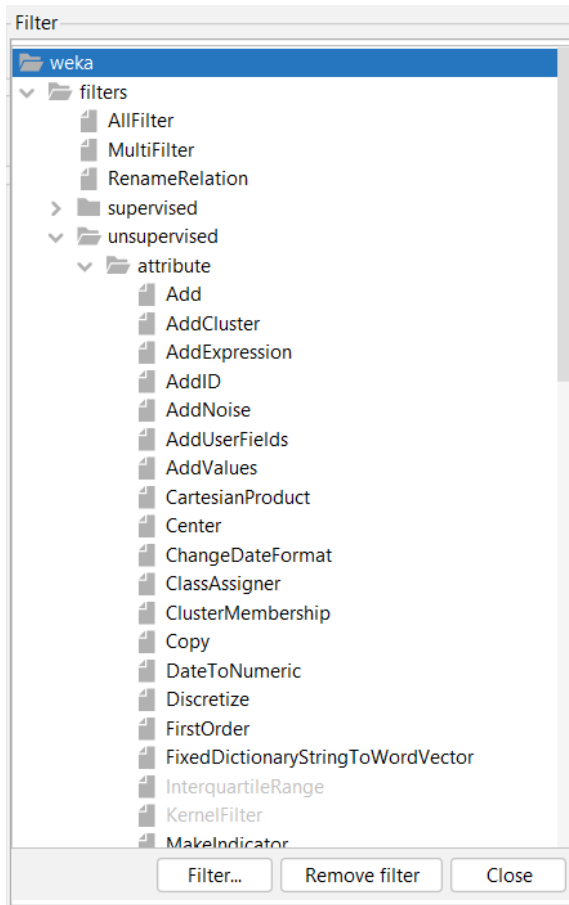
Selected attribute				
Name: breast-quad		Distinct: 5		Type: Nominal
Missing: 1 (0%)				Unique: 0 (0%)
No.	Label	Count	Weight	
1	left_up	97	97	
2	left_low	110	110	
3	right_up	33	33	
4	right_low	24	24	
5	central	21	21	

- Thông thường sẽ có 2 cách giải quyết dữ liệu bị mất:
 - Remove missing values: loại bỏ dữ liệu bị mất trong trường hợp dữ liệu đó không quan trọng hoặc số lượng dữ liệu bị mất quá ít (chiếm khoảng dưới 3% tổng số lượng mẫu).
 - Replace missing values: thay thế dữ liệu bị mất bằng một giá trị khác. Đối với trường hợp missing values là biến số thì có thể thay thế bằng các giá trị 0, mean, median, ... tùy vào trường hợp nhất định. Còn trong trường hợp missing values là biến categorical thì có thể nhóm chúng vào chung một nhóm và đặt tên nhóm ví dụ là Missing.

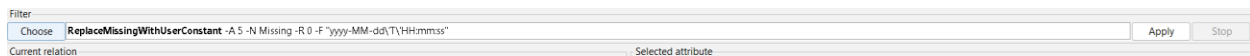
f. Let's propose solutions to the problem of missing values in the specific attribute.

Đối với thuộc tính node-caps, missing values chỉ chiếm 3% tỉ lệ nên có thể xử lý bằng cách loại bỏ dữ liệu. Do giá trị là biến categorical nên có thể nhóm chúng vào một nhóm. Ở đây em dùng cách nhóm missing values vào cùng một nhóm và đặt tên là missing.

Các bước thực hiện: Lần lượt nhấn nút Choose -> filters -> unsupervised -> attribute -> ReplaceMissingWithUserConstant.



Tiếp theo nhấn vào hộp filter để cài đặt:



Đặt các giá trị như sau: attributeIndices là **5** – số thứ tự của thuộc tính, nominalStringReplacementValue là **missing** – tên nhóm. Nhấn nút OK trên hộp thoại.

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant

About

Replaces all missing values for nominal, string, numeric and date attributes in the dataset with user-supplied constant values.

More

Capabilities

attributes 5

dateFormat yyyy-MM-dd'T'HH:mm:ss

dateReplacementValue

debug False

doNotCheckCapabilities False

ignoreClass False

nominalStringReplacementValue missing

numericReplacementValue 0

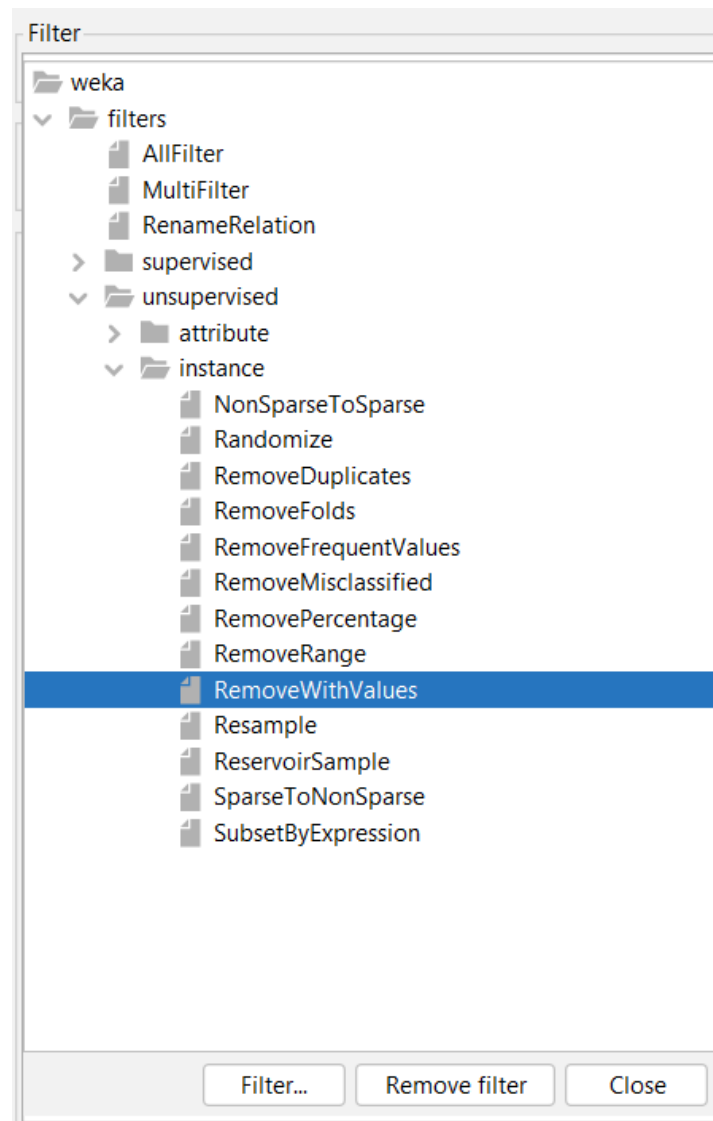
Open... Save... OK Cancel

Sau đó nhấn nút Apply và ta được kết quả:

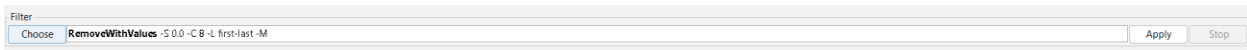
Selected attribute				
Name: node-caps		Distinct: 3		Type: Nominal
Missing: 0 (0%)				Unique: 0 (0%)
No.	Label	Count	Weight	
1	missing	8	8	
2	yes	56	56	
3	no	222	222	

Đối với thuộc tính breast-quad có missing values chiếm chưa đến 1% nên có thể loại bỏ nó.

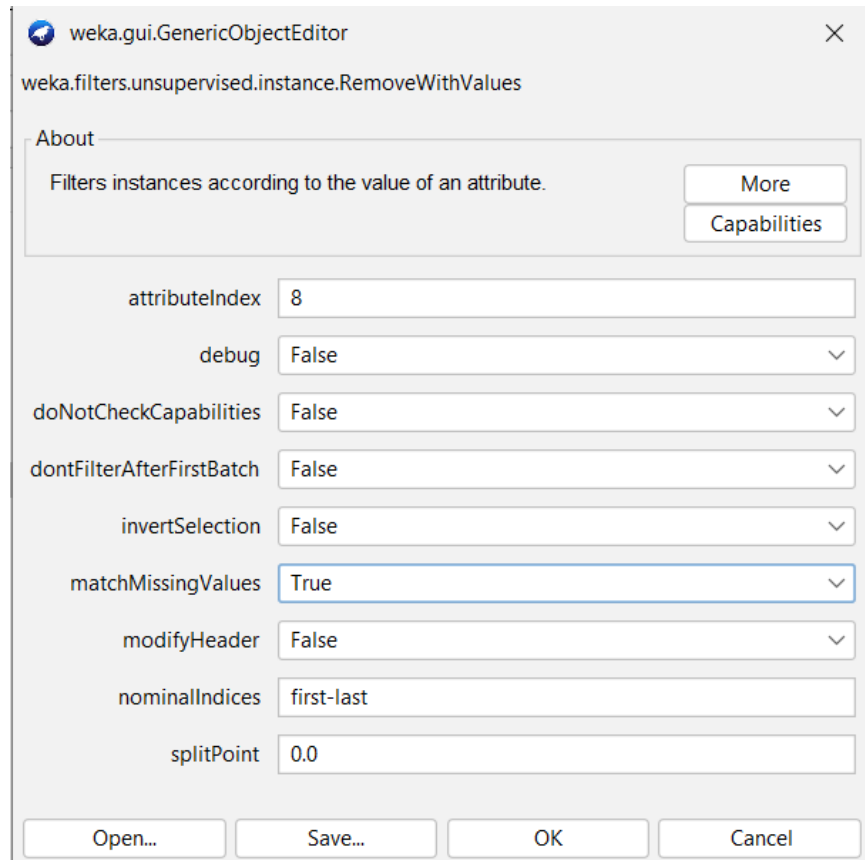
Các bước thực hiện: Nhấn chọn nút Choose -> filters ->unsupervised->instance->RemoveWithValues.



Tiếp theo nhấn vào hộp filter để cài đặt:



Đặt các giá trị như sau: attributeIndices là **8** – số thứ tự của thuộc tính, matchMissingValues là **True**. Nhấn nút OK trên hộp thoại.



weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.RemoveWithValues

About

Filters instances according to the value of an attribute. [More](#) [Capabilities](#)

attributeIndex: 8

debug: False

doNotCheckCapabilities: False

dontFilterAfterFirstBatch: False

invertSelection: False

matchMissingValues: True

modifyHeader: False

nominalIndices: first-last

splitPoint: 0.0

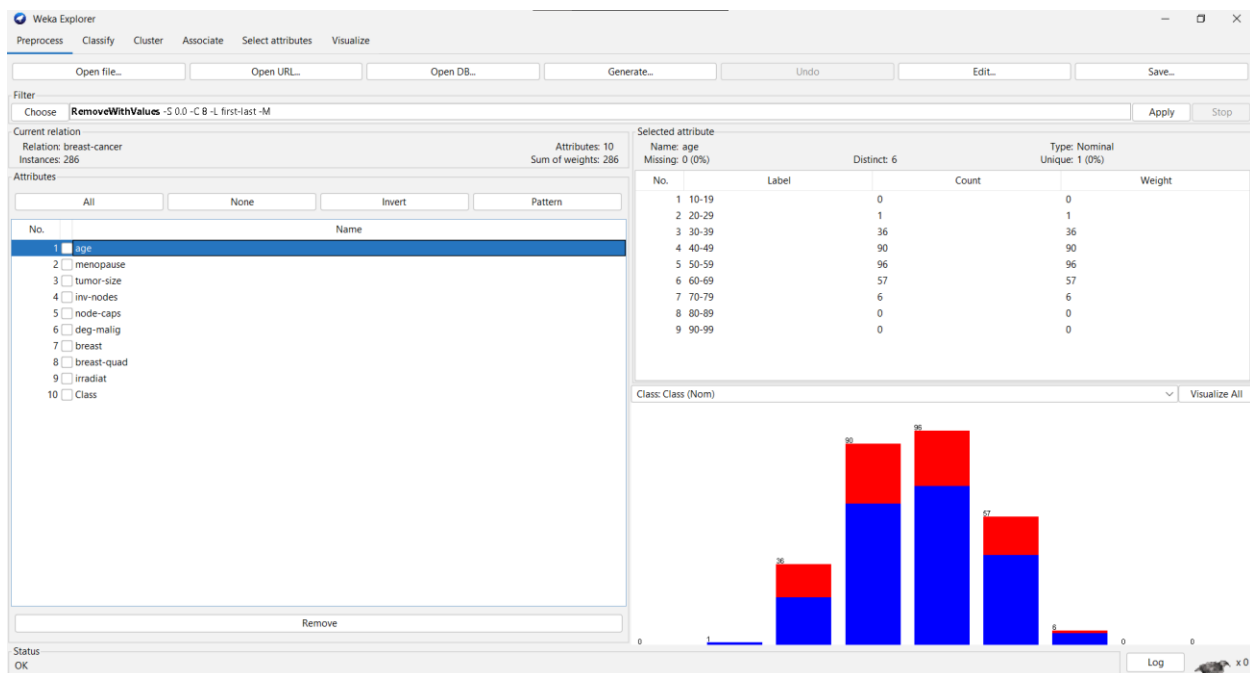
Open... Save... OK Cancel

Sau đó nhấn nút Apply và ta được kết quả:

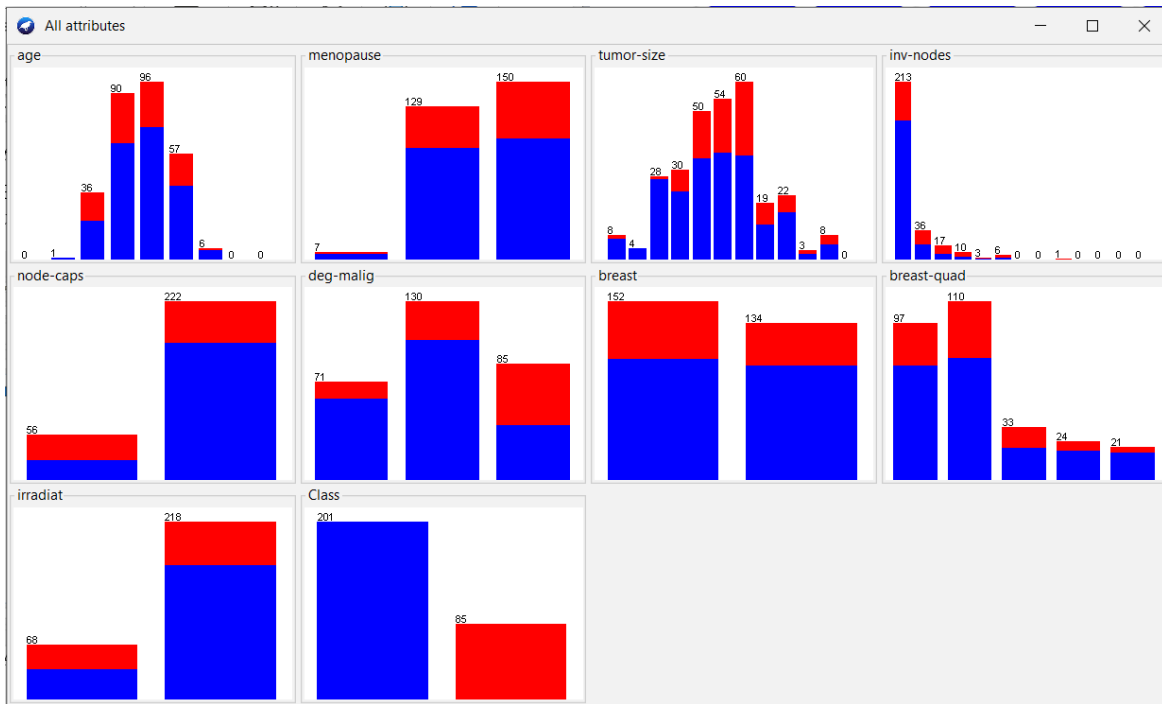
Selected attribute				
Name: breast-quad		Type: Nominal		
Missing: 0 (0%)		Unique: 0 (0%)		
No.	Label	Count	Weight	
1	left_up	0	0	
2	left_low	0	0	
3	right_up	0	0	
4	right_low	0	0	
5	central	0	0	

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

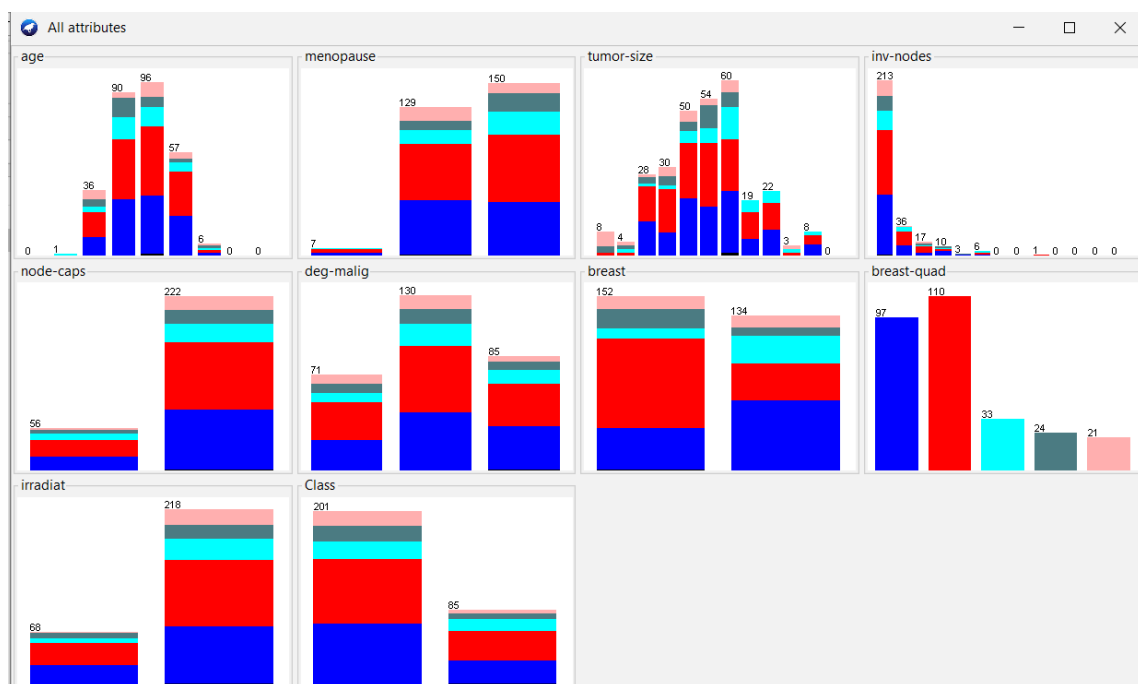
- Biểu đồ biểu diễn trong WEKA Explorer là mối quan hệ giữa một thuộc tính và thuộc tính lớp. Trong hình minh họa là thuộc tính **age** với **Class**.
- Biểu đồ của thuộc tính **Class** có 2 màu (xanh và đỏ) thể hiện 2 giá trị no-recurrence-events (màu xanh) và recurrence-events (màu đỏ). Ứng với mỗi thuộc tính trong tập dữ liệu (age, menopause, ...) thì các biểu đồ sẽ thể hiện mối quan hệ giữa thuộc tính đó với giá trị cần dự đoán trong thuộc tính **Class**. Đó là các biểu đồ **Histogram**.
- Ví dụ như hình dưới là: Biểu đồ phân bố có hay không có tái phát bệnh theo độ tuổi.



Khi dùng chức năng Visualize All ta có thể thấy được các biểu đồ phân bố của class theo từng thuộc tính:



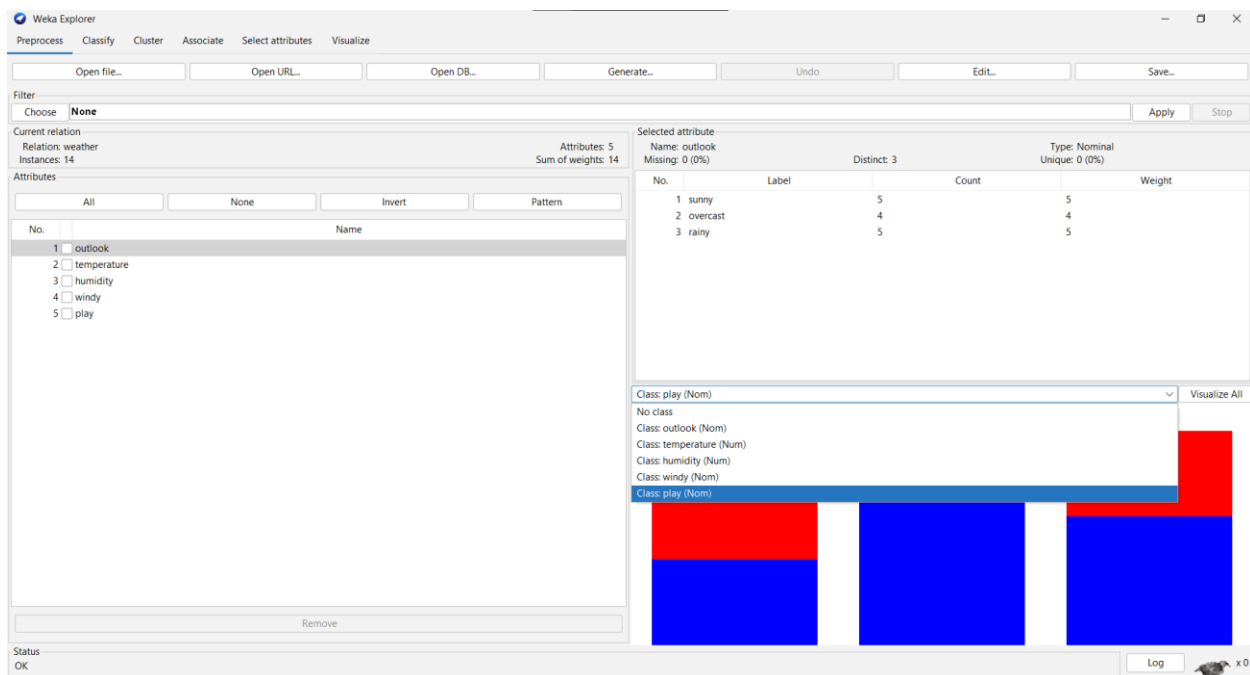
Nếu đổi thuộc tính **breast-quad** thành thuộc tính lớp thì ta sẽ có được các biểu đồ thể hiện mối quan hệ giữa các thuộc tính với 5 giá trị của breast-quad.



2.2. Khám phá tập dữ liệu Weather

a. *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?*

- Tập dữ liệu có 5 thuộc tính.
- Tập dữ liệu có 14 mẫu dữ liệu.
- Tập dữ liệu 3 thuộc tính thuộc dạng categorical: outlook, windy, và play.
- Tập dữ liệu có 2 thuộc tính thuộc dạng numerical: temperature và humidity.
- Có thể chọn thuộc các thuộc tính **outlook**, **windy**, **play** để làm label.



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

- Thuộc tính **temperature** có các giá trị:
 - Giá trị nhỏ nhất (Minimum): 64
 - Giá trị lớn nhất (Maximum): 85
 - Trung bình (Mean): 73.571
 - Độ lệch chuẩn (StdDev): 6.572

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	

- Thuộc tính **humidity** có các giá trị:
 - Giá trị nhỏ nhất (Minimum): 65
 - Giá trị lớn nhất (Maximum): 96
 - Trung bình (Mean): 81.643
 - Độ lệch chuẩn (StdDev): 10.285

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	

- Trong WEKA không cung cấp hết tất cả các số trong five-number summary, cụ thể là chỉ cung cấp min và max, còn lại Q1, median, và Q2 sẽ phải tự tìm. Ở đây em dùng hàm **describe()** trong thư viện **pandas** python để tính cho 2 thuộc tính trên:

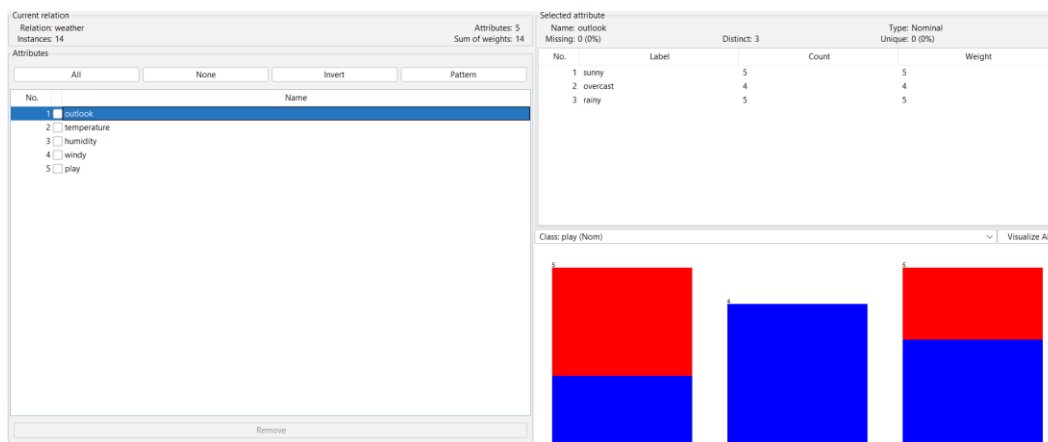
Vậy ta có kết quả như sau:

```
data_df[['temperature', 'humidity']].describe()
```

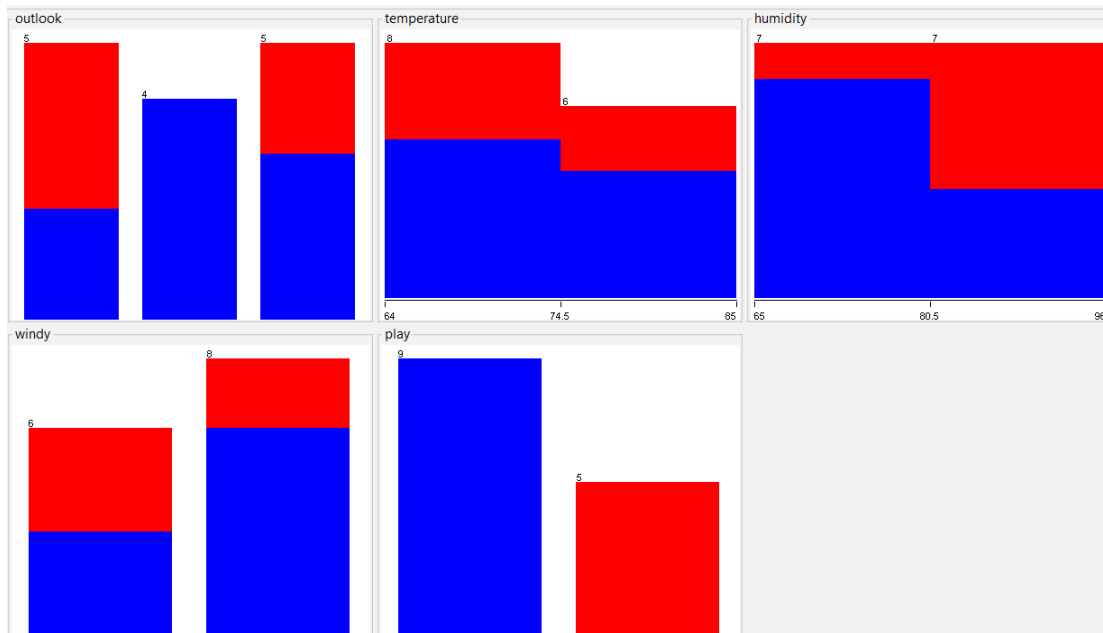
	temperature	humidity
count	14.000000	14.000000
mean	73.571429	81.642857
std	6.571667	10.285218
min	64.000000	65.000000
Q1	69.250000	71.250000
Median	72.000000	82.500000
Q2	78.750000	90.000000
max	85.000000	96.000000

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

- Thuộc tính class là play, được dùng để dự đoán xem ngày đó có thể đi chơi hay không (yes/no). Đồ thị trong cửa sổ Explorer thể hiện phân bố của kết quả cần dự đoán yes hoặc no của một thuộc tính cụ thể nào đó. Đó là các biểu đồ **Histogram**.
- Biểu đồ của thuộc tính **play** có 2 cột với 2 màu (xanh và đỏ) thể hiện 2 giá trị yes (màu xanh) và no (màu đỏ). Ứng với mỗi trường hợp trong tập dữ liệu (outlook, temperature, ...), ta có thể dùng để dự đoán liệu có nên đi chơi hay không.
- Như hình bên dưới là: **Biểu đồ phân bố dựa theo outlook.**

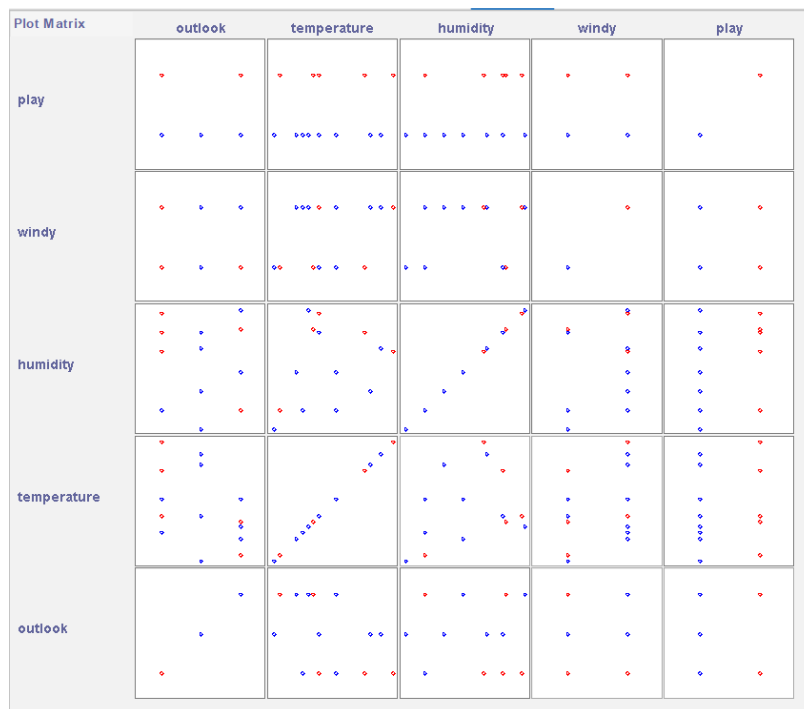


- Tương tự ở các biểu đồ khác.



d. 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?

Tên của biểu đồ trong Visualize tag là: **scatterplot matrix**.



Theo em, không có mối tương quan giữa các cặp thuộc tính khác nhau.

2.3. Khám phá tập dữ liệu Credit in Germany

a. *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).*

- Một file arff là một tệp văn bản ASCII mô tả danh sách các trường hợp của 1 tập hợp các thuộc tính trong một tập dữ liệu. Tệp arff có hai phần: Header và Data.
- Phần Header khai báo các relation và attribute. Nội dung phần comments ở phần Header trong credit-g.arff chứa thông tin mô tả chi tiết các trường dữ liệu của các thuộc tính.
- Tập dữ liệu có 1000 mẫu, và 21 thuộc tính.

Current relation

Relation: german_credit
Instances: 1000

Attributes: 21
Sum of weights: 1000

- Mô tả 5 thuộc tính bất kì:
 - Duration (liên tục): thời hạn trong 1 tháng.
 - Credit_amount (liên tục): số tiền tín dụng.
 - Checking_status (rời rạc): trạng thái hiện tại của các tài khoản, được tính theo khoảng lương trong ít nhất 1 năm. Các khoảng giá trị là <0 , $0 \leq x < 200$, ≥ 200 , và no checking. 'DM' viết trong file credit-g.arff chỉ một loại đơn vị tiền tệ, viết tắt của từ Deutsche Mark.
 - Purpose (rời rạc): mục đích sau khi mở tín dụng là new car (mua xe mới), used car (mua xe đã qua sử dụng), furniturre/equipment (trang thiết bị), radio/TV, domestic appliances (đồ gia dụng), repairs (sửa chữa), education (giáo dục), vacation (kì nghỉ), retraining (học lại), business (kinh doanh), others (những mục đích khác). Đây là những biến rời rạc trong thuộc tính purpose.

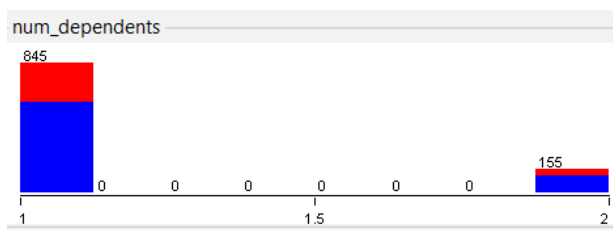
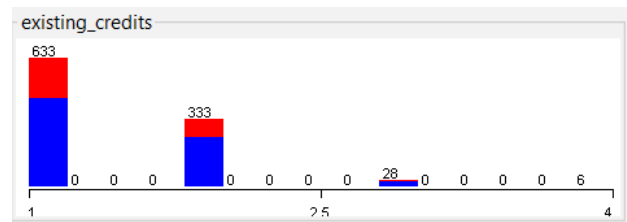
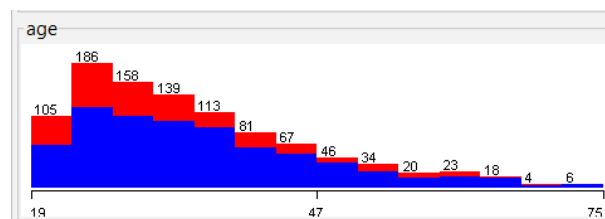
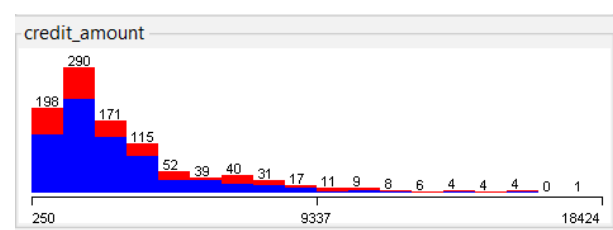
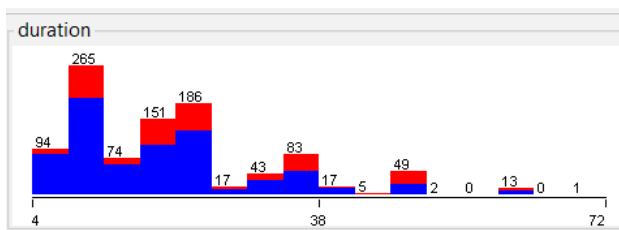
- Personal_status (rời rạc): tình trạng hôn nhân và giới tính. Các trường hợp là: male div/sep (nam đã ly hôn/ly thân), female div/dep/mar (nữ đã ly hôn/ly thân/kết hôn), male single (nam còn độc thân), male mar/wid (nam đã kết hôn và mất vợ), female single (nữ độc thân). Đây là những biến rời rạc trong thuộc tính personal_status.

b. Which attribute is used for the label?

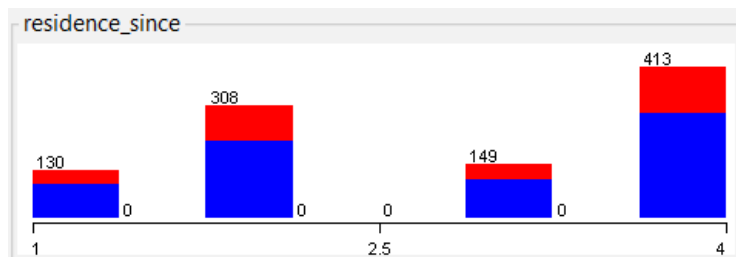
Các thuộc tính được dùng cho label là *checking_status*, *credit_history*, *purpose*, *savings_status*, *employment*, *personal_status*, *other_parties*, *property_magnitude*, *other_payment_plants*, *housing*, *job*, *own_telephone*, *foreign_worker*, *class*.

c. Let's describe the distribution of continuous attributes. (Left skewed or right skewed)

- *duration*, *credit_amount*, *age*, *existing_credits*, *num_dependents* có phân phối lệch trái.

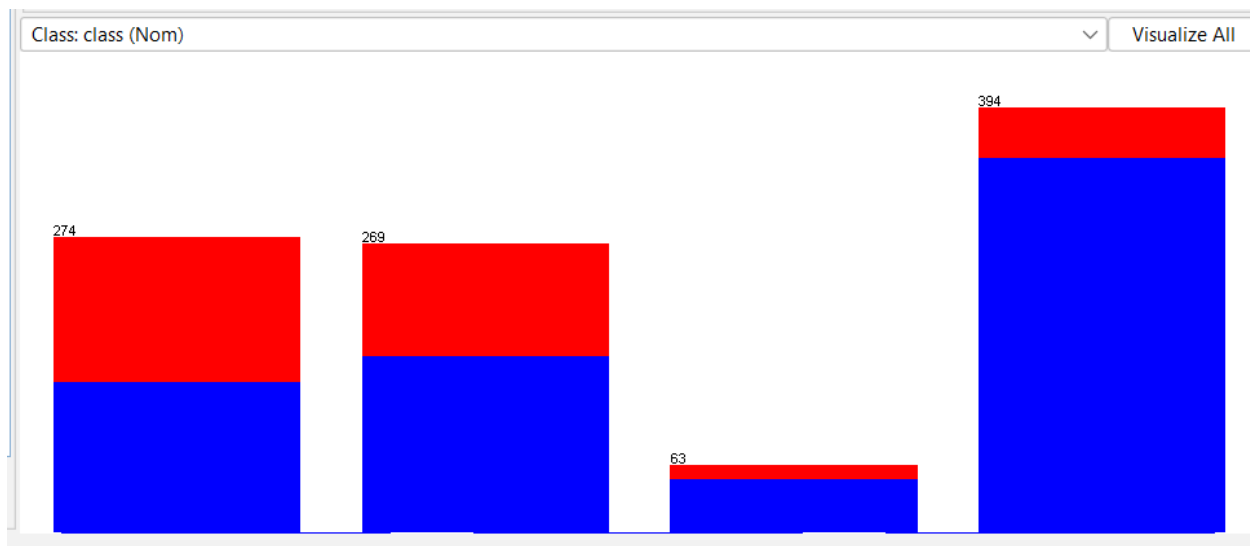


- *Residence_since* có phân phối lệch phải.

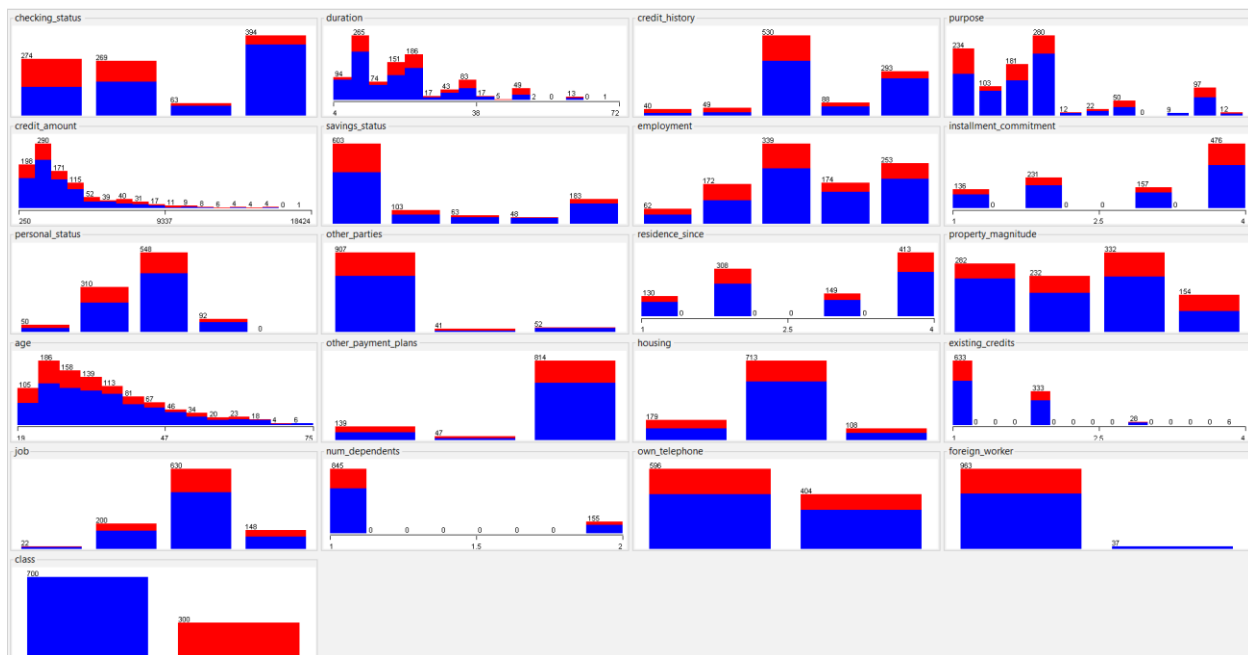


d. *Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.*

- Thuộc tính class được dùng để dự đoán xem người khách hàng sẽ thuộc diện good hay bad. Đồ thị trong cửa sổ Explorer thể hiện phân bố của kết quả cần dự đoán dựa theo một thuộc tính cụ thể nào đó. Đó là các biểu đồ **Histogram**.
- Biểu đồ của thuộc tính **class** có 2 cột với 2 màu (xanh và đỏ) thể hiện 2 giá trị good (màu xanh) và bad (màu đỏ). Ứng với mỗi label trong tập dữ liệu (checking_status, credit_history, ...), ta có thể dùng để dự đoán về người khách hàng.
- Như hình bên dưới là: Biểu đồ phân bố dựa theo thuộc tính checking_status.



Tương tự, ta có các biểu đồ sau:



e. Let's move to the Select attributes tag. Describe all the options for attribute selection.

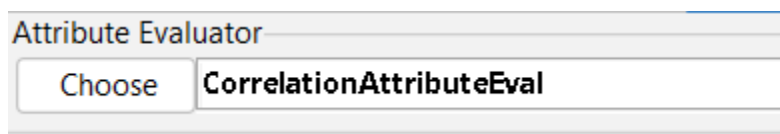
Trong tag Selection attributes có 2 phần tính năng: Attribute Evaluator và Search Method. Mỗi phần có nhiều lựa chọn khác nhau.

- Attributes Evaluator:
 - CfsSubsetEval: Đánh giá tập hợp con thuộc tính CFS.
 - ClassifierAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách sử dụng bộ phân loại do người dùng chỉ định
 - ClassifierSubsetEval: Sử dụng bộ phân loại để ước tính “giá trị” của một tập các thuộc tính.
 - CorrelationAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo mức độ tương quan giữa thuộc tính và class.
 - GainRatioAttributeEval: Đánh giá các thuộc tính riêng lẻ bằng cách đo tỉ lệ khuếch đại đối với class.
 - InfoGainAttributeEval: Đánh giá các thuộc tính riêng lẻ bằng cách đo mức độ thu được thông tin đối với class.

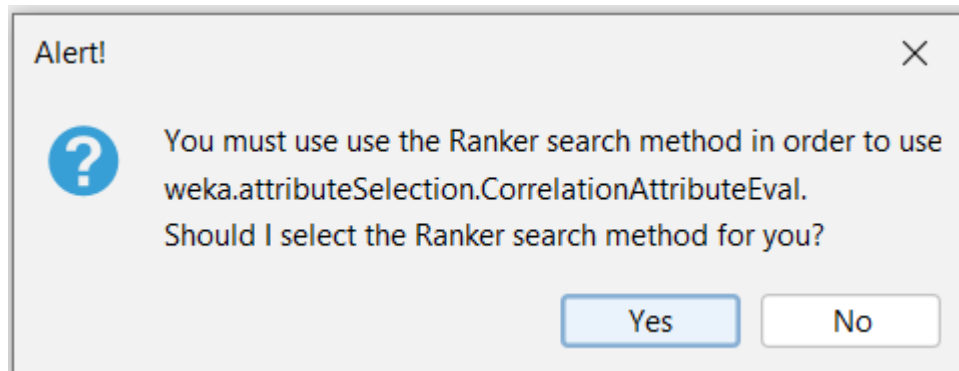
- OneRAttributeEval: Đánh giá các thuộc tính riêng lẻ bằng cách sử dụng trình phân loại OneR.
- PrincipalComponents: Thực hiện phân tích/chuyển đổi component chính.
- ReliefFAttributeEval: Đánh giá các thuộc tính riêng lẻ bằng cách sử dụng Relief.
- SymmetricalUncertAttributeEval: Đánh giá các thuộc tính riêng lẻ bằng cách đo độ không đối xứng với class.
- WrapperSubsetEval: Đánh giá tập hợp con thuộc tính wrapper.
- Search Method:
 - BestFirst: Tìm kiếm không gian của tập con các thuộc tính bằng phương pháp leo đồi tham lam được tăng cường bằng cơ sở quay lui.
 - GreedyStepwise: Thực hiện tìm kiếm tiến hoặc lùi trong không gian của tập hợp con các thuộc tính.
 - Ranker: Xếp hạng các thuộc tính dựa theo đánh giá riêng từng thuộc tính. Được sử dụng kết hợp với các bộ đánh giá thuộc tính (ReliefF, GainRatio, Entropy, ...).

f. 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)

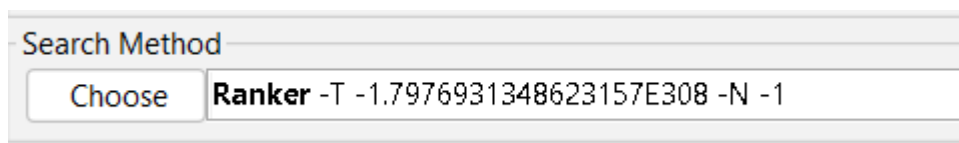
- Phương pháp được dùng là **CorrelationAttributeEval**.
- Các bước thực hiện:
 - Vào Select attributes tag, tại phần Attribute Evaluator, nhấn Choose rồi chọn CorrelationAttributeEval.



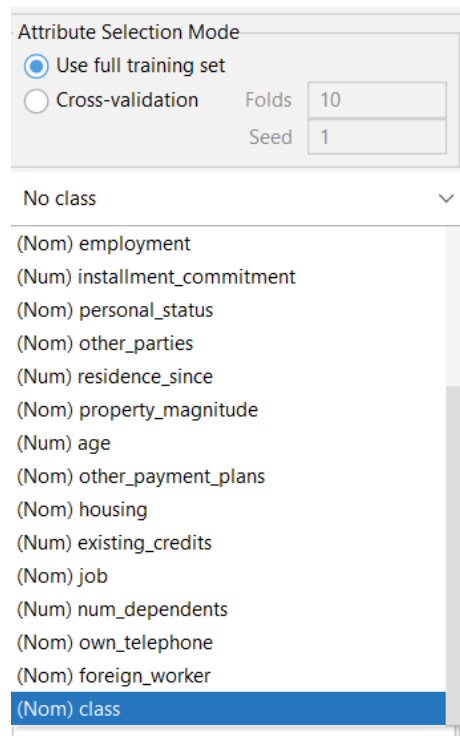
- Một hộp thoại hiện ra gợi ý sử dụng phương pháp tìm kiếm Ranker để có thể dùng lớp `weka.attributeSelection.CorrelationAttributeEval`. Nhấn Yes.



Search Method đã được cài sang Ranker.



- Tại phần Attribute Selection Mode, lúc này bảng chọn phía dưới đang là ‘No class’, chọn lại ‘(Nom) class’. Sau đó nhấn nút Start.



- Tại phần Attribute selection output, ta được kết quả 5 thuộc tính có giá trị đánh giá cao nhất lần lượt là: checking_status, duration, credit_amount, savings_status, housing.

Attribute selection output

```
Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
    Correlation Ranking Filter
Ranked attributes:
0.23276    1 checking_status
0.21493    2 duration
0.15474    5 credit_amount
0.13162    6 savings_status
0.12138   15 housing
0.108      14 other_payment_plans
0.09113    13 age
0.08988    3 credit_history
0.08208   20 foreign_worker
0.07494    4 purpose
0.0724     8 installment_commitment
0.07192    9 personal_status
0.05838   12 property_magnitude
0.0527     7 employment
0.04573   16 existing_credits
0.03647   19 own_telephone
0.01904   17 job
0.00612   10 other_parties
0.00301   18 num_dependents
0.00297   11 residence_since

Selected attributes: 1,2,5,6,15,14,13,3,20,4,8,9,12,7,16,19,17,10,18,11 : 20
```

3. Tiền xử lý dữ liệu với Python

Cú pháp nhập command line arguments của từng function:

- *Function 1: python preprocessing.py <datafilename.csv> <function>*

Vd: python preprocessing.py house-prices.csv function1

```
PS D:\HCMUS\HK6\CTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv function1
['LotFrontage', 'Alley', 'MasVnrType', 'MasVnrArea', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'FireplaceQu', 'GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageQu', 'GarageCond', 'PoolQC', 'Fence', 'MiscFeature']
```

- Function 2: `python preprocessing.py <datafilename.csv> <function>`

Vd: `python preprocessing.py house-prices.csv function2`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv function2
LotFrontage      173
Alley            941
MasVnrType       593
MasVnrArea       10
BsmtQual         27
BsmtCond         27
BsmtExposure     28
BsmtFinType1     27
BsmtFinType2     29
FireplaceQu      501
GarageType       60
GarageYrBlt      60
GarageFinish     60
GarageQual       60
GarageCond       60
PoolQC          1000
Fence            815
MiscFeature      963
dtype: int64
```

- Function 3: `python preprocessing.py <datafilename.csv> <function> --m <'mean'/'median'/'mode'> --out <outputfilename.csv>.`

Vd: `python preprocessing.py house-prices.csv func3 --m median --out function3.csv`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv func3 --m median --out function3.csv
Saved to function3.csv
```

- Function 4: `python preprocessing.py <datafilename.csv> <function> --x <percent> --out <outputfilename.csv>.`

Vd: `python preprocessing.py house-prices.csv fun4 --x 40 --out function4.csv`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv fun4 --x 40 --out function4.csv
Saved to function4.csv
```

Test với dữ liệu bị thiếu ở nhiều cột:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallScore	YearBuilt	YearRemo	RoofStyle	RoofMatl	Ex
1242	20 RL		83	9849 Pave			Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vn
484	120 RM		32	4500 Pave			Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Vn
392	60 RL		71	12209 Pave			IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vn
730	30 RM		52	6240 Pave		Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	Mi
255	20 RL		70	8400 Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	Mi
1094	20 RL		71	9230 Pave			Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	Mi
1021	20 RL		60	7024 Pave			Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vn
1341	20 RL		70	8294 Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	Mi
1025	20 RL			15498 Pave			IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	Str
848	20 RL		36	15523 Pave			IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Hc
457	70 RM		34	4571 Pave		Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	As
1266	160 FV		35	3735 Pave			Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	Mi
695	50 RM		51	6120 Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	W
24	120 RM		44	4224 Pave			Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg	Ce

Ta được kết quả, có thể thấy các dòng dữ liệu bị mất nhiều hơn 40% số thuộc tính đã bị xóa mất.

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallScore	YearBuilt	YearRemo	RoofStyle	RoofMatl	Ex
0	1242	20 RL		83	9849 Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Cc
7	484	120 RM		32	4500 Pave		Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Cc
8	392	60 RL		71	12209 Pave		IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Cc
9	730	30 RM		52	6240 Pave		Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	Cc
10	255	20 RL		70	8400 Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	Cc
11	1094	20 RL		71	9230 Pave		Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	Cc
12	1021	20 RL		60	7024 Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Cc
13	1341	20 RL		70	8294 Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	Cc
14	1025	20 RL			15498 Pave		IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	W
15	848	20 RL		36	15523 Pave		IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Cc
16	457	70 RM		34	4571 Pave		Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	Cc
17	1266	160 FV		35	3735 Pave		Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	Cc
18	695	50 RM		51	6120 Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	Cc
19	24	120 RM		44	4224 Pave		Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg	Cc
20	1314	60 RL		108	14774 Pave		IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg	Cc

- Function 5: `python preprocessing.py <datafilename.csv> <function> --x <percent> --out <outputfilename>.csv.`

Vd: `python preprocessing.py house-prices.csv func5 --x 50 --out function5.csv`

```
PS D:\HOMUS\HK6\KTDLLD\Lab01\Lab01\source> python preprocessing.py house-prices.csv func5 --x 50 --out function5.csv
Saved to function5.csv
```

Test thử với các cột mất hơn 50% dòng dữ liệu (trong hình chỉ thể hiện 1 phần):

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallScore	YearBuilt	YearRemo	RoofStyle	RoofMatl	Ex
1242	20 RL		83	9849 Pave			Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vn
1233	90 RL		70	9842 Pave			Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc
1401	50 RM		50	6000 Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W
1377	30 RL		52	6292 Pave			Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W
208	20 RL			12493 Pave							Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W
1392	90 RL		65	8944 Pave							Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Ph
980	20 RL		80	8816 Pave							Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vn
484	120 RM		32	4500 Pave							Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Vn
392	60 RL		71	12209 Pave							Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vn
730	30 RM		52	6240 Pave		Grvl					Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	Mi
255	20 RL		70	8400 Pave							Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	Mi
1094	20 RL		71	9230 Pave							Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	Mi
1021	20 RL		60	7024 Pave							Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vn
1341	20 RL		70	8294 Pave							Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	Mi
1025	20 RL			15498 Pave							Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	Str
848	20 RL		36	15523 Pave							Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Hc
457	70 RM		34	4571 Pave		Grvl					Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	As

Ta được kết quả: các cột thuộc tính LotShape, LandContour, Utilities, LotConfig đã bị xóa.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RoofStyle	RoofMatl	Exterior1st	Exterior2nd	MasVnrArea	ExterQual	Ex
0	1242	20	RL	83	9849	Pave	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	VinylSd	VinylSd	0	Gd	TA
1	1233	90	RL	70	9842	Pave	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	HdBoard	HdBoard	0	TA	TA
2	1401	50	RM	50	6000	Pave	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	WdShing	WdShng	0	TA	TA
3	1377	30	RL	52	6292	Pave	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	WdShng	WdShng	0	TA	TA
4	208	20	RL		12493	Pave	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	WdShng	WdShng	0	TA	TA
5	1392	90	RL	65	8944	Pave	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Plywood	Plywood	0	TA	TA
6	980	20	RL	80	8816	Pave	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	VinylSd	VinylSd	0	TA	TA
7	484	120	RM	32	4500	Pave	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	VinylSd	VinylSd	116	TA	TA
8	392	60	RL	71	12209	Pave	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	VinylSd	VinylSd	0	TA	TA
9	730	30	RM	52	6240	Pave	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	MetalSd	MetalSd	0	TA	TA
10	255	20	RL	70	8400	Pave	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	MetalSd	MetalSd	0	TA	Gc
11	1094	20	RL	71	9230	Pave	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	MetalSd	MetalSd	166	TA	TA
12	1021	20	RL	60	7024	Pave	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	VinylSd	VinylSd	0	TA	TA
13	1341	20	RL	70	8294	Pave	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	MetalSd	MetalSd	0	TA	TA
14	1025	20	RL		15498	Pave	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	Stone	HdBoard	0	Gd	TA

- Function 6: `python preprocessing.py <datafilename.csv> <function> --out <outputfilename.csv>`.

Vd: `python preprocessing.py house-prices.csv func6 --out function6.csv`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv func6 --out function6.csv
  Id  MSSubClass  MSZoning  LotFrontage  LotArea  Street  Alley  LotShape  ...  Fence  MiscFeature  MiscVal  MoSold  YrSold  SaleType  SaleCondition  SalePrice
0    1242         20      RL          83.0    9849   Pave   NaN      Reg  ...   NaN          NaN          0         6    2007      New      Partial      248328
1    1233         90      RL          70.0    9842   Pave   NaN      Reg  ...   NaN          NaN          0         3    2007      WD      Normal      101800
2    1401         50      RM          50.0    6000   Pave   NaN      Reg  ...   NaN          NaN          0         7    2008      WD      Normal      120000
3    1377         30      RL          52.0    6292   Pave   NaN      Reg  ...   NaN          NaN          0         4    2008      WD      Normal      91000
4     208         20      RL          NaN    12493   Pave   NaN      IR1  ...   GdLw          NaN          0         4    2008      WD      Normal      141000
..    ...         ...      ...          ...      ...   ...   ...      ...  ...   ...          ...          ...      ...      ...      ...      ...
986   985         90      RL          75.0    10125  Pave   NaN      Reg  ...   NaN          NaN          0         8    2009      COD      Normal      126000
989   582         20      RL          98.0    12704  Pave   NaN      Reg  ...   NaN          NaN          0         8    2009      New      Partial      253293
992   668         20      RL          65.0    8125   Pave   NaN      Reg  ...   NaN          NaN          0        10    2008      WD      Normal      193500
995  1190         60      RL          60.0    7500   Pave   NaN      Reg  ...   NaN          NaN          0         6    2010      WD      Normal      189000
996   192         60      RL          NaN    7472   Pave   NaN      IR1  ...   NaN          NaN          0         6    2007      WD      Normal      184000

[716 rows x 81 columns]
Saved to function6.csv
```

Sau khi xóa dữ liệu, tiến hành in ra màn hình thì thấy đã số dòng dữ liệu đã giảm đi do đã xóa bỏ những dòng bị trùng lặp.

- Function 7: `python preprocessing.py <datafilename.csv> <function> --col <column>`.

Vd: `python preprocessing.py house-prices.csv func7 --col SalePrice`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv func7 --col SalePrice
  Min-max  Z-score
0    0.369599  0.875745
1    0.115363 -0.951880
2    0.146941 -0.724874
3    0.096624 -1.086587
4    0.183378 -0.462944
..    ...      ...
995   0.266661  0.135754
996   0.257986  0.073390
997   0.280542  0.235537
998   0.521716  1.969267
999   0.166895 -0.581436

[716 rows x 2 columns]
```

- Function 8: `python preprocessing.py <datafilename.csv> <function> --cal <calculation> --col1 <column1> --col2 <column2>`

Vd: `python preprocessing.py house-prices.csv function8 --cal sub --col1 YearRemodAdd --col2 YearBuilt`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py house-prices.csv function8 --cal sub --col1 YearRemodAdd --col2 YearBuilt
0      0
1      0
2     21
3     20
4      0
..
995    0
996    32
997    0
998    1
999    0
Name: YearRemodAdd sub YearBuilt, Length: 1000, dtype: int64
```

Lưu ý:

- Khi người dùng không rõ cú pháp cần nhập cho từng thuộc tính thì có thể gõ dòng lệnh sau để hiển thị cú pháp của toàn bộ các chức năng:

`python preprocessing.py -h` hoặc `python preprocessing.py --help`

```
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py --h
Function 1: python preprocessing.py <datafilename.csv> <function>
Function 2: python preprocessing.py <datafilename.csv> <function>
Function 3: python preprocessing.py <datafilename.csv> <function> --m <'mean'/'median'/'mode'> --out <outputfilename.csv>
Function 4: python preprocessing.py <datafilename.csv> <function> --x <percent> --out <outputfilename.csv>
Function 5: python preprocessing.py <datafilename.csv> <function> --x <percent> --out <outputfilename>.csv
Function 6: python preprocessing.py <datafilename.csv> <function> --out <outputfilename.csv>
Function 7: python preprocessing.py <datafilename.csv> <function> --col <column>
Function 8: python preprocessing.py <datafilename.csv> <function> --cal <calculation> --col1 <column1> --col2 <column2>
PS D:\HCMUS\HK6\KTDLUD\Lab01\Lab01\source> python preprocessing.py --help
Function 1: python preprocessing.py <datafilename.csv> <function>
Function 2: python preprocessing.py <datafilename.csv> <function>
Function 3: python preprocessing.py <datafilename.csv> <function> --m <'mean'/'median'/'mode'> --out <outputfilename.csv>
Function 4: python preprocessing.py <datafilename.csv> <function> --x <percent> --out <outputfilename.csv>
Function 5: python preprocessing.py <datafilename.csv> <function> --x <percent> --out <outputfilename>.csv
Function 6: python preprocessing.py <datafilename.csv> <function> --out <outputfilename.csv>
Function 7: python preprocessing.py <datafilename.csv> <function> --col <column>
Function 8: python preprocessing.py <datafilename.csv> <function> --cal <calculation> --col1 <column1> --col2 <column2>
```

- Người dùng có thể nhập `function1` hoặc `func1` để thực hiện chức năng tương ứng.

IV. Tài liệu tham khảo

1. <http://people.sabanciuniv.edu/berrin/cs512/hws/hw1/WEKA%20Explorer%20Tutorial-REFERENCE.pdf>
2. <https://www.linkedin.com/pulse/using-machine-learning-techniques-predict-recurrence-breast-alva/>
3. <https://machinelearningmastery.com/how-to-handle-missing-values-in-machine-learning-data-with-weka/>
4. <https://www.slideshare.net/HoQuangThanh/la-chn-thuc-tnh-v-khai-ph-lut-kt-hp-trn-weka>
5. <http://bio.med.ucm.es/docs/weka/weka/core/OptionHandler.html>
6. <https://weka.sourceforge.io/doc.dev/weka/attributeSelection/ASSearch.html>
7. <https://www.geeksforgeeks.org/data-normalization-in-data-mining/>