

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TPHCM

Khoa Công nghệ thông tin

# KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

## BÀI TẬP 01

Tiền xử lý dữ liệu  
Preprocessing

Giảng viên hướng dẫn: Nguyễn Ngọc Thảo

Người thực hiện: Nhóm 37

Vũ Mạnh Hùng -18424029

Hà Tiến Đạt – 18424023

TP.HCM – 09/2019

## Mục lục

I)	Nội dung thực hiện báo cáo viết (25 điểm) .....	2
1)	Tích hợp dữ liệu (integration) (5 điểm) .....	2
2)	Tóm tắt dữ liệu mô tả (descriptive data summarization) (8.0 điểm) .....	7
3)	Chọn lọc dữ liệu (selection) (3.0 điểm) .....	19
4)	Làm sạch dữ liệu (cleaning) (5.0 điểm) .....	21
5)	Chuyển đổi dữ liệu (Transformation) (3.0 điểm) .....	31
6)	Rút gọn dữ liệu (Reduction) (1.0 điểm) .....	35
II)	Nội dung thực hiện cài đặt.....	36
1)	Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản (15 điểm)...	36
2)	Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước (10 điểm).....	37
III)	Nguồn tham khảo: .....	37
IV)	Phụ lục: .....	38

### I) Nội dung thực hiện báo cáo viết (25 điểm)

#### Dữ liệu thực nghiệm

Bài tập này làm việc trên tập dữ liệu bệnh tim **heart-h.arff** và **heart-c.arff** (tải dữ liệu tại <http://prdownloads.sourceforge.net/weka/datasets-UCI.jar> , 1.1MB)

Mục tiêu của việc khai thác dữ liệu từ các tập dữ liệu này là để hiểu rõ hơn các nhân tố nguy hiểm cho bệnh tim, cụ thể là ở thuộc tính thứ 14 – num (<50: không có bệnh, từ 50-1 đến 50-4 cho biết các mức tăng của bệnh).

Câu hỏi đặt ra là có thể dự đoán tình trạng bệnh tim của một bệnh nhân từ những dữ kiện sức khỏe khác của người này hay không? Tác vụ khai thác dữ liệu được chọn để trả lời câu hỏi này là phân lớp/dự đoán, và một vài giải thuật khác nhau sẽ được sử dụng để tìm ra phương án cho kết quả dự đoán tốt nhất.

#### Yêu cầu thực hiện

##### 1) Tích hợp dữ liệu (integration) (5 điểm)

Bước này thực hiện các thao tác cần thiết cho việc hợp nhất hai tập dữ liệu **heart-h.arff** và **heart-c.arff** thành một tập dữ liệu chung.

###### a) (1đ) Định nghĩa thế nào là tích hợp dữ liệu?

Tích hợp dữ liệu là gộp dữ liệu từ nhiều nguồn khác nhau. Nhằm mục đích giảm và tránh dư thừa hay mâu thuẫn dữ liệu trong tập dữ liệu kết quả. Từ đó giúp cải tiến độ chính xác và tốc độ trong những bước khai thác dữ liệu sau.

## Lab01-Preprocessing

- b) (1đ) Vấn đề nhận diện thực thể (*entity identification*) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?

Vấn đề nhận diện thực thể là làm sao có thể so khớp các dữ liệu thực tế từ nhiều nguồn với nhau. Làm sao biết được thuộc tính trong tập dữ liệu này tương ứng với thuộc tính kia trong tập dữ liệu khác.

Trong hai tập dữ liệu **heart-h.arff** và **heart-c.arff** có xảy ra vấn đề này

Khi so sánh các thuộc tính

```
291 % From: 7 To: reversible_defect
292 %
293 %
294 % Relabeled values in attribute 'num'
295 % From: '0' To: '<50'
296 % From: '1' To: '>50_1'
297 % From: '2' To: '>50_2'
298 % From: '3' To: '>50_3'
299 % From: '4' To: '>50_4'
300 %
301 #Selection Cleveland-14-heart-disease
302 #attribute 'age' real
303 #attribute 'sex' { female, male }
304 #attribute 'cp' { typ_angina, asympt, non_anginal, atyp_angina }
305 #attribute 'trestbps' real
306 #attribute 'chol' real
307 #attribute 'fbs' { t, f }
308 #attribute 'restecg' { left_vent_hyper, normal, st_t_wave_abnormality }
309 #attribute 'thal' { normal, stenosis, yes }
310 #attribute 'oldpeak' real
311 #attribute 'slope' { up, flat, down }
312 #attribute 'ca' real
313 #attribute 'thal' { fixed_defect, normal, reversible_defect }
314 #attribute 'num' { '<50', '>50_1', '>50_2', '>50_3', '>50_4' }
315 %
316 @data
317 #3, male, typ_angina, 145, 230, 6, left_vent_hyper, 150, no, 23, down, 0, fixed_defect, <50
318 #67, male, asympt, 160, 286, f, left_vent_hyper, 180, yes, 1, 5, flat, 3, normal, '>50_1'
319 #67, male, asympt, 120, 229, f, left_vent_hyper, 129, yes, 2, 6, flat, 2, reversible_defect, '
320 #37, male, non_anginal, 130, 250, f, normal, 187, no, 3, 5, down, 0, normal, '<50
321 #
322 #1, female, atyp_angina, 130, 204, f, left_vent_hyper, 172, no, 1, 4, up, 0, normal, '<50
323 #56, male, atyp_angina, 120, 238, f, normal, 178, no, 0, 8, up, 0, normal, '<50
324 #62, female, asympt, 140, 268, f, left_vent_hyper, 160, no, 3, 6, down, 2, normal, '>50_1
325 #57, female, asympt, 120, 354, f, normal, 163, yes, 0, 6, up, 0, normal, '<50
326 #63, male, asympt, 120, 210, f, left_vent_hyper, 150, no, 1, 4, flat, 0, normal, reversible_defect, '
327 #57, male, asympt, 140, 203, f, left_vent_hyper, 155, yes, 1, 5, flat, 0, normal, 0, reversible_defect, '
328 #57, male, asympt, 140, 182, f, normal, 145, no, 0, 4, flat, 0, fixed_defect, '<50
329 #56, female, atyp_angina, 140, 284, f, left_vent_hyper, 153, no, 1, 3, flat, 0, normal, '<50
330 #56, male, non_anginal, 130, 254, f, left_vent_hyper, 142, yes, 0, 6, flat, 1, fixed_defect, '
331 #44, male, atyp_angina, 120, 263, f, normal, 175, no, 0, up, 0, reversible_defect, '<50
332 #52, male, non_anginal, 172, 195, f, normal, 162, no, 0, 5, up, 0, reversible_defect, '<50
333 #57, male, non_anginal, 150, 185, f, normal, 174, no, 1, 6, up, 0, normal, '<50
334 #45, male, atyp_angina, 130, 225, f, normal, 165, no, 1, down, 0, reversible_defect, '>50_1
335 #56, male, non_anginal, 130, 250, f, normal, 175, no, 0, 7, 0, up, 0, normal, '<50
336 #49, female, non_anginal, 130, 275, f, normal, 139, no, 0, 2, up, 0, normal, '<50
337 #49, male, atyp_angina, 130, 266, f, normal, 171, no, 0, 6, up, 0, normal, '<50
338 #
339 #Selection Cleveland-14-heart-disease
340 #attribute 'age' real
341 #attribute 'sex' { female, male }
342 #attribute 'cp' { typ_angina, asympt, non_anginal, atyp_angina }
343 #attribute 'trestbps' real
344 #attribute 'chol' real
345 #attribute 'fbs' { t, f }
346 #attribute 'restecg' { left_vent_hyper, normal, st_t_wave_abnormality }
347 #attribute 'thal' { normal, stenosis, yes }
348 #attribute 'oldpeak' { down, flat, up }
349 #attribute 'ca' real
350 #attribute 'thal' { fixed_defect, normal, reversible_defect }
351 #attribute 'num' { '<50', '>50_1', '>50_2', '>50_3', '>50_4' }
352 %
353 @data
354 #3, male, typ_angina, 145, 230, 6, left_vent_hyper, 150, no, 23, down, 0, fixed_defect, <50
355 #67, male, asympt, 160, 243, f, normal, 165, no, 0, 7, 7, '<50
356 #67, male, asympt, 120, 229, f, normal, 170, no, 0, 7, 7, '<50
357 #30, male, atyp_angina, 170, 237, f, st_t_wave_abnormality, 170, no, 0, ?, ?, fixed_defect
358 #31, female, atyp_engine, 100, 219, f, st_t_wave_abnormality, 150, no, 0, ?, ?, '<50
359 #32, female, atyp_angina, 105, 198, f, normal, 168, no, 0, 7, 7, ?, '<50
360 #32, male, atyp_angina, 120, 225, f, normal, 188, no, 0, 7, 7, ?, '<50
361 #33, male, non_anginal, 120, 298, f, normal, 185, no, 0, 7, 7, ?, '<50
362 #34, female, atyp_angina, 140, 203, f, st_t_wave_abnormality, 168, no, 0, 7, 7, ?, '<50
363 #34, female, atyp_angina, 150, 214, f, st_t_wave_abnormality, 168, no, 0, 7, 7, ?, '<50
364 #34, male, atyp_angina, 88, 220, f, normal, 150, no, 0, 7, 7, ?, '<50
365 #35, female, atyp_angina, 120, 160, f, st_t_wave_abnormality, 185, no, 0, 7, 7, ?, '<50
366 #35, female, asympt, 140, 167, f, normal, 150, no, 0, 7, 7, ?, '<50
367 #35, male, atyp_angina, 150, 264, f, normal, 168, no, 0, 7, 7, ?, '<50
368 #36, male, atyp_angina, 120, 166, f, normal, 188, no, 0, 7, 7, ?, '<50
369 #36, male, non_anginal, 112, 340, f, normal, 189, no, 1, flat, 7, normal, '<50
370 #36, male, atyp_angina, 120, 166, f, normal, 188, no, 0, 7, 7, ?, '<50
371 #36, male, non_anginal, 150, 180, f, normal, 172, no, 0, 7, 7, ?, '<50
372 #37, female, atyp_engine, 120, 280, f, normal, 130, no, 0, 7, 7, ?, '<50
373 #
374 
```

Khi tích hợp

# Lab01-Preprocessing

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of the window. Use the up and down arrows to move through previous commands.

Command completion for classnames and files is initiated with <Tab>. In order to distinguish between files and classnames, file names must be either absolute or start with '.' or '~' (the latter is a shortcut for the home directory). <Alt+BackSpace> is used for deleting the text in the commandline in chunks.

Type 'help' followed by <Enter> to see an overview of all commands.

```
> java weka.core.Instances append D:/heart-h.arff D:/heart-c.arff > D:/heart-integration.arff

Finished redirecting output to 'D:/heart-integration.arff'.
java.lang.Exception: The two datasets have different headers:
Attributes differ at position 3:
Names differ: chest_pain != cp
weka.core.Instances.main(Instances.java:2602)
sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
sun.reflect.NativeMethodAccessorImpl.invoke(Unknown Source)
sun.reflect.DelegatingMethodAccessorImpl.invoke(Unknown Source)
java.lang.reflect.Method.invoke(Unknown Source)
weka.gui.SimpleCLIPanel$ClassRunner.run(SimpleCLIPanel.java:328)

at weka.core.Instances.main(Instances.java:2602)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(Unknown Source)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(Unknown Source)
at java.lang.reflect.Method.invoke(Unknown Source)
at weka.gui.SimpleCLIPanel$ClassRunner.run(SimpleCLIPanel.java:328)

The two datasets have different headers:
Attributes differ at position 3:
Names differ: chest_pain != cp
```

Giải quyết bằng cách tìm ra xem các thuộc tính khác nhau có mối liên hệ nào để tiến hành chọn lọc, kết hợp, ... Cụ thể ta thấy chest\_pain và cp là thuộc tính tương đương nhau có cùng tập giá trị {typ\_angina, asympt, non\_anginal, atyp\_angina} và đều là thuộc tính chest pain type (Loại đau ngực) nên ta sẽ kết hợp chúng lại thành thuộc tính mới cp\_ig

## Lab01-Preprocessing

- c) (1đ) Vấn đề dữ liệu dư thừa (*data redundancy*) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?

Vấn đề dữ thừa dữ liệu xảy ra khi có dữ liệu của thuộc tính này có thể suy ra từ dữ liệu của một hoặc một tập thuộc tính khác. Ngoài ra sự mâu thuẫn trong dữ liệu hay sự không đồng nhất kích thước thuộc tính cũng gây ra dữ thừa dữ liệu.

Trong hai tập dữ liệu **heart-h.arff** và **heart-c.arff** không tồn tại vấn đề này.

Nếu tồn tại vấn đề này thì cần nhắc loại bỏ dữ liệu dư thừa. Nhưng vẫn phải đảm bảo được tính đúng đắn của dữ liệu.

- d) (1đ) Vấn đề mâu thuẫn giá trị dữ liệu (*data value conflicts*) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?

Vấn đề mâu thuẫn giá trị dữ liệu là việc các thuộc tính biến thị cùng một khía cạnh nhưng lại có tập dữ liệu khác nhau. Điều này xảy ra khi thu thập dữ liệu ở những nơi có sự khác nhau về hệ thống đo lường, cách tính toán.

Trong hai tập dữ liệu **heart-h.arff** và **heart-c.arff** không tồn tại vấn đề này.

Nếu tồn tại vấn đề này thì xử lý bằng cách quy dữ liệu về chung một hệ thống đo lường, một cách tính

- e) (1đ) Tích hợp hai tập dữ liệu đã cho thành tập dữ liệu mới có tên là **heart-integration.arff**. Sử dụng WEKA để đọc tập dữ liệu tích hợp. Chụp màn hình cửa sổ Explorer, đánh dấu các vùng trong cửa sổ có thể cho biết số mẫu và số thuộc tính của dữ liệu.

➤ Tích hợp dữ liệu:

Sử dụng lệnh sau trong Weka/Simple CLI

```
java weka.core Instances append D:/heart-h.arff D:/heart-c.arff > D:/heart-integration.arff
```

## Lab01-Preprocessing

```
SimpleCLI

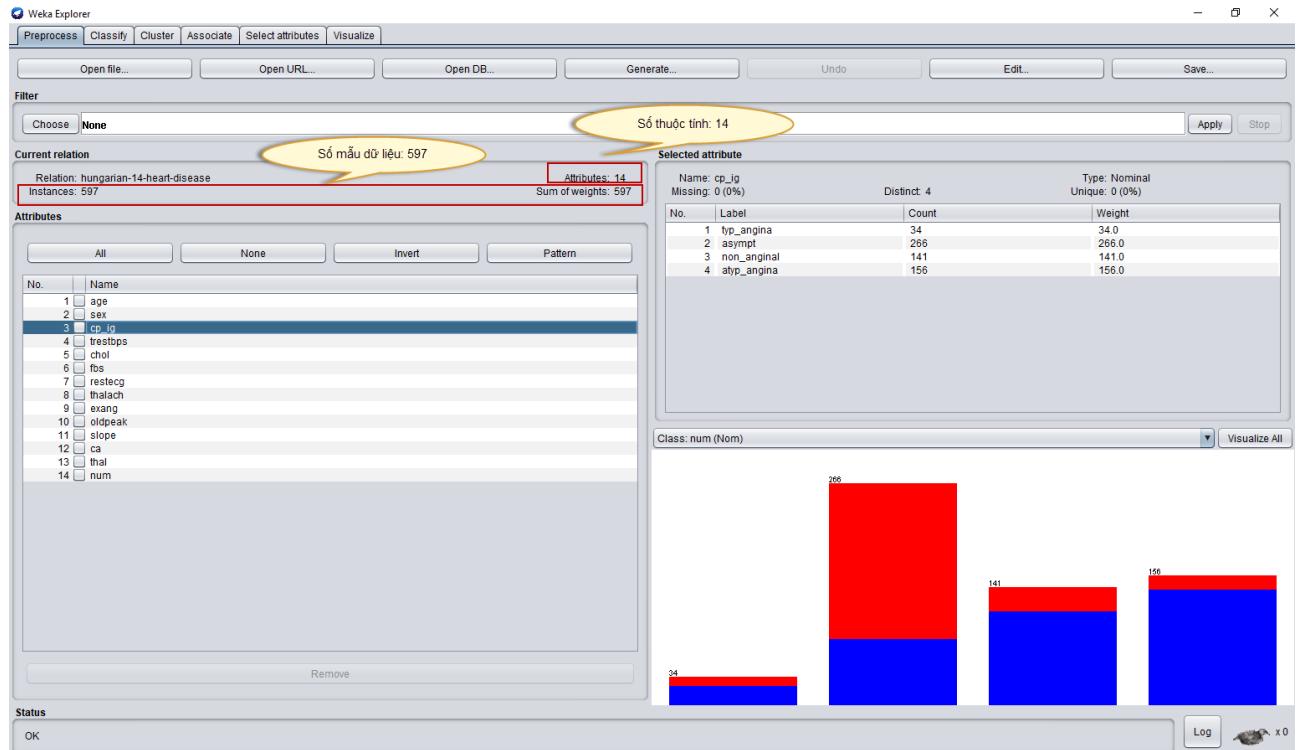
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.' or './'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

Type 'help' followed by <Enter> to see an overview
of all commands.
> java weka.core.Instances append D:/heart-h.arff D:/heart-c.arff > D:/heart-integration.arff

Finished redirecting output to 'D:/heart-integration.arff'.
```

### ➤ Đọc dữ liệu:



- Nộp lại tập tin **heart-integration.arff**. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).

File đính kèm

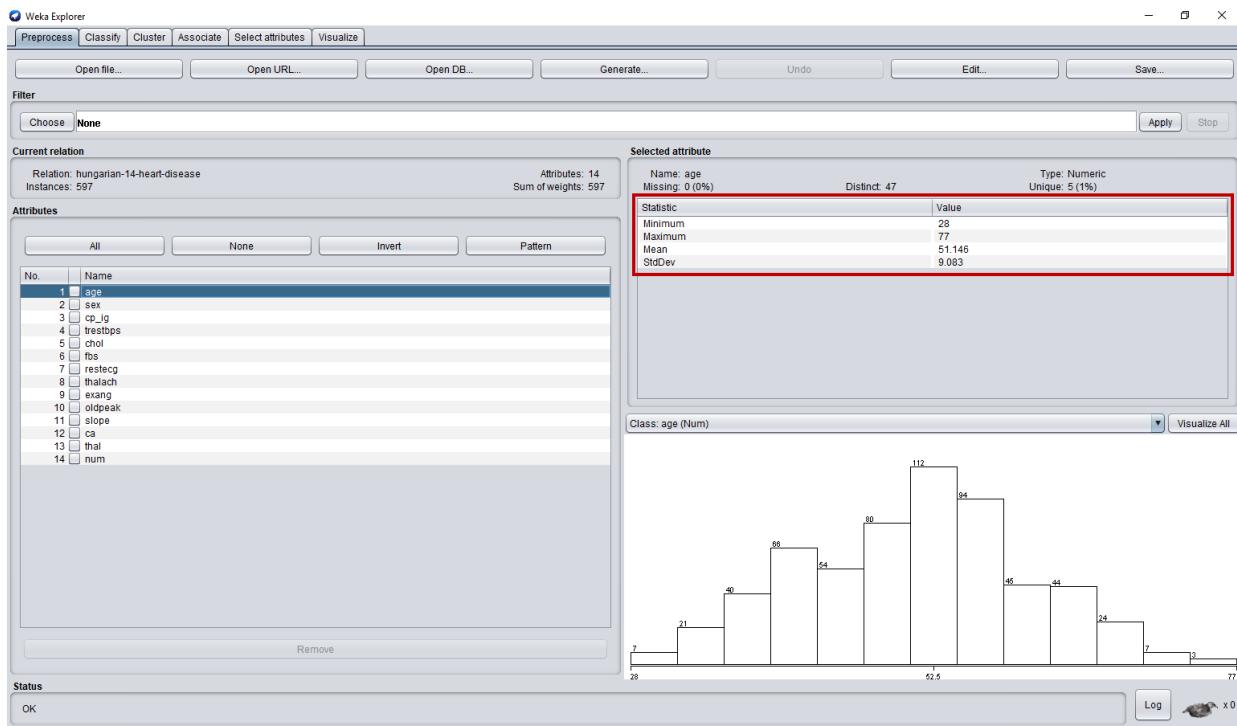
## Lab01-Preprocessing

### 2) Tóm tắt dữ liệu mô tả (descriptive data summarization) (8.0 điểm)

Một bước quan trọng trong tiền xử lý dữ liệu là làm quen với dữ liệu thông qua các giá trị thống kê, chúng cho ta biết những đặc tính phổ thông của dữ liệu.

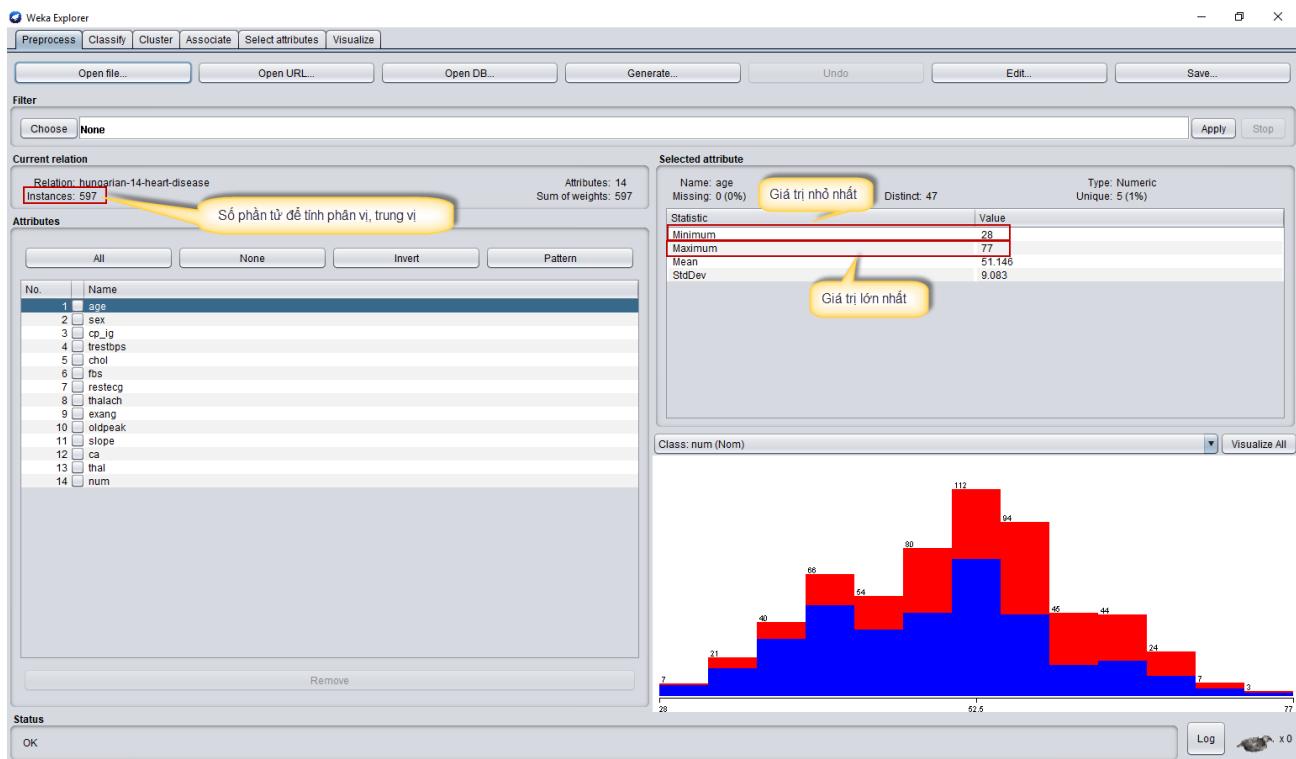
Mở tab **Preprocess** của cửa sổ WEKA Explorer, đọc tập dữ liệu **heart-integration.arff**.

- a) (1đ) Cho biết giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất và giá trị lớn nhất của thuộc tính **age**. Chụp màn hình cửa sổ Explorer, đánh dấu các vùng trong cửa sổ cho biết những thông tin này.



- Giá trị trung bình: 51.146
  - Độ lệch chuẩn: 9.083
  - Giá trị nhỏ nhất: 28
  - Giá trị lớn nhất: 77
- b) (1đ) Xác định **five-number summary** của thuộc tính **age**. Chụp màn hình cửa sổ Explorer hiển thị thông tin này nếu WEKA có cung cấp. Nếu không, bạn căn cứ vào những giá trị nào khác có trong WEKA để tính?
- **five-number summary** gồm:
    - Giá trị nhỏ nhất – Minimum: 28
    - Phân vị thứ nhất – Quantiles Q1: 44
    - Trung vị – Median: 52
    - Phân vị thứ ba – Quantiles Q3: 58
    - Giá trị lớn nhất – Maximum: 77
  - Căn cứ vào Instances để tính trung vị và phân vị

## Lab01-Preprocessing



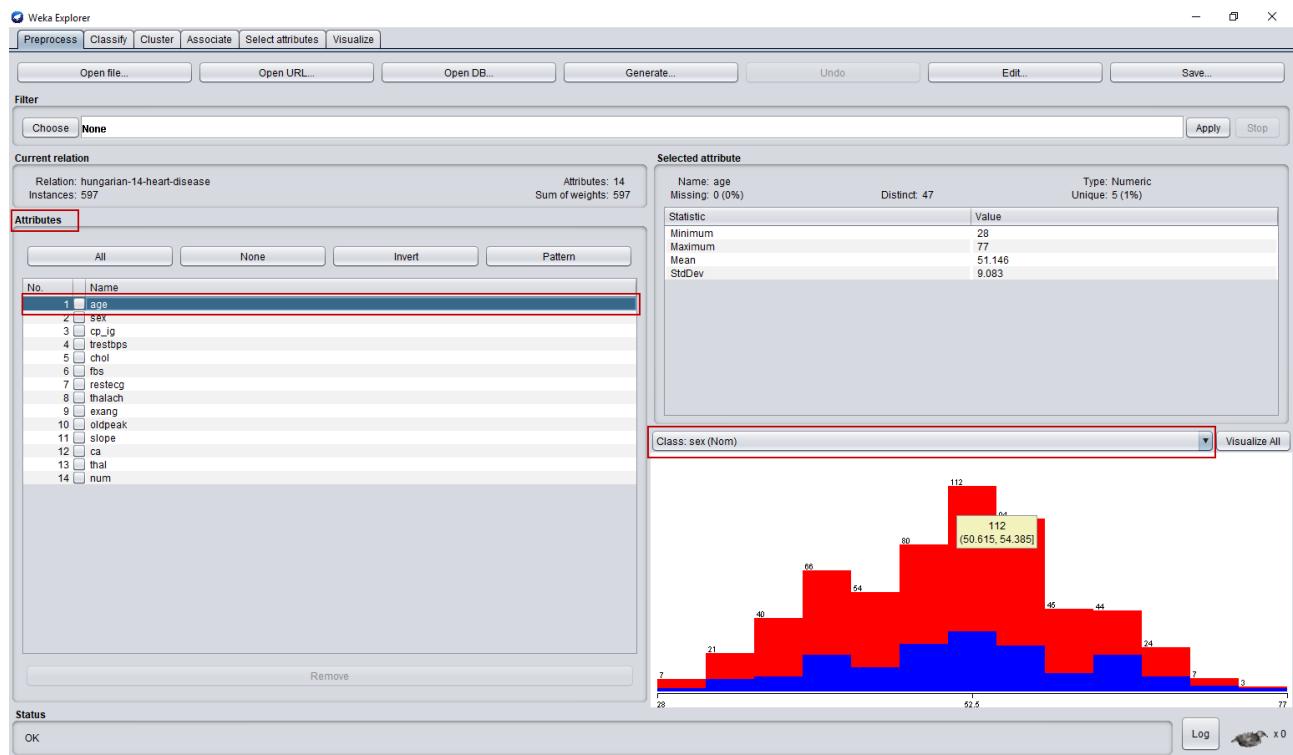
c) (1đ) Cho biết thuộc tính nào có kiểu thuộc tính số (*numeric*), kiểu rời rạc không có thứ tự (*categorical/nominal*), hoặc kiểu rời rạc có thứ tự (*ordinal*).

- Kiểu thuộc tính số (*numeric*): age, trestbps, chol, thalach, oldpeak, ca
- Kiểu rời rạc không có thứ tự (*categorical/nominal*): sex, cp\_ig (chest\_pain), fbs, restecg, exang, slope, thal, num
- Kiểu rời rạc có thứ tự (*ordinal*): Không có

d) (1đ) Giải thích ý nghĩa của đồ thị ở góc dưới bên phải của cửa sổ. Bạn gọi tên đồ thị này là gì? Đồ thị biểu diễn điều gì về tập dữ liệu? Màu xanh và màu đỏ có ý nghĩa gì (chú ý các pop-up hiện lên khi chuột di chuyển vào vùng đồ thị)?

- Đồ thị ở góc dưới bên phải của cửa sổ thể hiện sự tương quan giữa attributes ở bên trái cửa sổ và class bên phải cửa sổ và sự tương quan giữa các giá trị trong class
- Biểu đồ này là biểu đồ bar plot – biểu đồ phân phối
- Biểu diễn về các tập dữ liệu attributes và class. Trong đó trục hoành x thể hiện attributes và trục y thể hiện class.

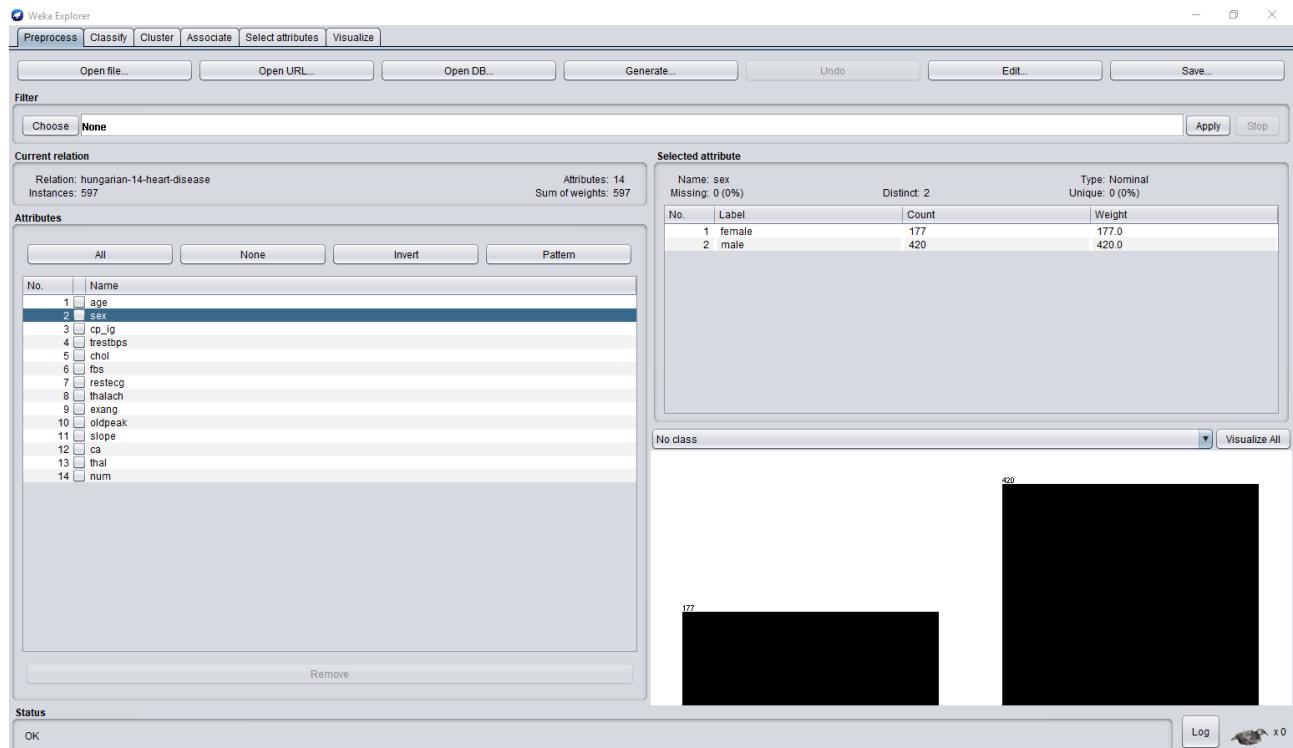
## Lab01-Preprocessing



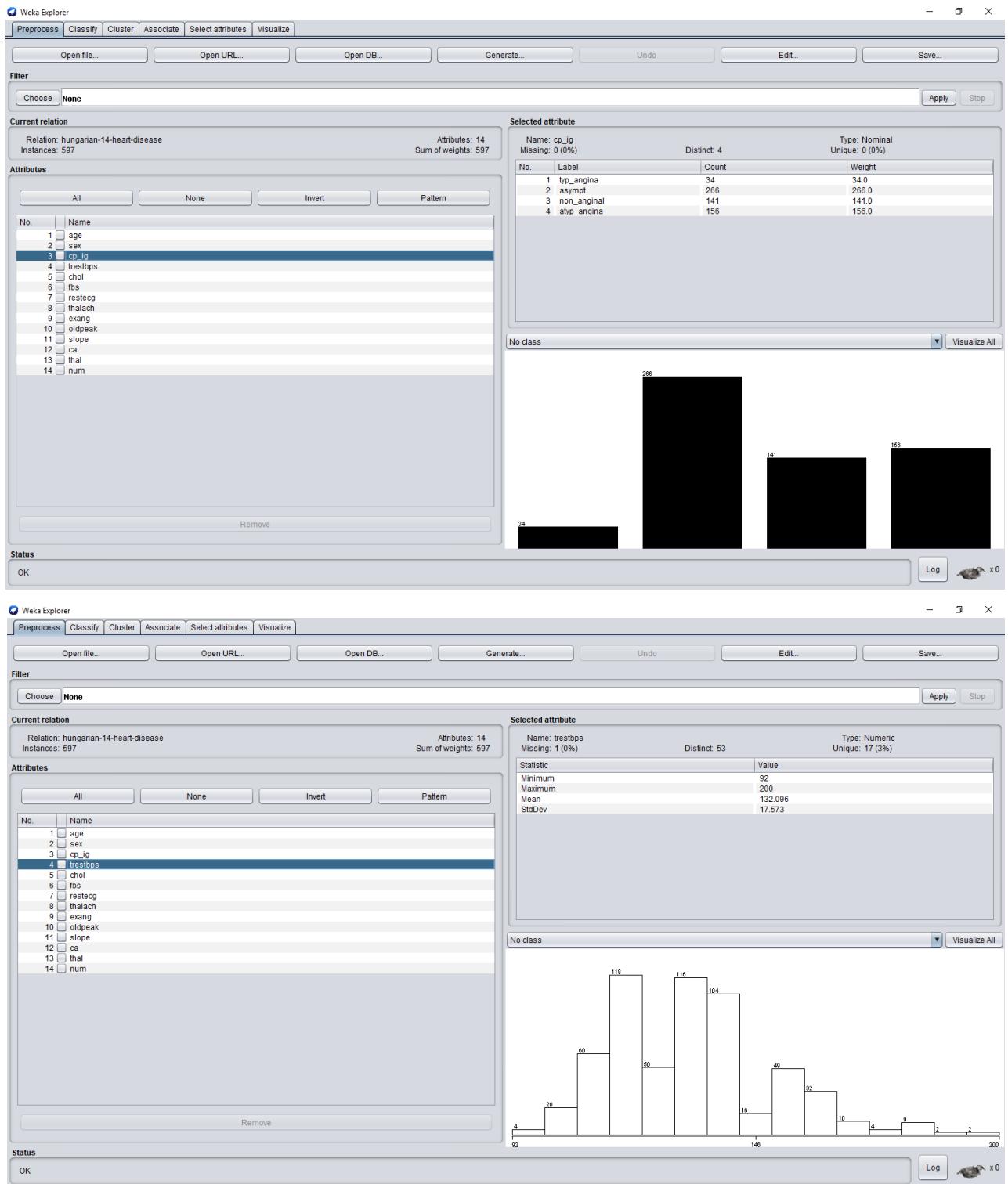
Như hình trên attributes age được chia thành 13 phần. Class sex chia thành 2 phần là male và female

- Màu đỏ, màu xanh thể hiện sự phân lớp dữ liệu do class thể hiện. Cụ thể hình trên màu đỏ là male, màu xanh là female. Các popup hiện lên thể hiện tần số xuất hiện của dữ liệu trong từng phần. Như hình trên độ tuổi trong khoảng [50.615, 54.385] xuất hiện 112 lần, tỷ lệ male lớn hơn female.

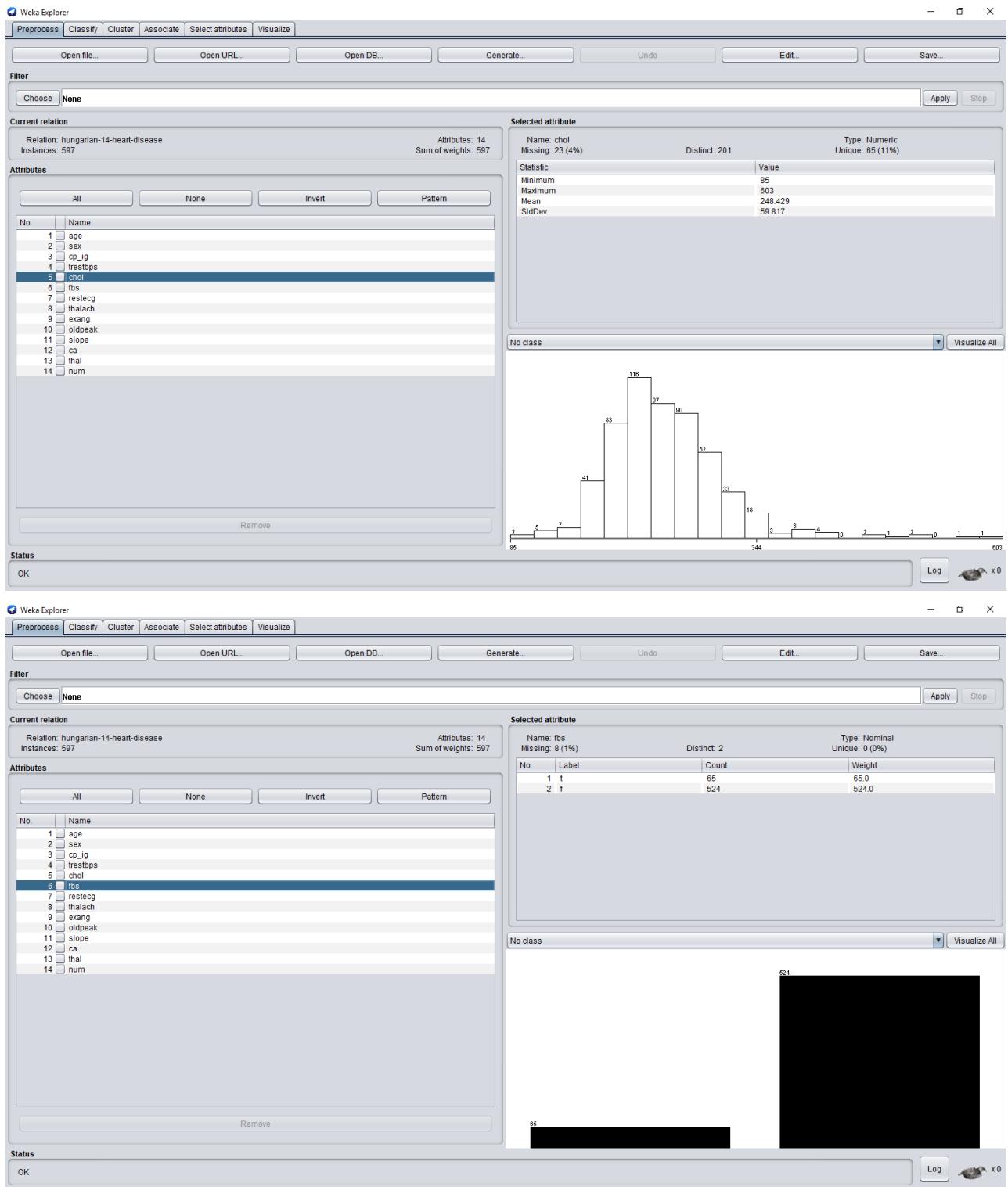
e) (1đ) Lần lượt xem xét các thuộc tính khác ngoài thuộc tính age. Chụp màn hình cửa sổ Explorer tương ứng với từng thuộc tính.



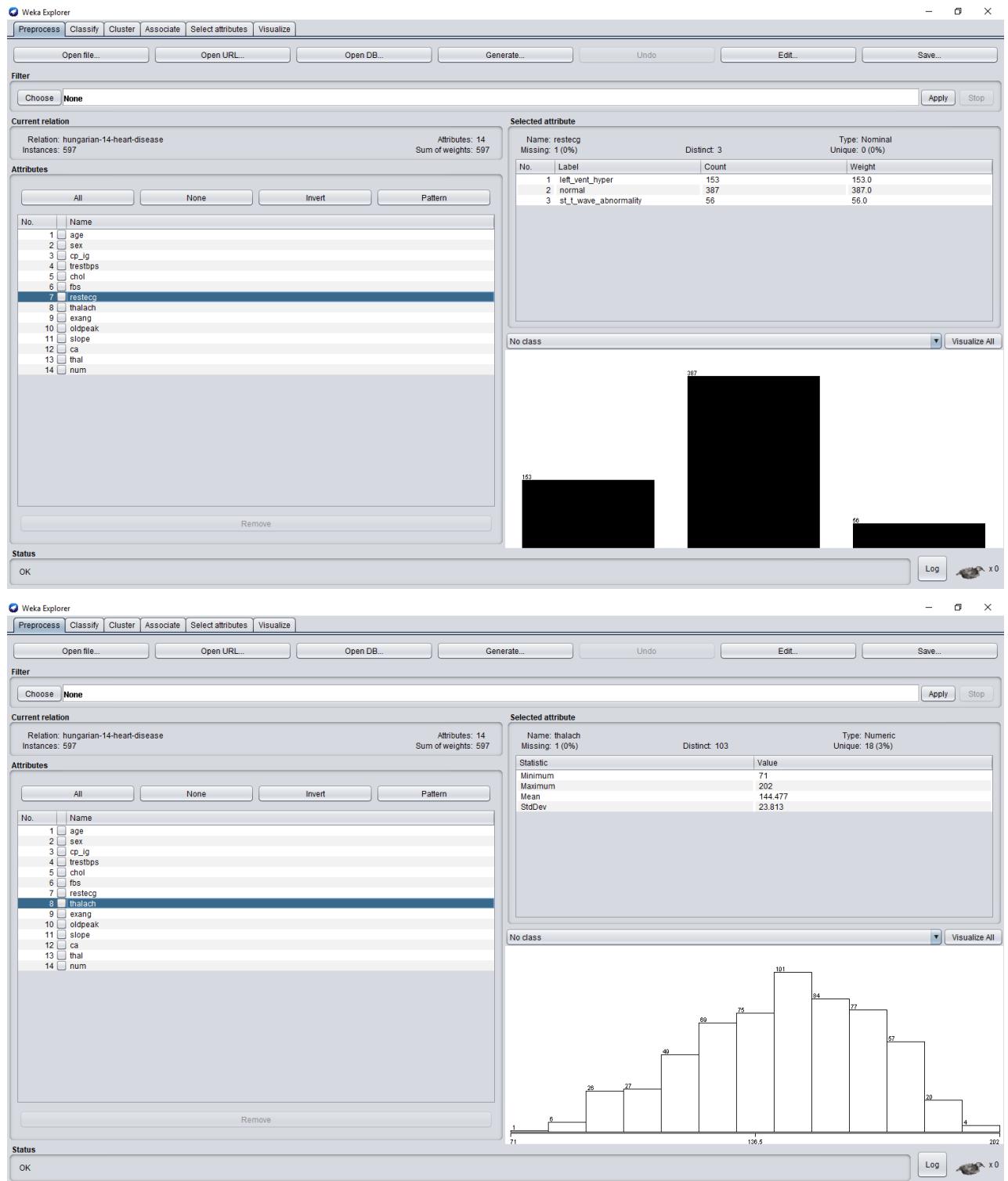
## Lab01-Preprocessing



## Lab01-Preprocessing



## Lab01-Preprocessing



## Lab01-Preprocessing

**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: hungarian-14-heart-disease  
Instances: 597 Attributes: 14 Sum of weights: 597

**Attributes**  
All None Invert Pattern

No.	Name
1	age
2	sex
3	cp.JG
4	trestbps
5	chol
6	fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	slope
12	ca
13	thal
14	num

Remove

**Selected attribute**  
Name: exang  
Missing: 1(0%) Distinct: 2 Unique: 0(0%)  
No. Label Count Weight  
1 no 408 408.0  
2 yes 188 188.0

No class Visualize All

**Status** OK Log x 0

**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: hungarian-14-heart-disease  
Instances: 597 Attributes: 14 Sum of weights: 597

**Attributes**  
All None Invert Pattern

No.	Name
1	age
2	sex
3	cp.JG
4	trestbps
5	chol
6	fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	slope
12	ca
13	thal
14	num

Remove

**Selected attribute**  
Name: oldpeak  
Missing: 0(0%) Distinct: 41 Unique: Numeric  
Statistic Value  
Minimum 0  
Maximum 6.2  
Mean 0.816  
StdDev 1.068

No class Visualize All

**Status** OK Log x 0

## Lab01-Preprocessing

**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: hungarian-14-heart-disease  
Instances: 597 Attributes: 14 Sum of weights: 597

**Attributes**

No.	Name
1	age
2	sex
3	cp_ig
4	trestbps
5	chol
6	fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	slope
12	ca
13	thal
14	num

Remove

**Selected attribute**  
Name: slope  
Missing: 190 (32%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	down	22	22.0
2	flat	231	231.0
3	up	154	154.0

No class Visualize All

**Status** OK Log x 0

**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: hungarian-14-heart-disease  
Instances: 597 Attributes: 14 Sum of weights: 597

**Attributes**

No.	Name
1	age
2	sex
3	cp_ig
4	trestbps
5	chol
6	fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	slope
12	ca
13	thal
14	num

Remove

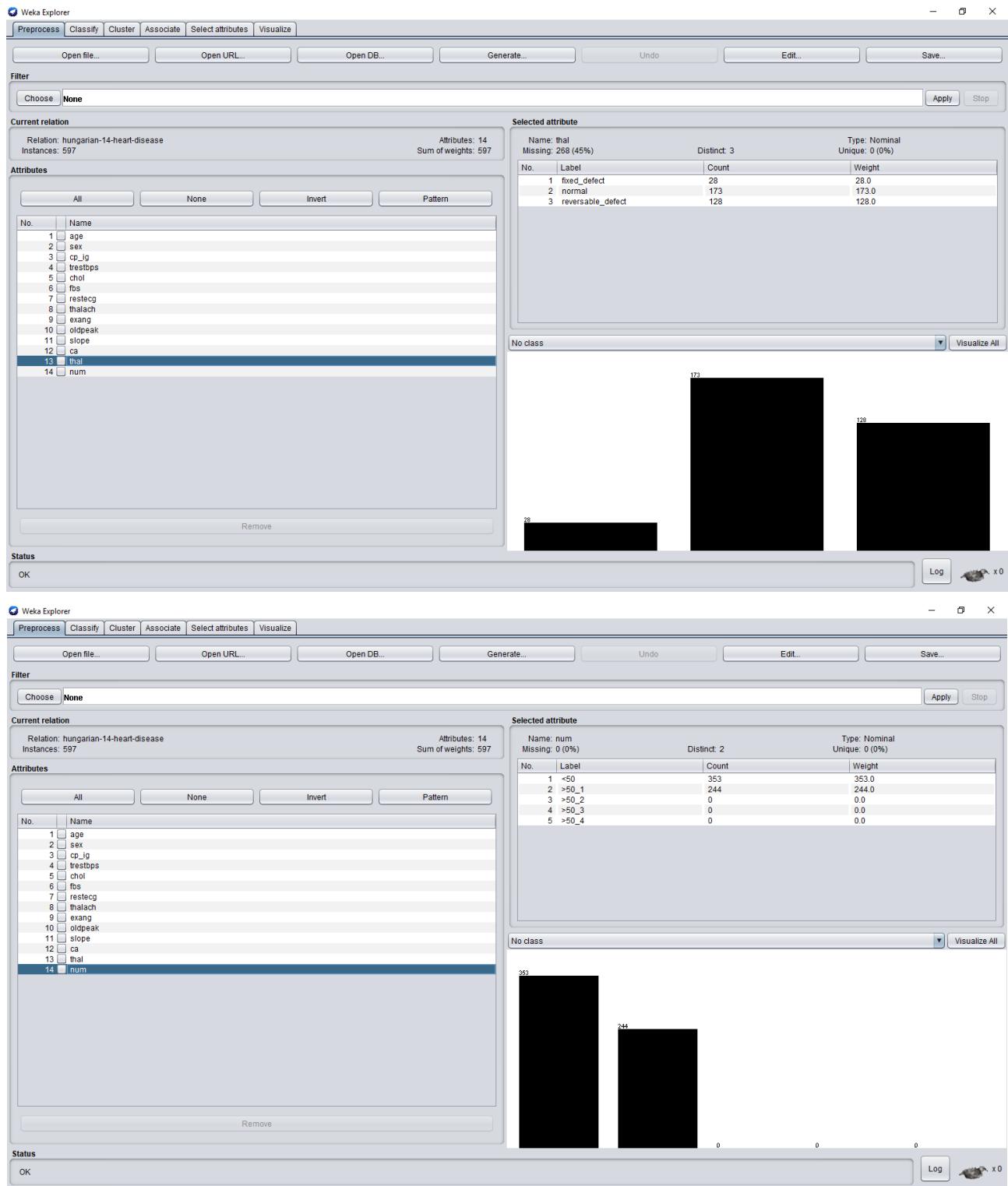
**Selected attribute**  
Name: ca  
Missing: 296 (50%) Distinct: 4 Type: Numeric Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	3
Mean	0.668
StdDev	0.936

No class Visualize All

**Status** OK Log x 0

## Lab01-Preprocessing



f) (1đ) Bạn có nhận xét gì từ những đồ thị trong câu e.?

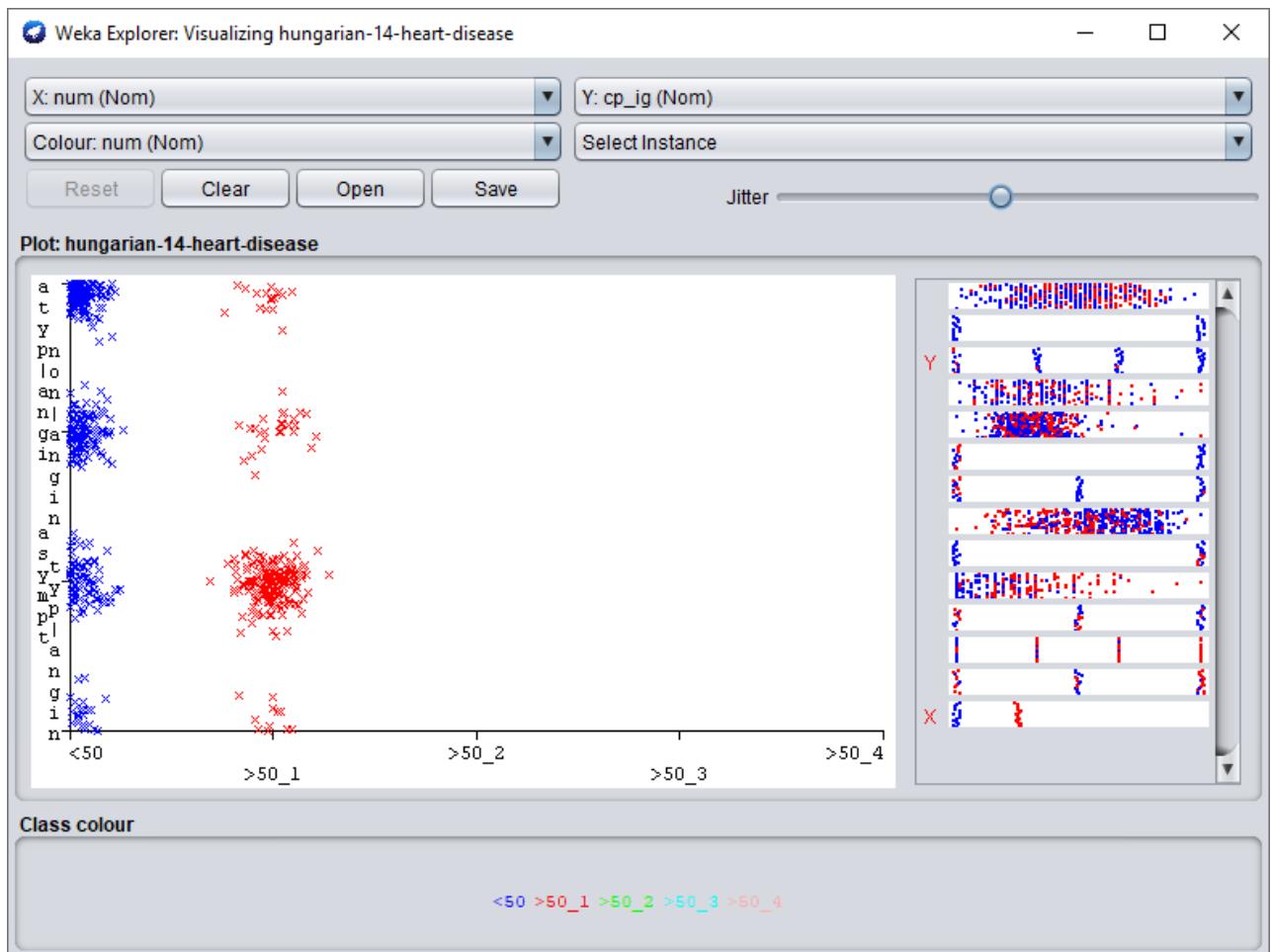
- Có nhiều cột dữ liệu bị trống → Có thể có dữ liệu bị mất
- Các cột dữ liệu ở nhiều thuộc tính có sự chênh lệch lớn → Có dữ liệu nhiễu

Bây giờ chúng ta sẽ chuyển sang tab **Visualize**

## Lab01-Preprocessing

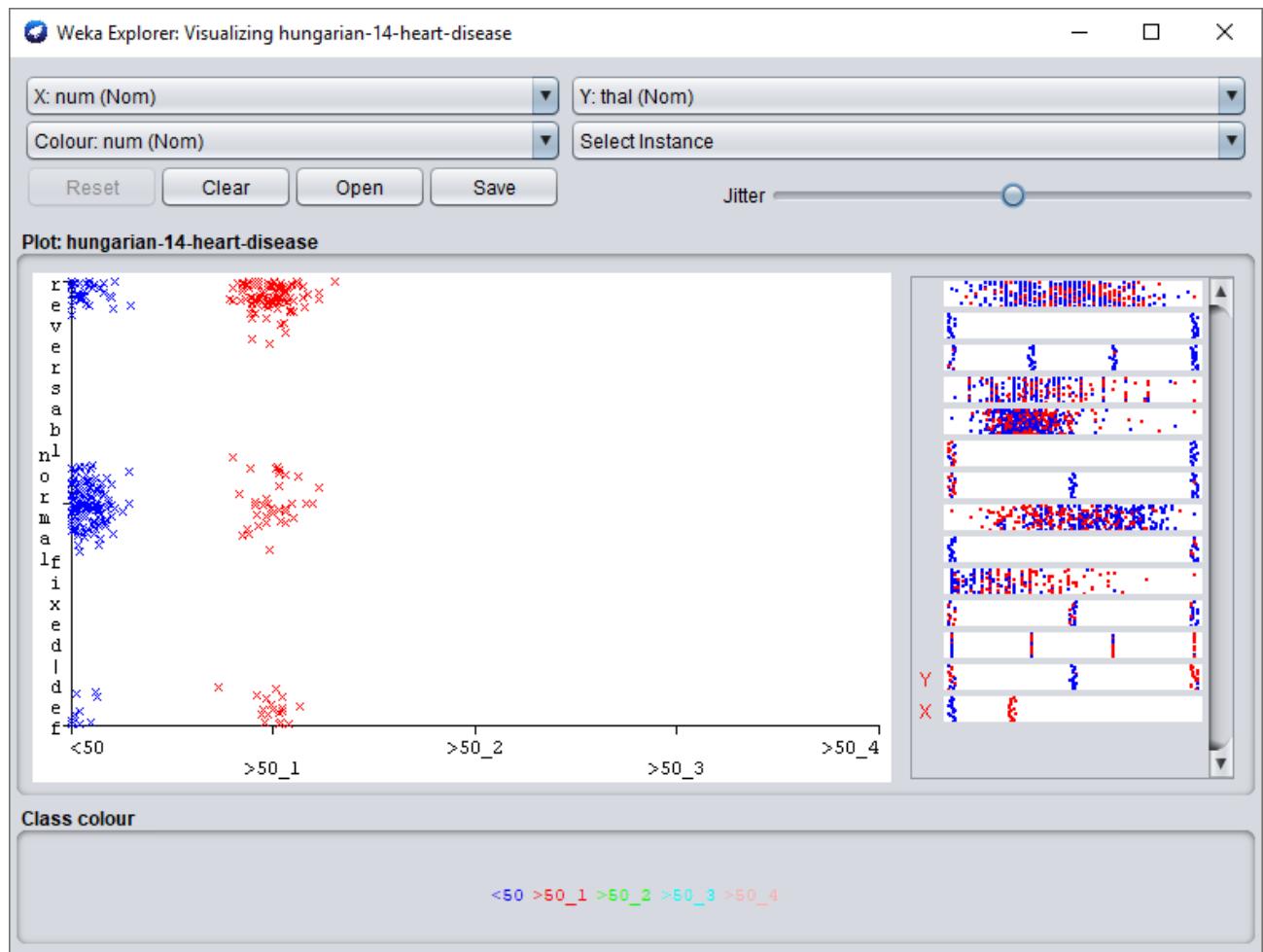
g) (1đ) Các đồ thị này được gọi tên bằng thuật ngữ gì trong textbook [4]? Chọn **jitter** tối đa, chú ý cột **num** (cột cuối cùng). Bạn cho rằng thuộc tính (Y) nào có khả năng dự đoán tốt nhất về bệnh tim như là một hàm của num (X)? Chụp hình đồ thị của cặp thuộc tính (X) – (Y) này.

- Trong textbook các đồ thị này được gọi là **scatter plot** (biểu đồ phân tán)
- Thuộc tính (Y) nào có khả năng dự đoán tốt nhất về bệnh tim như là một hàm của num (X):
  - chest pain type – Loại đau ngực



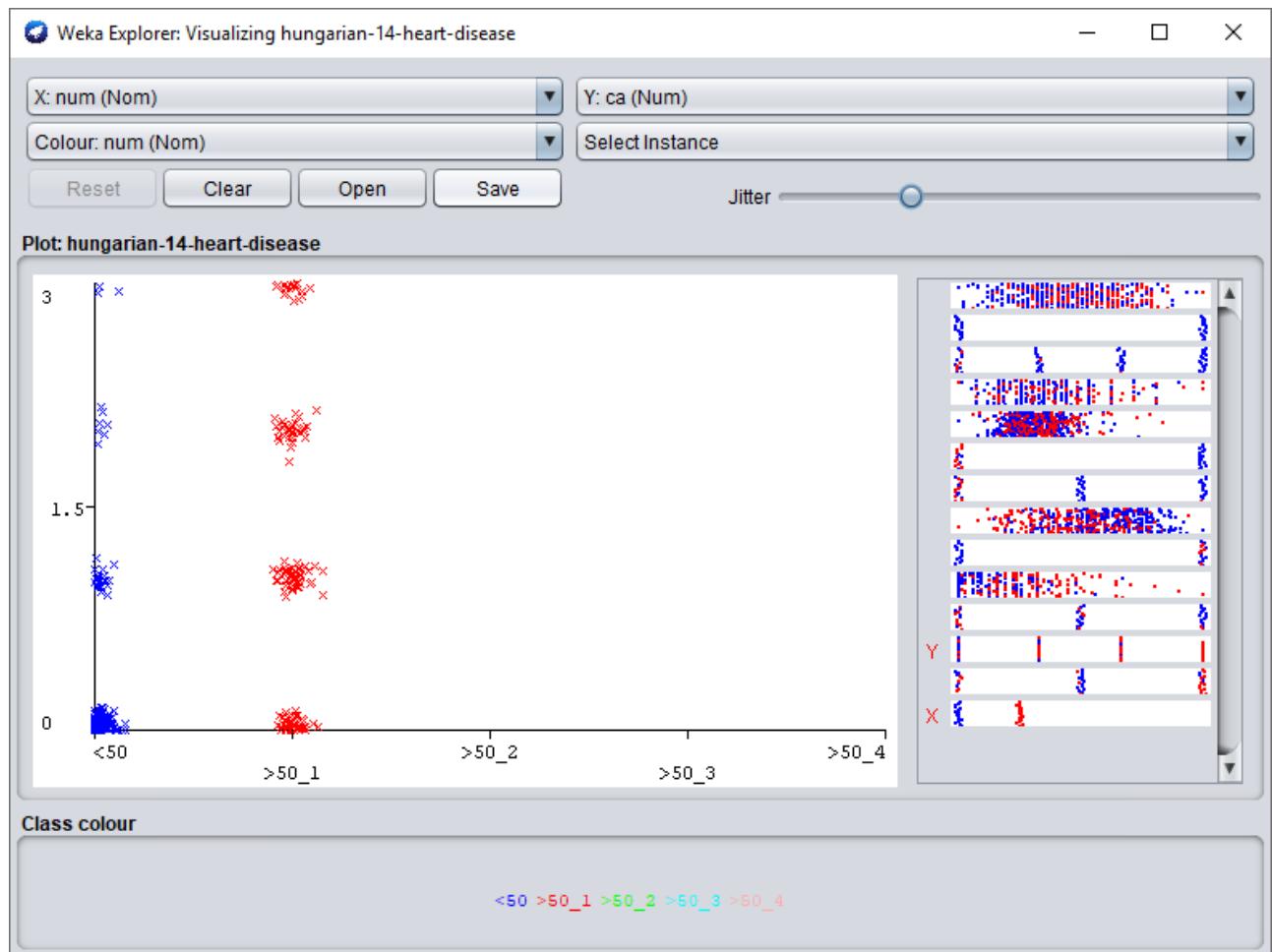
- thal

## Lab01-Preprocessing



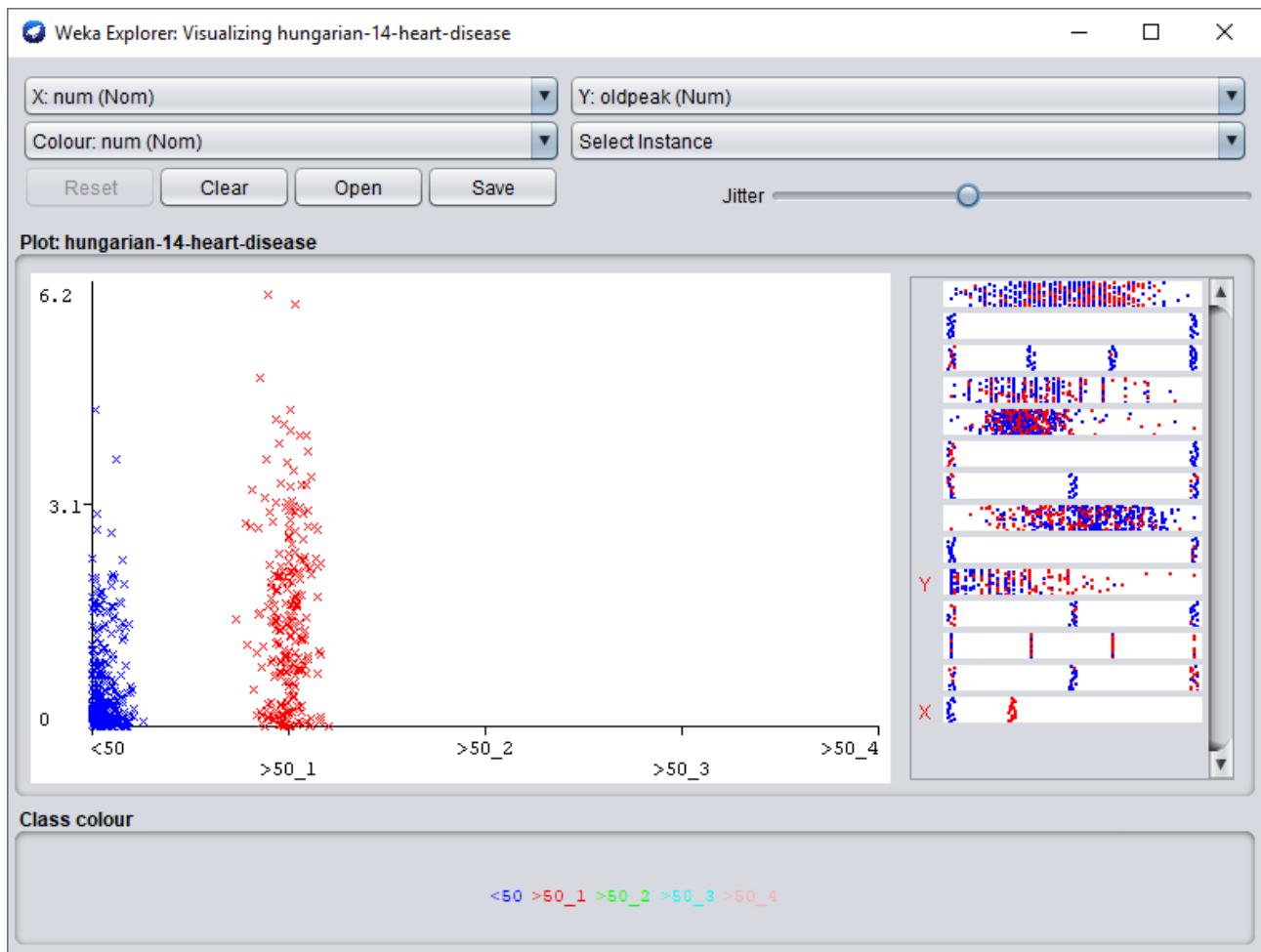
■ ca

## Lab01-Preprocessing



- oldpeak

# Lab01-Preprocessing



h) (1đ) Còn có những cặp thuộc tính nào khác có vẻ tương quan với nhau không?

Không có cặp thuộc tính nào khác có vẻ tương quan với nhau

### 3) Chọn lọc dữ liệu (selection) (3.0 điểm)

Các tập dữ liệu **heart-h.arff** và **heart-c.arff** khi công bố đã được xử lý bằng cách chọn ra tập hợp các thuộc tính liên quan đến mục tiêu khai thác dữ liệu.

a) (1đ) Dựa vào phần mô tả ở đầu tập tin arff, cho biết có bao nhiêu thuộc tính trong các tập dữ liệu heart-h và heart-c trước khi xử lý?

Số lượng thuộc tính trong các tập dữ liệu heart-h và heart-c trước khi xử lý là 76

# Lab01-Preprocessing

 Sử dụng tab **Select attributes** của cửa sổ WEKA Explorer

b) (1đ) Giải thích ngắn gọn từng phương pháp chọn lọc thuộc tính trong WEKA.

- Attribute Evaluator:
    - CfsSubsetEval: Đánh giá giá trị của một tập hợp con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng tính năng cùng với mức độ dư thừa giữa chúng.
    - ClassifierAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách sử dụng trình phân loại do người dùng chỉ định
    - ClassifierSubsetEval: Đánh giá các tập hợp thuộc tính trên dữ liệu huấn luyện hoặc kiểm tra riêng biệt
    - CorrelationAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo lường mối tương quan giữa thuộc tính và lớp.
    - GainRatioAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo tỷ lệ khuếch đại đối với lớp.
    - InfoGainAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo mức tăng thông tin đối với lớp
    - OneRAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách sử dụng trình phân loại OneR.
    - PrincipalComponents: Phân tích thành phần chính và chuyển đổi dữ liệu.
    - ReliefFAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách lấy mẫu xem xét giá trị của thuộc tính đã cho với giá trị gần nhất và khác lớp.
    - SymmetricalUncertAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo độ bất đối xứng của lớp.
    - WrapperSubsetEval: Đánh giá các tập thuộc tính bằng cách sử dụng sơ đồ học tập
  - Search Method:
    - BestFirst: Tìm kiếm không gian của các tập hợp thuộc tính bằng cách greedy hillclimbing.
    - GreedyStepwise: Thực hiện tìm kiếm tiến hoặc lùi tham lam thông qua tập hợp thuộc tính

## Lab01-Preprocessing

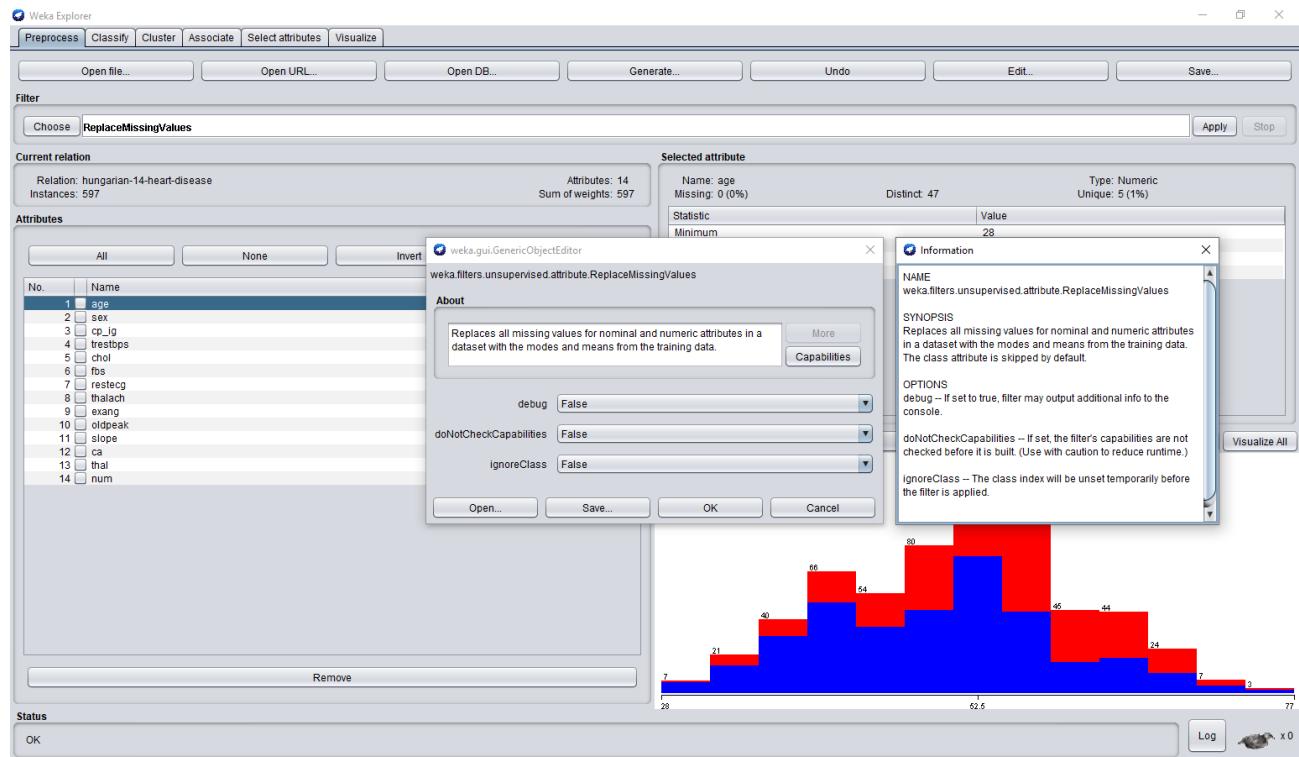
- Ranker: Xếp hạng các thuộc tính theo đánh giá cá nhân
- c) (1đ) So sánh với các phương pháp chọn lọc dữ liệu trong textbook. Phương pháp nào có trong textbook nhưng không có trong WEKA? Phương pháp nào có trong WEKA nhưng không có trong textbook?
- Các phương pháp chọn lọc trong textbox:
  - Stepwise forward selection: Chọn các thuộc tính tốt lần lượt đưa vào tập rỗng
  - Stepwise backward elimination: Lần lượt loại bỏ các thuộc tính xấu trong tập
  - Combination of forward selection and backward elimination: Kết hợp giữa lựa chọn thuộc tính tốt và loại bỏ thuộc tính xấu
  - Decision tree induction: Sử dụng cây quyết định để phân loại
- Các phương pháp có trong textbox đều có trong Weka

### 4) Làm sạch dữ liệu (cleaning) (5.0 điểm)

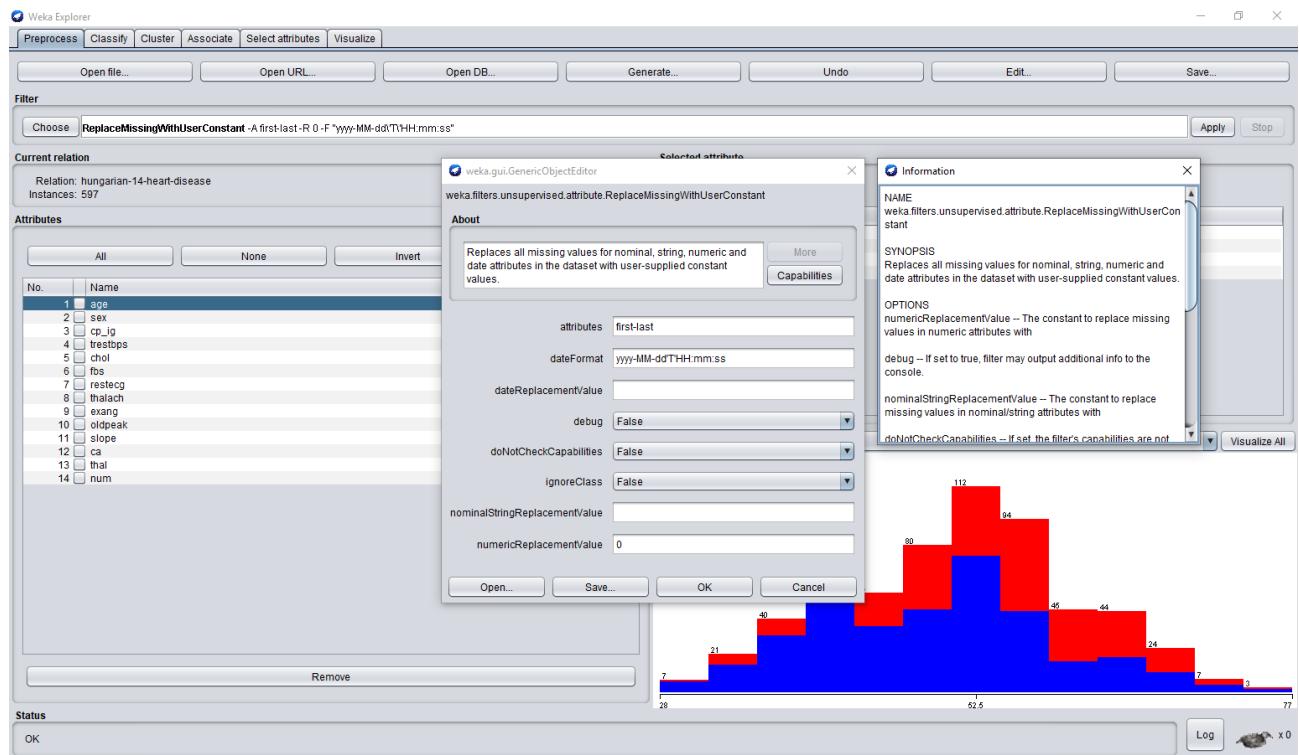
Dữ liệu thực tế thường phải đổi diện với tình trạng không đầy đủ, nhiễu hoặc không nhất quán. Chúng ta quay lại tab **Preprocess**, đọc tập dữ liệu tích hợp **heart-integration.arff** và sử dụng các bộ lọc để làm sạch tập dữ liệu này

- a) (1đ) Liệt kê các phương pháp đã học trong bài giảng để xử lý ván đề thiếu giá trị (*missing values*). WEKA hỗ trợ những phương pháp nào cho ván đề này?
  - Các phương pháp đã học trong bài giảng để xử lý ván đề thiếu giá trị:
    - Bỏ qua các mẫu tin có giá trị thiếu
    - Điền các giá trị thiếu bằng tay
    - Điền các giá trị thiếu tự động
      - ✓ Thay thế bằng hằng số chung
      - ✓ Thay thế bằng giá trị trung bình của thuộc tính
      - ✓ Thay thế bằng giá trị trung bình của thuộc tính trong một lớp
      - ✓ Thay thế bằng giá trị có nhiều khả năng nhất
  - WEKA hỗ trợ các phương pháp:
    - ReplaceMissingValues: Thay thế các giá trị thiếu dạng số, dạng rời rạc có không có thứ tự bằng dữ liệu huấn luyện

## Lab01-Preprocessing

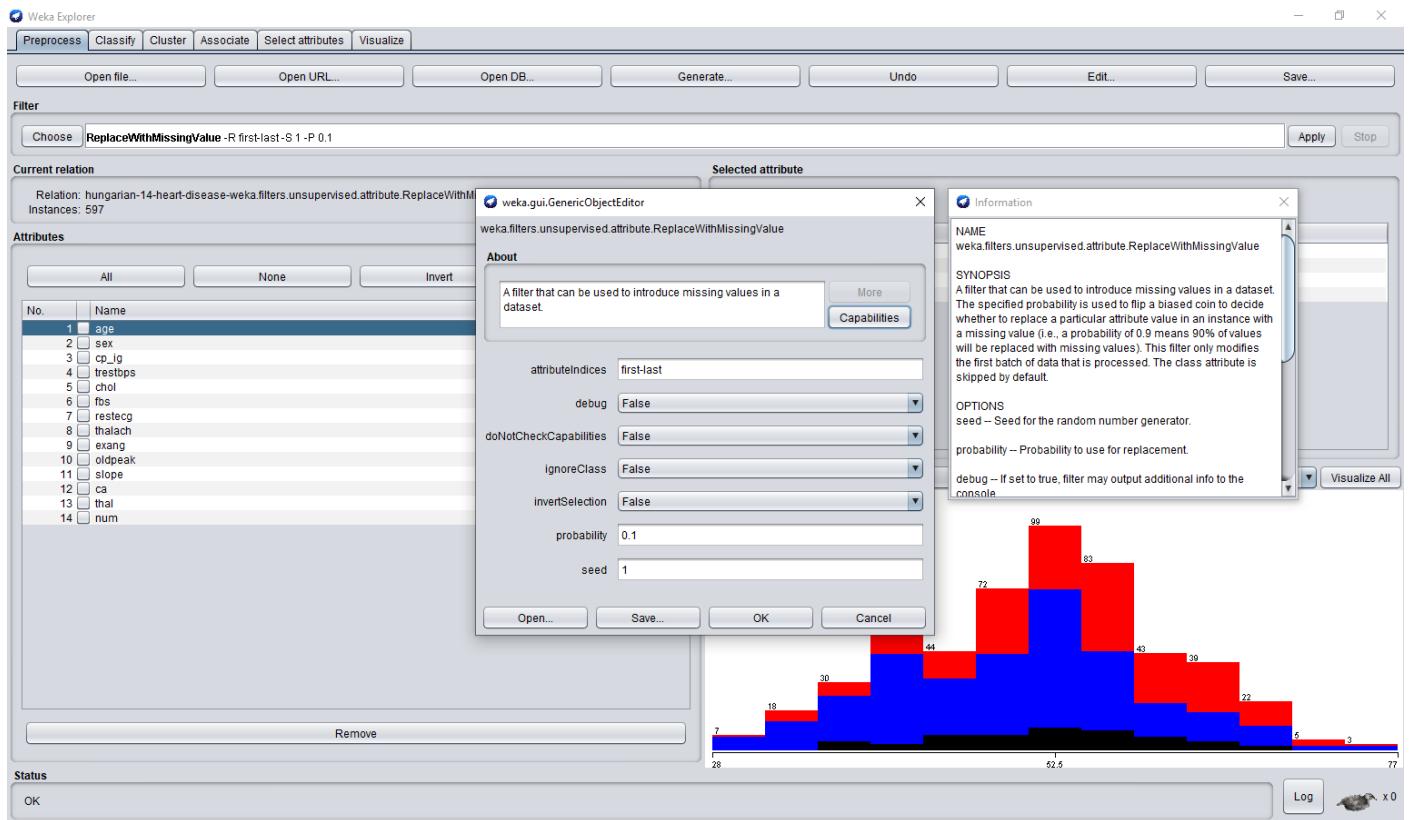


- ReplaceMissingWithUserConstant: Thay thế các giá trị thiếu dạng rác có không có thứ tự, chuỗi, số, ngày tháng bằng hằng số do người dùng tự định nghĩa

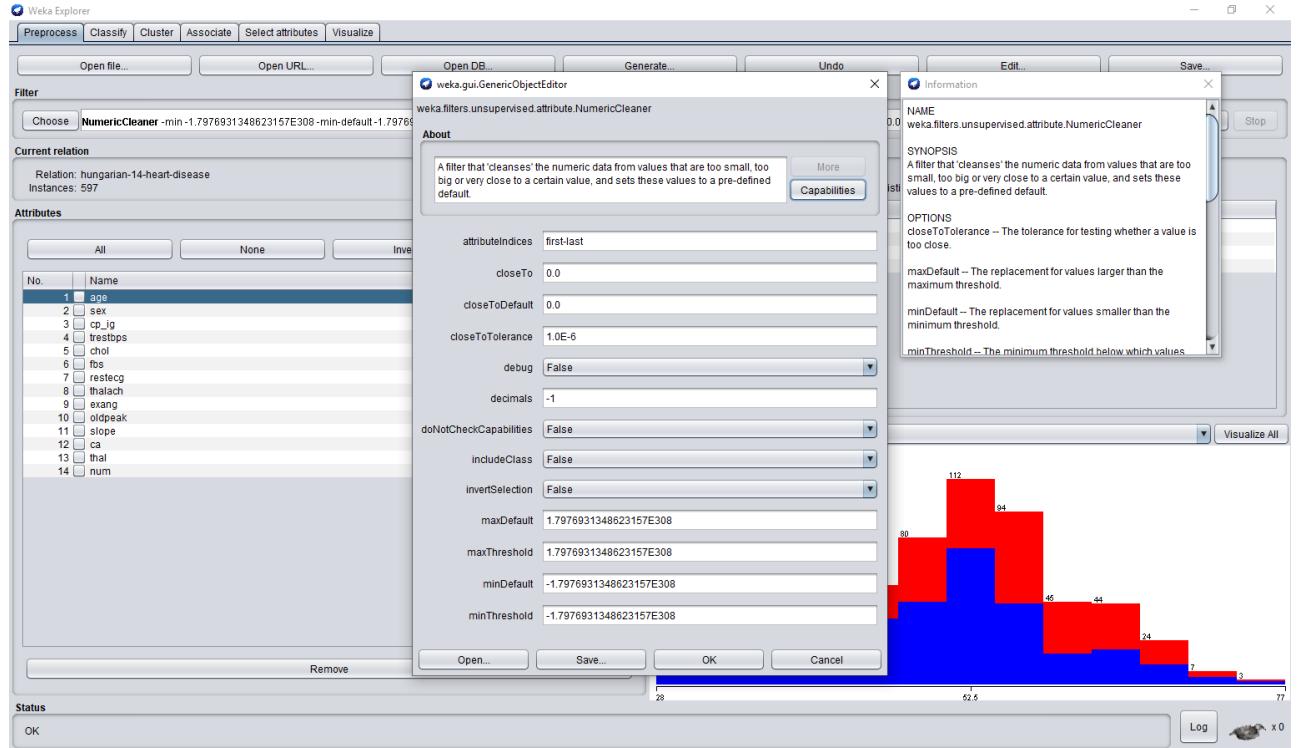


- ReplaceWithMissingValue: Lọc ra các giá trị bị thiếu

# Lab01-Preprocessing

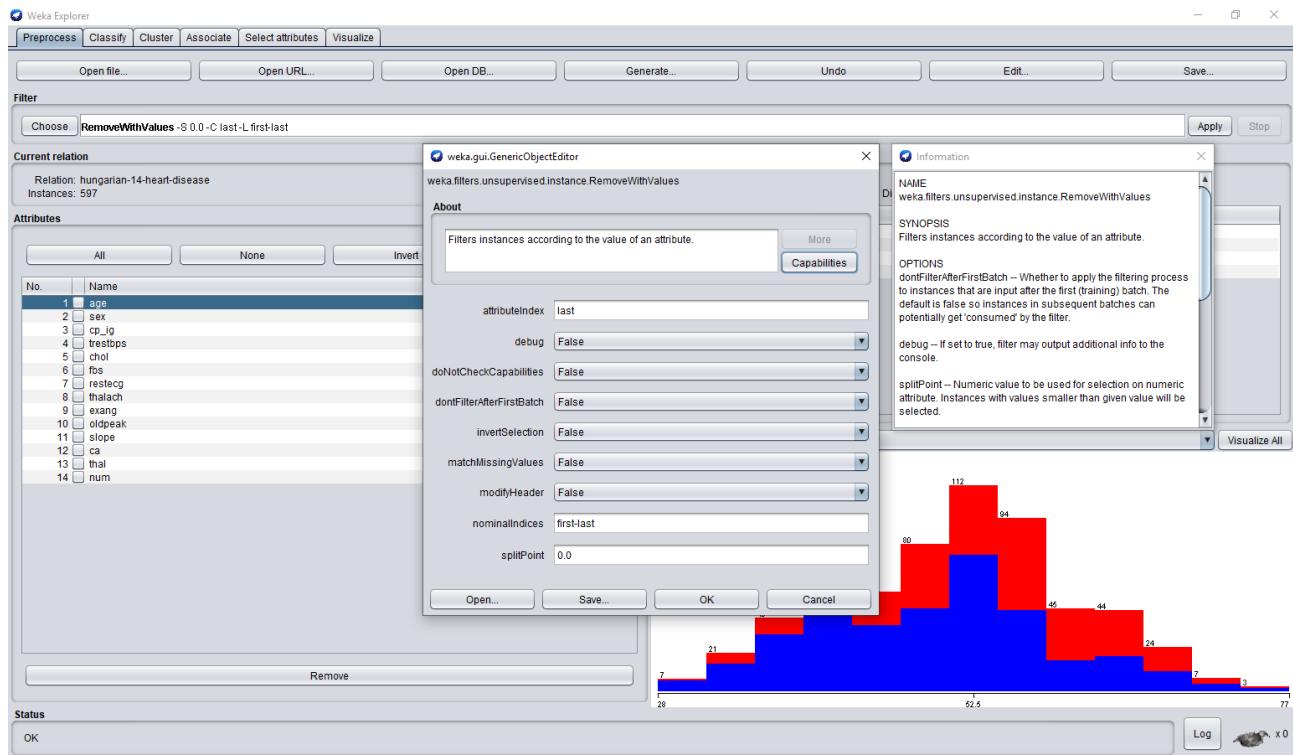


- Mark missing values use NumericalCleaner filter



- Remove Missing Data use RemoveWithValues filter

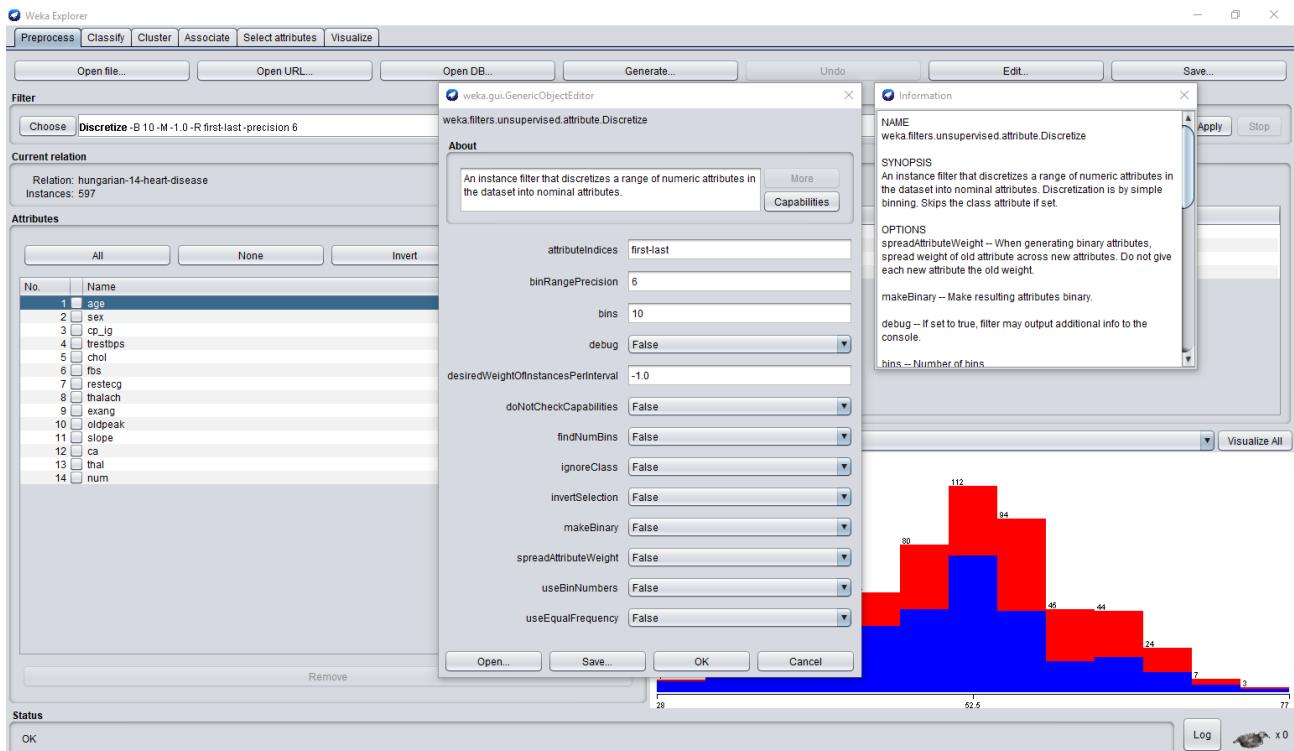
## Lab01-Preprocessing



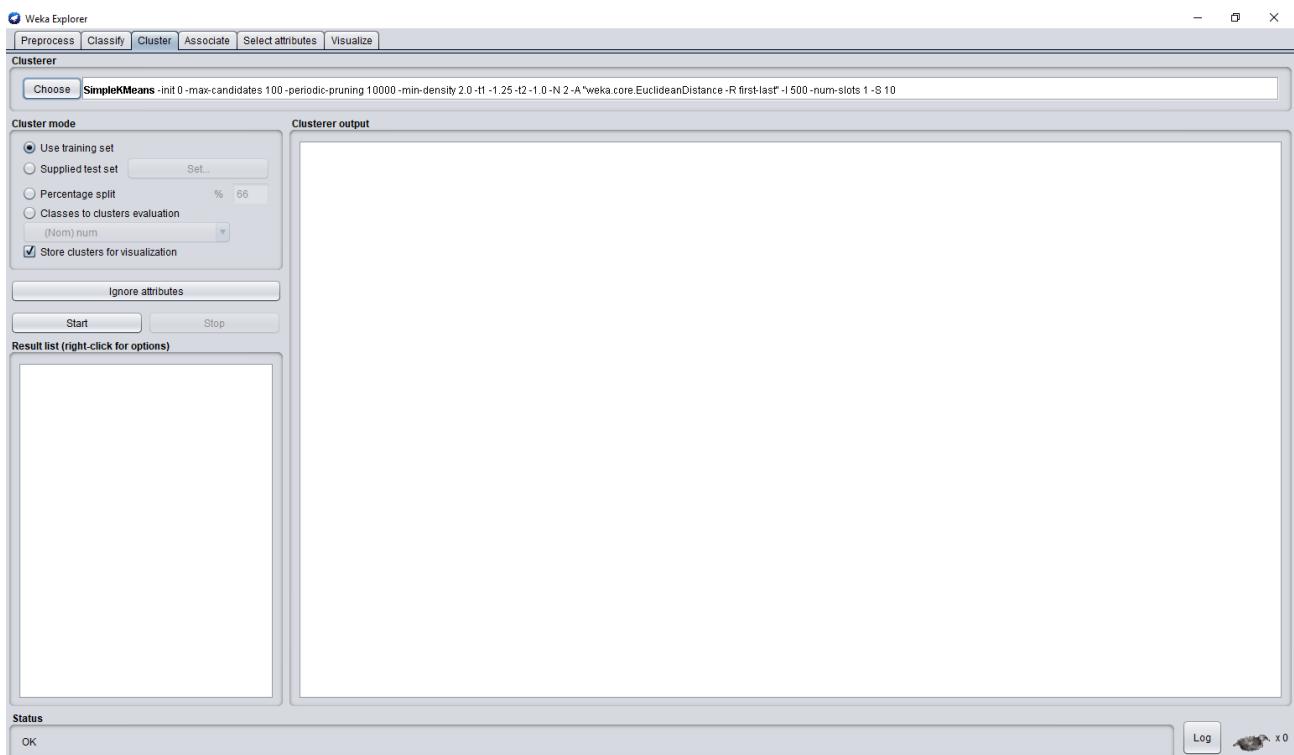
b) (1đ) Liệt kê các phương pháp đã học để loại bỏ dữ liệu nhiễu (*noisy data*). WEKA hỗ trợ những phương pháp nào cho vấn đề này?

- Các phương pháp đã học để loại bỏ dữ liệu nhiễu:
  - Phương pháp chia giỏ (Binning)
  - Phương pháp gom nhóm (Clustering)
  - Phương pháp hồi qui (Regression)
- WEKA hỗ trợ những phương pháp:
  - Discretize: Rời rạc hóa dữ liệu

# Lab01-Preprocessing

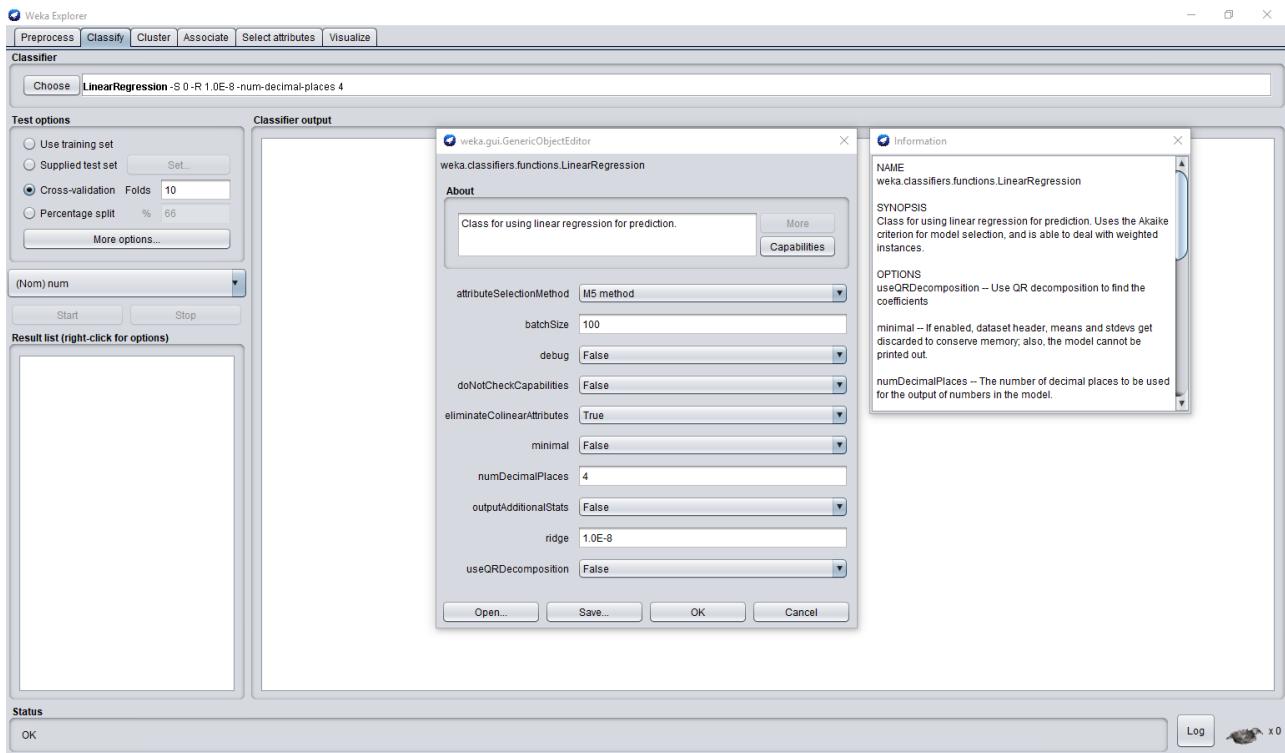


- Clusterers: Gom nhóm



- Regression Algorithms: Các thuật toán hồi quy hỗ trợ trong weka
  - ✓ Linear Regression
  - ✓ k-Nearest Neighbors
  - ✓ Decision Tree
  - ✓ Support Vector Machines
  - ✓ Multi-Layer Perceptron

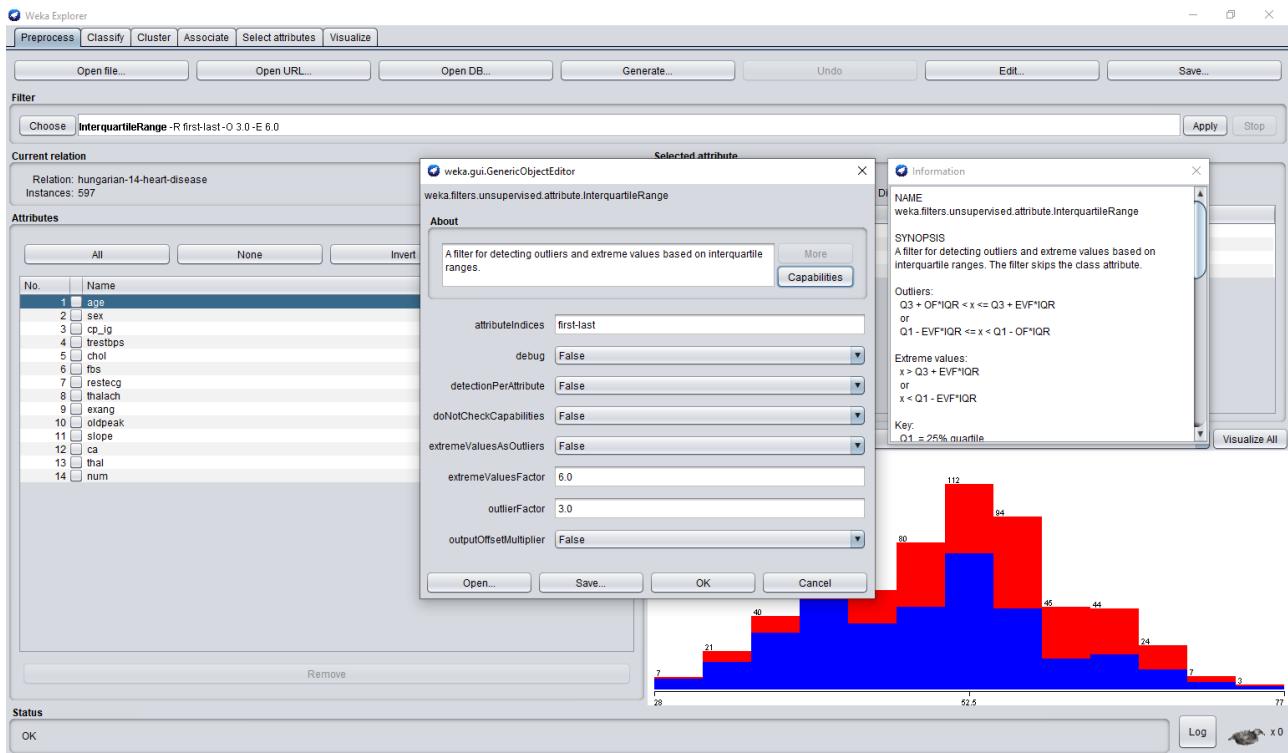
## Lab01-Preprocessing



c) (1đ) Liệt kê các phương pháp đã học để dò tìm dữ liệu tạp (*outlier detection*). WEKA hỗ trợ những phương pháp nào cho vấn đề này?

- Các phương pháp để dò tìm dữ liệu tạp:
  - Partitioning Methods: Phân vùng dữ liệu
  - Hierarchical Methods: Phân cấp dữ liệu
  - Classification Approach: Phương pháp gom nhóm
  - Statistical Approach: Phương pháp thống kê
  - Proximity based Approach: Phương pháp dựa trên mối liên quan
- WEKA hỗ trợ những phương pháp:
  - InterquartileRange: Phát hiện các ngoại lệ và cực trị

## Lab01-Preprocessing



d) (1đ) Tập dữ liệu có gặp phải các vấn đề nêu trên hay không? Nếu có, liệt kê một số giá trị đại diện cho từng trường hợp và mô tả lựa chọn của bạn để giải quyết vấn đề (bạn có thể chọn bộ lọc của WEKA hoặc tự đề xuất phương pháp riêng).

- Tập dữ liệu có gặp phải các vấn đề nêu trên
  - Missing values – thiếu giá trị:  
→ Sử dụng ReplaceMissingValues trong weka

## Lab01-Preprocessing

trestbps	chol	fb	restecg	thalach	exang	oldpeak	slope	ca	thal	num
130	132	f	left_vent_	185	no	0?	?	?	?	<50
120	243	f	normal	160	no	0?	?	?	?	<50
140	?	f	normal	170	no	0?	?	?	?	<50
170	237	f	st_t_wave	170	no	0?	?	?	fixed_def	<50
100	219	f	st_t_wave	150	no	0?	?	?	?	<50
105	198	f	normal	165	no	0?	?	?	?	<50
110	225	f	normal	184	no	0?	?	?	?	<50
125	254	f	normal	155	no	0?	?	?	?	<50
120	298	f	normal	185	no	0?	?	?	?	<50
130	161	f	normal	190	no	0?	?	?	?	<50
150	214	f	st_t_wave	168	no	0?	?	?	?	<50
98	220	f	normal	150	no	0?	?	?	?	<50
120	160	f	st_t_wave	185	no	0?	?	?	?	<50
140	167	f	normal	150	no	0?	?	?	?	<50
120	308	f	left_vent_	180	no	0?	?	?	?	<50
150	264	f	normal	168	no	0?	?	?	?	<50
120	166	f	normal	180	no	0?	?	?	?	<50
112	340	f	normal	184	no	1 flat	?	normal	?	<50
130	209	f	normal	178	no	0?	?	?	?	<50
150	160	f	normal	172	no	0?	?	?	?	<50
120	260	f	normal	130	no	0?	?	?	?	<50
130	211	f	normal	142	no	0?	?	?	?	<50
130	173	f	st_t_wave	184	no	0?	?	?	?	<50
130	283	f	st_t_wave	98	no	0?	?	?	?	<50
130	194	f	normal	150	no	0?	?	?	?	<50
120	223	f	normal	168	no	0?	?	?	normal	<50
130	315	f	normal	158	no	0?	?	?	?	<50
120	275	?	normal	129	no	0?	?	?	?	<50
140	297	f	normal	150	no	0?	?	?	?	<50
145	292	f	normal	130	no	0?	?	?	?	<50
110	182	f	st_t_wave	180	no	0?	?	?	?	<50

- Noisy data – dữ liệu nhiễu:
- Sử dụng Discretize trong weka

## Lab01-Preprocessing

1	age	sex	cp_ig	trestbps	chol	fbs	restecg	thalach	exang	oldpea	slope	ca	thal	num
546	50	male	asympt	140	341	f	st_t_wave_abnormality	125	yes	2.5	flat	?	?	>50_1
547	43	female	asympt	132	341	t	left_vent_hyper	136	yes	3	flat	0	reversible	>50_1
548	52	male	asympt	112	342	f	st_t_wave_abnormality	96	yes	1	flat	?	?	>50_1
549	56	male	asympt	155	342	t	normal	150	yes	3	flat	?	?	>50_1
550	55	female	atyp_angina	132	342	f	normal	166	no	1.2	up	0	normal	<50
551	55	female	atyp_angina	110	344	f	st_t_wave_abnormality	160	no	0	?	?	?	<50
552	57	female	asympt	180	347	f	st_t_wave_abnormality	126	yes	0.8	flat	?	?	<50
553	55	male	asympt	132	353	f	normal	132	yes	1.2	flat	1	reversible	>50_1
554	57	female	asympt	120	354	f	normal	163	yes	0.6	up	0	normal	<50
555	48	male	asympt	160	355	f	normal	99	yes	2	flat	?	?	>50_1
556	42	male	asympt	140	358	f	normal	170	no	0	?	?	?	<50
557	65	female	non_anginal	160	360	f	left_vent_hyper	151	no	0.8	up	0	normal	<50
558	54	male	asympt	150	365	f	st_t_wave_abnormality	134	no	1	up	?	?	<50
559	56	male	asympt	170	388	f	st_t_wave_abnormality	122	yes	2	flat	?	?	>50_1
560	40	female	asympt	150	392	f	normal	130	no	2	flat	?	fixed_def	>50_1
561	58	female	atyp_angina	180	393	f	normal	110	yes	1	flat	?	reversible	>50_1
562	55	female	atyp_angina	130	394	f	left_vent_hyper	150	no	0	?	?	?	<50
563	62	female	asympt	140	394	f	left_vent_hyper	157	no	1.2	flat	0	normal	<50
564	52	male	asympt	140	404	f	normal	124	yes	2	flat	?	?	>50_1
565	63	female	asympt	150	407	f	left_vent_hyper	154	no	4	flat	3	reversible	>50_1
566	56	female	asympt	134	409	f	left_vent_hyper	150	yes	1.9	flat	2	reversible	>50_1
567	44	male	asympt	150	412	f	normal	170	no	0	?	?	?	<50
568	65	female	non_anginal	140	417	t	left_vent_hyper	157	no	0.8	up	1	normal	<50
569	40	male	asympt	120	466	?	normal	152	yes	1	flat	?	fixed_def	>50_1
570	53	female	atyp_angina	113	468	?	normal	127	no	0	?	?	?	<50
571	44	male	asympt	135	491	f	normal	135	no	0	?	?	?	>50_1
572	53	male	non_anginal	145	518	f	normal	130	no	0	?	?	?	>50_1
573	32	male	asympt	118	529	f	normal	130	no	0	?	?	?	>50_1
574	67	female	non_anginal	115	564	f	left_vent_hyper	160	no	1.6	flat	0	reversible	<50
575	54	male	asympt	130	603	t	normal	125	yes	1	flat	?	?	>50_1
576	66	male	asympt	140	?	f	normal	94	yes	1	flat	?	?	>50_1

chol - serum cholesterol in mg/dl quá cao

- Outlier detection - dữ liệu tạp:
- Sử dụng InterquartileRange trong weka

## Lab01-Preprocessing

A	B	C	D	E	F	G	H	I	J	K	L	M	N
454	60	male	asympt	140	293 f	left_vent	170	no	1.2	flat	2	reversabl<50_1	
455	68	male	non_angir	118	277 f	normal	151	no	1	up	1	reversabl<50	
456	46	male	atyp_angi	101	197 t	normal	156	no	0	up	0	reversabl<50	
457	77	male	asympt	125	304 f	left_vent	162	yes	0	up	3	normal >50_1	
458	54	female	non_angir	110	214 f	normal	158	no	1.6	flat	0	normal <50	
459	58	female	asympt	100	248 f	left_vent	122	no	1	flat	0	normal <50	
460	48	male	non_angir	124	255 t	normal	175	no	0	up	2	normal <50	
461	57	male	asympt	132	207 f	normal	168	yes	0	up	0	reversabl<50	
462	52	male	non_angir	138	223 f	normal	169	no	0	up	?	normal <50	
463	54	female	atyp_angi	132	288 t	left_vent	159	yes	0	up	1	normal <50	
464	35	male	asympt	126	282 f	left_vent	156	yes	0	up	0	reversabl<50_1	
465	45	female	atyp_angi	112	160 f	normal	138	no	0	flat	0	normal <50	
466	70	male	non_angir	160	269 f	normal	112	yes	2.9	flat	1	reversabl<50_1	
467	53	male	asympt	142	226 f	left_vent	111	yes	0	up	0	reversabl<50	
468	59	female	asympt	174	249 f	normal	143	yes	0	flat	0	normal >50_1	
469	62	female	asympt	140	394 f	left_vent	157	no	1.2	flat	0	normal <50	
470	64	male	asympt	145	212 f	left_vent	132	no	2	flat	2	fixed_def>50_1	
471	57	male	asympt	152	274 f	normal	88	yes	1.2	flat	1	reversabl<50_1	
472	52	male	asympt	108	233 t	normal	147	no	0.1	up	3	reversabl<50	
473	56	male	asympt	132	184 f	left_vent	105	yes	2.1	flat	1	fixed_def>50_1	
474	43	male	non_angir	130	315 f	normal	162	no	1.9	up	1	normal <50	
475	53	male	non_angir	130	246 t	left_vent	173	no	0	up	3	normal <50	
476	48	male	asympt	124	274 f	left_vent	166	no	0.5	flat	0	reversabl<50_1	
477	56	female	asympt	134	409 f	left_vent	150	yes	1.9	flat	2	reversabl<50_1	
478	42	male	typ_angin	148	244 f	left_vent	178	no	0.8	up	2	normal <50	
479	59	male	typ_angin	178	270 f	left_vent	145	no	4.2	down	0	reversabl<50	
480	63	female	atyp_angi	140	195 f	normal	179	no	0	up	2	normal <50	
481	42	male	non_angir	120	240 t	normal	194	no	0.8	down	0	reversabl<50	
482	66	male	atyp_angi	160	246 f	normal	120	yes	0	flat	3	fixed_def>50_1	
483	54	male	atyp_angi	192	283 f	left_vent	195	no	0	up	1	reversabl<50_1	
484	69	male	non_angir	140	254 f	left_vent	146	no	2	flat	3	reversabl<50_1	
485	50	male	non_angir	129	196 f	normal	163	no	0	up	0	normal <50	
486	51	male	asympt	140	298 f	normal	122	yes	4.2	flat	3	reversabl<50_1	
487	43	male	asympt	132	247 t	left_vent	143	yes	0.1	flat	?	reversabl<50_1	
488	62	female	asympt	138	294 t	normal	106	no	1.9	flat	3	normal >50_1	
489	68	female	non_angir	120	211 f	left_vent	115	no	1.5	flat	0	normal <50	

e) (1đ) Lưu dữ liệu đã làm sạch vào tập tin **heart-cleaned.arff**. Chụp hình các phần dữ liệu có sự thay đổi trước và sau khi làm sạch.

- Làm sạch bằng ReplaceMissingValues trong weka

D	E	F	G	H	I	J	K	L	M	N	1	D	E	F	G	H	I	J	K	L	M	N
trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num	1	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
130	132 f	left_vent	185	no	0?	?	?	?	<50		2	130	132 f	left_vent	185	no	0	flat	0.66774	normal	<50	
120	243 f	normal	160	no	0?	?	?	?	<50		3	120	243 f	normal	160	no	0	flat	0.66774	normal	<50	
140	?	normal	170	no	0?	?	?	?	<50		4	140	248,4286 f	normal	170	no	0	flat	0.66774	normal	<50	
170	237 f	st_t_wave	170	no	0?	?	?	?	fixed_def	<50	5	170	237 f	st_t_wave	170	no	0	flat	0.66774	fixed_def	<50	
100	219 f	st_t_wave	150	no	0?	?	?	?	<50		6	100	219 f	st_t_wave	150	no	0	flat	0.66774	normal	<50	
105	198 f	normal	165	no	0?	?	?	?	<50		7	105	198 f	normal	165	no	0	flat	0.66774	normal	<50	
110	225 f	normal	184	no	0?	?	?	?	<50		8	110	225 f	normal	184	no	0	flat	0.66774	normal	<50	
125	254 f	normal	155	no	0?	?	?	?	<50		9	125	254 f	normal	155	no	0	flat	0.66774	normal	<50	
120	298 f	normal	185	no	0?	?	?	?	<50		10	120	298 f	normal	185	no	0	flat	0.66774	normal	<50	
130	161 f	normal	190	no	0?	?	?	?	<50		11	130	161 f	normal	190	no	0	flat	0.66774	normal	<50	
150	214 f	st_t_wave	168	no	0?	?	?	?	<50		12	150	214 f	st_t_wave	168	no	0	flat	0.66774	normal	<50	
98	220 f	normal	150	no	0?	?	?	?	<50		13	98	220 f	normal	150	no	0	flat	0.66774	normal	<50	
120	160 f	st_t_wave	185	no	0?	?	?	?	<50		14	120	160 f	st_t_wave	185	no	0	flat	0.66774	normal	<50	
140	167 f	normal	150	no	0?	?	?	?	<50		15	140	167 f	normal	150	no	0	flat	0.66774	normal	<50	
120	308 f	left_vent	180	no	0?	?	?	?	<50		16	120	308 f	left_vent	180	no	0	flat	0.66774	normal	<50	
150	264 f	normal	168	no	0?	?	?	?	<50		17	150	264 f	normal	168	no	0	flat	0.66774	normal	<50	
120	166 f	normal	180	no	0?	?	?	?	<50		18	120	166 f	normal	180	no	0	flat	0.66774	normal	<50	
112	340 f	normal	184	no	1	flat	?	?	normal		19	112	340 f	normal	184	no	1	flat	0.66774	normal	<50	
130	209 f	normal	178	no	0?	?	?	?	<50		20	130	209 f	normal	178	no	0	flat	0.66774	normal	<50	
150	160 f	normal	172	no	0?	?	?	?	<50		21	150	160 f	normal	172	no	0	flat	0.66774	normal	<50	
120	260 f	normal	130	no	0?	?	?	?	<50		22	120	260 f	normal	130	no	0	flat	0.66774	normal	<50	
130	211 f	normal	142	no	0?	?	?	?	<50		23	130	211 f	normal	142	no	0	flat	0.66774	normal	<50	
130	173 f	st_t_wave	184	no	0?	?	?	?	<50		24	130	173 f	st_t_wave	184	no	0	flat	0.66774	normal	<50	
130	283 f	st_t_wave	98	no	0?	?	?	?	<50		25	130	283 f	st_t_wave	98	no	0	flat	0.66774	normal	<50	
130	194 f	normal	150	no	0?	?	?	?	<50		26	130	194 f	normal	150	no	0	flat	0.66774	normal	<50	
120	223 f	normal	168	no	0?	?	?	?	<50		27	120	223 f	normal	168	no	0	flat	0.66774	normal	<50	
130	315 f	normal	158	no	0?	?	?	?	<50		28	130	315 f	normal	158	no	0	flat	0.66774	normal	<50	
120	275 ?	normal	129	no	0?	?	?	?	<50		29	120	275 f	normal	129	no	0	flat	0.66774	normal	<50	
140	297 f	normal	150	no	0?	?	?	?	<50		30	140	297 f	normal	150	no	0	flat	0.66774	normal	<50	
145	292 f	normal	130	no	0?	?	?	?	<50		31	145	292 f	normal	130	no	0	flat	0.66774	normal	<50	
110	182 f	st_t_wave	180	no	0?	?	?	?	<50		32	110	182 f	st_t_wave	180	no	0	flat	0.66774	normal	<50	

# Lab01-Preprocessing

- Làm sạch bằng Discretize trong weka

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1 age	sex	cp_ig	trestbps	chol	fb	restecg	thalach	exang	oldpeak	slope	ca	thal	num		
2 28	male	atyp_angi	130	132	f	left_vent	185	no	0	flat	0	normal	0		
3 29	male	atyp_angi	120	243	f	normal	160	no	0	flat	0	normal	0		
4 29	male	atyp_angi	140	248.426	f	normal	170	no	0	flat	0	normal	0		
5 30	female	typ_angin	170	237	f	st_t_wave	170	no	0	flat	0	male	5	'\(-inf-32.9)''	'\(-inf-32.9)''
6 31	female	atyp_angi	100	219	f	st_t_wave	150	no	0	flat	0	female	6	'\(-inf-32.9)''	'\(-inf-32.9)''
7 32	female	atyp_angi	105	198	f	normal	165	no	0	flat	0	female	7	'\(-inf-32.9)''	'\(-inf-32.9)''
8 32	male	atyp_angi	110	225	f	normal	184	no	0	flat	0	male	8	'\(-inf-32.9)''	'\(-inf-32.9)''
9 32	male	atyp_angi	125	254	f	normal	155	no	0	flat	0	male	9	'\(-inf-32.9)''	'\(-inf-32.9)''
10 33	male	non_angiu	120	298	f	normal	185	no	0	flat	0	male	10	'\(-inf-32.9)''	'\(-inf-32.9)''
11 34	female	atyp_angi	130	161	f	normal	190	no	0	flat	0	female	11	'\(-inf-32.9)''	'\(-inf-32.9)''
12 34	male	atyp_angi	150	214	f	st_t_wave	168	no	0	flat	0	male	12	'\(-inf-32.9)''	'\(-inf-32.9)''
13 34	male	atyp_angi	98	220	f	normal	150	no	0	flat	0	male	13	'\(-inf-32.9)''	'\(-inf-32.9)''
14 35	female	typ_angin	120	160	f	st_t_wave	185	no	0	flat	0	female	14	'\(-inf-32.9)''	'\(-inf-32.9)''
15 35	female	asympt	140	167	f	normal	150	no	0	flat	0	female	15	'\(-inf-32.9)''	'\(-inf-32.9)''
16 35	male	atyp_angi	120	308	f	left_vent	180	no	0	flat	0	male	16	'\(-inf-32.9)''	'\(-inf-32.9)''
17 35	male	atyp_angi	150	264	f	normal	168	no	0	flat	0	male	17	'\(-inf-32.9)''	'\(-inf-32.9)''
18 36	male	atyp_angi	120	166	f	normal	180	no	0	flat	0	male	18	'\(-inf-32.9)''	'\(-inf-32.9)''
19 36	male	non_angiu	112	340	f	normal	184	no	1	flat	0	male	19	'\(-inf-32.9)''	'\(-inf-32.9)''
20 36	male	non_angiu	130	209	f	normal	178	no	0	flat	0	male	20	'\(-inf-32.9)''	'\(-inf-32.9)''
21 36	male	non_angiu	150	160	f	normal	172	no	0	flat	0	male	21	'\(-inf-32.9)''	'\(-inf-32.9)''
22 37	female	atyp_angi	120	260	f	normal	130	no	0	flat	0	female	22	'\(-inf-32.9)''	'\(-inf-32.9)''
23 37	female	non_angiu	130	211	f	normal	142	no	0	flat	0	female	23	'\(-inf-32.9)''	'\(-inf-32.9)''
24 37	female	asympt	130	173	f	st_t_wave	184	no	0	flat	0	female	24	'\(-inf-32.9)''	'\(-inf-32.9)''
25 37	male	atyp_angi	130	283	f	st_t_wave	98	no	0	flat	0	male	25	'\(-inf-32.9)''	'\(-inf-32.9)''
26 37	male	non_angiu	130	194	f	normal	150	no	0	flat	0	male	26	'\(-inf-32.9)''	'\(-inf-32.9)''
27 37	male	asympt	120	223	f	normal	168	no	0	flat	0	male	27	'\(-inf-32.9)''	'\(-inf-32.9)''
28 37	male	asympt	130	315	f	normal	158	no	0	flat	0	male	28	'\(-inf-32.9)''	'\(-inf-32.9)''
29 38	female	atyp_angi	120	275	f	normal	129	no	0	flat	0	female	29	'\(-inf-32.9)''	'\(-inf-32.9)''
30 38	male	atyp_angi	140	297	f	normal	150	no	0	flat	0	male	30	'\(-inf-32.9)''	'\(-inf-32.9)''
31 38	male	non_angiu	145	292	f	normal	130	no	0	flat	0	male	31	'\(-inf-32.9)''	'\(-inf-32.9)''
32 39	female	non_angiu	110	182	f	st_t_wave	180	no	0	flat	0	female	32	'\(-inf-32.9)''	'\(-inf-32.9)''

- Làm sạch bằng InterquartileRange trong weka

U	V	E	F	G	H	I	J	K	L	M	N	O	P	
1 trestbps	chol	fb	restecg	thalach	exang	oldpeak	slope	ca	thal	num				
2 '\(124.4-1)'\(-inf-13)f	left_vent	'\(-inf-13)f												
3 '\(113.6-1)'\(240.4-2)f	normal	'\(-inf-13)f												
4 '\(135.2-1)'\(240.4-2)f	normal	'\(-inf-13)f												
5 '\(167.6-1)'\(188.6-2)f	st_t_wave	'\(-inf-13)f												
6 '\(inf-10)'\(188.6-2)f	st_t_wave	'\(-inf-13)f												
7 '\(102.8-1)'\(188.6-2)f	normal	'\(-inf-13)f												
8 '\(102.8-1)'\(188.6-2)f	normal	'\(-inf-13)f												
9 '\(124.4-1)'\(240.4-2)f	normal	'\(-inf-13)f												
10 '\(113.6-1)'\(292.2-3)f	normal	'\(-inf-13)f												
11 '\(124.4-1)'\(136.8-1)f	normal	'\(-inf-13)f												
12 '\(146-15c)'\(188.6-2)f	st_t_wave	'\(-inf-13)f												
13 '\(inf-10)'\(188.6-2)f	normal	'\(-inf-13)f												
14 '\(113.6-1)'\(136.8-1)f	st_t_wave	'\(-inf-13)f												
15 '\(135.2-1)'\(136.8-1)f	normal	'\(-inf-13)f												
16 '\(113.6-1)'\(292.2-3)f	left_vent	'\(-inf-13)f												
17 '\(146-15c)'\(240.4-2)f	normal	'\(-inf-13)f												
18 '\(113.6-1)'\(136.8-1)f	normal	'\(-inf-13)f												
19 '\(102.8-1)'\(292.2-3)f	normal	'\(-inf-13)f												
20 '\(124.4-1)'\(188.6-2)f	normal	'\(-inf-13)f												
21 '\(146-15c)'\(136.8-1)f	normal	'\(-inf-13)f												
22 '\(113.6-1)'\(240.4-2)f	normal	'\(-inf-13)f												
23 '\(124.4-1)'\(188.6-2)f	normal	'\(-inf-13)f												
24 '\(124.4-1)'\(136.8-1)f	st_t_wave	'\(-inf-13)f												
25 '\(124.4-1)'\(240.4-2)f	st_t_wave	'\(-inf-13)f												
26 '\(124.4-1)'\(188.6-2)f	normal	'\(-inf-13)f												
27 '\(113.6-1)'\(188.6-2)f	normal	'\(-inf-13)f												
28 '\(124.4-1)'\(292.2-3)f	normal	'\(-inf-13)f												
29 '\(113.6-1)'\(240.4-2)f	normal	'\(-inf-13)f												
30 '\(135.2-1)'\(292.2-3)f	normal	'\(-inf-13)f												
31 '\(135.2-1)'\(240.4-2)f	normal	'\(-inf-13)f												
32 '\(102.8-1)'\(136.8-1)f	st_t_wave	'\(-inf-13)f												

Cần nộp lại tập tin **heart-cleaned.arff**. Tập tin sai định dạng hoặc thiếu dữ liệu (-1đ). Không nộp tập tin (-2đ).

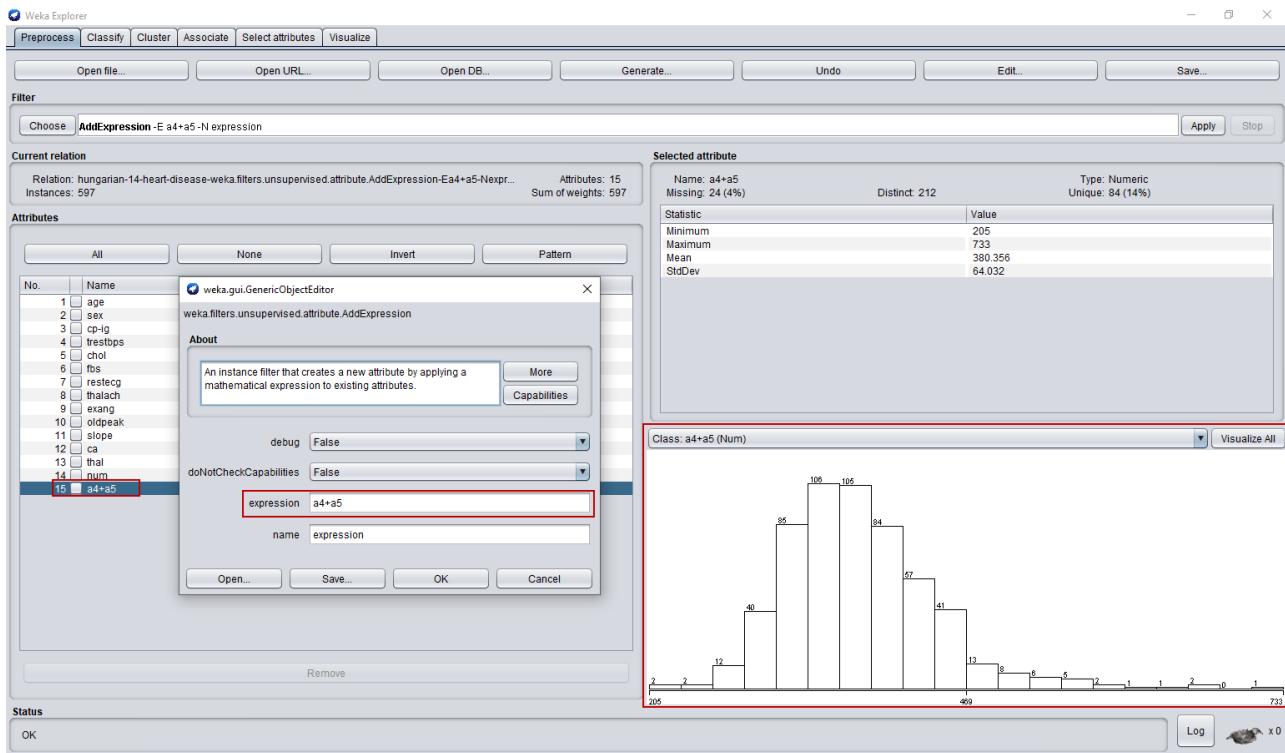
File đính kèm

## 5) Chuyển đổi dữ liệu (Transformation) (3.0 điểm)

Tìm hiểu các bộ lọc của WEKA hỗ trợ cho vấn đề chuyển đổi dữ liệu. Làm việc trên tập tin **heart-cleaned.arff**.

- (1đ) Bộ lọc nào của WEKA cho phép xây dựng thuộc tính (attribute construction), ví dụ, thêm một thuộc tính là tổng của 2 thuộc tính khác?
- Các bộ lọc Weka cho phép xây dựng thuộc tính (attribute construction):
  - Add: Thêm thuộc tính
  - AddID: Thêm thuộc tính ID
  - AddExpression: Thêm thuộc tính dựa trên các biểu thức chính quy

## Lab01-Preprocessing

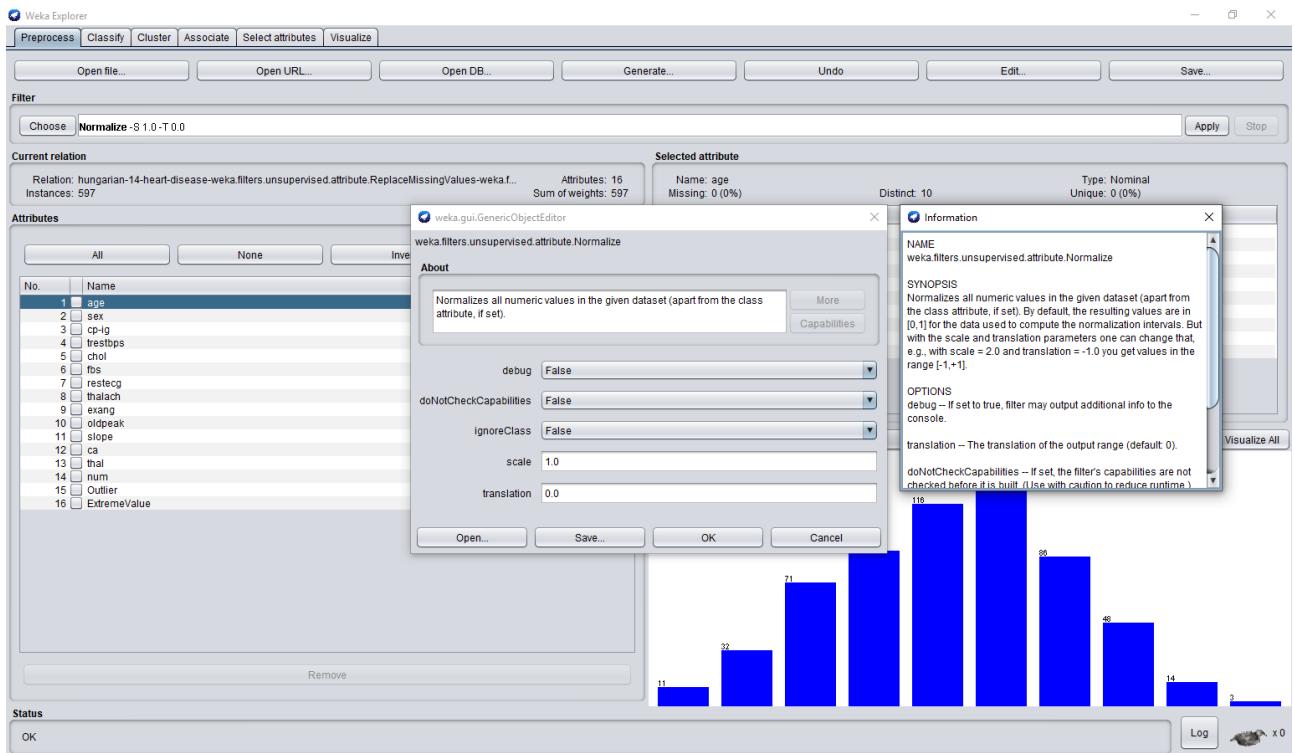


Thực hiện trên **heart-cleaned.arff** chưa được Discretize và InterquartileRange

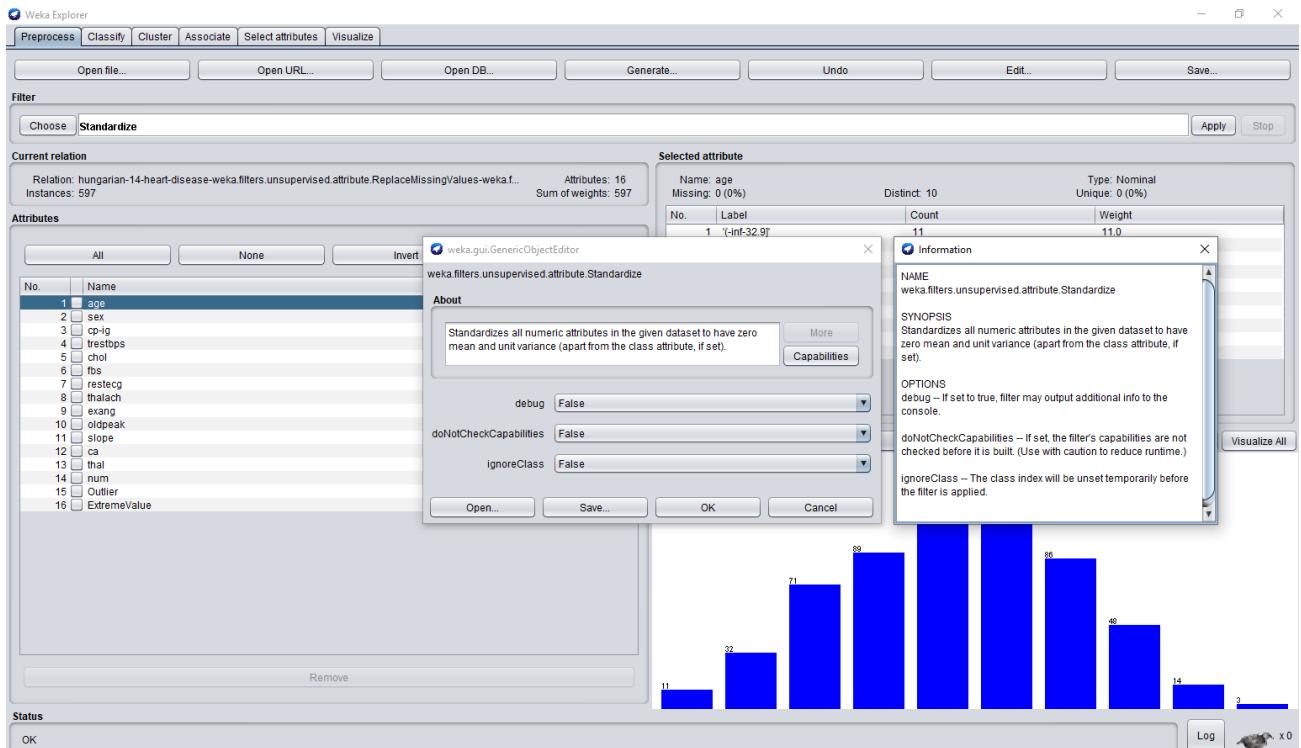
Thêm thuộc tính mới là tổng của cho là trestbps và chol

- b) (1đ) Bộ lọc nào của WEKA cho phép chuẩn hóa thuộc tính (*normalization*)? Bộ lọc này có thể chuẩn hóa Min-max, chuẩn hóa Z-score hay chuẩn hóa thập phân không? Nếu có, cho biết cụ thể cách thực hiện những chuẩn hóa này trong WEKA. Nếu không, mô tả giải pháp chuẩn hóa mà WEKA hỗ trợ.
- Bộ lọc WEKA cho phép chuẩn hóa thuộc tính (*normalization*) là
    - Normalize: Chuẩn hóa min - max

## Lab01-Preprocessing

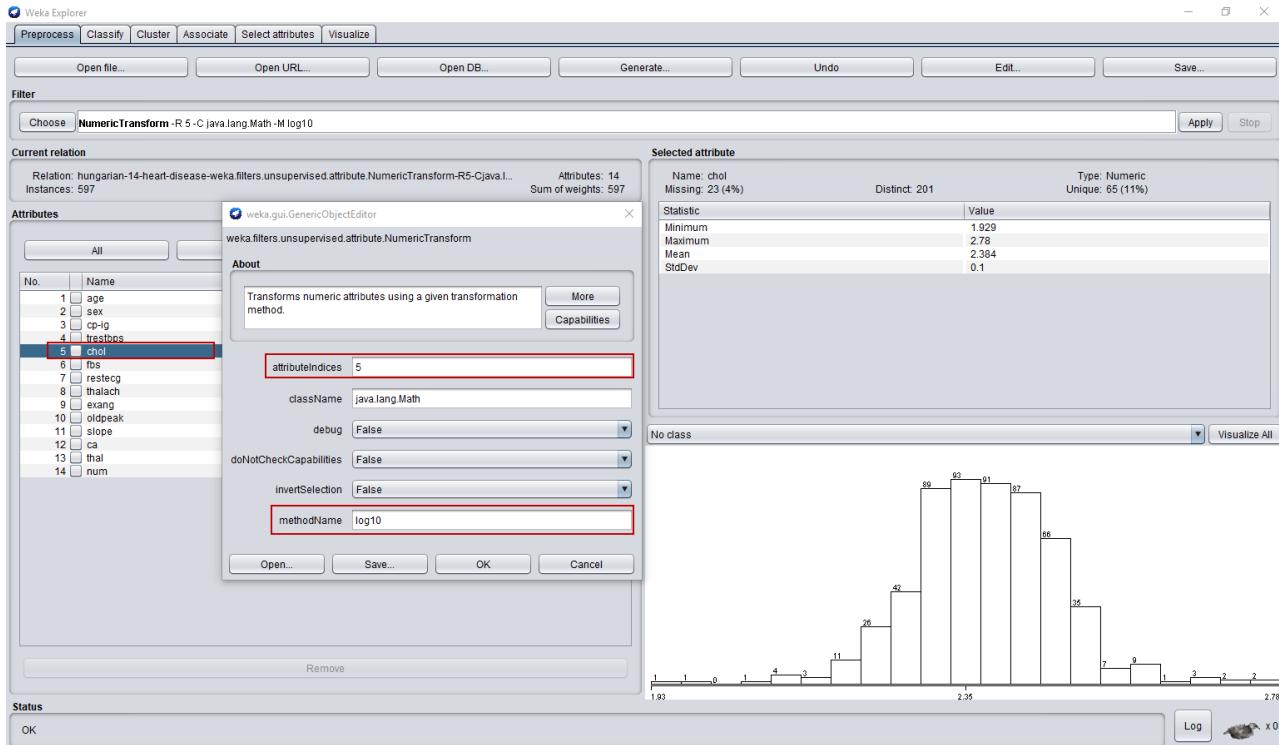


- ✓ scale: giới hạn
- ✓ translation: giá trị nhỏ nhất
- Standardize: Chuẩn hóa Z-score



- NumericTransform: Hỗ trợ chuẩn hóa thập

# Lab01-Preprocessing



- ✓ attributeIndices: Thuộc tính cần chuẩn hóa
- ✓ methodName: Phương thức chuẩn hóa

c) (1đ) Chọn một bộ lọc chuẩn hóa trong WEKA và tiến hành chuẩn hóa tất cả các thuộc tính là số thực. Lưu dữ liệu đã chuẩn hóa vào tập tin **heart-normal.arff**. Chụp hình ít nhất 10 dòng dữ liệu với tất cả thuộc tính số thực để thể hiện rõ sự thay đổi sau chuẩn hóa.

Thực hiện trên **heart-cleaned.arff** chưa được Discretize và InterquartileRange

Chọn Normalize (Chuẩn hóa min – max)

Chưa chuẩn hóa														Đã chuẩn hóa Normalize													
A	B	C	D	E	F	G	H	I	J	K	ca	A	B	C	D	E	F	G	H	I	J	K	ca				
1	age	sex	cp_ig	trestbps	chol	fb	restecg	thalach	exang	oldpeak	slope	0	1	0 male	atyp_angi	0.351852	0.090734 f	restecg	thalach	exang	oldpeak	slope	0	0			
2	28	male	atyp_angi	130	132 f	left_vent.	185	no	0 flat	0	0	0	2	0 female	atyp_angi	0.351852	0.090734 f	left_vent.	thalach	exang	oldpeak	slope	0	0			
3	29	male	atyp_angi	120	243 f	normal	160	no	0 flat	0	0	0	3	0.020408	male	atyp_angi	0.359259	0.030519 f	normal	0.679389	no	0 flat	0	0			
4	29	male	atyp_angi	140	248.4286 f	normal	170	no	0 flat	0	0	0	4	0.020408	male	atyp_angi	0.444444	0.315499 f	normal	0.755725	no	0 flat	0	0			
5	30	female	typ_angi	170	237 f	st_t_wave	170	no	0 flat	0	0	0	5	0.040816	female	typ_angi	0.722222	0.293436 f	st_t_wave	0.755725	no	0 flat	0	0			
6	31	female	atyp_angi	100	219 f	st_t_wave	150	no	0 flat	0	0	0	6	0.061224	female	atyp_angi	0.074074	0.238687 f	st_t_wave	0.603053	no	0 flat	0	0			
7	32	female	atyp_angi	105	198 f	normal	165	no	0 flat	0	0	0	7	0.081633	female	atyp_angi	0.12037	0.218147 f	normal	0.717557	no	0 flat	0	0			
8	32	male	atyp_angi	110	225 f	normal	184	no	0 flat	0	0	0	8	0.081633	male	atyp_angi	0.166667	0.270272 f	normal	0.862595	no	0 flat	0	0			
9	32	male	atyp_angi	125	254 f	normal	155	no	0 flat	0	0	0	9	0.081633	male	atyp_angi	0.305556	0.326255 f	normal	0.641221	no	0 flat	0	0			
10	33	male	non_angi	120	298 f	normal	185	no	0 flat	0	0	0	10	0.102041	male	non_angi	0.259259	0.411197 f	normal	0.870229	no	0 flat	0	0			
11	34	female	atyp_angi	130	161 f	normal	190	no	0 flat	0	0	0	11	0.122449	female	atyp_angi	0.351852	0.146718 f	normal	0.908397	no	0 flat	0	0			
12	34	male	atyp_angi	150	214 f	st_t_wave	168	no	0 flat	0	0	0	12	0.122449	male	atyp_angi	0.357037	0.249035 f	st_t_wave	0.740458	no	0 flat	0	0			
13	34	male	atyp_angi	98	220 f	normal	150	no	0 flat	0	0	0	13	0.122449	male	atyp_angi	0.055556	0.260618 f	normal	0.603053	no	0 flat	0	0			
14	35	female	typ_angi	120	160 f	st_t_wave	185	no	0 flat	0	0	0	14	0.142857	female	typ_angi	0.259259	0.144788 f	st_t_wave	0.870229	no	0 flat	0	0			
15	35	female	asympt	140	167 f	normal	150	no	0 flat	0	0	0	15	0.142857	female	asympt	0.444444	0.158301 f	normal	0.603053	no	0 flat	0	0			
16	35	male	atyp_angi	120	308 f	left_vent.	180	no	0 flat	0	0	0	16	0.142857	male	atyp_angi	0.259259	0.430502 f	left_vent.	0.832061	no	0 flat	0	0			
17	35	male	atyp_angi	150	264 f	normal	168	no	0 flat	0	0	0	17	0.142857	male	atyp_angi	0.537037	0.34556 f	normal	0.740458	no	0 flat	0	0			
18	36	male	atyp_angi	120	166 f	normal	180	no	0 flat	0	0	0	18	0.163265	male	atyp_angi	0.351852	0.156371 f	normal	0.832061	no	0 flat	0	0			
19	36	male	non_angi	112	340 f	normal	184	no	1 flat	0	0	0	19	0.163265	male	non_angi	0.185185	0.492278 f	normal	0.862595	no	0.161229	flat	0			
20	36	male	non_angi	130	209 f	normal	178	no	0 flat	0	0	0	20	0.163265	male	non_angi	0.351852	0.239382 f	normal	0.816794	no	0 flat	0	0			
21	36	male	non_angi	150	160 f	normal	172	no	0 flat	0	0	0	21	0.163265	male	non_angi	0.37037	0.144788 f	normal	0.770992	no	0 flat	0	0			
22	37	female	atyp_angi	120	260 f	normal	130	no	0 flat	0	0	0	22	0.183673	female	atyp_angi	0.259259	0.337838 f	normal	0.450382	no	0 flat	0	0			
23	37	female	non_angi	130	211 f	normal	142	no	0 flat	0	0	0	23	0.183673	female	non_angi	0.351852	0.243243 f	normal	0.541985	no	0 flat	0	0			
24	37	female	asympt	130	173 f	st_t_wave	184	no	0 flat	0	0	0	24	0.183673	female	asympt	0.351852	0.169884 f	st_t_wave	0.862595	no	0 flat	0	0			
25	37	male	atyp_angi	130	283 f	st_t_wave	98	no	0 flat	0	0	0	25	0.183673	male	atyp_angi	0.351852	0.382239 f	st_t_wave	0.206107	no	0 flat	0	0			
26	37	male	non_angi	130	194 f	normal	150	no	0 flat	0	0	0	26	0.183673	male	non_angi	0.351852	0.210425 f	normal	0.603053	no	0 flat	0	0			
27	37	male	asympt	120	223 f	normal	168	no	0 flat	0	0	0	27	0.183673	male	asympt	0.259259	0.266409 f	normal	0.740458	no	0 flat	0	0			
28	37	male	asympt	130	315 f	normal	158	no	0 flat	0	0	0	28	0.183673	male	asympt	0.351852	0.444015 f	normal	0.664122	no	0 flat	0	0			
29	38	female	atyp_angi	120	275 f	normal	129	no	0 flat	0	0	0	29	0.204082	female	atyp_angi	0.259259	0.336795 f	normal	0.442748	no	0 flat	0	0			
30	38	male	atyp_angi	140	297 f	normal	150	no	0 flat	0	0	0	30	0.204082	male	atyp_angi	0.444444	0.409266 f	normal	0.603053	no	0 flat	0	0			
31	38	male	non_angi	145	292 f	normal	130	no	0 flat	0	0	0	31	0.204082	male	non_angi	0.490741	0.399614 f	normal	0.450382	no	0 flat	0	0			
32	39	female	non_angi	110	182 f	st_t_wave	180	no	0 flat	0	0	0	32	0.22449	female	non_angi	0.166667	0.187259 f	st_t_wave	0.832061	no	0 flat	0	0			

# Lab01-Preprocessing

309	120	263 f	normal	173 no	0 up	0 reversible<50		309	0.259259	0.343629 f	normal	0.778626 no	0 up	0 reversible<50
310	172	199 t	normal	162 no	0.5 up	0 reversible<50		310	0.740741	0.220077 t	normal	0.694656 no	0.080645 up	0 reversible<50
311	150	168 f	normal	174 no	1.6 up	0 normal <50		311	0.537037	0.160232 f	normal	0.78626 no	0.258065 up	0 normal <50
312	110	229 f	normal	168 no	1 down	0 reversible>50_1		312	0.166667	0.277992 f	normal	0.740458 no	0.16129 down	0 reversible>50_1
313	140	239 f	normal	160 no	1.2 up	0 normal <50		313	0.444444	0.297297 f	normal	0.679389 no	0.193548 up	0 normal <50
314	130	275 f	normal	139 no	0.2 up	0 normal <50		314	0.351852	0.366795 f	normal	0.519084 no	0.032258 up	0 normal <50
315	130	266 f	normal	171 no	0.6 up	0 normal <50		315	0.351852	0.349421 f	normal	0.763359 no	0.096774 up	0 normal <50
316	110	211 f	left_vent_	144 yes	1.8 flat	0 normal <50		316	0.166667	0.243243 f	left_vent_	0.557252 yes	0.290323 flat	0 normal <50
317	150	283 t	left_vent_	162 no	1 up	0 normal <50		317	0.537037	0.382239 t	left_vent_	0.694656 no	0.16129 up	0 normal <50
318	120	284 f	left_vent_	160 no	1.8 flat	0 normal >50_1		318	0.259259	0.38417 f	left_vent_	0.679389 no	0.290323 flat	0 normal >50_1
319	132	224 f	left_vent_	173 no	3.2 up	2 reversible>50_1		319	0.37037	0.268324 f	left_vent_	0.778626 no	0.516129 up	0.666667 reversible>50_1
320	130	206 f	left_vent_	132 yes	2.4 flat	2 reversible>50_1		320	0.351852	0.233591 f	left_vent_	0.465649 yes	0.387097 flat	0.666667 reversible>50_1
321	120	219 f	normal	158 no	1.6 flat	0 normal <50		321	0.259259	0.256867 f	normal	0.664122 no	0.258065 flat	0 normal <50
322	120	340 f	normal	172 no	0 up	0 normal <50		322	0.259259	0.492278 f	normal	0.770992 no	0 up	0 normal <50
323	150	226 f	normal	114 no	2.6 down	0 normal <50		323	0.537037	0.272201 f	normal	0.328244 no	0.419355 down	0 normal <50
324	150	247 f	normal	171 no	1.5 up	0 normal <50		324	0.537037	0.312741 f	normal	0.763359 no	0.241935 up	0 normal <50
325	110	167 f	left_vent_	114 yes	2 flat	0 reversible>50_1		325	0.166667	0.158301 f	left_vent_	0.328244 yes	0.322581 flat	0 reversible>50_1
326	140	239 f	normal	151 no	1.8 up	2 normal <50		326	0.444444	0.297297 f	normal	0.610687 no	0.290323 up	0.666667 normal <50
327	117	230 t	normal	160 yes	1.4 up	2 reversible>50_1		327	0.231481	0.279923 t	normal	0.679389 yes	0.225806 up	0.666667 reversible>50_1
328	140	335 f	normal	158 no	0 up	0 normal >50_1		328	0.444444	0.482625 f	normal	0.664122 no	0 up	0 normal >50_1
329	135	234 f	normal	161 no	0.5 flat	0 reversible<50		329	0.398148	0.287645 f	normal	0.687023 no	0.080645 flat	0 reversible<50
330	130	233 f	normal	179 yes	0.4 up	0 normal <50		330	0.351852	0.285714 f	normal	0.824427 yes	0.064516 up	0 normal <50
331	140	226 f	normal	178 no	0 up	0 normal <50		331	0.444444	0.272201 f	normal	0.816794 no	0 up	0 normal <50
332	120	177 f	left_vent_	120 yes	2.5 flat	0 reversible>50_1		332	0.259259	0.177606 f	left_vent_	0.374046 yes	0.403226 flat	0 reversible>50_1
333	150	276 f	left_vent_	112 yes	0.6 flat	1 fixed_def>50_1		333	0.537037	0.368726 f	left_vent_	0.312977 yes	0.096774 flat	0.333333 fixed_def>50_1
334	132	353 f	normal	132 yes	1.2 flat	1 reversible>50_1		334	0.37037	0.517375 f	normal	0.465649 yes	0.193548 flat	0.333333 reversible>50_1
335	150	243 t	normal	137 yes	1 flat	0 normal <50		335	0.537037	0.305019 t	normal	0.503817 yes	0.16129 flat	0 normal <50
336	150	225 f	left_vent_	114 no	1 flat	3 reversible>50_1		336	0.537037	0.270272 f	left_vent_	0.328244 no	0.16129 flat	1 reversible>50_1
337	140	199 f	normal	178 yes	1.4 up	0 reversible<50		337	0.444444	0.220077 f	normal	0.816794 yes	0.225806 up	0 reversible<50
338	160	302 f	normal	162 no	0.4 up	2 normal <50		338	0.62963	0.418919 f	normal	0.694656 no	0.064516 up	0.666667 normal <50
339	150	212 t	normal	157 no	1.6 up	0 normal <50		339	0.537037	0.245174 t	normal	0.656489 no	0.258065 up	0 normal <50
340	130	330 f	left_vent_	169 no	0 up	0 normal >50_1		340	0.351852	0.472973 f	left_vent_	0.748092 no	0 up	0 normal >50_1

Cần nộp lại tập tin **heart-normal.arff**. Tập tin sai định dạng hoặc thiếu dữ liệu (-1đ). Không nộp tập tin (-2đ).

File đính kèm

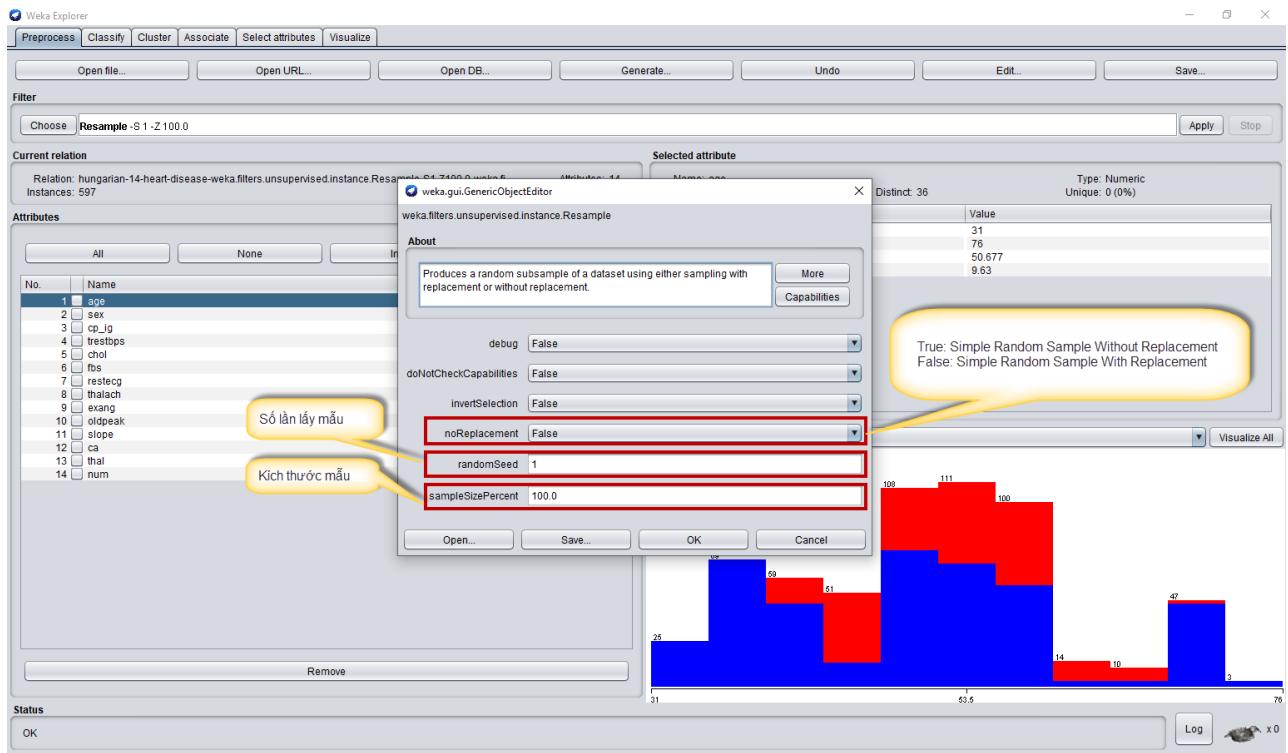
## 6) Rút gọn dữ liệu (Reduction) (1.0 điểm)

Các tập dữ liệu thường rất lớn, không thể thao tác trực tiếp được, và do đó cần áp dụng các kỹ thuật rút gọn dữ liệu khi tiền xử lý dữ liệu. Một chiến lược khác để rút gọn dữ liệu là lọc các dòng dữ liệu, hay còn gọi là lấy mẫu (*sampling*).

- a) (1.0đ) Bộ lọc nào của WEKA cho phép lấy mẫu? Nó có thể thực hiện **Simple Random Sample Without Replacement**, và **Simple Random Sample With Replacement** hay không? Nếu có, cho biết cụ thể cách thực hiện những kỹ thuật này trong WEKA. Nếu không, mô tả giải pháp lấy mẫu mà WEKA hỗ trợ.

- Bộ lọc **Resample** của WEKA cho phép lấy mẫu
- Có thể thực hiện **Simple Random Sample Without Replacement**, và **Simple Random Sample With Replacement**.
- Cách thực hiện:

# Lab01-Preprocessing



## II) Nội dung thực hiện cài đặt

Máy tính cần cài đặt python

### 1) Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản (15 điểm)

- Chương trình hỗ trợ các chức năng
  - a) Chuẩn hóa min-max trên danh sách thuộc tính chỉ định.
  - b) Chuẩn hóa Z-scores trên danh sách thuộc tính chỉ định.
  - c) Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ rộng trên danh sách thuộc tính chỉ định.
  - d) Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ sâu trên danh sách thuộc tính chỉ định.
  - e) Xóa các mẫu dữ liệu thiếu giá trị trên danh sách thuộc tính chỉ định.
  - f) Điền giá trị bị thiếu trên danh sách thuộc tính chỉ định, giá trị được điền là giá trị trung bình (mean) của thuộc tính nếu đó là thuộc tính số hoặc điền giá trị có tần số xuất hiện cao nhất (mode) nếu là thuộc tính rời rạc.

Cách thực thi chương trình chạy bằng cmd:

Có 2 cách để thực thi chương trình:

1. Thực thi bằng file .py:

python <ten chương trình .py> <file input .csv> <file output .csv> <CHUC NANG> <DANH SACH THUOC TINH> <OPTIONAL>

## Lab01-Preprocessing

2. Thực thi bằng file .exe theo đường dẫn

SourceCode\HorseColicDataset\37\_B1\Executable\dist\37\_B1:

<ten chương trình .exe> <file input .csv> <file output .csv> <CHUC NANG> <DANH SÁCH THUỐC TINH> <OPTIONAL>

Trong đó:

<CHUC NANG> : Chọn 1 trong các chức năng tu : a->f

<DANH SÁCH THUỐC TINH>: Chọn thuộc tính phù hợp với file input

<OPTIONAL>: chức năng c, d sẽ có thêm dữ liệu đầu vào là số lượng giỏ cần chia

Ví dụ:

Cách 1: python 37\_B1.py input.csv output.csv a age

Cách 2: 37\_B1.exe input.csv output.csv c age 3

## 2) Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước (10 điểm)

Cài đặt chương trình chuyển tập tin trên thành tập tin CSV (.csv), trong đó:

1. Xóa các mẫu rỗng.

2. Xóa các mẫu bị trùng lắp

3. Chuyển diện tích về km<sup>2</sup>

4. Sử dụng chương trình đã cài đặt ở phần B-1. để xóa các mẫu bị thiếu diện tích.

CHƯƠNG TRÌNH HỘ TRỢ FILE countries.txt ENCODING LÀ ANSI

Cách thực thi chương trình chạy bằng cmd:

Có 2 cách để thực thi chương trình:

1. Thực thi bằng file .py:

python <ten chương trình .py> <file input .csv>

2. Thực thi bằng file .exe theo đường dẫn

SourceCode\HorseColicDataset\37\_B1\Executable\dist\37\_B2:

<ten chương trình .exe> <file input .csv>

Ví dụ:

Cách 1: python 37\_B2.py countries.txt

Cách 2: 37\_B2.exe countries.txt

Vì yêu cầu quy định file output là :37\_B2.csv nên sẽ không có dữ liệu đầu vào cho file output

Sau khi có được file output 37\_B2.csv

Đem file 37\_B2.csv sang chương trình: 37\_B1 Để thực hiện chức năng "Xóa các mẫu bị thiếu diện tích"

## III) Nguồn tham khảo:

- J. Han and M. Kamber: Data Mining, Concepts and Techniques, Third Edition
- <https://www.youtube.com/watch?v=pNfqXhy0aCE>
- <http://weka.sourceforge.net/doc.dev/overview-summary.html>

## Lab01-Preprocessing

- <https://machinelearningmastery.com/how-to-handle-missing-values-in-machine-learning-data-with-weka/>
- <https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/>
- <https://pdfs.semanticscholar.org/526d/2be3a950b9d60edf87d9ccb6ffcc8f5d4a08.pdf>
- [https://storm.cis.fordham.edu/~yli/documents/CISC4631Spring16/LabOne\\_RemoveOutlier.pdf](https://storm.cis.fordham.edu/~yli/documents/CISC4631Spring16/LabOne_RemoveOutlier.pdf)
- <https://www.youtube.com/watch?v=lhzXFDbVNyc>
- <https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/>

## IV) Phụ lục:

Dataset **datasets-UCI.jar** là dữ liệu về bệnh tim của 4 trung tâm y tế.

Thu tập tất cả 76 thuộc tính nhưng chỉ quan tâm 14 thuộc tính trong hai data liệu **heart-h.arff** và **heart-c.arff**:

STT	STT trong dataset	Thuộc tính	Diễn giải	Giá trị
1	#3	(age)	Tuổi	Number
2	#4	(sex)	Giới tính	male female
3	#9	(chest_pain)	Loại đau ngực	typ_angina asympt non_anginal atyp_angina
4	#10	(trestbps)	Huyết áp	Number
5	#12	(chol)	Cholesterol trong máu	Number
6	#16	(fbs)	Đường huyết > 120 mg/dl	true false
7	#19	(restecg)	Điện tâm đồ	normal left_vent_hyper st_t_wave_abnormality
8	#32	(thalach)	Nhịp tim tối đa	Number

## Lab01-Preprocessing

9	#38	(exang)	Đau thắt ngực khi gắng sức	yes no
10	#40	(oldpeak)		Number
11	#41	(slope)		up flat down
12	#44	(ca)		Number
13	#51	(thal)		normal fixed_defect reversible_defect
14	#58	(num) (the predicted attribute)	Chẩn đoán bệnh	'<50' '>50_1' '>50_2' '>50_3' '>50_4'