



ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
Bài tập Thực hành 1

TIỀN XỬ LÝ DỮ LIỆU VỚI WEKA

Nhóm thực hiện

1. Hồng Thanh Hoài 1612855
 2. Huỳnh Minh Huân 1612858
-

Giáo viên lý thuyết
PGS.TS Lê Hoài Bắc

Giáo viên hướng dẫn
Nguyễn Ngọc Thảo, Lê Ngọc Thành

Tháng 04 năm 2019

Mục lục

1	Giới thiệu nhóm và phân công công việc	1
1.1	Giới thiệu nhóm	1
1.2	Phân công công việc	1
2	Nội dung	2
2.1	Tích hợp dữ liệu - Integration	2
2.1.1	Định nghĩa sự tích hợp dữ liệu	2
2.1.2	Kiểm tra nhận diện thực thể	2
2.1.3	Kiểm tra dữ liệu dư thừa	3
2.1.4	Kiểm tra mâu thuẫn dữ liệu	5
2.1.5	Tích hợp dữ liệu	5
2.2	Tóm tắt mô tả dữ liệu – Descriptive data summarization	5
2.2.1	Xét thuộc tính age	5
2.2.2	Five-number summary của thuộc tính age	6
2.2.3	Thông tin thuộc tính	7
2.2.4	Ý nghĩa của đồ thị trong cửa sổ Explorer	7
2.2.5	Đồ thị phân bố của các thuộc tính còn lại	9
2.2.6	Nhận xét	10
2.2.7	Đồ thị "scatter"	10
2.2.8	Những cặp thuộc tính có vẻ tương quan với nhau	11
2.3	Chọn lọc dữ liệu - Selection	11
2.4	Làm sạch dữ liệu - Cleaning	13
2.4.1	Missing values (dữ liệu thiếu)	13
2.4.2	Noisy data (dữ liệu nhiễu)	14
2.4.3	Outlier data (dữ liệu ngoại lệ/ dữ liệu tạp)	14
2.5	Chuyển đổi dữ liệu - Transformation	16
2.5.1	Xây dựng thuộc tính – <i>Attribute construction</i>	16
2.5.2	Chuẩn hóa – <i>Normalize</i>	19
2.5.3	Chọn phương pháp chuẩn hóa	21
2.6	Rút gọn dữ liệu - Reduction	22
2.6.1	Lấy mẫu dữ liệu với các bộ lọc Weka	22
2.6.2	Khả năng thực hiện của Weka	25
3	Đánh giá	26
	Tài liệu tham khảo	27

1 Giới thiệu nhóm và phân công công việc

1.1 Giới thiệu nhóm

Nhóm gồm 2 thành viên.

STT	Họ và tên	MSSV	Email	SĐT
1	Hồng Thanh Hoài	1612855	hthoai1006@gmail.com	0965596807
2	Huỳnh Minh Huân	1612858	minhhuanhuynh289@gmail.com	0824540646

1.2 Phân công công việc

STT	Họ và tên	Công việc
1	Hồng Thanh Hoài	Câu 1, 5, báo cáo.
2	Huỳnh Minh Huân	Câu 2, 3, 4, 6.

2 Nội dung

2.1 Tích hợp dữ liệu - Integration

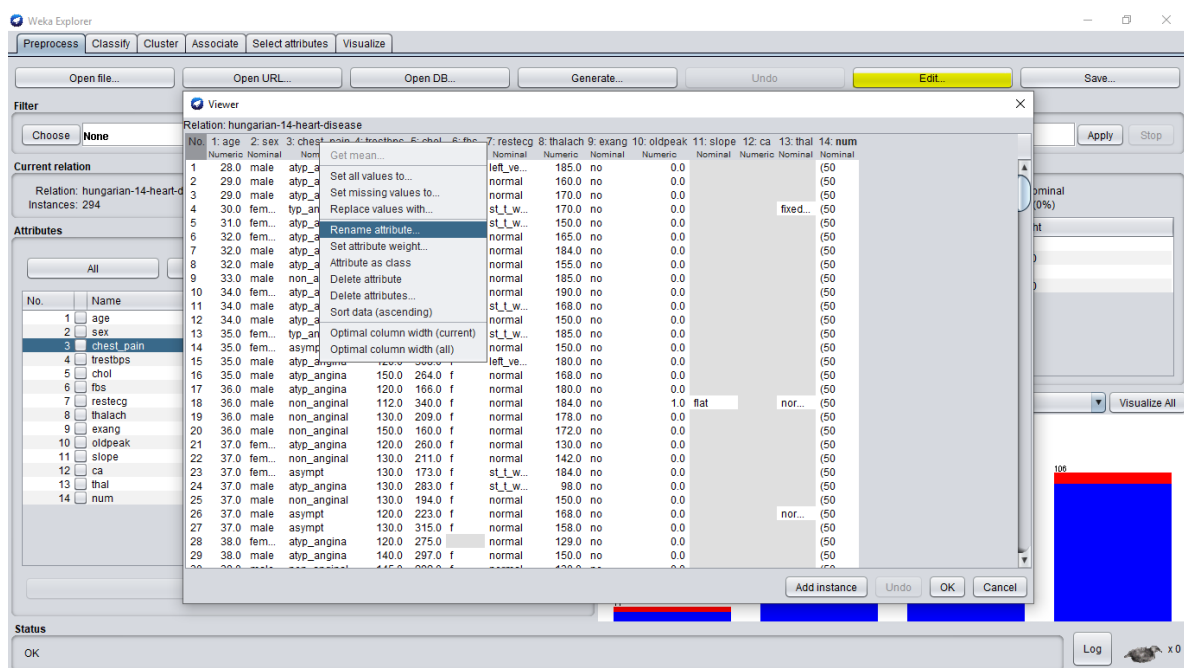
2.1.1 Định nghĩa sự tích hợp dữ liệu

Tích hợp dữ liệu là kết hợp dữ liệu từ nhiều nguồn vào một nơi lưu trữ thống nhất để tiện cho những thao tác về sau. Dữ liệu ở đây có thể bao gồm nhiều cơ sở dữ liệu, khối dữ liệu, hoặc các file dữ liệu.

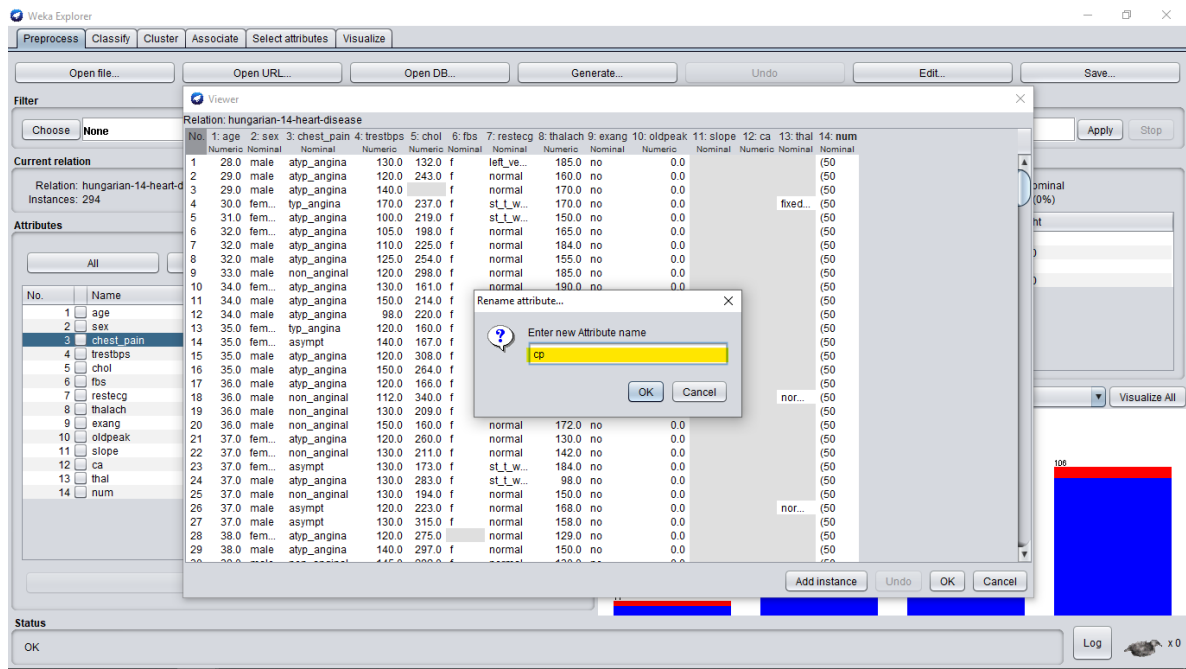
2.1.2 Kiểm tra nhận diện thực thể

Có vấn đề về nhận diện thực thể (*entity identification*) trong 2 dataset này. Cụ thể, ở file `heart-c.arff`, thuộc tính thứ 3 tên là `cp`, còn ở file `heart-h.arff`, thuộc tính thứ 3 tên là `chest_pain`.

Hướng giải quyết: Đổi tên thuộc tính `chest_pain` ở file `heart-h.arff` thành `cp`.



Hình 1: Ở tab *Preprocessing*, chọn *Edit*, chuột phải vào thuộc tính `cp` và chọn *Rename attribute...*



Hình 2: Điền tên cần đổi vào cửa sổ hiện lên và nhấn OK.

2.1.3 Kiểm tra dữ liệu dư thừa

Ta dùng *Karl Pearson* để kiểm tra sự tương quan giữa các thuộc tính số và *Chi-Square* để kiểm tra sự tương quan giữa các thuộc tính rời rạc.

Bảng 1: *Correlation* giữa các thuộc tính số của file `heart-c.arff`.

	age	trestbps	chol	thalach	oldpeak
trestbps	0.279				
chol	0.214	0.123			
thalach	-0.399	-0.047	0.010		
oldpeak	0.210	0.193	0.054	-0.344	
ca	0.365	0.103	0.122	-0.263	0.294

Bảng 2: *Chi-Square* giữa các thuộc tính rời rạc của file `heart-c.arff` (những giá trị gạch chân thể hiện hai thuộc tính có sự tương quan).

	sex	cp	fbs	restecg	exang	slope	thal
cp	6.822						
fbs	0.614	3.886					
restecg	3.697	9.688	2.297				
exang	6.081	<u>67.348</u>	0.2	2.976			
slope	0.648	<u>27.747</u>	3.373	10.947	<u>25.131</u>		
thal	<u>44.626</u>	<u>41.892</u>	5.542	3.526	<u>32.959</u>	<u>35.283</u>	
num	<u>23.914</u>	<u>81.686</u>	0.238	10.023	<u>57.799</u>	<u>47.507</u>	<u>85.304</u>

Bảng 3: *Correlation* giữa các thuộc tính số của file `heart-h.arff`.

	age	trestbps	chol	thalach
trestbps	0.245			
chol	0.091	0.084		
thalach	-0.459	-0.185	-0.128	
oldpeak	0.178	0.207	0.109	-0.303

Bảng 4: *Chi-Square* giữa các thuộc tính rời rạc của file `heart-h.arff` (những giá trị gạch chân thể hiện hai thuộc tính có sự tương quan).

	sex	cp	fbs	restecg	exang	slope	thal
cp	<u>19.042</u>						
fbs	2.593	1.290					
restecg	7.379	<u>29.535</u>	1.580				
exang	9.632	<u>82.540</u>	4.088	7.728			
slope	5.012	<u>66.018</u>	15.987	11.918	<u>176.898</u>		
thal	0.657	9.042	4.921	1.927	3.277	2.252	
num	<u>21.876</u>	<u>95.811</u>	8.084	2.758	<u>100.591</u>	<u>110.521</u>	10.201

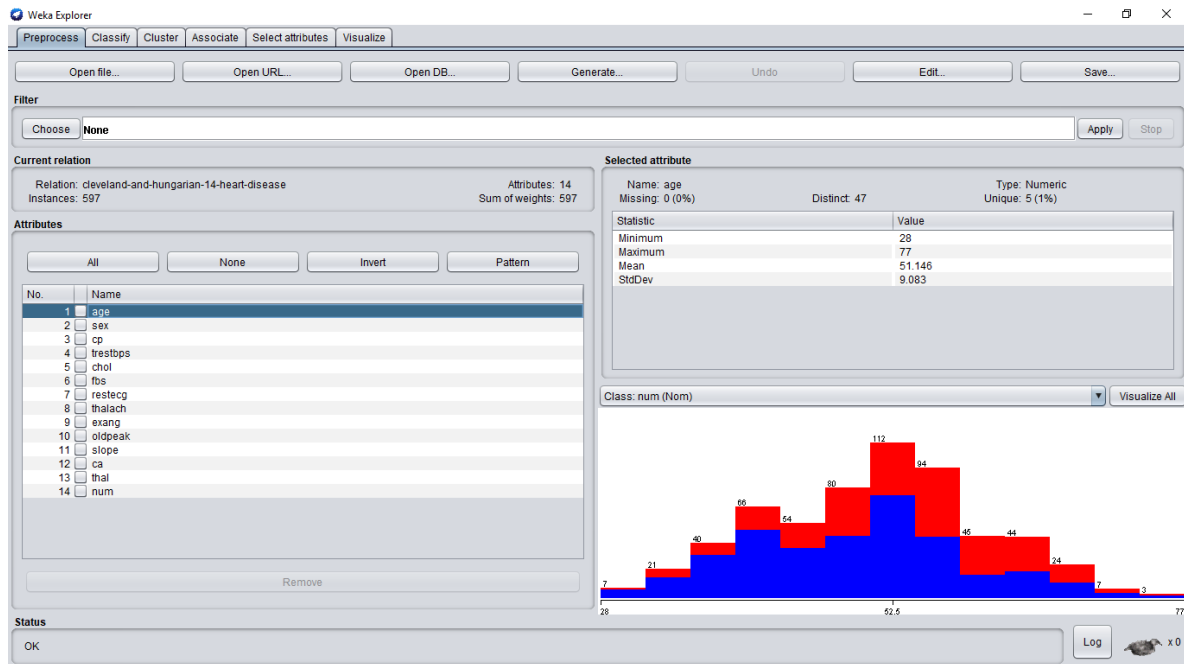
Ta thấy các thuộc tính có mức độ tương quan không quá cao nên không thể loại bỏ thuộc tính nào. Vậy, không có vấn đề dư thừa dữ liệu trong hai dataset này.

2.1.4 Kiểm tra mâu thuẫn dữ liệu

Kiểm tra từng thuộc tính của 2 dataset, ta thấy không có sự mâu thuẫn dữ liệu (data value conflicts) trong 2 dataset này.

2.1.5 Tích hợp dữ liệu

Sau khi tích hợp, dataset mới có 597 mẫu và 14 thuộc tính.

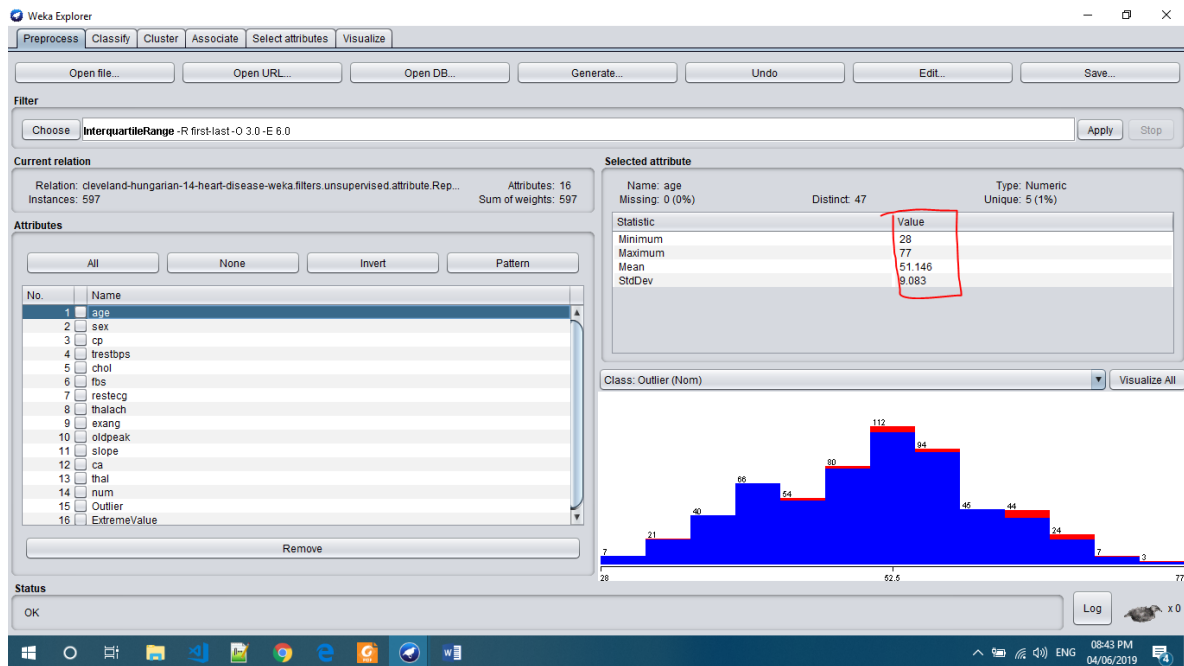


Hình 3: Dataset sau khi tích hợp.

2.2 Tóm tắt mô tả dữ liệu – Descriptive data summarization

2.2.1 Xét thuộc tính age

- Trung bình: mean = 51.146.
- Độ lệch chuẩn: sd = 9.038.
- Giá trị lớn nhất: maximum = 77.
- Giá trị nhỏ nhất: minimum = 28.



Hình 4: Thuộc tính age.

2.2.2 Five-number summary của thuộc tính age

- Minimum = 28.
- Q1 = 44.
- Median = 52.
- Q3 = 58.
- Maximum = 77.

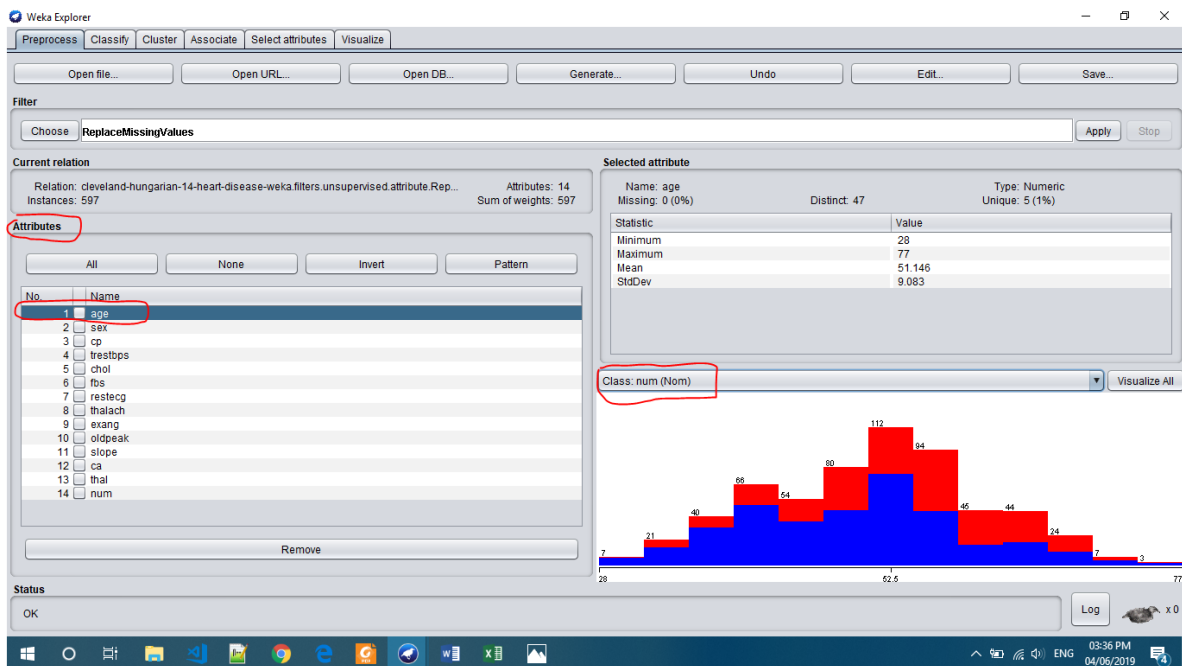
598	65	male	ε
599	44	Q1	
600	58	Q3	
601	28	MIN	
602	77	MAX	
603	52	MEDIAN	
604			
605			

Hình 5: Weka chỉ cung cấp minimum và maximum. Q1, Q3 và median được nhóm chuyển file arff sang csv sao đó sử dụng excel để tìm các giá trị Q1, Q3 và median.

2.2.3 Thông tin thuộc tính

- Thuộc tính số (numeric): age, trestbps, chol, thalach, oldspcak, ca (6 thuộc tính).
- Thuộc tính có thứ tự (ordinal): restecg, slope, thal, num (4 thuộc tính).
- Thuộc tính rời rạc/danh sách (categorical/nominal): sex, cp, fbs, exang (4 thuộc tính).

2.2.4 Ý nghĩa của đồ thị trong cửa sổ Explorer

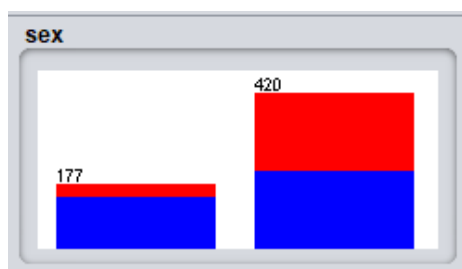


Hình 6: Tên đồ thị: Đồ thị phân bố của thuộc tính.

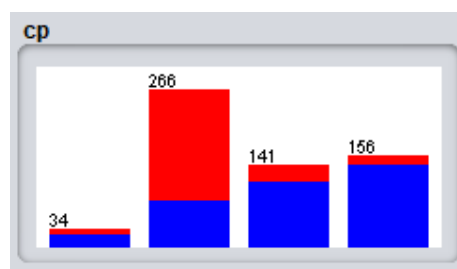
- Đồ thị này là đồ thị để thể hiện sự phân bố của thuộc tính hiện được chọn (nền xanh) ở khung Attributes và được phân lớp (classification) theo thuộc tính được chọn ở box Class.
- Tên đồ thị: Đồ thị phân bố của thuộc tính.
- Màu xanh ứng với num có label <50 (không mắc bệnh). Màu đỏ ứng với num có label >50_1 (mắc bệnh mức 1). (do mức 2, 3, 4 không có nên trên đồ thị không hiển thị).
- Mỗi cột biểu diễn một khoảng tuổi (đang xét theo thuộc tính tuổi), giá trị trên mỗi cột là số lượng instances có 'age' thuộc khoảng đó.
- Đồ thị biểu diễn phân bố tuổi và tỉ lệ giữa người có bệnh và không có bệnh trên mỗi khoảng tuổi được chia. Ví dụ, ở khoảng [28, 31.769] có 7 người tỉ lệ người không

mắc bệnh trong độ tuổi này chiếm cao hơn tỉ lệ người mắc bệnh mức 1 (màu xanh chiếm gần hết cột).

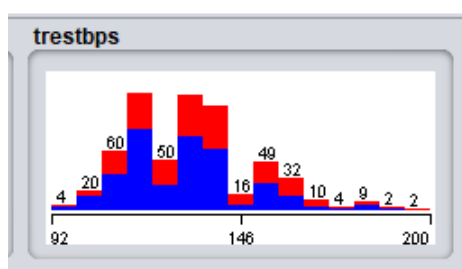
2.2.5 Đồ thị phân bố của các thuộc tính còn lại



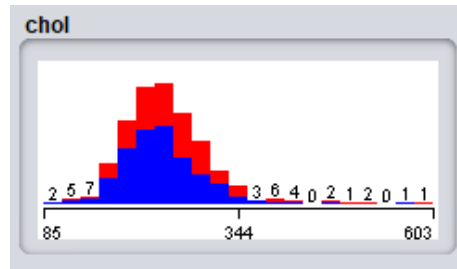
(a) sex



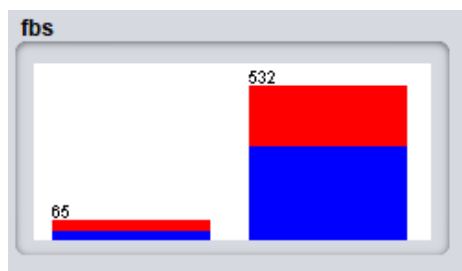
(b) chest_pain



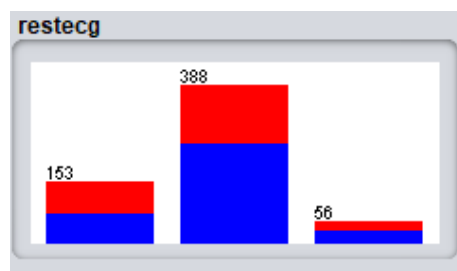
(c) trestbps



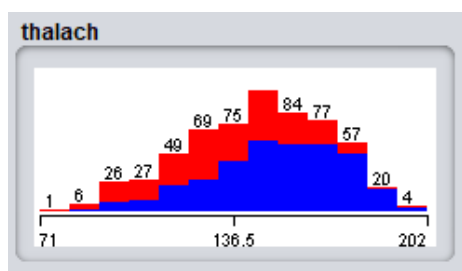
(d) chol



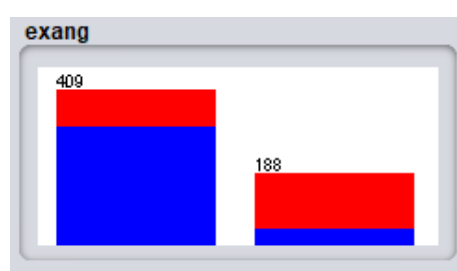
(e) fbs



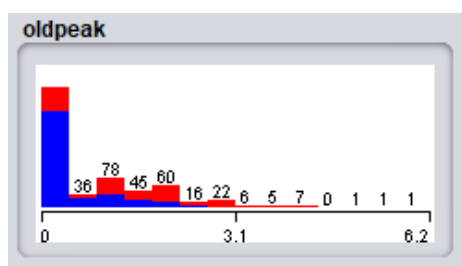
(f) restecg



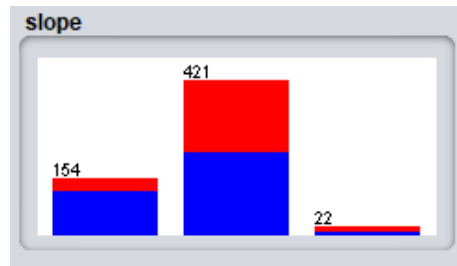
(g) thalach



(h) exang

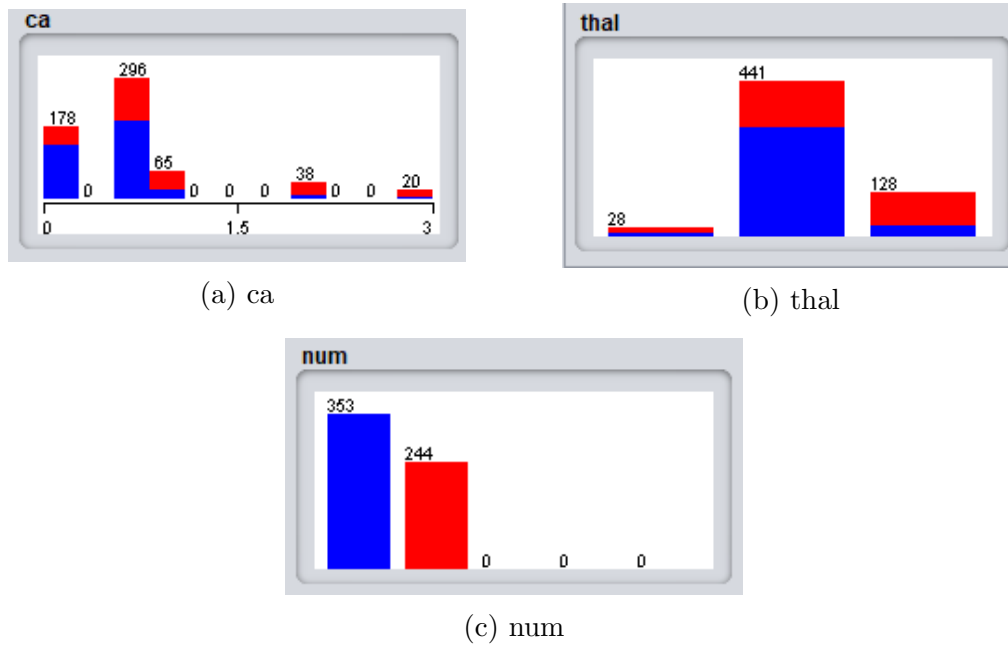


(i) oldpeak



(j) slop

Hình 7: Đồ thị phân bố của các thuộc tính còn lại (1).



Hình 8: Đồ thị phân bố của các thuộc tính còn lại (2).

2.2.6 Nhận xét

Đồ thị thể hiện trực quan mối liên hệ giữa 2 thuộc tính bất kỳ trong tập dữ liệu (ví dụ đang xét là giữa thuộc tính *num* với các thuộc tính còn lại). Từ đồ thị có thể nhận thấy được một số thuộc tính có khả năng dùng để phân loại xem có mắc bệnh hay không (sẽ được thể hiện cụ thể hơn ở phần Visualize mà Weka hỗ trợ nhưng có thể phần nào trực quan xem được).

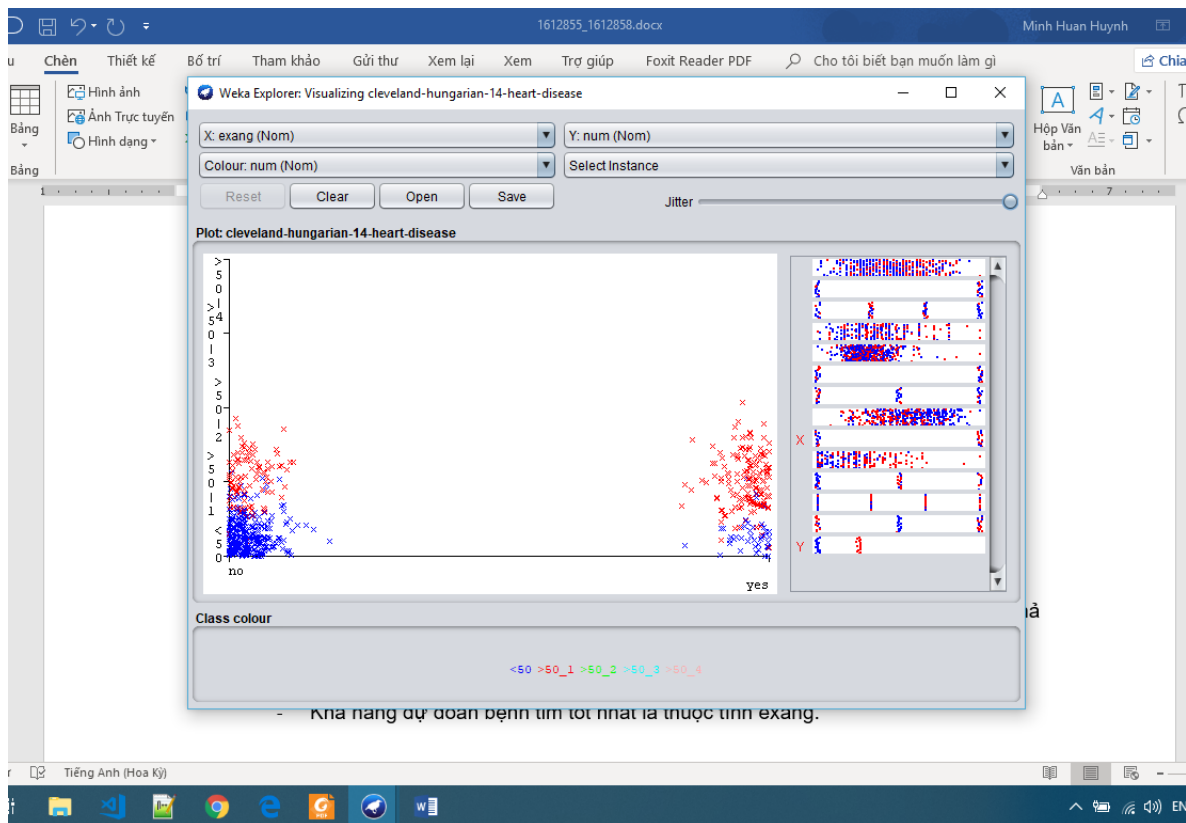
Đồ thị đối với dữ liệu rời rạc (nominal/categorize) thì các cột rời nhau, còn đồ thị có giá trị liên tục thì các cột kề nhau, mỗi cột là một khoảng (mỗi khoảng chứa một số lượng instance có giá trị thuộc khoảng đó).

Các giá trị của thuộc tính (min, max, mean) đối với dữ liệu số sẽ được thể hiện trực quan trên đồ thị.

2.2.7 Đồ thị "scatter"

Thuật ngữ sử dụng trong textbook để đặt tên cho đồ thị ở mục Visualize là "scatter plot".

- Theo nhóm, các thuộc tính có vẻ dẫn đến bệnh tim là: *oldspeak* (không mắc bệnh trong khoản 0 – 1.6, lớn hơn 1.6 khả năng mắc bệnh cao), *exang* (no – khả năng không mắc bệnh cao, yes – khả năng mắc bệnh cao), *thalach* (<136.5 – khả năng mắc bệnh cao, >136.5 – thường không mắc bệnh), *slope* (up – không mắc bệnh cao, flat – khả năng mắc bệnh cao).
- Khả năng dự đoán bệnh tim tốt nhất là thuộc tính *exang*.



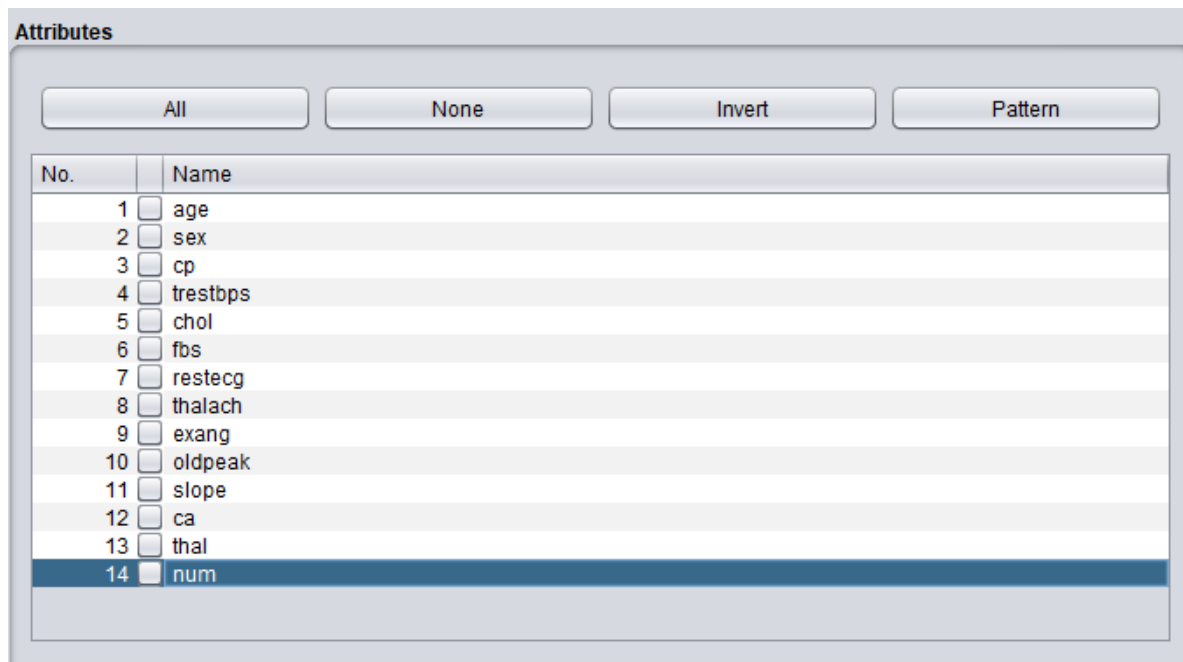
Hình 9: Thuộc tính exang dự đoán tốt nhất.

2.2.8 Những cặp thuộc tính có vẻ tương quan với nhau

Dựa vào quan sát đồ thị "scatter", ngoài những cặp thuộc tính giữa num và các thuộc tính khác như trên, thì những cặp sau đây có vẻ tương quan với nhau: (slop, cp), (slop, exang), (exang, cp).

2.3 Chọn lọc dữ liệu - Selection

- Có 14 thuộc tính trong datasets trước khi xử lý dữ liệu.



Hình 10: Có 14 thuộc tính trước khi xử lý dữ liệu.

- Các lựa chọn khác nhau để lựa chọn thuộc tính là:
 - Các phương pháp đánh giá:
 - * 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.
 - * ClassifierSubsetEval: Đánh giá các tập hợp thuộc tính trên dữ liệu huấn luyện hoặc một bộ kiểm tra riêng biệt.
 - * CorrelationAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo lường mối tương quan (Pearson's) giữa nó và lớp. Các thuộc tính danh nghĩa được xem xét trên một giá trị theo cơ sở giá trị bằng cách coi mỗi giá trị là một chỉ báo.
 - * GainRatioAttributeEval: Đánh giá theo độ đo tỉ lệ đạt được với các lớp. với công thức tính gain là $\text{GainR}(\text{Class}, \text{Attribute}) = (\text{H}(\text{Class}) - \text{H}(\text{Class} | \text{Attribute})) / \text{H}(\text{Attribute})$.
 - * OneRAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách sử dụng trình phân loại OneR.
 - * PrincipalComponents: Thực hiện phân tích thành phần chính (chọn ra các thành phần chính) nhằm biến đổi dữ liệu (transformation data).
 - * ReliefFAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách lặp lại việc lấy mẫu (sampling) của một instance và xem xét giá trị của thuộc tính đã cho với instance gần nhất của cùng một lớp và khác lớp.

- * SymmetricalUncertAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo độ không đảm bảo đối xứng đối với lớp.
- * WrapperSubsetEval: đánh giá bằng tập bao các phân loại (“wrapper” method wraps a classifier in cross-validation loop).
- Các phương pháp tìm kiếm (chọn thuộc tính theo nhu cầu):
 - * BestFirst: Tìm kiếm tập con không gian các thuộc tính bằng tham lam tăng cường với cơ sở quay lui.
 - * GreedyStepwise: Phương pháp tìm kiếm tham lam tiến hay lùi thông qua tập con không gian các thuộc tính.
 - * Ranker: xếp hạng các thuộc tính theo giá trị nó được đánh giá.
- So sánh các phương pháp giữa weka và textbook: trong textbook không thấy phương pháp tìm kiếm (chọn thuộc tính theo yêu cầu) là BestFirst và Ranker.

2.4 Làm sạch dữ liệu - Cleaning

2.4.1 Missing values (dữ liệu thiếu)

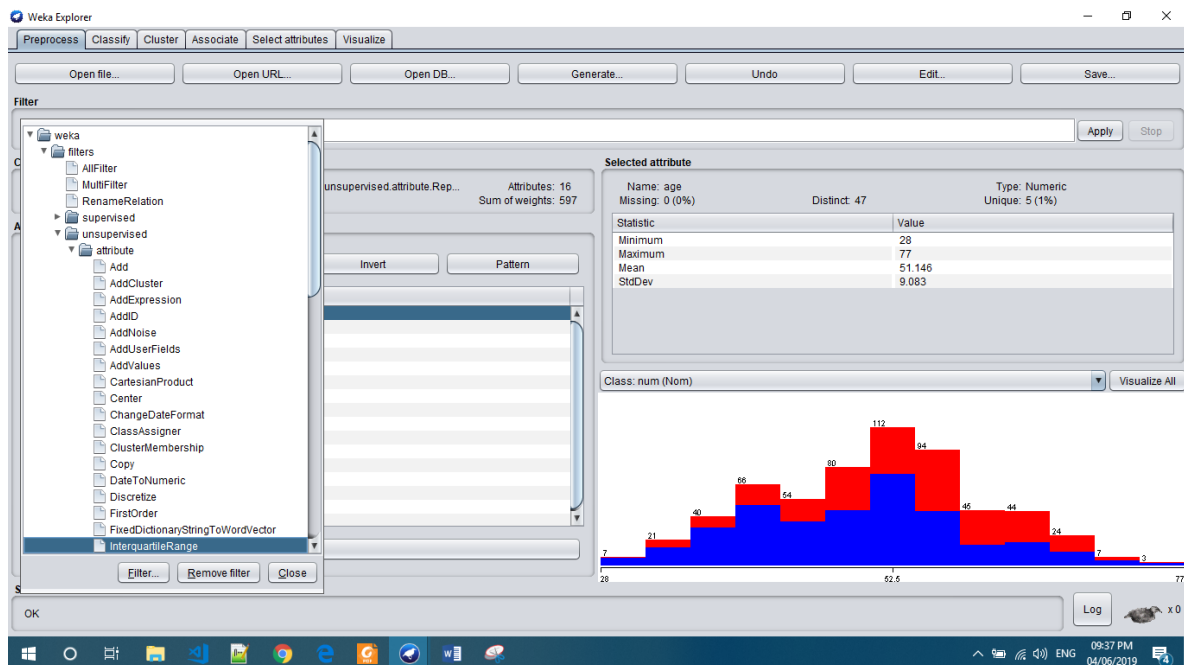
- Các phương pháp xử lý dữ liệu thiếu:
 - Bỏ qua bộ đó (ignore the tuple).
 - Điền các dữ liệu thiếu bằng phương pháp thủ công (manually).
 - Sử dụng giá trị hằng ngoại vi (global constant) để điền: “Unknown” hay “ ∞ ”.
 - Dùng thuộc tính trung bình (mean) để điền các giá trị thiếu.
 - Sử dụng thuộc tính trung bình cho tất cả các mẫu thuộc cùng một lớp với bộ dữ liệu đã cho.
 - Dùng giá trị có thể xảy ra nhất để điền vào giá trị còn thiếu.
- Weka đã cài đặt phương pháp: thay thế dữ liệu bằng giá trị trung bình hay mode, thay thế giá trị thiếu bằng hằng số do người dùng đặt, thay thế giá trị thiếu bằng giá trị có thể xảy ra nhất.
- Chọn phương pháp điền giá trị thiếu bằng mean hay mode. Vì:
 - không thể tiến hành điền manual vì dữ liệu thiếu nhiều ở thuộc tính ca (50%), thal (45%), slope (32%) và không thể biết được nên điền như thế nào.
 - Nếu chọn lược bỏ các tuple thiếu thì lại làm mất tính khách quan dữ liệu.
 - Chỉ có 1 số thuộc tính bị thiếu nên nhóm nghĩ tốt nhất là chọn ReplaceMissingValue by mean or mode sẽ đảm bảo.

2.4.2 Noisy data (dữ liệu nhiễu)

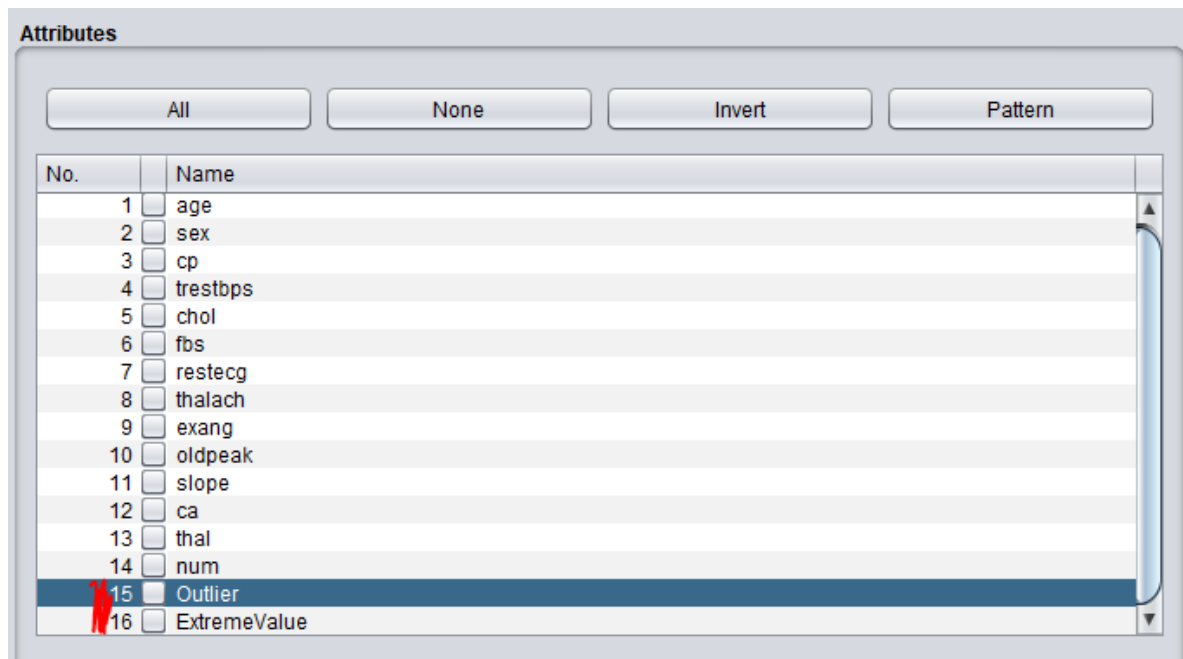
- Các phương pháp loại bỏ dữ liệu nhiễu:
 - Phương pháp phân khoảng hay chia giỏ (binning): Chia giỏ theo giá trị trung bình (bin mean), chia giỏ theo trung vị (bin median), chia giỏ theo biên (bin boundaries).
 - Hồi quy (regression)
 - Phân cụm, gom nhóm (clustering)
- Weka hỗ trợ: Hồi quy

2.4.3 Outlier data (dữ liệu ngoại lệ/ dữ liệu tạp)

- Các phương pháp:
 - Gom cụm (Clustering)
 - Numeric outlier (IQR): điểm dữ liệu tạp nằm ngoài khoảng interquartile (IQR).
 - Z-score
- Dò tìm dữ liệu tạp bằng Weka:



Hình 11: B1 - Mục *Filter* → *Choose* → *filters* → *unsupervised* → *attribute* → *InterquartileRange*.



Hình 12: B2 - Nhấn apply, sẽ xuất hiện thêm 2 thuộc tính trong mục Attributes là Outlier và ExtremeValue.

Selected attribute

Name: Outlier		Distinct: 2		Type: Nominal	
Missing: 0 (0%)				Unique: 0 (0%)	
No.	Label	Count		Weight	
1	no	572		572.0	
2	yes	25		25.0	

Hình 13: B3 - Chọn thuộc tính Outlier để xem kết quả, có dữ liệu tập hay không.

- Có dữ liệu tập trong dataset (như hình trên) là 25 (thuộc tính gần cuối là thuộc tính outlier, file được lưu lại sau khi thực hiện các bước trên).

```

21 63,male,typ_angina,145,233,t,left_vent_hyper,150,no,2.3,down,0,fixed_defect,<50,no,no
22 67,male,asympt,160,286,f,left_vent_hyper,108,yes,1.5,flat,3,normal,>50_1,yes,no
23 67,male,asympt,120,229,f,left_vent_hyper,129,yes,2.6,flat,2,reversable_defect,>50_1,no,no
24 37,male,non_anginal,130,250,f,normal,187,no,3.5,down,0,normal,<50,no,no
25 41,female,atyp_angina,130,204,f,left_vent_hyper,172,no,1.4,up,0,normal,<50,no,no
26 56,male,atyp_angina,120,236,f,normal,178,no,0.8,up,0,normal,<50,no,no

```



```

61 65,female,asympt,150,225,f,left_vent_hyper,114,no,1,flat,3,reversable_defect,>50_1,yes,no
62 40,male,typ_angina,140,199,f,normal,178,yes,1.4,up,0,reversable_defect,<50,no,no
63 71,female,atyp_angina,160,302,f,normal,162,no,0.4,up,2,normal,<50,no,no
64 59,male,non_anginal,150,212,t,normal,157,no,1.6,up,0,normal,<50,no,no

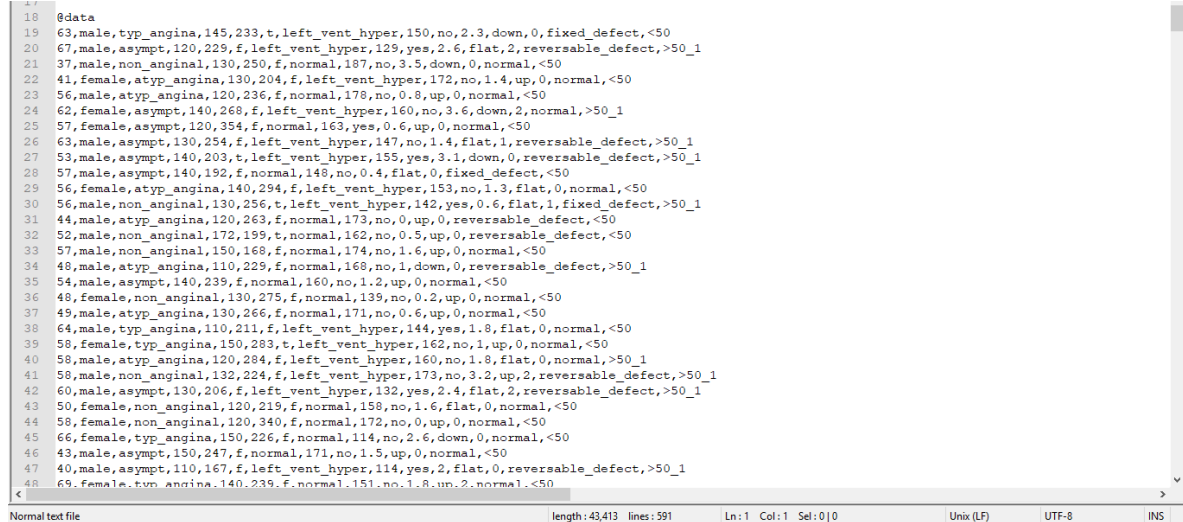
```

```

109 53,female,asympt,138,234,f,left_vent_hyper,160,no,0,up,0,normal,<50,no,no
110 51,female,non_anginal,130,256,f,left_vent_hyper,149,no,0.5,up,0,normal,<50,no,no
111 66,male,asympt,120,302,f,left_vent_hyper,151,no,0.4,flat,0,normal,<50,no,no
112 62,female,asympt,160,164,f,left_vent_hyper,145,no,6.2,down,3,reversable_defect,>50_1,yes,no
113 62,male,non_anginal,130,231,f,normal,146,no,1.8,flat,3,reversable_defect,<50,yes,no
114 44,female,non_anginal,108,141,f,normal,175,no,0.6,flat,0,normal,<50,no,no
115 63,female,non_anginal,135,252,f,left_vent_hyper,172,no,0,up,0,normal,<50,no,no

```

- Lưu heart-cleaned.arff.



```

18 @data
19 63,male,typ_angina,145,233,t,left_vent_hyper,150,no,2.3,down,0,fixed_defect,<50
20 67,male,asympt,120,229,f,left_vent_hyper,129,yes,2.6,flat,2,reversable_defect,>50_1
21 37,male,non_anginal,130,250,f,normal,187,no,3.5,down,0,normal,<50
22 41,female,atyp_angina,130,204,f,left_vent_hyper,172,no,1.4,up,0,normal,<50
23 56,male,atyp_angina,120,236,f,normal,178,no,0.8,up,0,normal,<50
24 62,female,asympt,140,268,f,left_vent_hyper,160,no,3.6,down,2,normal,>50_1
25 57,female,asympt,120,354,f,normal,163,yes,0.6,up,0,normal,<50
26 63,male,asympt,130,254,f,left_vent_hyper,147,no,1.4,flat,1,reversable_defect,>50_1
27 53,male,asympt,140,203,t,left_vent_hyper,155,yes,3.1,down,0,reversable_defect,>50_1
28 57,male,asympt,140,192,f,normal,148,no,0.4,flat,0,fixed_defect,<50
29 56,female,atyp_angina,140,294,f,left_vent_hyper,153,no,1.3,flat,0,normal,<50
30 56,male,non_anginal,130,256,t,left_vent_hyper,142,yes,0.6,flat,1,fixed_defect,>50_1
31 44,male,atyp_angina,120,263,t,normal,173,no,0,up,0,reversable_defect,<50
32 52,male,non_anginal,172,199,t,normal,162,no,0.5,up,0,reversable_defect,<50
33 57,male,non_anginal,150,168,f,normal,174,no,1.6,up,0,normal,<50
34 48,male,atyp_angina,110,229,f,normal,168,no,1,down,0,reversable_defect,>50_1
35 54,male,asympt,140,239,f,normal,160,no,1.2,up,0,normal,<50
36 48,female,non_anginal,130,275,f,normal,139,no,0.2,up,0,normal,<50
37 49,male,atyp_angina,130,266,f,normal,171,no,0.6,up,0,normal,<50
38 64,male,typ_angina,110,211,f,left_vent_hyper,144,yes,1.8,flat,0,normal,<50
39 58,female,typ_angina,150,283,t,left_vent_hyper,162,no,1,up,0,normal,<50
40 58,male,atyp_angina,120,284,f,left_vent_hyper,160,no,1.8,flat,0,normal,>50_1
41 58,male,non_anginal,132,224,f,left_vent_hyper,173,no,3.2,up,2,reversable_defect,>50_1
42 60,male,asympt,130,206,f,left_vent_hyper,132,yes,2.4,flat,2,reversable_defect,>50_1
43 50,female,non_anginal,120,219,f,normal,158,no,1.6,flat,0,normal,<50
44 58,female,non_anginal,120,340,f,normal,172,no,0,up,0,normal,<50
45 66,female,typ_angina,150,226,f,normal,114,no,2.6,down,0,normal,<50
46 43,male,asympt,150,247,f,normal,171,no,1.5,up,0,normal,<50
47 40,male,asympt,110,167,f,left_vent_hyper,114,yes,2,flat,0,reversable_defect,>50_1
48 69,female,typ_angina,140,239,f,normal,151,no,1.8,up,2,normal,<50

```

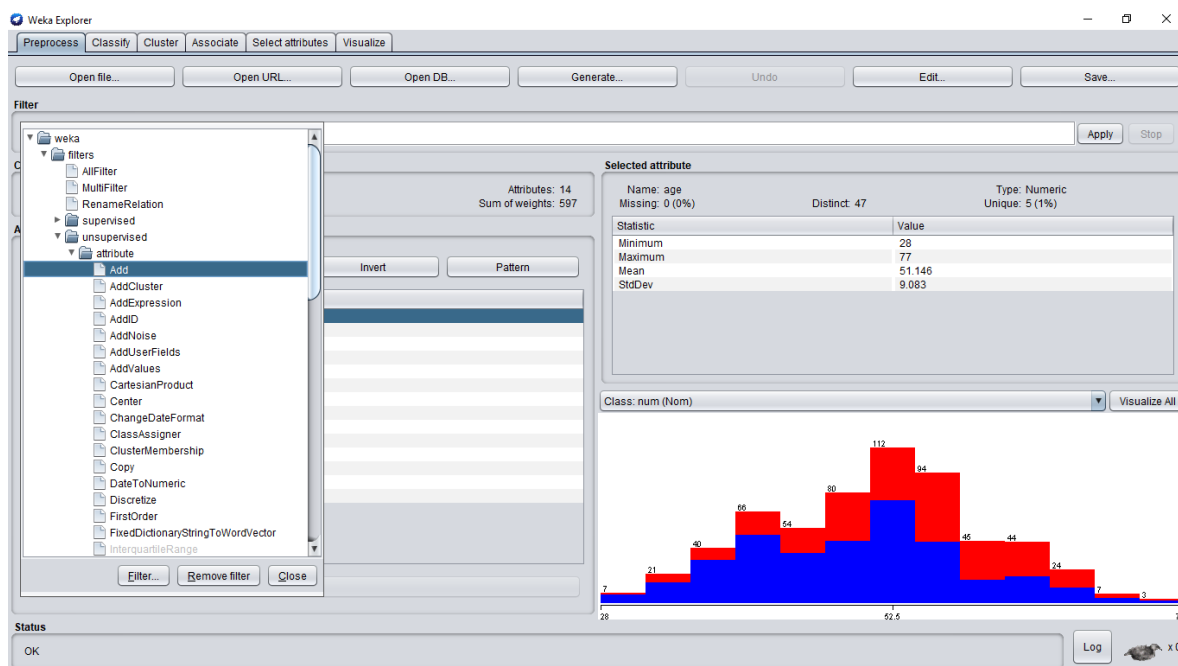
Hình 14: Sau khi làm sạch.

2.5 Chuyển đổi dữ liệu - Transformation

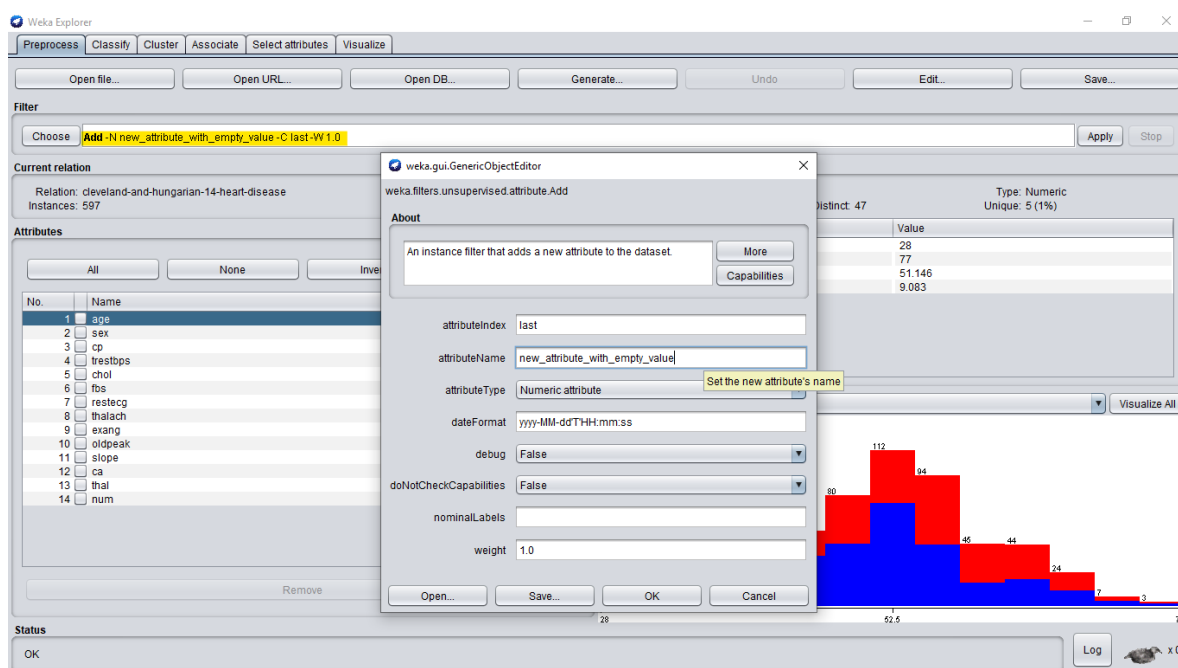
2.5.1 Xây dựng thuộc tính – *Attribute construction*

Weka hỗ trợ 4 bộ lọc để thêm một thuộc tính mới vào dataset, ta sẽ lần lượt tìm hiểu từng bộ lọc.

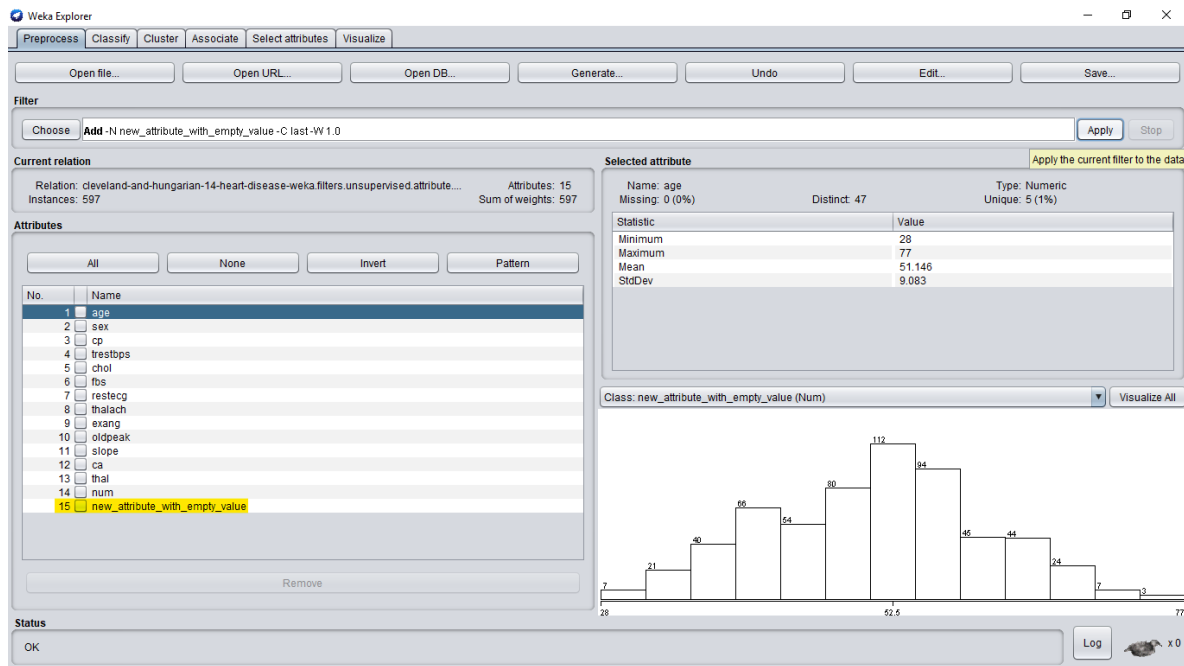
- attribute.Add



Hình 15: Ở thanh *Filter*, chọn *Choose* → *weka* → *filters* → *unsupervised* → *attribute* → *Add*



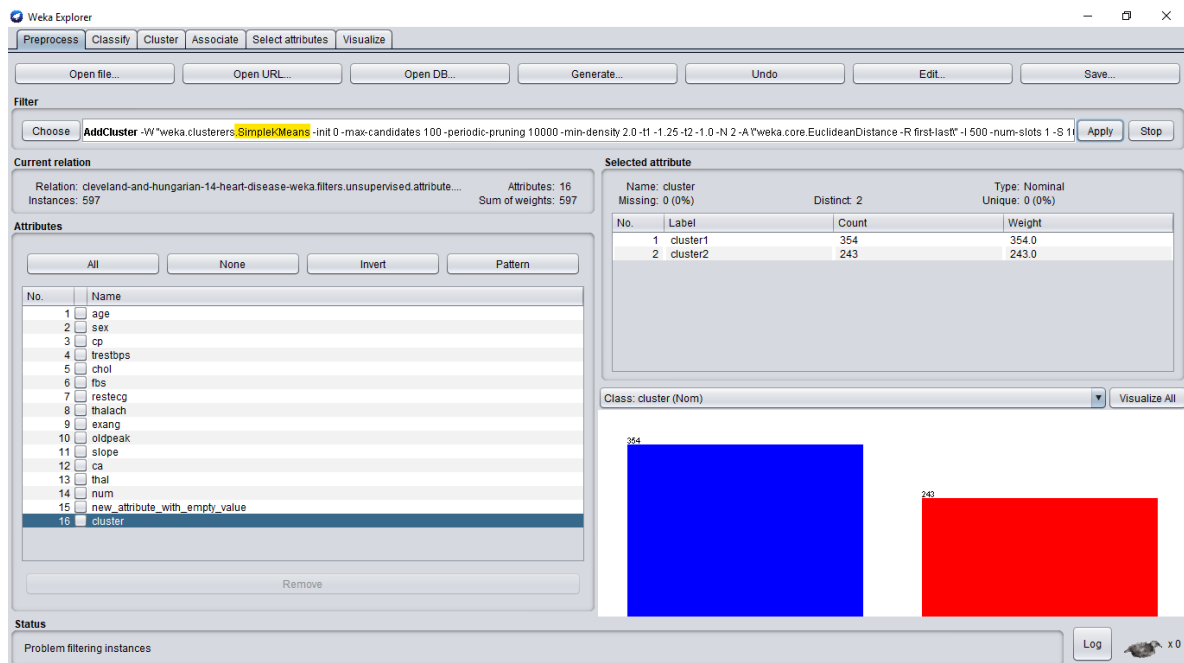
Hình 16: Nhảy vào thanh màu vàng và điền các thông số vào. Thuộc tính mới sẽ có dạng *missing value*.



Hình 17: Nhấn *Apply* để áp dụng bộ lọc, thuộc tính mới sẽ được thêm vào sau đó.

- `attribute.AddCluster`

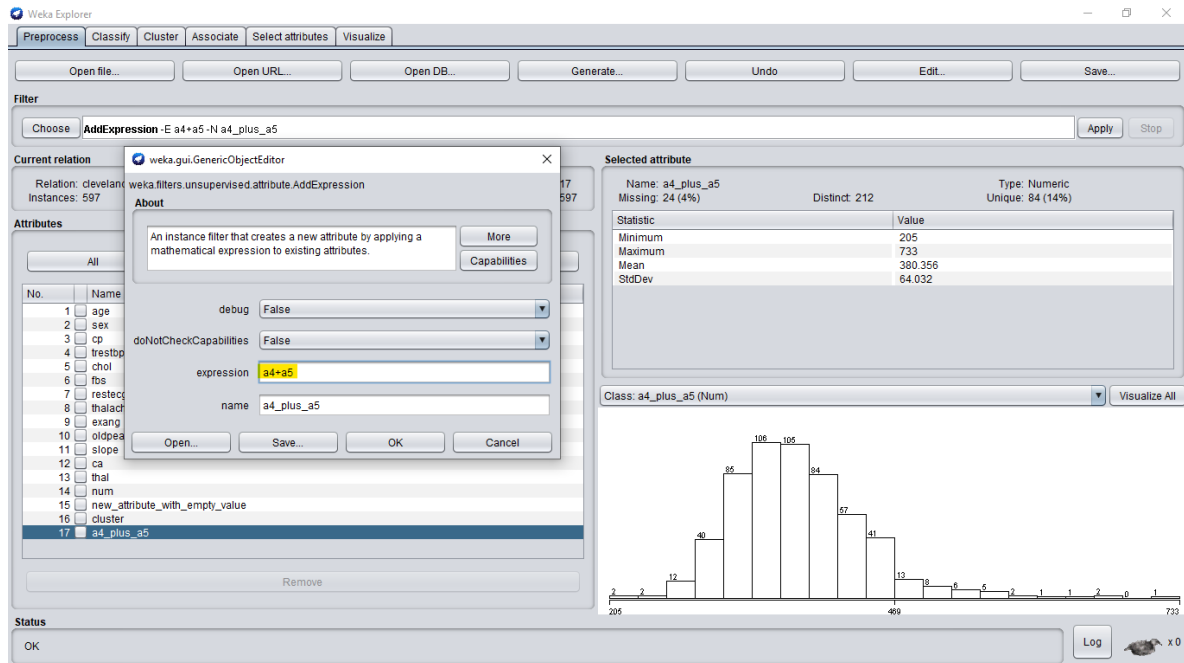
Ta thao tác tương tự như `attribute.Add`, bộ lọc này sẽ thêm vào một thuộc tính *nomial* thể hiện kết quả phân cụm cho các mẫu khi áp dụng một thuật toán phân cụm nào đó.



Hình 18: Thuộc tính *cluster* được thêm vào với thuật toán *SimpleKMeans*.

- `attribute.AddExpression`

Bộ lọc này sẽ thêm vào một thuộc tính mới bằng cách áp dụng một phép toán nào đó trên những thuộc tính sẵn có.



Hình 19: Thêm một thuộc tính là tổng của hai thuộc tính sẵn có, tên các biến có dạng $a + \text{số thứ tự của thuộc tính}$. Như ví dụ trên là $a4 + a5$.

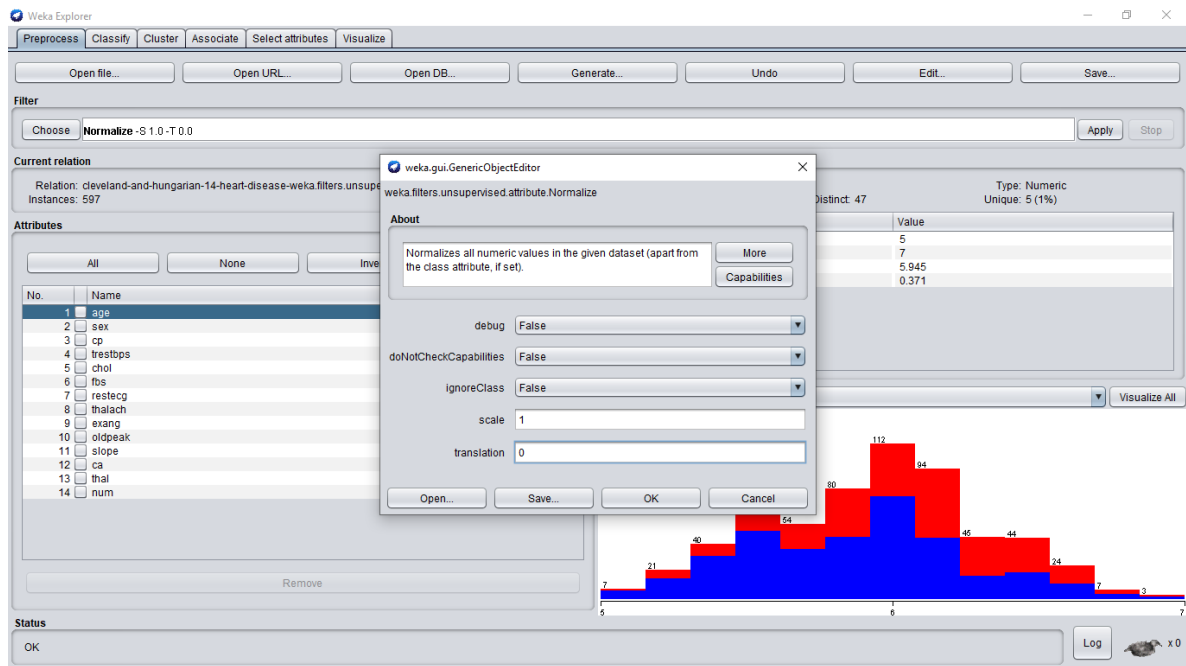
- `attribute.AddID`

Thêm thuộc tính ID vào dataset.

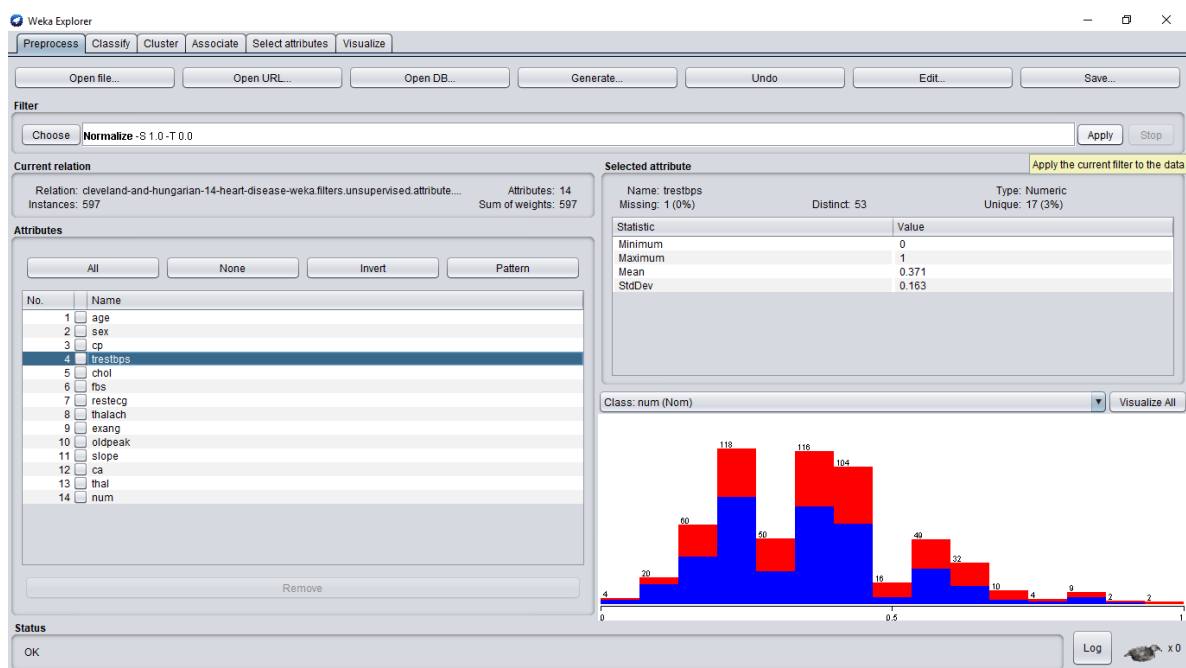
2.5.2 Chuẩn hóa – *Normalize*

Hai bộ lọc `attribute.Normalize` và `attribute.Standardize` cho phép chuẩn hóa Min-max và chuẩn hóa Z-score.

- Chuẩn hóa Min-max

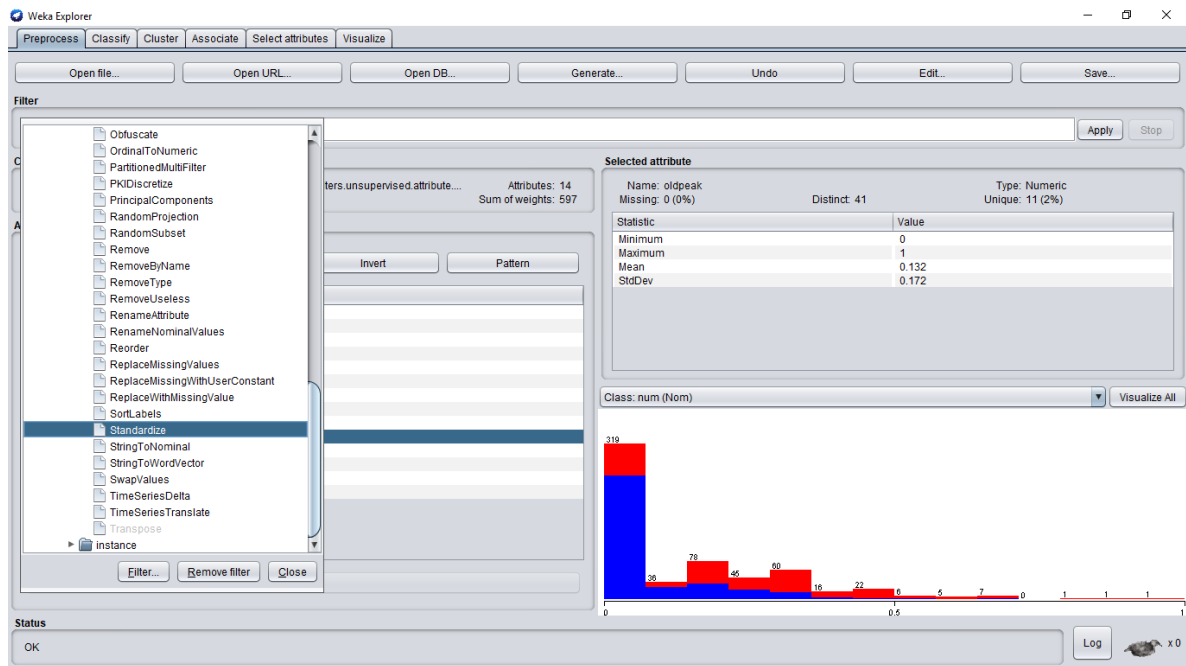


Hình 20: Ở thành *Filter*, chọn *Choose* → *weka* → *filters* → *unsupervised* → *attribute* → *Normalize*. Trong đó min là *translation*, còn max là *translation + scale*. Ví dụ, với *min* = 0, *max* = 1, ta điền như trên hình.

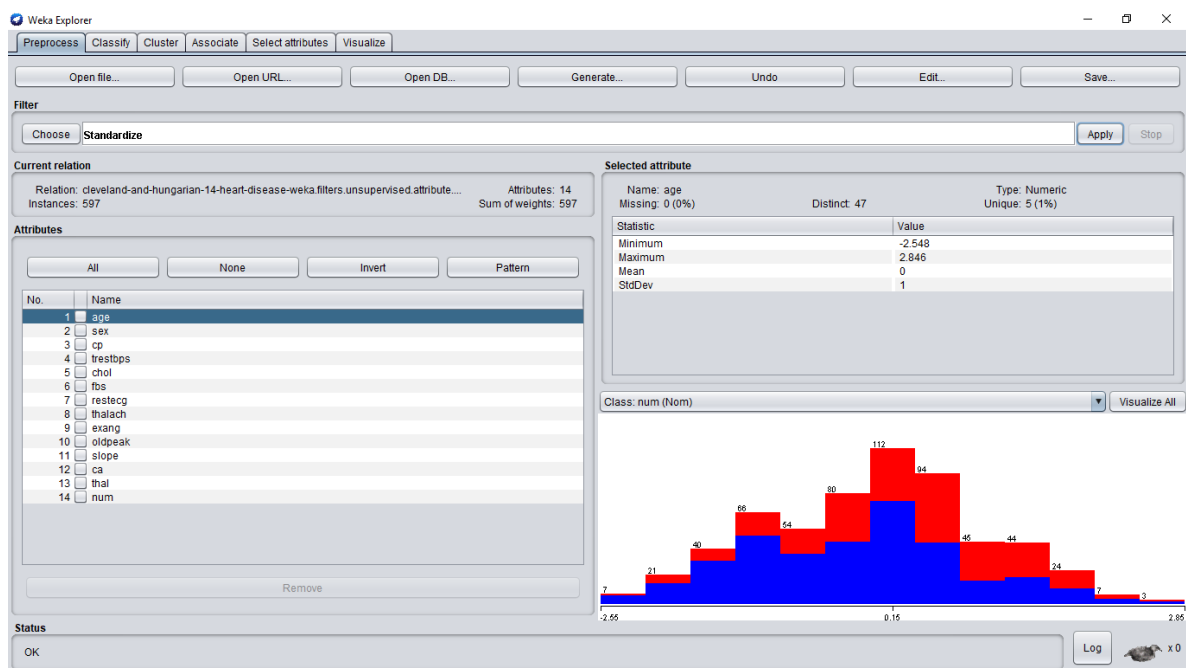


Hình 21: Nháy *Apply* để chuẩn hóa.

- Chuẩn hóa Z-score



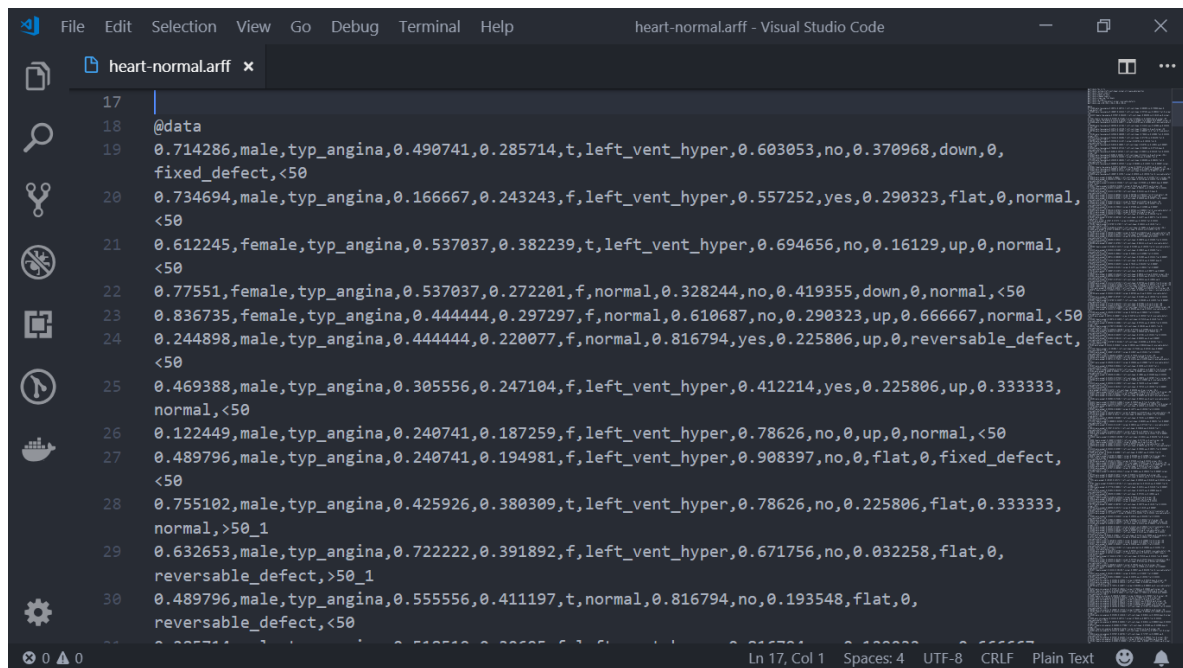
Hình 22: Ở thanh *Filter*, chọn *Choose* → *weka* → *filters* → *unsupervised* → *attribute* → *Standardize*.



Hình 23: Nháy *Apply* để chuẩn hóa.

2.5.3 Chọn phương pháp chuẩn hóa

Chọn Z-score vì sẽ thuận lợi trong trường hợp phân phối có Gaussian, đồng thời không từ chối dữ liệu mới nạp vào nằm ngoài min-max như trong chuẩn hóa Min-max.

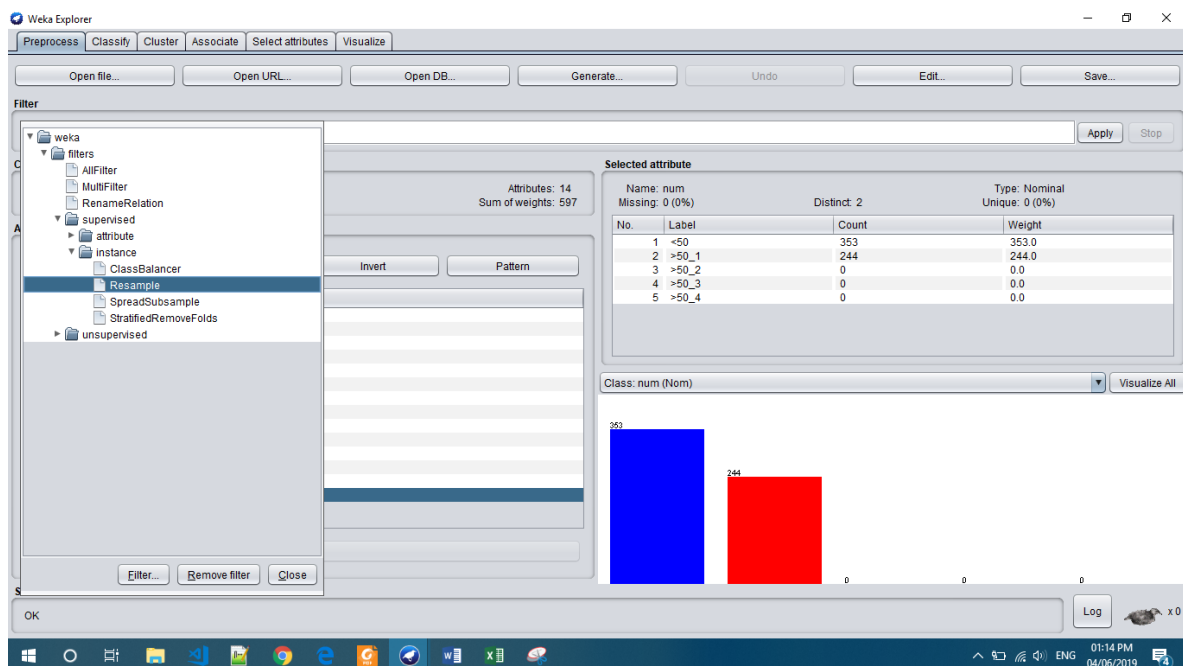


Hình 24: Sau khi chuẩn hóa Z-score.

2.6 Rút gọn dữ liệu - Reduction

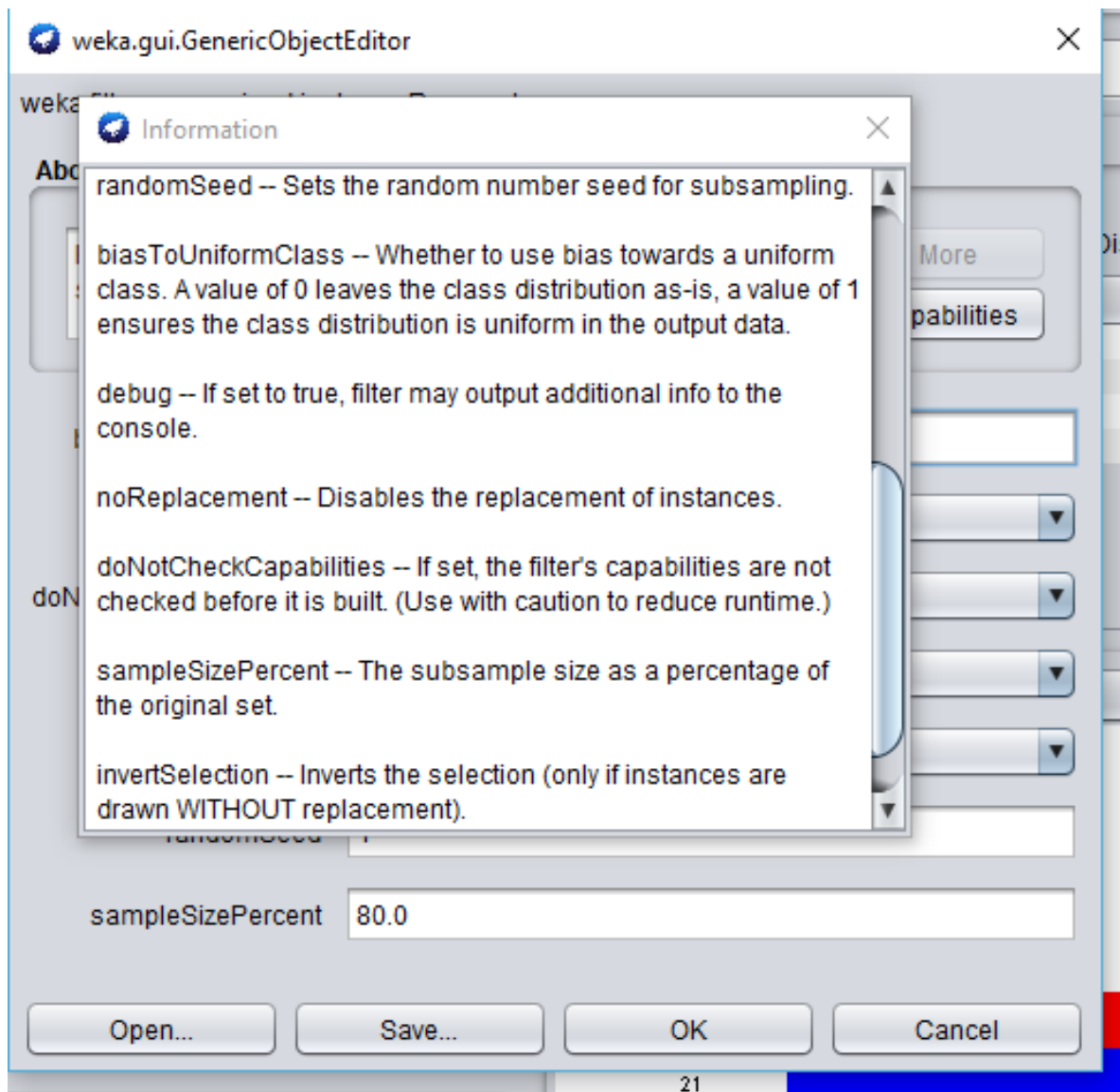
2.6.1 Lấy mẫu dữ liệu với các bộ lọc Weka

- B1: Ở mục Filter chọn nút Choose → supervised (hoặc unsupervised) → instance → Resample.

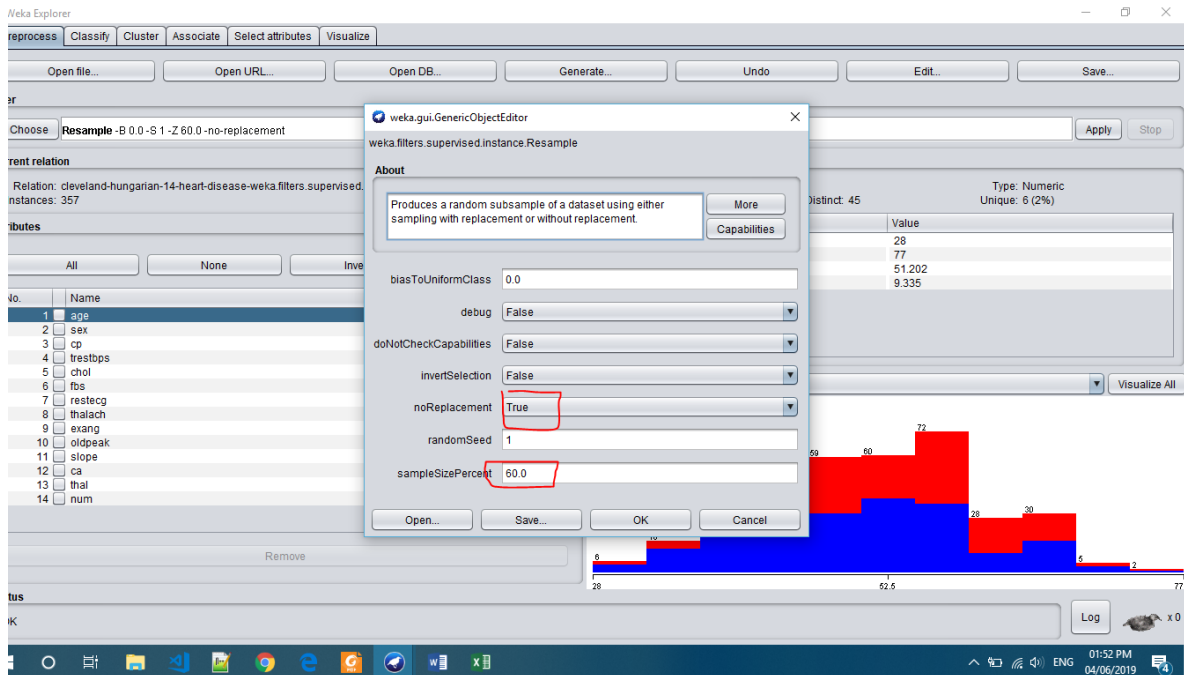


- B2: Nhấp chọn ô thông số bên cạnh nút Choose để chỉnh sửa các thông số:

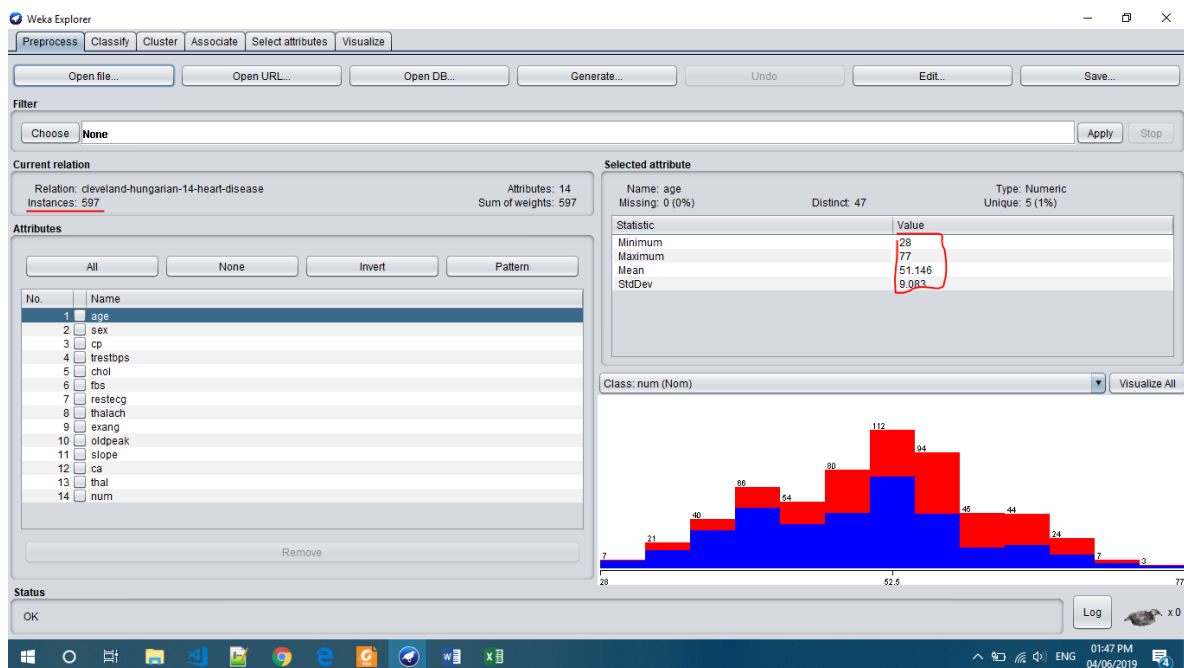
- randomSeed: đặt mầm (seed) ngẫu nhiên để lấy mẫu
- biasToUniformClass (thông số này chỉ có khi chọn mục supervised): giá trị 0 – phân phối của lớp như hiện trạng (như bộ dữ liệu đầu). Giá trị 1 – phân phối lớp của mẫu là phân phối chuẩn (uniform).
- noReplacement: True – with replacement (lấy tuple dữ liệu ra và trả lại nó trong bộ dữ liệu D, nghĩa là có thể nó sẽ vẫn có thể được rút ra trong lần lấy tuple sau). False – without replacement (lấy tuple dữ liệu ra (với xác suất rút $1/N$) và không trả nó lại trong bộ dữ liệu D, nghĩa là mỗi tuple chỉ xuất hiện 1 lần trong mẫu).
- sampleSizePercent: kích thước của mẫu so với bộ dữ liệu gốc bao nhiêu phần trăm.
- invertSelection: lấy mẫu với các bộ ngược lại (inverse) so với mẫu lấy được (chỉ áp dụng với mẫu không có lặp – without replacement).



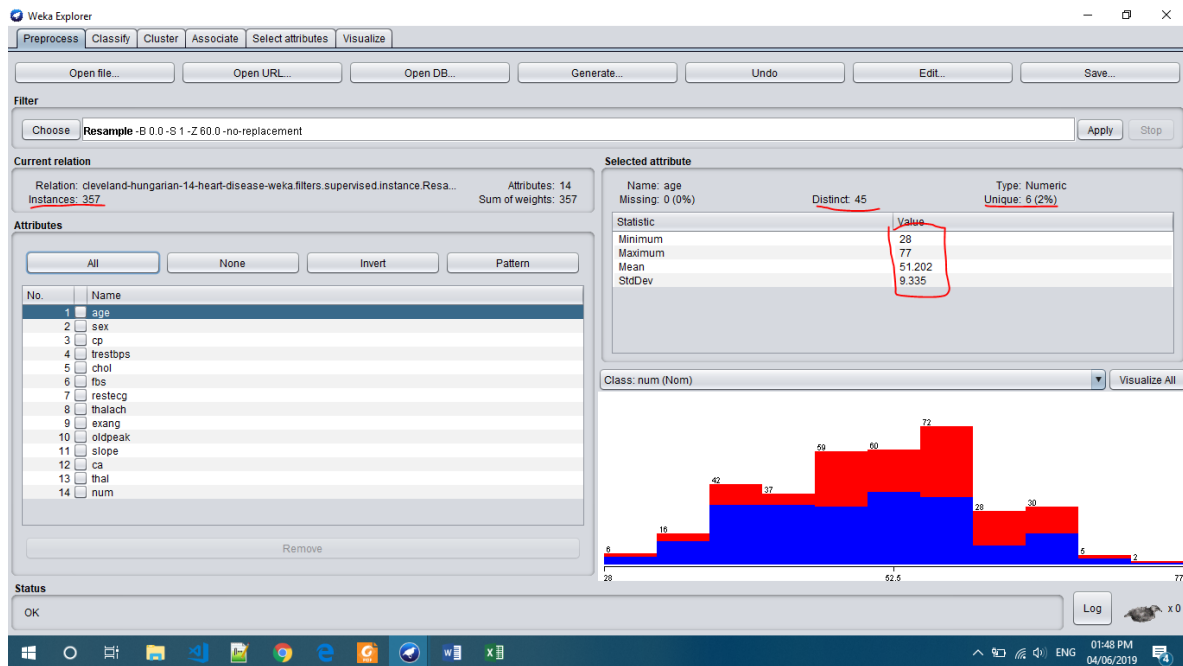
- B3: Sau khi chọn các thông số theo nhu cầu, nhấn nút Apply để lấy mẫu (sampling).



Hình 25: Ví dụ: ở đây nhóm lấy mẫu với các thông số: noReplacement = True, sampleSize = 60.



Hình 26: Ảnh bộ dữ liệu gốc.



Hình 27: Mẫu lấy được với các thông số như trên.

2.6.2 Khả năng thực hiện của Weka

Trong Weka, ta có thể thực hiện 2 phương pháp chính: Simple Random Sample Without Replace (SRSWOR) và Simple Random Sample With Replacement (SRSWR) bằng cách đặt thông số option noReplacement là True (không lặp) hoặc False (có lặp).

3 Đánh giá

STT	Nội dung	Hoàn thành
1	Tích hợp dữ liệu.	100%
2	Tóm tắt mô tả dữ liệu.	100%
3	Chọn lọc dữ liệu.	100%
4	Làm sạch dữ liệu.	100%
5	Chuyển đổi dữ liệu.	100%
6	Rút gọn dữ liệu.	100%
Mức độ hoàn thành tổng thể của bài tập:		100%

Tài liệu

- [1] Slide lý thuyết.
- [2] Trang chủ của Weka: <http://www.cs.waikato.ac.nz/ml/weka/>
- [3] J. Han and M. Kamber, *Data Mining, Concepts and Techniques, Second Edition*