

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



BÁO CÁO
BÀI TẬP THỰC HÀNH 1
TIỀN XỬ LÝ DỮ LIỆU VỚI WEKA

CHUYÊN NGÀNH:
KHỌC HỌC MÁY TÍNH

Thành phố Hồ Chí Minh, ngày 1 tháng 4 năm 2019

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



BÁO CÁO
BÀI TẬP THỰC HÀNH 1
TIỀN XỬ LÝ DỮ LIỆU VỚI WEKA

|Giảng viên hướng dẫn|

Thầy, cô: Lê Ngọc Thành, Nguyễn Ngọc Thảo

|Sinh viên|

Phạm Phong Hòa

1612176

MÔN HỌC: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Thành phố Hồ Chí Minh, ngày 1 tháng 4 năm 2019

Mục lục

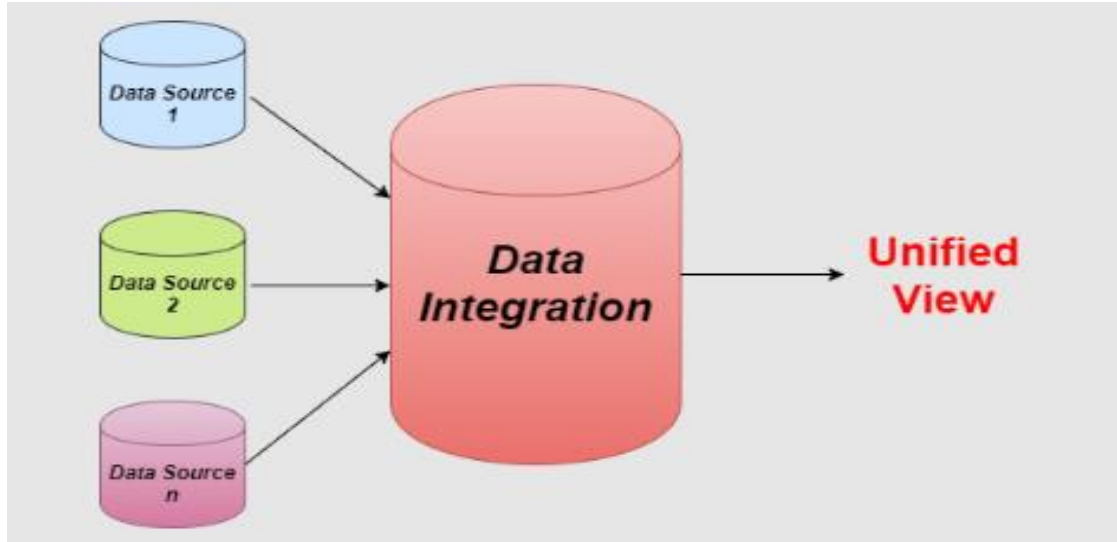
I. TÍCH HỢP DỮ LIỆU (INTEGRATION)	5
1. Định nghĩa sự tích hợp dữ liệu:	5
2. Vấn đề về nhận diện thực thể (<i>entity identification</i>) và cách giải quyết:	6
3. Vấn đề dữ liệu dư thừa (<i>redundancy</i>) và cách giải quyết:	9
4. Sự mâu thuẫn dữ liệu (<i>data value conflicts</i>) và cách giải quyết: ^[5]	10
5. Tích hợp các dataset và số lượng mâu thuẫn khi tích hợp và số lượng thuộc tính:	10
6. Chụp màn hình của cửa sổ Explorer:	11
II. TÓM TẮT MÔ TẢ DỮ LIỆU – DESCRIPTIVE DATA SUMMARIZATION:	13
1. Trong tab Preprocess, xem xét thuộc tính age:	13
2. Five-number summary:	14
3. Các loại thuộc tính:	14
4. Ý nghĩa của đồ thị trong cửa sổ Explorer:	14
5. Xem xét các thuộc tính khác của dataset dưới dạng đồ thị.	15
6. Nhận xét:	20
7. Tab Visualize.	20
8. Những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau:	21
III. CHỌN LỌC DỮ LIỆU (SELECTION):	21
1. Thuộc tính trong những dataset trước khi xử lý:	21
2. Tab Select attributes và những lựa chọn khác nhau của Weka để chọn lọc thuộc tính: [7]	24
3. So sánh với các phương pháp chọn lọc dữ liệu trong textbook và phương pháp không có trong Weka:	25
IV. LÀM SẠCH DỮ LIỆU (CLEANING):	26
1. Các giá trị thiếu (<i>Missing values</i>):	26
2. Dữ liệu nhiễu (<i>Noisy data</i>):	27
3. Detect tìm dữ liệu tạp (<i>Outlier detection</i>): ^[12]	28
4. File heart-cleaned.arff:	33
V. CHUYỂN ĐỔI DỮ LIỆU (TRANSFORMATION):	34
1. Xây dựng thuộc tính – Attribute construction:	35
2. Chuẩn hóa – Normalize một thuộc tính:	35

3.	<i>Tiến hành chuẩn hóa tất cả các thuộc tính là số thực:</i>	37
4.	<i>File heart-normal.arff:</i>	37
VI.	RÚT GỌN DỮ LIỆU (REDUCTION):	38
VII.	TỔNG KẾT:	40

I. TÍCH HỢP DỮ LIỆU (INTEGRATION)

1. Định nghĩa sự tích hợp dữ liệu:

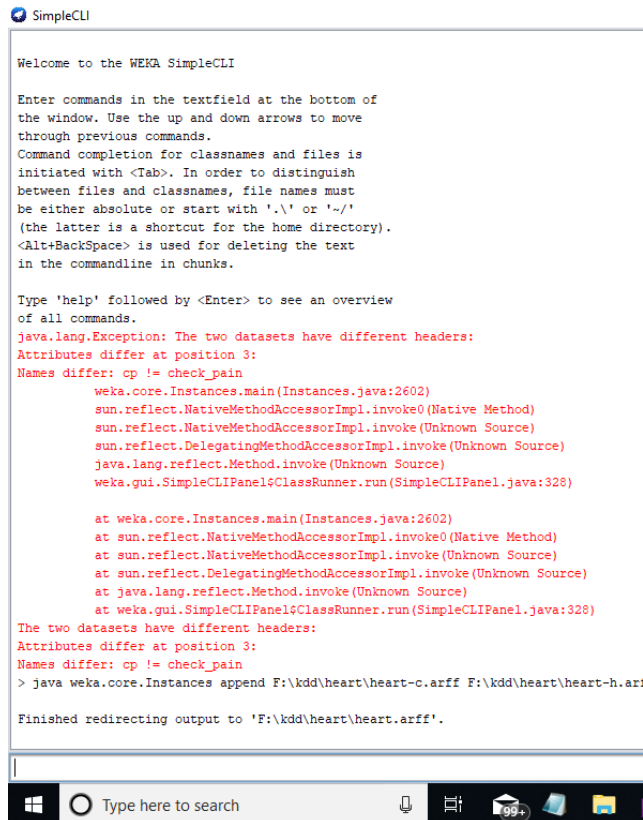
- Khái niệm: Là một kỹ thuật tiền xử lý dữ liệu mà trong đó dữ liệu từ nhiều nguồn khác nhau được kết hợp lại và sau đó cung cấp cho người dùng một cái nhìn thống nhất.



- Các nguồn ở đây có thể được bao gồm nhiều cơ sở dữ liệu, data cubes hoặc flat files. Một trong những hướng cài đặt có phần chiếm ưu thế hiện nay là xây dựng một kho chứa dữ liệu doanh nghiệp.
- Ta thấy rằng lợi ích của kho chứa là cho phép doanh nghiệp biến diễn những phân tích dựa trên dữ liệu có sẵn trong kho. ^[1]
- Việc tích hợp dữ liệu cẩn thận giúp ta giảm và tránh sự dư thừa và tính không thống nhất của dữ liệu kết quả. ^[2]
- Có 3 vấn đề được xem xét xuyên suốt trong quá trình tích hợp dữ liệu: ^[3]
 - Sơ đồ tích hợp
 - Sự dư thừa dữ liệu
 - Phát hiện và tái giải quyết các giá trị gặp mâu thuẫn.

2. Vấn đề về nhận diện thực thể (entity identification) và cách giải quyết:

- Khi mở hai dataset bằng notepad++ ta thấy, ở thuộc tính thứ ‘cp’ của ‘heart-c.arff’ và ‘check-pain’ của ‘heart-h.arff’ thực chất chỉ là một thuộc tính duy nhất và chúng có tên khác nhau. Do đó, khi tích hợp dữ liệu bằng Simple CLI của weka đã thông báo lỗi này. Như vậy suy ra có vấn đề nhận thực thể ở đây.



```
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.

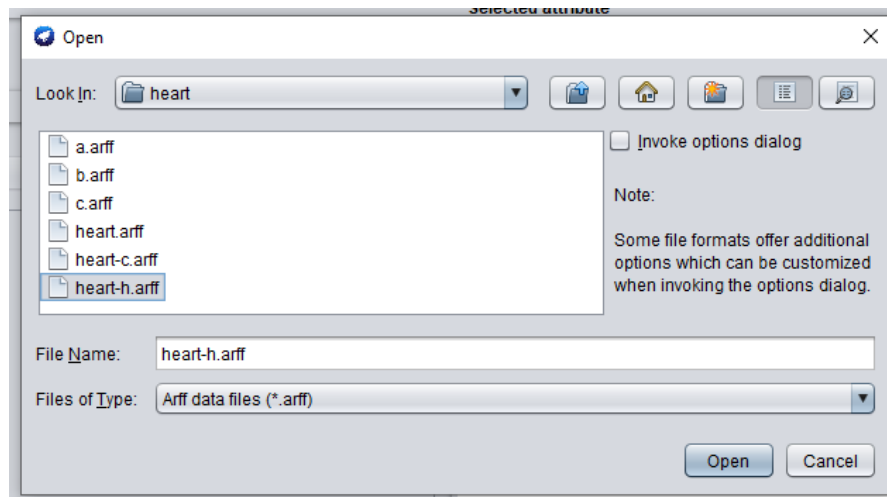
Type 'help' followed by <Enter> to see an overview
of all commands.
java.lang.Exception: The two datasets have different headers:
Attributes differ at position 3:
Names differ: cp != check_pain
    weka.core.Instances.main(Instances.java:2602)
    sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    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: cp != check_pain
> java weka.core.Instances append F:\kdd\heart\heart-c.arff F:\kdd\heart\heart-h.arff

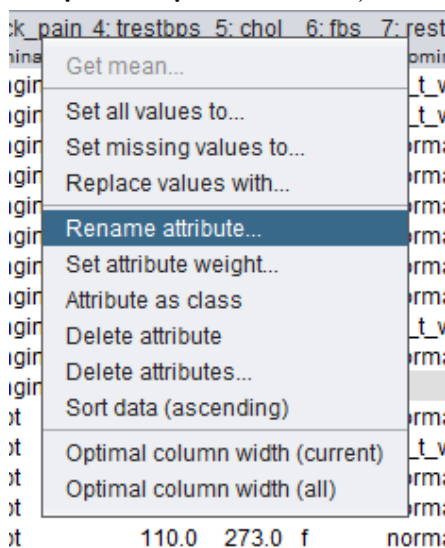
Finished redirecting output to 'F:\kdd\heart\heart.arff'.
```

File heart-c.arff:
@attribute 'cp' {typ_angina, asympt, non_anginal, atyp_angina}
File heart-h.arff:
@attribute 'chest_pain' {typ_angina, asympt, non_anginal, atyp_angina}
File heart.arff (file tích hợp từ heart-c.arff và heart-h.arff)
@attribute 'cp' {typ_angina, asympt, non_anginal, atyp_angina}

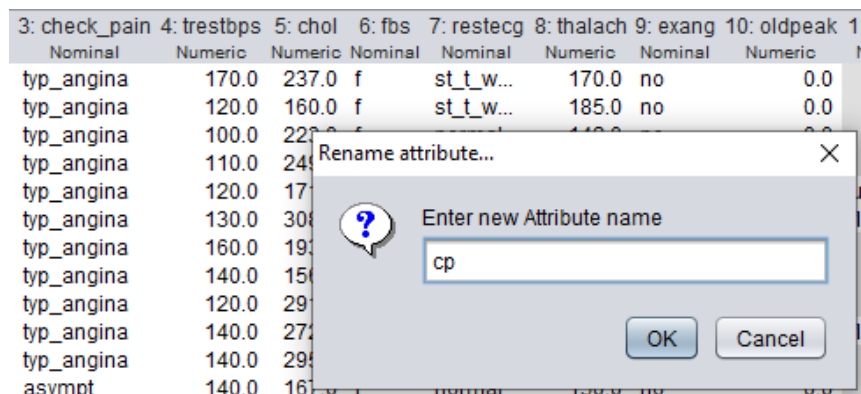
- Trường hợp nếu có thì có cách giải quyết như sau:
 - Mở hai dataset lên bằng notepad++ và sửa trực tiếp để thống nhất một tên gọi duy nhất.
 - Ở đây sẽ thống nhất chọn tên là ‘cp’.
- Một cách khác để tiến hành sửa tên thuộc tính thứ 3 như thông báo ở trên là tiến hành sửa bằng giao diện Explorer như sau:
 - Mở cửa sổ Explorer:
Trong tab Preprocessing, chọn Open File → chọn tiếp tập tin cần hiệu chỉnh → chọn tiếp Open



- Chọn tiếp button Edit, cửa mới hiện lên các cột và các dòng record, sau đó rê chuột vào cột cần đổi tên, rồi nhấn chuột phải chọn Rename Attribute ...



- Chọn Rename attribute ... sau đó sửa lại tên thuộc tính theo đúng như giá trị thống nhất. Ở đây sẽ gõ vào là 'cp' thay thế cho 'check_pain'. Chọn tiếp OK. Chọn tiếp OK thì cửa sổ view sẽ tự nhiên lưu lại và sửa lại tên thuộc tính.



- Sau khi sửa xong ta chạy lại thì lại báo lỗi khác:

```
SimpleCLI
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: cp != check_pain
> java weka.core.Instances append F:\kdd\heart\heart-c.arff F:\kdd\heart\heart-h.arff > F:\kdd\heart\heart.arff

Finished redirecting output to 'F:\kdd\heart\heart.arff'.
> java weka.core.Instances append F:\kdd\heart\heart-c.arff F:\kdd\heart\heart-h.arff > F:\kdd\heart\heart.arff

Finished redirecting output to 'F:\kdd\heart\heart.arff'.
java.lang.Exception: The two datasets have different headers:
Attributes differ at position 11:
Labels differ at position 1: up != down
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 11:
Labels differ at position 1: up != down
```

- Do thứ tự các thuộc tính trong tập giá trị của thuộc tính ‘slope’ bị khác nhau (tuy vẫn cùng là một tập nhưng lại không thể tích hợp được).
- Cách ở đây là mở notepad++ rồi sửa trực tiếp trên file heart-h.arff chỗ thuộc tính slope thành {up, flat, down}
- Tới đây thì ta chạy lại thì lệnh tích hợp báo thành công.

```
> java weka.core.Instances append F:\kdd\heart\heart-c.arff F:\kdd\heart\heart-h.arff > F:\kdd\heart\heart.arff

Finished redirecting output to 'F:\kdd\heart\heart.arff'.
```

- Ta mở file và kiểm tra file heart.arff:


```

F:\kdd\heart\heart.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 13 new 14 Course h SMS h Info h new 15 heart.c.arff heart.arff
1 @relation cleveland-14-heart-disease
2
3 @attribute age numeric
4 @attribute sex {female,male}
5 @attribute cp {typ_angina,asympt,non_anginal,atyp_angina}
6 @attribute trestbps numeric
7 @attribute chol numeric
8 @attribute fbs {t,f}
9 @attribute restecg {left_vent_hyper,normal,st_t_wave_abnormality}
10 @attribute thalach numeric
11 @attribute exang {no,yes}
12 @attribute oldpeak numeric
13 @attribute slope {up,flat,down}
14 @attribute ca numeric
15 @attribute thal {fixed_defect,normal,reversable_defect}
16 @attribute num {<50,>50_1,>50_2,>50_3,>50_4}
17
18 @data
19
20 63,male,typ_angina,145,233,t,left_vent_hyper,150,no,2.3,down,0,fixed_defect,<50
21 67,male,asympt,160,286,f,left_vent_hyper,108,yes,1.5,flat,3,normal,>50_1
22 67,male,asympt,120,229,f,left_vent_hyper,129,yes,2.6,flat,2,reversable_defect,>50_1
23 37,male,non_anginal,130,250,f,normal,187,no,3.5,down,0,normal,<50
24 41,female,atyp_angina,130,204,f,left_vent_hyper,172,no,1.4,up,0,normal,<50
25 56,male,atyp_angina,120,236,f,normal,178,no,0.8,up,0,normal,<50
26 62,female,asympt,140,268,f,left_vent_hyper,160,no,3.6,down,2,normal,>50_1
27 57,female,asympt,120,354,f,normal,163,yes,0.6,up,0,normal,<50
28 63,male,asympt,130,254,f,left_vent_hyper,147,no,1.4,flat,1,reversable_defect,>50_1
29 53,male,asympt,140,203,t,left_vent_hyper,155,yes,3.1,down,0,reversable_defect,>50_1
30 57,male,asympt,140,192,f,normal,148,no,0.4,flat,0,fixed_defect,<50
31 56,female,atyp_angina,140,294,f,left_vent_hyper,153,no,1.3,flat,0,normal,<50
32 56,male,non_anginal,130,256,t,left_vent_hyper,142,yes,0.6,flat,1,fixed_defect,>50_1
Normal text file length: 40,880 lines: 617 Ln: 12 Col:
Type here to search

```

3. Vấn đề dữ liệu dư thừa (redundancy) và cách giải quyết:

- Không tồn tại các vấn đề dư thừa.
- Nếu có ta giải quyết theo cách như sau: Ta phân tích tương quan (correlation analysis): ^[4]
 - Dựa trên dữ liệu hiện có, kiểm tra khả năng dẫn ra một thuộc tính B từ thuộc tính A.
 - Đối với các thuộc tính số (numerical attributes), đánh giá tương quan giữa hai thuộc tính với các hệ số tương quan (correlation coefficient, aka Pearson's product moment coefficient).
 - Đối với các thuộc tính rời rạc (categorical/discrete attributes), đánh giá tương quan giữa hai thuộc tính với phép kiểm thử chisquare (χ^2).
- Mặt khác tồn tại vấn đề trùng các record với nhau:
 - Khi tích hợp ta thấy có 597 dòng record:

Current relation

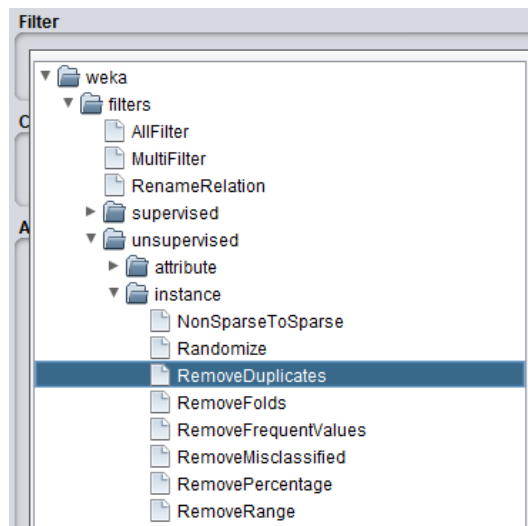
Relation: cleveland-14-heart-disease

Instances: 597

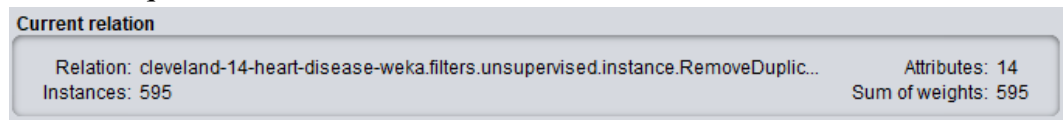
Attributes: 14

Sum of weights: 597

- Để biết có bị tình trạng trùng record hay không, ta làm các bước như sau:
 Vào choose, chọn filter, chọn tiếp unsupervised, chọn instance, chọn RemoveDuplicate.



- Trong khung Attributes, chọn All, sau đó chọn Apply trên dòng Choose. Ta được kết quả như sau:



- Số dòng record còn lại, suy ra đã xuất hiện tình trạng duplicate.

4. Sự mâu thuẫn dữ liệu (data value conflicts) và cách giải quyết: ^[5]

- Không xuất hiện sự mâu thuẫn.
- Nếu có giải quyết như sau, ta sẽ xét từng trường hợp như sau.
 - Representation: “2004/12/25” với “25/12/2004”: *Thực chất ở đây đang là xử lý lại chuỗi kí tự cho đạt một thống nhất chung.*
 - Scaling: thuộc tính weight trong các hệ thống đo khác nhau với các đơn vị đo khác nhau, thuộc tính price trong các hệ thống tiền tệ khác nhau với các đơn vị tiền tệ khác nhau: *Thực chất ở đây xử lý biến đổi lại giá trị thuộc tính bằng một công thức toán học. Cho nên buộc phải lập trình trực tiếp để biến đổi.*
 - Encoding: “yes” và “no” với “1” và “0”: *Đây là xử lý chuỗi và kiểu Boolean, buộc phải lập trình tùy vào trường hợp cụ thể.*

5. Tích hợp các dataset và số lượng mâu thuẫn khi tích hợp và số lượng thuộc tính:

- Ta có câu lệnh tích hợp hai dataset trên như sau

```
java weka.core.Instances append <path file 1>
<khoảng trắng> <path file 2> .... > <khoảng
trắng><path file output>
```

- Về quy trình thì đã được trình bày trong mục 3 “Vấn đề về nhận diện thực thể (entity identification) và cách giải quyết”
- File heart.arff sau khi được tích hợp là:

```

1 @relation cleveland-14-heart-disease
2
3 @attribute age numeric
4 @attribute sex {female,male}
5 @attribute cp {typ_angina,asympt,non_anginal,atyp_angina}
6 @attribute trestbps numeric
7 @attribute chol numeric
8 @attribute fbs {t,f}
9 @attribute restecg {left_vent_hyper,normal,st_t_wave_abnormality}
10 @attribute thalach numeric
11 @attribute exang {no,yes}
12 @attribute oldpeak numeric
13 @attribute slope {up,flat,down}
14 @attribute ca numeric
15 @attribute thal {fixed_defect,normal,reversable_defect}
16 @attribute num {<50,>50_1,>50_2,>50_3,>50_4}
17
18 @data
19
20 63,male,typ_angina,145,233,t,left_vent_hyper,150,no,2.3,down,0,fixed_defect,<50
21 67,male,asympt,160,286,f,left_vent_hyper,108,yes,1.5,flat,3,normal,>50_1
22 67,male,asympt,120,229,f,left_vent_hyper,129,yes,2.6,flat,2,reversable_defect,>50_1
23 37,male,non_anginal,130,250,f,normal,187,no,3.5,down,0,normal,<50
24 41,female,atyp_angina,130,204,f,left_vent_hyper,172,no,1.4,up,0,normal,<50
25 56,male,atyp_angina,120,236,f,normal,178,no,0.8,up,0,normal,<50
26 62,female,asympt,140,268,f,left_vent_hyper,160,no,3.6,down,2,normal,>50_1
27 57,female,asympt,120,354,f,normal,163,yes,0.6,up,0,normal,<50
28 63,male,asympt,130,254,f,left_vent_hyper,147,no,1.4,flat,1,reversable_defect,>50_1
29 53,male,asympt,140,203,t,left_vent_hyper,155,yes,3.1,down,0,reversable_defect,>50_1
30 57,male,asympt,140,192,f,normal,148,no,0.4,flat,0,fixed_defect,<50
31 56,female,atyp_angina,140,294,f,left_vent_hyper,153,no,1.3,flat,0,normal,<50
32 56,male,non_anginal,130,256,t,left_vent_hyper,142,yes,0.6,flat,1,fixed_defect,>50_1
33 44,male,atyp_angina,120,263,f,normal,173,no,0,up,0,reversable_defect,<50
34 52,male,non_anginal,172,199,t,normal,162,no,0.5,up,0,reversable_defect,<50
35 57,male,non_anginal,150,168,f,normal,174,no,1.6,up,0,normal,<50
36 48,male,atyp_angina,110,229,f,normal,168,no,1,down,0,reversable_defect,>50_1
37 54,male,asympt,140,239,f,normal,160,no,1.2,up,0,normal,<50
38 58,female,non_anginal,130,275,f,normal,135,no,0.2,up,0,normal,<50

```

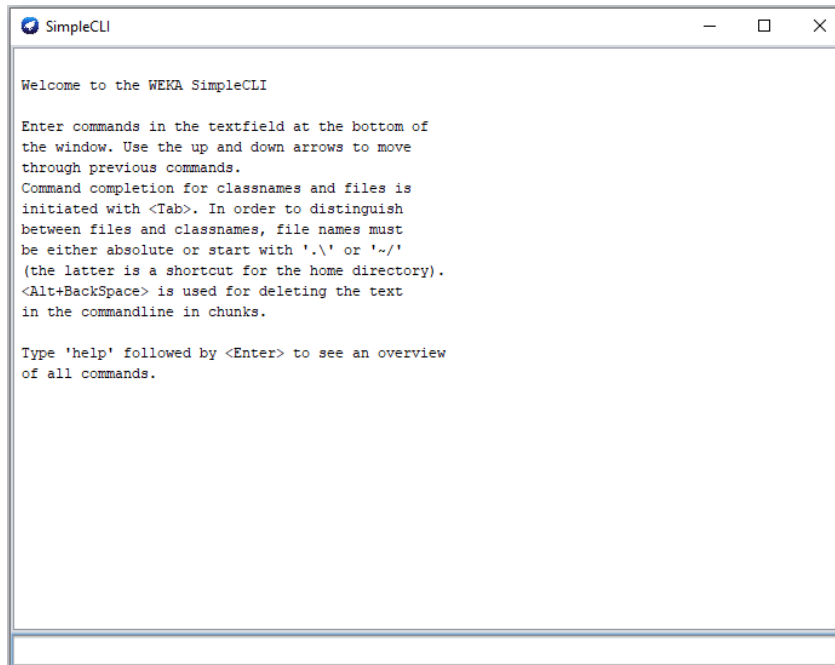
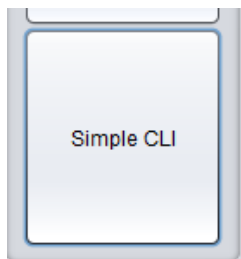
- Số lượng mẫu thuận: 0
- Số lượng thuộc tính: 14 thuộc tính

STT	Tên thuộc tính	Miền giá trị
1	age	numeric
2	sex	{female,male}
3	cp	{typ_angina,asympt,non_anginal,atyp_angina}
4	trestbps	numeric
5	chol	numeric
6	fbs	{t,f}
7	restecg	{left_vent_hyper,normal,st_t_wave_abnormality}
8	thalach	numeric
9	exang	{no,yes}
10	oldpeak	numeric
11	slope	{up,flat,down}
12	ca	numeric
13	thal	{fixed_defect,normal,reversable_defect}
14	num	{<50,>50_1,>50_2,>50_3,>50_4}

6. Chụp màn hình của cửa sổ Explorer:

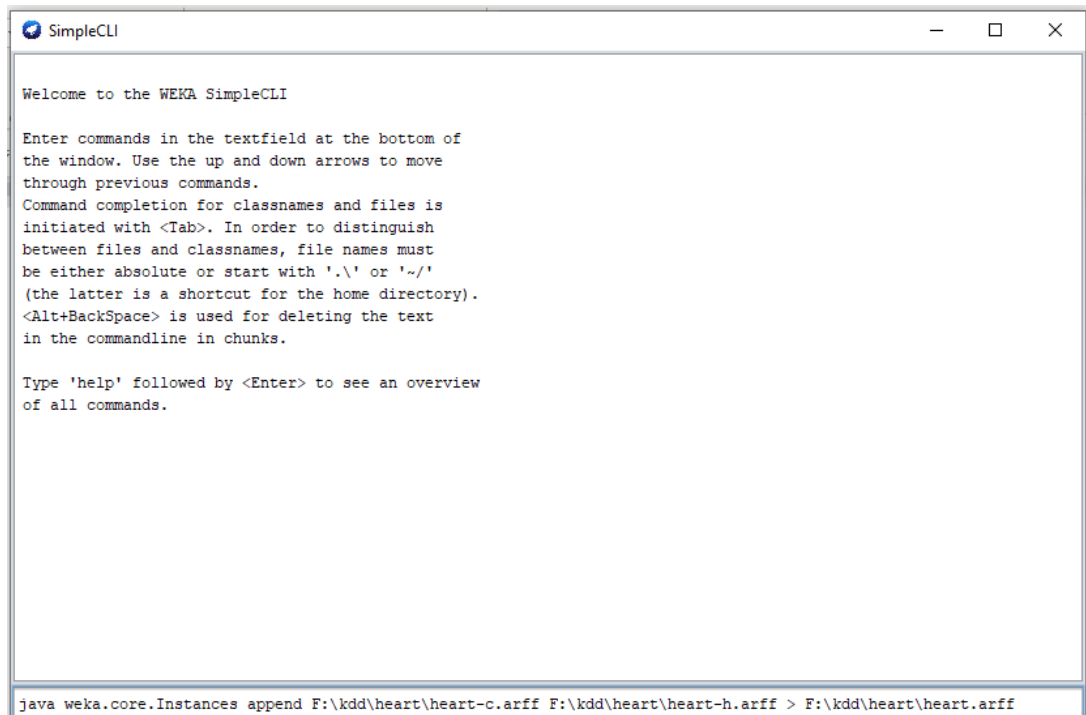
Sau đây là quy trình chi tiết cho việc mở file và tích hợp dữ liệu.

- Bước 1: Xử lý dữ liệu thuộc tính sao cho các dataset cần tích hợp phải thống nhất với nhau về thuộc tính, số lượng thuộc tính, trật tự các thuộc tính ở mỗi record, miền giá trị các thuộc tính phải như nhau và có cùng một trật tự.
- Bước 2: Mở weka chọn Simple CLI:

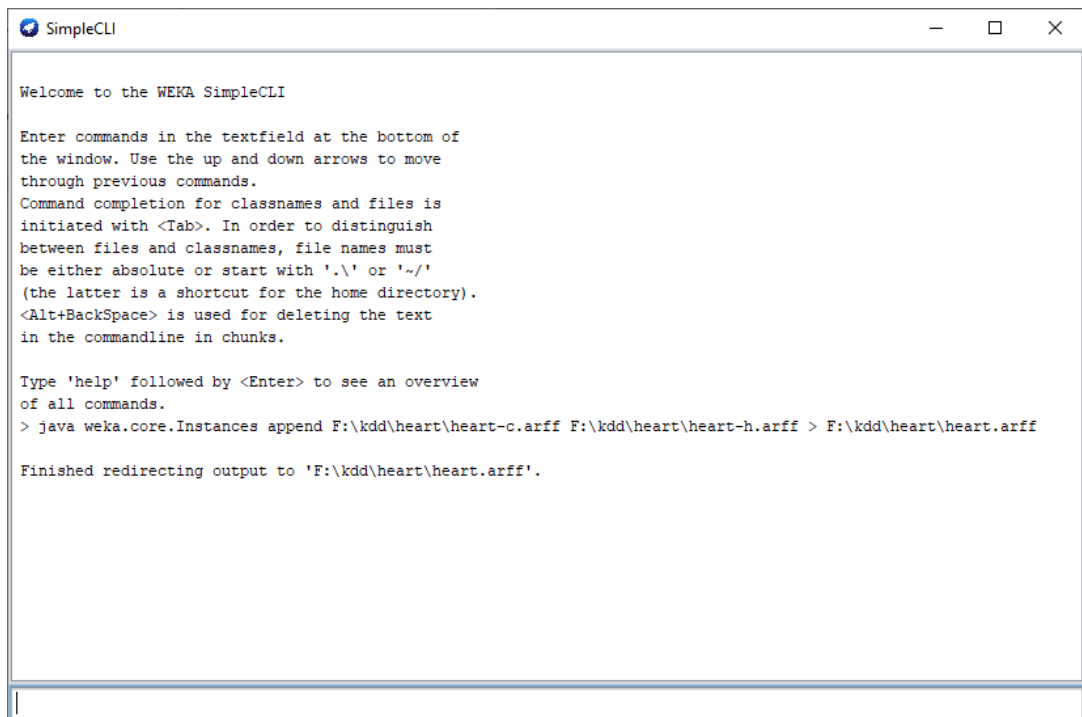


- Bước 3: Gõ dòng lệnh:

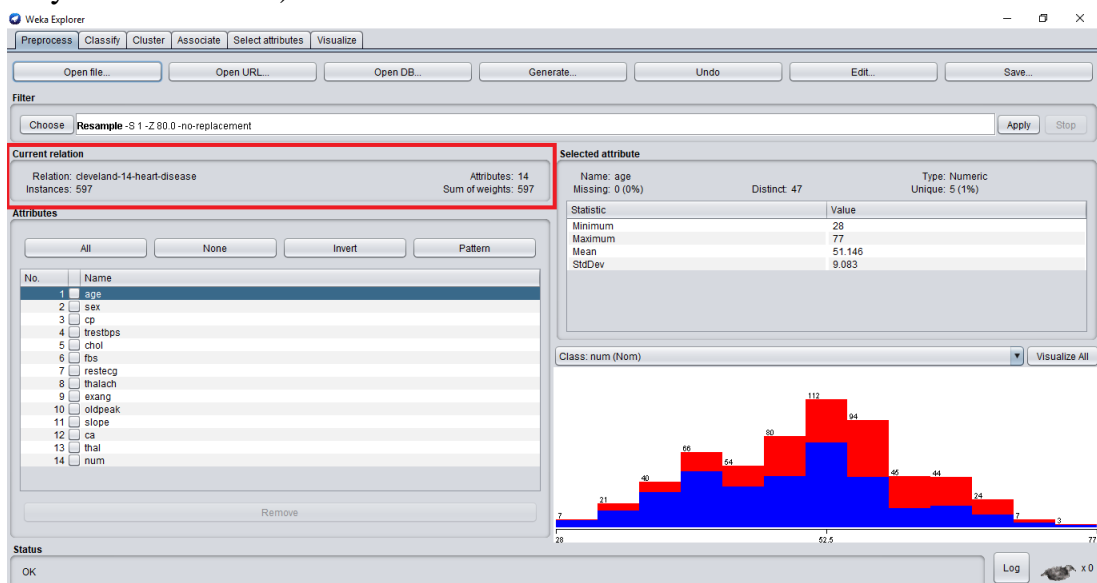
```
java weka.core.Instances append F:\kdd\heart\heart-c.arff  
F:\kdd\heart\heart-h.arff > F:\kdd\heart\heart.arff
```



Enter, màn hình sẽ hiện:



- Bước 4: Mở file bằng Explorer để kiểm tra xem file đã tích hợp chưa (Để ý thấy có 597 records).



II. TÓM TẮT MÔ TẢ DỮ LIỆU – DESCRIPTIVE DATA SUMMARIZATION:

1. Trong tab Preprocess, xem xét thuộc tính age:

Kết quả kiểm tra thấy:

Selected attribute	
Name: age	Type: Numeric
Missing: 0 (0%)	Distinct: 47
	Unique: 5 (1%)
Statistic	Value
Minimum	28
Maximum	77
Mean	51.171
StdDev	9.082

- Trung bình: Mean = 51.171
- Độ lệch chuẩn: StdDev = 9.082
- Giá trị nhỏ nhất: Minimum = 28
- Giá trị lớn nhất: Maximum = 77

2. Five-number summary:

- Ta chuyển file arff thành file csv, mở bằng công cụ excel, sau đó chuyển từ file csv sang file.xlsx.
- Khi chuyển file arff thì tất cả các cột có kiểu numeric vào excel sẽ được định dạng chuỗi, do đó ta chuyển sang định dạng số thực để dễ thao tác.
- Kế đến, ta thực hiện tính five-number summary bằng excel như sau:

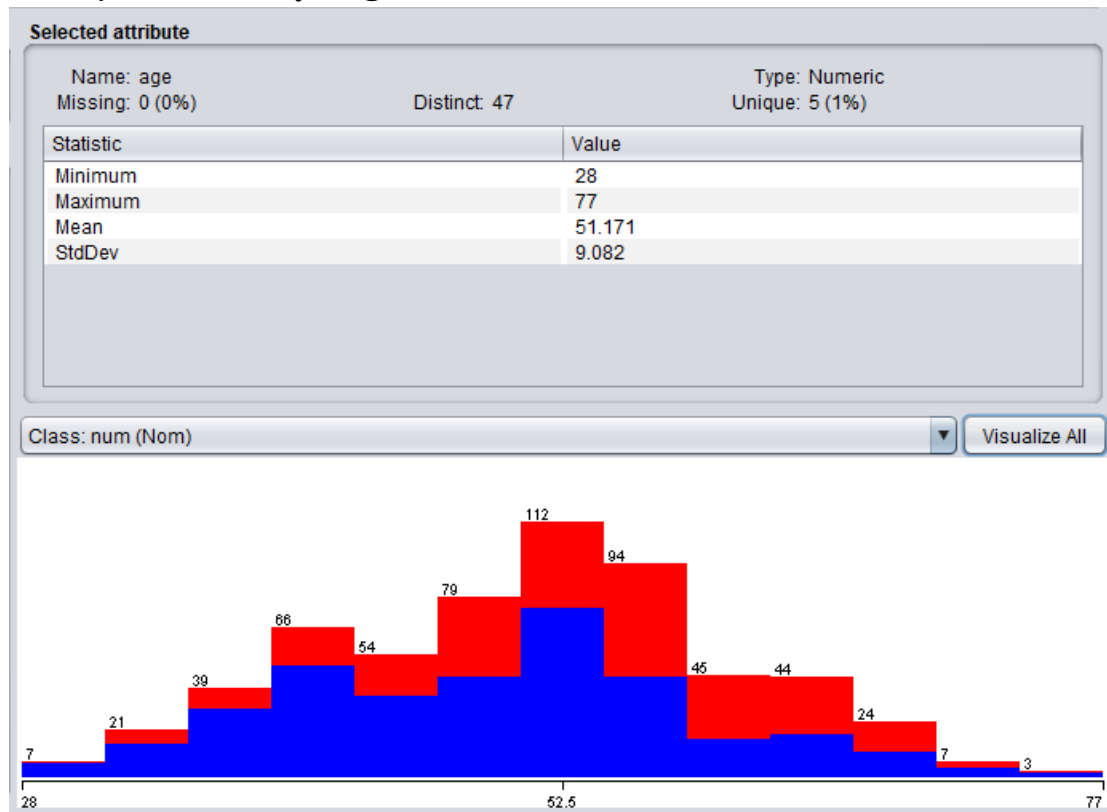
The sample minimum (smallest observation):	Max	= Max (B4:B600)	77.00
The lower quartile or first quartile:	Min	= Min (B4:B600)	28.00
The median (the middle value):	Median	= Median (B4:B600)	52.00
The upper quartile or third quartile:	Q1	= Quartile (B4:B600,1)	44
The sample maximum (largest observation):	Q3	= Quartile (B4:B600,3)	57.5

3. Các loại thuộc tính:

- Số (numeric): age, trestbps, thalach, oldpeak, ca.
- Thuộc tính có thứ tự (ordinal): Không có
- Thuộc tính nào là rời rạc/danh sách (categorical/nominal): sex, cp, fbs, restecg, exang, slope, thal, num

4. Ý nghĩa của đồ thị trong cửa sổ Explorer:

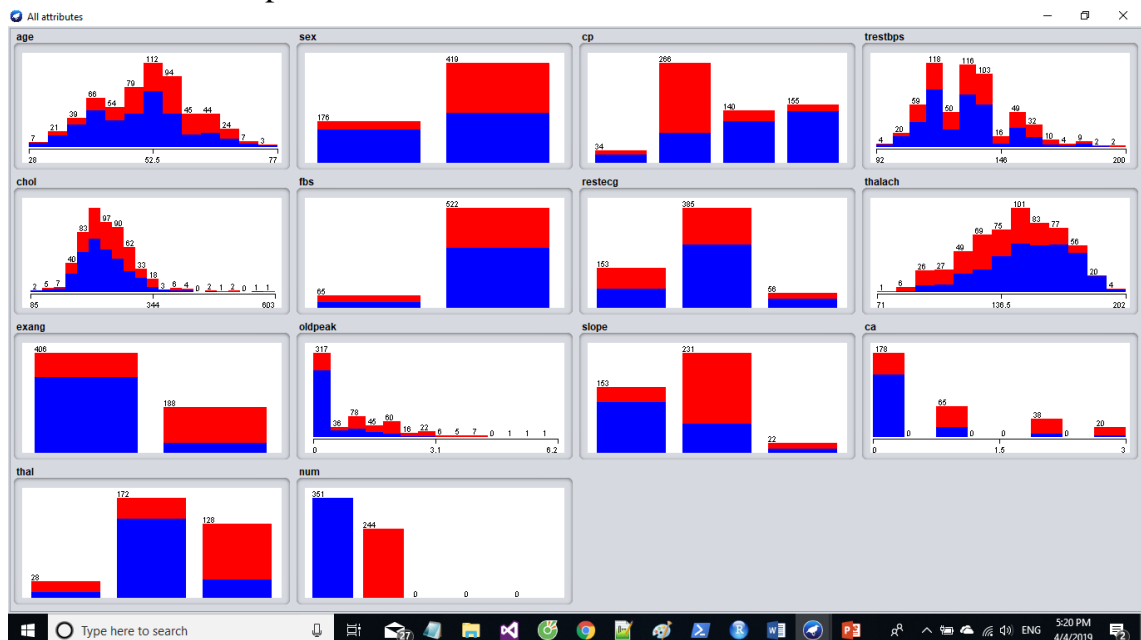
Ta chọn buổi đồ này để giải thích:



- Tên cho đồ thị: Histogram thể hiện số lượng yếu tố num lên 13 khoảng từ 28 đến 77.
- Ý nghĩa màu sắc:
 - Màu xanh: Thể hiện số lượng record có thuộc tính num = '<50'
 - Màu đỏ: thể hiện số lượng record có thuộc tính num = '>50_1'
 - Ở đây ta thấy đáng lý đồ thị này phải có 5 màu, thể hiện cho 5 giá trị '<50', '>50_1', '>50_2', '>50_3', '>50_4'. Tuy nhiên có 3 ba giá trị '>50_2', '>50_3', '>50_4' có số lượng record đều là 0. Do đó, ta chỉ thấy có hai màu xanh và đỏ.
- Đồ thị này biểu diễn cho cái gì?
 - Ta thấy đồ thị dạng cột chồng thể hiện từ giá trị 28 đến 77 của tuổi, ta có:
 - Mỗi cột có hai màu xanh đỏ chồng lên nhau, mỗi phần xanh hay đỏ thể hiện số lượng record tại thuộc tính num có giá trị '<50' đối với màu xanh, và '>50_1' đối với màu đỏ.
 - Các cột nằm sát bên nhau không có ranh giới, chứng tỏ đây là kiểu dữ liệu dạng liên tục.
 - Màu xanh ở dưới mà đỏ ở trên theo thứ tự '<50' rồi đến '>50_1'.
 - Đồ thị thể hiện mức độ hay tỉ lệ nhìn trực quan của giữa record có num = '<50' và num = '>50_1'
 - Mỗi cột sẽ hiện lên khoảng chia trên khi ta di chuột vào, điển hình ở đây có 13 cột, cho nên t có thể đoán ra khoảng chia là $(77-28)/13 \sim 3.76$

5. Xem xét các thuộc tính khác của dataset dưới dạng đồ thị.

Trước hết, ta xét phần Vizualize all:



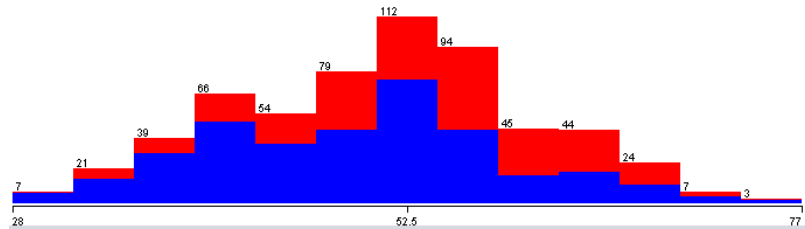
Sau đó, xét chi tiết từng đồ thị với các thông số chi tiết như sau: (Ở đây thuộc tính num luôn luôn được thể hiện dưới dạng màu sắc và cột chồng trên đồ thị).

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 47
Type: Numeric
Unique: 5 (1%)

Statistic	Value
Minimum	28
Maximum	77
Mean	51.171
StdDev	9.082

Class: num (Nom) Visualize All

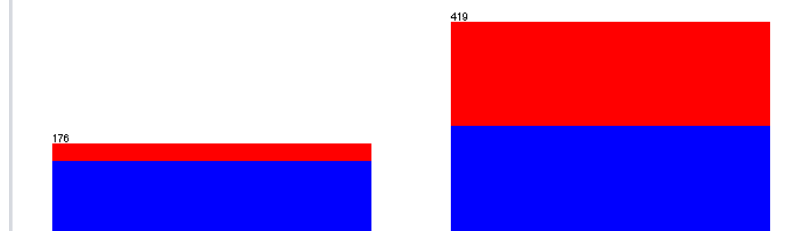


Selected attribute

Name: sex
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	female	176	176.0
2	male	419	419.0

Class: num (Nom) Visualize All

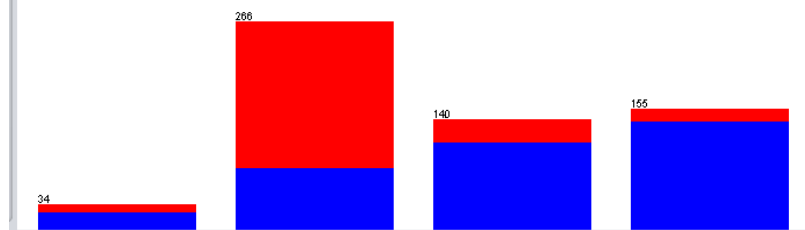


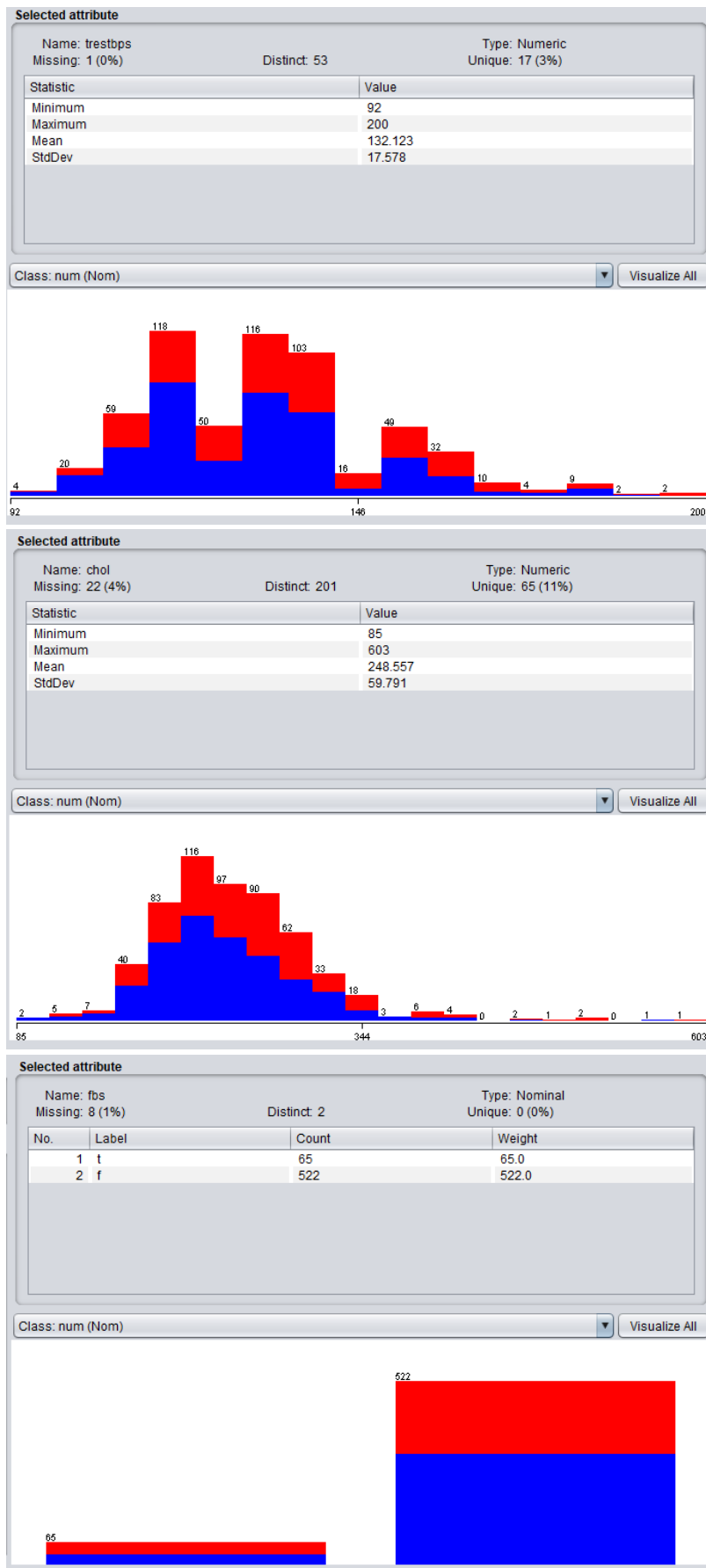
Selected attribute

Name: cp
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	typ_angina	34	34.0
2	asympt	266	266.0
3	non_anginal	140	140.0
4	atyp_angina	155	155.0

Class: num (Nom) Visualize All





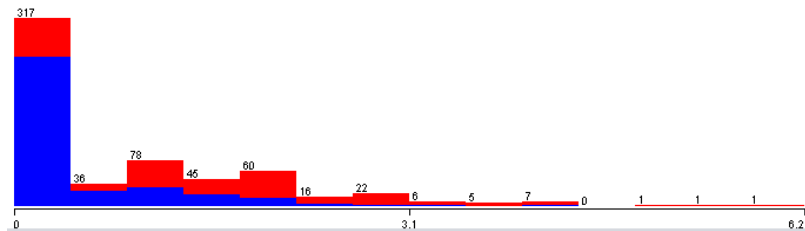


Selected attribute

Name: oldpeak
Missing: 0 (0%)
Distinct: 41
Type: Numeric
Unique: 11 (2%)

Statistic	Value
Minimum	0
Maximum	6.2
Mean	0.819
StdDev	1.069

Class: num (Nom) Visualize All

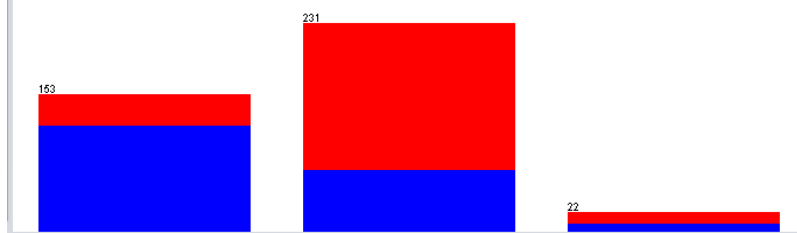


Selected attribute

Name: slope
Missing: 189 (32%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	up	153	153.0
2	flat	231	231.0
3	down	22	22.0

Class: num (Nom) Visualize All

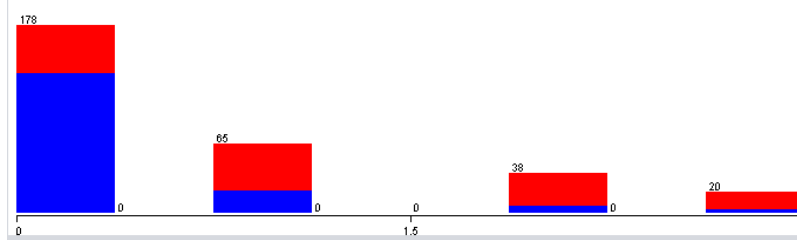


Selected attribute

Name: ca
Missing: 294 (49%)
Distinct: 4
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	3
Mean	0.668
StdDev	0.936

Class: num (Nom) Visualize All



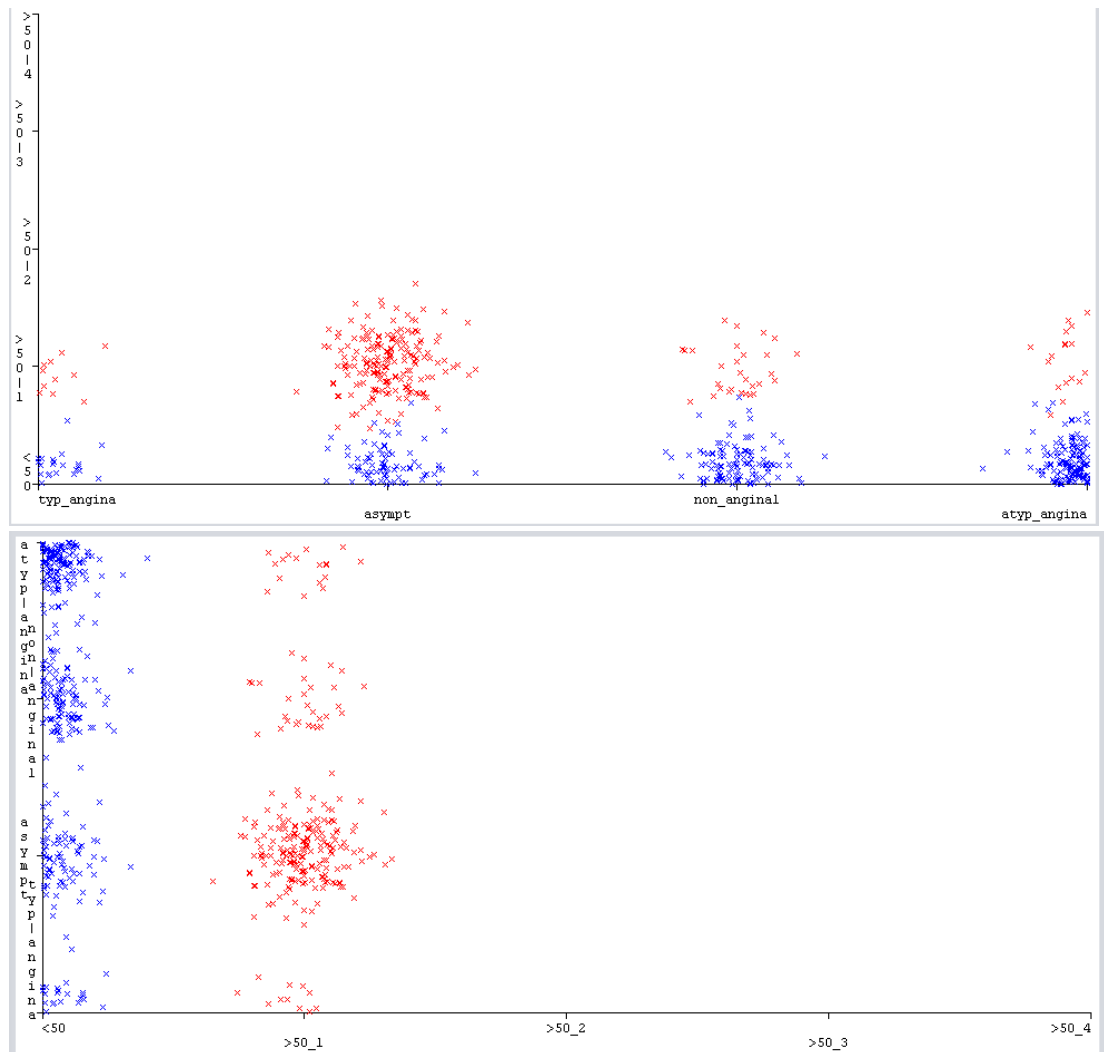


6. Nhận xét:

- Đồ thị chỉ có hai màu xanh và đỏ (vì 3 giá trị ‘>50_2’, ‘>50_3’, ‘>50_4’, đều có số record tương ứng là 0).
- Từ các đồ thị chỉ có hai màu ta có thể nói rằng từ hai khảo sát này số người không mắc bệnh tim và số người mắc bệnh tim cấp độ 1 chiếm ưu thế.
- Dựa vào đồ thị thuộc tính num ta thấy số record có num = ‘<50’ cao nhất, rồi đến num = ‘>50_1’

7. Tab Visualize.

- Thuật ngữ sử dụng trong textbook để đặt tên cho đồ thị là: **Plot Matrix**
- Các thuộc tính nào có vẻ như dẫn đến bệnh tim nhiều nhất:
Chọn thuộc tính cp.
- Thuộc tính có khả năng dự đoán bệnh tim tốt nhất (Y) như là một hàm của num(X).
Hình thứ nhất cp là trục hoành, num là trục tung; hình thứ hai num là trục hoành, cp là trục tung.



8. Những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau:

Ta có những cặp thuộc tính sau:

STT	Thuộc tính 1	Thuộc tính 2
1	cp	num
2	exang	num
3	cp	exang
4	fbs	num
5	fbs	thai
6	fbs	ca
7	fbs	chol

III. CHỌN LỌC DỮ LIỆU (SELECTION):

Các dataset sử dụng trong bài tập đã được xử lý bằng các chọn ra tập các thuộc tính liên quan đến mục tiêu khai thác dữ liệu.

1. Thuộc tính trong những dataset trước khi xử lý:

Các thuộc tính trước khi xử lý: 76 thuộc tính.

No.	Name and Description
1	id: patient identification number
2	ccf: social security number (I replaced this with a dummy value of 0)
3	age: age in years
4	sex: sex (1 = male; 0 = female)
5	painloc: chest pain location (1 = substernal; 0 = otherwise)
6	painexer (1 = provoked by exertion; 0 = otherwise)
7	relrest (1 = relieved after rest; 0 = otherwise)
8	pncaden (sum of 5, 6, and 7)
9	cp: chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
10	trestbps: resting blood pressure (in mm Hg on admission to the hospital)
11	Htn
12	chol: serum cholesterol in mg/dl
13	smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
14	cigs (cigarettes per day)
15	years (number of years as a smoker)
16	fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
17	dm (1 = history of diabetes; 0 = no such history)
18	famhist: family history of coronary artery disease (1 = yes; 0 = no)
19	restecg: resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
20	ekgmo (month of exercise ECG reading)
21	ekgday (day of exercise ECG reading)
22	ekgyr (year of exercise ECG reading)
23	dig (digitalis used during exercise ECG: 1 = yes; 0 = no)
24	prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25	nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
26	pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27	diuretic (diuretic used during exercise ECG: 1 = yes; 0 = no)

28	proto: exercise protocol 1 = Bruce 2 = Kottus 3 = McHenry 4 = fast Balke 5 = Balke 6 = Noughton 7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!) 8 = bike 125 kpa min/min 9 = bike 100 kpa min/min 10 = bike 75 kpa min/min 11 = bike 50 kpa min/min 12 = arm ergometer
29	thaldur: duration of exercise test in minutes
30	thaltme: time when ST measure depression was noted
31	met: mets achieved
32	thalach: maximum heart rate achieved
33	thalrest: resting heart rate
34	tpeakbps: peak exercise blood pressure (first of 2 parts)
35	tpeakbpd: peak exercise blood pressure (second of 2 parts)
36	Dummy
37	trestbpd: resting blood pressure
38	exang: exercise induced angina (1 = yes; 0 = no)
39	xhypo: (1 = yes; 0 = no)
40	oldpeak = ST depression induced by exercise relative to rest
41	slope: the slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
42	rldv5: height at rest
43	rldv5e: height at peak exercise
44	ca: number of major vessels (0-3) colored by flourosopy
45	restckm: irrelevant
46	exerckm: irrelevant
47	restef: rest raidonuclid (sp?) ejection fraction
48	restwm: rest wall (sp?) motion abnormality 0 = none 1 = mild or moderate 2 = moderate or severe 3 = akinesis or dyskmem (sp?)
49	exeref: exercise radinalid (sp?) ejection fraction
50	exerwm: exercise wall (sp?) motion
51	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

52	thalsev : not used
53	thalpul : not used
54	earlobe : not used
55	cmo : month of cardiac cath (sp?) (perhaps "call")
56	cday : day of cardiac cath (sp?)
57	cyr : year of cardiac cath (sp?)
58	num : diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)
59	Lmt
60	Ladprox
61	Laddist
62	Diag
63	Cxmain
64	Ramus
65	om1
66	om2
67	Rcaprox
68	Rcadist
69	lvx1 : not used
70	lvx2 : not used
71	lvx3 : not used
72	lvx4 : not used
73	lvf : not used
74	cathef : not used
75	junk : not used
76	name : last name of patient

2. *Tab Select attributes và những lựa chọn khác nhau của Weka để chọn lọc thuộc tính:* [7]

- **Phương pháp Best First:** Dựa trên cơ sở tìm kiếm không gian của các tập hợp của các thuộc tính bằng cách tăng cường leo đồi tham lam dựa trên cơ sở quay lui.

Đặt số lượng nút không cải thiện liên tiếp để kiểm soát các bước độ quay lui được thực hiện. Best First có thể bắt đầu với bộ thuộc tính trống và tìm kiếm chuyển tiếp hoặc bắt đầu với các record thuộc tính đầy đủ và tìm kiếm ngược hoặc bắt đầu tại bất kỳ điểm nào và tìm kiếm theo cả hai hướng (bằng cách xem xét tất cả các bổ sung và xóa thuộc tính tại điểm đã cho).

- **Phương pháp Greedy Step Wise:** Thực hiện tìm kiếm tiến hoặc lùi tham lam trong không gian của các tập hợp thuộc tính

Có thể bắt đầu bằng từ 0 hoặc tất cả các thuộc tính hoặc từ một điểm tùy ý trong không gian. Dừng khi có hành động bổ sung hoặc xóa bất kỳ thuộc tính còn lại dẫn đến sự suy hệt trong tính toán. Cũng có thể tạo danh sách các thuộc tính xếp hạng bằng cách di chuyển không gian từ bên này sang bên kia và lưu lại thứ tự các thuộc tính được chọn.

- **Phương pháp Ranker:** Xếp loại những đánh giá riêng lẻ (ReliefF, Entropy, GainRatio, ...). Ta sẽ giao lại các đánh giá với nhau.

3. So sánh với các phương pháp chọn lọc dữ liệu trong textbook và phương pháp không có trong Weka:

- **Best First:** Tìm kiếm ưu tiên tối ưu sẽ kết hợp 2 phương pháp trên cho phép ta đi theo một con đường duy nhất tại một thời điểm, nhưng đồng thời vẫn "quan sát" được những hướng khác. Nếu con đường đang đi "có vẻ" không triển vọng bằng những con đường ta đang "quan sát" ta sẽ chuyển sang đi theo một trong số các con đường này. Để tiện lợi ta sẽ dùng chữ viết tắt BFS thay cho tên gọi tìm kiếm ưu tiên tối ưu. ^[9]

- **Greedy Step Wise:**

Tiến:^[8]

Tập thuộc tính được khởi tạo: {A1, A2, A3, A4, A5, A6}

Tập rút gọn: {} => {A1} => {A1, A4} => Rút gọn: {A1, A4, A6}

Lùi:^[8]

Tập thuộc tính khởi tạo:

{A1, A2, A3, A4, A5, A6} => {A1, A3, A4, A5, A6} => {A1, A4, A5, A6}

=> Tập rút gọn: {A1, A4, A6}

- **Ranker:** Ta sẽ tập trung theo công thức tính độ đo entropy:^[10]

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

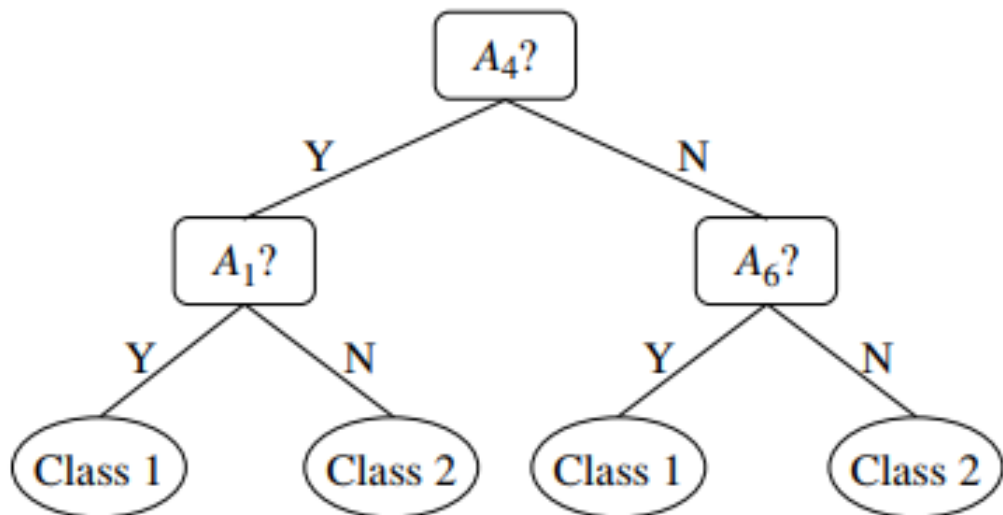
$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D)$$

$$Gain(A) = Info(D) - Info_A(D)$$

- **Decision tree** (không có trong weka):^[8]

Tập thuộc tính khởi tạo:

{A1, A2, A3, A4, A5, A6}



=> Tập rút gọn:
 {A1, A4, A6}

IV. LÀM SẠCH DỮ LIỆU (CLEANING):

Xử lý các dữ liệu thiếu, nhiễu, và mâu thuẫn. Sử dụng các bộ lọc trong Weka để làm sạch dữ liệu.

1. Các giá trị thiếu (Missing values):

- Các phương pháp xử lý dữ liệu bị thiếu:^[11]
 - Bỏ qua các record có giá trị thuộc tính bị thiếu.
 - Điền các giá trị thiếu bằng tay:
 - Điền các giá trị tự động:
 - Thay thế bằng hàm số chung.
 - Thay thế bằng giá trị trung bình của thuộc tính trong một lớp
 - Thay thế bằng giá trị có nhiều khả năng nhất: Suy ra từ công thức Bayesian, Decision Tree hoặc thuật giải EM (Expectation Maximization).
- Weka có những phương pháp sau:
 - Thay thế bằng giá trị trung bình:

Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data

The class attribute is skipped by default.

CAPABILITIES
 Class -- Binary class, Date class, Empty nominal class, Missing class values, No class, Nominal class, Numeric class, Relational class, String class, Unary class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

Interfaces -- Sourceable, UnsupervisedFilter, WeightedAttributesHandler, WeightedInstancesHandler

Additional
 Minimum number of instances: 0

- Thay thế bằng giá trị người dung nhập:

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

CAPABILITIES
 Class -- Binary class, Date class, Empty nominal class, Missing class values, No class, Nominal class, Numeric class, Relational class, String class, Unary class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

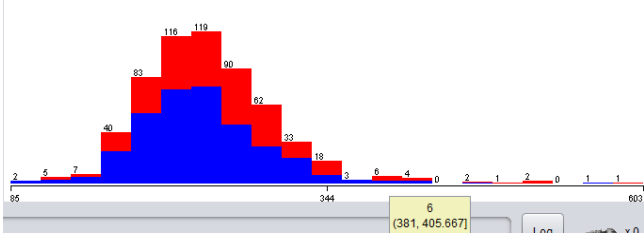
Interfaces -- StreamableFilter, UnsupervisedFilter, WeightedAttributesHandler, WeightedInstancesHandler

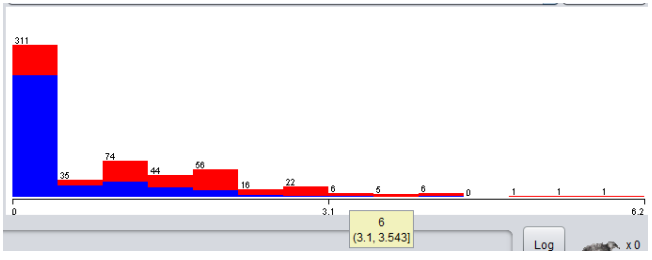
Additional
 Minimum number of instances: 0

- Chọn phương pháp: Thay thế bằng giá trị trung bình.
 Lý do: Vì khi ta thay thế bằng giá trị trung bình thì lúc này trung bình của cột không bị thay đổi.
- Cài đặt phương pháp khác không có trong weka (Cây quyết định):

2. Dữ liệu nhiễu (Noisy data):

- Các phương pháp khử nhiễu như sau:
 - Phương pháp chia giỏ (binning):
 - Sắp xếp và chia dữ liệu vào các giỏ có cùng độ sâu (equal-depth)
 - Khử nhiễu bằng giá trị trung bình, trung tuyến, biên giỏ, ...
 - Gom nhóm (Clustering):
 - Phát hiện và loại bỏ các khác biệt
 - Phương pháp hồi quy (Regression):
 - Đưa dữ liệu vào hàm hồi quy.
 - Kết hợp sự kiểm tra giữa máy tính và con người (Computer/human inspection):
 - Phát hiện giá trị nghi ngờ và kiểm tra bởi con người.
- Các phương pháp weka đã cài đặt:
 - Thực hiện việc xóa dữ liệu gây nhiễu một cách bán tự động.
 - Xem từng thuộc tính trong biểu đồ, nếu xuất hiện giá trị thuộc tính nào có số lượng record chỉ đạt từ 1 đến 4% (khoảng này do em tự thiết kế sau khi xem kĩ dữ liệu) và cộng thêm với điều kiện là dữ liệu lại có yes có no lẫn lộn nhau (tức là độ phân hóa khá cao) thì sẽ tiến hành xóa những record này ra khỏi dataset.
 - Cụm thể ở đây em xét được như sau:

Chọn thuộc tính để khử nhiễu	Phân tích yếu tố nhiễu
Thuộc tính: chol Ta có biểu đồ như sau: 	Nhìn vào biểu đồ ta thấy rằng: Từ giá trị 381 trở về sau có vằn đề khá rõ. Thứ nhất dữ liệu phân bố khá thưa: 17/595 records. Thứ hai, thuộc tính num lại phân bố một cách xen kẽ là rời rạc không tập trung => các thuộc tính phân lớp sau

	này sẽ khó khăn trong việc phân lớp ở đây. Cách thức khắc phục: Xóa những dòng record này.
<p>Ta xét thuộc tính: oldpeak.</p> 	<p>Thứ nhất, ta nhận thấy giá trị thuộc tính từ 3.1 trở về sau thì số lượng record chỉ có 20/578 records.</p> <p>Thứ hai cái record lại bị phân hóa, không đồng nhất với nhau về mặt kết quả.</p>

- Kết quả còn lại sau khi khử nhiễu số record còn lại là: 558 records.

3. Dò tìm dữ liệu tạp (Outlier detection):^[12]

Phương pháp dò tìm dữ liệu tạp đã học:

– Phương pháp Numeric Outlier:

- Đây là phương pháp đơn giản nhất, là phương pháp phát hiện outlier không cần tham số. Dựa trên khoảng cách giữa các tứ phân vị
- Ta có phân vị thứ 1 và thứ 3 được tính toán. Một điểm dữ liệu tạp (outlier) là điểm nằm ngoài IQR (Interquartile Range).
- Gọi x_i là điểm outlier, ta có:

$$x_i > Q3 + k(IQR) \vee x_i < Q1 - k(IQR)$$

Tại:

$$IQR = Q3 - Q1, k \geq 0$$

– Phương pháp Z-Score:

- Là phương pháp có tham số trong một hoặc nhiều không gian đặc trưng.
- Dựa vào phân phối đều (phân phối Gaussian)
- Điểm dữ liệu tạp (điểm ngoại lai) là điểm cách xa giá trị trung bình. Khoảng cách từ điểm x_i tùy thuộc vào ngưỡng, được tính theo công thức:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Tại:

- x_i : là điểm ngoại lai
- μ, σ : lần lượt là kỳ vọng và độ lệch chuẩn.
- Ta có điều kiện: $|z_i| > z_{thr}$, z_{thr} là ngưỡng và z_i phụ thuộc vào điểm đặt của z_{thr} .

– Phương pháp DBSCAN:

- Phương pháp này dựa trên phương pháp gom nhóm DBSCAN. DBSCAN là phương pháp phát hiện điểm ngoại lai dựa trên mật độ, không tham số trong không gian đặc trưng một hoặc nhiều chiều.
- Trong phương pháp DBSCAN, tất cả các điểm dữ liệu được xác định là Điểm lõi, Điểm biên hoặc Điểm nhiễu.

- Điểm cốt lõi là các điểm dữ liệu có ít nhất các điểm dữ liệu lân cận MinPts trong khoảng cách.
- Điểm biên là điểm nằm kế Điểm lõi trong khoảng cách nhưng có khoảng cách ϵ nhỏ hơn MinPts trong khoảng cách ϵ .
- Tất cả các điểm dữ liệu khác là Điểm nhiễu, cũng được xác định là ngoại lai (hay điểm dữ liệu tạp).
- Do đó việc phát hiện điểm ngoại lai phụ thuộc vào số lượng MinPts, khoảng cách ϵ và độ đo khoảng cách (Vd: Euclidean hoặc Mahattan)
- **Phương pháp Isolation forest:**
 - Đây là một phương pháp không tham số cho các bộ dữ liệu lớn trong không gian đặc trưng một hoặc nhiều chiều.
 - Một khái niệm quan trọng trong phương pháp này là số cách ly.
 - Số cách ly là số lượng phân tách cần thiết để cô lập một điểm dữ liệu. Số lượng phân chia này được xác định bằng cách làm theo các bước sau:
 - Một điểm “a” cô lập được chọn ngẫu nhiên
 - Một điểm dữ liệu ngẫu nhiên “b” được chọn ở khoảng giữa min và max và phải khác điểm “a”.
 - Nếu giá trị “b” thấp hơn giá trị của “a”, thì lập “b” là giá trị chặn dưới.
 - Nếu giá trị “b” cao hơn giá trị của “a”, thì lập “b” là giá trị chặn trên.
 - Các bước này thực hiện lặp đi lặp lại miễn là có điểm dữ liệu không phải là một điểm khác nhau giữa chặn trên và chặn dưới.
 - Ta phải đòi hỏi ít sự phân tách hơn để cô lập được điểm ngoại lai hay điểm ngoại lai có số cách ly nhỏ hơn so với ngưỡng.
 - Ngưỡng được định nghĩa dựa trên sự ước lượng theo tỉ số phần trăm của điểm ngoại lai trong dữ liệu cũng chính là điểm bắt đầu của thuật toán này.
- Dò tìm dữ liệu tạp bằng weka:
 - Ta sẽ dùng chức năng InterquartileRange: Do phương pháp này dễ cài đặt, thao tác tính toán nhanh, đồng thời đây là phương pháp không cần tính toán bằng tham số truyền vào.
 - Mô tả như sau (Dựa trên tứ phân vị):

A filter for detecting outliers and extreme values based on interquartile ranges

The filter skips the class attribute.

Outliers:
 $Q3 + OF * IQR \leq x < Q3 + EVF * IQR$
or
 $Q1 - EVF * IQR \leq x < Q1 - OF * IQR$

Extreme values:
 $x > Q3 + EVF * IQR$
or
 $x < Q1 - EVF * IQR$

Key:
Q1 = 25% quartile
Q3 = 75% quartile
IQR = Interquartile Range, difference between Q1 and Q3
OF = Outlier Factor
EVF = Extreme Value Factor

CAPABILITIES
Class -- Binary class, Date class, Empty nominal class, Missing class values, No class, Nominal class, Numeric class, Relational class, String class, Unary class

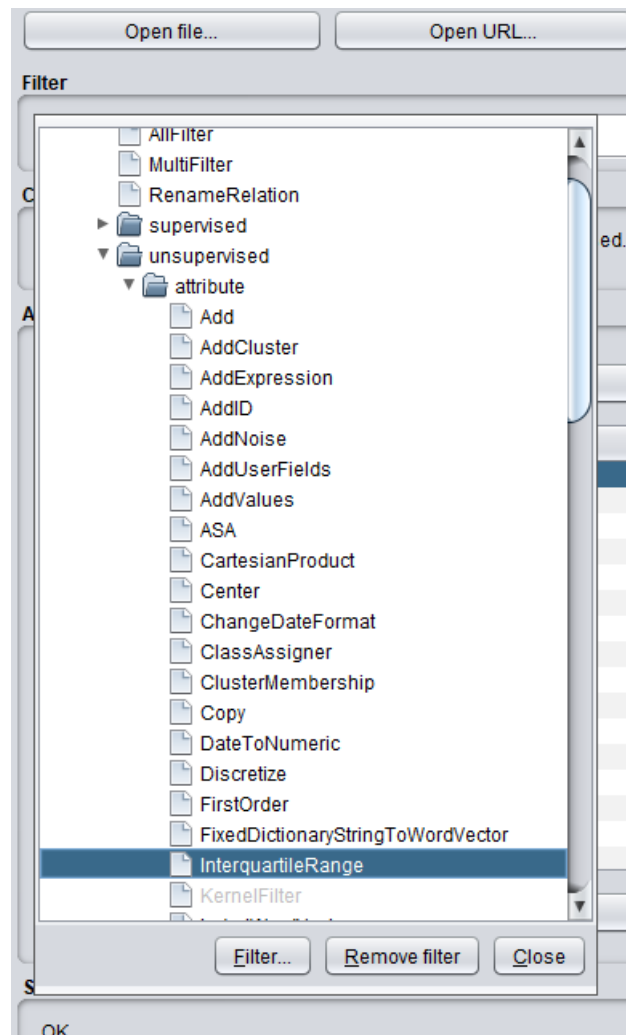
Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

Interfaces -- WeightedAttributesHandler

Additional
Minimum number of instances: 0

- Có dữ liệu tạp trong dataset này không? → Có tồn tại dữ liệu tạp trong dataset.

- Các thao tác được tiến hành xác định như sau:
 - Trong tab Preprocessing, chọn Choose, tiếp đến chọn Filter, chọn tiếp unsupervised, chọn tiếp Attributes, chọn tiếp InterquartileRange, rồi Apply.



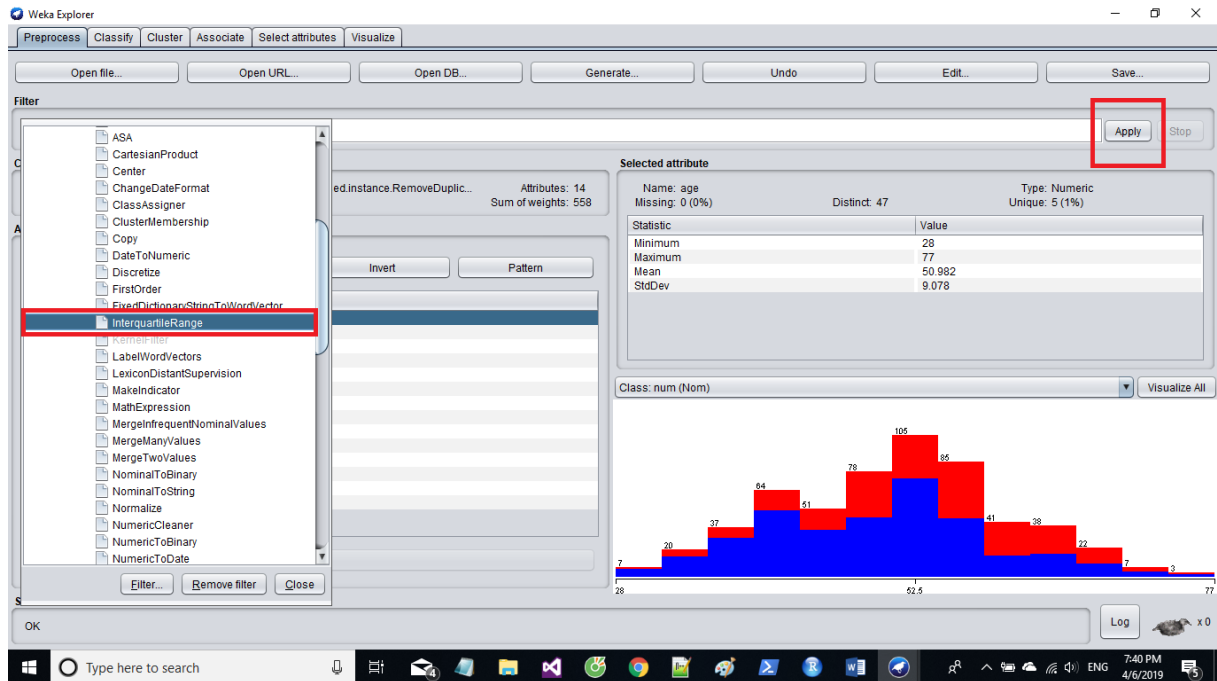
- Liệt kê một số dữ liệu tập:

1	53	male	non_anginal	130	246	t	left_vent_hyper	173	no	0	up	3	normal	<50
2	66	male	atyp_angina	160	246	f	normal	120	yes	0	flat	3	fixed_defect	>50_1
3	77	male	asympt	125	304	f	left_vent_hyper	162	yes	0	up	3	normal	>50_1
4	45	male	asympt	142	309	f	left_vent_hyper	147	yes	0	flat	3	reversable_defect	>50_1
5	52	male	asympt	108	233	t	normal	147	no	0.1	up	3	reversable_defect	<50
6	49	male	non_anginal	118	149	f	left_vent_hyper	126	no	0.8	up	3	normal	>50_1
7	65	female	asympt	150	225	f	left_vent_hyper	114	no	1	flat	3	reversable_defect	>50_1
8	57	male	asympt	165	289	t	left_vent_hyper	124	no	1	flat	3	reversable_defect	>50_1
9	67	male	asympt	160	286	f	left_vent_hyper	108	yes	1.5	flat	3	normal	>50_1
10	62	male	non_anginal	130	231	f	normal	146	no	1.8	flat	3	reversable_defect	<50
11	63	male	asympt	130	330	t	left_vent_hyper	132	yes	1.8	up	3	reversable_defect	>50_1
12	62	female	asympt	138	294	t	normal	106	no	1.9	flat	3	normal	>50_1
13	49	male	non_anginal	120	188	f	normal	139	no	2	flat	3	reversable_defect	>50_1
14	69	male	non_anginal	140	254	f	left_vent_hyper	146	no	2	flat	3	reversable_defect	>50_1
15	58	male	asympt	128	216	f	left_vent_hyper	131	yes	2.2	flat	3	reversable_defect	>50_1
16	70	male	asympt	130	322	f	left_vent_hyper	109	no	2.4	flat	3	normal	>50_1

****Lưu ý: khi Apply InterquartileRange xong thì ta được 16 cột. Ở cột Outlier ta sắp xếp lại, rồi xóa hết những record có Outlier = yes. “Attribute as class”***

thuộc tính num. Sau đó sẽ thực hiện thao tác xóa hết hai cột ExtremeValue và Outlier. Cụ thể như sau:

Bước 1: Chọn InterquartileRange → Chọn Apply:



Bước 2: Mở edit, xem dữ liệu → Sắp xếp lại cột Outlier:

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised...

No.	1: age	2: sex	3: cp	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num	15: Outlier	16: ExtremeValue
...	60.0	male	asym...	125.0	258.0	f	normal	141.0	yes	2.8	flat	1.0	reve...	50_1	no	no
...	60.0	male	asym...	145.0	282.0	f	normal	142.0	yes	2.8	flat	2.0	reve...	50_1	no	no
...	70.0	male	non...	160.0	269.0	f	normal	112.0	yes	2.9	flat	1.0	reve...	50_1	no	no
...	60.0	male	non...	140.0	185.0	f	normal	155.0	no	3.0	flat	0.0	nor...	50_1	no	no
...	48.0	male	asym...	160.0	193.0	f	normal	102.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	55.0	male	asym...	140.0	201.0	f	normal	130.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	45.0	male	asym...	104.0	208.0	f	normal	148.0	yes	3.0	flat	0.0	nor...	50_1	no	no
...	57.0	male	asym...	150.0	255.0	f	normal	92.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	58.0	male	asym...	128.0	259.0	f	normal	130.0	yes	3.0	flat	2.0	reve...	50_1	no	no
...	53.0	male	asym...	124.0	260.0	f	normal	112.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	36.0	male	atyp...	120.0	267.0	f	normal	160.0	no	3.0	flat	0.66...	nor...	50_1	no	no
...	44.0	male	atyp...	150.0	288.0	f	normal	150.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	47.0	male	asym...	160.0	291.0	f	normal	158.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	57.0	male	asym...	110.0	335.0	f	normal	143.0	yes	3.0	flat	1.0	reve...	50_1	no	no
...	41.0	male	asym...	120.0	336.0	f	normal	118.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	43.0	fem...	asym...	132.0	341.0	t	normal	136.0	yes	3.0	flat	0.0	reve...	50_1	no	no
...	56.0	male	asym...	155.0	342.0	t	normal	150.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	53.0	male	asym...	140.0	203.0	t	normal	155.0	yes	3.1	down	0.0	reve...	50_1	no	no
...	53.0	male	non...	130.0	246.0	t	normal	173.0	no	0.0	up	3.0	nor...	50_1	yes	no
...	66.0	male	atyp...	160.0	246.0	f	normal	120.0	yes	0.0	flat	3.0	fixed...	50_1	yes	no
...	77.0	male	asym...	125.0	304.0	f	normal	162.0	yes	0.0	up	3.0	nor...	50_1	yes	no
...	45.0	male	asym...	142.0	309.0	f	normal	147.0	yes	0.0	flat	3.0	reve...	50_1	yes	no
...	52.0	male	asym...	108.0	233.0	t	normal	147.0	no	0.1	up	3.0	reve...	50_1	yes	no
...	49.0	male	non...	118.0	149.0	f	normal	126.0	no	0.8	up	3.0	nor...	50_1	yes	no
...	65.0	fem...	asym...	150.0	225.0	f	normal	114.0	no	1.0	flat	3.0	reve...	50_1	yes	no
...	57.0	male	asym...	165.0	289.0	t	normal	124.0	no	1.0	flat	3.0	reve...	50_1	yes	no
...	67.0	male	asym...	160.0	286.0	f	normal	108.0	yes	1.5	flat	3.0	nor...	50_1	yes	no
...	62.0	male	non...	130.0	231.0	f	normal	146.0	no	1.8	flat	3.0	reve...	50_1	yes	no
...	63.0	male	asym...	130.0	330.0	t	normal	132.0	yes	1.8	up	3.0	reve...	50_1	yes	no
...	62.0	fem...	asym...	138.0	294.0	t	normal	106.0	no	1.9	flat	3.0	nor...	50_1	yes	no
...	49.0	male	non...	120.0	188.0	f	normal	139.0	no	2.0	flat	3.0	reve...	50_1	yes	no
...	69.0	male	non...	140.0	254.0	f	normal	146.0	no	2.0	flat	3.0	reve...	50_1	yes	no
...	58.0	male	asym...	128.0	216.0	f	normal	131.0	yes	2.2	flat	3.0	reve...	50_1	yes	no
...	70.0	male	asym...	130.0	322.0	f	normal	109.0	no	2.4	flat	3.0	nor...	50_1	yes	no

Bước 3: Xóa toàn bộ phần record có Outlier = yes:

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised...

No.	1: age	2: sex	3: cp	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num	15: Outlier	16: ExtremeValue
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal
...	60.0	male	asy...	125.0	258.0	f	left_ve...	141.0	yes	2.8	flat	1.0	reve...	50_1	no	no
...	60.0	male	asy...	145.0	282.0	f	left_ve...	142.0	yes	2.8	flat	2.0	reve...	50_1	no	no
...	70.0	male	non...	160.0	269.0	f	normal	112.0	yes	2.9	flat	1.0	reve...	50_1	no	no
...	60.0	male	non...	140.0	185.0	f	left_ve...	155.0	no	3.0	flat	0.0	nor...	50_1	no	no
...	48.0	male	asy...	160.0	193.0	f	normal	102.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	55.0	male	asy...	140.0	201.0	f	normal	130.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	45.0	male	asy...	104.0	208.0	f	left_ve...	148.0	yes	3.0	flat	0.0	nor...	50_1	no	no
...	57.0	male	asy...	150.0	255.0	f	normal	92.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	58.0	male	asy...	128.0	259.0	f	left_ve...	130.0	yes	3.0	flat	2.0	reve...	50_1	no	no
...	53.0	male	asy...	124.0	260.0	f	st_t_w...	112.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	36.0	male	atyp...	120.0	267.0	f	normal	160.0	no	3.0	flat	0.66...	nor...	50_1	no	no
...	44.0	male	atyp...	150.0	288.0	f	normal	150.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	47.0	male	asy...	160.0	291.0	f	st_t_w...	158.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	57.0	male	asy...	110.0	335.0	f	normal	143.0	yes	3.0	flat	1.0	reve...	50_1	no	no
...	41.0	male	asy...	120.0	336.0	f	normal	118.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	43.0	fem...	asy...	132.0	341.0	t	left_ve...	136.0	yes	3.0	flat	0.0	reve...	50_1	no	no
...	56.0	male	asy...	155.0	342.0	t	normal	150.0	yes	3.0	flat	0.66...	nor...	50_1	no	no
...	53.0	male	asy...	140.0	203.0	t	left_ve...	155.0	yes	3.1	down	0.0	reve...	50_1	no	no
...	53.0	male	non...	130.0	246.0	t	left_ve...	173.0	no	50_1	yes	no
...	66.0	male	atyp...	160.0	246.0	f	normal	120.0	yes	50_1	yes	no
...	77.0	male	asy...	125.0	304.0	f	left_ve...	162.0	yes	50_1	yes	no
...	45.0	male	asy...	142.0	309.0	f	left_ve...	147.0	yes	50_1	yes	no
...	52.0	male	asy...	108.0	233.0	t	normal	147.0	no	50_1	yes	no
...	49.0	male	non...	118.0	149.0	f	left_ve...	126.0	no	50_1	yes	no
...	65.0	fem...	asy...	150.0	225.0	f	left_ve...	114.0	no	50_1	yes	no
...	57.0	male	asy...	165.0	289.0	t	left_ve...	124.0	no	50_1	yes	no
...	67.0	male	asy...	160.0	286.0	f	left_ve...	108.0	yes	50_1	yes	no
...	62.0	male	non...	130.0	231.0	f	normal	146.0	no	50_1	yes	no
...	63.0	male	asy...	130.0	330.0	t	left_ve...	132.0	yes	50_1	yes	no
...	62.0	fem...	asy...	138.0	284.0	t	normal	106.0	no	50_1	yes	no
...	49.0	male	non...	120.0	188.0	f	normal	139.0	no	50_1	yes	no
...	69.0	male	non...	140.0	254.0	f	left_ve...	146.0	no	50_1	yes	no
...	58.0	male	asy...	128.0	216.0	f	left_ve...	131.0	yes	50_1	yes	no
...	70.0	male	asy...	130.0	322.0	f	left_ve...	109.0	no	50_1	yes	no

Undo
Copy
Search...
Clear search
Delete selected instance
Delete ALL selected instances
Insert new instance
Set instance weight

Add instance Undo OK Cancel

Bước 4: Set cột num thành cột phân lớp, Xóa cột Outlier và ExtremeValue:

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised...

No.	1: age	2: sex	3: cp	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num	15: Outlier	16: ExtremeValue
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal
1	56.0	male	asy...	120.0	85.0	f	normal	140.0	no	0.0	flat	0.66...	nor...	(50	no	no
2	52.0	male	atyp...	140.0	100.0	f	normal	138.0	yes	0.0	flat	0.66...	nor...	(50	no	no
3	50.0	male	asy...	140.0	129.0	f	normal	135.0	no	0.0	flat	0.66...	nor...	(50	no	no
4	28.0	male	atyp...	130.0	132.0	f	left_ve...	185.0	no	0.0	flat	0.66...	nor...	(50	no	no
5	39.0	male	non...	160.0	147.0	t	normal	160.0	no	0.0	flat	0.66...	nor...	(50	no	no
6	42.0	male	non...	160.0	147.0	f	normal	146.0	no	0.0	flat	0.66...	nor...	(50	no	no
7	41.0	male	atyp...	120.0	157.0	f	normal	182.0	no	0.0	up	0.0	nor...	50	no	no
8	45.0	fem...	atyp...	112.0	160.0	f	normal	138.0	no	0.0	flat	0.0	nor...	(50	no	no
9	35.0	fem...	typ...	120.0	160.0	f	st_t_w...	185.0	no	0.0	flat	0.66...	nor...	(50	no	no
10	36.0	male	non...	150.0	160.0	f	normal	172.0	no	0.0	flat	0.66...	nor...	(50	no	no
11	34.0	fem...	atyp...	130.0	161.0	f	normal	190.0	no	0.0	flat	0.66...	nor...	(50	no	no
12	46.0	male	non...	150.0	163.0	f	normal	116.0	no	0.0	flat	0.66...	nor...	(50	no	no
13	36.0	male	atyp...	120.0	166.0	f	normal	180.0	no	0.0	flat	0.66...	nor...	(50	no	no
14	35.0	fem...	asy...	140.0	167.0	f	normal	150.0	no	0.0	flat	0.66...	nor...	(50	no	no
15	50.0	male	atyp...	120.0	168.0	f	normal	160.0	no	0.0	flat	0.0	nor...	(50	no	no
16	37.0	fem...	asy...	130.0	173.0	f	st_t_w...	184.0	no	0.0	flat	0.66...	nor...	(50	no	no
17	38.0	male	non...	138.0	175.0	f	normal	173.0	no	0.0	up	0.66...	nor...	50_1	no	no
18	60.0	fem...	non...	120.0	178.0	t	normal	96.0	no	0.0	up	0.0	nor...	(50	no	no
19	51.0	male	asy...	130.0	179.0	f	normal	100.0	no	0.0	flat	0.66...	reve...	(50	no	no
20	58.0	male	non...	140.0	179.0	f	normal	160.0	no	0.0	flat	0.66...	nor...	50_1	no	no
21	42.0	male	non...	130.0	180.0	f	normal	150.0	no	0.0	up	0.0	nor...	(50	no	no
22	34.0	male	typ...	118.0	182.0	f	left_ve...	174.0	no	0.0	up	0.0	nor...	(50	no	no
23	39.0	fem...	non...	110.0	182.0	f	st_t_w...	180.0	no	0.0	flat	0.66...	nor...	(50	no	no
24	53.0	male	asy...	130.0	182.0	f	normal	148.0	no	0.0	flat	0.66...	nor...	(50	no	no
25	41.0	fem...	atyp...	125.0	184.0	f	normal	180.0	no	0.0	flat	0.66...	nor...	(50	no	no
26	56.0	male	atyp...	130.0	184.0	f	normal	100.0	no	0.0	flat	0.66...	nor...	(50	no	no
27	52.0	male	typ...	118.0	186.0	f	left_ve...	190.0	no	0.0	flat	0.0	fixed...	(50	no	no
28	43.0	fem...	atyp...	150.0	186.0	f	normal	154.0	no	0.0	flat	0.66...	nor...	50_1	no	no
29	49.0	male	non...	140.0	187.0	f	normal	172.0	no	0.0	flat	0.66...	nor...	(50	no	no

Get mean...
Set all values to...
Set missing values to...
Replace values with...
Rename attribute...
Set attribute weight...
Attribute as class
Delete attribute
Delete attributes...
Sort data (ascending)
Optimal column width (current)
Optimal column width (all)

Add instance Undo OK Cancel

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised...

No.	1: age	2: sex	3: cp	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: Outlier	15: ExtremeValue	16: num
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	56.0	male	asy...	120.0	85.0	f	normal	140.0	no	0.0	flat	0.66...	nor...	no	no	(50
2	52.0	male	atyp...	140.0	100.0	f	normal	138.0	yes	0.0	flat	0.66...	nor...	no	no	(50
3	50.0	male	asy...	140.0	129.0	f	normal	135.0	no	0.0	flat	0.66...	nor...	no	no	(50
4	28.0	male	atyp...	130.0	132.0	f	left_ve...	185.0	no	0.0	flat	0.66...	nor...	no	no	(50
5	39.0	male	non...	160.0	147.0	t	normal	160.0	no	0.0	flat	0.66...	nor...	no	no	(50
6	42.0	male	non...	160.0	147.0	f	normal	146.0	no	0.0	flat	0.66...	nor...	no	no	(50
7	41.0	male	atyp...	120.0	157.0	f	normal	182.0	no	0.0	flat	0.66...	nor...	no	no	(50
8	45.0	fem...	atyp...	112.0	160.0	f	normal	138.0	no	0.0	flat	0.66...	nor...	no	no	(50
9	35.0	fem...	typ...	120.0	160.0	f	st_t_w...	185.0	no	0.0	flat	0.66...	nor...	no	no	(50
10	36.0	male	non...	150.0	160.0	f	normal	172.0	no	0.0	flat	0.66...	nor...	no	no	(50
11	34.0	fem...	atyp...	130.0	161.0	f	normal	190.0	no	0.0	flat	0.66...	nor...	no	no	(50
12	46.0	male	non...	150.0	163.0	f	normal	116.0	no	0.0	flat	0.66...	nor...	no	no	(50
13	36.0	male	atyp...	120.0	166.0	f	normal	180.0	no	0.0	flat	0.66...	nor...	no	no	(50
14	35.0	fem...	asy...	140.0	167.0	f	normal	150.0	no	0.0	flat	0.66...	nor...	no	no	(50
15	50.0	male	atyp...	120.0	168.0	f	normal	160.0	no	0.0	flat	0.66...	nor...	no	no	(50
16	37.0	fem...	asy...	130.0	173.0	f	st_t_w...	184.0	no	0.0	flat	0.66...	nor...	no	no	(50
17	38.0	male	non...	138.0	175.0	f	normal	173.0	no	0.0	flat	0.66...	nor...	no	no	(50
18	60.0	fem...	non...	120.0	178.0	t	normal	96.0	no	0.0	up	0.0	fixed...	no	no	(50
19	51.0	male	asy...	130.0	179.0	f	normal	100.0	no	0.0	flat	0.66...	nor...	no	no	(50
20	58.0	male	non...	140.0	179.0	f	normal	160.0	no	0.0	flat	0.66...	nor...	no	no	(50
21	42.0	male	non...	130.0	180.0	f	normal	150.0	no	0.0	flat	0.66...	nor...	no	no	(50
22	34.0	male	typ...	118.0	182.0	f	left_ve...	174.0	no	0.0	up	0.0	normal	no	no	(50
23	39.0	fem...	non...	110.0	182.0	f	st_t_w...	180.0	no	0.0	flat	0.66...	nor...	no	no	(50
24	53.0	male	asy...	130.0	182.0	f	normal	148.0	no	0.0	flat	0.66...	nor...	no	no	(50
25	41.0	fem...	atyp...	125.0	184.0	f	normal	180.0	no	0.0	flat	0.66...	nor...	no	no	(50
26	56.0	male	atyp...	130.0	184.0	f	normal	100.0	no	0.0	flat	0.66...	nor...	no	no	(50
27	52.0	male	typ...	118.0	186.0	f	left_ve...	190.0	no	0.0	flat	0.66...	nor...	no	no	(50
28	43.0	fem...	atyp...	150.0	186.0	f	normal	154.0	no	0.0	flat	0.66...	nor...	no	no	(50
29	49.0	male	non...	140.0	187.0	f	normal	172.0	no	0.0	flat	0.66...	nor...	no	no	(50
30	54.0	male	asy...	165.0	188.0	f	normal	145.0	no	0.0	flat	0.66...	nor...	no	no	(50

Select items

ExtremeValue

Outlier

age

ca

chol

cp

exang

fbs

Select

Pattern

Cancel

Add instance

Undo

OK

Cancel

- Kết quả sau thao tác này sẽ nhận được 542 dòng record.

4. File heart-cleaned.arff:

- Lưu với tên: heart-cleaned.arff
- Sau khi lưu, ta kiểm tra lại bằng cách vào mục Edit để xem toàn bộ record.

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.InterquartileRar

No.	1: age	2: sex	3: cp	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal
1	56.0	male	asympt	120.0	85.0	f	normal	140.0	no	0.0	flat	0.667774	normal	(50
2	52.0	male	atyp_angina	140.0	100.0	f	normal	138.0	yes	0.0	flat	0.667774	normal	(50
3	50.0	male	asympt	140.0	129.0	f	normal	135.0	no	0.0	flat	0.667774	normal	(50
4	28.0	male	atyp_angina	130.0	132.0	f	left_vent_hyper	185.0	no	0.0	flat	0.667774	normal	(50
5	39.0	male	non_anginal	160.0	147.0	t	normal	160.0	no	0.0	flat	0.667774	normal	(50
6	42.0	male	non_anginal	160.0	147.0	f	normal	146.0	no	0.0	flat	0.667774	normal	(50
7	41.0	male	atyp_angina	120.0	157.0	f	normal	182.0	no	0.0	up	0.0	normal	(50
8	45.0	female	atyp_angina	112.0	160.0	f	normal	138.0	no	0.0	flat	0.667774	normal	(50
9	35.0	female	typ_angina	120.0	160.0	f	st_t_wave_abnormality	185.0	no	0.0	flat	0.667774	normal	(50
10	36.0	male	non_anginal	150.0	160.0	f	normal	172.0	no	0.0	flat	0.667774	normal	(50
11	34.0	female	atyp_angina	130.0	161.0	f	normal	190.0	no	0.0	flat	0.667774	normal	(50
12	46.0	male	non_anginal	150.0	163.0	f	normal	116.0	no	0.0	flat	0.667774	normal	(50
13	36.0	male	atyp_angina	120.0	166.0	f	normal	180.0	no	0.0	flat	0.667774	normal	(50
14	35.0	female	asympt	140.0	167.0	f	normal	150.0	no	0.0	flat	0.667774	normal	(50
15	50.0	male	atyp_angina	120.0	168.0	f	normal	160.0	no	0.0	flat	0.0	normal	(50
16	37.0	female	asympt	130.0	173.0	f	st_t_wave_abnormality	184.0	no	0.0	flat	0.667774	normal	(50
17	38.0	male	non_anginal	138.0	175.0	f	normal	173.0	no	0.0	up	0.667774	normal	(50
18	60.0	female	non_anginal	120.0	178.0	t	normal	96.0	no	0.0	up	0.0	normal	(50
19	51.0	male	asympt	130.0	179.0	f	normal	100.0	no	0.0	flat	0.667774	reversible_defect	(50
20	58.0	male	non_anginal	140.0	179.0	f	normal	160.0	no	0.0	flat	0.667774	normal	(50
21	42.0	male	non_anginal	130.0	180.0	f	normal	150.0	no	0.0	up	0.0	normal	(50
22	34.0	male	typ_angina	118.0	182.0	f	left_vent_hyper	174.0	no	0.0	up	0.0	normal	(50
23	39.0	female	non_anginal	110.0	182.0	f	st_t_wave_abnormality	180.0	no	0.0	flat	0.667774	normal	(50
24	53.0	male	asympt	130.0	182.0	f	normal	148.0	no	0.0	flat	0.667774	normal	(50
25	41.0	female	atyp_angina	125.0	184.0	f	normal	180.0	no	0.0	flat	0.667774	normal	(50
26	56.0	male	atyp_angina	130.0	184.0	f	normal	100.0	no	0.0	flat	0.667774	normal	(50
27	52.0	male	typ_angina	118.0	186.0	f	left_vent_hyper	190.0	no	0.0	flat	0.0	fixed_defect	(50
28	43.0	female	atyp_angina	150.0	186.0	f	normal	154.0	no	0.0	flat	0.667774	normal	(50
29	49.0	male	non_anginal	140.0	187.0	f	normal	172.0	no	0.0	flat	0.667774	normal	(50
30	51.0	male	atyp_angina	125.0	188.0	f	normal	145.0	no	0.0	flat	0.667774	normal	(50
31	51.0	female	non_anginal	110.0	190.0	f	normal	120.0	no	0.0	flat	0.667774	normal	(50
32	35.0	male	atyp_angina	122.0	192.0	f	normal	174.0	no	0.0	up	0.0	normal	(50
33	62.0	female	typ_angina	160.0	193.0	f	normal	116.0	no	0.0	flat	0.667774	normal	(50
34	37.0	male	non_anginal	130.0	194.0	f	normal	150.0	no	0.0	flat	0.667774	normal	(50
35	51.0	female	atyp_angina	160.0	194.0	f	normal	170.0	no	0.0	flat	0.667774	normal	(50
36	63.0	female	atyp_angina	140.0	195.0	f	normal	179.0	no	0.0	up	2.0	normal	(50
37	48.0	female	non_anginal	120.0	195.0	f	normal	125.0	no	0.0	flat	0.667774	normal	(50
38	53.0	male	non_anginal	120.0	195.0	f	normal	140.0	no	0.0	flat	0.667774	normal	(50

Add

- Khi mở bằng Notepad++ thì được như sau:

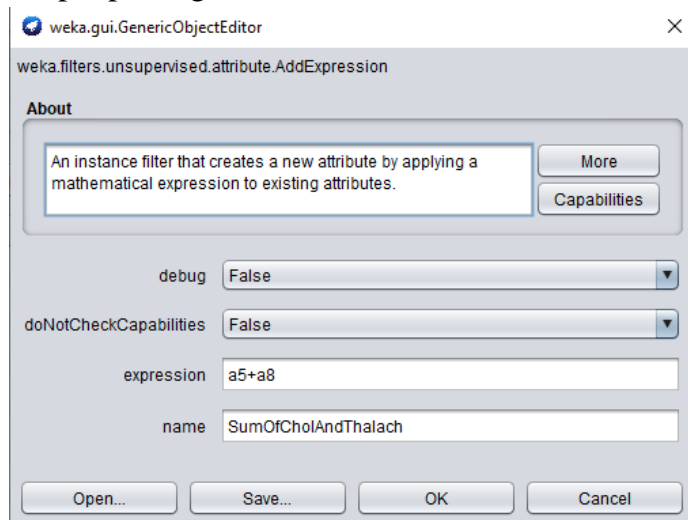
```
F:\kdd\heart\heart-cleaned.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1612176.R new 12 new 13 new 14 Course.h SMS.h info.h new 15 heart-c.arff heart.arff heart-cleaned.arff
1 @relation 'cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsup
2
3 @attribute age numeric
4 @attribute sex {female,male}
5 @attribute cp {typ_angina,asympt,non_anginal,atyp_angina}
6 @attribute trestbps numeric
7 @attribute chol numeric
8 @attribute fbs {t,f}
9 @attribute restecg {left_vent_hyper,normal,st_t_wave_abnormality}
10 @attribute thalach numeric
11 @attribute exang {no,yes}
12 @attribute oldpeak numeric
13 @attribute slope {up,flat,down}
14 @attribute ca numeric
15 @attribute thal {fixed_defect,normal,reversible_defect}
16 @attribute num {<50,>50_1,>50_2,>50_3,>50_4}
17
18 @data
19 56,male,asympt,120,85,f,normal,140,no,0,flat,0.667774,normal,<50
20 52,male,atyp_angina,140,100,f,normal,138,yes,0,flat,0.667774,normal,<50
21 50,male,asympt,140,129,f,normal,135,no,0,flat,0.667774,normal,<50
22 28,male,atyp_angina,130,132,f,left_vent_hyper,185,no,0,flat,0.667774,normal,<50
23 39,male,non_anginal,160,147,t,normal,160,no,0,flat,0.667774,normal,<50
24 42,male,non_anginal,160,147,f,normal,146,no,0,flat,0.667774,normal,<50
25 41,male,atyp_angina,120,157,f,normal,182,no,0,up,0,normal,<50
26 45,female,atyp_angina,112,160,f,normal,138,no,0,flat,0,normal,<50
27 35,female,typ_angina,120,160,f,st_t_wave_abnormality,185,no,0,flat,0.667774,normal,<50
28 36,male,non_anginal,150,160,f,normal,172,no,0,flat,0.667774,normal,<50
29 34,female,atyp_angina,130,161,f,normal,190,no,0,flat,0.667774,normal,<50
30 46,male,non_anginal,150,163,f,normal,116,no,0,flat,0.667774,normal,<50
31 36,male,atyp_angina,120,166,f,normal,180,no,0,flat,0.667774,normal,<50
32 35,female,asympt,140,167,f,normal,150,no,0,flat,0.667774,normal,<50
33 50,male,atyp_angina,120,168,f,normal,160,no,0,flat,0,normal,<50
34 37,female,asympt,130,173,f,st_t_wave_abnormality,184,no,0,flat,0.667774,normal,<50
35 38,male,non_anginal,138,175,f,normal,173,no,0,up,0.667774,normal,<50
36 60,female,non_anginal,120,178,t,normal,96,no,0,up,0,normal,<50
37 51,male,asympt,130,179,f,normal,100,no,0,flat,0.667774,reversible_defect,<50
<
Normal text file length: 41,074 lines: 561 Ln: 1 Col: 1
Type here to search
F:\kdd\heart\heart-cleaned.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1612176.R new 12 new 13 new 14 Course.h SMS.h info.h new 15 heart-c.arff heart.arff heart-cleaned.arff
526 56,male,asympt,132,184,f,left_vent_hyper,105,yes,2.1,flat,1,fixed_defect,>50_1
527 59,male,non_anginal,126,218,t,normal,134,no,2.2,flat,1,fixed_defect,>50_1
528 64,male,asympt,120,246,f,left_vent_hyper,96,yes,2.2,down,1,normal,>50_1
529 54,male,asympt,124,266,f,left_vent_hyper,109,yes,2.2,flat,1,reversible_defect,>50_1
530 60,male,asympt,130,206,f,left_vent_hyper,132,yes,2.4,flat,2,reversible_defect,>50_1
531 38,male,asympt,92,117,f,normal,134,yes,2.5,flat,0.667774,normal,>50_1
532 43,male,asympt,120,177,f,left_vent_hyper,120,yes,2.5,flat,0,reversible_defect,>50_1
533 58,male,non_anginal,112,230,f,left_vent_hyper,165,no,2.5,flat,1,reversible_defect,>50_1
534 52,male,asympt,160,331,f,normal,94,yes,2.5,flat,0.667774,normal,>50_1
535 50,male,asympt,140,341,f,st_t_wave_abnormality,125,yes,2.5,flat,0.667774,normal,>50_1
536 70,male,asympt,145,174,f,normal,125,yes,2.6,down,0,reversible_defect,>50_1
537 67,male,asympt,120,229,f,left_vent_hyper,129,yes,2.6,flat,2,reversible_defect,>50_1
538 61,male,typ_angina,134,234,f,normal,145,no,2.6,flat,2,normal,>50_1
539 50,male,asympt,150,243,f,left_vent_hyper,128,no,2.6,flat,0,reversible_defect,>50_1
540 60,female,asympt,150,258,f,left_vent_hyper,157,no,2.6,flat,2,reversible_defect,>50_1
541 44,male,asympt,120,169,f,normal,144,yes,2.8,down,0,fixed_defect,>50_1
542 58,female,asympt,170,225,t,left_vent_hyper,146,yes,2.8,flat,2,fixed_defect,>50_1
543 54,male,asympt,110,239,f,normal,126,yes,2.8,flat,1,reversible_defect,>50_1
544 65,male,asympt,135,254,f,left_vent_hyper,127,no,2.8,flat,1,reversible_defect,>50_1
545 60,male,asympt,125,258,f,left_vent_hyper,141,yes,2.8,flat,1,reversible_defect,>50_1
546 60,male,asympt,145,282,f,left_vent_hyper,142,yes,2.8,flat,2,reversible_defect,>50_1
547 70,male,non_anginal,160,269,f,normal,112,yes,2.9,flat,1,reversible_defect,>50_1
548 60,male,non_anginal,140,185,f,left_vent_hyper,155,no,3,flat,0,normal,>50_1
549 48,male,asympt,160,193,f,normal,102,yes,3,flat,0.667774,normal,>50_1
550 55,male,asympt,140,201,f,normal,130,yes,3,flat,0.667774,normal,>50_1
551 57,male,asympt,150,255,f,normal,92,yes,3,flat,0.667774,normal,>50_1
552 58,male,asympt,128,259,f,left_vent_hyper,130,yes,3,flat,2,reversible_defect,>50_1
553 36,male,atyp_angina,120,267,f,normal,160,no,3,flat,0.667774,normal,>50_1
554 44,male,atyp_angina,150,288,f,normal,150,yes,3,flat,0.667774,normal,>50_1
555 47,male,asympt,160,291,f,st_t_wave_abnormality,158,yes,3,flat,0.667774,normal,>50_1
556 57,male,asympt,110,335,f,normal,143,yes,3,flat,1,reversible_defect,>50_1
557 41,male,asympt,120,336,f,normal,118,yes,3,flat,0.667774,normal,>50_1
558 43,female,asympt,132,341,t,left_vent_hyper,136,yes,3,flat,0,reversible_defect,>50_1
559 56,male,asympt,155,342,t,normal,150,yes,3,flat,0.667774,normal,>50_1
560 53,male,asympt,140,203,t,left_vent_hyper,155,yes,3.1,down,0,reversible_defect,>50_1
561
Normal text file length: 41,074 lines: 561 Ln: 27 Col: 83 Sel: 0|0
Type here to search
```

V. CHUYỂN ĐỔI DỮ LIỆU (TRANSFORMATION):

Trong số các kỹ thuật chuyển đổi dữ liệu, sử dụng các bộ lọc của Weka để tìm hiểu các kỹ thuật sau:

1. Xây dựng thuộc tính – Attribute construction:

- Bộ lọc của weka cho phép thêm một thuộc tính
- Chọn kiểu dữ liệu của thuộc tính, đặt tên thuộc tính, chọn giá trị khởi tạo mặc định.
- Thông qua bộ lọc Add Expression.
- Ở đây ví dụ ta muốn thêm một cột có giá trị là tổng của cột chol (a5) và thalach(a8):
- Cú pháp trong bộ lọc như sau:



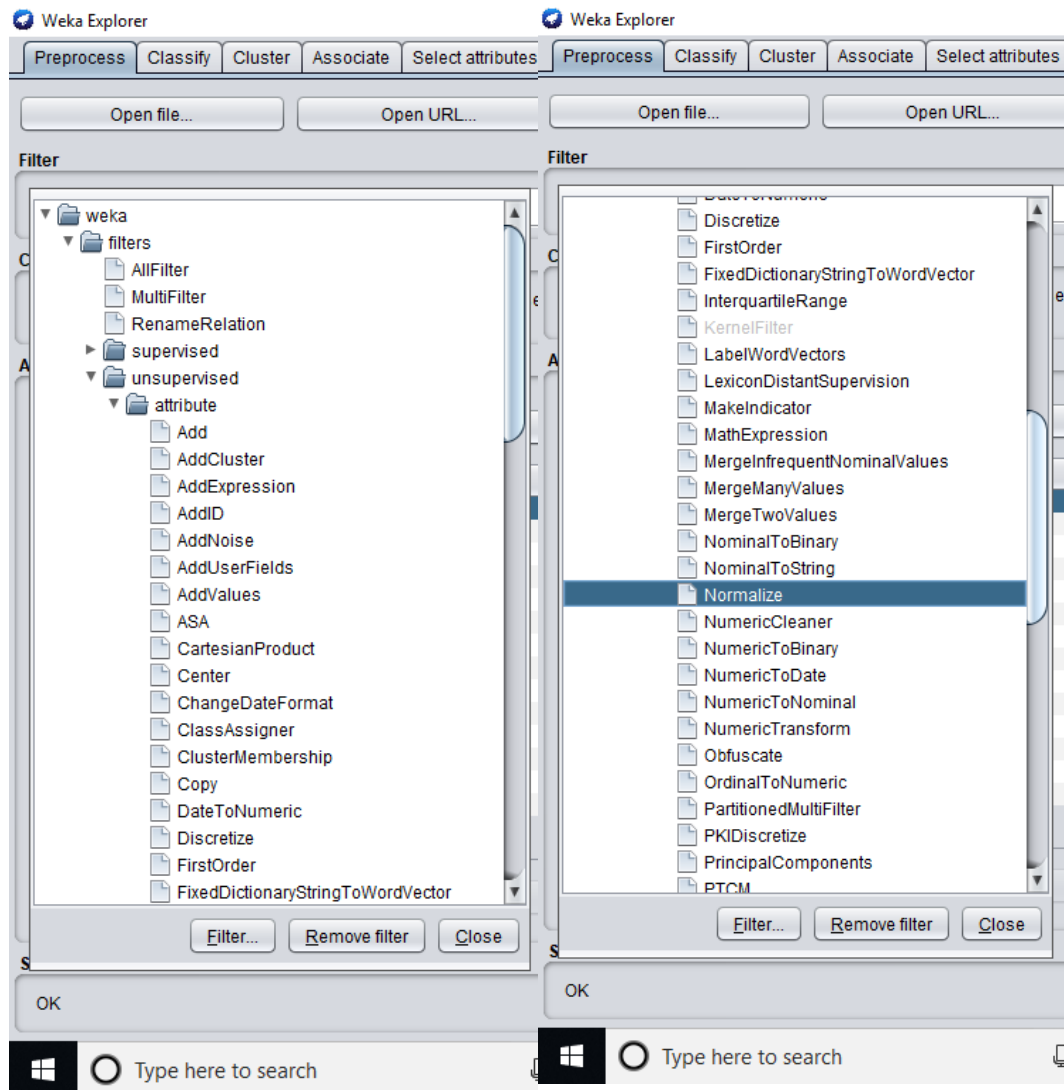
- Kết quả:

5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num	15: SumOfCholAndThalach
Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Numeric
85.0	f	normal	140.0	no	0.0	flat	0.667774	normal	(50	225.0

2. Chuẩn hóa – Normalize một thuộc tính:

- Bộ lọc để chuẩn hóa: Normalize, Standardize
- Bộ lọc chuẩn hóa min-max: Normalize.
 - Cách thức thực hiện:

- Chọn Choose → Chọn Filter → Chọn unsupervised → Chọn Attribute → chọn Normalize.



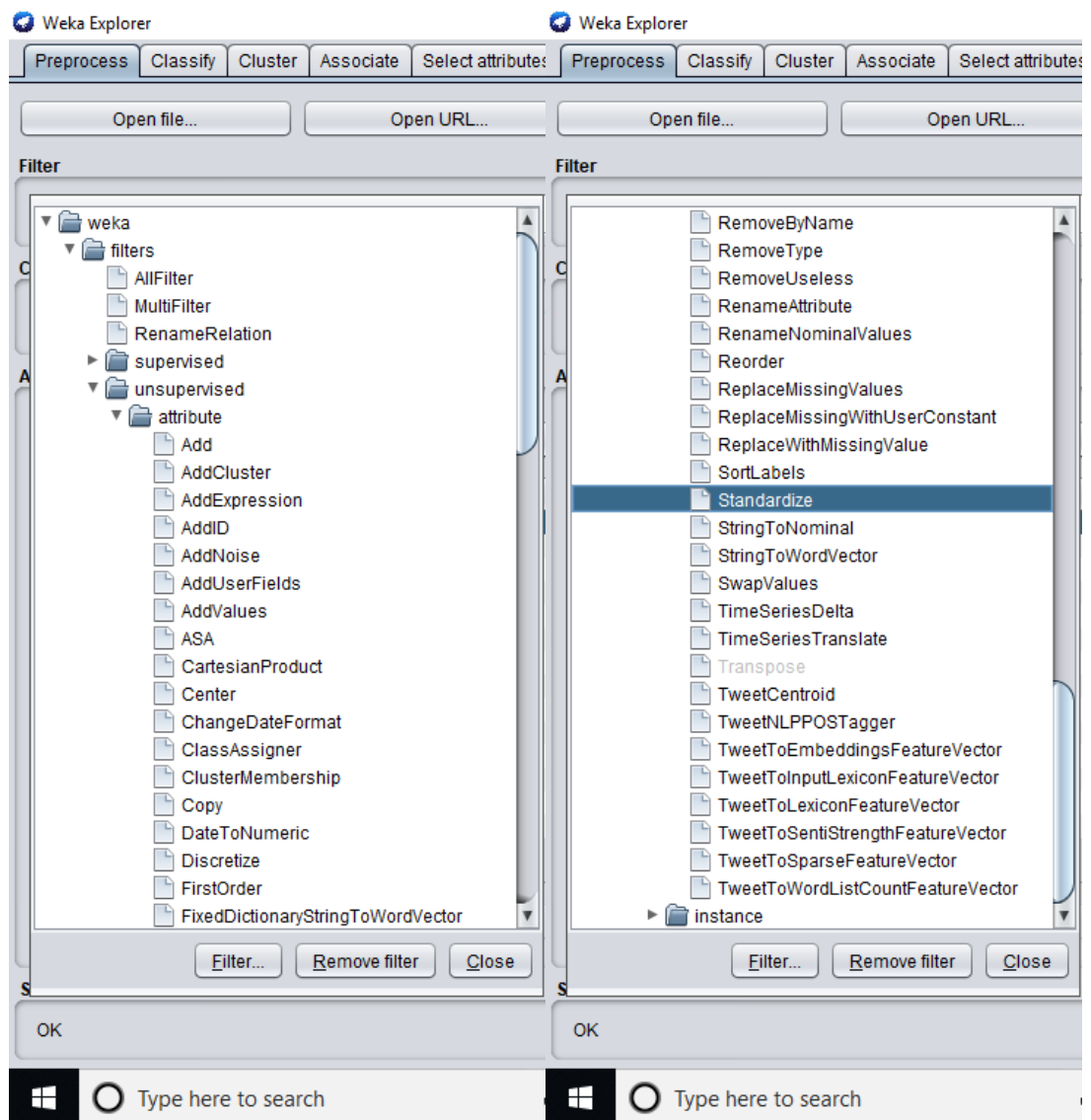
- Chọn Apply, ta thấy dữ liệu đã được chuẩn hóa theo min-max, như sau:

Selected attribute		
Name: age	Distinct: 46	Type: Numeric
Missing: 0 (0%)		Unique: 4 (1%)
Statistic	Value	
Minimum	0	
Maximum	1	
Mean	0.473	
StdDev	0.186	

- Bộ lọc chuẩn hóa Z-score: Standardize

- Cách thức thực hiện:

- Chọn Choose → Chọn Filter → Chọn unsupervised → Chọn Attribute → chọn Standardize.



- Chọn Apply, ta thấy thấy dữ liệu được chuẩn hóa theo phân phối chuẩn (phân phối Gaussian).
 - Bộ lọc chuẩn hóa thập phân: Không tồn tại trong weka
- 3. Tiến hành chuẩn hóa tất cả các thuộc tính là số thực, giải thích sự lựa chọn:
 - Chọn phương pháp min-max
 - Lý do: Dễ thao tác, các con số có vẻ thân thiện hơn với Z-score. Các giá trị từ 0 đến 1, người xử lý dễ dàng xác định xác suất hay tỷ lệ giá trị so với max.
- 4. File heart-normal.arff:
 - Trong Edit:

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.InterquartileRange

No.	1: age	2: sex	3: cp	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	0.583333	male	asympt	0.259259	0.0	f	normal	0.526718	no	0.0	flat	0.333887	normal	(50
2	0.5	male	atyp_angina	0.444444	0.053571	f	normal	0.51145	yes	0.0	flat	0.333887	normal	(50
3	0.458333	male	asympt	0.444444	0.157143	f	normal	0.48855	no	0.0	flat	0.333887	normal	(50
4	0.0	male	atyp_angina	0.351852	0.167857	f	left_vent_hyper	0.870229	no	0.0	flat	0.333887	normal	(50
5	0.229167	male	non_anginal	0.62963	0.221429	t	normal	0.679389	no	0.0	flat	0.333887	normal	(50
6	0.291667	male	non_anginal	0.62963	0.221429	f	normal	0.572519	no	0.0	flat	0.333887	normal	(50
7	0.270833	male	atyp_angina	0.259259	0.257143	f	normal	0.847328	no	0.0	up	0.0	normal	(50
8	0.354167	female	atyp_angina	0.185185	0.267857	f	normal	0.51145	no	0.0	flat	0.0	normal	(50
9	0.145833	female	typ_angina	0.259259	0.267857	f	st_t_wave_abnormality	0.870229	no	0.0	flat	0.333887	normal	(50
10	0.166667	male	non_anginal	0.537037	0.267857	f	normal	0.770992	no	0.0	flat	0.333887	normal	(50
11	0.125	female	atyp_angina	0.351852	0.271429	f	normal	0.908397	no	0.0	flat	0.333887	normal	(50
12	0.375	male	non_anginal	0.537037	0.278571	f	normal	0.343511	no	0.0	flat	0.333887	normal	(50
13	0.166667	male	atyp_angina	0.259259	0.289286	f	normal	0.832061	no	0.0	flat	0.333887	normal	(50
14	0.145833	female	asympt	0.444444	0.292857	f	normal	0.603053	no	0.0	flat	0.333887	normal	(50
15	0.458333	male	atyp_angina	0.259259	0.296429	f	normal	0.679389	no	0.0	flat	0.0	normal	(50
16	0.1875	female	asympt	0.351852	0.314286	f	st_t_wave_abnormality	0.862595	no	0.0	flat	0.333887	normal	(50
17	0.208333	male	non_anginal	0.425926	0.321429	f	normal	0.778626	no	0.0	up	0.333887	normal	(50
18	0.666667	female	non_anginal	0.259259	0.332143	t	normal	0.19084	no	0.0	up	0.0	normal	(50
19	0.479167	male	asympt	0.351852	0.335714	f	normal	0.221374	no	0.0	flat	0.333887	reversible_defect	(50
20	0.625	male	non_anginal	0.444444	0.335714	f	normal	0.679389	no	0.0	flat	0.333887	normal	(50
21	0.291667	male	non_anginal	0.339286	0.339286	f	normal	0.603053	no	0.0	up	0.0	normal	(50
22	0.125	male	typ_angina	0.240741	0.346429	f	left_vent_hyper	0.78626	no	0.0	up	0.0	normal	(50
23	0.229167	female	non_anginal	0.166667	0.346429	f	st_t_wave_abnormality	0.832061	no	0.0	flat	0.333887	normal	(50
24	0.520833	male	asympt	0.351852	0.346429	f	normal	0.587786	no	0.0	flat	0.333887	normal	(50
25	0.270833	female	atyp_angina	0.305556	0.353571	f	normal	0.832061	no	0.0	flat	0.333887	normal	(50
26	0.583333	male	atyp_angina	0.351852	0.353571	f	normal	0.221374	no	0.0	flat	0.333887	normal	(50
27	0.5	male	typ_angina	0.240741	0.360714	f	left_vent_hyper	0.908397	no	0.0	flat	0.0	fixed_defect	(50
28	0.3125	female	atyp_angina	0.537037	0.360714	f	normal	0.633588	no	0.0	flat	0.333887	normal	(50
29	0.4375	male	non_anginal	0.444444	0.364286	f	normal	0.770992	no	0.0	flat	0.333887	normal	(50
30	0.479167	male	atyp_angina	0.305556	0.367857	f	normal	0.564885	no	0.0	flat	0.333887	normal	(50
31	0.479167	female	non_anginal	0.166667	0.375	f	normal	0.374046	no	0.0	flat	0.333887	normal	(50
32	0.145833	male	atyp_angina	0.277778	0.382143	f	normal	0.78626	no	0.0	up	0.0	normal	(50
33	0.708333	female	typ_angina	0.62963	0.385714	f	normal	0.343511	no	0.0	flat	0.333887	normal	(50
34	0.1875	male	non_anginal	0.351852	0.389286	f	normal	0.603053	no	0.0	flat	0.333887	normal	(50
35	0.479167	female	atyp_angina	0.62963	0.389286	f	normal	0.755725	no	0.0	flat	0.333887	normal	(50
36	0.729167	female	atyp_angina	0.444444	0.392857	f	normal	0.824427	no	0.0	up	1.0	normal	(50
37	0.416667	female	non_anginal	0.259259	0.392857	f	normal	0.412214	no	0.0	flat	0.333887	normal	(50
38	0.520833	male	non_anginal	0.259259	0.392857	f	normal	0.526718	no	0.0	flat	0.333887	normal	(50

Add Instance Undo OK Cancel

– Trong notepad++:

```

F:\kdd\heart\heart-normal.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1 @relation "cleveland-14-heart-disease-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.InterquartileRange"
2
3 @attribute age numeric
4 @attribute sex {female,male}
5 @attribute cp {typ_angina,asympt,non_anginal,atyp_angina}
6 @attribute trestbps numeric
7 @attribute chol numeric
8 @attribute fbs {t,f}
9 @attribute restecg {left_vent_hyper,normal,st_t_wave_abnormality}
10 @attribute thalach numeric
11 @attribute exang {no,yes}
12 @attribute oldpeak numeric
13 @attribute slope {up,flat,down}
14 @attribute ca numeric
15 @attribute thal {fixed_defect,normal,reversible_defect}
16 @attribute num {<50,>50_1,>50_2,>50_3,>50_4}
17
18 @data
19 0.583333,male,asympt,0.259259,0,f,normal,0.526718,no,0,flat,0.333887,normal,<50
20 0.5,male,atyp_angina,0.444444,0.053571,f,normal,0.51145,yes,0,flat,0.333887,normal,<50
21 0.458333,male,asympt,0.444444,0.157143,f,normal,0.48855,no,0,flat,0.333887,normal,<50
22 0,male,atyp_angina,0.351852,0.167857,f,left_vent_hyper,0.870229,no,0,flat,0.333887,normal,<50
23 0.229167,male,non_anginal,0.62963,0.221429,t,normal,0.679389,no,0,flat,0.333887,normal,<50
24 0.291667,male,non_anginal,0.62963,0.221429,f,normal,0.572519,no,0,flat,0.333887,normal,<50
25 0.270833,male,atyp_angina,0.259259,0.257143,f,normal,0.847328,no,0,up,0,normal,<50
26 0.354167,female,atyp_angina,0.185185,0.267857,f,normal,0.51145,no,0,flat,0,normal,<50
27 0.145833,female,typ_angina,0.259259,0.267857,f,st_t_wave_abnormality,0.870229,no,0,flat,0.333887,normal,<50
28 0.166667,male,non_anginal,0.537037,0.267857,f,normal,0.770992,no,0,flat,0.333887,normal,<50
29 0.125,female,atyp_angina,0.351852,0.271429,f,normal,0.908397,no,0,flat,0.333887,normal,<50
30 0.375,male,non_anginal,0.537037,0.278571,f,normal,0.343511,no,0,flat,0.333887,normal,<50
31 0.166667,male,atyp_angina,0.259259,0.289286,f,normal,0.832061,no,0,flat,0.333887,normal,<50
32 0.145833,female,asympt,0.444444,0.292857,f,normal,0.603053,no,0,flat,0.333887,normal,<50
33 0.458333,male,atyp_angina,0.259259,0.296429,f,normal,0.679389,no,0,flat,0,normal,<50
34 0.1875,female,asympt,0.351852,0.314286,f,st_t_wave_abnormality,0.862595,no,0,flat,0.333887,normal,<50
35 0.208333,male,non_anginal,0.425926,0.321429,f,normal,0.778626,no,0,up,0.333887,normal,<50
36 0.666667,female,non_anginal,0.259259,0.332143,t,normal,0.19084,no,0,up,0,normal,<50
37 0.479167,male,asympt,0.351852,0.335714,f,normal,0.221374,no,0,flat,0.333887,reversible_defect,<50
38

```

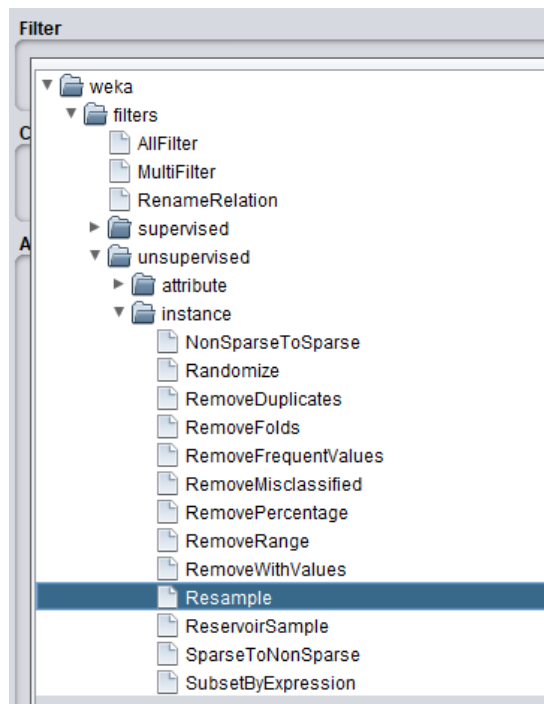
Normal text file length: 53,241 lines: 561 Ln: 14 Col: 22 Sel: 0 | 0

VI. RÚT GỌN DỮ LIỆU (REDUCTION):

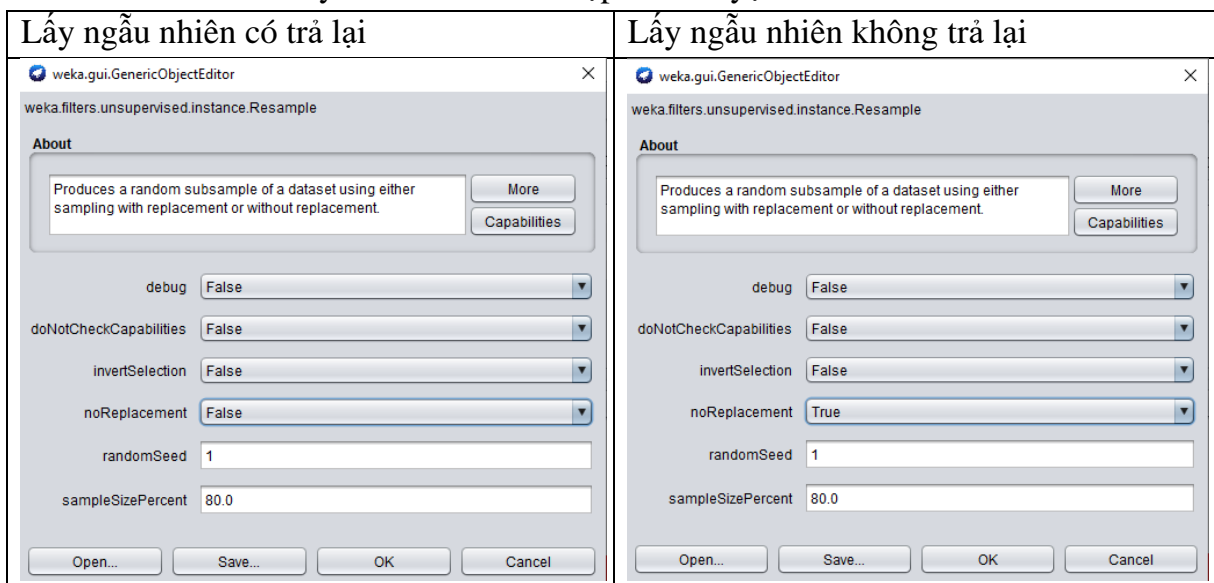
Các cơ sở dữ liệu thường rất lớn, không thể thao tác trực tiếp được. Các kỹ thuật rút gọn dữ liệu được áp dụng để tiền xử lý dữ liệu. Trong tab Preprocess, bên cạnh việc chọn lọc thuộc tính, một phương pháp để rút gọn dữ liệu là chọn lọc các dòng trong một dataset, hay còn gọi là lấy mẫu (sampling).

– Cách để lấy mẫu với các bộ lọc của Weka:

- Chọn Choose → Chọn unsupervised → Chọn instance → Chọn Resample.



- Sau đó, chọn khung quy định của resample với mục đích là tiến hành sửa lại số record lấy làm mẫu để làm tập huấn luyện.



- Con số 80%: Ở đây theo quyển “Hands-On Machine Learning with Scikit-Learn and TensorFlow” cuối trang 29, 80 sẽ cho tập huấn luyện và 20% sẽ cho tập kiểm tra.
- Hai phương pháp chính để lấy mẫu:
 - Simple Random Sample Without Replacement: Lấy mẫu ngẫu nhiên không trả lại.
 - Simple Random Sample With Replacement: Lấy mẫu ngẫu nhiên có trả lại.

VII. TỔNG KẾT:

STT	Tên công việc	Tỷ lệ hoàn thành	Ghi chú
1	Báo cáo	100	
2	Tích hợp	100	
3	Tóm tắt	100	
4	Chọn lọc	100	
5	Làm sạch	96	Mục IV: phần 1: gạch đầu dòng cuối cùng: phần thao tác điền dữ liệu còn thiếu bằng thuật toán cây quyết định
6	Chuyển đổi	96	Mục IV: phần 2: gạch đầu dòng cuối cùng: Bộ lọc chuẩn hóa thập phân
7	Rút gọn	100	

Tham khảo:

[1]: <http://www.lastnightstudy.com/Show?id=39/Data-Integration-In-Data-Mining>

[2]: Data Mining. Concepts and Techniques, 3rd Edition

[3]: <https://www.slideshare.net/kavithamuneeshwaran/data-integration-and-transformation-in-data-mining>

[4] [5]: http://scholar.vimaru.edu.vn/sites/default/files/thinhnv/files/dm_-_chapter_2_-_preprocessing_0.pdf

[6]: <https://voer.edu.vn/m/tim-kiem-uu-tien-toi-uu-best-first-search/5f9cfb74>

[7]: Tham khảo chú thích trong weka.

[8]: trang 103 đến trang 104

[9]: <https://voer.edu.vn/m/tim-kiem-uu-tien-toi-uu-best-first-search/5f9cfb74>

[10]: Trang 336, 337

[11]: Khai thác dữ liệu: Làm sạch dữ liệu - Data Cleaning:

<https://www.youtube.com/watch?v=PidRQCyUKPg&list=PL67CJL04EcjN4hIkgZT3dgyxETTXNOiZb&index=12>

[12]: <https://www.kdnuggets.com/2018/12/four-techniques-outlier-detection.html>

Quyển: [https://github.com/devakar/deep-learning-](https://github.com/devakar/deep-learning-books/blob/master/Hands%20on%20Machine%20Learning%20with%20Scikit%20Learn%20and%20TensorFlow.pdf?fbclid=IwAR0wGce2E8Z8FVLBrx1RcH4S2SWek1PjGPTLzeKFNaPuKikglUKadLYf5Jo)

[books/blob/master/Hands%20on%20Machine%20Learning%20with%20Scikit%20Learn%20and%20TensorFlow.pdf?fbclid=IwAR0wGce2E8Z8FVLBrx1RcH4S2SWek1PjGPTLzeKFNaPuKikglUKadLYf5Jo](https://github.com/devakar/deep-learning-books/blob/master/Hands%20on%20Machine%20Learning%20with%20Scikit%20Learn%20and%20TensorFlow.pdf?fbclid=IwAR0wGce2E8Z8FVLBrx1RcH4S2SWek1PjGPTLzeKFNaPuKikglUKadLYf5Jo)