



TIỀN XỬ LÝ DỮ LIỆU 1

Giảng viên: Nguyễn Tu Trung
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2021

Nội dung

- ❖ Chuẩn bị dữ liệu
- ❖ Nạp dữ liệu (Loading the Data)
- ❖ Hiển thị dữ liệu
- ❖ Loại bỏ-lựa chọn thuộc tính
- ❖ Rời rạc hóa dữ liệu (Discretization)
- ❖ Chỉnh sửa dữ liệu

Chuẩn bị dữ liệu

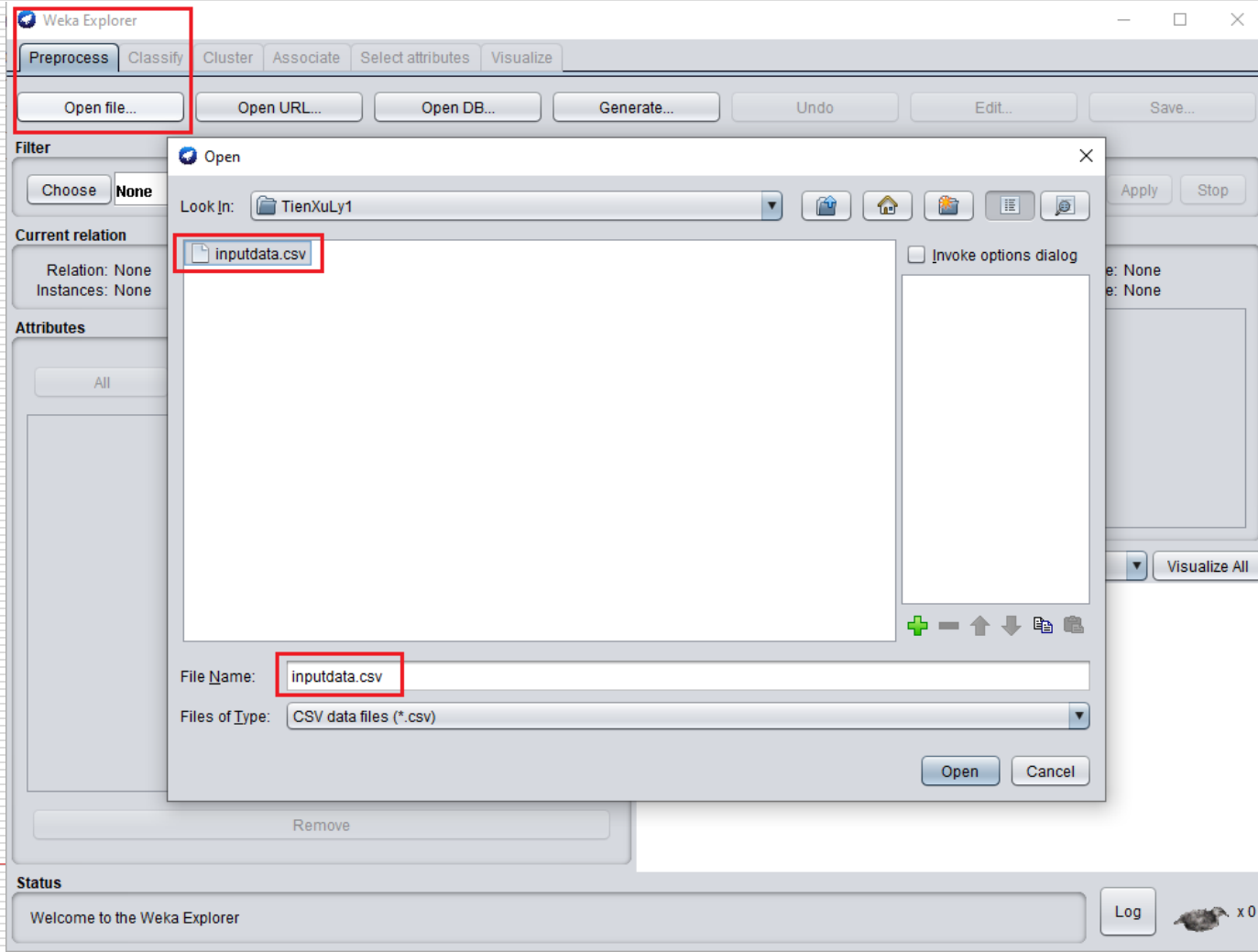
- ❖ File dữ liệu: `inputdata.csv`
- ❖ Gồm các trường:
 - ❖ Id: Mã khách hàng, Age: Tuổi khách hàng
 - ❖ Sex, Region: Nơi cư trú
 - ❖ Incom: Thu nhập, Married: Tình trạng hôn nhân
 - ❖ Children: Số con, Car: Có xe hơi không?
 - ❖ Save_Act: Có tk tiết kiệm không?
 - ❖ Current_Act: Hiện tại có tk không?
 - ❖ Mortgage: Có thể chấp không?
 - ❖ Pep: Có kế hoạch trả nợ không?

Chuẩn bị dữ liệu

	id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
2	ID12101	48	FEMALE	INNER_CITY	17546	NO	1	NO	NO	NO	NO	YES
3	ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
4	ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	YES	NO	NO
5	ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
6	ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
7	ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES
8	ID12107	22	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES
9	ID12108	58	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO
0	ID12109	37	FEMALE	SUBURBAN	25304.3	YES	2	YES	NO	NO	NO	NO
1	ID12110	54	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO
2	ID12111	66	FEMALE	TOWN	59803.9	YES	0	NO	YES	YES	NO	NO
3	ID12112	52	FEMALE	INNER_CITY	26658.8	NO	0	YES	YES	YES	YES	NO
4	ID12113	44	FEMALE	TOWN	15735.8	YES	1	NO	YES	YES	YES	YES
5	ID12114	66	FEMALE	TOWN	55204.7	YES	1	YES	YES	YES	YES	YES
6	ID12115	36	MALE	RURAL	19474.6	YES	0	NO	YES	YES	YES	NO
7	ID12116	38	FEMALE	INNER_CITY	22342.1	YES	0	YES	YES	YES	YES	NO
8	ID12117	37	FEMALE	TOWN	17729.8	YES	2	NO	NO	NO	YES	NO
9	ID12118	46	FEMALE	SUBURBAN	41016	YES	0	NO	YES	NO	YES	NO
0	ID12119	62	FEMALE	INNER_CITY	26909.2	YES	0	NO	YES	NO	NO	YES
1	ID12120	31	MALE	TOWN	22522.8	YES	0	YES	YES	YES	NO	NO
2	ID12121	61	MALE	INNER_CITY	57880.7	YES	2	NO	YES	NO	NO	YES
3	ID12122	50	MALE	TOWN	16497.3	YES	2	NO	YES	YES	NO	NO
4	ID12123	54	MALE	INNER_CITY	38446.6	YES	0	NO	YES	YES	NO	NO

Nạp dữ liệu (Loading the Data)

- ❖ Trong Weka Explorer, chọn tab Preprocess
- ❖ Chọn Open file => Chọn file inputdata.csv



Hiển thị dữ liệu

- ❖ Bên trái: Danh sách các thuộc tính
- ❖ Bên phải: Thống kê tương ứng với trường được chọn

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply Stop

Current relation: Relation: inputdata Instances: 600 Attributes: 12 Sum of weights: 600

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> id
2	<input checked="" type="checkbox"/> age
3	<input type="checkbox"/> sex
4	<input type="checkbox"/> region
5	<input type="checkbox"/> income
6	<input type="checkbox"/> married
7	<input type="checkbox"/> children
8	<input type="checkbox"/> car
9	<input type="checkbox"/> save_act
10	<input type="checkbox"/> current_act
11	<input type="checkbox"/> mortgage
12	<input type="checkbox"/> pep

Remove

Selected attribute

Name: age Missing: 0 (0%) Distinct: 50 Type: Numeric Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	67
Mean	42.395
StdDev	14.425

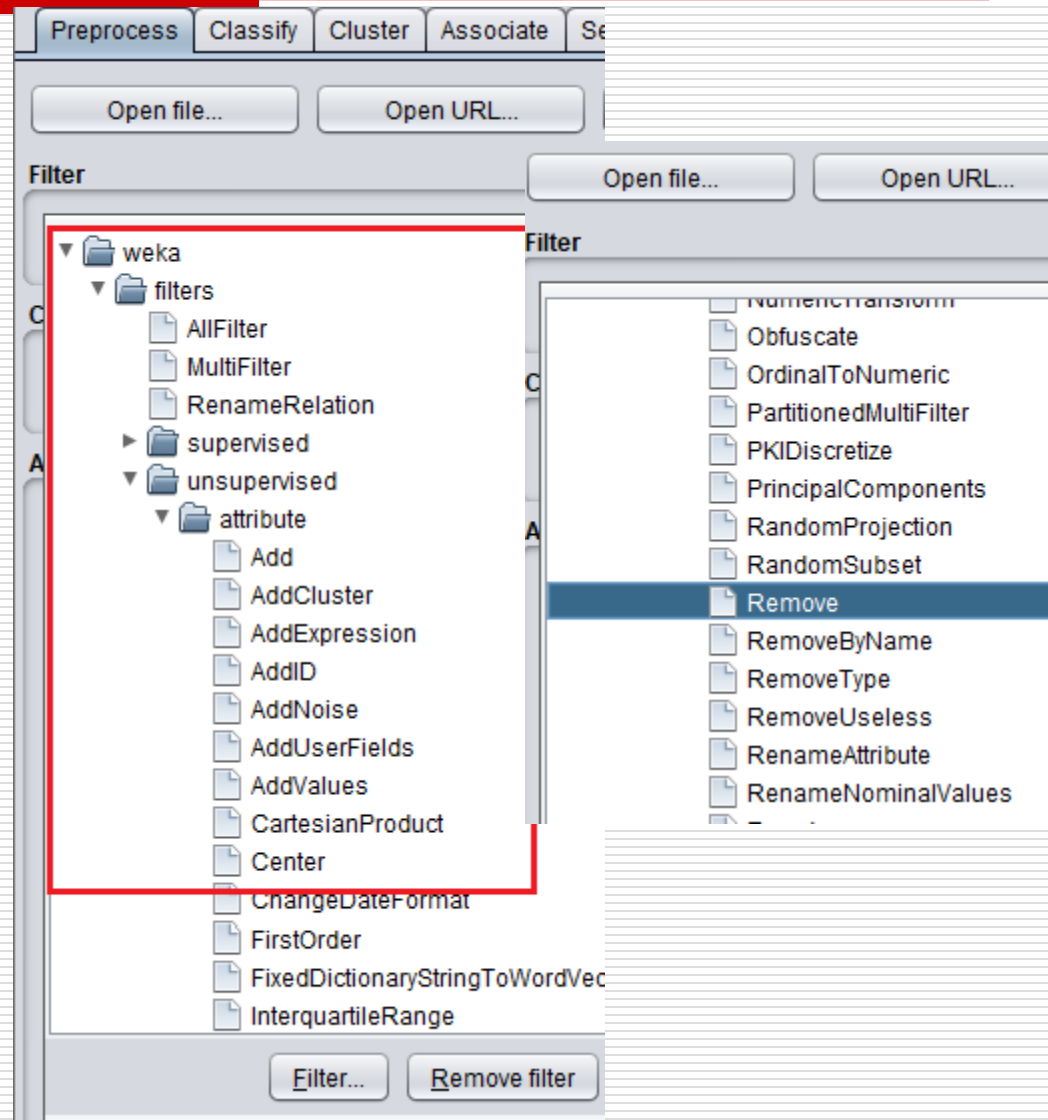
Class: pep (Nom) Visualize All

Bar chart showing the distribution of the 'age' attribute. The x-axis represents age values, and the y-axis represents frequency. The chart is divided into two series: red (top) and blue (bottom).

Status: OK Log x 0

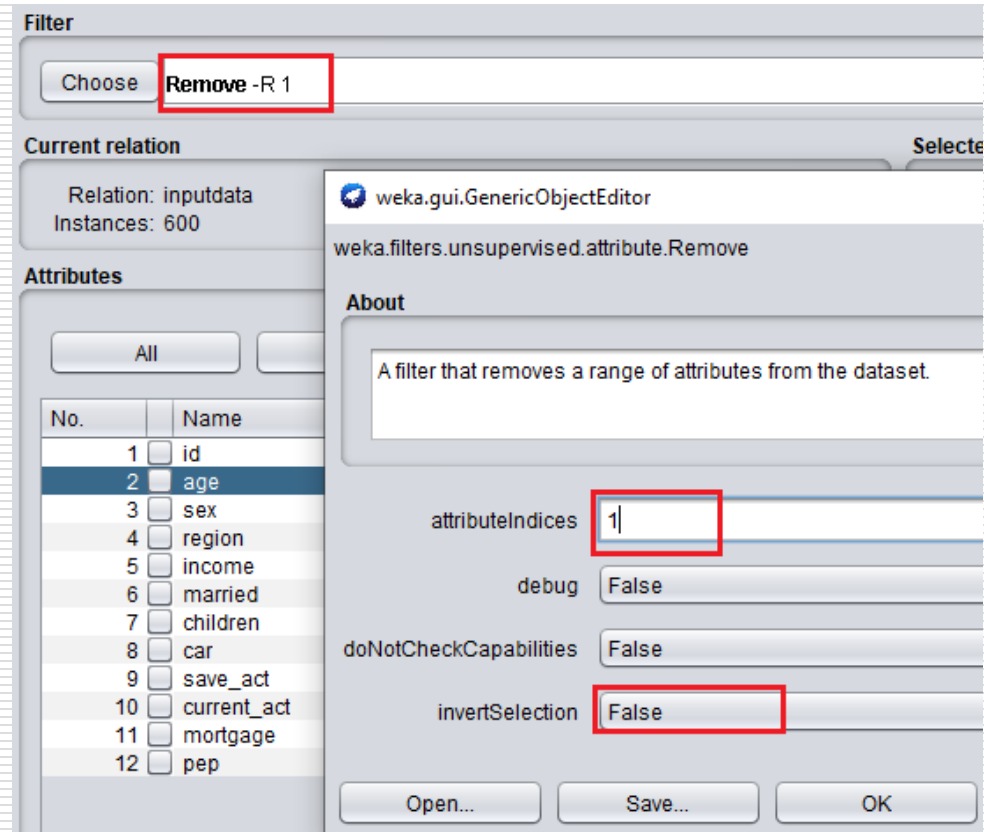
Loại bỏ-lựa chọn thuộc tính

- ❖ Khi thực hiện khai phá dữ liệu, có thể:
 - ❖ Chỉ cần một số trường cần thiết (lựa chọn)
 - ❖ Và loại bỏ một số trường khác
- ❖ Ví dụ: Loại bỏ trường **id**
- ❖ Trong group Filter, nhấn nút “Choose” => *filters > unsupervised > attribute > Remove*



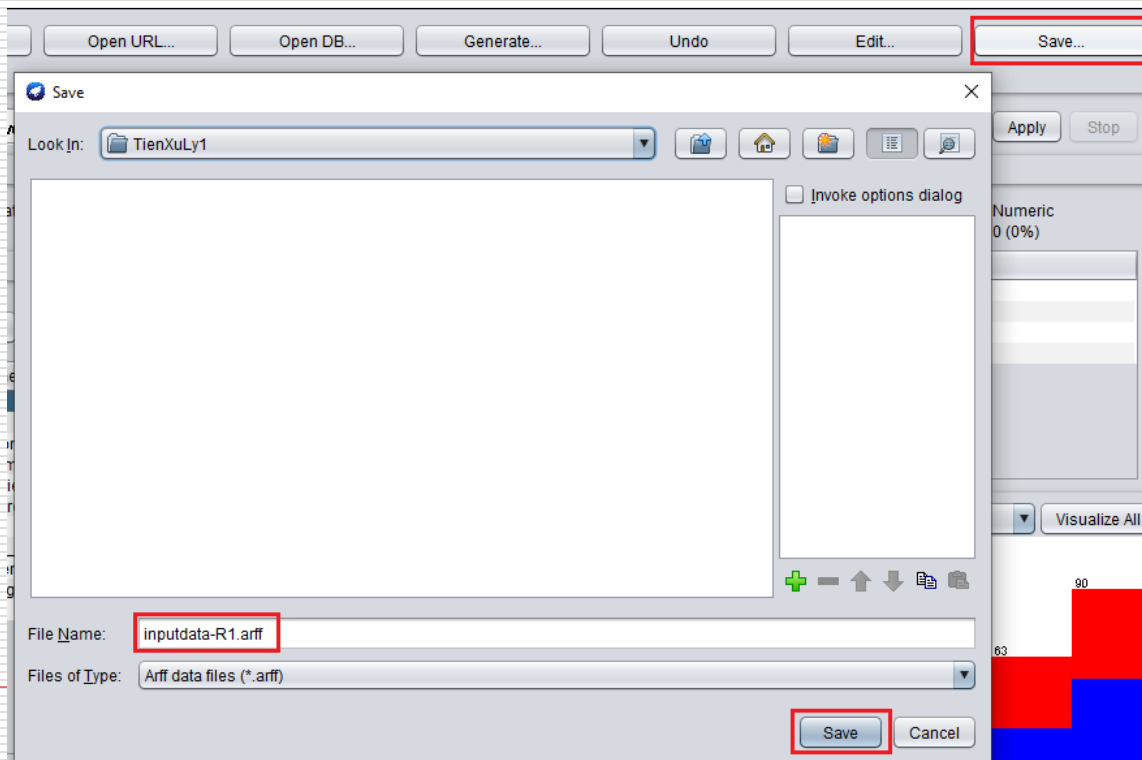
Loại bỏ-lựa chọn thuộc tính

- ❖ Nhấn chuột trái vào Textbox bên phải nút “Choose”
- ❖ Trong hộp thoại hiện ra
- ❖ Nhập “1” - ứng với chỉ số trường Id vào hộp attributeIndices
- ❖ Nhập “False” vào hộp invertSelection – không đảo lựa chọn
- ❖ Nhấn OK
- ❖ Nhấn Apply



Loại bỏ-lựa chọn thuộc tính

- ❖ Kết quả: Thuộc tính Id đã được loại bỏ
- ❖ Tương tự có thể thực hiện loại bỏ thuộc tính khác không cần thiết với mục đích khai phá dữ liệu
- ❖ Lưu lại với tên: [inputdata1.arff](#)



Loại bỏ-lựa chọn thuộc tính

- ❖ Nội dung file: [inputdata-R1.arff](#)
- ❖ Trường Id đã bị loại khỏi phần mô tả thuộc tính và dữ liệu

```
inputdata-R1.arff
1 @relation bank-data-weka.filters.unsupervised.attribute.Remove-R1
2
3 @attribute age numeric
4 @attribute sex {FEMALE,MALE}
5 @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
6 @attribute income numeric
7 @attribute married {NO,YES}
8 @attribute children numeric
9 @attribute car {NO,YES}
10 @attribute save_act {NO,YES}
11 @attribute current_act {NO,YES}
12 @attribute mortgage {NO,YES}
13 @attribute pep {YES,NO}
14
15 @data
16 48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES
17 40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
18 51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
19 23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
20 57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
21 57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES
22 22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES
23 58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO
```

Rời rạc hóa dữ liệu (Discretization)

- ❖ Một số kỹ thuật khai phá dữ liệu như khái phá luật kết hợp **chỉ có thể thực hiện trên các dữ liệu phân loại** (categorical/ nominal data)
- ❖ Nếu muốn sử dụng kỹ thuật đó => Phải rời rạc hóa trên các thuộc tính có kiểu dữ liệu liên tục (như kiểu numeric)
- ❖ File **inputdata-R1.arff** có 3 trường-thuộc tính kiểu số: “age”, “income”, và “children”
- ❖ Rời rạc hóa:
 - ❖ Thuộc tính: “children”
 - ❖ Các thuộc tính: “age” và “income”

Rời rạc hóa thuộc tính “children”

- ❖ Phạm vi giá của nó chỉ có thể là 0,1,2 và 3 => Có thể giữ lại các giá trị của thuộc tính này
- ❖ Mở file “inputdata-R1.arff” bằng bất kỳ text editor nào (ví dụ: WordPad, Notepad)
- ❖ Thay từ khóa “numeric” bằng các giá trị rời rạc {0,1,2,3}
- ❖ Lưu kết quả lại với tên file “inputdata2.arff”

```
@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
```

Rời rạc hóa thuộc tính “age” và “income”

- ❖ Mở file dữ liệu “inputdata2.arff”
- ❖ Thuộc tính “age” và “income” có kiểu dữ liệu là numeric
- ❖ Chọn thuộc tính “children” => kiểu dữ liệu đã chuyển về nominal

The screenshot shows the Weka GUI. On the left, the 'Current relation' panel displays 'Relation: bank-data-weka.filters.unsupervised.attribute...' and 'Instances: 600'. Below it, the 'Attributes' list shows 11 attributes: age, sex, region, income, married, children, car, save_act, current_act, mortgage, and pep. The 'children' attribute is selected. On the right, the 'Selected attribute' panel shows details for 'children': Name: children, Missing: 0 (0%), Distinct: 4, Type: Nominal, and Unique: 0 (0%). A table below shows the distribution of values for 'children':

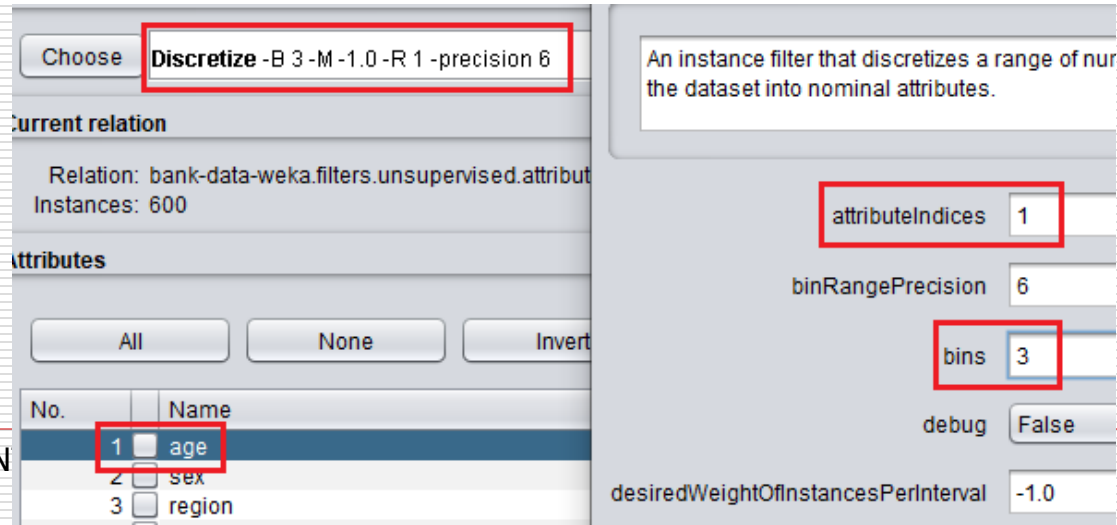
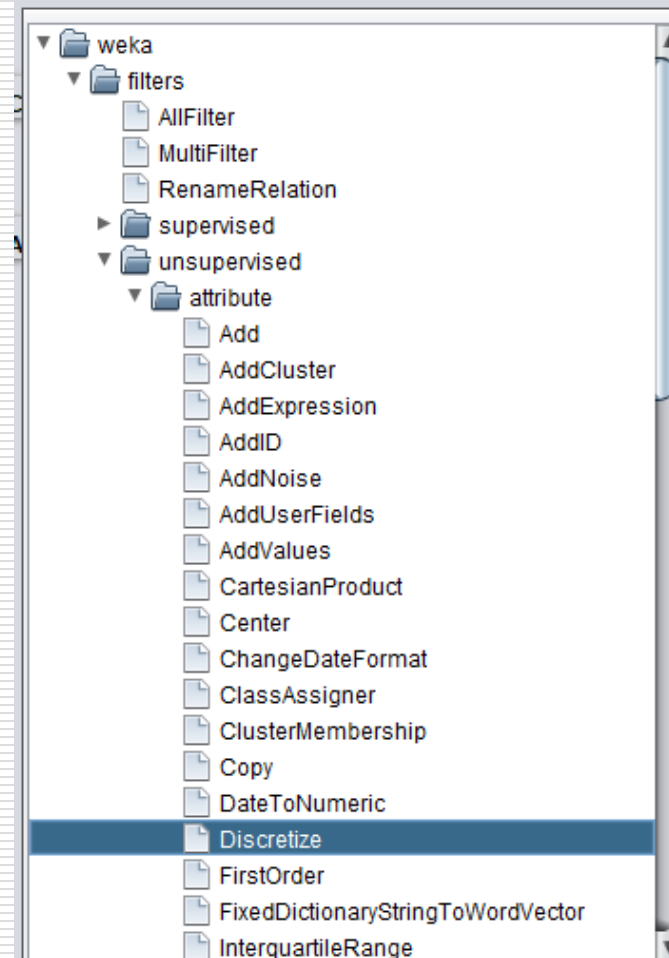
No.	Label	Count	Weight
1	0	263	263.0
2	1	135	135.0
3	2	134	134.0
4	3	68	68.0

At the bottom, the 'Class' is set to 'pep (Nom)'.

- ❖ Bộ lọc *Discretize*: chia nhỏ (binning), sắp xếp và chia dữ liệu vào các giỏ có cùng độ rộng (equal-width) => Chia vùng giá trị thành N khoảng cùng kích thước
- ❖ Độ rộng của từng khoảng = (giá trị lớn nhất – giá trị nhỏ nhất)/N => Mặc định, Weka gán N=10

Rời rạc hóa thuộc tính “age” và “income”

- ❖ Nhấn nút Filter => Chọn filters > **unsupervised** > attribute > Discretize
- ❖ Nhấn Textbox bên phải nút “Choose” để mở hộp thoại thiết lập tham số
- ❖ Nhập 1 tương ứng với index của thuộc tính “age” trong textbox attributeIndices
- ❖ Nếu muốn chia giá trị tuổi về 3 khoảng => Nhập 3 trong textbox bins => Nhấn OK



Rời rạc hóa thuộc tính “age” và “income”

- ❖ Click "Apply" để thực hiện
- ❖ Kết quả được tạo ra trong một working relation mới
 - ❖ Các giá trị liên tục trong thuộc tính “age” được tự động chia vào 3 khoảng có nhãn lần lượt là "(-inf-34.333333]", "(34.333333-50.666667]" "(50.666667- inf)"

The screenshot shows the Weka Discretize tool interface. The "Choose" button is set to "Discretize -B 3 -M -1.0 -R 1 -precision 6". The "Apply" button is highlighted with a red box. The "Selected attribute" section shows "Name: age", "Missing: 0 (0%)", "Distinct: 3", and "Type: Nominal". A table below shows the resulting intervals and their counts and weights.

No.	Label	Count	Weight
1	'(-inf-34.333333]'	195	195.0
2	'(34.333333-50.666667]'	214	214.0
3	'(50.666667-inf)'	191	191.0

In the "Attributes" section, the "age" attribute is selected, highlighted with a red box.

Rời rạc hóa thuộc tính “age” và “income”

- ❖ Tương tự thực hiện với thuộc tính “income”

The screenshot shows the WEKA Discretize filter configuration and its results. On the left, the filter settings are: attributeIndices set to 4, binRangePrecision set to 6, bins set to 3, and debug set to False. On the right, the 'Selected attribute' panel shows the 'income' attribute has been discretized into 3 nominal bins. The results table shows the following data:

No.	Label	Count	Weight
1	'(-inf-24386.173333]'	285	285.0
2	'(24386.173333-437...'	235	235.0
3	'(43758.136667-inf)'	80	80.0

- ❖ Lưu kết quả lại với tên file “inputdata3.arff”

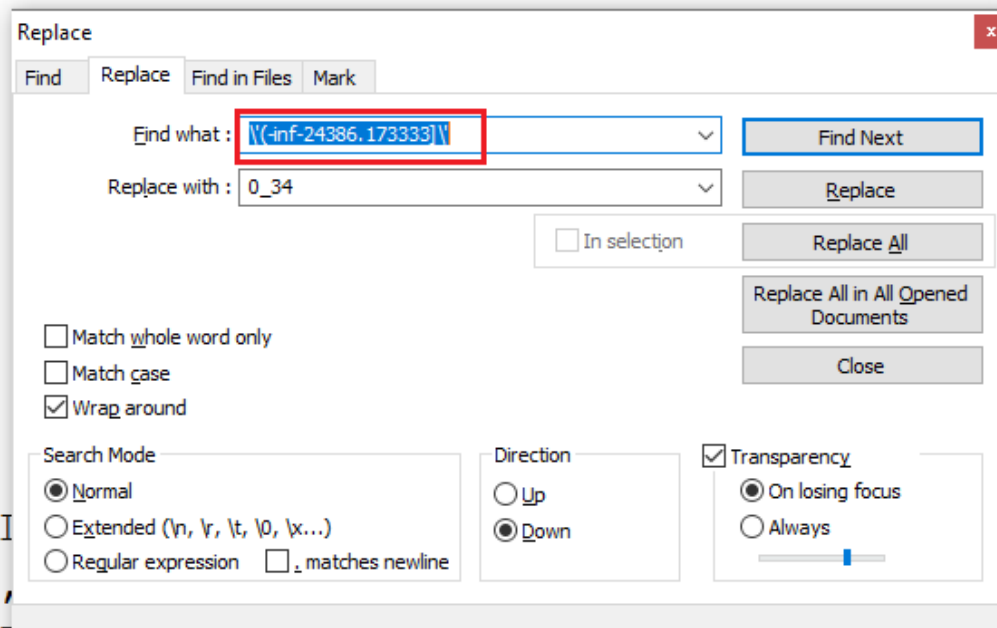
Chỉnh sửa dữ liệu

❖ Để các nhãn dễ hiểu hơn:

- ❖ Mở file “inputdata3.arff” với Notepad
- ❖ Với thuộc tính Age: Thay nhãn “(-inf-34.333333]” bằng 0_34, nhãn “(34.333333-50.666667]” bằng 35_51 và nhãn “(50.666667- inf)” bằng 52-max
- ❖ Với thuộc tính Age: Thay nhãn (-inf-24386.173333] bằng 0_24386, nhãn (24386.173333-43758.136667] bằng 24387_43758 và nhãn (43758.136667-inf) bằng 43759_max

Chỉnh sửa dữ liệu

```
{'\ '(-inf-34.333333]\'', '\ '(34.333333-50.666667]\'', '\ '(50.666667-inf)\''}  
{FEMALE,MALE}  
on {INNER_CITY,TOWN,RURAL,SUBURBAN}  
ne {'\ '(-inf-24386.173333]\'', '\ '(24386.173333-43758.136667]\'', '\ '(43758.13  
ied {NO,YES}  
dren {0,1,2,3}  
{NO,YES}  
_act {NO,YES}  
ent_act {NO,YES}  
age {NO,YES}  
{YES,NO}  
  
0.666667]\'', FEMALE  
0.666667]\'', MALE,  
nf)\'', FEMALE, INNER_CITY, '\ '(-inf-24386.173333]\'', YES, YES,  
YES, YES, YES, NO, NO
```



❖ Lưu lại file dữ liệu cuối cùng có tên “**intputdata4.arff**”