



CHUẨN BỊ DỮ LIỆU TRAIN TEST

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

Hà Nội, 2021

Nội dung

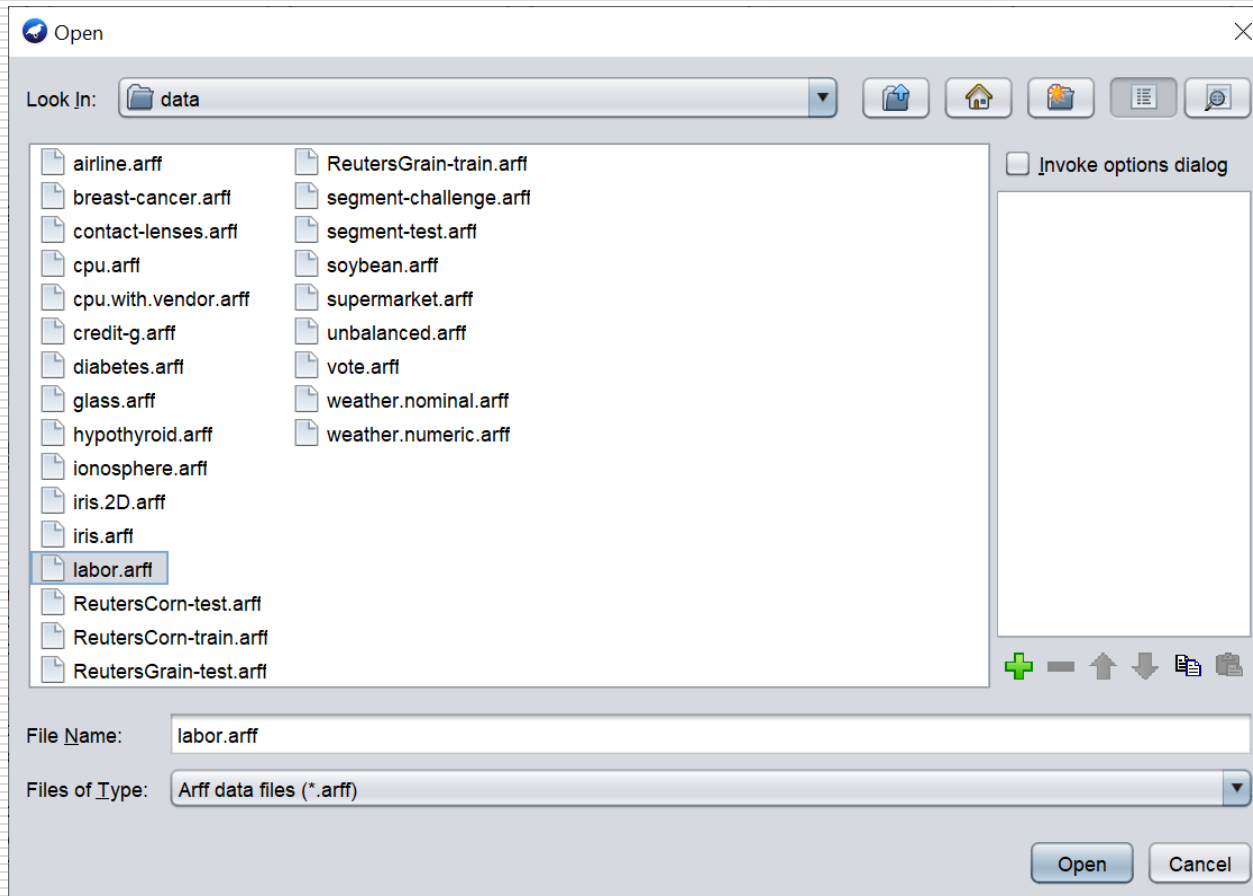
- ❖ Dùng bộ lọc RemovePercentage
- ❖ Dùng bộ lọc resample

Dùng bộ lọc RemovePercentage

- ❖ Mở file labor.arff
- ❖ Chọn bộ lọc RemovePercentage
- ❖ Thiết lập tham số
- ❖ Tạo dữ liệu huấn luyện
- ❖ Tạo dữ liệu thử nghiệm

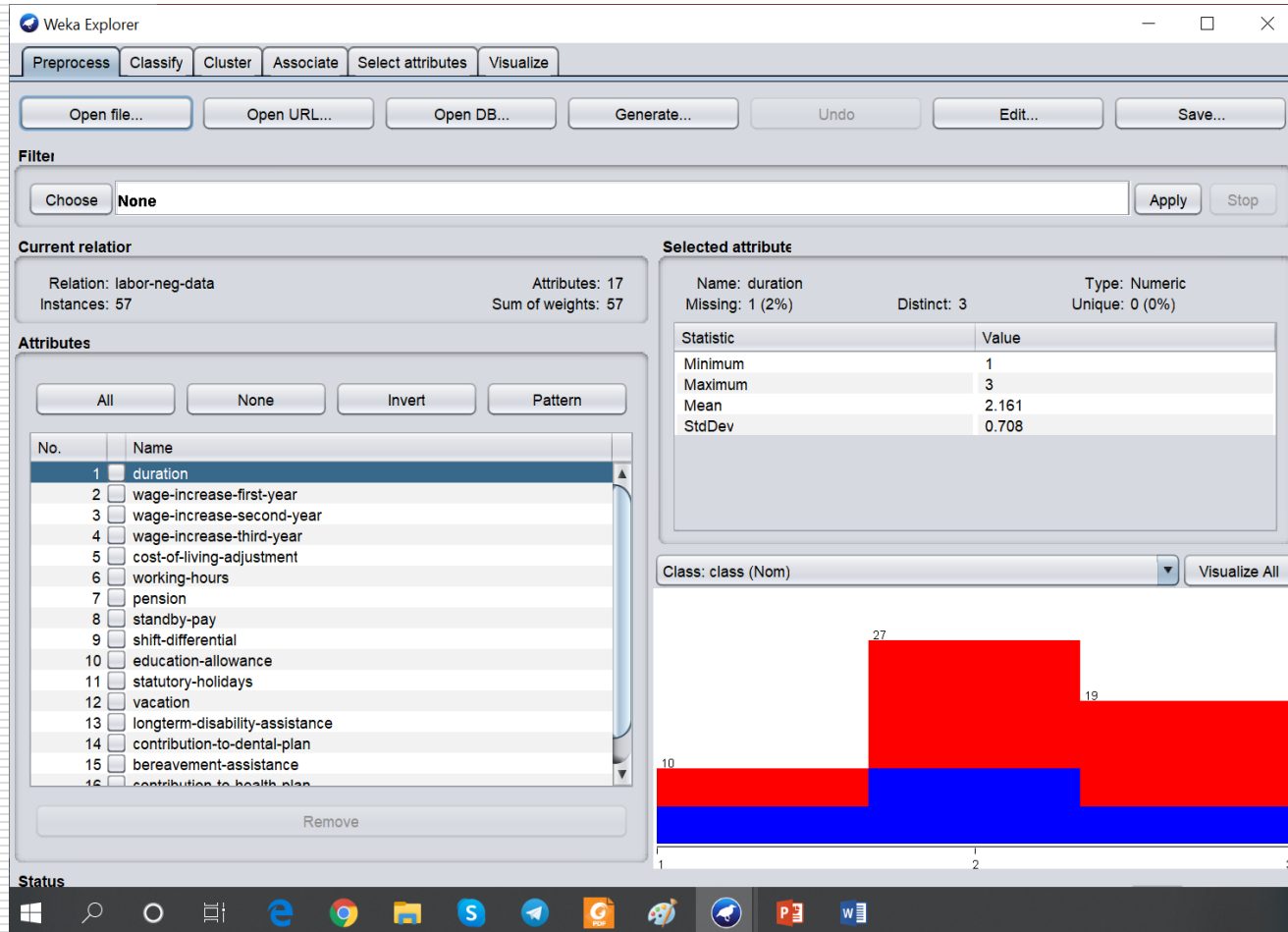
Mở file labor.arff

- ❖ Trong Tab Preprocessing, chọn Open File
- ❖ Chọn file labor.arff



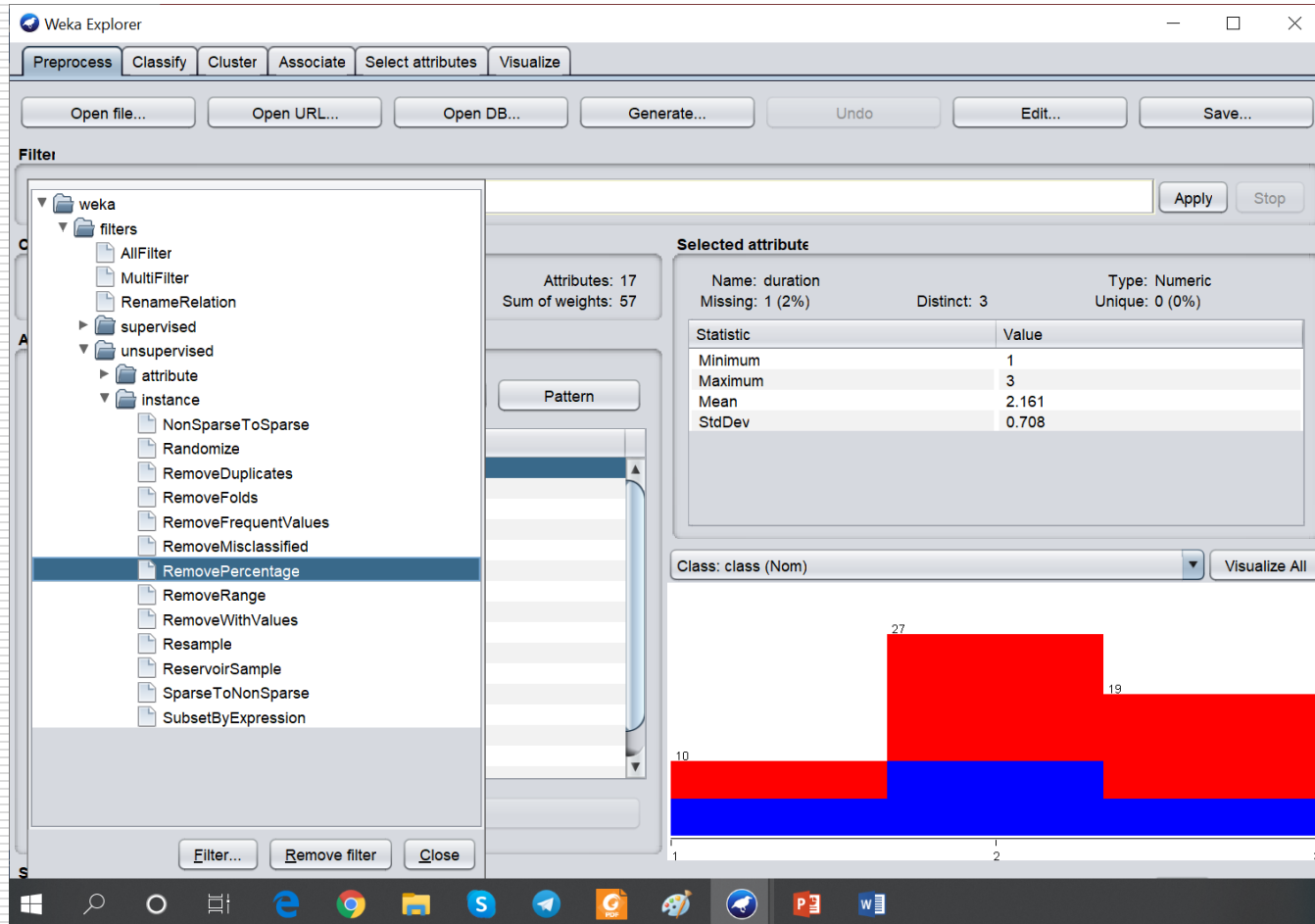
Mở file labor.arff

❖ Kết quả:



Chọn bộ lọc RemovePercentage

❖ Nhấn nút Choose, chọn đến RemovePercentage



Thiết lập tham số

- ❖ Nhấn vào hộp đóng khung bởi chữ nhật đỏ để mở thiết lập:

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. In the 'Filter' section, the 'RemovePercentage -P 50.0' filter is applied to the 'duration' attribute. The 'Current relation' section shows 'Relation: labor-neg-data' with 'Instances: 57' and 'Attributes: 17'. The 'Selected attribute' section shows 'Name: duration' with 'Missing: 1 (2%)', 'Distinct: 3', and 'Type: Numeric'. The 'Attributes' list on the left shows 'duration' selected. The 'Visualize All' button is visible. The bottom status bar shows 'Status'.

Weka Explorer

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

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

Filter

Choose **RemovePercentage -P 50.0** Apply Stop

Current relation

Relation: labor-neg-data
Instances: 57
Attributes: 17
Sum of weights: 57

Selected attribute

Name: duration
Missing: 1 (2%)
Distinct: 3
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	3
Mean	2.161
StdDev	0.708

Class: class (Nom) Visualize All

Attributes

All | None | Invert | Pattern

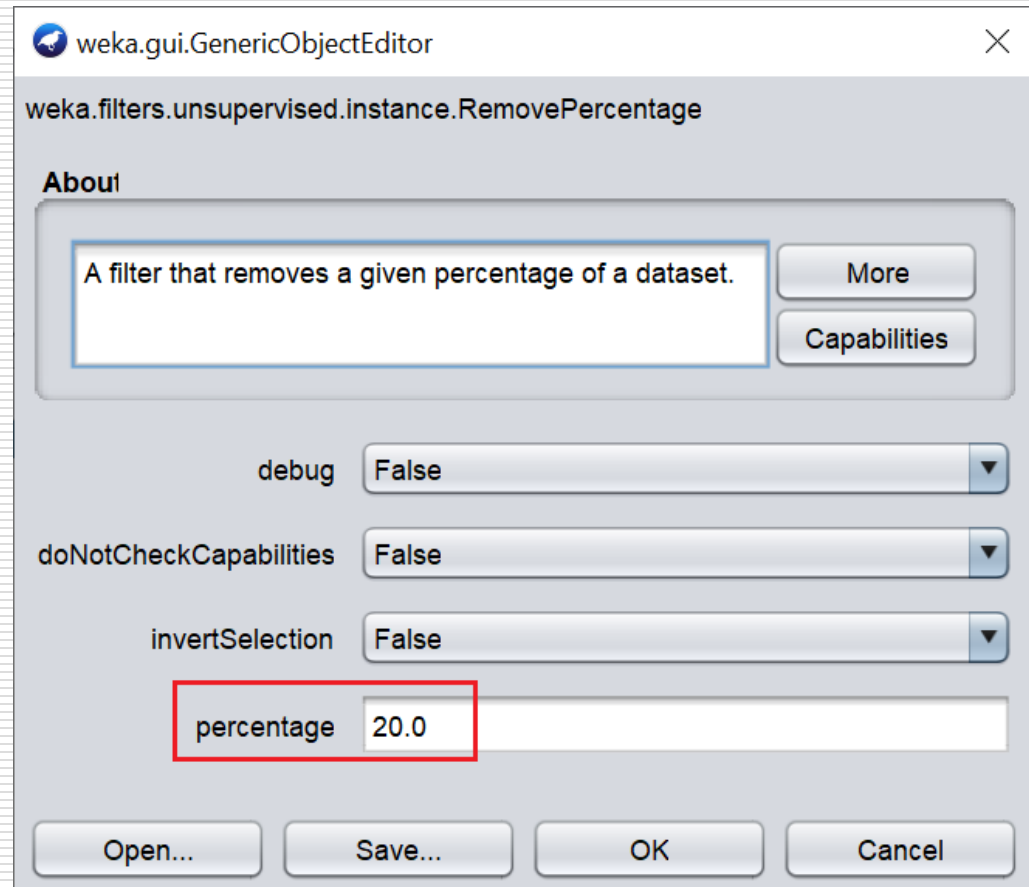
No.	Name
1	<input checked="" type="checkbox"/> duration
2	<input type="checkbox"/> wage-increase-first-year
3	<input type="checkbox"/> wage-increase-second-year
4	<input type="checkbox"/> wage-increase-third-year
5	<input type="checkbox"/> cost-of-living-adjustment
6	<input type="checkbox"/> working-hours
7	<input type="checkbox"/> pension
8	<input type="checkbox"/> standby-pay
9	<input type="checkbox"/> shift-differential
10	<input type="checkbox"/> education-allowance
11	<input type="checkbox"/> statutory-holidays
12	<input type="checkbox"/> vacation
13	<input type="checkbox"/> longterm-disability-assistance
14	<input type="checkbox"/> contribution-to-dental-plan
15	<input type="checkbox"/> bereavement-assistance
16	<input type="checkbox"/> contribution-to-health-plan

Remove

Status

Thiết lập tham số

- ❖ Nhập 20% (loại 20% dữ liệu gốc): Tập train là 80%
- ❖ insertSelection: True
- ❖ Nhấn OK



Tạo dữ liệu huấn luyện

❖ Nhấn Apply để lấy tập huấn luyện

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. In the 'Filter' section, 'RemovePercentage -P 20.0' is chosen, and the 'Apply' button is highlighted with a red box. The 'Current relation' section shows 'Relation: labor-neg-data-weka.filters.unsupervised.instance...' with 17 attributes and 46 instances. The 'Attributes' list on the left includes 'duration', 'wage-increase-first-year', 'wage-increase-second-year', 'wage-increase-third-year', 'cost-of-living-adjustment', 'working-hours', 'pension', 'standby-pay', 'shift-differential', 'education-allowance', 'statutory-holidays', 'vacation', 'longterm-disability-assistance', 'contribution-to-dental-plan', 'bereavement-assistance', and 'contribution-to-health-plan'. The 'Selected attribute' section for 'duration' shows statistics: Minimum 1, Maximum 3, Mean 2.13, and StdDev 0.687. The 'Class' is set to 'class (Nom)'. A bar chart at the bottom shows the distribution of the 'duration' attribute across three classes (1, 2, 3), with counts 8, 24, and 14 respectively.

Weka Explorer

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

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

Filter

Choose **RemovePercentage -P 20.0** **Apply** Stop

Current relation

Relation: labor-neg-data-weka.filters.unsupervised.instance... Attributes: 17
Instances: 46 Sum of weights: 46

Attributes

All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> duration
2	<input type="checkbox"/> wage-increase-first-year
3	<input type="checkbox"/> wage-increase-second-year
4	<input type="checkbox"/> wage-increase-third-year
5	<input type="checkbox"/> cost-of-living-adjustment
6	<input type="checkbox"/> working-hours
7	<input type="checkbox"/> pension
8	<input type="checkbox"/> standby-pay
9	<input type="checkbox"/> shift-differential
10	<input type="checkbox"/> education-allowance
11	<input type="checkbox"/> statutory-holidays
12	<input type="checkbox"/> vacation
13	<input type="checkbox"/> longterm-disability-assistance
14	<input type="checkbox"/> contribution-to-dental-plan
15	<input type="checkbox"/> bereavement-assistance
16	<input type="checkbox"/> contribution-to-health-plan

Remove

Selected attribute

Name: duration
Missing: 0 (0%)
Distinct: 3
Type: Numeric
Unique: 0 (0%)

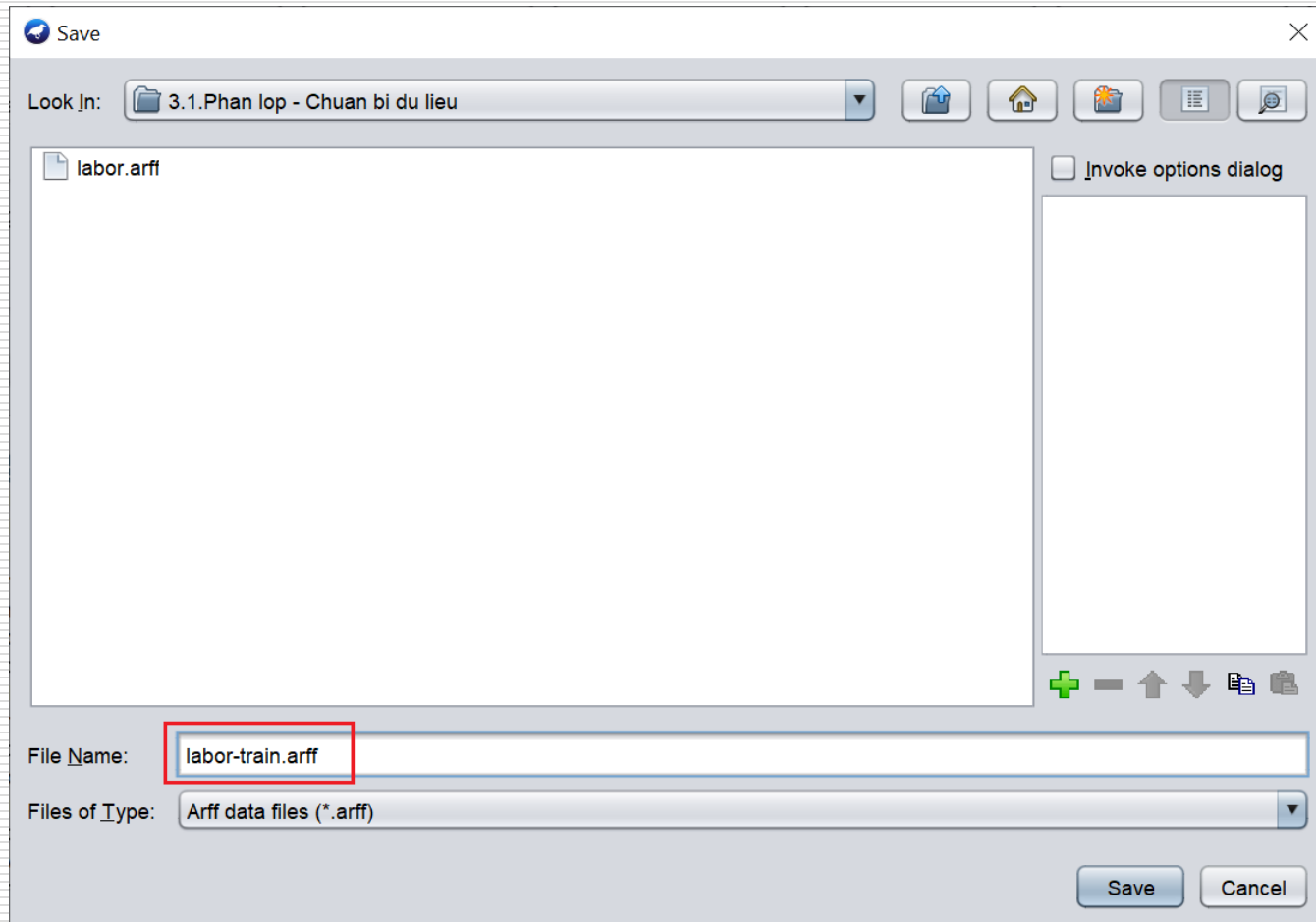
Statistic	Value
Minimum	1
Maximum	3
Mean	2.13
StdDev	0.687

Class: class (Nom) Visualize All

Status

Tạo dữ liệu huấn luyện

- ❖ Nhấn Save để mở hộp thoại để ghi ra file train



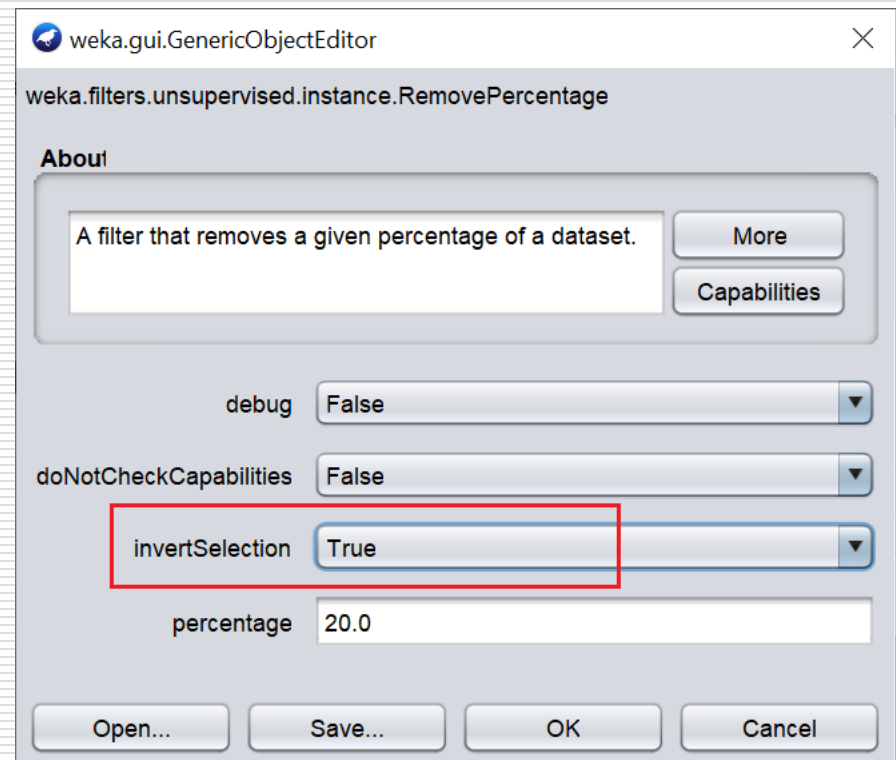
Tạo dữ liệu thử nghiệm

- ❖ Nhấn nút Undo để khôi phục lại tập bản ghi ban đầu

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. In the top toolbar, the 'Undo' button is highlighted with a red rectangle. Below the toolbar, the 'Filter' section shows 'RemovePercentage -P 20.0' selected. The 'Current relation' section displays 'Relation: labor-neg-data-weka.filters.unsupervised.instance...' and 'Instances: 46'. The 'Attributes' section on the left lists various attributes, with 'duration' selected. The 'Selected attribute' section on the right shows statistics for 'duration': Minimum 1, Maximum 3, Mean 2.13, and StdDev 0.687. At the bottom right, a bar chart visualizes the data distribution for the 'duration' attribute, showing three bars with heights 8, 24, and 14, colored red and blue.

Tạo dữ liệu thử nghiệm

- ❖ Tiếp tục nhấn hộp RemovePercentage (đóng khung hình chữ nhật đỏ) để mở hộp thoại thiết lập
- ❖ Chọn invertSelection là True (để lấy tập test gồm 20% dữ liệu ban đầu)
- ❖ Nhấn OK



Tạo dữ liệu thử nghiệm

- ❖ Nhấn Apply để lấy dữ liệu test
- ❖ Nhấn Save để ghi lại dữ liệu test

Weka Explorer

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

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

Filter

Choose **RemovePercentage -P 20.0 -V** **Apply** Stop

Current relation

Relation: labor-neg-data-weka.filters.unsupervise... Attributes: 17
Instances: 9 Sum of weights: 9

Attributes

All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> duration
2	<input type="checkbox"/> wage-increase-first-year
3	<input type="checkbox"/> wage-increase-second-year
4	<input type="checkbox"/> wage-increase-third-year
5	<input type="checkbox"/> cost-of-living-adjustment
6	<input type="checkbox"/> working-hours
7	<input type="checkbox"/> pension
8	<input type="checkbox"/> standby-pay

Remove

Selected attribute

Name: duration
Missing: 0 (0%)
Distinct: 3
Type: Numeric
Unique: 1 (11%)

Statistic	Value
Minimum	1
Maximum	3
Mean	1.667
StdDev	0.707

Class: class (Nom) Visualize All

Status

OK Log x 0

Dùng bộ lọc resample

- ❖ Các bước tương tự như với bộ lọc RemovePercentage

The screenshot shows the Weka Explorer interface. The 'Filter' tab is active, and the 'Resample' filter is selected in the 'filters' list. The 'Selected attribute' table shows the following statistics for the 'duration' attribute:

Statistic	Value
Minimum	1
Maximum	3
Mean	1.667
StdDev	0.707

Below the table, the 'Class: class (Nom)' is selected, and a bar chart visualizes the distribution. The chart shows two bars: a red bar for class 1 with a value of 8, and a blue bar for class 2 with a value of 1. The x-axis is labeled 1, 2, 3 and the y-axis is labeled 1, 2, 3, 4, 5, 6, 7, 8.

Dùng bộ lọc resample

- ❖ Các bước tương tự như với bộ lọc RemovePercentage
- ❖ Chỉ thay đổi bộ lọc và thiết lập thuộc tính

- ❖ Tính năng invertSelection hiện hoạt động không ổn định

