CHUẨN BỊ DỮ LIỆU TRAIN TEST

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

Hà Nội, 2021

Nội dung

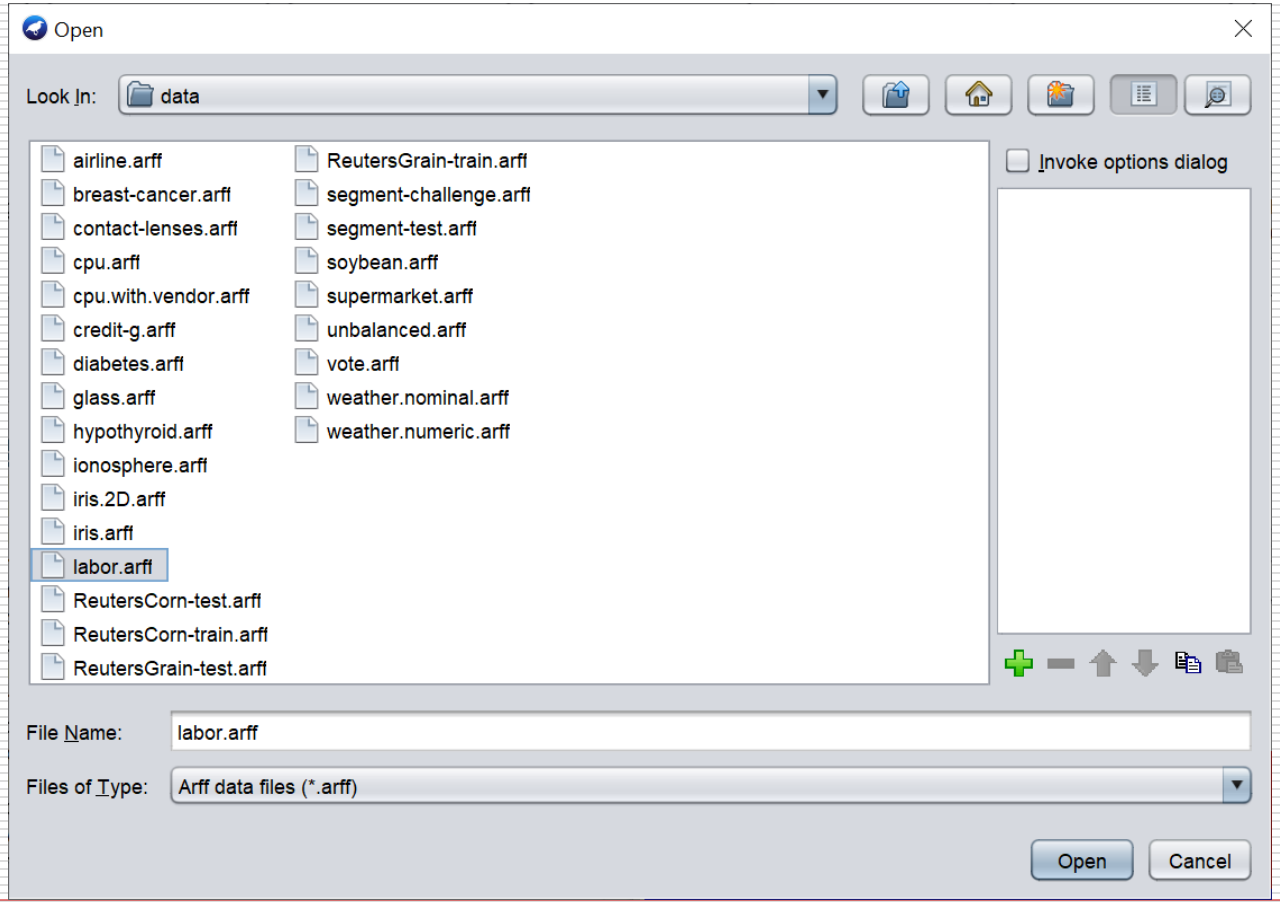
- Dùng bộ lọc RemovePercentage
- Dùng bộ lọc resample

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

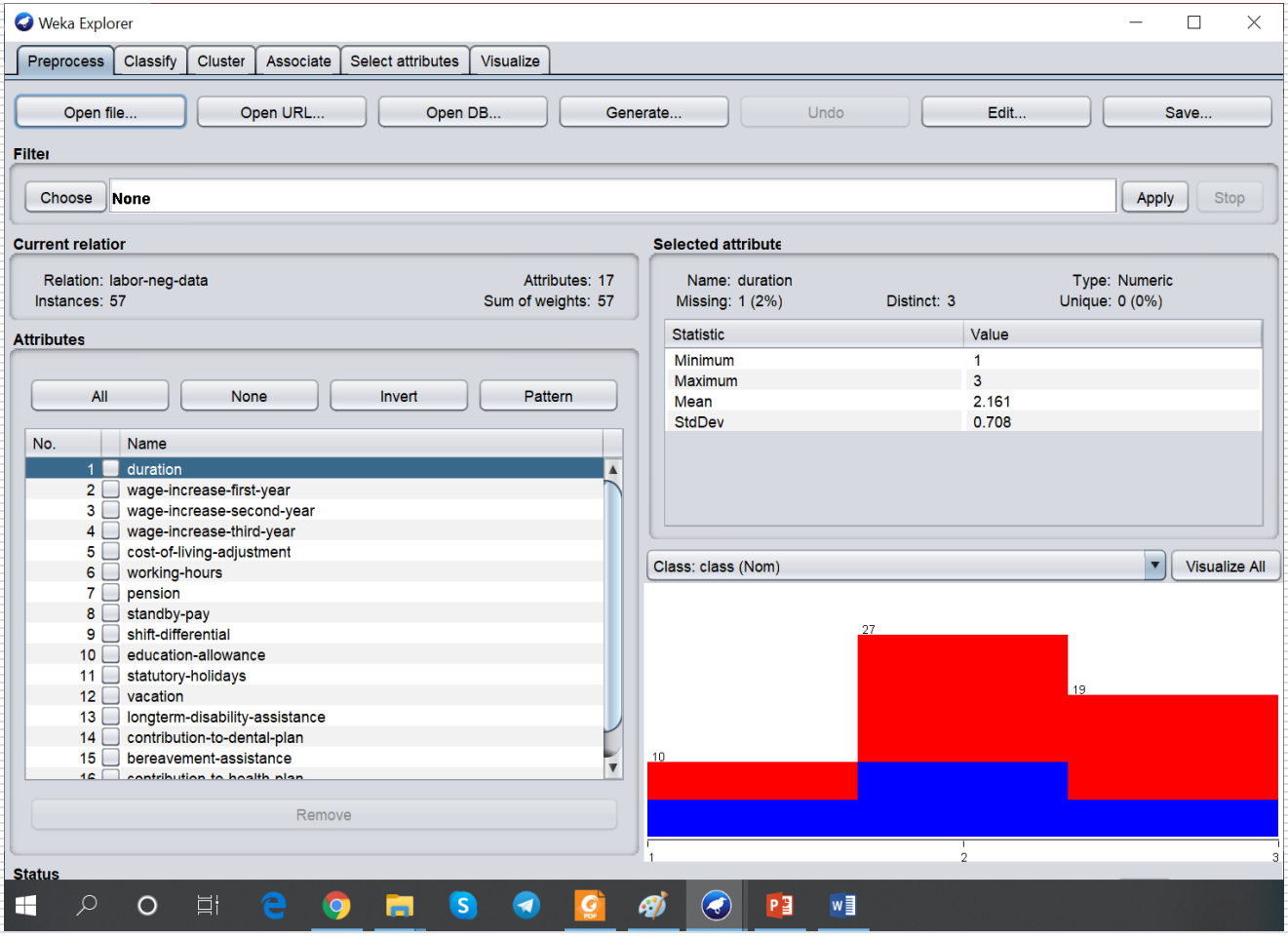
Mở file labor.arff

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



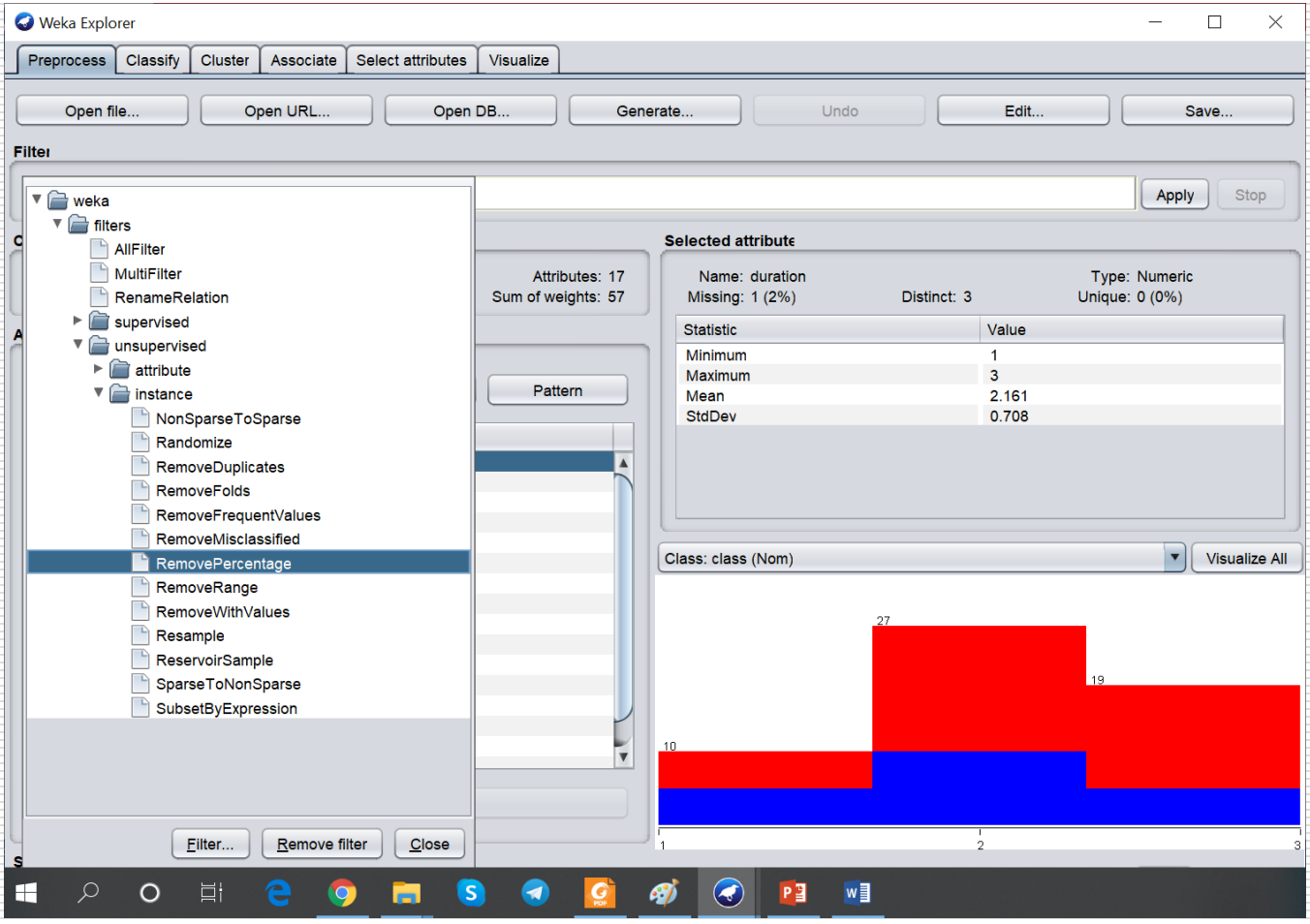
Mở file labor.arff

- Kết quả:



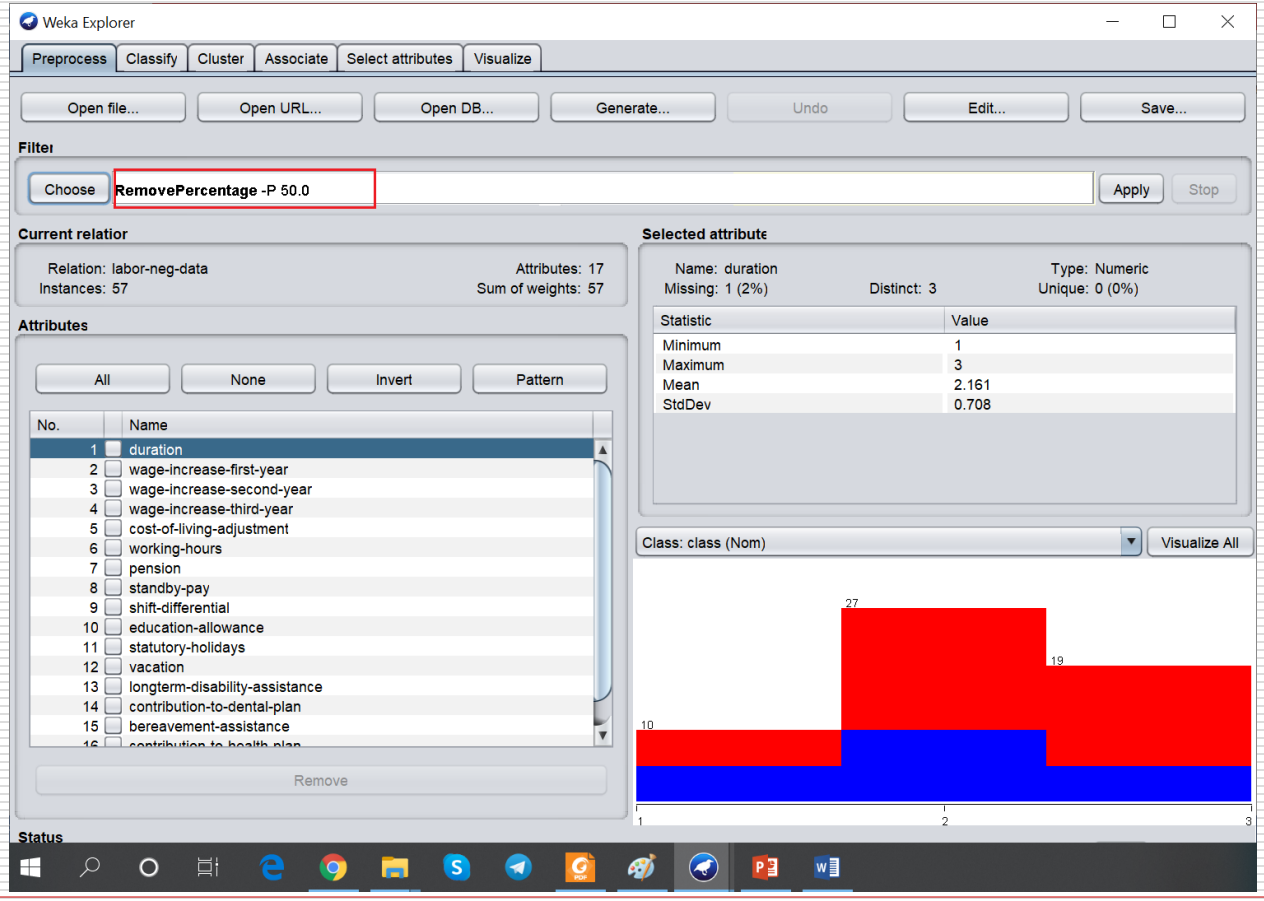
Chọn bộ lọc RemovePercentage

- Nhấn nút Choose, chọn đến RemovePercentage



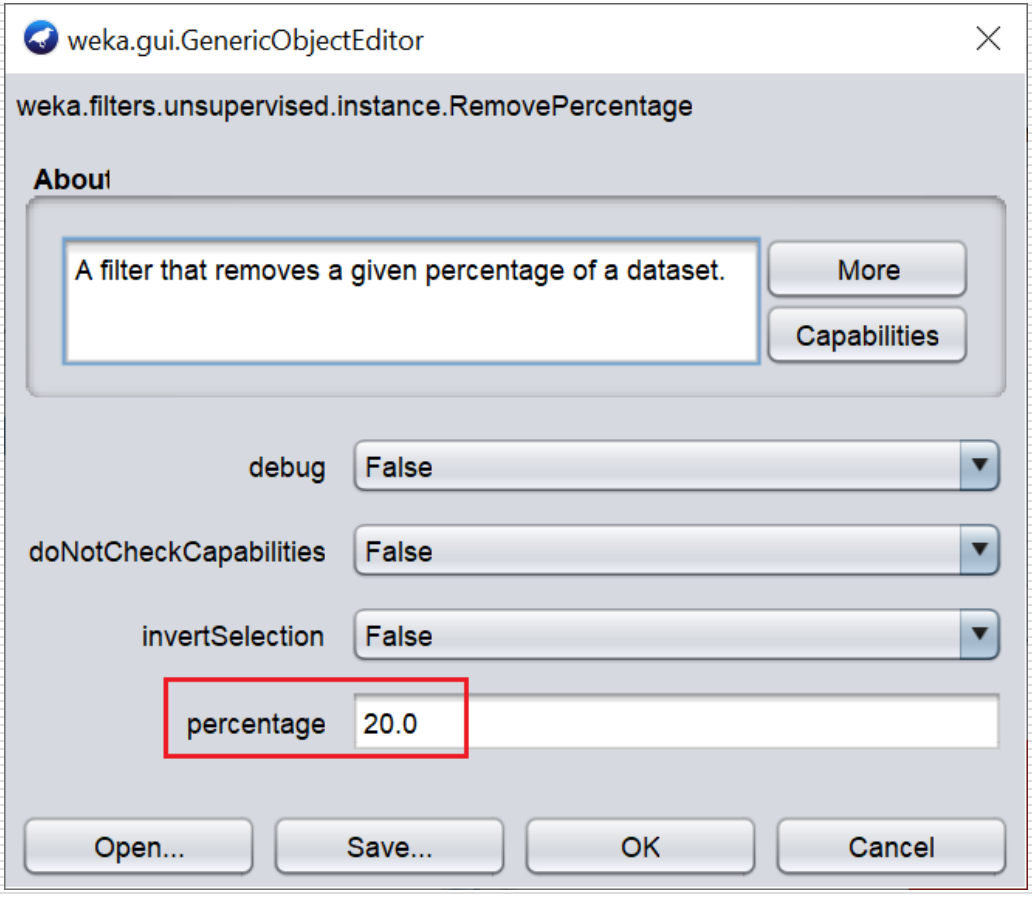
Thiết lập tham số

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



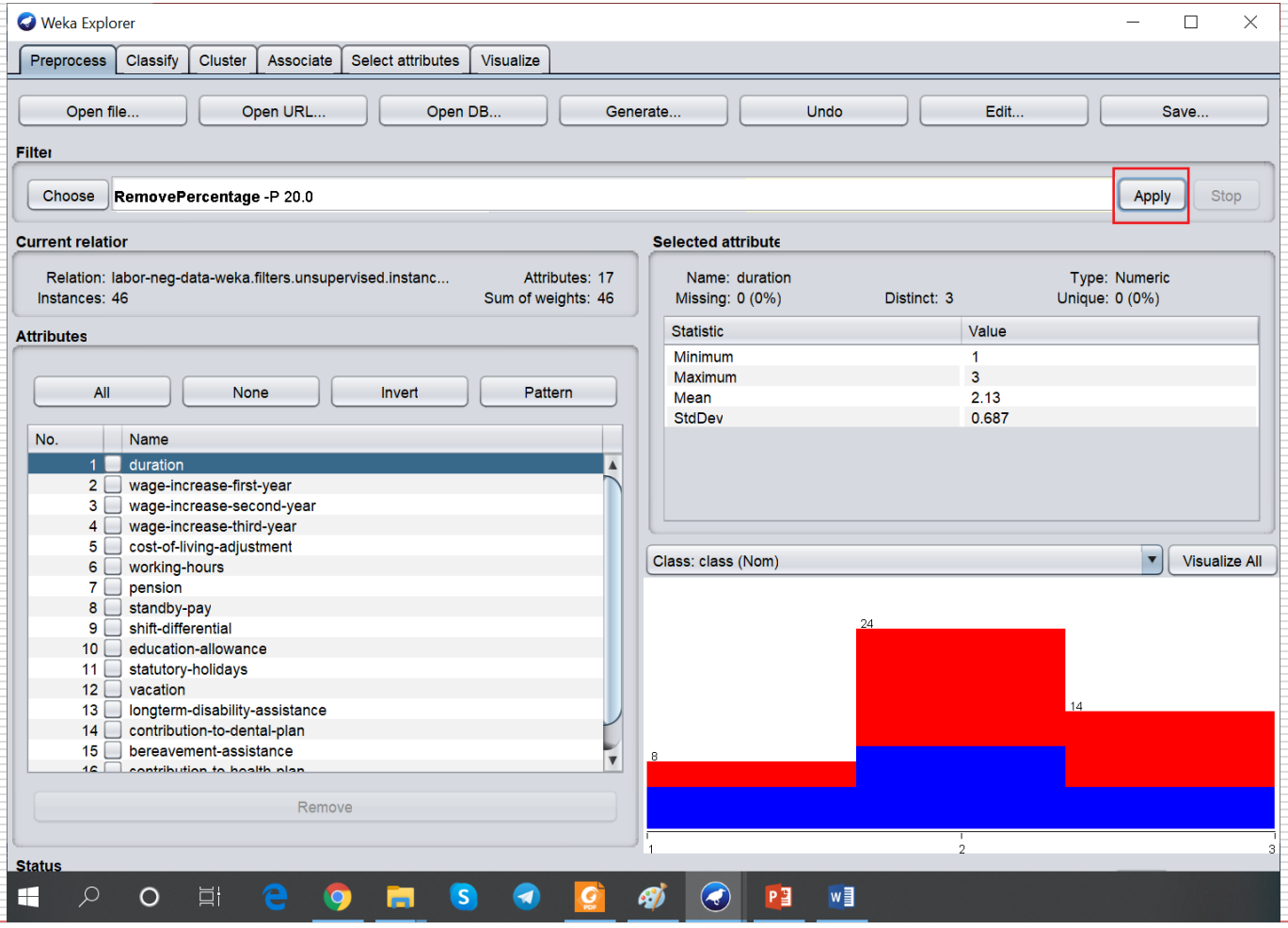
Thiết lập tham số

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



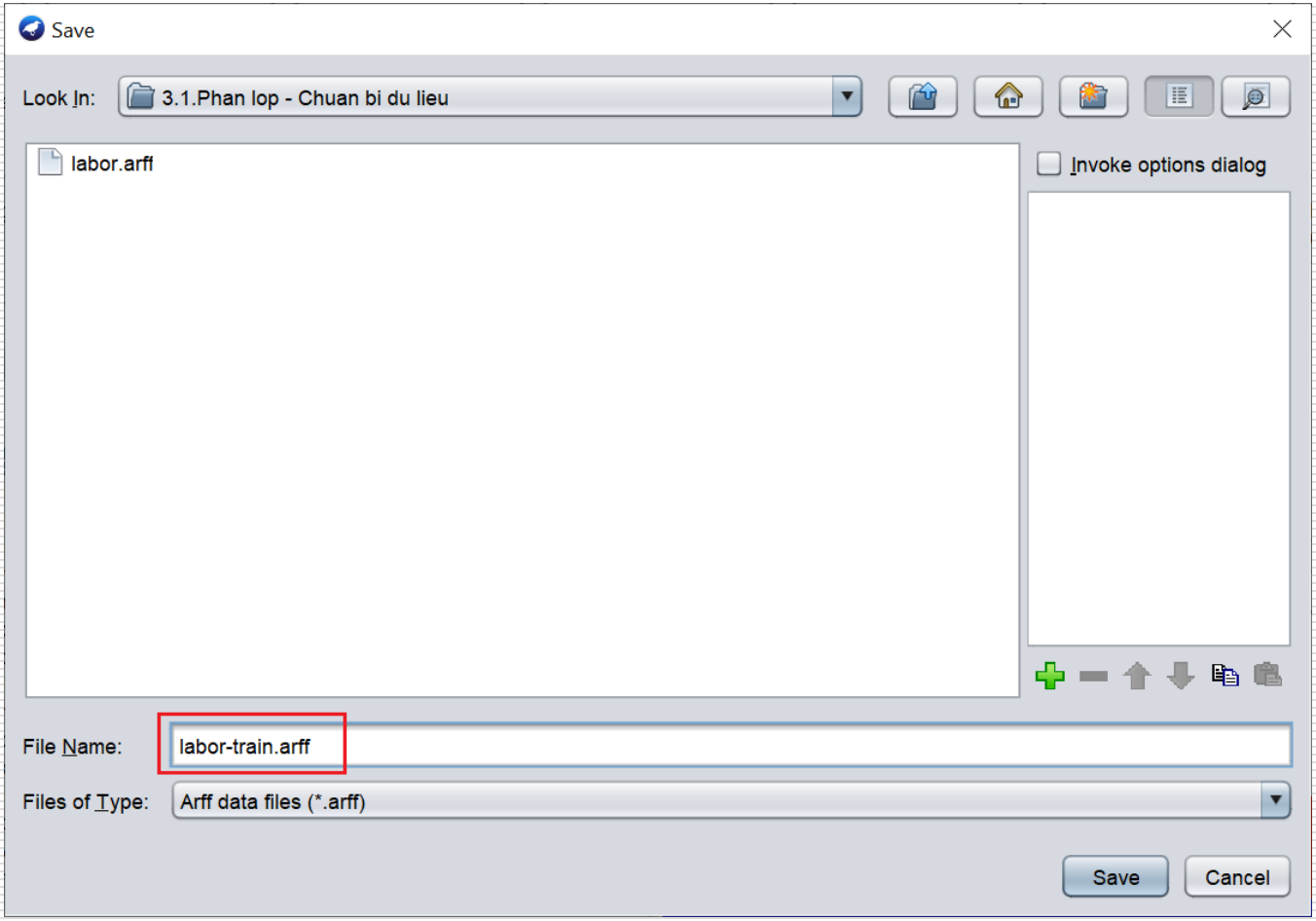
Tạo dữ liệu huấn luyện

- Nhấn Apply để lấy tập huấn luyện



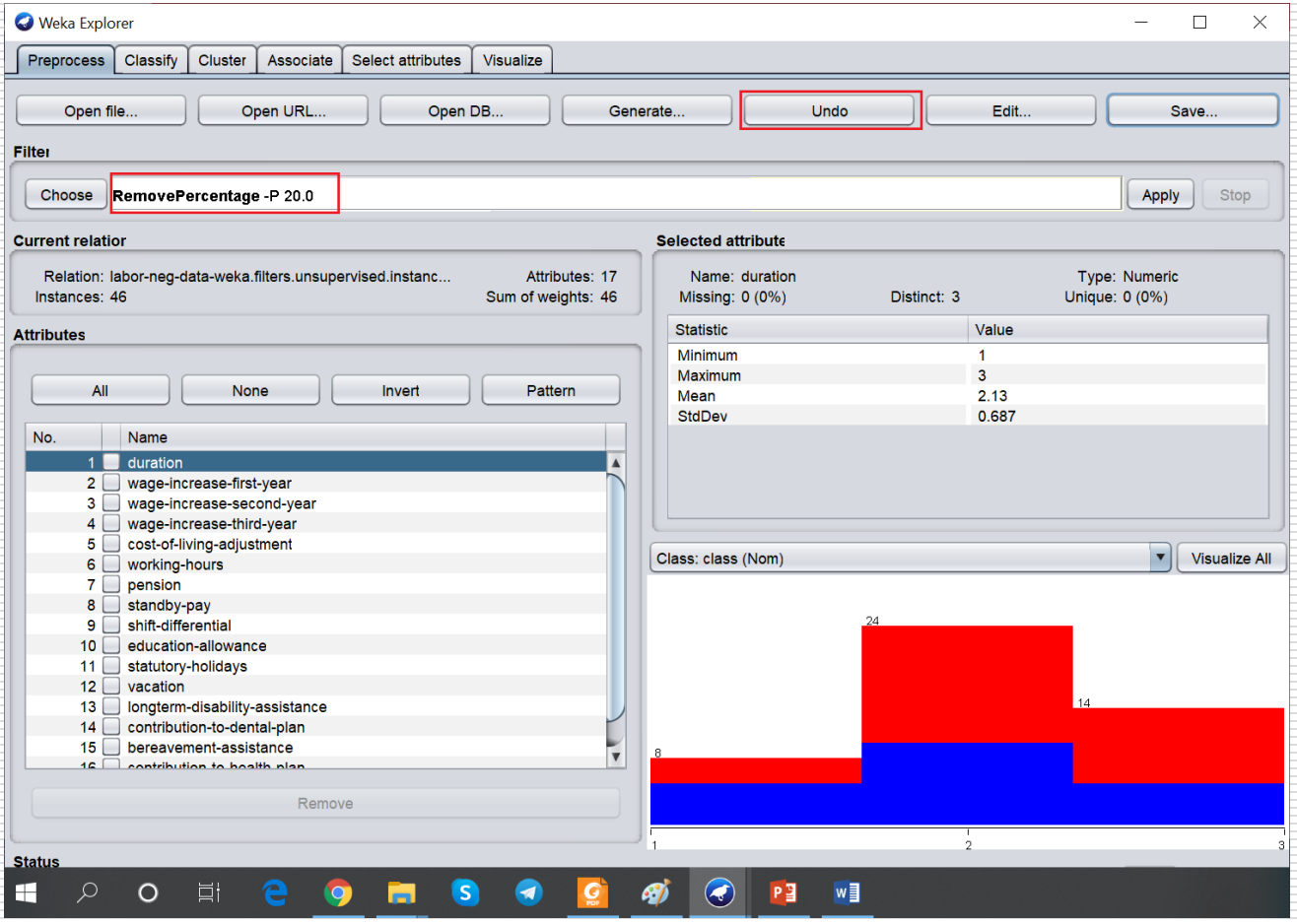
Tạo dữ liệu huấn luyện

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



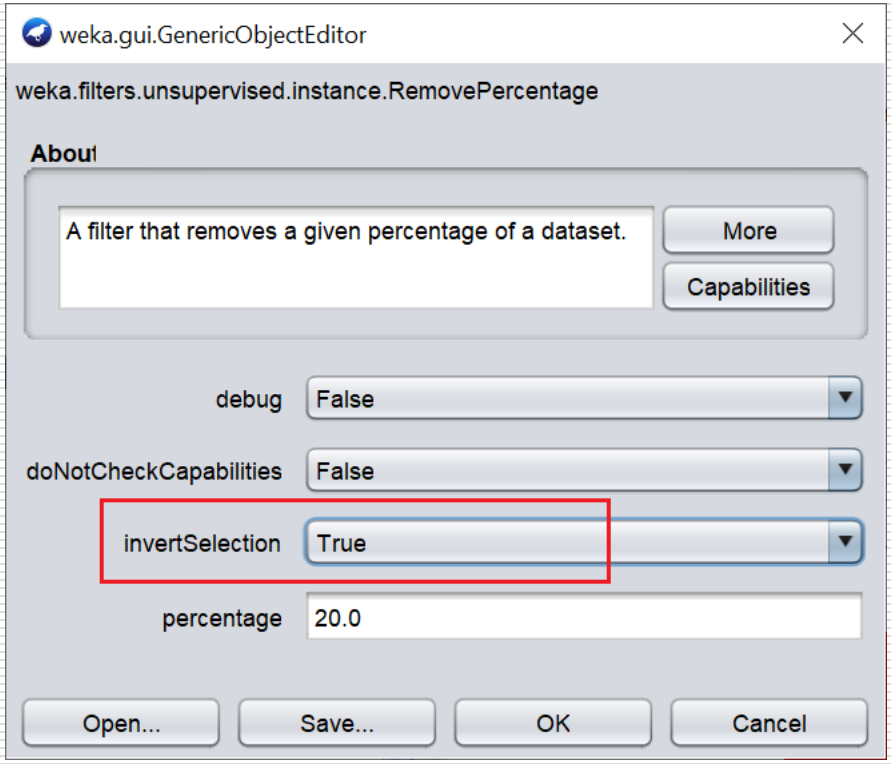
Tạo dữ liệu thử nghiệm

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



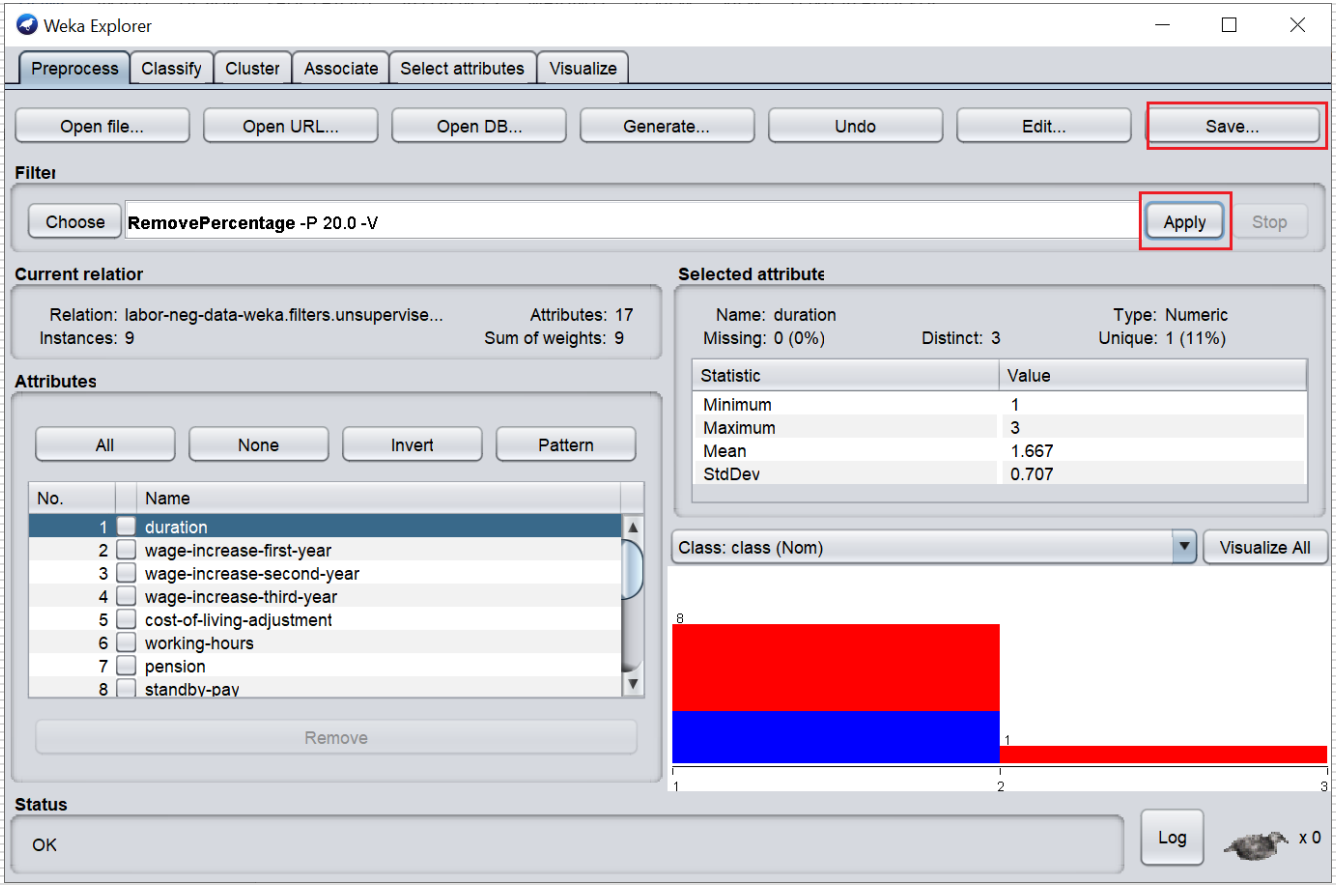
Tạo dữ liệu thử nghiệm

- 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



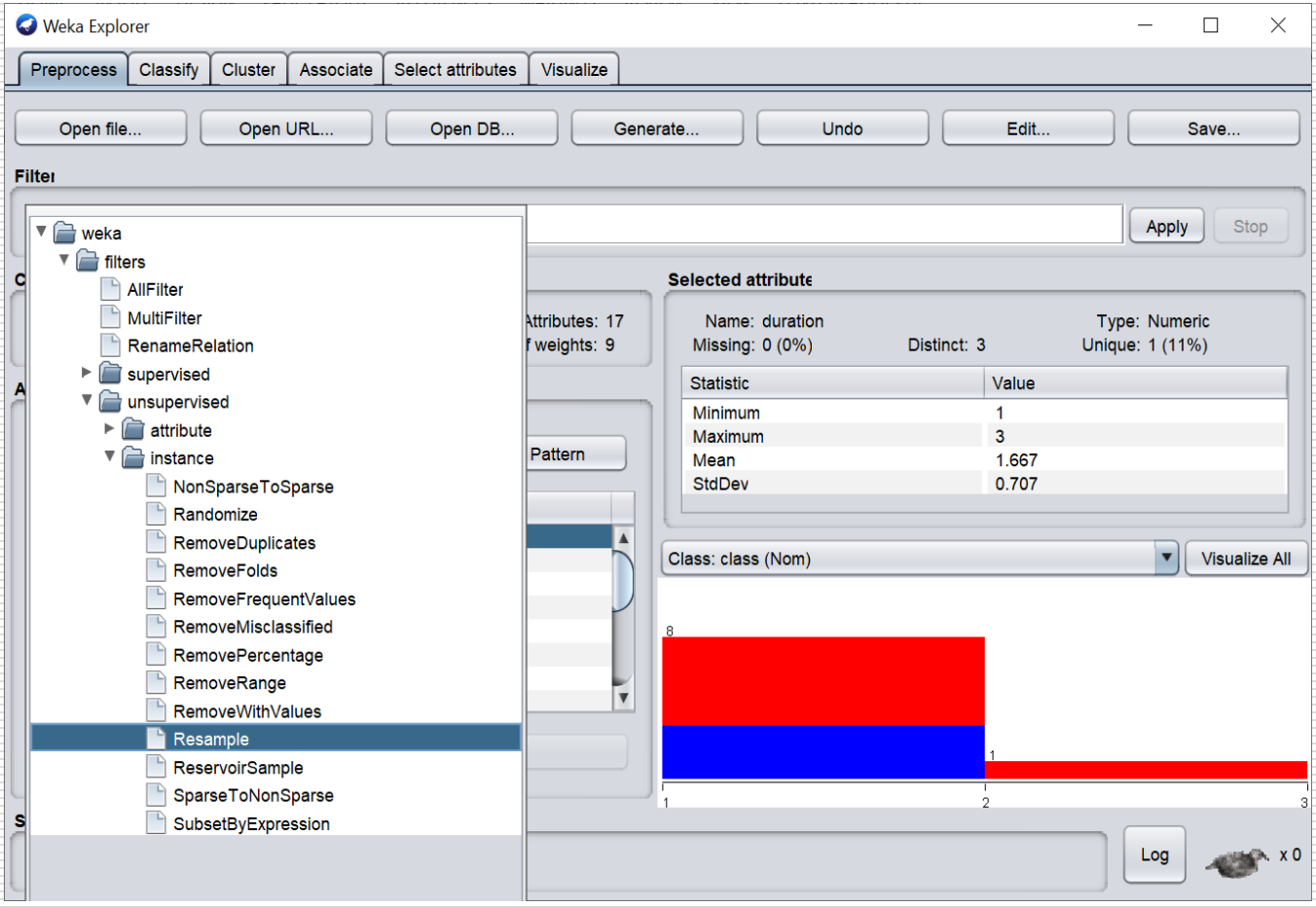
Tạo dữ liệu thử nghiệm

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



Dùng bộ lọc resample

- Các bước tương tự như với bộ lọc RemovePercentage



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