
# PHÂN CỤM DỮ LIỆU

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

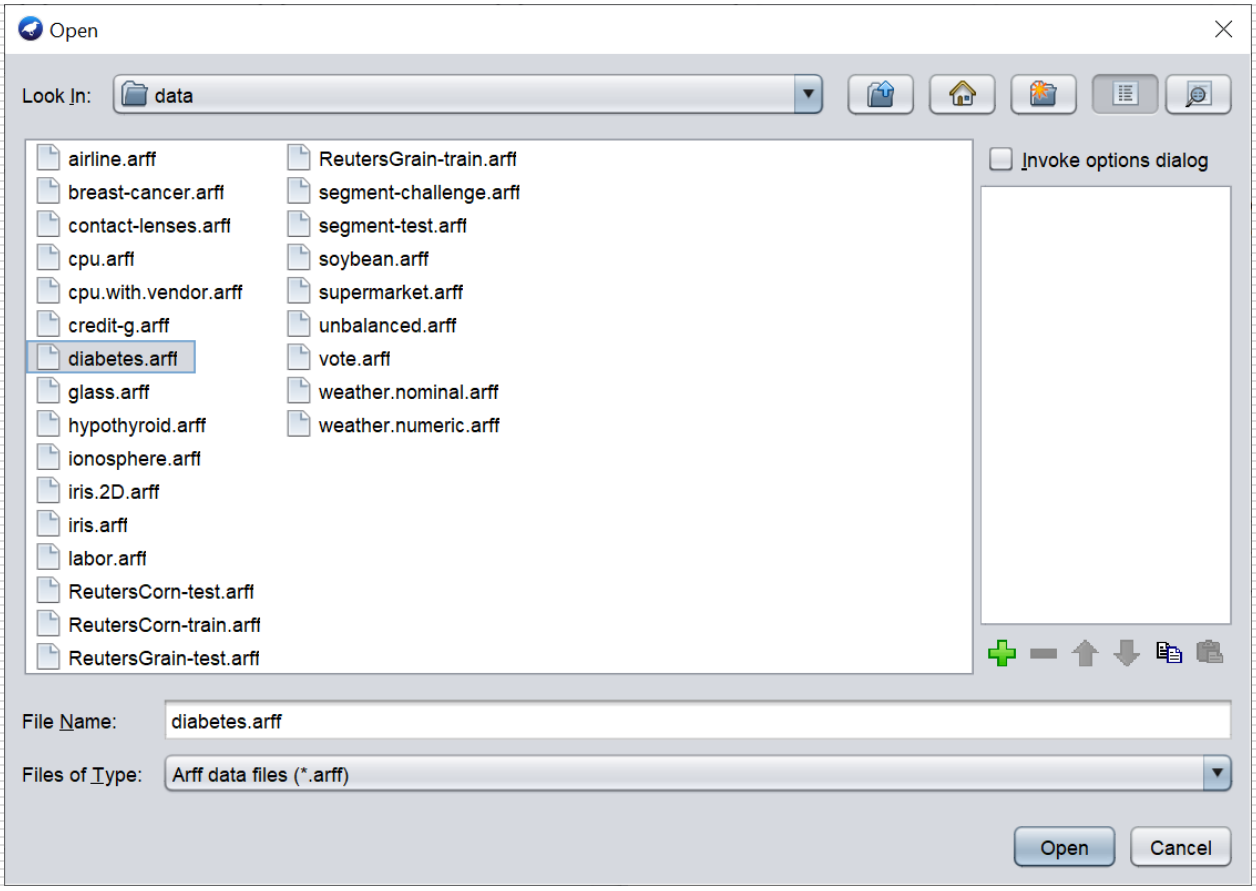
Hà Nội, 2021

# Nội dung

- Mở file diabetes.arff
- Loại bỏ thuộc tính class (phân lớp)
- Phân cụm k-Means
- So sánh kết quả phân cụm và trường class
- Phân cụm với thuật toán EM

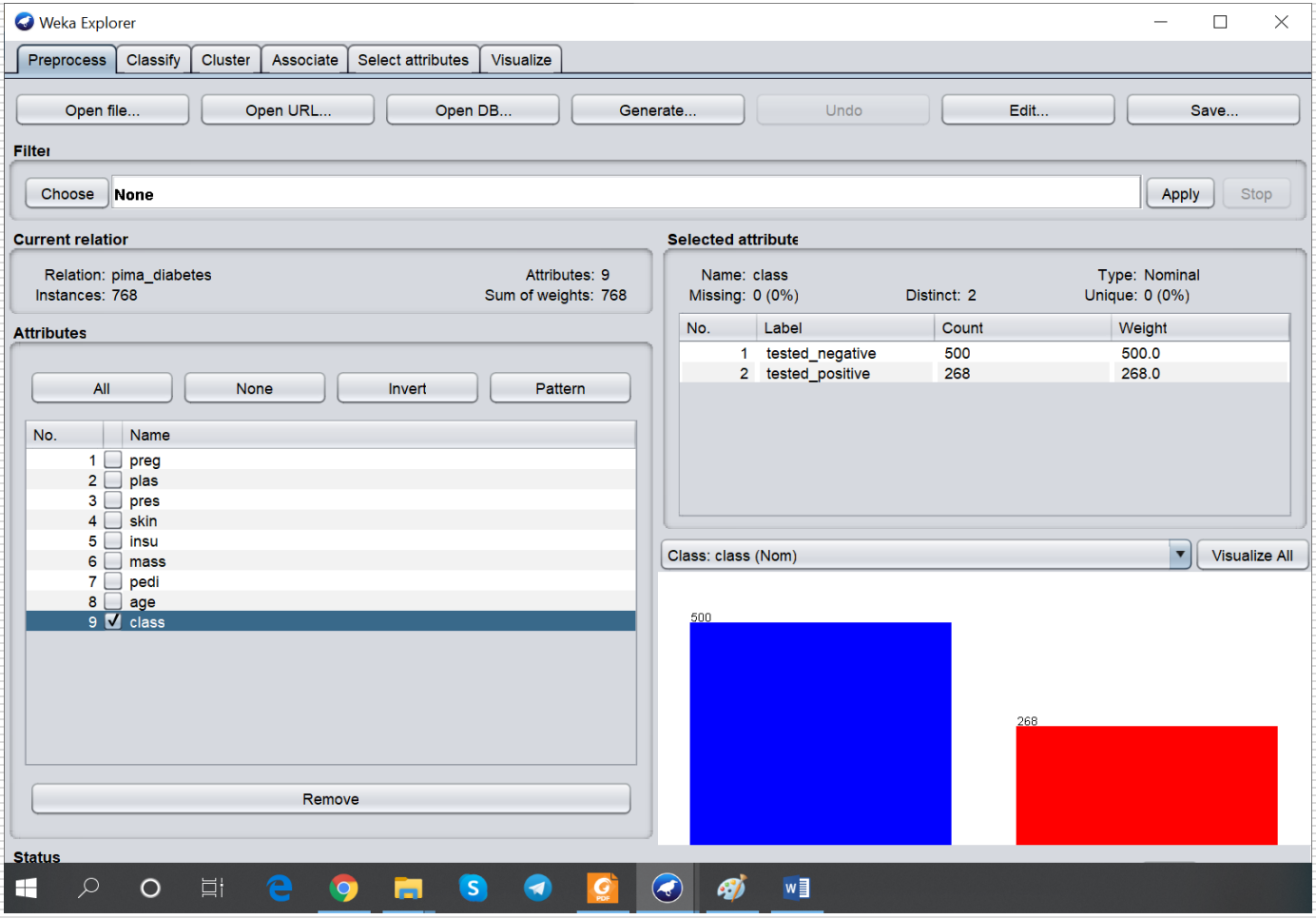
# Mở file diabetes.arff

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



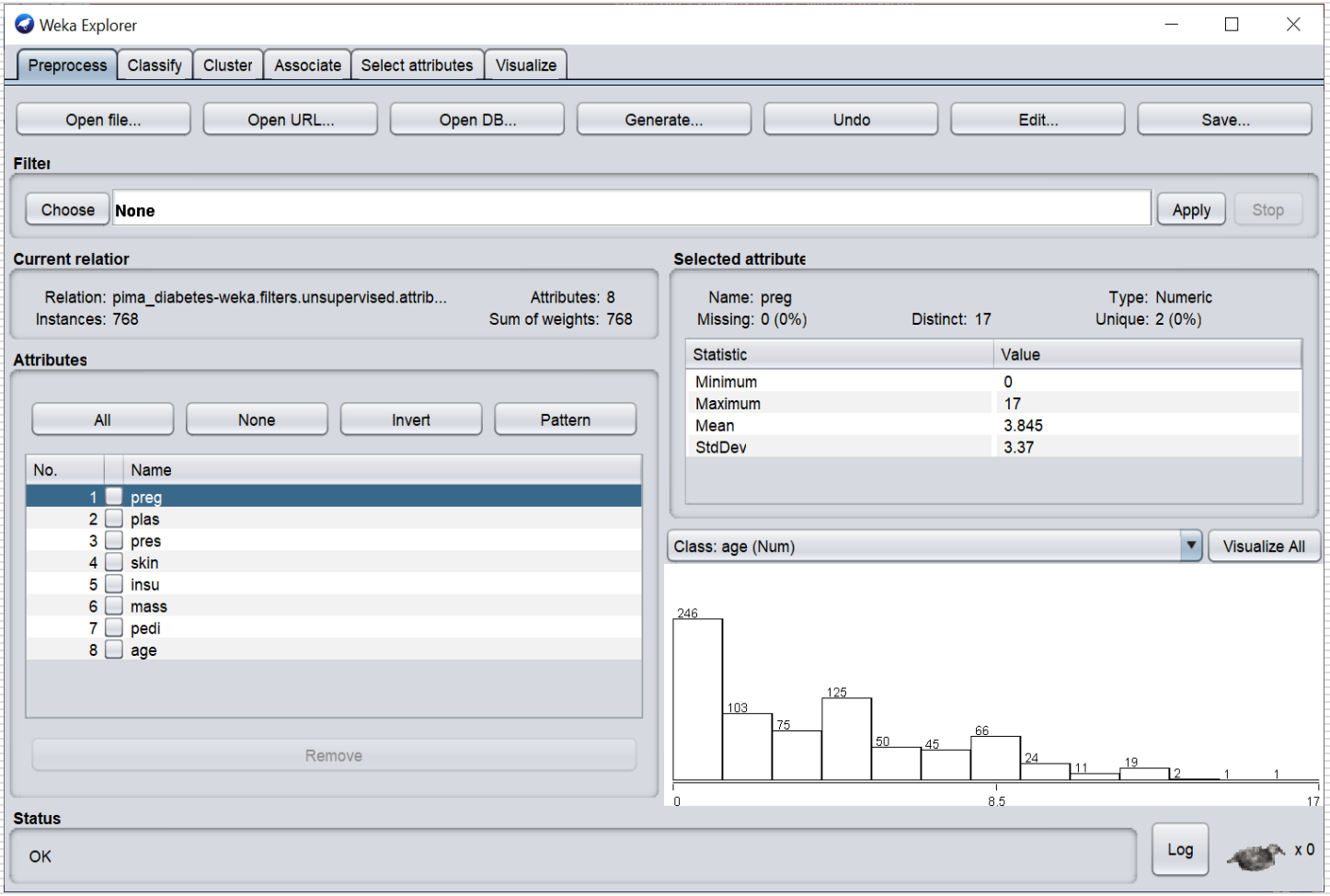
# Loại bỏ thuộc tính class (phân lớp)

- Tick chọn các trường class và nhấn Remove để loại bỏ:



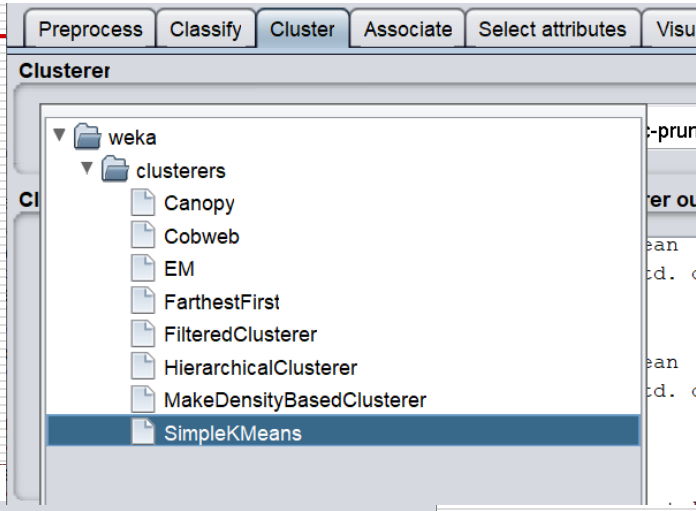
# Loại bỏ thuộc tính class (phân lớp)

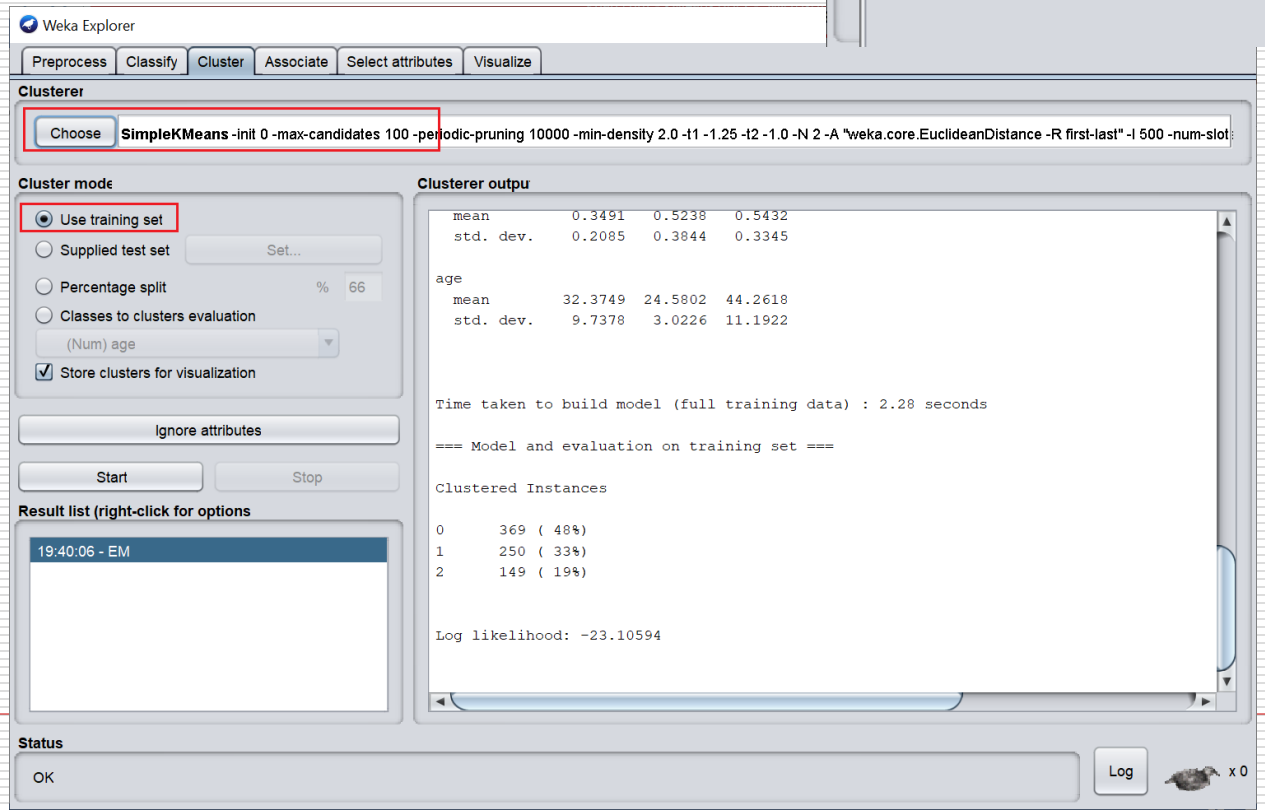
- Kết quả:



# Phân cụm k-Means

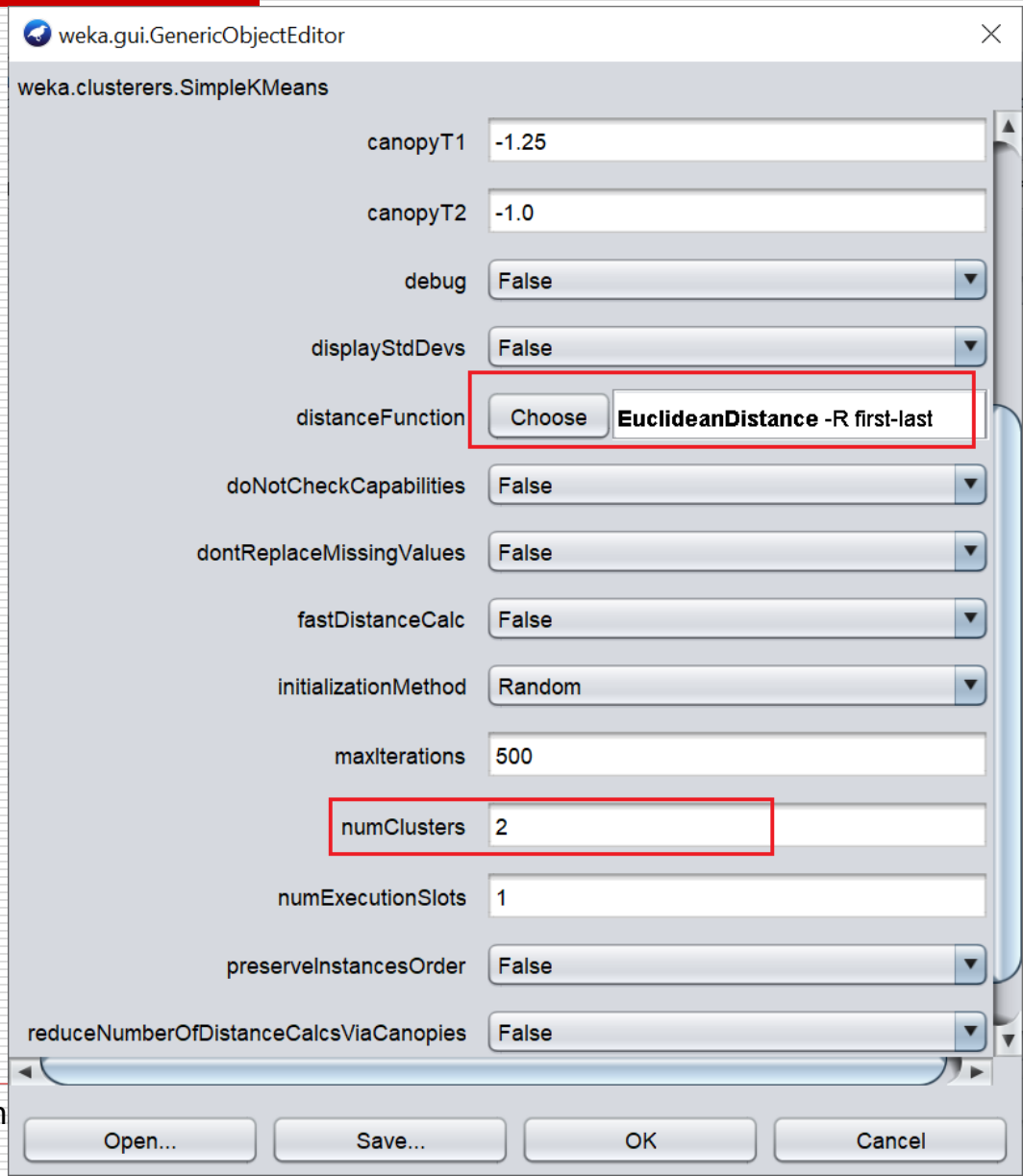
- Chuyển tab Cluster
- Nhấn nút Choose để chọn thuật toán
- Chọn "Use training set"



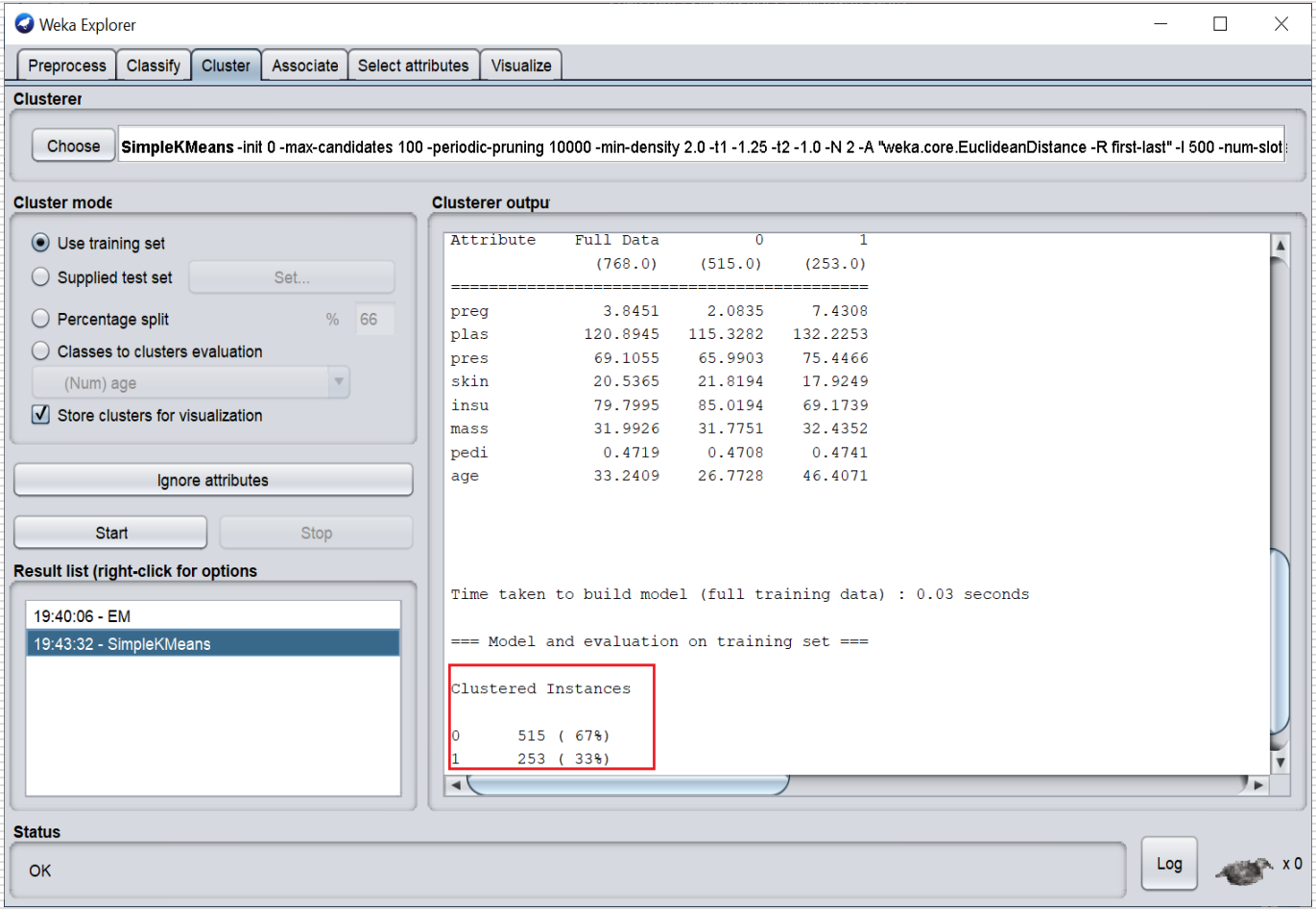
# Phân cụm k-Means

- Nhấn chuột trái vào Hộp thông tin thuật toán bên cạnh để mở cửa sổ cấu hình thuật toán
- Cấu hình tham số cho kMeans
- Nhấn OK để đóng cửa sổ



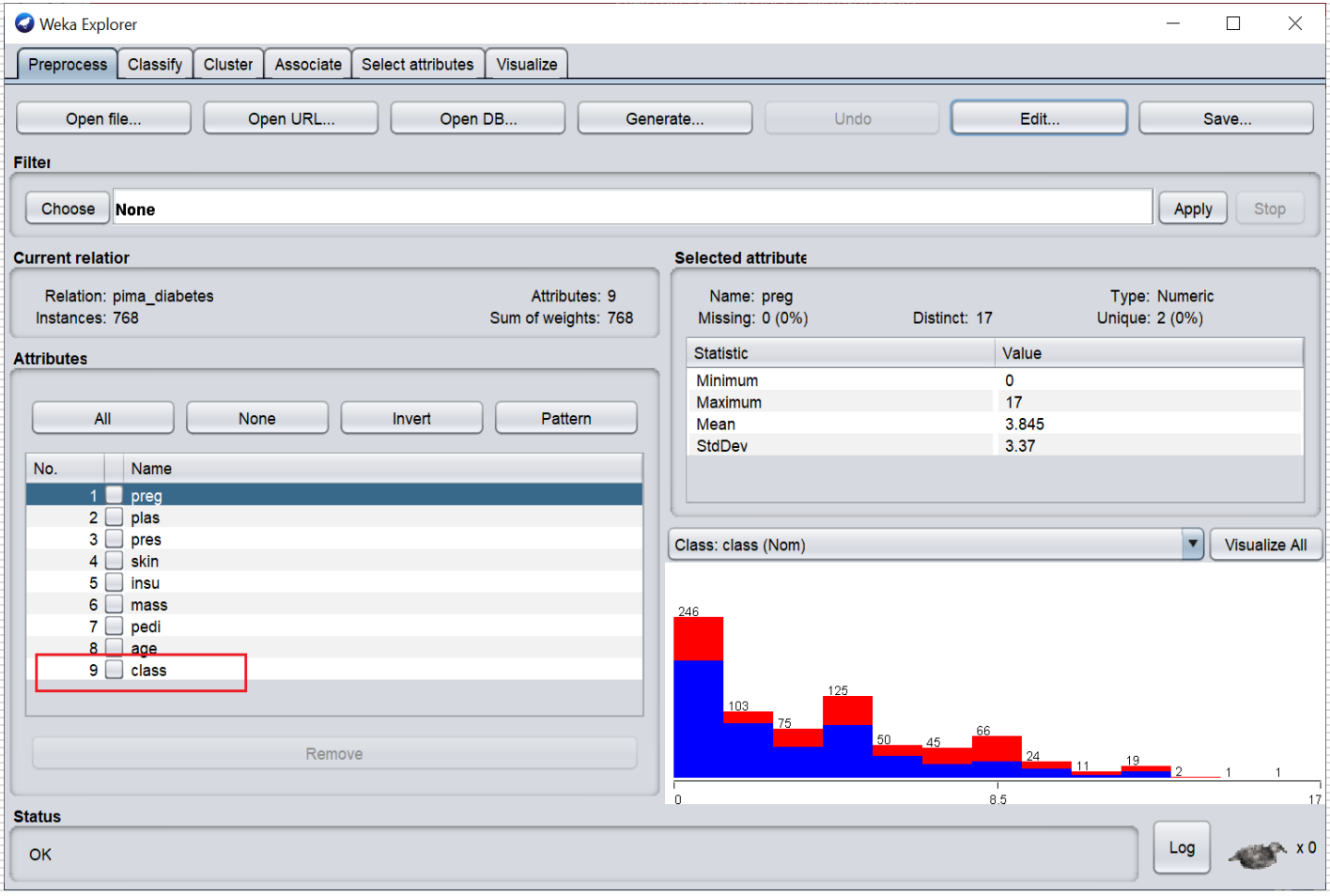
# Phân cụm k-Means

- Nhấn nút Start và xem kết quả:



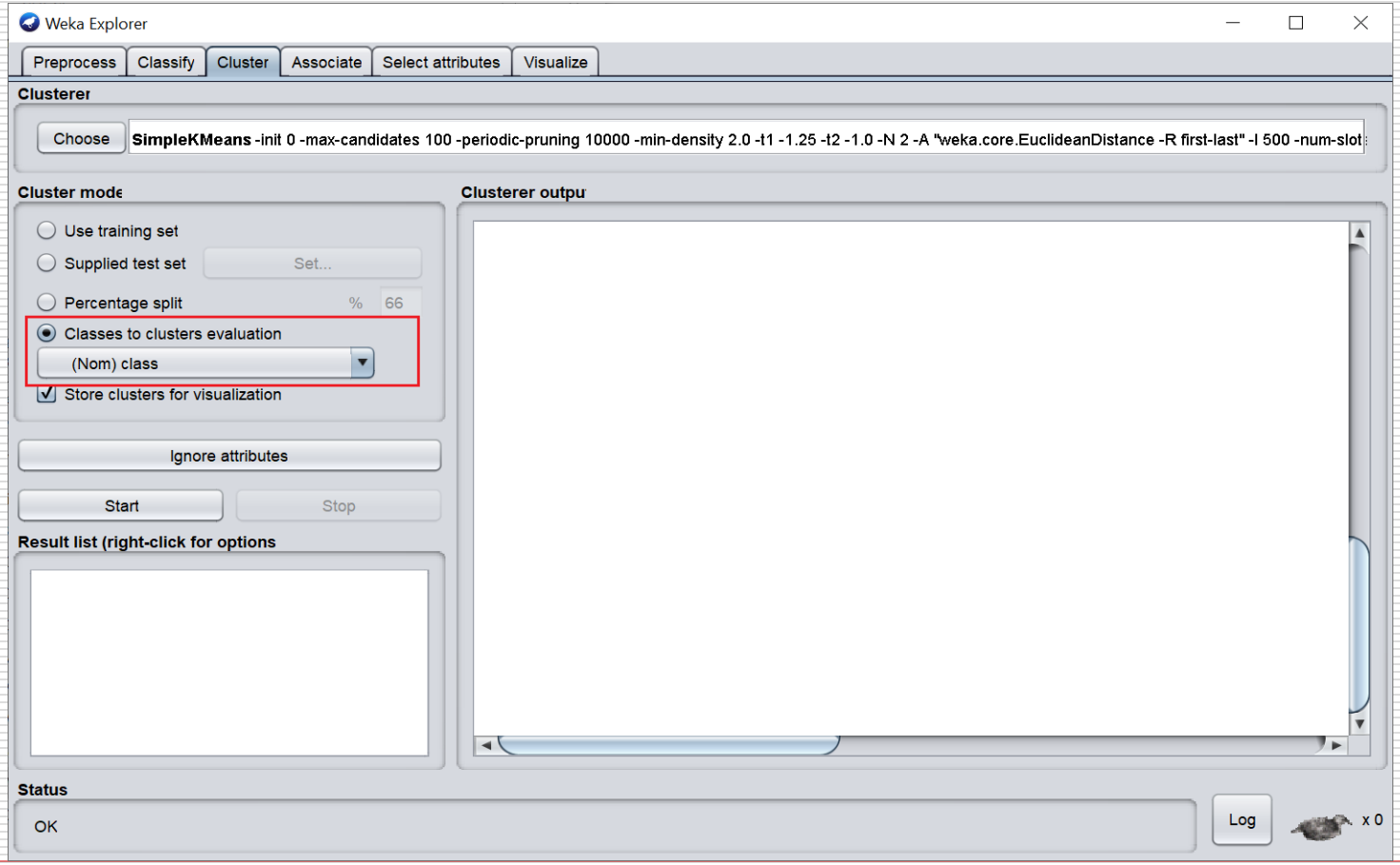
# So sánh kết quả phân cụm và trường class

- Giữ nguyên trường class:



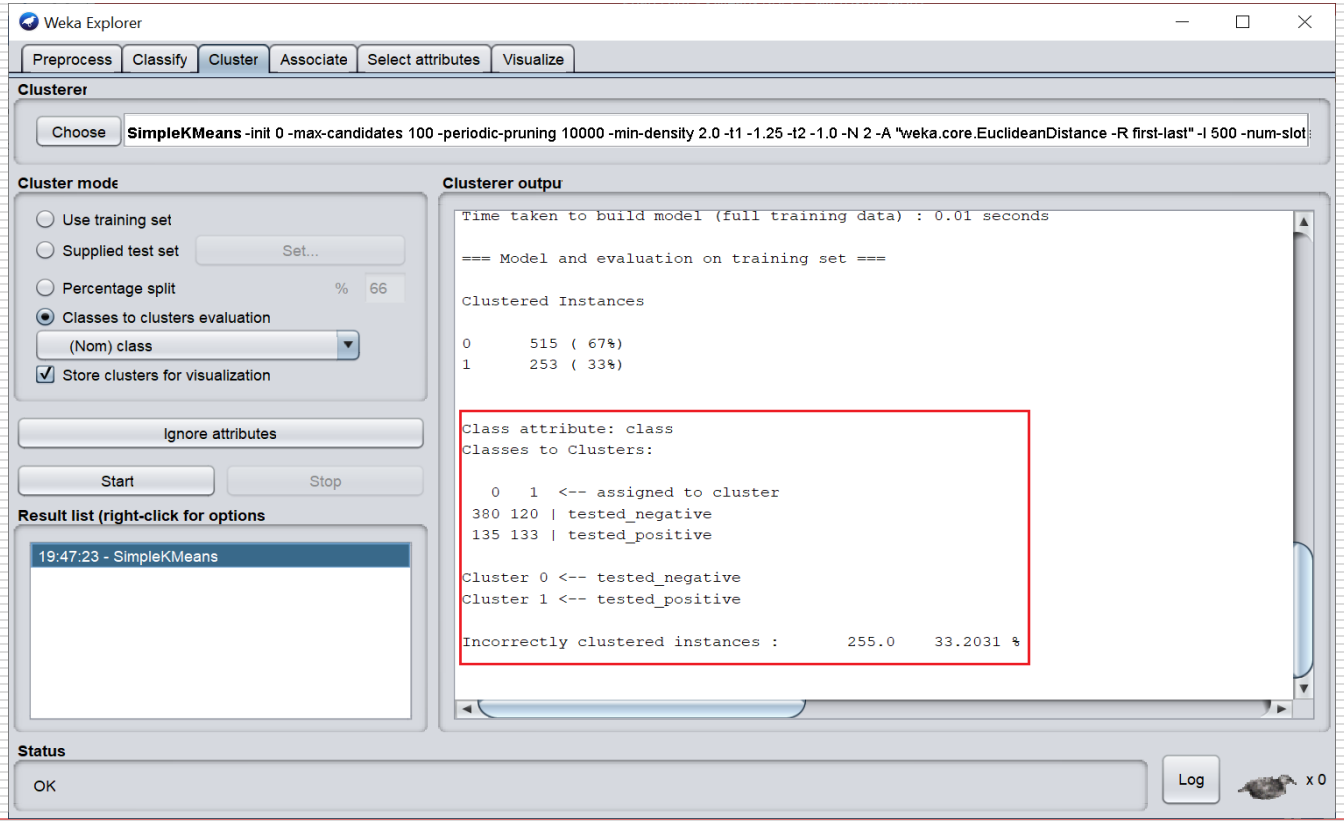
# So sánh kết quả phân cụm và trường class

- Trong tab cluster, ko chọn training set mà chọn trường để so khớp:



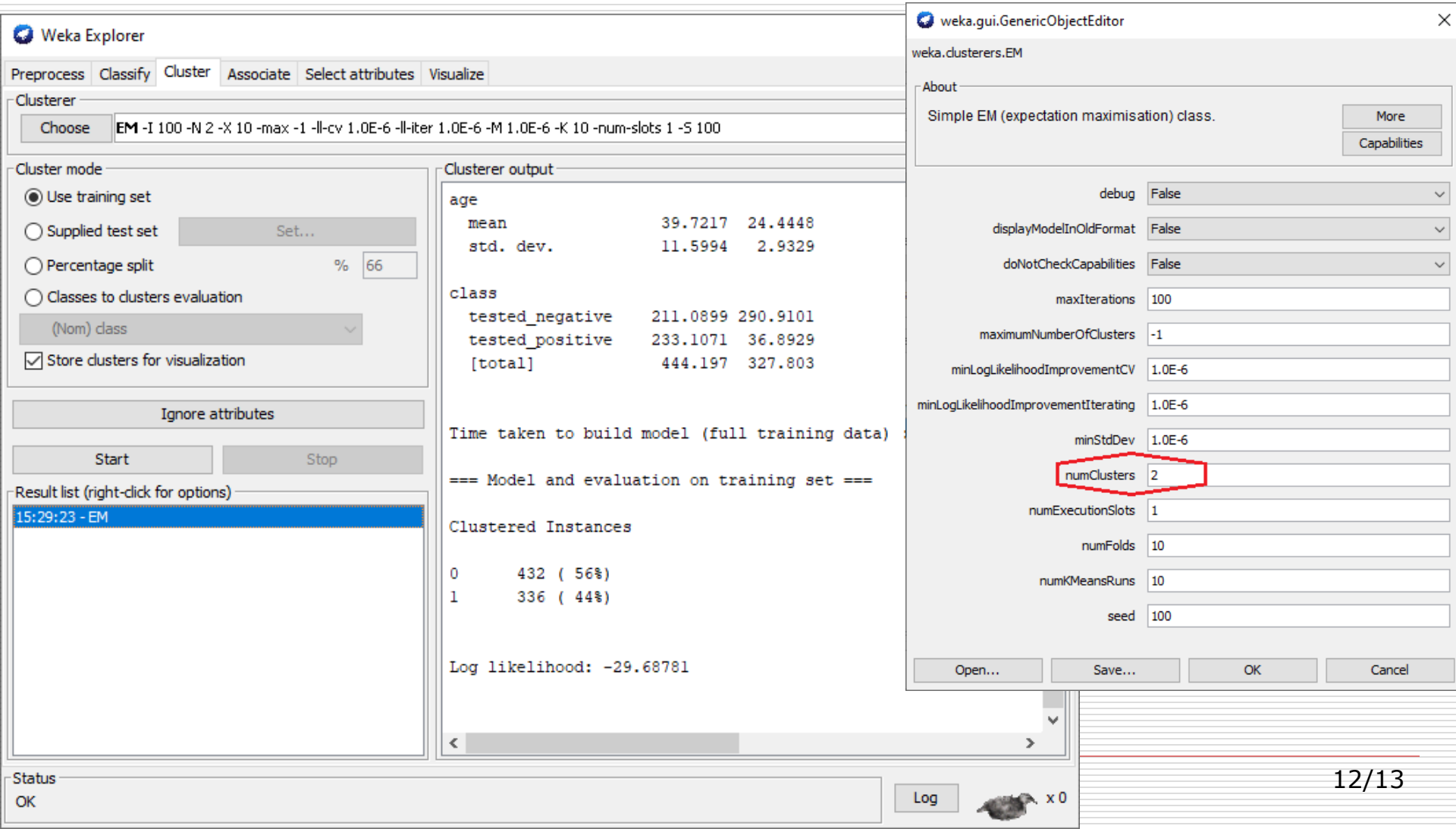
# So sánh kết quả phân cụm và trường class

- Nhấn Start và xem kết quả:
  - Trong 515 về cụm 0 thì có 135 đúng
  - Trong 253 về cụm 1 thì có 133 đúng



# Phân cụm với thuật toán EM

- Cách làm giống kMeans, chỉ khác về tham số mô hình



# Phân cụm với thuật toán EM

- Có thể chọn số cụm là -1 và EM tự động chọn số cụm tốt