



ĐẠI HỌC QUỐC GIA TP.HCM  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

# BÁO CÁO

## LAB01 - Preprocessing

Môn: Khai thác dữ liệu và ứng dụng

**Tên nhóm: 9**

**Thành viên nhóm:**

1742032 - Đỗ Nguyễn Minh Luân

1742041 - Trần Thế Ngọc

1742079 - Huỳnh Thư Tú

## Nội dung thực hiện báo cáo viết

### 1. Tích hợp dữ liệu (integration) (5.0 điểm)

a. (1.0đ) Định nghĩa thế nào là tích hợp dữ liệu?

\_ Tích hợp dữ liệu là quá trình kết hợp dữ liệu từ nhiều nguồn thông tin khác nhau. Các nguồn này gồm nhiều cơ sở dữ liệu (multiple databases), dữ liệu khối (data cubes), các dữ liệu được lưu trữ trong các tập tin và thư mục mà không phải lưu trữ database (flat files). Nhằm cung cấp cho người dùng một cái nhìn tổng quan và duy nhất về các dữ liệu này. Là một đặc tính quan trọng nhất của data warehouse.

b. (1.0đ) Vấn đề nhận diện thực thể (entity identification) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?

\_ Vấn đề nhận diện thực thể (entity identification) có xảy ra trong hai tập dữ liệu.  
\_ Để giải quyết vấn đề này người ta sử dụng siêu dữ liệu (metadata).

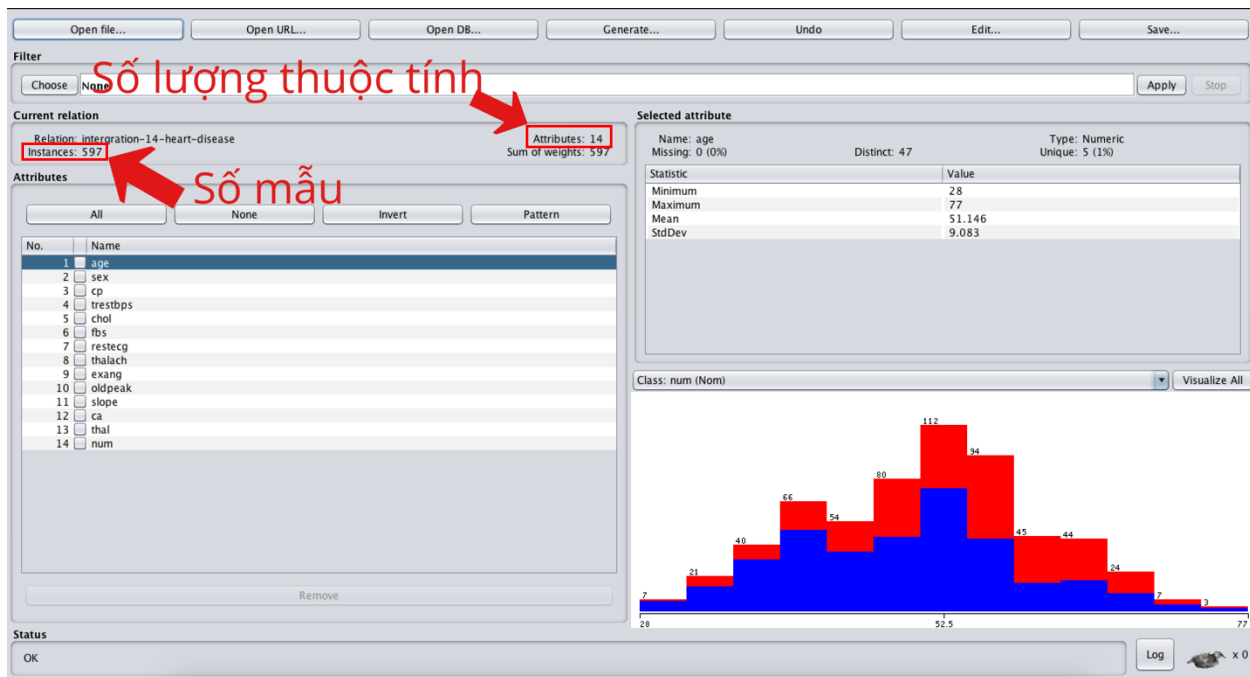
c. (1.0đ) Vấn đề dữ liệu dư thừa (data redundancy) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?

\_ Vấn đề dữ liệu dư thừa (data redundancy) có xảy ra trong hai tập dữ liệu.  
\_ Để giải quyết vấn đề này ta loại bỏ những dữ liệu dư thừa như sau:  
+ Một thuộc tính là thừa nếu nó có thể suy ra từ các thuộc tính khác.  
+ Cùng một thuộc tính có thể có nhiều tên trong các cơ sở dữ liệu khác nhau.  
+ Một số mẫu tin dữ liệu bị lặp lại.  
+ Dùng phép phân tích tương quan.  
•  $r = 0$ : X và Y không tương quan  
•  $r > 0$ : tương quan thuận.  $X \uparrow \leftrightarrow Y \uparrow$   
•  $r < 0$ : tương quan nghịch.  $X \downarrow \leftrightarrow Y \uparrow$

d. (1.0đ) Vấn đề mâu thuẫn giá trị dữ liệu (data value conflicts) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?

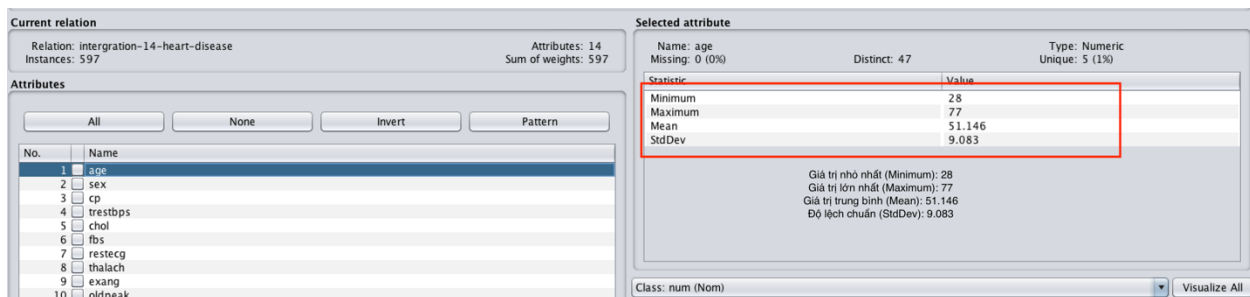
\_ Vấn đề mâu thuẫn giá trị dữ liệu (data value conflicts) có xảy ra trong hai tập dữ liệu  
\_ Để giải quyết vấn đề này ta xác định chuẩn và ánh xạ dựa trên siêu dữ liệu.

e. (1.0đ) Tích hợp hai tập dữ liệu đã cho thành tập dữ liệu mới có tên là heart-integration.arff. Sử dụng WEKA để đọc tập dữ liệu tích hợp. Chụp màn hình cửa sổ Explorer, đánh dấu các vùng trong cửa sổ có thể cho biết số mẫu và số thuộc tính của dữ liệu.



## 2. Tóm tắt dữ liệu mô tả (descriptive data summarization) (8.0 điểm)

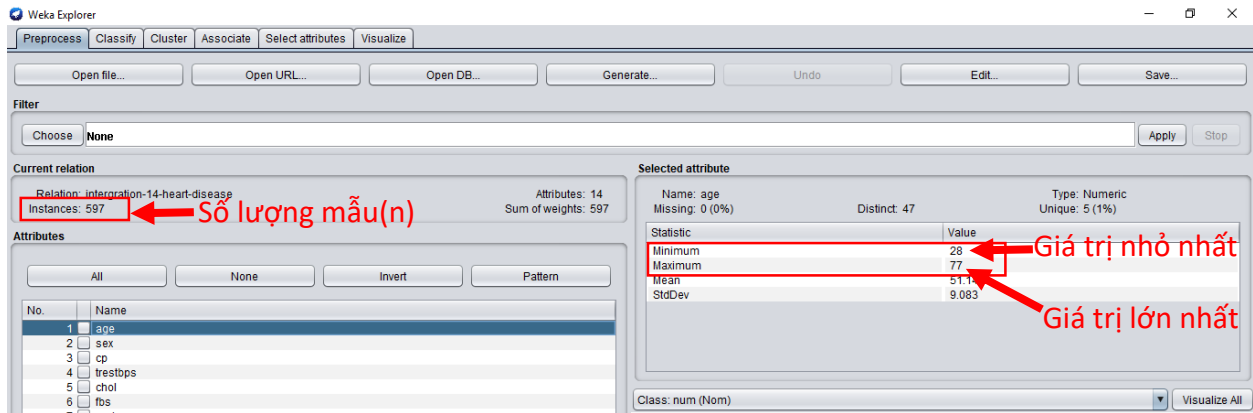
a. (1.0đ) Cho biết giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất và giá trị lớn nhất của thuộc tính age. Chụp màn hình cửa sổ Explorer, đánh dấu các vùng trong cửa sổ cho biết những thông tin này



b. (1.0đ) Xác định five-number summary của thuộc tính age. Chụp màn hình cửa sổ Explorer hiển thị thông tin này nếu WEKA có cung cấp. Nếu không, bạn cần cú vào những giá trị nào khác có trong WEKA để tính?

\_ Để xác định five-number summary của thuộc tính age ta cần biết các giá trị: giá trị lớn nhất, giá trị nhỏ nhất, trung vị, phân vị thứ nhất, phân vị thứ ba.

\_ Weka cung cấp một số thông tin để tính five-number summary của thuộc tính age:



\_ Để xác định five-number summary của thuộc tính age ta cần căn cứ vào những giá trị khác có trong WEKA để tính như: giá trị nhỏ nhất, giá trị lớn nhất, số lượng mẫu.

c. (1.0đ) Cho biết thuộc tính nào có kiểu thuộc tính số (numeric), kiểu rời rạc không có thứ tự (categorical/nominal), hoặc kiểu rời rạc có thứ tự (ordinal)

\_ Numeric: age, trestbps, chol, thalach, oldpeak, ca

\_ Nominal: sex, cp, fbs, restecg, exang, slope, thal, num

d. (1.0đ) Giải thích ý nghĩa của đồ thị ở góc dưới bên phải của cửa sổ. Bạn gọi tên đồ thị này là gì? Đồ thị biểu diễn điều gì về tập dữ liệu? Màu xanh và màu đỏ có ý nghĩa gì (chú ý các pop-up hiện lên khi chuột di chuyển vào vùng đồ thị)?

\_ Giải thích ý nghĩa của đồ thị: Cho phép chúng ta kiểm tra dạng phân phối (chẳng hạn, phân phối chuẩn), điểm dị biệt, độ trôi, độ nhọn của tập dữ liệu.

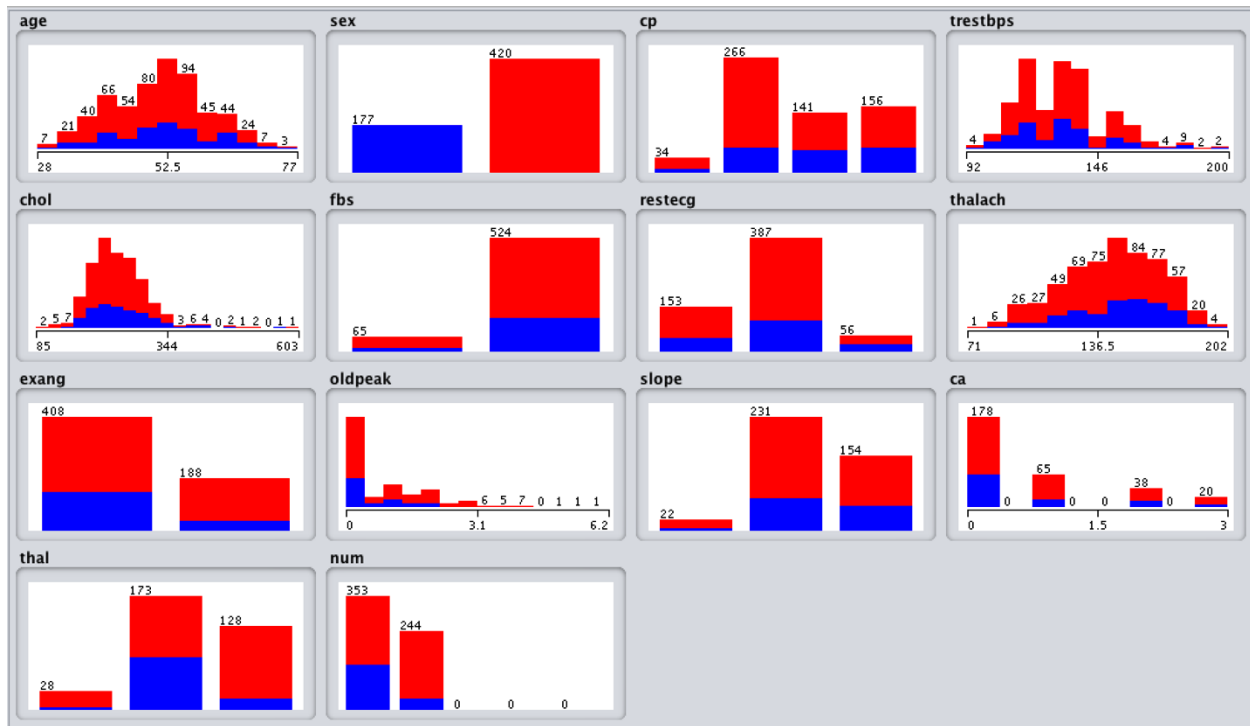
\_ Đồ thị này tên là Histogram.

\_ Đồ thị biểu diễn rõ hơn các nhân tố nguy hiểm cho bệnh tim.

\_ Màu xanh hiển thị lớp <50

\_ Màu đỏ hiển thị lớp >50\_1

e. (1.0đ) Lần lượt xem xét các thuộc tính khác ngoài thuộc tính age. Chụp màn hình cửa sổ Explorer tương ứng với từng thuộc tính



f. (1.0đ) Bạn có nhận xét gì từ những đồ thị trong câu e.?

- \_ Có 2 dạng đồ thị tương ứng với 2 kiểu thuộc tính (numeric, nominal)
- \_ Đối với đồ thị dạng nominal (Histogram):  
Mỗi cột tương ứng với số lượng (count) label trong thuộc tính đó
- \_ Đối với đồ thị dạng Numeric (Barplot):  
Phân bố đều trong khoảng (Minimum, Maximum)
- \_ Có các giá trị trung bình (Mean) và độ lệch chuẩn (StdDev)

g. (1.0đ) Các đồ thị này được gọi tên bằng thuật ngữ gì trong textbook [4]? Chọn jitter tối đa, chú ý cột num (cột cuối cùng). Bạn cho rằng thuộc tính (Y) nào có khả năng dự đoán tốt nhất về bệnh tim như là một hàm của num (X)? Chụp hình đồ thị của cặp thuộc tính (X) – (Y) này

- \_ Các đồ thị này được gọi bằng tên: Scatter Plot

### 3. Chọn lọc dữ liệu (selection) (3.0 điểm)

a. (1.0đ) Dựa vào phần mô tả ở đầu tập tin arff, cho biết có bao nhiêu thuộc tính trong các tập dữ liệu heart-h và heart-c trước khi xử lý?

- \_ Có 14 thuộc tính trong các tập dữ liệu heart-h và heart-c trước khi xử lý.

b.(1.0đ) Giải thích ngắn gọn từng phương pháp chọn lọc thuộc tính trong WEKA.

\_ Use full training set: Mức giá trị của tập thuộc tính được xác định thông qua việc sử dụng toàn bộ dữ liệu được training.

\_ Cross validation: Mức giá trị của tập thuộc tính được xác định bằng quá trình xác thực chéo. Phần Fold và Seed dùng set giá trị folds (gấp) và số lượng random seed dùng để trộn dữ liệu

c.1.0đ) So sánh với các phương pháp chọn lọc dữ liệu trong textbook. Phương pháp nào có trong textbook nhưng không có trong WEKA? Phương pháp nào có trong WEKA nhưng không có trong textbook?

\_ Các phương pháp trong textbook:

Stepwise forward selection, Stepwise backward elimination, Combination of forward selection and backward elimination, Decision tree induction

\_ Các phương pháp trong weka:

Use full training set, Cross validation

\_ Điểm giống nhau giữa các phương pháp trong textbook và weka:

Đều dựa trên mức độ giá trị của thuộc tính để chọn lọc

\_ Điểm khác nhau giữa textbook và weka:

+ Textbook:

Đối với phương pháp Stepwise forward selection, việc chọn lọc bắt đầu với việc khởi tạo tập các thuộc tính rỗng. Xác định các thuộc tính tốt trong mảng ban đầu và thêm vào tập khởi tạo.

Đối với phương pháp Stepwise backward elimination, việc chọn lọc bắt đầu với việc loại bỏ các thuộc tính xấu ra khỏi tập thuộc tính ban đầu

Đối với phương pháp Decision tree induction, sử dụng các giải thuật để chọn lọc ra các thuộc tính tốt

+ Weka:

Chọn lọc thuộc tính dựa trên mức giá trị của thuộc tính thông qua việc training data hoặc xác thực chéo

#### 4. Làm sạch dữ liệu (cleaning) (5.0 điểm)

a. (1.0đ) Liệt kê các phương pháp đã học trong bài giảng để xử lý vấn đề thiếu giá trị (missing values). WEKA hỗ trợ những phương pháp nào cho vấn đề này?

\_ Liệt kê các phương pháp đã học trong bài giảng để xử lý vấn đề thiếu giá trị (missing values):

- + Bỏ qua các mẫu tin có giá trị thiếu.
- + Điền các giá trị thiếu bằng tay.
- + Điền các giá trị thiếu tự động.

\_ Các phương pháp mà WEKA hỗ trợ cho vấn đề này là:

- + Set missing values to..

b. (1.0đ) Liệt kê các phương pháp đã học để loại bỏ dữ liệu nhiễu (noisy data). WEKA hỗ trợ những phương pháp nào cho vấn đề này?

\_ Liệt kê các phương pháp đã học để loại bỏ dữ liệu nhiễu (noisy data):

- + Phương pháp chia giỏ.
- + Phương pháp gom nhóm.
- + Phương pháp hồi quy

\_ Các phương pháp mà WEKA hỗ trợ cho vấn đề này là:

- + Use training set
- + Supplied test set
- + Cross-validation
- + Percentage split

d. (1.0đ) Tập dữ liệu có gặp phải các vấn đề nêu trên hay không? Nếu có, liệt kê một số giá trị đại diện cho từng trường hợp và mô tả lựa chọn của bạn để giải quyết vấn đề (bạn có thể chọn bộ lọc của WEKA hoặc tự đề xuất phương pháp riêng).

\_ Tập dữ liệu gặp phải vấn đề thiếu dữ liệu (Missing values)

\_ Giải pháp:

1. Thay thế các trường dữ liệu bị thiếu (Replace missing values)

Sử dụng bộ lọc của WEKA: Unsupervised > Attribute > Replace missing values

Mục đích: Thay thế các giá trị bị thiếu cho 2 tập kiểu dữ liệu Numeric và Nominal với các giá trị trung bình trong tập dữ liệu

2. Thêm các dữ liệu rác (Add noise)

Sử dụng bộ lọc của WEKA: Unsupervised > Attribute > Add Noise

Mục đích: Thêm các dữ liệu rác để thay thế các trường dữ liệu bị thiếu

e. (1.0đ) Lưu dữ liệu đã làm sạch vào tập tin heart-cleaned.arff. Chụp hình các phần dữ liệu có sự thay đổi trước và sau khi làm sạch.

Replace missing values

Trước khi làm sạch:

Current relation		Selected attribute	
Relation: intergration-14-heart-disease Instances: 597		Name: chol Missing: 23 (4%) Distinct: 201 Type: Numeric Unique: 65 (11%)	
Attributes		Statistic	
All None Invert Pattern		Value	
No.	Name		
1	age	Minimum	
2	sex	Maximum	
3	cp	Mean	
4	trestbps	StdDev	
5	chol		
6	fbs		
7	restecg		
8	thalach		
9	exang		
10	oldpeak		
		Class: num (Nom)	
		Visualize All	

Sau khi làm sạch:

Current relation		Selected attribute	
Relation: intergration-14-heart-disease-weka.filters.unsupervised.attribute.ReplaceMiss... Instances: 597		Name: chol Missing: 0 (0%) Distinct: 202 Type: Numeric Unique: 65 (11%)	
Attributes		Statistic	
All None Invert Pattern		Value	
No.	Name		
1	age	Minimum	
2	sex	Maximum	
3	cp	Mean	
4	trestbps	StdDev	
5	chol		
6	fbs		
7	restecg		
8	thalach		

## 5. Chuyển đổi dữ liệu (Transformation) 5 (3.0 điểm)

a. (1.0đ) Bộ lọc nào của WEKA cho phép xây dựng thuộc tính (attribute construction), ví dụ, thêm một thuộc tính là tổng của 2 thuộc tính khác?

\_ Bộ lọc Add, AddCluster, AddValues, AddID, AddNoise, AddUserFields, AddExpression

b. (1.0đ) Bộ lọc nào của WEKA cho phép chuẩn hóa thuộc tính (normalization)? Bộ lọc này có thể chuẩn hóa Min-max, chuẩn hóa Z-score hay chuẩn hóa thập phân không? Nếu có, cho biết cụ thể cách thực hiện những chuẩn hóa này trong WEKA. Nếu không, mô tả giải pháp chuẩn hóa mà WEKA hỗ trợ.

\_ Bộ lọc cho phép chuẩn hóa thuộc tính: Unsupervised.Attribute.Normalize

\_ Bộ lọc này không thể chuẩn hóa min-max, Z-score hoặc chuẩn hóa thập phân

\_ WEKA cung cấp các bộ lọc tương ứng để hỗ trợ:

+ Chuẩn hóa min-max: weka.filters.unsupervised.attribute.MathExpression

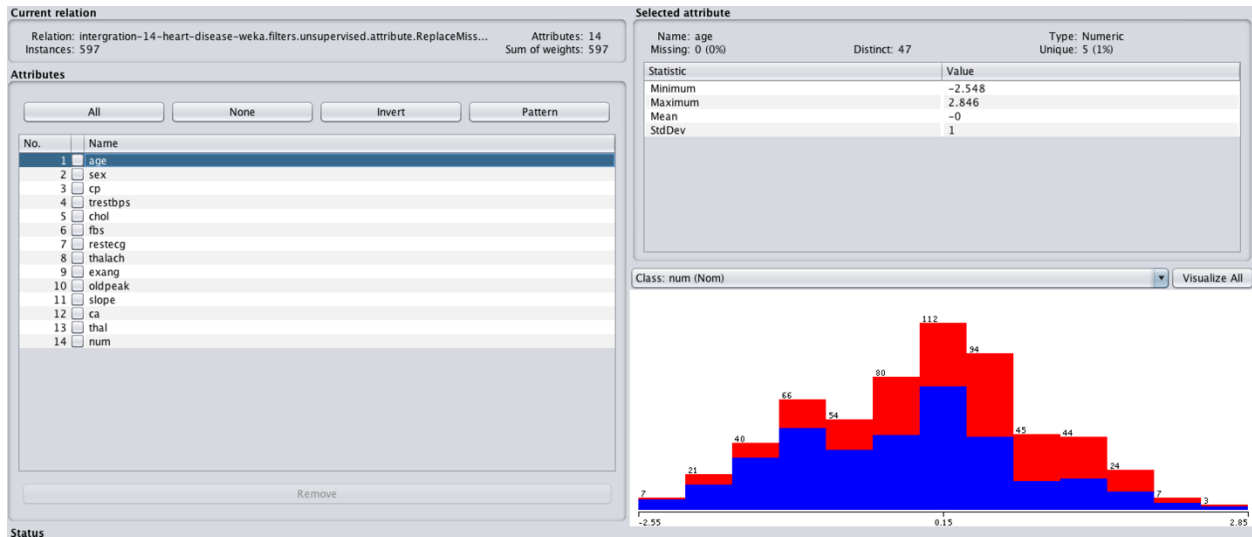
+ Chuẩn hóa Z-score: weka.filters.unsupervised.attribute.Standardlize

+ Chuẩn hóa thập phân:



c. (1.0đ) Chọn một bộ lọc chuẩn hóa trong WEKA và tiến hành chuẩn hóa tất cả các thuộc tính là số thực. Lưu dữ liệu đã chuẩn hóa vào tập tin heart-normal.arff. Chụp hình ít nhất 10 dòng dữ liệu với tất cả thuộc tính số thực để thể hiện rõ sự thay đổi sau chuẩn hóa.

Sau khi chuẩn hóa



The screenshot shows the WEKA interface with the 'heart-normal.arff' file open. The file contains 14 attributes: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and num. The data is shown in a table format with 14 columns and 14 rows. The first row of data is: 28,male,atyp\_angina,130,132,f,left\_vent\_hyper,185,no,0,flat,0.667774,norm,1.0. The last row of data is: 38,female,atyp\_angina,140,207,f,norm,158,no,0,flat,0.667774,norm,1.0.

## 6. Rút gọn dữ liệu (Reduction) (1.0 điểm)

a. (1.0đ) Bộ lọc nào của WEKA cho phép lấy mẫu? Nó có thể thực hiện Simple Random Sample Without Replacement, và Simple Random Sample With Replacement hay không? Nếu có, cho biết cụ thể cách thực hiện những kỹ thuật này trong WEKA. Nếu không, mô tả giải pháp lấy mẫu mà WEKA hỗ trợ

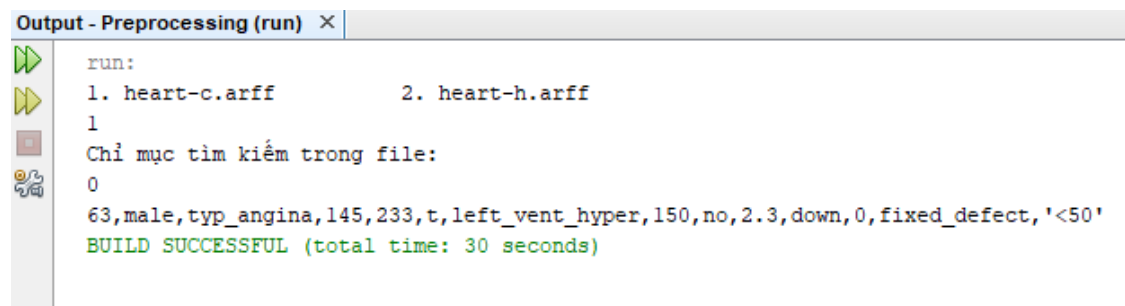
\_ Bộ lọc `weka.filters.unsupervised.instance.Resample` cho phép lấy mẫu có thể thực hiện Simple Random Sample Without Replacement và With Replacement.

\_ Cách thực hiện của kỹ thuật:

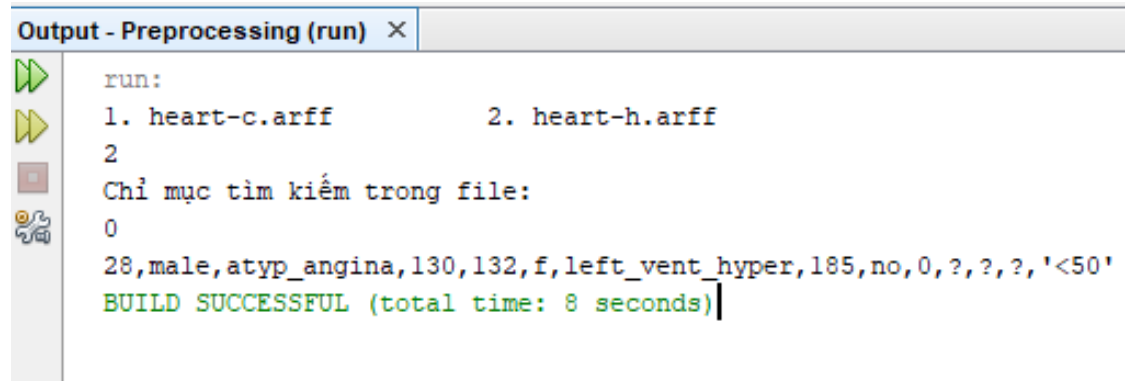
Tạo ra một mảng dataset các mẫu phụ ngẫu nhiên dùng một trong 2 cách Without Replacement và With Replacement. Kích thước tập dữ liệu gốc phải nằm vừa trong bộ nhớ. Số lượng instances được tạo ra bởi tập dữ liệu phải được cung cấp cụ thể. Tập dữ liệu phải có một thuộc tính của Nominal class. Nếu không có thì phải dùng bản filter trong Unsupervised package. Bộ lọc được tạo để duy trì sự phân phối lớp trong các mẫu phụ hoặc ảnh hưởng trực tiếp đến việc phân phối các hình mẫu. Khi sử dụng trong batch mode (vd: `FilteredClassifier`), các dữ liệu tiếp theo sẽ không được lấy mẫu lại.

## Nội dung thực hiện cài đặt

a. (5.0đ) Đọc hai tập dữ liệu `heart-h.arff` và `heart-c.arff` (2.0đ). Cho phép người dùng truy vấn một dòng dữ liệu bất kỳ trong các tập dữ liệu này bằng cách nhập chỉ mục của dòng dữ liệu (tính từ 0) (3.0đ).



```
run:
1. heart-c.arff      2. heart-h.arff
1
Chỉ mục tìm kiếm trong file:
0
63,male,typ_angina,145,233,t,left_vent_hyper,150,no,2.3,down,0,fixed_defect,<50'
BUILD SUCCESSFUL (total time: 30 seconds)
```



```
run:
1. heart-c.arff      2. heart-h.arff
2
Chỉ mục tìm kiếm trong file:
0
28,male,atyp_angina,130,132,f,left_vent_hyper,185,no,0,?, ?, ?, <50'
BUILD SUCCESSFUL (total time: 8 seconds)
```

b. (5.0đ) Tích hợp tự động hai tập dữ liệu heart-h.arff và heart-c.arff thành tập dữ liệu mới heart-integration-auto.arff theo các kỹ thuật đã lựa chọn trong phần báo cáo viết (2.0đ). Đối chiếu với tập tin heart-integration.arff (3.0đ).

Dữ liệu ở tập tin heart-integration-auto.arff và tập tin heart-integration.arff là giống nhau.

c. (5.0đ) Làm sạch tự động tập dữ liệu heart-integration-auto.arff thành tập dữ liệu mới heart-cleaned-auto.arff theo các kỹ thuật đã lựa chọn trong phần báo cáo viết (2.0đ). Đối chiếu với tập tin heart-cleaned.arff (3.0đ).

Dữ liệu ở tập tin heart-cleaned-auto.arff và tập tin heart-cleaned.arff là giống nhau.

d. (5.0đ) Chuẩn hóa tự động tập dữ liệu heart-cleaned-auto.arff thành tập dữ liệu mới heart-normal-auto.arff theo các kỹ thuật đã lựa chọn trong phần báo cáo viết (2.0đ). Đối chiếu với tập tin heart-normal.arff (3.0đ).

Dữ liệu ở tập tin heart-normal-auto.arff và tập tin heart-normal.arff là giống nhau.