**A – Nội dung thực hiện báo cáo viết:**

**Yêu cầu 1: Tiền xử lý dữ liệu:**

**1. 1 Liệt kê các thuộc tính có xảy ra trường hợp dữ liệu thiếu. Cho biết số giá trị thiếu ở mỗi thuộc tính đã nhận diện.**

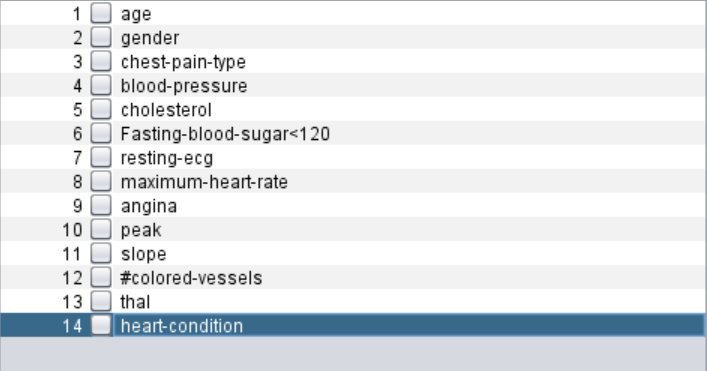
**Trả lời:**

- Các thuộc tính có xảy ra trường hợp thiếu và giá trị thiếu cho mỗi thuộc tính:

* age : 3
* slope: 5

**1.2 Để tránh việc thiếu dữ liệu ảnh hưởng đến chất lượng phân tích, hãy áp dụng bộ lọc thích hợp trong nhóm Unsupervised/Attribute để thay thế mọi giá trị thiếu trong tập dữ liệu gốc. Đồng thời, đổi tên thuộc tính class thành heart-condition. Lưu dữ liệu sau tiền xử lý vào tập tin cardiology-cleaned.arff và nộp lại tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu.**

**Trả lời:**



* Nhóm em đã xử lý dữ liệu còn thiếu và đã có lưu lại file theo yêu cầu, được gửi đính kèm.

**1.3 Các giá trị thiếu đã lần lượt được thay thế bằng những giá trị gì? Giải thích lý do vì sao chúng được thay thế như vậy?**

**Trả lời:**

- Thuộc tính có kiểu là nonimal thì được thay bằng giá trị mode. Bởi vì ta lấy giá trị xuất hiện nhiều nhất gán cho những dòng còn thiếu có phần trăm làm sai kết quả sẽ thấp. Giá trị xuất hiện nhiều sẽ mang một ý nghĩa nào đó .

- Thuộc tính có kiểu là numberic thì được thay bằng giá trị mean. Bởi vì ta lấy giá trị trung bình thay cho giá trị còn thiếu thì sẽ làm cho cân bằng giá trị trong trong dữ liệu, khi thu thập dữ liệu sẽ có những giá trị cực kì cao, đồng thời cũng sẽ có những giá trị thấp.

**Yêu cầu 2: Gom cụm có xét thông tin phân phân lớp:**

**2.1 Áp dụng giải thuật SimpleKMeans lên dữ liệu tiền xử lý (cardiology-cleaned) với thông số chỉ định như bên dưới**

**• Cluster mode: Use training set**

**• Xét cả thuộc tính lớp heart-condition**

**• Số lượng cụm: 3**

**• Display the standard deviation**

**• Distance function: Euclidean Distance**

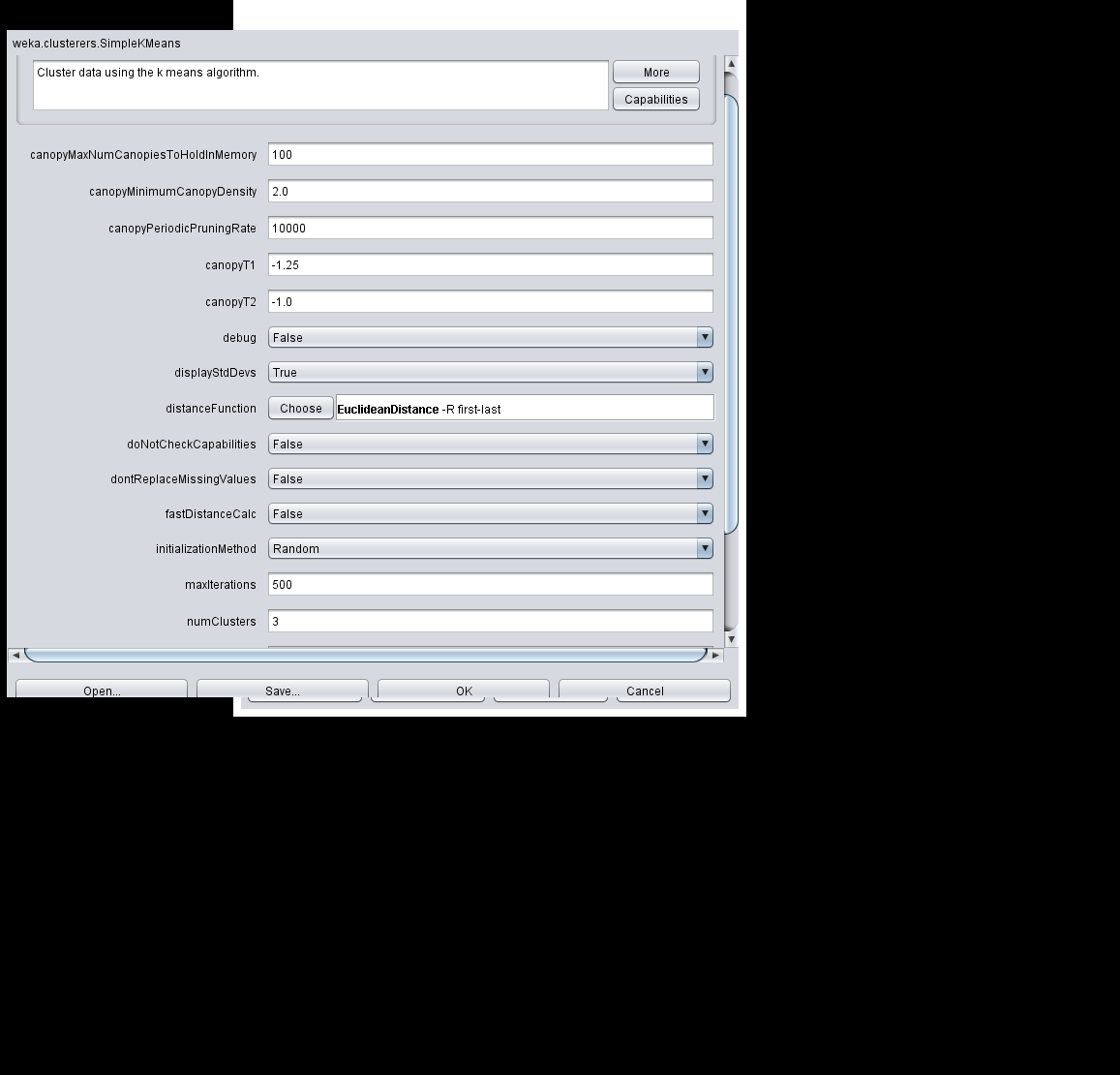
**Dựa vào kết quả thực thi, hãy thông dịch kết quả tâm cụm tìm được để rút ra nhận định về các vấn đề sau**

**a. Nam giới có nguy cơ mắc bệnh về tim cao hơn.**

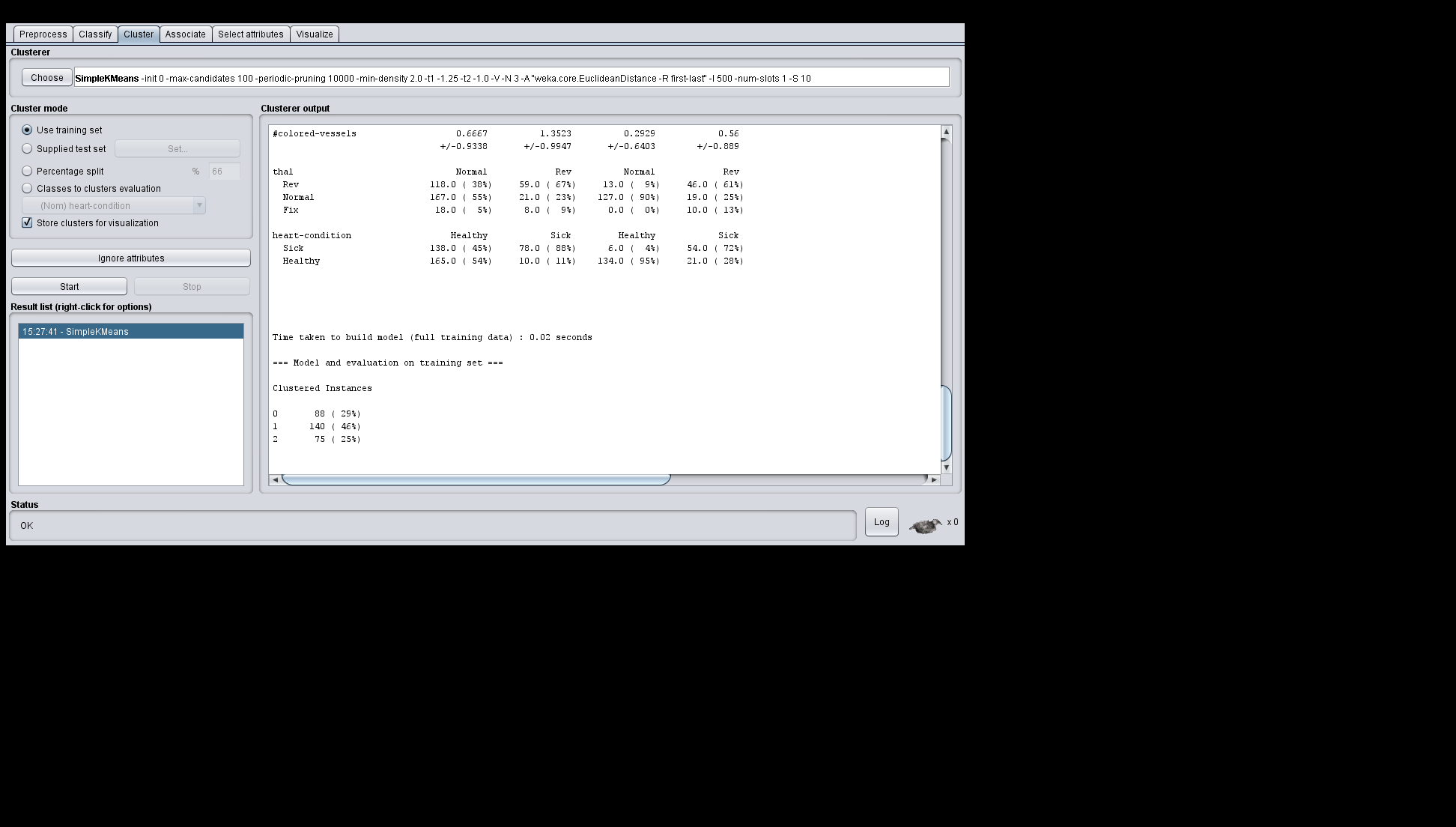
**b. Nữ giới có nguy cơ mắc bệnh về tim thấp hơn.**

**Trả lời:**

* Thiết lập chức năng theo yêu cầu:



* Kết quả sau thực hiện được



**Cluster 0:**   
57.5528,Male,Asymptomatic,135.4886,256.0795,FALSE,Hyp,133.2955,TRUE,1.8114,Flat,  
1.3523,Rev,Sick

**Cluster 1:**

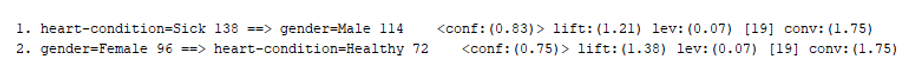
52.0595,Female,NoTang,128.6714,245.65,FALSE,Normal,FALSE,0.49,Up,0.2929,Normal,

Healthy

**Cluster 2:**  
54.76,Male,Asymtomatic,132.6,235.8933,FALSE,Normal,149.12,FALSE,1.16,Flat,0.56,Rev,  
Sick

Weka sinh ra ba cụm 0,1 và 2 và sau khi phân cụm như hình trên (có đánh dấu) thì cụm 1 gồm 140 (46%) mẫu là nữ có độ tuổi trung bình 57 không bị bệnh tim so với 2 cụm 0 và 1 đều là nam mắc bệnh tim chiến 54% do đó nữ mắc bệnh tim ít hơn nam.

**2.2 Nhận định từ kết quả gom cụm có nhất quán với mối liên hệ giữa giới tính và bệnh tim được Apriori cung cấp như trên? Giải thích lý do.**

**Trả lời:**

* Kết quả gom cụm nhất quán với mối quan hệ giữa giới tính và bệnh tim được Apriori cung cấp như đề bài vì: theo như luật kết hợp (1) thì có 138 người bị bệnh tim thì có hết 114 là nam bị bệnh tim và theo kết quả gom cụm thì phần trăm nam mắc bệnh tim cao hơn nữ.

**2.3 Khảo sát với giá trị k thay đổi từ 2 đến 5. Lưu lại thông tin hiển thị ở Clusterer Output vào các tập tin <k>-clusters.txt, và nộp lại các tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).**

**Yêu cầu 3: Đánh giá kết quả gom cụm:**

**3.1 Áp dụng giải thuật k-means lên dữ liệu tiền xử lý (cardiology-cleaned) với thông số chỉ định như bên dưới để thu được 3 cụm dữ liệu**

**• Cluster mode: Use training set**

**• Bỏ qua thuộc tính lớp heart-condition**

**• Số lượng cụm: 3**

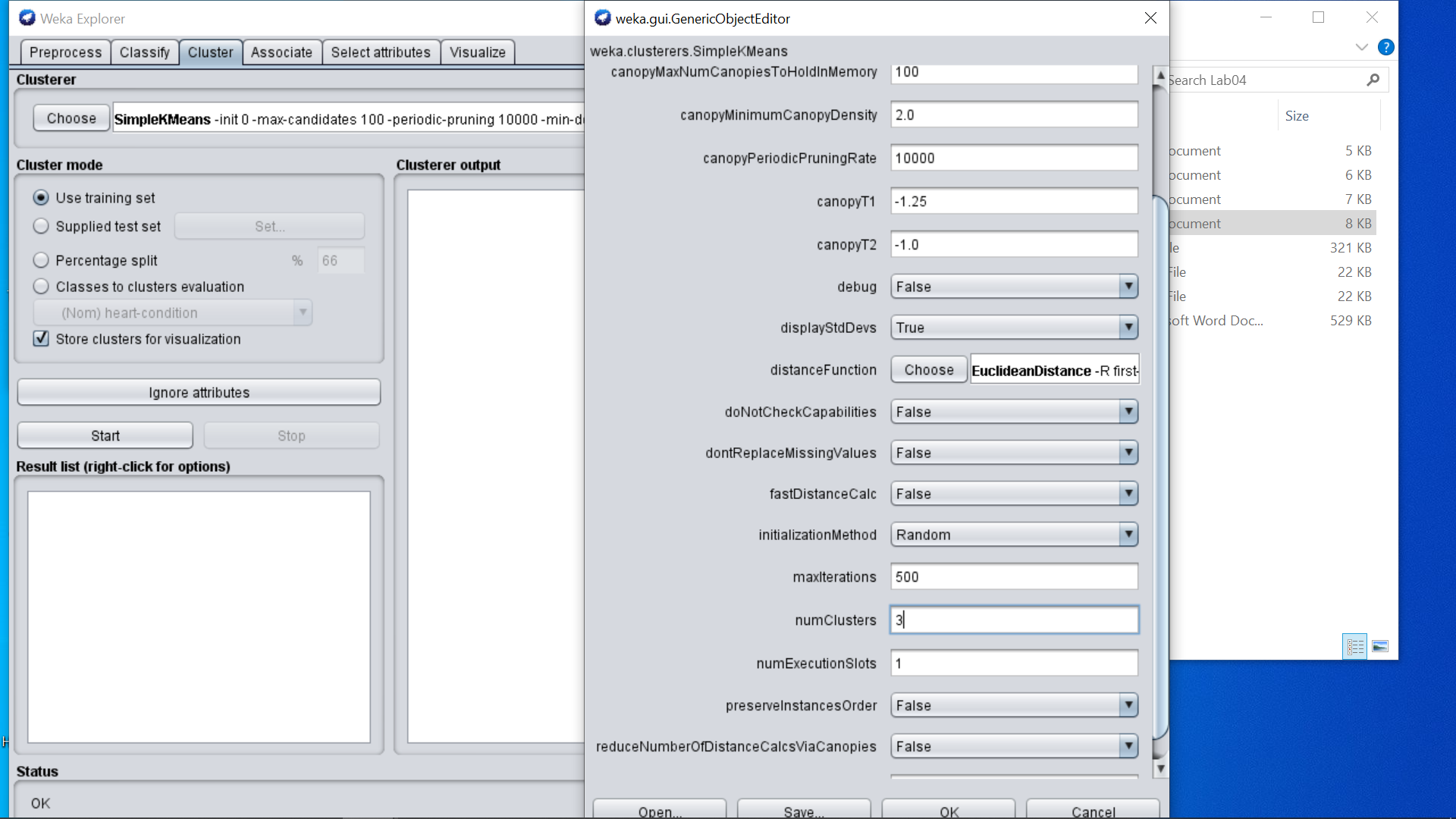
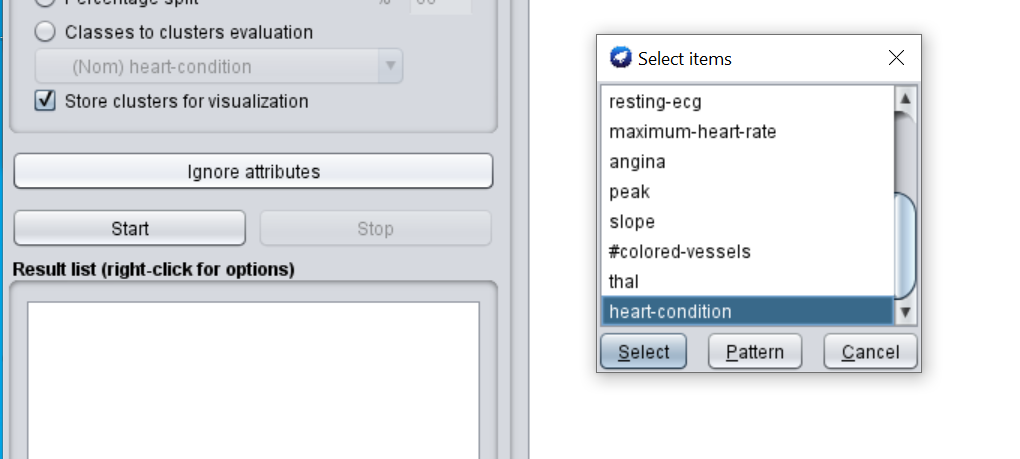
**• Display the standard deviation**

**• Distance function: Euclidean Distance**

**Lưu lại thông tin Clusterer Output vào tập tin 3-ts-classignored.txt và nộp lại tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).**

**Trả lời:**

* Nhóm có gửi file kèm theo như yêu cầu.
* Hình ảnh đặt tham số



**3.2 Áp dụng giải thuật k-means lên dữ liệu tiền xử lý (cardiology-cleaned) với thông số chỉ định như bên dưới để thu được 3 cụm dữ liệu**

**• Cluster mode: Classes to cluster evaluation**

**• Bỏ qua thuộc tính lớp heart-condition**

**• Số lượng cụm: 3**

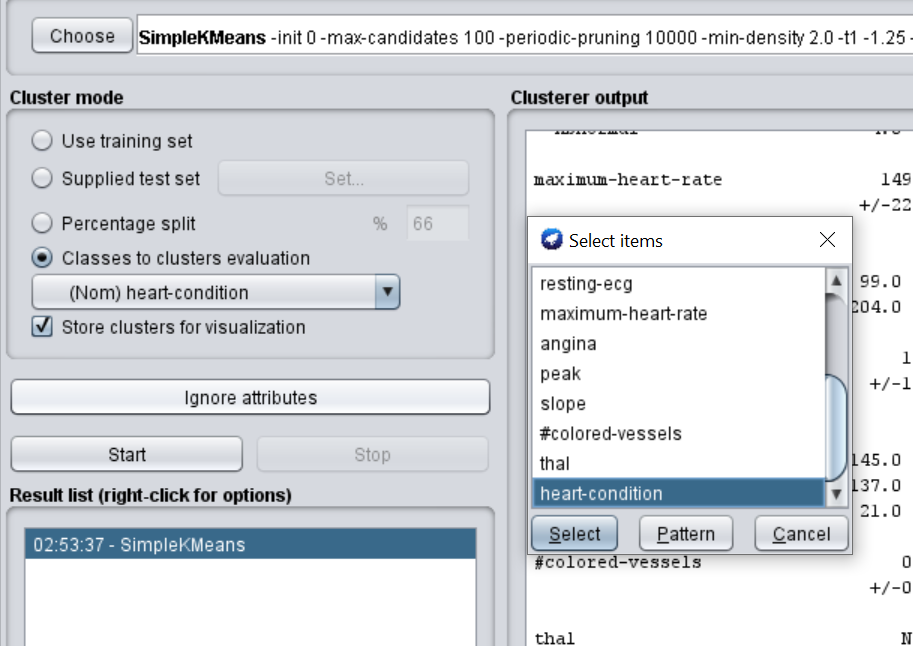
**• Display the standard deviation**

**• Distance function: Euclidean Distance Lưu lại thông tin Clusterer Output vào tập tin 3-ce-classignored.txt và nộp lại tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).**

**Trả lời:**

- Nhóm có gửi kèm file như đề yêu cầu.

- Hình ảnh đặt tham số theo yêu cầu.



**3.3 Giả sử bạn nhận được kết quả gom cụm như bên dưới. Hãy diễn giải thông tin này.**

**Clustered Instances**

1. **113 ( 37%)**
2. **81 ( 27%)**
3. **109 ( 36%)**

**Trả lời:**

- Theo kết quả gom cụm như trên thì số cụm gom là 3 tương ứng các cụm: 0, 1 và 2

- Cụm 0 gom được 113 mẫu chiếm 37% trong tập dữ liệu

- Cụm 1 gom được 81 mẫu chiếm 27% trong tập dữ liệu

- Cụm 2 gom được 109 mẫu chiếm 36% trong tập dữ liệu

**3.4 Giả sử bạn nhận được kết quả gom cụm như bên dưới. Hãy diễn giải thông tin này.**

**Class attribute: heart-condition**

**Classes to Clusters:**

1. **1 2 <-- assigned to cluster**

**22 20 96 | Sick**

**91 61 13 | Healthy**

**Cluster 0 <-- Healthy Cluster**

**Cluster 1 <-- No class Cluster**

**Cluster 2 <-- Sick**

**Trả lời:**

* Thuộc tính đưa vào gom cụm đó là: **heart-condition.**
* Gồm có ba cụm 0, 1 và 2.
* Vì class **heart-condition**  chỉ có hai thuộc tính **Healthy** và **Sick** mà lại chia thành ba cụm do đó theo như kết quả trong hình thì cụm 0 sẽ là cụm Healthy giá trị nhận ở cột hear-condition là Healthy, cụm 1 thì không thuộc hai giá của class **heart-condition,** cụm 2 là cụm **Sick** nhận giá trị cột Healthy-condition là Sick.
* Cụm 0 gom được 22 mẫu.

**B – Nội dung thực hành:**

Trong thư mục có các file:

cardiology-cleaned.arff: data input cho

3.py: file chạy chính

Chạy câu lệnh với cú pháp:

<ID nhom> <input> <output\_model><k>

Ví dụ:

<ID nhom>: tên của tập tin thực thi chương trình là ID của nhóm

<input>: tập tin dữ liệu đầu vào

<output\_model>: tập tin đầu ra

<k>: số lượng cụm cần gom

**Ví dụ:**

python 3.py cardiology-cleaned.arff clusters.txt 3