TRƯỜNG ĐẠI HỌC THỦY LỢI

KHOA CÔNG NGHỆ THÔNG TIN

**BỘ MÔN HỆ THỐNG THÔNG TIN**

-----0 0 0-----



**BÀI TẬP LỚN**

**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI**

**PHÂN LỚP DỰ ĐOÁN CƠN ĐAU TIM SỬ DỤNG THUẬT TOÁN NAIVE BAYES**

|  |  |
| --- | --- |
| Sinh viên thực hiện: | Trần Việt Phương - 2051063533  Trần Thanh Thảo – 2051063511  Nguyễn Hữu Trường - 2051060752  Vũ Thanh Sơn – 2051063469 |
| Nhóm thực hiện: | Nhóm 06 |
| Lớp: | 62TH-NB |
| Giảng viên phụ trách môn học: | Nguyễn Tu Trung |

**Hà Nội, năm 2023**

**MỤC LỤC**

[Chương 1: Tổng quan đề tài. 4](#_Toc150186092)

[1.1 Đặt vấn đề 4](#_Toc150186093)

[1.2 Cơ sở hình thành đề tài 4](#_Toc150186094)

[1.3 Mục tiêu đề tài. 4](#_Toc150186095)

[1.4 Đối tượng và phương pháp nghiên cứu 4](#_Toc150186096)

[Chương 2: Tiền xử lý dữ liệu 5](#_Toc150186097)

[2.1 Mô tả dữ liệu 5](#_Toc150186098)

[2.2 Tiền xử lý dữ liệu bằng Weka 6](#_Toc150186099)

[2.3 Dữ liệu sau khi xử lý 11](#_Toc150186103)

[Chương 3: Thuật toán Naive Bayes 12](#_Toc150186104)

[3.1 Thuật toán Bayes 12](#_Toc150186105)

[3.2 Mô tả bài toán 14](#_Toc150186109)

[3.3 Giải bài toán bằng thuật toán Bayes 16](#_Toc150186114)

[Chương 4: Thực hiện giải thuật Bayes trên Weka 20](#_Toc150186117)

[4.1 Giới thiệu Weka 20](#_Toc150186118)

[4.2 Các tính năng của phần mềm Weka 20](#_Toc150186119)

[4.3 Các chức năng 21](#_Toc150186120)

[4.4 Giải trên Weka 21](#_Toc150186121)

[Chương 5: Thực nghiệm và đánh giá trên C++ 28](#_Toc150186126)

[KẾT LUẬN 29](#_Toc150186127)

[TÀI LIỆU THAM KHẢO 29](#_Toc150186128)

# **Chương 1: Tổng quan đề tài.**

## **1.1 Đặt vấn đề**

Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác suất, thống kê, máy học nhằm tìm kiếm các tri thức tiềm ẩn trong các kho dữ liệu có kích thước lớn mà người dùng khó có thể nhận biết bằng những kỹ thuật thông thường. Nguồn dữ liệu các chỉ số cơ thể rất lớn, áp dụng khai phá dữ liệu trong lĩnh vực này sẽ mang lại nhiều ý nghĩa cho xã hội. Nó sẽ cung cấp những thông tin quý giá nhằm hỗ trợ trong việc chẩn đoán khả năng mắc phải cơn đau tim của bệnh nhân.

Ứng dụng kỹ thuật phân lớp dữ liệu trong khai phá dữ liệu nhằm phân lớp khả năng gặp phải cơn đau tim là một trong những hướng nghiên cứu chính của đề tài. Đề tài đề xuất phương pháp phân lớp dựa trên mật độ và thuật toán Bayes để tìm ra quy luật tiềm ẩn trong dữ liệu.

## **1.2 Cơ sở hình thành đề tài**

Những năm gần đây, ngành y tế rất chú trọng đến việc ứng dụng công nghệ thông tin trong các hoạt động điều hành, chẩn đoán tình trạng bệnh nhân. Việc ứng dụng CNTT trong công tác quản lý, CSKH, tính toán khả năng mắc cơn đau tim từ lâu đã trở thành yêu cầu cấp thiết.

Tuy nhiên vẫn có khó khăn, đó là khi xây dựng và triển khai những hệ thống tổng hợp số liệu, do một số hệ thống thành phần chưa đảm bảo tính liên thông, tích hợp nên phải thực hiện qua nhiều bước chuyển đổi trung gian. Bên cạnh đó, các hệ thống thông tin trong lĩnh vực điều khiển, tự động hóa còn thiếu, và chưa phát triển đồng bộ. Vì vậy đề tài “*Sử dụng Bayes phân lớp dự đoán cơn đau tim”* rất đáng được quan tâm, giúp cho hoạt động chẩn đoán của bệnh viện trở nên thuận lợi và chính xác hơn đối với người dân. Đề tài áp dụng với hơn 300 bộ dữ liệu đã thu thập được.

## **1.3 Mục tiêu đề tài.**

Đề tài tập chung vào nghiên cứu kỹ thuật phân lớp trong khai phá dữ liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Sau khi phân tích đặc điểm của dữ liệu thu thập được và lựa chọn giải thuật phù hợp với dữ liệu, việc phân lớp khả năng mắc cơn đau tim, độ hiệu quả của hệ thống cũng là mục tiêu chính của đề tài.

## **1.4 Đối tượng và phương pháp nghiên cứu**

Đề tài tập chung vào nghiên cứu kỹ thuật phân lớp trong khai phá dữ liệu. Cụ thể là nghiên cứu thuật toán Bayes để áp dụng vào việc phân tích cơ sở dữ liệu khả năng mắc cơn đau tim. Thu thập dữ liệu đầu vào của bài toán được lấy từ Kaggle. Sử dụng phương pháp phân lớp Bayes tiến hành nghiên cứu và thực hiện bài toán trong khai phá dữ liệu.

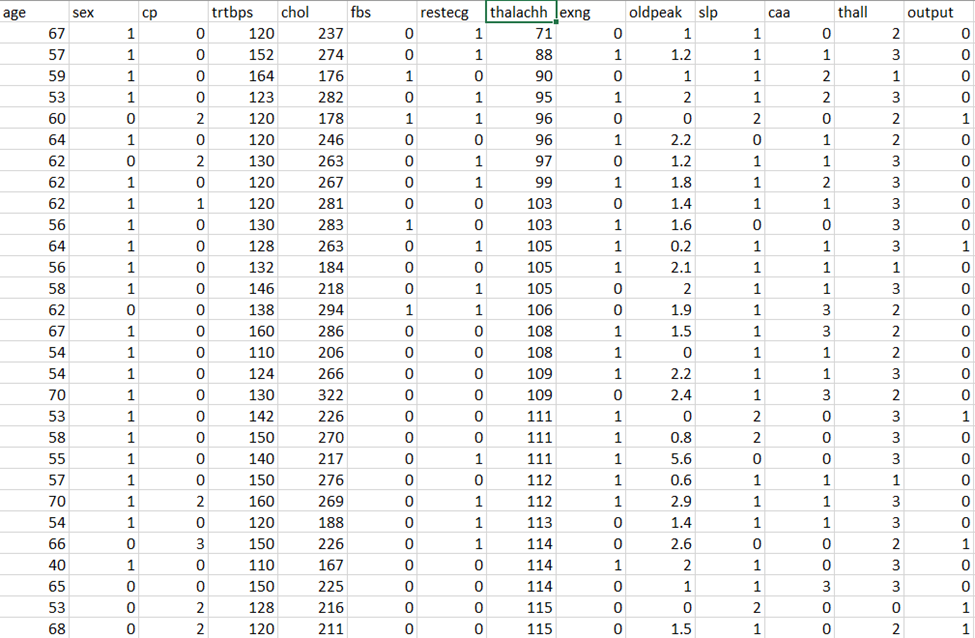
# **Chương 2: Tiền xử lý dữ liệu**

## **Mô tả dữ liệu**

* Thông tin dữ liệu:
* Dữ liệu thô từ Kaggle: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>.
* Mục tiêu: Dự đoán cơn đau tim qua các chỉ số của cơ thể
* Thông tin thuộc tính:

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Mô tả** | **Giá trị** |
| Age | Tuổi | 26 🡪 77 |
| Sex | Giới tính | * 1: nam * 0: nữ |
| Cp | Loại đau ngực | * Giá trị 0: đau thắt ngực điển hình * Giá trị 1: đau thắt ngực không điển hình * Giá trị 2: đau không thắt ngực * Giá trị 3: không có triệu chứng |
| Trbps | Huyết áp khi nghỉ ngơi (tính bằng mm Hg). | 94 🡪 200 |
| Chol | Cholestoral tính bằng mg/dl được tải qua cảm biến BMI. | 126 🡪 564 |
| Fbs | Đường huyết lúc đói > 120 mg/dl | * 1: đúng * 0: sai |
| Restecg | Kết quả điện tâm đồ khi nghỉ ngơi | * Giá trị 0: bình thường * Giá trị 1: có sóng ST-T bất thường (sóng T đảo ngược và/hoặc ST chênh lên hoặc chênh xuống > 0,05 mV) * Giá trị 2: cho thấy phì đại thất trái có thể xảy ra hoặc xác định theo tiêu chí của Estes |
| Thalachh | Nhịp tim đạt tối đa | 71🡪 202 |
| Exng | Đau thắt ngực do gắng sức | * 1: có * 0: không |
| Oldpeak | Đỉnh trước điện tâm đồ | 0 🡪 6.2 |
| Slp | Dốc của điện tâm đồ. | 0 🡪 2 |
| Caa | Số lượng mạch chính. | 0 🡪 4 |
| thall | Tỷ lệ thal | 1 🡪 3 |
| Output | Biến đầu ra: | * 0: ít nguy cơ đau tim * 1: nhiều nguy cơ đau tim |

* Dữ liệu thô ban đầu: 303 mẫu dữ liệu chưa được xử lý.



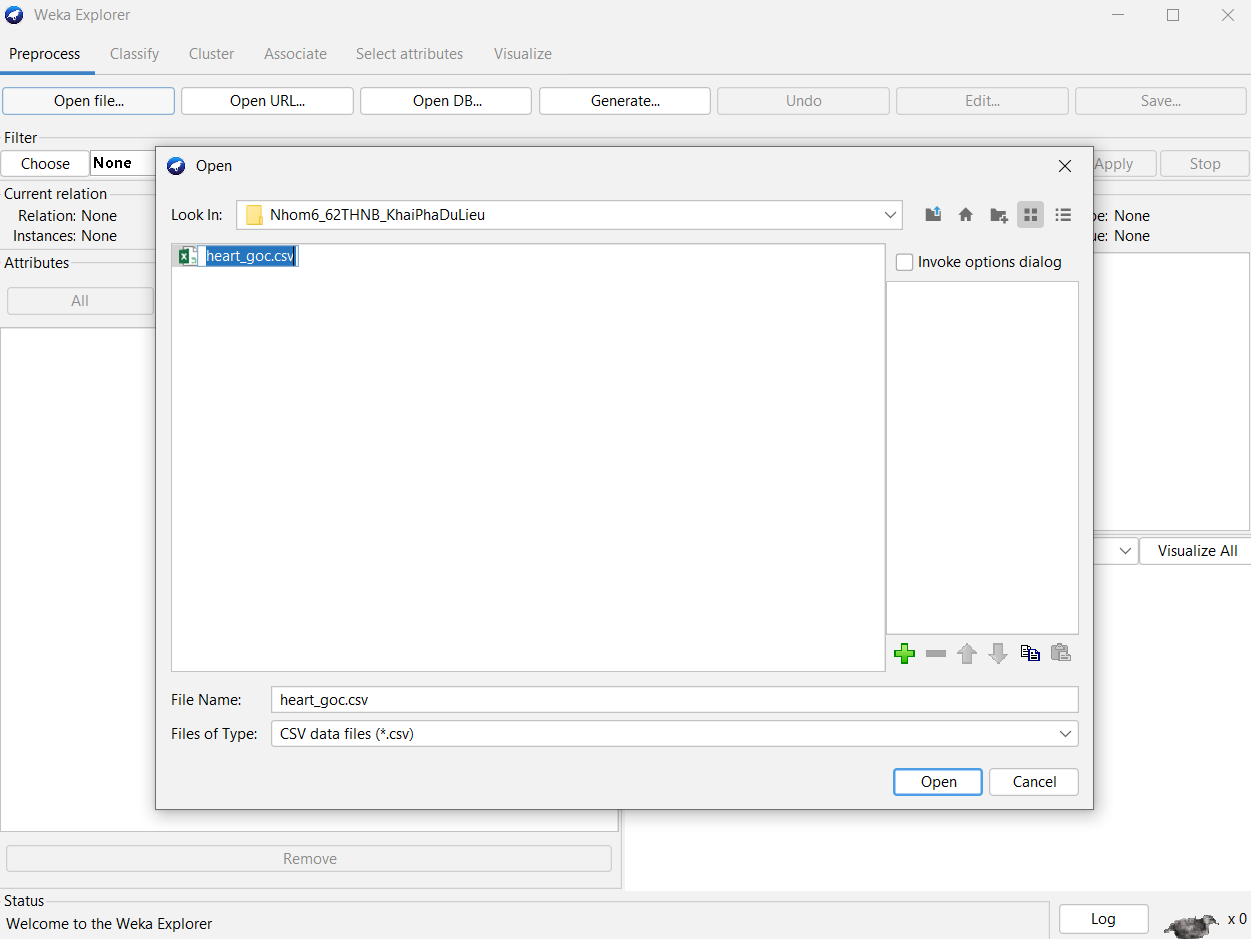
*Ảnh minh họa dữ liệu thô*

## **Tiền xử lý dữ liệu bằng Weka**

### **Nạp dữ liệu**

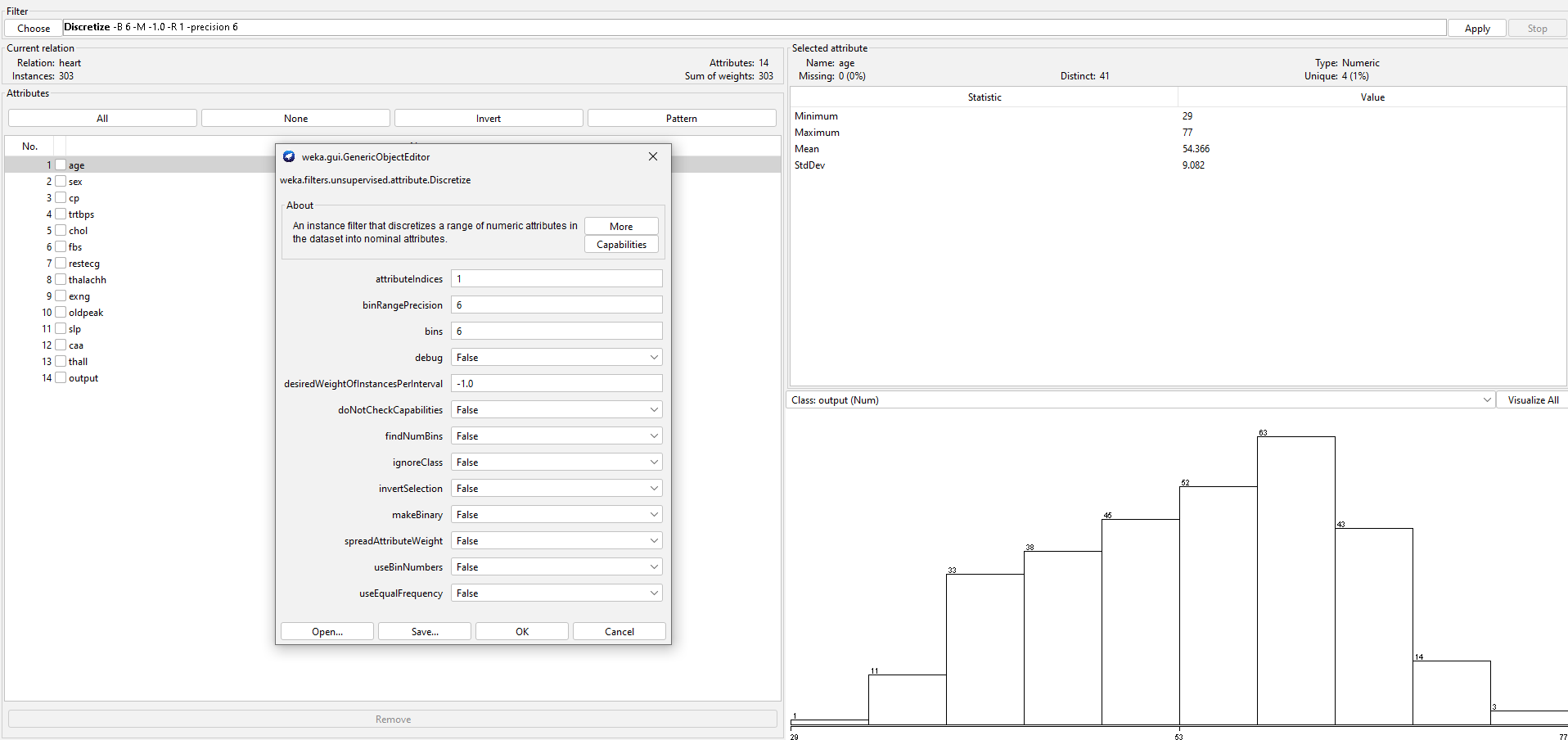
* Bước 1: Tại tab Preprocessing, click vào nút Open file
* Bước 2: Tìm đến thư mục lưu file **heart\_goc.csv** và chọn Open

(Lưu ý: Trường File of Type để là: CSV data file (“.CSV”)

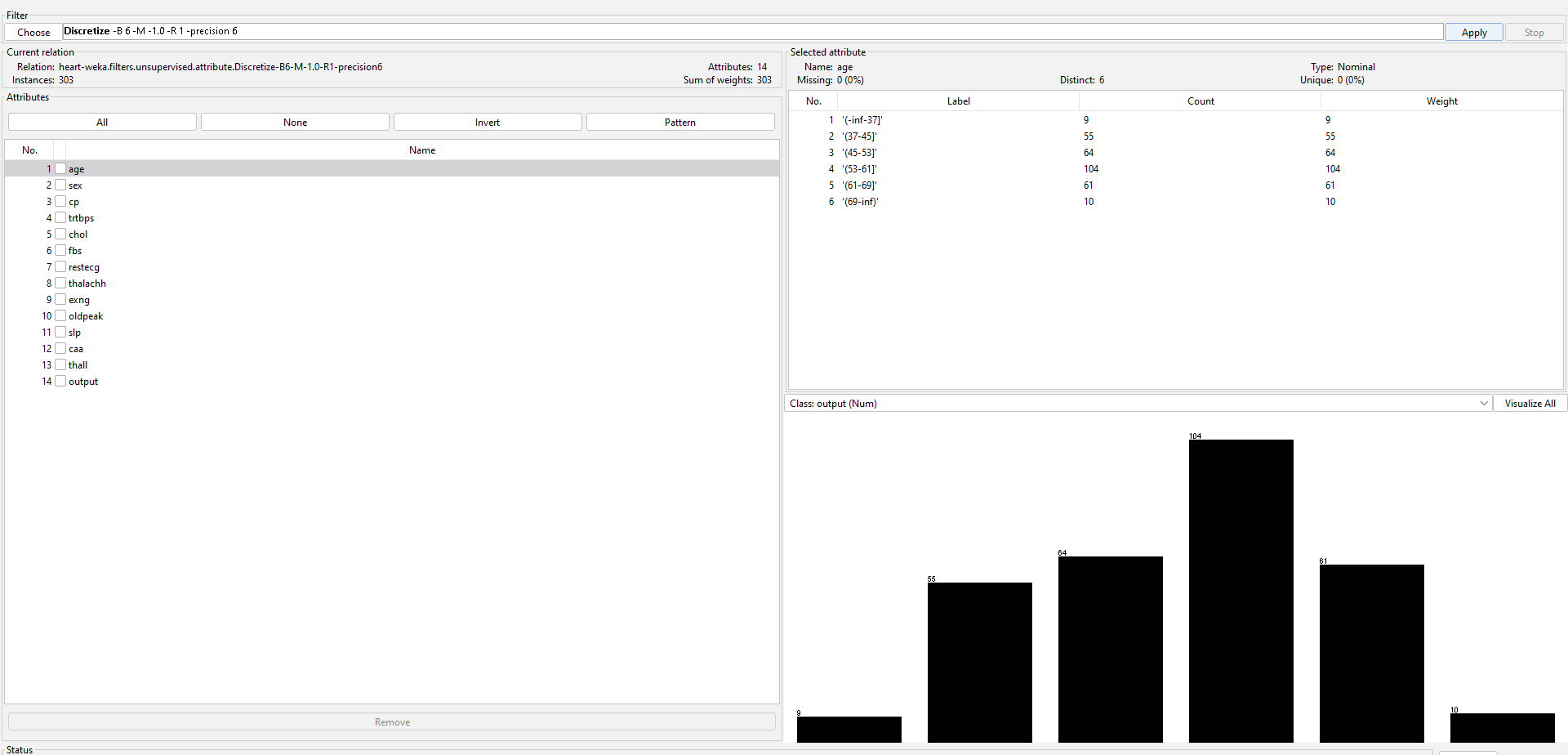


## **Rời rạc hóa dữ liệu**

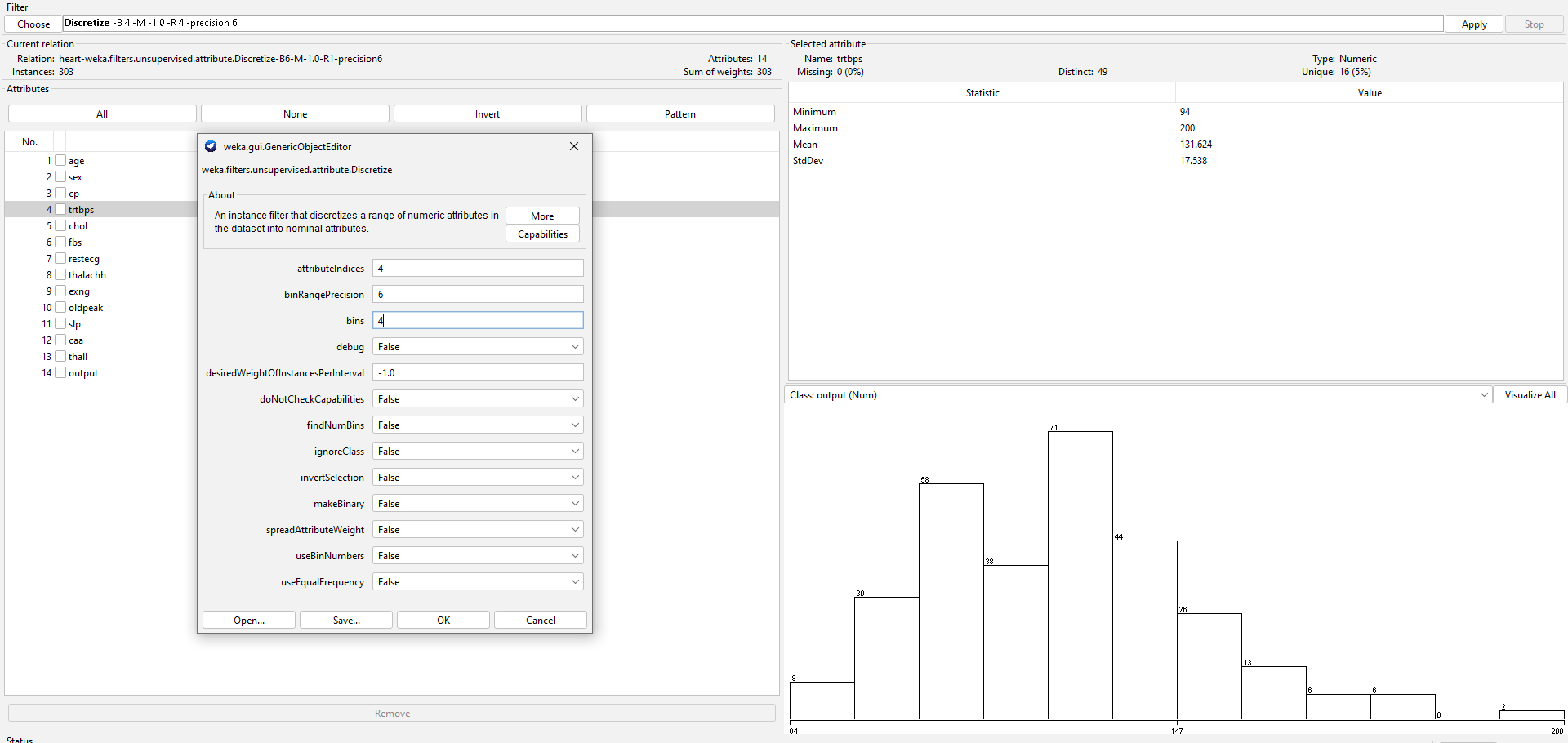
* Vấn đề: Khi dữ liệu đầu vào là liên tục, việc áp dụng thuật toán Bayes có thể khó khăn vì phải tính toán xác suất liên tục. Trong trường hợp này, rời rạc hóa dữ liệu có thể được sử dụng để chuyển đổi các biến liên tục thành các biến rời rạc trước khi áp dụng thuật toán Bayes. Việc chuyển đổi này giúp thuật toán Bayes dễ dàng tính toán xác suất cho từng khoảng dữ liệu rời rạc thay vì tính toán trên dữ liệu liên tục.
* Nhận xét: ta thấy các thuộc tính : “age”, “trtbps”, “chol”, “thalach”, “oldpeak” dầu vào là các giá trị liên tục. Ta sẽ chuyển đổi thành các biến rời rạc
* Age:



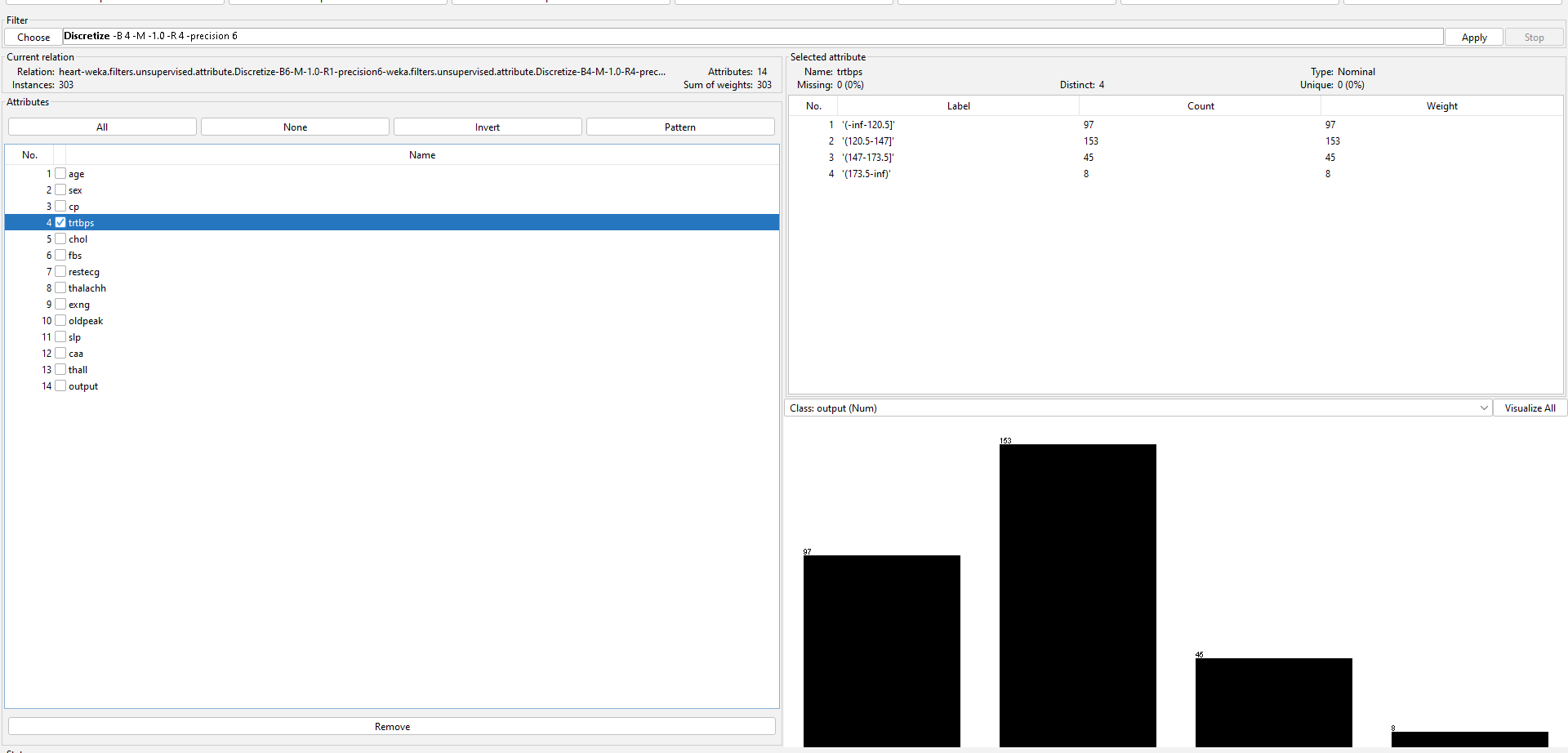
* Thuộc tính Age được chia thành 6 bins:



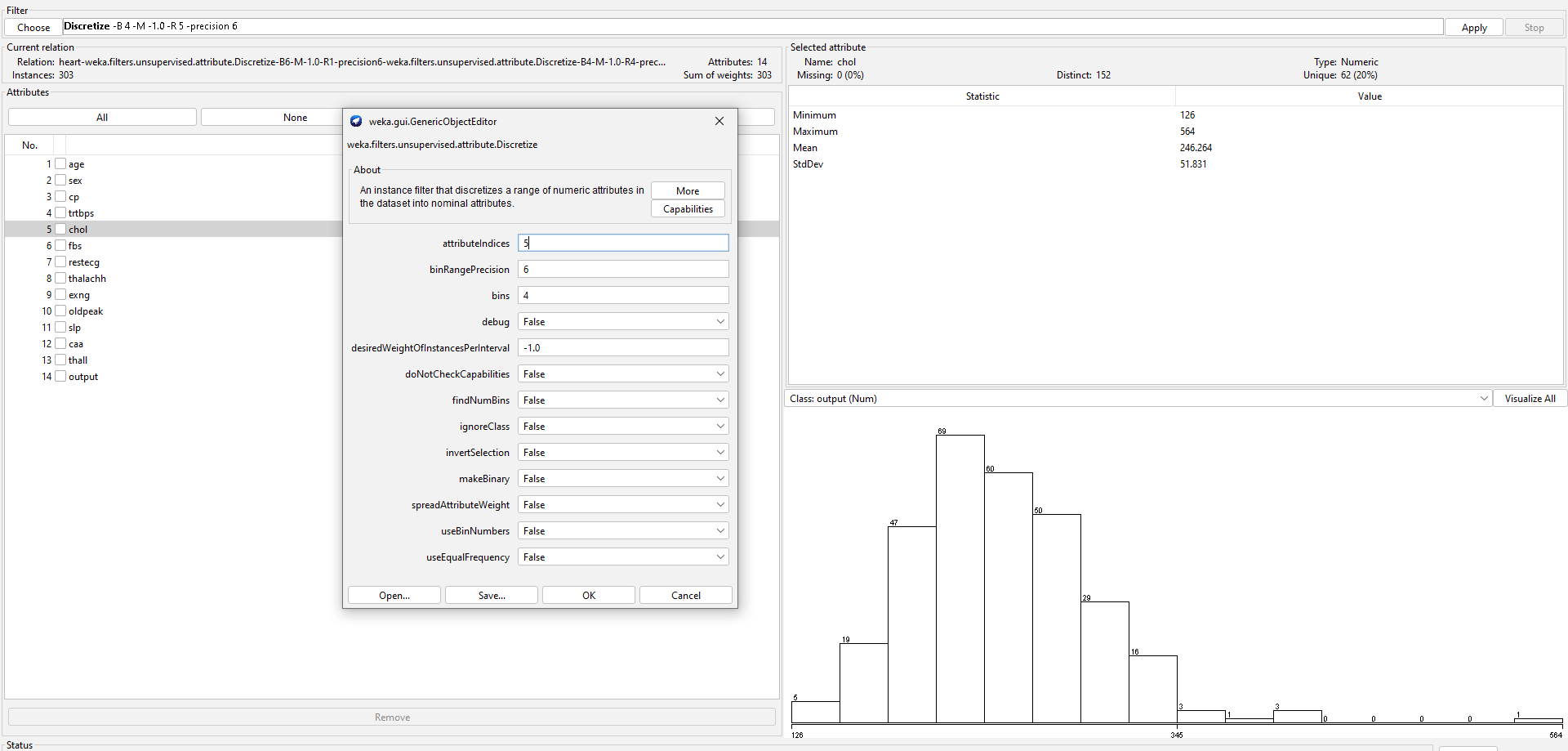
* Trtbps:



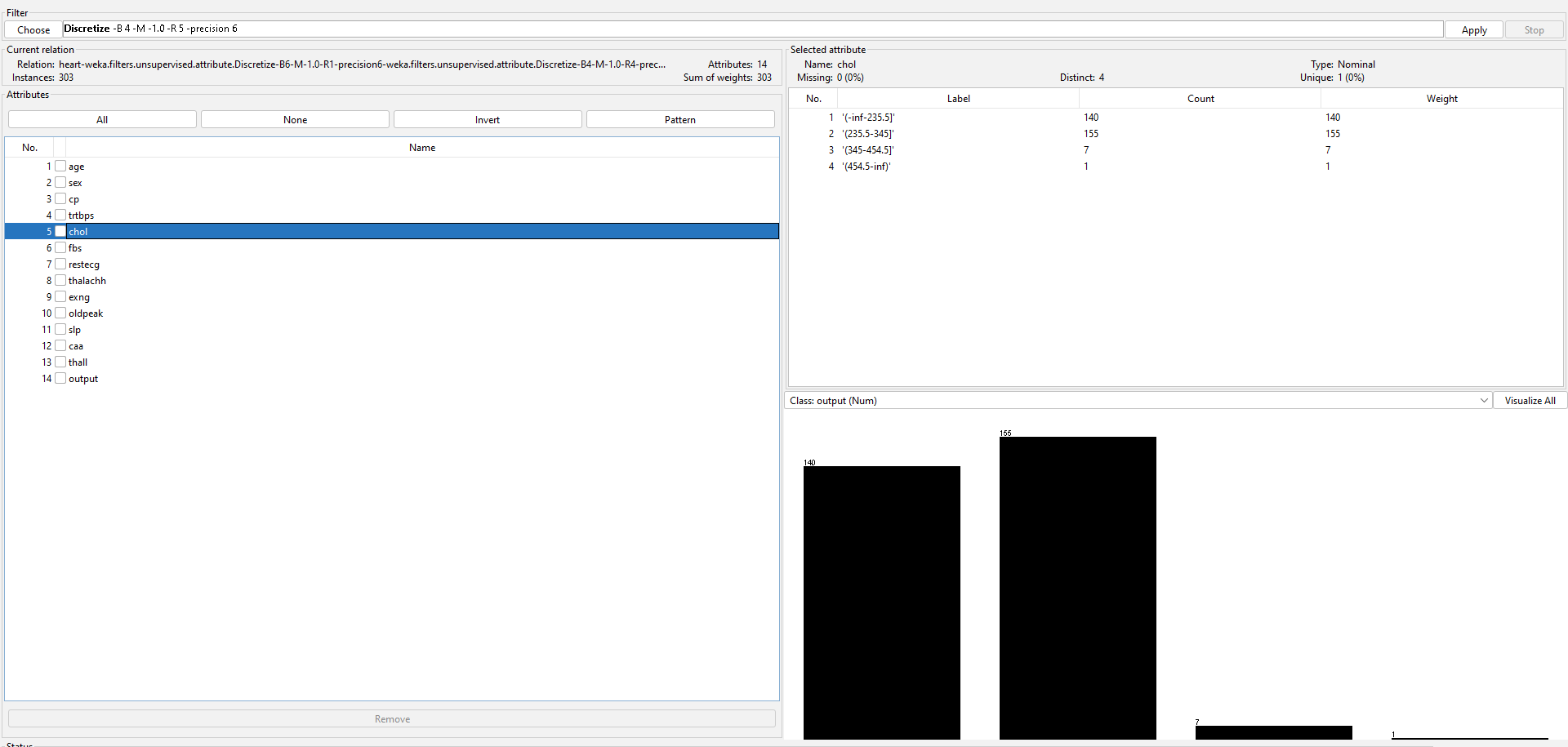
* Thuộc tính Trbps được chia thành 4 bins:



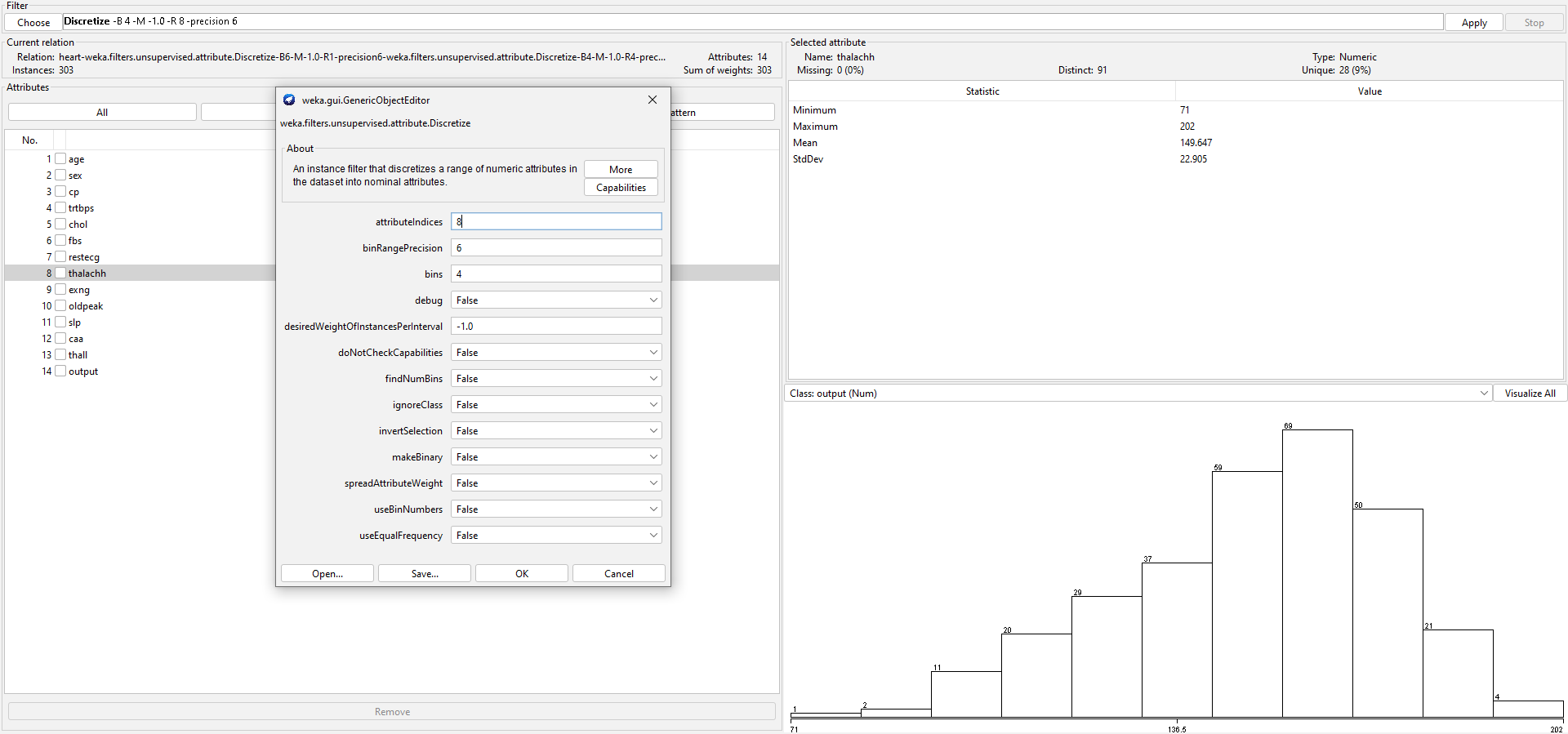
* Chol:



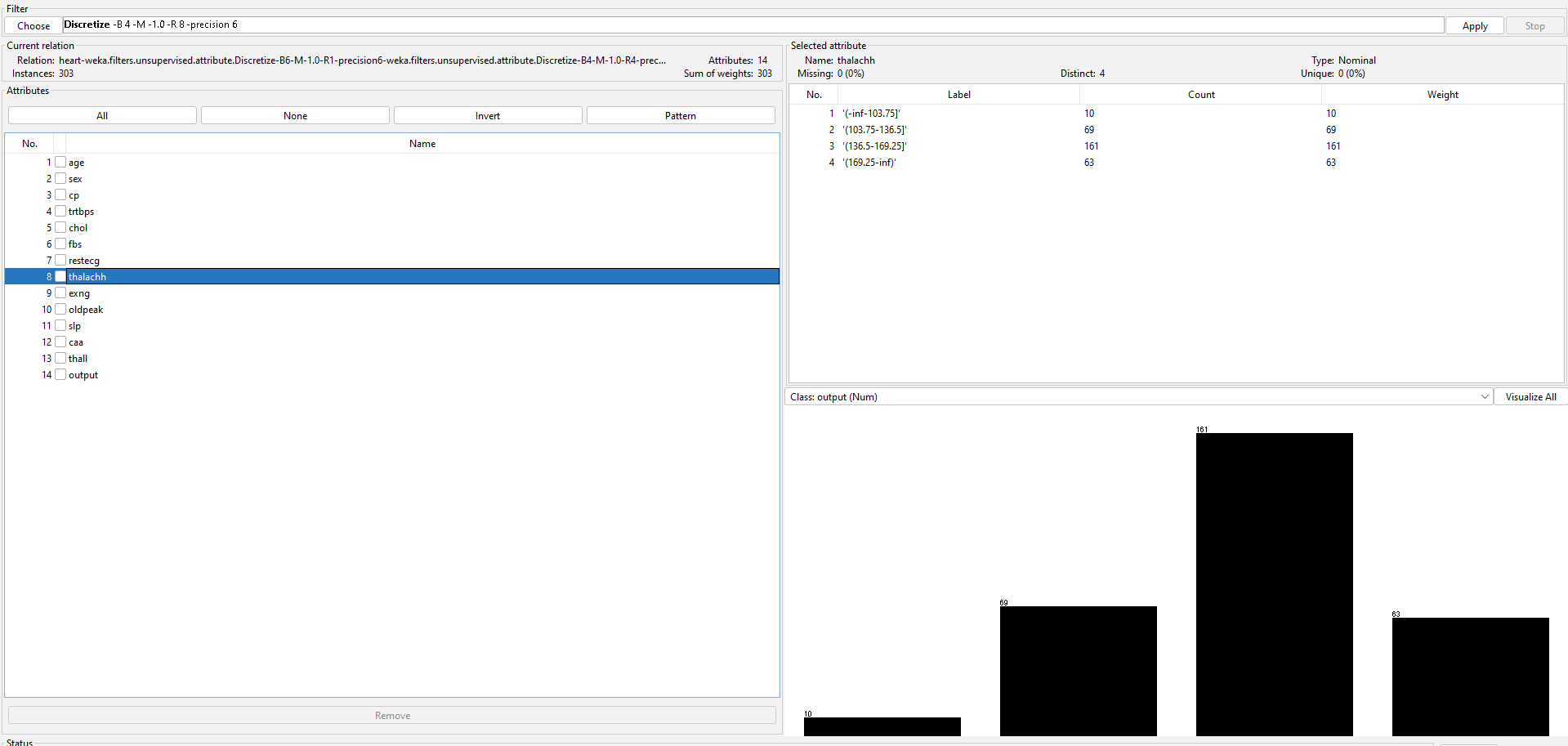
* Thuộc tính Chol được chia thành 4 bins:



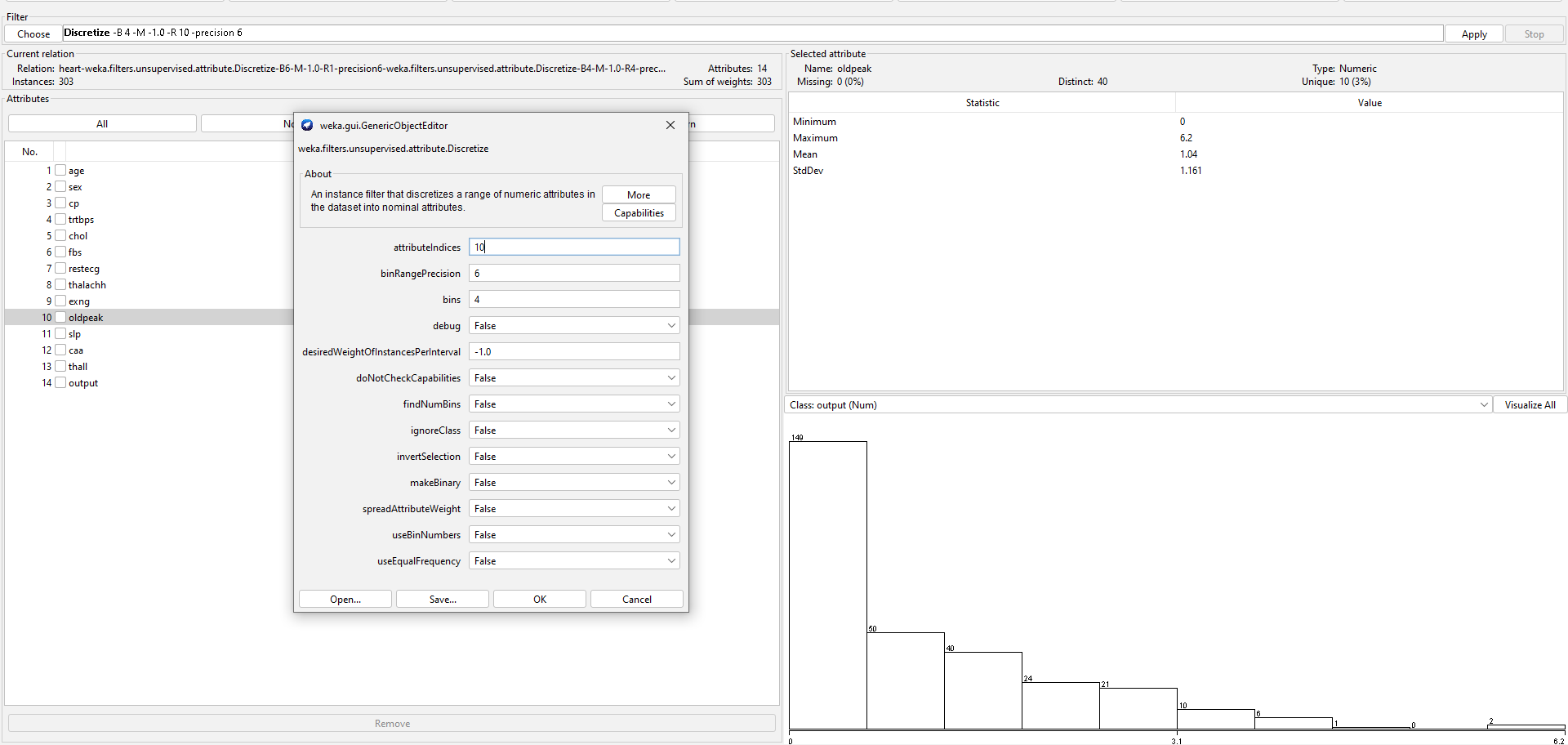
* Thalach:



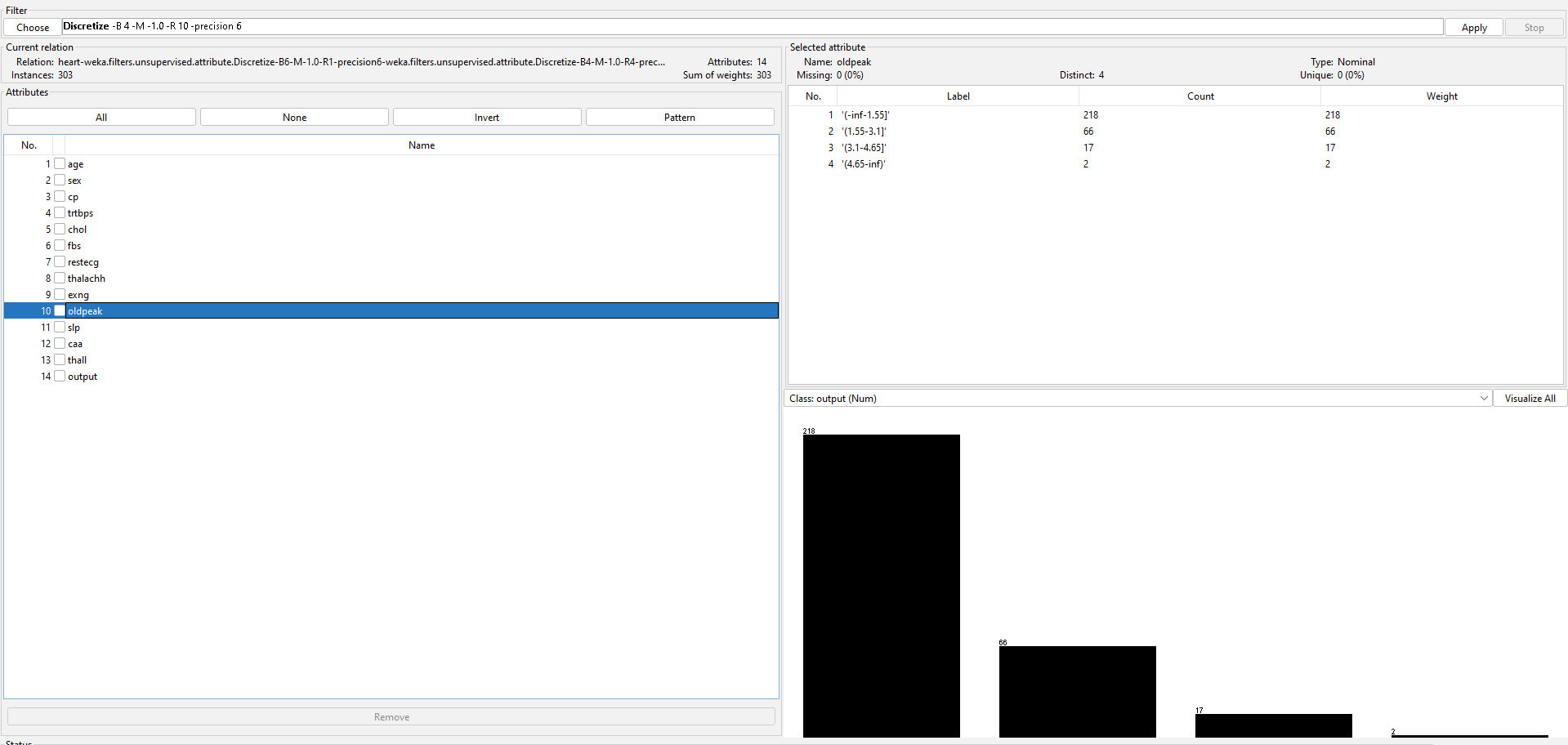
* Thuộc tính Thalach được chia thành 4 bins:



* Oldpeak:

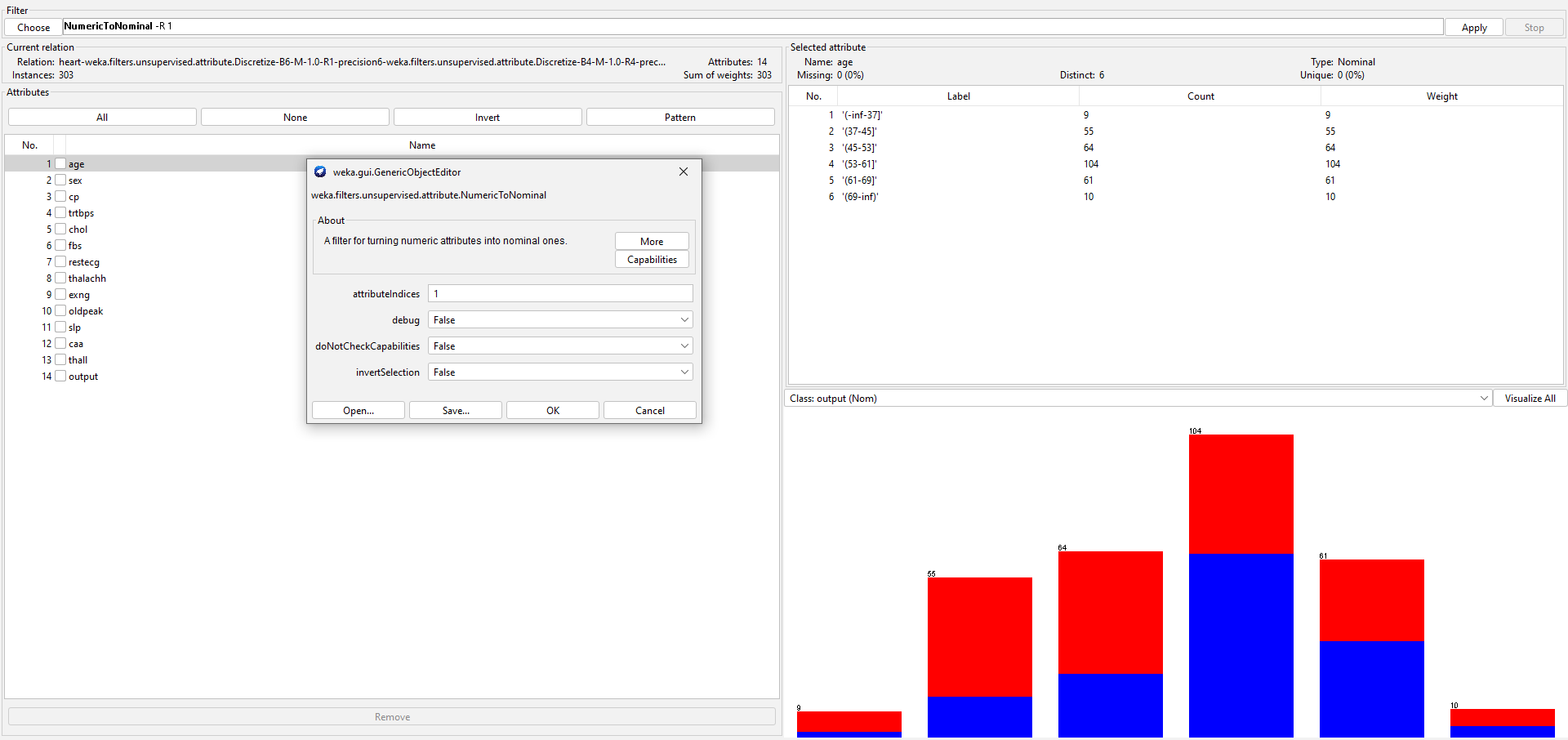


* Thuộc tính Oldpeak được chia thành 4 bins:

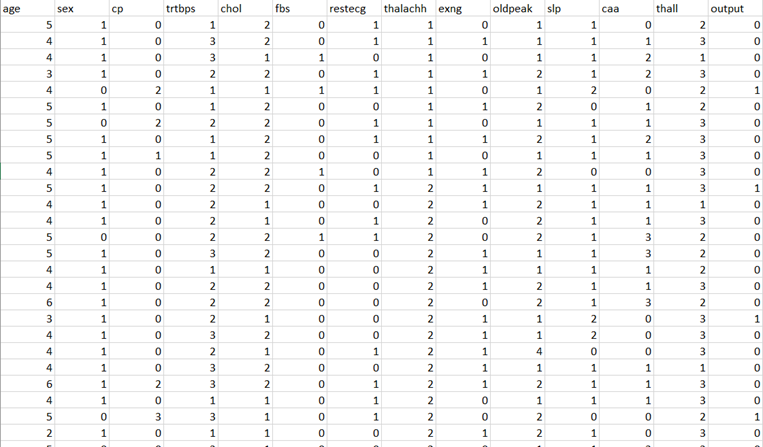


### **Chuyển đổi dữ liệu.**

* Một số thuật toán phân loại (Bayes) yêu cầu dữ liệu đầu vào là dữ liệu thể hiện dưới dạng các thuộc tính nomial (thuộc tính phân loại, không phải thuộc tính số). Vì vậy, nếu bạn muốn sử dụng các thuật toán này trong Weka, bạn cần phải chuyển đổi các thuộc tính số sang dạng nomial để thuật toán có thể sử dụng được.



## **Dữ liệu sau khi xử lý**



# **Chương 3: Thuật toán Naive Bayes**

## **3.1 Thuật toán Bayes**

## **Khái niệm**

Bayes là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân lớp. Tư tưởng chính của thuật toán Bayes là tìm cách phân nhóm các đối tượng (objects) đã cho vào K lớp (K là số các lớp được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

### **Quy trình khai phá dữ liệu**

* ***Bước 1***: Trích chọn dữ liệu (data selection): Là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases,data warehouses).
* ***Bước 2*** : Tiền xử lý dữ liệu (data preprocessing): Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,...), rút gọn dữ liệu (sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu,...), rời rạc hóa dữ liệu (dựa vào histogram, entropy, phân khoảng,…). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và được rời rạc hóa.
* ***Bước 3***: Biến đổi dữ liệu (data transformation): Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai phá ở bước sau.
* ***Bước 4***: Khai phá dữ liệu (data mining): Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình khám phá tri thức, áp dụng các kỹ thuật khai phá (phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.
* ***Bước 5***: Đánh giá và biểu diễn tri thức (knowledge representation & evaluation): Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin (tri thức) và mối liên hệ đặc biệt trong dữ liệu đã được khai phá ở bước trên biểu diễn theo dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật,… Đồng thời, bước này cũng đánh giá những tri thức khai phá được theo những tiêu chí nhất định.
* Trong giai đoạn KPDL, có thể cần sự tương tác của người dùng để điều chỉnh và rút ra các tri thức cần thiết nhất. Các **tri thức** nhận được cũng có thể được lưu và sử dụng lại.

### **Các bước thuật toán Bayes**

* **Định lý Bayes được phát biểu như sau:**

**𝑃 (𝑌| 𝑋) = 𝑃(𝑋|𝑌).𝑃(𝑌) /𝑃(𝑋)**

* P(Y): Xác suất của sự kiện Y xảy ra
* P(X): Xác suất của sự kiện X xảy ra
* P(X|Y): Xác suất (có điều kiện) của sự kiện X xảy ra, nếu biết rằng sự kiện Y đã xảy ra
* P(Y|X): Xác suất (có điều kiện) của sự kiện Y xảy ra, nếu biết rằng sự kiện X đã xảy ra
* **Thuật toán Naive Bayes thực hiện qua các bước chính sau:**
* Bước 1:Huấn luyện Bayes (dựa vào tập dữ liệu)

+Tính xác suất P(Ci)

+Tính xác suất P(Xk|Ci)

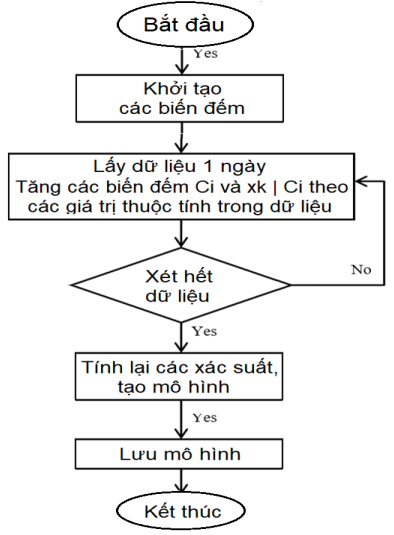
* Bước 2: Phân lớp Xnew

+Tính F(Xnew,Ci) =𝑃(𝐶𝑖) 𝑘=1𝑛P(𝑥𝑘|𝐶𝑖)

+Xnew được gán vào lớp Cq saocho

𝐹𝑋𝑛𝑒𝑤,𝐶𝑞=max(𝐹𝑋𝑛𝑒𝑤,𝐶𝑖)

* **Sơ đồ thuật toán**



*Sơ đồ thuật toán Naive Bayes*

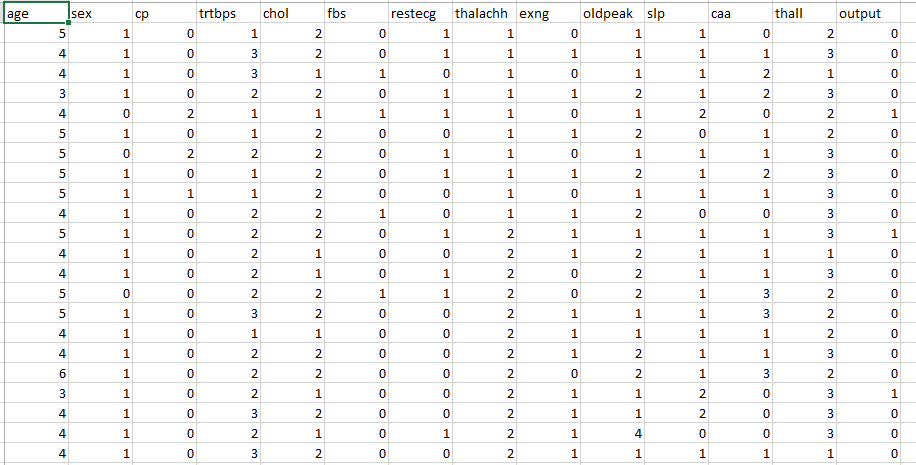
## **3.2 Mô tả bài toán**

## **Bài toán**

* Phân lớp khả năng mắc cơn đau tim bằng phương pháp Bayes.
* Phương pháp sử dụng: Phân lớp Bayes.

### **Dữ liệu đầu vào Input**

Dữ liệu đầu vào của bài toán được lấy từ Kaggle



**Mô tả tập dữ liệu bao gồm các thuộc tính:**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Mô tả** |
|  | Age | Tuổi của đối tượng |
|  | Sex | Giới tính của đối tượng. Trong đó: - 1: nam - 0: nữ |
|  | Cp | Loại đau ngực. Trong đó:   * 0: đau thắt ngực điển hình * 1: đau thắt ngực không điển hình * 2: đau không đau thắt ngực * 3: không có triệu chứng |
|  | Trtbps | Huyết áp khi nghỉ ngơi (tính bằng mmHg) |
| 1. e | Chol | chỉ số cholesterol tính bằng mg/dl được tải qua cảm biến BMI |
| 6. | Fbs | Đường huyết lúc đói > 120 mg/dl. Trong đó   * 1: có * 0: không |
| 7. | Restecg | Kết quả điện tâm đồ khi nghỉ ngơi. Trong đó   * 0: Bình thường * 1: Có sóng ST-T bất thường (sóng T đảo ngược và/hoặc ST chênh lên hoặc xuống > 0.05 mV) * 2: Cho thấy phì đại thất trái có thể xảy ra hoặc xác định theo tiêu chí của Estes |
| 8. | Thalachh | Nhịp tim đạt tối đa |
| 9. | Exng | Đau thắt ngực do gắng sức. Trong đó:   * 1: Có * 0: Không |
| 10. | Oldpeak | Đỉnh trước điện tâm đồ |
| 11. | Slp | Dốc của điện tâm đồ |
| 12. | Caa | Số lượng mạch chính |
| 13. | Thall | Tỷ lệ thal |
| 14. | Output | Kết quả chẩn đoán. Trong đó:   * 0: Ít nguy cơ đau tim * 1: Nhiều nguy cơ đau tim |

### **Dữ liệu đầu ra Output**

Sử dụng dữ liệu về khả năng mắc cơn đau tim từ đó dự đoán khả năng mắc cơn đau tim của một đối tượng nào đó

* 0: Ít nguy cơ đau tim
* 1: Nhiều nguy cơ đau tim

### **Quá trình xử lý**

Gồm các bước như sau:

* Thu thập dữ liệu: Đầu tiên, ta thu thập và chọn lọc dữ liệu để huấn luyện mô hình. Dữ liệu này sẽ phải được gán nhãn (label) để mô hình có thể học được cách phân loại các đối tượng.
* Tiền xử lý dữ liệu: Sau khi thu thập dữ liệu, ta tiến hành tiền xử lý nhằm loại bỏ các dữ liệu không cần thiết, xử lý dữ liệu thiếu, chuẩn hóa dữ liệu, ..
* Chọn mô hình phân loại: Tiếp theo, ta chọn một mô hình phân loại phù hợp với bài toán và huấn luyện nó với dữ liệu đã tiền xử lý.
* Đánh giá mô hình: Sau khi huấn luyện xong, ta đánh giá hiệu suất của mô hình bằng cách đo lường độ chính xác (accuracy) hoặc các độ đo khác.
* Sử dụng mô hình: Cuối cùng, ta sử dụng mô hình đã huấn luyện để phân loại đối tượng mới. Mô hình sẽ nhận đầu vào là một đối tượng, tiến hành xử lý và trả về nhãn (label) tương ứng.

## **3.3 Giải bài toán bằng thuật toán Bayes**

## **Tiền xử lý để đưa ra những thuộc tính có giá trị liên tục về 1 khoảng**

a. Age (29-77: 6 khoảng):

* <37: 1
* 38-45: 2
* 46-53: 3
* 54-61: 4
* 62-69: 5
* >70 : 6

b. Trtbps (94-200: 4 khoảng):

* <120: 1
* 121-147: 2
* 148-173: 3
* >174: 4

c. Chol (126-564: 4 khoảng):

* <235: 1
* 236-345: 2
* 346-454: 3
* >455: 4

d. Thalach (71-202: 4 khoảng):

* <103: 1
* 104-137: 2
* 138-169: 3
* >170: 4

e. Oldpeak (0-6.2: 4 khoảng):

* [0-1.55]: 1
* (1.55-3.1]: 2
* (3.1-4.65]: 3
* (4.65-max): 4

### **Tiến hành giải bài toán**

* **Thuộc tính Age:**

|  |  |
| --- | --- |
| * P( 1| 0) =2/138 * P( 1| 1)=7/165 * P(2| 0)= 14/138 * P(2|1)=41/165 * P(3|0)= 22/138 * P(3|1)=42 /165 | * P(4|0)= 63/138 * P(4|1)= 41/165 * P(5|0)= 33/138 * P(5|1)= 28/165 * P(6|0)= 4/138 * P(6|1)= 6/165 |

* **Thuộc tính Sex:**
* P( 0| 0) = 24/138
* P( 0| 1) = 72/165
* P( 1|0) = 114/138
* P( 1| 1) =93/165
* **Thuộc tính Trtbps:**

|  |  |
| --- | --- |
| * P( 1| 0) =37/138 * P( 1| 1) =60/165 * P(2| 0)= 71/138 * P(2|1)=82/165 | * P(3|0)=24/138 * P(3|1)=21/165 * P(4|0)=6/138 * P(4|1)=2/165 |

* **Thuộc tính Chol:**

|  |  |
| --- | --- |
| * P( 1| 0) =55/138 * P( 1| 1) =85/165 * P(2| 0)= 80/138 * P(2|1)=75/165 | * P(3|0)=3/138 * P(3|1)=4/165 * P(4|0)=0/138 * P(4|1)=1/165 |

* **Thuộc tính Fbs:**

• P( 0| 0) =116/138

• P( 0| 1) =142/165

• P(1| 0)= 22/138

• P(1|1)=23/165

* **Thuộc tính Rectegr:**

• P( 0| 0) =79/138

• P( 0| 1) =68/165

• P(1| 0)= 56/138

• P(1|1)=96/165

* **Thuộc tính CP:**

|  |  |
| --- | --- |
| • P( 0| 0) =104/138  • P( 0| 1) =39/165  • P(1| 0)= 9/138  • P(1|1)=41/165 | • P(2| 0)= 18/138  • P(2|1)=69/165  • P(3|0)=7/138  • P(3|1)=16/165 |

* **Thuộc tính Thaclachh:**

|  |  |
| --- | --- |
| • P( 1| 0) =9/138  • P( 1| 1) =1/165  • P(2| 0)= 50/138  • P(2|1)=19/165 | • P(3|0)=69/138  • P(3|1)=92/165  • P(4|0)=10/138  • P(4|1)=53/165 |

* **Thuộc tính Exng:**

• P( 0| 0) = 62/138

• P( 0| 1) = 142/165

• P( 1|0) = 76/138

• P( 1| 1) =23/165

* **Thuộc tính Oldpeak:**

|  |  |
| --- | --- |
| • P( 1| 0) =74/138  • P( 1| 1) =144/165  • P(2| 0)= 47/138  • P(2|1)=19/165 | • P(3|0)=15/138  • P(3|1)=2/165  • P(4|0)=2/138  • P(4|1)=0/165 |

* **Thuộc tính Slp:**

|  |  |
| --- | --- |
| • P( 0| 0) =12/138  • P( 0| 1) =9/165  • P(1| 0)= 91/138 | • P(1|1)=49/165  • P(2| 0)= 35/138  • P(2|1)=107/165 |

* **Thuộc tính Caa:**

|  |  |
| --- | --- |
| • P( 0| 0) =45/138  • P( 0| 1) =130/165  • P(1| 0)= 44/138  • P(1|1)=21/165 | • P(2| 0)= 31/138  • P(2|1)=7/165  • P(3|0)=17/138  • P(3|1)=3/165 |

* **Thuộc tính Thall:**

|  |  |
| --- | --- |
| • P( 0| 0) =1/138  • P( 0| 1) =1/165  • P(1| 0)= 12/138  • P(1|1)=6/165 | • P(2| 0)= 36/138  • P(2|1)=130/165  • P(3|0)=89/138  • P(3|1)=28/1653 |

# **Chương 4: Thực hiện giải thuật Bayes trên Weka**

## **4.1 Giới thiệu Weka**

Weka là một phần mềm mã nguồn mở mạnh mẽ và phổ biến trong lĩnh vực học máy và khai phá dữ liệu. Được phát triển tại Đại học Waikato ở New Zealand, Weka cung cấp một loạt các công cụ và thư viện để thực hiện nhiều nhiệm vụ trong lĩnh vực xử lý dữ liệu, phân loại, phân cụm, học tập có giám sát và không có giám sát, và nhiều ứng dụng khác liên quan đến học máy và khai phá dữ liệu.

Weka là lựa chọn lý tưởng cho cả những người mới bắt đầu và những chuyên gia trong lĩnh vực này. Nó có giao diện đồ họa thân thiện, giúp người dùng dễ dàng tạo, huấn luyện và đánh giá các mô hình học máy mà không cần phải viết mã nguồn. Đồng thời, Weka cũng hỗ trợ các ngôn ngữ lập trình như Java để người dùng có thể tùy chỉnh và mở rộng các tính năng theo nhu cầu riêng.

Với sự đa dạng và tính linh hoạt của nó, Weka đã trở thành một công cụ quan trọng trong cộng đồng nghiên cứu và công nghiệp trong việc giải quyết các vấn đề phức tạp liên quan đến dữ liệu và học máy.

## **4.2 Các tính năng của phần mềm Weka**

- *Mã nguồn mở và miễn phí*: Weka là một phần mềm mã nguồn mở, cho phép người dùng truy cập và sửa đổi mã nguồn theo nhu cầu. Điều này giúp cộng đồng phát triển và cải thiện Weka liên tục.

- *Giao diện đồ họa thân thiện:* Weka cung cấp một giao diện đồ họa dễ sử dụng cho người dùng, cho phép họ thao tác với các công cụ và mô hình học máy mà không cần kiến thức chuyên sâu về lập trình.

- *Hỗ trợ nhiều thuật toán học máy*: Weka đi kèm với một bộ sưu tập đa dạng các thuật toán học máy và khai phá dữ liệu. Người dùng có thể thử nghiệm và so sánh hiệu suất của nhiều thuật toán khác nhau.

- *Kết hợp nhiều công cụ:* Weka không chỉ là một công cụ học máy mà còn cung cấp nhiều tính năng liên quan đến tiền xử lý dữ liệu, trực quan hóa, và đánh giá mô hình. - *Tích hợp với ngôn ngữ lập trình Java:* Weka được xây dựng trên nền tảng Java, cho phép người dùng tích hợp dễ dàng với các ứng dụng Java khác và tận dụng khả năng mở rộng của ngôn ngữ này.

*- Hệ sinh thái mở rộng*: Cộng đồng Weka đã tạo ra nhiều plugin và phần mở rộng mạnh mẽ để bổ sung tính năng và thuật toán mới vào Weka, mở ra nhiều khả năng tùy chỉnh. - Học máy có giám sát và không có giám sát: Weka hỗ trợ cả học máy có giám sát (supervised learning) và không có giám sát (unsupervised learning), cho phép người dùng xử lý nhiều loại vấn đề khác nhau.

- *Hệ thống tài liệu và cộng đồng hỗ trợ:* Weka có một hệ thống tài liệu phong phú và một cộng đồng sẵn sàng giúp đỡ thông qua các diễn đàn và tài liệu hướng dẫn.

🡪 Nhờ những tính năng này, Weka trở thành một công cụ hữu ích cho những người làm việc trong lĩnh vực học máy, nghiên cứu dữ liệu, và khai phá dữ liệu để xử lý và phân tích dữ liệu một cách hiệu quả.

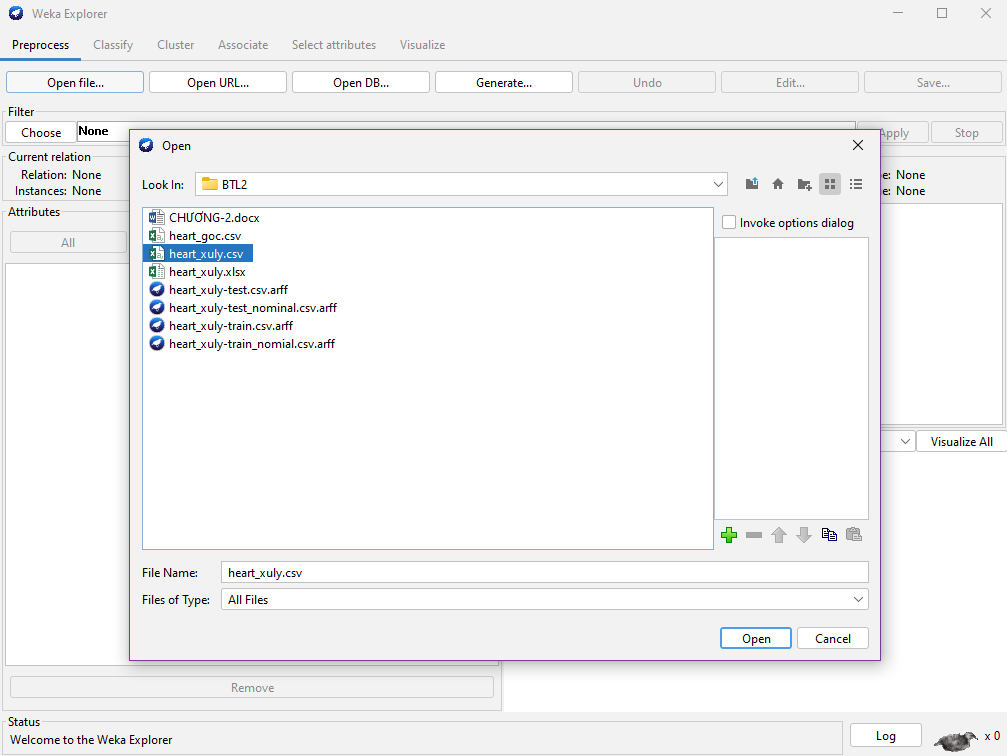
## **4.3 Các chức năng**

* Tiền xử lý dữ liệu
* Lựa chọn đặc trưng
* Huấn luyện mô hình học máy
* Khai phá dữ liệu không có giám sát
* Đánh giá mô hình
* Trực quan hóa dữ liệu
* Kết hợp với ngôn ngữ lập trình Java và các ngôn ngữ khác
* Tùy chỉnh và mở rộng
* Hỗ trợ tài liệu và cộng đồng

## **4.4 Giải trên Weka**

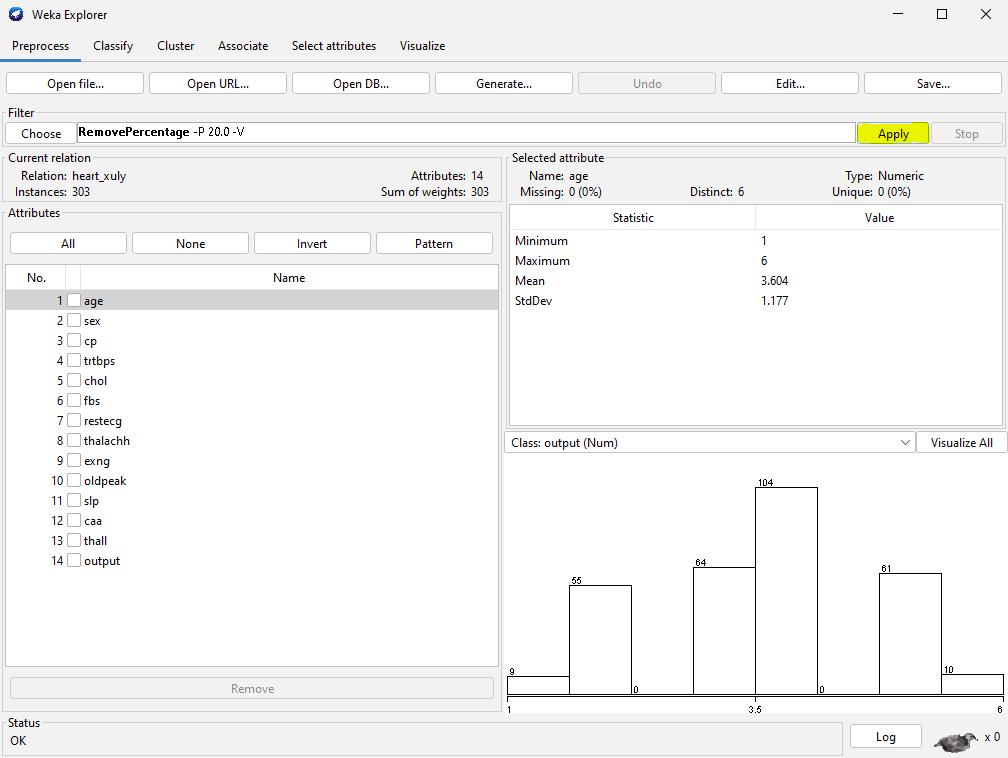
### **Chia tập Train**

* Bước 1: Mở file dữ liệu định dạng file CSV

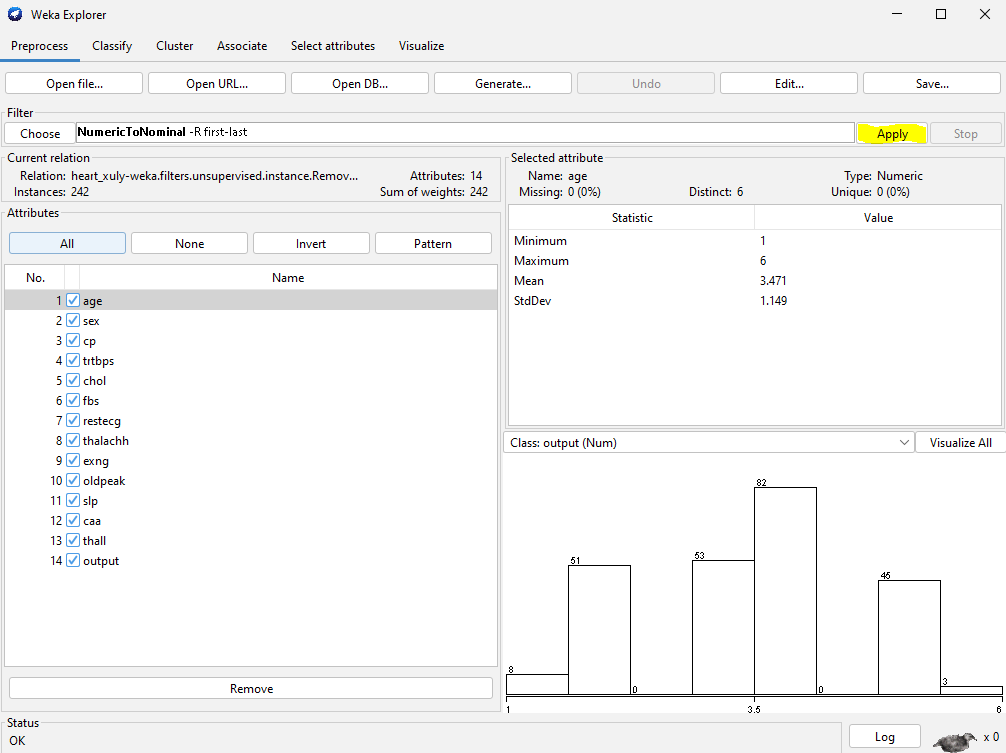


* Bước 2: Chia dữ liệu bằng RemovePercentage
* Nhập field Percentage 20.0 (loại 20% dữ liệu gốc): Tập train là 80%
* Chọn invertSelection: False
* Nhấn OK

+ Chọn “Apply”



* Bước 3: Chuyển đổi Numeric to Nomial



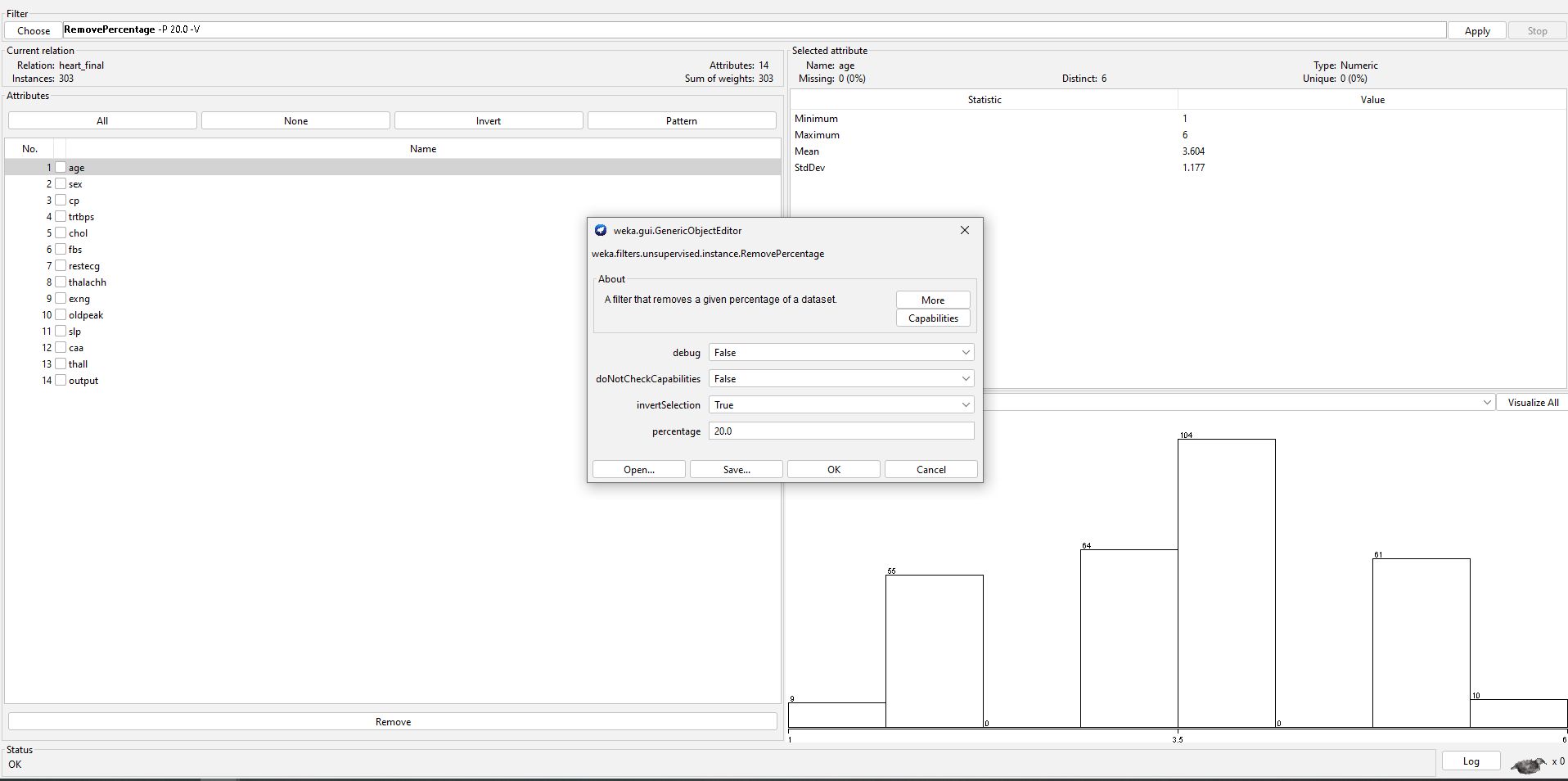
* Bước 4: Lưu file tên “**heart\_xuly-train\_nomial**”

Graphical user interface, text

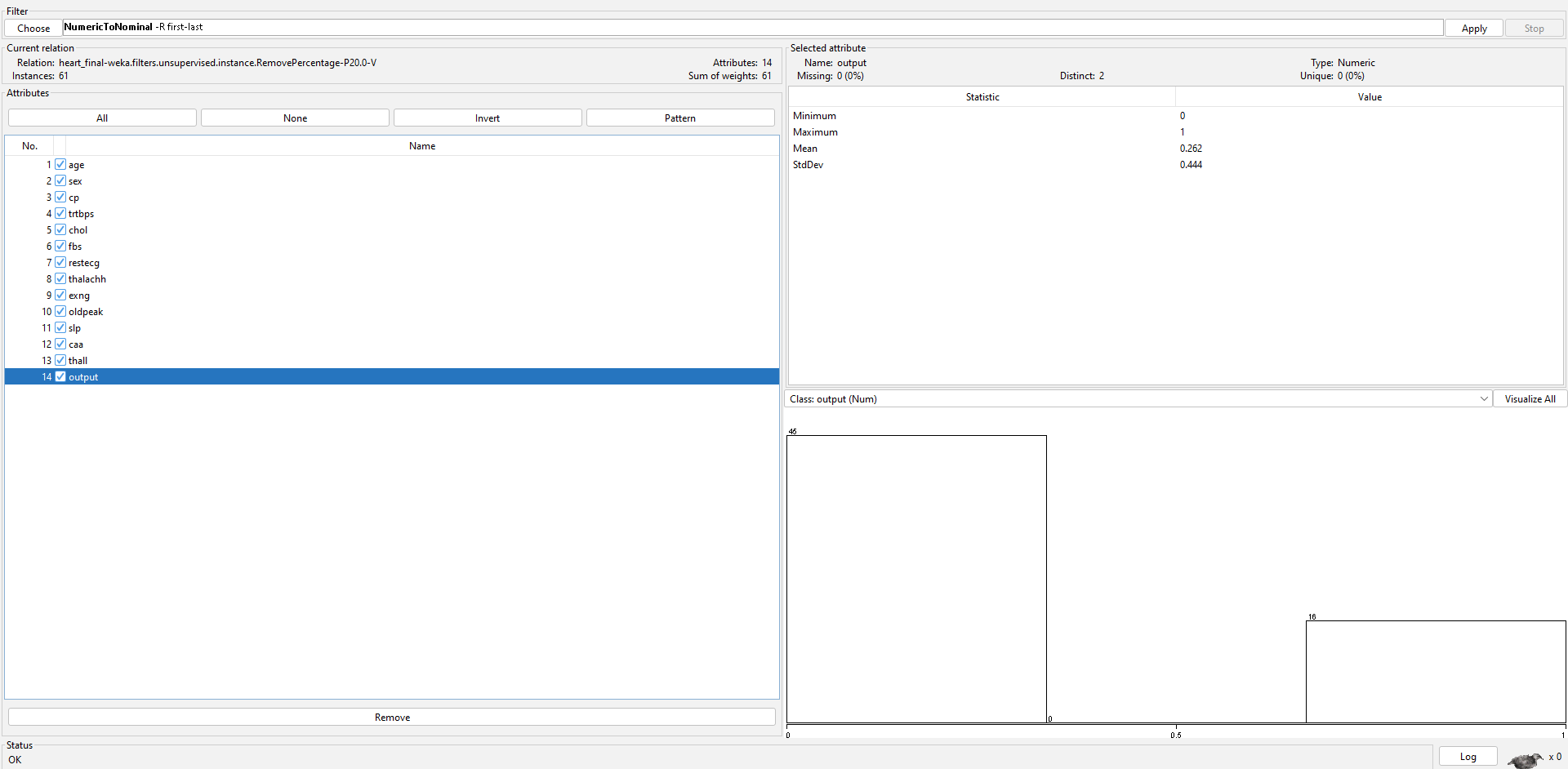
Description automatically generated

### **Chia tập Test**

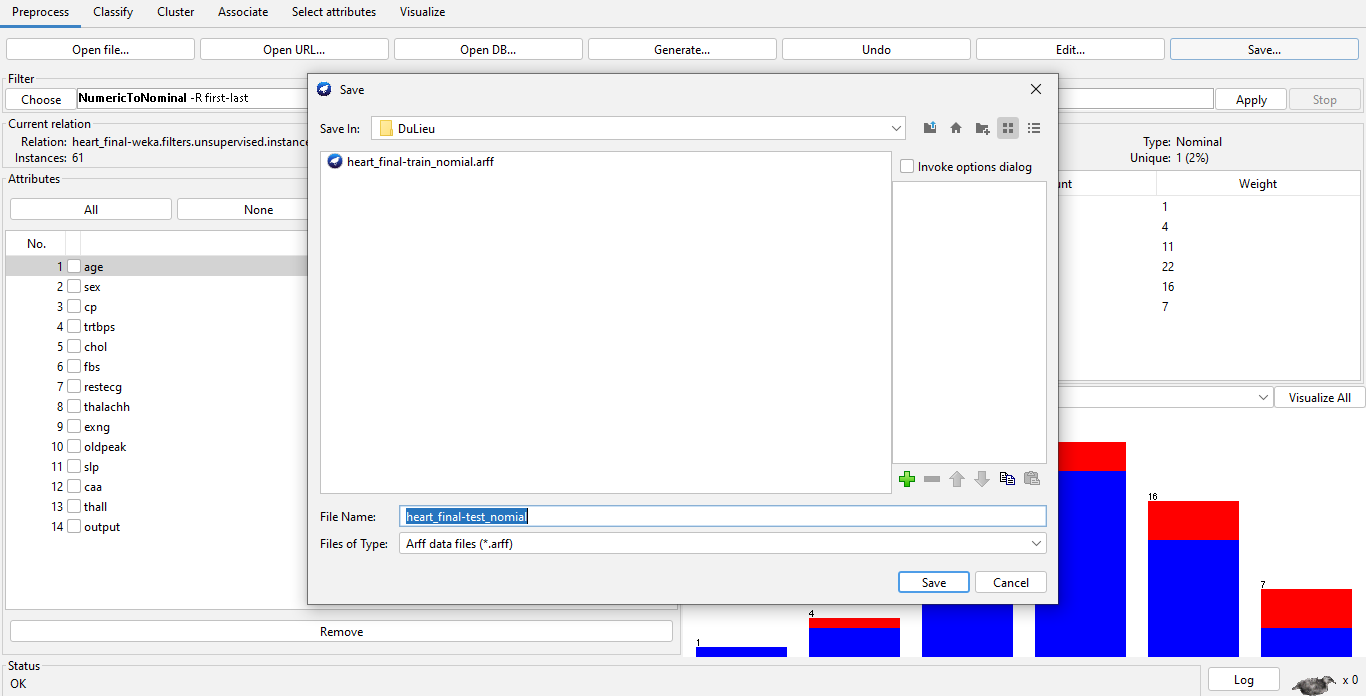
* Bước 1: Chia tập test bằng RemovePercentage
* Chọn invertSelection: True
* Nhập Percentage: 20%
* Chọn “Apple”



* Bước 2: Chuyển đổi Numeric to Nomial

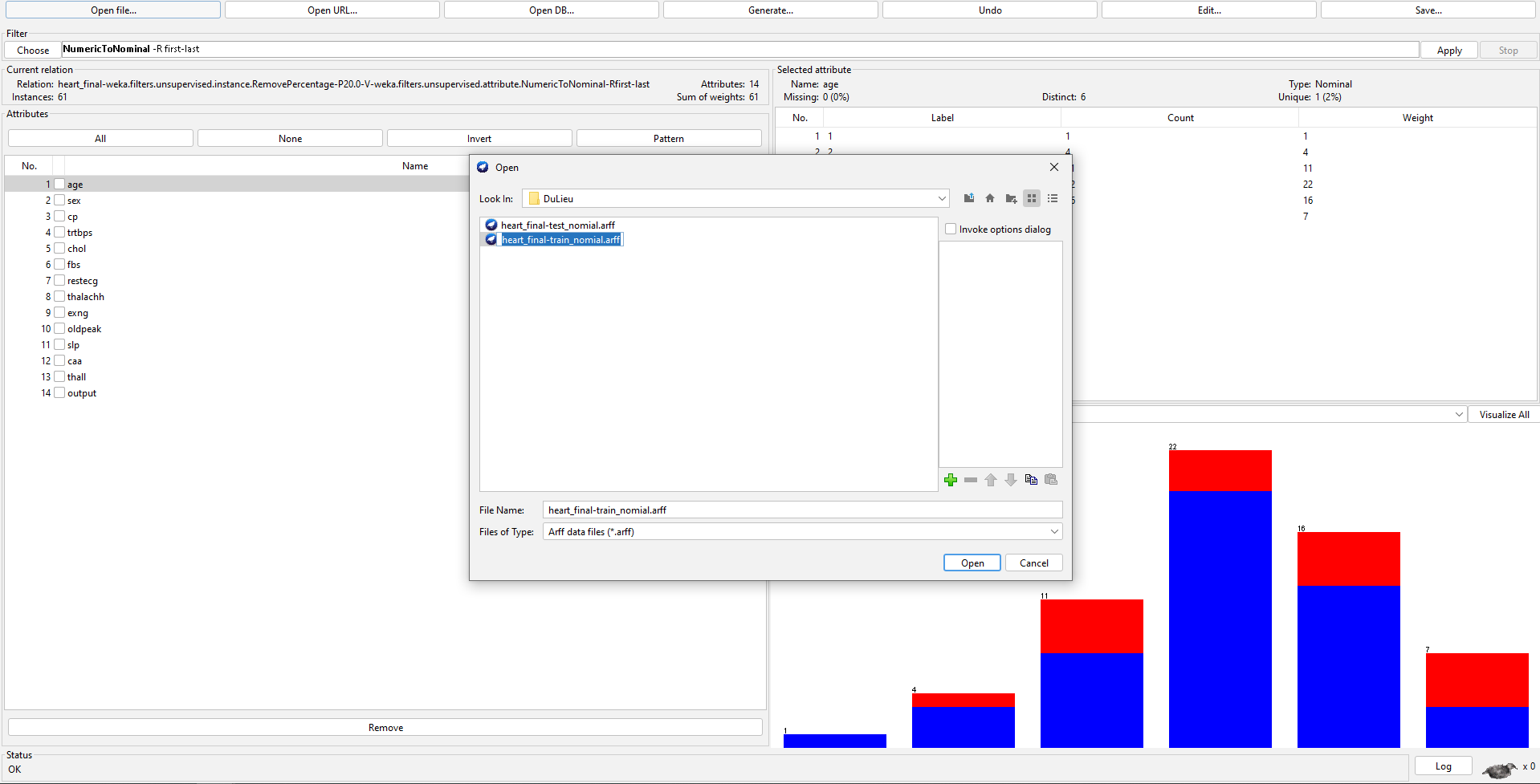


* Bước 3: Lưu file tên “**heart\_final-test\_nomial**”



### **Thực hiện huấn luyện Bayes**

* Bước 1: Mở file dữ liệu Train

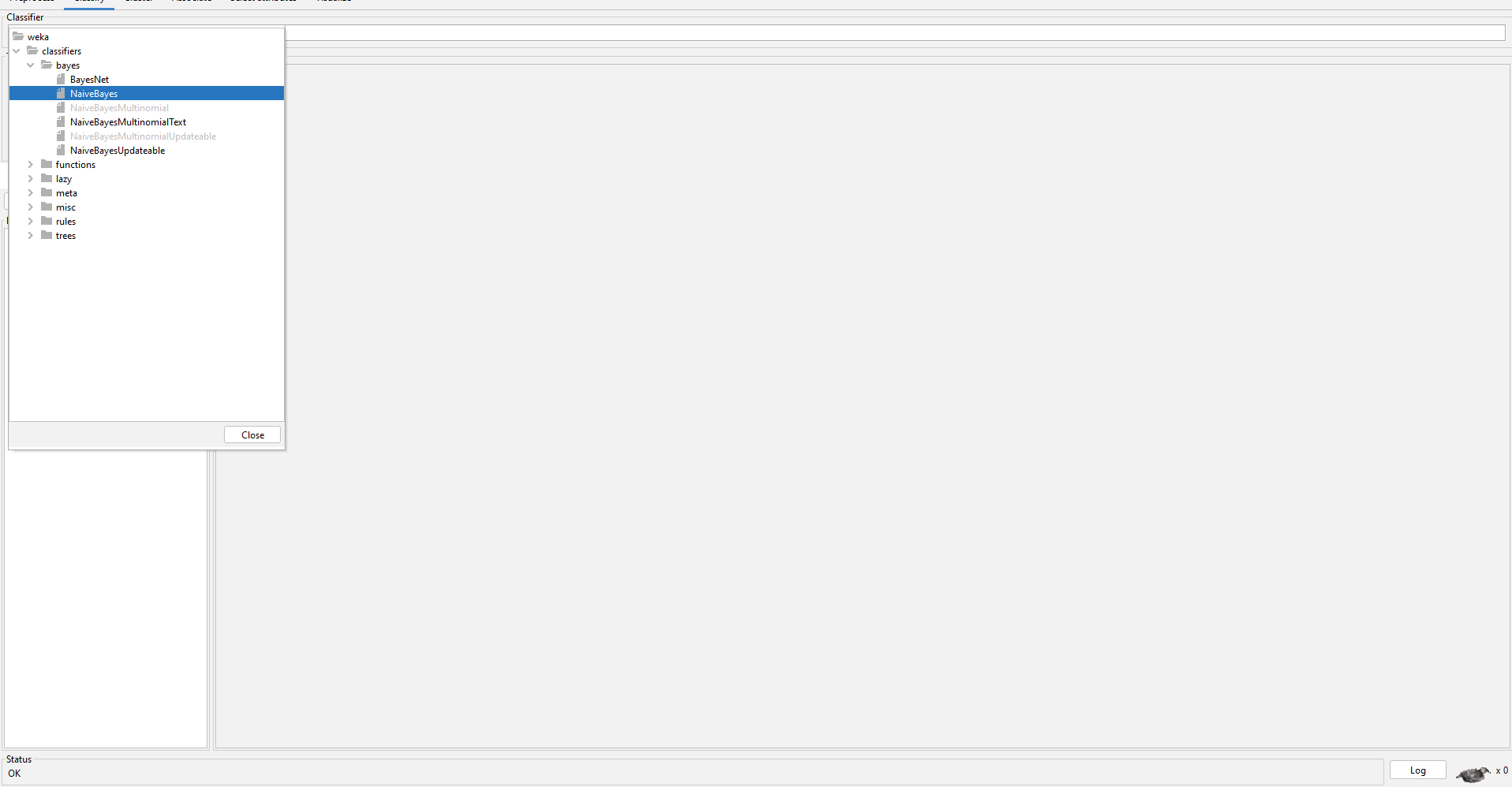


* Bước 2: Thực hiện phân lớp Bayes

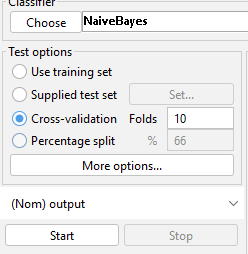
+ Chuyển cửa sổ Clasify

+ Chọn loại phân lớp NaiveBayes

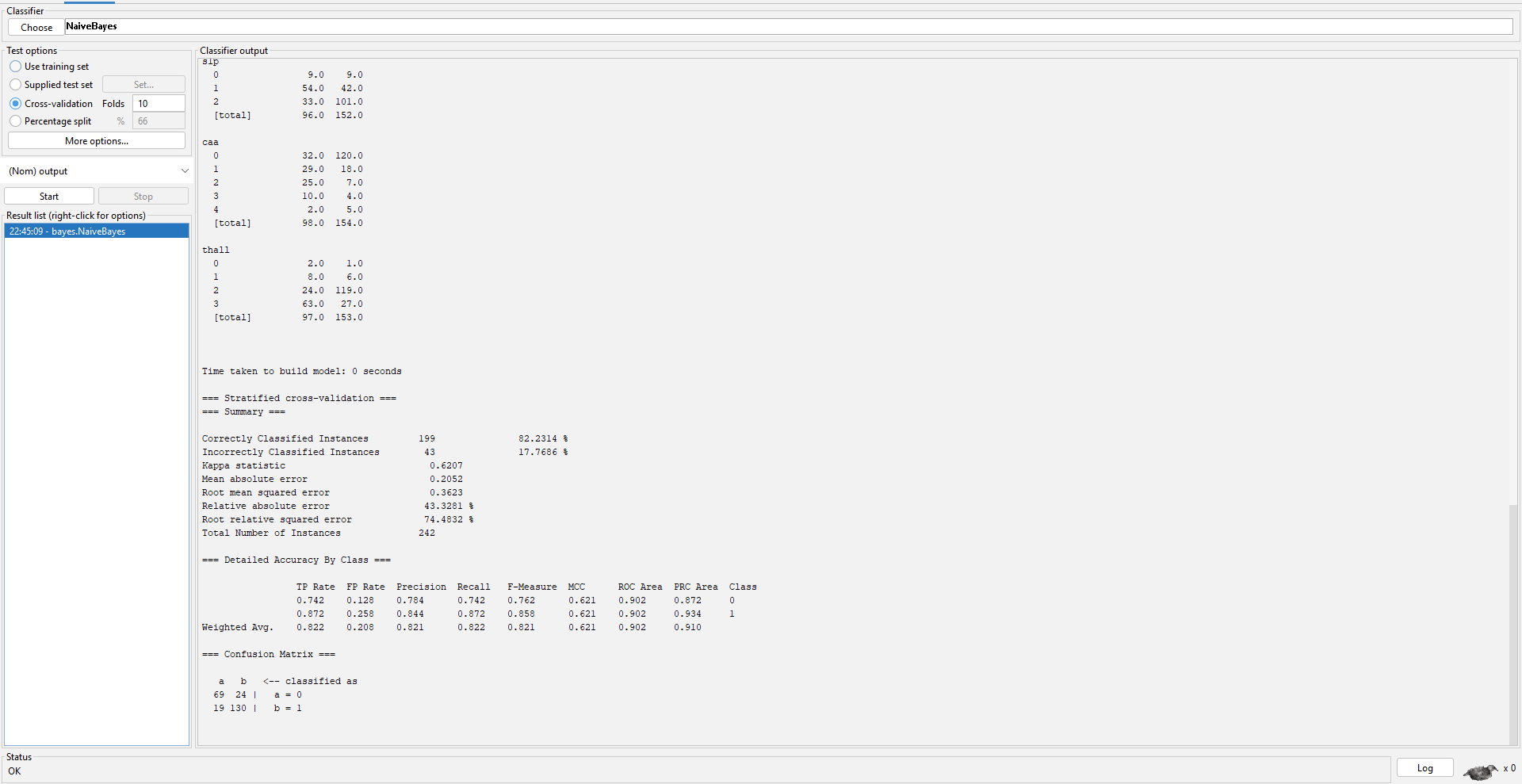
+ Chọn Cross-validation



+ Nhấn Start

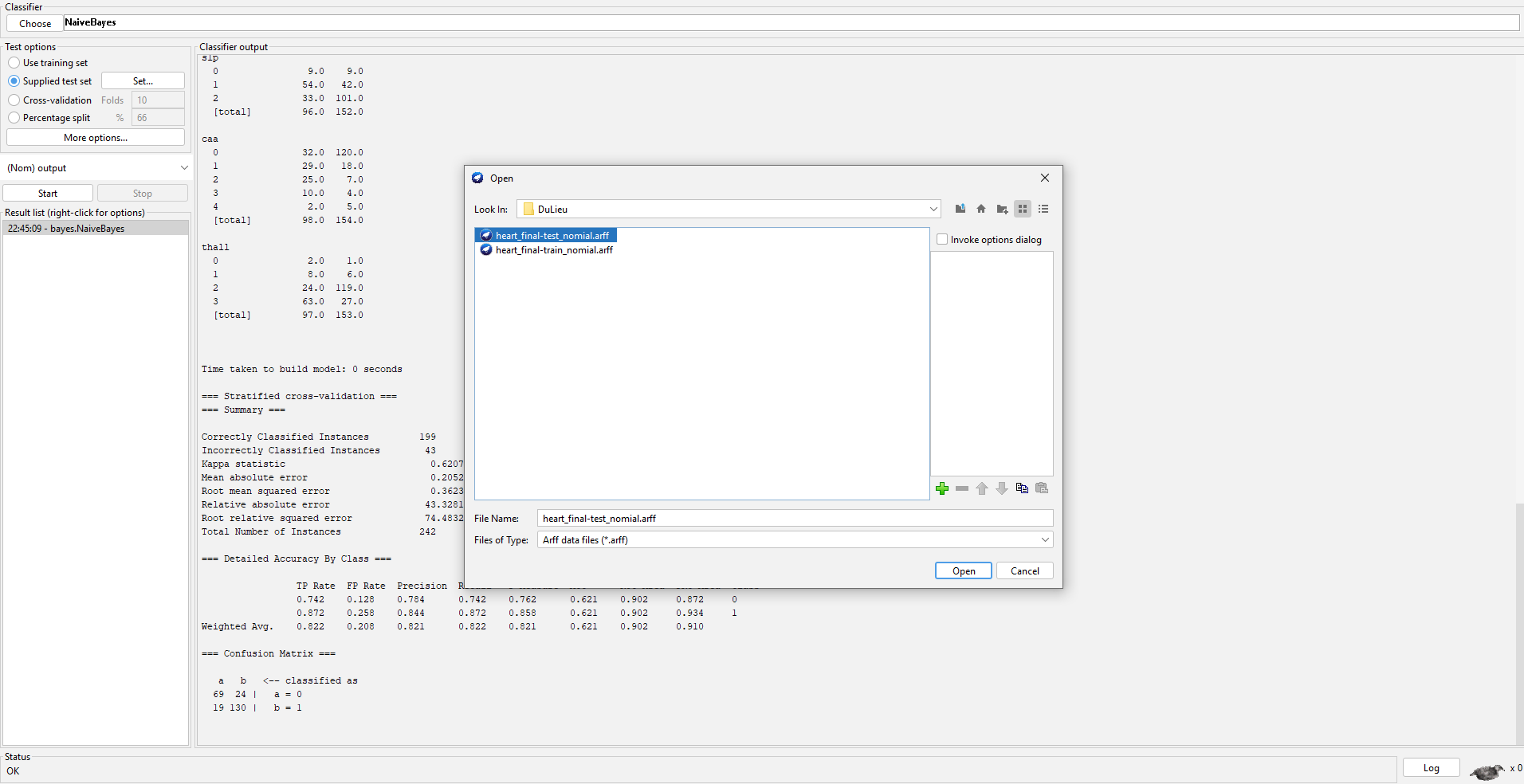


+ Kết quả

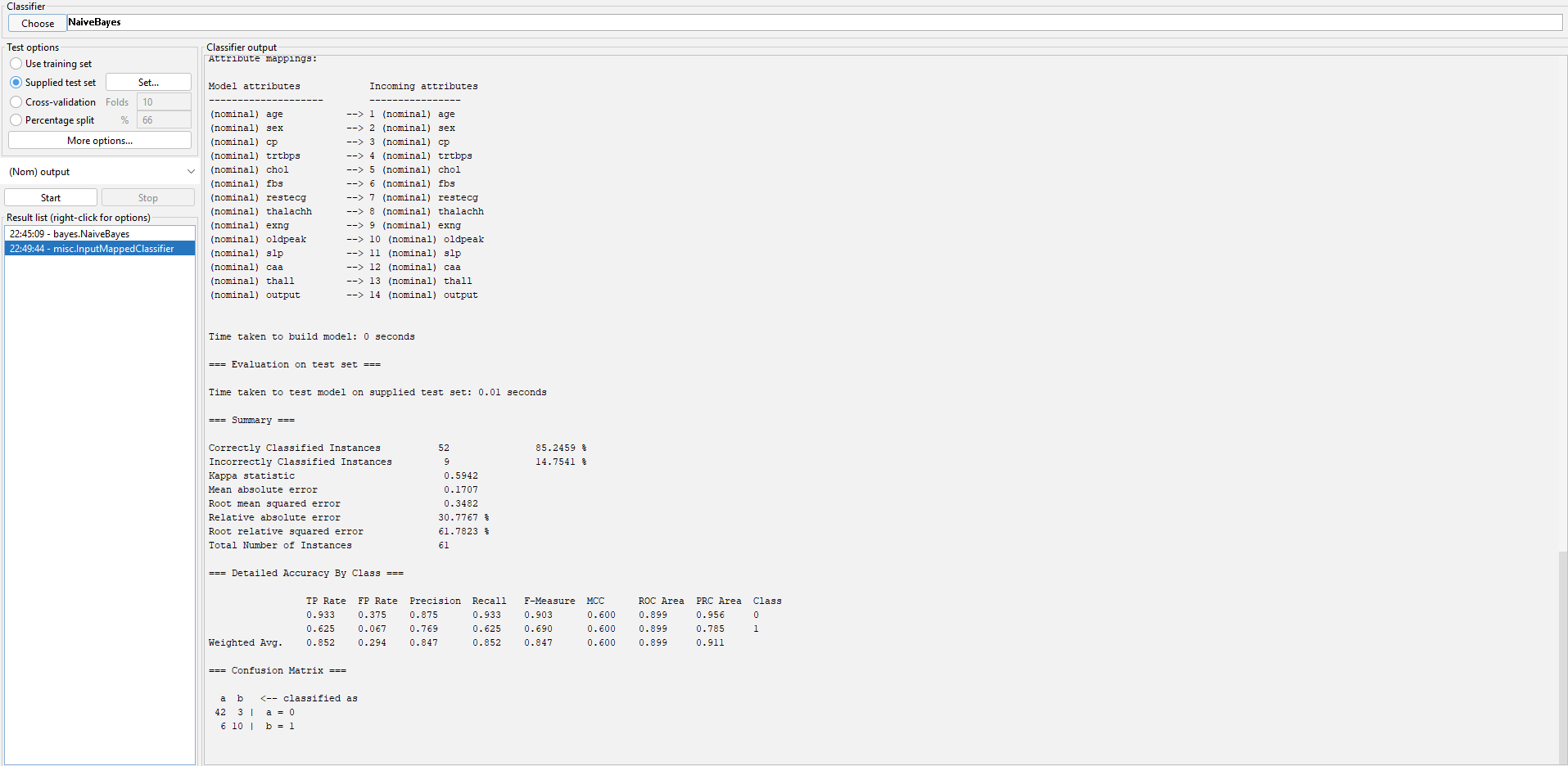


### **Thực hiện kiểm thử Bayes**

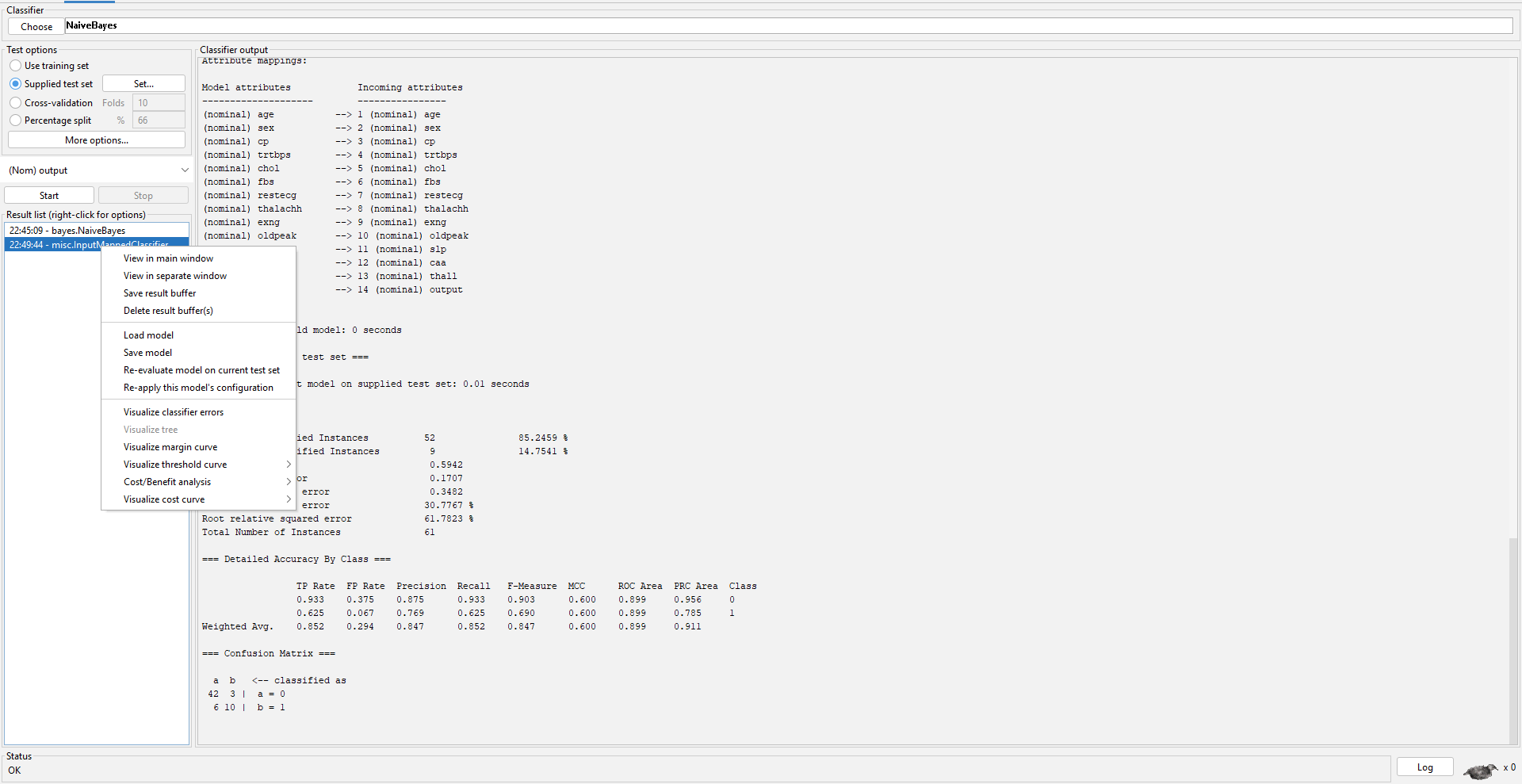
* Bước 1: Chọn file Test
* Chọn “Supplied test set” ở Test Options
* Chọn “Set”
* Chọn file cần kiểm thử và ấn “Open”

****

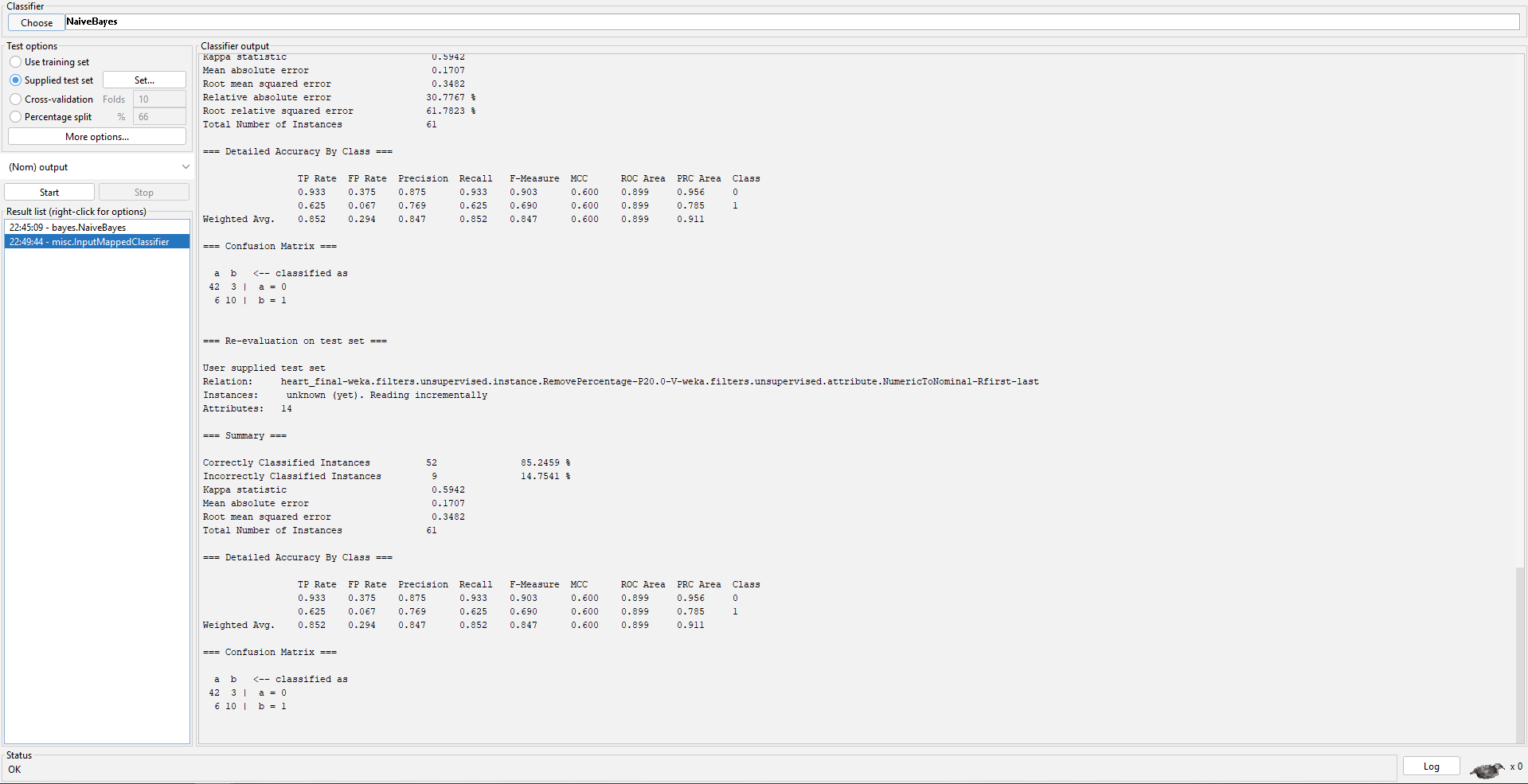
* Bước 2: Chọn “Start” để kiểm thử
* Kết quả



* Bước 3: Xem kết quả
* Chuột phải vào dòng model muốn sử dụng trong Result list
* Chọn “Re-evaluate model on current test set”



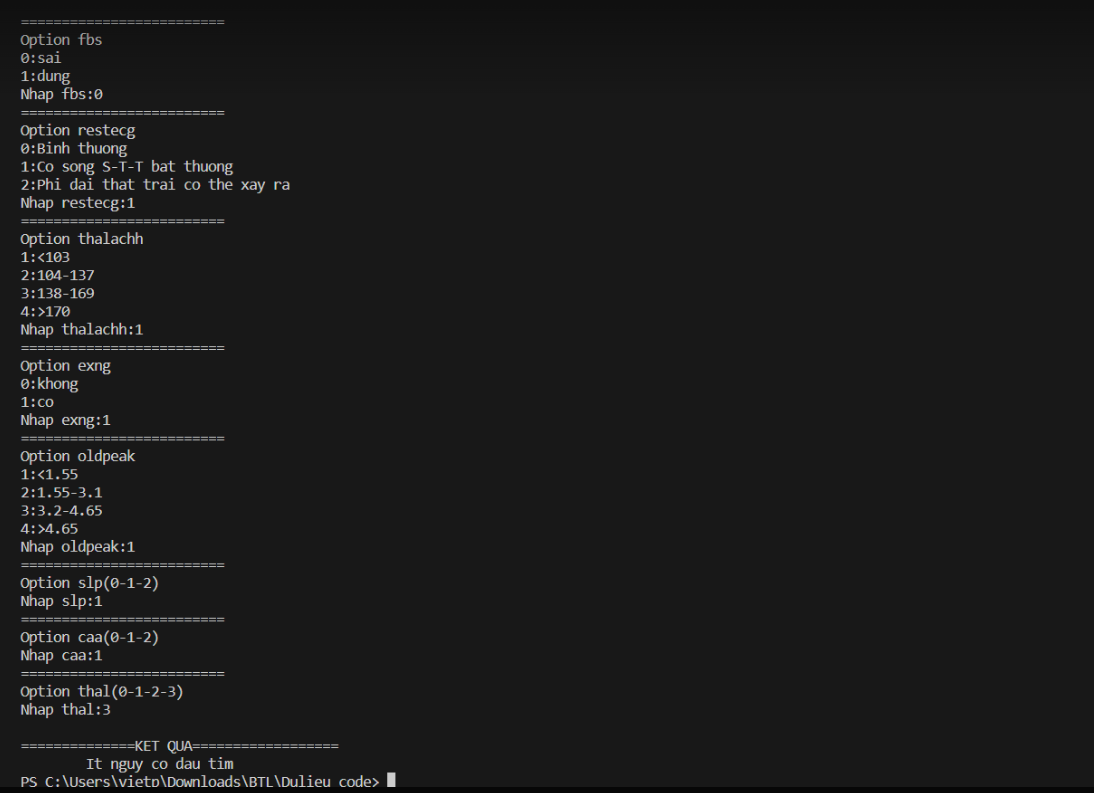
* Kết quả



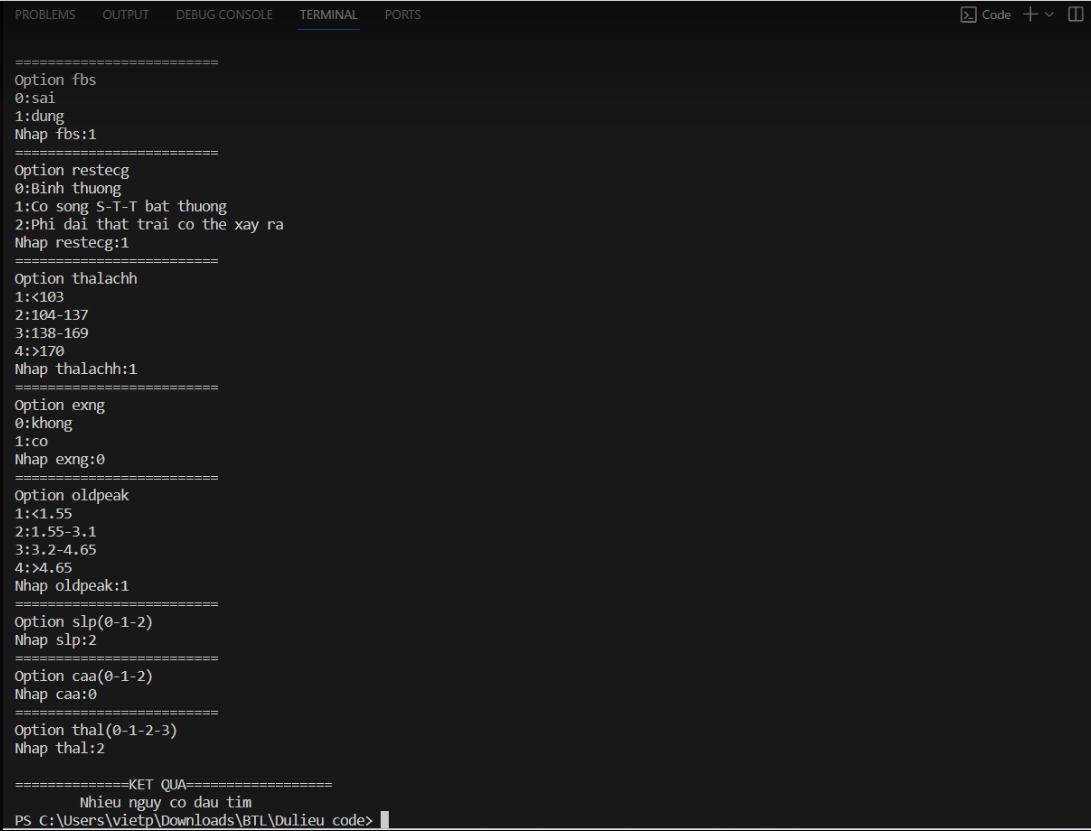
# **Chương 5: Thực nghiệm và đánh giá trên C++**

**Kết quả khi chạy trên Dev-C++**

* Nhập các dãy số lần lượt theo thứ tự nhập của code từ tập test: 4,1,0,3,2,0,1,1,1,1,1,1,3,0



* Kiểm tra với **heart\_final.csv** có thể thấy khớp với giá trị ít đau tim
* Nhập các dãy số lần lượt theo thứ tự nhập của code từ tập test: 4,0,2,1,1,1,1,1,0,1,2,0,2,1:



* So sánh kết quả vscode với **heart\_final.csv** ta có thể thấy khớp với giá trị nhiều nguy cơ đau tim.

# **KẾT LUẬN**

Sau thời gian thực hiện, chúng em đã thực hiện được một số kết quả sau:

- Tìm hiểu được về khai phá dữ liệu.

- Vai trò của khai phá dữ liệu.

- Tìm hiểu về thuật toán Bayes.

- Tìm hiểu về Bayes giải quyết bài toán phân lớp.

Chúng em đã tìm hiểu lý thuyết phân lớp đến thuật toán Bayes. Tuy độ chính xác còn chưa cao do bản chất của phương pháp cũng như tập dữ liệu chưa đủ lớn mong thầy cô giúp đỡ để bài toán của chúng em được hoàn thiện hơn, đặc biệt là thầy Nguyễn Tu Trung - giảng viên hướng dẫn chúng em môn Khai phá dữ liệu.

Chúng em xin chân thành cảm ơn!

# **TÀI LIỆU THAM KHẢO**

[1] Slide môn học Khai phá dữ liệu

[2] Giáo trình Khai phá dữ liệu.