

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI

**---**🙢🙢🕮🙠🙢**---**



**Báo cáo môn học**

**Khai phá dữ liệu**

**Đề tài: Khai phá phân tích dữ liệu hồ sơ lâm sàng bệnh suy tim sử dụng thuật toán**

**phân lớp, phân cụm**

|  |  |
| --- | --- |
| Giảng viên hướng dẫn: | TS. Lương Hồng Lan |
| Nhóm sinh viên thực hiện: | Phạm Văn Hà - 1851061760  Đinh Tất Hiển - 1851061360  Phạm Hoàng Minh - 1851061718 |

Hà Nội, ngày 10 tháng 03 năm 2022

**MỤC LỤC**

[Lời nói đầu 4](#_Toc100700468)

[CHƯƠNG 1: Tổng quan về Khai phá dữ liệu 5](#_Toc100700469)

[1.1. Khái niệm cơ bản 5](#_Toc100700470)

[1.2. Một số kỹ thuật Khai phá dữ liệu 6](#_Toc100700471)

[1.2.1 Kỹ thuật khai phá luật kết hợp 6](#_Toc100700472)

[1.2.2 Kỹ thuật phân lớp 7](#_Toc100700473)

[1.2.3 Kỹ thuật phân cụm 7](#_Toc100700474)

[CHƯƠNG 2 Dữ liệu đầu vào của tập dữ liệu hồ sơ lâm sàng bệnh suy tim 8](#_Toc100700475)

[CHƯƠNG 3: Phân lớp cho tập dữ liệu hồ sơ lâm sàng bệnh suy tim 9](#_Toc100700476)

[3.1 Khái niệm 9](#_Toc100700477)

[3.2 Phân lớp bằng cây quyết định 9](#_Toc100700478)

[3.2.1 Thuật toán xây dựng cây quyết định 10](#_Toc100700479)

[3.3. Tiền xử lý dữ liệu cho thuật toán phân lớp 11](#_Toc100700480)

[3.4. Làm sạch dữ liệu: 14](#_Toc100700481)

[3.5. Khai phá dữ liệu 17](#_Toc100700482)

[3.5.1. Thuật toán - cây quyết định 17](#_Toc100700483)

[3.6. Phát hiện tri thức sau khi dùng thuật toán cây quyết định 19](#_Toc100700484)

[CHƯƠNG 4: Phân cụm cho tập dữ liệu hồ sơ lâm sàng bệnh suy tim dựa trên thuật toán K-means clustering 19](#_Toc100700485)

[4.1 Khái niệm 19](#_Toc100700486)

[4.2 Các phương pháp phân cụm 19](#_Toc100700487)

[4.3 Thuật toán K-means 20](#_Toc100700488)

[Kết luận 22](#_Toc100700489)

[TÀI LIỆU THAM KHẢO 23](#_Toc100700490)

# Lời nói đầu

Trong thời buổi hiện đại ngày nay, công nghệ thông tin cũng như những ứng dụng của nó không ngừng phát triển, lượng thông tin và cơ sở dữ liệu được thu thập và lưu trữ cũng tích lũy ngày một nhiều lên. Con người cũng vì thế mà cần có thông tin với tốc độ nhanh nhất để đưa ra quyết định dựa trên lượng dữ liệu khổng lồ đã có. Các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế, vì thế, một khuynh hướng kỹ thuật mới là Kỹ thuật phát hiện tri thức và khai phá dữ liệu nhanh chóng được phát triển. Khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới. Ở Việt Nam, kỹ thuật này đang được nghiên cứu và dần đưa vào ứng dụng. Khai phá dữ liệu là một bước trong quy trình phất hiện tri thức. Hiện nay, mọi người không ngừng tìm tòi các kỹ thuật để thực hiện khai phá dữ liệu một cách nhanh nhất và có được kết quả tốt nhất. Trong bài tập lớn này, chúng em tìm hiểu và trình bày về một kỹ thuật trong khai phá dữ liệu để phân lớp dữ liệu cũng như tổng quan về khai phá dữ liệu, với đề tài “**Khai phá phân tích dữ liệu hồ sơ lâm sàng bệnh suy tim sử dụng thuật toán phân lớp, phân cụm**”. Trong quá trình làm bài tập lớn này, chúng em xin gửi lời cảm ơn đến TS.Lương Hồng Lan. Cô đã rất tận tình hướng dẫn chi tiết cho chúng em, những kiến thức cô cung cấp rất hữu ích. Chúng em rất mong nhận được những góp ý từ cô. Chúng em xin chân thành cảm ơn!

# CHƯƠNG 1: Tổng quan về Khai phá dữ liệu

## 1.1. Khái niệm cơ bản

Khai phá dữ liệu là một quá trình xác định các mẫu tiềm ẩn có tính hợp lệ, mới lạ, có ích và có thể hiểu được trong một khối dữ liệu rất lớn.

Khai phá tri thức từ CSDL (Knowledge Discovery in Database) Khai phá tri thức từ CSDL gồm 5 bước:

* B1: Lựa chọn CSDL
* B2: Tiền xử lý
* B3: Chuyển đổi
* B4: Khai phá dữ liệu
* B5: Diễn giải và đánh giá

Khai phá dữ liệu là 1 bước trong quá trình khai phá tri thức từ CSDL. Các ứng dụng của khai phá dữ liệu Phát hiện tri thức và khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực: thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, thuật toán, tính toán song song và tốc độ cao, thu thập tri thức cho các hệ chuyên gia, quan sát dữ liệu... Đặc biệt phát hiện tri thức và khai phá dữ liệu rất gần gũi với lĩnh vực thống kê, sử dụng các phương pháp thống kê để mô hình dữ liệu và phát hiện các mẫu, luật ... Ngân hàng dữ liệu (Data Warehousing) và các công cụ phân tích trực tuyến (OLAP- On Line Analytical Processing) cũng liên quan rất chặt chẽ với phát hiện tri thức và khai phá dữ liệu. Khai phá dữ liệu có nhiều ứng dụng trong thực tế, ví dụ như:

Bảo hiểm, tài chính và thị trường chứng khoán: phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán. Danh mục vốn và giá, lãi suất, dữ liệu thẻ tín dụng, phát hiện gian lận, ... Thống kê, phân tích dữ liệu và hỗ trợ ra quyết định. Điều trị y học và chăm sóc y tế: một số thông tin về chuẩn đoán bệnh lưu trong các hệ thống quản lý bệnh viện. Phân tích mối liên hệ giữa các triệu chứng bệnh, chuẩn đoán và phương pháp điều trị (chế độ dinh dưỡng, thuốc, ...) Sản xuất và chế biến: Quy trình, phương pháp chế biến và xử lý sự cố. Text mining và Web mining: Phân lớp văn bản và các trang Web, tóm tắt văn bản,... Lĩnh vực khoa học: Quan sát thiên văn, dữ liệu gene, dữ liệu sinh vật học, tìm kiếm, so sánh các hệ gene và thông tin di truyền, mối liên hệ gene và một số bệnh di truyền, ... Mạng viễn thông: Phân tích các cuộc gọi điện thoại và hệ thống giám sát lỗi, sự cố, chất lượng dịch vụ, ... Các bước của quá trình khai phá dữ liệu Quy trình phát hiện tri thức thường tuân theo các bước sau:

* Bước thứ nhất: Hình thành, xác định và định nghĩa bài toán. Là tìm hiểu lĩnh vực ứng dụng từ đó hình thành bài toán, xác định các nhiệm vụ cần phải hoàn thành. Bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.
* Bước thứ hai: Thu thập và tiền xử lý dữ liệu. Là thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu (làm sạch dữ liệu), xử lý việc thiếu dữ liệu (làm giàu dữ liệu), biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức. Do dữ liệu được lấy từ nhiều nguồn khác nhau, không đồng nhất, … có thể gây ra các nhầm lẫn. Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hoá.
* Bước thứ ba: Khai phá dữ liệu, rút ra các tri thức. Là khai phá dữ liệu, hay nói cách khác là trích ra các mẫu hoặc/và các mô hình ẩn dưới các dữ liệu. Giai đoạn này rất quan trọng, bao gồm các công đoạn như: chức năng, nhiệm vụ và mục đích của khai phá dữ liệu, dùng phương pháp khai phá nào? Thông thường, các bài toán khai phá dữ liệu bao gồm: các bài toán mang tính mô tả - đưa ra tính chất chung nhất của dữ liệu, các bài toán dự báo - bao gồm cả việc phát hiện các suy diễn dựa trên dữ liệu hiện có. Tuỳ theo bài toán xác định được mà ta lựa chọn các phương pháp khai phá dữ liệu cho phù hợp.
* Bước thứ tư: Sử dụng các tri thức phát hiện được. Là hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện. Các kết quả của quá trình phát hiện tri thức có thể được đưa vào ứng dụng trong các lĩnh vực khác nhau do các kết quả có thể là các dự đoán.

## 1.2. Một số kỹ thuật Khai phá dữ liệu

### 1.2.1 Kỹ thuật khai phá luật kết hợp

Trong khai phá dữ liệu, mục đích của luật kết hợp là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Để khai phá luật kết hợp có rất nhiều thuật toán, nhưng dùng phổ biến nhất là thuật toán Apriori. Đây là thuật toán khai phá tập phổ biến trong dữ liệu giao dịch để phát hiện các luật kết hợp dạng khẳng định nhị phân và được sử dụng để xác định, tìm ra các luật kết hợp trong dữ liệu giao dịch. Ngoài ra, còn có các thuật toán FP-growth, thuật toán Partition,…

### 1.2.2 Kỹ thuật phân lớp

Trong kỹ thuật phân lớp gồm có các thuật toán:

* Phân lớp bằng cây quyết định (giải thuật ID3, J48): phân lớp dữ liệu dựa trên việc lập nên cây quyết định, nhìn vào cây quyết định có thể ra quyết định dữ liệu thuộc phân lớp nào. Phân lớp dựa trên xác suất (Naïve Bayesian): dựa trên việc giả định các thuộc tính độc lập mạnh với nhau qua việc sử dụng định lý Bayes.
* Phân lớp dựa trên khoảng cách (giải thuật K – láng giềng): làm như láng giềng làm, dữ liệu sẽ được phân vào lớp của k đối tượng gần với dữ liệu đó nhất. Phân lớp bằng SVM: phân lớp dữ liệu dựa trên việc tìm ra một siêu phẳng “tốt nhất” để tách các lớp dữ liệu trên không gian nhiều chiều hơn.

### 1.2.3 Kỹ thuật phân cụm

Phân cụm dữ liệu là cách phân bố các đối tượng dữ liệu vào các nhóm/ cụm sao cho các đối tượng trong một cụm thì giống nhau hơn các phần tử khác cụm, gồm có một số phương pháp phân cụm cơ bản như:

* Phân cụm bằng phương pháp K-mean: tìm ra tâm của các cụm mà khoảng cách của tâm đó đến các đối tượng, dữ liệu khác là ngắn.
* Phân cụm trên đồ thị, ngoài ra khai phá dữ liệu có rất nhiều kỹ thuật, nhưng đây là những kỹ thuật cơ bản và đơn giản trong khai phá dữ liệu mà chúng em được tìm hiểu.

# CHƯƠNG 2 Dữ liệu đầu vào của tập dữ liệu hồ sơ lâm sàng bệnh suy tim

**(nguồn UCI)**

Bộ dữ liệu được sử dụng để khai phá dữ liệu là bộ dữ liệu dự báo nguy cơ tiểu đường *heart\_failure\_clinical\_records\_dataset.csv* được cung cấp bởi *http://archive.ics.uci.edu/ml/machine-learning-databases/00519/*

Số lượng thuộc tính: 13

Số lượng mẫu: 299

*Bảng 2.1 Liệt kê các thuộc tính và các giá trị của từng thuộc tính trong tập dữ liệu*

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Giá trị | Kiểu |
| Age (Tuổi của bệnh nhân) | 40 tuổi đến 95 tuổi | Numeric |
| Anaemia (Giảm hồng cầu hoặc hemoglobin) | N – Y | String |
| Creatinine phosphokinase (CPK)  (Mức độ enzym CPK trong máu) | 23 - 7861 | Numeric |
| Diabetes (Tiểu đường) | N - Y | String |
| Ejection franction (Phần trăm máu rời tim mỗi lần tim co bóp) | 14 - 80 | Numeric |
| High blood pressure (Tăng huyết áp) | N - Y | String |
| Platelets (Tiểu cầu) | 25100 - 850000 | Numeric |
| Serum creatinine (Hàm lượng creatinine trong huyết thanh) | 0.5 – 9.4 | Numeric |
| Serum sodium (Hàm lượng natri trong huyết thanh) | 113 - 148 | Numeric |
| Sex (Giới tính) | Male - Female | String |
| Smoking (Trạng thái hút thuốc lá) | N - Y | String |
| Time (Thời gian theo dõi) | 4 - 285 | Numeric |
| Death event (Target) (Trạng thái của bệnh nhân trong thời gian theo dõi) | N - Y | String |

# CHƯƠNG 3: Phân lớp cho tập dữ liệu hồ sơ lâm sàng bệnh suy tim

## 3.1 Khái niệm

Phân lớp (Classification/Categorization) thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho các dự báo tiếp theo. Một số phương pháp điển hình là cây quyết định, luật phân lớp, mạng neuron. Nội dụng của phân lớp chính là học một hàm ánh xạ các dữ liệu vào một trong một số lớp đã biết.

Bản chất của bài toán phân lớp là dự đoán các nhãn (hay lớp) của các phần tử dữ liệu đầu vào và các nhãn (hay lớp) này là các giá trị rời rạc. Thông thường, các giải thuật phân lớp thường hoạt động thông qua 2 bước. Bước đầu tiên nó sẽ phân tích tập dữ liệu đã gán nhãn để tìm ra mô hình phù hợp mô tả tập dữ liệu đó. Bước này được gọi là bước học (learning step) hay pha học (learning phase) và tập dữ liệu gán nhãn phục vụ quá trình học này được gọi là dữ liệ huấn luyện (training data). Dữ liệu huấn luyện là một tập các phần tử dữ liệu (data point) có gán nhãn, hay còn được gọi là bản ghi (tuple) mô tả dữ liệu và nhãn (hay lớp) tương ứng của bản ghi đó. Bước thứ hai là bước kiểm tra hiệu năng của chúng. Để tránh hiện tượng quá phù hợp (overfit), một tập dữ liệu khác gọi là tập dữ liệu kiểm thử (testing set) sẽ được sử dụng để đo độ chính xác của thuật toán.

## 3.2 Phân lớp bằng cây quyết định

Giải thuật học dựa trên cây quyết định hoạt động trên tập dữ liệu được biểu diễn bằng các thuộc tính có giá trị liên tục thì cần thực hiện bước rời rạc hóa. Các giải thuật phân lớp bằng cây quyết định (ID3, CART, C4.5) đều áp dụng cách tiếp cận ăn tham (greedy) (một thuật toán không quay lui) (non-backtracking) để xây dựng cây theo hướng từ trên xuống. Tập dữ liệu huấn luyện được sẽ được chia thành các tập nhỏ hơn trong quá trình xây dựng cây theo cơ chế chia để trị (devide-and-conquer).

* Mô tả sơ lược về cây quyết định:

Để nạp dữ liệu vào Weka, ta cần thêm thông tin các thuộc tính nằm trong file mô tả vào tập dữ liệu, sau đó chuyển đuôi mở rộng thành \*.csv

Dữ liệu đầu vào để dự báo nguy cơ mắc bệnh tiểu đường giai đoạn đầu lấy link sau:

*http://archive.ics.uci.edu/ml/machine-learning-databases/00519/*

* Tập dữ liệu mô tảdấu hiệu của bệnh nhân mới mắc bệnh tiểu đường hoặc sẽ là bệnh nhân tiểu đường:
  + Số lượng mẫu: 299
  + Số lượng thuộc tính: 13
  + Kiểu của mỗi thuộc tính: numeric, string.
  + Sự phân bố của dữ liệu vào các phân lớp khá cân bằng. Số lượng các phân lớp không áp đảo nhau (imbalanced).

### 3.2.1 Thuật toán xây dựng cây quyết định

Đầu vào: Tập D chứa dữ liệu huấn luyện

Attribute\_list chứa danh sách các thuộc tính ứng cử

Đầu ra: Cây quyết định

**Generate\_decision\_tree (D, attribute\_list)**

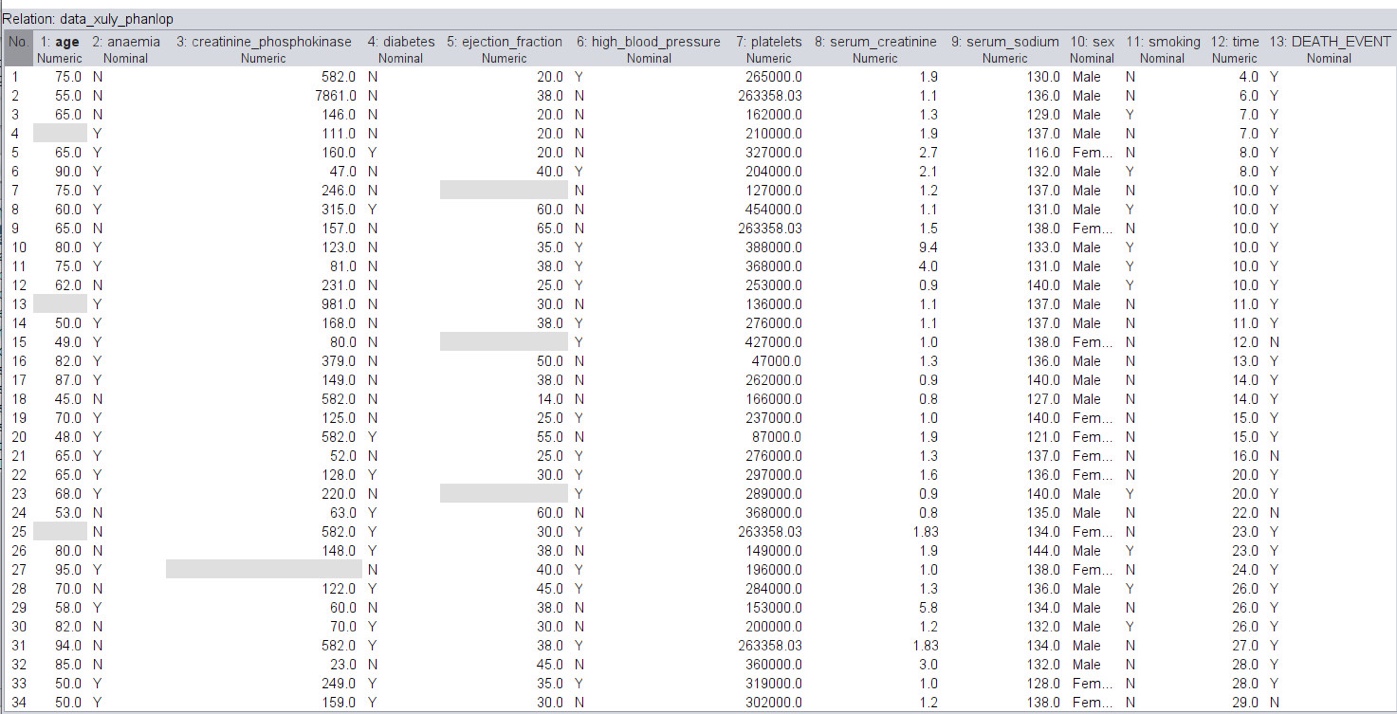
1. Tạo một nút gốc N cho cây quyết định
2. If toàn bộ dữ liệu trong D đều thuộc lớp C, return nút N là nút lá có nhãn C
3. If attribute\_list là rỗng, return nút N với nhãn là lớp xuất hiện nhiều nhất trong D
4. Splitting\_attribute = attribute\_selection\_method ( D, attribute\_list ) tìm thuộc tính phân chia tốt nhất
5. Gán cho nút N nhãn là splitting\_attribute
6. Attribute\_list  attribute\_list \ {splitting\_attribute} (loại bỏ thuộc tính splitting\_attribute khỏi attribute\_list)
7. For each giá trị j của thuộc tính splitting\_attribute
   1. Gọi Dj là tập chứa các phần tử dữ liệu mà thuộc tính splitting\_attribute có giá trị j
   2. If Dj là rỗng thì thêm một nút lá Nj cho nút N có nhãn là nhãn phổ biến nhất xuất hiện trong D
   3. Else gắn cây trả về bởi Generate\_decision\_tree (Dj, attribute\_list) vào nút

N

1. Return N

## 3.3. Tiền xử lý dữ liệu cho thuật toán phân lớp

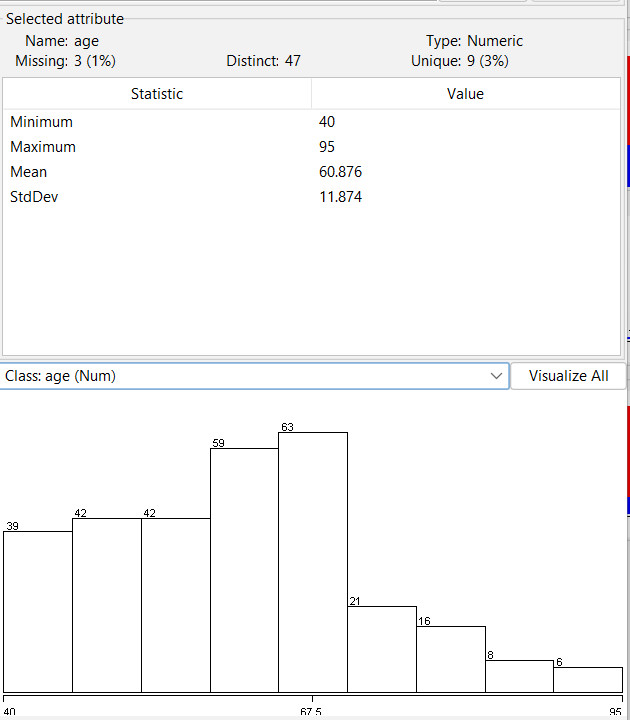
Tập dữ liệu gồm 13 thuộc tính và cả 13 thuộc tính đều không rời rạc. Các thuộc tính điều có dạng numberic và string nên ta phải tiền xử lý dữ liệu, rời rạc hóa các thuộc tính để phù hợp với thuật toán phân lớp.



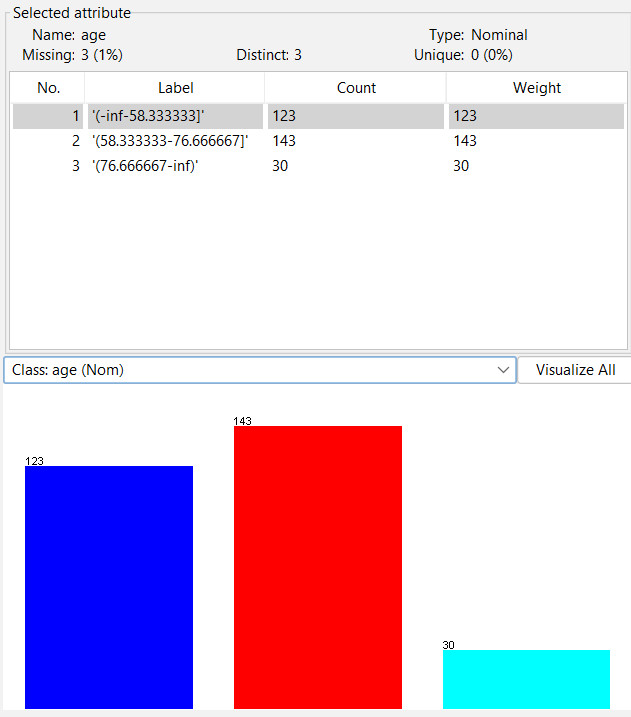
*Hình 3.0 Bảng dữ liệu thô ban đầu*

Đối với thuộc tính Age (tuổi) là dữ liệu liên tục (từ 40 đến 95 tuổi). Ta cần rời rạc hóa dữ liệu thuộc tính tuổi bằng phương pháp bins (đóng thùng), tức là ta sẽ phân độ tuổi từ 40 đến 95 thành 3 loại lần lượt là:

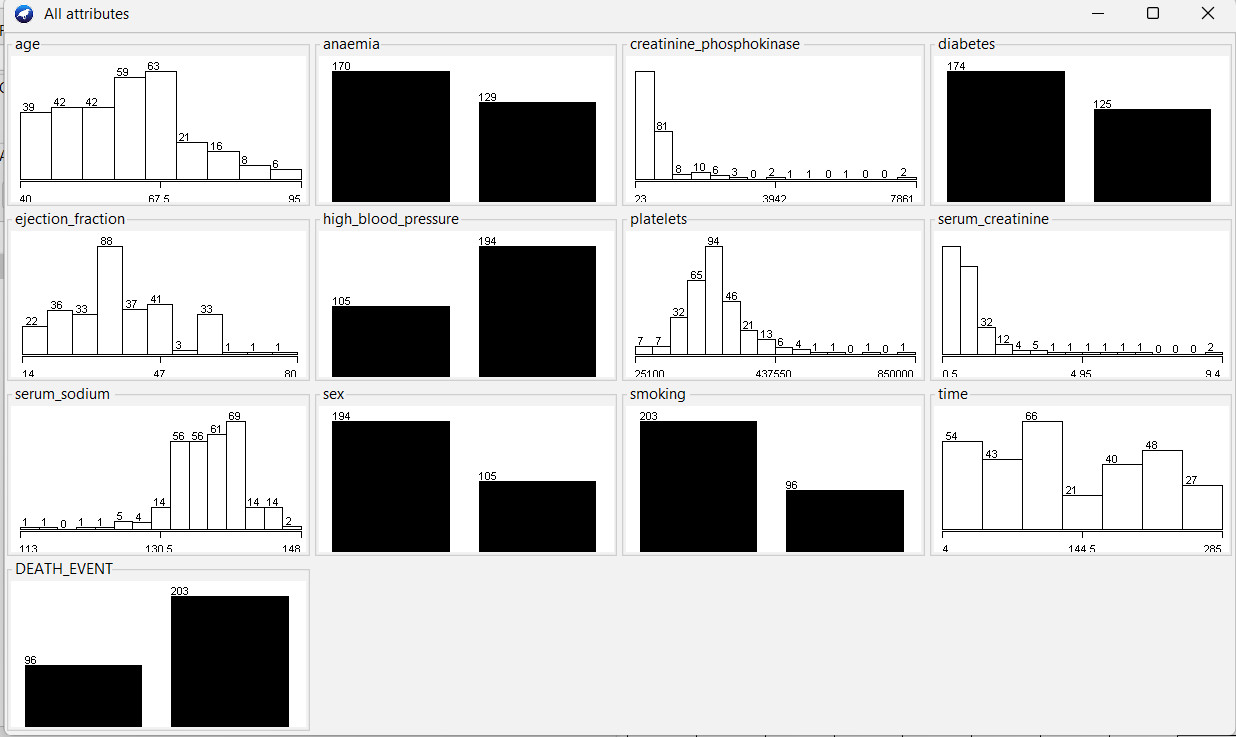
1. Lớp tuổi 40 - 58.
2. Lớp tuổi 59 – 76.
3. Lớp tuổi 77 – 95.



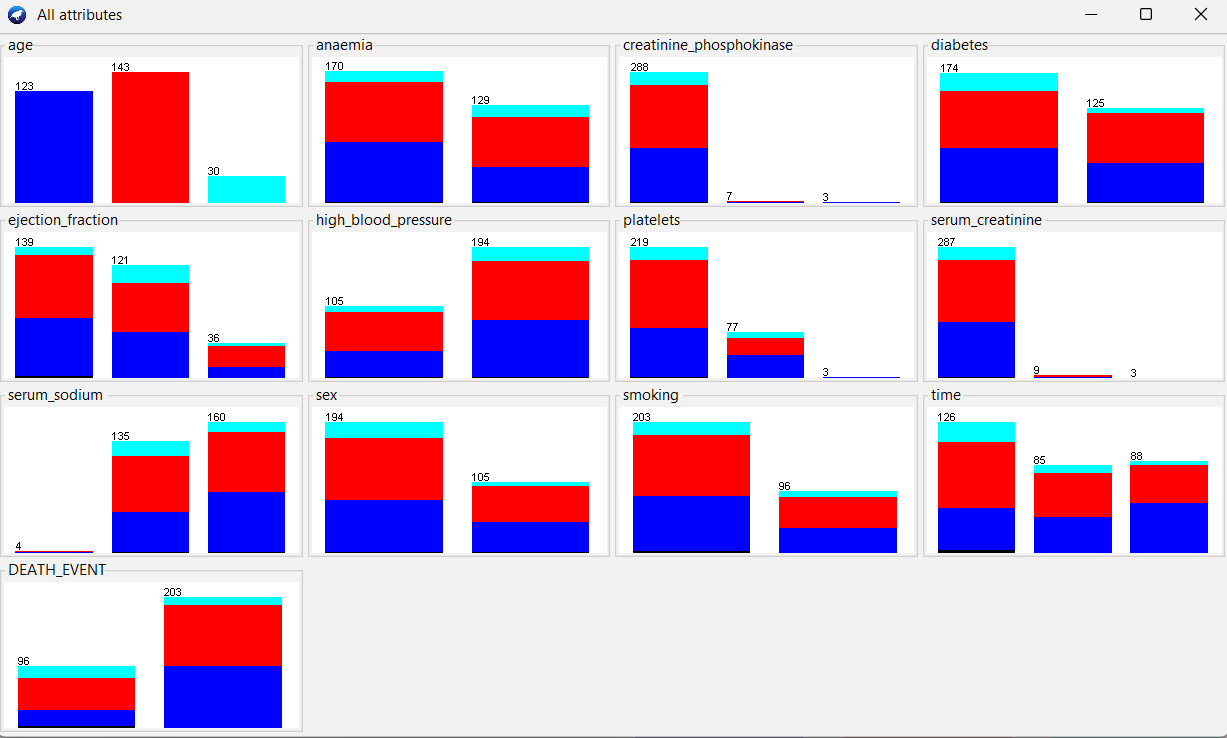
*Hình 3.1 Thuộc tính Age trước khi rời rạc hóa dữ liệu (độ tuổi min = 40 và max = 95)*



*Hình 3.2 Thuộc tính Age (tuổi) sau khi được rời rạc hóa dữ liệu thành 3 bins.*



*Hình 3.3 Các thuộc tính liên tục trước khi thực hiện rời rạc hóa dữ liệu.*

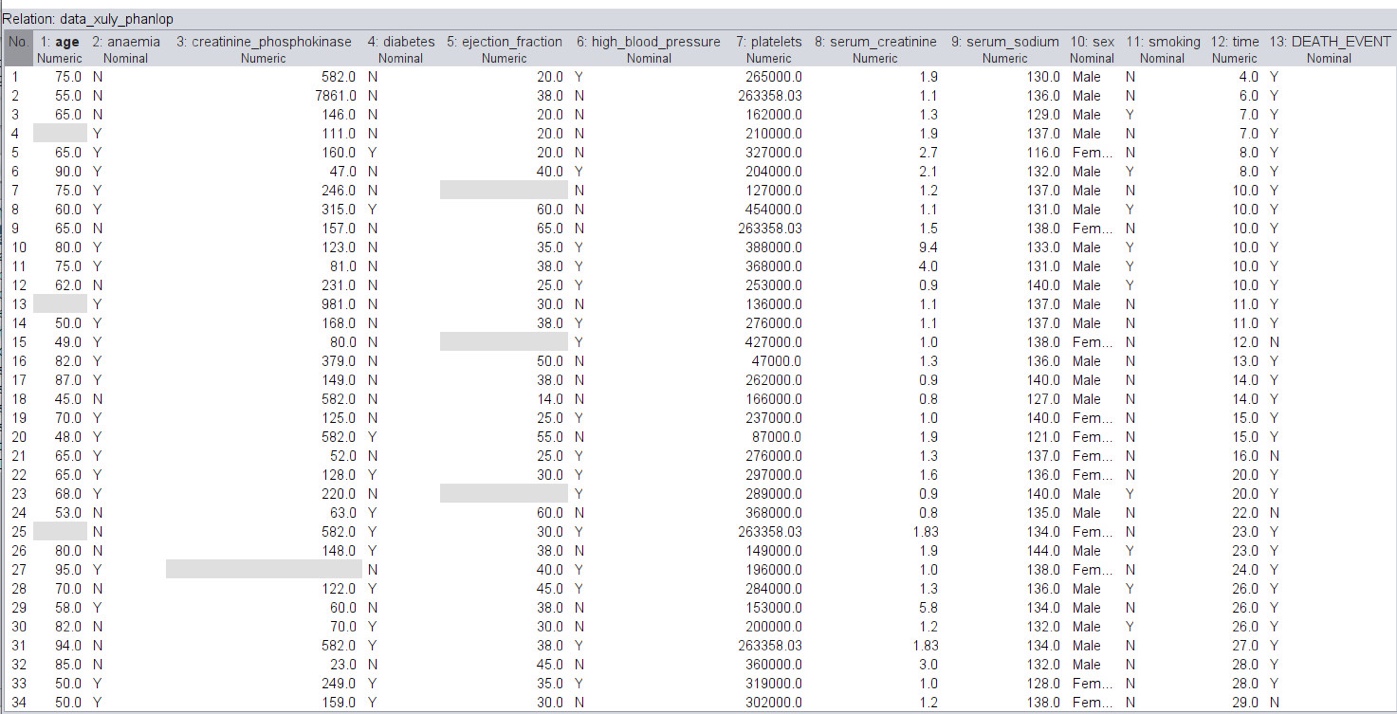


*Hình 3.4 Các thuộc tính liên tục sau khi thực hiện rời rạc hóa dữ liệu.*

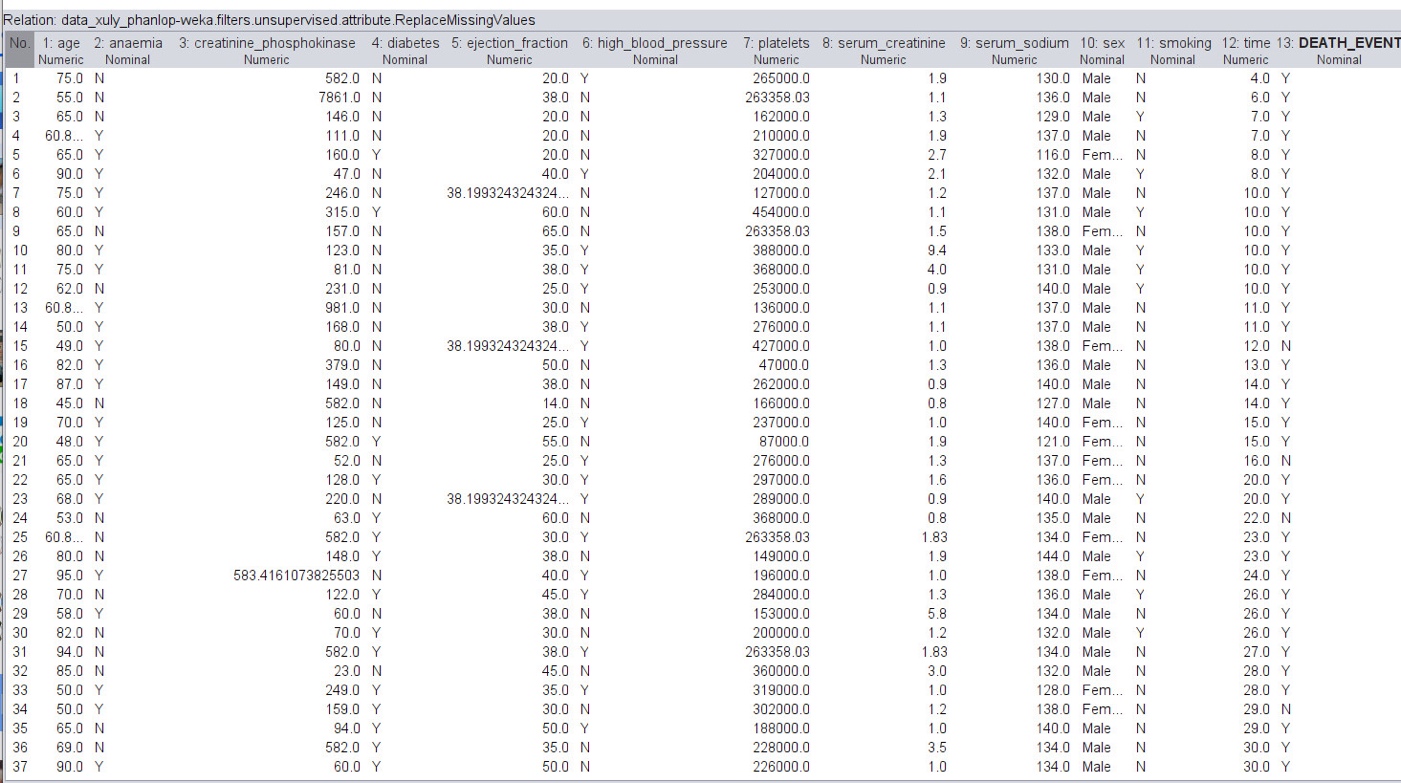
**3.4. Làm sạch dữ liệu:**

- Là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiếu (missing data) xử lý dữ liệu bị nhiễu (noisy data) và không nhất quán.

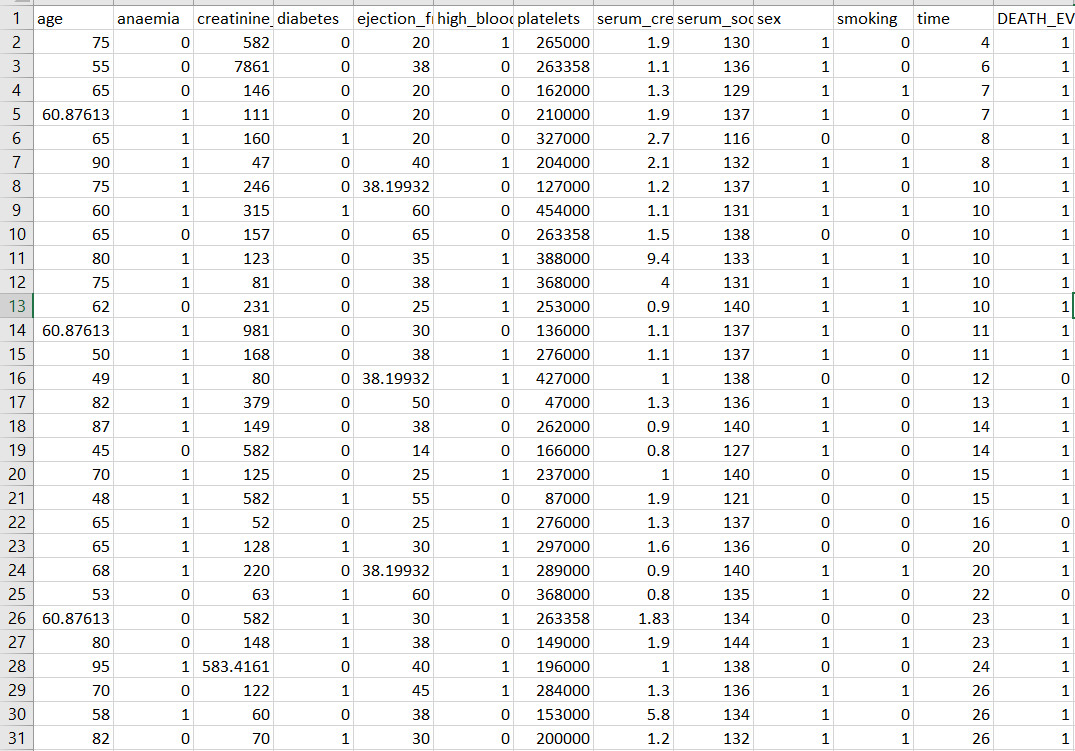
- Đọc file dữ liệu bằng weka:



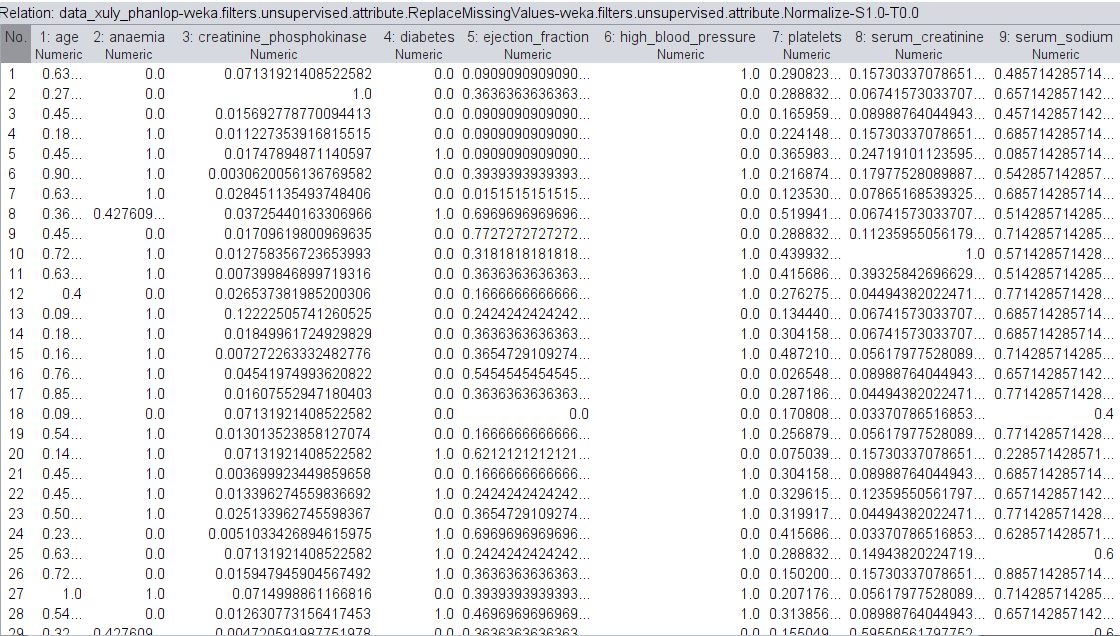
- Các dữ liệu thiếu được thay thế bằng giá trị trung bình của thuộc tính dùng bộ lọc *ReplaceMissingValue*



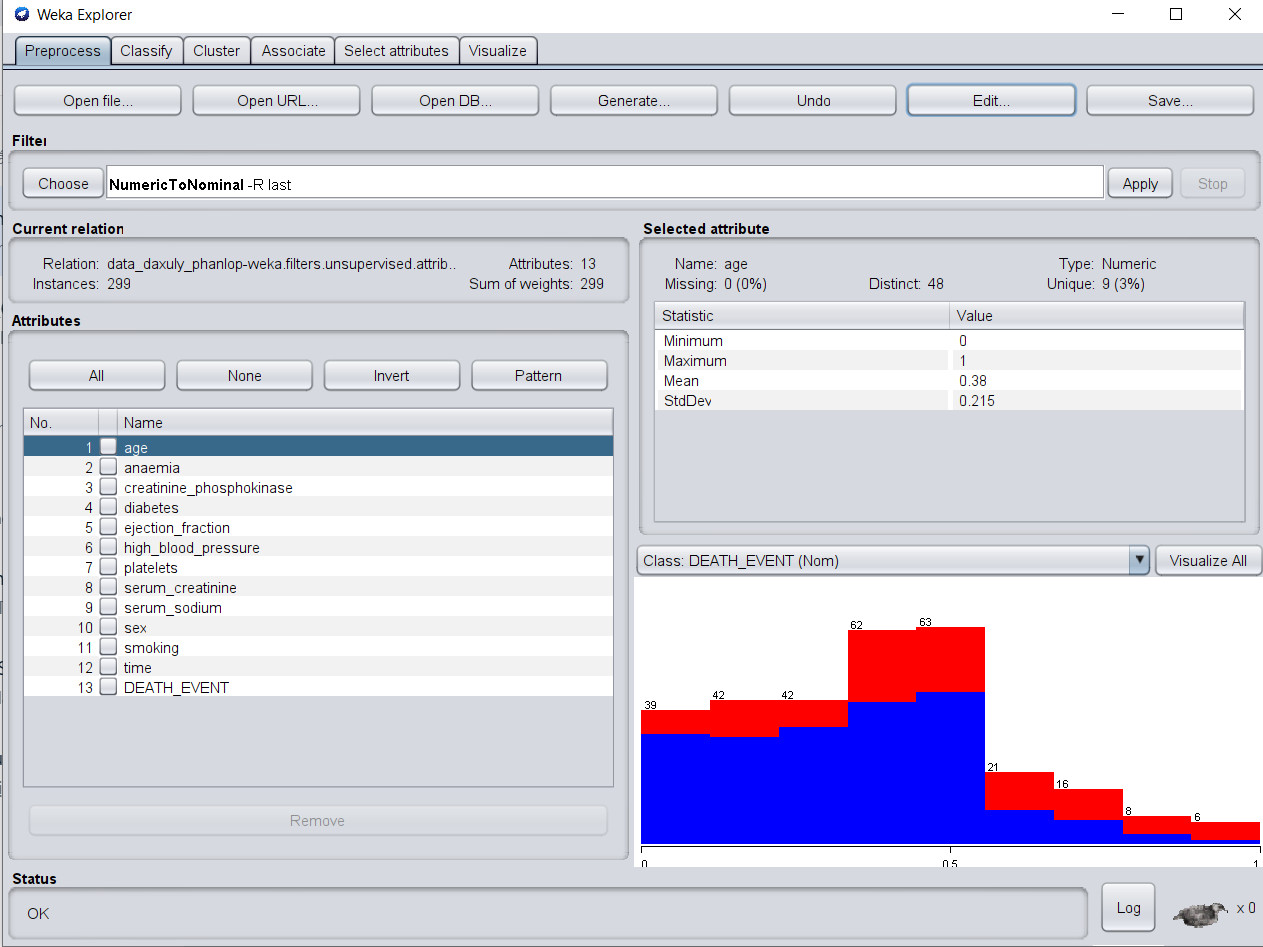
**-** Dùng Excel ta chuyển các giá trị chữ về dạng 0,1*.*

**

**-** T chuẩn hóa các thuộc tính số về đoạn [0, 1] bằng phương pháp chuẩn hóa min-max bằng bộ lọc *Normalize.*



**-** Chuẩn hóa dữ liệu cột phân lớp từ dạng Numeric sang Nominal, sử dụng bộ lọc *NumericToNominal* và hiệu chỉnh phạm vi áp dụng từ *first-last* sang *last.*



## 3.5. Khai phá dữ liệu

Sau khi đã làm xong bước tiền xử lý dữ liệu, ta dùng thuật toán Cây Quyết định để khai phá dữ liệu:

### 3.5.1. Thuật toán - cây quyết định

1. Chia dữ liệu thành các phần dựa và values, in ra để biết thêm chi tiết

def test\_split(index, value, dataset):

    left, right = list(), list()

    for row in dataset:

        if row[index] < value:

            left.append(row)

        else:

            right.append(row)

    return left, right

2. Tạo cây quyết định

def build\_tree(train, max\_depth, min\_size):

    root = get\_split(train)

    split(root, max\_depth, min\_size, 1)

    return root

3.In cây quyết định

def print\_tree(node, depth=0):

    if isinstance(node, dict):

        print('%s[X%d < %.3f]' % ((depth\*' ', (node['index']+1), node['value'])))

        print\_tree(node['left'], depth+1)

        print\_tree(node['right'], depth+1)

    else:

        print('%s[%s]' % ((depth\*' ', node)))

tree = build\_tree(dt, 2, 1)

print\_tree(tree)

Kết Quả:

[X12 < 0.249]

[X9 < 0.686]

[1.0]

[1.0]

[X8 < 0.124]

[0.0]

[0.0]

4. Sử dụng thư viện

from sklearn import tree

clf = tree.DecisionTreeClassifier(max\_depth=2)

X\_train = [row[:-1] for row in train]

y\_train = [row[-1] for row in train]

clf = clf.fit(X\_train, y\_train)

pre = clf.predict(X\_test)

score1 = 0.0

for y\_pre, y\_r in zip(pre, y\_test):

    if(y\_pre ==  y\_r):

        score1 +=1

print("Accuracy: " + str(round(score1\*100/len(test), 3)) + "%")

Kết quả bài toán:

Accuracy: 68.889%

## 3.6. Phát hiện tri thức sau khi dùng thuật toán cây quyết định

Sau khi nhìn vào cây quyết định được sinh ra từ thuật toán ta có thể phát hiện ra được các tri thức như sau:

• Nếu người nào thuộc khoảng tuổi từ 59 - 76 có phần trăm máu lưu thông mỗi lần tim co bóp trong khoảng 36-58%, chỉ số natri trong máu nằm trong 124 - 136 và các chỉ số creatinine thấp thì có nguy cơ tử vong cao.

• Nếu người nào thuộc khoảng tuổi từ 59 - 76 mắc bệnh cao huyết áp, tiểu đường, chỉ số tiểu cầu thấp, có phần trăm máu lưu thông mỗi lần tim co bóp trong khoảng 36- 58%, chỉ số natri trong máu trên 137 và các chỉ số creatinine thấp thì có nguy cơ tử vong.

# CHƯƠNG 4: Phân cụm cho tập dữ liệu hồ sơ lâm sàng bệnh suy tim dựa trên thuật toán K-means clustering

## 4.1 Khái niệm

Phân cụm (Clustering) thực hiện việc nhóm dữ liệu thành các “cụm” ( có thể nói là các lớp mới) để có thể phát hiện được các mẫu phân bố dữ liệu trong miền ứng dụng. Phân cụm là một bài toán mô tả hướng tới việc nhận biết một tập hữu hạn các cụm hoặc các lớp để mô tả dữ liệu. Các cụm (lớp) có thể tách rời nhau và toàn phần (tạo nên một phân hoạch cho tập dữ liệu) hoặc được trình bày đẹp hơn như phân lớp có thứ bậc hoặc có thể chồng lên nhau ( giao nhau) . Ví dụ như bài toán phát hiện các nhóm người tiêu dùng trong CSDL tiếp thị hoặc nhận biết các loại quang phổ trong tập phép đo không gian hồng ngoại… Thông thường , mục tiêu định hướng của bài toán phân cụm là cực đại tính tương đồng giữa các phần tử trong mỗi cụm và cực tiểu tính tương đồng giữa các phần tử thuộc các cụm khác nhau.

## 4.2 Các phương pháp phân cụm

Việc phân loại các giải thuật phân cụm là bài toán không đơn giản, lý do là có nhiều tiêu chí phân loại, hơn nữa có nhiều giải thuật có cùng một số đặc trưng nên việc phân loại cũng không thể tách bạch được. Hay nói cách khác, giữa các phân loại có sự giao nhau.

Dưới đây chúng em xin liệt kê một số cách phân loại bài toán phân cụm:

* Phân cụm phẳng và phân cụm phân cấp.
* Phân cụm dựa vào mật độ (density-based).
* Phương pháp phân cụm dựa trên lưới (grid-based).
* Phương pháp phân cụm dựa trên mô hình (model).
* Phân cụm đơn định (deterministic) và phân cụm xác suất (probability).
* Phân cụm dựa trên ràng buộc (constraint-based).
* Phân cụm dữ liệu có số chiều lớn (high-dimensional data).
* Phân cụm theo lô (batch) và phân cụm gia tăng (incremental).

## 4.3 Thuật toán K-means

Giải thuật K-means thuộc lớp phân cụm phẳng, đầu vào cho thuật toán K-means là tập dữ liệu D gồm n phần tử dữ liệu, số lượng các cụm đầu ra *k*. Đầu ra của giải thuật là *k* cụm dữ liệu. Giải thuật k-means được trình bày như sau:

*Đầu vào*: Tập dữ liệu *D*, số lượng các cụm *k.*

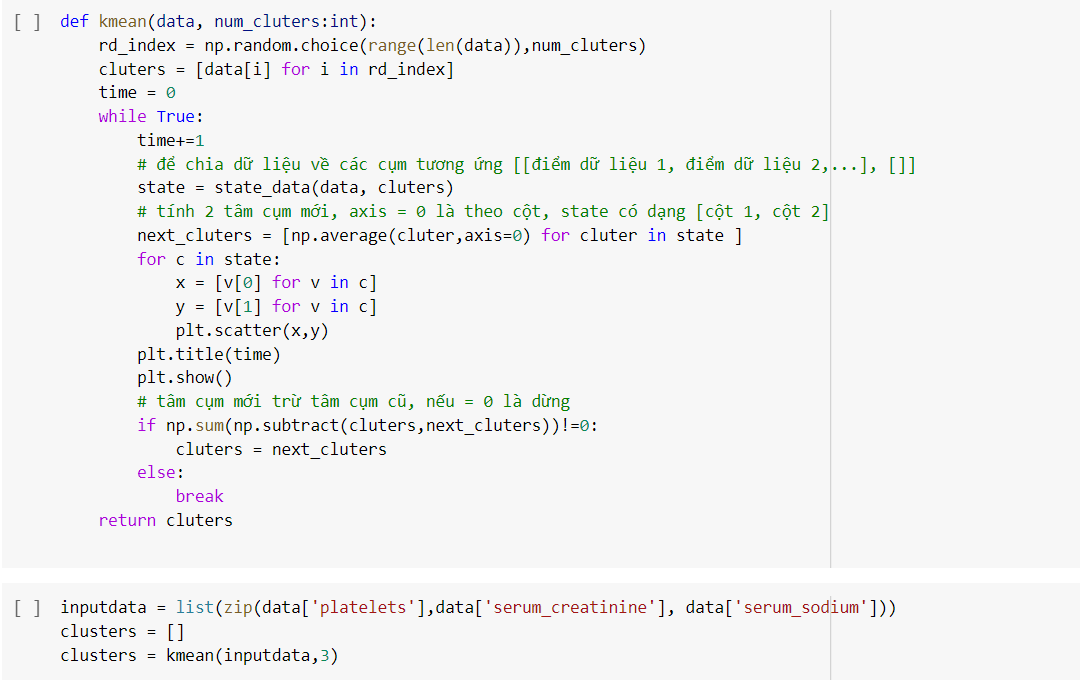
*Đầu ra*: Tập dữ liệu đã được phân thành *k* cụm.

**Thuật toán k-means**

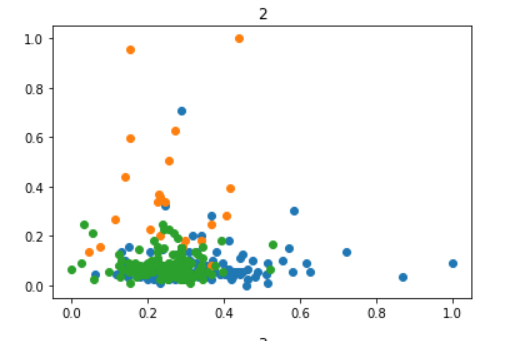
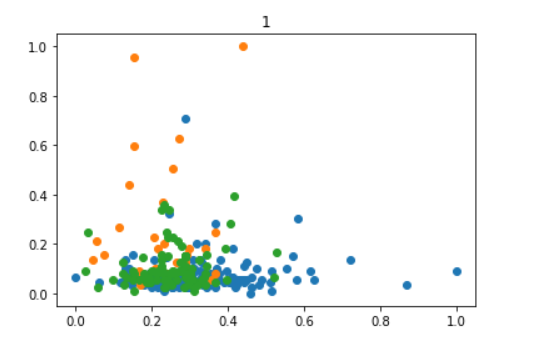
1. Chọn ngẫu nhiên *k* phần tử trong *D* làm trọng tâm ban đầu cho các cụm.
2. Phân các phần tử dữ liệu trong *D* vào các cụm dựa vào độ tương đồng của nó với trọng tâm của các cụm. Phần tử dữ liệu sẽ được phân vào cụm có độ tương đồng lớn nhất.
3. Tính lại trọng tâm của các cụm.
4. Nhảy đến bước 2 cho đến khi quá trình hội tụ (không có sự gán lại các phần tử dữ liệu giữa các cụm , hay trọng tâm của các cụm là không đổi).

**PHÂN CỤM BÀI TOÁN VỚI K-MEAN:**

Sau khi áp dụng K-means với k = 3 ta được:



Kết quả :



Sau khi sử dụng weka để phân cụm, ta thấy rằng k = 3 có độ tương đồng giữa các phần tử lớn nhất . Điều này thể hiện rằng trong bài toán này có thể xác định và nhóm được các cụm bệnh nhân mắc bệnh suy tim có nguy cơ tử vong cao dựa trên các thuộc tính từ đó đưa ra được những dự đoán cũng như hướng điều trị phù hợp.

# Kết luận

Sự bùng nổ thông tin cùng với sự phát triển ứng dụng ngày càng rộng rãi của công nghệ thông tin trong mọi lĩnh vực đã khiến nhu cầu xử lý những khỗi dữ liệu khổng lồ để kết xuất ra những thông tin, tri thức hữu ích cho người sử dụng một cách tự động, nhanh chóng và chính xác, trở thành nhân tố quan trọng hàng đầu cho mọi lĩnh vực. Để kết xuất được những thông tin, tri thức quý giá đó thì việc sử dụng khai phá dữ liệu trong việc xử lý dữ liệu là cần thiết. Khai phá dữ liệu là một trong những kỹ thuật quan trọng, mang tính thời sự không chỉ với lĩnh vực công nghệ thông tin mà còn quan trọng đối với rất nhiều lĩnh vực khác trong đó có cả lĩnh vực y tế.

Trong lĩnh vực y tế, khai phá dữ liệu cho ta những tri thức về bệnh tật. Tuy việc áp dụng kỹ thuật này vào y tế còn một số hạn chế như việc lựa chọn thuật toán phù hợp để khai phá dữ liệu, cần một lượng lớn dữ liệu về bệnh nhân, tốc độ xử lý của máy tính. Nhưng cũng đã phần nào hỗ trợ cho việc chẩn đoán bệnh trở lên nhanh chóng và tăng thêm độ chính xác cho các bác sĩ trong việc chẩn đoán bệnh nhân.

# TÀI LIỆU THAM KHẢO

1. *- Giáo trình của Lương Hồng Lan*
2. *- https://archive.ics.uci.edu/ml/datasets.php*
3. *- Giáo trình Khai Phá Dữ Liệu – Nguyễn Hà Nam – Nguyễn Trí Thành – Hà Quang Thụy*