**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙡🕮🙣-----

****

**ĐỀ TÀI:**

**KHAI PHÁ DỮ LIỆU BỆNH NHÂN BẰNG PHƯƠNG PHÁP HỒI QUY VÀ PHÂN LỚP**

**Nhóm sinh viên thực hiện:**

1. Vũ Thị Hường – 1551060742 – 57TH2

2. Lương Thị Thu – 1551060920 – 57TH2

3. Nguyễn Văn Nguyên – 1451061969 - 56TH-PM

**Giảng viên hướng dẫn:** TS. Đặng Thị Thu Hiền

**Hà Nội, 4/2019**

**MỤC LỤC**

[MỞ ĐẦU 3](#_Toc4995488)

[CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 4](#_Toc4995489)

[1.1 Phát hiện tri thức và khai phá dữ liệu. 4](#_Toc4995490)

[1.2 Quy trình khám phá tri thức trong CSDL. 4](#_Toc4995491)

[1.3 Mô tả bài toán chuẩn đoán bệnh. 5](#_Toc4995492)

[1.3.1 Tổng quan bài toán. 5](#_Toc4995493)

[1.3.2 Phân tích dữ liệu thô. 5](#_Toc4995494)

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 7](#_Toc4995495)

[2.1 Làm sạch dữ liệu. 7](#_Toc4995496)

[2.2 Tích hợp dữ liệu. 8](#_Toc4995497)

[2.3 Biến đổi dữ liệu. 8](#_Toc4995498)

[CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN HỒI QUY 11](#_Toc4995499)

[3.1 Giới thiệu về kỹ thuật hồi quy. 11](#_Toc4995500)

[3.2 Thuật toán hồi quy tuyến tính đa biến. 11](#_Toc4995501)

[CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN LỚP\ 19](#_Toc4995502)

[4.1 Giới thiệu về bài toán phân lớp. 19](#_Toc4995503)

[4.2 Thuật toán phân lớp ID3. 19](#_Toc4995504)

[4.4 Đánh giá mô hình. 33](#_Toc4995505)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 34](#_Toc4995506)

[3.1 Kết luận. 34](#_Toc4995507)

[3.2 Hướng phát triển. 34](#_Toc4995508)

[TÀI LIỆU THAM KHẢO 35](#_Toc4995509)

# **MỞ ĐẦU**

Trong những năm gần đây cùng với phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nỗ về tri thức. Kho dữ liệu, nguồn tri thức của nhân loại cũng trở nên đồ sộ, vô tận làm cho vấn đề khai thác các nguồn tri thức đó ngày càng trở nên nóng bỏng và đặt ra thách thức lớn cho nền công nghệ thông tin thế giới.

Nhu cầu về tìm kiếm và xử lý thông tin, cùng với yêu cầu về khả năng kịp thời khai thác chúng để mạng lại những năng suất và chất lượng cho công tác quản lý, hoạt động kinh doanh... đã trở nên cấp thiết trong xã hội hiện đại. Để đáp ứng phần nào yêu cầu này, người ta đã xây dựng các công cụ tìm kiếm và xử lý thông tin nhằm giúp cho người dùng tìm kiếm được các thông tin cần thiết cho mình.

Với các phương pháp khai thác cơ sở dữ liệu truyền thống chưa đáp ứng được các yêu cầu đó. Để giải quyết vấn đề này, một hướng đi mới đó là nghiên cứu và áp dụng kỹ thuật khai phá dữ liệu và khám phá tri thức trong môi trường Web. Do đó, việc nghiên cứu các mô hình dữ liệu mới và áp dụng các phương pháp khai phá dữ liệu trong khai phá tài nguyên Web là một xu thế tất yếu vừa có ý nghĩa khoa học vừa mang ý nghĩa thực tiễn cao.

Vì vậy chúng em chọn đề tài: “***Khai phá dữ liệu bệnh nhân bằng phương pháp hồi quy và phân lớp”***, để làm báo cáo kết thúc môn học của mình

**Báo cáo gồm 3 chương:**

Chương 1: Tổng quan về khai phá dữ liệu.

Chương 2: Tiền xử lý dữ liệu.

Chương 3; Khai phá dữ liệu bằng thuật toán hồi quy.

Chương 4: Khai phá dữ liệu bằng thuật toán phân lớp.

Chương 5. Kết luận và hướng phát triển.

# **CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU**

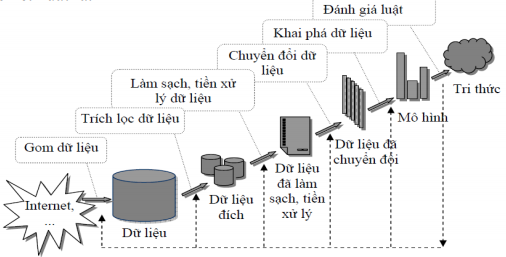
* 1. **Phát hiện tri thức và khai phá dữ liệu.**

Phát hiện tri thức (*Knowledge Discovery*) trong các cơ sở dữ liệu là một qui trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được. [1]

Khai phá dữ liệu (Data mining) được định nghĩa như sau: “*Data mining là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn*”.

Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin.

**1.2 Quy trình khám phá tri thức trong CSDL.**



**Hình 1.1 Quá trình khai phá dữ liệu từ cơ sở dữ liệu**

Quá trình phát hiện tri thức bao gồm các bước:

* **Làm sạch dữ liệu:** Các nhiễu và dữ liệu không nhất quán sẽ được loại bỏ.
* **Tích hợp dữ liệu:** Dữ liệu từ nhiều nguồn khác có thể được tổ hợp lại.
* **Lựa chọn dữ liệu:** Những dữ liệu thích hợp với nhiệm vụ phân tích sẽ được trích rút ra từ CSDL.
* **Chuyển đổi dữ liệu:** Dữ liệu sau khi được chọn lọc sẽ được chuyển đổi hay hợp nhất về dạng thích hợp cho việc khai phá.
* **Khai phá dữ liệu:** Quá trình cốt lõi, tất yếu trong đó các phương pháp thông minh sẽ được áp dụng nhằm trích rút ra các mẫu dữ liệu.
* **Đánh giá mẫu:** Các nhà phân tích dữ liệu sẽ dựa trên một số độ đo nào đó để xác định lợi ích thực sự, độ quan trọng của các mẫu biểu diễn tri thức.
* **Biểu diễn tri thức:** Giai đoạn này các kĩ thuật biểu diễn và hiển thị tri thức sẽ được sử dụng để đưa tri thức đã lấy ra đến người dùng.
  1. **Mô tả bài toán chuẩn đoán bệnh.**
     1. **Tổng quan bài toán.**

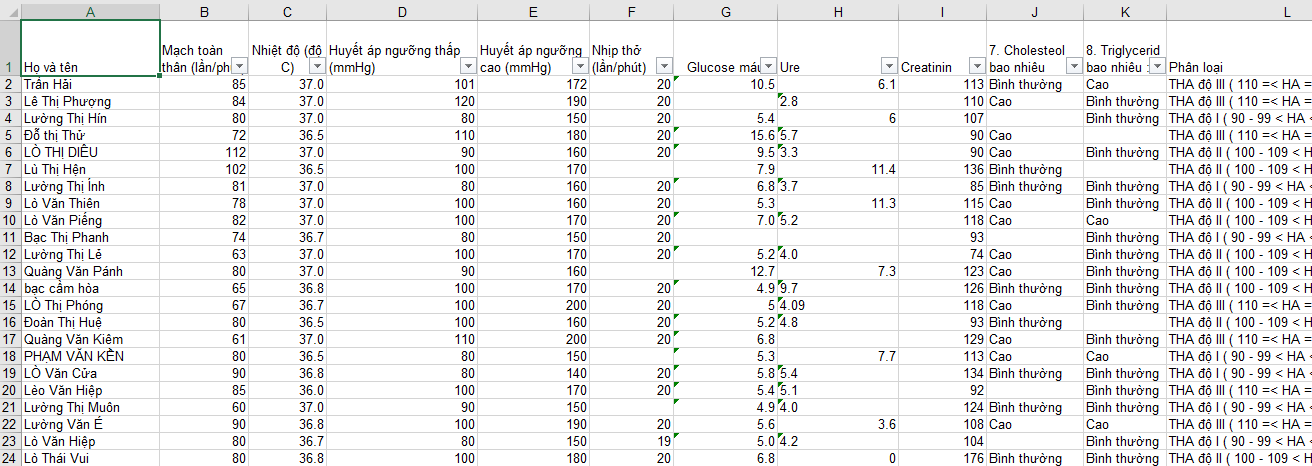
Dataset gồm các mô tả về các thuộc tính tương ứng với chuẩn đoán bệnh tăng huyết áp. Áp dụng các thuật toán để xác định xem đối tượng mắc bệnh tăng huyết áp độ: THA độ I, THA độ II, THA độ III.

* + 1. **Phân tích dữ liệu thô.**

Nguồn dữ liệu thô: Bệnh viện đa khoa Thái Bình.

*+ Hiểu dữ liệu:* Dữ liệu xét nghiệm chuẩn đoán bệnh huyết áp. Phân loại độ tăng ghuyết áp dựa trên giá trị các trọng sộ xét nghiệm.

+ *Dữ liệu gồm:* Dữ liệu bao gồm 399 bản ghi cùng 12 thuộc tính chuẩn đoán bệnh huyết áp.



**Hình 1.2 Dữ liệu ban đầu.**

*Hiểu các thuộc tính:*

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa thuộc tính** |
| 1 | Họ và tên | Họ và tên của bệnh nhân |
| 2 | Mạch toàn thân | Mạch toàn thân của bệnh nhân |
| 3 | Nhiệt độ | Nhiệt độ cơ thể của bệnh nhân |
| 4 | Huyết áp ngưỡng thấp | Huyết áp ngưỡng thấp của bệnh nhân |
| 5 | Huyết áp ngưỡng cao | Huyết áp ngưỡng cao của bệnh nhân |
| 6 | Nhịp thở | Nhịp thở của bệnh nhân |
| 7 | Glucose máu | Nồng độ glucose trong máu |
| 8 | Ure | Nồng độ ure đo được trong máu |
| 9 | Creatinin | Nồng độ creatinin trong máu |
| 10 | Cholesterol | Cholesterol của bệnh nhân |
| 11 | Triglycerid | Triglycerid của bệnh nhân |
| 12 | Phân loại | Phân loại xem bệnh nhân mắc bệnh hoặc không mắc bệnh.  Có các lớp bệnh:  + THA độ I (90 - 99 < HA < 140 - 159) + THA độ II (100 - 109 < HA < 160 - 179)  + THA độ III (110 =< HA =< 180)  + HA bình thường cao (85 - 89 < HA < 130 - 139) |

# **CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU**

## **2.1 Làm sạch dữ liệu.**

Là quá trinh 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. [2]

1. Xử lý dữ liệu bị thiếu (missing data)
2. Xử lý dữ liệu liệu, không nhất quán (inconsistent data).

Thực hiện:

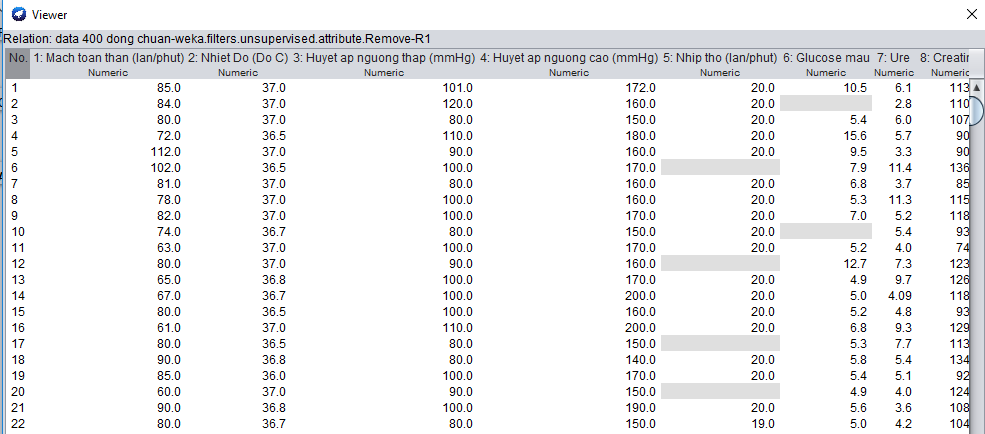
* Xử lý trên exell:

+ Xử lý nhất quán dữ liệu.

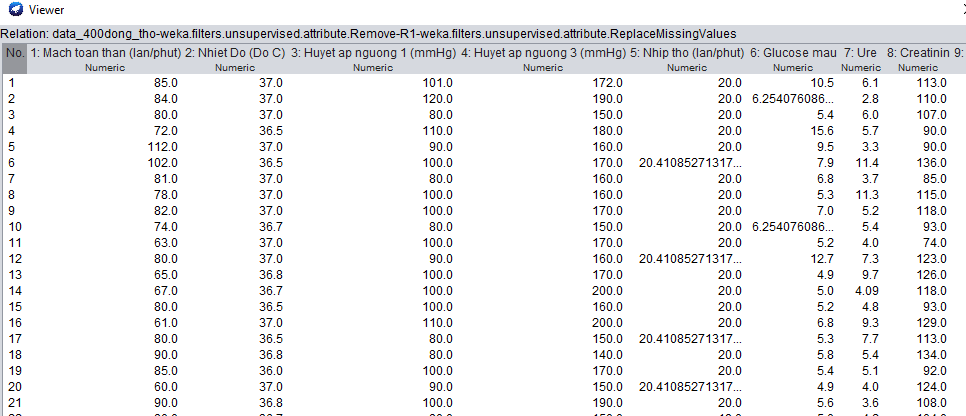
+ Loại bỏ dấu lưu file dưới định dạng .csv, tiền xử lý trên Weka.

* Đọc dữ liệu vào weka:

+ Loại bỏ thuộc tính họ và tên không ảnh hưởng tới phân loại bệnh.



+ Các dữ liệu thiếu thay thế bằng giá trị trung bình của thuộc tính dùng bộ lọc ReplaceMissingValue.



## **2.2 Tích hợp dữ liệu.**

Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu.

1. Tích hợp lược đồ và so trùng đối tượng.
2. Vấn đề dư thừa.
3. Phát hiện và xử lý mẫu thuẫn giá trị dữ liệu.

=> Dữ liệu lấy từ một nguồn nên không cần thực hiện quá trình này.

## **2.3 Biến đổi dữ liệu.**

Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu.

Các chiến lược thu giảm:

+ Làm trơn dữ liệu.

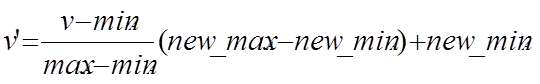
+ Kết hợp dữ liệu.

+ Tổng quát hóa dữ liệu.

+ Chuẩn hóa dữ liệu.

+ Xây dựng thuộc tính đặc tính.

* *Thực hiện chuẩn hóa dữ liệu:*



*Trong đó:* v = [minA, maxA] là giá trị cũ.

v’ = [0, 1] là giá trị mới.

Ví dụ:

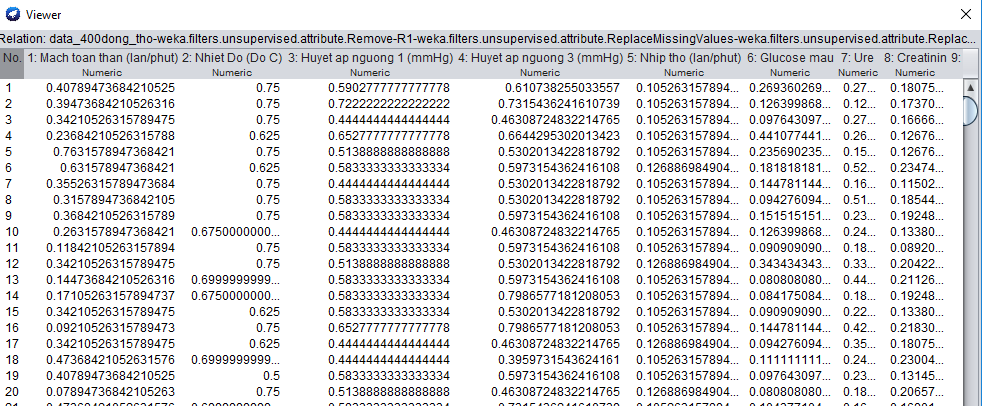
Thuộc tính Mạch toàn thân: Có v = [54, 130].

Chuẩn hóa về giá trị: v’ = [0, 1].

Với v = 85.

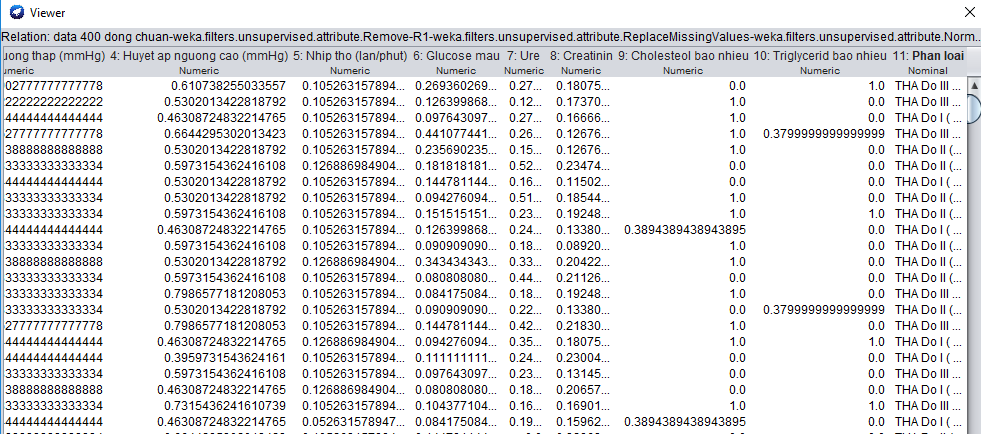
* v’= (85 – 54)/(130-54)\*(1-0) + 0 = 0.40789

Tiến hành chuẩn hóa các thuộc tính số về đoạn [0, 1] bằng phương pháp chuẩn hóa min-max bằng bộ lọc Normalize, lưu file lại dưới định dạng csv.

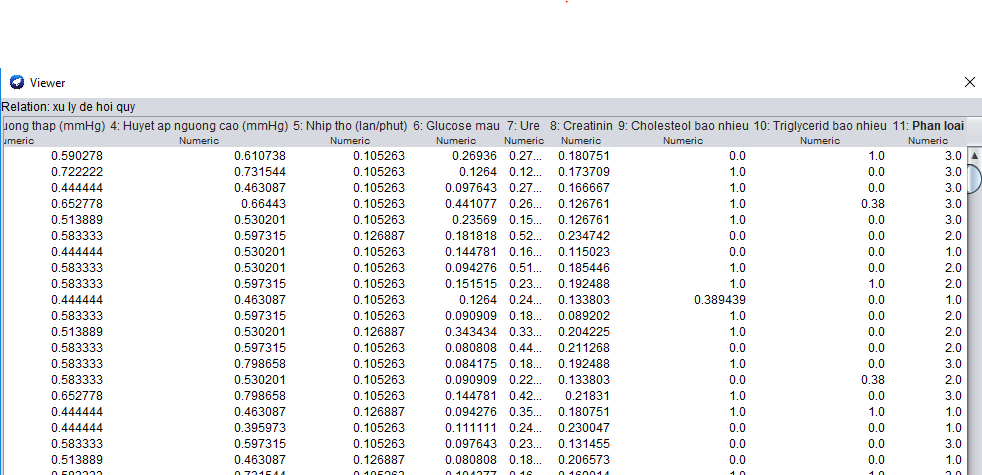


* *Chuẩn hóa với mô hình hồi quy tuyến tính.*
* Xử lý trên exell:

+ Nhìn vào tập dữ liệu ta thấy có 2 thuộc tính rời rạc (norminal): cholesterol và triglycerid chuyển về kiểu numeric bằng cách thay thế giá trị Cao = 3, Trung binh = 2, Thap = 1.



+ Đưa cột Phân loại về dạng numeric đối với áp dụng thuật toán hồi quy tuyến tính:



* **Chuẩn hóa với mô hình phân lớp:**

Dữ liệu của thuộc tính rời rạc cột phân loại đã là dạng norminal rồi nên vẫn giữ nguyên.

# **CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN HỒI QUY**

## **3.1 Giới thiệu về kỹ thuật hồi quy.**

* **Khái niệm:**

*+* Theo R.D.snee (1977): *Hồi quy là kỹ thuật thống kê trong lĩnh vực phân tích dữ liệu và xây dựng các mô hình từ thực nghiệm*, cho phép mô hình hồi quy vừa được khám phá được dùng cho mục đích dự báo, điều khiển, hay học cơ chế đã tạo ra dữ liệu.

*+* Theo Wiki (2009): *Hồi quy là kỹ thuật thống kê cho phép ước* lượng các mối liên kết giữa các biến. [1]

* **Mô hình và phương trình:**

+ Mô hình mô tả mỗi liên kết giữa một tập các biến dự báo/độc lập và một hay nhiều biến đáp ứng/phụ thuộc.

**Y = f (X, β)**

Trong đó:

- X: Các biến dự báo/độc lập. Dùng để giải thích sự biến đổi của các đáp ứng Y.

- Y: Các biến đáp ứng/phụ thuộc. Dùng để mô tả các hiện tượng được quan tâm/giải thích.

- β: Các hệ số hồi quy, mô tả ảnh hưởng của X đối với Y.

## **3.2 Thuật toán hồi quy tuyến tính đa biến.**

* **Định nghĩa:** Hồi quy tuyến tính đa biến: Phân tích mỗi quan hệ giữa biến phụ thuộc và hai hay nhiều biến độc lập. [4]

Phương trình: **Y = B0 + B1\*X1 + B2\*X2 + …**

Trong đó:

+ I = 1…n với n là số đối tượng đã quan sát.

+Y: Biến phụ thuộc.

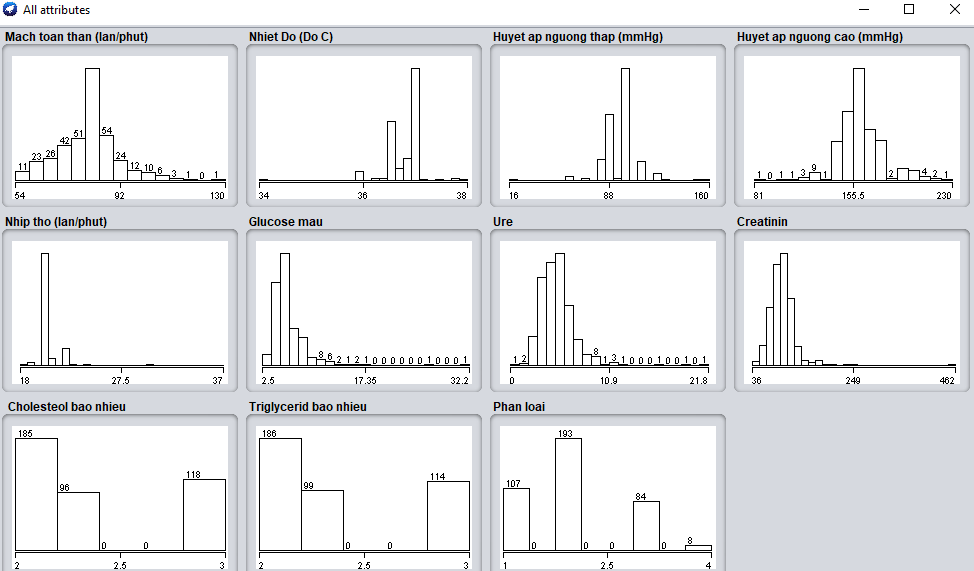
+ X: Biến độc lập

+ B0: Giá trị của Y khi X = 0.

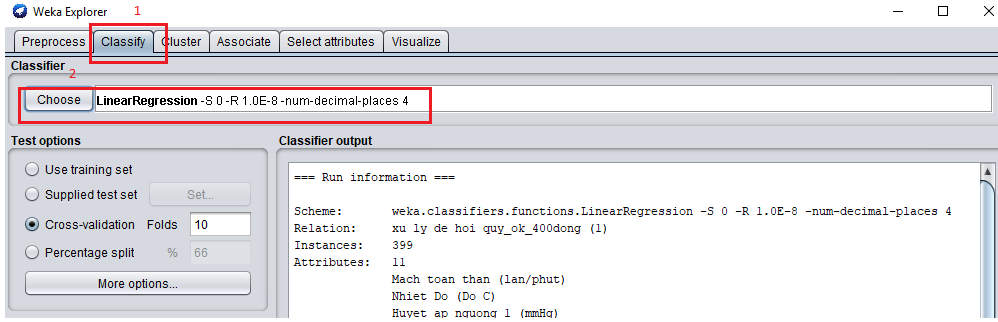
+ B1 …k: Trị số các hệ số hồi quy.

* **Các bước:**

**Bước 1:** Đọc dữ liệu vào weka. Dữ liệu trong dataset như sau:



**Bước 2:** Trên giao diện weka chọn Classify and chọn LinearRegression.



**Bước 3: Tiến hành đánh giá hiệu quả hồi quy của thuật toán đối với tập dữ liệu được dùng theo hai phương pháp:**

*Phương pháp 1: Theo Cross-validation.*

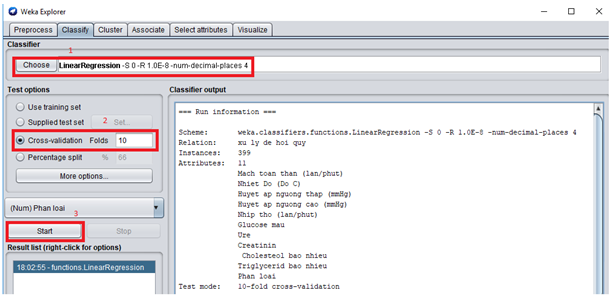
Tập dữ liệu sẽ được chia thành k tập có kích thước xấp xỉ nhau, và phương trình học được sẽ được đánh giá bởi phương pháp cross-validation.

* Dữ liệu gốc sẽ được chia thành n phần bằng nhau. Quá trình Train/Test model sẽ được thực hiện lặp lại n lần. 1 phần dữ liệu dùng để Test và (n-1) phần dùng để training.

*Phương pháp 2: Theo Percentage split.*

* Cho biết tỷ lệ phân chia là bao nhiêu phần trăm thì đạt hiệu quả phân lớp cao nhất.

**(1) Cross-validation:**



**Kết quả chạy:**

*=== Run information ===*

*Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4*

*Relation: xu ly de hoi quy*

*Instances: 399*

*Attributes: 11*

*Mach toan than (lan/phut)*

*Nhiet Do (Do C)*

*Huyet ap nguong thap (mmHg)*

*Huyet ap nguong cao (mmHg)*

*Nhip tho (lan/phut)*

*Glucose mau*

*Ure*

*Creatinin*

*Cholesteol bao nhieu*

*Triglycerid bao nhieu*

*Phan loai*

*Test mode: 10-fold cross-validation*

*=== Classifier model (full training set) ===*

*Linear Regression Model*

*Phan loai =*

*0.0058 \* Mach toan than (lan/phut) +*

*0.0225 \* Huyet ap nguong cao (mmHg) +*

*0.042 \* Ure +*

*0.1197 \* Cholesteol bao nhieu +*

*-2.6113*

*Time taken to build model: 0.15 seconds*

*=== Cross-validation ===*

*=== Summary ===*

*Correlation coefficient 0.5187*

*Mean absolute error 0.4498*

*Root mean squared error 0.6454*

*Relative absolute error 83.1135 %*

*Root relative squared error 84.9725 %*

*Total Number of Instances 392*

*Ignored Class Unknown Instances 7*

**Giải thích kết quả:**

=> Kết quả trả về theo 3 vùng dữ liệu:

**Run information. -------- Thông tin chạy.**

* Cho biết thông tin về dữ liệu nguồn.

*+ Đề án sử dụng:*

Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

*+ Cơ sở dữ liệu:* xu ly de hoi quy

+ Số trường :14.

+ Số thuộc tính 399.

+ Chế độ kiểm tra:

*10-fold cross-validation*

**Classifier model ------- Mô hình.**

* Vùng hiển thị kết quả training. Chế độ phân lớp toàn bộ kết quả training.
* Mô hình hồi quy tuyến tính sau khi phân lớp:

Phan loai =

     0.0058 \* Mach toan than (lan/phut) +

     0.0225 \* Huyet ap nguong cao (mmHg) +

     0.042 \* Ure +

     0.1197 \*  Cholesteol bao nhieu +

    -2.6113

* Thời gian tiến hành: 0.15s.

**Cross-validation -------- Đánh giá mô hình.**

*Correlation coefficient 0.5187*

*Mean absolute error 0.4498*

*Root mean squared error 0.6454*

+ Hệ số tương quan: 0.5187.

+ Sai số tuyệt đối trung bình: 0.4498.

+ Sai số bình phương trung bình: 0.6454.

+ Độ sai tuyệt đối trung bình: 83.1135%.

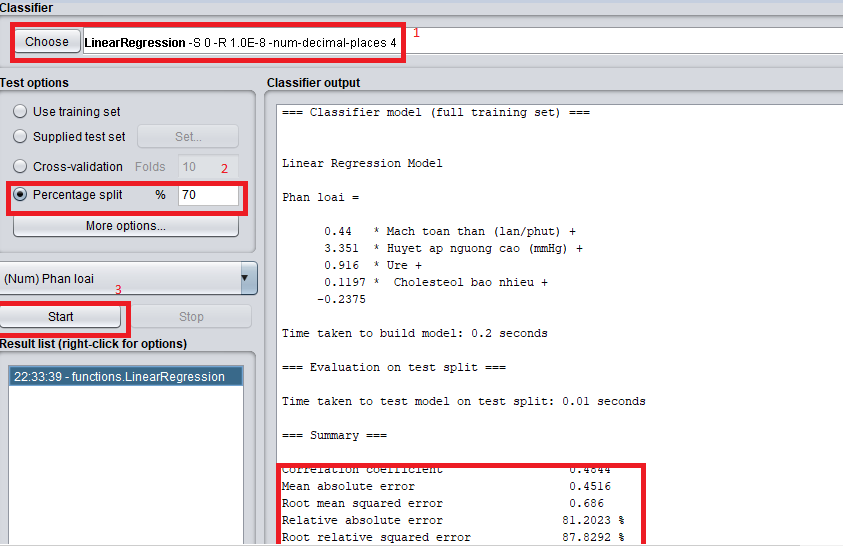
+ Sai số bình phương trung bình: 84.9725%

+ Tổng số bản ghi: 392.

+ Bỏ qua trường hợp không xác định: 7.

**(2) Percentage split:**

Train/Test: 70/30%.



**Kết quả chạy cho thấy:**

+ Sai số tuyệt đối trung bình: 0.4516

+ Sai số bình phương trung bình: 0.686.

* **Nhận xét:**

***+*** *Kết quả* cho thấy chạy bằng phương pháp cross-validation có kết quả sai số tốt nhất.

***+*** *Tri thức thu được* từ 1 tập dữ liệu đầu vào gồm có 4 giá trị: trung bình mạch toàn thân, trung bình huyết áp ngưỡng cao, trung bình ure, trung bình choresterol. Từ đó ta có thể tính được mức bị bệnh nhân huyết áp bằng mô hình hồi quy tuyến tính thu được sau khi chạy thuật toán.

# **CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN LỚP\**

## **4.1 Giới thiệu về bài toán phân lớp.**

Phân lớp dữ liệu là xếp các đối tượng DL vào một trong các lớp đã được xác định trước. Gồm 2 bước: [1]

*Bước 1: Xây dựng mô hình.*

+ Mô tả tập các lớp xác định trước.

+ Tập học/huấn luyện: Các mẫu dành cho xây dựng mô hình.

+ Mỗi mẫu thuộc về 1 lớp đã định nghĩa trước.

+ Tìm luật phân lớp, cây quyết định hoặc công thức toán mô tả lớp.

*Bước 2: Vận hành mô hình.*

+ Phân lớp các đối tượng chưa biết.

+ Xác định độ chính xác của mô hình, sử dụng tập dữ liệu kiểm ta độc lập.

+ Độ chính xác chấp nhận được -> áp dụng mô hình để phân lớp các mẫu chưa xác định được nhãn lớp.

## **4.2 Thuật toán phân lớp J48.**

**Giới thiệu về cây quyết định**

Cây quyết định (Decision tree) có cấu trúc được dùng để phân lớp các đối tượng

dựa vào dãy các luật (series of rules).

Các thuộc tính của đối tượng (ngoại trừ thuộc tính phân lớp – Category attribute) có thể thuộc các kiểu dữ liệu khác nhau (Binary, Nominal, ordinal quantitive values) trong khi đó thuộc tính phân lớp phải có dữ liệu là Binary hoặc Ordinal. [3]

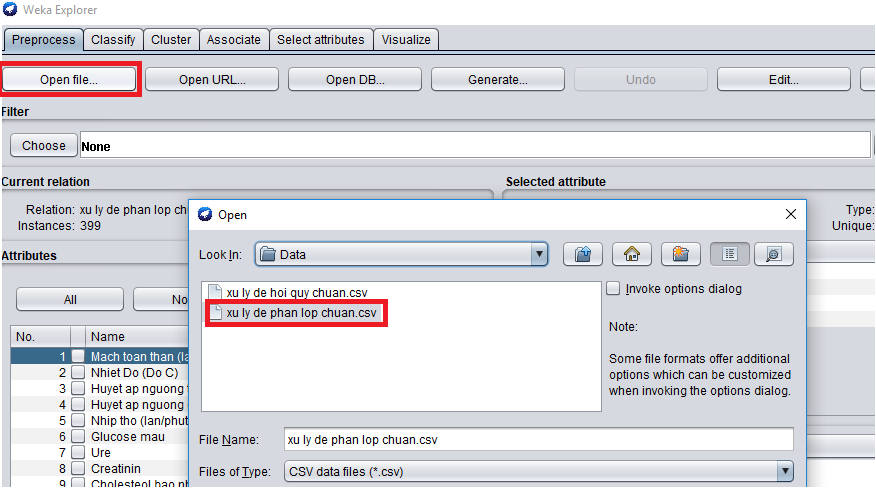
* Các thuộc tính cùng với lớp của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết

Quy trình train và test của một classifer:

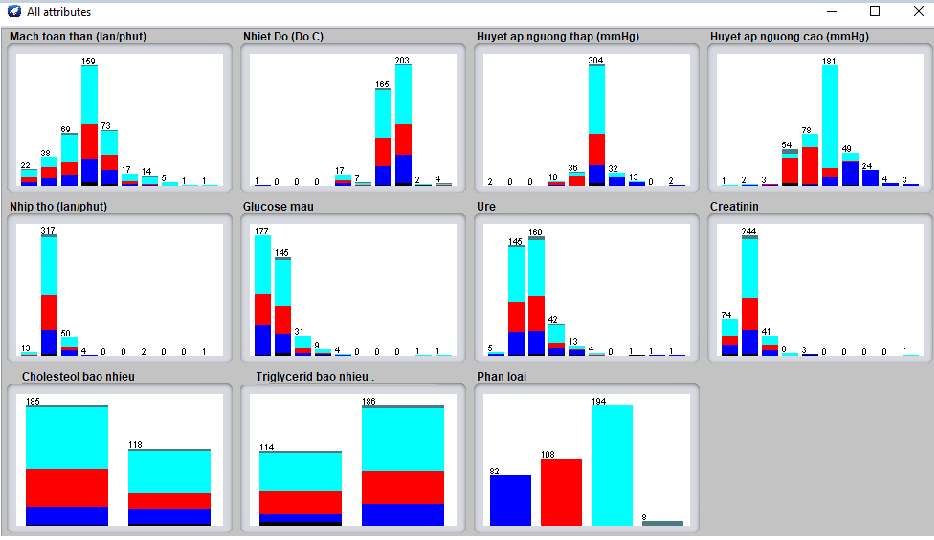
* Dữ liệu để xây dựng mô hình: dữ liệu gốc dữ liệu này phải có thuộc tính phân lớp gọi là Categorical attribute
* Dữ liệu gốc sẽ được chia thành 2 phần là Train set và Testing set
* Cuối cùng là tính toán lỗi để đánh giá model.

**Các bước:**

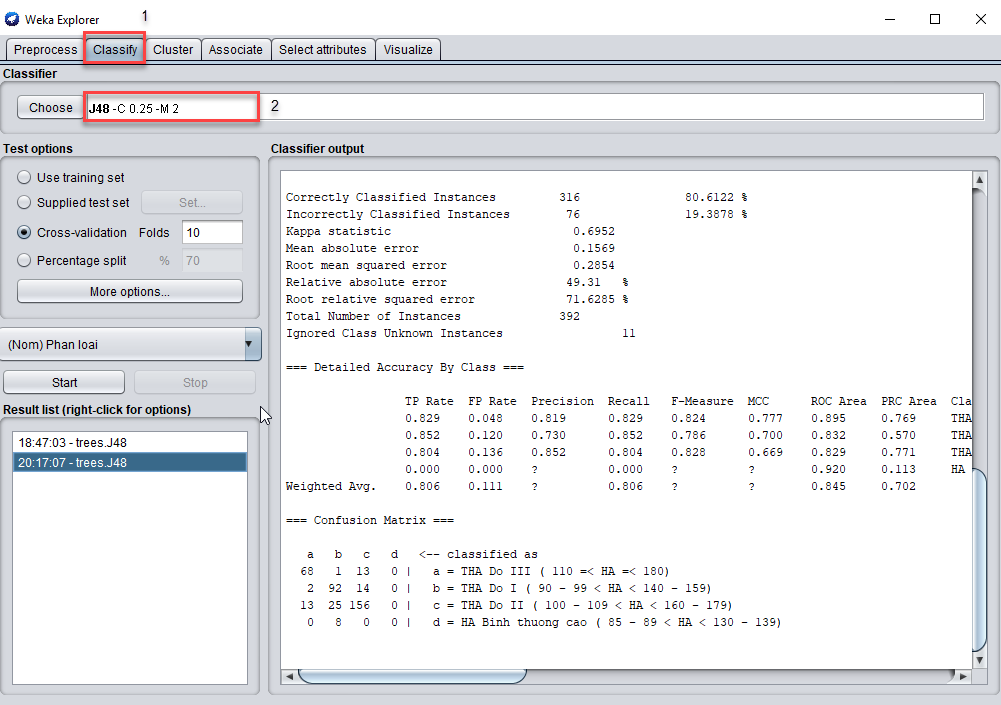
**Bước 1:** Mở file .csv đã tiền xử lý bằng Weka trước đó trên Weka



Dataset như sau:



**Bước 2:** Trên giao diện Weka chọn Classify sau đó chọn J48.

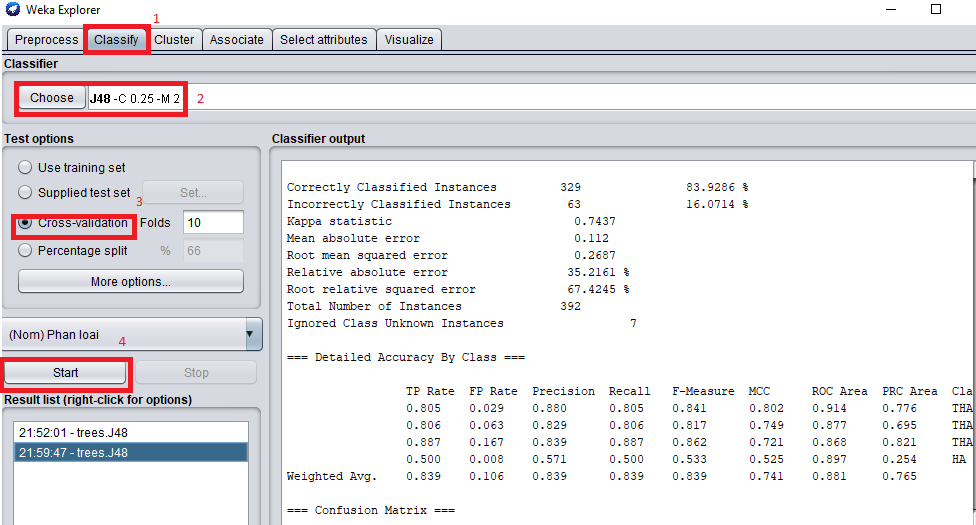


**Bước 3: Tiến hành đánh giá hiệu quả phân lớp của thuật toán đối với tập dữ liệu được dùng theo hai phương pháp:**

*Phương pháp 1: Theo Cross-validation.*

*Phương pháp 2: Theo Percentage split.*

**(1) Cross-validation:**



**Kết quả chạy :**

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: xu ly de phan lop chuan

Instances: 399

Attributes: 11

Mach toan than (lan/phut)

Nhiet Do (Do C)

Huyet ap nguong thap (mmHg)

Huyet ap nguong cao (mmHg)

Nhip tho (lan/phut)

Glucose mau

Ure

Creatinin

Cholesteol bao nhieu

Triglycerid bao nhieu

Phan loai

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

Huyet ap nguong cao (mmHg) <= 0.503356

| Huyet ap nguong thap (mmHg) <= 0.548611

| | Huyet ap nguong cao (mmHg) <= 0.261745

| | | Creatinin <= 0.14108: THA Do III ( 110 =< HA =< 180) (3.0)

| | | Creatinin > 0.14108: THA Do II ( 100 - 109 < HA < 160 - 179) (3.0/1.0)

| | Huyet ap nguong cao (mmHg) > 0.261745

| | | Glucose mau <= 0.127946: THA Do I ( 90 - 99 < HA < 140 - 159) (76.0/5.0)

| | | Glucose mau > 0.127946

| | | | Huyet ap nguong cao (mmHg) <= 0.395973

| | | | | Creatinin <= 0.159624: HA Binh thuong cao ( 85 - 89 < HA < 130 - 139) (8.0/1.0)

| | | | | Creatinin > 0.159624: THA Do I ( 90 - 99 < HA < 140 - 159) (4.0)

| | | | Huyet ap nguong cao (mmHg) > 0.395973: THA Do I ( 90 - 99 < HA < 140 - 159) (13.0/1.0)

| Huyet ap nguong thap (mmHg) > 0.548611

| | Nhiet Do (Do C) <= 0.675

| | | Nhip tho (lan/phut) <= 0.126887

| | | | Mach toan than (lan/phut) <= 0.526316: THA Do I ( 90 - 99 < HA < 140 - 159) (4.0)

| | | | Mach toan than (lan/phut) > 0.526316: THA Do II ( 100 - 109 < HA < 160 - 179) (2.0)

| | | Nhip tho (lan/phut) > 0.126887: THA Do II ( 100 - 109 < HA < 160 - 179) (2.0)

| | Nhiet Do (Do C) > 0.675

| | | Nhip tho (lan/phut) <= 0.157895: THA Do II ( 100 - 109 < HA < 160 - 179) (15.0)

| | | Nhip tho (lan/phut) > 0.157895: THA Do I ( 90 - 99 < HA < 140 - 159) (3.0/1.0)

Huyet ap nguong cao (mmHg) > 0.503356

| Huyet ap nguong cao (mmHg) <= 0.604027

| | Huyet ap nguong thap (mmHg) <= 0.513889

| | | Huyet ap nguong thap (mmHg) <= 0.375: THA Do I ( 90 - 99 < HA < 140 - 159) (2.0)

| | | Huyet ap nguong thap (mmHg) > 0.375

| | | | Nhiet Do (Do C) <= 0.65

| | | | | Nhiet Do (Do C) <= 0.6: THA Do II ( 100 - 109 < HA < 160 - 179) (2.0)

| | | | | Nhiet Do (Do C) > 0.6: THA Do I ( 90 - 99 < HA < 140 - 159) (6.0/1.0)

| | | | Nhiet Do (Do C) > 0.65: THA Do II ( 100 - 109 < HA < 160 - 179) (31.0/6.0)

| | Huyet ap nguong thap (mmHg) > 0.513889

| | | Huyet ap nguong thap (mmHg) <= 0.583333: THA Do II ( 100 - 109 < HA < 160 - 179) (121.0/5.0)

| | | Huyet ap nguong thap (mmHg) > 0.583333

| | | | Ure <= 0.321101: THA Do II ( 100 - 109 < HA < 160 - 179) (15.0/3.0)

| | | | Ure > 0.321101: THA Do III ( 110 =< HA =< 180) (4.0)

| Huyet ap nguong cao (mmHg) > 0.604027

| | Creatinin <= 0.241784: THA Do III ( 110 =< HA =< 180) (72.0/7.0)

| | Creatinin > 0.241784: THA Do II ( 100 - 109 < HA < 160 - 179) (6.0/1.0)

Number of Leaves : 20

Size of the tree : 39

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 329 83.9286 %

Incorrectly Classified Instances 63 16.0714 %

Kappa statistic 0.7437

Mean absolute error 0.112

Root mean squared error 0.2687

Relative absolute error 35.2161 %

Root relative squared error 67.4245 %

Total Number of Instances 392

Ignored Class Unknown Instances 7

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.805 0.029 0.880 0.805 0.841 0.802 0.914 0.776 THA Do III ( 110 =< HA =< 180)

0.806 0.063 0.829 0.806 0.817 0.749 0.877 0.695 THA Do I ( 90 - 99 < HA < 140 - 159)

0.887 0.167 0.839 0.887 0.862 0.721 0.868 0.821 THA Do II ( 100 - 109 < HA < 160 - 179)

0.500 0.008 0.571 0.500 0.533 0.525 0.897 0.254 HA Binh thuong cao ( 85 - 89 < HA < 130 - 139)

Weighted Avg. 0.839 0.106 0.839 0.839 0.839 0.741 0.881 0.765

=== Confusion Matrix ===

a b c d <-- classified as

66 1 15 0 | a = THA Do III ( 110 =< HA =< 180)

1 87 18 2 | b = THA Do I ( 90 - 99 < HA < 140 - 159)

8 13 172 1 | c = THA Do II ( 100 - 109 < HA < 160 - 179)

0 4 0 4 | d = HA Binh thuong cao ( 85 - 89 < HA < 130 - 139)

**Đọc nội dung kết quả** : kết quả được trả về theo 3 vùng dữ liệu

* Vùng Run Information: Cho biết thông tin về nguồn dữ liệu
* Đề án sử sử dụng: weka.classifiers.trees.J48 -C 0.25 -M 2
* Cơ sở dữ liệu: xu ly de phan lop chuan
* Các trường: 399
* Thuộc tính: 11

Mach toan than (lan/phut)

Nhiet Do (Do C)

Huyet ap nguong thap (mmHg)

Huyet ap nguong cao (mmHg)

Nhip tho (lan/phut)

Glucose mau

Ure

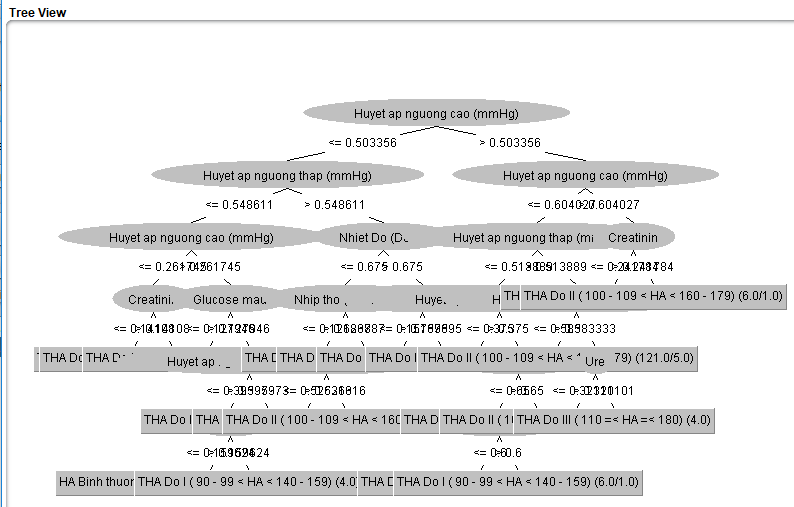
Creatinin

Cholesteol bao nhieu

Triglycerid bao nhieu :

Phan loai

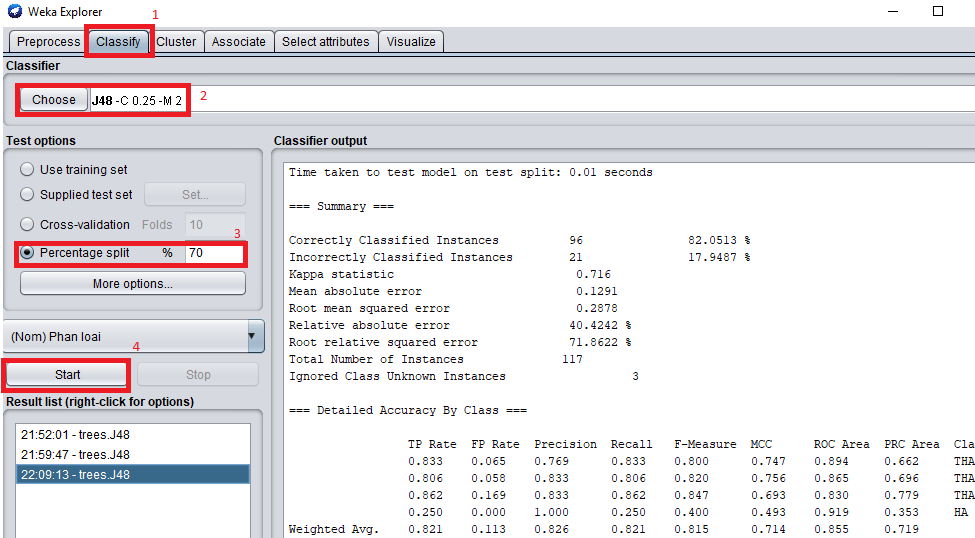
* Chế độ kiểm tr : 10-fold cross-validation
* Vùng hiển thị kết quả trainning :
* Chế độ phân lớp toàn bộ dữ liệu
* Cây J48 sau khi tiến hành trainning:
* Số lượng lá: 20
* Kích thước cây: 39



* **Tỷ lệ TP** : Tỷ lệ dương tính thực (các trường hợp được phân loại chính xác như một lớp đã cho)
* **Tỷ lệ FP** : Tỷ lệ dương tính giả (các trường hợp được phân loại sai là một lớp nhất định)
* **Độ chính xác** : Tỷ lệ các trường hợp thực sự là một lớp chia cho tổng số các trường hợp được phân loại là lớp đó
* **Recall**: Tỷ lệ các trường hợp được phân loại là một lớp nhất định chia cho tổng số thực tế trong lớp đó (tương đương với tỷ lệ TP)
* **F-Measure** : Một biện pháp kết hợp cho độ chính xác và thu hồi được tính bằng 2 \* Precision \* Recall / (Precision + Recall).
* Tóm tắt kết quả xác nhận phân lớp:
* Trường hợp phân lớp chính xác: 329 83.9216%
* Trường hợp phân lớp không chính xác: 63 19.3878%
* Thông kê lỗi.

**(2) Percentage split:**

Train: 70%, Test: 30%.



**Kết quả chạy cho thấy:**

* Tóm tắt kết quả xác nhận phân lớp:
* Trường hợp phân lớp chính xác: 96 82.0513%
* Trường hợp phân lớp không chính xác: 21 17.9487 %
* **Nhận xét:**

Kết quả cho thấy, sử dụng phương pháp cross-validation cho kết quả chuẩn xác nhất.

Một số tri thức thu được:

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp < = 0.5486 và creatinin <= 0.14108 thì THA độ III (110 =< HA =< 180).

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp < = 0.5486 và creatinin > 0.14108 thì THA độ II (100 - 109 < HA < 160 – 179).

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp < = 0.5486 và Glucose mau <= 0.1279 thì THA độ I (90 - 99 < HA < 140 – 159).

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp < = 0.5486 và Glucose mau > 0.127946 và creatinin <= 0.159624 thì HA bình thường cao (85 - 89 < HA < 130 – 139).

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp < = 0.5486 và Glucose mau > 0.127946 và creatinin > 0.159624 thì THA độ I.

+ Huyết áp ngưỡng cáo > 0.395973 thì THA độ I (90 - 99 < HA < 140 – 159).

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp > 0.548611 và nhiệt độ < 0.675 và nhịp thở <= 0.126887 và mạch toàn thân <= 0.526316 thì THA độ I.

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp > 0.548611 và nhiệt độ < 0.675 và nhịp thở <= 0.126887 và mạch toàn thân > 0.526316 thì THA độ II.

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp > 0.548611 và nhiệt độ < 0.675 và nhịp thở > 0.126887 thì THA độ II.

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp > 0.548611 và nhiệt độ > 0.675 và nhịp thở <= 0.157895 thì THA độ II.

+ Huyết áp ngưỡng cao <= 0.50336 và huyết áp ngưỡng thấp > 0.548611 và nhiệt độ > 0.675 và nhịp thở > 0.157895 thì THA độ I.

+ Huyết áp ngưỡng cao > 0.50336 và huyết áp ngưỡng thấp <= 0.375 và nhiệt độ <= 0.6 thì THA độ II.

+ Huyết áp ngưỡng cao > 0.50336 và huyết áp ngưỡng thấp <= 0.375 và nhiệt độ >6 thì THA độ I.

+ Huyết áp ngưỡng cao > 0.50336 và huyết áp ngưỡng thấp <= 0.375 và nhiệt độ > 0.65 thì THA độ II.

+ Huyết áp ngưỡng cao > 0.50336 và huyết áp ngưỡng thấp <= 0.583333 thì THA độ II.

+ Huyết áp ngưỡng cao > 0.50336 và huyết áp ngưỡng thấp >0.583333 thì THA độ II và ure <= 0.321101 thì THA độ II.

+ Huyết áp ngưỡng cao > 0.50336 và huyết áp ngưỡng thấp >0.583333 thì THA độ II và ure > 0.321101 thì THA độ III.

+ Huyết áp ngưỡng cao > 0.50336 và creatinin <= 0.241784 thì THA độ III.

+ Huyết áp ngưỡng cao > 0.50336 và creatinin > 0.241784 thì THA độ II.

**So sánh độ đo của J48 với thuật toán Naïve Bayes: Phương pháp cross-validation.**

|  |  |  |
| --- | --- | --- |
| Độ do | Cây quyết định | Navie Bayes |
| MAE | 0.112 | 0.1732 |
| RMSE | 0.2687 | 0.3152 |

## **4.4 Đánh giá mô hình.**

Từ 2 giải thuật đã đưa ra được tập luật và mô hình để phân lớp bệnh tăng huyết áp. Có thể dễ dàng nhận thấy sai số của mô hình sử dụng thuật toán J48 có độ chính xác và độ đo đánh giá mô hình tốt hơn mô hình của thuật toán Navie Bayes. Vì vậy mô hình cây quyết định sử dụng thuật toán J48 có tính ứng dụng cao.

# **CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **3.1 Kết luận.**

Hồi quy tuyến tính và phân lớp là 2 lĩnh vực khá quan trọng trong khai phá dữ liệu nhằm đưa ra những dự đoán, xu hướng trong tương lai, nó được ứng dụng trong nhiều ngành như y tế, thương mại, ngân hàng, giáo dục ... Trong quá trình tìm hiểu và hoàn thành bài tập lớn với đề tài “*Khai phá dữ liệu bệnh nhân bằng phương pháp hồi quy và tuyến tính*”. **Nhóm em đã đạt được một số kết quả như sau:**

* Tìm hiểu tổng quan về khai phá dự liệu, bài toán phân lớp, phương pháp hồi quy tuyến tính và thuật toán J48 để từ đó xây dựng mô hình hồi quy và mô hình phân lớp hỗ trợ chẩn đoán bệnh.
* Thu thập dữ liệu bệnh nhân, tiền xử lý dữ liệu bằng exell và weka. Xây dựng nên mô hình hồi quy và cây quyết định trên phần mềm weka.
* So sánh kết quả tỷ lệ train/test để lựa chọn tỷ lệ đánh giá mô hình tốt nhất.
* Đánh giá mô hình phân lớp so sánh giữa 2 phương pháp J48 và Navie bayes.

**Tuy nhiên bài tập nhóm vẫn còn một số hạn chế:**

* Việc thu thập dữ liệu chưa đầy đủ,chi tiết với nhiều bệnh khác nhau mà chỉ tập trung vào bệnh Tăng huyết áp, do vậy các căn bệnh khác chưa hỗ trợ chẩn đoán được bệnh.
* Kết quả dự đoán tương đối cao nhưng vẫn chưa được tốt nhất.
  1. **Hướng phát triển.**
* Xây dựng, cải tiến mô hình chẩn đoán bệnh với phương pháp học máy khác như SVM, KNN, Random Forest,…
* Phân tích, đánh giá phù hợp với bệnh dựa trên các mô hình học máy để áp dụng vào thực tiễn trợ giúp các bác sĩ trong quá trình chẩn đoán bệnh được tốt hơn.

Trong quá trình hoàn thành bài tập lớn, chúng em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em sẽ không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của quý thầy cô và các bạn để báo cáo và kỹ năng của chúng em ngày được hoàn thiện hơn và có thế áp dụng được trong thực tiễn.

# **TÀI LIỆU THAM KHẢO**

**1. Tài liệu tiếng việt.**

[1] TS.Đặng Thị Thu Hiền, (2019), Bài giảng Khai Phá dữ liệu.

[2] <https://ongxuanhong.wordpress.com/2015/08/20/tien-xu-ly-du-lieu-horse-colic-dataset/>

[3] <https://www.slideshare.net/tenzou2411/tiu-lun-khai-ph-d-liu-s-dng-wekazphn-lp-trn-dataset-weather-arff>

[4] <https://cuongndh.blogspot.com/p/khai-pha-du-lieu.html>

[5] <https://machinelearningcoban.com/2016/12/28/linearregression/>

[6] https://vi.wikipedia.org/wiki/Định\_lý\_Bayes.

**2. Tài liệu tiếng anh.**

[7] <https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/>

[8] <https://en.wikipedia.org/wiki/Mean_squared_error>.

[9] <https://en.wikipedia.org/wiki/Mean_absolute_error>.