MỞ ĐẦU

MỞ ĐẦU 3
CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 4
1.1 Phát hiện tri thức và khai phá dữ liệu. 4
1.2 Quy trình khám phá tri thức trong CSDL 4
1.3 Mô tả bài toán chuẩn đoán bệnh 4
1.3.1 Tổng quan bài toán 4
1.3.2 Phân tích dữ liệu thô 5
CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 7
2.1 Làm sạch dữ liệu. 7
2.2 Tích hợp dữ liệu. 8
2.3 Biến đổi dữ liệu. 8
CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG MÔ HÌNH HỒI QUY 11
3.1 Giới thiệu về kỹ thuật hồi quy 11
3.2 Mô hình hồi quy tuyến tính đa biến. 11
CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN LỚP\ 18
4.1 Giới thiệu về bài toán phân lớp. 18
4.2 Thuật toán phân lớp J48 18
4.4 Đánh giá mô hình 30
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 31
3.1 Kết luận. 31
3.2 Hướng phát triển. 31
TÀI LIỆU THAM KHẢO 32

MỞ ĐẦU

Trong những năm gần đây cùng với phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nổ 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 5 chương:

Chương 1: Tổng quan về khai phá dữ liệu.

Chương 2: Tiền xử lý dữ liệu.

Chương 3; Khai phá dữ liệu bằng mô hình 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

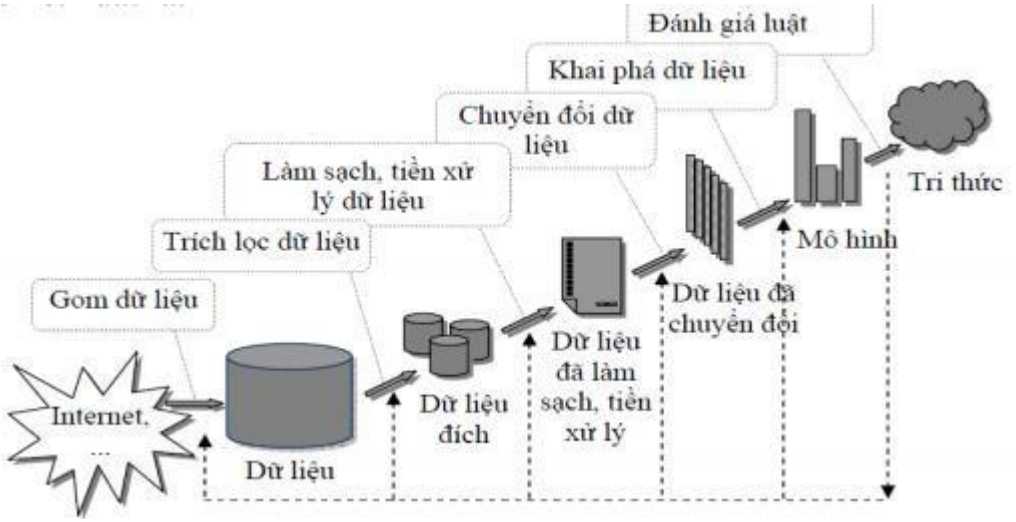
1.1 Phát hiện tri thức và khai phá dữ liệu.

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

1.3 Mô tả bài toán chuẩn đoán bệnh.

1.3.1 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, HA bình thường cao.

1.3.2 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 huyết áp dựa trên giá trị các trọng số xét nghiệm.

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

	A	B	C	D	E	F	G	H	I	J	K	L
1	Họ và tên	Mạch toàn thân (lần/ph	Nhiệt độ (độ C)	Huyết áp ngưỡng thấp (mmHg)	Huyết áp ngưỡng cao (mmHg)	Nhịp thở (lần/phút)	Glucose máu	Ure	Creatinin	7. Cholesteol bao nhiêu	8. Triglycerid bao nhiêu	Phân loại
2	Trần Hải	85	37.0	101	172	20	10.5	6.1	113	Bình thường	Cao	THA độ III (110 =< HA =
3	Lê Thị Phượng	84	37.0	120	190	20	2.8		110	Cao	Bình thường	THA độ III (110 =< HA =
4	Lường Thị Hín	80	37.0	80	150	20	5.4	6	107		Bình thường	THA độ I (90 - 99 < HA <
5	Đỗ thị Thử	72	36.5	110	180	20	15.6	5.7	90	Cao		THA độ III (110 =< HA =
6	LÒ THỊ DIỀU	112	37.0	90	160	20	9.5	3.3	90	Cao	Bình thường	THA độ II (100 - 109 < H
7	Lù Thị Hện	102	36.5	100	170		7.9	11.4	136	Bình thường		THA độ II (100 - 109 < H
8	Lường Thị Ính	81	37.0	80	160	20	6.8	3.7	85	Bình thường	Bình thường	THA độ I (90 - 99 < HA <
9	Lò Văn Thiên	78	37.0	100	160	20	5.3	11.3	115	Cao	Bình thường	THA độ II (100 - 109 < H
10	Lò Văn Piếng	82	37.0	100	170	20	7.0	5.2	118	Cao	Cao	THA độ II (100 - 109 < H
11	Bạc Thị Phanh	74	36.7	80	150	20			93		Bình thường	THA độ I (90 - 99 < HA <
12	Lường Thị Lê	63	37.0	100	170	20	5.2	4.0	74	Cao	Bình thường	THA độ II (100 - 109 < H
13	Quàng Văn Pánh	80	37.0	90	160		12.7	7.3	123	Cao	Bình thường	THA độ II (100 - 109 < H
14	bạc cầm hòa	65	36.8	100	170	20	4.9	9.7	126	Bình thường	Bình thường	THA độ II (100 - 109 < H
15	LÒ Thị Phóng	67	36.7	100	200	20	5	4.09	118	Cao	Bình thường	THA độ III (110 =< HA =
16	Đoàn Thị Huệ	80	36.5	100	160	20	5.2	4.8	93	Bình thường		THA độ II (100 - 109 < H
17	Quàng Văn Kiêm	61	37.0	110	200	20	6.8		129	Cao	Bình thường	THA độ III (110 =< HA =
18	PHẠM VĂN KÊN	80	36.5	80	150		5.3	7.7	113	Cao	Cao	THA độ I (90 - 99 < HA <
19	LÒ Văn Cửa	90	36.8	80	140	20	5.8	5.4	134	Bình thường	Bình thường	THA độ I (90 - 99 < HA <
20	Lèo Văn Hiệp	85	36.0	100	170	20	5.4	5.1	92		Bình thường	THA độ III (110 =< HA =
21	Lường Thị Muôn	60	37.0	90	150		4.9	4.0	124	Bình thường	Bình thường	THA độ I (90 - 99 < HA <
22	Lường Văn Ế	90	36.8	100	190	20	5.6	3.6	108	Cao	Cao	THA độ III (110 =< HA =
23	Lò Văn Hiệp	80	36.7	80	150	19	5.0	4.2	104		Bình thường	THA độ I (90 - 99 < HA <
24	Lò Thái Vui	80	36.8	100	180	20	6.8	0	176	Bình thường	Bình thường	THA độ II (100 - 109 < H

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) + HA bình thường.

CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

2.1 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. [2]

(1) Xử lý dữ liệu bị thiếu (missing data)

(2) Xử lý dữ liệu liệu, không nhất quán (inconsistent data).

Thực hiện:

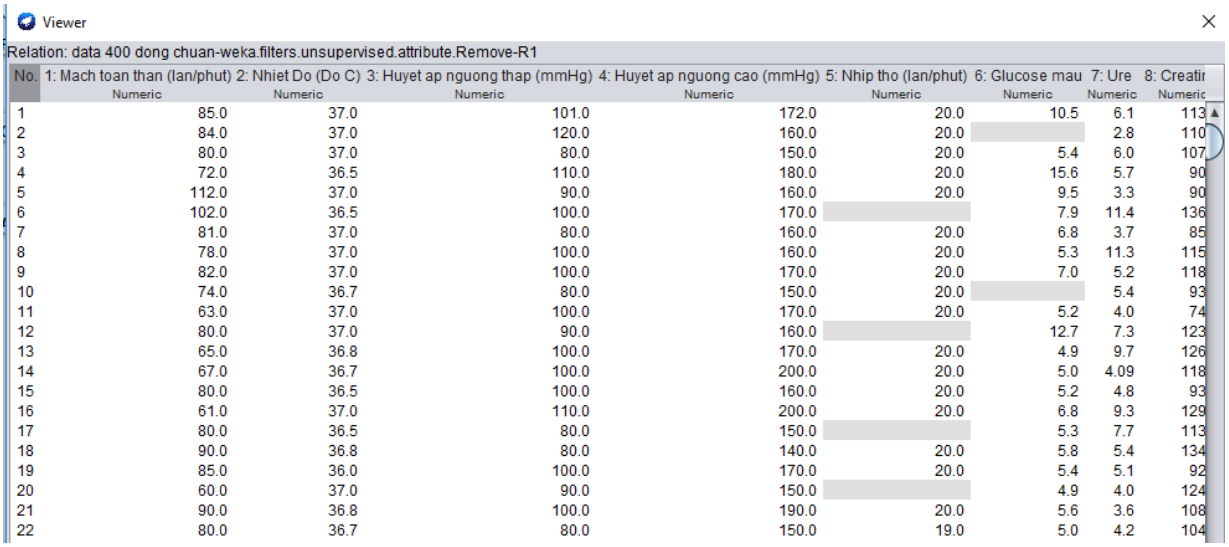
- Xử lý trên excel:

+ Xử lý nhất quán dữ liệu.

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

- Đọc dữ liệu vào weka:

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

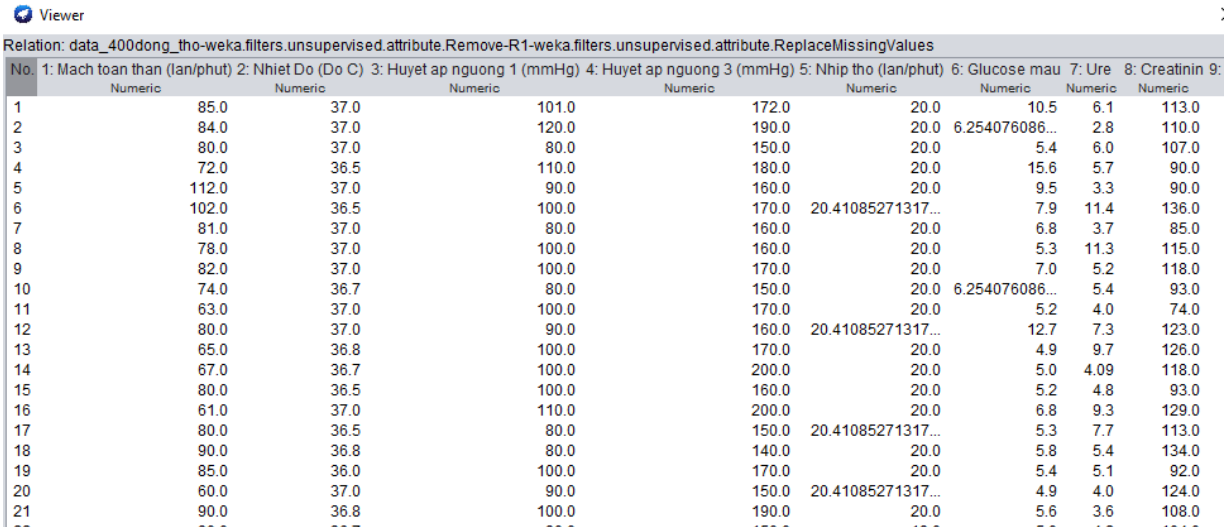


Viewer

Relation: data 400 dong chuan-weka.filters.unsupervised.attribute.Remove-R1

No.	1: Mach toan than (lan/phut)	2: Nhiet Do (Do C)	3: Huyet ap nguong thap (mmHg)	4: Huyet ap nguong cao (mmHg)	5: Nhip tho (lan/phut)	6: Glucose mau	7: Ure	8: Creatir
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	85.0	37.0	101.0	172.0	20.0	10.5	6.1	113
2	84.0	37.0	120.0	160.0	20.0		2.8	110
3	80.0	37.0	80.0	150.0	20.0	5.4	6.0	107
4	72.0	36.5	110.0	180.0	20.0	15.6	5.7	90
5	112.0	37.0	90.0	160.0	20.0	9.5	3.3	90
6	102.0	36.5	100.0	170.0		7.9	11.4	136
7	81.0	37.0	80.0	160.0	20.0	6.8	3.7	85
8	78.0	37.0	100.0	160.0	20.0	5.3	11.3	115
9	82.0	37.0	100.0	170.0	20.0	7.0	5.2	118
10	74.0	36.7	80.0	150.0	20.0		5.4	93
11	63.0	37.0	100.0	170.0	20.0	5.2	4.0	74
12	80.0	37.0	90.0	160.0		12.7	7.3	123
13	65.0	36.8	100.0	170.0	20.0	4.9	9.7	126
14	67.0	36.7	100.0	200.0	20.0	5.0	4.09	118
15	80.0	36.5	100.0	160.0	20.0	5.2	4.8	93
16	61.0	37.0	110.0	200.0	20.0	6.8	9.3	129
17	80.0	36.5	80.0	150.0		5.3	7.7	113
18	90.0	36.8	80.0	140.0	20.0	5.8	5.4	134
19	85.0	36.0	100.0	170.0	20.0	5.4	5.1	92
20	60.0	37.0	90.0	150.0		4.9	4.0	124
21	90.0	36.8	100.0	190.0	20.0	5.6	3.6	108
22	80.0	36.7	80.0	150.0	19.0	5.0	4.2	104

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



Viewer

Relation: data_400dong_tho-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: Mach toan than (lan/phut) Numeric	2: Nhiet Do (Do C) Numeric	3: Huyet ap nguong 1 (mmHg) Numeric	4: Huyet ap nguong 3 (mmHg) Numeric	5: Nhip tho (lan/phut) Numeric	6: Glucose mau Numeric	7: Ure Numeric	8: Creatinin Numeric	9:
1	85.0	37.0	101.0	172.0	20.0	10.5	6.1	113.0	
2	84.0	37.0	120.0	190.0	20.0	6.254076086...	2.8	110.0	
3	80.0	37.0	80.0	150.0	20.0	5.4	6.0	107.0	
4	72.0	36.5	110.0	180.0	20.0	15.6	5.7	90.0	
5	112.0	37.0	90.0	160.0	20.0	9.5	3.3	90.0	
6	102.0	36.5	100.0	170.0	20.41085271317...	7.9	11.4	136.0	
7	81.0	37.0	80.0	160.0	20.0	6.8	3.7	85.0	
8	78.0	37.0	100.0	160.0	20.0	5.3	11.3	115.0	
9	82.0	37.0	100.0	170.0	20.0	7.0	5.2	118.0	
10	74.0	36.7	80.0	150.0	20.0	6.254076086...	5.4	93.0	
11	63.0	37.0	100.0	170.0	20.0	5.2	4.0	74.0	
12	80.0	37.0	90.0	160.0	20.41085271317...	12.7	7.3	123.0	
13	65.0	36.8	100.0	170.0	20.0	4.9	9.7	126.0	
14	67.0	36.7	100.0	200.0	20.0	5.0	4.09	118.0	
15	80.0	36.5	100.0	160.0	20.0	5.2	4.8	93.0	
16	61.0	37.0	110.0	200.0	20.0	6.8	9.3	129.0	
17	80.0	36.5	80.0	150.0	20.41085271317...	5.3	7.7	113.0	
18	90.0	36.8	80.0	140.0	20.0	5.8	5.4	134.0	
19	85.0	36.0	100.0	170.0	20.0	5.4	5.1	92.0	
20	60.0	37.0	90.0	150.0	20.41085271317...	4.9	4.0	124.0	
21	90.0	36.8	100.0	190.0	20.0	5.6	3.6	108.0	

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:*

$$v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$$

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

v’ = [0, 1] là giá trị mới.

Ví dụ:

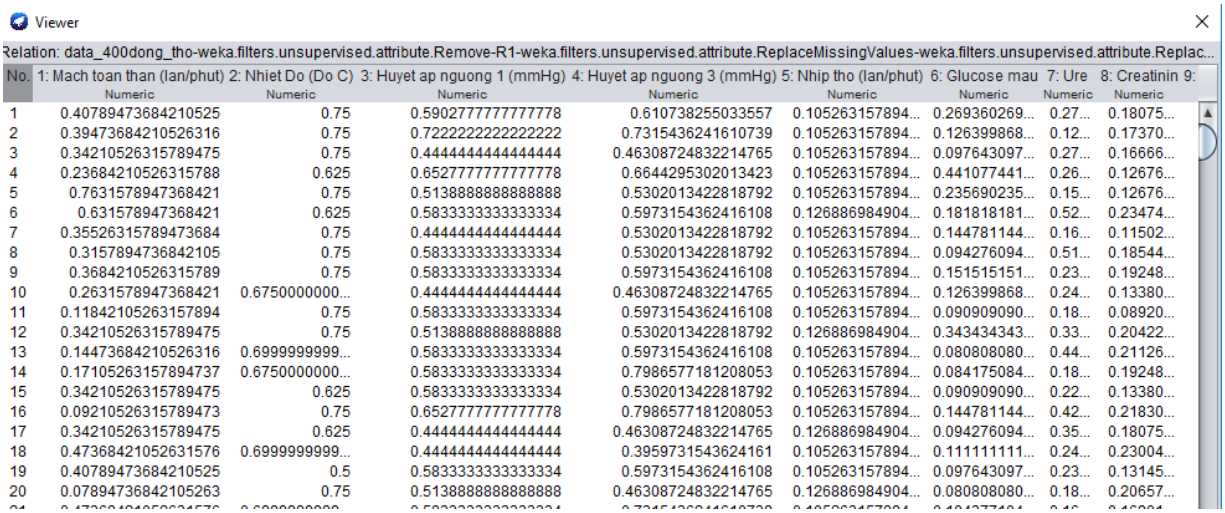
Thuộc tính Mạch toàn thân: Có v = [54, 130].

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.



Viewer

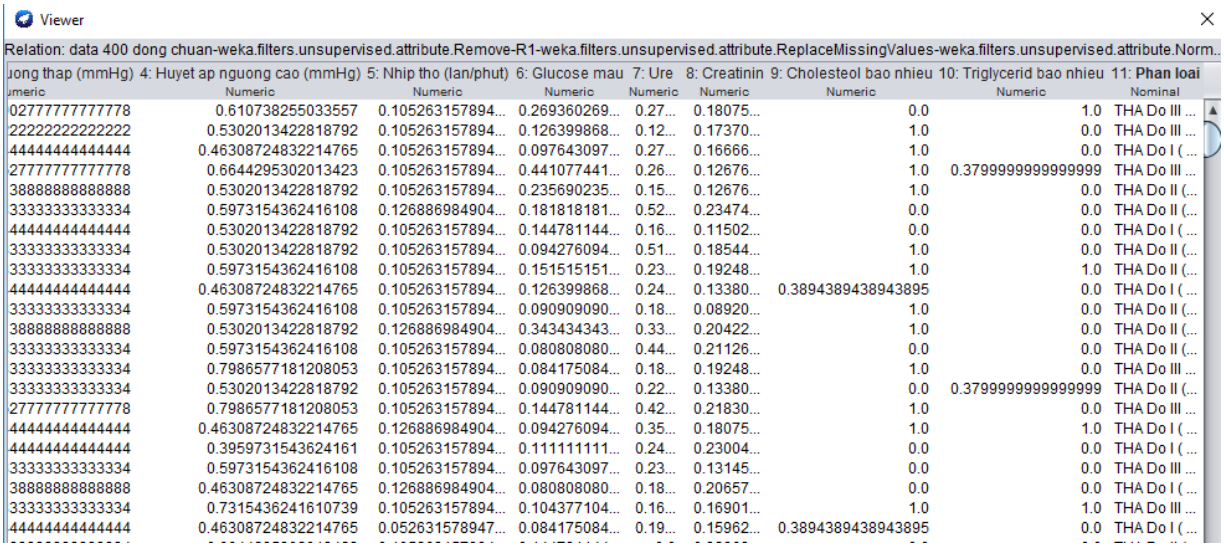
Relation: data_400dong_tho-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Replac...

No.	1: Mach toan than (lan/phut) Numeric	2: Nhiet Do (Do C) Numeric	3: Huyet ap nguong 1 (mmHg) Numeric	4: Huyet ap nguong 3 (mmHg) Numeric	5: Nhip tho (lan/phut) Numeric	6: Glucose mau Numeric	7: Ure Numeric	8: Creatinin Numeric	9:
1	0.40789473684210525	0.75	0.5902777777777778	0.610738255033557	0.105263157894...	0.269360269...	0.27...	0.18075...	
2	0.39473684210526316	0.75	0.7222222222222222	0.7315436241610739	0.105263157894...	0.126399868...	0.12...	0.17370...	
3	0.34210526315789475	0.75	0.4444444444444444	0.46308724832214765	0.105263157894...	0.097643097...	0.27...	0.16666...	
4	0.23684210526315788	0.625	0.6527777777777778	0.6644295302013423	0.105263157894...	0.441077441...	0.26...	0.12676...	
5	0.7631578947368421	0.75	0.5138888888888888	0.5302013422818792	0.105263157894...	0.235690235...	0.15...	0.12676...	
6	0.631578947368421	0.625	0.5833333333333334	0.5973154362416108	0.126886984904...	0.181818181...	0.52...	0.23474...	
7	0.35526315789473684	0.75	0.4444444444444444	0.5302013422818792	0.105263157894...	0.144781144...	0.16...	0.11502...	
8	0.3157894736842105	0.75	0.5833333333333334	0.5302013422818792	0.105263157894...	0.094276094...	0.51...	0.18544...	
9	0.3684210526315789	0.75	0.5833333333333334	0.5973154362416108	0.105263157894...	0.151515151...	0.23...	0.19248...	
10	0.2631578947368421	0.6750000000...	0.4444444444444444	0.46308724832214765	0.105263157894...	0.126399868...	0.24...	0.13380...	
11	0.11842105263157894	0.75	0.5833333333333334	0.5973154362416108	0.105263157894...	0.090909090...	0.18...	0.08920...	
12	0.34210526315789475	0.75	0.5138888888888888	0.5302013422818792	0.126886984904...	0.343434343...	0.33...	0.20422...	
13	0.14473684210526316	0.6999999999...	0.5833333333333334	0.5973154362416108	0.105263157894...	0.080808080...	0.44...	0.21126...	
14	0.17105263157894737	0.6750000000...	0.5833333333333334	0.7986577181208053	0.105263157894...	0.084175084...	0.18...	0.19248...	
15	0.34210526315789475	0.625	0.5833333333333334	0.5302013422818792	0.105263157894...	0.090909090...	0.22...	0.13380...	
16	0.09210526315789473	0.75	0.6527777777777778	0.7986577181208053	0.105263157894...	0.144781144...	0.42...	0.21830...	
17	0.34210526315789475	0.625	0.4444444444444444	0.46308724832214765	0.126886984904...	0.094276094...	0.35...	0.18075...	
18	0.47368421052631576	0.6999999999...	0.4444444444444444	0.3959731543624161	0.105263157894...	0.111111111...	0.24...	0.23004...	
19	0.40789473684210525	0.5	0.5833333333333334	0.5973154362416108	0.105263157894...	0.097643097...	0.23...	0.13145...	
20	0.07894736842105263	0.75	0.5138888888888888	0.46308724832214765	0.126886984904...	0.080808080...	0.18...	0.20657...	

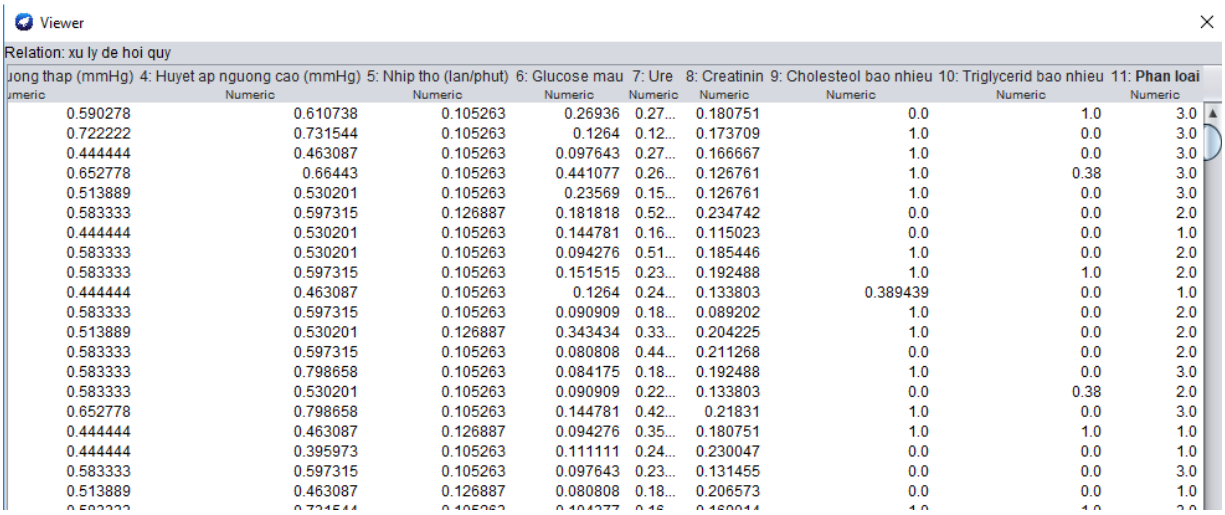
- *Chuẩn hóa với mô hình hồi quy tuyến tính.*

- Xử lý trên excel:

+ Nhìn vào tập dữ liệu ta thấy có 2 thuộc tính rời rạc (nominal): 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 nominal rồi nên vẫn giữ nguyên.

CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG MÔ HÌNH 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.

$$\mathbf{Y = f\ (X,\ \beta)}$$

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 Mô hình 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.

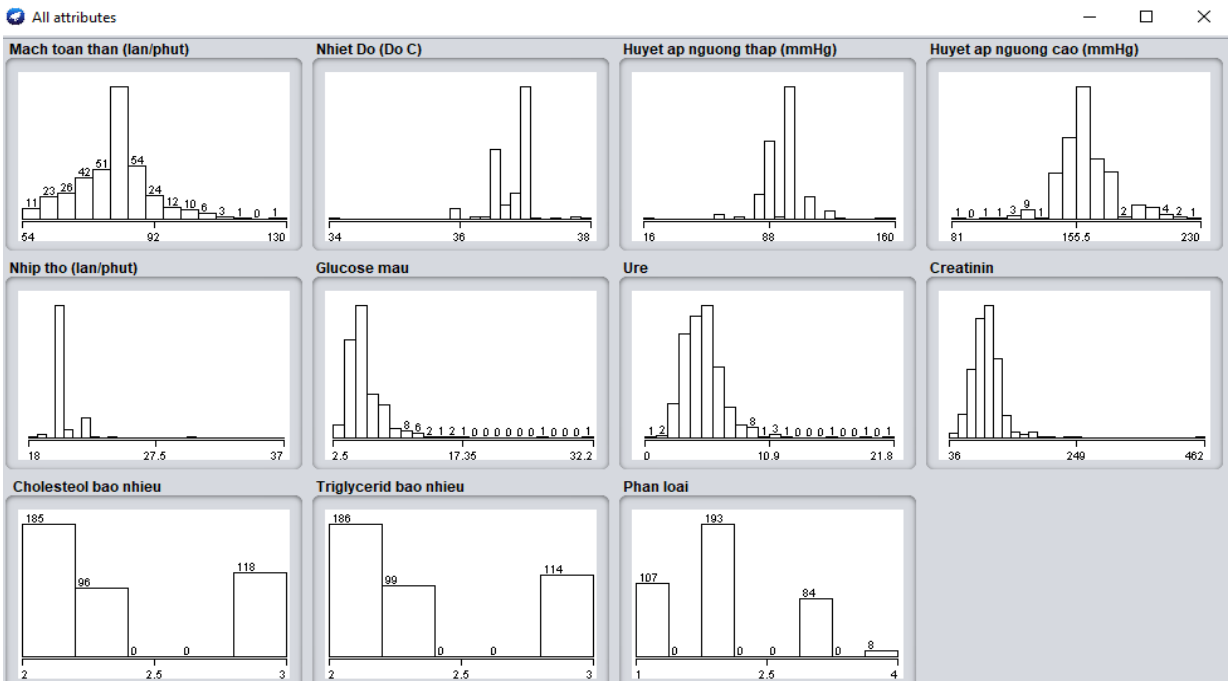
+ X: Biến độc lập

+ B0: Giá trị của Y khi X = 0.

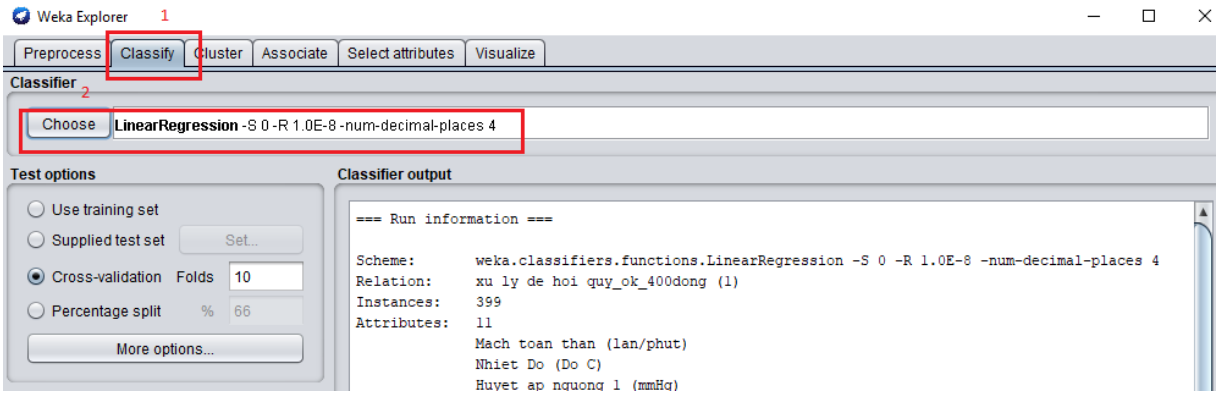
+ B1 …k: Trị số các hệ số hồi quy.

- **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.



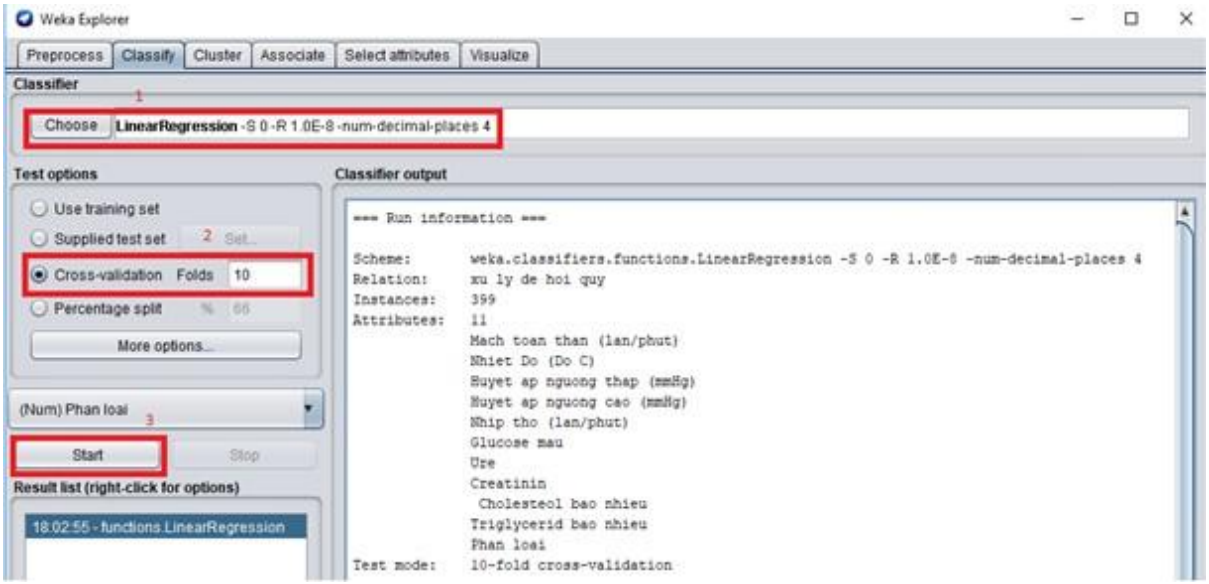
Bước 3: Tiến hành đánh giá hiệu quả của mô hình hồi quy đối với tập dữ liệu đã sử dụng theo hai phương pháp:

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.

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:

Classifier model:

- 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 +

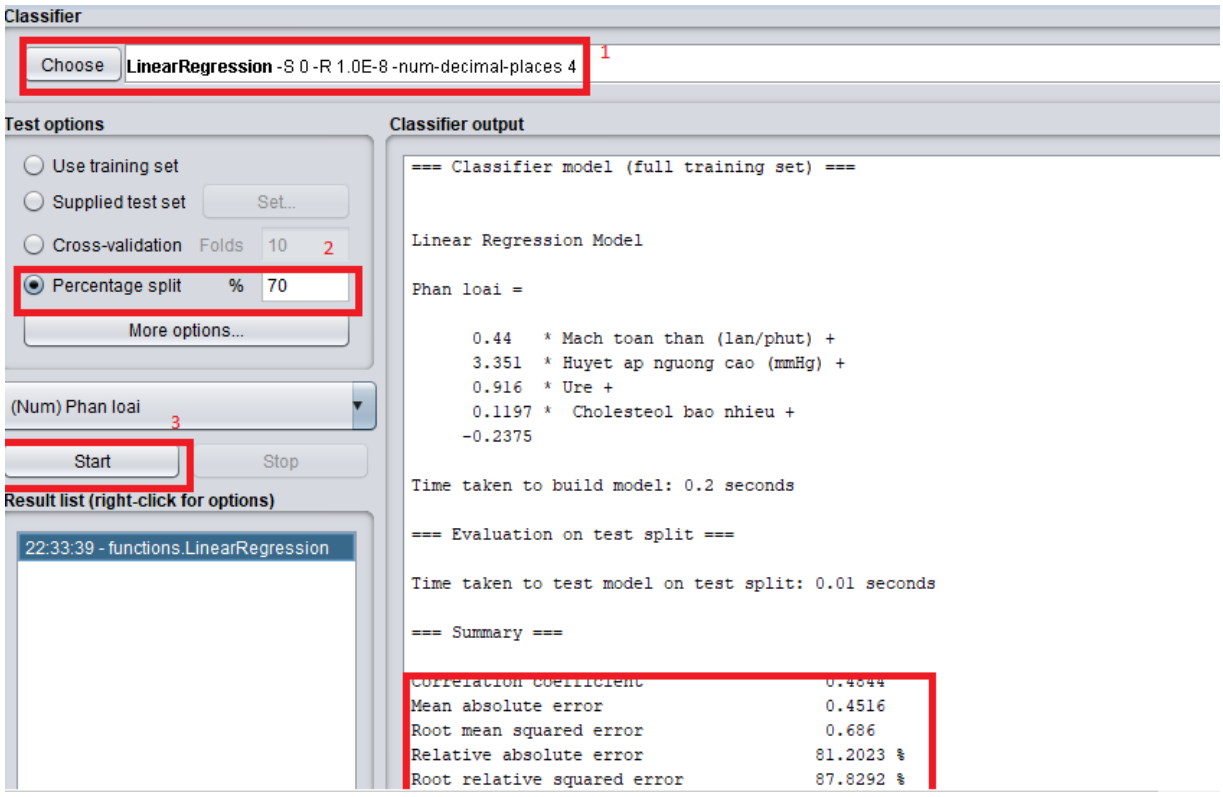
0.1197 * Cholesteol bao nhieu +

-2.6113

Cross-validation:

+ Hệ số tương quan: 0.5187.
+ Sai số tuyệt đối trung bình: 0.4498.
+ Sai số bình phương trung bình: 0.6454.
+ Sai số tuyệt đối tương đối: 83.1135%.
+ Sai số bình phương tương đối: 84.9725%
+ Tổng số bản ghi: 392.
+ Bỏ qua trường hợp không xác định: 7.

(2) Percentage split: Train/Test: 70/30%.



Thống kê lỗi:

+ Hệ số tương quan: 0.4844.
+ Sai số tuyệt đối trung bình: 0.4516

+ Sai số bình phương trung bình: 0.686.

+ Sai số tuyệt đối tương đối: 81.2023 %

+ Sai số bình phương tương đối: 87.8292 %

- **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 tăng huyết áp của bệnh nhân bằng mô hình hồi quy tuyến tính thu được sau khi chạy mô hình.

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

Trong lĩnh vực máy học, **cây quyết định** là một kiểu mô hình dự báo (*predictive model*), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong (*internal node*) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó

Cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó. Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa

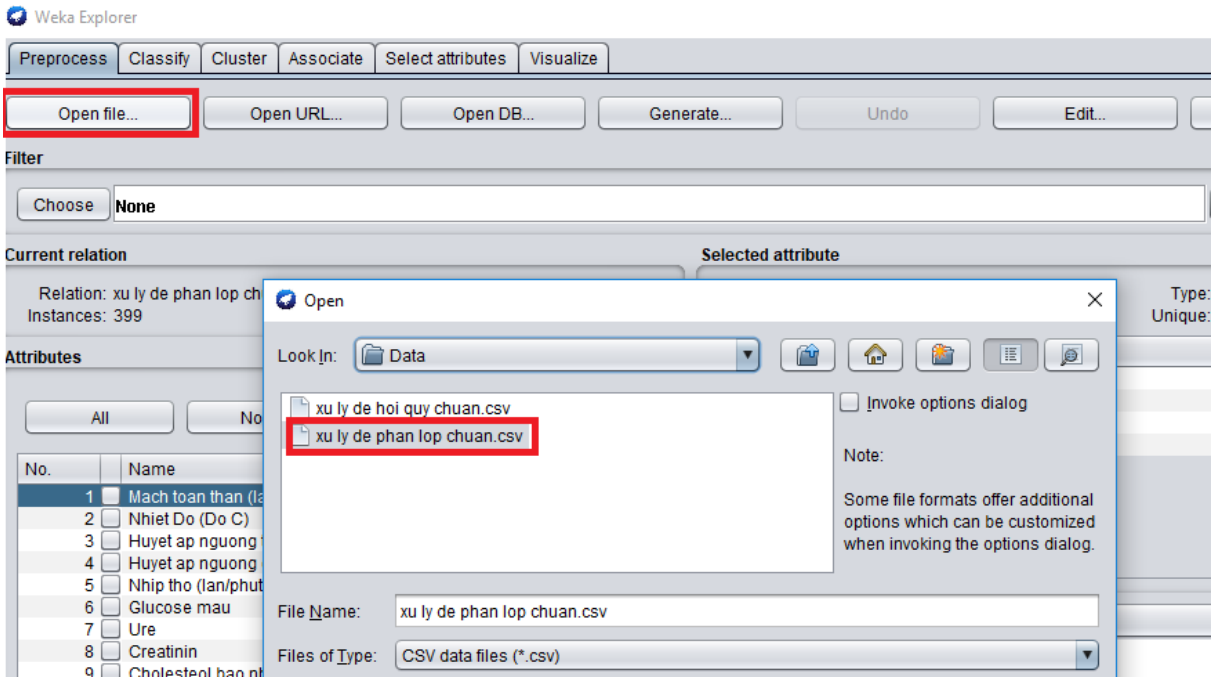
theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết [6]

Quy trình train và test của một classifer:

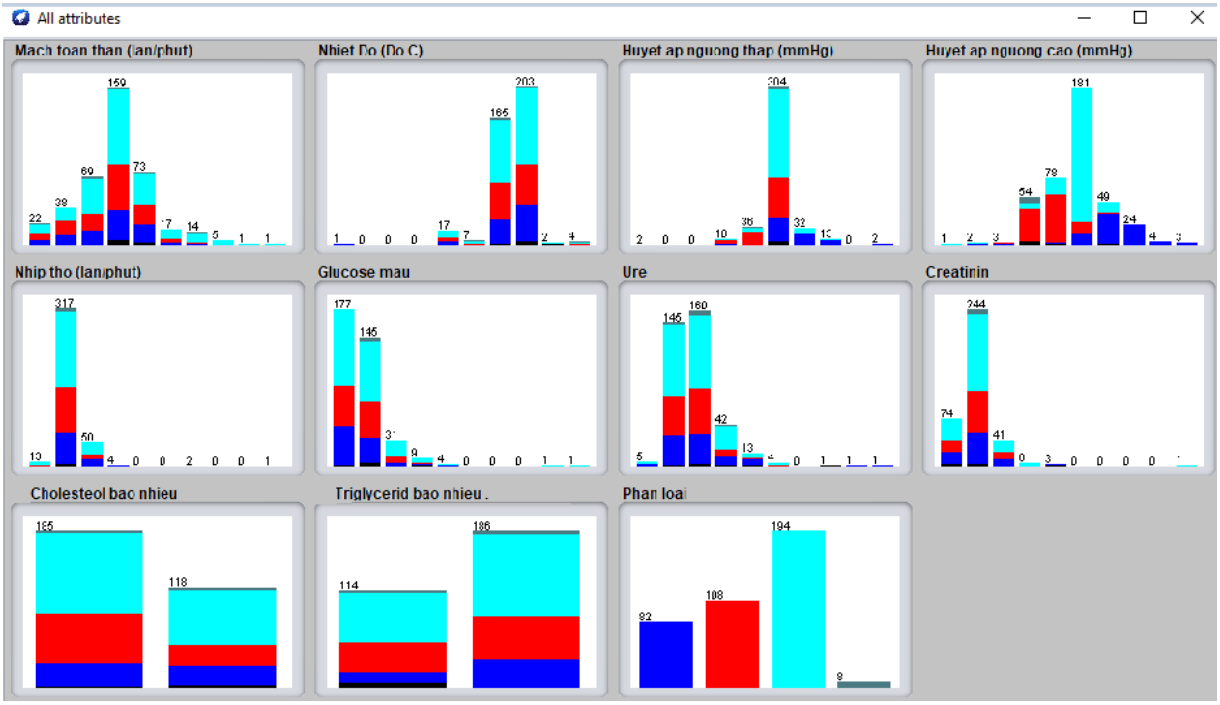
- Dữ liệu để xây dựng mô hình: dữ liệu gốc dữ liệu này phải có thuộc tính phân lớp gọi là Categorical attribute
- Dữ liệu gốc sẽ được chia thành 2 phần là Train set và Testing set
- Cuối cùng là tính toán lỗi để đánh giá model.

Các bước:

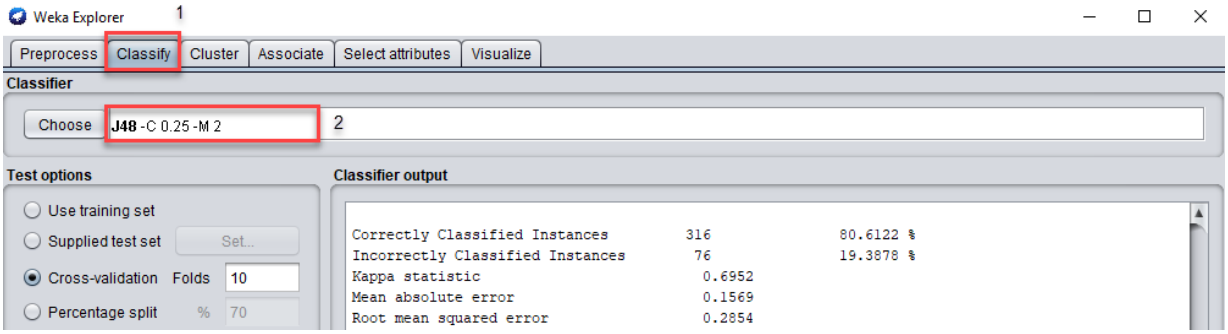
Bước 1: Mở file .csv đã tiền xử lý bằng Weka trước đó trên Weka



Dataset như sau:



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

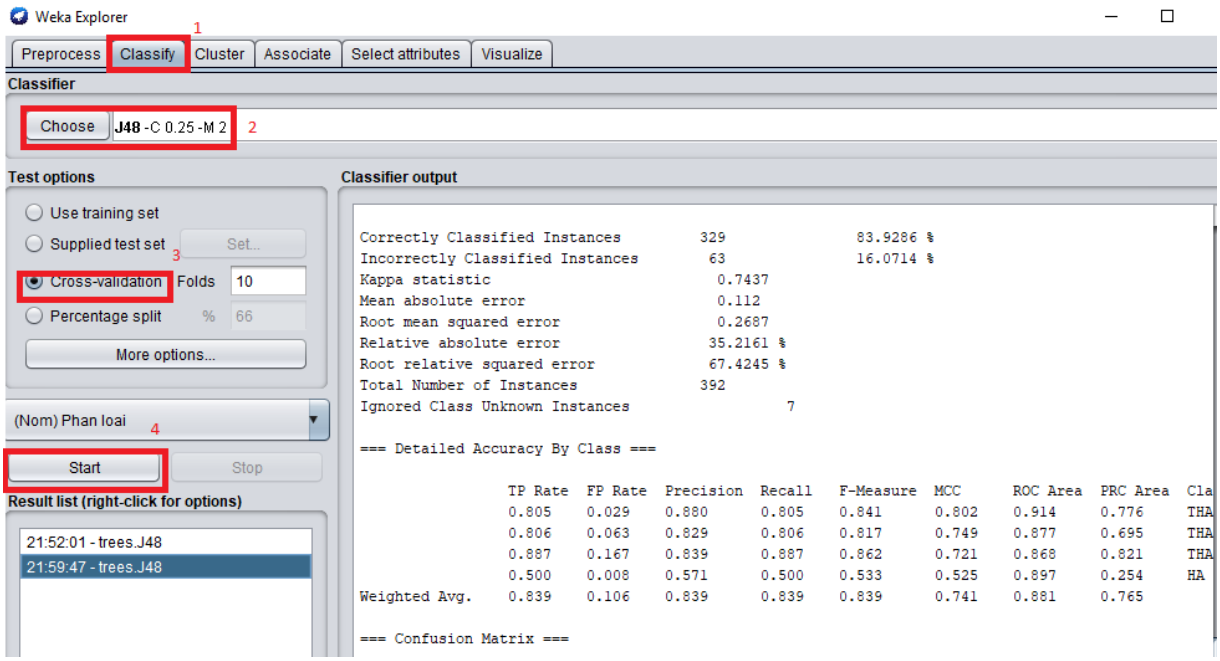


Bước 3: Tiến hành đánh giá hiệu quả phân lớp của thuật toán đối với tập dữ liệu được dùng theo hai phương pháp:

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

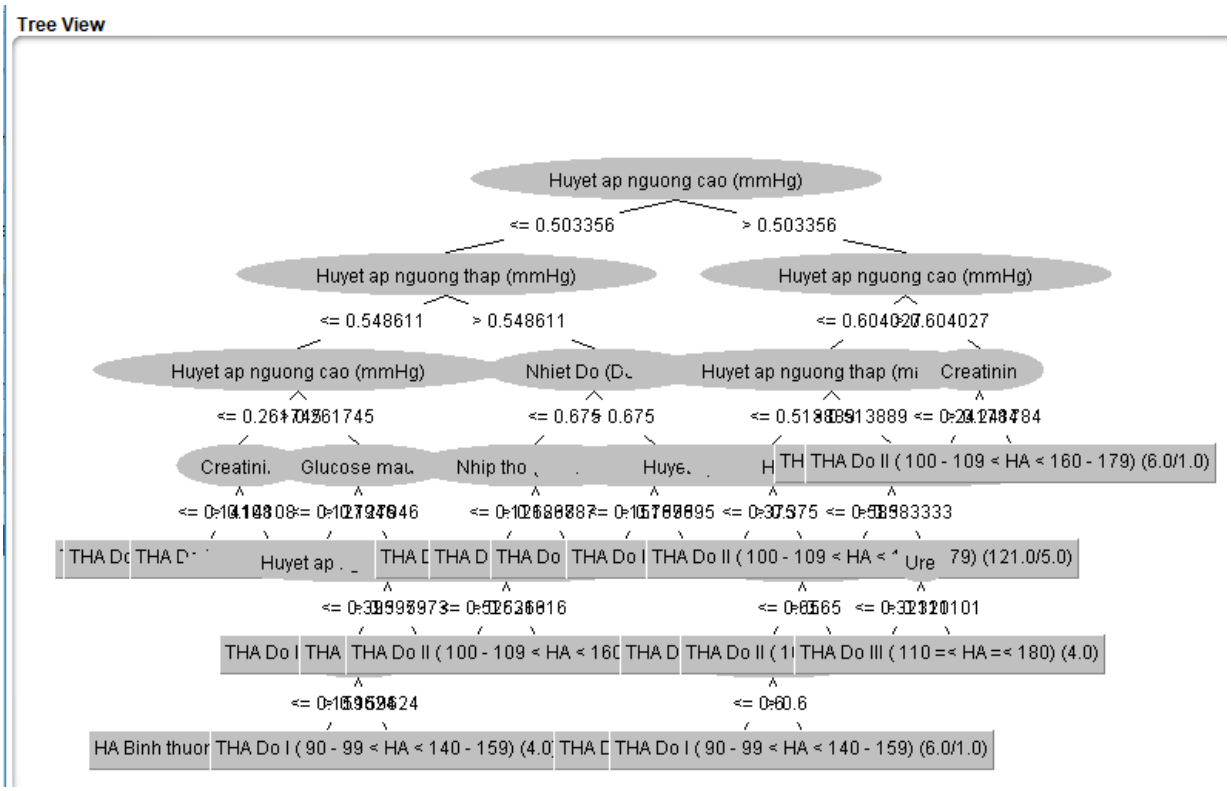
    Triglycerid bao nhieu :

    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



- Đánh giá độ chính xác:

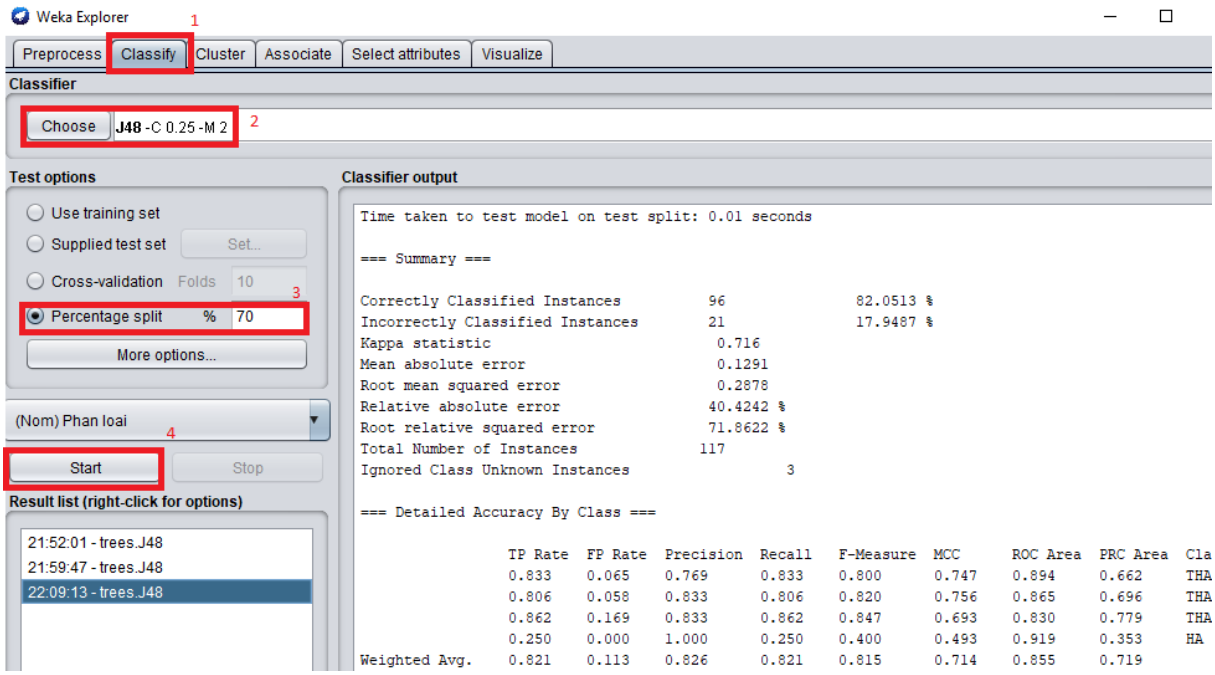
- **Tỷ lệ TP** : Tỷ lệ dương tính thực (các trường hợp được phân loại chính xác như một lớp đã cho)
- **Tỷ lệ FP** : Tỷ lệ dương tính giả (các trường hợp được phân loại sai là một lớp nhất định)
- **Độ chính xác** : Tỷ lệ các trường hợp thực sự là một lớp chia cho tổng số các trường hợp được phân loại 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.

Tri thức thu được:

+ Nếu Huyết áp ngưỡng cao <= 0.261745 và huyết áp ngưỡng thấp < = 0.5486 và creatinin <= 0.14108 thì THA độ III (110 =< HA =< 180).

+ Nếu Huyết áp ngưỡng cao <= 0.261745 và huyết áp ngưỡng thấp < = 0.5486 và creatinin > 0.14108 thì THA độ II (100 - 109 < HA < 160 – 179).

+ Nếu 0.261745 < Huyết áp ngưỡng cao <= 0.503356 và huyết áp ngưỡng thấp < = 0.5486 và Glucose mau <= 0.1279 thì THA độ I (90 - 99 < HA < 140 – 159).

+ Nếu 0.261745 < Huyết áp ngưỡng cao <= 0.395973 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).

+ Nếu 0. 261745 < Huyết áp ngưỡng cao <= 0. 395973 và huyết áp ngưỡng thấp < = 0.5486 và Glucose mau > 0.127946 và creatinin > 0.159624 thì THA độ I.

+ Nếu 0. 395973 < Huyết áp ngưỡng cao > 0.503356 thì THA độ I (90 - 99 < HA < 140 – 159).

+ Nếu 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.

+ Nếu 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.

+ Nếu 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.

+ Nếu 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.

+ Nếu 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.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.604027và Huyết áp ngưỡng cao (mmHg) <= 0.375 thì THA Do I.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.604027 và 0.375 < huyết áp ngưỡng thấp <= 0.513889 và nhiệt độ <= 0.6 thì THA độ II.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.50336 0.604027 và 0.375 < huyết áp ngưỡng thấp <= 0.513889 và 0.6 < Nhiệt độ <= 0.65 thì THA độ I.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.604027 và huyết áp ngưỡng thấp <= 0.375 và nhiệt độ > 0.65 thì THA độ II.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.604027 và huyết áp ngưỡng thấp <= 0.583333 thì THA độ II.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.604027 và huyết áp ngưỡng thấp > 0.583333 và ure <= 0.321101 thì THA độ II.

+ Nếu 0.503356 < Huyết áp ngưỡng cao <= 0.604027 và huyết áp ngưỡng thấp > 0.583333 và ure > 0.321101 thì THA độ III.

+ Nếu Huyết áp ngưỡng cao > 0.604027 và creatinin <= 0.241784 thì THA độ III.

+ Nếu Huyết áp ngưỡng cao > 0.604027 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

=> 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.

4.4 Đánh giá mô hình.

Từ 2 giải thuật hồi quy và phân lớp đã đưa ra mô hình và các luật để chuẩn đoán bệnh THA của bệnh nhân tại Bệnh Viện Đa Khoa Thái Bình. Sai số của mô hình là tương đối tốt nên có thể cho kết quả chuẩn đoán tương đối chính xác, có tính ứng dụng thực tế.

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 xu hướng trong tương lai, nó được ứng dụng trong nhiều ngành như y tế, thương mại... Hoàn thành 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 excel 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 thuật toán 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.

3.2 Hướng phát triển.

- ✔ Xây dựng, cải tiến mô hình chẩn đoán bệnh với phương pháp học máy khác như SVM, KNN, Random Forest,…
- ✔ Áp dụng mô hình học máy vào một số lĩnh vực khác và đi vào áp dụng thực.

Trong quá trình hoàn thành bài tập lớn, nhóm 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 thể 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. & https://vi.wikipedia.org/wiki/Cây_quyết_định

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.