ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

------✶------

KHÓA LUẬN TỐT NGHIỆP

Tên đề tài:

ỨNG DỤNG DATA MINING XÂY DỰNG HỆ HỖ TRỢ RA QUYẾT ĐỊNH TRONG KHÁM CHỮA BỆNH TIỂU ĐƯỜNG

Giảng viên hướng dẫn: TS Nguyễn Đình Thuân

Lớp: HTTT03

Khóa: 2008 - 2013

Sinh viên thực hiện:

Ung Quốc Bình 08520029

Nguyễn Văn Lâm 08520193

Tp.HCM, ngày 19 tháng 12 năm 2012

LỜI CẢM ƠN

Đầu tiên, nhóm thực hiện xin gửi lời cám ơn sâu sắc tới tập thể thầy cô Trường Đại học Công Nghệ Thông Tin – Đại học Quốc Gia TP. HCM và quý thầy cô khoa Hệ Thống Thông Tin đã truyền đạt những kiến thức nền tảng trong suốt quá trình đào tạo vừa qua để nhóm có thể thực hiện đề tài này. Và nhóm cũng xin chân thành cảm ơn tập thể các y bác sĩ tại bệnh viện Quận Thủ Đức và bệnh viện Đa Khoa Khu Vực Thủ Đức đã nhiệt tình chỉ dẫn những kiến thức ngoài chuyên ngành nhưng lại vô cùng quan trọng đối với đề tài của nhóm thực hiện.

Đặc biệt, nhóm thực hiện xin gởi lời cảm ơn và lòng biết ơn sâu sắc nhất đến Tiến Sĩ Nguyễn Đình Thuân, người đã tận tình hướng dẫn và tạo điều kiện tốt nhất cho nhóm trong suốt quá trình thực hiện khóa luận tốt nghiệp vừa qua.

Trong khoảng thời gian hơn năm tháng thực hiện đề tài, nhóm đã có dịp vận dụng tất cả những kiến thức đã tích lũy đồng thời kết hợp với những kiến thức thu được trong suốt thời gian thực hiện đề tài để có thể hoàn thành tốt một báo cáo khóa luận. Tuy nhiên, trong quá trình thực hiện vẫn có thể xảy ra những thiếu sót không thể tránh khỏi. Vì thế, nhóm thực hiện ra mong nhận được sự đóng góp từ phía thầy cô nhằm hoàn thiện những kiến thức mà nhóm đã tích lũy đồng thời cũng sẽ là hành trang của nhóm để thực hiện tiếp những đề tài nghiên cứu khác trong thời gian tương lại.

Xin chân thành cảm ơn quý thầy cô

Nhóm sinh viên thực hiện

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

MỤC LỤC

LỜI CẢM ƠN ..... **ii**

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN ..... **iii**

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN ..... **iv**

DANH MỤC HÌNH ..... **vii**

DANH MỤC BẢNG ..... **viii**

DANH MỤC CÁC THUẬT NGỮ, TỪ VIẾT TẮT ..... **x**

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI **1**

1.1. Đặt vấn đề 1

1.2. Mục tiêu 5

1.3. Phạm vi khóa luận 5

1.4. Phương pháp nghiên cứu và nội dung thực hiện 6

1.5. Kết quả dự kiến 6

1.6. Bố cục báo cáo 7

CHƯƠNG 2: DATA MINING VÀ NHỮNG THÀNH TỰU TRONG Y HỌC **9**

2.1. Data mining là gì 9

2.2. Những nghiên cứu đầu tiên 9

2.3. Những ứng dụng của Data mining trong y học 11

2.4. Những khó khăn trong việc ứng dụng Data mining vào y học 14

CHƯƠNG 3: HỆ HỖ TRỢ RA QUYẾT ĐỊNH LÂM SÀNG **15**

3.1. Hệ hỗ trợ ra quyết định 15

3.1.1. Hệ hỗ trợ ra quyết định là gì? 15

3.1.2. Kiến trúc chung của hệ hỗ trợ ra quyết định 17

3.2. Hệ hỗ trợ ra quyết định lâm sàng 18

3.2.1. Khái niệm 18

3.2.2. Các dạng của hệ hỗ trợ ra quyết định lâm sàng 18

3.2.3. Tính năng tiêu biểu 22
3.2.4. Những thử thách trong quá trình triển khai 22
3.2.5. Tương lai của hệ hỗ trợ ra quyết định lâm sàn 23
3.3. Phương pháp tiếp cận 24
3.3.1. Học có giám sát 25
3.3.2. Học không giám sát 26
CHƯƠNG 4: TRIỂN KHAI 27
4.1. Sơ lược về thành phần hệ thống 27
4.1.1. Dữ liệu đầu vào 27
4.1.2. Kết quả đầu ra 49
4.2. Phương pháp thực hiện 49
4.2.1. Tiền xử lý dữ liệu 49
4.2.2. Cài đặt giải thuật 51
4.2.3. Phương pháp đánh giá 66
CHƯƠNG 5: KẾT QUẢ THỰC NGHIỆM 68
5.1. Ứng dụng thực tế 68
5.1.1. Tiền xử lý dữ liệu 68
5.1.2. Xây dựng mô hình 71
5.1.3. Chẩn đoán 73
5.2. Đánh giá kết quả 75
5.2.1. Đánh giá dữ liệu 75
5.2.2. Đánh giá giải thuật 77
5.3. Kết luận và hướng phát triển 83
5.3.1. Kết luận 83
5.3.2. Hướng phát triển 85
TÀI LIỆU THAM KHẢO 86

DANH MỤC HÌNH

Hình 2.1. Bản đồ đồ thị 10
Hình 2.2. Biểu đồ phân cực 11
Hình 3.1. Mô hình kiến trúc của Hệ hỗ trợ ra quyết định 17
Hình 3.2. Mô hình học có giám sát 25
Hình 3.3. Mô hình học không giám sát 26
Hình 4.1. Mô hình cửa sổ của phương pháp làm sạch dữ liệu bán tự động 50
Hình 4.2. Mô hình đơn giản của giải thuật cây quyết định 54
Hình 4.3. Mã giả của giải thuật C4.5 56
Hình 4.4. Ví dụ minh họa giải thuật Naïve Bayes 58
Hình 4.5. Mặt siêu phẳng tách các mẫu dương khỏi các mẫu âm 63
Hình 5.1. Màn hình Tiền xử lý dữ liệu – Làm sạch dữ liệu 69
Hình 5.2. Màn hình Tiền xử lý dữ liệu – Rời rạc hóa dữ liệu – Binning 70
Hình 5.3. Màn hình Tiền xử lý dữ liệu – Rời rạc hóa dữ liệu – Tùy chỉnh 71
Hình 5.4. Màn hình Xây dựng mô hình 72
Hình 5.5. Màn hình Xem mô hình 73
Hình 5.6. Màn hình Chẩn đoán 74

DANH MỤC BẢNG

Bảng 3.1. So sách các hệ thống xử lý dữ liệu 15
Bảng 3.2. Một số hệ hỗ trợ ra quyết định lâm sàng đã được sử dụng 24
Bảng 4.1. Dữ liệu của hệ thống 29
Bảng 4.2. Dữ liệu thông tin cá nhân của bệnh nhân 29
Bảng 4.3. Dữ liệu xét nghiệm máu mỡ 30
Bảng 4.4. Cholesterol 31
Bảng 4.5. High Density Lipoprotein 32
Bảng 4.6. Low Density Lipoprotein 33
Bảng 4.7. Trigleceride 34
Bảng 4.8. Dữ liệu xét nghiệm sinh hóa 34
Bảng 4.9. Dữ liệu xét nghiệm men gan 36
Bảng 4.10 Dữ liệu huyết đồ 39
Bảng 4.11 Mean Corpuscular Volume 43
Bảng 4.12 Mean Corpuscular Hemoglobin Concentration 44
Bảng 4.13 Red Distribution Witdh 44
Bảng 4.14 Platelet Count 45
Bảng 4.15 Dữ liệu xét nghiệm điện giải 47
Bảng 4.16. Dữ liệu phân lớp 49
Bảng 4.17. Các chỉ số liên qua đến Precision và Recall 66
Bảng 5.1. Thống kê dữ liệu đã thu thập 75
Bảng 5.2. Thống kê hiện trạng bệnh tiểu đường bằng dữ liệu thu thập 76
Bảng 5.3. Kết quả đánh giá giải thuật Naïve Bayes tự cài đặt 78
Bảng 5.4. Kết quả đánh giá giải thuật Naïve Bayes áp dụng Framework 79
Bảng 5.4. Kết quả đánh giá giải thuật C4.5 80

Bảng 5.5. Các tập luật tiểu biểu của phân lớp “True” 81

Bảng 5.6. Các tập luật tiêu biểu của phân lớp “False” 81

Bảng 5.7. Kết quả đánh giá giải thuật SVM 82

DANH MỤC CÁC THUẬT NGỮ, TỪ VIẾT TẮT

STT	Tiếng Việt	Viết tắt TV	Tiếng Anh	Viết tắt TA
1	Khai phá dữ liệu		Data Mining	DM
2	Hệ hỗ trợ ra quyết định	HHTRQĐ	Decision Support Systems	DSS
3	Hệ hỗ trợ ra quyết định lâm sàng	HHTRQĐLS	Clinical Decision Support Systems	CDSS
4	Cơ sở dữ liệu	CSDL	Database	
5	Hệ cơ sở trị thức		Knowledge Based Systems	KBS
6	Hệ phi cơ sở tri thức		Non-knowledge Based Systems	NBS
7	Mạng Nơron		Artificial Neurol Networks	ANN
8	Giải thuật di truyền		Genetic Algorithms	GA
9	Máy hỗ trợ vector		Support Vector Machine	SVM

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

Trong chương mở đầu này, nhóm thực hiện sẽ giới thiệu tổng quát về đề tài. Nhóm thực hiện sẽ trình bày nội dung sau: Lý do chọn đề tài? Nhu cầu thực tế của đề tài? Mục tiêu và phạm vi của đề tài? Bài toán chính của đề tài? Cuối cùng là phương pháp để thực hiện và giải quyết bài toán đó ra sao?

1.1. Đặt vấn đề

Bệnh tiểu đường là một trong những căn bệnh phổ biến nhất của thế kỉ 21, là một trong những nguyên nhân chính dẫn đến các bệnh hiểm nghèo như bệnh tim, tai biến, suy thận, mù mắt, hoại thư…Bệnh tiểu đường thường gây nguy hiểm nhiều nhất cho người già và những người béo phì.

Bệnh tiểu đường (còn được gọi là bệnh đái tháo đường) là một nhóm bệnh rối loạn chuyển hóa carbodydrates khi hóc môn insulin của tuyến tụy bị thiếu hoặc giảm tác động trong cơ thể[1]. Sự sản sinh sẽ được điều chỉnh bởi lượng glucose trong máu. Nó có trong chức năng chuyển hóa glucose (từ cardbohydrates) trong tế bào để cung cấp cho quá trình trao đổi chất và tạo thành năng lượng cho cơ thể. Sự thiếu hụt insulin hoặc không sử dụng được insulin sẽ làm giảm khả năng hấp thụ glucose và vì thế glucose sẽ tích tụ trong gan và các tế báo chất béo dẫn đến việc tăng mức đường huyết và đường trong nước tiểu Đồng thời những nhân tố như gen di truyền, chế độ dinh dưỡng không tốt, bị stress, ít vận động và thừa cân là những yếu tố quan trọng có thể dẫn đến việc mắc bệnh tiểu đường.

Bệnh nhân tiểu đường với lượng đường huyết cao trong máu gây tổn thương tế bào vi mạch thận, làm giảm chức năng lọc, bài tiết nước tiểu của thận. Bệnh nặng dẫn đến suy thận và hủy hoại chức năng của thận, dẫn đến việc đi tiểu với lượng đường cao trong nước tiểu thường thấy ở bệnh nhân tiểu đường. Với lượng đường huyết cao trong mạch máu, khiến cho những mạch máu nhỏ tại võng mạc bị nghẽn, có thể bị vỡ

[1] http://vi.wikipedia.org/wiki/Ti%E1%BB%83u_%C4%91%C6%B0%E1%BB%9Dng

gây tấy đỏ, sưng ứ gây ra tổn thương mất và các bệnh võng mạc. Ngoài ra, các biến chứng của bệnh tiểu đường còn gây ra đục thủy tinh thể, tăng nhãn áp, gây mù lòa. Thêm vào đó là các biến chứng nguy hiểm về các mạch máu và tim. Khi các dấu hiệu tổn thương mạch máu, tim ngày càng nặng thì bệnh nhân rất dễ bị cao huyết áp, xơ cứng động mạch, nhồi máu cơ tim, tai biến mạch máu não gây bại liệt hoặc tử vong.

Hiện nay có 2 loại bệnh tiểu đường:

- Bệnh tiểu đường dạng một (hay còn gọi là tiểu đường tuýp một): Ở dạng này, tuyến tụy của bệnh nhân hầu như hoặc không có khả năng sinh ra insulin. Nguyên nhân là do hệ miễn dịch tự hủy hoại các tế bào beta trong tuyến tụy có nhiệm vụ sản sinh ra insulin. Chỉ có khoảng 5-10% tổng số bệnh nhân mắc bệnh tiểu đường dạng một, phần lớn xảy ra ở trẻ em và người trẻ tuổi (dưới 20 tuổi). Các triệu chứng thường khởi phát độ ngột và tiến triển nhanh hơn nếu không điều trị.

- Bệnh tiểu đường dạng hai (hay còn gọi là tiểu đường tuýp hai): Với những người mắc bệnh tiểu đường dạng hai, lượng insulin sản sinh ra ban đầu hoàn toàn bình thường nhưng các tế bào đã không hoặc kém nhạy cảm với sự có mặt insulin. Đó là hiện tượng nhờn insulin (kháng insulin). Lượng đường trong máu không được chuyển hóa thành năng lượng nên giữ ở mức cao, cơ thể bệnh nhân phản ứng bằng cách tăng sản xuất insulin lên dẫn đến việc quá tải cho tuyến tụy và lượng insulin được tiết ra giảm dần. Tỉ lệ người mắc bệnh tiểu đường dạng 2 khoảng từ 90% đến 95%.

    - Bệnh tiểu đường dạng hai có nguyên nhân tiềm ẩn trong cấu tạo gen, nó làm cho bệnh phát triển nhanh. Nếu những người mang trong mình gen tạo mầm mống cho bệnh tiểu đường sớm biết được điều đó và có biện pháp phòng ngừa bằng cách sống và ăn uống tốt thì bệnh sẽ không xuất hiện và phát triển. Bệnh tiểu đường trong trường hợp này sẽ giữ ở dạng

tiềm ẩn. Nhưng ngược lại, với cách sống không khoa học, căn bệnh sẽ phát triển rất nhanh.

- Số bệnh nhân mắc tiểu đường dạng hai chiếm khoảng 90 – 95% trong tổng số bệnh nhân, thường gặp ở lứa tuổi trên 40 nhưng gần đây xuất hiện ngày càng nhiều ở lứa tuổi 30, thậm chí ở cả lứa tuổi thanh thiếu niên. Bệnh nhân thường ít có triệu chứng và chỉ thường được phát hiện khi một trong các biến chức bắt đầu bộc phát hoặc chỉ phát hiện tình cờ khi đi xét nghiệm máu trước khi mổ.

Vào khoảng giữa những năm 80 của thế kỉ trước, tổng số người mắc bệnh tiểu đường trên thế giới vào khoảng 30 triệu. Ngày nay con số này đã lên tới 246 triệu và theo dự đoán tới năm 2025 số người mắc bệnh sẽ lên tới 380 triệu. Mỗi năm, thế giới có khoảng 3,2 triệu người chết vì bệnh tiểu đường, tương dương với số người chết hàng năm vì bệnh HIV/AIDS. Theo thống kê của WHO, cứ mỗi 30 giây lại có một người mắc bệnh tiểu đường bị cắt cụt chi, mỗi ngày có khoảng 5000 người mắt khả năng nhìn do biến chứng về mắt của bệnh tiểu đường. Căn bệnh này làm ảnh hưởng lớn tới nền kinh tế thế giới. Ước tính, mỗi năm trên thế giới người ta bỏ ra khoảng 215 đến 375 triệu đô la để điều trị căn bệnh này[2]. Cũng theo WHO, chi phí để điều trị cho những người mắc bệnh tiểu đường gấp từ 2 – 3 lần người chưa có bệnh. Và chi phí cho việc phòng bệnh tiểu đường lại thấp hơn nhiều lần so với chi phí điều trị cho các biến chứng của bệnh tiểu đường.

Riêng tại Việt Nam, năm 2011, theo công bố của Hiệp Hội Đái tháo đường Thế giới IDF Diabetes Atlas, Việt Nam có 1,7 triệu người mắc bệnh tiểu đường (tương đương với 3,2% dân số trong độ tuổi từ 20 đến 79). Dự đoán đến năm 2030, số người mắc bệnh tiểu đường sẽ tăng lên 3,1 triệu người (tương đương 3,5% tổng dân số trưởng thành).

[2] www.diabetes.org

Việt Nam hiện nay đang là một nước đang phát triển, trong khi các bệnh nhiễm trùng lây lan còn đang phổ biến thì nay các bệnh của một xã hội công nghiệp – bệnh không lây lan lại bùng phát với tốc độ đáng lo ngại. Do những thay đổi đột ngột về kinh tế, xã hội kéo theo những thay đổi về lối sống làm cho tỉ lệ bệnh không lây lan tăng nhanh, trong khi đó chúng ta lại chưa có kinh nghiệm trong lĩnh vực này. Thậm chí hầu hết ở một số địa phương trong nước không có bác sĩ chuyên khoa chuyên ngành nội tiết và rối loạn chuyển hóa. Theo thống kê, báo cáo của Vụ điều trị - Bộ y tế năm 2005, 100% người mắc các bệnh rối loạn chuyển hóa phải chuyển lên tuyến trên. Về mặt dự phòng, nước ta cũng chưa có hệ thống để phát hiện sớm, ngăn ngừa khả năng tiến tới bệnh tiểu đường ở những nhóm người có những yếu tố mắc bệnh cao. Đó cũng là nguyên nhân tại sao tỷ lệ người mắc bệnh tiểu đường ở Việt Nam còn cao.

Thêm vào đó, khi ý thức phòng bệnh của người dân còn kém (có tới 70-80% số người dân không hiểu biết về bệnh và cách phòng bệnh) thì việc ngăn ngừa bệnh rất khó khăn. Nhất là ở những vùng nông thôn khi mà điều kiện vật chất ở bệnh viện vẫn chưa thể đảm bảo cho việc thực hiện những xét nghiệm cần thiết.

Trong thời gian thực hiện đề tài, nhóm thực hiện đã có cơ hội khảo sát và thu thập dữ liệu tại các bệnh viện lớn trên địa bàn thành phố Hồ Chí Minh như bệnh viện Chợ Rẫy, bệnh viện Quận Thủ Đức, bệnh viên Đa Khoa Thủ Đức, bệnh viện Quân Dân Miền Đông…Sau thời gian khảo sát và làm việc với bệnh viện, nhóm đã được cho biết rằng, đa số các bệnh viện này đều đã tự xây dưng riêng cho mình một hệ thống quản lý riêng nhưng vẫn có một số điểm không hoàn chỉnh như sau: bệnh viên Quân Dân Miền Đông chưa xây dựng được hệ thống quản lý nội trú, bệnh án của bệnh nhân điều không được lưu vào cơ sở dữ liệu, chương trình quản lý không tích hợp hệ hỗ trợ ra quyết định.

Và một khi hệ thống quản lý vẫn chưa được xây dựng hoàn chỉnh thì khái niệm ứng dụng Data mining vào y học hoàn toàn lạ lẫm đối với các bác sĩ tại các bệnh viện lớn này.

Do đó, nhóm đã quyết định ứng dụng Data mining vào y học để xây dựng nên một hệ hỗ trợ ra quyết định trong khám chữa bệnh tiểu đường riêng biệt dựa trên những dữ liệu bao gồm các mẫu xét nghiệm đa số dành cho những bệnh nhân có nguy cơ mắc bệnh tiểu đường như: đường huyết, máu mỡ, men gan…

1.2. Mục tiêu

Nghiên cứu, phân tính, cài đặt và đánh giá các giải thuật đã sử dụng. Qua đó chọn nên một thuật toán thích hợp nhất để xây dựng nên ứng dụng chẩn đoán bệnh tiểu đường.

Đối tượng được chẩn đoán ở đây là những bệnh nhân đã thực hiện qua những xét nghiệm dành cho những bệnh nhân có nghi ngờ mắc bệnh tiểu đường. Để thực hiện được điều này, nhóm đã đến từng bệnh viện trên địa bàn Tp Hồ Chí Minh để thu thập dữ liệu của tất cả những bệnh nhân đã trải qua các ca xét nghiệm.

Các bước xây dựng ứng dụng thực tế:

- Bước 1: Thu thập các bộ dữ liệu khám bệnh và xét nghiệm của các bệnh nhân.
- Bước 2: Phân tích và cài đặt những thuật toán đã được chọn trước.
- Bước 3: Đánh giá kết quả của từng thuật toán.
- Bước 4: Áp dụng giải thuật thích hợp nhất để xây dựng ứng dụng.

1.3. Phạm vi khóa luận

Phạm vi của khóa luận dừng lại ở việc nghiên cứu, phân tích, cài đặt và đánh giá các giải thuật dựa trên các bộ dữ liệu của các bệnh nhân đã thực hiện những xét nghiệm. Qua đó tìm ra được những giải thuật thích hợp cho việc xây dựng nên hệ thống hỗ trợ ra quyết định trong chẩn đoán bệnh tiểu đường.

Đối tượng nghiên cứu: Những bệnh nhân đã từng thực hiện xét nghiệm và chẩn đoán bệnh tại các bệnh viên trên địa bàn Tp Hồ Chí Minh.

Đối tượng chẩn đoán: Những bệnh nhân chưa được thực hiện chẩn đoán bệnh tiểu đường.

1.4. Phương pháp nghiên cứu và nội dung thực hiện

Phương pháp tiếp cận:

- Tìm hiểu về hệ thống hỗ trợ ra quyết định lâm sàng (Clinical Decision Support System).
  - Các kỹ thuật khai phá dữ liệu trong hệ hỗ trợ ra quyết định.
  - Các phương pháp triển khai.
  - Các phương pháp đánh giá
- Tiến hành ứng dụng hệ hỗ trợ ra quyết định

Nội dung thực hiện:

- Nghiên cứu những tài liệu có liên quan đến CSDD.
- Liên hệ với các bệnh viện để thực hiện thu thập dữ liệu.
- Nghiên cứu các giải thuật đã từng được áp dụng trong CSDD.
- Tiến hành xây dựng hệ hỗ trợ ra quyết định lâm sàn.
- Thực hiện kiểm thử và đánh giá.

1.5. Kết quả dự kiến

Cài đặt thành công những giải thuật có thể đưa ra những chẩn đoán với độ chính xác cao nhất.

Xây dựng một ứng dụng đã được cài đặt giải thuật có độ chính xác cao nhất và tương tác với người dùng thông qua:

- Đầu vào:

- Dữ liệu của bệnh nhân cần được thực hiện chẩn đoán.

- Đầu ra:

  - Kết quả chẩn đoán.
  - Chỉ ra được những thuộc tính có khả năng dẫn đến khả năng mắc bệnh tiểu đường.
  - Độ chính xác của giải thuật vừa sử dụng.

1.6. Bố cục báo cáo

Chương 2: *Data mining và những thành tựu trong y học*

Trong chương này, nhóm thực hiện sẽ giới thiệu chung về Data mining và những ứng dụng tiêu biểu nhất của Data mining trong nền y học thế giới. Nêu lên những điểm nỗi bật cũng như những khó khăn của việc ứng dụng Data mining vào y học.

Chương 3: *Hệ hỗ trợ ra quyết định lâm sàng*

Sau chương 2, nhóm thực hiện sẽ giới thiệu về hệ hỗ trợ ra quyết định và hệ hỗ trợ ra quyết định lâm sàng. Đồng thời nếu lên những tính năng tiêu biểu của hệ hỗ trợ ra quyết định lâm sàng cùng với những thử thách trong qua trình triển khai và tương lai của hệ này. Sau cùng sẽ là các phương pháp tiếp cận thường được sử dụng trong quá trình xây dựng hệ.

Chương 4: *Triển khai*

Ở chương này, nhóm sẽ nêu lên những thành phần của hệ thống bao gồm dữ liệu đầu vào và kết quả đầu ra. Phương pháp thực hiện trong các quá trình tiền xử lý dữ liệu, cài đặt giải thuật.

Chương 5: *Kết quả thực nghiệm*

Trong chương cuối, nhóm thực hiện sẽ trình bày tổng quát về ứng dụng mà nhóm đã xây dựng. Những đánh giá về kết quả thực nghiệm và đưa ra kết luận cùng với hướng phát triển của đề tài.

CHƯƠNG 2: DATA MINING VÀ NHỮNG THÀNH TỰU TRONG Y HỌC

Trong chương này, nhóm sẽ nêu lên những khái niệm về Data mining và những thành tựu nổi bật của Data mining trong nên y học thế giới. Qua đó đưa ra những điểm nổi bật, những khó khăn của việc ý dụng Data mining trong y học. Và cuối cùng là những phương pháp khai phá dữ liệu đã được ứng dụng.

2.1. Data mining là gì

Data mining (còn được gọi là khai phá dữ liệu) là một quá trình trích xuất những thông tin trong một tập dữ liệu lớn. Nội dung của thông tin bao gồm các lĩnh vực như nhận dạng mô hình, thống kê, khoa học máy tính và quản lý cơ sở dữ liệu. Vì vậy, định nghĩa về Data mining chủ yếu phụ thuộc vào quan điểm của người đưa ra định nghĩa đó**[**14**][**13**]**.

Mục tiêu chính của Data mining là trích xuất những thông tin hữu ích và các mô hình mới từ cơ sở dữ liệu. Việc đưa ra một mô hình mới nhằm phục vụ cho hai mục đích chính là: dự đoán và mô tả.

Data mining đã được ứng dụng khá nhiều trong lĩnh vực kinh tế như tìm ra các giao dịch gian lận, marketing, bán lẻ…Nhưng ngoài ra, nó còn rất thích hợp cho việc hỗ trợ ra quyết định trong y học. Khi lượng dữ liệu được tạo ra trong các cơ sở y tế ngày càng lớn thì các tổ chức y tế đã bắt đầu quan tâm đến khai phá dữ liệu để nâng cao chất lượng dịch vụ và tránh lãng phí tài nguyên.

2.2. Những nghiên cứu đầu tiên

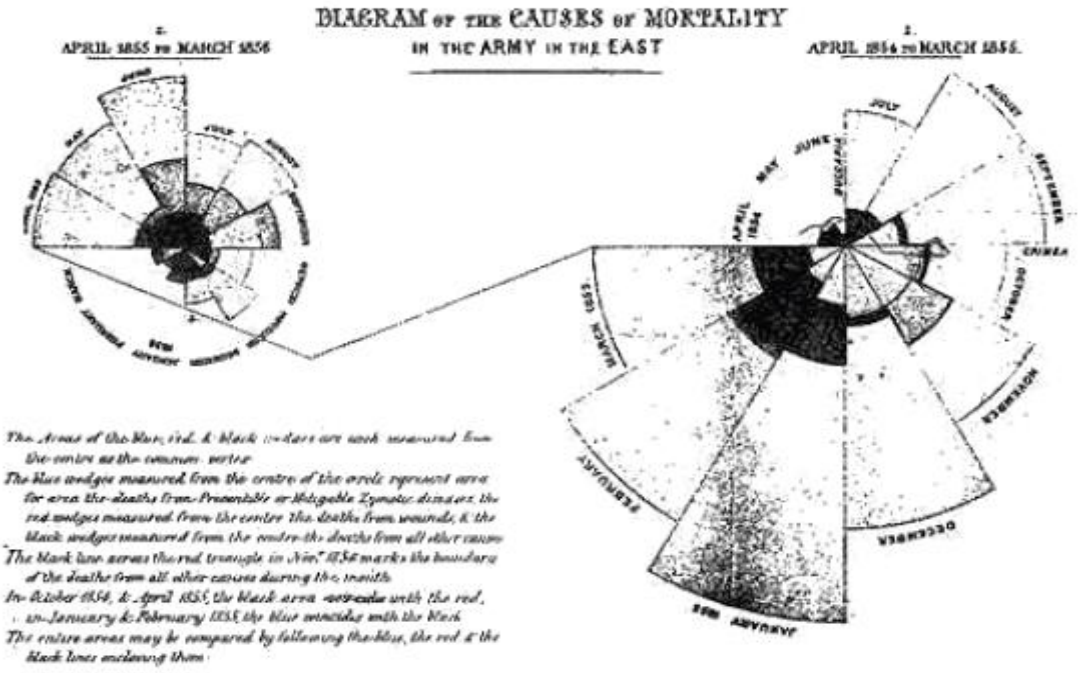
Dựa trên dữ liệu y học đưa ra những thông tin hữu ích (còn được gọi là Evidence Based Medicine – EBM) đã tồn tại trong nhiều thế kỉ trước**[**15**]**.

John Snow, được biết đến là cha đẻ của dịch tễ học hiện đãi, đã sử dụng các bản đồ đồ thị vào năm 1854 để phát hiện ra nguồn bệnh thổ tả và đã chứng minh rằng bệnh này lây lan qua hệ thống cấp nước. Ông đã đếm số lượng các bệnh nhân và vẽ sơ

đồ vị trí của các bệnh nhân trên bản đồ bằng các thanh màu đen. Khi đó ông phát hiện rằng hầu hết các ca tử vong đều nằm xung quanh một điểm bơm nước xác định tại London.

Hình 2.1. Bản đồ đồ thị

Năm 1855, Florence Nightngale đã phát minh ra biểu đồ phân cực để chỉ ra nhiều ca tử vong của binh lính do sự mất vệ sinh trong khám lâm sàng và hoàn toàn có thể ngăn ngừa được. Bà đã sử dụng biểu đồ này để thuyết phục các nhà lãnh đạo thực hiện các chính sách cải cách nhằm giảm số lượng các ca tử vong.



Hình 2.2. Biểu đồ phân cực

Trong những nguyên cứu này, Snow và Nightngale đã chính mình thực hiện việc thu thập dữ liệu, sàng lọc và phân tích thông qua các dữ liệu về tỉ lệ tử vong trong suốt thời gian nghiên cứu vì số lượng dữ liệu có thể quản lý được. Ngày nay, dân số trở nên đông đúc, tốc độ phát bệnh của bệnh dịch làm cho việc thao táo dữ liệu bằng những phương pháp trước đây hoàn toàn không thể thực hiện.

Data mining và những ứng dụng của nó trong y học và sức khỏe cộng đồng là một lĩnh vực còn non trẻ nhưng hết sức hữu ích. Tuy phát triển khá chậm nhưng vẫn được áp dụng để giải quyết các vấn đề khác nhau của việc khai phá tri thức trong lĩnh vực này.

2.3. Những ứng dụng của Data mining trong y học

Ngăn ngừa các lỗi hay mắc phải trong quá trình chăm sóc bệnh nhân: Khi các tổ chức ứng dụng Data mining vào dữ liệu thì nhiều tri thức hữu ích sẽ được khai phá và khả năng giảm bớt các ca tử vong lại tăng thêm. Một nghiên cứu cho thấy có đến 87% ca tử vong trong các bệnh viện lớn tại Mỹ có thể ngăn ngừa được nếu các nhân

viên bệnh viện (kể cả bác sĩ) không mắc phải các sai sót trong quá trình chăm sóc sức khỏe cho bệnh nhân (HeathGrades Hospitals Study 2007)**[**15**]**.

Hoạch định chính sách trong y tế: Lavrac đã kết hợp cơ sở dữ liệu không gian và Data mining để phân tích điểm giống nhau giữa các trung tâm y tế tại Slovenia. Ông đã khai phá được các mô hình trong trung tâm y tế để có thể đưa ra các chính sách khuyến nghị đến Bộ Y Tế nhằm cải thiện các chính sách.

Phát hiện điểm bất thường trong bảo hiểm y tế: Tiết kiệm tiền và chi phí là một trong những mục tiêu quan trọng mà các tổ chức y tế hướng đến khi áp dụng Data mining. Nếu như trong lĩnh vực kinh tế, Data mining đã được áp dụng để tìm ra các gian lận trong thẻ tín dụng và bảo hiểm thì hiện nay, chúng còn được dùng để tìm ra những điểm bất thường trong bảo hiểm y tế.

Phát hiện và ngăn chặn bệnh dịch:

- Cheng đã đưa ra dẫn chứng trong việc sử dụng các thuật toán phân lớp để phát hiện ra các trường hợp mắc bệnh tim – một loại bệnh được quan tâm nhất thế giới.

- Cao cũng đã áp dụng Data mining như một công cụ nhằm kiểm soát các thí nghiệm vac xin lâm sàn. Việc phát hiện các bệnh nhân có biểu hiện khác thường trở nên dễ dàng hơn nếu chỉ nhìn vào các tập dữ liệu.

Ngăn ngừa, quản lý bệnh dịch và đưa ra các chính sách trong y tế:

- Kellogg đã nêu ra một kỹ thuật kết hợp mô hình không gian và khai phá dữ liệu không gian để tìm ra các điểm bùng phát dịch bệnh. Phân tích kết quả được trích xuất từ khai phá dữ liệu để đưa ra các chính sách nhằm phát hiện và quản lý dịch bệnh bùng phát.

- Wong đã giới thiệu WSARE, một giải thuật phát hiện dịch bệnh khi vừa ở giai đoạn đầu. WSARE được tạo nên dựa trên các luật kết hợp và mạng Bayesian. Áp dụng WSARE trên có mô hình giả lập có thể cho ra những

kết quả tương đối chính xác. Đây làm một giải thuật đầy tiềm năng đang đợi đến ngày được ứng dụng vào thực tế.

Các hệ thống ra quyết định không gây tổn thương cho bệnh nhân: Một số chẩn đoán và xét nghiệm có thể gây tổn thương đến một bộ phận nhất định nào của bệnh nhân. Ví dụ như sinh thiết, một loại xét nghiệm ở phụ nữ nhằm phát hiện bệnh ung thư cổ tử cung.

- Thangavel đã sử dụng thuật toán gom cụm K – means để phân tích các bệnh nhân ung thư cổ tử cung và đã nhận thấy việc gom cụm dữ liệu có thể đưa ra những kết quả chẩn đoán chính xác hơn các phương pháp hiện có mà không gây tổn thương cho bệnh nhân. Ông cũng đã tìm thấy một số bộ thuộc tính có thể cung cấp cho các bác sĩ như một hỗ trợ trong việc quyết định đưa một bệnh nhân có khả năng mắc bệnh ung thư đi tiến hành sinh thiết hay không.

- Gorunesca đã đưa một phương pháp khác là sử dụng máy tính CAD và siêu âm nội soi Elastographỵ đã được tính hợp Data mining để giúp các bác sĩ đưa ra quyết định xem bệnh nhân có phải đi tiến hành sinh thiết hay không. Trong khi đó phương pháp truyền thống là các bác sĩ sẽ nhìn vào các bộ phim siêu âm và thiết bị để đưa ra quyết định.

Phân loại thuốc có hại: Một số thuốc và hóa chất tuy đã được đánh giá rằng sẽ không gây hại cho con người nhưng sau một thời gian dài sử dụng thì kết quả là ngược lại. Tổ chức US Food và Drug Administration đã sử dụng Data mining để tìm ra các loại thuốc có hại khi sử dụng lâu dài trong cơ sở dữ liệu của họ. Giải thuật đó có tên là MGPS (Multi-item Gamma Poisson Shrinker) đã tìm ra 67% loại thuốc có hại và kết quả đưa ra sớm hơn 5 năm nếu dùng những cách thông thường

Qua những liệt kệ trên, chúng ta đã có thể thấy được tầm quan trọng của Data mining trong y học. Nhưng vẫn còn rất nhiều cách sử dụng Data mining khác như trong lĩnh vực này hiện đang được các chuyên gia đào sâu nghiên cứu.

2.4. Những khó khăn trong việc ứng dụng Data mining vào y học

Việc áp dụng Data mining vào lĩnh vực y tế gặp rất nhiều khó khăn do chính đặc trưng, khác biệt giữa 2 lĩnh vực này. Trong khi Data mining chỉ quan tâm đến việc mô tả dữ liệu thì y học lại quan tâm đến những giải thích dù một chi tiết nhỏ nhất cũng có thể là một ranh giới giữa sống và chết.

Cho dù các kết quả thi được từ quá trình khai phá dữ liệu có đáng tin cậy nhưng việc thay đổi thói quen của các bác sĩ, y tá cũng là một điều không dễ dàng. Ayres đã cho biết đã có nhiều trường hợp các bác sĩ từ chối thay đổi các chính sách của bệnh viện. Đã có một trường hợp xảy ra khi một bác sĩ quên rửa tay sau khi khám nghiệm tử thi và gây ra nhiều ca tử vong ngay sau đó. Đến lúc này, các bác sĩ mới thay đổi thói quen của họ.

Không chỉ thế, đa số các bác sĩ thường chỉ xin lời khuyên từ các bác sĩ cấp trên có nhiều kinh nghiệm hơn họ hơn là chỉ ngồi nhìn vào những mô hình được khai phá từ cơ sở dữ liệu.

Ngoài ra những dữ liệu riêng tư của bệnh nhân cũng là một vật cản lớn trong việc ứng dụng khai phá dữ liệu vào y học vì để đưa ra một kết quả chính xác nhất thì cần một lượng lớn những dữ liệu cần thiết. Nhưng chỉ có những dữ liệu riêng tư này mới có thể giúp con người ta tránh được những bệnh chết người

CHƯƠNG 3: HỆ HỖ TRỢ RA QUYẾT ĐỊNH LÂM SÀNG

Trong chương này, nhóm sẽ giới thiệu về hệ hỗ trợ ra quyết định và một nhánh khác của nó là hệ hỗ trợ ra quyết định lâm sàng. Đồng thời nêu lên những tính năng tiêu biểu mà một hệ hỗ trợ cần có, những thử thách trong quá trình triển khai, tương lai hệ hỗ trợ ra quyết định lâm sàng và những phương pháp tiếp cận của hệ.

3.1. Hệ hỗ trợ ra quyết định

3.1.1. Hệ hỗ trợ ra quyết định là gì?

Đầu thập kỷ 70, Gorry và Scott – Morton (1971) định nghĩa HHTRQĐ là các hệ thống dựa trên hệ thống tương tác với máy tính cho các nhà ra quyết định dùng các dữ liệu và mô hình để giải quyết các vấn đề phi cấu trúc.

Little đưa ra một định nghĩa khác về HHTRQĐ là tập các cơ sở mô hình chứa các thủ tục xử lý dữ liệu giúp các nhà quản lý ra quyết định, hệ thống cần phải đơn giản, dễ điều khiển, thích nghi và dễ liên lạc với nhau.

Alter (1980) đưa ra khái niệm HHTRQĐ bằng cách so sánh các hệ thống xử lý dữ liệu:

Khía cạnh	Hệ hỗ trợ ra quyết định	Hệ thống xử lý dữ liệu
Sử dụng	Chủ động	Bị động
Người sử dụng	Nhà quản lý	Văn phòng
Mục tiêu	Tính hiệu quả, tính linh hoạt	Hiệu quả máy móc, tính phi mâu thuẫn
Phạm vi về thời gian	Hiện tại và tương lai	Quá khứ
Mục đích, tiêu đề	Tính linh hoạt	Tính phi mâu thuẫn

Bảng 3.1. So sách các hệ thống xử lý dữ liệu

Moore và Chang (1980) chỉ ra rằng khái niệm cấu trúc không đủ ý nghĩa trong trường hợp tổng quá, một bài toán có thể được mô tả là có cấu trúc hoặc không có cấu trúc chỉ liên quan đến người ra quyết định. Do đó, HHTRQĐ là:

- Hệ thống có khả năng mở rộng
- Có khả năng trợ giúp phân tích dữ liệu và mô hình hóa quyết định
- Được sử dụng cho những hoàn cảnh và thời gian bất thường

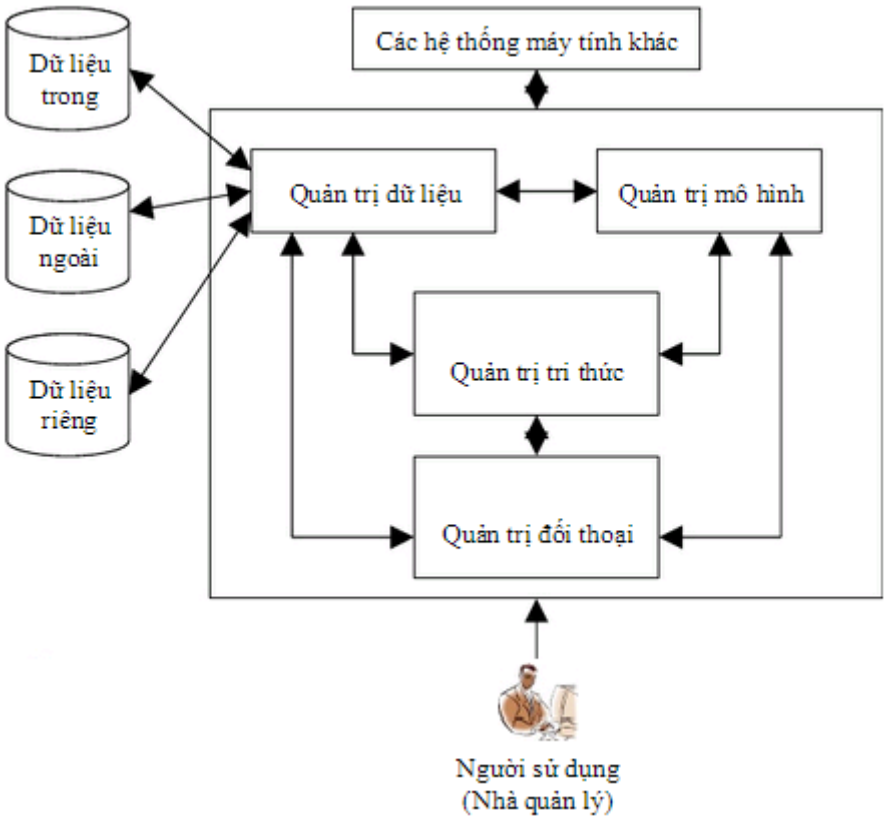
Bonzek, Holsapple, Whinston (1980) đưa ra khái niệm tổng quá hơn về HHTRQĐ gồm các thành phần chính:

- Có một hệ ngôn ngữ là cơ chế cho phép tương tác giữa người dùng và các thành phần khác của hệ
- Một hệ tri thức chứa các tri thức về lĩnh vực bao gồm dữ liệu và các loại thủ tục
- Hệ xử lý bài toán, chứa đựng các khả năng xử lý bài toán và người ra quyết định cần đến

Keen (1980) cho rằng HHTRQĐ là sản phẩm của quá trình phát triển trong đó người sử dụng HHTRQĐ, người tạo ra HHTRQĐ và bản thân HHTRQĐ có khả năng ảnh hướng tác động đến sự phát triển của hệ thống và các thành phần

Như vậy, có nhiều cách định nghĩa HHTRQĐ khác nhau. Nhưng về tổng quát, HHTRQĐ là hệ thống thông tin hỗ trợ bằng máy tính có thích nghi linh hoạt và tương tác với nhau, đặc biệt được phát triển để hỗ trợ một vấn đề quản lý không có cấu trúc nhằm cải tiến việc ra quyết định. Nó tập hợp dữ liệu cung cấp cho người sử dụng một giao diện thân thiện và cho phép tự ra quyết định một cách sáng suốt. Nó hỗ trợ tất cả các giai đoạn của việc ra quyết định và bao gồm cả một cơ sở tri thức.

3.1.2. Kiến trúc chung của hệ hỗ trợ ra quyết định



Hình 3.1. Mô hình kiến trúc của Hệ hỗ trợ ra quyết định

Trong đó:

- Quản trị dữ liệu: bao gồm các CSDL chứa dữ liệu và được quản lý bởi một phần mềm là hệ quản trị CSDL (quản lý và khai thác)
- Quản trị mô hình: cho phép khai thác và quản lý các mô hình định lượng (xử lý) khác nhau, cung cấp khả năng phân tích cho hệ thống
- Quản trị đối thoại: cung cấp giao diện cho người dùng để liên lạc và ra lệnh cho HHTRQĐ
- Quản trị tri thức: hoạt động như một thành phần độc lập, hoặc có thể trợ giúp cho bất kỳ một trong ba hệ thống nói trên

3.2. Hệ hỗ trợ ra quyết định lâm sàng

3.2.1. Khái niệm

Hệ hỗ trợ ra quyết định lâm sàng là một dạng khác của HHTRQĐ. HHTRQĐLS là một hệ thống máy tính được xây dựng để hỗ trợ cho các bác sĩ và các chuyên gia sức khỏe trong việc đưa ra quyết định dựa trên dữ liệu của bệnh nhân.

3.2.2. Các dạng của hệ hỗ trợ ra quyết định lâm sàng

HHTRQĐLS đã được phát triển trong suốt 30 năm qua, chúng có thể là một hệ độc lập hoặc là một phần của hệ thống phi thương mại được xây dựng trên dữ liệu bệnh nhân. Trong những năm gần đây, nhiều hệ thống phi thương mại trở nên phổ biến trên thị trường và nhiều nhà đầu tư đã bắt đầu tích hợp HHTRQĐLS vào hệ thống dữ liệu của họ**[**12**]**.

Hiện nay, HHTRQĐLS chia thành hai dạng chính:

- Knowledge – Based Systems.
- Nonknowledge – Based Systems.

3.2.2.1. Knowledge – Based Systems

KBS là một chương trình máy tính được thiết kế để mô hình hóa khả năng giải quyết vấn đề của một chuyên gia con người.

KBS là hệ thống dựa trên tri thức, cho phép mô hình hóa tri thức của chuyên gia, dùng tri thức để giải quyết vấn phức tạp thuộc cùng lĩnh vực.

Hai yếu tố quan trọng trong KBS là tri thức chuyên gia và lập luận, tương ứng với hệ thống có 2 khối chính là cơ sở tri thức và động cơ suy diễn.

- Cơ sở tri thức (Knowledge – Based): chức các tri thức chuyên sâu về lĩnh vực như chuyên gia. Cơ sở tri thức bao gồm các sự kiện, các luật, các khái niệm và các quan hệ. Các luật được biểu diễn dưới dạng IF…THEN…

- Động cơ suy diễn: bộ xử lý tri thức theo mô hình hóa theo cách lập luận của chuyên gia. Động cơ hoạt động dựa trên thông tin về vấn đề đang xét, so sánh với trị thức lưu trong cơ sở tri thức rồi rút ra kết luận.

Y học là một lĩnh vực màu mỡ trong việc ứng dụng các khái niệm trên. Trong 20 năm trở lại đây, các lập trình viên của những hệ thống này bắt đầu được hỗ trợ trong việc xây dựng các chương trình hỗ trợ chăm sóc sức khỏe.

Một bất tiện đối với những KBS là một hệ độc lập và dữ liệu bệnh nhân cần được người dùng nạp trực tiếp. Còn đối với những KBS được tính hợp vào hệ thống thì dữ liệu của bệnh nhân có thể được lấy dữ hệ thống dữ liệu.

Nhìn chung, Việt Nam hiện nay cũng đã bắt đầu phát triển những hệ chuyên gia riêng cho mình. Trong số đó gồm những hệ chuyên gia nổi tiếng như:

- *AMD* – hệ chuyên gia chẩn đoán bệnh lâm sàn của tác giả Ngô Thắng Lợi – giải nhất của cuộc thi sáng tạo thanh thiếu niên, nhi đồng tình Thừa Thiên Huế lần thứ IV (2010 – 2011)

- *Medinfo* – hệ chuyên gia chẩn đoán và phân loại bệnh lâm sàng của tác giả Nguyễn Tấn Tôn Thất Đỗ Vũ

- *Chương trình chẩn đoán vành mạch và suy tim* của hai tác giả Văn Thế Thành và Trần Minh Bảo thuộc trường Đại học Công nghệ thực phẩm Tp. HCM[7] .

3.2.2.2. Nonknowledge – Based Systems

Khác với KBS, NBS sử dụng một dạng của trí tuệ nhân tạo được gọi là máy học, cho phép máy tính học những kinh nghiệm trong quá khứ để có thể nhận dạng các dữ mô hình trong dữ liệu lâm sàn. Mạng Neural nhân tạo (Artificial Neurol Networks) và các giải thuật di truyền(Genetic Algorithms) là hai dạng của NBS.

- Mạng Neural nhân tạo (ANN):

- Mạng neural nhân tạo là một mô phỏng xử lý thông tin, được nghiên cứu ra từ hệ thống thần kinh của sinh vật, giống như bộ não để xử lý thông tin. Nó bao gồm các mối gắn kết cao cấp để xử lý các yếu tố làm việc trong mối liên hệ giải quyết vấn đề rõ ràng. ANN giống như con người, được học bởi kinh nghiệm, lưu những kinh nghiệm hiểu biết và sử dụng trong những tình huống phù hợp.

- ANN đầu tiên được giới thiệu vào năm 1943 bởi nhà thần kinh học Warren McCulloch và nhà logic học Walter Pits. Nhưng với những kỹ thuật trong thời gian này chưa cho phép họ nghiên cứu nhiều. Những năm gần đây, mô phỏng ANN xuất hiện, phát triển và các nghiên cứu đã được thực hiện trong ngành: điện, điện tử, kỹ thuật chế tạo, y học, quân sự, kinh tế…

- Khi được ứng dụng vào y học, ANN khá giống với KBS. Thay vì lấy tri thức từ các sách y học hoặc từ một chuyên gia lâm sàng thì ANN lại phân tích các mô hình trong dữ liệu bệnh nhân để thu thập những mối liên hệ từ những dấu hiệu và triệu chứng của bệnh nhân. KBS thường bao hàm một lượng lớn các bệnh thường gặp ở con người, dữ liệu đầu vào là những triệu chứng, dấu hiệu và dữ liệu đầu ra là bệnh nhân có khả năng mắc bệnh đó hay không. ANN cũng tương tự với dữ liệu đầu vào và đầu ra nhưng chỉ tập trung vào một bệnh nhất định như nhồi máu cơ tim, ung thư, tiểu đường…

- Ưu điểm chung của ANN và NBS là loại bỏ đi tất cả các luật IF…THEN… và không phải nhập tri thức từ các chuyên gia. Ưu điểm lớn nhất của ANN là có thể thao tác trên cả những dữ liệu không hoàn thiện. Ngoài ra, ANN cũng không cần học một lượng lớn dữ liệu để đưa ra kết quả dữ đoán nhưng khi cơ sở dữ liệu huấn luyện càng hoàn thiện thì độ chính xác của ANN càng tăng cao.

- Tuy nhiên ANN vẫn có một số nhược điểm riêng của nó, đặc biệt là quá trình huấn luyện dữ liệu có thể làm tiêu tốn khá nhiều thời gian. Kết quả huấn luyện thường rất khó hiểu làm cho độ tin cậy của ANN trờ thành một vấn đề lớn. Dù là vậy nhưng ANN vẫn được ứng dụng rất nhiều vào lĩnh vực y khoa như chẩn đoán viêm ruột thừa, đau lưng, mất trí nhớ, nhồi máu cơ tim, các trường hợp tâm thần…

- Giải thuật di truyền (GA):

  - Một phương pháp phi cơ sở tri thức khác thường được sử dụng để tạo nên HHTRQĐLS là giải thuật di truyền. Được pháp triển bởi John Holland ở những năm 1940 tại Massachusetts Institute of Technology dự trên học thuyết tiến hóa của Darwin: sự thích nghi và chọn lọc tự nhiên. Từ tập các lời giải ban đầu, thông qua nhiều bước tiến hóa để hình thành những lời giải tốt hơn, cuối cùng sẽ tìm ra lời giải tối ưu nhất.

  - GA sử dụng các thuật ngữ lấy từ di truyền học:

    - Một tập hợp các lời giải được gọi là một lớp hay một quần thể (Population)

    - Mỗi lời giải được biểu diễn bởi một nhiễm sắc thể hay cá thể (chromosome)

    - Nhiễm sắc thể tạo thành từ các gien.

  - Một quá trình tiến hóa được thử hiện trên một quần thể tương đương với sự tìm kiếm trên không gian các lời giải có thể của bài toán. Quá trình tìm kiếm này luôn đòi hỏi sự cân bằng giữa hai mục tiêu: *Khai thác lời giải tốt nhất và xem xét toàn bộ không gian tìm kiếm.*

- GA thực hiện tìm kiếm theo nhiều hướng bằng cách duy trì tập hợp các lời giải thích có thể, khuyến khích sự hình thành và trao đổi thông giữa các hướng. Tập lời giải phải trải qua nhiều bước tiến hóa, tại mỗi thế hệ, một tập hợp mới các cá thể được tạo ra chứa các phần của những cá thể thích nghi nhất trong thế hệ cũ. Đồng thời GA khai thác một cách hiệu quả thông tin trước đó để suy xét trên điểm tìm kiếm với mong muốn có được sự cải thiên qua từng thế hệ. Như vậy, các đặc trưng được đánh giá tốt sẽ có cơ hội phát triển và các tính chất tồi (không thích nghi với môi trường) sẽ có xu hướng biến mất.

- Hiện nay giải thuật di truyền được áp dụng ngày càng nhiều trong kinh doanh, khoa học và kỹ thuật vì tính chất không quá phức tạp mà hiểu quả của nó. Hơn nữa, giải thuật di truyền không đòi hỏi khắt khe với không gian tìm kiếm như giả định về sự liên tục, sự có đạo hàm…Dù cũng được áp dụng vào y học nhưng số lượng phần mềm ứng dụng GA hoàn toàn ít hơn so với ANN. Tuy nhiên GA cũng đã làm nên tên tuổi của mình trong việc xây dựng nên hệ thống chẩn đoán các bệnh về tiết niệu ở phụ nữ.

3.2.3. Tính năng tiêu biểu

Hệ thống sẽ báo cho người dùng khả năng xảy ra lỗi khi kê đơn thuốc, chẩn đoán bệnh đến bác sĩ và bộ chẩn đoán sẽ cải thiện khả năng chẩn đoán của họ.

Ngoài ra chúng còn có khả năng giảm thiểu mức độ nghiêm trọng và ngăn ngừa biến chứng của các bệnh nguy hiểm, đưa ra các tác dụng phụ của thuốc có thể làm ảnh hướng đến cả giá thành và chất lượng của chăm sóc y tế.

3.2.4. Những thử thách trong quá trình triển khai

Dữ liệu bao giờ cũng là một trong những thử thách đầu tiên của HHTRQĐ. Nhiều hệ thống yêu cầu người dùng phải thực hiện truy vấn hoặc nhập dữ liệu làm tiêu tốn khá nhiều thời gian đồng thời làm giảm đi tính tiện dụng của những hệ

HHTRQĐLS. Tuy nhiên vẫn có một số hệ thống được tích hợp liền với CSDL của bệnh viện nhưng hiện vẫn chưa có một phương pháp chẩn cho việc tích hợp nên việc triển khai hệ thống này cũng gặp không ít khó khăn.

Khi triển khai những hệ thống ra quyết định độc lập thường người ta phải đối mặt với câu hỏi: Ai sẽ là người nhập liệu? Bác sĩ là nhân tố chính trong việc đưa ra quyết định nhưng họ lại không phải là người thường xuyên tương tác với hệ thống.

Một thử thách khác là vấn đề về từ vựng. Điều này chỉ xảy ra khi bác sĩ sử dụng một hệ thống với những ngôn từ hoàn toàn không thuộc chuyên ngành của họ. Đa số các bác sĩ đều muốn một hệ thống được xây dựng dưới cái nhìn y học của họ.

Nhìn chung, một hệ thống muốn phát huy được toàn bộ khả năng của nó chỉ khi có sự kết hợp những chuyên gia xây dựng hệ thống và các y bác sĩ.

3.2.5. Tương lai của hệ hỗ trợ ra quyết định lâm sàn

Tuy còn gặp nhiều thử thách trong việc tích hợp nhưng chúng ta không thể phủ nhận được tiềm năng của HHTRQĐLS trong việc cải thiện chất lượng dịch vụ y tế. Ngày nay, nhiều HHTRQĐLS dành cho các bệnh nhân liên tục được ra đời đồng thời vấn đề an toàn của bệnh nhân được quan tâm nhiều hơn tạo ra tiền đề cho việc phát triển các hệ thống này. Nhiều bác sĩ cũng bắt đầu sử dụng HHTRQĐLS theo yêu cầu của công động để nâng cao chất lượng mặc dù trước đây họ đã từng từ chối không sử dụng chúng.

Công dụng của hệ thống thông tin và HHTRQĐLS trong việc cải thiển chất lượng dịch vụ y tế đã được cộng đồng y khoa chú ý hơn nhờ những thành công đáng có của những hệ thống này.

Tuy vẫn chưa có văn bản pháp lý nào yêu cầu việc sử dụng những hệ thống này nhưng nếu chúng không vi phạm y đức và cả khả năng cải thiện chất lượng dịch vụ thì không thể bỏ qua được những lợi ích mà chúng mang lại.

3.3. Phương pháp tiếp cận

Với sự phát triển của điện toán và công nghệ y học, các bộ dữ liệu lớn cũng như các phương pháp phân loại dữ liệu trở nên đa dạng, phức tạp đã bắt đầu được phát triển và nghiên cứu. Đồng thời làm cho khai phá dữ liệu nhận được nhiều sự chú ý trong vào thập kỉ vừa qua, và đã có một số lượng lớn các ứng dụng bao gồm cả khai phá dữ liệu và các HHTRQĐLS. Bảng sau liệt kê một số HHTRQĐLS đã từng được sử dụng để cải thiện chất lượng y tế**[**12**]**:

Hệ thống	**Mô tả**
Hệ thống nhận diện và diễn giải hình ảnh y khoa	
Máy tính hỗ trợ chẩn đoán ung thư vú	Sự khác biệt giữa các nốt vú lành tính và ác tính, dựa trên nhiều tính năng siêu âm
Chẩn đoán của chứng rối loạn thần kinh cơ	Phân loại tín hiệu điện đồ (EMG), dựa trên các hình dạng và tỷ lệ phát của đơn vị động cơ hành động tiềm năng(MUAPs)
Hệ thống giáo dục	
Khai thác tài liệu sinh y học	Hệ thống tự động để khai phá MEDLINE cho các tham khảo về gen và protein và đánh giá sự phù hợp của mỗi tham khảo được phân công
Phân tích biểu hiện của gen và protein	
Phát hiện các tế bào ung thư vú	Nhận dạng các nhóm tế bào ung thứ vú dựa vào việc phân tích các biểu hiện của gen

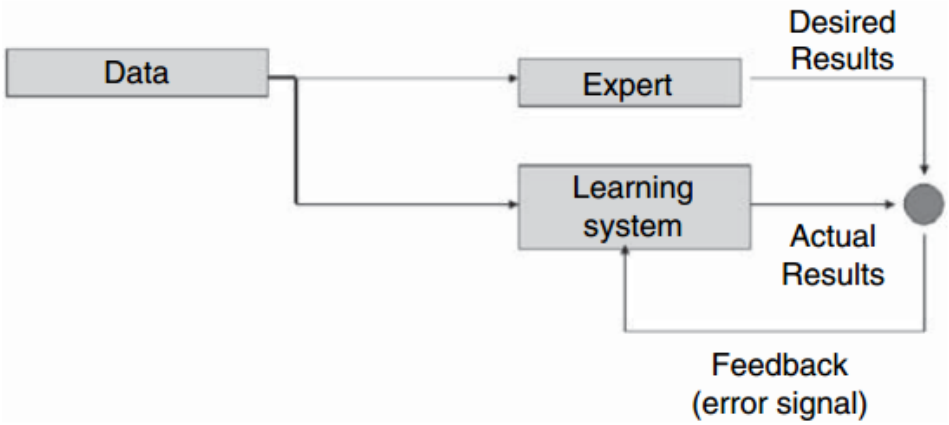
Bảng 3.2. Một số hệ hỗ trợ ra quyết định lâm sàng đã được sử dụng

Khai phá dữ liệu có thể xem như một việc học từ dữ liệu. Do đó, khai phá dữ liệu chia thành hai hướng là học có giám sát và học không có giám sát

3.3.1. Học có giám sát

Học có giám sát, còn được gọi là khai phá dữ liệu định hướng. Trong đó, dữ liệu đã được phân lớp sẵn và tri thức thu được từ những dữ liệu này phải trải qua một quá trình được gọi là huấn luyện. Dữ liệu dùng để thực hiện qua trình huấn luyện được gọi là mẫu huấn luyện (*Training Set*).Mẫu huấn luyện lại bao gồm các biến phụ thuộc hoặc biến mục tiêu, biến độc lập hoặc biến đầu vào.

Hệ thống được điều chỉnh dựa trên các mẫu huấn luyện và tín hiệu lỗi (sự khác biệt giữa kết quả mong đợi và kết quả thực tế). Nói cách khác, học có giám sát có thể được xem như là một hoạt động làm giảm sự khác biệt giữa các giá trị mong đợi và giá trị thực tế như là sự tiến bộ của quá trình đào tạo. Nếu mẫu huấn luyện có đầy đủ dữ liệu thì sự khác biệt này sẽ được giảm thiểu và quá trình nhận dạng mô hình sẽ ngày càng chính xác hơn.



Hình 3.2. Mô hình học có giám sát

Mục tiêu của phương pháp này là để thiết lập một mối quan hệ hoặc dự đoán mô hình giữa biến độc lập và biến phụ thuộc. Trong mô hình đó, các biến độc lập có nhiệm vụ dự đoán và mô tả quá trình dự đoán (tùy vào giải thuật) kết quả của biến phụ thuộc. Mô hình này dùng được dùng để dự đoán giá trị trong tương lai hoặc hành vi của một đối tượng hoặc thực thể.

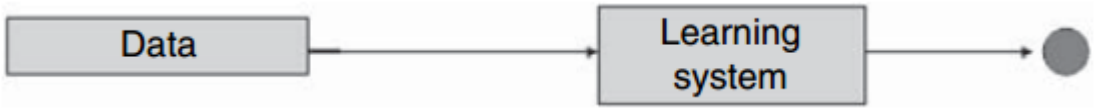
Một mô hình được gọi là mô hình phân loại nếu biến mục tiêu rời rạc và ngược lại nếu biến mục tiêu là liên tục thì mô hình đó được gọi là mô hình hồi quy.

Các giải thuật sử dụng phương pháp tiếp cận là học có giám sát: Cây quyết định, Naïve Bayes, mạng Neural, hồi quy tuyến tính, SVM…

3.3.2. Học không giám sát

Học không giám sát còn được gọi là học vô hướng, hệ thống được trình bày với một tập dữ liệu không được phân nhóm. Dựa trên sự tương đồng của những bộ dữ liệu mà hệ thống thực hiện việc gom các lớp hoặc các cụm cho đến khi một tập các mô hình bắt đầu xuất hiện.

Phương pháp này không có biến mục tiêu, tất cả các biến được xử lý theo cùng một cách nên không có sự phân biệt giữa các biến phục thuộc và biến độc lập.



Hình 3.3. Mô hình học không giám sát

Các giải thuật sử dụng phương pháp học không giám sát là: các thuật toán gom cụm dữ liệu…

CHƯƠNG 4: TRIỂN KHAI

Sau đây nhóm thực hiện sẽ giới thiệu về thành phần hệ thống bao gồm những dữ liệu đầu vào và kết quả đầu ra. Và cuối cùng là những phương pháp mà nhóm đã sử dụng để xử lý những dữ liệu đó.

4.1. Sơ lược về thành phần hệ thống

4.1.1. Dữ liệu đầu vào

Nhóm thực hiện đã sử dụng một bộ dữ liệu được đề xuất bởi hai tác giả *Dominick Doust* và *Zack Walsh* kết hợp với quá trình thực hiện thu thập dữ liệu tại các bệnh viện trên địa bàn thành phố Hồ Chí Minh. Nhóm đã xây dựng nên một bộ dữ liệu xét nghiệm thường dùng tại các bệnh viện**[**11.

STT	Mô tả	Bộ	Đơn vị tính
1.	Mã bệnh nhân	Thông tin cá nhân	Không
2.	Giới tính		
3.	Năm sinh		
4.	Ngày khám		
5.	Cholesterol	Xét nghiệm máu mỡ	mg/dL
6.	HDL_Cholesterol		mg/dL
7.	Triglyceride		mg/dL
8.	LDL_Cholesterol		mg/dL
9.	Glucose	Xét nghiệm sinh hóa và đường huyết	mg/dL
10.	Urea		mg/dL

11.	SGOT	Xét nghiệm men gan	U/L
12.	SGPT		U/L
13.	WBC	Xét nghiệm huyết đồ	G/L
14.	LYM#		G/L
15.	MONO#		G/L
16.	GRAN#		G/L
17.	LYM%		%
18.	MONO%		%
19.	GRAN%		%
20.	HGB		g/dL
21.	RBC		T/L
22.	HCT		%
23.	MCV		fL
24.	MCH		Pg
25.	MCHC		g/dL
26.	RDW_CV		%
27.	PLT		G/L
28.	MPV		fL
29.	PDW		%

30.	PCT		%
31.	Na	Xét nghiệm điện phân	mmol/l
32.	K		mmol/l
33.	Cl		mmol/l
34.	Ca		mmol/l
35.	Tiểu đường (Có hay không)	Phân lớp dữ liệu	Không

Bảng 4.1. Dữ liệu của hệ thống

Bảng dữ liệu trên được chia thành 7 bộ dữ liệu nhỏ: bộ thông tin cá nhân, bộ xét nghiệm máu mỡ, bộ xét nghiệm sinh hóa, bộ xét nghiệm men gan, bộ xét nghiệm huyết đồ, bộ xét nghiệm điện giải[1][3] và bộ dữ liệu phân lớp.

4.1.1.1. Thông tin cá nhân

STT	Thông tin cá nhân
1.	Mã bệnh nhân
2.	Giới tính
3.	Năm sinh
4.	Ngày khám

Bảng 4.2. Dữ liệu thông tin cá nhân của bệnh nhân

Thông tin cá nhân của bệnh nhân cũng quyết định khả năng mắc bệnh tiểu đường. Hơn 80% số lượng bệnh nhân mắc bệnh tiểu đường trong bộ dữ liệu mà nhóm tác giả dùng là những người có độ tuổi từ 60 – 70.

Do thời gian hạn hẹp nên nhóm không thể thu thập thêm thông tin bệnh nhân nằm trong hồ sơ bệnh án, đa số các bệnh viện vẫn chưa xây dựng được hệ thống bệnh án điện tử. Trong đó có một số thuộc tính rất quan trọng đó là: huyết áp, mức độ hoạt động và tiền sử mắc bệnh tiểu đường của gia đình bệnh nhân.

4.1.1.2. Máu mỡ

STT	Máu mỡ	Đơn vị tính
1.	Cholesterol	mg/dL
2.	HDL_Cholesterol	mg/dL
3.	Triglyceride	mg/dL
4.	LDL_Cholesterol	mg/dL

Bảng 4.3. Dữ liệu xét nghiệm máu mỡ

Rối loạn máu mỡ, các xơ vữa động mạch, tăng huyết áp, hội chứng thận hư…là một trong những triệu chứng thường thấy của người mắc bệnh tiểu đường.

Trong đó, rối loạn máu mỡ (*Dyslipidemia*) là thường gặp nhất. Đây là tên gọi chung của một số bệnh do xáo trộn các chất mỡ trong máu: hoặc quá nhiều hoặc quá ít các chất lipoprotein. Bệnh thường biểu hiện qua độ tăng cholesterol, tăng loại lipoprotein "xấu" (LDL), tăng lại triglyceride hoặc thiếu loại lipoprotein "tốt" (HDL).

4.1.1.2.1. Cholesterol

Cholesterol là một chất béo steroid, có ở màng tế bào của tất cả các mô trong cơ thể, và được vận chuyển trong huyết tương của mọi động vật. Hầu hết cholesterol không có nguồn gốc từ thức ăn mà nó được tổng hợp bên trong cơ thể. Cholesterol hiện diện với nồng độ cao ở các mô tổng hợp nó hoặc có mật độ màng dày đặc, như gan, tủy sống, não và mảng xơ vữa động mạch.

Cholesterol là một phần quan trọng của cơ thể, được dùng trong cấu tạo của màng tế báo, của một số hormone và một số các công dụng khác trong cơ thể. Nhưng có quá nhiều cholesterol trong máu là một nguy cơ lớn, có khả năng gây bệnh về tim mạch, nhất là nhồi máu cơ tim và tai biến mạch máu não.

Vì là chất mỡ, không hòa tan trong nước được, cholesterol và các chất mỡ như triglycerides, phải kết hợp với những khối tạp dễ tan trong nước là lipoprotein để dễ di chuyển trong máu. Vì thế, khi xét nghiệm lượng mỡ trong máu, ngoài tổng số cholesterol, người ta còn phân tích cholesterol theo các loại lipoprotein trong máu.

Tổng mức Cholesterol trong máu:

Nồng độ mg/DL	Giải thích
Dưới 200 mg/DL	Lý tưởng (Nguy cơ thấp)
Từ 200 đến 239 mg/DL	Chạm ngưỡng cao (Nguy cơ cao hơn)
Từ 240 mg/DL trở lên	Cholesterol trong máu cao (tăng gấp đôi nguy cơ so với mức lý tưởng)

Bảng 4.4. Cholesterol

4.1.1.2.2. Lipoprotein

Lipoprotein là tập hợp những khối tạp gồm mỡ và đạm trong máu dùng cho việc chuyển tải cholesterol và triglycerides. Chất mỡ phospholipid bọc bên ngoài có khả năng hòa nước, chất mỡ nằm bên trong lõi có kèm chất apoprotein. Các mô trong có thể có thể nhận ra chất apoprotein và tiếp nhận lipoprotein.

Lipoprotein được chia nhiều loại tùy theo tỷ trọng (*density*):

- HDL_Cholesterol (*High Density Lipoprotein*): khoảng 1 phần 3 tổng số cholesterol được mang trong HDL. Giới y học thường cho rằng HDL thường đem cholesterol ra khỏi động mạch trở về gan và sau đó bài tiết ra

khỏi cơ thể. Một số khác cho rằng HDL "hốt" cholesterol ứ thừa trong các màng xơ vữa và làm chậm sự phát triển của những mảng này. Vì thế, HDL thường có mệnh danh là loại "Cholesterol có ích". HDL càng thấp thì cơ hội bị bệnh tim mạch càng cao, và ngược lại, HDL cao có thể làm giảm khả năng bị bệnh tim mạch.

Nồng độ mg/DL	**Giải thích**
Dưới 40 mg/DL (Nam) Dưới 50 mg/DL (Nữ)	HDL thấp (nguy cơ cao hơn)
Từ 40 đến 59 mg/DL	HDL trung bình (nguy cơ khá thấp)
Từ 60 mg/DL trở lên	HDL cao (nguy cơ rất thấp)

Bảng 4.5. High Density Lipoprotein

- ILD_Cholesterol (*Intermediate Density Lipoprotein*) .

- LDL_Cholesterol (*Low Density Lipoprotein*): Ngược lại với HDL, LDL có mệnh danh là "Cholesterol xấu". Khi có quá nhiều LDL, cholesterol bị đưa vào các mảng của động mạch, dần dần làm hẹp đường kính của mạch. Sau đó, kết hợp với các chất khác trong mà của thành động mạch tạo thành những mảng xơ vữa (*atherosclerosis*). Những mảng này có thể bị rạn nứt làm cho thành động mạch không được trơn tru. Khi chảy qua những chỗ "gồ ghề" này, dòng máu dễ bị hỗn loạn không đều, trì trệ và dễ đọn lại thành cục máu động (*thrombus*). Cục máu đông này có thể phát triển theo kiểu "phù sa bồi đắp", lớp lớp chồng lên nhau, có lúc đầy đủ để làm nghẽn động mạch. Nếu trường hợp này xảy ra trong động mạch vành tim thì kết quả là nghẽn mạch tim, gây chứng nhồi máu cơ tim. Trong trường hợp cục máu đông bị sút ra khỏi thành động mạch, trôi theo dòng máu cho đến khi kẹt vào một mạch có đường kính nhỏ làm nghẽn mạch ấy. Nếu chẳng may đấy là mạch dẫn máu của não thì kết quả là chứng tai biến mạch máu não.

Nồng độ mg/DL	Giải thích
Dưới 70 mg/DL (Nam)	Nguy cơ rất cao bị đau tim và tử vong do cơn đau tim
Dưới 100 mg/DL (Nữ)	Tối ưu đối với những người mắc bệnh tim hoặc tiểu đường
100 đến 129 mg/DL	Gần hoặc trên mức tối ưu
130 đến 159 mg/DL	Chạm ngưỡng cao
160 đến 189 mg/DL	Cao
Từ 190 mg/DL trở lên	Rất cao

Bảng 4.6. Low Density Lipoprotein

- VLDL_Cholesterol (*Very Low Density Lipoprotein*)

4.1.1.2.3. Triglyceride

Triglyceride hay còn gọi là chất béo trung tính, triacylglycerol, TAG hay triacylgleceride là 1 este có người gốc từ glyxêtin và 3 axit béo. Nó là một thành phần chính của dầu thực vật và mỡ động vật.

Ở cơ thể người, mức độ cao triglyceride trong mạch máu dẫn đến xơ vữa động mạch gây nguy cơ về bệnh tim mạch và đột quỵ. Tuy nhiên, ảnh hưởng tiêu cực của triglyceride đến việc nâng cao LDL, HDL đến nay vẫn chưa được xác định rõ ràng. Mối nguy hiểm có thể được cho là sự tương quan tỷ lệ nghịch giữa nồng độ triglyceride và nồng độ HDL.

Nồng độ mg/DL	Giải thích
Dưới 150 mg/DL	Bình thường, nguy cơ thấp
Từ 150 đến 199 mg/DL	Khá cao
Từ 200 đến 499 mg/DL	Cao
Trên 500 mg/DL	Rất cao: nguy cơ cao

Bảng 4.7. Trigleceride

4.1.1.3. Sinh hóa

STT	Sinh hóa	Đơn vị tính
1.	Glucose	mg/dL
2.	Urea	mg/dL

Bảng 4.8. Dữ liệu xét nghiệm sinh hóa

4.1.1.3.1. Glucose

Đặc điểm:

- Glucose được tạo thành từ 3 nguồn chính: thức ăn, do phân hủy glycogen, do quá trình tân tạo đường từ các thành phần khác. Glucose là nguồn năng lượng chủ yếu của não và cơ.
- Glucose huyết luôn hằng định do cơ thể điều hòa thần kinh – nội tiết. Các hormone điều hòa glucose huyết được phân thành hai nhóm đối lập: một bên là insulin làm giảm, một bên là những hormone làm tăng glucose huyết.

Ý nghĩa:

- Trị số bình thường: lúc đói 70 – 110 mg/dl; SI = 3.9 – 6.1 mmol/l.

- Tăng: hay gặp nhất là tăng đường huyết do đái tháo đường. Nồng độ glucose huyết lúc đói cao hơn 125 mg/dl (7.0 mmol/l) được coi là bệnh lý. Đường hết cao tới 290 – 310 mg/dl (16 – 17 mmol) có nguy cơ gây hôn mê đái tháo đường. Tuy nhiên không thể nêu lên một giới hạn cụ thể vì trị số này thay đổi khá nhiều với từng ca bệnh.
- Ngoài đái tháo đường, tăng đường huyết còn do một số bệnh nội tiết khác: hội chứng Cushing (cường năng vỏ thượng thân), tăng năng tuyết giấp.
- Tăng đường huyết cũng có thể do dùng một số loại thuốc: glucocorticoid, thuốc lợi tiểu thiazid, phenytoin…
- Giảm: hạ đường huyết dưới 45 mg/dl (2.5 mmol/l) cũng rất nguy hiểm. Nguyên nhân thường liên quan đến dùng insulin và các thuốc uống trong điều trị đái tháo đường.
- Hạ đường huyết còn do một số nguyên nhân khác như u tủy tạng, suy gan, thiểu năng tuyến yên, thiểu năng tuyết giáp, thiểu năng vỏ thượng thận.

4.1.1.3.2. Urea

Đặc điểm:

- Urea là sản phẩm thoái hóa của protein, được tạo thành ở gan thông qua chu trình Urea. Urea có thể khuếch tán dễ dàng qua phần lớn các màng tế bào và phân tách rộng khắp các dịch nội và ngoại bào trong cơ thể. Urea được đào thải chủ yếu qua thận, sau khi lọc qua cầu thận, một phần Urea được hấp thu ở ống thận. Ngoài ra còn được thải trừ một phần nhỏ qua mồ hôi và qua ruột

Ý nghĩa:

- Trị số bình thường: 20 – 40 mg/dl; SI 3.3 – 6.6 mmol/l

- Giảm: giảm Urea máu hiếm gặp, thường gặp ở giai đoạn cuối của thiếu năng gan do suy giảm tổng hợp Urea
- Tăng: Urea huyết tăng cao có thể là nguyên nhân trước thận, sau thận, hoặc tại thận.
  - Nguyên nhân trước thận như mất nước, nôn mửa, tia chảy, giảm lưu lượng máu, sốc, suy tim
  - Nguyên nhân sau thân như tắc đường tiết niệu (sỏi)
  - Nguyên nhân tại thân như viêm cầu thận cấp hoặc mạn, viêm ống thận cấp do nhiễm độc

4.1.1.4. Men gan

STT	Men gan	Đơn vị tính
1.	SGOT	U/L
2.	SGPT	U/L

Bảng 4.9. Dữ liệu xét nghiệm men gan

4.1.1.4.1. SGOT

Đặc điểm:

- SGOT (*Serum Glutamic Oxaloacetic Transaminase*) còn có tên khác là ASAT (*Aspartat Amino Transferase*). Đây là enzyme có vai trò vận chuyển nhóm amin. Enzym này có nhiều ở mô tin và gan, ở các mô khác ít gặp.

Ý nghĩa:

- Trị số bình thường: 0 – 35 U/l; SI 0 – 0.58 µkat/l
- Tăng:

- Nhồi máu cơ tim: SGOT là enzyme thứ hai tăng sớm trong huyết thanh sau nhồi máu cơ tim, tăng bắt đầu sau 6 – 8h, đặt đỉnh cao sau 24 giờ rồi bình thường sau 4 – 6 ngày.

- Tổn thương tế bào gan: SGOT tăng trong các bệnh có tổn thương tế bào gan, đặc biệt trong viêm gan virus hoặc do nhiễm độc. Trường hợp này SGOT và SGPT huyết thanh tăng sớm trước các biểu hiện lâm sàng gấp hàng chục lần bình thường. Trường hợp viêm gan mạn, xơ gan hay ứ mật, hoạt độ SGOP tăng vừa phải tùy theo mức độ tiêu hủy tế bào.

Nhiều thuốc có thể gây tăng SGOP vì gây tổn thương tế bào gan, ví dụ isoniazid, đặc biệt khi kết hợp với rifampicin. Khi tiếp tục uống thuốc mà enzyme vẫn tiếp tục tăng, ví dụ gấp hơn ba lần giới hạn cao của bình thường thì cần ngưng tạm thời hoặc vĩnh viễn thuốc đó.

4.1.1.4.2. SGPT

Đặc điểm:

- SGPT (*Serum Glutamic Pyruvic Transaminase*) còn có tên gọi khác là ALAT (*Alanin Amino Transferase*). Đây cũng là enzyme có vai trò chuyển vận nhóm amin. Enzym này chủ yếu tập trung ở tế bào nhu mô gan.

Ý nghĩa:

- Trị số bình thường: 0 – 35U/l; SI 0 – 0.58 µkat/l.

- Tăng:

  - Tổn thương tế bào gan: SGPT tăng chủ yếu trong các bệnh có tổn thương tế bào gan. Mặc dù cả hai enzyme SGOT và SGPT đều tăng trong các bệnh về gan nhưng SGPT được coi là enzyme đặc hiệu với gan hơn vì thường ít khi tăng trong các bệnh ngoài nhu mô gan.

4.1.1.5. Huyết đồ

STT	Mô tả	Đơn vị tính
1.	WBC	G/L
2.	LYM#	G/L
3.	MONO#	G/L
4.	GRAN#	G/L
5.	LYM%	%
6.	MONO%	%
7.	GRAN%	%
8.	HGB	g/dL
9.	RBC	T/L
10.	HCT	%
11.	MCV	fL
12.	MCH	Pg
13.	MCHC	g/dL
14.	RDW_CV	%
15.	PLT	G/L
16.	MPV	fL
17.	PDW	%

18.	PCT	%

Bảng 4.10 Dữ liệu huyết đồ

Huyết đồ còn được gọi là công thức máu, là một trong những xét nghiệm thường quy được sử dụng nhiều nhất trong các xét nghiệm huyết học cũng như xét nghiệm y khoa [3].

Trước đây công thức máu được thực hiện bằng dụng cụ đếm tay, để xác định số lượng của từng loại tế bào máu, ngày nay mẫu máu được đưa vào và nhờ các máy đếm tự động, do vậy việc thực hiện công thức máu trở nên đơn giản hơn nhiều.

Huyết đồ là xét nghiệm quan trọng cung cấp cho người thầy thuốc những thông tin hữu ích về tình trạng của bệnh nhân hoặc của người được xét nghiệm. Tuy nhiên phải biết rằng chỉ riêng công thức máu thì không thể cho phép đưa ra một chẩn đoán chính xác về nguyên nhân gây bệnh, nó chỉ có tính chất định hướng hoặc gợi ý mà thôi

4.1.1.5.1. WBC

WBC (*White Blood Cell*) là số lượng bạch cầu có trong một đơn vị máu. Giá trị bình thường của bạch cầu là 3200 – 9800/mm^3; SI 3.2 – 9.8 x 109/L.

Bạch cầu giúp cơ thể chống đỡ lại tác nhân gây bệnh bằng quá trình thực bào hoặc bằng quá trình miễn dịch.

Số lượng bạch câu trên 10000/mm^3 được coi là tăng bạch cầu. Khi có số lượng xuống dưới 3000/mm^3 coi là giảm bạch cầu.

Tăng bạch cầu gặp trong cách trường hợp:

- Trong đại đa số các bệnh nhiễm khuẩn gây mủ.
- Trong các bệnh nhiễm độc.

[3] http://vi.wikipedia.org/wiki/C%C3%B4ng_th%E1%BB%A9c_m%C3%A1u

- Khi có sang chấn thương tổn tế bào, sau phẫu thuật.
- Đặc biết, bạch cầu tăng rất cao trong bệnh ung thư dòng bạch cầu.

Giảm bạch cầu trong các trường hợp:

- Sốt rét.
- Thương hàn.
- Bệnh do virus.
- Chứng mất bạch cầu hạt, giảm sản hoặc tủy xương.

4.1.1.5.2. LYM#

LYM# (*Lymphocyte Count*) là số lượng bạch cầu lympho. Trị số bình thường 0.6 – 3.4 Giga/L.

4.1.1.5.3. MONO#

MONO# (*Monocyte Count*) là số lượng bạch cầu mono. Trị số bình thường 0.0 – 0.9 Giga/L.

4.1.1.5.4. LYM%

LYM% (*% lymphocytes*) tỷ lệ % bạch cầu lympho. Lympho là những tế bào có khả năng miễn dịch của cơ thể, chúng có thể trở thành những tế bào "nhớ" sau khi tiếp xúc với tác nhân gây bệnh và tồn tại cho đến khi tiếp xúc lần nữa với cùng tác nhân ấy, khi ấy chúng sẽ gây ra những phản ứng miễn dịch mạnh mẽ, nhanh và kéo dài hơn sơ với lần đầu. Trị số bình thường 17 – 48%.

Tăng trong nhiễm khuẩn mạn, chứng tăng bạch cầu đơn do nhiễm khuẩn và nhiễm virus khác, bệnh bạch cầu dòng lympho mạn, bệnh Hodgkin, viêm loét đại tràng, suy tuyến thượng thận, ban xuất huyết do giảm tiểu cầu tự phát.

Giảm trong hội chứng suy giảm miễn dịch mắc phải (AIDS), ức chế tủy xương do các hóa chất trị liệu, thiếu máu bất sản, các ung thư, các steroid, tăng chức năng vỏ thượng thận, các rối loạn thần kinh (bệnh xơ cứng rải rác, nhược cơ, hội chứng thần kinh ngoại biên do rối loạn tự miễn Guillain)

4.1.1.5.5. MONO%

MONO% (*% monocytes*) tỷ lệ % bạch cầu Mono. Mono bào là dạng chưa trưởng thành của đại thực bào trong máu vì vậy chưa có khả năng thực bào. Đại thực bào là những tế bào có vai trò bảo vệ bằng cách thực bào, khả năng này của nó mạnh hơn của bạch cầu đa trung tính. Chúng sẽ phân bố đến các mô của cơ thể, tồn tại đó hàng tháng, hàng năm cho đến khi được huy động đi làm các chức năng bảo vệ. Vì vậy mono bào sẽ tăng trong các bệnh nhiễm khuẩn mãn tính như lao, viêm vòi trứng mãn tính…Trị số bình thường 4 – 8 %.

Tăng trong các trường hợp bệnh nhiễm virus, nhiễm ký sinh trùng, nhiễm khuẩn, các ung thư, viêm ruột, bệnh bạch cầu dòng monocyte, u lympho, u tủy, sorcoidosis…

Giảm trong các trường hợp thiếu máu do bất sản, bệnh bạch cầu dòng lympho, sử dụng glucocorticoid.

4.1.1.5.6. HGB

HGB hoặc Hb (*Hemoglobin*) hay huyết tố cầu là một protein phức tạp chứa phần tử sắt có khả năng thu thập, lưu giữ và phóng thích Ôxy trong cơ thể động vật hữu nhũ và một số động vật khác. Trị số bình thường 12 – 16.5 g/dl

Tăng trong mất nước, bệnh tim và bệnh phổi.

Giảm trong thiếu máu, chảy máu và các phản ứng gây tan máu

4.1.1.5.7. RBC

RBC (*Red Blood Cell*) hay hồng cầu là loại tế bào máu có chức năng chính là hô hấp, chuyển chở HGB, qua đó đưa Ôxy từ phổi đến các mô. Hồng cầu có hình đĩa lõm hai mặt nên tỷ lệ giữa diện tích của màng bao bọc tế bào so với các thành phần chứa bên trong tế bào rất lớn. Hồng cầu cũng có thể thay đổi hình dạng khi đi qua các mau mạch. Trị số trung bình 3.8 – 5.8 Tera/l

Tăng: thiếu máu do nhiều nguyên nhân – có thể do giảm tổng hợp (suy tủy, rối loạn tổng hợp porphyrin,…), tăng phá hủy (thiếu máu, tan máu) hoặc do mất máu.

Giảm: Trong trường hợp mô bị thiếu Ôxy, sẽ có quá trình điều hòa kích thích tạo hồng cầu ở tủy xương. Nguyên nhân gây thiếu Ôxy ở mô có thể do sống ở vùng cao, suy tim, các bệnh đường hô hấp…và những nguyên nhân này có thể gây tăng hồng cầu thứ phát và số lượng hồng cầu có thể tăng đến 6 -8 triệu/mm^3. Bên cạnh đó, còn có các trường hợp tăng hồng cầu do bệnh lý, vì một số nguyên nhân nào đó, tủy xương sản xuất ra quá nhiều hồng cầu, trong trường hợp này số lượng bạch cầu và tiểu cầu đều tăng.

4.1.1.5.8. HCT

HCT (*Hematocrit*) hay dung tích hồng cầu, đây là phầm trăm thể tích của má mà các tế bào máu (chủ yếu là hồng cầu) chiếm. Trị số bình thường ở nam 39 – 49%, ở nữ 33 – 43%.

Tăng trong các rối loạn dị ứng, chứng tăng hồng cầu, hút thuốc lá, bệnh phổi tắc nghẽn mạn tính, bệnh mạch vành, ở trên núi cao, mất nước, chứng giảm lưu lượng máu.

Giảm trong chảy máu, tan máu, thiếu máu và thai nghén.

4.1.1.5.9. MCV

MCV (*Mean Corpuscular Volume*) hay thể tích trung bình của một hồng cầu. Trị số bình thường 85 – 95 fL.

MCV được tính bằng công thức: MCV = HCT / RBC.

Giá trị MCV cho phép phân biệt các loại thiếu máu

Thể tích fL	Giải thích
Dưới 85 fL	Thiếu máu hồng cầu nhỏ
Từ 85 đến 95 fL	Thiếu máu hồng cầu bình
Trên 95 fL	Thiếu máu hồng cầu đại

Bảng 4.11 Mean Corpuscular Volume

Tăng trong thiếu hụt vitamin B12, thiếu acid folic, bệnh gan, nghiện rượt, chứng tăng hồng cầu, suy tuyết giáp, bất sản tủy xương, xơ hóa tủy xương.

Giảm trong thiếu hụt sắt, hội chứng thalassemia và các bệnh hemoglobin khác, thiếu máu trong bệnh mãn tính, thiếu máu nguyên hồng cầu (*sideroblastic anemia*), suy thận mãn tính, nhiễm độc chì

4.1.1.5.10. MCH

MCH (*Mean Corpuscular Hemoglobin*) – Lượng HGB trung bình hồng cầu. MCH được tính theo công thức MCH = HGB / RBC. Trị số bình thường 26 – 32 pg.

Tăng trong thiếu máu sắc hồng cầu bình thường, chứng hồng cầu hình tròn di truyền nặng, sự có mặt của các yếu tố ngưng kết lạnh.

Giảm trong bắt đầu thiếu máu thiếu sát, thiếu máu nói chung, thiếu máu đang tái tạo

4.1.1.5.11. MCHC

MCHC (*Mean Corpuscular Hemoglobin Concentration*) – nồng độ HGB trung bình hồng cầu. MCHC được tính theo công thức: MCHC = HGB / HCT = MCH / MCV

Nồng đồ g/dL	Giải thích
Dưới 32 g/dL	Thiếu máu nhược sắc
Từ 32 đến 36 g/dL	Thiếu máu đẳng sắc
Trên 36 g/dL	Hiếm khi xảy ra vì khả năng bảo hòa của hồng cầu chỉ đến đây là hết.

Bảng 4.12 Mean Corpuscular Hemoglobin Concentration

Tăng trong thiếu máu sắc hồng cầu bình thường, chứng hồng cầu hình tròn di truyền nặng, sự có mặt của các yếu tố ngưng kết lạnh.

Giảm trong thiếu máu đang tái tạo, có thể bình thường hoặc giảm trong thiếu máu do giảm folate hoặc vitamin B12, xơ gan, nghiệm rượt

4.1.1.5.12. RDW_CV

RDW_CV hay RDW (*Red Distribution Width*) – độ phân bố hồng cầu. Nói lên sự thay đổi của kích thước hồng cầu. Con số càng lớn nói lên sự thay đổi kích thước hồng cầu càng nhiều. Trị số bình thường là 11 – 15%.

Tỷ lệ %	Giải thích
Dưới 11%	Hiếm gặp
Từ 11 – 15%	Hồng cầu kích thước đồng điều
Trên 15%	Hồng cầu to nhỏ không đầu đều

Bảng 4.13 Red Distribution Witdh

RDW nằm trong trị số bình thường và:

- MCV tăng khi thiếu máu bất sản, trước bệnh bạch cầu.

- MCV bình thường gặp trong thiếu máu trong các bệnh mãn tính, mất máu hoặc tan máu cấp tính, bệnh enzyme hoặc bệnh hemoglobin không thiếu máu.
- MCV giảm khi thiếu máu trong các bệnh mãn tính, bệnh thalassemia dị hợp tử.

RDW tăng và:

- MCV tăng khi thiếu hụt vitamin B12, thiếu hụt folate, thiếu máu tan huyết do miễn dịch, ngưng kết lạnh, bệnh bạch cầu lympho mạn.
- MCV bình thường khi thiếu sắt giai đoạn sớm, thiếu hụt vitamin B12 giai đoạn sớm, thiếu hụt folate giai đoạn sớm, thiếu mau do bệnh globin.
- MCV giảm khi thiếu sắt, sự phân mảnh hồng cầu, thalassemia

4.1.1.5.13. PLT

PLT (*Platelet Count*) – số lượng tiểu cầu cho biết số lượng tiểu cầu trong một đơn vi máu, đóng vai trò quan trọng trong quá trình đông máu.

Thể tích G/L	Giải thích
Dưới 150 G/L	Nguy cơ xuất huyết tăng cao
Từ 150 – 400 G/L	Bình thường
Trên 400 G/L	Nguy cơ đột quỵ, nhồi máu cơ tim…

Bảng 4.14 Platelet Count

Tăng trong những rối loạn tăng sinh tủy xương: chứng tăng hồng cầu, bệnh bạch cầu dòng tủy mạn, chứng tăng tiểu cầu vô căn, xơ hóa tủy xương, sau chảy máu, sau phẫu thuật cắt bỏ lách, chứng tăng tiểu cầu dẫn đến các bệnh viêm.

Giảm trong ức chế hay thay thế tủy xương, các chất hóa trị liệu, chứng phì đại lách, sự đông máu trong lòng mạch rải rác, các kháng thể tiểu cầu (ban xuất huyết do giảm tiểu cầu tự phát, sốt Dengue, ban xuất huyết sau truyền máu, giảm tiểu cầu do miễn dịch đồng loại ở trẻ sơ sinh)

4.1.1.5.14. MPV

MPV (*Mean Platelet Volume*) thể tích trung bình tiểu cầu. Trị số trung bình 6.5 – 11 fL

Tăng trong bệnh tim mạch, tiểu đường, hút thuốc lá, stress, nhiễm độc do tuyến giáp…

Giảm trong thiếu máu do bất sản, thiếu nguyên nhân hồng cầu khổng lồ, hóa trị liệu ung thư, bạch cầu cấp…

4.1.1.5.15. PDW

PDW (*Platelet Distribution Width*) – độ phân bố tiểu cầu. Trị số bình thường 6 -18%.

Tăng trong ung thư phổi, bệnh hồng cầu liềm, nhiễm khuẩn huyết gram dương, gram âm.

Giảm trong nghiện rượu.

4.1.1.5.16. PCT

PCT (*Plateletcrit*) – khối tiểu cầu. Trị số bình thường 0.1 – 0.5%

Tăng trong ung thư đại thực tràng.

Giảm trong nghiện rượu, nhiễm nội độc tố.

4.1.1.6. Điện giải

STT	Máu mỡ	Đơn vị tính
1.	Na	mmol/l
2.	K	mmol/l
3.	Cl	mmol/l
4.	Ca	mmol/l

Bảng 4.15 Dữ liệu xét nghiệm điện giải

4.1.1.6.1. Na

Na hay Natri có mặt chủ yếu ở dịch ngoại bào, cùng với Clo, Bicarnobat…Duy trì áp suất thẩm thấu cho dịch ngoại bào. Chuyển hóa Natri chịu ảnh hưởng của hormone steroid vỏ thượng thận. Natri trong tế bào luôn được đổi mới do sự trao đổi Natri giữa trong và tế bào. Trị số bình thường 135 – 145 mmol/l

Tăng trong ưu năng vỏ thượng thận, tăng aldosteron tiên phát, đái tháo nhạt, hôn mê trong tăng áp lực thẩm thấu trong đái tháo đường

Giảm trong mất muối nhiều qua đường tiêu hóa, nước tiểu, mồ hôi, thiểu năng vỏ thượng thận, tổn thương ống thận nặng, suy thận mạn.

4.1.1.6.2. K

K hay Kali được coi là ion chủ yếu trong khu vực tế bào, cùng với một số ion khác của nội bào tạo nên áp suất thẩm thấu cho nội bào. Kali đóng vai tròn quan trọng trong co sơ, dẫn truyền thần kinh, hoạt động enzyme, và chức năng mà tế bào…Trị số bình thường 3.5 – 4.5 mmol/l.

Nồng độ Kali cao gây ức chết dẫn truyền, ngừng tim ở thì tâm trương.

Nồng độ Kali thấp gây ngừng tim ở thì tâm thu.

Nồng độ Kali bất tường có ảnh hưởng đến điện thế của màng cơ tim, phản ánh qua điện tâm đồ.

Nồng độ Kali cao hay thấp đều làm ảnh hưởng đến sự co các cơ vân và cơ trơn, gây nên liệt mềm.

Kali tăng khi suy thận, từ tế bào ra: sốc phản vệ, chấn thương nặng, bỏng nặng, tiểu cơ vân…Nhiễm toan, tan máu, suy vỏ thượng thận.

Giảm Kali đưa vào ít (nhịn đói, nghiện rượu, truyền dịch kéo dài không có kali…), hấp thu kém, mất nhiều do đường tiêu hóa: nôn mửa, do thận…Bệnh liệu chu kì di truyền Westphal…

4.1.1.6.3. Cl

Cl hay Clo chủ yếu ở dịch ngoại bào, cùng với các ion khác Clo tạo nên áp suất thẩm thấu của cơ thể. Nhưng thay đổi của Clo thường đi kèm với sự thay đổi của Natri. Trị số bình thường 90 – 110 mmol/l

Tăng Clo do mất nước, ưu năng vỏ thượng thận, đái tháo nhạt, tăng áp lực thẩm thấu trong đái tháo đường.

Giả Clo do ăn nhạt, mất muối, thiểu năng vỏ thượng thận.

4.1.1.6.4. Ca

Ca hay Canxi ion hóa là một hợp chất Canxi trong đó Ca++ ion được hoạt động tối đa. Trị số bình thường 1.17 – 1.29 mmol/l.

Tăng trong ưu năng tuyến cận giáp, dùng nhiều vitamin D, ung thư (xương, vú, phế quản), đa u tủy xương.

Giảm trong thiếu năng tuyết cận giáp, gây có giật, thiếu vitamin D, còi xương, các bệnh về thận, viêm tụy cấp, thưa xương, loãng xương…

Lưu ý: Các trị số bình thường trên được thống kê dựa trên người Việt. Các trị số bình thường này còn thay đổi tùy theo máy làm xét nghiệm, theo lứa tuổi, theo chủng tộc của người được làm xét nghiệm [4] [5].

4.1.2. Kết quả đầu ra

STT	Kết quả chẩn đoán	Đơn vị tính
1.	Tiểu đường (Có/Không)	Không

Bảng 4.16. Dữ liệu phân lớp

Dựa trên 34 thuộc tính của 5 loại xét nghiệm và 1 bộ thông tin cá nhân của bệnh nhân, nhóm thực hiện sẽ cài đặt các giải thuật cơ bản trong Data mining để chẩn đoán cho những bệnh nhân có nguy cơ mắc bệnh tiểu đường đồng thời đưa ra được các tập luật hay các biểu hiện thường thấy trong các trường hợp bệnh nhân mắc bệnh tiểu đường thông qua các thuộc tính trên.

4.2. Phương pháp thực hiện

4.2.1. Tiền xử lý dữ liệu

4.2.1.1. Làm sạch dữ liệu

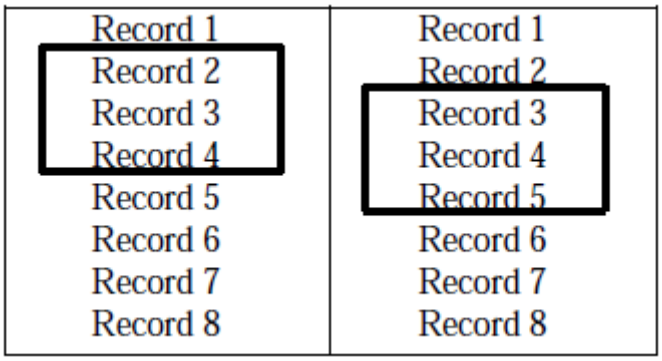
Trong quá trình làm sạch dữ liệu, nhóm thực hiện đã nhận thấy rằng việc bổ sung những dữ liệu bị thiếu là một điều không hợp lí. Vì dữ liệu xét nghiệm của mỗi bệnh nhân thường không liên quan đến nhau. Do đó trong quá trình làm sạch, nhóm đã chọn phương pháp duy nhất đó chính là loại bỏ những dòng dữ liệu bị thiếu hoặc nhiễu.

[4] http://www.thuocmoi.com.vn/index.php?option=com_k2&view=item&id=10214:y-ngh%C4%A9a-c%C3%A1c-th%C3%B4ng-s%E1%BB%91-sinh-h%C3%B3a-m%C3%A1u&catid=103:thuoc-khang-sinh&Itemid=24

[5] http://www.benhhoc.com/chu-de/1044-Phan-tich-huyet-do.html

Về phần những dòng dữ liệu trùng lặp. Nhóm đã dùng phương pháp *Làm sạch dữ liệu bán tự động* (*Semi – Automatic Data Cleaning*)**[**17**]** nhưng có phần đơn giản hơn vì nhóm thực hiện chỉ dùng hai thuộc tính để quyết định đó là "mã bệnh nhân" và "ngày khám bệnh".

Ý tưởng của phương pháp này là tạo một cửa sổ *k* lớn hơn 1. Chương trình sẽ so sánh các dòng dữ liệu trong cửa sổ đó để loại bỏ bớt đi dữ liệu dư thừa.



Hình 4.1. Mô hình cửa sổ của phương pháp làm sạch dữ liệu bán tự động

Nhóm thực hiện không những đã áp dụng phương pháp này mà còn cải tiến thêm bằng cách sắp xếp dữ liệu theo "mã bệnh nhân" và cửa sổ *k* sẽ là số lượng dòng dữ liệu có trùng "mã bệnh nhân".

4.2.1.2. Rời rạc hóa dữ liệu

Phương pháp rời rạc hóa dữ liệu đơn giản (*Simple Discretization Methods: Binning*)[2]

- Phân hoạch cân bằng theo bề rộng (*Equal-width distance partitioning*):
  - Chia miền giá trị: *N* đoạn dài như nhau
  - Miền giá trị từ A (nhỏ nhất) tới B (lớn nhất) → W = (B – A) / N
  - Đơn giản nhất nhưng không xử lý tốt khi dữ liệu không cân bằng (đều)

4.2.2. Cài đặt giải thuật

4.2.2.1. Naïve Bayes

4.2.2.1.1. Giới thiệu

Naïve Bayes là phương pháp phân lớp dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học [Mitchell, 1996] [Joachms, 1997] [Jason, 2001], được sử dụng lần đầu tiên trong lĩnh vực phân lớp bởi Maron vào năm 1996 [Maron, 1961] sau đó trở nên phổ biến trong nhiều lĩnh vực như trong các công cụ tìm kiếm [Rijsbergen, 1970], các bộ lọc email [Sahami, 1998]…**[**16**]**

4.2.2.1.2. Giải thuật

Giải thuật Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Áp dụng trong bài toán phân lớp, các dữ liệu gồm có:

- D: tập dữ liệu huấn luyện đã được vector hóa dưới dạng vector $x=(x_1,x_2,..,x_n)$
- C_i: phân lớp i, với i= {1, 2, …,m}
- Các thuộc tính độc lập điều kiện đôi một với nhau

Theo định lý Bayes:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Theo tính chất độc lập điều kiện:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Trong đó:

- $P(C_i|X)$ là xác suất thuộc phân lớp i khi biết trước mẫu X
- $P(C_i)$ là xác suất của phân lớp i
- $P(x_k|C_i)$ là xác suất của thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân lớp i

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu huấn luyện). Tính $P(C_i)$ và $P(x_k|C_i)$

Bước 2: Phân lớp vector $x= (x_1,x_2,..,x_n)$, ta cần tính xác suất thuộc từng phân lớp khi đã biết trước x. x được gán vào lớp có xác suất lớn nhất theo công thức.

$$\max_{C_i \in C}\left(P(C_i)\prod_{k=1}^{n}P(x_k|C_i)\right)$$

4.2.2.1.3. Đánh giá

Các giải thuật Bayes tính toán khả năng cụ thể đối với các giả định, chẳng hạn như bộ phân lớp Naïve Bayes là một trong các cách tiếp cận thực tế nhất đối với các kiểu học chắc chắn. Các nhà nghiên cứu đã chứng minh rằng bộ phân lớp Naïve Bayes là tương đối mạnh và trong một số trường hợp có còn cho kết quả tốt hơn. Bên cạnh đó, việc thiết kế một hệ thống phân lớp Naïve Bayes trong thực tế thường dễ dàng hơn so với phương pháp phân lớp khác vì tính đơn giản của nó và thời gian thực thi khá nhanh.

Các đặc điểm:

- Phương pháp học này tính xác suất rõ ràng cho các giả định bằng cách đếm tần suất của các kết hợp dữ liệu khác nhau trong tập dữ liệu huấn luyện để tính xác suất.

- Mỗi mẫu học quan sát được có thể giảm hoặc tăng xác suất dự đoán một giả định là đúng đắn. Điều này cho ta một cách tiếp cận thiết kế của một hệ thống học uyển chuyển hơn các giải thuật khác và loại trừ hoàn toàn một giả đính nếu nó có sự không nhất quán với bất kỳ mẫu đơn lẻ nào.
- Kiến thức biến trước có thể kết hợp với dữ liệu quan sát được để xác định khả năng của một giả định. Trong đó kiến thức biết trước có được bằng cách phân phối xác suất trên toàn bộ dữ liệu quan sát được cho mỗi giả định có thể có.
- Phương pháp Naïve Bayes cung cấp xác suất dự đoán. Mỗi thể hiện mới được phân lớp bằng cách kết hợp dự đoán của nhiều giả định.
- Phương pháp học Naïve Bayes có tính tăng trưởng:
  - Mỗi mẫu huấn luyện có thể tặng/ giảm dần khả năng đúng của một giả thiết.
  - Tri thức ưu tiên có thể kết hợp với dữ liệu quan sát được.
- Ngay cả khi các phương pháp Naïve Bayes khó trong tính toán chúng vẫn có thể cung cấp một chuẩn để tạo quyết định tối ưu so với các phương pháp khác.
- Các thuộc tính trong tập mẫu học phải là độc lập điều kiện.
- Độ chính xác của giải thuật phụ thuộc nhiều vào tập dữ liệu học ban đầu.

Ưu điểm:

- Dễ cài đặt
- Sử dụng được cho cả biến rời rạc và biến liên tục
- Thời gian huấn luyện ngắn

- Độ chính xác cao

Nhược điểm:

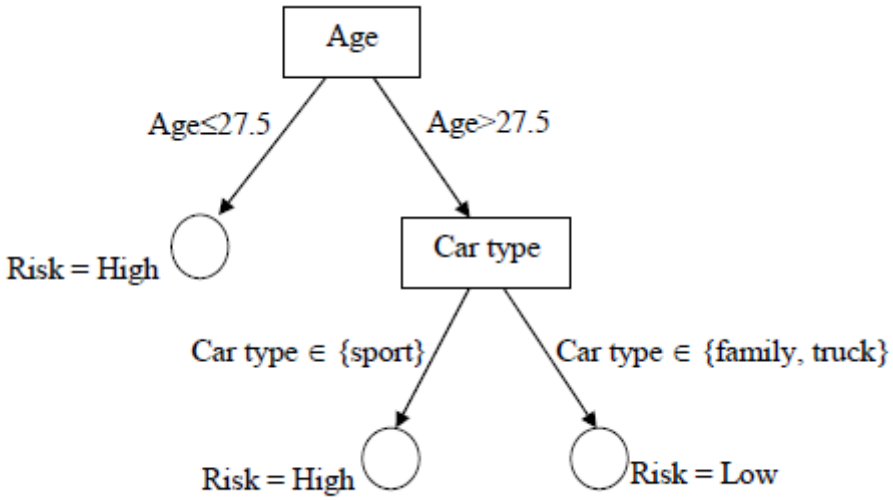
- Giả thiết về tính độc lập điều kiện của các thuộc tính làm giảm độ chính xác.

4.2.2.2. Decision Tree

4.2.2.2.1. Giới thiệu

Cây quyết định với những ưu điểm riêng của mình được đánh giá là một công cụ mạnh, phổ biến và đặc biệt thích hợp cho phân lớp dữ liệu.[16][5]

Cây quyết định là biểu đồ phát triển có cấu trúc dạng cây như hình sau:



Hình 4.2. Mô hình đơn giản của giải thuật cây quyết định

Trong cây quyết định gồm có:

- Gốc: là node trên cùng của cây
- Node trong: biểu diễn một kiểm tra trên một thuộc tính đơn (hình chữ nhật)
- Nhánh: biểu diễn các kết quả của kiểm tra trên node trong (mũi tên)

- Node lá: biểu diễn lớp hay sự phân phối lớp (hình tròn)

Để phân lớp mẫu dữ liệu chưa biết, giá trị các thuộc tính của mẫu được đưa vào kiểm tra trên cây quyết định. Mỗi mẫu tương ứng có một đường đi từ gốc đến lá và biểu diễn dự đoán giá trị phân lớp mẫu đó.

Giải thuật cây quyết định nhóm nghiên cứu có tên là C4.5 – sự kế thừa của giải thuật máy học bằng cây quyết định dựa trên nền tảng là kết quản nghiên cứu của Hunt và các cộng sự của ông trong nửa cuối thập kỷ 50 và nửa đầu những năm 60 (Hunt 1962). Phiên bản đầu tiên ra đời là ID3 (Quinlan 1979) là một hệ thống đơn giản ban đầu chứa khoảng 600 dòng lệnh Pascal, và tiếp theo là C4 (Quinlan 1987). Năm 1993, J.Ross Quinlan đã thừa kế các kết quả đó phát triển thành C4.5 với 9000 dòng lệnh C chứa trong một đĩa mềm. Mặc dù đã có một phiên bản phát triển từ C4.5 là C5.0 – một hệ thống tạo ra lợi nhuận từ Rule Quest Research, nhưng nhiều tranh luận, nghiên cứu vẫn tập trung vào C4.5 vì mã nguồn của nó là sẵn dùng. Trong các thuật toán phân lớp dữ liệu dựa trên cây quyết định, C4.5 là thuật toán hiểu quả và được dùng rộng rãi nhất trong các ứng dụng phân lớp với lượng dữ liệu nhỏ cỡ vài trăm ngàn bản ghi.

4.2.2.2.2. Giải thuật

Với những đặc điểm C4.5 là giải thuật phân lớp dữ liệu dựa trên cây quyết định hiệu quả và phổ biến trong những ứng dụng khai quá cơ sở dữ liệu nhỏ. C4.5 sử dụng cơ chế lưu trữ dữ liệu thường trú trong bộ nhớ, chính đặc điểm này làm cho C4.5 chỉ thích hợp với những cơ sở dữ liệu nhỏ, và cơ chế sắp xếp lại dữ liệu mỗi node trong quá trình phát triển cây quyết định C4.5 còn chứa một kỹ thuật cho phép biểu diễn tại mỗi node trong quá trình phát triển cây quyết định C4.5 còn chứa một kỹ thuật cho phép biểu diễn lại cây quyết định dưới dạng một danh sách sắp thứ tự các luật if-then. Kỹ thuật này cho phép làm giảm bớt kích thước tập luật và đơn giản hóa các luật và đơn giản hóa các luật mà độ chính xác so với những nhánh tương ứng cây quyết định là tương đương.

```
(1)   ComputerClassFrequency(T);
(2)      if OneClass or FewCases
return a leaf;
Create a decision node N;
(3)      ForEach Attribute A
ComputeGain(A);
(4)      N.test=AttributeWithBestGain;
(5)      if N.test is continuous
find Threshold;
(6)      ForEach T' in the splitting of T
(7)      if T' is Empty
Child of N is a leaf
else
(8)    Child of N=FormTree(T');
(9)      ComputeErrors of N;
return N
```

Hình 4.3. Mã giả của giải thuật C4.5

Phần lớn các hệ thống máy học đều cố gắng để tạo ra một cây càng nhỏ càng tốt vì phần lớn cây nhỏ hơn thì dễ hiểu hơn và dễ đặt được độ chính xác dự đoạn cao hơn. Do không thể đảm bảo sự cực tiểu của cây quyết định, C4.5 dựa vào nghiên cứu tối ưu hóa, và sự lựa chọn phân cách chia ra mà có *độ lựa chọn thuộc tính* đặt giá trị cực đại.

Hai độ đo thường được sử dụng trong C4.5 là *Information Gain* và *Gain Ratio*. *RF*(C_j,*S*) biểu diễn tần xuất (*Relative Frequency*) các *case* trong *S* thuộc về lớp C_j.

$$RF\ (Cj,\ S) = |Sj| \ /\ |S|$$

Với $|S_j|$ là kích thước tập các *case* có giá trị phân lớp là C_j. |*S*| là kích thước tập dữ liệu đào tạo.

Chỉ số thông tin cần thiết cho sự phân lớp: *I*(*S*) với *S* là tập cần xét sự phân lớp được tính bằng:

$$I(S) = -\sum_{j=1}^{x} RF(C_j, S)\log(RF(C_j, S)).$$

Sau khi *S* được phân chia thành các tập con S_1,S_2...S_t bởi *Test B* thì *Information Gain* được tính bằng:

$$G(S, B) = I(S) - \sum_{i=1}^{t} \frac{|S_i|}{|S|} I(S_i).$$

Test B sẽ được chọn nếu có *G*(*S*,*B*) đạt giá trị lớn nhất.

Tuy nhiên có một vấn đề khi sử dụng *G*(*S*,*B*) ưu tiên test có số lượng lớn kết quả, ví dụ *G*(*S*,*B*) đặt cực đại với test mà từng S_i chỉ chứa một *case* đơn. Tiêu chuẩn *Gain Ratio* giải quết được vấn đề này bằng việc đưa vào *thông tin tiềm năng* (*Potential Information*) của bản thân mỗi phân hoạch.

$$P(S, B) = -\sum_{i=1}^{t} \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right).$$

Test B sẽ được chọn nếu có tỉ số giá trị *Gain Ratio* = *G*(*S*,*B*)/*P*(*S*,*B*) lớn nhất.

Trong mô hình phân lớp C4.5, có thể dùng một trong hai loại chỉ số *Information Gain* hay *Gain Ratio* để xác định thuộc tính tốt nhất. Trong đó *Gain Ratio* là lựa chọn mặc định.

Mô tả cách tính *Information Gain* với dữ liệu rời rạc. Cho bảng dữ liệu sau:

rid	age	income	student	credit_rating	Class: buys_computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Hình 4.4. Ví dụ minh họa giải thuật Naïve Bayes

Trong tập dữ liệu trên: s_1 là tập những bản ghi có giá trị phân lớp là *yes*, s_2 là tập những bản ghi có giá trị phân lớp là *no*. Khi đó:

- $I(S)=I(s_1,s_2)=I(9,5)= -9/14*\log 29/14 – 5/14* \log 25/14 = 0.940$
- Tính G(S,A) với A lần lượt là từng thuộc tính:
  - A = age. Thuộc tính age đã được rời rạc hóa thành các giá trị <30, 30-40 và 40>.
    - Với age= "<30": I (S1) = (s11,s21) = -2/5log22/5 –3/5log23/5 =0,971
    - Với age =" 30-40": I (S2) = I(s12,s22) = 0
    - Với age =" >40": I (S3) = I(s13,s23) = 0.971. Σ |Si| / |S|* I(Si) = 5/14* I(S1) + 4/14 * I(S2) + 5/14 * I(S3) =0. 694
    - **Gain (S, age)** = I(s1,s2) – Σ |Si| / |S|* I(Si) = 0.246. Tính tương tự với các thuộc tính còn lại.
  - A = income: Gain (S, income) = 0.029
  - A = student: Gain (S, student) = 0.151

- A = credit_rating: Gain (S, credit_rating) = 0.048

- Thuộc tính *age* là thuộc tính có độ đo *Information Gain* lớn nhất. Do vậy *age* được chọn làm thuộc tính phát triển tại node đang xét

Với dữ liệu liên tục việc xử lý chúng đòi hỏi nhiều tài nguyên tính toán hơn dữ liệu rời rạc. Gồm các bước sau:

- Kỹ thuật *Quick sort* được sử dụng để sắp xếp các case trong tập dữ liệu đào tạo theo thứ tự tăng dần hoặc giảm dần các giá trị của thuộc tính liên tục V đang xét. Được tập giá trị V = {v1, v2, …, vm}

- Chia tập dữ liệu thành hai tập con theo ngưỡng θi = (vi + vi+1)/2 nằm giữa hai giá trị liền kề nhau vi và vi+1. Test để phân chia dữ liệu là test nhị phân dạng V <= θi hay V > θi. Thực thi test đó ta được hai tập dữ liệu con: V1 = {v1, v2, …, vi} và V2 = {vi+1, vi+2, …, vm}

- Xét (m-1) ngưỡng θi có thể có ứng với m giá trị của thuộc tính V bằng cách tính *Information gain* hay G*ain ratio* với từng ngưỡng đó. Ngưỡng có giá trị của *Information gain* hay *Gain ratio* lớn nhất sẽ được chọn làm ngưỡng phân chia của thuộc tính đó. Việc tìm ngưỡng (theo cách tuyến tính như trên) và sắp xếp tập training theo thuộc tính liên tục đang xem xét đôi khi gây ra thắt cổ chai vì tốn nhiều tài nguyên tính toán

4.2.2.2.3. Đánh giá

Ưu điểm:

- *Khả năng tạo ra các quy tắc có thể hiểu được*: Cây quyết định có khả năng sinh ra các quy tắc có thể chuyển đổi được sang dạng tiếng Anh, hoặc các câu lệnh SQL. Đây là ưu điểm nổi bật của kỹ thuật này. Thậm chí với những tập dữ liệu lớn khiến cho hình dáng cây quyết định lớn và phức tạp, việc đi theo bất cứ đường nào trên cây là dễ dàng theo nghĩa phổ biến và rõ

ràng. Do vậy sự giải thích cho bất cứ một sự phân lớp hay dự đoán nào đều tương đối minh bạch.

- *Khả năng thực thi trong những lĩnh vực hướng quy tắc*: Điều này có nghe có vẻ hiển nhiên, nhưng quy tắc quy nạp nói chung và cây quyết định nói riêng là lựa chọn hoàn hảo cho những lĩnh vực thực sự là các quy tắc. Rất nhiều lĩnh vực từ di truyền tới các quá trình công nghiệp thực sự chứa các quy tắc ẩn, không rõ ràng (*underlying rules*) do khá phức tạp và tối nghĩa bởi những dữ liệu lỗi (*noisy)*. Cây quyết định là một sự lựa chọn tự nhiên khi chúng ta nghi ngờ sự tồn tại của các quy tắc ẩn, không rõ ràng.

- *Dễ dàng tính toán trong khi phân lớp*: Mặc dù như chúng ta đã biết, cây quyết định có thể chứa nhiều định dạng, nhưng trong thực tế, các thuật toán sử dụng để tạo ra cây quyết định thường tạo ra những cây với số phân nhánh thấp và các test đơn giản tại từng node. Những test điển hình là: so sánh số, xem xét phần tử của một tập hợp, và các phép nối đơn giản. Khi thực thi trên máy tính, những test này chuyển thành các toán hàm logic và số nguyên là những toán hạng thực thi nhanh và không đắt. Đây là một ưu điểm quan trọng bởi trong môi trường thương mại, các mô hình dự đoán thường được sử dụng để phân lớp hàng triệu thậm trí hàng tỉ bản ghi.

- *Khả năng xử lý với cả thuộc tính liên tục và rời rạc*: Cây quyết định xử lý "tốt" như nhau với thuộc tính liên tục và thuộc tính rời rạc. Tuy rằng với thuộc tính liên tục cần nhiều tài nguyên tính toán hơn. Những thuộc tính rời rạc đã từng gây ra những vấn đề với mạng neural và các kỹ thuật thống kê lại thực sự dễ dàng thao tác với các *tiêu chuẩn phân chia* (splitting criteria) trên cây quyết định: mỗi nhánh tương ứng với từng phân tách tập dữ liệu theo giá trị của thuộc tính được chọn để phát triển tại node đó. Các thuộc tính liên tục cũng dễ dàng phân chia bằng việc chọn ra một số gọi là ngưỡng trong tập các giá trị đã sắp xếp của thuộc tính đó. Sau khi chọn

được ngưỡng tốt nhất, tập dữ liệu phân chia theo test nhị phân của ngưỡng đó.

- *Thể hiện rõ ràng những thuộc tính tốt nhất*: Các thuật toán xây dựng cây quyết định đưa ra thuộc tính mà phân chia tốt nhất tập dữ liệu đào tạo bắt đầu từ node gốc của cây. Từ đó có thể thấy những thuộc tính nào là quan trọng nhất cho việc dự đoán hay phân lớp.

Khuyết điểm:

- *Dễ xảy ra lỗi khi có quá nhiều lớp*: Một số cây quyết định chỉ thao tác với những lớp giá trị nhị phân dạng *yes/no hay accept/reject.* Số khác lại có thể chỉ định các bản ghi vào một số lớp bất kỳ, nhưng dễ xảy ra lỗi khi số ví dụ đào tạo ứng với một lớp là nhỏ. Điều này xảy ra càng nhanh hơn với cây mà có nhiều tầng hay có nhiều nhánh trên một node.
- *Chi phí tính toán đắt để huấn luyện*: Điều này nghe có vẻ mâu thuẫn với khẳng định ưu điểm của cây quyết định ở trên. Nhưng quá trình phát triển cây quyết định đắt về mặt tính toán. Vì cây quyết định có rất nhiều node trong trước khi đi đến lá cuối cùng. Tại từng node, cần tính một *độ đo (hay tiêu chuẩn phân chia)* trên từng thuộc tính, với thuộc tính liên tục phải thêm thao tác xắp xếp lại tập dữ liệu theo thứ tự giá trị của thuộc tính đó. Sau đó mới có thể chọn được một thuộc tính phát triển và tương ứng là một phân chia tốt nhất. Một vài thuật toán sử dụng tổ hợp các thuộc tính kết hợp với nhau có trọng số để phát triển cây quyết định. Quá trình cắt cụt cây cũng "đắt" vì nhiều cây con ứng cử phải được tạo ra và so sánh.
- *Không cho ra kết quả tốt với những bộ dữ liệu huấn luyện nhỏ.*

4.2.2.3. Suport Vector Machine

4.2.2.3.1. Giới thiệu

SVM (*Support Vector Machine*) phương pháp phân lớp rất hiệu quả được Vladimir N.Vapnik và Corinna Cortes giới thiệu vào năm 1995 để giải quyết nhận mẫu hai lớp sử dụng nguyên lý *Cực tiểu hóa Rủi ro Cấu trúc* (*Structural Risk Minimization*) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để nhận dạng và phân loại[10] .

Ý tưởng chính của SVM: là chuyển tập mẫu từ không gian biểu diễn Rn của chúng sang một không gian Rd có số chiều lớn hơn. Trong không gian Rd, tìm một siêu phẳng tối ưu để phân hoạch tập mẫu này dựa trên phân lớp của chúng, cũng có nghĩa là tìm ra miền phân bố của từng lớp trong không gian Rn để từ đó xác định được phân lớp của 1 mẫu cần nhận dạng.

4.2.2.3.2. Giải thuật

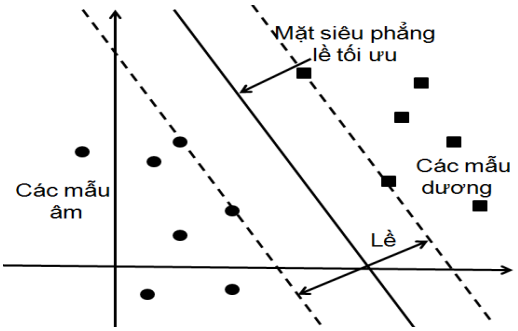
Đặc trưng cơ bản quyết định khả năng nhận diện của hệ thống là hiệu suất tổng quát hóa, hay là khả năng nhận diện những dữ liệu mới dựa vào những tri thức đã tích lũy được trong quá trình huấn luyện. Thuật toán huấn luyện được đánh giá là tốt nếu sau quá trình huấn luyện, hiệu suất tổng quát hóa của hệ thống nhận được cao. Hiệu suất tổng quát hóa phụ thuộc vào hai tham số là sai số huấn luyện và năng lực của máy học. Trong đó sai số huấn luyện là tỷ lệ lỗi phân loại trên tập dữ liệu huấn luyện. Còn năng lực của máy học được xác định bằng kích thước Vapnik-Chervonenkis (kích thước VC). Kích thước VC là một khái niệm quan trọng đối với một họ hàm phân tách (hay là bộ phân loại). Đại lượng này được xác định bằng số điểm cực đại mà họ hàm có thể phân tách hoàn toàn trong không gian đối tượng. Một hệ thống nhận diện tốt là hệ thống có năng lực thấp nhất (có nghĩa là đơn giản nhất) và đảm bảo sai số huấn luyện nhỏ. Phương pháp SVM được xây dựng dựa trên ý tưởng này. Xét bài toán phân loại đơn giản nhất phân loại hai phân lớp với tập dữ liệu mẫu:

$$\{(x_i, y_i) \mid i = 1, 2, \ldots, N, xi \in Rm \}$$

Trong đó mẫu là các vector đối tượng được phân loại thành các mẫu dương và mẫu âm:

- Các mẫu dương là các mẫu x_i thuộc lĩnh vực quan tâm và được gán nhãn $y_i = 1$

- Các mẫu âm là các mẫu x_i không thuộc lĩnh vực quan tâm và được gán nhãn $y_i = -1$



Hình 4.5. Mặt siêu phẳng tách các mẫu dương khỏi các mẫu âm

Trong trường hợp này, bộ phân loại SVM là mặt siêu phẳng phân tách các mẫu dương khỏi các mẫu âm với độ chênh lệch cực đại, trong đó độ chênh lệch – còn gọi là lề (margin) xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất (hình). Mặt siêu phẳng này được gọi là mặt siêu phẳng lề tối ưu.

Các mặt siêu phẳng trong không gian đối tượng có phương trình là $w^Tx + b = 0$, trong đó *w* là vector trọng số, *b* là độ dịch. Khi thay đổi w và b, hướng và khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi. Bộ phân loại SVM được định nghĩa như sau:

$$f(x) = sign(w_Tx + b)$$

Trong đó:

sign(z) = +1 nếu $z \geq 0$,

sign(z) = −1 nếu $z < 0$.

Nếu f(x) = +1 thì *x* thuộc về lớp dương (lĩnh vực được quan tâm), và ngược lại, nếu f(x) = −1 thì *x* thuộc về lớp âm (các lĩnh vực khác).

Máy học SVM là một họ các mặt siêu phẳng phụ thuộc vào các tham số w và b. Mục tiêu của phương pháp SVM là ước lượng w và b để cực đại hóa lề giữa các lớp dữ liệu dương và âm. Các giá trị khác nhau của lề cho ta các họ mặt siêu phẳng khác nhau, và lề càng lớn thì năng lực của máy học càng giảm. Như vậy, cực đại hóa lề thực chất là việc tìm một máy học có năng lực nhỏ nhất. Quá trình phân loại là tối ưu khi sai số phân loại là cực tiểu. Nếu tập dữ liệu huấn luyện là khả tách tuyến tính, ta có các ràng buộc sau:

$w^T x_i + b \geq +1$ nếu $y_i = +1$

$w^T x_i + b \leq -1$ nếu $y_i = -1$

Hai mặt siêu phẳng có phương trình là $w^T x_i + b = \pm 1$ được gọi là các mặt siêu phẳng hỗ trợ (các đường nét đứt trên hình). Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán quy hoạch toàn phương sau: Cực đại hóa:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j X_i^T X_j)$$

với các ràng buộc:

$\alpha_i \geq 0$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

trong đó các hệ số Lagrange α_i, i = 1, 2, ..., N, là các biến cần được tối ưu hóa. Vector w sẽ được tính từ các nghiệm của bài toán toàn phương nói trên như sau:

$$W = \sum_{i=1}^{N} \alpha_i y_i X_i$$

Để xác định độ dịch b, ta chọn một mẫu xi sao cho với $\alpha_i > 0$, sau đó sử dụng điều kiện Karush–Kuhn–Tucker (KKT) như sau:

$$\alpha_i [] y_i (w^T x_i + b) - 1] = 0$$

Các mẫu x_i tương ứng với $\alpha_i > 0$ là những mẫu nằm gần mặt siêu phẳng quyết định nhất (thỏa mãn dấu đẳng thức trong (2), (3)) và được gọi là các vector hỗ trợ. Những vector hỗ trợ là những thành phần quan trọng nhất của tập dữ liệu huấn luyện. Bởi vì nếu chỉ có các vector hỗ trợ, ta vẫn có thể xây dựng mặt siêu phẳng lề tối ưu như khi có một tập dữ liệu huấn luyện đầy đủ.

4.2.2.3.3. Đánh giá

Ưu điểm:

- *Giảm thiểu biến thiên trên các lỗi chính xác và làm cho hệ thống tin cậy hơn*. Nguồn gốc của SVM dựa trên sự chắc chắn về lỗi chính xác, có thể phân loại ngâu nhiên các mẫu đối tượng được chọn mà lỗi được sử dụng sao cho nhỏ nhất.
- *Giải quyết hiệu quả các bài toán về dữ liệu có số chiều lớn* (ảnh của dữ liệu biểu diễn gene, protein, tế bào).
- *Có khả năng giải quyết dữ liệu nhiễu và tách dời nhóm hoặc dữ liệu huấn luyện quá ít.*
- *SVM là một phương pháp phân lớp nhanh, có hiểu suất tổng hợp tốt và hiệu suất tính toán cao.*

Nhược điểm:

- *Tốc độ phân lớp chậm*: tùy thuốc vào số lượng vector thu được sau khi huấn luyện.
- *Quá trình huấn luyện đòi hỏi không gian nhớ lớn*: do đó việc huấn luyện đối với các bài toán có số lượng mẫu lớn sẽ gặp trở ngại trong vấn đề lưu trữ.
- *Giải quyết các bài toán phân loại nhiều lớp rất khó khăn*: có nhiều chiến lược được đề xuất mở rộng SVM cho bài toán phân loại nhiều lớp với

những điểm mạnh, yếu khác nhau tùy thuộc vào từng loại dữ liệu cụ thể, Cho đến nay, việc lựa chọn các chiến lược phân lớp vẫn thường được tiến hành trên cơ sở thực nghiệm.

4.2.3. Phương pháp đánh giá

Nhóm thực hiện sử dụng 4 thông số để đánh giá các giải thuật đã sử dụng là Precision, Recall, F-Measure và Accuracy[10] [4] .

	Kết quả mong đợi	
Kết quả thực nghiệm	TP (*True Positive*) Phần tử dương được phân loại dương	FP (*False Positive*) Phần tử âm được phân loại dương
	FN (*Fasle Negative*) Phần tử dương được phân loại âm	TN(*True Negative*) Phần tử âm được phân loại âm

Bảng 4.17. Các chỉ số liên qua đến Precision và Recall

Chỉ số Precision

$$Precision = TP / (TP + FP)$$

- *Định nghĩa:* Là số phần tử dương được phân loại dương trên tổng số các phần tử được phân loại dương.
- *Đánh giá dựa trên kết quả của Precision:*
  - Chỉ số Precision có giá trị từ 0 → 1
  - Giá trị Precision càng cao thể hiện xác suất để một kết quả được đưa ra là đúng cao.

Chỉ số Recall

$$Recall = TP / (TP + FN)$$

- *Định nghĩa*: là số phần tử dương được phân loại dương trên tổng số phần tử dương.
- *Đánh giá dựa trên kết quả của Recall*:
  - Chỉ số Recall có giá trị từ 0 → 1
  - Giá trị Recall càng cao thể hiện khả năng đưa ra một kết quả đúng của giải thuật càng cao.

Các chỉ số khác

$$\text{True Negative Rate} = TN / (TN + FP)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- *True Negative Rate* là số phần tử âm được phân loại âm trên tổng số các phần tử âm.
- *Accuracy* là số phần tử phân loại đúng trên tổng số phần tử - phản ánh độ chính xác của giải thuật.

Chỉ số F-Measure

$$\text{F-Measure} = 2.(\text{Precision}.\text{Recall}) / (\text{Precision} + \text{Recall})$$

- *Định nghĩa:* Là chỉ số nhằm đánh giá độ chính xác thông qua quá trình kiểm thử dựa trên sự xem xét đến hai chỉ số là Precision và Recall.
- *Đánh giá dựa trên F-Measure*: chỉ số F-Measure càng cao phản ánh động chính xác càng cao.

CHƯƠNG 5: KẾT QUẢ THỰC NGHIỆM

Nhằm đưa ra một đánh giá chính xác hơn về những giải thuật và phương pháp đã nghiêm cứu. Nhóm thực hiện đã áp dụng những hiểu biết về hệ hỗ trợ ra quyết định lâm sàng để xây dựng nên một hệ hỗ trợ ra quyết định trong khám chữa bệnh tiểu đường. Từ việc xây dựng nên hệ thống, nhóm có thể thu được mức độ hiệu quả của những giải thuật trong quá trình nghiệm thu.

5.1. Ứng dụng thực tế

Dựa trên kiến trúc của hệ hộ trợ ra quyết định, chương trình mà nhóm thực hiện đã xây dựng được chia làm 3 phần chính:

- Tiền xử lý dữ liệu
- Xây dựng mô hình
- Chẩn đoán

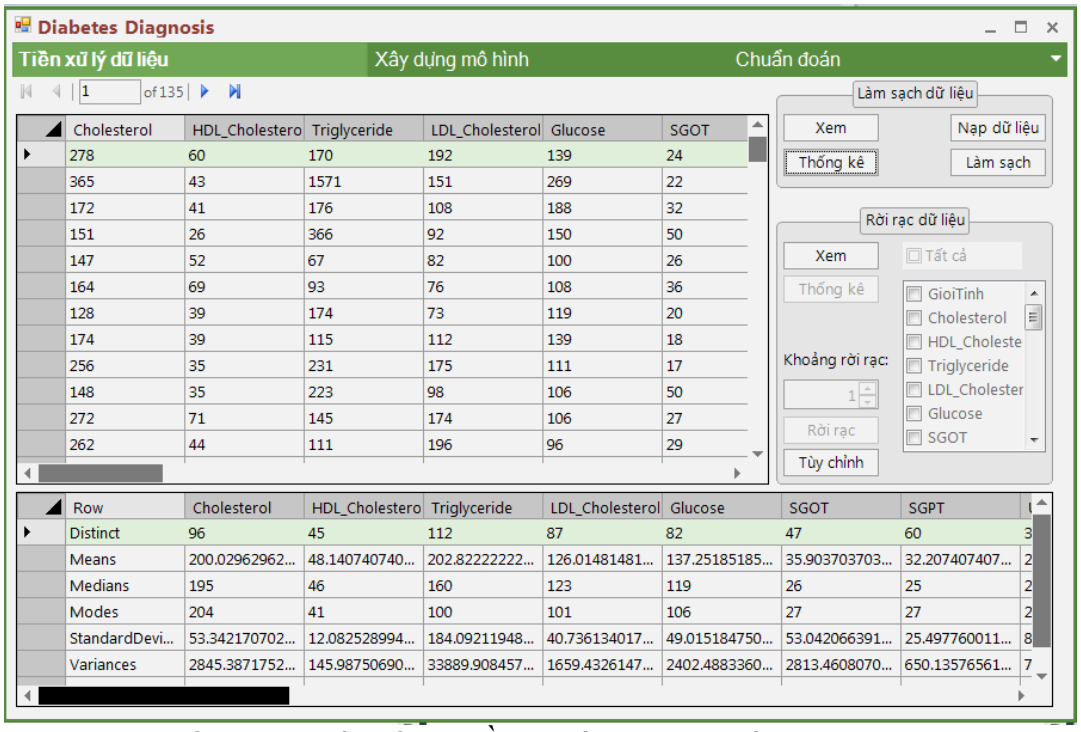
5.1.1. Tiền xử lý dữ liệu

Trong phần này, người dùng thực hiện hai công việc chính như sau:

- Làm sạch dữ liệu
- Rời rạc hóa dữ liệu

5.1.1.1. Làm sạch dữ liệu

Khi vừa khởi động chương trình, người dùng sẽ được tiếp xúc với giao diện:



Hình 5.1. Màn hình Tiền xử lý dữ liệu – Làm sạch dữ liệu

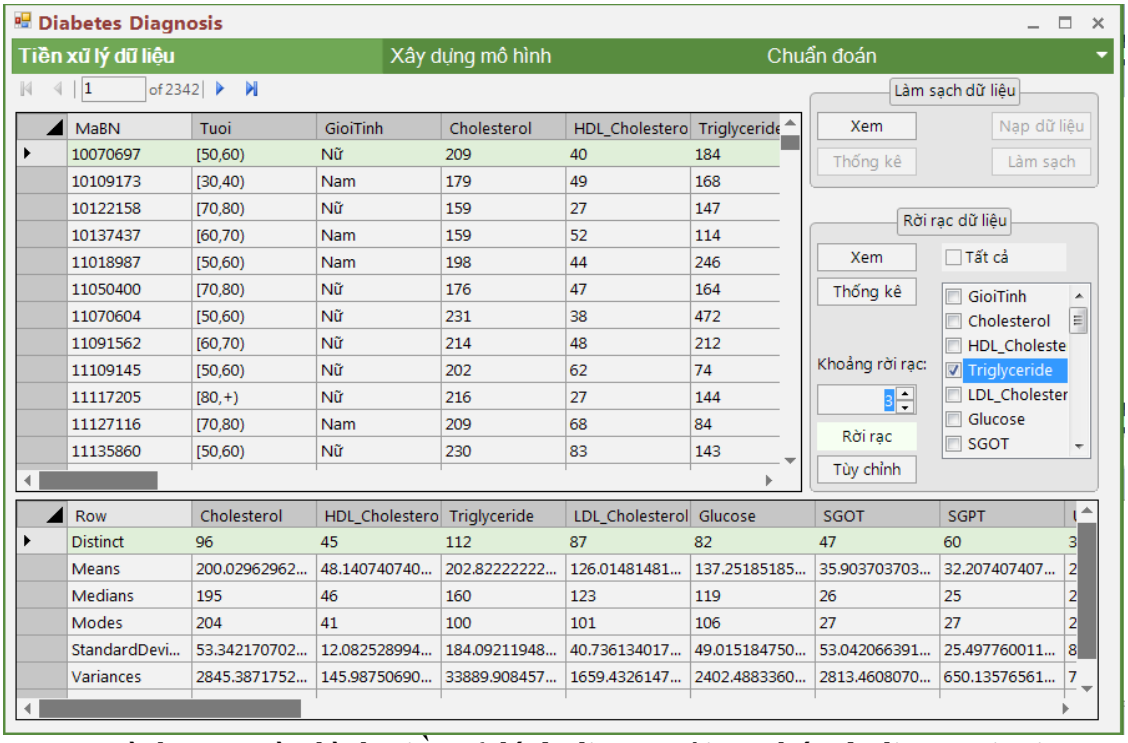
Màn hình sẽ tự động lấy cơ sở dữ liệu mà người dùng đã nạp trước đây và hiển thị, ta tạm gọi đây là *dữ liệu nguyên mẫu*. Trong màn hình này, người dùng có thể thực hiện các thao tác sau:

- *Xem*: Nút dùng để quay lại màn hình làm sạch dữ liệu khi người sử dụng đang ở màn hình rời rạc hóa dữ liệu.
- *Thống kê*: Nút dùng để thực hiện những thống kê về *Means, Medians, Standard Deviation, Variances…*trên *dữ liệu nguyên mẫu*.
- *Nạp dữ liệu*: Nút dùng để nạp một *dữ liệu nguyên mẫu* mới. Khi nạp dữ liệu mới thì người dùng phải có những lưu ý sau:
  - Dữ liệu đầu vào phải nằm trong tập tin Excel 2003 và các cột dữ liệu phải được sắp xếp theo yêu cầu của nhóm thực hiện đưa ra. Nếu xảy ra sai sót thì chương trình sẽ không thể thực hiện thao tạc nạp dữ liệu và tự động sử dụng lại bộ dữ liệu cũ.

- Một khi đã nạp dữ liệu thành công thì hệ thống sẽ tự động làm mới lại tất cả dữ liệu. Đồng nghĩa với việc là tất cả các mô hình mà người dùng đã xây dựng sẽ bị xóa bỏ

- *Làm sạch*: Nút dùng để thực hiện đồng thời hai thao tác xóa đi dữ liệu nhiễu và khởi động phương pháp *Làm sạch dữ liệu bán tự động*[17] để loại bỏ dữ liệu trùng lặp

5.1.1.2. Rời rạc hóa dữ liệu



MaBN	Tuoi	GioiTinh	Cholesterol	HDL_Cholestero	Triglyceride
10070697	[50,60)	Nữ	209	40	184
10109173	[30,40)	Nam	179	49	168
10122158	[70,80)	Nữ	159	27	147
10137437	[60,70)	Nam	159	52	114
11018987	[50,60)	Nam	198	44	246
11050400	[70,80)	Nữ	176	47	164
11070604	[50,60)	Nữ	231	38	472
11091562	[60,70)	Nữ	214	48	212
11109145	[50,60)	Nữ	202	62	74
11117205	[80,+)	Nữ	216	27	144
11127116	[70,80)	Nam	209	68	84
11135860	[50,60)	Nữ	230	83	143

Row	Cholesterol	HDL_Cholestero	Triglyceride	LDL_Cholesterol	Glucose	SGOT	SGPT
Distinct	96	45	112	87	82	47	60
Means	200.02962962...	48.140740740...	202.82222222...	126.01481481...	137.25185185...	35.903703703...	32.207407407...
Medians	195	46	160	123	119	26	25
Modes	204	41	100	101	106	27	27
StandardDevi...	53.342170702...	12.082528994...	184.09211948...	40.736134017...	49.015184750...	53.042066391...	25.497760011...
Variances	2845.3871752...	145.98750690...	33889.908457...	1659.4326147...	2402.4883360...	2813.4608070...	650.13576561...

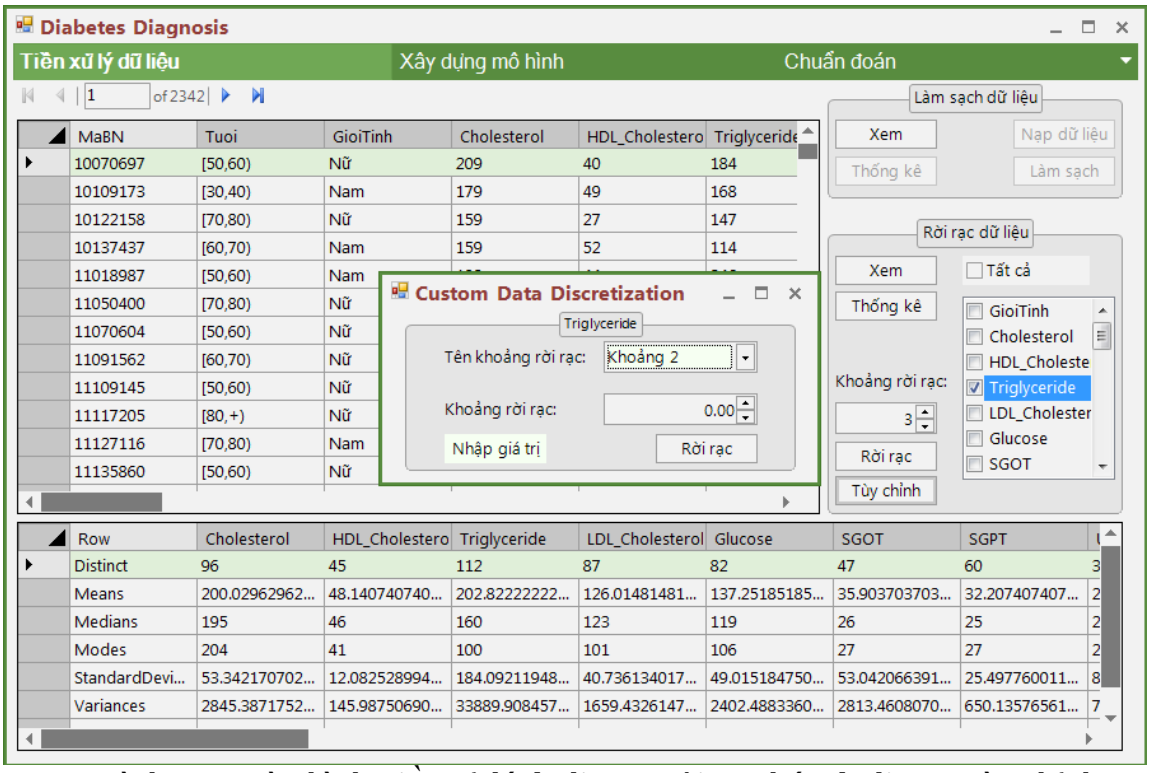
Hình 5.2. Màn hình Tiền xử lý dữ liệu – Rời rạc hóa dữ liệu – Binning

Màn hình sẽ lấy cơ sở dữ liệu mà người dùng đã thực hiện rời rạc hóa và hiển thị, ta tạm gọi là một *View*. Hiện màn hình đang thực hiện thao tác rời rạc hóa dữ liệu bằng phương pháp *Binning*, ngoài ra người dùng có thể thực hiện các thao tác khác như:

- *Thống kê*: Nút dùng để thống kê một thuộc tính trong *View*.

- *Rời rạc*: Nút dùng để thực hiện thao tác rời rạc hóa dữ liệu theo phương pháp *Binning* (hình trên). Với phương pháp này, người dùng chỉ việc nhập số khoảng cần rời rạc và chọn thuộc tính muốn rời rạc (có thể thực hiện trên nhiều thuộc tính).

- *Tùy chỉnh*: Nút dùng để thực hiện thao tác rời rạc hóa dữ liệu theo tùy chỉnh của người dùng. Với phương pháp này, người dùng cần phải nhập số khoảng cần, chọn thuộc tính và có thể tùy chỉnh số liệu giữa các khoảng. Tuy nhiên, việc tùy chỉnh phải tuân theo quy định của hệ thống đó là giá trị sau phải lớn hơn giá trị trước và mỗi lần chỉ có thể thực hiện thao tác trên một thuộc tính.

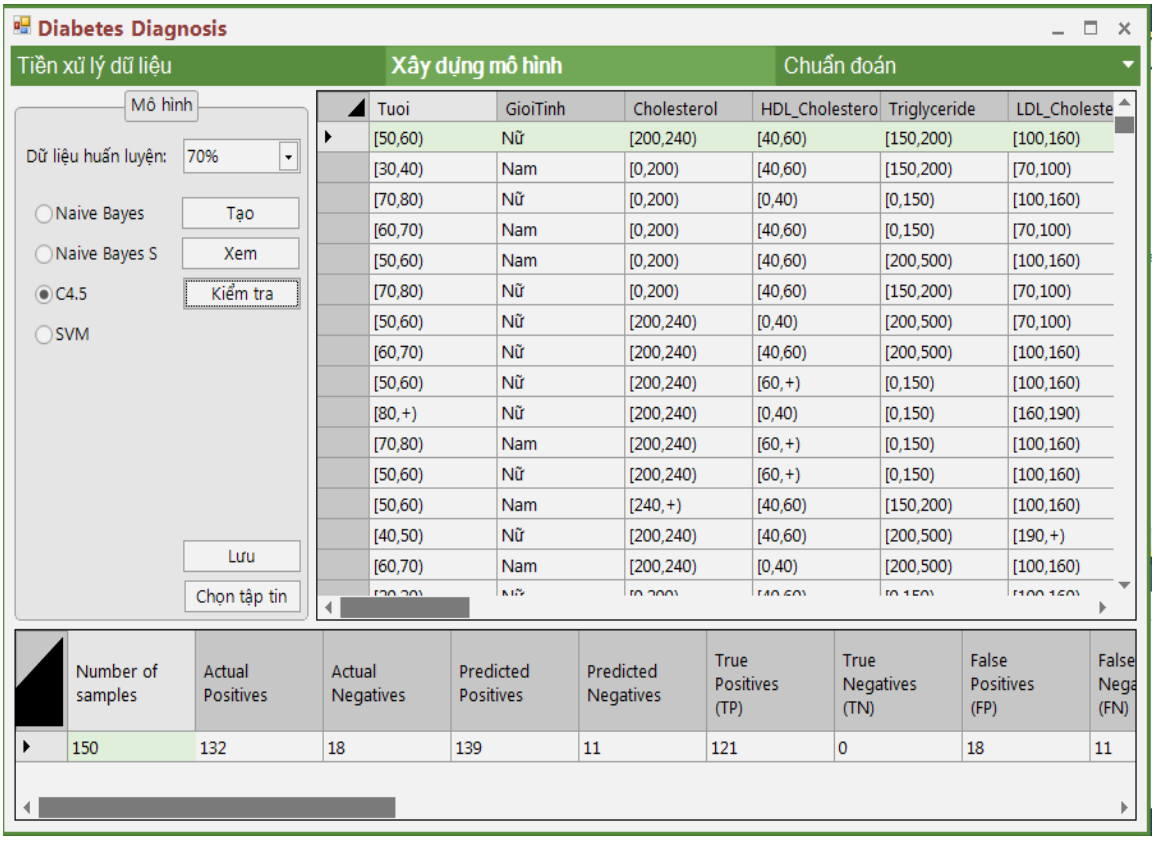


Hình 5.3. Màn hình Tiền xử lý dữ liệu – Rời rạc hóa dữ liệu – Tùy chỉnh

5.1.2. Xây dựng mô hình

Đây là một trong những phần quan trọng nhất của hệ thống, nó giữ chức năng chính là xây dựng, kiểm thử và lưu trữ những mô hình huấn luyện thu được từ việc cài đặt các giải thuật.

Kết hợp với Accord.NET Framework[6], nhóm thực hiện đã tự cài đặt giải thuật Naïve Bayes và cũng giải thuật này với 2 giải thuật khác là cây quyết định C4.5, SVM với sự trợ giúp của Accord.NET.



Hình 5.4. Màn hình Xây dựng mô hình

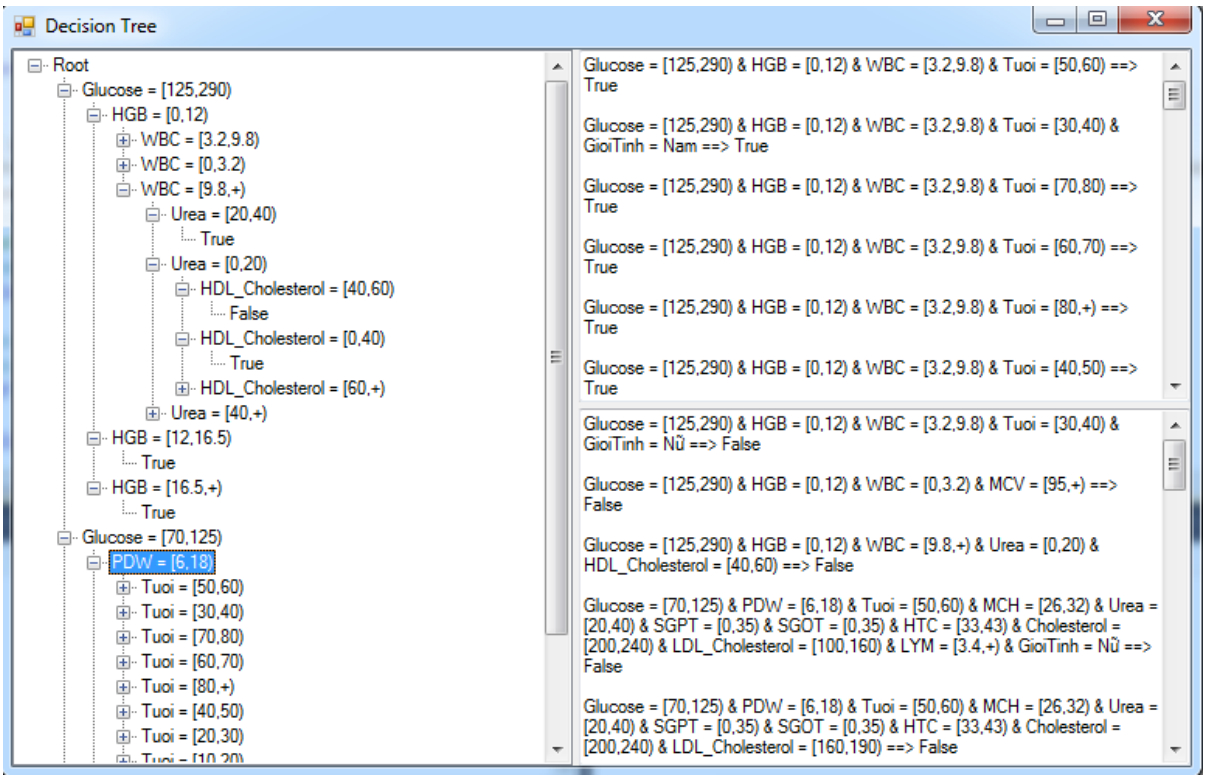
Màn hình hiển thị *View* cùng với 4 sự lựa chọn về giải thuật như sau:

- Naïve Bayes (Tự cài đặt)
- Naïve Bayes (Framework)
- C4.5 (Framework)
- SVM (Framework)

Đi kèm với giải thuật là các chức năng sau:

[6] http://accord-net.en.softonic.com/

- *Tạo*: Nút để tạo một mô hình mới. Người dùng cần nhập tỷ lệ dữ liệu huấn luyện. Ví dụ nếu người dùng nhập 70% nghĩa là 70% dữ liệu trong *View* sẽ được dùng để huấn luyện và 30% dữ liệu còn lại dùng để kiểm tra mô hình.
- *Xem*: Nút để xem một mô hình đã huấn luyện. Chắc năng này chỉ có áp dụng cho giải thuật C4.5. Vì chỉ có giải thuật này mới đưa ra được các tập luật.

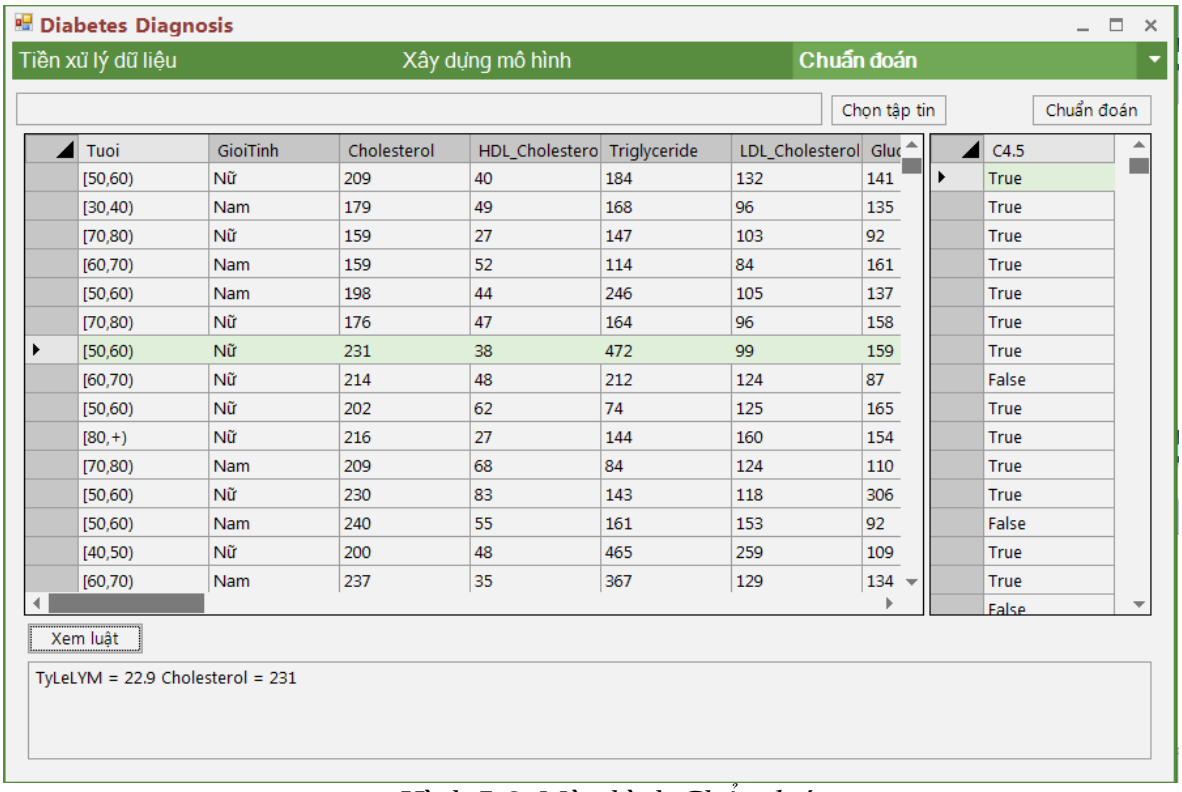


Hình 5.5. Màn hình Xem mô hình

- *Kiểm tra*: Nút để xem những đánh giá về mô hình thu được từ việc huấn luyện. Phương pháp đánh giá đã được nhóm nêu ở chương trên.
- *Lưu*: Nút đề thực hiện thao tác lưu lại mô hình đã huấn luyện.
- *Chọn tập tin*: Nút dùng để thực hiện thao tác nộp một mô hình đã được huấn luyện trước đây vào để sử dụng.

5.1.3. Chẩn đoán

Đây là chức năng mà mọi hệ hỗ trợ ra quyết định đều hướng tới. Phần cốt lõi của hệ thống sẽ đưa trao cho các y bác sĩ nhằm hỗ trợ họ ra quyết định.



Hình 5.6. Màn hình Chẩn đoán

Tại màn hình này, người dùng có thể thực hiện các chức năng sau:

- *Chọn tập tin*: Nút dùng để nạp dữ liệu xét nghiệm của bệnh nhân cần chẩn đoán. Chức năng này được thiết kế cho việc thực hiện chẩn đó cho nhiều bệnh nhân cùng một lúc. Tập tin nạp vào là Excel 2003 và số cột dữ liệu phải được sắp theo mẫu của nhóm thực hiện đề ra. Sau khi người dùng chọn tập tin, hệ thống sẽ thực hiện quá trình tiền xử lý dữ liệu sau đó hiển thị lên màn hình.

- *Chẩn đoán*: Nút dùng để thực hiện chẩn đoán những dòng dữ liệu đã nạp vào. Kết quả chẩn đoán được hiển thị ngay bên phải màn hình dữ liệu cần chẩn đoán.

- *Xem luật*: Nút dùng để hiển thị tập luật khớp nhất với dòng dữ liệu mà người dùng đã chọn.

5.2. Đánh giá kết quả

Khi muốn đánh giá một giải thuật, thì điều cần quan tâm là hiệu suất của giải thuật khi được áp dụng với dữ liệu mới từ đó có thể đánh giá được hiểu quả cũng như độ sai sót của giải thuật. Do đó với mỗi thao tác tạo mới mô hình luôn đi kèm với việc kiểm tra độ chính xác của mô hình đó. Ngoài ra, lượng dữ liệu nằm trong tập dữ liệu huấn luyện (*Training Dataset*) và tập kiểm thử (*Testing Dataset*) cũng tùy thuộc vào quyết định của người sử dụng. Dữ liệu kiểm thử dùng để đánh giá mức độ hiểu quả của mô hình và đồng thời nó cũng hoàn toàn độc với dữ liệu huấn luyện để có thể đưa ra một đánh giá khách quan về mô mình đang dùng.

Quá trình đánh giá giải thuật bao hàm cả việc đánh giá dữ liệu thu thập được. Sau đây, nhóm thực hiện sẽ tập trung vào đánh giá dữ liệu và kết quả kiểm thử

5.2.1. Đánh giá dữ liệu

Kết thúc qua trình thu thập dữ liệu, nhóm thực hiện đã thực hiện công việc thống kê vào đưa ra được những kết quả như sau:

Địa điểm	**Năm 2011**		**Năm 2012**	
	Dữ liệu xét nghiệm (dòng)	Dữ liệu khám bệnh (dòng)	Dữ liệu xét nghiệm (dòng)	Dữ liệu khám bệnh (dòng)
Bệnh viện Quận Thủ Đức	3290	65535	6791	40503
Bệnh viện Đa khoa khu vực Thủ Đức	1026	22263	758	24901

Bảng 5.1. Thống kê dữ liệu đã thu thập

Trong quá trình làm sạch dữ liệu, nhóm đã thực hiện làm một thống kê trên cơ sở dữ liệu được và đưa ra kết quả về hiện trạng bệnh tiểu đường tại 2 bệnh viện này như sau:

Tên thống kê	Tiểu đường (bệnh nhân)	Không tiểu đường (bệnh nhân)	Tổng
Số người mắc bệnh tiểu đường	4670	3421	**8091**
Số người mắc bệnh cao huyết áp (CHA)	882	46	**928**
Số người mắc bệnh rối loạn máu mỡ (RLMM)	2144	188	**2332**
Số người mắc bệnh về tim mạch	271	40	**311**
Số người mắc bệnh về gan	285	60	**345**
Số người mắc bệnh về thận	190	18	**208**
Số người mắc bệnh CHA và RLMM	644	12	**656**
Số người mắc bệnh CHA, RLMM và tim mạch	69	0	**69**
Số người mắc bệnh CHA, RLMM và gan	20	2	**22**

Bảng 5.2. Thống kê hiện trạng bệnh tiểu đường bằng dữ liệu thu thập

- Các bệnh về tim mạch chủ yếu là các bệnh: nhồi máu cơ tim, thiếu máu cơ tim cục bộ, sơ vữa động mạch…
- Các bệnh về gan: suy gan, tăng men gan, gan nhiễm mỡ…
- Các bệnh về thận: suy thận, suy thượng thận…
- Ngoài ra còn các bệnh khác như tai biến mạch máu não, suy van tĩnh mạch chi dưới, thiểu năng tuần hoàn não…

Sau khi thực hiện quá trình tiền xử lý dữ liệu, nhóm đưa ra một kết quả như sau:

- Tổng số dòng dữ liệu thu thập được: **11865 dòng** (dữ liệu xét nghiệm), **153202 dòng** (dữ liệu khám bệnh)
- Tổng số dòng dữ liệu có thể sử dụng được: **10081 dòng** (dữ liệu xét nghiệm), **106038 dòng** (dữ liệu khám bệnh)
- Tổng số dòng dữ liệu sau quá trình tiền xử lý: **2000 dòng** trong đó có **1774** bệnh nhân mắc bệnh tiểu đường và **226** bệnh nhân không mắc bệnh.

Đánh giá dữ liệu:

- Dữ liệu thu thập được tuy nhiều nhưng số dữ liệu có thể dùng cho việc cài đặt chương trình lại không được như mong muốn.
- Dữ liệu trùng lập khá nhiều ngoài ra dữ liệu nhiễu rất nhiều, đa số điều nằm ở bộ dữ liệu xét nghiệm. Điều này chứng tỏ rằng, bệnh nhân chỉ thực hiện các xét nghiệm khi phát hiện các biến chứng xuất hiện.

5.2.2. Đánh giá giải thuật

Để thực hiện việc kiểm thử, nhóm đã chia dữ liệu thành 3 bộ theo thứ tự 500, 500, 1000 dòng dữ liệu được chia theo tỉ lệ từ 70% dữ liệu huấn luyện và 30% dữ liệu kiểm thử.

- Bộ 1: 500 dòng dữ liệu trong đó có 443 dòng dữ liệu phân lớp “True” và 57 dòng dữ liệu phân lớp “False”
- Bộ 2: 500 dòng dữ liệu trong đó có 444 dữ liệu phân lớp “True” và 56 dòng “dữ liệu phân lớp “False”
- Bộ 2: 1000 dòng dữ liệu, trong đó có 887 dữ liệu phân lớp “True” và 113 dòng dữ liệu phân lớp “False”

Sau đây là kết quả mà nhóm thực hiện đã thu được.

5.2.2.1. Naïve Bayes

Giải thuật này được chia làm 2 phần là phần tự cài đặt và áp dụng Framework.

Sau đây là kết quả của giải thuật Naïve Bayes tự cài đặt

Các chỉ số đánh giá	Bộ 1	Bộ 2	Bộ 3
Số dòng dữ liệu kiểm thử	150	150	300
Mắc bệnh tiểu đường	132	134	264
Không mắc bệnh	18	16	36
Chẩn đoán mắc bệnh	149	143	292
Chẩn đoán không mắc bệnh	9	7	8
True Positive	126	129	261
True Negative	3	2	5
False Possive	15	14	31
False Negative	6	5	3
Precision	0.894	0.9	0.894
Recall	0.955	0.963	0.989
F – Measure	0.923	0.931	0.939
Accuracy	0.86	0.873	0.887

Bảng 5.3. Kết quả đánh giá giải thuật Naïve Bayes tự cài đặt

Kết quả của việc áp dụng Framework

Các chỉ số đánh giá	Bộ 1	Bộ 2	Bộ 3
Số dòng dữ liệu kiểm thử	150	150	300
Mắc bệnh tiểu đường	132	134	264
Không mắc bệnh	18	16	36
Chẩn đoán mắc bệnh	141	142	293
Chẩn đoán không mắc bệnh	9	8	7
True Positive	126	128	261
True Negative	3	2	4
False Positive	15	14	32
False Negative	6	6	3
Precision	0.894	0.9	0.891
Recall	0.955	0.955	0.989
F – Measure	0.923	0.928	0.937
Accuracy	0.86	0.867	0.883

Bảng 5.4. Kết quả đánh giá giải thuật Naïve Bayes áp dụng Framework

Nhận xét:

- Naïve Bayes là một trong những giải thuật có độ chính xác cao.
- Dễ dàng cài đặt.
- Chỉ cho kết quả chính xác nếu dữ liệu phân lớp đồng đều.

5.2.2.2. C4.5

Các chỉ số đánh giá	Bộ 1	Bộ 2	Bộ 3
Số dòng dữ liệu kiểm thử	150	150	300
Mắc bệnh tiểu đường	132	134	264
Không mắc bệnh	18	16	36
Chẩn đoán mắc bệnh	118	128	266
Chẩn đoán không mắc bệnh	32	22	34
True Positive	108	115	234
True Negative	8	3	4
False Positive	10	13	32
False Negative	24	19	30
Precision	0.915	0.898	0.879
Recall	0.818	0.858	0.886
F – Measure	0.864	0.878	0.883
Accuracy	0.773	0.787	0.793

Bảng 5.4. Kết quả đánh giá giải thuật C4.5

Sau đây là một số tập luật tiêu biểu thu được sau khi thực hiện huấn luyện dữ liệu bằng giải thuật C4.5

Phân lớp mắc bệnh tiểu đường
Glucose = [290,+)
Glucose = [125,290) & Tuoi = [80,+)
Glucose = [125,290) & Tuoi = [50,60)
Glucose = [125,290) & Tuoi = [30,40) & LDL_Cholesterol = [0,100)
Glucose = [125,290) & Tuoi = [70,80) & Urea = [20,40)
Glucose = [0,125) & PDW = [6,18) & Tuoi = [70,80)
Glucose = [0,125) & PDW = [6,18) & Tuoi = [20,30) & SGOT = [35,+)
Glucose = [0,125) & PDW = [6,18) & Tuoi = [80,+) & MCH = [26,32)

Bảng 5.5. Các tập luật tiểu biểu của phân lớp "True"

Phân lớp mắc bệnh tiểu đường
Glucose = [125,290) & Tuoi = [20,30) & Cholesterol = [200,240)
Glucose = [125,290) & Tuoi = [30,40) & LDL_Cholesterol = [130,160)
Glucose = [125,290) & Tuoi = [70,80) & Urea = [0,20) & Cholesterol = [200,240)
Glucose = [125,290) & Tuoi = [40,50) & WBC = [9.8,+) & GioiTinh = Nữ
Glucose = [0,125) & PDW = [6,18) & Tuoi = [30,40) & RBC = [5.8,+)
Glucose = [0,125) & PDW = [6,18) & Tuoi = [10,20) & HDL_Cholesterol = [60,+)

Bảng 5.6. Các tập luật tiêu biểu của phân lớp "False"

Nhận xét:

- Tuy là giải thuật phức tạp nhưng đã được hỗ trợ sẵn nên người dùng chỉ việc sử dụng.

- Dù độ chính xác không cao bằng các giải thuật khác nhưng lại có thể đưa ra được cái tập luật nhằm giải thích được lý do chẩn đoán. C4.5 nói riêng và cây quyết định nói chung là giải thuật thích hợp nhất cho việc ứng dụng khai phá dữ liệu trong y học.

5.2.2.3. SVM

Các chỉ số đánh giá	Bộ 1	Bộ 2	Bộ 3
Số dòng dữ liệu kiểm thử	150	150	300
Mắc bệnh tiểu đường	132	134	264
Không mắc bệnh	18	16	36
Chẩn đoán mắc bệnh	150	150	300
Chẩn đoán không mắc bệnh	0	0	0
True Positive	132	134	264
True Negative	18	0	0
False Positive	0	16	36
False Negative	18	0	0
Precision	0.88	0.893	0.88
Recall	1	1	0.88
F – Measure	0.936	0.943	1
Accuracy	0.88	0.893	0.936

Bảng 5.7. Kết quả đánh giá giải thuật SVM

Nhận xét:

- Thu được kết quả không mong muốn như trên là một thiếu sót không đáng có trong quá trình thu thập dữ liệu. Độ chệnh lệch giữa 2 phân lớp trong dữ liệu mà nhóm thực hiện sử dụng là quá cao.
- Tuy nhiên, qua đó ta có thể thấy được giải thuật SVM chỉ có thể hoạt động thật tốt khi các phân lớp dữ liệu là đồng đều với nhau.

5.3. Kết luận và hướng phát triển

5.3.1. Kết luận

Đề tài về "Hệ hỗ trợ ra quyết định lâm sàng" là một đề tài mới tại môi trường Việt Nam. Do đó, đi kèm với đề tài là những khó khăn vẫn còn tồn đọng trong các công trình nghiên cứu của các nước tiên tiến và những bất cấp khi được áp dụng vào các bệnh viện tại Việt Nam. Mỗi hệ hỗ trợ tuy thường hướng vào một chủ đề khác nhau nhưng mục đích chung vẫn là nhằm hỗ trợ cho các y bác sĩ có thêm nhiều lựa chọn trong quá trình khám chữa bệnh và năng cao chất lượng dịch vụ y tế cộng đồng. Sau sáu tháng nghiên cứu, nhóm thực hiện nhận thấy rằng đây là một đề tài hay và sẽ đạt được hiệu suất cao khi được ứng dụng vào thực tế.

Trong thời gian nghiên cứu đề tài, nhóm đã có thêm thời gian để củng cố và đồng thời bổ sung kiến thức. Do đó, qua khóa luận này nhóm đã đạt được khá nhiều thành công về mặt kiến thức lẫn thực tiễn. Tuy nhiên do thời gian đầu chưa nắm vững kiến thức nên vì một số lý do khách quan cũng như chủ quan nên đề tài vẫn có một số hạn chế nhất định.

5.3.1.1. Kết quả

Thực hiện khảo sát và thu thập dữ liệu lại các bệnh viện trên địa bàn Tp Hồ Chí Minh nhằm nằm được mô hình dữ liệu và hiện trạng bệnh tiểu đường của bệnh nhân.

Sau quá trình thu thập dữ liệu, nhóm thực hiện đã xây dựng được một bộ dữ liệu riêng dành cho giai đoạn cài đặt dữ liệu.

Nghiên cứu và cài cặt Hệ hỗ trợ ra quyết định lâm sàng dựa trên cấu trúc của Hệ hỗ trợ ra quyết định.

Nghiên cứu và cài đặt thành công các giải thuật thường dùng trong việc xây dựng các hệ hỗ trợ là Naïve Bayes, Decision Tree C4.5, Support Vector Machine…Riêng đối với giải thuật C4.5 đã được nhóm đặc biệt quan tâm vì giải thuật này có thể đưa ra được các mô hình mà cả bác sĩ lẫn người bệnh đều có thể hiểu được. Do đó khả năng áp dụng vào thực tế cũng như xây dựng những hệ hỗ trợ khác rất cao.

Kết quả đánh giá và kiểm thử giải thuật đã cho thấy kết quả mà nhóm đạt được khá phù hợp với các công trình nghiên cứu trước đây mà nhóm đã tham khảo.

5.3.1.2. Hạn chế

Do thời gian nghiên cứu và thực hiện đề tại còn giới hạn (thời gian thực hiện những nghiên cứu của các tác giả trước đây thường là 2 đến 3 năm) nên đề tài vẫn còn một số hạn chế nhất định.

Hạn chế đầu tiên đó là quá trình thu thập dữ liệu. Do không được tiếp xúc với cơ sở dữ liệu của bệnh viện mà phải thông qua nhân viên của bệnh viên nên bộ dữ liệu thu về không được đầy đủ như bộ dữ liệu đã được đề nghị. Ngoài ra, dữ liệu xét nghiệm của từng bệnh nhân lại không đầy đủ. Chỉ có khoảng 1/5 bệnh nhân thực hiện đủ hết các xét nghiệm. Thêm vào đó, là số lượng bệnh nhân thực hiện các xét nghiệm này đa số là đã mắc phải bệnh tiểu đường là cho tỉ lệ bệnh nhân mắc bệnh tiểu đường cao gấp nhiều lần so với số bệnh nhân không mắc bệnh trong dữ liệu thu thập được. Đây cũng chính là nguyên nhân dẫn đến thất bại của nhóm trong việc cài đặt giải thuật SVM.

Do khoảng thời gian eo hẹp nên nhóm đã không thể đưa chương trình mà nhóm xây dựng vào sử dụng tại các bệnh viện. Vì vậy, kết quả mà nhóm thực hiện thu được tuy độ chính xác khá cao nhưng chỉ là chủ quản. Đề tài của nhóm thực hiện đã mắc phải một hạn chế chung của tất cả các nghiên cứu trước đây đó là kết quả thực

nghiệm thu được hoàn toàn đi ngược với các cơ sở pháp y. Do đó việc triển khai dùng thử nghiệm tại các bệnh viện để thu được nhận xét từ các bác sĩ là rất cần thiết.

5.3.2. Hướng phát triển

Như đã trình bày ở phần trên, nhóm thực hiện đã thấy rằng còn rất nhiều điểm cần hoàn thiện trong chương trình và dữ liệu. Vì thế, sau đây là những điểm cần phát triển của đề tài:

- Xây dựng bộ dữ liệu phù hợp hơn với tình hình bệnh tiểu đường tại Việt Nam
- Nghiên cứu thêm các giải thuật có độ chính xác cao hơn và có khả năng giải thích những tập luật tìm thấy được chi tiết hơn.
- Cải thiện hệ thống mà nhóm đã xây hơn theo chuẩn của một hệ hỗ trợ ra quyết định.
- Phát triển thêm những tính năng cần thiết như: dự đoán thời gian mắc bệnh, dự đoán các biến chứng…

TÀI LIỆU THAM KHẢO

Tiếng Việt:

[1] PGS.TS Nguyễn Đạt Anh, DS.CKII Nguyễn Thị Hương (2012) el at, *Các xét nghiệm thường quy áp dụng trong thực hành lâm sàng*, nhà xuất bản Y Học

[2] TS. Võ Thị Ngọc Châu (2011), *Giáo trình điện tử ngành khoa học máy tính – Các vấn đều tiền xừ lý dự liệu*, Trường Đại học Bách Khóa Tp.HCM

[3] PGS.TS. Hoàng Thị Kim Huyền (2007) el at, *Hóa dược – Dược lý III (Dược lâm sàng)*, nhà xuất bản Y Học – Hà Nội

[4] Phạm Nguyên Khang, Trần Cao Đệ (2012), *Phân loại văn bản với máy học Vector hỗ trợ và cây quyết định*, Tạp chí Khoa học, Trường Đại học Cần Thơ, số 21a, 52-63

[5] Nguyễn Thị Thùy Linh (2005), *Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định*, Khóa luận tốt nghiệp, Trường Đại học Công Nghệ Hà Nội

[6] TS. Mai Văn Nam (2010), *Giáo trình nguyên lý thống kê kinh tế*, Nhà xuất bản Văn Hóa Thông Tin

[7] Văn Thế Thành, Trần Minh Bảo (2012), *Xây dựng hệ hỗ trợ ra quyết định chuẩn đoán bệnh*, Tạp chí Khoa học, Đại học Huế, Tập 74A, Số 5, 129-139

[8] Dương Thị Hiền Thanh (2008), *Kỹ Thuật Mạng Nơron và giải thuật di truyền trong khai phá dữ liệu và thử nghiệm ứng dụng*, Luận văn thạc sỹ, Trường Đại học Bách Khoa Hà Nội

[9] Huỳnh Tùng, Nguyễn Thị Kim Quy (2012), *Xây dựng hệ thống khuyến nghị lựa chọn sản phẩm*, Khóa luận tốt nghiệp, Trường Đại học Công Nghệ Thông Tin Tp. HCM

[10] Phan Thành Vũ, Ngô Đình Thế Hoàn (2012), *Hệ thống nhận diện tên riêng: Công cụ và Dữ liệu*, Khóa luận tốt nghiệp, Trường Đại học Công Nghệ Thông Tin Tp. HCM

Tiếng Anh:

[11**]** Doust Dominick, Walsh Zack (2011), *Data Mining Clustering: A Healthcare Application*, MCIS 2011 Proceedings, Paper 65

[12**]**Eta S.Berner el at (2007), *Clinical Decision Support Systems: Theory and Practice* (*Second Edition*), Springer

[13**]** Ian H. Witten, Eibe Frank, Mark A.Hall el at (2011), *Data Mining: Practical Machine Learning Tools and Techniques* (*Third Edition*), Elsevier

[14**]** Jiawei Han, Micheline Kamber el at (2006), *Data Mining: Concepts and Techniques*, Elsevier

[15**]** Ruben D.Canlas Jr. (2009), *Data Mining in Heathcare: Current Applications and Issues*, Master of Science in Information Technology, Carnegie Mellon University - Australia

[16**]** Xindong Wu, Vipin Kumar et al (2009), *The Top Ten Algorithms in Data Mining*, Taylor & Francis Group

[17**]** Wynne Hsu, Mong Li Lee, Bing Liu el at (2000), *Exploration Mining in Diabetic Patients Databases: Findings and Conclusions*, Subject Project, National University of Singapore