

## Lời cam đoan

Tôi cam đoan rằng luận văn này : “**ỨNG DỤNG KHAI PHÁ DỮ LIỆU XÂY DỰNG HỆ HỖ TRỢ CHẨN ĐOÁN Y KHOA**” là bài nghiên cứu của chính tôi. Ngoại trừ những tài liệu tham khảo được trích dẫn trong luận văn này, tôi cam đoan rằng toàn phần hay những phần nhỏ của luận văn này chưa từng được công bố hay được sử dụng để nhận bằng cấp ở những nơi khác.

Không có sản phẩm/nghiên cứu nào của người khác được sử dụng trong luận văn này mà không được trích dẫn theo đúng quy định.

Luận văn này chưa bao giờ được nộp để nhận bất kỳ bằng cấp nào tại các trường đại học hoặc cơ sở đào tạo khác.

Tp.HCM, ngày 10 tháng 04 năm 2014

Tác giả luận văn

A handwritten signature in blue ink, appearing to read "TÔNG ĐỨC PHONG".

Tông Đức Phong

## **Lời cảm ơn**

Lời cảm ơn đầu tiên tôi xin được gửi đến TS. Nguyễn Thanh Hiên – Giảng viên Trường Đại học Tôn Đức Thắng Tp.HCM, cảm ơn thầy đã truyền đạt kiến thức, kinh nghiệm và những gợi ý giúp tôi hoàn thành luận văn này.

Tiếp theo tôi muốn gửi lời cảm ơn đến thạc sĩ Dương Ngọc Hiếu – Giảng viên CNTT Trường Đại học Bách Khoa Tp.HCM đã giúp đỡ tôi rất nhiều trong cách thức thu thập số liệu và các kiến thức liên quan để hoàn thành luận văn.

Tôi cũng bày tỏ lòng biết ơn các điều dưỡng và các bác sĩ chuyên khoa tại Bệnh viện Bệnh Nhiệt Đới, Bệnh viện Nguyễn Tri Phương. Các anh, chị đã rất nhiệt tình giải thích vấn đề chuyên môn giúp tôi hoàn thành tốt công việc của mình.

Xin chân thành cảm ơn Ban Giám hiệu, quý Thầy Cô, cảm ơn sự hỗ trợ và giúp đỡ nhiệt thành của Phòng Quản lý Sau Đại học Trường Đại học Hồng Bàng Tp.HCM trong thời gian tôi thực hiện luận văn này.

Cuối cùng, chân thành cảm ơn người thân, bạn bè luôn bên cạnh động viên, hỗ trợ về mặt tinh thần để tôi vượt qua khó khăn và hoàn thành tốt luận văn.

# Tóm tắt luận văn

Ngành y tế và giáo dục luôn là vấn đề sống còn của bất kỳ quốc gia nào trên thế giới. Trong những năm gần đây, chính phủ Việt Nam đặc biệt đầu tư cho hai ngành mũi nhọn này thông qua các chính sách, nguồn vốn dành cho trang bị hạ tầng và nghiên cứu khoa học. Trong lĩnh vực nghiên cứu khoa học, càng ngày càng có nhiều công trình khoa học về y tế. Tuy nhiên các nghiên cứu khoa học về ứng dụng công nghệ thông tin để giải quyết các bài toán về y tế là không nhiều. Do đặc điểm về vị trí địa lý của Việt Nam là một nước nhiệt đới nên có rất nhiều loại bệnh liên quan đến sốt siêu vi trong đó sốt xuất huyết là bệnh rất nguy hiểm đồng thời chưa có vaccine chủng ngừa và chưa có thuốc đặc trị, vì vậy đề tài nghiên cứu các qui luật chẩn đoán bệnh sốt xuất huyết tại Việt Nam bằng kỹ thuật khai phá dữ liệu. Dựa vào các triệu chứng lâm sàng và cận lâm sàng có thể phân lớp bệnh của bệnh nhân nhằm giúp các bác sĩ chẩn đoán và điều trị tốt hơn cho bệnh nhân.

Nghiên cứu tiến hành theo 4 bước chính : (1) Tìm hiểu nghiệp vụ y tế liên quan đến bệnh sốt xuất huyết; (2) Thu thập và tiền xử lý dữ liệu; (3) Tìm hiểu bài toán phân lớp trong khai phá dữ liệu, lựa chọn thuật toán phù hợp với yêu cầu bài toán đặt ra và dữ liệu thu thập được; (4) Hiện thực chương trình máy tính và đánh giá ý nghĩa thực tiễn.

Ngoài ra đề tài cũng đề xuất một phương pháp phối hợp giữa các chuyên gia của lĩnh vực Công nghệ thông tin và Y tế để xây dựng mô hình hỗ trợ chẩn đoán cho các loại bệnh khác nhau nhằm hỗ trợ các tuyến y tế vùng sâu vùng xa, những nơi chăm sóc sức khỏe ban đầu còn thiếu về năng lực chuyên môn lẫn trang thiết bị.

# **Abstract**

The medical branch and education are always the principle problems of every countries in the world. In the recent years, Vietnamese government has specially invested for these two main areas through capital and policy for equipping infrastructure and studying science. In the science, there have been more and more researches about medicine. However, there are not many science researches of applying communication to solve medical problems. In Vietnam, because of the geography position of a tropical country, there are so many diseases related to ultravirous fever, such as petechial fever – a very dangerous disease. This subject studies the laws of diagnosing the petechial fever through techniques of discovering data. Based on the clinical signs and near clinical signs, we can subclass diseases of the patients to help the doctors diagnose and treat them better.

This research follows four main stages : First, finding out the medical specialist skills relating to petechial fever. Next, collecting and pre-processing the data. Then learning the “ math of subclassing ” in discovering data to choose the algorithm which is suitable to the inquiries and the collected data. Finally, performing the computer program and evaluating reality meanings.

Besides this subject also puts forward a method of co-ordinating the communication experts and medical experts to build a model which can help the doctors in diagnosing different diseases in order to help medical branches in rural and remote areas where there are still lack of ability and medical equipment for the first aid.

# Mục lục

Lời cam đoan.....	i
Lời cảm ơn .....	ii
Tóm tắt luận văn.....	iii
Abstract .....	iv
Mục lục .....	v
Danh mục chữ viết tắt.....	viii
Danh mục hình .....	ix
Danh mục bảng .....	xi
Danh mục công thức.....	xii
<b>Chương 1. TỔNG QUAN ĐỀ TÀI .....</b>	<b>1</b>
1.1. Đặt vấn đề .....	1
1.2. Cơ sở hình thành đề tài.....	2
1.3. Một số kết quả nghiên cứu trong và ngoài nước.....	2
1.3.1. Kết quả nghiên cứu trên thế giới .....	2
1.3.2. Kết quả nghiên cứu trong nước .....	2
1.4. Mục tiêu luận văn.....	3
1.5. Đối tượng và phương pháp nghiên cứu .....	3
1.6. Ý nghĩa của đề tài.....	3
1.6.1. Ý nghĩa khoa học .....	3
1.6.2. Ý nghĩa thực tiễn .....	4
1.7. Bố cục luận văn.....	4
<b>Chương 2. CƠ SỞ LÝ THUYẾT.....</b>	<b>6</b>

2.1.	Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining).....	6
2.1.1.	Khái niệm về khai phá dữ liệu.....	6
2.1.2.	Các giai đoạn của quá trình khai phá dữ liệu [4]: .....	6
2.2.	Tổng quan về hệ hỗ trợ ra quyết định .....	8
2.3.	Bài toán Phân lớp trong Khai phá dữ liệu .....	8
2.3.1.	Khái niệm về phân lớp .....	8
2.3.2.	Quá trình phân lớp dữ liệu .....	9
2.3.3.	Phân lớp dữ liệu bằng cây quyết định.....	12
2.3.4.	Đánh giá hiệu quả phân lớp.....	16
2.3.5.	Thuật toán C4.5 xây dựng cây quyết định .....	17
2.4.	Cơ sở dữ liệu Y khoa.....	21
2.4.1.	Sơ lược bệnh Sốt xuất huyết .....	21
2.4.2.	Diễn biến lâm sàng bệnh sốt xuất huyết dengue [19].....	22
2.4.3.	Chẩn đoán [19] .....	24
<b>Chương 3.</b>	<b>XÂY DỰNG HỆ HỖ TRỢ CHẨN ĐOÁN Y KHOA .....</b>	<b>26</b>
3.1.	Cơ sở dữ liệu xây dựng mô hình.....	26
3.1.1.	Kho chứa dữ liệu bệnh án điện tử.....	27
3.1.2.	Tiền xử lý dữ liệu .....	30
3.1.3.	Phân tích dữ liệu bệnh án điện tử .....	33
3.1.4.	Các qui luật chẩn đoán .....	35
3.1.5.	Bệnh án mẫu .....	36
3.1.6.	Chẩn đoán.....	36
3.2.	Xây dựng ứng dụng .....	36
3.2.1.	Giới thiệu chương trình.....	36

3.2.2. Cách thức vận hành chương trình.....	37
<b>Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ .....</b>	<b>55</b>
4.1. Thủ nghiệm.....	55
4.1.1. Thủ nghiệm tập dữ liệu với ít thuộc tính: .....	55
4.1.2. Thủ nghiệm với tập dữ liệu đầy đủ thuộc tính.....	56
4.2. Đánh giá.....	61
<b>Chương 5. TỔNG KẾT .....</b>	<b>62</b>
5.1. Kết luận.....	62
5.2. Hạn chế của đề tài .....	63
5.3. Hướng phát triển .....	63
<b>TÀI LIỆU THAM KHẢO.....</b>	<b>64</b>
PHỤ LỤC 1.....	68
PHỤ LỤC 2.....	72

# Danh mục chữ viết tắt

<b>KPDL</b>	(Data mining)	:	Khai phá dữ liệu
<b>DSS</b>	(Decision support system)	:	Hệ hỗ trợ ra quyết định
<b>CNTT</b>		:	Công nghệ thông tin
<b>IT</b>	(Information technology)	:	Công nghệ thông tin
<b>CSDL</b>		:	Cơ sở dữ liệu
<b>SXH</b>		:	Sốt xuất huyết
<b>HCT</b>	(Hematocrit)	:	Dung tích hồng cầu
<b>PLT</b>	(plaquette)	:	Tiêu cầu
<b>WBC</b>	(White blood cell)	:	Bạch cầu
<b>BS</b>		:	Bác sĩ
<b>BMI</b>	(Body mass index)	:	Chỉ số khối của cơ thể
<b>WHO</b>	(World Health Organization):	Tổ chức Y tế Thế giới	
<b>BVBND</b>		:	Bệnh viện bệnh Nhiệt Đới
<b>ĐHYKPNT</b>		:	Đại học Y khoa Phạm Ngọc Thạch

# **Danh mục hình**

Hình 2.1 : Kết quả quá trình phân lớp .....	9
Hình 2.2 : Xây dựng mô hình phân lớp .....	10
Hình 2.3 : Bước phân lớp.....	11
Hình 2.4 : Mô hình cây quyết định trong phân lớp dữ liệu về thời tiết .....	13
Hình 3.1 : Mô hình xây dựng giải pháp hỗ trợ chẩn đoán bệnh .....	26
Hình 3.2 : Tập dữ liệu thu thập được.....	28
Hình 3.3 : Tập dữ liệu huấn luyện đưa vào hệ thống .....	33
Hình 3.4 : NS1 được chọn vì có độ đo tốt nhất.....	35
Hình 3.5 : Màn hình khởi động chương trình .....	37
Hình 3.6 : Màn hình chọn file dữ liệu .....	37
Hình 3.7 : Màn hình chọn dữ liệu từ kho dữ liệu.....	38
Hình 3.8 : Màn hình chọn bảng dữ liệu .....	38
Hình 3.9 : Màn hình chọn thuộc tính.....	39
Hình 3.10: Màn hình tạo cây quyết định .....	40
Hình 3.11 : Màn hình tạo cây quyết định .....	41
Hình 3.12: Màn hình lấy cây đã lưu dạng xml.....	42
Hình 3.13: Màn hình thống kê tập luật của tập dữ liệu .....	42
Hình 3.14: Màn hình thống kê 10 luật.....	43

Hình 3.15 : Mô hình kiểm tra kết quả.....	44
Hình 3.16 : Màn hình kiểm tra kết quả.....	45
Hình 3.17 : Kết quả kiểm tra 30% dữ liệu.....	45
Hình 3.18 : Màn hình lấy file dữ liệu kiểm tra .....	46
Hình 3.19 : Màn hình lấy dữ liệu từ file kiểm tra .....	47
Hình 3.20 : Màn hình kết quả kiểm tra từ file Excel.....	48
Hình 3.21 : Màn hình kiểm tra chéo (Cross validation).....	49
Hình 3.22 : Màn hình kết quả kiểm tra chéo (Cross validation).....	50
Hình 3.23 : Màn hình chẩn đoán bệnh.....	51
Hình 3.24 : Kết quả chẩn đoán từ cây có sẵn.....	52
Hình 4.1 : Hình vẽ SXH Dengue có dấu hiệu cảnh báo .....	58
Hình 4.2 : Hình vẽ SXH Dengue.....	59
Hình 4.3 : Hình vẽ SXH Dengue nặng .....	60

# **Danh mục bảng**

Bảng 2.1 : Bảng dữ liệu về thời tiết .....	13
Bảng 2.2 : Ví dụ về thời tiết trong 3 ngày.....	14
Bảng 2.3 : Kết quả phân lớp dữ liệu cho bảng 2.2 .....	15
Bảng 2.4 : Huấn luyện với thuộc tính phân lớp là buys computer .....	20
Bảng 3.1 : Bảng phân loại thể trạng cơ thể theo chỉ số BMI .....	29
Bảng 3.2 : Bảng kiểu dữ liệu của các thuộc tính .....	32
Bảng 4.1 : Bảng danh sách Bác sĩ đánh giá chương trình.....	61

## **Danh mục công thức**

Công thức (2.1) : Tính chỉ số thông tin (Information) .....	19
Công thức (2.2) : Tính chỉ số thông tin mong muôn (Entropy).....	19
Công thức (2.3) : Tính độ lợi thông tin (Information Gain).....	19
Công thức (2.4) : Thông tin tiềm năng (potential information).....	19
Công thức (2.5) : Tính tỉ lệ độ lợi thông tin (Gain ratio) .....	19
Công thức (3.1) : Tính chỉ số sức khỏe .....	29

# Chương 1. TỔNG QUAN ĐỀ TÀI

## 1.1. Đặt vấn đề

Ứng dụng công nghệ thông tin vào việc lưu trữ và xử lý thông tin ngày nay được áp dụng hầu hết trong mọi lĩnh vực, điều này đã tạo ra một lượng lớn dữ liệu được lưu trữ với kích thước tăng lên không ngừng. Đây chính là điều kiện tốt cho việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập bảng biểu và khai phá dữ liệu.

Khai phá dữ liệu (KPDL) là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác suất, thống kê, máy học nhằm tìm kiếm các tri thức tiềm ẩn trong các kho dữ liệu có kích thước lớn mà người dùng khó có thể nhận biết bằng những kỹ thuật thông thường. Nguồn dữ liệu y khoa rất lớn, nếu áp dụng KPDL trong lĩnh vực này sẽ mang lại nhiều ý nghĩa cho ngành y tế. Nó sẽ cung cấp những thông tin quý giá nhằm hỗ trợ trong việc chẩn đoán và điều trị sớm giúp bệnh nhân thoát được nhiều căn bệnh hiểm nghèo.

Trong lĩnh vực Y khoa ở Việt Nam, hiện nay các tuyến y tế phường, xã, vùng sâu, vùng xa còn thiếu nhân lực y tế có trình độ chuyên môn và thiếu các trang thiết bị cần thiết trong chẩn đoán bệnh. Vì vậy xây dựng hệ hỗ trợ chẩn đoán rất cần thiết cho ngành y tế hiện nay ở Việt Nam. Hệ hỗ trợ sẽ kết hợp với cán bộ y tế giúp chẩn đoán sớm một số bệnh phát hiện sớm được những bệnh nguy hiểm và giảm gánh nặng kinh tế cho gia đình bệnh nhân và cho xã hội. Để minh chứng cho những lợi ích mà hệ hỗ trợ chẩn đoán mang lại, đề tài chọn dữ liệu bệnh sốt xuất huyết để thử nghiệm và đánh giá.

Ứng dụng kỹ thuật phân lớp dữ liệu trong khai phá dữ liệu nhằm xây dựng hệ thống hỗ trợ chẩn đoán là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật giải cũng như đặc điểm của dữ liệu thu thập được về

bệnh sốt xuất huyết, đề tài đề xuất ứng dụng mô hình phân lớp bằng cây quyết định với thuật toán C4.5 để tìm ra các qui luật tìm ẩn trong dữ liệu.

## 1.2. Cơ sở hình thành đề tài

Theo thông báo của Tổ chức Y tế thế giới, trên thế giới có 2,5 tỷ người sống trong vùng sốt xuất huyết (SXH) lưu hành thì có tới 1,8 tỷ người thuộc khu vực châu Á Thái Bình Dương [1]. Việt Nam là nước có bệnh SXH lưu hành rộng, SXH luôn là một trong những bệnh truyền nhiễm có số mắc cao hàng đầu mỗi năm Việt Nam vẫn có khoảng trên 100.000 bệnh nhân SXH và gần 100 người tử vong vì bệnh này và Bộ Y tế Việt Nam luôn quan tâm đến những nhiệm vụ trọng tâm của chương trình quốc gia phòng chống SXH [1]. Vì vậy xây dựng hệ hỗ trợ chẩn đoán y khoa để góp phần chẩn đoán nhanh và phát hiện sớm những nguy cơ dịch bệnh là vấn đề quan trọng của gia đình và xã hội. Đề tài áp dụng công nghệ thông tin xây dựng hệ hỗ trợ chẩn đoán với dữ liệu thu thập được từ bệnh SXH.

## 1.3. Một số kết quả nghiên cứu trong và ngoài nước

### 1.3.1. Kết quả nghiên cứu trên thế giới

Trên thế giới đã cho ra nhiều ứng dụng từ hệ hỗ trợ để chẩn đoán nhanh và điều trị bệnh tốt hơn như Hệ thống chẩn đoán y tế Caduceus của Harry Pope [17]; Hệ thống chuyên gia y tế DiagnosisPro [18]; MYCIN (1973) hệ hỗ trợ chẩn đoán bệnh nhiễm trùng máu [6]; PUFF (1982) dùng để phân tích kết quả xét nghiệm chức năng phổi [7]; PSG-Expert (2000) chẩn đoán bệnh mất ngủ [8]; BI-RADS(2007) chẩn đoán ung thư vú [9]; Naser xây dựng một hệ chẩn đoán bệnh về da (2008) [10]; Compete quản lý bệnh nhân tăng huyết áp, tiểu đường, bệnh mạn tính ...

### 1.3.2. Kết quả nghiên cứu trong nước

Ở Việt Nam tình hình ứng dụng công nghệ thông tin trong y tế còn tương đối ít, vào cuối năm 1980 cũng có những nghiên cứu hệ hỗ trợ bác sĩ chẩn đoán bệnh nội

khoa, châm cứu và chẩn trị đông y[2], hệ hỗ trợ ra quyết định trong việc chẩn đoán lâm sàng [3] ... Tuy vậy những nghiên cứu về chẩn đoán y khoa nhằm xây dựng các hệ hỗ trợ quyết định vẫn còn hạn chế.

## 1.4. Mục tiêu luận văn

Đề tài tập trung vào nghiên cứu bài toán phân lớp dữ liệu trong KPDL, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Ngoài ra, việc thu thập dữ liệu bệnh của một bệnh cụ thể cũng được quan tâm và đề tài đề xuất sử dụng dữ liệu bệnh sốt xuất huyết. Sau khi đã phân tích đặc điểm của dữ liệu thu thập được và lựa chọn thuật giải phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ hỗ trợ chẩn đoán cũng là mục tiêu chính của đề tài.

## 1.5. Đối tượng và phương pháp nghiên cứu

Đề tài tập trung nghiên cứu kỹ thuật phân lớp trong KPDL (cụ thể là nghiên cứu thuật toán C4.5) để áp dụng vào việc phân tích cơ sở dữ liệu y khoa. Luận văn thu thập dữ liệu bệnh SXH của tất cả bệnh nhân (không phân biệt tuổi, giới tính) đến khám và điều trị tại Bệnh viện Nguyễn Tri Phương và Bệnh viện Bệnh Nhiệt Đới.

Sử dụng phương pháp nghiên cứu hồi cứu với sự hỗ trợ chuyên môn của các bác sĩ chuyên khoa, đề tài tiến hành nghiên cứu trên cơ sở các thuật toán phân lớp dữ liệu trong KPDL.

## 1.6. Ý nghĩa của đề tài

### 1.6.1. Ý nghĩa khoa học

Với sự trợ giúp của máy tính, đề tài đóng góp một biện pháp thực hiện hỗ trợ các cán bộ y tế chẩn đoán bệnh cho bệnh nhân. Kết quả, kinh nghiệm thu được khi

thực hiện đề tài này sẽ giúp các cán bộ y tế phát hiện sớm bệnh cho bệnh nhân, đồng thời mong muốn những người đang công tác trong lĩnh vực Y học và Khoa học máy tính ngoài lại với nhau để tìm ra những giải pháp tốt hơn trong vấn đề chẩn đoán và điều trị bệnh bằng cách kết hợp giữa 2 lĩnh vực Y học và Khoa học máy tính.

### **1.6.2. Ý nghĩa thực tiễn**

Chẩn đoán bệnh và phát hiện bệnh là cả một quá trình, đòi hỏi các cán bộ y tế không những phải thật vững chuyên môn mà còn có đầy đủ các trang thiết bị y tế mới có thể chẩn đoán chính xác bệnh cho bệnh nhân. Nếu chẩn đoán bệnh sai sẽ đưa đến điều trị sai, không phát hiện được bệnh cho bệnh nhân... sẽ dẫn đến những tổn thất lớn về tinh thần lẫn vật chất cho bệnh nhân và gia đình họ. Việc phát hiện bệnh sớm thì khả năng thất bại trong điều trị sẽ giảm hoặc có thể giúp bệnh nhân và gia đình họ đưa ra những quyết định điều trị thích hợp. Vì vậy chẩn đoán và phát hiện được bệnh sớm sẽ giúp cán bộ y tế đưa ra những phác đồ điều trị hiệu quả đồng thời theo dõi, cảnh báo và tư vấn giúp bệnh nhân tránh những biến chứng nguy hiểm, giảm gánh nặng kinh tế cho gia đình và xã hội.

## **1.7. Bố cục luận văn**

Luận văn bao gồm các phần sau:

### **Chương 1: Tổng quan đề tài**

Giới thiệu về những vấn đề liên quan đến phân lớp dữ liệu trong Khai phá dữ liệu (Data mining), cơ sở hình thành đề tài, mục tiêu, phạm vi nghiên cứu, ý nghĩa thực tiễn và bố cục luận văn.

### **Chương 2: Cơ sở lý thuyết**

Chương 2 nói lên cách tiếp cận và giải quyết vấn đề của luận văn. Trình bày cơ sở toán học và áp dụng lý thuyết vào bài toán. Trình bày kiến thức cơ bản về bệnh SXH, ý nghĩa các chẩn đoán (thuộc tính).

**Chương 3: Xây dựng hệ hỗ trợ chẩn đoán.**

Trong chương này trình bày đặc điểm của dữ liệu, các bước tiền xử lý dữ liệu trước khi đưa vào hệ thống. Xây dựng ứng dụng chẩn đoán dựa vào dữ liệu về bệnh SXH.

**Chương 4: Thực nghiệm và đánh giá.**

Chạy chương trình với tập dữ liệu huấn luyện. Vẽ sơ đồ cây quyết định và rút ra các tập luật. Kiểm nghiệm đánh giá chương trình với tập dữ liệu kiểm tra.

**Chương 5: Tổng kết.**

Ý nghĩa thực tiễn mang tính cộng đồng, những hạn chế và hướng phát triển của luận văn.

## Chương 2. CƠ SỞ LÝ THUYẾT

### 2.1. Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining)

#### 2.1.1. Khái niệm về khai phá dữ liệu

KPDL thu hút sự chú ý của nền công nghiệp thông tin và xã hội trong những năm gần đây. Với sự phát triển của công nghệ thông tin, dữ liệu lưu trữ mỗi ngày trở thành một cơ sở dữ liệu rất lớn. Dựa vào khối lượng dữ liệu này, ta dùng những kỹ thuật KPDL để chuyển đổi thành những thông tin có ích hoặc rút ra những tri thức mới từ dữ liệu thu thập được[11]. Giáo sư Tom Mitchell định nghĩa Khai phá dữ liệu như sau: “*Khai phá dữ liệu là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai*” [12].

Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng, ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin. Khai phá dữ liệu trong lĩnh vực y tế sẽ tạo ra các hệ chẩn đoán bệnh hỗ trợ bác sĩ trong vấn đề chẩn đoán bệnh cho bệnh nhân.

#### 2.1.2. Các giai đoạn của quá trình khai phá dữ liệu [4]:

##### a. Tìm hiểu nghiệp vụ và dữ liệu

Giai đoạn này ta cần xác định vấn đề cần giải quyết, tìm hiểu kiến thức về bài toán đang thực hiện bao gồm các tri thức của các chuyên gia trong lĩnh vực cần nghiên cứu từ đó xác định chính xác nguồn dữ liệu để thu thập đồng thời phải hiểu được cấu trúc dữ liệu, ý nghĩa và tầm quan trọng của nó để từ đó ta đưa ra bài toán cụ thể để giải quyết vấn đề.

### b. Chuẩn bị dữ liệu

Giai đoạn này ta dùng các kỹ thuật tiền xử lý dữ liệu để xử lý dữ liệu đã thu thập được sao cho các giải thuật KPDL có thể hiểu được. Tiền xử lý dữ liệu bao gồm:

- + Xử lý dữ liệu bị thiếu hoặc mất: Các giá trị bị thiếu hoặc mất sẽ được thay thế bằng các giá trị thích hợp hơn hoặc xóa những dữ liệu sai miền giá trị và giải quyết sự không nhất quán.
- + Khử sự trùng lắp dữ liệu : Loại bỏ những dữ liệu bị trùng.
- + Giảm nhiễu dữ liệu: Các dữ liệu bị nhiễu sẽ được điều chỉnh hoặc loại ra khỏi cơ sở dữ liệu.
- + Rời rạc hóa dữ liệu: Các dữ liệu số sẽ được rời rạc hóa ra dạng phù hợp cho khai phá dữ liệu.
- + Giảm chiều: Loại bỏ các thuộc tính chứa ít thông tin để tiết kiệm thời gian và tài nguyên của máy tính.

### c. Mô hình hóa dữ liệu

Dùng các giải thuật của KPDL để tìm ra các qui luật của dữ liệu, quan trọng nhất trong giai đoạn này là tìm được giải thuật phù hợp để giải quyết vấn đề đã đặt ra.

### d. Hậu xử lý và đánh giá mô hình

Đây là giai đoạn biến đổi từ những luật rút ra được (của giai đoạn trước) từ tập huấn luyện sang dạng phù hợp với nghiệp vụ của bài toán đang nghiên cứu. Đồng thời cũng sẽ là giai đoạn đánh giá của các chuyên gia tư vấn dựa trên tập dữ liệu thử. Dựa vào nhận xét và hỗ trợ của các chuyên gia khi đó sẽ điều chỉnh kịp thời các mô hình của các giai đoạn trước. Các mô hình đạt yêu cầu với các chuyên gia sẽ được sử dụng.

### e. Triển khai mô hình

Các mô hình đạt yêu cầu sẽ được xây dựng thành chương trình ứng dụng thực tế nhằm hỗ trợ đưa ra quyết định theo yêu cầu của người dùng.

## 2.2. Tổng quan về hệ hỗ trợ ra quyết định

Hệ hỗ trợ ra quyết định là một hệ thống thuộc hệ thống thông tin, có nhiệm vụ cung cấp các thông tin hỗ trợ cho việc ra quyết định để tham khảo và giải quyết vấn đề. Hệ hỗ trợ ra quyết định có thể dùng cho cá nhân hay tổ chức và có thể hỗ trợ gián tiếp hoặc trực tiếp [13].

Trong lĩnh vực y tế, hệ hỗ trợ ra quyết định dựa vào tri thức đã học sẽ cung cấp thông tin chẩn đoán bệnh cho nhân viên y tế. Thông tin này được trích lọc để cung cấp một cách thông minh có giá trị cho quá trình chẩn đoán, theo dõi và điều trị bệnh hiệu quả hơn, từ đó ta thấy một số lợi ích của hệ hỗ trợ ra quyết định trong y tế như sau:

- Tăng cường chất lượng chẩn đoán, chăm sóc bệnh nhân.
- Giảm nguy cơ sai sót để tránh các tình huống nguy hiểm cho bệnh nhân.
- Tăng cường hiệu quả ứng dụng Công nghệ thông tin vào lĩnh vực y tế để giảm bớt những thủ tục giấy tờ không cần thiết,...

## 2.3. Bài toán Phân lớp trong Khai phá dữ liệu

### 2.3.1. Khái niệm về phân lớp

Phân lớp là một hình thức phân tích dữ liệu nhằm rút ra những mô hình mô tả những lớp trong dữ liệu. Những mô hình này gọi là mô hình phân lớp (classifier hoặc classification model) được dùng để dự đoán những nhãn lớp có tính phân loại (categorical), rời rạc và không có thứ tự cho những đối tượng dữ liệu mới.

### 2.3.2. Quá trình phân lớp dữ liệu

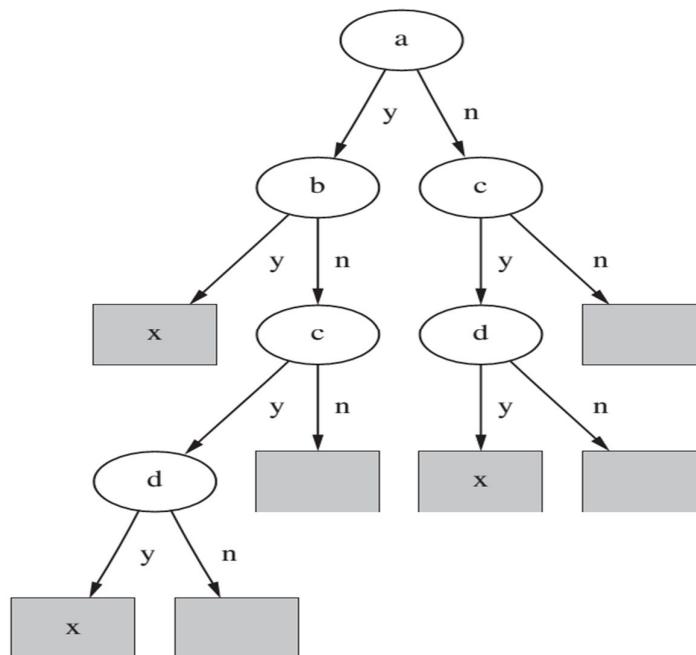
Một quá trình phân lớp gồm 2 bước:

#### a. Bước thứ nhất: Học/Huấn luyện [11]

Quá trình học nhằm xây dựng một mô hình phân lớp (Classifier) bao gồm các lớp dữ liệu đã được khái niệm trước từ tập dữ liệu đầu vào. Bước học (hay giai đoạn huấn luyện) dùng một giải thuật phân lớp (Classification Algorithms) để phân lớp các bản ghi của dữ liệu huấn luyện. Trong đó tập huấn luyện là một tập dữ liệu có cấu trúc với các thuộc tính và bộ dữ liệu tương ứng với các thuộc tính.

#### b. Bước thứ 2: Phân lớp (Classification) [11]

Ở bước thứ hai (Hình 2.1), mô hình tìm được ở bước thứ nhất sẽ được dùng cho việc phân loại những dữ liệu mới. Ta dùng một tập kiểm tra, bao gồm các bản ghi kiểm tra và các nhãn lớp liên kết với chúng để so sánh kết quả đầu ra của bộ phân lớp. Các bản ghi kiểm tra này chưa được dùng để xây dựng mô hình phân lớp ở bước 1. Kết quả của mô hình phân lớp như sơ đồ sau:



Hình 2.1: Kết quả quá trình phân lớp [11]

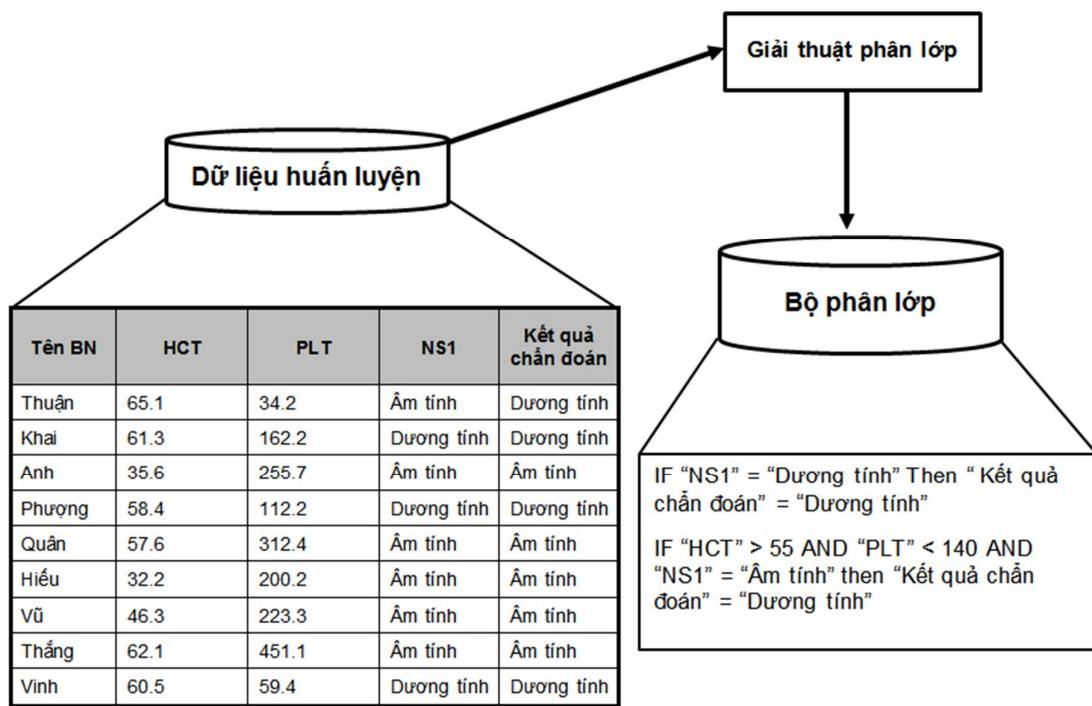
IF  $a = y$  and  $b = y$  then class x

IF  $a = n$  and  $c = y$  and  $d = y$  then class x

### Ví dụ minh họa :

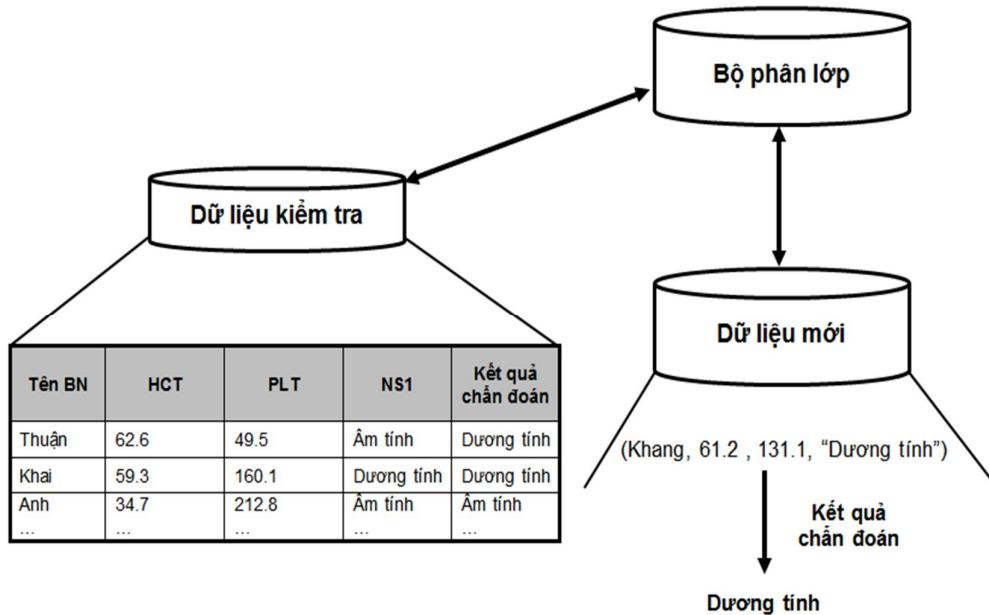
#### Bước 1 Xây dựng mô hình

Mục đích : Phân lớp bệnh nhân vào 2 lớp: “Dương tính” và “Âm tính” trong bộ phân lớp có nhãn “KẾT QUẢ CHẨN ĐOÁN”. Mỗi bệnh nhân có các thuộc tính dùng để phân lớp như sau: HCT, PLT, NS1. Sau khi huấn luyện, ta được mô hình phân lớp.



Hình 2.2: Xây dựng mô hình phân lớp

## Bước 2: Phân lớp



Hình 2.3: Bước phân lớp

Đánh giá kết quả mô hình ở bước 1, ta dùng tập dữ liệu kiểm tra. Với một mẫu mới, dùng bộ phân lớp để phân lớp mẫu này vào một trong các lớp được rút ra từ mô hình ở bước 1. Trong dữ liệu kiểm tra của hình 2.3, bệnh nhân Khai có các giá trị : HCT = 59.3; PLT = 160.1; NS1 = “Đương tính” thì mô hình sẽ phân lớp cho trường hợp này là “Kết quả chẩn đoán” = “Đương tính” (hình 2.3).

### Một số vấn đề bộ phân lớp cần quan tâm giải quyết:

- + Độ chính xác: Độ tin cậy của một luật dựa vào độ chính xác khi phân lớp.
- + Tốc độ: Trong một số tình huống, tốc độ phân lớp được xem như là một yếu tố quan trọng.
- + Dễ hiểu: Một bộ phân lớp dễ hiểu sẽ tạo cho người sử dụng tin tưởng hơn vào hệ thống, đồng thời giúp cho người sử dụng tránh được việc hiểu lầm kết quả của một luật được đưa ra bởi hệ thống.
- + Đơn giản: Kết quả đưa ra cây quyết định liên quan kích thước của nó.

- + Thời gian để học: Khi hệ thống hoạt động trong môi trường thay đổi thường xuyên, điều đó yêu cầu hệ thống phải học rất nhanh một luật phân lớp hoặc nhanh chóng điều chỉnh một luật đã được học cho phù hợp với thực tế.

### Các kỹ thuật phân lớp :

- + Mô hình phân lớp dùng cây quyết định (Decision tree classification)
- + Phân lớp dùng mạng Neural
- + Phân lớp dùng mạng Bayesian
- + Phân lớp với K-nearest neighbor classifier
- + Phân tích thống kê
- + Các thuật toán di truyền
- + Phương pháp tập thô (Rough set Approach)

#### 2.3.3. Phân lớp dữ liệu bằng cây quyết định

##### a. Định nghĩa:

Cây quyết định là kết quả của quá trình huấn luyện một tập dữ liệu với các bản ghi đã có thuộc tính. Cây quyết định là một công cụ phổ biến trong KPDL và phân lớp dữ liệu[14].

Đặc điểm của cây quyết định là một cây có cấu trúc kiểu lưu đồ, trong đó :

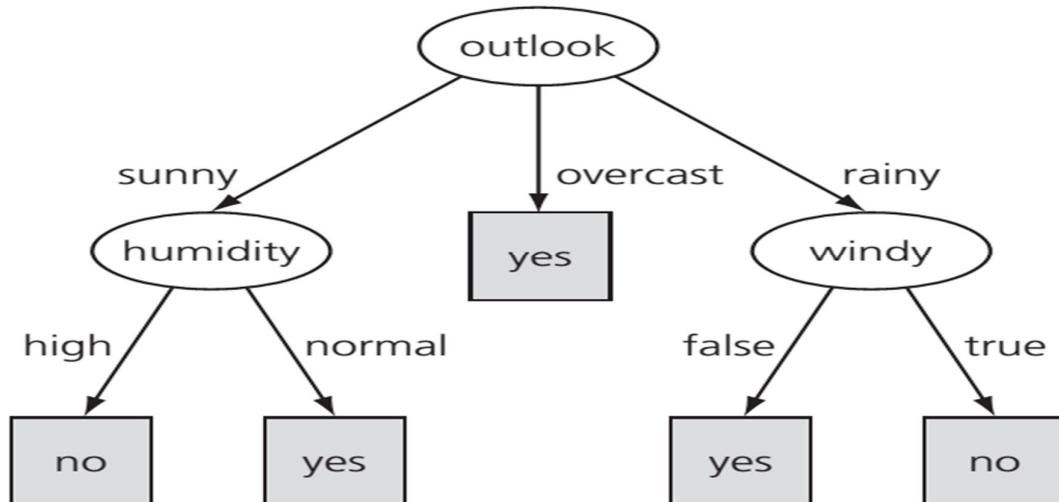
- + Gốc : Là nút trên cùng của cây.
- + Nút trong : Biểu diễn một kiểm tra trên một thuộc tính đơn (hình Oval).
- + Nhánh : Biểu diễn các kết quả của kiểm tra trên nút.
- + Nút lá : Biểu diễn lớp hay sự phân phối lớp (hình vuông hoặc chữ nhật)

Ví dụ: ta có bảng dữ liệu về 14 đối tượng và 5 thuộc tính trong bảng 2.1. Trong đó thuộc tính Play là thuộc tính phân lớp.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Bảng 2.1: Bảng dữ liệu về thời tiết [11]

Dựa vào quá trình phân lớp, ta có thể tạo cây quyết định như sau



Hình 2.4: Mô hình cây quyết định trong phân lớp dữ liệu về thời tiết [11]

### b. Sử dụng cây quyết định trong dự đoán lớp các dữ liệu chưa biết:

Mục đích của cây quyết định là dùng để dự đoán lớp (xác định lớp) cho các đối tượng dữ liệu chưa biết. Giả sử ta biết thời tiết trong 3 ngày với các giá trị dữ liệu đã biết về các thuộc tính “Outlook”, “Temperature”, “Humidity”, và “Windy”. Tuy nhiên ta chưa biết người chơi thể thao có quyết định ( thuộc tính phân lớp “Play”) như thế nào (Yes hoặc No). Như vậy với 4 thuộc tính đã cho ta sử dụng cây quyết định được tạo ra ở trên (hình 2.4) để dự đoán họ sẽ quyết định như thế nào ?

<b>Outlook</b>	<b>Temperature</b>	<b>Humidity</b>	<b>Windy</b>	<b>Play</b>
Sunny	Hot	Normal	False	?
Overcast	Cool	Normal	False	?
Rainy	Cool	High	True	?

**Bảng 2.2:** Ví dụ về thời tiết trong 3 ngày

Ta bắt đầu từ nút gốc của cây (hình 2.4) từ thuộc tính “Outlook”, nếu “Outlook” là **Overcast** thì quyết định của người chơi sẽ là **Yes**. Nếu “Outlook” là **Sunny** thì cây quyết định xét tiếp đến thuộc tính “Humidity”, nếu “Humidity” là **High** thì quyết định **No** hoặc **Normal** thì quyết định là **Yes**. Tiếp theo ta thấy nếu “Outlook” là “Rainy” cây quyết định xét tiếp đến thuộc tính Windy, nếu “Windy” là **False** thì quyết định **Yes** còn **True** thì quyết định là **No**.

Tóm lại theo cây quyết định trên thì các luật (Series of rules) được sinh ra như sau:

Luật 1: if outlook is Overcast then Play = Yes

Luật 2: if outlook is Sunny and Humidity is High then Play = No

Luật 3: if outlook is Sunny and Humidity is Normal then Play = Yes

Luật 4: if outlook is Rainy and Windy is False then Play = Yes

Luật 5: if outlook is Rainy and Windy is True then Play = No

Dựa vào các luật này ta có kết quả phân lớp cho dữ liệu trong bảng 2.2 như sau:

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	Normal	False	Yes
Overcast	Cool	Normal	False	Yes
Rainy	Cool	High	True	No

**Bảng 2.3:** Kết quả phân lớp dữ liệu cho bảng 2.2 [11]

Qua ví dụ trên ta thấy phân lớp bằng cây quyết định rất dễ hiểu. Tuy nhiên cây quyết định phụ thuộc rất lớn vào dữ liệu huấn luyện và thuật toán phân lớp.

### c. Đánh giá cây quyết định trong lĩnh vực khai phá dữ liệu:

#### Ưu điểm :

- + Quá trình xây dựng cây quyết định không dùng kiến thức về lĩnh vực dữ liệu đang nghiên cứu hoặc thông số đầu vào nào.
- + Kết quả của quá trình huấn luyện (học) được biểu diễn dưới dạng cây nên dễ hiểu và gần gũi với con người.
- + Nhìn chung, các giải thuật cây quyết định cho kết quả có độ chính xác khá cao.

#### Khuyết điểm :

- + Đôi với các tập dữ liệu có nhiều thuộc tính thì cây quyết định sẽ lớn (về chiều sâu cả chiều ngang), vì vậy làm giảm độ dễ hiểu.
- + Việc xếp hạng các thuộc tính để phân nhánh dựa vào lần phân nhánh trước đó và bỏ qua sự phụ thuộc lẫn nhau giữa các thuộc tính.
- + Khi dùng độ lợi thông tin (Information Gain) để xác định thuộc tính rẽ nhánh, các thuộc tính có nhiều giá trị thường được ưu tiên chọn.

#### d. Các thuật toán của cây quyết định:

- + ID3 (Decision tree)
- + C4.5
- + Cart (Classification and Regression Trees)
- + SLIQ (Supervised Learning In Quest)
- + Sprint (Scalable PaRallelization INduction of decision Trees)

#### 2.3.4. Đánh giá hiệu quả phân lớp

Mô hình phân lớp sau khi được tạo ra cần phải được đánh giá hiệu quả của mô hình đó. Để đánh giá mô hình, đề tài đề cập đến 2 phương pháp đánh giá phổ biến là holdout và k-fold cross-validation. Cả 2 kỹ thuật này đều dựa trên các phân hoạch ngẫu nhiên tập dữ liệu ban đầu. Thông thường, dữ liệu huấn luyện được chia làm 2 phần theo tỉ lệ 70% dùng để huấn luyện và 30% dùng để kiểm tra [20] tuy nhiên tỉ lệ này có thể thay đổi tùy ý.

##### **Phương pháp holdout:**

Dữ liệu gốc được chia ngẫu nhiên thành 2 phần (70% dùng để huấn luyện và 30% dùng để kiểm tra) như đã nói ở trên, sau quá trình huấn luyện thì mô hình được hình thành và chương trình sẽ dựa vào dữ liệu kiểm tra để đánh mô hình đúng được bao nhiêu phần trăm trong tổng số dữ liệu kiểm tra.

##### **Phương pháp k-fold cross validation:**

Là một phương pháp quan trọng trong việc đánh giá và phát triển mô hình huấn luyện. Cũng như phương pháp holdout, tập dữ liệu gốc cũng chia ra là 2 phần nhưng với k tập con (fold). Khi đó dữ liệu huấn luyện sẽ thay đổi ngẫu nhiên k lần để tạo ra k mô hình khác nhau. Mỗi mô hình được hình thành từ tập huấn luyện sẽ được đánh giá với tập dữ liệu kiểm tra tương ứng. Mô hình cho kết quả tốt nhất sẽ được chọn.

Ngoài ra có thể đánh giá hiệu quả phân lớp từ một tập dữ liệu kiểm tra độc lập đã được cấu hình tương tự như tập huấn luyện.

### 2.3.5. Thuật toán C4.5 xây dựng cây quyết định

#### a. Tóm quan:

C4.5 (thuật toán cải tiến của ID3) là một thuật toán phân lớp tạo ra cây quyết định được phát triển bởi J. Ross Quinlan[15]. Cây quyết định được tạo ra bởi thuật toán C4.5 có đặc điểm đơn giản, dễ sử dụng, dễ hiểu bởi các luật tạo ra ở nút lá của cây có thể biểu diễn dưới dạng câu lệnh If- then.

### b. Mã giả của thuật toán C4.5

```

1.   Function C45_builder(tập_A, tập_thuộc_tính)
2.   {
3.       if (mọi record trong tập_A đều nằm trong cùng một lớp)
4.       {
5.           return một nút lá được gán nhãn bởi lớp đó
6.       }
7.   else
8.   {
9.       if (tập_thuộc_tính là rỗng )
10.      {
11.          return nút lá được gán nhãn bởi tuyển của tất cả các
12.          lớp trong tập_A;
13.      }
14.   else
15.   {
16.       Chọn một thuộc tính P, lấy nó làm gốc cho cây hiện
17.       tại;
18.       Xóa P ra khỏi tập_thuộc_tính;
19.       For each (giá trị V của P)
20.       {
21.           Tạo một nhánh của cây gán nhãn V;
22.           Đặt vào phân_vùng V các ví dụ trong tập_A có
23.           giá trị V tại thuộc tính P;
24.           Gọi C45_builder (phân_vùng V, tập_thuộc_tính),
25.           gắn kết quả vào nhánh V;
26.       }
27.   }

```

### c. Thuật toán C4.5 dùng Gain-entropy

Cây được tạo ra sau khi huấn luyện có kích thước càng nhỏ thì độ chính xác càng cao và càng dễ hiểu nên thuật toán C4.5 dựa vào độ đo để lựa chọn thuộc tính tốt nhất. Hai độ đo được sử dụng trong C4.5 **information gain** hoặc **gain ratio**.

S: Tập training

$S_i$ : Lớp của tập các lớp  $C_i$  ( $i=1, \dots, m$ )

$a_j$ : Giá trị thuộc tính A ( $j=1, \dots, v$ )

**Chỉ số thông tin (Information) [16] cho sự phân lớp:**

$$I(S_1, S_2, \dots, S_m) = - \sum_{i=1}^m \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (2.1)$$

**Chỉ số thông tin mong muốn (Entropy) [16] cho sự phân lớp:**

Giả sử thuộc tính A được chọn để huấn luyện,  $A=\{S'_1, S'_2, \dots, S'_3\}$  khi đó Entropy của A được tính theo công thức như sau:

$$Ent(A) = \sum_{j=1}^v \frac{S'_{ij}}{S} \left( - \sum_{i=1}^m \frac{S'_{ij}}{S'_{ij}} \log_2 \frac{S'_{ij}}{S'_{ij}} \right) \quad (2.2)$$

Trong đó  $S'_{ij}$  là các trường hợp phân lớp của S'

**Chỉ số độ lợi thông tin (Information Gain) [16] cho phân lớp:**

Độ lợi thông tin (Information Gain) có được bởi việc phân nhánh trên thuộc tính A được tính như sau:

$$Gain(A) = I(S_1, S_2, \dots, S_m) - Ent(A) \quad (2.3)$$

Thuộc tính có Information Gain lớn nhất được chọn làm tiêu chí phân chia.

**Tỉ lệ độ lợi thông tin (Gain ratio) [16]:**

$$IV(A) = - \sum_{j=1}^v \frac{S'_{ij}}{S} \log_2 \frac{S'_{ij}}{S} \quad (2.4)$$

$$Gain\_Ratio = \frac{Gain(A)}{IV(A)} \quad (2.5)$$

### Mô tả cách tính information gain

Ví dụ ta có dữ liệu về Buy Computer như 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

Bảng 2.4: Huấn luyện với thuộc tính phân lớp là buys computer [16]

Trong tập dữ liệu huấn luyện trên:

$$S_1: \text{Buys\_computer} = \text{"yes"}. S_2: \text{Buys\_computer} = \text{"no"}.$$

Các thuộc tính: age; income; student và credit\_rating. Ta tính :

+ **Information:**

$$I(S) = I(S_1, S_2) = I(9, 5) = -9/14 * \log_2 9/14 - 5/14 * \log_2 5/14 = 0.940$$

+ **Entropy :** Tính Entropy của tất cả các thuộc tính trong bảng 2.4 như sau:

**Với thuộc tính Age:**

Thuộc tính age đã được rời rạc hóa thành các giá trị S'\_1, S'\_2, S'\_3 tương ứng là age < 30, age từ 30 đến 40, và age > 40

Với  $age = S'_1 = <30$  thì  $S'_{11} = 2; S'_{21} = 3$

$$I(S'_1) = I(S'_{11}, S'_{21}) = -2/5 \log 2/5 - 3/5 \log 3/5 = 0,971$$

Với  $age = S'_2 = 30-40$  thì  $S'_{12} = 4; S'_{22} = 0$

$$I(S'_2) = I(S'_{12}, S'_{22}) = 0$$

Với  $age = S'_3 = >40$  thì  $S'_{13} = 3; S'_{23} = 2$

$$I(S'_3) = I(S'_{13}, S'_{23}) = 0.971$$

Khi đó **Entropy** của thuộc tính Age là:

$$\text{Ent(age)} = \sum |S_i| / |S| * I(S_i) = 5/14 * I(S'_1) + 4/14 * I(S'_2) + 5/14 * I(S'_3) = 0.694$$

+ **Information Gain**: Tính **Information Gain** cho từng thuộc tính

$$\text{Gain (age)} = I(S) - \text{Ent(age)} = 0.246$$

Tính **Entropy** và **Information Gain** tương tự với các thuộc tính còn lại:

$$\text{Gain (income)} = 0.029$$

$$\text{Gain (student)} = 0.151$$

$$\text{Gain (credit\_rating)} = 0.048$$

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

## 2.4. Cơ sở dữ liệu Y khoa

### 2.4.1. Sơ lược bệnh Sốt xuất huyết

Sốt xuất huyết Dengue là một bệnh do Virus Dengue gây ra (Virus Dengue có 4 chủng khác nhau), bệnh xảy ra ở những vùng có khí hậu nhiệt đới và cận nhiệt đới. Bệnh được lan truyền bởi muỗi Aedes aegypti. Hiện nay SXH chưa có vaccine phòng ngừa và thuốc đặc trị. Bệnh này có thể gây ra cho từng người hoặc gây ra dịch trong một vùng dân cư. Khởi bệnh dễ nhầm lẫn với một số bệnh khác chính vì vậy

nếu không chẩn đoán và điều trị kịp thời sẽ gây ra những biến chứng rất nguy hiểm và có thể dẫn đến tử vong.

### **2.4.2. Diễn biến lâm sàng bệnh sốt xuất huyết dengue [19]**

Bệnh SXH Dengue rất đa dạng. Bệnh xảy ra đột ngột ở một người bình thường và viễn tiến bệnh từ nhẹ đến nặng rất nhanh qua 3 giai đoạn: sốt, giai đoạn nguy hiểm và giai đoạn hồi phục.

#### **a. Giai đoạn sốt**

##### **Lâm sàng**

- + Sốt cao đột ngột, liên tục.
- + Nhức đầu, chán ăn, buồn nôn.
- + Da xung huyết.
- + Đau cơ, đau khớp, nhức hai hố mắt.
- + Nghiệm pháp dây thắt dương tính.
- + Thường có châm xuất huyết ở dưới da, chảy máu chân răng hoặc chảy máu cam.

##### **Cận lâm sàng**

- + Dung tích hồng cầu (Hematocrit) bình thường.
- + Số lượng tiểu cầu bình thường hoặc giảm dần.
- + Số lượng bạch cầu thường giảm.

#### **b. Giai đoạn nguy hiểm: Thường vào ngày thứ 3-7 của bệnh**

##### **Lâm sàng**

- + Người bệnh có thể còn sốt hoặc đã giảm sốt.

- + Biểu hiện thoát huyết tương do tăng tính thâm thành mạch
- + Tràn dịch màng phổi, mô kẽ, màng bụng, gan to, có thể đau.
- + Nếu thoát huyết tương nhiều -> huyết áp kẹt (hiệu số huyết áp tối đa và tối thiểu  $\leq 20$  mmHg), tụt huyết áp hoặc không đo được huyết áp, tiểu ít.
- + Xuất huyết dưới da.
- + Xuất huyết ở niêm mạc.
- + Xuất huyết nội tạng như tiêu hóa, phổi, não là biểu hiện nặng.
- + Một số trường hợp nặng có thể có biểu hiện suy tạng như viêm gan nặng, viêm não, viêm cơ tim.

### Cận lâm sàng

- + HCT tăng.
- + PLT  $< 100.000/\text{mm}^3$ .
- + Siêu âm hoặc xquang có thể phát hiện tràn dịch màng bụng, màng phổi.

### c. Giai đoạn hồi phục

#### Lâm sàng

- + Sau 24-48 giờ của giai đoạn nguy hiểm, có hiện tượng tái hấp thu dần dịch từ mô kẽ vào bên trong lòng mạch. Giai đoạn này kéo dài 48-72 giờ.
- + Người bệnh hết sốt, toàn trạng tốt lên, thèm ăn, huyết động ổn định và tiểu nhiều.
- + Có thể có nhịp tim chậm và thay đổi về điện tâm đồ.
- + Trong giai đoạn này, nếu truyền dịch quá mức có thể gây ra phù phổi hoặc suy tim.

### Cận lâm sàng

- + Hematocrit trở về bình thường hoặc có thấp hơn do hiện tượng pha loãng máu khi dịch được tái hấp thu trở lại.
- + Số lượng bạch cầu máu thường tăng lên sớm sau giai đoạn hạ sốt.
- + Số lượng tiểu cầu dần trở về bình thường, muộn hơn so với số lượng bạch cầu.

#### 2.4.3. Chẩn đoán [19]

Bệnh SXH Dengue được chia làm 3 mức độ (Theo Tổ chức Y tế Thế giới năm 2009):

- + SXH Dengue.
- + SXH Dengue có dấu hiệu cảnh báo.
- + SXH Dengue nặng.

##### a. Sốt xuất huyết Dengue

#### Lâm sàng

Sốt cao đột ngột, liên tục từ 2-7 ngày và có ít nhất 2 trong các dấu hiệu sau:

- + Có xuất huyết dưới da.
- + Nhức đầu, chán ăn, buồn nôn.
- + Đau cơ.

#### Cận lâm sàng

- + HCT bình thường hoặc tăng.
- + PLT bình thường hoặc hơi giảm.
- + WBC thường giảm.

### b. Sốt xuất huyết Dengue có dấu hiệu cảnh báo

**Lâm sàng:** Bao gồm các triệu chứng lâm sàng của SXH Dengue, kèm theo các dấu hiệu cảnh báo sau:

- + Vật vã, lù đù, li bì.
- + Đau bụng vùng gan hoặc ấn đau vùng gan.
- + Gan to  $> 2$  cm.
- +Ói nhiều.
- + Xuất huyết niêm mạc.
- + Tiêu ít.

#### Cận lâm sàng

- + Hematocrit tăng cao.
- + Tiêu cầu giảm nhanh chóng.

### c. Sốt xuất huyết Dengue nặng

- + Thoát huyết tương nặng dẫn đến sốc.
- + Xuất huyết nặng.
- + Suy tạng.

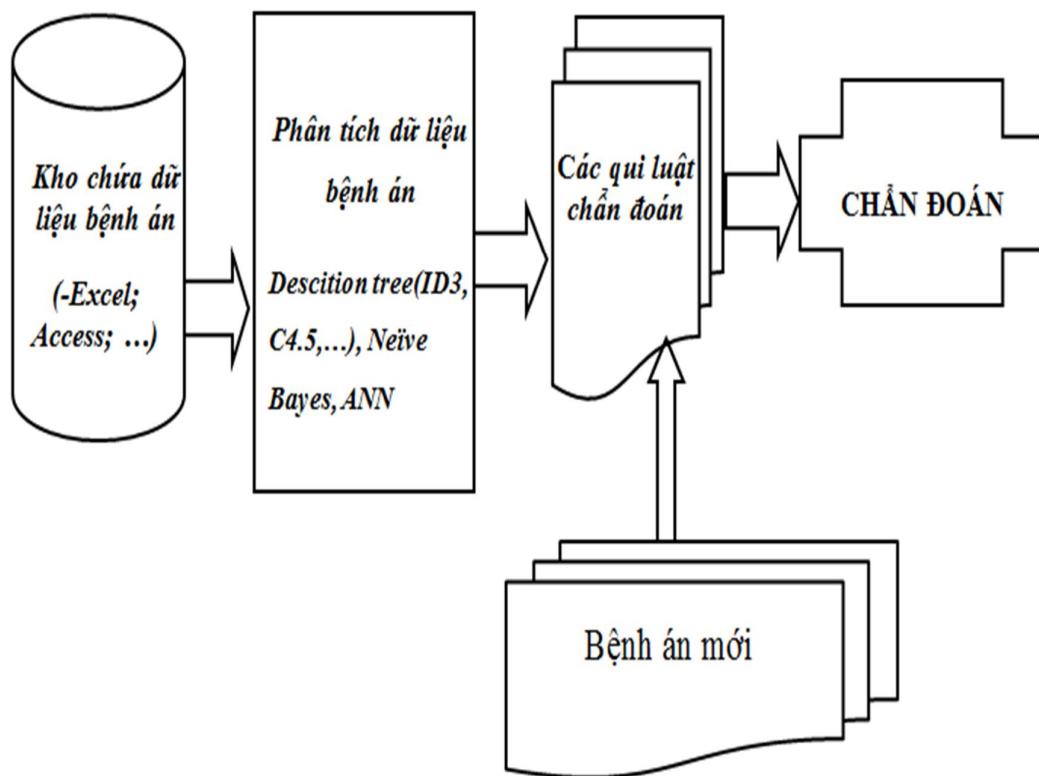
### d. Tìm kháng nguyên và kháng thể vi rút Dengue

- + Tìm kháng nguyên NS1.
- + Tìm kháng thể IgM.
- + Tìm kháng thể IgG.

# Chương 3. XÂY DỰNG HỆ HỖ TRỢ CHẨN ĐOÁN Y KHOA

## 3.1. Cơ sở dữ liệu xây dựng mô hình

Sau khi thu thập dữ liệu ta cần xây dựng cơ sở dữ liệu, lưu trữ các thông tin cần thiết cho bộ điều khiển theo mô hình sau:



Hình 3.1 : Mô hình xây dựng giải pháp hỗ trợ chẩn đoán bệnh

### 3.1.1. Kho chứa dữ liệu bệnh án điện tử

Bệnh án điện tử bao gồm thông tin cá nhân của bệnh nhân như giới tính, tuổi, sinh hiệu ... và thông tin điều trị bệnh.

#### a. Đối tượng nghiên cứu

**Dân số chọn mẫu:** Tất cả các bệnh nhân trên 15 tuổi được chẩn đoán ban đầu là SXH với các mức độ nặng nhẹ khác nhau nhập vào Bệnh viện.

**Tiêu chuẩn chọn mẫu:** Chọn mẫu khi bệnh nhân được chẩn đoán:

- + SXH Dengue.
- + SXH Dengue có dấu hiệu cảnh báo.
- + SXH Dengue nặng.

Tập dữ liệu chứa các thông tin về các bệnh nhân bị SXH từ SXH Dengue cho đến Sốt xuất huyết Dengue nặng được nhập viện điều trị tại Bệnh viện.

### b. Thu thập Dữ liệu:

Dữ liệu thu thập được lưu trữ dưới dạng file Excel

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T				
1	HoTen	Tuoi	Chieu	Tuo	Chieu	BMI	Chieu	Mach	NhiетDo	HuyetAp	TamTr	Lech	Huyete	tech	Huyet	BachCau	ChanDoan	KhoangBachCau	TieuCau	KhoangTieuCau	IGM	IGG	NS1
2	TRAN VAN N	36	2	19.2277	3	82	37.50	100	60	4	3	3.700	SXH	Dengue			3	128,000	3	TRUE	FALSE	FALSE	
3	DUONG T Q.N	18	2	18.026	3	80	37.00	110	70	4	3	3.700	SXH	Dengue			3	107,000	3	FALSE	FALSE	TRUE	
4	HOANG T THANH TA	62	4	22.8928	2	110	37.00	100	60	4	3	3.700	SXH	Dengue			3	75,000	3	FALSE	FALSE	FALSE	
5	VO VAN PHO	28	2	22.6563	2	80	38.00	110	70	4	3	3.100	SXH	Dengue			3	160,000	1	FALSE	FALSE	TRUE	
6	TRINH MINH NH	22	2	21.6311	3	82	37.50	110	70	4	3	3.100	SXH	Dengue			3	62,000	3	TRUE	TRUE	TRUE	
7	NGUYEN V	27	2	23.4375	2	80	37.00	110	70	4	3	1.900	SXH	Dengue			3	65,000	3	FALSE	FALSE	TRUE	
8	DO DANG H	22	2	21.875	3	80	37.00	100	60	4	3	4.000	SXH	Dengue	có dấu hiệu cảnh báo		3	83,000	3	FALSE	FALSE	TRUE	
9	NGUYEN VAN DI	23	2	19.7777	3	82	37.00	90	70	2	1	1.700	SXH	Dengue nặng			3	9,000	3	TRUE	TRUE	TRUE	
10	PHAM THI HA	37	2	18.3143	3	80	37.00	100	60	4	3	2.600	SXH	Dengue			3	100,000	3	FALSE	FALSE	TRUE	
11	NGUYEN THI T	22	2	16.8889	1	80	37.50	110	70	4	3	2.500	SXH	Dengue	có dấu hiệu cảnh báo		3	91,000	3	TRUE	TRUE	FALSE	
12	LE VAN MA	22	2	21.0938	3	80	39.00	110	70	4	3	7.500	SXH	Dengue			3	160,000	1	FALSE	FALSE	TRUE	
13	BUI DUC VI	22	2	18.026	3	75	37.00	110	70	4	3	7.400	NSV				3	160,000	1	FALSE	FALSE	FALSE	
14	NGUYEN THI XUAN KHA	29	2	18.7328	3	70	38.50	110	70	2	1	2.100	SXH	Dengue			3	38,000	3	TRUE	FALSE	FALSE	
15	NGUYEN DUY D	26	2	23.3377	2	70	38.50	80	60	5	4	1.000	SXH	Dengue nặng			3	47,000	3	TRUE	TRUE	TRUE	
16	LE NGUYEN QUOC PHO	26	2	17.9688	1	74	37.00	120	70	4	3	7.100	SXH	Dengue	có dấu hiệu cảnh báo		3	77,000	3	TRUE	TRUE	FALSE	
17	BUI THANH TO	23	2	17.2111	1	74	37.00	110	70	4	3	3.200	SXH	Dengue			3	130,000	3	TRUE	TRUE	FALSE	
18	LE XUAN TUL	22	2	24.1588	2	75	38.50	110	70	4	3	15.000	SXH	Dengue	có dấu hiệu cảnh báo		3	77,000	3	FALSE	FALSE	TRUE	
19	CAO QUOC TU	18	2	16.6933	1	74	37.00	110	70	4	3	3.400	SXH	Dengue	có dấu hiệu cảnh báo		3	33,000	3	TRUE	TRUE	FALSE	
20	NGUYEN THANH A	25	2	18.5901	3	74	37.00	100	60	2	1	4.800	SXH	Dengue	có dấu hiệu cảnh báo		3	24,000	3	TRUE	TRUE	FALSE	
21	BUI THI THU TH	20	2	20	3	110	37.50	100	80	4	3	1.100	SXH	Dengue nặng			3	10,000	3	TRUE	TRUE	FALSE	
22	HO MINH P	26	2	17.9982	1	75	37.00	110	70	3	2	5.500	NSV				3	160,000	1	FALSE	FALSE	FALSE	
23	NGO ANH KY	32	2	19.8413	3	80	38.00	110	70	4	3	4.900	SXH	Dengue			3	38,000	3	TRUE	TRUE	TRUE	
24	NGUYEN VAN DUO	33	2	18.424	3	80	37.80	90	60	3	2	2.200	SXH	Dengue	có dấu hiệu cảnh báo		3	21,000	3	TRUE	TRUE	FALSE	
25	VY THI LY	37	2	20.4294	3	70	37.00	110	70	4	3	3.800	SXH	Dengue			3	77,000	3	FALSE	FALSE	TRUE	
26	LUU THI THO	20	2	18.3143	3	80	38.00	110	80	5	4	3.800	SXH	Dengue			3	103,000	3	TRUE	TRUE	TRUE	
27	TRAN VAN N	21	2	24.2188	2	82	37.00	100	60	4	3	1.700	SXH	Dengue			3	78,000	3	FALSE	FALSE	TRUE	
28	NGUYEN THI H	23	2	16.8919	1	80	37.00	120	70	4	3	4.200	SXH	Dengue			3	73,000	3	TRUE	FALSE	FALSE	
29	TRAN NGOC A	17	1	21.3039	3	68	39.00	110	70	4	3	3.300	SXH	Dengue	có dấu hiệu cảnh báo		3	59,000	3	TRUE	TRUE	FALSE	
30	TRAN T TRUONG THA	20	2	18.4911	3	82	37.00	110	70	4	3	2.000	SXH	Dengue nặng			3	32,000	3	TRUE	FALSE	FALSE	
31	NGUYEN QUOC QU	25	2	22.2656	2	100	38.00	100	60	5	4	3.200	SXH	Dengue			3	30,000	3	FALSE	FALSE	TRUE	
32	HOANG THI TA	30	2	18.2222	3	80	37.00	110	70	4	3	3.900	NSV				3	153,000	1	FALSE	FALSE	FALSE	
33	TRAN THI XUAN HUO	30	2	23.3091	2	82	37.00	110	70	4	3	2.600	NSV				3	200,000	1	FALSE	FALSE	FALSE	
34	THANH CHE PHUO	22	2	19.9594	3	75	40.00	100	60	4	3	1.800	SXH	Dengue			3	38,000	3	FALSE	FALSE	TRUE	
35	DIEU NGOC A	19	2	20.5499	3	75	37.00	120	70	4	3	2.800	SXH	Dengue			3	54,000	3	TRUE	TRUE	FALSE	
36	HOANG THI TH	23	2	20.1348	3	80	38.50	110	70	3	2	3.400	SXH	Dengue			3	90,000	3	FALSE	FALSE	TRUE	

Hình 3.2: Tập dữ liệu thu thập được

**Thông tin bệnh nhân bao gồm :** Mã số bệnh nhân (SoHS), họ tên, tuổi, giới tính (qui định Nữ: 0; Nam:1) và chỉ số sức khỏe BMI. Công thức tính BMI như sau :

$$BMI = \frac{\text{cân nặng (kg)}}{\text{chiều cao}^2(m)} \quad (3.1)$$

BMI	Thể trạng
Bình thường	18.5 – 22.9
Nhẹ cân	< 18.5
Thừa cân	>=23

Bảng 3.1 Bảng phân loại thể trạng cơ thể theo chỉ số BMI [5]

### Triệu chứng lâm sàng

- + Mạch : Chậm, bình thường, nhanh, quá nhanh.
- + Nhiệt độ : Trung bình, sốt.
- + Huyết áp : Huyết áp tâm trương, huyết áp tâm thu.
- + Dấu hiệu cơ năng: Nhức đầu, đau cơ, đau bụng, ói, ho, tiêu lỏng.
- + Triệu chứng khác: Xuất huyết, vàng da, sốc.

### Cận lâm sàng:

- + Siêu âm : Gan to hoặc bình thường.
- + Hematocrite ( HCT ) : Bình thường, cao, thấp.
- + Bạch cầu (WBC) : Bình thường, cao, thấp.
- + Tiêu cầu (PLT) : Bình thường, cao, thấp.
- + Tìm kháng nguyên NS1: Dương tính hoặc âm tính
- + Tìm kháng thể IgM : Dương tính hoặc Âm tính
- + Tìm kháng thể IgG : Dương tính hoặc Âm tính

### 3.1.2. Tiền xử lý dữ liệu

Dữ liệu sau khi thu thập sẽ lưu vào file excel sau đó import vào database của **dulieuSXH.mdb**. Dữ liệu được nhập trên một hàng (tuple) bao gồm các thuộc tính:

*STT, SoHS, tuoi, gioi, HATThu, HATTRuong, nhucdau, dauco, xuathuyet, daubung, oi, ho, tieulong, trigiac, vangda, Ganto, NS1, TGM, IGG, HCT, Hongcau, Bachcau, Tieucau, DolechHA, KhoangHC, KhoangBC, KhoangTC, Chandoan, ChandoanID.*

Để thực hiện mô hình khai phá luật kết hợp ta cần hiệu chỉnh lại dữ liệu và loại bỏ các thuộc tính không cần thiết:

- + Loại bỏ các thuộc tính mà dữ liệu bị thiếu hoặc bị nhiều quá nhiều.
- + Loại bỏ thuộc tính “SoHS”, “Gioitinh” vì các thuộc tính này không dùng trong mô hình, gọi là lọc thuộc tính
- + Rời rạc hóa dữ liệu:

#### Rời rạc dữ liệu HCT

HCT			
ID	Ý Nghĩa	Từ	Đến
1	Bình thường	35	45
2	Cao	45	
3	Thấp	0	35

#### Rời rạc dữ liệu tiểu cầu

Tiểu cầu			
ID	Ý Nghĩa	Từ	Đến
1	Bình thường	140.000	400.000
2	Cao	400.000	
3	Thấp	0	140.000

### Rời rạc dữ liệu bạch cầu

Bạch cầu			
ID	Ý Nghĩa	Từ	Đến
1	Bình thường	5.000	10.000
2	Cao	10.000	
3	Thấp	0	5.000

**Bảng kiểu dữ liệu của các thuộc tính:**

STT	Thuộc tính	Kiểu	Ví dụ	Mô tả
1	SoHS	Numeric	13.04436	Định danh bệnh nhân
2	Tuoi	Numeric	36	Tuổi bệnh nhân
3	Giới tính	Nominal	True, False	Giới tính bệnh nhân
4	HATThu	Numeric	12	Huyết áp trên
5	HATTRuong	Numeric	8	Huyết áp dưới
6	HCT	Numeric	54.1	Dung tích hồng cầu
7	Bachcau	Numeric	3.2	Chỉ số bạch cầu
8	Tieucau	Numeric	89	Chỉ số tiểu cầu
9	DolechHA	Numeric	1,2,3	Độ lệch huyết áp
10	KhoangHCT	Numeric	1,2,3	Bình thường, cao thấp
11	KhoangBachcau	Numeric	1,2,3	Bình thường, cao thấp
12	KhoangTieucau	Numeric	1,2,3	Bình thường, cao thấp
13	NS1	Nominal	True, False	Dương tính, Âm tính
14	IGM	Nominal	True, False	Dương tính, Âm tính
15	IGG	Nominal	True, False	Dương tính, Âm tính
16	Dauco	Nominal	True, False	Có ,không
17	Nhucdau	Nominal	True, False	Có ,không
18	Xuathuyet	Nominal	True, False	Có ,không
19	Daubung	Nominal	True, False	Có ,không
20	Oi	Nominal	True, False	Có ,không
21	Ho	Nominal	True, False	Có ,không
22	Ganto	Nominal	True, False	Có ,không
23	Vangda	Nominal	True, False	Có ,không
24	Chandoan	Nominal	SXHDengue,...	Tình trạng bệnh
25	ChandoanID	Numeric	1,2,3,4,5,6	Phân lớp chẩn đoán

Bảng 3.2: Bảng kiểu dữ liệu của các thuộc tính

Sau bước tiền xử lý, dữ liệu sẽ được đưa vào hệ thống như hình 3.3

The screenshot shows a software application window titled "HỆ THỐNG HỖ TRỢ CHẨN ĐOÁN Y KHOA". The interface has several tabs at the top: "Xử lý dữ liệu", "Cây quyết định", "Kiểm thử kết quả", "Chẩn đoán", and "Kiểm tra chéo (CV)". Below these tabs, there is a sub-menu "Tiền xử lí dữ liệu" and a path "D:\De tai CH\soft\DuDoan\_SoXuatHuyet\_timluat - Cc". In the center, there is a button "Lấy dữ liệu" (Get data). The main area contains a table titled "Bảng dữ liệu" with the identifier "tbl\_sxh\_SData". The table has columns: Ho ten, HuyetApTT, HuyetAPTrong, HCT, BachCau, TieuCau, DoLechHu, KhoangHCT, KhoangBachCau, IGM, and KhoangTie NS1. The table lists numerous patient records with names like NGUYỄN THỊ..., ĐẶNG THỊ TR..., DƯƠNG THỊ..., etc., along with their corresponding values for each parameter. To the right of the table, a vertical list of features is shown with checkboxes, including ID, STT, SHS, Ho ten, HuyetApTT, HuyetAPTrong, HCT, BachCau, TieuCau, DoLechHu, KhoangHCT, KhoangBachCau, IGM, KhoangTie, NS1, DauCo, IGG, NhucDau, XuatHuyet, DauBung, Oi, Ho, GanTo, TieuLong, VangDa, ChanDoan, and ChanDoanID. Below this list, there is a section "Thuộc tính phân loại" with a dropdown menu set to "ChanDoanID". At the bottom of the interface, there are fields for "Tên thuộc tính", "Kiểu", "Bảng danh mục", "Trọng số", and a "Thoát" (Exit) button.

Hình 3.3: Tập dữ liệu huấn luyện đưa vào hệ thống

### 3.1.3. Phân tích dữ liệu bệnh án điện tử

Phân tích dữ liệu bệnh án điện tử để tìm quy luật chẩn đoán khi được cung cấp các thông tin về triệu chứng lâm sàng và cận lâm sàng của bệnh nhân.

Các giải thuật được sử dụng để xây dựng hệ thống là sử dụng giải thuật tiêu biểu: Cây quyết định với thuật toán C4.5 dùng Gain – Entropy làm độ đo để lựa chọn thuộc tính tốt nhất.

**Cây quyết định được xử lý như sau:**

1. Bảng dữ liệu ở đây là tập dữ liệu bệnh nhân SXH lấy từ các bệnh án, dữ liệu viết bằng tay...xử lý lại các thông tin có ích lợi cho hệ thống (tập dữ liệu dulieuSXH.mdb).
2. Gọi hàm (T)
3. Hàm S ban đầu là Hàm (T)

Chương trình bắt đầu làm việc như sau:

4. Nếu tất cả các dữ liệu trong S có cùng một lớp thì kết thúc.
  - + Dữ liệu trong tập S có cùng một lớp đây là cùng một bệnh SXH.
  - + Dùng chương trình
5. Với mỗi thuộc tính trung gian trong tập dữ liệu S. Tính giá trị cho thuộc tính đó.
  - + Đây là các yếu tố liên quan đến việc chẩn đoán cho bệnh nhân như: Mạch, Huyết áp, nhiệt độ hay các chỉ số xét nghiệm
  - + Ta đưa các chỉ số đó vào để chương trình tính xác suất.
6. Chọn thuộc tính có độ đo tốt nhất

Ví dụ: Như ở trên ta đưa ra 4 thuộc tính để cây quyết định xây dựng là Độ lệch huyết áp, chỉ số xét nghiệm là IgG, IgM, NS1, ta được:

RE1 ---> Độ lệch Huyết áp

RE1 ---> IgM

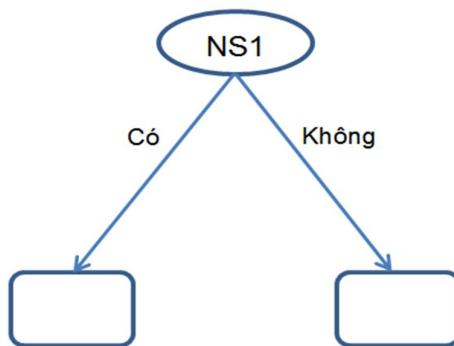
RE1 ---> IgG

RE2 ---> NS1 (NS1 được chọn là thuộc tính tốt nhất)

**NS1** có độ đo tốt nhất nên lấy **NS1** làm nút gốc và có 2 giá trị là True (Có), False (Không).

7. Nếu thuộc tính có K giá trị sẽ chia  $K \rightarrow S$ , sau đó gọi đệ quy để phân tách trên miền giá trị đó.

Đến đây (với ví dụ trên) chương trình sẽ phân **NS1** thành 2 thành phần (2 nhánh) để xét đệ quy đến khi được kết quả cuối cùng sẽ gán nút lá vào và ta được phân lớp chẩn đoán.



**Hình 3.4:** NS1 được chọn vì có độ đo tốt nhất

Phần lớn các hệ thống học máy đều cố gắng tạo ra 1 cây càng nhỏ càng tốt, vì cây càng nhỏ thì dễ đạt được độ chính xác dự đoán cao hơn.

Do không thể đảm bảo được cực tiểu của cây quyết định, C4.5 dựa vào nghiên cứu tối ưu hóa, và sự lựa chọn cách phân chia mà có độ đo lựa chọn thuộc tính đạt giá trị cực đại.

Hai độ đo được sử dụng trong C4.5 là information gain hoặc gain ratio.

### 3.1.4. Các qui luật chẩn đoán

Các qui luật chẩn đoán được xây dựng nhờ giải thuật phân lớp của cây quyết định trong KPDL. Qui luật này được rút ra từ tập dữ liệu huấn luyện, mỗi nút lá là một luật. Có được thông tin về lâm sàng và cận lâm sàng, bác sĩ điều trị sẽ dựa vào các thông tin này để đưa ra kết luận bệnh cuối cùng cho bệnh nhân.

**Cận lâm sàng :** Bệnh nhân được khẳng định nhiễm virus Dengue với các xét nghiệm HCT tăng hơn 20% giá trị HCT trước đó, PLT (tiểu cầu) giảm, có dấu hiệu xuất huyết, dương tính với NS1, IgM...

### Tiêu chuẩn loại trừ

Bệnh nhân có tiền sử bệnh tim, phổi, gan, thận... có bệnh lý rối loạn lipid máu, bệnh mạch vành đã biết trước hoặc phát hiện trong quá trình nhập viện.

#### 3.1.5. Bệnh án mẫu

Bệnh án mẫu là hồ sơ bệnh án của bệnh nhân bao gồm các thông tin về các triệu chứng lâm sàng, cận lâm sàng và kết luận chẩn đoán của bệnh nhân. Bệnh án mẫu được dùng để đưa các thông tin về lâm sàng và cận lâm sàng của bệnh nhân vào hệ thống, các thông tin này sẽ được phân tích dựa trên các qui luật chẩn đoán của hệ thống.

#### 3.1.6. Chẩn đoán

Sau khi đưa thông tin từ bệnh án mẫu vào hệ thống, chương trình dựa vào các luật đã được rút ra trong quá trình huấn luyện sẽ phân bệnh án mẫu vào lớp tương ứng của hệ thống để cho ra kết quả chẩn đoán. Kết quả này sẽ được so sánh với kết luận chẩn đoán trong bệnh án mẫu.

### 3.2. Xây dựng ứng dụng

#### 3.2.1. Giới thiệu chương trình

Chương trình được thiết kế và vận hành theo yêu cầu của đề tài nghiên cứu. Phần công nghệ của chương trình là dựa vào thế mạnh của Data mining với kỹ thuật phân lớp của cây quyết định và thuật toán C4.5 để khai phá dữ liệu về bệnh SXH nhằm rút ra các qui luật chẩn đoán bệnh SXH.

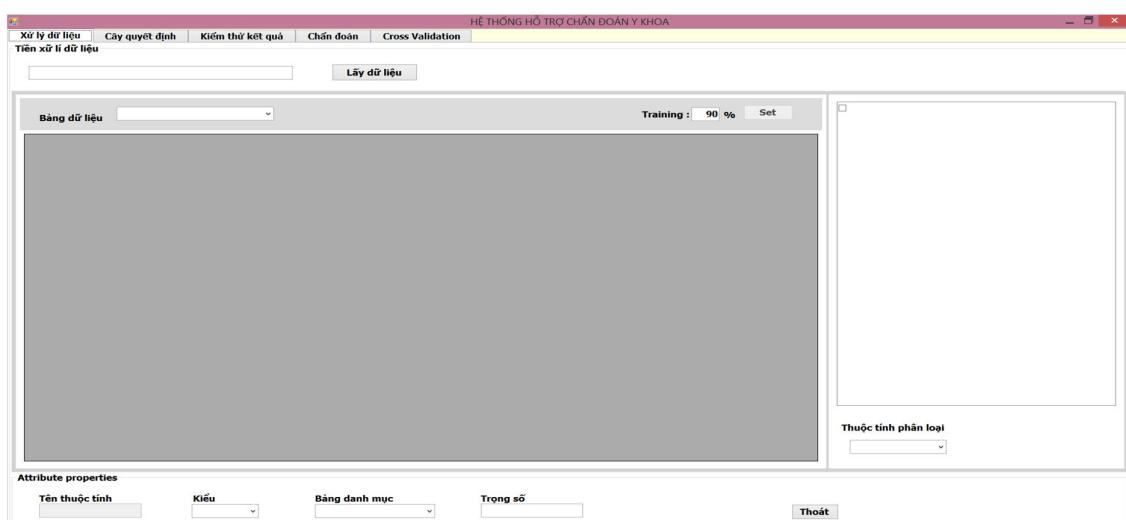
Chương trình sẽ tạo ra cây quyết định, đi từ nút gốc đến nút lá của một nhánh nào đó của cây sẽ cho ta biết một qui luật chẩn đoán. Cây quyết định sau khi được

tạo ra sẽ tự động lưu lại dưới dạng “\*.xml” để sử dụng cho các lần chẩn đoán sau mà không cần phải học lại toàn bộ dữ liệu để tạo cây. Điều này sẽ rút ngắn thời gian chẩn đoán. Chương trình chỉ học lại khi có thay đổi hoặc cập nhật dữ liệu.

Chương trình được xây dựng bằng ngôn ngữ C#, Net Framework 4.5

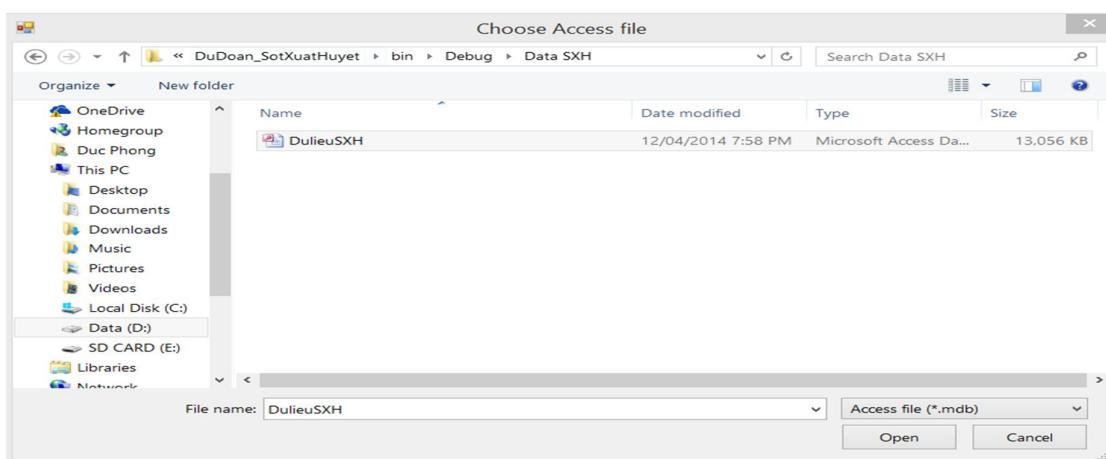
### 3.2.2. Cách thức vận hành chương trình

**Bước 1:** Chương trình được khởi động từ file exe



Hình 3.5: Màn hình khởi động chương trình

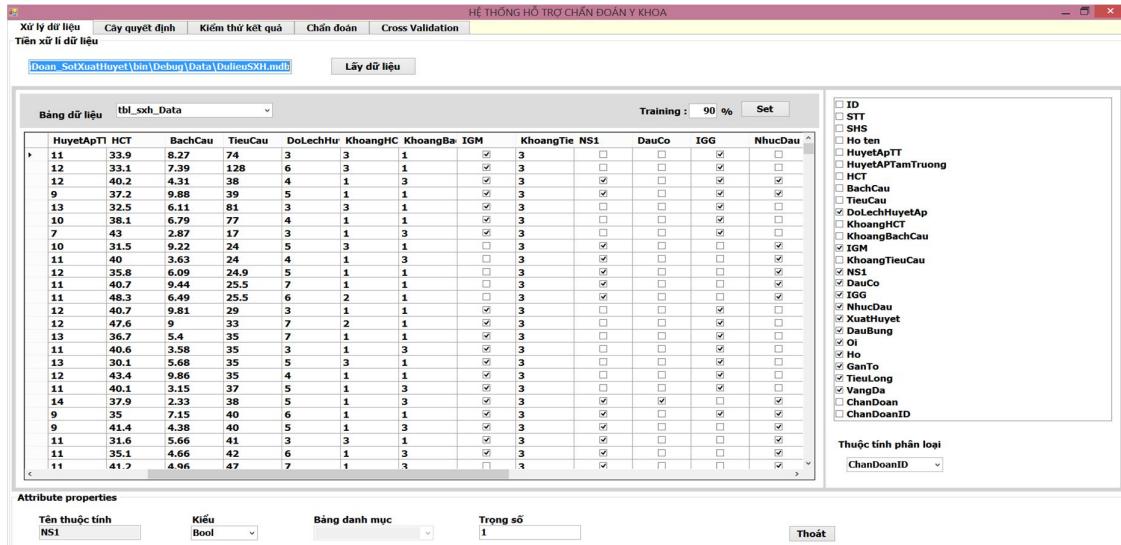
**Bước 2:** Chọn đường dẫn đến kho dữ liệu bệnh án điện tử bằng cách click nút “Lấy dữ liệu” và chọn file dữ liệu để đưa vào chương trình.



Hình 3.6 Màn hình chọn file dữ liệu

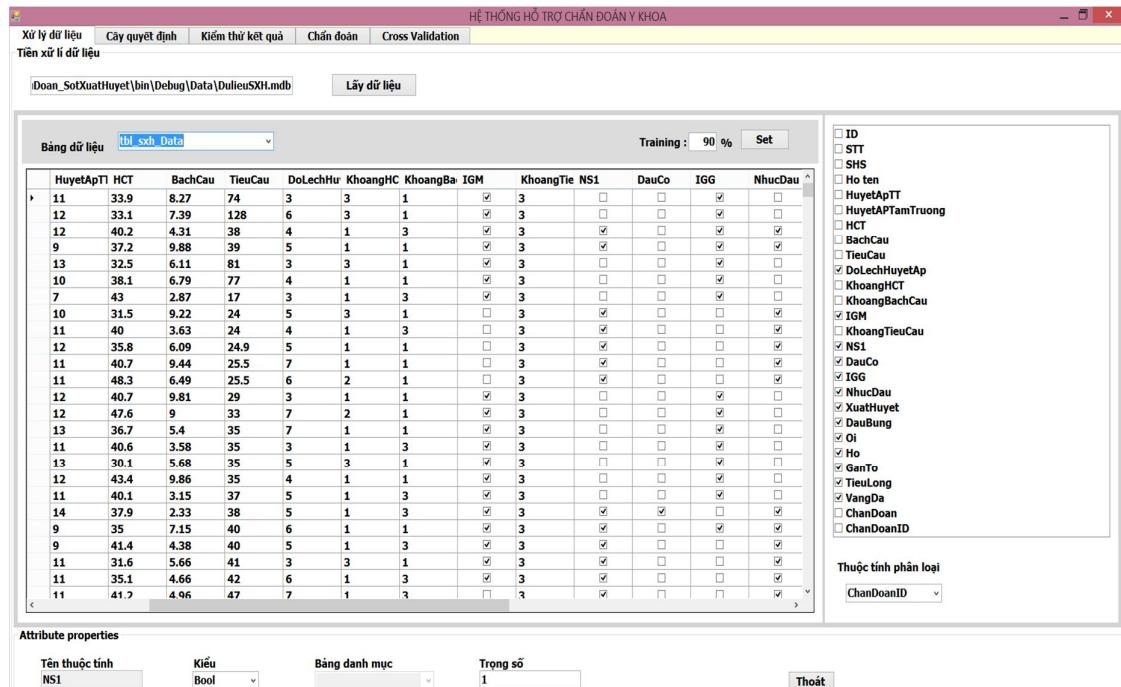
### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa

Chọn file DulieuSXH.mdb, click Open



Hình 3.7: Màn hình chọn dữ liệu từ kho dữ liệu

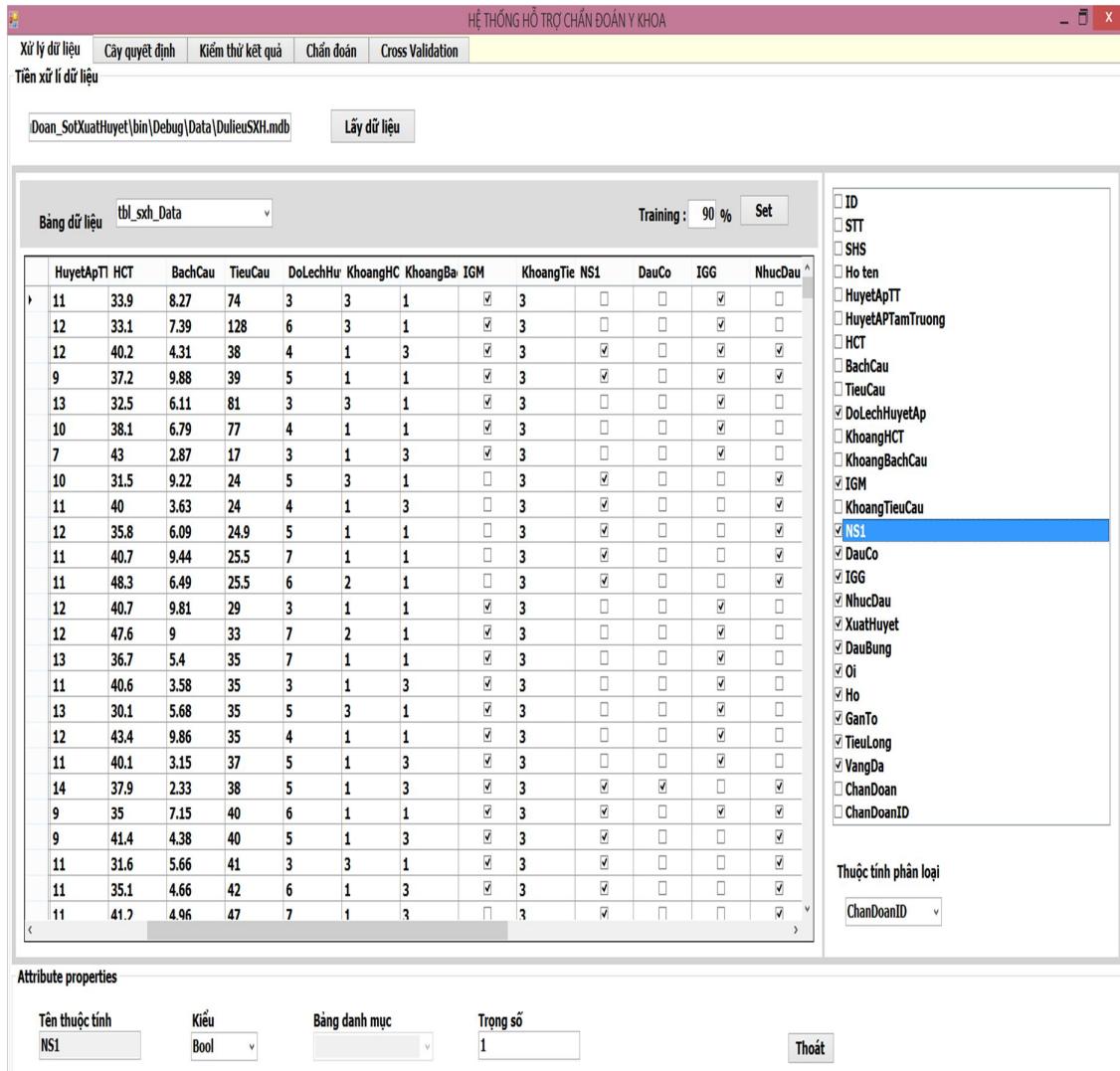
**Bước 3:** Chọn bảng dữ liệu (Ví dụ: chọn bảng tbl\_sxh\_Data)



Hình 3.8: Màn hình chọn bảng dữ liệu

Khi chọn bảng dữ liệu xong, toàn bộ dữ liệu sẽ được đưa vào hệ thống.

**Bước 4:** Chọn thuộc tính để xây dựng cây quyết định



Hình 3.9: Màn hình chọn thuộc tính

Chọn các thuộc tính như: DoLechHuyetAp, IgM, IgG, NS1,...

Độ lệch huyết áp (DoLechHuyetAp ) là hiệu số giữa huyết áp tâm trương và huyết áp tâm thu. Ví dụ: Huyết áp tâm trương là 11, huyết áp tâm thu 7 vậy độ lệch huyết áp là 4. Khi chọn thuộc tính, ta cần chú ý chọn kiểu dữ liệu (bảng 3.2) cho phù hợp.

**Bước 5:** Chọn số lượng dữ liệu ( **% training** ) để máy học và click nút “**SET**”

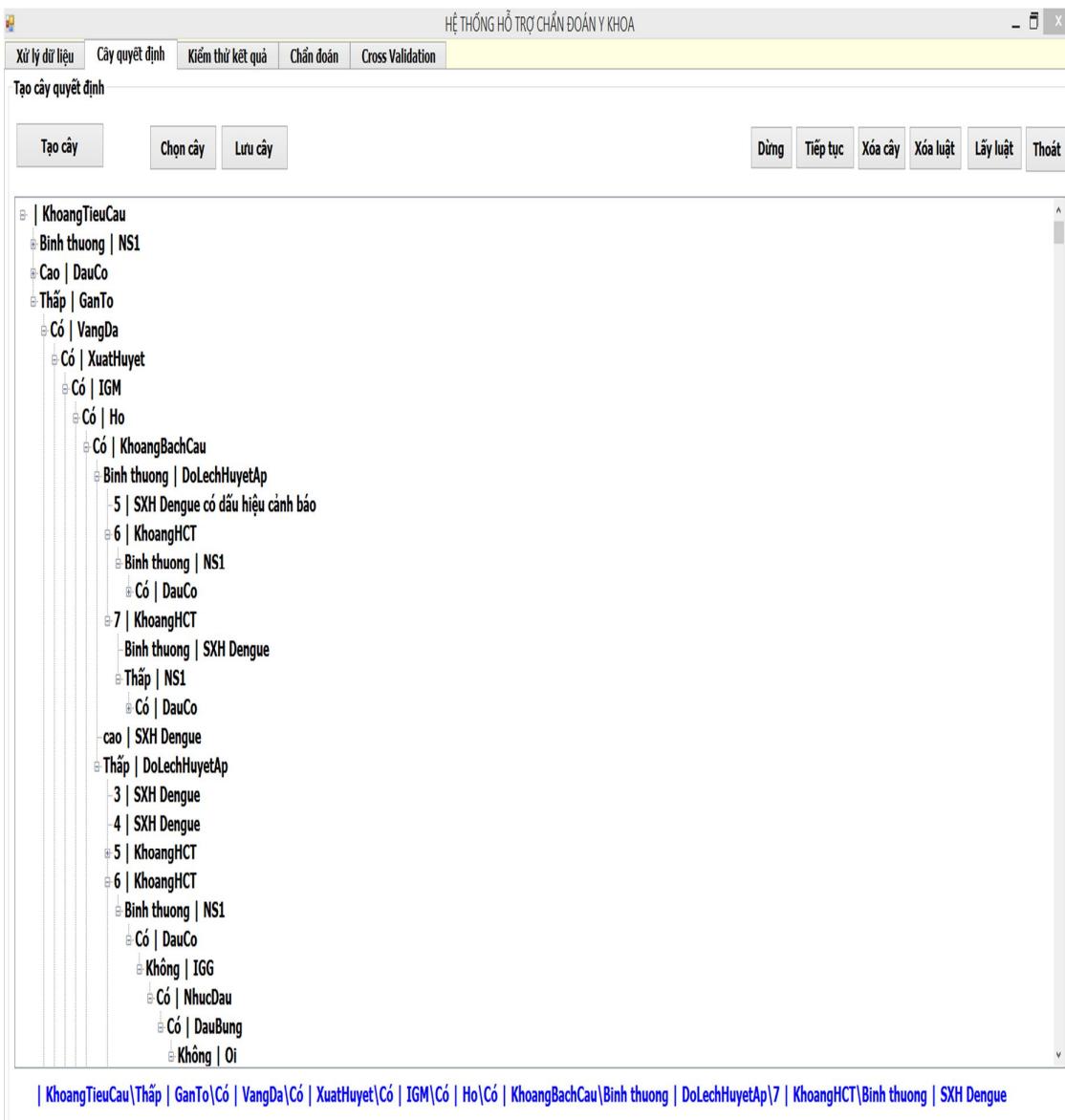
**Bước 6:** Chọn tab “**Cây quyết định**”, click nút “**Tạo cây**” để tiến hành xây dựng cây quyết định từ dữ liệu được chọn ở bước 4 và 5. Dữ liệu được tải từ kho bệnh án điện tử và các cấu hình của thuộc tính được chọn để xây dựng cây quyết định.



Hình 3.10 Màn hình tạo cây quyết định

Click nút “**Tạo cây**”, chương trình thực hiện thuật toán phân lớp để tạo ra cây quyết định như hình 3.11.

### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa

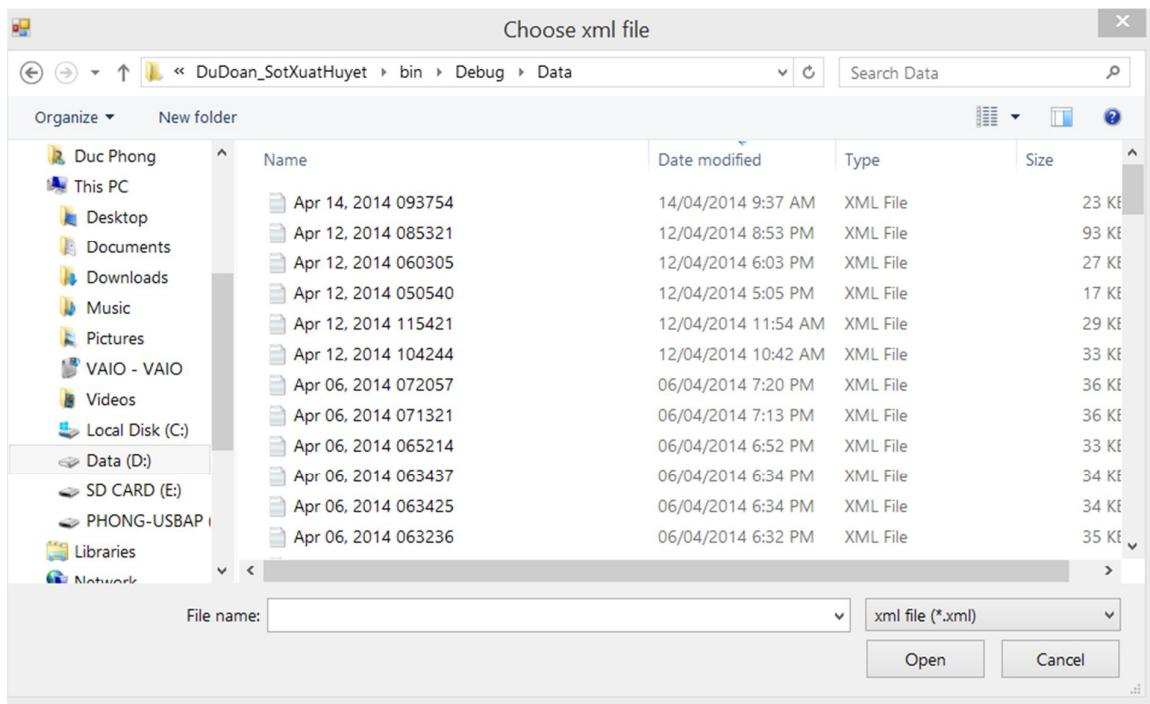


Hình 3.11: Màn hình tạo cây quyết định

Sau khi xây dựng xong cây quyết định chương trình sẽ tự động lưu lại cây ở dạng file XML để sau này có thể tải lên chương trình sử dụng lại. Điều này sẽ giảm thời gian học cho máy và máy chỉ học khi nào có sự cập nhật dữ liệu. Double click vào bất kỳ nút lá nào của cây hệ thống sẽ xuất hiện luật tương ứng với nút lá đó.

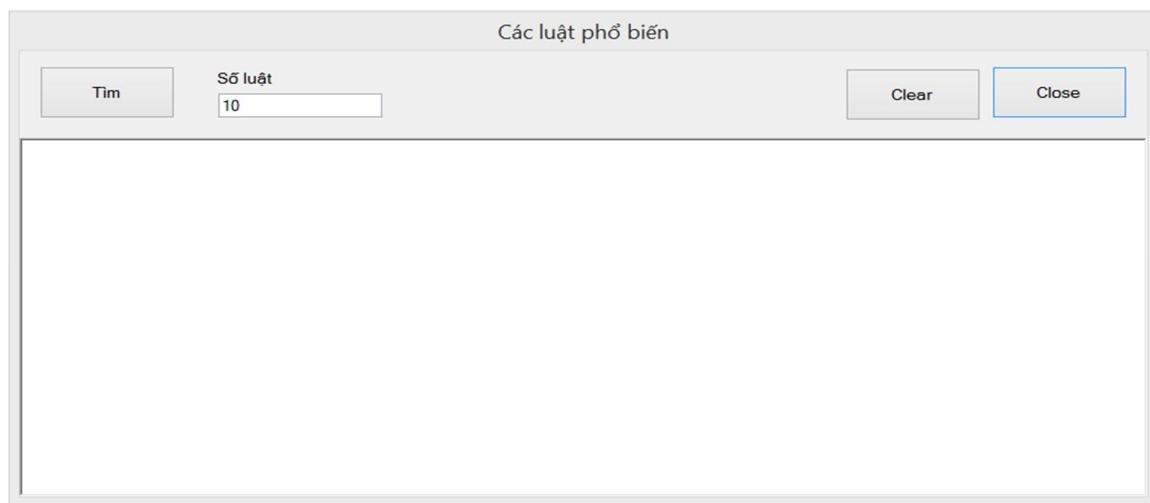
Click vào nút “**Chọn cây**” để lấy một cây quyết định đã được tạo trước đây (lưu dưới dạng xml) để phân lớp cho các thuộc tính mới.

### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa



Hình 3.12 Màn hình lấy cây đã lưu dạng xml

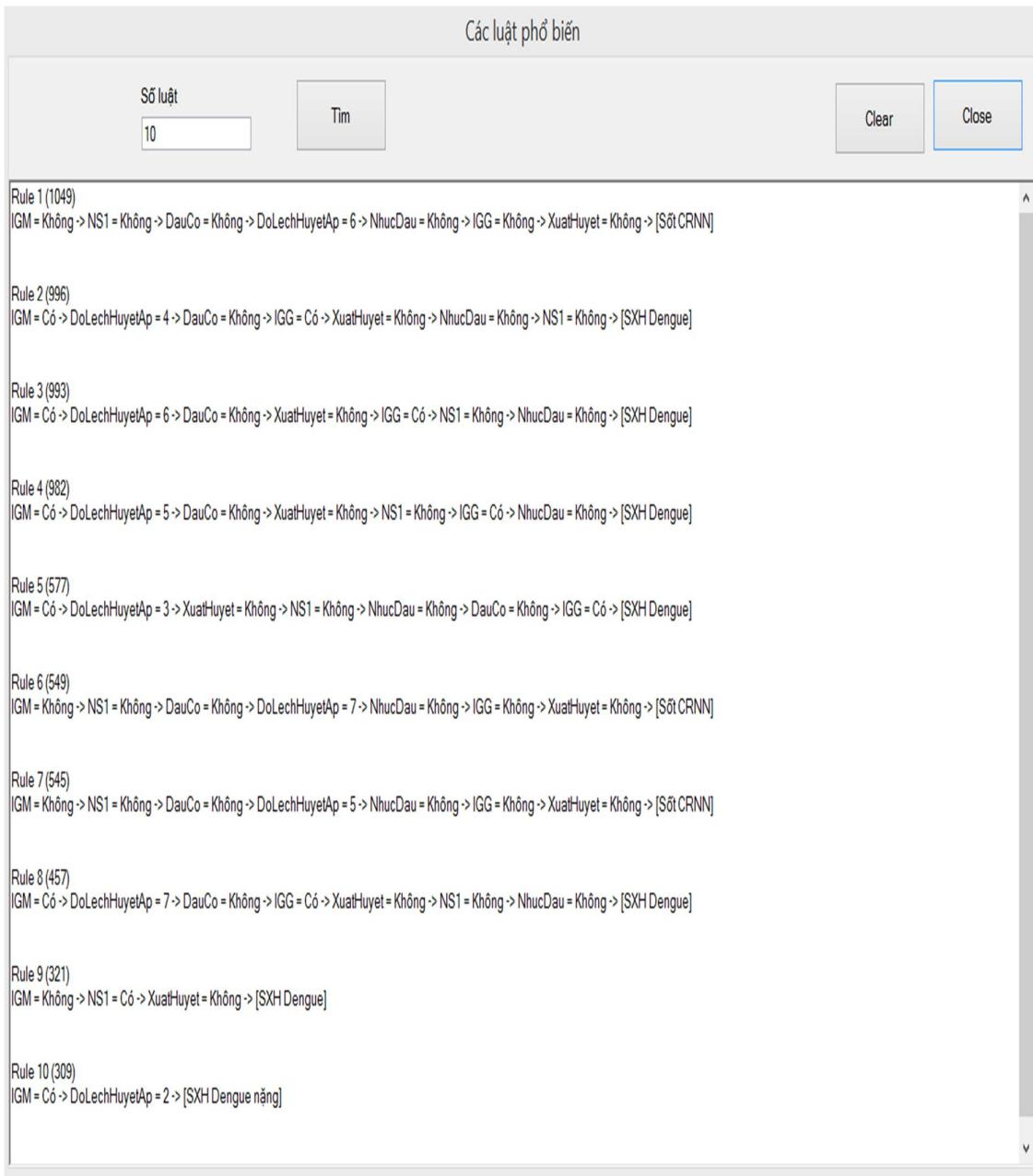
Click vào nút “Lấy luật“ để xem tập luật được rút sau khi huấn luyện.



Hình 3.13 Màn hình thống kê tập luật của tập dữ liệu

Nhập vào số luật cần thống kê vào ô “**Số luật**” và click nút “**Tìm**”. Ví dụ ta cần tìm 10 luật, chương trình sẽ thống kê 10 luật với số lần xuất hiện nhiều nhất trong tập luật và liệt kê ra màn hình.

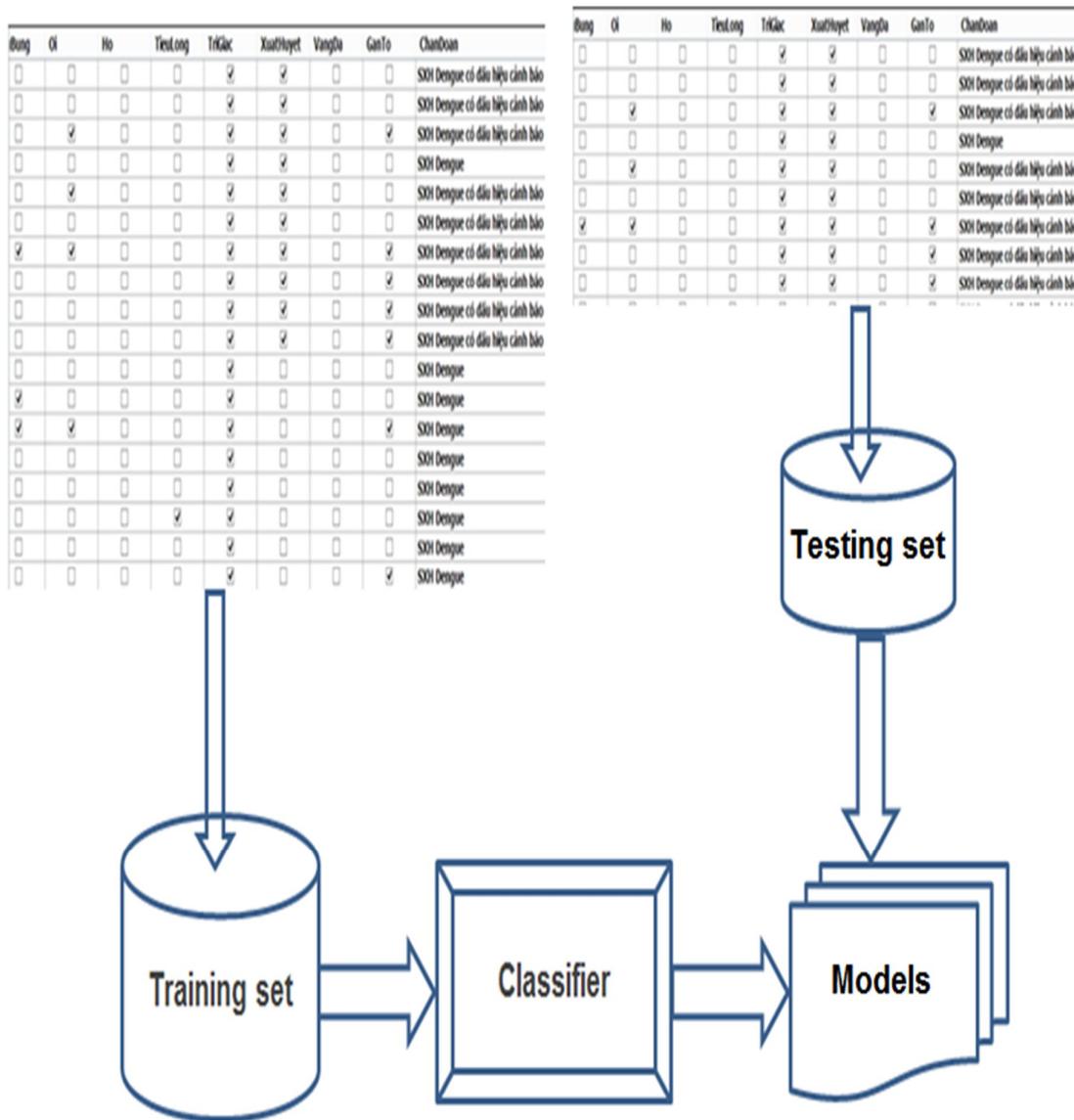
### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa



Hình 3.14 màn hình thông kê 10 luật

**Bước 7:** Chọn tab “**Kiểm thử kết quả**” để tiến hành đánh giá mô hình vừa được xây dựng. Kiểm tra kết quả ta phải có dữ liệu kiểm tra (Test set). Tập dữ liệu này có thể import từ file Excel hoặc từ số lượng dữ liệu gốc chưa được huấn luyện.

Qui trình huấn luyện và kiểm tra có mô hình như sau:



Hình 3.15: Mô hình kiểm tra kết quả

Nếu dùng 70% dữ liệu gốc để huấn luyện (training set) thì dữ liệu kiểm tra (testing set) còn 30%, khi đó click vào nút “**Test Data**” trong Tab “**Kiểm thử kết quả**” để tiến hành kiểm tra. Kết quả kiểm tra cho ta biết tỉ lệ phần trăm đúng với mô hình vừa tạo được, qua đó đánh giá được mô hình cây quyết định vừa xây dựng.

### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa



Hình 3.16 Màn hình kiểm tra kết quả

Click nút “**30 % Data test**” để tiến hành kiểm tra.

STT	SHS	Ho ten	HuyetApT1	HuyetAPTh	HCT	BachCau	TieuCau	DoLechHu	KhoangHC	KhoangBa	KhoangTie	NS1	IGM	IGG	NhucDau	DauCo	XuatHuyet
92	13.18767	CHÂU TI...	7	6	37	5.38	14	2	1	1	3	False	True	True	False	False	False
93	13.18813	ĐÔ THỊ ...	9	6	38.5	2.2	14	3	1	3	3	False	True	True	False	False	False
94	13.18813	ĐÔ THỊ ...	5	4	42.8	5.21	14	2	1	1	3	False	True	True	False	False	False
95	13.18983	ĐINH T...	9	8	44.2	2.17	15	2	1	3	3	False	True	True	False	False	False
96	13.18983	ĐINH T...	9	7	32	1.32	15	2	3	3	3	False	True	True	False	False	False
97	13.18983	ĐINH T...	8	6	41.1	4.95	16	2	1	3	3	False	True	True	False	False	False
98	13.18983	ĐINH T...	7	5	47.7	3.41	16	2	2	3	3	False	True	True	False	False	False
99	13.18983	ĐINH T...	9	7	42.1	4.39	16	2	1	3	3	False	True	True	False	False	False
100	13.19039	ĐINH HI...	7	5	45.6	1.66	16	2	2	3	3	False	True	True	False	False	False
101	13.19039	ĐINH HI...	6	5	44.3	11.09	16	2	1	2	3	False	True	True	False	False	False
102	13.19039	ĐINH HI...	8	5	41.7	5.15	16	3	1	1	3	False	True	True	False	False	False
103	13.19046	ĐÀO THI...	8	5	39.2	6.36	16	3	1	1	3	False	True	True	False	False	False
104	13.19046	ĐÀO THI...	7	39.7	4	16	2	1	3	3	False	True	True	False	False	False	False
105	13.19131	BÙI THỊ ...	7	5	41.8	3.96	17	2	1	3	3	False	True	True	False	False	False
106	13.19131	BÙI THỊ ...	8	6	38.6	3.53	17	2	1	3	3	False	True	True	False	False	False
107	13.19501	ĐÀO KI...	13	8	35.6	7.71	20	5	1	1	3	True	True	True	False	True	True
108	13.19501	ĐÀO KI...	5	3	47	2.24	20	2	2	3	3	False	True	True	False	False	False
109	13.19501	ĐÀO KI...	11	6	43.6	7.13	20	5	1	1	3	True	True	True	False	True	True
110	13.19501	ĐÀO KI...	11	5	44.8	8.62	20	6	1	1	3	True	True	True	False	True	True
111	13.19529	ĐÔ THỊ ...	11	6	35.7	12.08	21	5	1	2	3	True	False	False	True	False	True
112	13.19529	ĐÔ THỊ ...	12	7	42.2	6.4	21	5	1	1	3	True	True	True	False	True	True
113	13.19529	ĐÔ THỊ ...	13	8	40.7	4.64	21	5	1	3	3	True	True	False	True	False	True
114	13.19529	ĐÔ THỊ ...	10	5	39.6	4.7	21	5	1	3	3	True	True	False	True	False	True
115	13.19743	ĐÀNG T...	13	8	41.2	6.19	22	5	1	1	3	True	False	False	True	False	True
116	13.19743	ĐÀNG T...	13	7	43.9	6.92	22	6	1	1	3	True	False	False	True	False	True
117	13.19939	ĐÔ THỊ ...	10	6	37.9	4.16	23	4	1	3	3	True	False	False	True	True	True

Thông kê kết quả  
Tổng số dữ liệu: 899  
Kết quả đúng: 887 (98%)

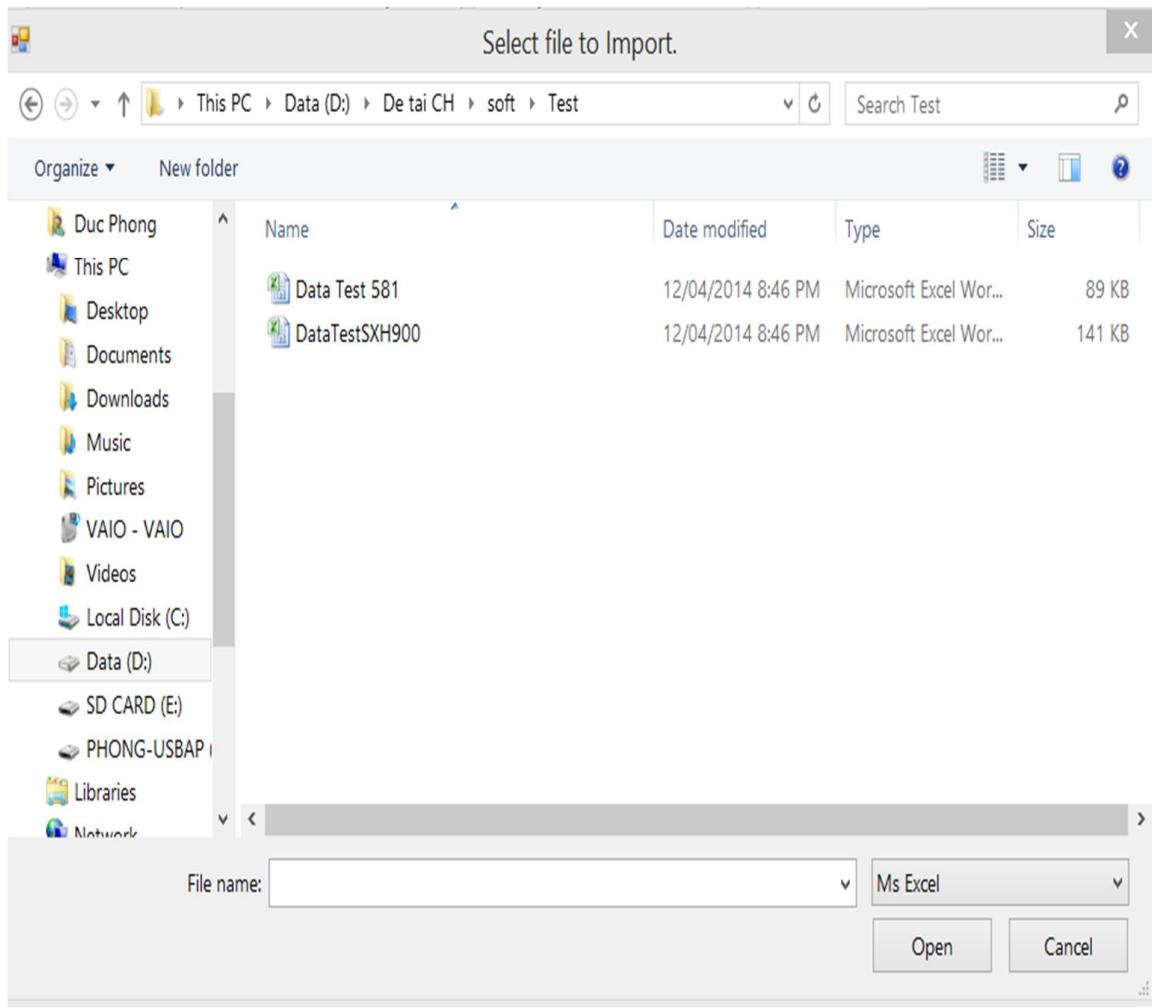
Hình 3.17 Kết quả kiểm tra dữ liệu

Ngoài ra có thể lấy tập dữ liệu kiểm tra từ file Excel để kiểm tra mô hình. Tập kiểm dữ liệu tra cũng được tiền xử lý để có cấu trúc giống như dữ liệu của tập huấn luyện. Kết quả cũng cho ta biết tỉ lệ phần trăm đúng ứng với mô hình đang xét của hệ thống.

Các bước lấy file Excel kiểm tra như sau :

1. Click “**Lấy File dữ liệu kiểm tra**” và chọn file Excel cần kiểm tra.
2. Click nút “**Kiểm tra file dữ liệu**”

Xem hình các bước thực hiện như sau :



Hình 3.18 Màn hình lấy file dữ liệu kiểm tra

### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa

Chọn File cần kiểm tra và click Open. Kết quả ở hình 3.19

HỆ THỐNG HỖ TRỢ CHẨN ĐOÁN Y KHOA																		
Xử lý dữ liệu	Cây quyết định	Kiểm thử kết quả	Chẩn đoán	Kiểm tra chéo (CV)														
Kiểm tra cây quyết định																		
10 % Data test				Lấy file dữ liệu kiểm tra	Kiểm tra file dữ liệu	Xóa màn hình												
<b>STT</b> <b>SHS</b> <b>Họ tên</b> <b>HuyetApTT</b> <b>HuyetAPTr</b> <b>HCT</b> <b>BachCau</b> <b>TieuCau</b> <b>DoLechHu</b> <b>KhoangHC</b> <b>KhoangBa</b> <b>KhoangTie NS1</b> <b>IGM</b> <b>IGG</b> <b>NhuCoDau</b> <b>DauCo</b> <b>XuatHuyet</b>																		
9	13.05362	TRẦN T...	11	7	40	3.63	24	4	1	3	3	True	False	False	True	False	True	True
17	13.05699	BÙI THI ...	13	8	30.1	5.68	35	5	3	1	3	False	True	True	False	False	False	False
29	13.05868	PHẠM T...	12	7	37.8	1.95	55	5	1	3	3	True	False	False	True	False	False	False
40	13.06109	LÊ THỊ T...	13	8	38.8	2.49	69	5	1	3	3	False	True	True	False	False	False	False
48	13.06144	DOAN T...	11	5	30.7	4.86	106	6	3	3	3	False	True	True	False	False	False	False
49	13.06144	DOAN T...	10	4	42.4	4.95	107	6	1	3	3	False	True	True	False	False	False	False
61	13.06222	NGUYỄN...	10	3	43.1	4.57	135	7	1	3	3	False	True	True	False	False	False	False
92	13.06436	BÙI THỊ ...	5	3	35.4	5.8	11	2	1	1	3	False	True	True	False	False	False	False
104	13.06545	DƯƠNG ...	6	4	44.1	6.27	13	2	1	1	3	False	True	True	False	False	False	False
112	13.06919	LÂM KL...	9	7	37.2	1.4	15	2	1	3	3	False	True	True	False	False	False	False
135	13.07254	HỒ KHÀ...	6	4	41.8	5.86	18	2	1	1	3	False	True	True	False	False	False	False
154	13.07965	NGÔ XU...	11	7	39.6	7.58	20	4	1	1	3	True	True	False	True	False	True	True
156	13.08194	NGUYỄN...	12	7	39.8	5.34	20	5	1	1	3	True	True	True	True	False	True	True
178	13.08676	TRƯỜNG...	10	4	38.7	6.1	23	6	1	1	3	True	False	False	True	False	True	True
179	13.08676	TRƯỜNG...	13	9	38.3	2.68	23	4	1	3	3	True	False	False	True	False	True	True
196	13.09008	THANG ...	12	7	44.1	2.88	26	5	1	3	3	False	True	True	False	False	False	False
240	13.09775	NGUYỄN...	9	5	37.4	3.23	31	4	1	3	3	False	True	True	False	False	False	False
252	13.10303	NGUYỄN...	13	8	41.8	5.41	33	5	1	1	3	False	True	True	False	False	False	False
272	13.10938	HÚA TH...	11	6	40.2	2.53	36	5	1	3	3	False	True	True	False	False	False	False
287	13.11699	HÀ THỊ ...	13	7	42.1	3.03	38	6	1	3	3	True	True	False	True	False	True	True
298	13.12124	LƯƠNG ...	13	9	36.4	7.39	41	4	1	1	3	True	True	False	True	False	True	True
309	13.12259	HỒ PHẠ...	11	6	38	1.76	42	5	1	3	3	True	True	True	True	False	True	True
315	13.12376	NGUYỄN...	9	3	37.1	2.4	43	6	1	3	3	True	True	True	True	False	True	True
324	13.12663	TRẦN T...	12	7	38.2	5.3	45	5	1	1	3	True	False	False	True	False	True	True
326	13.12667	NGUYỄN...	11	6	37.9	8.7	47	5	1	1	3	True	False	False	True	True	True	True
328	13.12748	HỒ THỊ ...	13	10	39.8	6.75	47	3	1	1	3	True	False	False	True	False	True	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Hình 3.19 Màn hình dữ liệu kiểm tra được lấy từ file Excel

Click nút “ Kiểm tra file dữ liệu”

HỆ THỐNG HỖ TRỢ CHẨN ĐOÁN Y KHOA																								
Xử lý dữ liệu		Cây quyết định		Kiểm thử kết quả		Chẩn đoán		Kiểm tra chéo (CV)																
Kiểm tra cây quyết định																								
10 % Data test					Lấy file dữ liệu kiểm tra					Kiểm tra file dữ liệu					Xóa màn hình									
STT	SHS	Ho ten	HuyetApT1	HuyetApT2	HCT	BachCau	TieuCau	DoLechHu	KhoangHC	KhoangBa	KhoangTie	NS1	IGM	IGG	NhuDau	DauCo	XuatHuyet							
9	13.05362	TRẦN T...	11	7	40	3.63	24	4	1	3	3	True	False	False	True	False	True							
17	13.05699	BÙI THỊ ...	13	8	30.1	5.68	35	5	3	1	3	False	True	True	False	False	False							
29	13.05868	PHẠM T...	12	7	37.8	1.95	55	5	1	3	3	True	False	False	True	False	False							
40	13.06109	LÊ THỊ T...	13	8	38.8	2.49	69	5	1	3	3	False	True	True	False	False	False							
48	13.06144	ĐOÀN T...	11	5	30.7	4.86	106	6	3	3	3	False	True	True	False	False	False							
49	13.06144	ĐOÀN T...	10	4	42.4	4.95	107	6	1	3	3	False	True	True	False	False	False							
61	13.06222	NGUYỄN...	10	3	43.1	4.57	135	7	1	3	3	False	True	True	False	False	False							
92	13.06436	BÙI THỊ ...	5	3	35.4	5.8	11	2	1	1	3	False	True	True	False	False	False							
104	13.06545	DƯƠNG ...	6	4	44.1	6.27	13	2	1	1	3	False	True	True	False	False	False							
112	13.06919	LÂM KI...	9	7	37.2	1.4	15	2	1	3	3	False	True	True	False	False	False							
135	13.07254	HỒ KHÁ...	6	4	41.8	5.86	18	2	1	1	3	False	True	True	False	False	False							
154	13.07965	NGÔ XU...	11	7	39.6	7.58	20	4	1	1	3	True	True	False	True	False	True							
156	13.08194	NGUYỄN...	12	7	39.8	5.34	20	5	1	1	3	True	True	True	False	True								
178	13.08676	TRƯỜNG...	10	4	38.7	6.1	23	6	1	1	3	True	False	False	True	False	True							
179	13.08676	TRƯỜNG...	13	9	38.3	2.68	23	4	1	3	3	True	False	False	True	False	True							
196	13.09008	THANG ...	12	7	44.1	2.88	26	5	1	3	3	False	True	True	False	False	False							
240	13.09775	NGUYỄN...	9	5	37.4	3.23	31	4	1	3	3	False	True	True	False	False	False							
252	13.10303	NGUYỄN...	13	8	41.8	5.41	33	5	1	1	3	False	True	True	False	False	False							
272	13.10938	HÙA TH...	11	6	40.2	2.53	36	5	1	3	3	False	True	True	False	False	False							
287	13.11699	HÀ THỊ ...	13	7	42.1	3.03	38	6	1	3	3	True	True	False	True	False	True							
298	13.12124	LƯƠNG ...	13	9	36.4	7.39	41	4	1	1	3	True	True	False	True	False	True							
309	13.12259	HỒ PHẠ...	11	6	38	1.76	42	5	1	3	3	True	True	True	True	False	True							
315	13.12376	NGUYỄN...	9	3	37.1	2.4	43	6	1	3	3	True	True	True	True	False	True							
324	13.12663	TRẦN T...	12	7	38.2	5.3	45	5	1	1	3	True	False	False	True	False	True							
326	13.12667	NGUYỄN...	11	6	37.9	8.7	47	5	1	1	3	True	False	False	True	True	True							
328	13.12748	HỒ THỊ ...	13	10	39.8	6.75	47	3	1	1	3	True	False	False	True	False	True							
329	13.12770	YẾN THỊ ...	10	6	39.8	6.75	47	3	1	1	3	True	False	False	True	False	True							

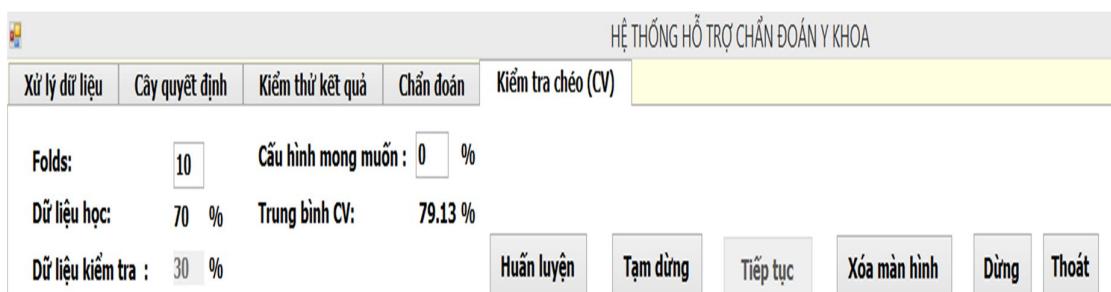
Hình 3.20 Màn hình kết quả kiểm tra dữ liệu từ file Excel

Chương trình sẽ cho ta biết :

- Tổng số dữ liệu được kiểm tra.

- Số kết quả đúng (%).

Trường hợp xây dựng mô hình và kiểm tra chéo bằng công cụ Cross validation thì ta cũng thực hiện các bước từ 1 đến 5, sau đó chọn tab “**Kiểm tra chéo (CV)**”. Màn hình kiểm tra chéo như sau :

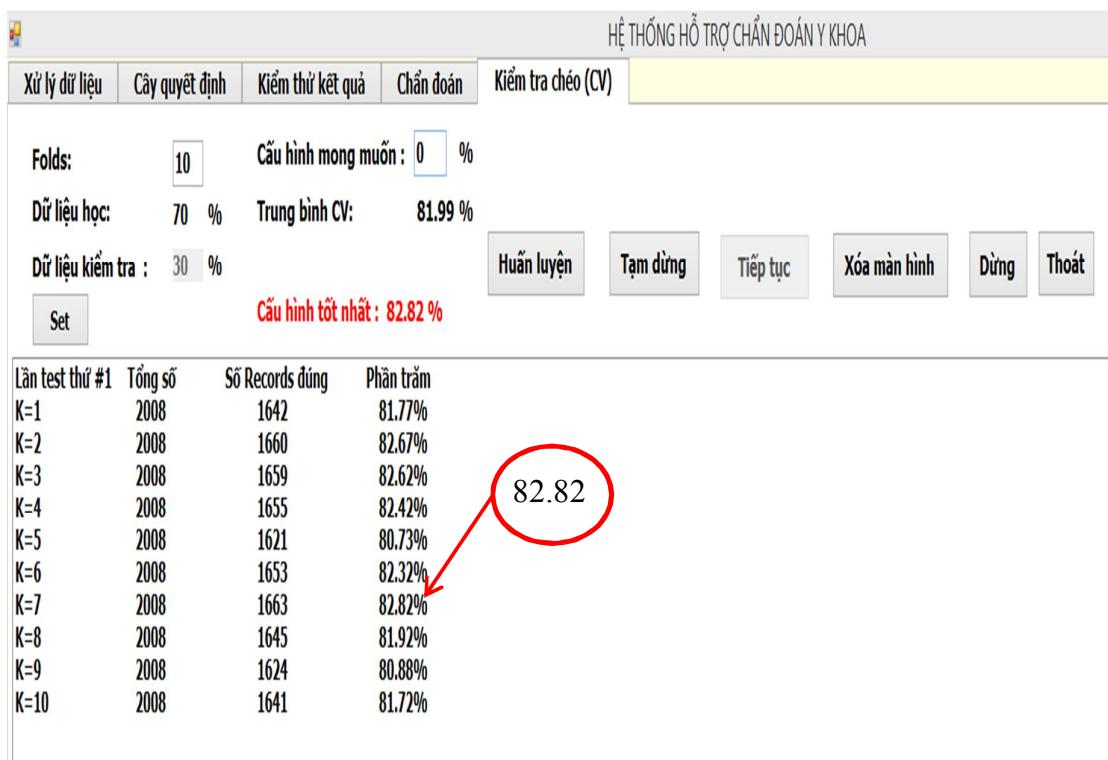


Hình 3.21 Màn hình kiểm tra chéo (Cross validation)

Nhập số k vào ô Folds.

Nhập phần trăm dữ liệu học (huấn luyện) vào ô “Dữ liệu học”

Click nút “**Huấn luyện**” để tiến hành thực hiện Cross validation. Mô hình có kết quả kiểm tra cao nhất sẽ được chọn.

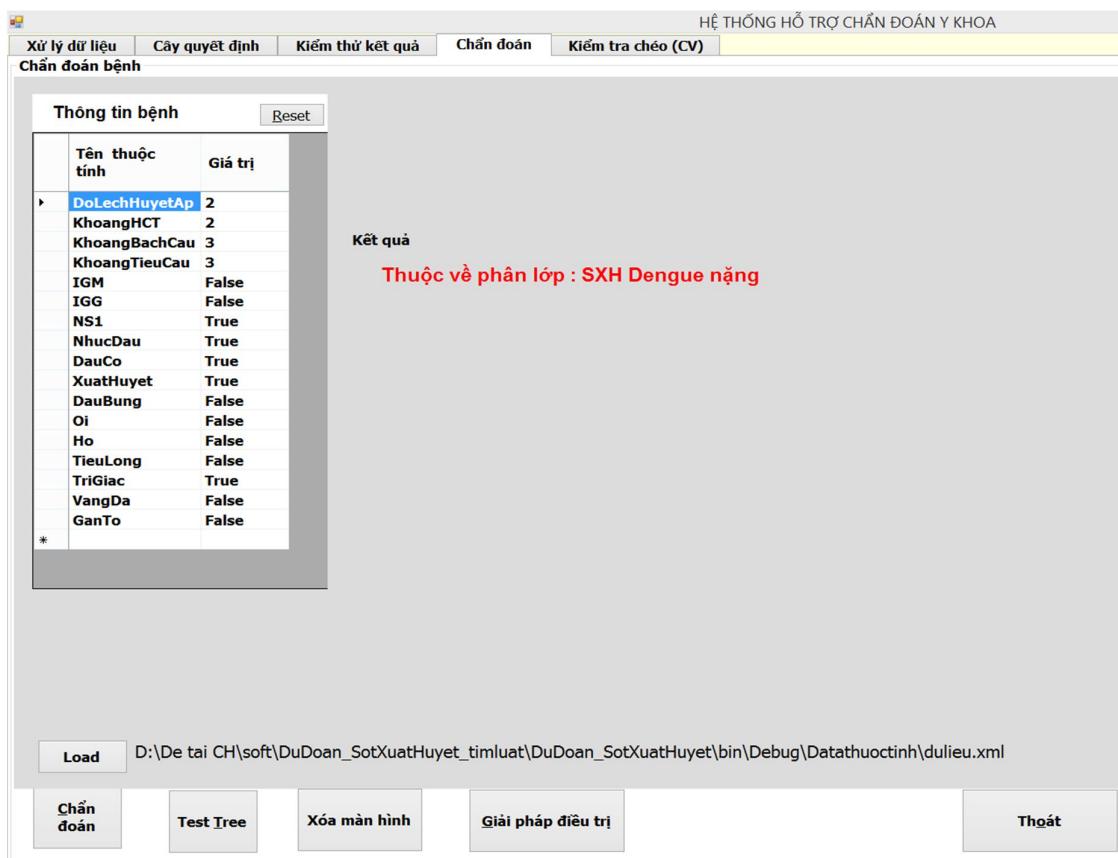


Hình 3.22 Màn hình kết quả kiểm tra chéo (Cross validation)

Kết quả 82.82% là tỉ lệ tốt nhất nên sẽ được chọn làm mô hình trong đợt huấn luyện này.

**Bước 8:** Chọn tab “**Chẩn đoán**”. Danh sách các thông tin (thuộc tính) dùng để chẩn đoán đã được thiết lập sẵn tương ứng với các thuộc tính đã được chọn ở bước 4 (các thông tin này được thiết lập sẵn và lưu thành file dulieu.xml, file này chứa cấu trúc bao gồm các thuộc tính của một bệnh án mẫu). Nhập vào kết quả lâm sàng và cận lâm sàng cần kiểm tra của bệnh nhân hoặc bệnh án mẫu vào cột “Giá trị” tương ứng với cột “Tên thuộc tính”, chương trình sẽ dựa vào mô hình của cây quyết định vừa xây dựng để phân lớp (chẩn đoán) cho dữ liệu cần kiểm tra. Việc thiết kế thông tin mẫu của bệnh nhân bằng file XML rất linh động. Nếu sau này cấu trúc bệnh án điện tử có thay đổi, chỉ cần hiệu chỉnh lại file XML cho phù hợp là chương trình có thể vận hành bình thường mà không cần phải hiệu chỉnh lại mã nguồn của chương trình.

### Chương 3 : Xây dựng hệ hỗ trợ chẩn đoán y khoa



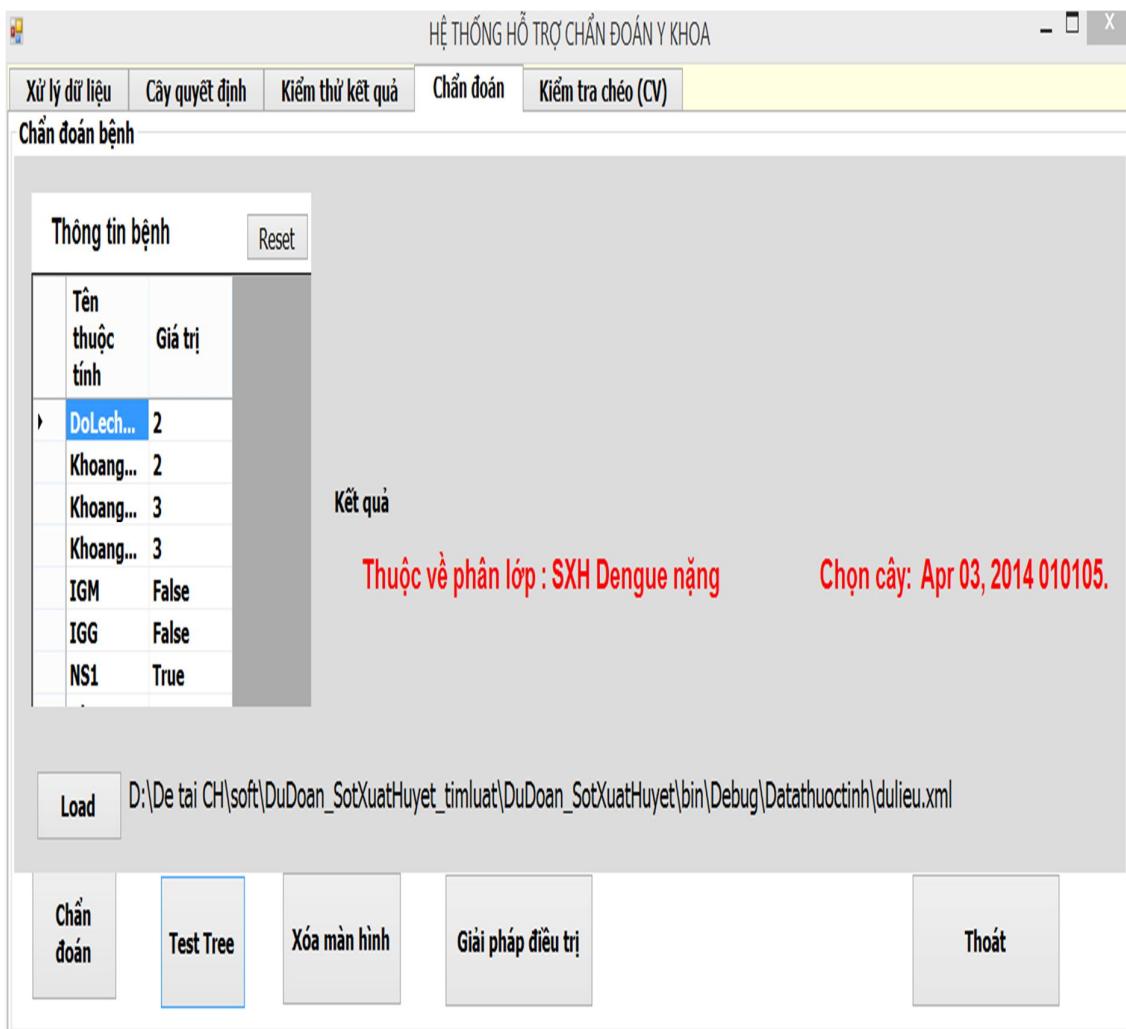
Hình 3.23: Màn hình chẩn đoán bệnh

Màn hình chẩn đoán có 2 phần: “**Tên thuộc tính**” và phần nhập “**Giá trị**”. Ví dụ nhập vào ô giá trị của các thuộc tính:

- + Độ lệch huyết áp là 2,3,4,5
- + Các triệu chứng: Nhức đầu, đau cơ, đau bụng, ói chọn true hoặc false

Sau khi nhập phần giá trị cho các thuộc tính, click vào nút “**Chẩn đoán**” để phân lớp các thuộc tính vừa nhập.

Trường hợp không tạo cây quyết định mà muốn sử dụng các cây quyết định đã tạo trước đây thì ta click nút “**Test tree**”. Khi đó các thuộc tính (bệnh án mới) muốn kiểm tra sẽ được dò tìm trong các cây trước đây và cho kết quả phân lớp. Việc này máy thực hiện rất nhanh. Kết quả chẩn đoán trường hợp này như hình 3.24



Hình 3.24 Kết quả chẩn đoán từ cây có sẵn

Sau khi chẩn đoán, chương trình sẽ cho ta biết kết quả phân lớp và được dùng mô hình nào (cây quyết định đã được lưu)

**Bước 9:** Sau khi hệ thống đã cho kết quả phân lớp (chẩn đoán), click vào tab “**Giải pháp điều trị**” để mở phác đồ điều trị tương ứng với chẩn đoán của hệ thống (các file phác đồ điều trị được soạn sẵn tương ứng với các chẩn đoán và được lưu trữ ở dạng file word).

Ví dụ phác đồ điều trị SXH Dengue [19]:

Phần lớn các trường hợp đều được điều trị ngoại trú và theo dõi tại y tế cơ sở, chủ yếu là điều trị triệu chứng và phải theo dõi chặt chẽ phát hiện sớm sốc xảy ra để xử trí kịp thời.

1. Điều trị triệu chứng:

- Nếu sốt cao  $\geq 39^{\circ}\text{C}$ , cho thuốc hạ nhiệt, nới lỏng quần áo và lau mát bằng nước ấm.
- Thuốc hạ nhiệt chỉ dùng Paracetamol đơn chất, liều dùng 10 – 15mg/kg cân nặng/lần cách nhau mỗi 4-6 giờ.
- Chú ý:
  - + Tổng liều không quá 60mg/kg cân nặng / 24 giờ.
  - + Không dùng aspirin (acetyl salicylic acid), analgin, ibuprofen để điều trị vì có thể gây xuất huyết, toan huyết.

2. Bù dịch sớm bằng đường uống:

Khuyến khích người bệnh uống nhiều nước Oresol hoặc nước sôi để nguội, nước trái cây (nước dù, cam, chanh,...) hoặc nước cháo loãng với muối.

## THỐNG KÊ THỜI GIAN PHÂN LỚP VÀ THỜI GIAN HỖ TRỢ CHẨN ĐOÁN

Số thuộc tính	Thời gian phân lớp	Số luật	Kết quả kiểm tra	Thời gian hỗ trợ chẩn đoán
16	3 phút 14 giây	341	98%	< 1 giây
10	1 phút 13 giây	276	95%	< 1 giây
8	39 giây	226	93%	< 1 giây

Bảng 3.3: Bảng thống kê kết quả thực hiệu ứng dụng

Bảng thống kê thể hiện rõ ràng nếu bệnh nhân cung cấp cho nhân viên y tế đầy đủ các triệu chứng lâm sàng đồng thời nhân viên y tế thực hiện các cận lâm sàng cần thiết cho bệnh nhân thì kết quả chẩn đoán càng cao (trường hợp số thuộc tính là 16 và kết quả kiểm tra (hỗ trợ chẩn đoán) là 98%). Nếu số thuộc tính chỉ là 10 thì kết quả giảm xuống còn 95%). Qua đó ta thấy phù hợp với thực tế.

## Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Luận văn sẽ tiến hành thực nghiệm nhằm kiểm chứng và đánh giá các phương pháp thực hiện cũng như các kết quả thực tế thu thập được.

Việc đánh giá tính hiệu quả của chương trình hoàn toàn dựa vào các thông tin thu thập được trong suốt quá trình thực nghiệm. việc đánh giá này dựa trên một số tiêu chí:

- + Tính tiện dụng của ứng dụng.
- + Độ chính xác của thông tin thu thập được.
- + Rút ngắn thời gian chẩn đoán bệnh.

### 4.1. Thử nghiệm

#### 4.1.1. Thử nghiệm tập dữ liệu với ít thuộc tính:

Với một số thuộc tính về lâm sàng và cận lâm sàng (độ lệch huyết áp; khoảng tiêu cầu; khoảng HCT; khoảng bạch cầu; xuất huyết; NS1; IGM) trong tập dữ liệu. Một số luật được khai phá như sau:

- + Nếu KhoangTieuCau = “Thấp”  $\wedge$  DoLechHuyetAp = 5  $\wedge$  IGM = “Có”  $\wedge$  XuatHuyet = “Không”  $\wedge$  KhoangHCT = “Bình thường”  $\wedge$  KhoangBachCau = “Thấp”  $\wedge$  NS1 = “Không”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”

- + Nếu KhoangTieuCau = “Thấp”  $\wedge$  DoLechHuyetAp = 6  $\wedge$  IGM = “Không”  $\wedge$  NS1 = “Có”  $\wedge$  XuatHuyet = “Không”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”
- + Nếu KhoangTieuCau = Thấp  $\wedge$  DoLechHuyetAp = 4  $\wedge$  XuatHuyet = Có  $\wedge$  IGM = Không  $\wedge$  KhoangBachCau = Thấp  $\wedge$  KhoangHCT = Bình thường  $\wedge$  NS1 = Có  $\rightarrow$  Thuộc phân lớp “SXH Dengue”
- + Nếu KhoangTieuCau = Thấp  $\wedge$  DoLechHuyetAp = 3  $\wedge$  KhoangHCT = Thấp  $\wedge$  XuatHuyet = Không  $\wedge$  NS1 = Không  $\wedge$  KhoangBachCau = Bình thường  $\wedge$  IGM = Có  $\rightarrow$  Thuộc phân lớp “SXH Dengue”
- + Nếu KhoangTieuCau = Thấp  $\wedge$  DoLechHuyetAp = 3  $\wedge$  KhoangHCT = Cao  $\wedge$  IGM = Có  $\wedge$  XuatHuyet = Không  $\wedge$  KhoangBachCau = Bình thường  $\wedge$  NS1 = Không  $\rightarrow$  Thuộc phân lớp “SXH Dengue nồng”
- + Nếu KhoangTieuCau = Thấp  $\wedge$  DoLechHuyetAp = 6  $\wedge$  IGM = Có  $\wedge$  KhoangHCT = Cao  $\wedge$  KhoangBachCau = Bình thường  $\wedge$  NS1 = Không  $\wedge$  XuatHuyet = Không  $\rightarrow$  Thuộc phân lớp “SXH Dengue có dấu hiệu cảnh báo”
- + Nếu KhoangTieuCau = Thấp  $\wedge$  DoLechHuyetAp = 3  $\wedge$  KhoangHCT = Bình thường  $\wedge$  XuatHuyet = Có  $\wedge$  IGM = Không  $\wedge$  KhoangBachCau = cao  $\wedge$  NS1 = Có  $\rightarrow$  Thuộc phân lớp “SXH Dengue có dấu hiệu cảnh báo”

#### **4.1.2. Thủ nghiệm với tập dữ liệu đầy đủ thuộc tính**

Với đầy đủ các thuộc tính về biểu hiện lâm sàng và cận lâm sàng trong tập dữ liệu, một số luật được khai phá như sau:

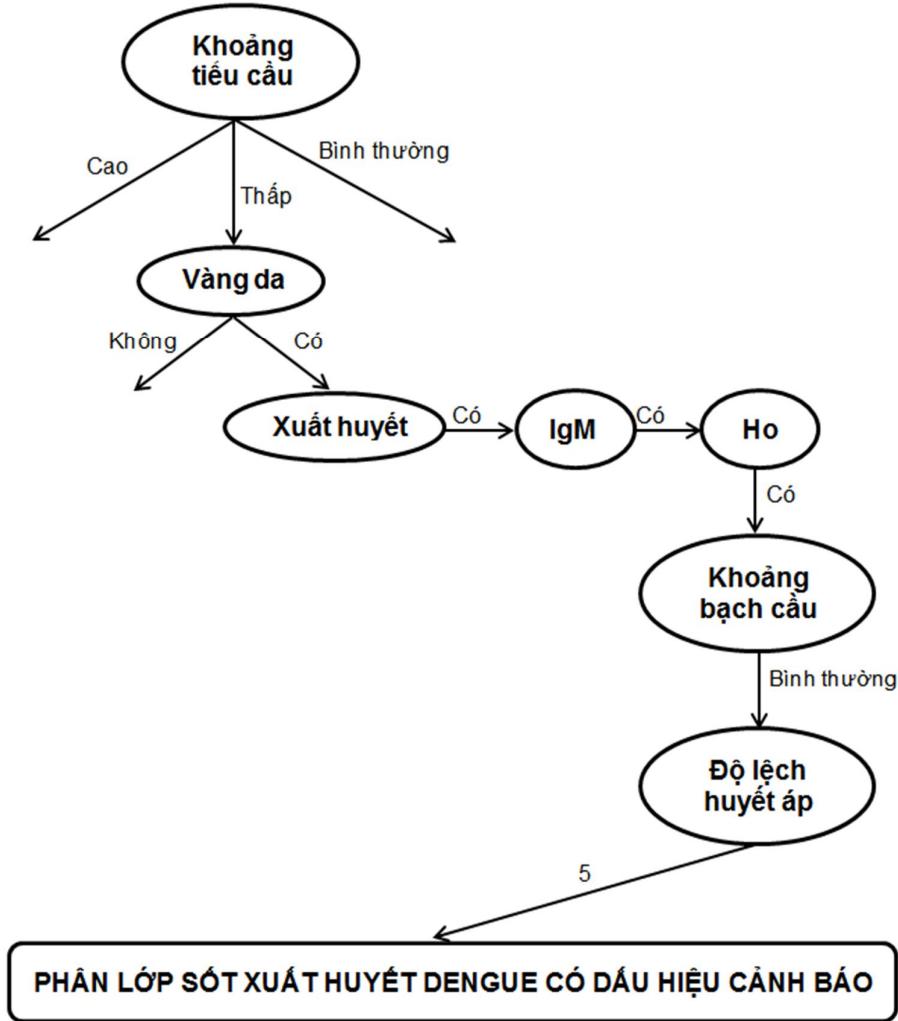
- + Nếu bệnh nhân có Khoangtieucau = “thấp”  $\wedge$  DoLechHuyetAp = 3  $\wedge$  KhoangHCT = “Bình thường”  $\wedge$  Xuathuyet = “Có”  $\wedge$  KhoangBachCau = “Bình thường”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”

- + Nếu KhoangTieuCau = “Tháp”  $\wedge$  vangda = “Có”  $\wedge$  XuatHuyet = “Có”  $\wedge$  IGM = “Có”  $\wedge$  Ho = “Có”  $\wedge$  KhoangBachCau = “Bình thường”  $\wedge$  DoLechHuyetAp = 5  $\rightarrow$  Thuộc phân lớp “SXH Dengue có dấu hiệu cảnh báo”
- + Nếu KhoangTieuCau = “Tháp”  $\wedge$  vangda = “Có”  $\wedge$  XuatHuyet = “Có”  $\wedge$  IGM = “Có”  $\wedge$  Ho = “Có”  $\wedge$  KhoangBachCau = “Bình thường”  $\wedge$  DoLechHuyetAp = 7  $\wedge$  KhoangHCT = “Bình thường”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”
- + Nếu KhoangTieuCau = “Tháp”  $\wedge$  vangda = “Có”  $\wedge$  XuatHuyet = “Có”  $\wedge$  IGM = “Có”  $\wedge$  Ho = “Có”  $\wedge$  KhoangBachCau = “cao”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”
- + Nếu KhoangTieuCau = “Tháp”  $\wedge$  VangDa = “Có”  $\wedge$  XuatHuyet = “Có”  $\wedge$  IGM = “Có”  $\wedge$  Ho = “Không”  $\wedge$  KhoangHCT = “Tháp”  $\wedge$  IGG = “Không”  $\wedge$  DoLechHuyetAp = 4  $\wedge$  KhoangBachCau = “Bình thường”  $\wedge$  NS1 = “Có”  $\wedge$  DauCo = “Không”  $\wedge$  NhucDau = “Có”  $\wedge$  DauBung = “Có”  $\wedge$  Oi = “Không”  $\wedge$  GanTo = “Có”  $\wedge$  TieuLong = “Không”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”
- + Nếu KhoangTieuCau = “Tháp”  $\wedge$  VangDa = “Có”  $\wedge$  XuatHuyet = “Có”  $\wedge$  IGM = “Không”  $\wedge$  KhoangBachCau = “Tháp”  $\wedge$  DauBung = “Không”  $\wedge$  KhoangHCT = “Bình thường”  $\wedge$  DoLechHuyetAp = 5  $\wedge$  Ho = “Không”  $\wedge$  NS1 = “Có”  $\wedge$  DauCo = “Không”  $\wedge$  IGG = “Không”  $\wedge$  NhucDau = “Có”  $\wedge$  Oi = “Không”  $\wedge$  GanTo = “Có”  $\wedge$  TieuLong = “Không”  $\rightarrow$  Thuộc phân lớp “SXH Dengue có dấu hiệu cảnh báo”

### Vẽ ví dụ cây quyết định

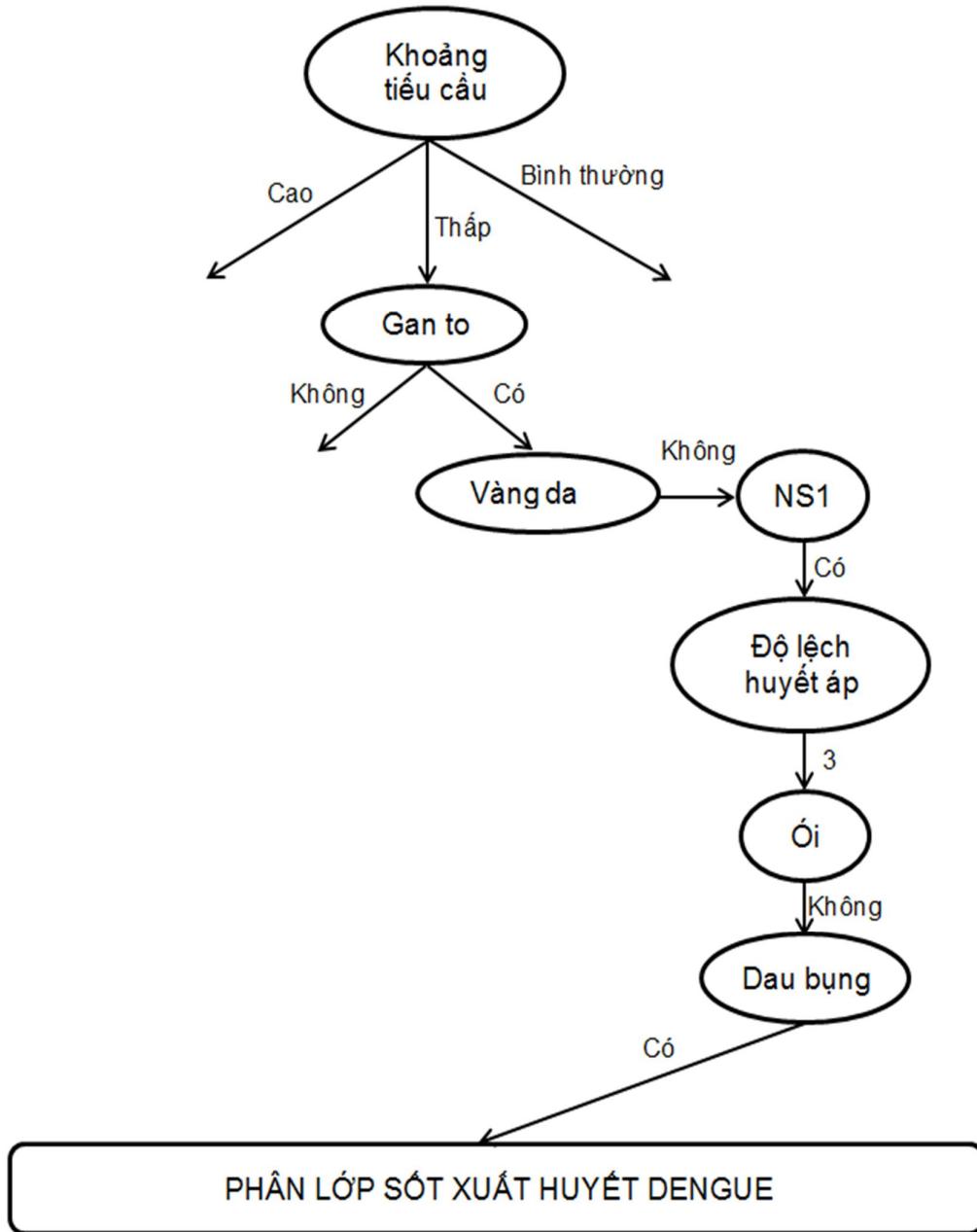
Vẽ phân lớp “SXH Dengue có dấu hiệu cảnh báo” với các thuộc tính: Nếu khoảng tiêu cầu = “Tháp”  $\wedge$  vàng da = “Có”  $\wedge$  Xuất huyết = “Có”  $\wedge$  IGM = “Có”  $\wedge$

$Ho = \text{"Có"} \wedge \text{khoảng bạch cầu} = \text{"Bình thường"} \wedge \text{độ lệch huyết áp} = 5 \rightarrow \text{Thuộc phân lớp "SXH Dengue có dấu hiệu cảnh báo"}$



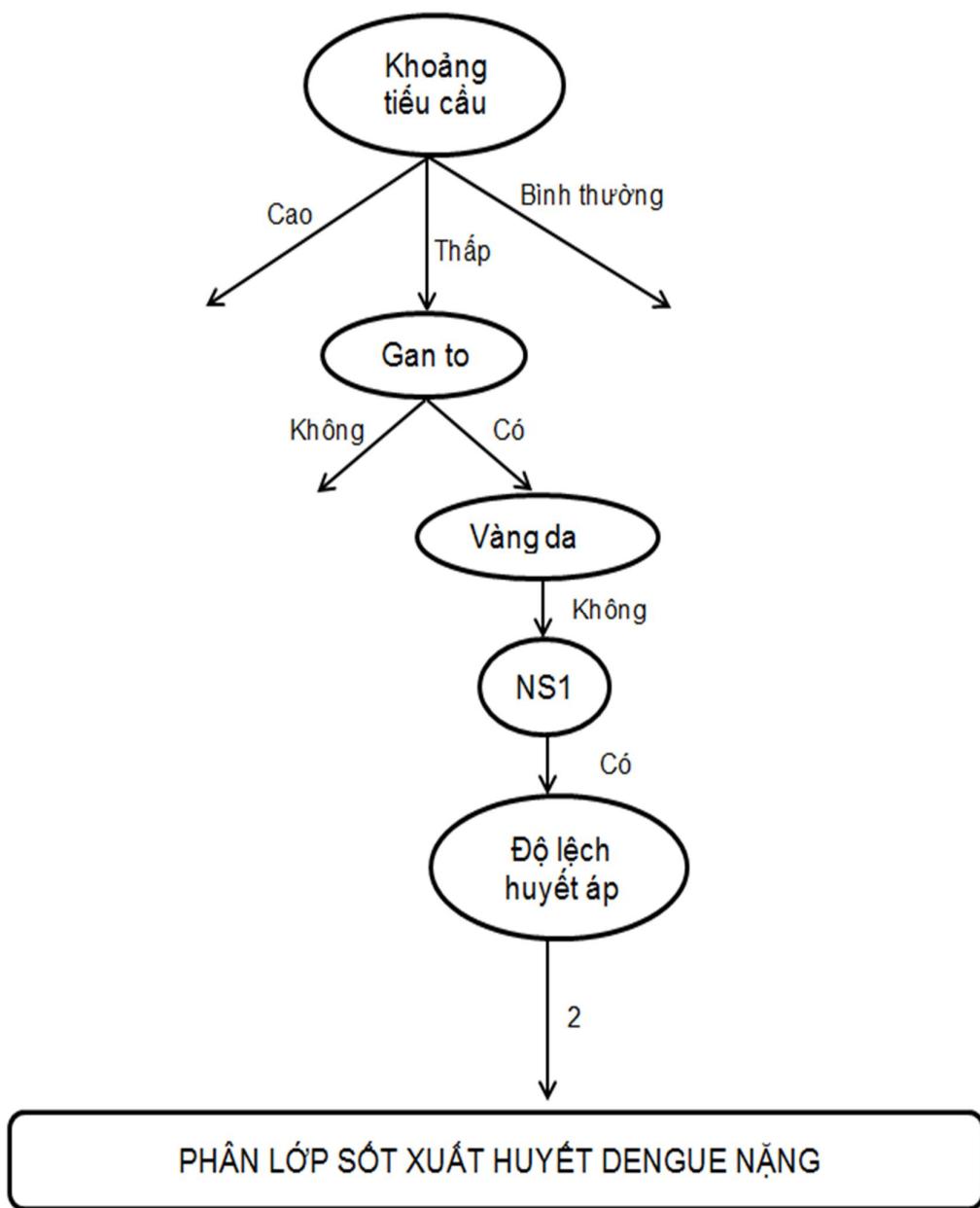
Hình 4.1: Hình vẽ SXH Dengue có dấu hiệu cảnh báo

Vẽ phân lớp “SXH Dengue” với các thuộc tính : Nếu khoảng tiêu cầu = “Thấp”  $\wedge$  Gan to = “Có”  $\wedge$  Vàng da = “Không”  $\wedge$  NS1 = “Có”  $\wedge$  độ lệch huyết áp = 3  $\wedge$  Ói = “Không”  $\wedge$  Đau bụng = “Có”  $\rightarrow$  Thuộc phân lớp “SXH Dengue”



Hình 4.2 hình vẽ SXH Dengue

Vẽ phân lớp “SXH Dengue nặng” với các thuộc tính: Nếu khoảng tiêu cầu = “Thấp”  $\wedge$  Gan to = “Có”  $\wedge$  Vàng da = “Không”  $\wedge$  NS1 = “Có”  $\wedge$  độ lệch huyết áp = 2  $\rightarrow$  Thuộc phân lớp “SXH Dengue nặng”



Hình 4.3 hình vẽ SXH Dengue nặng

## 4.2. Đánh giá

Sau khi được trình bày tính năng của ứng dụng đồng thời triển khai thử nghiệm đã được bác sĩ chuyên khoa (kèm theo danh sách) tại Bệnh viện Bệnh Nhiệt Đới TPHCM đánh giá khá cao.

Danh sách bác sĩ tham gia đánh giá ứng dụng

STT	Họ & tên bác sĩ	Đơn vị công tác
1	BS. CK1. Nguyễn Lê Như Tùng	BVBND
2	BS. Lý Quốc Công	BVBND
3	BS. Nguyễn Thanh Liêm	BVBND
4	BS. Trần Quốc Tân	BVBND
5	BS. Ngô Chí Nguyên	BVBND
6	BS. Nguyễn Thanh Phong	BVBND
7	BS. Nguyễn Thanh Tùng	BVBND
8	BS. Phạm Thị Hải Mến	BVBND
9	BS. Dư Tác Tạo	BVBND
10	Ths. BS. Lê Đức Vinh	BVBND - ĐHYKPNT

Bảng 4.1 : Bảng danh sách Bác sĩ đánh giá chương trình

# Chương 5. TỔNG KẾT

## 5.1. Kết luận

Luận văn đưa ra một cách nhìn về sự kết hợp Công nghệ thông tin vào lĩnh vực Y tế. Kết quả đề tài là hệ hỗ trợ chẩn đoán mang tính chất cộng đồng, sẽ giúp rất nhiều về mặt chuyên môn đối với các tuyến y tế chưa có đội ngũ bác sĩ có trình độ chuyên môn cao, thiếu trang thiết bị y tế, cũng như các tuyến y tế vùng sâu vùng xa. Ngoài ra, đối với sinh viên y khoa và các bác sĩ trẻ, hệ hỗ trợ chẩn đoán giúp ôn lại các kiến thức về bệnh nhiễm ở vùng nhiệt đới cụ thể là cho biết sau khi bệnh nhân bị SXH sẽ biết được mức độ của bệnh để có hướng điều trị chính xác.

Kết quả đề tài vẫn chưa thật sự tốt, kết quả chẩn đoán của hệ hỗ trợ chẩn đoán vẫn còn nhiều trường hợp chưa đúng. Tuy nhiên sau khi triển khai thử nghiệm cho các bác sĩ chuyên khoa tại Bệnh viện Bệnh Nhiệt Đới, kết quả đề tài được đánh giá cao. Theo nhận định của các bác sĩ, nếu tiếp tục được đầu tư và phát triển, hệ hỗ trợ chẩn đoán sẽ giúp ích rất nhiều cho bác sĩ trong việc chẩn đoán và điều trị bệnh cho bệnh nhân.

Ngoài ra đề tài có thể được phát triển để chẩn đoán nhanh và chính xác hơn khi hệ hỗ trợ chẩn đoán được kết nối trực tiếp với các hệ thống thông tin quản lý bệnh án tại cơ sở y tế. Hơn nữa, hệ hỗ trợ chẩn đoán có thể áp dụng cho nhiều loại bệnh khác nhau.

## 5.2. Hạn chế của đề tài

Về mặt công nghệ, đề tài sử dụng thuật toán C4.5 có một số hạn chế về vấn đề xử lý dữ liệu. Trong trường hợp dữ liệu có quá nhiều lớp thuật toán sẽ dễ gây ra lỗi và dữ liệu càng nhiều, thời gian huấn luyện càng lâu.

Về vấn đề thực tiễn, dữ liệu của đề tài nghiên cứu thu thập tại một thời điểm cắt ngang chính vì vậy không quan sát rõ được diễn tiến bệnh của bệnh nhân. Đề đạt được kết quả tốt hơn ta cần thu thập dữ liệu từ khi bệnh nhân có những triệu chứng ban đầu đến khi phát bệnh và khỏi bệnh. Đồng thời cần lấy thêm dữ liệu của quá trình điều trị bệnh của bệnh nhân, điều đó sẽ cho ta hiểu hơn quá trình sinh bệnh và giúp cho chương trình đạt được hiệu quả thực tiễn cao.

## 5.3. Hướng phát triển

Khai phá dữ liệu là bài toán được nhiều nhà nghiên cứu quan tâm bởi nó được ứng dụng rộng rãi trong các lĩnh vực cũng như chứa đựng nhiều hướng mở rộng khác nhau. Tuy nhiên để mở rộng ứng dụng và được đưa vào thực tiễn thì ta cần làm thêm một số công việc sau:

- + Số liệu bệnh nhân phải được thu thập nhiều hơn.
- + Thu thập số liệu diễn tiến bệnh về lâm sàng và cận lâm sàng.
- + Thu thập diễn tiến phác đồ điều trị.
- + Xử lý dữ liệu tốt hơn để tăng hiệu suất thực thi chương trình.
- + Xây dựng hệ thống chẩn đoán bệnh cho nhiều loại bệnh khác nhau.
- + Cần sự hợp tác chuyên môn của các chuyên gia công nghệ thông tin và y tế.
- + Tìm hiểu thuật toán khác như C5.0, mạng Bayesian hoặc mạng Neuron nếu có hiệu quả hơn.
- + Tìm hiểu thuật toán ILA để so sánh các kết quả đạt được đối với thuật toán C4.5.

## TÀI LIỆU THAM KHẢO

### Tài liệu tiếng Việt

- [1] “Hội thảo khu vực hướng ứng ngày ASEAN phòng chống sốt xuất huyết” Hà Nội 14/6/2013
- [2] Nguyễn Thanh Thủy, Hệ thống trợ giúp và kiểm tra đơn thuốc chữa bệnh tăng huyết áp ES-TENSION, Tạp chí tin học và điều khiển học, Viện Công nghệ thông tin, 12(3), (1996), 10-18.
- [3] Đỗ Văn Thành, Một cách tiếp cận ra quyết định trong chẩn đoán lâm sàng, Tạp chí Tin học và điều khiển học, Viện công nghệ thông tin, 16(1),(2000), 52-58
- [4] Nguyễn Đức Cường, “Slide bài giảng môn học BI & DM: Business Intellegent and Data Mining”, 2011-2012
- [5] Tạ Văn Bình. Những nguyên lý nền tảng đái tháo đường – tăng glucose máu. NXB Y học, Hà Nội 2007

### Tài liệu tiếng Anh

- [6] Buchanan B.G. (1984), Shortliffe E.H, Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, pp. 209-232.
- [7] Aikins J. S., Kunz J. C., Shortliffe E. H., and Fallat K. J. (1983), “PUFF: An Expert System for Interpretation of Pulmonary Function Data”, *Comput Biomed* 16, pp. 199-208.
- [8] Fred A., Filipe J., Partinen M., Paiva T. (2000), "PSG-Expert: An Expert System for the Diagnosis of Sleep Disorders", *IOS Press* 78, pp. 127-147.

- [9] Ngah U. K., Aziz S. A. (2007), "A BI-RADS Based Expert Systems for the Diagnoses of Breast Diseases", *American Journal of Applied Sciences* 4 (11), pp. 867-875. 33
- [10] Naser S.S.A, Akkila A.N. (2008), "A Proposed Expert System for Skin Diseases Diagnosis", *Journal of Applied Sciences Research* 4(12): pp. 1682-1693.
- [11] J. Han and Micheline Kamber. *Data Mining:Concepts and Techniques*, 3<sup>rd</sup> Edition. Morgan Kaufmann Publishers, 2011.
- [12] T. Mitchell, Machine Learning and Data Mining, Communications of the ACM, Vol. 42 (1999), No. 11, pp. 30--36.s
- [13] Manuel Mora Autonomous,Guisseppi A. Forgionne, JatinderN. D. Gupta.“Decision Making Support Systems:Achievements, Trends and Challenges for the New Decade”, pp. 1-5, (2003)
- [14] John Shafer, Rakesh Agrawal, Manish Mehta. “Sprint – A Scalable Classifier for Data mining” in Predeeings of the 22<sup>nd</sup> International Conference on very large database, India 1996.
- [15] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993
- [16] Ming Li “Data mining. Chepter 10 : Predictive Modeling”, Department of Computer Science and Technology Nanjing University, 2011

## **Tài liệu Internet**

- [17] [http://en.wikipedia.org/wiki/Clinical\\_decision\\_support\\_system](http://en.wikipedia.org/wiki/Clinical_decision_support_system)
- [18] <http://en.diagnosispro.com/>

[19] <http://thuvienphapluat.vn/archive/Quyet-dinh/Quyet-dinh-458-QD-BYT-huong-dan-chan-doan-dieu-tri-sot-xuat-huyet-Dengue-vb120583t17.aspx>

[20] <http://technet.microsoft.com/en-us/library/bb895174.aspx>