TRƯỜNG ĐẠI HỌC DUY TÂN

KHOA SAU ĐẠI HỌC

Tiểu luận môn

HỆ CHUYÊN GIA

HỆ THỐNG HỖ TRỢ CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

Hướng dẫn : PGS.TS Hoàng Văn Dũng

Thực hiện : Phạm Minh Tuấn

Võ Đình Hiếu

Nguyễn Anh Quân

Lớp : K22MCS (Khoa học máy tính)

Đà nẵng, 07/2021

BẢNG PHÂN CHIA CÔNG VIỆC

Phạm Minh Tuấn

+ Viết nội dung Lời mở đầu
+ Viết nội dung Chương 3
+ Viết nội dung Chương 4
+ Phát triển Ứng dụng
+ Làm nội dung Slide chương 3

Võ Đình Hiếu

+ Tìm hiểu về Hệ chuyên gia và Cây quyết định
+ Tính toán và phân tích dữ liệu sinh để sinh ra tập luật
+ Viết nội dung Chương 1
+ Làm nội dung Slide chương 1

Nguyễn Anh Quân

+ Tìm hiểu về Cây quyết định và Thuật toán C4.5
+ Tính toán và phân tích dữ liệu sinh để sinh ra tập luật
+ Viết nội dung Chương 2
+ Làm nội dung Slide chương 2

MỤC LỤC

DANH MỤC HÌNH VẼ 1
DANH MỤC BẢNG 2
LỜI MỞ ĐẦU 3
1. Lý do chọn đề tài 3
2. Mục tiêu nghiên cứu 3
3. Phương pháp nghiên cứu 3
4. Bố cục tiểu luận 4
CHƯƠNG 1: GIỚI THIỆU HỆ CHUYÊN GIA 5
1.1. TỔNG QUAN VỀ HỆ CHUYÊN GIA 5
1.1.1. Khái niệm hệ chuyên gia 5
1.1.2. Kiến trúc của hệ chuyên gia 5
1.1.3. Đặc điểm của hệ chuyên gia 6
1.1.4. Lợi ích của hệ chuyên gia 7
1.1.5. Hoạt động của hệ chuyên gia 7
1.1.6. Biểu diễn và suy diễn tri thức trong hệ chuyên gia 8
1.1.7. Ứng dụng của hệ chuyên gia 9
1.1.8. Các bước xây dựng hệ chuyên gia 10
1.1.9. Hạn chế của hệ chuyên gia 11
1.2. GIỚI THIỆU CÂY QUYẾT ĐỊNH 11
1.2.1. Cách sử dụng Cây quyết định 12
1.2.2. Duyệt cây và phân lớp dữ liệu 12
1.2.3. Giới thiệu thuật toán C4.5 14
CHƯƠNG 2: BỆNH TIỂU ĐƯỜNG VÀ THUẬT TOÁN C4.5 16
2.1. GIỚI THIỆU BỆNH TIỂU ĐƯỜNG 16
2.2. DỮ LIỆU BỆNH TIỂU ĐƯỜNG 16
2.3. MÔ TẢ THUỘC TÍNH DỮ LIỆU 17
2.4. CÀI ĐẶT THUẬT TOÁN TRÊN TẬP DỮ LIỆU 19
2.5. DANH SÁCH LUẬT 40
CHƯƠNG 3: ỨNG DỤNG CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG 48
3.1. GIỚI THIỆU ỨNG DỤNG 48

3.2. GIAO DIỆN VÀ TÍNH NĂNG 48
3.2.1. Truy cập ứng dụng 48
3.2.2. Tính năng Tra cứu bệnh 48
3.2.3. Tính năng Quản trị 50
3.2.4. Tính năng Giới thiệu 53
CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 54
TÀI LIỆU THAM KHẢO 55

DANH MỤC HÌNH VẼ

Ký hiệu	Nội dung	Trang
Hình 1.1	Kiến trúc của hệ chuyên gia	3
Hình 1.2	Hoạt động của hệ chuyên gia	6
Hình 1.3	Suy diễn tiến	7
Hình 1.4	Suy diễn lùi	7
Hình 1.5	Cấu trúc cây quyết định	10
Hình 2.1	Cây quyết định tại thuộc tính Polyuria	27
Hình 2.2	Cây quyết định tại thuộc tính Alopecia	37
Hình 2.3	Cây quyết định với bảng dữ liệu mẫu	38
Hình 3.1	Truy cập ứng dụng Hỗ trợ chuẩn đoán bệnh tiểu đường	46
Hình 3.2	Giao diện trang chủ website	47
Hình 3.3	Giao diện trang kết quả nếu không mắc bệnh	47
Hình 3.4	Giao diện trang kết quả người dùng mắc bệnh	48
Hình 3.5	Giao diện trang đăng nhập quản trị	48
Hình 3.6	Giao diện trang quản trị dữ liệu	49
Hình 3.7	Giao diện trang thêm mới dữ liệu	50
Hình 3.8	Giao diện trang giới thiệu	51

DANH MỤC BẢNG

Ký hiệu	Nội dung	Trang
Hình 2.1	Các thuộc tính và giá trị của dữ liệu mẫu	13
Hình 2.2	Bảng dữ liệu mẫu	14
Hình 2.3	Mẫu dữ liệu với thuộc tính Polyuria có giá trị Yes	25
Hình 2.4	Mẫu dữ liệu với thuộc tính Alopecia có giá trị Yes	34

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Theo số liệu thống kê từ Liên đoàn Đái tháo đường thế giới (IDF) cho thấy, cứ mỗi giờ có thêm hơn 1.000 bệnh nhân đái tháo đường (ĐTĐ) mắc mới, và cứ mỗi 8 giây có 1 người chết do ĐTĐ. IDF chỉ ra, bệnh ĐTĐ hiện nay có thể coi là một loại bệnh dịch toàn cầu với 415 triệu người trưởng thành bị bệnh chiếm 8,8% dân số thế giới. Tại Việt Nam, số liệu từ Hội nội tiết và ĐTĐ (VADE) cho biết, hiện có tới 3,53 triệu người đang "chung sống" với căn bệnh ĐTĐ và mỗi ngày có ít nhất 80 trường hợp tử vong vì các biến chứng liên quan. Dự báo, số người bắc bệnh có thể tăng lên 6,3 triệu vào năm 2045. Cùng nằm trong xu hướng đó, Việt Nam được xếp trong 10 quốc gia có tỷ lệ gia tăng bệnh nhân ĐTĐ cao nhất thế giới với tỷ lệ tăng 5,5% mỗi năm.

Bệnh tiểu đường rất nguy hiểm và cần được điều trị suốt đời. Bệnh do hệ thống miễn dịch bị phá hủy các tế bào beta sản xuất insulin trong tuyến tụy. Nếu không được kiểm soát chặt chẽ sẽ dẫn tới những biến chứng rất nguy hiểm.

Nhận thấy sự cần thiết về việc phổ cập kiến thức cũng như giúp cho mọi người có thể dễ dàng chuẩn đoán sớm xem mình có khả năng mắc bệnh hay không là lý do chính để thực hiện đề tài.

2. Mục tiêu nghiên cứu

Được tiếp cận kiến thức môn Hệ chuyên gia và các môn học khác cùng với sự hướng dẫn của PGS. TS. Hoàng Văn Dũng, nhóm thực hiện mong muốn xây dựng được một ứng dụng Hệ chuyên gia có tính thực tiễn.

Dựa vào nguồn dữ liệu thống kê về bệnh tiểu đường, nhóm thực hiện mong muốn ứng dụng các giải thuật đã học nhằm khai phá dữ liệu để có thể làm chủ được kiến thức về Hệ chuyên gia và các môn khác có liên quan.

3. Phương pháp nghiên cứu

Tìm hiểu kiến thức về Hệ chuyên gia nhằm nắm bắt kiến trúc hệ thống, các giải thuật xử lý dữ liệu cũng như các bước tiến hành để xây dựng ứng dụng.

Nghiên cứu phương pháp phân tích thiết kế, các công cụ hỗ trợ trong phạm vi đề tài để phục vụ cho việc thiết kế.

Tìm hiểu kiến thức về ngôn ngữ lập trình, cơ sở dữ liệu và các công cụ hỗ trợ khác để phát triển hệ thống.

4. Bố cục tiểu luận

Nội dung của bài tiểu luận được trình bày với bố cục gồm 4 chương như sau:

Chương 1: Tìm hiểu về hệ chuyên và giải thuật C4.5.

Chương 2: Thông tin về dữ liệu bệnh tiểu đường và áp dụng giải thuật C4.5

Chương 3: Ứng dụng thực tiễn chuẩn đoán bệnh tiểu đường

Chương 4: Kết luận trình bày kết quả đạt được của tiểu luận và tính thực tế cũng như định hướng phát triển trong tương lai.

Chương 1
GIỚI THIỆU HỆ CHUYÊN GIA

1.1. TỔNG QUAN VỀ HỆ CHUYÊN GIA

1.1.1. Khái niệm hệ chuyên gia

Hệ chuyên gia là một hệ thống chương trình máy tính chứa các thông tin tri thức và các quá trình suy diễn về một lĩnh vực cụ thể nào đó để giải quyết các bài toán khó mà đòi hỏi sự uyên bác của các chuyên gia trong ngành.

Hệ chuyên gia sử dụng các khả năng lập luận để đi tới các kết luận hoặc gợi ý. Tùy theo thiết kế chương trình mà hệ chuyên gia đưa ra trình tự các hành động cần thực hiện để giải quyết vấn đề.

Ví dụ về các hệ chuyên gia nổi tiếng:

MYCIN: Chương trình cung cấp cho thầy thuốc các ý kiến chữa trị liên quan đến liệu pháp kháng sinh.

DENDRAL: Hệ chuyên gia được sử dụng để phân tích hóa học để dự đoán cấu trúc phân tử.

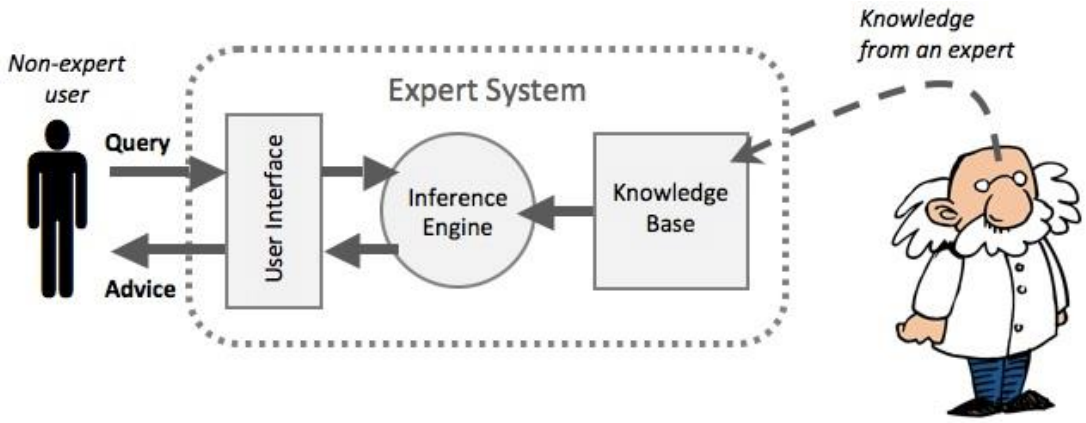
PXDES: Hệ chuyên gia dùng để dự đoán mức độ và loại ung thư phổi.

CaDet: Hệ chuyên gia có thể xác định ung thư ở giai đoạn đầu.

1.1.2. Kiến trúc của hệ chuyên gia

Một hệ chuyên gia gồm ba thành phần chính:

- Cơ sở tri thức (Knowledge Base)
- Máy suy diễn hay mô tơ suy diễn (Inference Engine)
- Hệ thống giao tiếp (User Interface)



Hình 1.1: Kiến trúc của hệ chuyên gia

❖ ***Cơ sở tri thức (Knowledge Base)***

Là một kho dữ kiện, nó lưu trữ tất cả tri thức trong một lĩnh vực nào đó, tri thức này do chuyên gia con người chuyển giao.

Nó bao gồm các mục sau: các khái niệm cơ bản, các sự kiện, các luật và mối quan hệ giữa chúng.

Ví dụ:

- Tri thức về bệnh tiểu đường do chuyên gia bệnh tiểu đường chuyển giao.
- Tri thức về bất động sản do các chuyên gia về bất động sản chuyển giao.

❖ ***Máy suy diễn hay mô tơ suy diễn (Inference Engine)***

Là bộ xử lý cho tri thức, được mô hình sao cho giống với việc suy luận của chuyên gia con người.

Bộ xử lý này làm việc dựa trên thông tin mà người dùng mô tả về vấn đề, kết hợp với Cơ sở tri thức để đưa ra kết luận hay khuyến nghị.

❖ ***Hệ thống giao tiếp (User Interface)***

Là bộ phận giao tiếp với người dùng, cung cấp các câu hỏi và các tùy chọn khác để người dùng trả lời hoặc lựa chọn để giao tiếp với hệ chuyên gia.

1.1.3. Đặc điểm của hệ chuyên gia

❖ ***Tính chuyên môn cao***

Hệ chuyên gia vô cùng thông minh và tính chuyên môn cao. Nó cung cấp khả năng giải quyết vấn đề hiệu quả, độ chính xác và giàu trí tưởng tượng.

❖ ***Thời gian trả lời thỏa đáng***

Thời gian trả lời hợp lý, bằng hoặc nhanh hơn so với chuyên gia con người trong cùng một quyết định. Hệ chuyên gia là một hệ thống thời gian thực.

❖ ***Độ tin cậy cao***

Không thể xảy ra sự cố hoặc giảm sút độ tin cậy khi sử dụng.

❖ ***Hiệu quả cao***

Khả năng trả lời với mức độ tinh thông bằng hoặc cao hơn so với chuyên gia trong cùng lĩnh vực.

❖ ***Dễ hiểu***

Hệ chuyên gia giải thích các bước suy luận một cách dễ hiểu và nhất quán, không giống như cách trả lời bí ẩn của các hộp đen.

1.1.4. Lợi ích của hệ chuyên gia

❖ ***Phổ cập (increased availability)***

Là sản phẩm chuyên gia, được phát triển không ngừng với hiệu quả sử dụng không thể phủ nhận.

❖ ***Giảm giá thành (reduced cost)***

Đầu tư một lần sử dụng nhưng sử dụng được lâu dài, kinh phí phát triển sẽ rẻ hơn so với kinh phí thuê chuyên gia con người.

❖ ***Giảm rủi ro (reduced dangers)***

Giúp con người tránh được trong các môi trường rủi ro, nguy hiểm.

❖ ***Tính thường trực (Permanance).***

Bất kể lúc nào cũng có thể khai thác sử dụng, trong khi con người có thể mệt mỏi, nghỉ ngơi hay vắng mặt. Một hệ chuyên gia có thể có nhiều lĩnh vực khác nhau và được khai thác đồng thời bất kể thời gian sử dụng.

❖ ***Độ tin cậy (increased relialility)***

Luôn đảm bảo độ tin cậy khi khai thác. Khả năng giảng giải (explanation). Câu trả lời với mức độ tinh thông được giảng giải rõ ràng chi tiết, dễ hiểu.

❖ ***Khả năng trả lời (fast reponse)***

Trả lời theo thời gian thực, khách quan. Tính ổn định, suy luận có lý và đầy đủ mọi lúc mọi nơi.

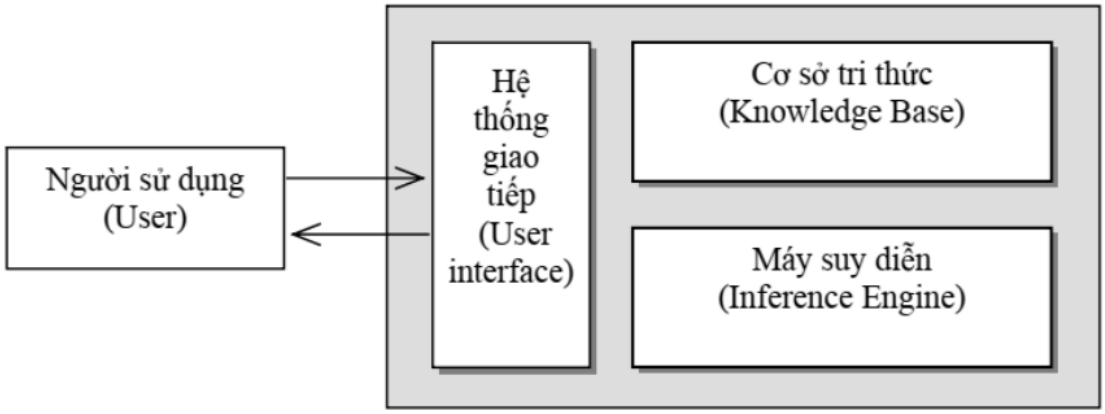
❖ ***Trợ giúp thông minh như một người hướng dẫn.***

❖ ***Có thể truy cập như là một cơ sở dữ liệu thông minh***

1.1.5. Hoạt động của hệ chuyên gia

Giao diện người dùng là phần quan trọng nhất của Hệ chuyên gia. Người sử dụng đặt câu hỏi cho hệ chuyên gia bằng cách: Cung cấp sự kiện là những gì đã biết, đã có thật hay những tri thức có ích cho hệ chuyên gia.

Máy suy diễn hay mô tơ suy diễn làm việc dựa trên thông tin mà người dùng mô tả về vấn đề, kết hợp với Cơ sở tri thức để cho ra các câu trả lời là những lời khuyên hay những gợi ý đúng đắn cho người sử dụng qua hệ thống giao tiếp.



Hình 1.2: Hoạt động của hệ chuyên gia

1.1.6. Biểu diễn và suy diễn tri thức trong hệ chuyên gia

❖ ***Biểu diễn tri thức***

Tri thức của một Hệ chuyên gia có thể được biểu diễn theo nhiều phương pháp khác nhau. Tùy theo từng các thiết kế và cấu trúc, người ta có thể sử dụng một hoặc đồng thời nhiều phương pháp biểu diễn.

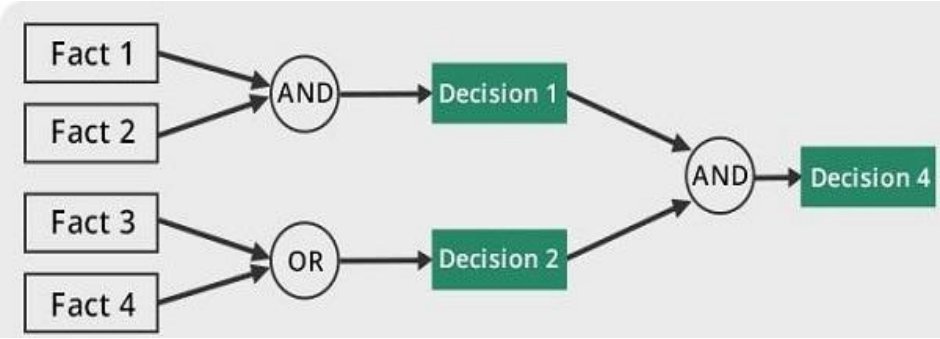
Các cách biểu diễn tri thức khác như:

- Biểu diễn tri thức bởi các luật sản xuất
- Biểu diễn tri thức nhờ mệnh đề logic
- Biểu diễn tri thức nhờ mạng ngữ nghĩa
- Biểu diễn tri thức nhờ ngôn ngữ nhân tạo
- Biểu diễn tri thức nhờ các sự kiện không chắc chắn
- Biểu diễn nhờ bộ ba: đối tượng, thuộc tính và giá trị.
- Biểu diễn nhờ khung…

❖ ***Suy diễn tiến***

Suy diễn tiến (forward chaining) là quá trình suy diễn bắt đầu từ tập sự kiện đã biết, rút ra những sự kiện mới và cứ như vậy cho đến khi có được sự kiện cần chứng minh hoặc không có luật nào sinh ra các sự kiện mới (tập sự kiện đúng là cực đại).

Hình thức suy diễn này thường áp dụng cho các dạng hệ thống về: hoạch định, giám sát, điều khiển, diễn dịch.

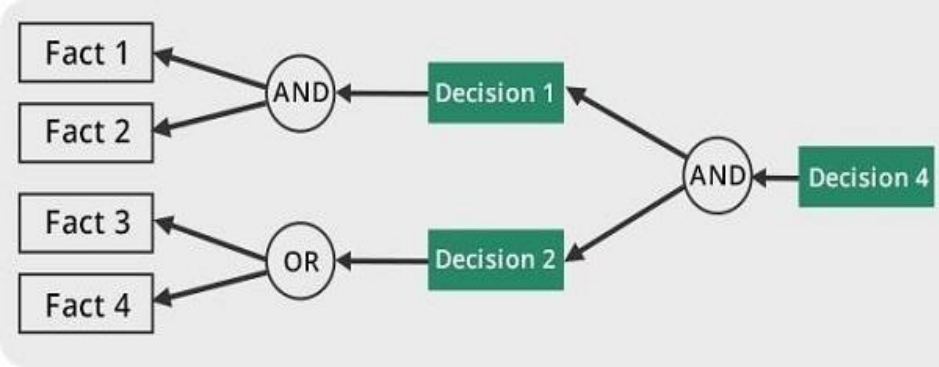


Hình 1.3: Suy diễn tiến

❖ ***Suy diễn lùi***

Suy diễn lùi (backward chaining) là quá trình bắt đầu với một danh sách các dữ liệu đích và thực hiện suy luận ngược từ kết quả cho đến các tiên đề suy ra nó.

Hình thức suy diễn này thường áp dụng cho các dạng: chuẩn đoán, kê toa, gỡ rối.



Hình 1.4: Suy diễn lùi

1.1.7. Ứng dụng của hệ chuyên gia

Hệ chuyên gia ngày nay được ứng dụng rộng rãi ở rất nhiều lĩnh vực. Hiệu quả của hệ chuyên gia mang lại là rất lớn nên ngày càng nhiều có những ứng dụng mới tiên tiến hơn, hiện đại hơn được ra đời để đáp ứng nhu cầu xã hội. Các lĩnh vực mà hệ chuyên gia đang ứng dụng rất nhiều như là:

- Diễn giải (Interpretation): Đưa ra kết luận hay mô tả dễ hiểu từ những tập dữ liệu thô.
- Dự báo (Prediction): Đưa ra hậu quả có thể xảy ra trước một tình huống.
- Chẩn đoán (Diagnosis): Xác định nguyên nhân của những sự cố trong các tình huống phức tạp dựa trên các triệu chứng quan sát được.
- Thiết kế (Design): Tìm ra cấu hình cho các thành phần hệ thống đáp ứng các mục tiêu trong khi vẫn thỏa mãn một tập hợp các ràng buộc về thiết kế.

- Lập kế hoạch (Planing): Tìm ra một chuỗi các hành động để đạt được một tập hợp các mục tiêu, khi được cho trước các điều kiện khởi đầu và ràng buộc trong thời gian chạy.
- Theo dõi (Monotoring): So sánh những hành vi quan sát được của hệ thống với hành vi mong đợi.
- Gỡ rối (Debugging and Repair): Chỉ định và cài đặt phương pháp chữa trị cho những sự cố, rủi ro.
- Hướng dẫn (Instruction): Phát hiện và sửa chữa những thiếu sót trong quan niệm của học viên về một chủ đề lĩnh vực nào đó.
- Điều khiển (Control): Chỉ đạo hành vi trong một môi trường phức tạp.

1.1.8. Các bước xây dựng hệ chuyên gia

Hệ chuyên gia được xây dựng qua 6 giai đoạn như sau:

❖ ***Giai đoạn 1: Khởi tạo dự án***

- Định nghĩa vấn đề
- Đánh giá nhu cầu
- Đánh giá thay thế các giải pháp
- Xác minh rằng phương pháp tiếp cận hệ thống chuyên gia là phù hợp
- Xem xét vấn đề quản lý

❖ ***Giai đoạn 2: Phân tích và thiết kế hệ thống***

- Trình bày thiết kế ý tưởng
- Quyết định sự phát triển chiến lược
- Quyết định nguồn của kiến thức và đảm bảo hợp tác
- Chọn tài nguyên máy tính
- Thực hiện một nghiên cứu khả thi
- Thực hiện phân tích chi phí - lợi ích

❖ ***Giai đoạn 3: Tạo mẫu***

- Xây dựng một nguyên mẫu nhỏ
- Kiểm tra, cải thiện và mở rộng nó
- Chứng minh và phân tích tính khả thi

- Hoàn thành thiết kế.

❖ ***Giai đoạn 4: Phát triển hệ thống***

- Xây dựng nền tảng kiến thức
- Kiểm tra, đánh giá và nâng cao kiến thức nền tảng
- Lập kế hoạch tích hợp

❖ ***Giai đoạn 5: Triển khai***

- Đảm bảo người dùng chấp nhận
- Cài đặt, trình diễn và triển khai hệ thống
- Sắp xếp định hướng và đào tạo cho người dùng
- Đảm bảo an ninh
- Cung cấp tài liệu, Sắp xếp để tích hợp và thử nghiệm hiện trường

❖ ***Giai đoạn 6: Hậu triển khai***

- Hoạt động
- Bảo trì
- Nâng cấp
- Đánh giá định kỳ

1.1.9. Hạn chế của hệ chuyên gia

Không thể đưa ra phản ứng sáng tạo trong một tình huống bất thường

Những sai sót trong cơ sở kiến thức có thể dẫn đến quyết định sai lầm

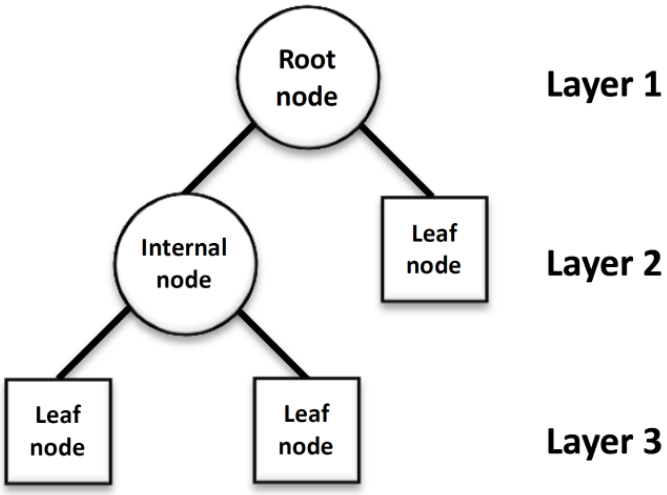
Chi phí bảo trì của một hệ thống chuyên gia quá đắt

Mỗi vấn đề là khác nhau do đó giải pháp từ chuyên gia con người cũng có thể khác và sáng tạo hơn

1.2. GIỚI THIỆU CÂY QUYẾT ĐỊNH

Cây quyết định là một cấu trúc phân cấp của các nút và các nhánh, nút bên trong được gán nhãn bằng các thuộc tính. Các nhánh bắt nguồn từ một nút có nhãn là thuộc tính A sẽ được gán nhãn bằng mỗi giá trị có thể có của thuộc tính A. Các nút lá của cây biểu diễn nhãn lớp hoặc sự phân bố của lớp. Để phân lớp một mẫu chưa biết chúng ta duyệt nó từ nút gốc đến nút lá, với mỗi thuộc tính bắt gặp nhánh tương ứng với giá

trị của mẫu cho thuộc tính đó sẽ được đi theo cho đến khi gặp nút lá, phân lớp mẫu này tương ứng với nút lá đó sẽ được trả về.



Hình 1.5: Cấu trúc cây quyết định

1.2.1. Cách sử dụng Cây quyết định

Kiểm tra những giá trị thuộc tính của từng nút bắt đầu từ nút gốc của cây quyết định. Từ các nhánh chứa các giá trị của thuộc tính, ta tìm lần đến một phân lớp cuối cùng và từ đây ta có thể suy ra các luật tương ứng để mô tả cho quá trình khám phá tri thức từ các mẫu dữ liệu.

- Mỗi một đường dẫn từ gốc đến lá trong cây tạo thành một luật.
- Mỗi cặp giá trị thuộc tính trên một đường dẫn tạo nên một sự liên kết.
- Nút lá giữ quyết định phân lớp dự đoán.
- Các luật tạo được dễ hiểu hơn các cây

1.2.2. Duyệt cây và phân lớp dữ liệu

❖ ***Lựa chọn tiêu chuẩn phân lớp***

Ta có thể chọn bất kỳ thuộc tính nào làm nút của cây, điều này có khả năng xuất hiện nhiều cây quyết định khác nhau cùng biểu diễn một tập mẫu, có cây xuất hiện nhiều nút hoặc cây đơn giản, điều quan trọng là chọn thuộc tính nào để có thể phân lớp tốt dữ liệu sau này, một cách trực quan là ta nên chọn thuộc tính có độ phân biệt cao lên gần với nút gốc của cây, tức là chọn thuộc tính cho cây quyết định nhỏ nhất theo các cách sau:

- Tạo ra các nhóm sao cho một lớp chiếm ưu thế trong từng nhóm.
- Thuộc tính được chọn là thuộc tính cho độ đo tốt nhất, có lợi nhất cho quá trình phân lớp.

Độ đo để đánh giá chất lượng phân chia là độ đo đồng nhất, có 3 tiêu chuẩn hay dùng nhất trong việc lựa chọn:

- Entropy (Information Gain)
- Information Gain Ratio
- Gini Index

❖ ***Điều kiện để dừng việc phân chia***

- Tất cả những mẫu huấn luyện thuộc về cùng một lớp.
- Không còn thuộc tính còn lại nào để phân chia tiếp.
- Không còn mẫu nào còn lại.

Các thuật toán trên cây quyết định điểm khác biệt chính là tiêu chuẩn phân chia như liệt kê bên trên, ở đây chúng ta áp dụng thuật toán C4.5 nên trong nội dung tiểu luận chỉ đề cập đến độ lợi thông tin để chọn lựa thuộc tính phân lớp

❖ ***Độ lợi thông tin (Information Gain)***

Information Gain là đại lượng được sử dụng để lựa chọn thuộc tính có độ lợi thông tin lớn nhất để phân lớp dữ liệu. Giả sử cho P, N là hai lớp và S là tập dữ liệu chứa p phần tử của lớp P và n phần tử của lớp N. Khối lượng của thông tin cần để quyết định một mẫu tùy ý trong S thuộc về lớp P hoặc N được định nghĩa như sau:

$$I(p,n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n} \quad \text{(Công thức 1.1)}$$

Giả sử rằng sử dụng thuộc tính A để phân hoạch tập hợp S thành những tập hợp {S1, S2, ... ,Sv}. Nếu Si chứa những pi mẫu của lớp P và ni mẫu của N, Entropy hay thông tin mong đợi cần để phân lớp những đối tượng trong tất cả các cây con Si là:

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i) \qquad \text{(Công thức 1.2)}$$

Độ lợi thông tin nhận được bởi việc phân nhánh trên thuộc tính A là:

$$Gain(A) = I(p,n) - E(A) \qquad \text{(Công thức 1.3)}$$

1.2.3. Giới thiệu thuật toán C4.5

Nhiệm vụ của giải thuật C4.5 là học cây quyết định từ một tập các dữ liệu huấn luyện bằng cách xét từng thuộc tính của tập dữ liệu huấn luyện để tìm ra thuộc tính có độ lợi thông tin cao nhất và phân nhánh cho thuộc tính đó. Biểu diễn này cho phép chúng ta xác định phân loại một đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó.

Ý tưởng giải thuật C4.5 như sau:

Đầu vào: Một tập hợp các mẫu huấn luyện. Mỗi mẫu huấn luyện bao gồm các thuộc tính mô tả một tình huống, hay một đối tượng nào đó, và một giá trị phân loại của nó.

Đầu ra: Cây quyết định có khả năng phân loại đúng đắn các mẫu huấn luyện trong tập dữ liệu rèn luyện, và phân loại đúng cho cả các bộ chưa gặp trong tương lai.

Giải thuật:

```
Function induce_tree (tập_mẫu_huấn_luyện, tập_thuộc_tính)
Begin
    If mọi mẫu trong tập_mẫu_huấn_luyện đều nằm trong cùng một lớp Then
        Return một nút lá được gán nhãn bởi lớp đó
    Else If tập_thuộc_tính là rỗng Then
        return nút lá được gán nhãn bởi tuyển của tất cả các lớp trong tập_mẫu_huấn_luyện
    Else
    Begin  chọn một thuộc tính P, lấy nó làm gốc cho cây hiện tại;
        xóa P ra khỏi tập_thuộc_tính;
        với mỗi giá trị V của P
        Begin   tạo một nhánh của cây gán nhãn V;
          Đặt vào phân_vùng V các mẫu trong tập_mẫu_huấn_luyện có giá trị V tại thuộc
          tính P;  Gọi induce_tree(phân_vùngV, tập_thuộc_tính), gắn kết quả vào nhánh V
        End
    End
End
```

Để xây dựng cây quyết định, tại mỗi nút của cây thì thuật toán đều tính lượng thông tin nhận được trên các thuộc tính và chọn thuộc tính có lượng thông tin tốt nhất làm nút phân tách trên cây.

Information Gain là đại lượng được sử dụng để lựa chọn thuộc tính có độ lợi thông tin lớn nhất để phân lớp dữ liệu. Giả sử cho P, N là hai lớp và S là tập dữ liệu chứa p phần tử của lớp P và n phần tử của lớp N. Khối lượng thông tin cần để quyết định một mẫu tùy ý trong S thuộc về lớp P hoặc N được định nghĩa như sau:

$$I(p,n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n}$$

Entropy là khái niệm để đo tính thuần nhất của một tập huấn luyện. Một tập huấn luyện là thuần nhất nếu như tất cả các phần tử của tập huấn luyện đều thuộc cùng một loại, hay nói cách khác tập huấn luyện này có độ pha trộn là thấp nhất.

Giả sử rằng sử dụng thuộc tính A để phân hoạch tập hợp S thành những tập hợp {S1, S2, ... ,Sv}. Nếu Si chứa những pi mẫu của lớp P và ni mẫu của N, entropy hay thông tin mong đợi cần để phân lớp những đối tượng trong tất cả các cây con Si là:

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i)$$

Độ lợi thông tin nhận được bởi việc phân nhánh trên thuộc tính A là:

$$Gain(A) = I(p,n) - E(A)$$

Tuy nhiên thuộc tính có nhiều giá trị không phải lúc nào cũng cho việc phân lớp tốt nhất, vì vậy ta cần chuẩn hóa độ đo Gain.

Tính thông tin trung bình của từng thuộc tính, để hạn chế xu hướng chọn thuộc tính có nhiều giá trị, thông tin trung bình của thuộc tính A được tính như sau:

$$\text{SplitInfo(A)} = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} log_2 (\frac{|D_j|}{|D|})$$

Việc chọn thuộc tính để phân nhánh dựa vào độ đo Gain Ration được tính theo công thức sau:

$$\text{GainRatio(A)} = \text{Gain(A)} / \text{SplitInfo(A)}$$

Chương 2
BỆNH TIỂU ĐƯỜNG VÀ THUẬT TOÁN C4.5

2.1. GIỚI THIỆU BỆNH TIỂU ĐƯỜNG

Theo báo cáo của Tổ chức Y tế Thế giới (WHO), bệnh tiểu đường là một trong những căn bệnh mãn tính đe dọa tính mạng phát triển nhanh nhất, đã ảnh hưởng đến 422 triệu người trên toàn thế giới, theo báo cáo của Tổ chức Y tế Thế giới (WHO), vào năm 2018. Do sự hiện diện của giai đoạn không có triệu chứng tương đối dài, nên việc phát hiện sớm bệnh tiểu đường là luôn mong muốn cho một kết quả có ý nghĩa về mặt lâm sàng. Khoảng 50% tất cả những người mắc bệnh tiểu đường không được chẩn đoán vì giai đoạn không có triệu chứng kéo dài của nó.

Việc chẩn đoán sớm bệnh tiểu đường chỉ có thể thực hiện được bằng cách đánh giá đúng các triệu chứng dấu hiệu phổ biến và ít phổ biến hơn, có thể được tìm thấy trong các giai đoạn khác nhau từ khi bắt đầu phát bệnh cho đến khi chẩn đoán.

Kỹ thuật phân loại khai thác dữ liệu đã được các nhà nghiên cứu chấp nhận tốt cho mô hình dự báo rủi ro của bệnh. Để dự đoán khả năng mắc bệnh tiểu đường cần một bộ dữ liệu, trong đó chứa dữ liệu của bệnh nhân tiểu đường mới hoặc sẽ là bệnh nhân tiểu đường.

Trong nghiên cứu này, nhóm chúng tôi xây dựng hỗ trợ chuẩn đoán bệnh tiểu đường giúp mọi người có thể tự đánh giá được mình có đang mắc nguy cơ tiểu đường hay không để đi khám chữa bệnh kịp thời.

2.2. DỮ LIỆU BỆNH TIỂU ĐƯỜNG

Dữ liệu bệnh tiểu đường được được lấy từ website:

https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#

Dữ liệu được tổng hợp từ các bệnh nhân bị bệnh tiểu đường ở Bệnh viện ở Sylhet, Bangladesh bởi 4 bác sĩ:

1) M M Faniqul Islam, Đại học Queen Mary của London, Vương quốc Anh, m.islam '@' smd17.qmul.ac.uk

2) Rahatara Ferdousi, Đại học Metropolitan Sylhet, Bangladesh, rahatara '@' metrouni.edu.bd
3) Sadikur Rahman, and Humayra, Đại học Metropolitan Sylhet, Bangladesh, rahmansadik004 '@' gmail.com
4) Yasmin Bushra, Đại học Thủ đô Sylhet, Bangladesh, humayrabushra234 '@' gmail.com

2.3. MÔ TẢ THUỘC TÍNH DỮ LIỆU

Đặc điểm của tập dữ liệu: Đa biến

Số lượng bản ghi: 520

Số thuộc tính: 17

Ngày cập nhật: 12/07/2020

STT	Thuộc tính	Kiểu	Giá trị	Diễn giải
1	Age	Numeric	16-90	Tuổi
2	Gender	Norminal	Male, Female	Giới tính bệnh nhân
3	Polyuria	Norminal	Yes, No	Triệu chứng đi tiểu nhiều (khối lượng >= 2.5 lít trong vòng 24 giờ ở người lớn)
4	Polydipsia	Norminal	Yes, No	Triệu chứng khát nước, luôn cảm muốn uống nước bất kể uống bao nhiêu nước vẫn thấy khô miệng
5	Sudden Weight Loss	Norminal	Yes, No	Cơ thể bị sụt cân không rỏ nguyên nhân trong một thời gian ngắn
6	Weakness	Norminal	Yes, No	Triệu chứng mệt mỏi, cơ thể luôn cảm thấy mệt mỏi, kiệt sức, đặc biệt khi vừa ngủ dậy
7	Polyphagia	Norminal	Yes, No	Triệu chứng đói quá mức, luôn muốn ăn cho dù vừa ăn xong, lúc nào cũng cảm thấy đói

8	Genital Thrush	Norminal	Yes, No	Bị bệnh tưa miệng, xuất hiện các mảng trắng trong miệng hoặc lưỡi đỏ khác thường
9	Visual Blurring	Norminal	Yes, No	Triệu chứng mờ mắt, thị lực giảm suốt, có hiện tượng xuất huyết, phù nề trong mắt
10	Itching	Norminal	Yes, No	Triệu chứng ngứa, da bị khô, bong tróc và nứt nẻ
11	Irritability	Norminal	Yes, No	Triệu chứng khó chịu, cơ thể luôn bứt rứt, khó chịu và hay cáu gắt
12	Delayed Healing	Norminal	Yes, No	Triệu chứng khó lành vết thương, xuất hiện các biến chứng khác trong quá trình hồi phục
13	Partial Paresis	Norminal	Yes, No	Triệu chứng liệt, cơ thể sẽ bị liệt một bộ phận nào đó
14	Muscle Stiffness	Norminal	Yes, No	Triệu chứng cứng cơ, cảm thấy cơ xương khớp bị cứng, khó vận động
15	Alopecia	Norminal	Yes, No	Triệu chứng rụng tóc, tóc rất yếu, mỏng và rụng nhiều
16	Obesity	Norminal	Yes, No	Mắc bệnh béo phì
17	Class	Norminal	Positive, Negative	Dương tính, Âm tính

Bảng 0.1: Các thuộc tính và giá trị của dữ liệu mẫu

2.4. CÀI ĐẶT THUẬT TOÁN TRÊN TẬP DỮ LIỆU

Từ dữ liệu lưu trữ ta rút trích 21 mẫu dữ liệu theo bảng sau:

Age	Gender	Polyuria	Polydipsia	SuddenWeightLoss	Weakness	Polyphagia	GenitalThrush	VisualBlurring	Itching	Irritability	DelayedHealing	PartialParesis	MuscleStiffness	Alopecia	Obesity	Class
57	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	No	No	Positive
47	Male	No	No	No	No	No	No	No	Yes	No	No	No	No	Yes	No	Negative
45	Male	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	Negative
57	Male	No	No	No	No	Yes	No	Yes	No	No	No	No	Yes	No	No	Negative
72	Male	Yes	No	No	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Negative
30	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
27	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
38	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
43	Male	No	No	No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	No	Negative
40	Male	No	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes	Negative
47	Male	Yes	Yes	No	No	No	No	No	No	No	No	No	Yes	No	No	Positive
62	Male	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes	No	Positive
49	Male	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No	No	Positive
53	Male	Yes	No	Yes	No	No	No	No	No	No	Yes	Yes	No	No	No	Positive
68	Male	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	Positive
61	Male	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Positive
39	Male	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	Positive
38	Male	Yes	No	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	Positive
44	Male	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	Yes	Negative
36	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
43	Male	No	No	No	Yes	No	Yes	No	Yes	No	No	No	No	Yes	No	Negative

Bảng 0.2: Bảng dữ liệu mẫu

Ta áp dụng tính độ đo GainRatio cho các thuộc tính theo bảng dữ liệu mẫu trên để xác định thuộc tính nào được chọn trong quá trình tạo cây quyết định.

Bộ mẫu dữ liệu của chúng ta có 02 miền giá trị {d, a} (d ứng với "Positive" và a ứng với "Negative")

Trước tiên, ta tính lượng thông tin trên tất cả mẫu dữ liệu S theo bảng trên:

$$I(S) = -\frac{9}{21} log_2\left(\frac{9}{21}\right) - \frac{12}{21} log_2\left(\frac{12}{21}\right) = 0.985$$

❖ Tính GainRatio cho thuộc tính Polyuria:

Bảng Entropy của thuộc tính Polyuria				
STT	Polyuria	d_i	a_i	$I(d_i, a_i)$
1	Yes (11)	9	2	0.684
2	No (10)	0	10	0

Ta có:

$$E(Polyuria) = \frac{11}{21} * I(d_1, a_1) + \frac{10}{21} * I(d_2, a_2) = \frac{11}{21} * 0.684 + \frac{10}{21} * 0 = 0.358$$

Trong đó:

$$I(d_1, a_1) = -\frac{9}{11} * \log_2 \frac{9}{11} - \frac{2}{11} * \log_2 \frac{2}{11} = 0.684$$

$$I(d_2, a_2) = -\frac{0}{10} * \log_2 \frac{0}{10} - \frac{10}{10} * \log_2 \frac{10}{10} = 0$$

Do đó:

$$Gain(Polyuria) = I(S) - E(Polyuria) = 0.985 - 0.358 = 0.627$$

Tính độ đo SplitInfo cho thuộc tính Polyuria:

$$SplitInfo(Polyuria) = -\frac{11}{21}\log_2 \frac{11}{21} - \frac{10}{21}\log_2 \frac{10}{21} = 0.998$$

Vậy ta tính được độ đo GainRatio cho thuộc tính Polyuria:

$$GainRatio(Polyuria) = Gain(Polyuria)/SplitInfo(Polyuria) = 0.627/0.998 = 0.628$$

❖ Tính GainRatio cho thuộc tính Polydipsia:

$$I(S) = -\frac{9}{21} log_2 \left(\frac{9}{21}\right) - \frac{12}{21} log_2 \left(\frac{12}{21}\right) = 0.985$$

Bảng Entropy của thuộc tính Polydipsia				
STT	Polydipsia	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	6	0	0
2	No (15)	3	12	0.722

Ta có:

$$E(Polydipsia) = \frac{6}{21} * I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 0 + \frac{15}{21} * 0.722 = 0.516$$

Trong đó:

$$I(d_1, a_1) = -\frac{6}{6} * \log_2 \frac{6}{6} - \frac{0}{6} * \log_2 \frac{0}{6} = 0$$

$$I(d_2, a_2) = -\frac{3}{15} * \log_2 \frac{3}{15} - \frac{12}{15} * \log_2 \frac{12}{15} = 0.722$$

Do đó:

Gain(Polydipsia) = I(S) – E(Polydipsia) = 0.985-0.516=0.470

Tính độ đo SplitInfo cho thuộc tính Polydipsia:

$$\text{SplitInfo(Polydipsia)} = -\frac{6}{21}\log_2\frac{6}{21} - \frac{15}{21}\log_2\frac{15}{21} = 0.863$$

Vậy ta tính được độ đo GainRatio cho thuộc Polydipsia:

GainRatio(Polydipsia) = Gain(Polydipsia)/SplitInfo(Polydipsia)

= 0.470/0.863=0.544

❖ Tính GainRatio cho thuộc tính SuddenWeightLoss:

Bảng Entropy của thuộc tính SuddenWeightLoss				
STT	SuddenWeightLoss	d_i	a_i	$I(d_i, a_i)$
1	Yes (5)	3	2	0.971
2	No (16)	6	10	0.954

Ta có:

$$E(\text{SuddenWeightLoss}) = \frac{5}{21} * I(d_1, a_1) + \frac{16}{21} * I(d_2, a_2) = \frac{5}{21} * 0.971 + \frac{16}{21} * 0.954 = 0.958$$

Gain(SuddenWeightLoss) = 0.985 - 0.958 = 0.027

$$\text{SplitInfo(SuddenWeightLoss)} = -\frac{5}{21}\log_2\frac{5}{21} - \frac{16}{21}\log_2\frac{16}{21} = 0.792$$

GainRatio(SuddenWeightLoss) = 0.027/0.792 = 0.034

❖ Tính GainRatio cho thuộc tính Weakness:

Bảng Entropy của thuộc tính Weakness				
STT	Weakness	d_i	a_i	$I(d_i, a_i)$
1	Yes (5)	2	3	0.971
2	No (16)	7	9	0.989

Ta có:

E(Weakness) = $\frac{5}{21}$ * I(d_1, a_1) + $\frac{16}{21}$ * I(d_2, a_2) = $\frac{5}{21}$ * 0.971+ $\frac{16}{21}$ * 0.989= 0.984

Gain(Weakness) = 0.985 - 0.984 = 0.001

SplitInfo(Weakness) = $-\frac{5}{21}\log_2\frac{5}{21} - \frac{16}{21}\log_2\frac{16}{21} = 0.792$

GainRatio(Weakness) = 0.001/0.792 = 0.001

❖ Tính GainRatio cho thuộc tính Polyphagia:

Bảng Entropy của thuộc tính Polyphagia				
STT	Polyphagia	d_i	a_i	I(d_i, a_i)
1	Yes (7)	4	3	0.985
2	No (14)	5	9	0.940

Ta có:

E(Polyphagia) = $\frac{7}{21}$ * I(d_1, a_1) + $\frac{14}{21}$ * I(d_2, a_2) = $\frac{7}{21}$ * 0.985+ $\frac{14}{21}$ * 0.940= 0.955

Gain(Polyphagia) = 0.985- 0.955 = 0.030

SplitInfo(Polyphagia) = $-\frac{7}{21}\log_2\frac{7}{21} - \frac{14}{21}\log_2\frac{14}{21} = 0.918$

GainRatio(Polyphagia) = 0.030/0.918= 0.033

❖ Tính GainRatio cho thuộc tính GenitalThrush:

Bảng Entropy của thuộc tính GenitalThrush				
STT	GenitalThrush	d_i	a_i	I(d_i, a_i)
1	Yes (7)	3	4	0.985
2	No (14)	6	8	0.985

Ta có:

$E(GenitalThrush) = \frac{7}{21} * I(d_1, a_1) + \frac{14}{21} * I(d_2, a_2) = \frac{7}{21} * 0.985 + \frac{14}{21} * 0.985 = 0.985$

Gain(GenitalThrush) = 0.985– 0.985= 0

$SplitInfo(GenitalThrush) = -\frac{7}{21}\log_2\frac{7}{21} - \frac{14}{21}\log_2\frac{14}{21} = 0.918$

GainRatio(GenitalThrush) = 0/0.918 = 0

❖ Tính GainRatio cho thuộc tính VisualBlurring:

Bảng Entropy của thuộc tính VisualBlurring				
STT	VisualBlurring	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	4	2	0.918
2	No (15)	5	10	0.918

Ta có:

$E(VisualBlurring) = \frac{6}{21} * I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 0.918 + \frac{15}{21} * 0.918 = 0.918$

Gain(VisualBlurring) = 0.985– 0.918 = 0.067

$SplitInfo(VisualBlurring) = -\frac{6}{21}\log_2\frac{6}{21} - \frac{15}{21}\log_2\frac{15}{21} = 0.863$

GainRatio(VisualBlurring) = 0.067/0.863 = 0.078

❖ Tính GainRatio cho thuộc tính Itching:

Bảng Entropy của thuộc tính Itching				
STT	Itching	d_i	a_i	$I(d_i, a_i)$
1	Yes (10)	5	5	1
2	No (11)	3	8	0.845

Ta có:

$E(Itching) = \frac{10}{21} * I(d_1, a_1) + \frac{11}{21} * I(d_2, a_2) = \frac{10}{21} * 1 + \frac{11}{21} * 0.845 = 0.919$

Gain(Itching) = 0.985 – 0.919 = 0.066

SplitInfo(Itching) = $-\frac{10}{21}\log_2\frac{10}{21} - \frac{11}{21}\log_2\frac{11}{21} = 0.998$

GainRatio(Itching) = 0.066/0.998 = 0.066

❖ Tính GainRatio cho thuộc tính Irritability:

Bảng Entropy của thuộc tính Irritability				
STT	Irritability	d_i	a_i	$I(d_i, a_i)$
1	Yes (3)	3	0	0
2	No (18)	6	12	0.918

Ta có:

E(Irritability) = $\frac{3}{21}$ * $I(d_1, a_1)$ + $\frac{18}{21}$ * $I(d_2, a_2)$ = $\frac{3}{21}$ * 0 + $\frac{18}{21}$ * 0.918= 0.787

Gain(Irritability) = 0.985 – 0.787 = 0.198

SplitInfo(Irritability) = $-\frac{3}{21}\log_2\frac{3}{21} - \frac{18}{21}\log_2\frac{18}{21} = 0.592$

GainRatio(Irritability) = 0.198/0.592 = 0.335

❖ Tính GainRatio cho thuộc tính DelayedHealing:

Bảng Entropy của thuộc tính DelayedHealing				
STT	DelayedHealing	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	3	3	1
2	No (15)	6	9	0.971

Ta có:

E(DelayedHealing) = $\frac{6}{21}$ * $I(d_1, a_1)$ + $\frac{15}{21}$ * $I(d_2, a_2)$ = $\frac{6}{21}$ * 1 + $\frac{15}{21}$ * 0.971= 0.979

Gain(DelayedHealing) = 0.985 – 0.979 = 0.006

SplitInfo(DelayedHealing) = $-\frac{6}{21}\log_2\frac{6}{21} - \frac{15}{21}\log_2\frac{15}{21} = 0.863$

GainRatio(DelayedHealing) = 0.006/0.863 = 0.007

❖ Tính GainRatio cho thuộc tính PartialParesis:

Bảng Entropy của thuộc tính PartialParesis				
STT	PartialParesis	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	5	1	0.650
2	No (15)	4	11	0.837

Ta có:

E(PartialParesis) = $\frac{6}{21}$ * $I(d_1, a_1)$ + $\frac{15}{21}$ * $I(d_2, a_2)$ = $\frac{6}{21}$ * 0.650+ $\frac{15}{21}$ * 0.837= 0.738

Gain(PartialParesis) = 0.985 – 0.738 = 0.202

SplitInfo(PartialParesis) = $-\frac{6}{21}\log_2\frac{6}{21} - \frac{15}{21}\log_2\frac{15}{21} = 0.863$

GainRatio(PartialParesis) = 0.202/0.863 = 0.234

❖ Tính GainRatio cho thuộc tính MuscleStiffness:

Bảng Entropy của thuộc tính MuscleStiffness				
STT	MuscleStiffness	d_i	a_i	$I(d_i, a_i)$
1	Yes (5)	3	2	0.971
2	No (16)	6	10	0.954

Ta có:

E(MuscleStiffness) = $\frac{5}{21}$ * $I(d_1, a_1)$ + $\frac{16}{21}$ * $I(d_2, a_2)$ = $\frac{5}{21}$ * 0.971 + $\frac{16}{21}$ * 0.954 = 0.958

Gain(MuscleStiffness) = 0.985 – 0.958= 0.027

SplitInfo(MuscleStiffness) = $-\frac{5}{21}\log_2\frac{5}{21} - \frac{16}{21}\log_2\frac{16}{21} = 0.792$

GainRatio(MuscleStiffness) = 0.027/0.792 = 0.034

❖ Tính GainRatio cho thuộc tính Alopecia:

Bảng Entropy của thuộc tính Alopecia				
STT	Alopecia	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	1	5	0.650
2	No (15)	8	7	0.997

Ta có:

E(Alopecia) = $\frac{6}{21}$ * $I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 0.650 + \frac{15}{21} * 0.997 = 0.898$

Gain(Alopecia) = 0.985– 0.898 = 0.088

SplitInfo(Alopecia) = $-\frac{6}{21}\log_2\frac{6}{21} - \frac{15}{21}\log_2\frac{15}{21} = 0.863$

GainRatio(Alopecia) = 0.088/0.863 = 0.101

❖ Tính GainRatio cho thuộc tính Obesity:

Bảng Entropy của thuộc tính Obesity				
STT	Obesity	d_i	a_i	$I(d_i, a_i)$
1	Yes (4)	2	2	1
2	No (17)	7	10	0.977

Ta có:

E(Obesity) = $\frac{4}{21}$ * $I(d_1, a_1) + \frac{17}{21} * I(d_2, a_2) = \frac{4}{21} * 1 + \frac{17}{21} * 0.977 = 0.982$

Gain(Obesity) = 0.985– 0.982 = 0.003

SplitInfo(Obesity) = $-\frac{4}{21}\log_2\frac{4}{21} - \frac{17}{21}\log_2\frac{17}{21} = 0.702$

GainRatio(Obesity) = 0.003/0.977= 0.005

❖ Tính GainRatio cho thuộc tính Gender:

Bảng Entropy của thuộc tính Gender				
STT	Gender	d_i	a_i	$I(d_i, a_i)$
1	Male (21)	9	12	0.985
2	Female (0)	0	0	0

Ta có:

E(Gender) = $\frac{9}{21}$ * $I(d_1, a_1)$ + $\frac{12}{21}$ * $I(d_2, a_2)$ = $\frac{21}{21}$ * 0.985 + $\frac{0}{21}$ * 0 = 0.985

Gain(Gender) = 0.985 – 0.985 = 0

SplitInfo(Gender) = $-\frac{21}{21}\log_2\frac{21}{21} - \frac{0}{21}\log_2\frac{0}{21} = 0$

GainRatio(Gender) = 0

❖ Tính GainRatio cho thuộc tính Age:

Bảng Entropy của thuộc tính Age				
STT	Age	d_i	a_i	$I(d_i, a_i)$
1	57(2)	1	1	1
2	47(2)	1	1	1
3	45(1)	0	1	0
4	72(1)	0	1	0
5	30(1)	0	1	0

6	27(1)	0	1	0
7	38(2)	1	1	1
8	43(2)	0	2	0
9	40(1)	0	1	0
10	62(1)	1	0	0
11	49(1)	1	0	0
13	53(1)	1	0	0
13	68(1)	1	0	0
14	61(1)	1	0	0
15	39(1)	1	0	0
16	44(1)	0	1	0
17	36(1)	0	1	0

Ta có:

$E(Age) = \frac{2}{21} * I(d_1, a_1) + \frac{2}{21} * I(d_2, a_2) + \frac{1}{21} * I(d_3, a_3) + \frac{1}{21} * I(d_4, a_4) + \frac{1}{21} * I(d_5, a_5) + \frac{1}{21} * I(d_6, a_6) + \frac{2}{21} * I(d_7, a_7) + \frac{2}{21} * I(d_8, a_8) + \frac{1}{21} * I(d_9, a_9) + \frac{1}{21} * I(d_{10}, a_{10}) + \frac{1}{21} * I(d_{11}, a_{11}) + \frac{1}{21} * I(d_{12}, a_{12}) + \frac{1}{21} * I(d_{13}, a_{13}) + \frac{1}{21} * I(d_{14}, a_{14}) + \frac{1}{21} * I(d_{15}, a_{15}) + \frac{1}{21} * I(d_{16}, a_{16}) + \frac{1}{21} * I(d_{17}, a_{17})$

$= \frac{2}{21} *1 + \frac{2}{21} *1 + 0 + 0 + 0 + 0 + \frac{2}{21} *1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 0.286$

Gain(Age) = 0.985 – 0.286 = 0.699

$SplitInfo(Age) = -\frac{2}{21}\log_2\frac{2}{21} - \frac{2}{21}\log_2\frac{2}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{2}{21}\log_2\frac{2}{21} - \frac{2}{21}\log_2\frac{2}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} = 4.011$

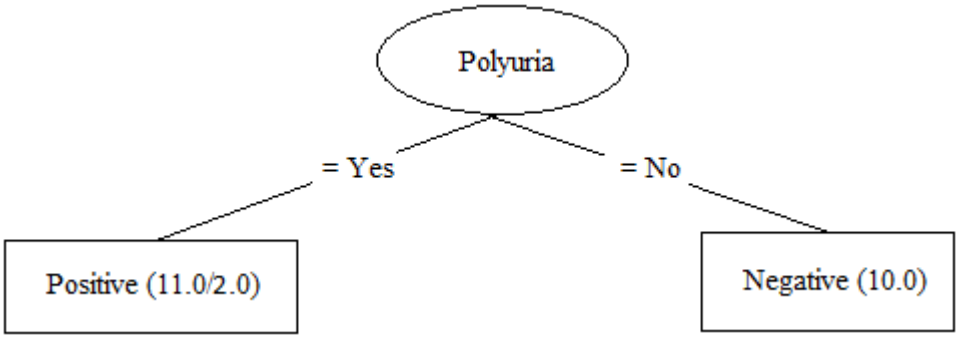
GainRatio(Age) = 0.699/4.011= 0.174

Độ đo GainRatio của các thuộc tính được sắp xếp giảm dần

STT	Thuộc tính	GainRatio
1	Polyuria	0.628
2	Polydipsia	0.544
3	PartialParesis	0.234
4	Irritability	0.335
5	Age	0.174
6	Alopecia	0.101
7	VisualBlurring	0.078
8	Itching	0.066
9	MuscleStiffness	0.034
10	SuddenWeightLoss	0.034
11	Polyphagia	0.033
12	DelayedHealing	0.007
13	Obesity	0.005
14	Weakness	0.001
15	Gender	0
16	GenitalThrush	0

Như vậy thuộc tính có độ đo GainRatio lớn nhất là “Polyuria”.

Cây phân nhánh theo thuộc tính “Polyuria” có hình dạng như sau:



Hình 2.1: Cây quyết định tại thuộc tính Polyuria

Nhận xét: Sau khi phân nhánh cây theo thuộc tính “Polyuria”, ở nút con có thuộc tính “No” có tất cả các mẫu thuộc về một lớp, tuy nhiên thuộc tính “Yes” vẫn chưa có

mẫu nào thuộc về một lớp. Vì vậy ta lập bảng dữ liệu phân theo giá trị tương ứng theo từng nút và tiếp tục phân nhánh cây quyết định theo từng nút này.

Tiếp tục áp dụng thuật toán C4.5 cho từng nút tương ứng bảng dữ liệu sau:

Age	Gender	Polyuria	Polydipsia	SuddenWeightLoss	Weakness	Polyphagia	GenitalThrush	VisualBlurring	Itching	Irritability	DelayedHealing	PartialParesis	MuscleStiffness	Alopecia	Obesity	Class
57	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	No	No	Positive
72	Male	Yes	No	No	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Negative
47	Male	Yes	Yes	No	No	No	No	No	No	No	No	No	Yes	No	No	Positive
62	Male	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes	No	Positive
49	Male	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No	No	Positive
53	Male	Yes	No	Yes	No	No	No	No	No	No	Yes	Yes	No	No	No	Positive
68	Male	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	Positive
61	Male	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Positive
39	Male	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	Positive
38	Male	Yes	No	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	Positive
44	Male	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	Yes	Negative

Bảng 0.3: Mẫu dữ liệu với thuộc tính Polyuria có giá trị Yes(S1)

Tính lượng thông tin trên tất cả mẫu dữ liệu S1:

$$I(S) = -\frac{9}{11} log_2\left(\frac{9}{11}\right) - \frac{2}{11} log_2\left(\frac{2}{11}\right) = 0.684$$

❖ Tính GainRatio cho thuộc tính Gender:

Bảng Entropy của thuộc tính Gender				
STT	Gender	d_i	a_i	$I(d_i, a_i)$
1	Male (11)	9	2	0.684
2	Female (0)	0	0	0

Ta có:

$$E(Gender) = \frac{9}{11} * I(d_1, a_1) + \frac{12}{11} * I(d_2, a_2) = \frac{11}{11} * 0.684 + \frac{0}{11} * 0 = 0.684$$

$$Gain(Gender) = 0.684 - 0.684 = 0$$

$$SplitInfo(Gender) = -\frac{11}{11}\log_2\frac{11}{11} - \frac{0}{11}\log_2\frac{0}{11} = 0$$

$$GainRatio(Gender) = 0$$

❖ Tính GainRatio cho thuộc tính Polydipsia:

Bảng Entropy của thuộc tính Polydipsia				
STT	Polydipsia	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	6	0	0
2	No (5)	3	2	0.971

Ta có:

$$E(\text{Polydipsia}) = \frac{6}{11} * I(d_1, a_1) + \frac{5}{11} * I(d_2, a_2) = \frac{6}{11} * 0 + \frac{5}{11} * 0.971 = 0.441$$

$$\text{Gain(Polydipsia)} = 0.684 - 0.441 = 0.243$$

$$\text{SplitInfo(Polydipsia)} = -\frac{6}{11}\log_2\frac{6}{11} - \frac{5}{11}\log_2\frac{5}{11} = 0.994$$

$$\text{GainRatio(Polydipsia)} = 0.243/0.944 = 0.244$$

❖ Tính GainRatio cho thuộc tính SuddenWeightLoss:

Bảng Entropy của thuộc tính SuddenWeightLoss				
STT	SuddenWeightLoss	d_i	a_i	$I(d_i, a_i)$
1	Yes (4)	3	1	0.811
2	No (7)	6	1	0.592

Ta có:

$$E(\text{SuddenWeightLoss}) = \frac{4}{11} * I(d_1, a_1) + \frac{7}{11} * I(d_2, a_2) = \frac{4}{11} * 0.811 + \frac{7}{11} * 0.592 = 0.672$$

$$\text{Gain(SuddenWeightLoss)} = 0.684 - 0.672 = 0.012$$

$$\text{SplitInfo(SuddenWeightLoss)} = -\frac{4}{11}\log_2\frac{4}{11} - \frac{7}{11}\log_2\frac{7}{11} = 0.946$$

$$\text{GainRatio(SuddenWeightLoss)} = 0.684/0.946 = 0.013$$

❖ Tính GainRatio cho thuộc tính Weakness:

Bảng Entropy của thuộc tính Weakness				
STT	Weakness	d_i	a_i	$I(d_i, a_i)$
1	Yes (3)	2	1	0.918
2	No (8)	7	1	0.544

Ta có:

$E(\text{Weakness}) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = \frac{3}{11} * 0.918 + \frac{8}{11} * 0.544 = 0.646$

$\text{Gain(Weakness)} = 0.684 - 0.646 = 0.038$

$\text{SplitInfo(Weakness)} = -\frac{3}{11}\log_2\frac{3}{11} - \frac{8}{11}\log_2\frac{8}{11} = 0.845$

$\text{GainRatio(Weakness)} = 0.038/0.845 = 0.045$

❖ Tính GainRatio cho thuộc tính Polyphagia:

Bảng Entropy của thuộc tính Polyphagia				
STT	Polyphagia	d_i	a_i	$I(d_i, a_i)$
1	Yes (5)	4	1	0.722
2	No (6)	5	1	0.650

Ta có:

$$E(\text{Polyphagia}) = \frac{5}{11} * I(d_1, a_1) + \frac{6}{11} * I(d_2, a_2) = 0.683$$

$\text{Gain(Polyphagia)} = 0.001$

$\text{SplitInfo(Polyphagia)} = -\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11} = 0.994$

$\text{GainRatio(Polyphagia)} = 0.001$

❖ Tính GainRatio cho thuộc tính GenitalThrush:

Bảng Entropy của thuộc tính GenitalThrush				
STT	GenitalThrush	d_i	a_i	$I(d_i, a_i)$
1	Yes (4)	3	1	0.811
2	No (7)	6	1	0.592

Ta có:

$$E(\text{GenitalThrush}) = \frac{4}{11} * I(d_1, a_1) + \frac{7}{11} * I(d_2, a_2) = 0.672$$

Gain(GenitalThrush) = 0.012

$$\text{SplitInfo(GenitalThrush)} = -\frac{4}{11}\log_2\frac{4}{11} - \frac{7}{11}\log_2\frac{7}{11} = 0.946$$

GainRatio(GenitalThrush) = 0.013

❖ Tính GainRatio cho thuộc tính VisualBlurring:

Bảng Entropy của thuộc tính VisualBlurring				
STT	VisualBlurring	d_i	a_i	$I(d_i, a_i)$
1	Yes (5)	4	1	0.722
2	No (6)	5	1	0.650

Ta có:

$$E(\text{VisualBlurring}) = \frac{5}{11} * I(d_1, a_1) + \frac{6}{11} * I(d_2, a_2) = 0.683$$

Gain(VisualBlurring) = 0.001

$$\text{SplitInfo(VisualBlurring)} = -\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11} = 0.994$$

GainRatio(VisualBlurring) = 0.001

❖ Tính GainRatio cho thuộc tính Itching:

Bảng Entropy của thuộc tính Itching				
STT	Itching	d_i	a_i	$I(d_i, a_i)$
1	Yes (7)	5	2	0.863
2	No (4)	4	0	0

Ta có:

$$E(Itching) = \frac{7}{11} * I(d_1, a_1) + \frac{4}{11} * I(d_2, a_2) = 0.549$$

Gain(Itching) = 0.135

$$SplitInfo(Itching) = -\frac{7}{11}\log_2\frac{7}{11} - \frac{4}{11}\log_2\frac{4}{11} = 0.946$$

GainRatio(Itching) = 0.143

❖ Tính GainRatio cho thuộc tính Irritability:

Bảng Entropy của thuộc tính Irritability				
STT	Irritability	d_i	a_i	$I(d_i, a_i)$
1	Yes (3)	3	0	0
2	No (8)	6	2	0.811

Ta có:

$$E(Irritability) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = 0.590$$

Gain(Irritability) = 0.094

$$SplitInfo(Irritability) = -\frac{3}{11}\log_2\frac{3}{11} - \frac{8}{11}\log_2\frac{8}{11} = 0.845$$

GainRatio(Irritability) = 0.111

❖ Tính GainRatio cho thuộc tính DelayedHealing:

Bảng Entropy của thuộc tính DelayedHealing				
STT	DelayedHealing	d_i	a_i	$I(d_i, a_i)$
1	Yes (5)	3	2	0.971
2	No (6)	6	0	0

Ta có:

$$E(DelayedHealing) = \frac{5}{11} * I(d_1, a_1) + \frac{6}{11} * I(d_2, a_2) = 0.441$$

Gain(DelayedHealing) = 0.243

$$SplitInfo(DelayedHealing) = -\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11} = 0.994$$

GainRatio(DelayedHealing) = 0.244

❖ Tính GainRatio cho thuộc tính PartialParesis:

Bảng Entropy của thuộc tính PartialParesis				
STT	PartialParesis	d_i	a_i	$I(d_i, a_i)$
1	Yes (6)	5	1	0.650
2	No (5)	4	1	0.722

Ta có:

$$E(PartialParesis) = \frac{6}{11} * I(d_1, a_1) + \frac{5}{11} * I(d_2, a_2) = 0.683$$

Gain(PartialParesis) = 0.001

$$SplitInfo(PartialParesis) = -\frac{6}{11}\log_2\frac{6}{11} - \frac{5}{11}\log_2\frac{5}{11} = 0.994$$

GainRatio(PartialParesis) = 0.001

❖ Tính GainRatio cho thuộc tính MuscleStiffness:

Bảng Entropy của thuộc tính MuscleStiffness				
STT	MuscleStiffness	d_i	a_i	$I(d_i, a_i)$
1	Yes (4)	3	1	0.811
2	No (7)	6	1	0.592

Ta có:

$$E(\text{MuscleStiffness}) = \frac{4}{11} * I(d_1, a_1) + \frac{7}{11} * I(d_2, a_2) = 0.672$$

Gain(MuscleStiffness) = 0.012

$$\text{SplitInfo(MuscleStiffness)} = -\frac{4}{11}\log_2\frac{4}{11} - \frac{7}{11}\log_2\frac{7}{11} = 0.946$$

GainRatio(MuscleStiffness) = 0.013

❖ Tính GainRatio cho thuộc tính Alopecia:

Bảng Entropy của thuộc tính Alopecia				
STT	Alopecia	d_i	a_i	$I(d_i, a_i)$
1	Yes (3)	1	2	0.918
2	No (8)	8	0	0

Ta có:

$$E(\text{Alopecia}) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = 0.250$$

Gain(Alopecia) = 0.434

$$\text{SplitInfo(Alopecia)} = -\frac{3}{11}\log_2\frac{3}{11} - \frac{8}{11}\log_2\frac{8}{11} = 0.845$$

GainRatio(Alopecia) = 0.513

❖ Tính GainRatio cho thuộc tính Obesity:

Bảng Entropy của thuộc tính Obesity				
STT	Obesity	d_i	a_i	$I(d_i, a_i)$
1	Yes (3)	2	1	0.918
2	No (8)	7	1	0.544

Ta có:

$$E(Obesity) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = 0.646$$

Gain(Obesity) = 0.038

$$SplitInfo(Obesity) = -\frac{3}{11}\log_2\frac{3}{11} - \frac{8}{11}\log_2\frac{8}{11} = 0.845$$

GainRatio(Obesity) = 0.045

❖ Tính GainRatio cho thuộc tính Age:

Bảng Entropy của thuộc tính Age				
STT	Age	d_i	a_i	$I(d_i, a_i)$
1	57	1	0	0
2	72	0	1	0
3	47	1	0	0
4	62	1	0	0
5	49	1	0	0
6	53	1	0	0
7	68	1	0	0
8	61	1	0	0
9	39	1	0	0

10	38	1	0	0
11	44	0	1	0

Ta có:

$$E(Age) = \frac{1}{11} * I(d_1, a_1) + \frac{1}{11} * I(d_2, a_2) + \frac{1}{11} * I(d_3, a_3) + \frac{1}{11} * I(d_4, a_4) +$$

$$\frac{1}{11} * I(d_5, a_5) + \frac{1}{11} * I(d_6, a_6) + \frac{1}{11} * I(d_4, a_4) + \frac{1}{11} * I(d_7, a_7) +$$

$$\frac{1}{11} * I(d_8, a_8) + \frac{1}{11} * I(d_9, a_9) + \frac{1}{11} * I(d_{10}, a_{10}) + \frac{1}{11} * I(d_{11}, a_{11}) = 0$$

Gain(Age) = 0.684

SplitInfo(Age) = 3.459

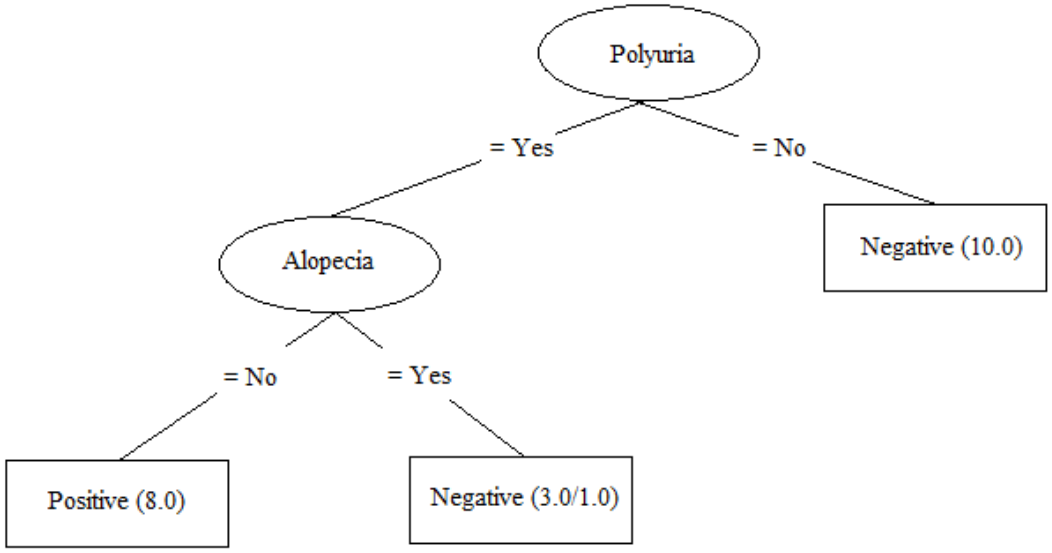
GainRatio(Age) = 0.198

Độ đo GainRatio của các thuộc tính được sắp xếp giảm dần

STT	Thuộc tính	GainRatio
1	Alopecia	0.513
2	Polydipsia	0.244
3	DelayedHealing	0.244
4	Age	0.198
5	Itching	0.143
6	Irritability	0.111
7	Weakness	0.045
8	Obesity	0.045
9	SuddenWeightLoss	0.013
10	GenitalThrush	0.013
11	MuscleStiffness	0.013
12	Polyphagia	0.001
13	VisualBlurring	0.001
14	PartialParesis	0.001
15	Gender	0

Như vậy thuộc tính có độ đo GainRatio lớn nhất là “Alopecia”.

Cây phân nhánh theo thuộc tính “Alopecia” có hình dạng như sau:



Hình 2.2: Cây quyết định tại thuộc tính Alopecia

Tiếp tục áp dụng thuật toán cho dữ liệu thuộc tính Alopecia nhánh “Yes”. Ta có bảng dữ liệu S2 sau:

Age	Gender	Polydipsia	SuddenWeightLoss	Weakness	Polyphagia	GenitalThrush	VisualBlurring	Itching	Irritability	DelayedH	PartialParesis	MuscleStiffness	Obesity	Class
72	Male	No	No	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	Negative
62	Male	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	No	No	Positive
44	Male	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	Negative

Bảng 0.4: Mẫu dữ liệu với thuộc tính Alopecia có giá trị Yes(S2)

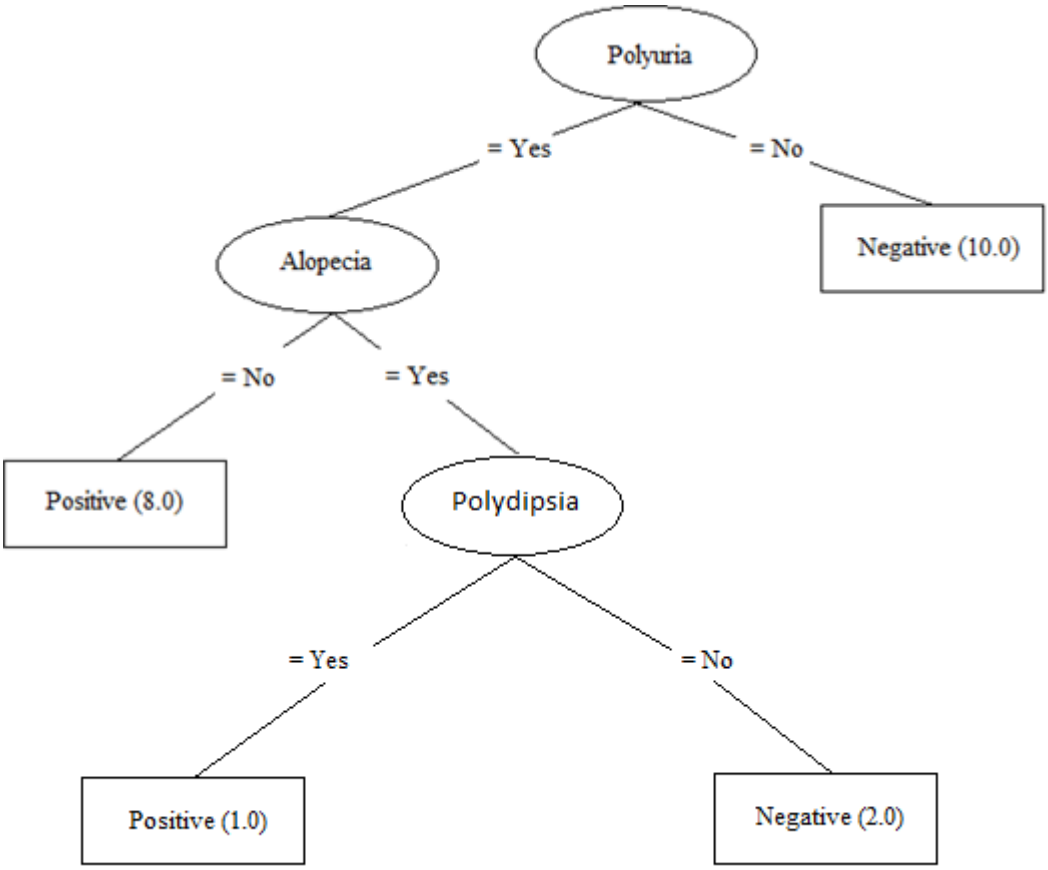
Tương tự tính độ đo GainRatio của các thuộc tính được sắp xếp giảm dần như sau:

STT	Thuộc tính	GainRatio
1	Polydipsia	1
2	Itching	1
3	Irritability	1
4	Age	0.579
5	PartialParesis	0.274
6	VisualBlurring	0.274
7	Weakness	0.274
8	Obesity	0.274

9	SuddenWeightLoss	0.274
10	GenitalThrush	0.274
11	MuscleStiffness	0.274
12	Polyphagia	0.274
13	Gender	0
14	DelayedHealing	0

Như vậy thuộc tính có độ đo GainRatio lớn nhất là “Polydipsia”, “Itching”, “Irritability”. Ta chọn thuộc tính “Polydipsia”.

Cây quyết định với thuật toán C4.5 có dạng như sau:



Hình 2.3: Cây quyết định với bảng dữ liệu mẫu

2.5. DANH SÁCH LUẬT

Sau khi dùng thuật toán C4.5 thao tác trên dữ liệu mẫu ta rút ra được tổng cộng 33 luật sau huấn luyện như sau:

Positive =: (Polyuria == Yes) && (Polydipsia == Yes)

Positive =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == Yes)

Positive =: (Polyuria == Yes) && (Polydipsia == No) && (Itching == No)

Positive =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == No) && (Age <= 38.5)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == Yes) && (GenitalThrush == Yes)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Female) && (Alopecia == Yes) && (Age <= 25.5)

Positive =: (Polyuria == Yes) && (Polydipsia == No) && (Itching == Yes) && (Age <= 69.5) && (Obesity == No)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Female) && (Alopecia == No) && (Irritability == No) && (Age > 34.5)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == Yes) && (Weakness == No)

Positive =: (Polyuria == Yes) && (Polydipsia == No) && (Itching == Yes) && (Age <= 69.5) && (Obesity == Yes) && (DelayedHealing == No)

Positive =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == No) && (Age > 38.5) && (Alopecia == Yes) && (DelayedHealing == No)

Positive =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == No) && (Age > 38.5) && (Alopecia == No) && (Itching == No)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Female) && (Alopecia == No) && (Irritability == No) && (Age <= 34.5) && (VisualBlurring == Yes)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == Yes) && (GenitalThrush == No) && (Age <= 42.5) && (Polyphagia == Yes)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == No) && (DelayedHealing == Yes) && (Age <= 37.5)

Positive =: (Polyuria == Yes) && (Polydipsia == No) && (Itching == Yes) && (Age <= 69.5) && (Obesity == Yes) && (DelayedHealing == Yes) && (Gender == Female)

Positive =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == No) && (Age > 38.5) && (Alopecia == Yes) && (DelayedHealing == Yes) && (Polyphagia == No)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == Yes) && (Weakness == Yes) && (MuscleStiffness == No) && (Age > 45.5)

Positive =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == No) && (DelayedHealing == Yes) && (Age > 37.5) && (Alopecia == No) && (SuddenWeightLoss == Yes)

Negative =: (Polyuria == Yes) && (Polydipsia == No) && (Itching == Yes) && (Age > 69.5)

Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Female) && (Alopecia == Yes) && (Age > 25.5)

Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Female) && (Alopecia == No) && (Irritability == Yes)

Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == No) && (DelayedHealing == No)

Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == Yes) && (GenitalThrush == No) && (Age > 42.5)

Negative =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == No) && (Age > 38.5) && (Alopecia == No) && (Itching == Yes)

Negative =: (Polyuria == Yes) && (Polydipsia == No) && (Itching == Yes) && (Age <= 69.5) && (Obesity == Yes) && (DelayedHealing == Yes) && (Gender == Male)

Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == Yes) && (GenitalThrush == No) && (Age <= 42.5) && (Polyphagia == No)

Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Female) && (Alopecia == No) && (Irritability == No) && (Age <= 34.5) && (VisualBlurring == No)
Negative =: (Polyuria == No) && (Polydipsia == Yes) && (Irritability == No) && (Age > 38.5) && (Alopecia == Yes) && (DelayedHealing == Yes) && (Polyphagia == Yes)
Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == Yes) && (Weakness == Yes) && (MuscleStiffness == Yes)
Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == No) && (DelayedHealing == Yes) && (Age > 37.5) && (Alopecia == Yes)
Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == Yes) && (Weakness == Yes) && (MuscleStiffness == No) && (Age <= 45.5)
Negative =: (Polyuria == No) && (Polydipsia == No) && (Gender == Male) && (Irritability == No) && (PartialParesis == No) && (DelayedHealing == Yes) && (Age > 37.5) && (Alopecia == No) && (SuddenWeightLoss == No)

Dựa vào bảng các tập luật trên ta có thể diễn giải ra các tập luật như sau:

1. Nếu bệnh nhân có triệu chứng tiểu nhiều, có triệu chứng khát nhiều thì mắc bệnh tiểu đường.
2. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nhiều, có triệu chứng khó chịu thì mắc bệnh tiểu đường.
3. Nếu bệnh nhân có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, không có triệu chứng ngứa thì mắc bệnh tiểu đường.
4. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nhiều, không có triệu chứng khó chịu và tuổi nhỏ hơn hoặc bằng 38.5 thì mắc bệnh tiểu đường.

5. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, có triệu trứng khó chịu, bị bệnh tưa miệng thì mắc bệnh tiểu đường.
6. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nữ, có triệu chứng rụng tóc, tuổi nhỏ hơn hoặc bằng 25.5 thì mắc bệnh tiểu đường.
7. Nếu bệnh nhân có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, có triệu chứng ngứa, tuổi nhỏ hơn hoặc bằng 69.5, có mắc bệnh béo phì thì mắc bệnh tiểu đường.
8. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nữ, không có triệu chứng rụng tóc, không có triệu chứng khó chịu, tuổi lớn hơn 34.5 thì mắc bệnh tiểu đường.
9. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, có triệu chứng liệt, không có triệu chứng mệt mỏi thì mắc bệnh tiểu đường.
10. Nếu bệnh nhân có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, có triệu chứng ngứa, tuổi nhỏ hơn hoặc bằng 69.5, có mắc bệnh béo phì, không có triệu chứng khó lành vết thương thì mắc bệnh tiểu đường.
11. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nhiều, không có triệu chứng khó chịu, tuổi lớn hơn 38.5, có triệu chứng rụng tóc, không có triệu chứng khó lành vết thương thì mắc bệnh tiểu đường.
12. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nước, không có triệu chứng khó chịu, tuổi lớn hơn 38.5, không có triệu chứng rụng tóc, không có triệu chứng ngứa thì mắc bệnh tiểu đường.
13. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nữ, không có triệu chứng rụng tóc, không có triệu chứng khó chịu, tuổi nhỏ hơn hoặc bằng 34.5, có triệu chứng mờ mắt thì mắc bệnh tiểu đường.
14. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, có triệu chứng khó chịu, không bị tưa miệng, tuổi

nhỏ hơn hoặc bằng 42.5, có triệu chứng đói quá mức thì mắc bệnh tiểu đường.

15. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, không có triệu chứng liệt, có triệu chứng khó lành vết thương, tuổi nhỏ hơn hoặc bằng 37.5 thì mắc bệnh tiểu đường.
16. Nếu bệnh nhân có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, có triệu chứng ngứa, tuổi nhỏ hơn hoặc bằng 69.5, có mắc bệnh béo phì, có triệu chứng khó lành vết thương, giới tính là nữ thì mắc bệnh tiểu đường.
17. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nhiều, không có triệu chứng khó chịu, tuổi lớn hơn 38.5, có triệu chứng rụng tóc, có triệu chứng khó lành vết thương, không có triệu chứng đói quá mức thì mắc bệnh tiểu đường.
18. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, có triệu chứng liệt, có triệu chứng mệt mỏi, không có triệu chứng cứng cơ, tuổi lớn hơn 45.5 thì mắc bệnh tiểu đường.
19. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, không có triệu chứng liệt, có triệu chứng khó lành vết thương, tuổi lớn hơn 37.5, không có triệu chứng rụng tóc, có triệu chứng sụt cân không rõ nguyên nhân thì mắc bệnh tiểu đường.
20. Nếu bệnh nhân có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, có triệu chứng ngứa, tuổi lớn hơn 69.5 thì không mắc bệnh tiểu đường.
21. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nữ, có triệu chứng rụng tóc, tuổi lớn hơn 25.5 thì không mắc bệnh tiểu đường.
22. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nữ, không có triệu chứng rụng tóc, có triệu chứng khó chịu thì không mắc bệnh tiểu đường.

23. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, không có triệu chứng liệt, không có triệu chứng khó lành vết thương thì không mắc bệnh tiểu đường.
24. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, có triệu chứng khó chịu, không bị bệnh tưa miệng, tuổi lớn hơn 42.5 thì không mắc bệnh tiểu đường.
25. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nhiều, không có triệu chứng khó chịu, tuổi lớn hơn 38.5, không có triệu chứng rụng tóc, có triệu chứng ngứa thì không mắc bệnh tiểu đường.
26. Nếu bệnh nhân có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, có triệu chứng ngứa, tuổi nhỏ hơn hoặc bằng 69.5, có mắc bệnh béo phì, có triệu chứng khó lành vết thương, giới tính là nam thì không mắc bệnh tiểu đường.
27. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, có triệu chứng khó chịu, không bị bệnh tưa miệng, tuổi nhỏ hơn hoặc bằng 42.5, không có triệu chứng đói quá mức thì không mắc bệnh tiểu đường.
28. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nữ, không có triệu chứng rụng tóc, không có triệu chứng khó chịu, tuổi nhỏ hơn hoặc bằng 34.5, không có triệu chứng mờ mắt thì không mắc bệnh tiểu đường.
29. Nếu bệnh nhân không có triệu chứng tiểu nhiều, có triệu chứng khát nhiều, không có triệu chứng khó chịu, tuổi lớn hơn 38.5, có triệu chứng rụng tóc, có triệu chứng khó lành vết thương, có triệu chứng đói quá mức thì không mắc bệnh tiểu đường.
30. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, có triệu chứng liệt, có triệu chứng mệt mỏi, có triệu chứng cứng cơ thì không mắc bệnh tiểu đường.

31. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, không có triệu chứng liệt cơ, có triệu chứng khó lành vết thương, tuổi lớn hơn 37.5, có triệu chứng rụng tóc thì không mắc bệnh tiểu đường.
32. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, có triệu chứng liệt, có triệu chứng mệt mỏi, không có triệu chứng cứng cơ thì không mắc bệnh tiểu đường.
33. Nếu bệnh nhân không có triệu chứng tiểu nhiều, không có triệu chứng khát nhiều, giới tính là nam, không có triệu chứng khó chịu, không có triệu chứng liệt, có triệu chứng khó lành vết thương, tuổi lớn hơn 37.5, không có triệu chứng rụng tóc, không có triệu chứng sụt cân không rõ nguyên nhân thì không mắc bệnh tiểu đường.

Chương 3
ỨNG DỤNG CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

3.1. GIỚI THIỆU ỨNG DỤNG

Ứng dụng ESMedical là website dùng để quản lý thông tin dữ liệu về triệu chứng, tạo ra các bộ luật và hỗ trợ người dùng tra cứu thông tin về bệnh tiểu đường.

ESMedical phát triển trên nền tảng Microsoft .Net Framework 4.0, sử dụng ngôn ngữ C#, JavaScript để lập trình.

ESMedical sử dụng thư viện mã nguồn mở http://accord-framework.net/ để khai phá dữ liệu và tạo luật sử dụng để chuẩn đoán.

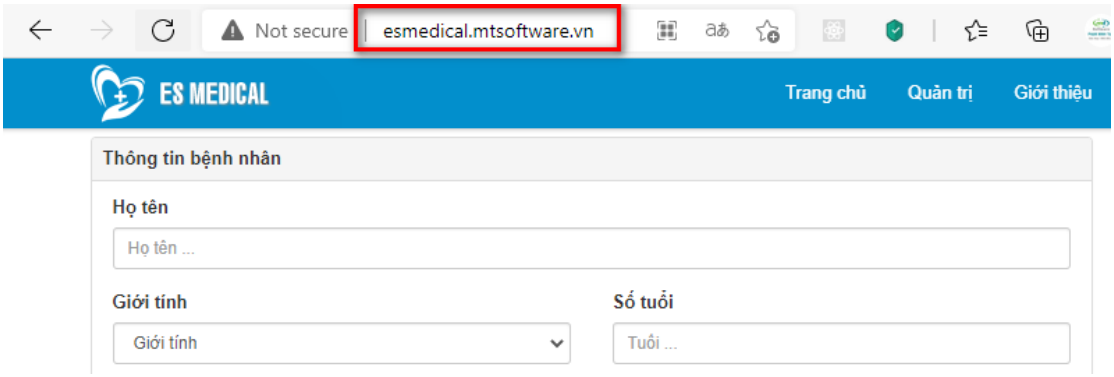
Về mặt giao diện sử dụng Bootstrap 3.7 và các thành phần phụ thuộc khác để triển khai ứng dụng.

3.2. GIAO DIỆN VÀ TÍNH NĂNG

3.2.1. Truy cập ứng dụng

Ứng dụng đã được triển khai trên internet nên để truy cập trên máy tính hoặc điện thoại chỉ cần mở trình duyệt và truy cập đường dẫn sau để vào ứng dụng.

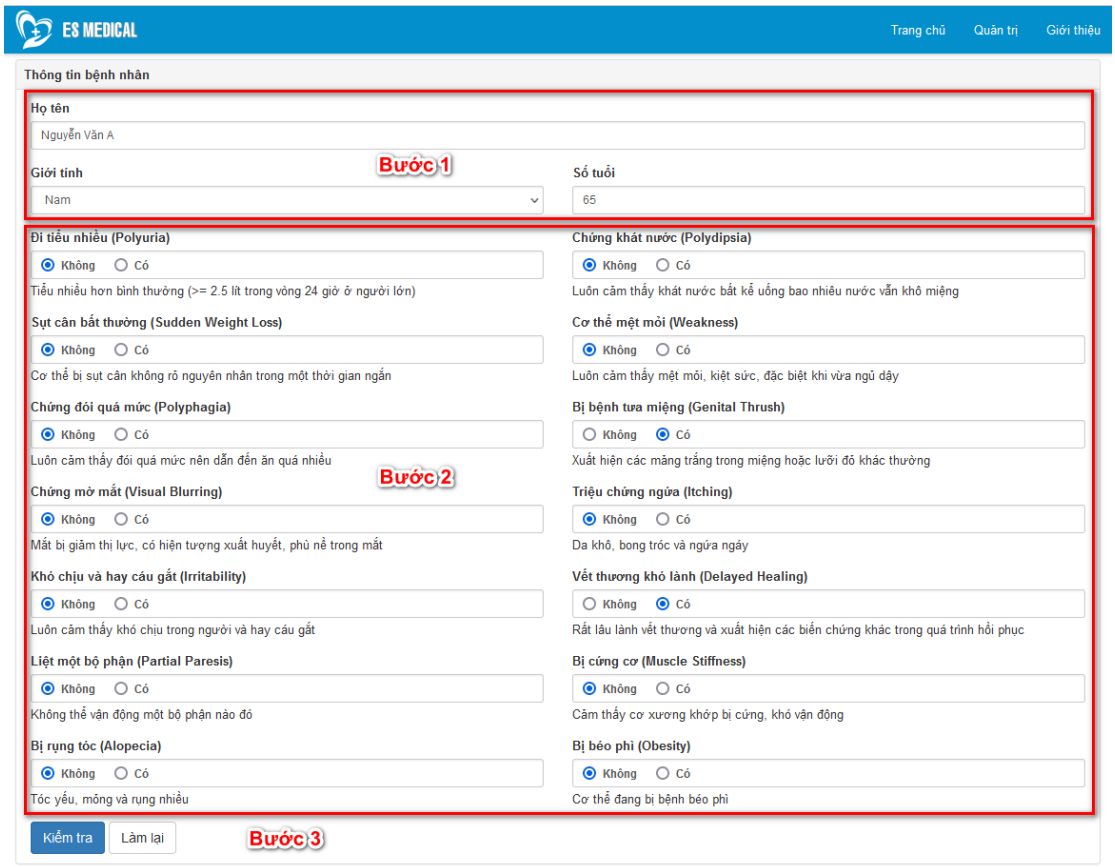
http://esmedical.mtsoftware.vn/



Hình 3.1: Truy cập ứng dụng Hỗ trợ chuẩn đoán bệnh tiểu đường

3.2.2. Tính năng Tra cứu bệnh

Để thực hiện tra cứu có mắc bệnh tiểu đường hay không, sau khi truy cập vào website sẽ hiển thị giao diện trang chủ như hình sau (Hình 3.2):



Hình 3.2: Giao diện trang chủ website

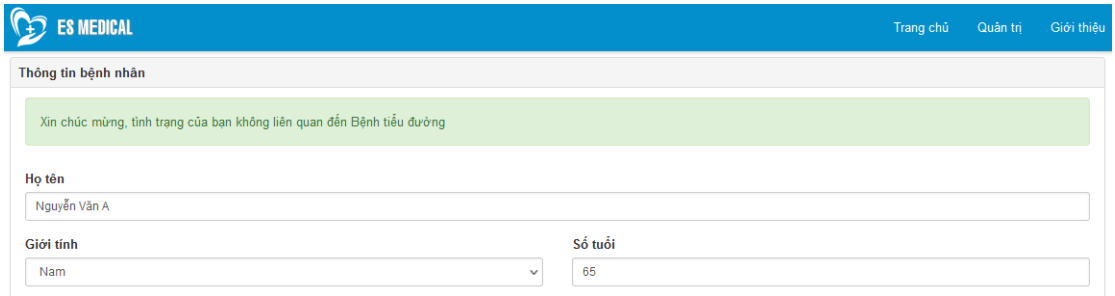
Tại giao diện trang chủ, để tra cứu thông tin người dùng cần thao tác theo thứ tự các bước như sau:

Bước 1: Nhập Họ tên, Giới tính, Tuổi.

Bước 2: Chọn các triệu chứng mà người bệnh đang gặp.

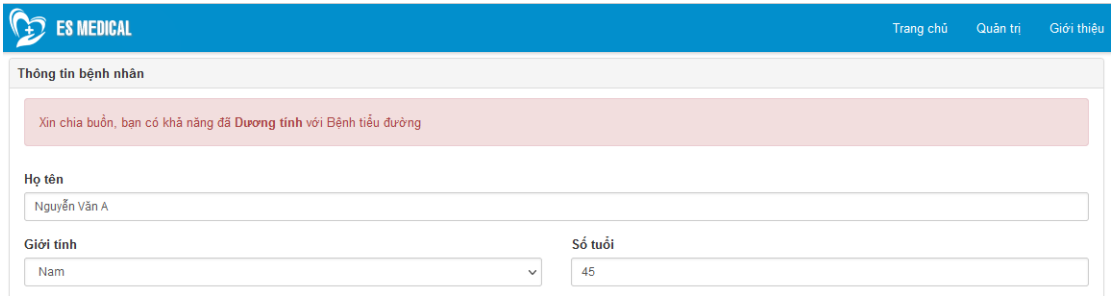
Bước 3: Bấm nút Kiểm tra để xem kết quả.

Nếu không mắc bệnh, ứng dụng sẽ trả kết quả như bên dưới (Hình 3.3):



Hình 3.3: Giao diện trang kết quả nếu không mắc bệnh

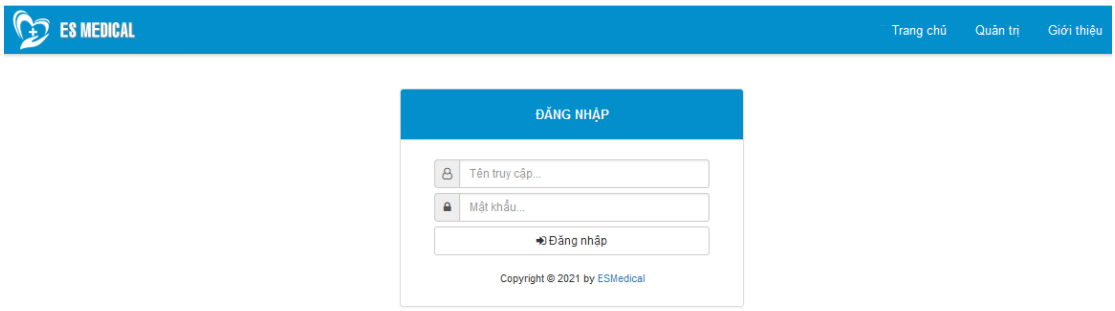
Nếu người dùng mắc bệnh, ứng dụng sẽ trả kết quả như bên dưới (hình 3.4):



Hình 3.4: Giao diện trang kết quả người dùng mắc bệnh

3.2.3. Tính năng Quản trị

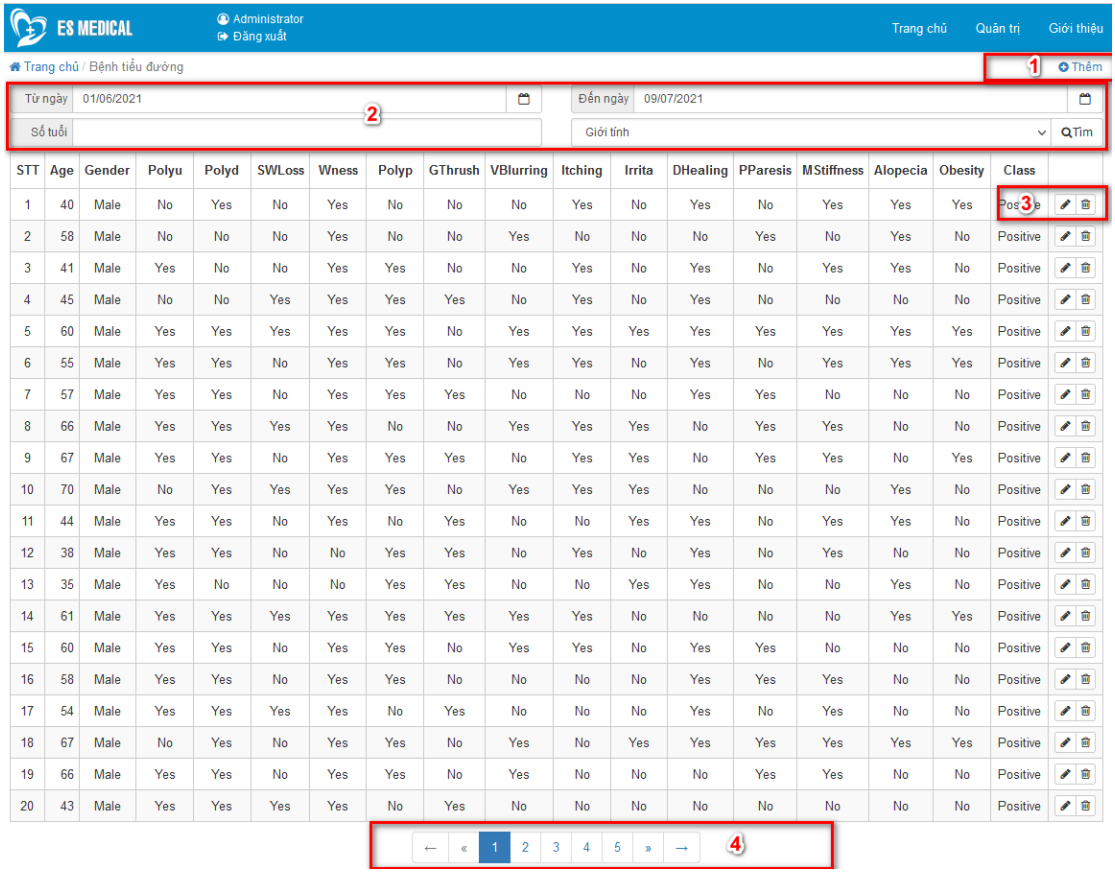
Để quản lý thông tin dữ liệu về triệu chứng, ta kích vào liên kết Quản trị ở menu chính. Nếu chưa đăng nhập thì website sẽ hiển thị trang đăng nhập để người dùng nhập thông tin như hình vẽ dưới (Hình 3.5):



Hình 0.5: Giao diện trang đăng nhập quản trị

Nếu đăng nhập thành công, website sẽ hiển thị trang quản trị thông tin dữ liệu như bên dưới (hình 3.6):

- ***Mục 1:*** Nếu muốn Thêm mới triệu chứng, kích Thêm để mở trang thêm mới.
- ***Mục 2:*** Các tùy chọn tìm kiếm như Ngày tháng, Tuổi, Giới tính. Sau khi chọn để tìm kiếm ta kích vào nút Tìm kiếm.
- ***Mục 3:*** 2 nút Sửa và Xóa dòng thông tin tương ứng
- ***Mục 4:*** Nếu có nhiều trang, chọn vào số trang tương ứng để hiển thị dữ liệu



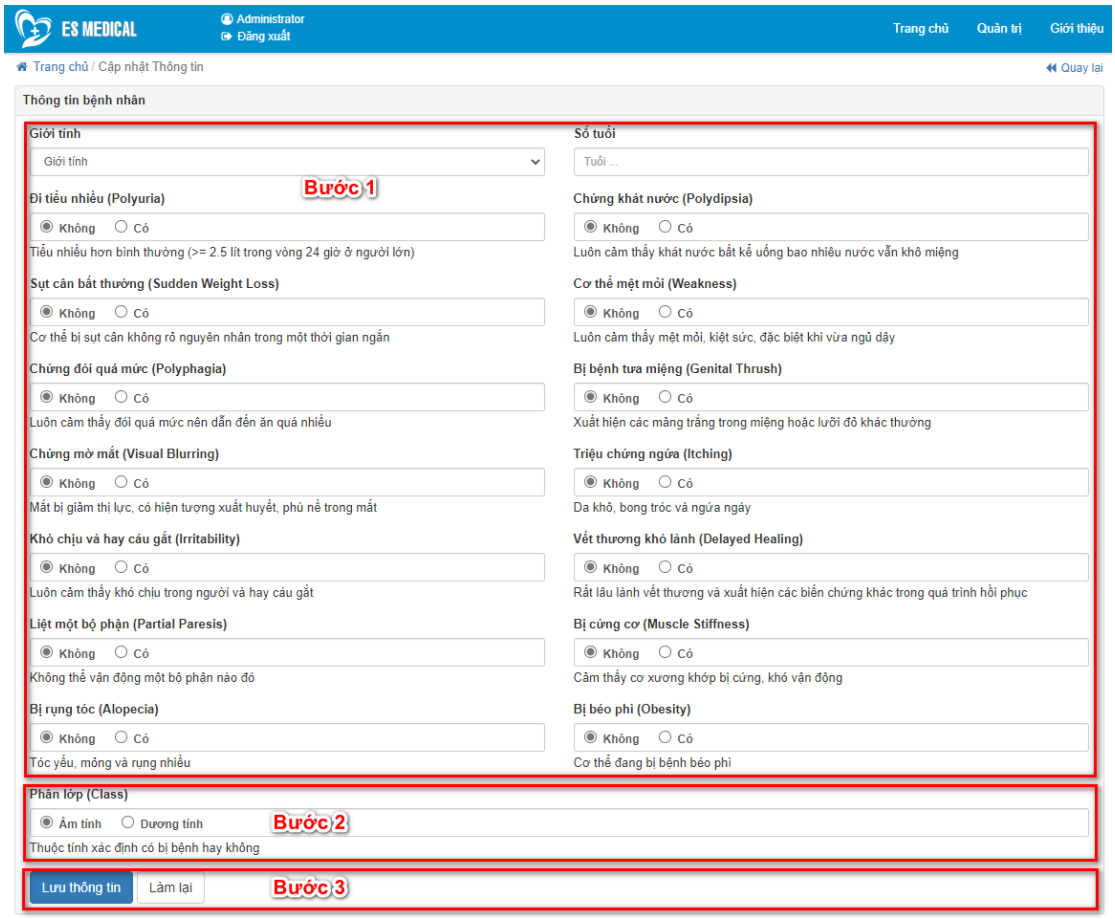
STT	Age	Gender	Polyu	Polyd	SWLoss	Wness	Polyp	GThrush	VBlurring	Itching	Irrita	DHealing	PParesis	MStiffness	Alopecia	Obesity	Class
1	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
2	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
3	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
4	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
5	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
6	55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Positive
7	57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Positive
8	66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Positive
9	67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Positive
10	70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	Positive
11	44	Male	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No	Positive
12	38	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Positive
13	35	Male	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	Positive
14	61	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Positive
15	60	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	Positive
16	58	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No	Positive
17	54	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	Yes	No	No	Positive
18	67	Male	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Positive
19	66	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Positive
20	43	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No	No	No	Positive

Hình 0.6: Giao diện trang quản trị dữ liệu

❖ ***Thêm mới Triệu chứng***

Để thêm mới triệu chứng, ta kích vào liên kết Thêm ở mục 1 bên trên, website sẽ hiển thị Trang thêm mới như bên dưới (hình 3.7):

- ***Bước 1:*** Nhập và chọn thông tin như Tuổi, Giới tính, Triệu chứng.
- ***Bước 2:*** Phân loại cho dữ liệu là có bệnh hay không có bệnh.
- ***Bước 3:*** Bấm Lưu thông tin để lưu trữ, bấm Làm lại nếu muốn nhập tiếp.



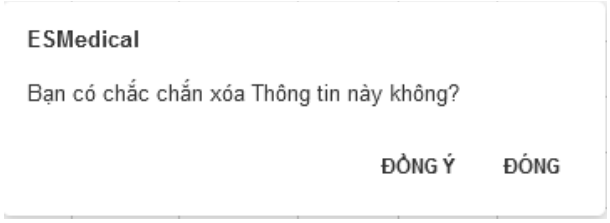
Hình 0.7: Giao diện trang thêm mới dữ liệu

❖ *Sửa đổi Triệu chứng*

Để sửa đổi triệu chứng, từ giao diện quản trị triệu chứng (hình 3.6) ta kích vào nút có hình cái Bút  của dòng dữ liệu muốn Sửa đổi để mở Trang thay đổi dữ liệu, giao diện Trang sửa đổi tương tự như Trang thêm mới (hình 3.7). Thực hiện 3 bước như mục Thêm mới để sửa đổi lại thông tin triệu chứng.

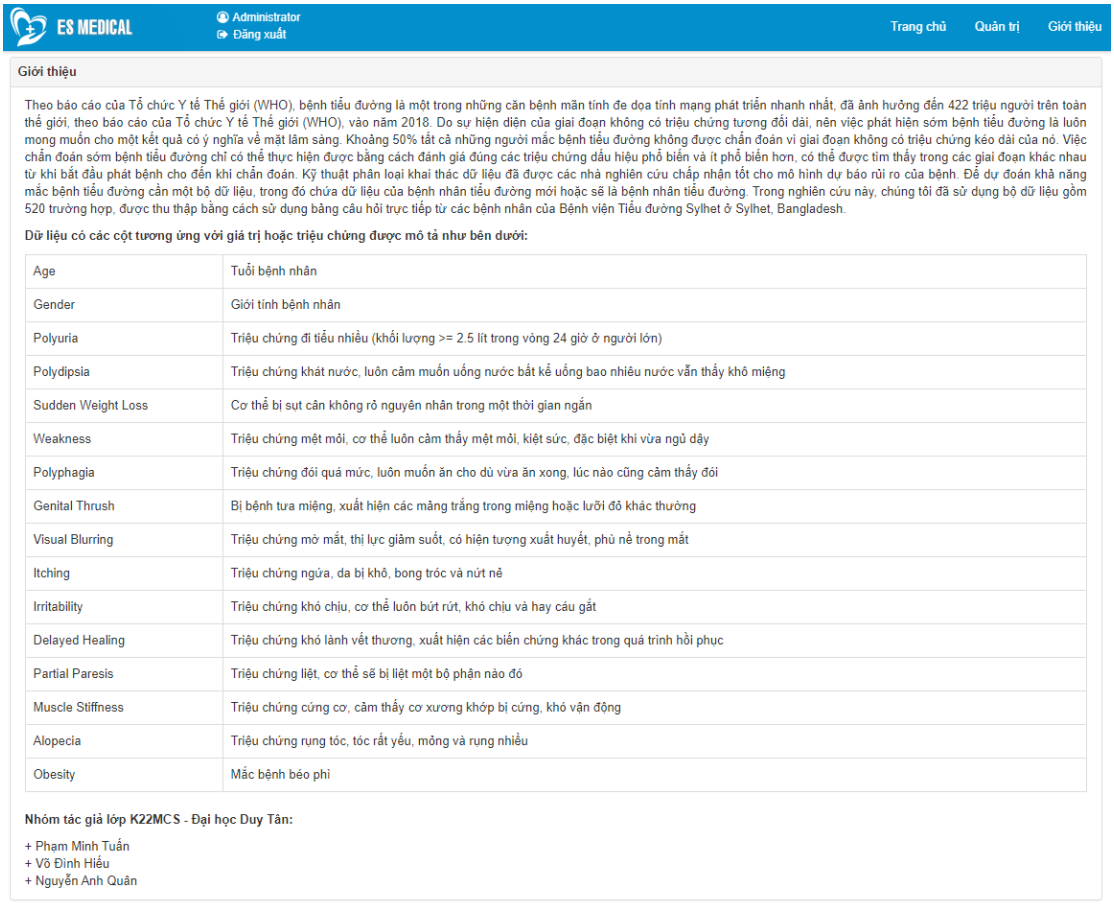
❖ *Xóa Triệu chứng*

Để xóa triệu chứng, từ giao diện quản trị triệu chứng (hình 3.6) ta kích vào nút có hình cái Sọt rác  của dòng dữ liệu muốn xóa. Sau khi kích vào Website sẽ hiển thị thông báo như hình bên dưới, nếu chọn Đồng ý sẽ xóa dòng đang chọn.



3.2.4. Tính năng Giới thiệu

Khi truy cập vào trang giới thiệu, ứng dụng sẽ giới thiệu khái quát về mức độ nguy hiểm của Bệnh tiểu đường, nguồn gốc, thông tin và diễn giải các trường trong cơ sở dữ liệu và thông tin về nhóm tác giả phát triển (hình 3.8).



Hình 0.8: Giao diện trang giới thiệu

Chương 4
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bệnh tiểu đường rất nguy hiểm nên việc có một ứng dụng để phổ cập và hỗ trợ chuẩn đoán là vô cùng thiết thực. Việc thực hiện đề tài mang ý nghĩa nhằm giúp cho người sử dụng hiểu biết hơn về bệnh tiểu đường, tiết kiệm được nhiều thời gian cũng như tự đánh giá được tình trạng sức khỏe để đi khám và điều trị kịp thời.

Thông qua quá trình nghiên cứu về mô hình cây quyết định và kiến thức môn học Hệ chuyên gia, tiểu luận đã tiến hành giải quyết bài toán thực tế về hỗ trợ chuẩn đoán bệnh tiểu đường. Cụ thể, tiểu luận đã đi sâu nghiên cứu và làm rõ những nội dung sau:

Tìm hiểu được kiến tức về môn Hệ chuyên gia, cấu trúc, cách xây dựng và sự cần thiết của Hệ chuyên gia trong đời sống.

Tìm hiểu về thuật toán C4.5 để ứng dụng vào việc phân tích dữ liệu được lấy tại website https://archive.ics.uci.edu/ml/index.php. Tập luật sinh ra từ dữ liệu được ứng dụng để chuẩn đoán bệnh tiểu đường.

Phát triển được ứng dụng website quản lý thông tin triệu chứng và hỗ trợ chuẩn đoán bệnh tiểu đường dựa trên tập luật sinh ra từ thuật toán C4.5.

Tiểu luận đã cho thấy sự hữu ích của việc phân tích và khai phá dữ liệu, giải quyết các bài toán thực tế có các đặc tính được lặp đi lặp lại ví dụ như triệu chứng về bệnh tiểu đường. Tuy nhiên, do một số nguyên nhân khách quan và chủ quan, tiểu luận vẫn còn tồn tại một số hạn chế sau:

Dữ liệu thu thập được còn ít nên công tác dự báo mới chỉ dừng lại ở phạm vi hỗ trợ, nhiều trường hợp còn sai số.

Chưa tìm hiểu được các giải thuật khác để hỗ trợ và khắc phục những điểm yếu cho giải thuật cây quyết định trong việc phân tích dữ liệu nhằm đưa ra kết quả phân tích và chuẩn đoán được chính xác hơn.

Để khắc phục những hạn chế nêu trên, trong thời gian tới, hướng nghiên cứu sẽ tiếp tục mở rộng phạm vi thu thập dữ liệu, nghiên cứu sâu hơn về các thuật toán và các công cụ hỗ trợ khác để tiến hành nâng cấp ứng dụng và triển khai rộng rãi hơn.

TÀI LIỆU THAM KHẢO

1) Giáo trình HỆ CHUYÊN GIA – PGS.TS Hoàng Văn Dũng
2) Damanik, I. S., Windarto, A. P., Wanto, A., Andani, S. R., & Saputra, W. (2019). Decision Tree Optimization in C4. 5 Algorithm Using Genetic Algorithm. Paper presented at the Journal of Physics: Conference Series.
3) Lakshmi, B., Indumathi, T., & Ravi, N. (2016). A Study on C. 5 decision tree classification algorithm for risk predictions during pregnancy. Procedia Technology, 24, 1542-1549.
4) Tayefi, M., Tajfard, M., Saffar, S., Hanachi, P., Amirabadizadeh, A. R., Esmaeily, H, Ghayour-Mobarhan, M. (2017). hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree
5) Trương, T. Q. (2018). Hướng xây dựng cây quyết định với chi phí hiệu quả. Trường Đại học Bách khoa-Đại học Đà Nẵng,
6) Hoan, Nguyen Quang, et al. "MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION, PREDICTION." UTEHY Journal of Science and Technology 17 (2018): 62-66.
7) https://archive.ics.uci.edu/ml/index.php
8) Joesph C. Giarratano and Gary D. Riley (2005), "Expert Systems: Principles and Programming", Fourth Edition, Thomson Course Technology.
9) Peter Jackson (1999), Introduction to Expert Systems, Third Edition, Addison-Wesley Longman, Harlow, England.