

TRƯỜNG ĐẠI HỌC DUY TÂN
KHOA SAU ĐẠI HỌC

Tiểu luận môn

HỆ HỖ TRỢ RA QUYẾT ĐỊNH

HỆ THỐNG HỖ TRỢ CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

Hướng dẫn : PGS.TS Lê Đắc Nhường

Thực hiện : Phạm Minh Tuấn

Võ Đình Hiếu

Nguyễn Anh Quân

Lớp : K22MCS (Khoa học máy tính)

Đà nẵng, 01/2021

MỤC LỤC

LỜI MỞ ĐẦU	1
CHƯƠNG 1: CÂY QUYẾT ĐỊNH VÀ THUẬT TOÁN C4.5	3
1.1. KHÁI NIỆM CÂY QUYẾT ĐỊNH	3
1.2. CÁC LOẠI CÂY QUYẾT ĐỊNH	3
1.3. TẠO CÂY QUYẾT ĐỊNH	4
1.4. SỬ DỤNG CÂY QUYẾT ĐỊNH	6
1.5. DUYỆT CÂY VÀ PHÂN LỚP DỮ LIỆU	7
1.5.1. Lựa chọn tiêu chuẩn phân lớp	7
1.5.2. Điều kiện để dừng việc phân chia	7
1.5.3. Độ lợi thông tin (Information Gain)	7
1.5.4. Vấn đề quá khớp trong phân lớp dữ liệu	10
1.5.5. Cây quyết định với cơ sở dữ liệu lớn	11
1.6. THUẬT TOÁN DUYỆT CÂY C4.5	11
CHƯƠNG 2: DỮ LIỆU BỆNH TIỂU ĐƯỜNG VÀ THUẬT TOÁN C4.5	14
2.1. GIỚI THIỆU BỆNH TIỂU ĐƯỜNG	14
2.2. THÔNG TIN DỮ LIỆU	14
2.3. MÔ TẢ THUỘC TÍNH	15
2.4. CÀI ĐẶT THUẬT TOÁN TRÊN TẬP DỮ LIỆU	16
2.5. DANH SÁCH LUẬT	37
CHƯƠNG 3: ỨNG DỤNG THỰC TẾ	42
3.1. GIỚI THIỆU ỨNG DỤNG	42
3.2. TRUY CẬP ỨNG DỤNG	42
3.3. GIAO DIỆN VÀ TÍNH NĂNG	42
Chương 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	47
TÀI LIỆU THAM KHẢO	48

DANH MỤC HÌNH VẼ

Ký hiệu	Nội dung	Trang
Hình 1.1	Cấu trúc cây quyết định	3
Hình 1.2	Cây quyết định dữ liệu mua máy tính	6
Hình 2.1	Cây quyết định tại thuộc tính Polyuria	28
Hình 2.2	Cây quyết định của một phần dữ liệu mẫu	37
Hình 2.3	Cây quyết định dữ liệu bệnh tiểu đường	41
Hình 3.1	Cách truy cập ứng dụng Hỗ trợ chuẩn đoán bệnh tiểu đường	42
Hình 3.2	Giao diện trang chủ website	43
Hình 3.3	Giao diện trang kết quả nếu không mắc bệnh	44
Hình 3.4	Giao diện trang kết quả người dùng mắc bệnh	45
Hình 3.5	Giao diện trang giới thiệu	46

DANH MỤC BẢNG

Ký hiệu	Nội dung	Trang
Bảng 1.1	Dữ liệu mua máy tính	5
Bảng 2.1	Các thuộc tính và giá trị của dữ liệu bệnh tiểu đường	16
Bảng 2.2	Dữ liệu mẫu bệnh tiểu đường	17

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 mắc bệnh có thể tăng lên 6,3 triệu vào năm 2045. Với những số liệu nói trên, Việt Nam được xếp nằm 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 nguy hiểm.

2. Mục tiêu, phương pháp nghiên cứu

Nhận thấy được mức độ nguy hiểm của bệnh tiểu đường nhưng để phổ cập một cách rộng rãi thì chúng ta cần một giải pháp hoàn thiện để có thể phổ biến kiến thức và hỗ trợ chuẩn đoán sớm những người có nguy cơ mắc bệnh.

Với những kiến thức có được khi tiếp cận môn học Hệ hỗ trợ ra quyết định, nhóm chúng em thấy khả năng xây dựng một hệ thống hỗ trợ chuẩn đoán bệnh tiểu đường để người dùng có thể tìm hiểu và tra cứu khả năng mắc bệnh là hoàn toàn khả thi. Vì vậy chúng em đã mạnh dạn thực hiện đề tài: “Hỗ trợ chuẩn đoán bệnh tiểu đường”. Đề tài có ý nghĩa thực tiễn cao 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.

Được sự hướng dẫn tận tình của PGS. TS. Lê Đắc Như, với kiến thức của nhóm thực hiện đã xây dựng khá hoàn thiện một trang website để có thể chuẩn đoán và đưa ra kết quả.

3. Bố cục tiểu luận

Nội dung của bài tiểu luận được trình bày với bố cục gồm 04 chương như sau:

Chương 1: Khái quát về cây quyết định, giới thiệu về giải thuật C4.5.

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

Chương 3: Phát triển website dựa trên tập luật và dữ liệu

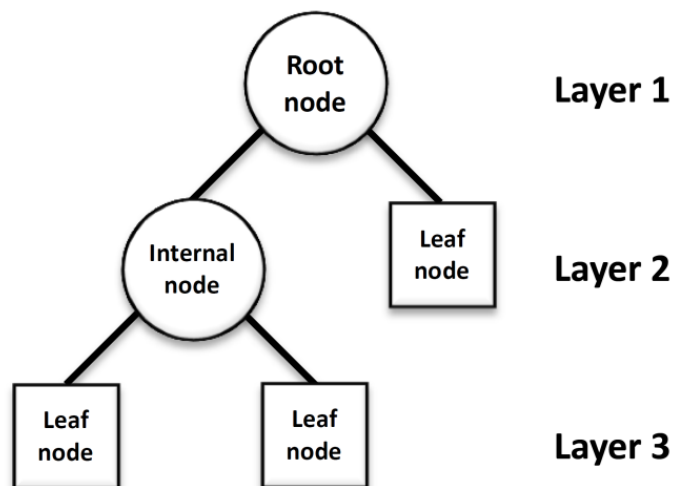
Chương 4: Kết luận: Trình bày kết quả đạt được của tiểu luận và tính thực tế cũng như định hướng phát triển trong tương lai.

Chương 1

CÂY QUYẾT ĐỊNH VÀ THUẬT TOÁN C4.5

1.1. KHÁI NIỆM 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 1.1: Cấu trúc cây quyết định

1.2. CÁC LOẠI CÂY QUYẾT ĐỊNH

Cây quyết định có 2 loại cơ bản sau đây:

Cây hồi quy (Regression tree): ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện).

Cây phân loại (Classification tree): nếu y là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

1.3. TẠO CÂY QUYẾT ĐỊNH

Có nhiều thuật toán khác nhau để xây dựng cây quyết định như: CLS, CART, ID3, C4.5, SLIQ, SPRINT, EC4.5, C5.0... Nhưng nói chung quá trình xây dựng cây quyết định đều được chia ra làm 3 giai đoạn cơ bản:

Tạo cây: Cây quyết định được tạo thành bằng cách lần lượt chia (theo phương pháp đệ quy) một tập dữ liệu thành các tập dữ liệu con, mỗi tập con được tạo thành từ các phần tử của cùng một lớp. Các nút (không phải là nút lá) là các điểm phân nhánh của cây. Việc phân nhánh tại các nút có thể dựa trên việc kiểm tra một hay nhiều thuộc tính để xác định việc phân chia dữ liệu.

Quá trình xây dựng một cây quyết định cụ thể bắt đầu bằng một nút rỗng bao gồm toàn bộ các đối tượng huấn luyện và làm như sau:

- 1) Nếu tại nút hiện thời, tất cả các đối tượng huấn luyện đều thuộc vào một lớp nào đó thì nút này chính là nút lá có tên là nhãn lớp chung của các đối tượng.
- 2) Trường hợp ngược lại, sử dụng một độ đo, chọn thuộc tính điều kiện phân chia tốt nhất tập mẫu huấn luyện có tại nút.
- 3) Tạo một lượng nút con của nút hiện thời bằng số các giá trị khác nhau của thuộc tính được chọn. Gán cho mỗi nhánh từ nút cha đến nút con một giá trị của thuộc tính rồi phân chia các đối tượng huấn luyện vào các nút con tương ứng.
- 4) Nút con K được gọi là thuần nhất, trở thành lá, nếu tất cả các đối tượng mẫu tại đó đều thuộc vào cùng một lớp.
- 5) Lặp lại các bước 1 - 3 đối với mỗi nút chưa thuần nhất.

Tỉa cây: Sau giai đoạn tạo cây chúng ta có thể dùng phương pháp “Độ dài mô tả ngắn nhất” (Minimum Description Length) hay giá trị tối thiểu của IG để tỉa cây (chúng ta có thể chọn giá trị tối thiểu của IG trong giai đoạn tạo cây đủ nhỏ để cho cây phát triển tương đối sâu, sau đó lại nâng giá trị này lên để tỉa cây).

Đánh giá cây: Dùng để đánh giá độ chính xác của cây kết quả. Tiêu chí đánh giá là tổng số mẫu được phân lớp chính xác trên tổng số mẫu đưa vào.

Việc tạo cây quyết định bao gồm 2 giai đoạn: Tạo cây và tỉa cây

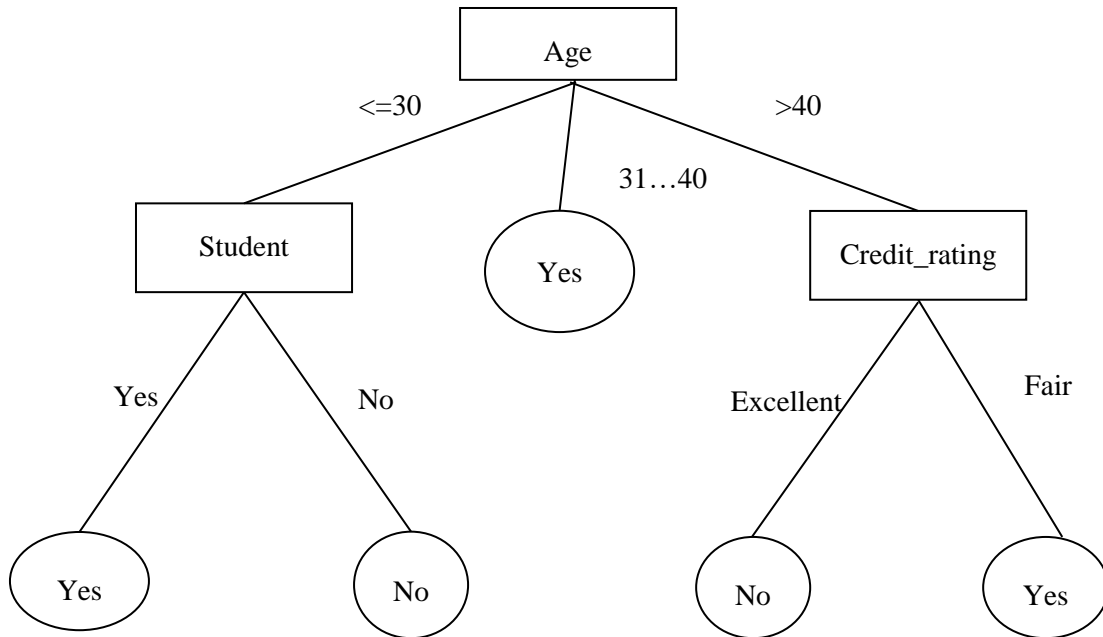
Ví dụ: Tạo cây quyết định theo bảng dữ liệu sau:

No	Age	Income	Student	Credit_ratingg	Buy_computer
1	<=30	High	No	Fair	No
2	<=30	High	No	Fair	No
3	31..40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31..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	31..40	Medium	No	Excellent	Yes
13	31..40	High	Yes	Fair	Yes

Bảng 1.1: Dữ liệu mua máy tính

Bảng dữ liệu này nhằm mô tả việc mua máy tính hay không dựa vào các thuộc tính tuổi (age), mức thu nhập (income), sinh viên (student), tỷ lệ tín dụng (credit_rating) và thuộc tính nhãn lớp mô tả việc mua máy tính hay không (Buy_Computer).

Cây quyết định thu được với bảng dữ liệu trên:



Hình 1.2: Cây quyết định dữ liệu mua máy tính

1.4. 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

Ví dụ: Nhìn cây quyết định ở Hình 1.2, ta suy ra được các luật tương ứng theo từng nút của cây như sau:

- 6) Nếu (Age<=30) và (Student=yes) Thì buy_computer=yes
- 7) Nếu (Age từ 31..40) Thì buy_computer=yes
- 8) Nếu (Age >40) và (Credit_rating=fair) Thì buy_computer=yes

1.5. DUYỆT CÂY VÀ PHÂN LỚP DỮ LIỆU

1.5.1. 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

1.5.2. 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

1.5.3. 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 $\{S_1, S_2, \dots, S_v\}$. Nếu S_i 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 S_i là:

$$E(A) = \sum_{i=1}^v \frac{P_i + n_i}{P + n} I(P_i, n_i) \quad (\text{công thức 1.2})$$

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

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

Ví dụ: Tính độ lợi thông tin theo bảng dữ liệu: *Bảng 1.1*

Thừa nhận:

- Lớp P: Buy_computer = “yes”
- Lớp N: Buy_computer = “no”
- Thông tin cần thiết để phân lớp một mẫu được cho là:

$$I(p, n) = I(9, 5) = -\frac{9}{9+5} \log_2 \frac{9}{9+5} - \frac{5}{9+5} \log_2 \frac{5}{9+5} = 0.940$$

Sau đó ta tính entropy (thông tin mong đợi cần để phân lớp những đối tượng trong tất cả các cây con) cho từng thuộc tính của bảng dữ liệu trên:

$$E(Age) = \frac{5}{14} I(2, 3) + \frac{4}{14} I(4, 0) + \frac{5}{14} I(3, 2) = \frac{5}{14} 0.971 + 0 + \frac{5}{14} 0.971 = 0.694$$

$$\text{Do đó: } Gain(Age) = I(9, 5) - E(Age) = 0.246$$

Tương tự:

$$Gain(Income) = 0.029$$

$$Gain(Student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

Như vậy, độ lợi thông tin của thuộc tính tuổi là lớn nhất, ta chọn thuộc tính này để phân lớp dữ liệu. Ta nhận thấy độ đo Gain có xu hướng chọn các thuộc tính có nhiều giá trị, 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, việc chọn thuộc tính không chỉ dựa vào độ đo Gain mà còn phụ thuộc vào độ đo Gain Ration:

$$\text{GainRatio}(A) = \text{Gain}(A) / \text{SplitInfo}(A) \quad (\text{công thức 1.4})$$

Trong đó công thức bên dưới là cách tính thông tin để phân nhánh trên cây quyết định:

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

(v : số giá trị của thuộc tính A).

Đây là công thức tính độ đo Gain Ratio cho thuộc tính A trên cơ sở dữ liệu D , sau đó ta chọn thuộc tính nào có độ đo Gain Ratio lớn nhất để phân lớp theo thuộc tính đó.

Lấy ví dụ trên, ta tính độ đo SplitInfo và Gain Ratio cho các thuộc tính:

$$\text{SplitInfo}(\text{Age}) = -\frac{5}{14} \log_2 \frac{5}{14} - \frac{4}{14} \log_2 \frac{4}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 1.576$$

Tương tự:

$$\text{SplitInfo}(\text{Income}) = 1.555$$

$$\text{SplitInfo}(\text{Student}) = 1$$

$$\text{SplitInfo}(\text{Credit-rating}) = 0.984$$

Do đó:

$$\text{Gain Ratio}(\text{Age}) = \text{Gain}(\text{Age}) / \text{SplitInfo}(\text{Age}) = 0.246 / 1.576 = 0.156$$

$$\text{Gain Ratio}(\text{Income}) = 0.019$$

$$\text{Gain Ratio}(\text{Student}) = 0.151$$

$$\text{Gain Ratio}(\text{Credit_rating}) = 0.049$$

Độ đo Gain Ratio của thuộc tính Age có giá trị lớn nhất nên ta chọn thuộc tính này để phân lớp dữ liệu.

1.5.4. Vấn đề quá khớp trong phân lớp dữ liệu

Cây được tạo ra có thể quá khớp với dữ liệu huấn luyện. Việc quá khớp có thể xảy ra trong những trường hợp sau đây:

Khi có độ nhiều của dữ liệu, một số có thể phản ánh sự dị thường vì những phần tử hỗn loạn hoặc những phần tử nằm ngoài phân lớp, dữ liệu huấn luyện quá ít tạo thành các giá trị tối đa cục bộ trong tìm kiếm tham lam, đôi lúc mỗi mẫu dữ liệu cho ta một khái niệm. Kết quả nhận được cây quá nhiều nhánh, thiếu chính xác đối với những mẫu chưa biết. Vì thế, trong thuật toán qui nạp chúng ta đã dùng các độ đo để chọn thuật tính phân lớp cây vừa đủ sâu và cho kết quả phân lớp tốt nhất. Tuy nhiên, ngay cả sử dụng các độ đo vẫn còn có quá nhiều các khái niệm nhất quán với không gian giả thuyết trên tập huấn luyện, có những trường hợp chỉ có vài mẫu dữ liệu cho một khái niệm, như vậy cây quyết định được kết hợp để phân lớp tất cả các trường hợp của tập huấn luyện một cách chính xác có thể thực hiện một cách nghèo nàn trên các mẫu mới mà đã không được sử dụng để xây dựng cây quyết định, ta nói đây là trường hợp quá khớp với dữ liệu huấn luyện bởi vì dữ liệu huấn luyện chỉ là một tập con của tất cả các mẫu trong kho dữ liệu. Vấn đề làm thế nào xây dựng được mô hình tốt nhất dự đoán cho các mẫu chưa biết.

Có hai cách tiếp cận để tránh quá khớp dữ liệu như sau:

Tỉa trước: cách tiếp cận này dùng để dừng sự tăng trưởng của cây sớm hơn trước khi nó tìm đến một điểm mà tại đó phân lớp hoàn hảo tập dữ liệu huấn luyện. Điều đó có nghĩa là không tiếp tục phân vùng một nút nếu điều này tạo kết quả ở dưới một ngưỡng theo một hệ đánh giá nhất định. Khi dừng lại một nút thì nút đó trở thành nút lá và nó có thể chứa hầu hết tần suất xuất hiện các lớp giữa các tập con của mẫu hoặc phân phối xác suất của toàn bộ mẫu. Khó khăn nhất trong việc tỉa trước là tạo ra một ngưỡng thích hợp để dừng việc phân chia tại một nút.

Tỉa sau: đây là cách tiếp cận phổ biến nhất, cho phép một cây tăng trưởng đầy đủ, sau đó ta mới tiến hành cắt tỉa bằng cách duyệt từ dưới lên. Tại mỗi nút trong của cây, ta tính tỉ lệ sai số kỳ vọng khi nó bị cắt bỏ và khi chưa cắt. Tỉ lệ sai số khi nó bị cắt được tính dựa vào hợp nhất các thể hiện ở các nhánh con của nó. Tỉ lệ sai số khi nó chưa bị cắt được tính theo tỉ lệ sai số ở mỗi nhánh kết hợp với trọng số của mỗi nhánh.

Nếu việc cắt bỏ một nút dẫn đến tỉ lệ sai số trông đợi lớn hơn thì nút đó được giữ lại, ngược lại thì cắt bỏ. Nút bị cắt bỏ sẽ trở thành nút lá và nhãn lớp được thay bằng hầu hết tần suất xuất hiện giữa các lớp trong các nhánh tạo thành nó.

1.5.5. Cây quyết định với cơ sở dữ liệu lớn

Sự phân lớp là một vấn đề cổ điển được nghiên cứu một cách mở rộng bởi những nhà thống kê và những nhà nghiên cứu máy học, chúng có tính co giãn vì vậy phân lớp các tập dữ liệu có hàng triệu mẫu và hàng trăm thuộc tính với tốc độ chấp nhận được.

Quy nạp cây quyết định được đánh giá cao trong khai phá dữ liệu lớn vì những nguyên nhân sau:

- Tốc độ học tương đối nhanh so với những phương pháp phân loại khác.
- Có thể hoán chuyển được thành những luật phân lớp đơn giản và dễ hiểu.
- Có thể sử dụng truy vấn SQL để truy xuất cơ sở dữ liệu.
- Sự chính xác phân lớp có thể so sánh được với những phương pháp khác.

Giải thuật cây quyết định là mô hình dạng cây. Không có giới hạn cho khối lượng dữ liệu đầu vào cũng như số lượng thuộc tính được gán vào giải thuật, khối lượng càng lớn cây càng lớn – càng rộng và càng sâu hơn.

1.6. THUẬT TOÁN DUYỆT CÂY 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 tuyến 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_vùngV các mẫu trong tập_mẫu_huấn_luyện có giá trị V tại thuộc tính

P; Gọi **induce_tree**(phân_vùngV, tập_thuộc_tính), gán 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 $\{S_1, S_2, \dots, S_v\}$. Nếu S_i chứa những pi mẫu của lớp P và n_i mẫu của N, entropy hay thông tin mong đợi cần để phân lớp những đối tượng trong tất cả các cây con S_i 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:

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

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

$$GainRatio(A) = Gain(A) / SplitInfo(A)$$

Chương 2

DỮ LIỆU 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. THÔNG TIN DỮ LIỆU

Dữ liệu được lấy từ website:

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

Dữ liệu được tổng hợp 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

Đặc điểm của tập dữ liệu: Đa biến

Số lượng bản ghi: 520

Số thuộc tính: 17

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

7	Polyphagia	Norminal	Yes, No	Bị bệnh tưa miệng, xuất hiện các mảng trắng trong miệng hoặc lưỡi đỏ khác thường
8	Genital Thrush	Norminal	Yes, No	Triệu chứng mờ mắt, thị lực giảm sút, có hiện tượng xuất huyết, phù nề trong mắt
9	Visual Blurring	Norminal	Yes, No	Triệu chứng ngứa, da bị khô, bong tróc và nứt nẻ
10	Itching	Norminal	Yes, No	Triệu chứng khó chịu, cơ thể luôn bứt rứt, khó chịu và hay cáu gắt
11	Irritability	Norminal	Yes, No	Triệu chứng khó lành vết thương, xuất hiện các biến chứng khác trong quá trình hồi phục
12	Delayed Healing	Norminal	Yes, No	Triệu chứng liệt, cơ thể sẽ bị liệt một bộ phận nào đó
13	Partial Paresis	Norminal	Yes, No	Triệu chứng cứng cơ, cảm thấy cơ xương khớp bị cứng, khó vận động
14	Muscle Stiffness	Norminal	Yes, No	Triệu chứng rụng tóc, tóc rất yếu, mỏng và rụng nhiều
15	Alopecia	Norminal	Yes, No	Mắc bệnh béo phì
16	Obesity	Norminal	Yes, No	Giới tính bệnh nhân
17	Class	Norminal	Positive, Negative	Triệu chứng đi tiểu nhiều (khối lượng ≥ 2.5 lít trong vòng 24 giờ ở người lớn)

Bảng 2.1: Các thuộc tính và giá trị của dữ liệu bệnh tiểu đường

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 2.2: Dữ liệu mẫu bệnh tiểu đường

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:

$$I(S) = -\frac{9}{21} \log_2 \left(\frac{9}{21} \right) - \frac{12}{21} \log_2 \left(\frac{12}{21} \right) = 1.05$$

❖ 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	2.9
2	No (10)	0	10	0

Ta có:

$$E(\text{Polyuria}) = \frac{11}{21} * I(d1, a1) + \frac{10}{21} * I(d2, a2) = \frac{11}{21} * 2.9 + \frac{10}{21} * 0 = 1.519$$

Trong đó:

$$I(d1, a1) = -\frac{9}{11} * \log_2 \frac{9}{11} - \frac{2}{11} * \log_2 \frac{2}{11} = 2.9$$

$$I(d2, a2) = -\frac{0}{10} * \log_2 \frac{0}{10} - \frac{10}{10} * \log_2 \frac{10}{10} = 0$$

Do đó:

$$\text{Gain}(\text{Polyuria}) = I(S) - E(\text{Polyuria}) = 1.05 - 1.519 = -0.469$$

Tính độ đo SplitInfo cho thuộc tính Polyuria:

$$\text{SplitInfo}(\text{Polyuria}) = -\frac{11}{21} \log_2 \frac{11}{21} - \frac{10}{21} \log_2 \frac{10}{21} = 1$$

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

$$\text{GainRatio}(\text{Polyuria}) = \text{Gain}(\text{Polyuria}) / \text{SplitInfo}(\text{Polyuria}) = -0.469 / 1 = -0.469$$

❖ Tính GainRatio cho thuộc tính Polydipsia:

$$I(S) = -\frac{9}{21} \log_2 \left(\frac{9}{21} \right) - \frac{12}{21} \log_2 \left(\frac{12}{21} \right) = 1.05$$

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	2.57

Ta có:

$$E(\text{Polydipsia}) = \frac{6}{21} * I(d1, a1) + \frac{15}{21} * I(d2, a2) = \frac{6}{21} * 0 + \frac{15}{21} * 2.57 = 1.83$$

Trong đó:

$$I(d1, a1) = -\frac{6}{6} * \log_2 \frac{6}{6} - \frac{0}{6} * \log_2 \frac{0}{6} = 0$$

$$I(d_2, a_2) = -\frac{3}{15} * \log_2 \frac{3}{15} - \frac{12}{15} * \log_2 \frac{12}{15} = 2.57$$

Do đó:

$$\text{Gain(Polydipsia)} = I(S) - E(\text{Polydipsia}) = 1.05 - 1.83 = -0.78$$

Tính độ đo SplitInfo cho thuộc tính Polydipsia:

$$\text{SplitInfo(Polydipsia)} = -\frac{6}{21} \log_2 \frac{6}{21} - \frac{15}{21} \log_2 \frac{15}{21} = 1.62$$

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

$$\begin{aligned} \text{GainRatio(Polydipsia)} &= \text{Gain(Polydipsia)} / \text{SplitInfo(Polydipsia)} \\ &= -0.78 / 1.62 = -0.66 \end{aligned}$$

❖ 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	0.973
2	No (16)	6	10	1.186

Ta có:

$$E(\text{SuddenWeightLoss}) = \frac{5}{21} * I(d_1, a_1) + \frac{16}{21} * I(d_2, a_2) = \frac{5}{21} * 0.973 + \frac{16}{21} * 1.186 = 1.135$$

$$\text{Gain(SuddenWeightLoss)} = 1.05 - 1.135 = -0.085$$

$$\text{SplitInfo(SuddenWeightLoss)} = -\frac{5}{21} \log_2 \frac{5}{21} - \frac{16}{21} \log_2 \frac{16}{21} = 2.057$$

$$\text{GainRatio(SuddenWeightLoss)} = -0.085 / 2.057 = -0.041$$

❖ 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	0.435
2	No (16)	7	9	1.044

Ta có:

$$E(\text{Weakness}) = \frac{5}{21} * I(d_1, a_1) + \frac{16}{21} * I(d_2, a_2) = \frac{5}{21} * 0.435 + \frac{16}{21} * 1.044 = 0.899$$

$$\text{Gain}(\text{Weakness}) = 1.05 - 1.135 = 0.15$$

$$\text{SplitInfo}(\text{Weakness}) = -\frac{5}{21} \log_2 \frac{5}{21} - \frac{16}{21} \log_2 \frac{16}{21} = 2.057$$

$$\text{GainRatio}(\text{Weakness}) = 0.15/2.057 = 0.073$$

❖ 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	0.865
2	No (14)	5	9	1.248

Ta có:

$$E(\text{Polyphagia}) = \frac{7}{21} * I(d_1, a_1) + \frac{14}{21} * I(d_2, a_2) = \frac{7}{21} * 0.865 + \frac{14}{21} * 1.248 = 1.121$$

$$\text{Gain}(\text{Polyphagia}) = 1.05 - 1.121 = -0.071$$

$$\text{SplitInfo}(\text{Polyphagia}) = -\frac{7}{21} \log_2 \frac{7}{21} - \frac{14}{21} \log_2 \frac{14}{21} = 1.349$$

$$\text{GainRatio}(\text{Polyphagia}) = -0.071/1.349 = -0.052$$

❖ 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	0.489
2	No (14)	6	8	1.058

Ta có:

$$E(\text{GenitalThrush}) = \frac{7}{21} * I(d_1, a_1) + \frac{14}{21} * I(d_2, a_2) = \frac{7}{21} * 0.489 + \frac{14}{21} * 1.058 = 0.868$$

$$\text{Gain}(\text{GenitalThrush}) = 1.05 - 0.868 = 0.181$$

$$\text{SplitInfo}(\text{GenitalThrush}) = -\frac{7}{21} \log_2 \frac{7}{21} - \frac{14}{21} \log_2 \frac{14}{21} = 1.349$$

$$\text{GainRatio}(\text{GenitalThrush}) = 0.181/1.349 = 0.134$$

❖ 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	1.298
2	No (15)	5	10	1.349

Ta có:

$$E(\text{VisualBlurring}) = \frac{6}{21} * I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 1.298 + \frac{15}{21} * 1.349 = 1.335$$

$$\text{Gain}(\text{VisualBlurring}) = 1.05 - 1.335 = -0.285$$

$$\text{SplitInfo}(\text{VisualBlurring}) = -\frac{6}{21} \log_2 \frac{6}{21} - \frac{15}{21} \log_2 \frac{15}{21} = 1.629$$

$$\text{GainRatio}(\text{VisualBlurring}) = -0.285/1.629 = -0.175$$

❖ 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	0.65
2	No (11)	3	8	1.728

Ta có:

$$E(\text{Itching}) = \frac{10}{21} * I(d_1, a_1) + \frac{11}{21} * I(d_2, a_2) = \frac{10}{21} * 0.65 + \frac{11}{21} * 1.728 = 1.215$$

$$\text{Gain}(\text{Itching}) = 1.05 - 1.215 = -0.165$$

$$\text{SplitInfo}(\text{Itching}) = -\frac{10}{21} \log_2 \frac{10}{21} - \frac{11}{21} \log_2 \frac{11}{21} = 1$$

$$\text{GainRatio}(\text{Itching}) = -0.165/1 = -0.165$$

❖ 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	1.349

Ta có:

$$E(\text{Irritability}) = \frac{3}{21} * I(d_1, a_1) + \frac{18}{21} * I(d_2, a_2) = \frac{3}{21} * 0 + \frac{18}{21} * 1.349 = 1.157$$

$$\text{Gain}(\text{Irritability}) = 1.05 - 1.157 = -0.107$$

$$\text{SplitInfo}(\text{Irritability}) = -\frac{3}{21} \log_2 \frac{3}{21} - \frac{18}{21} \log_2 \frac{18}{21} = 3.9$$

$$\text{GainRatio}(\text{Irritability}) = -0.107/3.9 = -0.027$$

❖ 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	0.65
2	No (15)	6	9	1.116

Ta có:

$$E(\text{DelayedHealing}) = \frac{6}{21} * I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 0.65 + \frac{15}{21} * 1.116 = 0.983$$

$$\text{Gain}(\text{DelayedHealing}) = 1.05 - 0.983 = 0.066$$

$$\text{SplitInfo}(\text{DelayedHealing}) = -\frac{6}{21} \log_2 \frac{6}{21} - \frac{15}{21} \log_2 \frac{15}{21} = 1.629$$

$$\text{GainRatio}(\text{DelayedHealing}) = 0.066/1.629 = 0.04$$

❖ 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	3.297
2	No (15)	4	11	1.778

Ta có:

$$E(\text{PartialParesis}) = \frac{6}{21} * I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 3.297 + \frac{15}{21} * 1.778 = 2.212$$

$$\text{Gain}(\text{PartialParesis}) = 1.05 - 2.212 = -1.162$$

$$\text{SplitInfo}(\text{PartialParesis}) = -\frac{6}{21} \log_2 \frac{6}{21} - \frac{15}{21} \log_2 \frac{15}{21} = 1.629$$

$$\text{GainRatio}(\text{PartialParesis}) = -1.162/1.629 = -0.713$$

❖ 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	0.973
2	No (16)	6	10	1.186

Ta có:

$$E(\text{MuscleStiffness}) = \frac{5}{21} * I(d_1, a_1) + \frac{16}{21} * I(d_2, a_2) = \frac{5}{21} * 0.973 + \frac{16}{21} * 1.186 = 1.135$$

$$\text{Gain}(\text{MuscleStiffness}) = 1.05 - 1.135 = -0.085$$

$$\text{SplitInfo}(\text{MuscleStiffness}) = -\frac{5}{21} \log_2 \frac{5}{21} - \frac{16}{21} \log_2 \frac{16}{21} = 2.057$$

$$\text{GainRatio}(\text{MuscleStiffness}) = -0.085/2.057 = -0.041$$

❖ 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	0.13
2	No (15)	8	7	1.012

Ta có:

$$E(\text{Alopecia}) = \frac{6}{21} * I(d_1, a_1) + \frac{15}{21} * I(d_2, a_2) = \frac{6}{21} * 0.13 + \frac{15}{21} * 1.012 = 0.76$$

$$\text{Gain}(\text{Alopecia}) = 1.05 - 0.76 = 0.289$$

$$\text{SplitInfo}(\text{Alopecia}) = -\frac{6}{21} \log_2 \frac{6}{21} - \frac{15}{21} \log_2 \frac{15}{21} = 1.629$$

$$\text{GainRatio}(\text{Alopecia}) = 0.289/1.629 = 0.177$$

❖ 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	0.65
2	No (17)	7	10	1.09

Ta có:

$$E(\text{Obesity}) = \frac{4}{21} * I(d_1, a_1) + \frac{17}{21} * I(d_2, a_2) = \frac{4}{21} * 0.65 + \frac{17}{21} * 1.09 = 1$$

$$\text{Gain}(\text{Obesity}) = 1.05 - 1 = 0.05$$

$$\text{SplitInfo}(\text{Obesity}) = -\frac{4}{21} \log_2 \frac{4}{21} - \frac{17}{21} \log_2 \frac{17}{21} = 2.735$$

$$\text{GainRatio}(\text{Obesity}) = 0.05/2.735 = 0.015$$

❖ 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	0.489
2	Female (0)	0	0	0

Ta có:

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

$$\text{Gain}(\text{Gender}) = 0.985 - 0.985 = 0$$

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

$$\text{GainRatio}(\text{Gender}) = 0$$

❖ 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:

$$\begin{aligned}
 E(\text{Age}) &= \frac{2}{21} * I(d_1, a_1) + \frac{2}{21} * I(d_2, a_2) + \frac{1}{21} * I(d_3, a_3) + \frac{1}{21} * I(d_4, a_4) + \frac{1}{21} * I(d_5, a_5) + \frac{1}{21} * \\
 &I(d_6, a_6) + \frac{2}{21} * I(d_7, a_7) + \frac{2}{21} * I(d_8, a_8) + \frac{1}{21} * I(d_9, a_9) + \frac{1}{21} * I(d_{10}, a_{10}) + \frac{1}{21} * I(d_{11}, a_{11}) + \frac{1}{21} * \\
 &I(d_{12}, a_{12}) + \frac{1}{21} * I(d_{13}, a_{13}) + \frac{1}{21} * I(d_{14}, a_{14}) + \frac{1}{21} * I(d_{15}, a_{15}) + \frac{1}{21} * I(d_{16}, a_{16}) + \frac{1}{21} * I(d_{17}, a_{17}) \\
 &= \frac{2}{21} * 1 + \frac{2}{21} * 1 + 0 + 0 + 0 + 0 + \frac{2}{21} * 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 0.286
 \end{aligned}$$

$$\text{Gain}(\text{Age}) = 0.985 - 0.286 = 0.699$$

$$\begin{aligned} \text{SplitInfo(Age)} &= -\frac{2}{21}\log_2\frac{2}{21} - \frac{2}{21}\log_2\frac{2}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \\ &\frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{2}{21}\log_2\frac{2}{21} - \frac{2}{21}\log_2\frac{2}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \\ &\frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} - \frac{1}{21}\log_2\frac{1}{21} = 4.011 \end{aligned}$$

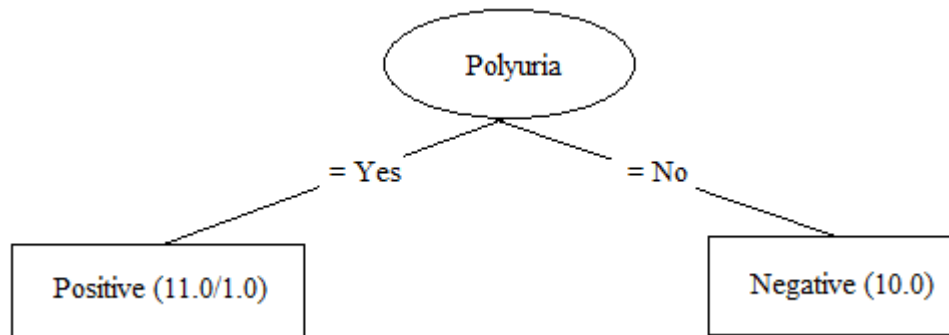
$$\text{GainRatio(Age)} = 0.699/4.011 = 0.174$$

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	0.233
4	Age	0.174
5	Irritability	0.113
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	0.006
14	Gender	0
15	Weakness	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

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:

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

$$\text{Gain}(\text{Gender}) = 0.985 - 0.985 = 0$$

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

$$\text{GainRatio}(\text{Gender}) = 0$$

❖ 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	0.970

Ta có:

$$E(\text{Polydipsia}) = \frac{6}{11} * I(d_1, a_1) + \frac{5}{11} * I(d_2, a_2) = \frac{6}{11} * 0 + \frac{5}{11} * 0.970 = 0.441$$

$$\text{Gain}(\text{Polydipsia}) = 0.684 - 0.516 = 0.168$$

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

$$\text{GainRatio}(\text{Polydipsia}) = 0.684/0.944 = 0.258$$

❖ 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:

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

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

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

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

❖ 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:

$$E(\text{Weakness}) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = \frac{3}{11} * 0.918 + \frac{8}{11} * 0.544 = 0.597$$

$$\text{Gain}(\text{Weakness}) = 0.684 - 0.597 = 0.087$$

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

$$\text{GainRatio}(\text{Weakness}) = 0.087/0.845 = 0.103$$

❖ 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	0.65
---	--------	---	---	------

Ta có:

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

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

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

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

❖ 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 (7)	6	1	0.592
2	No (4)	3	1	0.811

Ta có:

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

$$\text{Gain}(\text{GenitalThrush}) = 0.012$$

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

$$\text{GainRatio}(\text{GenitalThrush}) = 0.013$$

❖ 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	0.65
---	--------	---	---	------

Ta có:

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

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

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

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

❖ 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:

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

$$\text{Gain}(\text{Itching}) = 0.135$$

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

$$\text{GainRatio}(\text{Itching}) = 0.142$$

❖ 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:

$$E(\text{Irritability}) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = 0.59$$

$$\text{Gain}(\text{Irritability}) = 0.094$$

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

$$\text{GainRatio}(\text{Irritability}) = 0.111$$

❖ 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:

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

$$\text{Gain}(\text{DelayedHealing}) = 0.243$$

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

$$\text{GainRatio}(\text{DelayedHealing}) = 0.244$$

❖ 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	0.65

2	No (5)	4	1	0.722
---	--------	---	---	-------

Ta có:

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

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

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

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

❖ 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:

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

$$\text{Gain}(\text{MuscleStiffness}) = 0.012$$

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

$$\text{GainRatio}(\text{MuscleStiffness}) = 0.013$$

❖ 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:

$$E(\text{Alopecia}) = \frac{3}{11} * I(d_1, a_1) + \frac{8}{11} * I(d_2, a_2) = 0.25$$

$$\text{Gain}(\text{Alopecia}) = 0.434$$

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

$$\text{GainRatio}(\text{Alopecia}) = 0.513$$

❖ 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:

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

$$\text{Gain}(\text{Obesity}) = 0.038$$

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

$$\text{GainRatio}(\text{Obesity}) = 0.045$$

❖ 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:

$$\begin{aligned}
 E(\text{Age}) = & \frac{1}{11} * I(d_1, a_1) + \frac{1}{11} * I(d_2, a_2) + \frac{1}{11} * I(d_3, a_3) + \frac{1}{11} * I(d_4, a_4) + \\
 & \frac{1}{11} * I(d_5, a_5) + \frac{1}{11} * I(d_6, a_6) + \frac{1}{11} * I(d_7, a_7) + \\
 & \frac{1}{11} * I(d_8, a_8) + \frac{1}{11} * I(d_9, a_9) + \frac{1}{11} * I(d_{10}, a_{10}) + \frac{1}{11} * I(d_{11}, a_{11}) = 0
 \end{aligned}$$

$$\text{Gain}(\text{Age}) = 0.684$$

$$\text{SplitInfo}(\text{Age}) = 3.459$$

$$\text{GainRatio}(\text{Age}) = 0.198$$

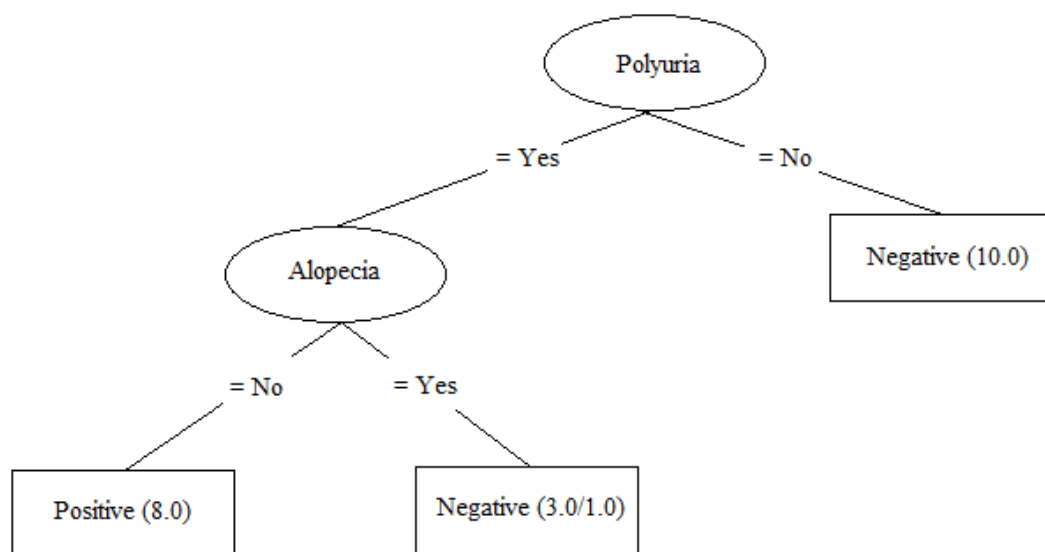
Độ đo GainRatio của các thuộc tính được sắp xếp giảm dần

STT	Thuộc tính	GainRatio
1	Alopecia	0.513
2	Polydipsia	0.258
3	DelayedHealing	0.244
4	Age	0.198
5	Itching	0.143

6	Irritability	0.111
7	Weakness	0.103
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 của một phần 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 22 luật sau huấn luyện như sau:

Polyuria = No

| Polydipsia = Yes

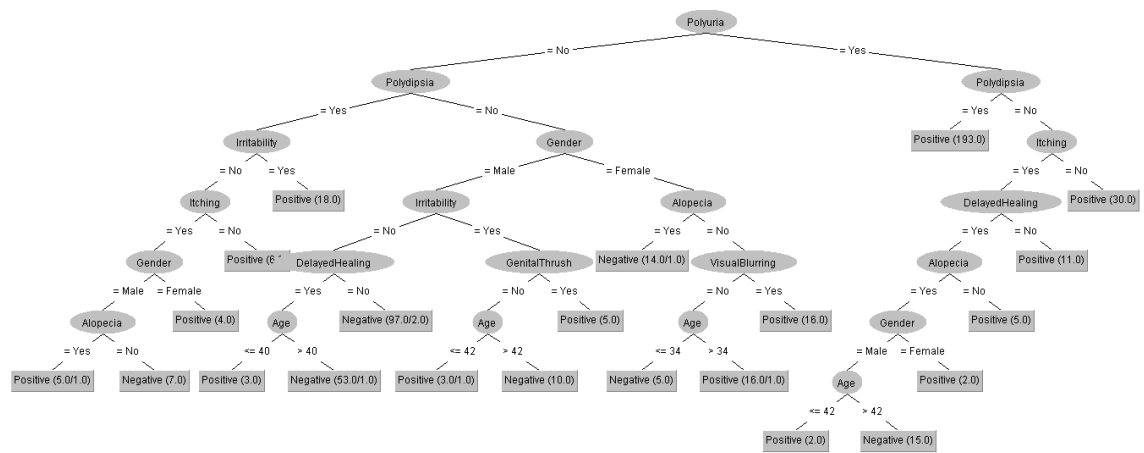
| | Irritability = No
| | | Itching = Yes
| | | | Gender = Male
| | | | | Alopecia = Yes: Positive
| | | | | Alopecia = No: Negative
| | | | Gender = Female: Positive
| | | Itching = No: Positive
| | Irritability = Yes: Positive
| Polydipsia = No
| | Gender = Male
| | | Irritability = No
| | | | DelayedHealing = Yes
| | | | | Age <= 40: Positive
| | | | | Age > 40: Negative
| | | | DelayedHealing = No: Negative
| | | Irritability = Yes
| | | | GenitalThrush = No
| | | | | Age <= 42: Positive
| | | | | Age > 42: Negative
| | | | GenitalThrush = Yes: Positive
| | Gender = Female
| | | Alopecia = Yes: Negative
| | | Alopecia = No
| | | | VisualBlurring = No
| | | | | Age <= 34: Negative
| | | | | Age > 34: Positive
| | | | VisualBlurring = Yes: Positive
Polyuria = Yes
| Polydipsia = Yes: Positive
| Polydipsia = No
| | Itching = Yes

| | | DelayedHealing = Yes
 | | | | Alopecia = Yes
 | | | | | Gender = Male
 | | | | | | Age \leq 42: Positive
 | | | | | | Age $>$ 42: Negative
 | | | | | Gender = Female: Positive
 | | | | Alopecia = No: Positive
 | | | DelayedHealing = No: Positive
 | | Itching = No: Positive

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

1. Nếu bệnh nhân không tiểu nhiều, có khát nhiều, không khó chịu, có ngứa, giới tính là nam, có rụng tóc thì mắc bệnh tiểu đường.
2. Nếu bệnh nhân không tiểu nhiều, có khát nhiều, không khó chịu, có ngứa, giới tính là nam, không rụng tóc thì không mắc bệnh tiểu đường.
3. Nếu bệnh nhân không tiểu nhiều, có khát nhiều, không khó chịu, có ngứa, giới tính là nữ thì mắc bệnh tiểu đường.
4. Nếu bệnh nhân không tiểu nhiều, có khát nhiều, không khó chịu, không ngứa thì mắc bệnh tiểu đường.
5. Nếu bệnh nhân không tiểu nhiều, có khát nhiều, có khó chịu thì mắc bệnh tiểu đường.
6. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nam, không khó chịu, có vết thương lâu lành, có tuổi \leq 40 thì mắc bệnh tiểu đường.
7. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nam, không khó chịu, có vết thương lâu lành, có tuổi $>$ 40 thì không mắc bệnh tiểu đường.
8. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nam, không khó chịu, không có vết thương lâu lành thì không mắc bệnh tiểu đường.
9. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nam, có khó chịu, không tưa miệng, có tuổi \leq 42 thì mắc bệnh tiểu đường.

10. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nam, có khó chịu, không tưa miệng, có tuổi >42 thì không mắc bệnh tiểu đường.
11. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nam, có khó chịu, có tưa miệng thì mắc bệnh tiểu đường.
12. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nữ, có rụng tóc thì không mắc bệnh tiểu đường.
13. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nữ, không rụng tóc, không mờ mắt, có tuổi ≤ 34 thì không mắc bệnh tiểu đường.
14. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nữ, không rụng tóc, không mờ mắt, có tuổi > 34 thì mắc bệnh tiểu đường.
15. Nếu bệnh nhân không tiểu nhiều, không khát nhiều, giới tính là nữ, không rụng tóc, có mờ mắt thì mắc bệnh tiểu đường.
16. Nếu bệnh nhân có tiểu nhiều, có khát nhiều thì mắc bệnh tiểu đường.
17. Nếu bệnh nhân có tiểu nhiều, không khát nhiều, có ngứa, có vết thương lâu lành, có rụng tóc, có giới tính là nam, có tuổi ≤ 42 thì mắc bệnh tiểu đường.
18. Nếu bệnh nhân có tiểu nhiều, không khát nhiều, có ngứa, có vết thương lâu lành, có rụng tóc, có giới tính là nam, có tuổi > 42 thì không mắc bệnh tiểu đường.
19. Nếu bệnh nhân có tiểu nhiều, không khát nhiều, có ngứa, có vết thương lâu lành, có rụng tóc, có giới tính là nữ thì mắc bệnh tiểu đường.
20. Nếu bệnh nhân có tiểu nhiều, không khát nhiều, có ngứa, có vết thương lâu lành, không rụng tóc thì mắc bệnh tiểu đường.
21. Nếu bệnh nhân có tiểu nhiều, không khát nhiều, có ngứa, không có vết thương lâu lành thì mắc bệnh tiểu đường.
22. Nếu bệnh nhân có tiểu nhiều, không khát nhiều, không ngứa thì mắc bệnh tiểu đường.



Hình 2.3: Cây quyết định dữ liệu bệnh tiểu đường

Chương 3

ỨNG DỤNG THỰC TẾ

3.1. GIỚI THIỆU ỨNG DỤNG

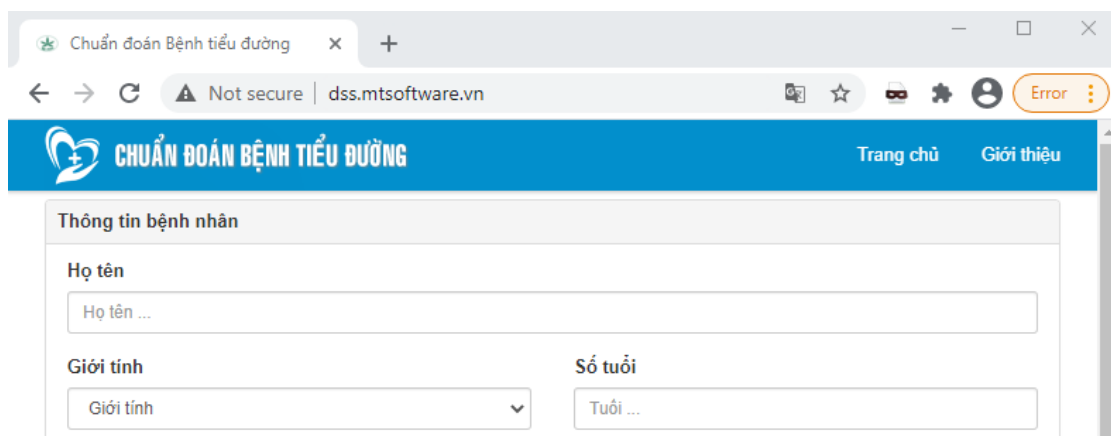
Chương trình “Hỗ trợ chuẩn đoán bệnh tiểu đường” được phát triển dạng website để tăng độ hiệu quả khi triển khai và dễ dàng nâng cấp ứng dụng khi có nhu cầu.

Website phát triển trên nền tảng Microsoft .Net Framework 4.0, sử dụng ngôn ngữ C#, JavaScript để lập trình. Về mặt giao diện sử dụng Bootstrap 3.7 và các thành phần phụ thuộc khác để triển khai ứng dụng.

3.2. 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://dss.mtsoftware.vn/>



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

3.3. GIAO DIỆN VÀ TÍNH NĂNG

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):


CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG
Trang chủ
Giới thiệu

Thông tin bệnh nhân

Họ tên

Giới tính

Số tuổi

Đi tiểu nhiều
☒ Không ☐ Có
 Tiểu nhiều hơn bình thường (≥ 2.5 lít trong vòng 24 giờ ở người lớn)

Chứng khát nước
☒ Không ☐ Có
 Luôn cảm thấy khát nước bất kể uống bao nhiêu nước vẫn khó miệng

Sụt cân bất thường
☒ Không ☐ Có
 Cơ thể bị sụt cân không rõ nguyên nhân trong một thời gian ngắn

Cơ thể mệt mỏi
☒ Không ☐ Có
 Luôn cảm thấy mệt mỏi, kiệt sức, đặc biệt khi vừa ngủ dậy

Chứng đói quá mức
☒ Không ☐ Có
 Luôn cảm thấy đói quá mức nên dẫn đến ăn quá nhiều

Bị bệnh tưa miệng
☒ Không ☐ Có
 Xuất hiện các mảng trắng trong miệng hoặc lưỡi đỏ khác thường

Chứng mờ mắt
☒ Không ☐ Có
 Mắt bị giảm thị lực, có hiện tượng xuất huyết, phù nề trong mắt

Triệu chứng ngứa
☒ Không ☐ Có
 Da khô, bong tróc và ngứa ngáy

Khó chịu và hay cáu gắt
☒ Không ☐ Có
 Luôn cảm thấy khó chịu trong người và hay cáu gắt

Vết thương khó lành
☒ Không ☐ Có
 Rất lâu lành vết thương và xuất hiện các biến chứng khác quá trình hồi phục

Liệt một bộ phận
☒ Không ☐ Có
 Không thể vận động một bộ phận nào đó

Bị cứng cơ
☒ Không ☐ Có
 Cảm thấy cơ xương khớp bị cứng, khó vận động

Bị rụng tóc
☒ Không ☐ Có
 Tóc yếu, mỏng và rụng nhiều

Bị béo phì
☒ Không ☐ Có
 Cơ thể đang bị bệnh béo phì

Nhóm tác giả lớp K22MCS - Đại học Duy Tân © 2021

Hình 3.2: Giao diện trang chủ website


Tại giao diện trang chủ người dùng cần thao tác theo thứ tự như sau:

Bước 1: Nhập Họ tên, Giới tính, Tuổi và chọn các triệu chứng

Bước 2: Bấm nút Kiểm tra để xem kết quả

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

Xin chào: Nguyễn Văn A, năm nay: 34 tuổi. Với các thông số bạn chọn và dựa vào dữ liệu tổng hợp của chúng tôi. Xin chúc mừng, tình trạng của bạn không liên quan đến Bệnh tiểu đường tuýp 1


CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG
Trang chủ
Giới thiệu

Thông tin bệnh nhân

Xin chào: Nguyễn Văn A, năm nay: 34 tuổi. Với các thông số bạn chọn và dựa vào dữ liệu tổng hợp của chúng tôi. Xin chúc mừng, tình trạng của bạn không liên quan đến Bệnh tiểu đường tuýp 1

Họ tên

Giới tính

Số tuổi

Đi tiểu nhiều
☒ Không ☐ Có
Tiểu nhiều hơn bình thường (≥ 2.5 lít trong vòng 24 giờ ở người lớn)

Chứng khát nước
☒ Không ☐ Có
Luôn cảm thấy khát nước bất kể uống bao nhiêu nước vẫn khô miệng

Sụt cân bất thường
☒ Không ☐ Có
Cơ thể bị sụt cân không rõ nguyên nhân trong một thời gian ngắn

Cơ thể mệt mỏi
☒ Không ☐ Có
Luôn cảm thấy mệt mỏi, kiệt sức, đặc biệt khi vừa ngủ dậy

Chứng đói quá mức
☒ Không ☐ Có
Luôn cảm thấy đói quá mức nên dẫn đến ăn quá nhiều

Bị bệnh tưa miệng
☒ Không ☐ Có
Xuất hiện các mảng trắng trong miệng hoặc lưỡi đỏ khác thường

Chứng mờ mắt
☒ Không ☐ Có
Mắt bị giảm thị lực, có hiện tượng xuất huyết, phù nề trong mắt

Triệu chứng ngứa
☒ Không ☐ Có
Da khô, bong tróc và ngứa ngáy

Khó chịu và hay cáu gắt
☒ Không ☐ Có
Luôn cảm thấy khó chịu trong người và hay cáu gắt

Vết thương khó lành
☒ Không ☐ Có
Rất lâu lành vết thương và xuất hiện các biến chứng khác trong quá trình hồi phục

Liệt một bộ phận
☒ Không ☐ Có
Không thể vận động một bộ phận nào đó

Bị cứng cơ
☒ Không ☐ Có
Cảm thấy cơ xương khớp bị cứng, khó vận động

Bị rụng tóc
☒ Không ☐ Có
Tóc yếu, mỏng và rụng nhiều


Bị béo phì
☒ Không ☐ Có
Cơ thể đang bị bệnh béo phì

Nhóm tác giả lớp K22MCS - Đại học Duy Tân © 2021

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):

Xin chào: Nguyễn Văn A, năm nay: 34 tuổi. Với các thông số bạn chọn và dựa vào dữ liệu tổng hợp của chúng tôi. Xin chia buồn, bạn đã Dương tính với Bệnh tiểu đường tuýp 1


CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

[Trang chủ](#)
[Giới thiệu](#)

Thông tin bệnh nhân

Xin chào: Nguyễn Văn A, năm nay: 34 tuổi. Với các thông số bạn chọn và dựa vào dữ liệu tổng hợp của chúng tôi. Xin chia buồn, bạn đã **Dương tính** với Bệnh tiểu đường tuýp 1

Họ tên

Giới tính

Số tuổi

Đi tiểu nhiều
☐ Không ☒ Có
Tiểu nhiều hơn bình thường (≥ 2.5 lít trong vòng 24 giờ ở người lớn)

Chứng khát nước
☐ Không ☒ Có
Luôn cảm thấy khát nước bất kể uống bao nhiêu nước vẫn khô miệng

Sụt cân bất thường
☐ Không ☒ Có
Cơ thể bị sụt cân không rõ nguyên nhân trong một thời gian ngắn

Cơ thể mệt mỏi
☐ Không ☒ Có
Luôn cảm thấy mệt mỏi, kiệt sức, đặc biệt khi vừa ngủ dậy

Chứng đói quá mức
☐ Không ☒ Có
Luôn cảm thấy đói quá mức nên dẫn đến ăn quá nhiều

Bị bệnh tưa miệng
☐ Không ☒ Có
Xuất hiện các mảng trắng trong miệng hoặc lưỡi đỏ khác thường

Chứng mờ mắt
☐ Không ☒ Có
Mắt bị giảm thị lực, có hiện tượng xuất huyết, phù nề trong mắt

Triệu chứng ngứa
☐ Không ☒ Có
Da khô, bong tróc và ngứa ngáy

Khó chịu và hay cáu gắt
☐ Không ☒ Có
Luôn cảm thấy khó chịu trong người và hay cáu gắt

Vết thương khó lành
☐ Không ☒ Có
Rất lâu lành vết thương và xuất hiện các biến chứng khác trong quá trình hồi phục

Liệt một bộ phận
☐ Không ☒ Có
Không thể vận động một bộ phận nào đó

Bị cứng cơ
☐ Không ☒ Có
Cảm thấy cơ xương khớp bị cứng, khó vận động

Bị rụng tóc
☐ Không ☒ Có
Tóc yếu, mỏng và rụng nhiều


Bị béo phì
☐ Không ☒ Có
Cơ thể đang bị bệnh béo phì

Nhóm tác giả lớp K22MCS - Đại học Duy Tân © 2021

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

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 thực hiện tiểu luận.

Sau khi truy cập sẽ hiển thị giống hình vẽ bên dưới (hình 3.5):



CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

Trang chủGiới thiệu

Giới thiệu

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, chúng tôi đã sử dụng bộ dữ liệu gồm 520 trường hợp, được thu thập bằng cách sử dụng bảng câu hỏi trực tiếp từ các bệnh nhân của Bệnh viện Tiểu đường Sylhet ở Sylhet, Bangladesh.

Dữ liệu có các cột tương ứng với giá trị hoặc triệu chứng được mô tả như bên dưới:

Age	Tuổi bệnh nhân
Gender	Giới tính bệnh nhân
Polyuria	Triệu chứng đi tiểu nhiều (khối lượng >= 2.5 lít trong vòng 24 giờ ở người lớn)
Polydipsia	Triệu chứng khát nước, luôn cảm muốn uống nước bất kể uống bao nhiêu nước vẫn thấy khô miệng
Sudden Weight Loss	Cơ thể bị sụt cân không rõ nguyên nhân trong một thời gian ngắn
Weakness	Triệu chứng mệt mỏi, cơ thể luôn cảm thấy mệt mỏi, kiệt sức, đặc biệt khi vừa ngủ dậy
Polyphagia	Triệu chứng đói quá mức, luôn muốn ăn cho dù vừa ăn xong, lúc nào cũng cảm thấy đói
Genital Thrush	Bị bệnh tưa miệng, xuất hiện các mảng trắng trong miệng hoặc lưỡi đỏ khác thường
Visual Blurring	Triệu chứng mờ mắt, thị lực giảm sút, có hiện tượng xuất huyết, phù nề trong mắt
Itching	Triệu chứng ngứa, da bị khô, bong tróc và nứt nẻ
Irritability	Triệu chứng khó chịu, cơ thể luôn bứt rứt, khó chịu và hay cáu gắt
Delayed Healing	Triệu chứng khó lành vết thương, xuất hiện các biến chứng khác trong quá trình hồi phục
Partial Paresis	Triệu chứng liệt, cơ thể sẽ bị liệt một bộ phận nào đó
Muscle Stiffness	Triệu chứng cứng cơ, cảm thấy cơ xương khớp bị cứng, khó vận động
Alopecia	Triệu chứng rụng tóc, tóc rất yếu, mỏng và rụng nhiều
Obesity	Mắc bệnh béo phì

Nhóm tác giả lớp K22MCS - Đại học Duy Tân:

+ Phạm Minh Tuấn
+ Võ Đình Hiếu
+ Nguyễn Anh Quân

Nhóm tác giả lớp K22MCS - Đại học Duy Tân © 2021

Hình 3.5: Giao diện trang giới thiệu

Phạm Minh Tuấn, Võ Đình Hiếu, Nguyễn Anh Quân - Lớp: K22MCS

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 hỗ trợ ra quyết định, 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:

Đưa ra cơ sở lý thuyết về mô hình cây quyết định và thuật toán C4.5 để ứng dụng vào việc phân tích dữ liệu để chuẩn đoán bệnh.

Phát triển website dựa trên tập luật sinh ra từ dữ liệu được tổng hợp và khai thác tại trang web <https://archive.ics.uci.edu/ml/index.php> để triển khai và vận hành thực tế.

Tiểu luận đã cho thấy sự hữu ích của việc phân tích dữ liệu để áp dụng, giải quyết các bài toán thực tế 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 hết tất cả các thuật toán về cây quyết định để áp dụng linh hoạt tùy theo từng trường hợp dữ liệu để tăng độ chính xác kết quả trả về.

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 hỗ trợ ra quyết định – PGS.TS Lê Đắc Như
- 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>