TRƯỜNG ĐẠI HỌC DUY TÂN

KHOA SAU ĐẠI HỌC

*Tiểu luận môn*

**CƠ SỞ DỮ LIỆU NÂNG CAO**

**HỆ THỐNG HỖ TRỢ CHUẨN ĐOÁN**

**BỆNH TIỂU ĐƯỜNG**

Hướng dẫn : TS. Lê Thanh Long

Thực hiện : Phạm Minh Tuấn

Võ Đình Hiếu

Nguyễn Anh Quân

Phạm Thị Thanh Thúy

Lớp : K22MCS (Khoa học máy tính)

Đà nẵng, 04/2021

# 

**MỤC LỤC**

[LỜI MỞ ĐẦU 1](#_Toc68351130)

[Chương 1:](#_Toc68351131) [CÂY QUYẾT ĐỊNH VÀ THUẬT TOÁN ID3 3](#_Toc68351132)

[1.1. Khái niệm cây quyết định 3](#_Toc68351133)

[1.2. Các loại cây quyết định 3](#_Toc68351134)

[1.3. Khởi tạo cây quyết định 4](#_Toc68351135)

[1.4. Sử dụng cây quyết định 6](#_Toc68351136)

[1.5. Duyệt cây và phân lớp dữ liệu 6](#_Toc68351137)

[1.5.1. Tiêu chuẩn phân lớp 6](#_Toc68351138)

[1.5.2. Điều kiện dừng phân lớp 7](#_Toc68351139)

[1.5.3. Hàm số Entropy 7](#_Toc68351140)

[1.5.4. Độ lợi thông tin (Information Gain) 8](#_Toc68351141)

[1.5.5. Vấn đề quá khớp trong phân lớp dữ liệu 9](#_Toc68351142)

[1.5.6. Cây quyết định với dữ liệu lớn 10](#_Toc68351143)

[1.6. Thuật toán duyệt cây id3 10](#_Toc68351144)

[1.7. Ưu và nhược điểm của cây quyết định 11](#_Toc68351145)

[Chương 2:](#_Toc68351146) [DỮ LIỆU BỆNH TIỂU ĐƯỜNG VÀ THUẬT TOÁN ID3 13](#_Toc68351147)

[2.1. Giới thiệu bệnh tiểu đường 13](#_Toc68351148)

[2.2. Thông tin dữ liệu 13](#_Toc68351149)

[2.3. Mô tả thuộc tính 14](#_Toc68351150)

[2.4. Cài đặt thuật toán trên tập dữ liệu 15](#_Toc68351151)

[2.5. Danh sách luật 40](#_Toc68351152)

[Chương 3:](#_Toc68351153) [ỨNG DỤNG THỰC TẾ 47](#_Toc68351154)

[3.1. Giới thiệu ứng dụng 47](#_Toc68351155)

[3.2. Truy cập ứng dụng 47](#_Toc68351156)

[3.3. Giao diện và tính năng 47](#_Toc68351157)

[Chương 4:](#_Toc68351158) [KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 52](#_Toc68351159)

[TÀI LIỆU THAM KHẢO 53](#_Toc68351160)

**DANH MỤC HÌNH VẼ**

[Hình 1.1: Cấu trúc cây quyết định 3](#_Toc68352498)

[Hình 1.2: Cây quyết định mua máy tính 5](#_Toc68352499)

[Hình 1.3: Hàm số Entropy 7](#_Toc68352500)

[Hình 2.1: Cây quyết định tại thuộc tính Polyuria 28](#_Toc68352501)

[Hình 2.2: Cây quyết định tại thuộc tính Alopecia 38](#_Toc68352502)

[Hình 2.3: Cây quyết định với bảng dữ liệu mẫu 40](#_Toc68352503)

[Hình 3.1: Truy cập ứng dụng Hỗ trợ chuẩn đoán bệnh tiểu đường 47](#_Toc68352504)

[Hình 3.2: Giao diện trang chủ ứng dụng 48](#_Toc68352505)

[Hình 3.3: Giao diện trang kết quả nếu không mắc bệnh 49](#_Toc68352506)

[Hình 3.4: Giao diện trang kết quả người dùng mắc bệnh 50](#_Toc68352507)

[Hình 3.5: Giao diện trang giới thiệu 51](#_Toc68352508)

**DANH MỤC BẢNG**

[Bảng 1.1: Dữ liệu mua máy tính 5](#_Toc68352487)

[Bảng 2.1: Các thuộc tính và giá trị của cơ sở dữ liệu 15](#_Toc68352488)

[Bảng 2.2: Cơ sở dữ liệu mẫu 16](#_Toc68352489)

[Bảng 2.3: Cơ sở dữ liệu mẫu sau rời rạc hóa giá trị 17](#_Toc68352490)

[Bảng 2.4: Mẫu dữ liệu với thuộc tính Polyuria có giá trị Yes(S1) 28](#_Toc68352491)

[Bảng 2.5: Mẫu dữ liệu với thuộc tính Alopecia có giá trị Yes(S2) 39](#_Toc68352492)

# 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. 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.

1. **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 Cơ sở dữ liệu nâng cao và các môn học khác có liên quan 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 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 TS. Lê Thanh Long cùng với kiến thức các môn đã được học nên nhóm đã xây dựng được khá hoàn thiện một ứng dụng hỗ trợ chuẩn đoán và đưa ra kết quả khá thiết thực.

1. **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 ID3

Chương 2: Thông tin về dữ liệu bệnh tiểu đường và áp dụng thuật toán ID3

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

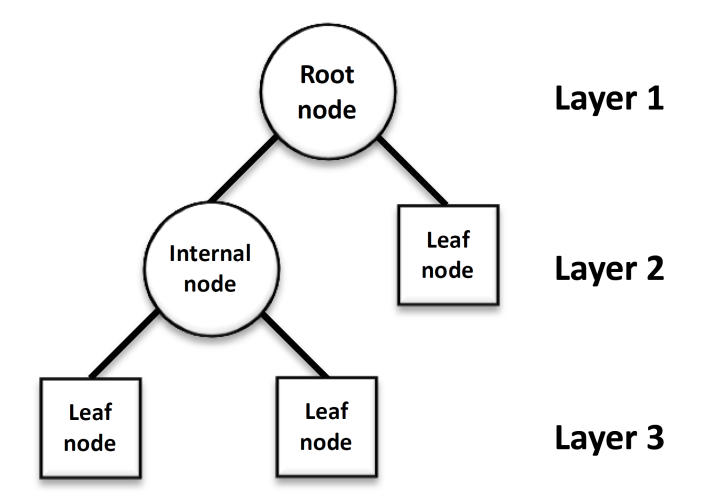
Chương 4: Kết luận: Trình bày kết quả đạt được của tiểu luận và tính thực tế cũng như định hướng phát triển trong tương lai.

# Chương 1

# CÂY QUYẾT ĐỊNH VÀ THUẬT TOÁN ID3

## ****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

## ****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).

## ****KHỞI TẠO CÂY QUYẾT ĐỊNH****

Có nhiều thuật toán khác nhau để xây dựng cây quyết định như: CLS, CART, ID3, ID3, SLIQ, SPRINT, EID3, 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 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 | CreditRating | BuyComputer |
| 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 (CreditRating) và thuộc tính nhãn lớp mô tả việc mua máy tính hay không (BuyComputer).

Cây quyết định thu được với bảng dữ liệu trên:

Age

Student

CreditRating

31 - 40

>40

<=30

No

Yes

Excellent

Fair

Hình 1.2: Cây quyết định mua máy tính

## ****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:

1. Nếu (Age<=30) và (Student=Yes) Thì BuyComputer = Yes
2. Nếu (Age từ 31 - 40) Thì BuyComputer = Yes
3. Nếu (Age >40) và (CreditRating=Fair) Thì BuyComputer = Yes

## ****DUYỆT CÂY VÀ PHÂN LỚP DỮ LIỆU****

### ****Tiêu chuẩn phân lớp****

Tại mỗi bước khi duyệt thuật toán, để lựa chọn dữ liệu thì ta chọn một thuộc tính tốt nhất theo một tiêu chí nào đó. Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các nút con tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi nút con. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn tham lam (greedy). Cách chọn này có thể không phải là tối ưu, nhưng cách làm này sẽ gần với cách làm tối ưu. Ngoài ra, cách làm này khiến cho bài toán cần giải quyết trở nên đơn giản hơn.

Phép phân lớp là nếu dữ liệu trong mỗi nút con hoàn toàn thuộc vào một lớp thì khi đó nút con này có thể được coi là một nút lá, tức ta không cần phân chia thêm nữa. Nếu dữ liệu trong các nút lá vẫn lẫn vào nhau theo tỉ lệ lớn, ta coi rằng phép phân chia đó chưa thực sự tốt.

Độ đo để đánh giá chất lượng phân chia là độ đo đồng nhất, ta có 2 tiêu chuẩn hay dùng nhất trong việc lựa chọn:

* Entropy
* Information Gain

### ****Điều kiện dừng phân lớp****

* 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 ID3 nên trong nội dung tiểu luận chỉ đề cập đến độ lợi để chọn lựa thuộc tính phân lớp.

### ****Hàm số Entropy****

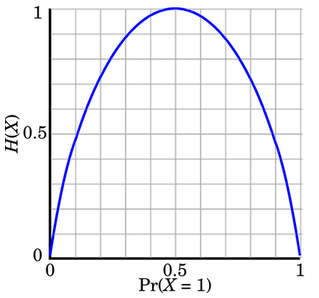
Cho một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1,x2,……,xn. Giả sử rằng xác suất để x nhận các giá trị này là pi=p(x=xi)

Ký hiệu phân phối này là p = (p1,p2,…,pn).

Entropy của phân phối này là:

(công thức 1.1a)

Hàm Entropy được biểu diễn dưới dạng đồ thị như sau:



Hình 1.3: Hàm số Entropy

Từ đồ thị ta thấy, hàm Entropy sẽ đạt giá trị nhỏ nhất nếu có một giá trị pi = 1, đạt giá trị lớn nhất nếu tất cả các pi bằng nhau.

Hàm Entropy càng lớn thì độ ngẫu nhiên của các biến rời rạc càng cao (càng không tinh khiết).

### ****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:

(công thức 1.1b)

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:

(công thức 1.2)

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

(công thức 1.3)

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

Thừa nhận:

* Lớp P: BuyComputer = “Yes”
* Lớp N: BuyComputer = “No”
* Thông tin cần thiết để phân lớp một mẫu được cho là:

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:

Do đó: Gain(Age) = E(9,5) - E(Age) = 0.246

Tương tự:

Gain(Income) = 0.029

Gain(Student) = 0.151

Gain(CreditRating) = 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.

### ****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 nhiễu 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.

### ****Cây quyết định với 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 dãn 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.

## ****THUẬT TOÁN DUYỆT CÂY ID3****

**Đầu vào:** Một tập hợp các ví dụ. Mỗi ví dụ 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 ví dụ trong tập dữ liệu rèn luyện, và hy vọng là phân loại đúng cho cả các ví dụ chưa gặp trong tương lai.

ID3 xây dựng cây quyết định theo giải thuật sau:

**Function InduceTree** (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 có giá trị V tại thuộc tính P;

Gọi **InduceTree** (phân vùngV, tập thuộc tính), gắn kết quả vào nhánh V

**End**

**End**

**End**

## ****ƯU VÀ NHƯỢC ĐIỂM CỦA CÂY QUYẾT ĐỊNH****

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích của nó nhưng đi kèm là các ưu và nhược điểm sau:

**Ưu điểm:**

* Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
* Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
* Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
* Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
* Có khả năng làm việc với dữ liệu lớn

**Nhược điểm:**

* Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
* Cây quyết định hay gặp vấn đề quá mức

# Chương 2

# DỮ LIỆU BỆNH TIỂU ĐƯỜNG VÀ THUẬT TOÁN ID3

* 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.

* 1. **THÔNG TIN DỮ LIỆU**

Dữ liệu được lấy từ website:

[https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#](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
   1. **MÔ TẢ THUỘC TÍNH**

Đặc điểm của tập dữ liệu: Đa biến

Số lượng bản ghi: 529

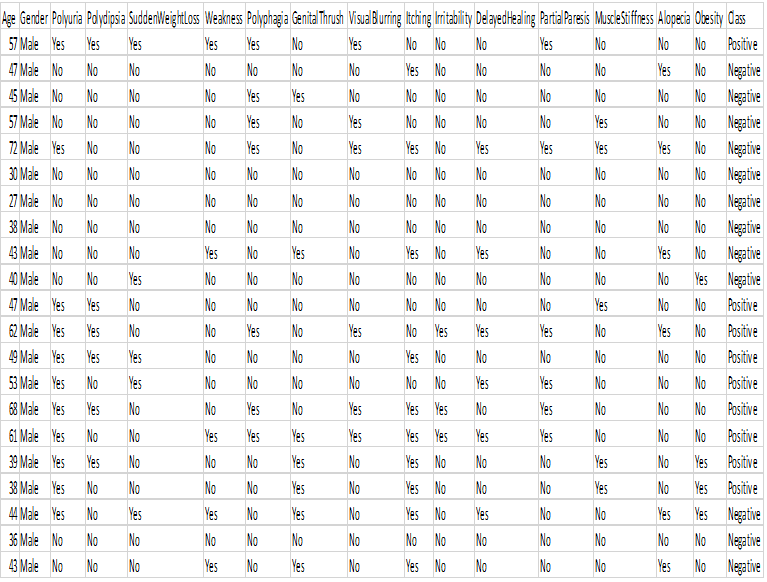
Số thuộc tính: 17

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Thuộc tính | Kiểu | Giá trị | Diễn giải |
| 1 | Age | Numeric | 16-90 | Tuổi bệnh nhân |
| 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 rõ 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 | Mắc bệnh, Không mắc bệnh |

Bảng 2.1: Các thuộc tính và giá trị của cơ sở dữ liệu

* 1. **CÀI ĐẶT THUẬT TOÁN TRÊN TẬP DỮ LIỆU**

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



Bảng 2.2: Cơ sở dữ liệu mẫu

Với bảng dữ liệu cơ sở mẫu, thuộc tính Age kiểu Numeric ta thực hiện rời rạc hóa thuộc tính Age thành kiểu Nominal để phù hợp với thuật toán ID3. Dữ liệu tuổi trong bảng dữ liệu mẫu bắt đầu từ 27-72. Ta chia làm 4 khoảng như sau: 27-38; 39-49; 50-60; 61-72. Ta được bảng cơ sở dữ liệu như sau:



Bảng 2.3: Cơ sở dữ liệu mẫu sau rời rạc hóa giá trị

Bộ mẫu dữ liệu của chúng ta có 02 miền giá trị {p, n} (p ứng với “Positive” và n ứng với “Negative”).

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

* Tính Gain cho thuộc tính Age:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Age | | | |
| STT | Age | pi | ni |
| 1 | 27-38 (5) | 1 | 4 |
| 2 | 39-49 (9) | 3 | 6 |
| 3 | 50-60 (3) | 2 | 1 |
| 4 | 61-72 (4) | 3 | 1 |

Ta có:

Do đó:

Gain(Age) = E(S) – E(Age) = 0.985 – 0.851 = 0.134

* Tính Gain cho thuộc tính Gender:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Gender | | | |
| STT | Gender | pi | ni |
| 1 | Male (21) | 9 | 12 |
| 2 | Female (0) | 0 | 0 |

Ta có:

Do đó:

Gain(Gender) = E(S) – E(Gender) = 0.985 – 0.985 = 0

* Tính Gain cho thuộc tính Polyuria:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Polyura | | | |
| STT | Polyuria | pi | ni |
| 1 | Yes (11) | 9 | 2 |
| 2 | No (10) | 0 | 10 |

Ta có:

Do đó:

Gain(Polyuria) = E(S) – E(Polyuria) = 0.985 – 0.358 = 0.627

* Tính Gain cho thuộc tính Polydipsia:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Polydipsia | | | |
| STT | Polydipsia | pi | ni |
| 1 | Yes (6) | 6 | 0 |
| 2 | No (15) | 3 | 12 |

Ta có:

Do đó:

Gain(Polydipsia) = E(S) – E(Polydipsia) = 0.985 - 0.516= 0.469

* Tính Gain cho thuộc tính SuddenWeightLoss:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính SuddenWeightLoss | | | |
| STT | SuddenWeightLoss | pi | ni |
| 1 | Yes (5) | 3 | 2 |
| 2 | No (16) | 6 | 10 |

Ta có:

Do đó:

Gain(SuddenWeightLoss) = E(S) - E(SuddenWeightLoss) = 0.985 - 0958 = 0.027

* Tính Gain cho thuộc tính Weakness:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Weakness | | | |
| STT | Weakness | pi | ni |
| 1 | Yes (5) | 2 | 3 |
| 2 | No (16) | 7 | 9 |

Ta có:

Do đó:

Gain(Weakness) = 0.985 - 0.985 = 0

* Tính Gain cho thuộc tính Polyphagia:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Polyphagia | | | |
| STT | Polyphagia | pi | ni |
| 1 | Yes (7) | 4 | 3 |
| 2 | No (14) | 5 | 9 |

Ta có:

Do đó:

Gain(Polyphagia) = 0.985 - 0.955 = 0.030

* Tính Gain cho thuộc tính GenitalThrush:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính GenitalThrush | | | |
| STT | GenitalThrush | pi | ni |
| 1 | Yes (7) | 3 | 4 |
| 2 | No (14) | 6 | 8 |

Ta có:

Do đó:

Gain(GenitalThrush) = 0.985 - 0.985 = 0

* Tính Gain cho thuộc tính VisualBlurring:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính VisualBlurring | | | |
| STT | VisualBlurring | pi | ni |
| 1 | Yes (6) | 4 | 2 |
| 2 | No (15) | 5 | 10 |

Ta có:

Do đó:

Gain(VisualBlurring) = 0.985 - 0.918 = 0.067

* Tính Gain cho thuộc tính Itching:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Itching | | | |
| STT | Itching | pi | ni |
| 1 | Yes (10) | 5 | 5 |
| 2 | No (11) | 3 | 8 |

Ta có:

Do đó:

Gain(Itching) = 0.985 - 0.919 = 0.066

* Tính Gain cho thuộc tính Irritability:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Irritability | | | |
| STT | Irritability | pi | ni |
| 1 | Yes (3) | 3 | 0 |
| 2 | No (18) | 6 | 12 |

Ta có:

Do đó:

Gain(Irritability) = 0.985 - 0.787 = 0.198

* Tính Gain cho thuộc tính DelayedHealing:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính DelayedHealing | | | |
| STT | DelayedHealing | pi | ni |
| 1 | Yes (6) | 3 | 3 |
| 2 | No (15) | 6 | 9 |

Ta có:

Do đó:

Gain(DelayedHealing) = 0.985 - 0.979 = 0.006

* Tính Gain cho thuộc tính PartialParesis:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính PartialParesis | | | |
| STT | PartialParesis | pi | ni |
| 1 | Yes (6) | 5 | 1 |
| 2 | No (15) | 4 | 11 |

Ta có:

Do đó:

Gain(PartialParesis) = 0.985 - 0.784 = 0.201

* Tính Gain cho thuộc tính MuscleStiffness:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính MuscleStiffness | | | |
| STT | MuscleStiffness | pi | ni |
| 1 | Yes (5) | 3 | 2 |
| 2 | No (16) | 6 | 10 |

Ta có:

Do đó:

Gain(MuscleStiffness) = 0.985 - 0.958 = 0.027

* Tính Gain cho thuộc tính Alopecia:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Alopecia | | | |
| STT | Alopecia | pi | ni |
| 1 | Yes (6) | 1 | 5 |
| 2 | No (15) | 8 | 7 |

Ta có:

Do đó:

Gain(Alopecia) = 0.985 - 0.898 = 0.087

* Tính Gain cho thuộc tính Obesity:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Obesity | | | |
| STT | Obesity | pi | ni |
| 1 | Yes (4) | 2 | 2 |
| 2 | No (17) | 7 | 10 |

Ta có:

Do đó:

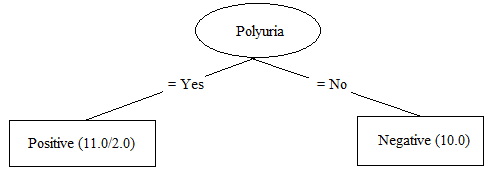
Gain(Obesity) = 0.985 - 0.981 = 0.004

Độ đo Gain của các thuộc tính được sắp xếp giảm dần:

|  |  |  |
| --- | --- | --- |
| STT | Thuộc tính | Gain |
| 1 | Polyuria | 0.627 |
| 2 | Polydipsia | 0.469 |
| 3 | PartialParesis | 0.201 |
| 4 | Irritability | 0.198 |
| 5 | Age | 0.134 |
| 6 | Alopecia | 0.087 |
| 7 | VisualBlurring | 0.067 |
| 8 | Itching | 0.066 |
| 9 | Polyphagia | 0.030 |
| 10 | SuddenWeightLoss | 0.027 |
| 11 | MuscleStiffness | 0.027 |
| 12 | DelayedHealing | 0.006 |
| 13 | Obesity | 0.004 |
| 14 | Gender | 0 |
| 15 | Weakness | 0 |
| 16 | GenitalThrush | 0 |

Như vậy thuộc tính có độ đo Gain 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 ID3 cho từng nút tương ứng bảng dữ liệu sau:



Bảng 2.4: Mẫu dữ liệu với thuộc tính Polyuria có giá trị Yes(S1)

Tính lượng Entropy thông tin trên tất cả mẫu dữ liệu S1 theo bảng 2.4:

* Tính Gain của cho thuộc tính Age:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Age | | | |
| STT | Age | pi | ni |
| 1 | 27-38 (1) | 1 | 0 |
| 2 | 39-49 (4) | 3 | 1 |
| 3 | 50-60 (2) | 2 | 0 |
| 4 | 61-72 (4) | 3 | 1 |

Ta có:

Do đó:

Gain(Age) = 0.684 - 0.590 = 0.094

* Tính Gain cho thuộc tính Gender:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Gender | | | |
| STT | Gender | pi | ni |
| 1 | Male (11) | 9 | 2 |
| 2 | Female (0) | 0 | 0 |

Ta có:

Do đó:

Gain(Gender) = 0.684 – 0.684 = 0

* Tính Gain cho thuộc tính Polydipsia:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Polydipsia | | | |
| STT | Polydipsia | pi | ni |
| 1 | Yes (6) | 6 | 0 |
| 2 | No (5) | 3 | 2 |

Ta có:

Do đó:

Gain(Polydipsia) = 0.684 - 0.441= 0.243

* Tính Gain cho thuộc tính SuddenWeightLoss:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính SuddenWeightLoss | | | |
| STT | SuddenWeightLoss | pi | ni |
| 1 | Yes (4) | 3 | 1 |
| 2 | No (7) | 6 | 1 |

Ta có:

Do đó:

Gain(SuddenWeightLoss) = 0.684 - 0.672 = 0.012

* Tính Gain cho thuộc tính Weakness:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Weakness | | | |
| STT | Weakness | pi | ni |
| 1 | Yes (3) | 2 | 1 |
| 2 | No (8) | 7 | 1 |

Ta có:

Do đó:

Gain(Weakness) = 0.684 - 0.646 = 0.038

* Tính Gain cho thuộc tính Polyphagia:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Polyphagia | | | |
| STT | Polyphagia | pi | ni |
| 1 | Yes (5) | 4 | 1 |
| 2 | No (6) | 5 | 1 |

Ta có:

Do đó:

Gain(Polyphagia) = 0.684 - 0.683 = 0.001

* Tính Gain cho thuộc tính GenitalThrush:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính GenitalThrush | | | |
| STT | GenitalThrush | pi | ni |
| 1 | Yes (7) | 6 | 1 |
| 2 | No (4) | 3 | 1 |

Ta có:

Do đó:

Gain(GenitalThrush) = 0.684 - 0.672 = 0.012

* Tính Gain cho thuộc tính VisualBlurring:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính VisualBlurring | | | |
| STT | VisualBlurring | pi | ni |
| 1 | Yes (5) | 4 | 1 |
| 2 | No (6) | 5 | 1 |

Ta có:

Do đó:

Gain(VisualBlurring) = 0.684 - 0.683 = 0.001

* Tính Gain cho thuộc tính Itching:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Itching | | | |
| STT | Itching | pi | ni |
| 1 | Yes (7) | 5 | 2 |
| 2 | No (4) | 4 | 0 |

Ta có:

Do đó:

Gain(Itching) = 0.684 – 0.549 = 0.135

* Tính Gain cho thuộc tính Irritability:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Irritability | | | |
| STT | Irritability | pi | ni |
| 1 | Yes (3) | 3 | 0 |
| 2 | No (8) | 6 | 2 |

Ta có:

Do đó:

Gain(Irritability) = 0.684 – 0.59 = 0.094

* Tính Gain cho thuộc tính DelayedHealing:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính DelayedHealing | | | |
| STT | DelayedHealing | pi | ni |
| 1 | Yes (5) | 3 | 2 |
| 2 | No (6) | 6 | 0 |

Ta có:

Do đó:

Gain(DelayedHealing) = 0.684 – 0.441 = 0.243

* Tính Gain cho thuộc tính PartialParesis:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính PartialParesis | | | |
| STT | PartialParesis | pi | ni |
| 1 | Yes (6) | 5 | 1 |
| 2 | No (5) | 4 | 1 |

Ta có:

Do đó:

Gain(PartialParesis) = 0.684 – 0.683 = 0.001

* Tính Gain cho thuộc tính MuscleStiffness:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính MuscleStiffness | | | |
| STT | MuscleStiffness | pi | ni |
| 1 | Yes (4) | 3 | 1 |
| 2 | No (7) | 6 | 1 |

Ta có:

Do đó:

Gain(MuscleStiffness) = 0.684 – 0.672 = 0.012

* Tính Gain cho thuộc tính Alopecia:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Alopecia | | | |
| STT | Alopecia | pi | ni |
| 1 | Yes (3) | 1 | 2 |
| 2 | No (8) | 8 | 0 |

Ta có:

Do đó:

E(Alopecia) = \* E(p1, n1)+ \* E(p2, n2) = 0.25

Gain(Alopecia) = 0.684 – 0.25 = 0.434

* Tính Gain cho thuộc tính Obesity:

|  |  |  |  |
| --- | --- | --- | --- |
| Bảng thuộc tính Obesity | | | |
| STT | Obesity | pi | ni |
| 1 | Yes (3) | 2 | 1 |
| 2 | No (8) | 7 | 1 |

Ta có:

Do đó:

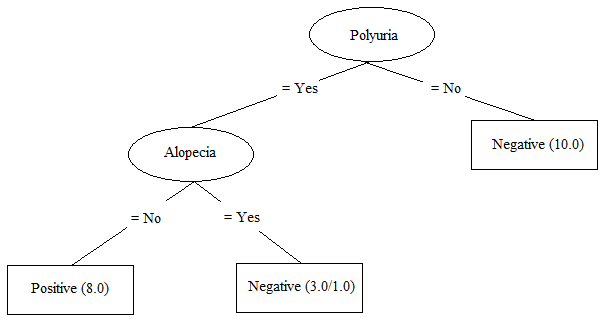
Gain(Obesity) = 0.684 – 0.646 = 0.038

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

|  |  |  |
| --- | --- | --- |
| STT | Thuộc tính | Gain |
| 1 | Alopecia | 0.434 |
| 2 | Polydipsia | 0.243 |
| 3 | DelayedHealing | 0.243 |
| 4 | Itching | 0.135 |
| 5 | Age | 0.094 |
| 6 | Irritability | 0.094 |
| 7 | Weakness | 0.038 |
| 8 | Obesity | 0.038 |
| 9 | SuddenWeightLoss | 0.012 |
| 10 | GenitalThrush | 0.012 |
| 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 Gain 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:



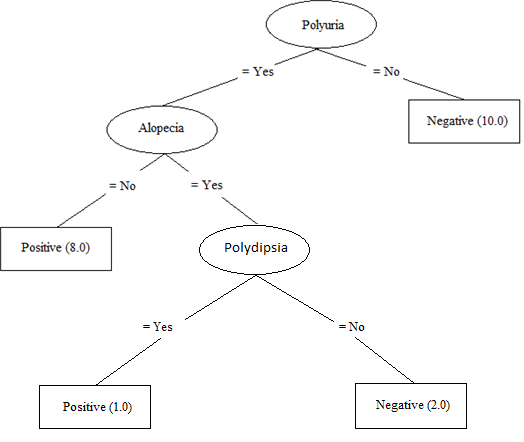
Bảng 2.5: Mẫu dữ liệu với thuộc tính Alopecia có giá trị Yes(S2)

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

|  |  |  |
| --- | --- | --- |
| STT | Thuộc tính | Gain |
| 1 | Polydipsia | 0.918 |
| 2 | Itching | 0.918 |
| 3 | Irritability | 0.918 |
| 4 | Age | 0.251 |
| 5 | Weakness | 0.251 |
| 6 | Obesity | 0.251 |
| 7 | SuddenWeightLoss | 0.251 |
| 8 | GenitalThrush | 0.251 |
| 9 | MuscleStiffness | 0.251 |
| 10 | Polyphagia | 0.251 |
| 11 | VisualBlurring | 0.251 |
| 12 | PartialParesis | 0.251 |
| 13 | Gender | 0 |
| 14 | DelayedHealing | 0 |

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

Cây quyết định với thuật toán ID3 có dạng như sau:



Hình 2.3: Cây quyết định với bảng dữ liệu mẫu

* 1. **DANH SÁCH LUẬT**

Với bảng dữ liệu cơ sở, thuộc tính Age kiểu Numeric ta thực hiện rời rạc hóa thuộc tính Age thành kiểu Nominal để phù hợp với thuật toán ID3. Dữ liệu tuổi trong bảng dữ liệu mẫu có khoảng bắt đầu từ 16-90. Ta chia làm 4 khoảng như sau: 16-34; 35-53; 54-71; 72-90.

Sau khi dùng thuật toán ID3 thao tác trên dữ liệu mẫu ta rút ra được tập luật sau huấn luyện như sau:

|  |
| --- |
| Polyuria = No  | Gender = Male  | | Polydipsia = Yes  | | | Irritability = No  | | | | MuscleStiffness = Yes  | | | | | VisualBlurring = No: Positive  | | | | | VisualBlurring = Yes: Negative  | | | | MuscleStiffness = No  | | | | | PartialParesis = No: Positive  | | | | | PartialParesis = Yes  | | | | | | Age = 35-53: Negative  | | | | | | Age = 54-71: Positive  | | | | | | Age = 16-34: Negative  | | | | | | Age = 72-90: Positive  | | | Irritability = Yes: Positive  | | Polydipsia = No  | | | Irritability = No  | | | | Age = 35-53  | | | | | DelayedHealing = Yes  | | | | | | Itching = Yes  | | | | | | | SuddenWeightLoss = No: Negative  | | | | | | | SuddenWeightLoss = Yes: Positive  | | | | | | Itching = No  | | | | | | | SuddenWeightLoss = No: Positive  | | | | | | | SuddenWeightLoss = Yes: Negative  | | | | | DelayedHealing = No: Negative  | | | | Age = 54-71  | | | | | PartialParesis = No: Negative  | | | | | PartialParesis = Yes  | | | | | | MuscleStiffness = Yes: Negative  | | | | | | MuscleStiffness = No: Positive  | | | | Age = 16-34: Negative  | | | | Age = 72-90  | | | | | Weakness = Yes: Positive  | | | | | Weakness = No: Negative  | | | Irritability = Yes  | | | | GenitalThrush = No  | | | | | Age = 35-53  | | | | | | Polyphagia = No: Negative  | | | | | | Polyphagia = Yes: Positive  | | | | | Age = 54-71: Negative  | | | | | Age = 16-34: Negative  | | | | | Age = 72-90: Positive  | | | | GenitalThrush = Yes: Positive  | Gender = Female  | | Alopecia = Yes  | | | DelayedHealing = Yes: Negative  | | | DelayedHealing = No: Positive  | | Alopecia = No  | | | Age = 35-53  | | | | Irritability = No: Positive  | | | | Irritability = Yes  | | | | | Polydipsia = Yes: Positive  | | | | | Polydipsia = No: Negative  | | | Age = 54-71: Positive  | | | Age = 16-34  | | | | VisualBlurring = No: Negative  | | | | VisualBlurring = Yes: Positive  | | | Age = 72-90: Positive  Polyuria = Yes  | Polydipsia = Yes: Positive  | Polydipsia = No  | | Itching = Yes  | | | DelayedHealing = Yes  | | | | Alopecia = Yes  | | | | | Gender = Male  | | | | | | Age = 35-53  | | | | | | | SuddenWeightLoss = No: Positive  | | | | | | | SuddenWeightLoss = Yes: Negative  | | | | | | Age = 54-71: Negative  | | | | | | Age = 16-34: Negative  | | | | | | Age = 72-90: Negative  | | | | | Gender = Female: Positive  | | | | Alopecia = No: Positive  | | | DelayedHealing = No: Positive  | | Itching = No: Positive |

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

|  |
| --- |
| 1. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, có bị cứng cơ, không bị mờ mắt thì mắc bệnh tiểu đường. 2. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, có bị cứng cơ, có bị mờ mắt thì không mắc bệnh tiểu đường. 3. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, không bị cứng cơ, không bị liệt một bộ phận thì mắc bệnh tiểu đường. 4. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, không bị cứng cơ, bị liệt một bộ phận, có tuổi trong khoảng 35-53 thì không mắc bệnh tiểu đường. 5. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, không bị cứng cơ, bị liệt một bộ phận, có tuổi trong khoảng 54-71 thì mắc bệnh tiểu đường. 6. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, không bị cứng cơ, bị liệt một bộ phận, có tuổi trong khoảng 16-34 thì không mắc bệnh tiểu đường. 7. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, không khó chịu, không bị cứng cơ, bị liệt một bộ phận, có tuổi trong khoảng 72-90 thì mắc bệnh tiểu đường. 8. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, có khát nhiều, có khó chịu thì mắc bệnh tiểu đường. 9. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 35-53, có vết thương lâu lành, có bị ngứa, không bị sút cân bất thường thì không mắc bệnh tiểu đường. 10. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 35-53, có vết thương lâu lành, có bị ngứa, bị sút cân bất thường thì mắc bệnh tiểu đường. 11. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 35-53, có vết thương lâu lành, không bị ngứa, không bị sút cân bất thường thì không mắc bệnh tiểu đường. 12. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 35-53, có vết thương lâu lành, không bị ngứa, bị sút cân bất thường thì mắc bệnh tiểu đường. 13. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 35-53, không bị vết thương lâu lành thì không mắc bệnh tiểu đường. 14. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 54-71, không bị liệt một bộ phận thì không mắc bệnh tiểu đường. 15. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 54-71, bị liệt một bộ phận, bị cứng cơ thì không mắc bệnh tiểu đường. 16. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 54-71, bị liệt một bộ phận, không bị cứng cơ thì mắc bệnh tiểu đường. 17. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 16-34 thì không mắc bệnh tiểu đường. 18. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 72-90, có cảm thấy mệt mỏi thì mắc bệnh tiểu đường. 19. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, không khó chịu, có tuổi trong khoảng 72-90, không cảm thấy mệt mỏi thì không mắc bệnh tiểu đường. 20. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, có bị khó chịu, không bị tưa miệng, có tuổi trong khoảng 35-53, không bị đói quá mức thì không mắc bệnh tiểu đường. 21. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, có bị khó chịu, không bị tưa miệng, có tuổi trong khoảng 35-53, bị đói quá mức thì mắc bệnh tiểu đường. 22. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, có bị khó chịu, không bị tưa miệng, có tuổi trong khoảng 54-71 thì không mắc bệnh tiểu đường. 23. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, có bị khó chịu, không bị tưa miệng, có tuổi trong khoảng 16-34 thì không mắc bệnh tiểu đường. 24. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, có bị khó chịu, không bị tưa miệng, có tuổi trong khoảng 72-90 thì mắc bệnh tiểu đường. 25. Nếu bệnh nhân không tiểu nhiều, giới tính là nam, không khát nhiều, có bị khó chịu, bị tưa miệng thì mắc bệnh tiểu đường. 26. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, bị rụng tóc, có vết thương lâu lành thì không mắc bệnh tiểu đường. 27. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, bị rụng tóc, không có vết thương lâu lành thì mắc bệnh tiểu đường. 28. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, không bị rụng tóc, có tuổi trong khoảng 35-53, không khó chịu thì mắc bệnh tiểu đường. 29. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, không bị rụng tóc, có tuổi trong khoảng 35-53, có khó chịu thì không mắc bệnh tiểu đường. 30. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, không bị rụng tóc, có tuổi trong khoảng 54-71 thì mắc bệnh tiểu đường. 31. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, không bị rụng tóc, có tuổi trong khoảng 16-34, không bị mờ mắt thì không mắc bệnh tiểu đường. 32. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, không bị rụng tóc, có tuổi trong khoảng 16-34, bị mờ mắt thì mắc bệnh tiểu đường. 33. Nếu bệnh nhân không tiểu nhiều, giới tính là nữ, không bị rụng tóc, có tuổi trong khoảng 72-90 thì mắc bệnh tiểu đường. 34. Nếu bệnh nhân bị tiểu nhiều, có khát nhiều thì mắc bệnh tiểu đường. 35. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, bị rụng tóc, giới tính là nam, có tuổi trong khoảng 35-53, không bị sụt cân bất thường thì mắc bệnh tiểu đường. 36. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, bị rụng tóc, giới tính là nam, có tuổi trong khoảng 35-53, bị sụt cân bất thường thì không mắc bệnh tiểu đường. 37. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, bị rụng tóc, giới tính là nam, có tuổi trong khoảng 54-71 thì không mắc bệnh tiểu đường. 38. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, bị rụng tóc, giới tính là nam, có tuổi trong khoảng 16-34 thì không mắc bệnh tiểu đường. 39. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, bị rụng tóc, giới tính là nam, có tuổi trong khoảng 72-90 thì không mắc bệnh tiểu đường. 40. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, bị rụng tóc, giới tính là nữ thì mắc bệnh tiểu đường. 41. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, có vết thương lâu lành, không bị rụng tóc thì không mắc bệnh tiểu đường. 42. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, có bị ngứa, không có vết thương lâu lành thì không mắc bệnh tiểu đường. 43. Nếu bệnh nhân bị tiểu nhiều, không khát nhiều, không bị ngứa thì không mắc bệnh tiểu đường. |

# Chương 3

# ỨNG DỤNG THỰC TẾ

* 1. **GIỚI THIỆU ỨNG DỤNG**

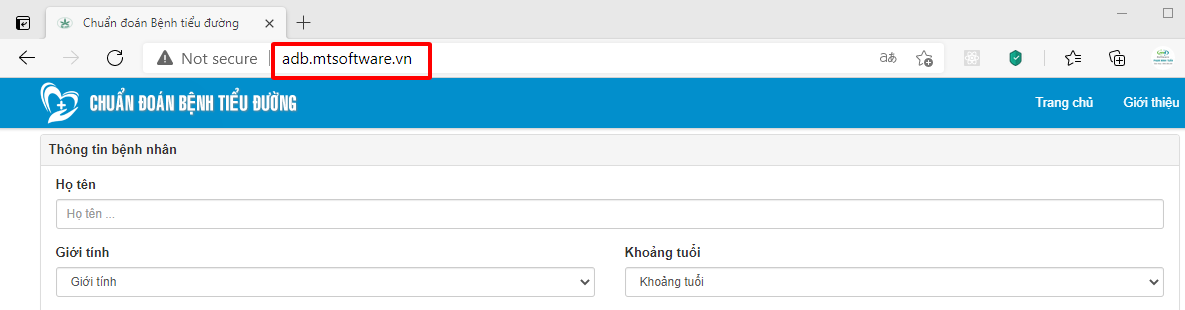
Chương trình “Hỗ trợ chuẩn đoán bệnh tiểu đường” được phát triển dạng website để tăng độ hiệu quả khi triển khai và dễ dàng nâng cấp ứng dụng khi có như cầu.

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.

* 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.

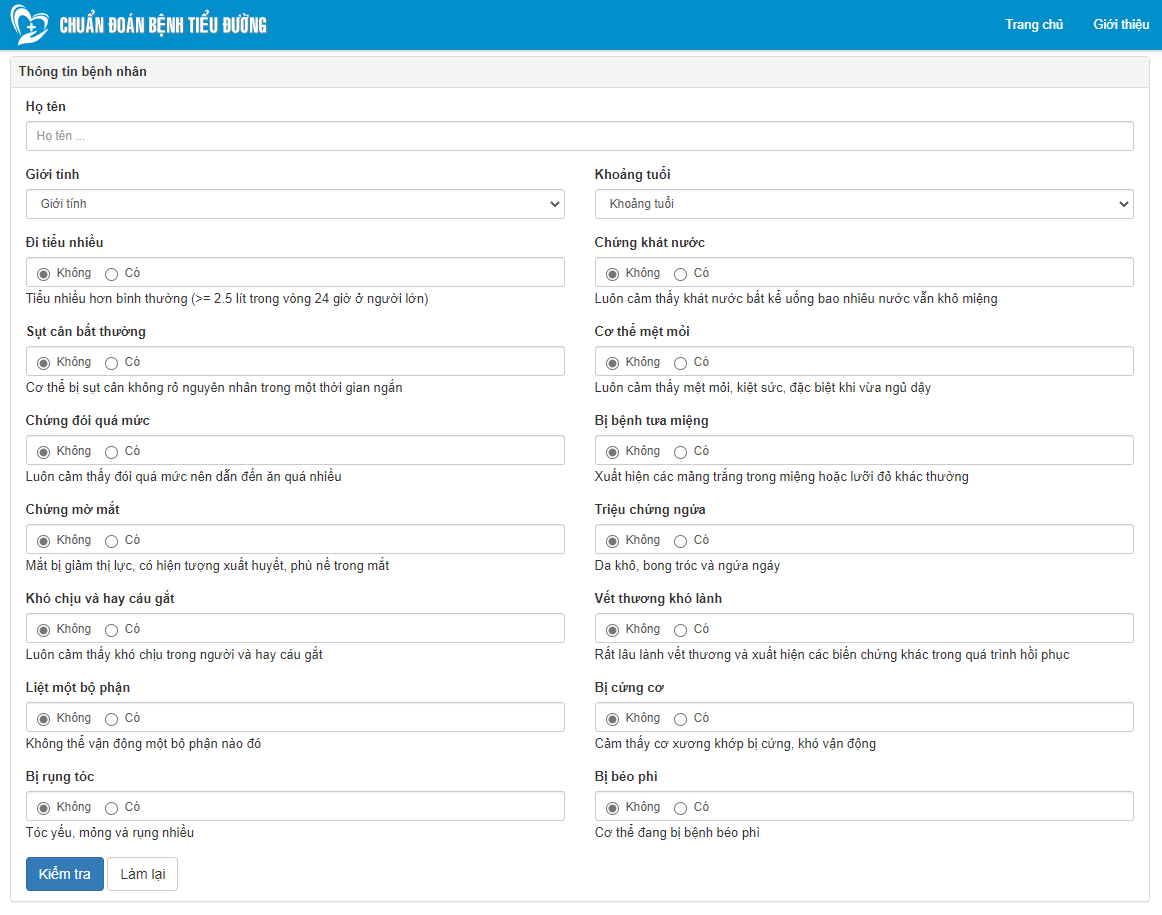
<http://adb.mtsoftware.vn/>



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

* 1. **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):



Hình 3.2: Giao diện trang chủ ứng dụng

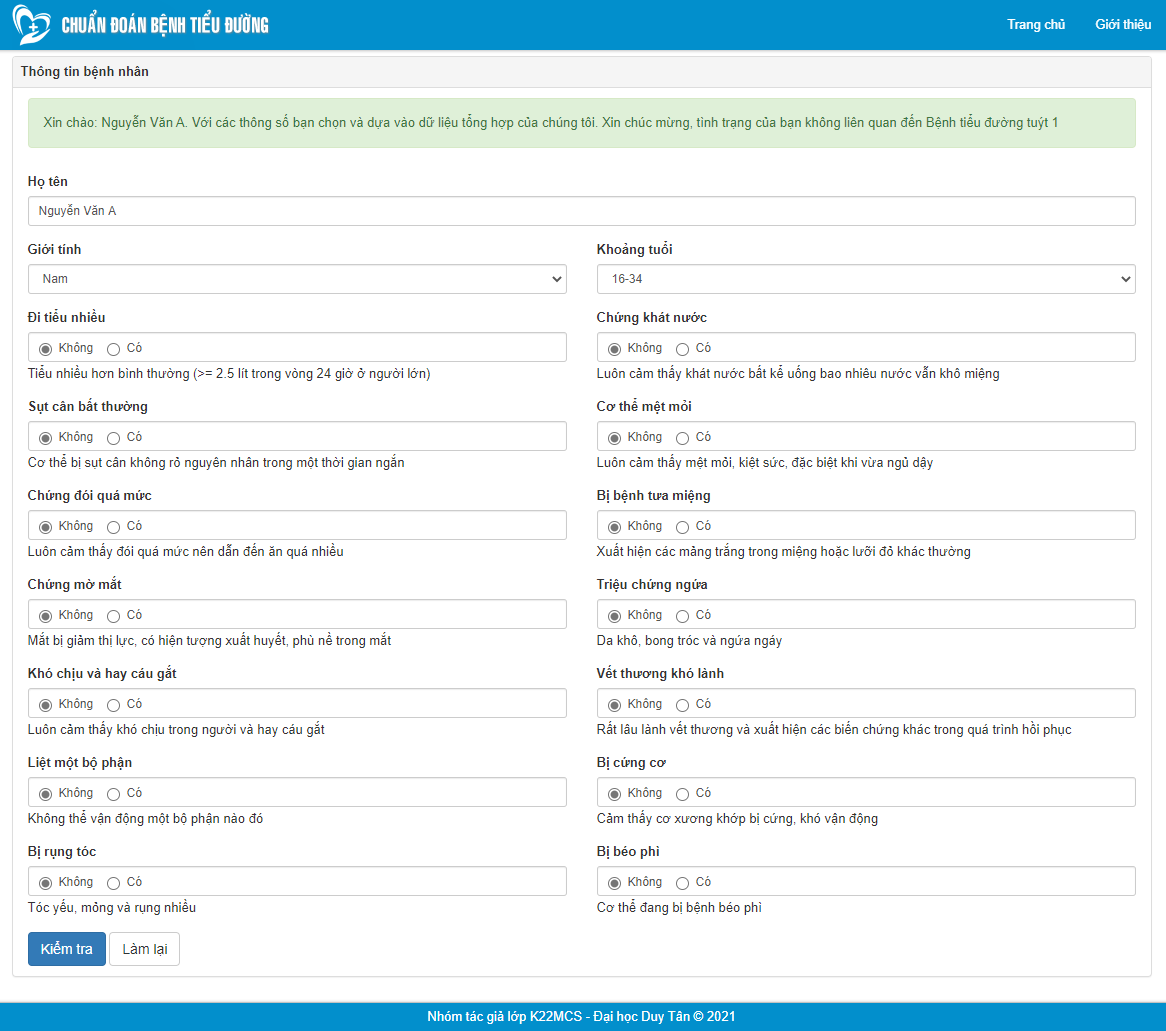
Tại giao diện trang chủ người dùng cần thao tác theo thứ tự như sau:

Bước 1: Nhập Họ tên, Giới tính, Khoảng 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):

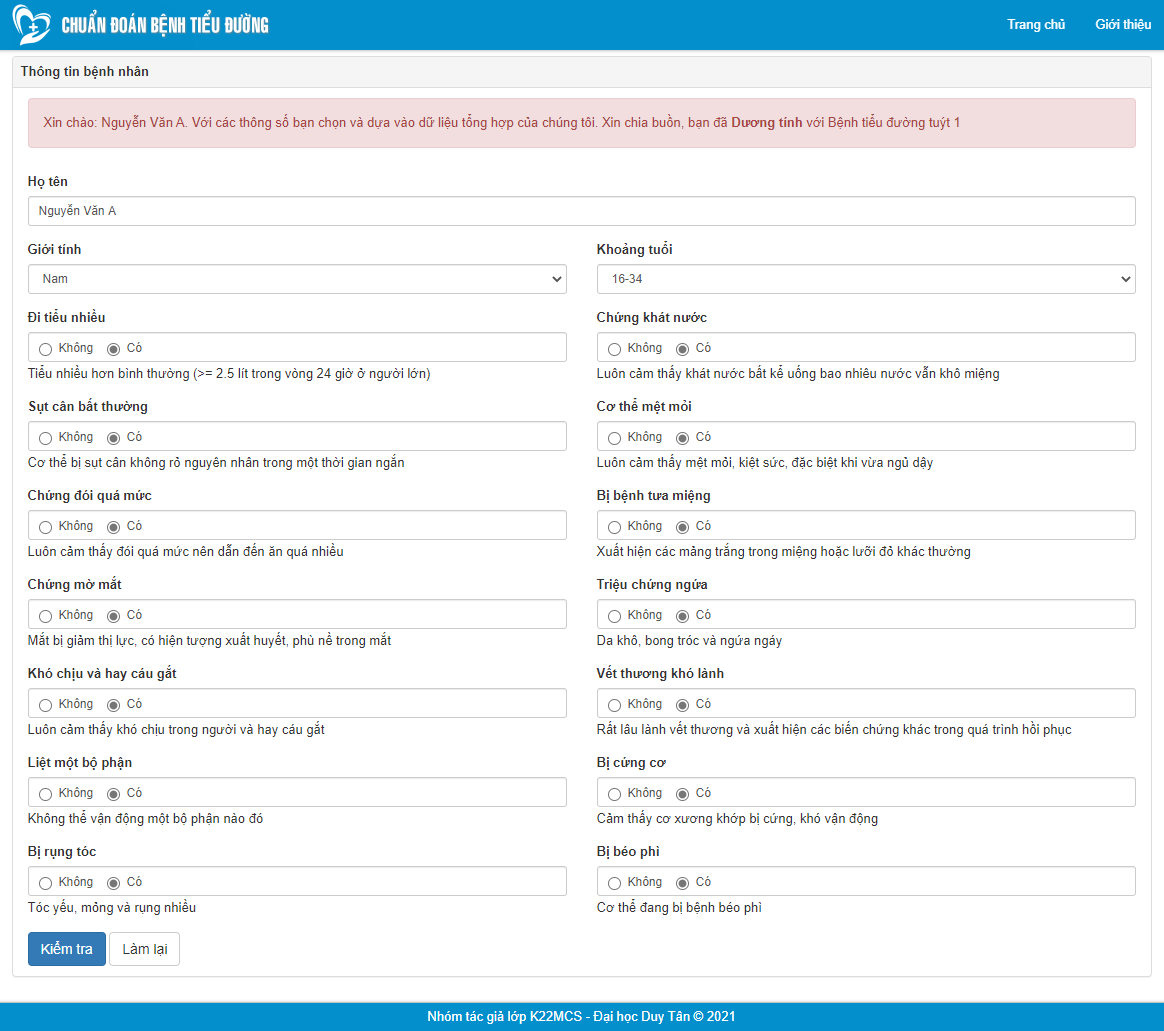
Xin chào: Nguyễn Văn A. Với các thông số bạn chọn và dựa vào dữ liệu tổng hợp của chúng tôi. Xin chúc mừng, tình trạng của bạn không liên quan đến Bệnh tiểu đường tuýt 1



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):

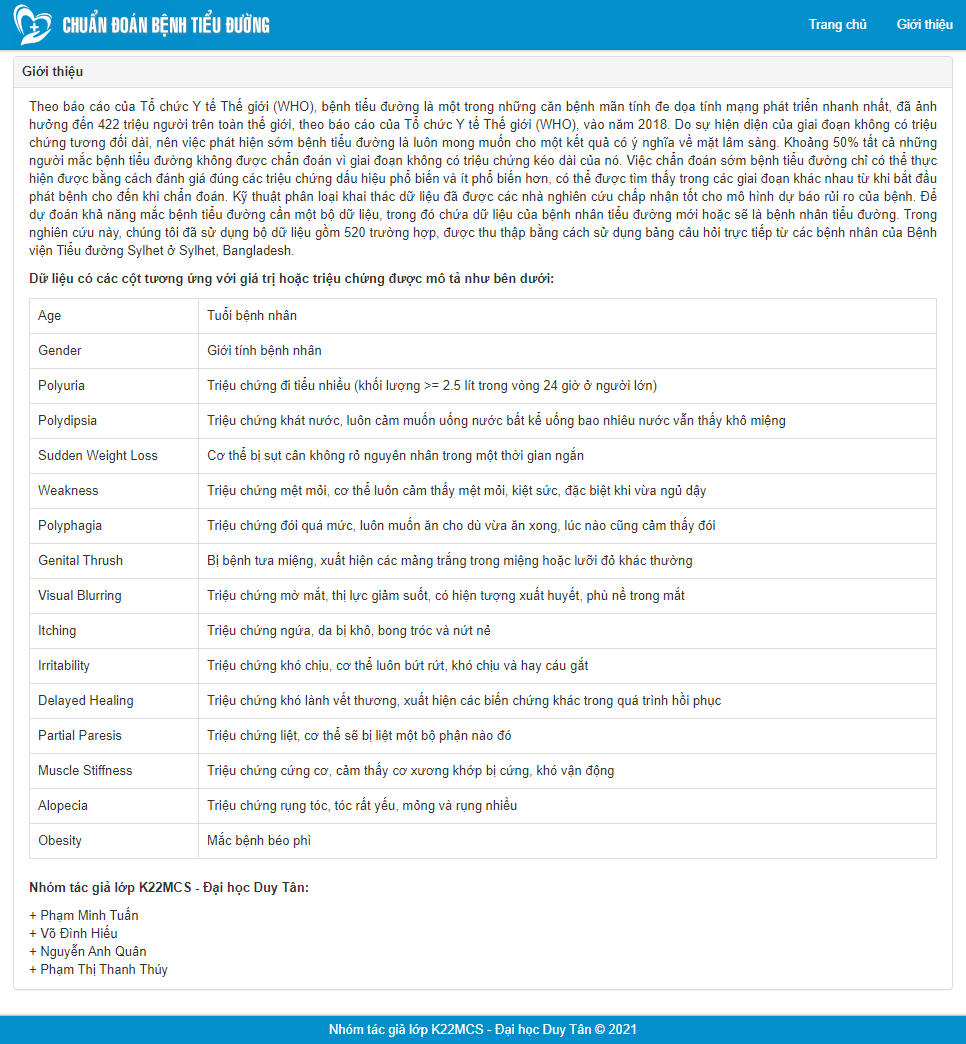
Xin chào: Nguyễn Văn A. Với các thông số bạn chọn và dựa vào dữ liệu tổng hợp của chúng tôi. Xin chia buồn, bạn đã Dương tính với Bệnh tiểu đường tuýt 1



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 phát triển.

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



Hình 3.5: 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 Cơ sở dữ liệu nâng cao, 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 ID3 để ứng dụng vào việc phân tích dữ liệu để chuẩn đoán bệnh.

Phát triển công cụ thực thi giải thuật ID3 để sinh ra tập luật từ tập dữ liệu bất kỳ có định dạng tập tin là \*.xlsx.

Phát triển website dựa trên tập luật sinh ra từ dữ liệu được lấy từ trang web <https://archive.ics.uci.edu/ml/index.php> sau đó chọn lọc xử lý sau đó 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 Cơ sở dữ liệu nâng cao – TS. Lê Thanh Long
2. <https://archive.ics.uci.edu/ml/index.php>
3. Peng, W., Chen, J., & Zhou, H. (2009). An implementation of ID3-decision tree learning algorithm. From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May, 13.
4. Bahety, A. (2014). Extension and evaluation of id3–decision tree algorithm. Entropy (S), 2(1), 1-8.
5. Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A new feature selection model based on ID3 and bees algorithm for intrusion detection system. Turkish Journal of Electrical Engineering & Computer Sciences, 23(2), 615-622.