**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

-----🙞🕮🙜-----

A yellow and red logo

Description automatically generated

**BÀI TẬP LỚN**

**MÔN: TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI:**

**TÌM HIỂU CÂY QUYẾT ĐỊNH VÀ ỨNG DỤNG DỰ ĐOÁN SỨC KHỎE**

**Sinh viên thực hiện: Bùi Tiến Đạt-2023604410**

**Thái Bách San-2023604234**

**Đỗ Thành Vinh-2023602405**

**Vũ Duy Việt-2023601997**

**Nhóm: 04**

**Lớp học phần: IT6094**

**Giảng viên hướng dẫn: Mai Thanh Hồng**

***Hà Nội, năm 2025***

# LỜI CẢM ƠN

Để hoàn thành tốt bài báo cáo "Tìm hiểu cây quyết định và ứng dụng dự đoán bệnh tiểu đường", chúng em xin gửi lời cảm ơn chân thành nhất đến:

❖ Giảng viên hướng dẫn chúng em: Cô Mai Thanh Hồng – người đã tạo điều kiện, định hướng và hỗ trợ chúng em xuyên suốt quá trình nghiên cứu. Cô không chỉ truyền đạt những kiến thức nền tảng về Trí tuệ nhân tạo mà còn hướng dẫn chúng em cách ứng dụng cây quyết định vào bài toán y tế cụ thể. Những góp ý chi tiết, sự nhiệt tình và tinh thần trách nhiệm của cô đã giúp nhóm khắc phục khó khăn, hoàn thiện bài báo cáo một cách tốt nhất.

❖ Các thành viên trong nhóm đã cùng nhau làm việc nghiêm túc, chia sẻ ý tưởng và hỗ trợ nhau trong quá trình thu thập dữ liệu, xây dựng mô hình cũng như viết báo cáo.

❖ Các nguồn tài liệu, nghiên cứu khoa học và bộ dữ liệu mở (như Pima Indians Diabetes Dataset) đã cung cấp cơ sở lý thuyết và dữ liệu thực tế để chúng em phát triển đề tài.

Chúng em mong nhận được sự đánh giá và góp ý từ cô để bài nghiên cứu được hoàn thiện hơn, đồng thời tích lũy kinh nghiệm cho những dự án tiếp theo.

# LỜI NÓI ĐẦU

Trong kỷ nguyên của cách mạng công nghiệp 4.0, Trí tuệ nhân tạo (AI) đã trở thành một trong những công nghệ đột phá nhất. AI đang dần khẳng định vai trò không thể thiếu trong hành trình phát triển của nhân loại. Đặc biệt, trong lĩnh vực chăm sóc sức khỏe, AI mang đến những giải pháp ưu việt giúp phát hiện sớm, dự đoán nguy cơ bệnh tật và hỗ trợ các quyết định y khoa quan trọng.

Bài báo cáo này được thực hiện với mục tiêu tìm hiểu sâu về thuật toán Cây quyết định – một trong những mô hình AI cơ bản nhưng mạnh mẽ, cùng ứng dụng của nó trong việc dự đoán sức khỏe.

Bệnh tiểu đường là một trong những căn bệnh mãn tính phổ biến nhất thế giới, gây ra gánh nặng lớn cho hệ thống y tế và chất lượng cuộc sống. Tại Việt Nam, tỷ lệ mắc bệnh đang gia tăng đáng báo động do lối sống ít vận động và chế độ ăn uống thiếu cân bằng. Việc ứng dụng AI để dự đoán sớm nguy cơ tiểu đường không chỉ giúp giảm chi phí điều trị mà còn mở ra hướng tiếp cận chủ động trong phòng ngừa bệnh.

Thông qua báo cáo này, chúng tôi hy vọng sẽ:

* Làm rõ cơ chế hoạt động của Cây quyết định trong bài toán phân loại y tế.
* Minh họa quy trình xử lý dữ liệu, huấn luyện mô hình và đo lường độ chính xác.
* Đề xuất hướng phát triển ứng dụng AI trong lĩnh vực chăm sóc sức khỏe tại Việt Nam, từ góc độ kỹ thuật đến chính sách.

Chúng tôi tin rằng, với sự phát triển không ngừng của AI cùng sự chung tay của cộng đồng y tế - công nghệ, những giải pháp như mô hình dự đoán tiểu đường sẽ góp phần hiện thực hóa mục tiêu "Y tế thông minh", nâng cao chất lượng cuộc sống người dân.

Mọi đóng góp ý kiến từ quý thầy cô và độc giả sẽ là động lực để nhóm hoàn thiện nghiên cứu này.

**MỤC LỤC**

[LỜI CẢM ƠN i](#_Toc200606764)

[LỜI NÓI ĐẦU ii](#_Toc200606765)

[DANH MỤC HÌNH VẼ v](#_Toc200606766)

[PHẦN A - MỞ ĐẦU 1](#_Toc200606767)

[PHẦN B: NỘI DUNG 5](#_Toc200606768)

[CHƯƠNG 1: TỔNG QUAN VỀ CÂY QUYẾT ĐỊNH 5](#_Toc200606769)

[1.1. Khái niệm và đặc điểm 5](#_Toc200606770)

[1.1.1. Định nghĩa 5](#_Toc200606771)

[1.1.2. Đặc điểm cấu trúc 5](#_Toc200606772)

[1.2. Các thuật toán cây quyết định phổ biến 6](#_Toc200606773)

[1.2.1. Thuật toán ID3 6](#_Toc200606774)

[1.2.2. Thuật toán C4.5 8](#_Toc200606775)

[1.3. Ưu điểm của thuật toán cây quyết định 10](#_Toc200606776)

[1.4. Nhược điểm của thuật toán cây quyết định 10](#_Toc200606777)

[CHƯƠNG 2: ỨNG DỤNG CÂY QUYẾT ĐỊNH TRONG Y TẾ 11](#_Toc200606778)

[2.1 Các bài toán dự đoán sức khỏe 11](#_Toc200606779)

[2.2 Case studies điển hình 12](#_Toc200606780)

[CHƯƠNG 3: THIẾT KẾ HỆ THỐNG DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG 15](#_Toc200606781)

[3.1 Tệp cơ sở dữ liệu 15](#_Toc200606782)

[3.1.1 Giới thiệu 15](#_Toc200606783)

[3.1.2 Cấu trúc tệp dữ liệu 15](#_Toc200606784)

[3.2 Môi trường và công cụ sử dụng 18](#_Toc200606785)

[3.2.1 Môi trường 19](#_Toc200606786)

[3.2.2 Công cụ và thư viện 19](#_Toc200606787)

[3.2.3 Công cụ phát triển 19](#_Toc200606788)

[3.3 Hệ thống sử dụng cây quyết định 20](#_Toc200606789)

[3.3.1 Phân tích dữ liệu 20](#_Toc200606790)

[3.3.2 Xử lý dữ liệu 21](#_Toc200606791)

[3.3.3: Xây dựng mô hình 22](#_Toc200606792)

[PHẦN C: KẾT LUẬN 28](#_Toc200606793)

[TÀI LIỆU THAM KHẢO 29](#_Toc200606794)

# DANH MỤC HÌNH VẼ

Hình 1.1: Mô hình cây quyết định dự đoán bệnh tim  [5](#_Toc200407962)

Hình 2.1: sự thay đổi của hàm entropy  [7](#_Toc200407962)

Hình 3.1: Dữ liệu dataset  [18](#_Toc200407962)

Hình 3.2: Xử lý giá trị bị thiếu [21](#_Toc200407962)

Hình 3.3: Hàm phân loại đặc trưng số [22](#_Toc200407962)

Hình 3.4: Các thư viện được sử dụng [23](#_Toc200407962)

Hình 3.5: Lệnh đọc dữ liệu [23](#_Toc200407962)

Hình 3.6: Lệnh gán nhãn đặc trưng và nhãn [23](#_Toc200407962)

Hình 3.7: Hàm tính Entropy [23](#_Toc200407962)

Hình 3.8: Hàm tính Gain [24](#_Toc200407962)

Hình 3.9: Hàm trả về giá trị Gain lớn nhất [24](#_Toc200407962)

Hình 3.10: Hàm xây dựng cây quyết định……………………………………..25

Hình 3.11: Huấn luyện mô hình cây quyết định [25](#_Toc200407962)

Hình 3.12: Tính độ chính xác của mô hình………………………………….....26

Hình 3.13: Ví dụ chẩn đoán bệnh nhân bất kì [26](#_Toc200407962)

Hình 3.14: Kết quả đánh giá của mô hình……………………………………...27

# PHẦN A - MỞ ĐẦU

**1. Lý do chọn đề tài**

Trong bối cảnh cuộc cách mạng công nghiệp 4.0, trí tuệ nhân tạo (AI) đang trở thành một trong những lĩnh vực then chốt, có tác động sâu rộng đến nhiều ngành nghề, đặc biệt là y tế. Việc ứng dụng AI vào chăm sóc sức khỏe không chỉ giúp nâng cao chất lượng chẩn đoán mà còn hỗ trợ dự phòng bệnh tật, giảm tải cho hệ thống y tế. Trong đó, cây quyết định là một trong những mô hình đơn giản nhưng hiệu quả, dễ triển khai trong các bài toán phân loại y khoa.

Nhận thấy tiềm năng ứng dụng thực tiễn của mô hình này, nhóm em lựa chọn đề tài "Tìm hiểu cây quyết định và ứng dụng dự đoán sức khỏe" với mong muốn nghiên cứu, làm rõ nguyên lý hoạt động của thuật toán, đồng thời xây dựng một hệ thống dự đoán nguy cơ bệnh lý dựa trên các yếu tố như tuổi tác, tiền sử bệnh án và thói quen sinh hoạt. Kết quả của đề tài không chỉ giúp củng cố kiến thức AI mà còn góp phần đề xuất giải pháp công nghệ thiết thực, hỗ trợ nâng cao chất lượng chăm sóc sức khỏe cộng đồng trong tương lai.

**2. Mục tiêu nghiên cứu**

Từ đề tài nghiên cứu này, chúng em mong muốn tìm hiểu rõ hơn về mô hình cây quyết định, từ lý thuyết đến ứng dụng thực tế trong lĩnh vực dự đoán sức khỏe. Qua đó, chúng em hy vọng có thể xây dựng được một hệ thống hỗ trợ đánh giá tình trạng sức khỏe dựa trên các yếu tố cơ bản như tuổi tác, tiền sử bệnh án và lối sống. Kết quả nghiên cứu không chỉ giúp củng cố kiến thức về trí tuệ nhân tạo mà còn góp phần tạo ra giải pháp công nghệ hữu ích, hỗ trợ nâng cao chất lượng chăm sóc sức khỏe cộng đồng. Đồng thời, đề tài cũng giúp chúng em và người đọc có cái nhìn rõ ràng hơn về tiềm năng ứng dụng của AI trong y tế, từ đó mở ra hướng phát triển mới cho các nghiên cứu sau này.

**3. Đối tượng nghiên cứu**

Như mục tiêu nghiên cứu đã đề cập, đối tượng chính của đề tài này là mô hình cây quyết định và ứng dụng cụ thể của nó trong lĩnh vực dự đoán sức khỏe. Nghiên cứu tập trung vào việc phân tích nguyên lý hoạt động, cách thức xây dựng và triển khai mô hình này để đánh giá các yếu tố ảnh hưởng đến sức khỏe như tuổi tác, tiền sử bệnh lý và thói quen sinh hoạt. Đồng thời, đề tài cũng hướng đến việc đánh giá tiềm năng ứng dụng thực tế của mô hình trong lĩnh vực y tế, góp phần hỗ trợ công tác chẩn đoán và phòng ngừa bệnh tật.

**4. Phạm vi nghiên cứu**

\*Không gian:

Nghiên cứu tập trung vào việc phân tích mô hình cây quyết định và ứng dụng cụ thể trong lĩnh vực dự đoán sức khỏe. Đề tài xem xét các yếu tố ảnh hưởng đến sức khỏe như tuổi tác, tiền sử bệnh lý, thói quen sinh hoạt và mối quan hệ của chúng với kết quả dự đoán từ mô hình.

\*Thời gian:

Nghiên cứu được thực hiện trong khoảng thời gian từ ngày [X] đến ngày [Y], bao gồm các giai đoạn: thu thập dữ liệu, phân tích mô hình, xây dựng hệ thống dự đoán và đánh giá kết quả.

**5. Kết quả mong muốn đạt được**

- Đề tài sẽ cung cấp một cái nhìn toàn diện về mô hình cây quyết định, từ lý thuyết cơ bản đến ứng dụng thực tiễn trong lĩnh vực dự đoán sức khỏe, giúp người đọc hiểu rõ nguyên lý hoạt động và giá trị của mô hình này trong y học.

- Kết quả nghiên cứu sẽ xây dựng được một hệ thống dự đoán sức khỏe dựa trên cây quyết định, có khả năng phân tích các yếu tố nguy cơ như tuổi tác, tiền sử bệnh lý và lối sống, từ đó hỗ trợ đưa ra các cảnh báo sớm về tình trạng sức khỏe.

- Đề tài sẽ góp phần nâng cao nhận thức về tiềm năng ứng dụng của trí tuệ nhân tạo trong lĩnh vực y tế, đồng thời đề xuất hướng phát triển cho các nghiên cứu tiếp theo nhằm tối ưu hóa hiệu quả của mô hình trong thực tế.

- Kết quả nghiên cứu cũng sẽ là tài liệu tham khảo hữu ích cho sinh viên và những ai quan tâm đến lĩnh vực AI ứng dụng trong chăm sóc sức khỏe, mở ra những ý tưởng mới để phát triển các giải pháp công nghệ phục vụ cộng đồng.

**6. Phương pháp nghiên cứu**

\*Phương pháp thu thập tài liệu:

+ Tìm kiếm và tổng hợp các tài liệu khoa học, sách báo, bài báo nghiên cứu liên quan đến cây quyết định và ứng dụng trong y tế

+ Thu thập các bộ dữ liệu sức khỏe từ các nguồn uy tín để phục vụ cho việc xây dựng mô hình

\*Phương pháp phân tích tài liệu:

+ Phân tích, đánh giá các thuật toán cây quyết định và các biến thể của nó

+ Nghiên cứu các ứng dụng thực tế của mô hình trong lĩnh vực dự đoán sức khỏe

+ Đánh giá ưu nhược điểm của phương pháp so với các mô hình AI khác

\*Phương pháp tổng hợp và trình bày:

+ Hệ thống hóa kiến thức về lý thuyết cây quyết định

+ Trình bày kết quả nghiên cứu dưới dạng báo cáo khoa học

+ Minh họa bằng các biểu đồ, sơ đồ quyết định và kết quả thử nghiệm cụ thể

\*Phương pháp thực nghiệm:

+ Xây dựng mô hình dự đoán sức khỏe bằng cây quyết định

+ Đánh giá hiệu quả mô hình thông qua các chỉ số như: độ chính xác, recall, F1-score

+ So sánh kết quả với các phương pháp truyền thống

**7. Cấu trúc bài báo cáo**

- PHẦN A: MỞ ĐẦU

- PHẦN B: NỘI DUNG CHÍNH

+ Chương I: Tổng quan về cây quyết định

+ Chương II: Ứng dụng của cây quyết định trong y tế

+ Chương III: Thiết kế hệ thống dự đoán bệnh tiểu đường

- PHẦN C: KẾT LUẬN

**8. Ý nghĩa bài báo cáo**

Bài báo cáo này không chỉ cung cấp cái nhìn toàn diện về mô hình cây quyết định và ứng dụng của nó trong lĩnh vực dự đoán sức khỏe, mà còn khẳng định giá trị thực tiễn của trí tuệ nhân tạo trong việc nâng cao chất lượng chăm sóc y tế. Nghiên cứu góp phần làm rõ khả năng ứng dụng của mô hình trong việc phân tích các yếu tố nguy cơ sức khỏe, từ đó hỗ trợ quá trình chẩn đoán sớm và phòng ngừa bệnh tật.

Bên cạnh đó, báo cáo còn đề xuất các hướng phát triển cụ thể để tối ưu hóa hiệu quả của mô hình, như cải thiện độ chính xác, mở rộng phạm vi ứng dụng và tích hợp với các công nghệ tiên tiến khác. Những kết quả đạt được không chỉ có ý nghĩa đối với các nhà nghiên cứu, sinh viên trong lĩnh vực AI, mà còn là tài liệu tham khảo hữu ích cho các chuyên gia y tế và nhà hoạch định chính sách.

Cuối cùng, báo cáo nhấn mạnh tầm quan trọng của việc phát triển các giải pháp công nghệ trong lĩnh vực sức khỏe, góp phần thúc đẩy sự kết hợp giữa khoa học dữ liệu và y học, hướng tới mục tiêu nâng cao chất lượng cuộc sống và sức khỏe cộng đồng một cách bền vững.

# PHẦN B: NỘI DUNG

# CHƯƠNG 1: TỔNG QUAN VỀ CÂY QUYẾT ĐỊNH

## 1.1. Khái niệm và đặc điểm

### 1.1.1. Định nghĩa

Cây quyết định (Decision Tree) là một mô hình học máy thuộc nhóm học có giám sát, dùng để giải quyết cả bài toán phân loại (classification) và hồi quy (regression). Mô hình này xây dựng một cấu trúc cây phân cấp, trong đó:

+ Nút gốc (root node): Đại diện cho toàn bộ tập dữ liệu.

+ Nút nhánh (internal node): Biểu diễn các thuộc tính/quyết định.

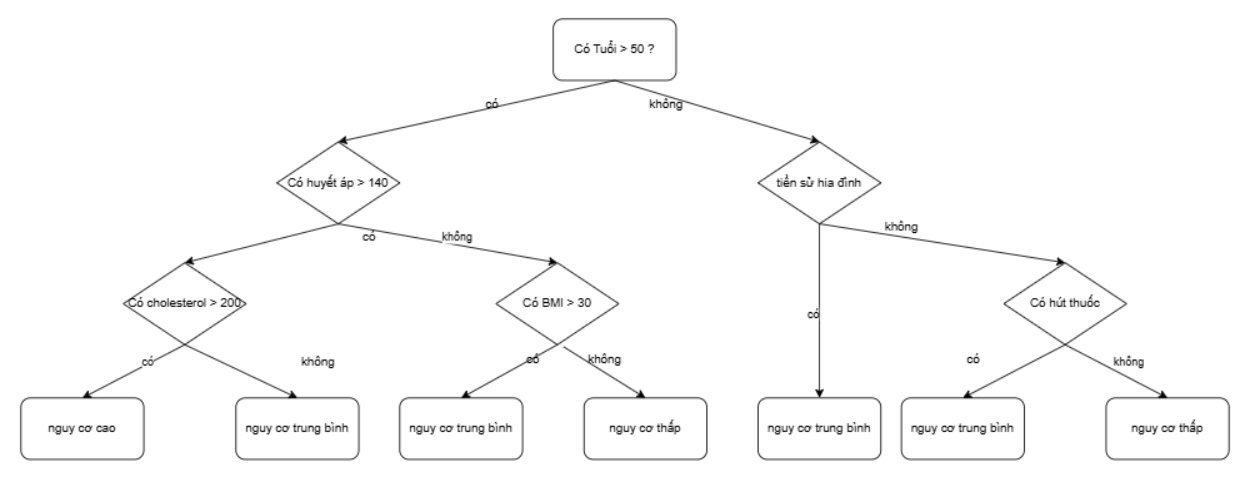
+ Nút lá (leaf node): Kết quả cuối cùng (lớp hoặc giá trị dự đoán).

### 1.1.2. Đặc điểm cấu trúc

- Phân chia đệ quy: Mỗi nút được chia thành các nhánh con dựa trên ngưỡng quyết định (threshold).

- Hướng tiếp cận Top-Down: Bắt đầu từ nút gốc, đi xuống các nút lá.

- Không yêu cầu chuẩn hóa dữ liệu: Do sử dụng các quy tắc phân chia trực tiếp.



Hình 1.1: Mô hình cây quyết định dự đoán bệnh tim

## 1.2. Các thuật toán cây quyết định phổ biến

### 1.2.1. Thuật toán ID3

**Mục tiêu**

Thuật toán ID3 được sử dụng để xây dựng cây quyết định (Decision Tree) nhằm phân loại các đối tượng dựa trên dữ liệu huấn luyện. Đây là thuật toán học có giám sát (supervised learning) phổ biến trong lĩnh vực trí tuệ nhân tạo.

**Đầu vào và đầu ra**

- Đầu vào:

+ Tập dữ liệu huấn luyện D, gồm các mẫu thuộc nhiều lớp.

+ Tập các thuộc tính đầu vào A={A1,A2,...,An}

- Đầu ra:

+ Một cây quyết định giúp phân loại dữ liệu mới dựa trên thuộc tính đầu vào.

**Nguyên lý hoạt động**

ID3 sử dụng các khái niệm Entropy và Information Gain để chọn ra thuộc tính tốt nhất tại mỗi nút phân tách. Thuật toán xây dựng cây theo cách đệ quy từ gốc đến các nút lá.

a.Entropy trong Cây quyết định (Decision Tree)

- Entropy là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với công thức như sau:

- Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1,x2,…,xn.

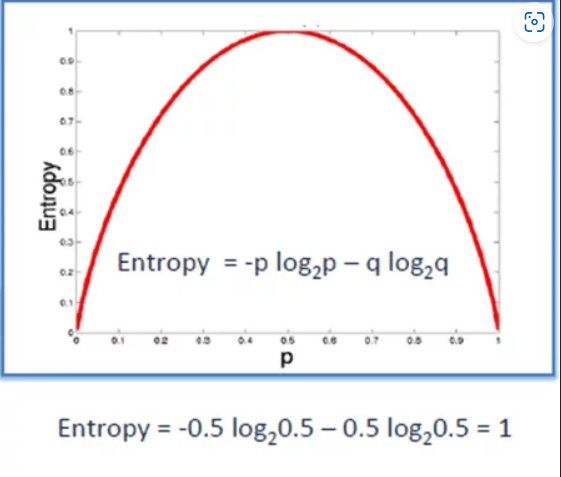
- Giả sử rằng xác suất để x nhận các giá trị này là pi=p(x=xi).

- Ký hiệu phân phối này là p=(p1 ,p2 ,…,pn). Entropy của phân phối này được định nghĩa là:

H(p)= – ∑nn=1 pi log(pi)

Giả sử bạn tung một đồng xu, entropy sẽ được tính như sau:

H = -[0.5 ln(0.5) + 0.5 ln(0.5)]



Hình 1.2: sự thay đổi của hàm entropy

Ta có thể thấy rằng, entropy đạt tối đa khi xác suất xảy ra của hai lớp bằng nhau.

P tinh khiết: pi = 0 hoặc pi = 1

P vẩn đục: pi = 0.5, khi đó hàm Entropy đạt đỉnh cao nhất

b.Information Gain – Lượng thông tin đạt được khi chia

Information Gain dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất.

Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:

•Bước 1: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với Nc phần tử thuộc lớp c cho trước:

H(S)= – ∑cc=1(Nc/N) log(Nc/N)

•Bước 2: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S1, S2, …, SK với số điểm trong mỗi child node lần lượt là m1, m2 ,…, mK , ta có:

H(x, S) = ∑kk=1 (mk / N) \* H(Sk)

Bước 3: Chỉ số Gain Information được tính bằng:

G(x, S) = H(S) – H(x,S)

### 1.2.2. Thuật toán C4.5

**Mục tiêu**

Trong thuật toán ID3, Information Gain được sử dụng làm độ đo. Tuy nhiên, phương pháp này lại ưu tiên những thuộc tính có số lượng lớn các giá trị mà ít xét tới những giá trị nhỏ hơn. Do vậy, để khắc phục nhược điểm trên, ta sử dụng độ đo Gain Ratio (trong thuật toán C4.5

**Đầu ra và đầu vào**

Giống ID3 nhưng C4.5 linh hoạt hơn, mở rộng thêm khả năng xử lý thuộc tính liên tục, giá trị thiếu, và giảm bias thông qua Gain Ratio.

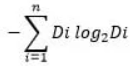
**Nguyên lí hoạt động**

- C4.5 xây dựng cây quyết định bằng cách chọn thuộc tính “tốt nhất” để phân chia dữ liệu tại mỗi nút, dựa trên tiêu chí Gain Ratio. Cây được xây theo hướng đệ quy từ trên xuống, và có thể dừng lại hoặc được cắt tỉa (pruning) để tránh overfitting.

- Đầu tiên, ta chuẩn hoá information gain với trị thông tin phân tách (split information):



Trong đó: Split Info được tính như sau:



- Giả sử chúng ta phân chia biến thành n nút cón và Di đại diện cho số lượng bản ghi thuộc nút đó. Do đó, hệ số Gain Ratio sẽ xem xét được xu hướng phân phối khi chia cây.

- Áp dụng cho ví dụ trên và với cách chia thứ nhất, ta có

Split Info = – ((4/7)\*log2(4/7)) – ((3/7)\*log2(3/7)) = 0.98

Gain Ratio = 0.09/0.98 = 0.092

- Tiêu chuẩn dừng

Trong các thuật toán Decision tree, với phương pháp chia trên, ta sẽ chia mãi các node nếu nó chưa tinh khiết. Như vậy, ta sẽ thu được một tree mà mọi điểm trong tập huấn luyện đều được dự đoán đúng (giả sử rằng không có hai input giống nhau nào cho output khác nhau). Khi đó, cây có thể sẽ rất phức tạp (nhiều node) với nhiều leaf node chỉ có một vài điểm dữ liệu. Như vậy, nhiều khả năng overfitting sẽ xảy ra.

- Để tránh trường họp này, ta có thể dừng cây theo một số phương pháp sau đây:

+ nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một class.

+ nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.

+ nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế chiều sâu của tree này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.

+ nếu tổng số leaf node vượt quá một ngưỡng nào đó.

+ nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

## 1.3. Ưu điểm của thuật toán cây quyết định

+ 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 là việc với dữ liệu lớn

## 1.4. Nhược điểm của thuật toán cây quyết định

+ Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm 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 overfitting(Là hiện tượng mô hình ghi nhớ quá tốt dữ liệu huấn luyện và phụ thuộc vào nó, việc này khiến cho mô hình không thể tổng quát hóa các quy luật để hoạt động với dữ liệu chưa từng được chứng kiến)

# CHƯƠNG 2: ỨNG DỤNG CÂY QUYẾT ĐỊNH TRONG Y TẾ

## 2.1 Các bài toán dự đoán sức khỏe

Cây quyết định là một công cụ hữu hiệu để giải quyết các bài toán dự đoán sức khỏe, giúp các chuyên gia y tế đưa ra dự đoán về tình trạng sức khỏe của bệnh nhân dựa trên dữ liệu đầu vào. Một số ứng dụng cụ thể bao gồm:

- Dự đoán nguy cơ mắc bệnh: Cây quyết định trong lĩnh vực dự đoán nguy cơ bệnh lý tích hợp phân tích đa yếu tố để định lượng khả năng phát triển các bệnh lý. Quy trình này dựa trên việc xử lý khối dữ liệu y khoa lớn nhằm thiết lập các mối tương quan thống kê giữa đặc điểm cá nhân và kết cục sức khỏe. Hệ thống tự động phân nhánh dựa trên các ngưỡng giá trị quan trọng, tạo ra cấu trúc phân cấp các điều kiện quyết định, từ đó phân tầng bệnh nhân vào các nhóm nguy cơ.

- Chẩn đoán bệnh: Trong quy trình chẩn đoán lâm sàng, phương pháp phân nhánh quyết định cung cấp khung logic để phân tích có hệ thống các thông số y khoa. Cơ chế hoạt động dựa trên việc tích hợp đồng thời nhiều nguồn thông tin bao gồm biểu hiện lâm sàng, chỉ số sinh hóa và yếu tố bệnh sử. Mỗi điểm nút trong cấu trúc thuật toán đại diện cho một quyết định phân loại dựa trên tiêu chí cụ thể, dẫn đến việc thu hẹp dần không gian chẩn đoán. Phương pháp này đặc biệt phát huy hiệu quả với các tình trạng bệnh lý có đặc điểm minh định và thường xuyên xuất hiện.

- Hỗ trợ chẩn đoán hình ảnh: Trong lĩnh vực chẩn đoán hình ảnh, Cây quyết định được sử dụng để đánh giá các thuộc tính định lượng và định tính từ dữ liệu hình ảnh, bao gồm các thông số về hình thái, cấu trúc mật độ và mối tương quan không gian. Cho phép phân loại các phát hiện theo mức độ quan trọng lâm sàng. Phương pháp này phát hiện được các mô hình bất thường mà con người không thể dễ dàng nhận biết được, tăng cường hiệu suất trong quy trình đánh giá hình ảnh và nâng cao độ tin cậy trong việc đưa ra các phát hiện sớm về bệnh. Điểm mạnh của cây quyết định nằm ở tính trực quan, giúp cả bác sĩ và bệnh nhân dễ dàng hiểu được logic đằng sau các dự đoán.

## 2.2 Case studies điển hình

Các nghiên cứu cho thấy cách cây quyết định được áp dụng thực tế trong y tế:

- Chẩn đoán bệnh: Trong trường hợp nghi ngờ viêm ruột thừa, cây quyết định có thể phân tích các triệu chứng như đau bụng dưới, sốt, và buồn nôn để đưa ra kết luận ban đầu, hỗ trợ quyết định phẫu thuật.

- Lựa chọn phương pháp điều trị: Với bệnh nhân ung thư, cây quyết định có thể xem xét giai đoạn bệnh, tuổi tác, và tình trạng sức khỏe để đề xuất phương án điều trị tốt nhất, như phẫu thuật, hóa trị, hay xạ trị.

Cây quyết định không chỉ hỗ trợ các chuyên gia y tế mà còn mang lại lợi ích cho cộng đồng, đặc biệt ở những khu vực thiếu thốn nguồn lực y tế. Tuy nhiên, để đạt độ chính xác cao nhất, việc sử dụng cây quyết định cần được kết hợp với đánh giá lâm sàng của bác sĩ hoặc các xét nghiệm chuyên sâu.

Dưới đây là hai case studies điển hình minh họa cách cây quyết định được áp dụng trong thực tế:

- Case Study 1: Dự đoán bệnh đái tháo đường

+ Mục đích: Xây dựng một hệ thống dự đoán nguy cơ mắc bệnh đái tháo đường nhằm hỗ trợ phát hiện sớm, đặc biệt tại các khu vực thiếu thiết bị y tế hiện đại như vùng sâu, vùng xa ở Việt Nam.

+ Dữ liệu sử dụng: Hệ thống khai thác thông tin từ nhiều thuộc tính của bệnh nhân, bao gồm:

Thông tin cá nhân: Tuổi (phân loại: <50, 50-54, 55-59, 60-64, >64), giới tính (nữ, nam), chỉ số BMI (<18, 18-22, 23-27, >27), vòng eo (<60, 60-70, 71-80, 81-90, >90 cm), vòng hông (<80, 80-90, 91-100, >100 cm).

Tiền sử gia đình: Bố mẹ, anh chị em ruột, hoặc ông bà nội có mắc bệnh đái tháo đường hay không (Có/Không).

Tiền sử bệnh lý: Bệnh nhân có mắc bệnh huyết áp (Có/Không), tiền sử đột quỵ, suy tim, loét bàn chân, hoặc các biến chứng khác (Có/Không).

Các yếu tố khác: Tình trạng mang thai (Đã/Chưa), cân nặng của con khi sinh (<1800, 1800-3600, >3600 gram).

- Phương pháp:

+ Sử dụng công cụ Business Intelligence Development Studio (BIDS) của Microsoft để xây dựng mô hình cây quyết định.

+ Dữ liệu được phân tích và phân loại dựa trên các thuộc tính đầu vào để tạo ra các nhánh quyết định.

- Kết quả:

+ Hệ thống phân loại bệnh nhân thành ba nhóm:

+ Bình thường (BT): Không có nguy cơ mắc bệnh.

+ Tiền đái tháo đường (TienDTD): Nguy cơ trung gian, cần theo dõi thêm.

+ Đái tháo đường (DTD): Nguy cơ cao hoặc đã mắc bệnh.

Ví dụ: Một bệnh nhân nữ, 55 tuổi, BMI >27, vòng eo >90 cm, và có tiền sử gia đình mắc đái tháo đường sẽ được hệ thống dự đoán thuộc nhóm DTD, từ đó bác sĩ có thể đưa ra các biện pháp can thiệp kịp thời như thay đổi lối sống hoặc điều trị y tế.

- Ý nghĩa:

+ Hệ thống này giúp các bác sĩ và nhân viên y tế tại các khu vực hạn chế về thiết bị chẩn đoán đưa ra quyết định nhanh chóng và chính xác hơn, từ đó cải thiện khả năng phát hiện sớm và quản lý bệnh đái tháo đường.

-Case Study 2: Chẩn đoán bệnh tự kỷ ở trẻ em

Mục đích: Phát triển một hệ thống chẩn đoán sơ bộ bệnh tự kỷ ở trẻ em dựa trên các biểu hiện hành vi và phát triển, hỗ trợ cha mẹ, giáo viên và chuyên gia y tế đánh giá nguy cơ.

Dữ liệu sử dụng:

Giao tiếp bằng lời (K): Mức độ bất thường trong khả năng sử dụng ngôn ngữ của trẻ.

Giao tiếp không lời (L: Mức độ bất thường trong cử chỉ, ánh mắt, hoặc các hình thức giao tiếp phi ngôn ngữ.

Mức độ hoạt động (M): Đánh giá sự bất thường trong hành vi vận động hoặc tương tác xã hội.

Đáp ứng trí tuệ (N): Mức độ phản ứng và nhận thức của trẻ với môi trường xung quanh.

# CHƯƠNG 3: THIẾT KẾ HỆ THỐNG DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG

## 3.1 Tệp cơ sở dữ liệu

### **3.1.1 Giới thiệu**

Tập dữ liệu (Dataset) đóng vai trò cực kỳ quan trọng trong quá trình phát triển và huấn luyện mô hình học máy và học sâu. Chất lượng của dữ liệu là yếu tố hàng đầu; một dataset chất lượng cao với dữ liệu chính xác và có liên quan là cần thiết để xây dựng mô hình hiệu quả.

Ngoài ra, dữ liệu cần phải đủ đại diện cho các tình huống mà mô hình sẽ gặp phải trong thực tế, giúp mô hình tổng quát tốt hơn. Kích thước của dataset cũng rất quan trọng; một dataset lớn thường giúp mô hình học được nhiều đặc trưng hơn và cải thiện độ chính xác, nhưng cần đảm bảo rằng dữ liệu vẫn giữ được chất lượng.

Hơn nữa, datasets cần cung cấp đủ thông tin để mô hình có thể học được các mối quan hệ và đặc trưng cần thiết, đồng thời không có thiếu sót nào có thể gây nhầm lẫn cho mô hình. Việc phân chia dữ liệu thành các tập huấn luyện, kiểm tra và xác thực là cần thiết để đánh giá hiệu suất của mô hình trên dữ liệu chưa thấy. Kỹ thuật tăng cường dữ liệu cũng có thể được sử dụng để tạo ra các phiên bản khác nhau của dữ liệu gốc, nâng cao tính đa dạng và khả năng tổng quát của mô hình.

Trong đề tài này, tập dữ liệu được sử dụng là Pima Indians Diabetes Database do Viện Y tế Quốc gia Hoa Kỳ cung cấp. Tập dữ liệu bao gồm các thông tin y tế được thu thập từ phụ nữ người da đỏ Pima, với các đặc trưng liên quan đến bệnh tiểu đường như: số lần mang thai, chỉ số BMI, huyết áp, mức insulin, độ dày da, mức đường huyết, tuổi và kết quả chẩn đoán (có hoặc không mắc bệnh).

### **3.1.2 Cấu trúc tệp dữ liệu**

Cơ sở dữ liệu của đề tài này được tổ chức như sau:

-Nguồn gốc: Tập dữ liệu này được thu thập từ National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

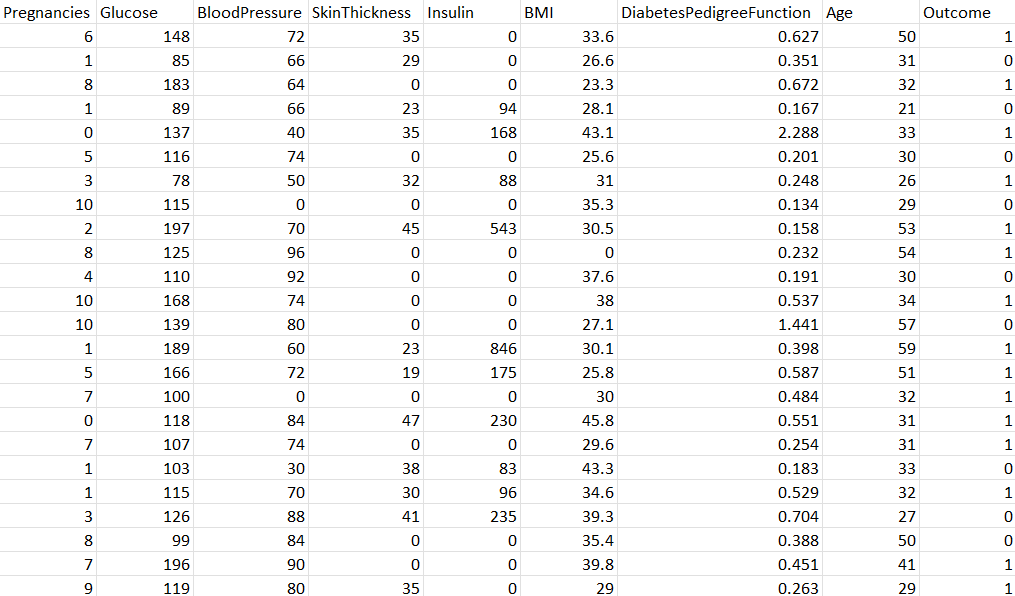
-Số lượng mẫu: 768 phụ nữ có độ tuổi từ 21 trở lên, thuộc dân tộc Pima Indian.

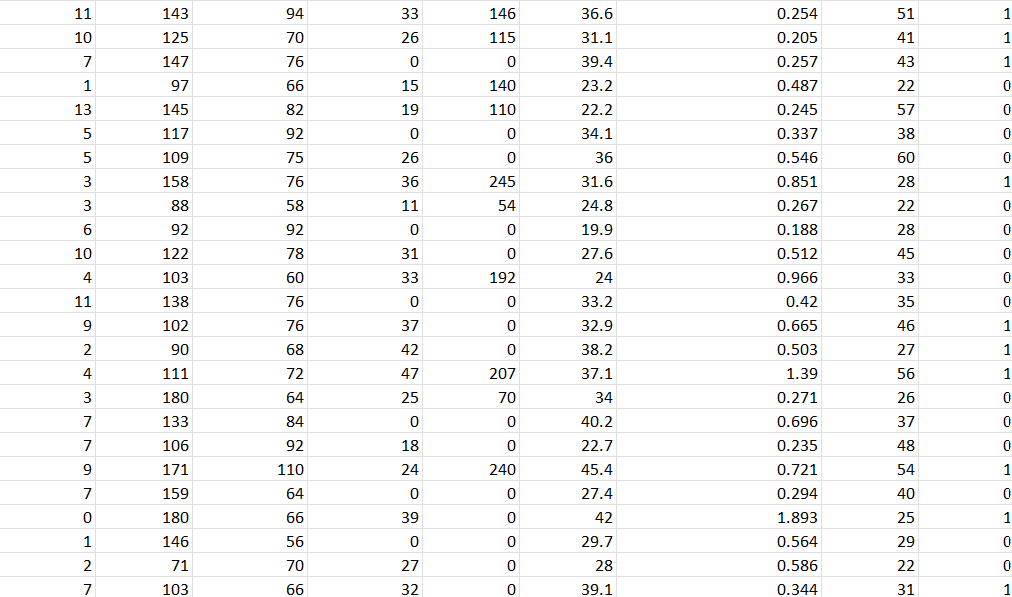
-Số lượng đặc trưng: 8 đặc trưng (biến độc lập) và 1 biến mục tiêu (Outcome).

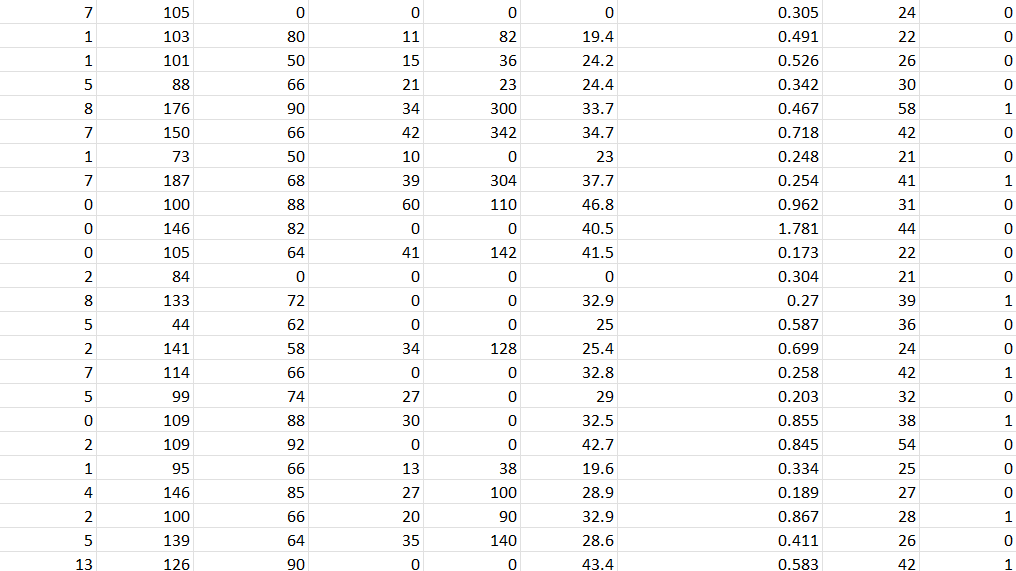
-Tên các biến: Pregnancies, Glucose,BloodPressure, SkinThickness, Insulin, BMI ,DiabetesPedigreeFunctionAge, Outcome (biến mục tiêu).

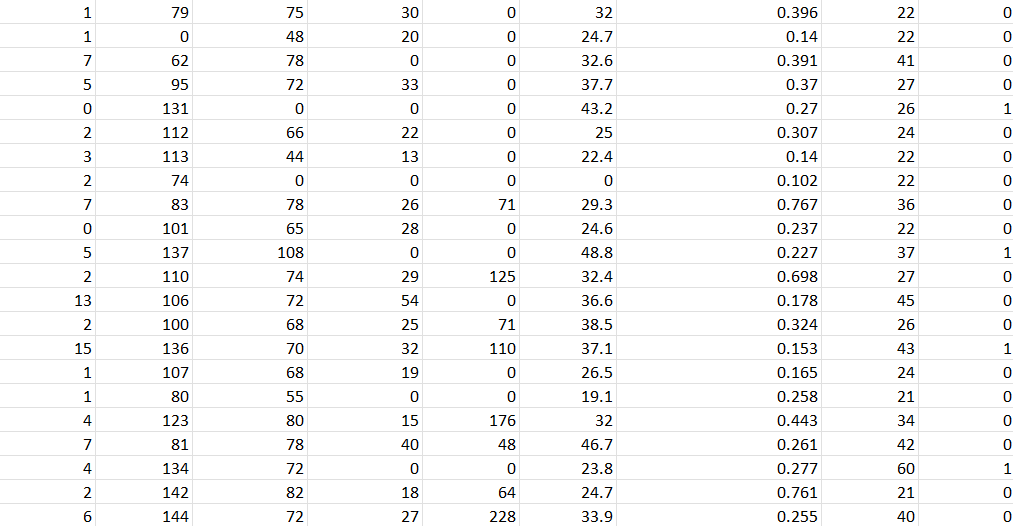
-Biến mục tiêu (Outcome): Biến nhị phân, với giá trị 0 (không mắc tiểu đường) và 1 (mắc tiểu đường)

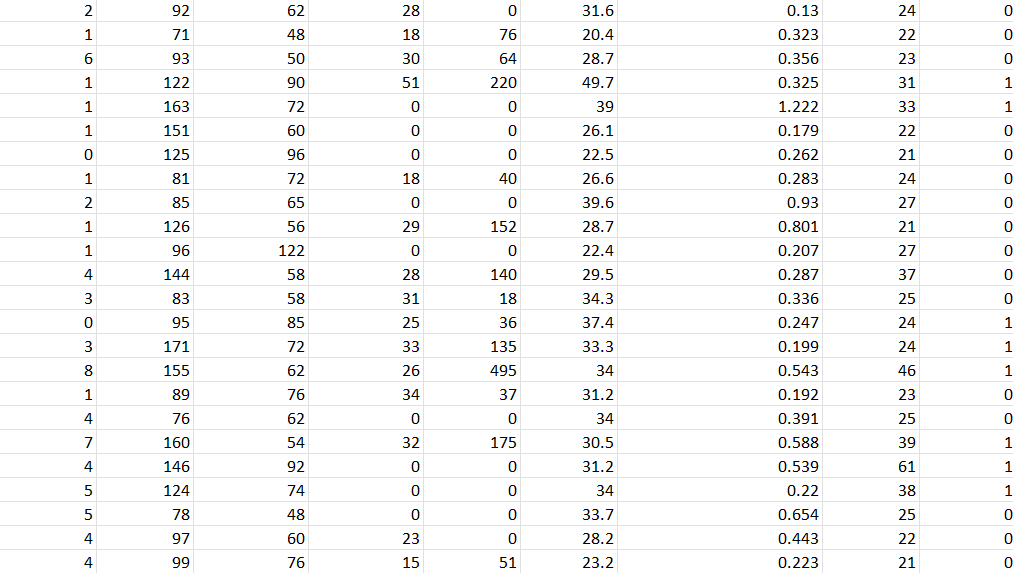
Bảng được hiển thị chính là tập dữ liệu chính, chứa thông tin về các đặc trưng y tế của một trong số 768 phụ nữ Pima Indian, nhằm mục tiêu dự đoán khả năng mắc bệnh tiểu đường theo dạng bảng.

****

****

****

****

****

Hình 3.1: Dữ liệu dataset

## 3.2 Môi trường và công cụ sử dụng

Trong đề tài chẩn đoán bệnh tiểu đường, việc lựa chọn môi trường và công cụ phát triển là rất quan trọng để đảm bảo chất lượng và hiệu suất của mô hình học máy. Môi trường này sẽ hỗ trợ trong việc xử lý dữ liệu, xây dựng mô hình, huấn luyện và đánh giá hiệu suất của mô hình

### **3.2.1 Môi trường**

Hệ điều hành: Có thể sử dụng Windows, macOS hoặc Linux. Trong lĩnh vực học máy, Linux thường được ưa chuộng do tính ổn định và khả năng tương thích tốt với các công cụ mã nguồn mở. Python: Python là ngôn ngữ lập trình chính được sử dụng trong dự án này nhờ vào cú pháp đơn giản, tính linh hoạt và sự hỗ trợ mạnh mẽ từ cộng đồng. Python là lựa chọn hàng đầu cho các ứng dụng học máy và khoa học dữ liệu.

Python là ngôn ngữ lập trình chính được sử dụng trong dự án này nhờ vào cú pháp đơn giản, tính linh hoạt và sự hỗ trợ mạnh mẽ từ cộng đồng. Python là lựa chọn hàng đầu cho các ứng dụng học máy và khoa học dữ liệu

### **3.2.2 Công cụ và thư viện**

− Pandas: Pandas là thư viện xử lý dữ liệu dạng bảng, được sử dụng để đọc, kiểm tra thông tin dữ liệu và làm sạch dữ liệu. Thư viện này rất hữu ích trong việc chuẩn bị dữ liệu đầu vào cho mô hình, đặc biệt là trong các tác vụ như phân tích hình ảnh võng mạc để phát hiện biến chứng tiểu đường.

− NumPy: NumPy là thư viện quan trọng cho xử lý mảng và tính toán số học, giúp quản lý dữ liệu hình ảnh và thực hiện các phép toán ma trận cần thiết cho quá trình xử lý và phân tích dữ liệu.

− Scikit-learn: Scikit-learn cung cấp các công cụ để phân chia và chuẩn hóa tập dữ liệu, cũng như các chức năng đánh giá mô hình. Thư viện này giúp dễ dàng thực hiện các tác vụ phân tích và xử lý dữ liệu.

### **3.2.3 Công cụ phát triển**

PyCharm là một môi trường phát triển tích hợp (IDE) dành cho Python, được phát triển bởi JetBrains. IDE này cung cấp các tính năng như gợi ý mã, kiểm tra lỗi cú pháp, và tích hợp với Git, giúp quản lý mã nguồn hiệu quả và hỗ trợ phát triển các dự án phân loại tin nhắn.

## 3.3 Hệ thống sử dụng cây quyết định

### **3.3.1 Phân tích dữ liệu**

-Bộ dữ liệu hiện có bao gồm 768 bản ghi, mỗi bản ghi có 8 đặc trưng và một biến kết quả nhị phân. Các đặc trưng chủ yếu là số, nhưng để phù hợp với thuật toán ID3, chúng sẽ được chuyển đổi thành các biến phân loại.

-Một số đặc trưng như Glucose, BloodPressure, SkinThickness, Insulin và BMI chứa các giá trị bằng 0, có thể biểu thị dữ liệu bị thiếu. Những giá trị này cần được xử lý trước khi huấn luyện mô hình. Bảng thống kê cơ bản của bộ dữ liệu được trình bày như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Đặc trưng** | **Trung bình** | **Trung vị** | **Tối thiểu** | **Tối đa** |
| Pregnancies | 3.85 | 3 | 0 | 17 |
| Glucose | 120.89 | 117 | 0 | 199 |
| BloodPressure | 69.11 | 72 | 0 | 122 |
| SkinThickness | 20.54 | 23 | 0 | 99 |
| Insulin | 79.80 | 30.5 | 0 | 846 |
| BMI | 31.99 | 32 | 0 | 67.1 |
| DiabetesPedigreeFunction | 0.4719 | 0.3725 | 0.078 | 2.42 |
| Age | 33.24 | 29 | 21 | 81 |
| Outcome | 0.348 | 0 | 0 | 1 |

Biến Outcome cho thấy khoảng 34,9% bệnh nhân mắc bệnh tiểu đường (Outcome=1), trong khi 65,1% không mắc (Outcome=0), cho thấy sự mất cân bằng nhẹ giữa hai lớp.

### **3.3.2 Xử lý dữ liệu**

-Xử lý dữ liệu bị thiếu: Các giá trị bằng 0 trong các cột Glucose, BloodPressure, SkinThickness, Insulin và BMI được thay thế bằng giá trị trung vị của từng cột tương ứng, nhằm giảm thiểu ảnh hưởng của các giá trị ngoại lai và đảm bảo tính toàn vẹn của dữ liệu.

A screen shot of a computer

Description automatically generated

Hình 3.2: Xử lý giá trị bị thiếu

-Phân loại đặc trưng số: Các đặc trưng được chuyển đổi thành các biến phân loại như sau:

+Pregnancies: Được phân loại thành '0', '1-2', '3-5', '6+' dựa trên số lần mang thai.

+Glucose: 'Normal' (<140), 'Prediabetes' (140-199), 'Diabetes' (>=200), dựa trên ngưỡng xét nghiệm dung nạp glucose đường uống trong 2 giờ.

+BloodPressure: 'Normal' (<80), 'Prehypertension' (80-89), 'Hypertension' (>=90), dựa trên huyết áp tâm trương.

+BMI: 'Underweight' (<18.5), 'Normal' (18.5-24.9), 'Overweight' (25-29.9), 'Obese' (>=30), dựa trên phân loại BMI tiêu chuẩn.

+Age: '20-29', '30-39', '40-49', '50-59', '60+', dựa trên nhóm tuổi.

+SkinThickness, Insulin, DiabetesPedigreeFunction: Được phân loại thành 'Low', 'Medium', 'High' bằng cách sử dụng phương pháp phân chia tần suất đều, với mỗi danh mục chứa khoảng một phần ba số điểm dữ liệu. Cụ thể, dữ liệu được sắp xếp, và các ngưỡng được xác định tại phân vị 33% và 66% để chia thành ba nhóm.

A computer screen with colorful text

Description automatically generated

Hình 3.3: Hàm phân loại đặc trưng số

### **3.3.3: Xây dựng mô hình**

Mô hình cây quyết định ID3 được xây dựng để phân loại khả năng mắc bệnh tiểu đường dựa trên các thông số y tế như chỉ số đường huyết, huyết áp, tuổi, v.v. Thuật toán ID3 là một phương pháp học có giám sát, hoạt động bằng cách chia tách dữ liệu dựa trên Entropy và Thông tin thu được (Information Gain) tại mỗi nút để xây dựng cấu trúc cây.

**Kiến trúc mô hình:**

Cây quyết định ID3 không có các tầng (layers) như mạng nơ-ron, mà bao gồm các nút quyết định (decision node) và nút lá (leaf node). Mỗi nút quyết định kiểm tra một thuộc tính (feature), và chia nhánh theo các giá trị có thể có của thuộc tính đó. Mô hình hoạt động như sau:

+ Gốc cây (Root node): là thuộc tính có thông tin thu được (information gain) cao nhất.

+ Nút quyết định (Decision Node): thực hiện phép kiểm tra trên một thuộc tính.

+ Nút lá (Leaf Node): chứa nhãn phân loại đầu ra, ví dụ: "Có bệnh", "Không bệnh".

Entropy và Information Gain là hai khái niệm trung tâm của ID3:

+ Entropy đo lường độ hỗn loạn của tập dữ liệu.

+ Information Gain đo mức độ giảm entropy sau khi phân chia dữ liệu theo một thuộc tính.

**Nhập các thư viện cần sử dụng**

A black background with white text

Description automatically generated

Hình 3.4: Các thư viện được sử dụng

**Tiến hành đọc dữ liệu**

A black background with white text

Description automatically generated

Hình 3.5: Lệnh đọc dữ liệu

-Đọc file CSV chứa dữ liệu bệnh tiểu đường

**Gán nhãn đặc trưng và nhãn**

A black background with green text

Description automatically generated

Hình 3.6: Lệnh gán nhãn đặc trưng và nhãn

features: Danh sách các **đặc trưng đầu vào** (biến độc lập).

target: Biến mục tiêu (Outcome) cho biết **có bị tiểu đường không (1) hay không (0)**.

**Hàm tính Entropy**

A screen shot of a computer

Description automatically generated

Hình 3.7: Hàm tính Entropy

-Hàm entropy tính toán độ hỗn loạn (entropy) của tập dữ liệu.

-Mục đích: Đo lường mức độ không đồng nhất của dữ liệu

**Hàm tính Gain**

A computer screen shot of a black screen

Description automatically generated

Hình 3.8: Hàm tính Gain

-Hàm info\_gain tính toán mức độ giảm entropy khi phân chia dữ liệu dựa trên một đặc trưng feature.

**Hàm trả về thuộc tính có Gain lớn nhất và xử lý trường hợp kết thúc**

A screen shot of a computer code

Description automatically generated

Hình 3.9: Hàm trả về giá trị Gain lớn nhất và xử lý

trường hợp kết thúc

-Hàm best\_split sẽ trả về đặc trưng có Gain cao nhất để rẽ nhánh.

-Hàm majority\_class sẽ trả về nhãn chiếm đa số khi không thể phân chia tiếp.

**Hàm xây dựng cây quyết định**

A computer screen shot of a program code

Description automatically generated

Hình 3.10: Hàm xây dựng cây quyết định

-Hàm xây dựng cây quyết định sử dụng thuật toán ID3 theo đệ quy.

-Điều kiện dừng:

+Nếu tất cả các mẫu trong tập hiện tại đều thuộc một lớp (Ví dụ đều là 1 hoặc 0): dừng lại và trả về nhãn đó.

+Nếu không còn đặc trưng nào để tách tiếp: dừng và chọn nhãn xuất hiện nhiều nhất làm dự đoán.

+Nếu cây đã phát triển đến mức độ sâu giới hạn: dừng lại, trả về nhãn đa số tại nút hiện tại.

**Huấn luyện mô hình**

A screen shot of a computer

Description automatically generated

Hình 3.11: Huấn luyện mô hình cây quyết định

**Đưa ra độ chính xác của mô hình**

A computer screen shot of white text

Description automatically generated

Hình 3.12: Tính độ chính xác của mô hình

**Minh họa việc dự đoán sức khỏe cho bệnh nhân**

A computer screen shot of a program

Description automatically generated

Hình 3.13: Ví dụ chẩn đoán bệnh nhân bất kì

**Kết quả đánh giá**

A screen shot of a computer

Description automatically generated

Hình 3.14: Kết quả đánh giá của mô hình

# PHẦN C: KẾT LUẬN

Trong bối cảnh y tế hiện đại, việc ứng dụng các mô hình trí tuệ nhân tạo như cây quyết định đã và đang mở ra những hướng đi đầy tiềm năng trong công tác chẩn đoán, dự phòng và điều trị bệnh. Thông qua đề tài "Tìm hiểu cây quyết định và ứng dụng dự đoán sức khỏe", đặc biệt là đối với bệnh tiểu đường, nhóm chúng em đã có cơ hội tiếp cận, phân tích và triển khai một trong những mô hình học máy cơ bản nhưng vô cùng hiệu quả.

Quá trình nghiên cứu đã giúp chúng em hiểu rõ nguyên lý hoạt động của thuật toán cây quyết định, đặc biệt là ID3, cách tính Entropy và Information Gain để xây dựng cấu trúc cây hợp lý. Đồng thời, việc áp dụng mô hình vào bộ dữ liệu thực tế Pima Indians Diabetes đã minh chứng tính khả thi trong việc phân loại và dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các yếu tố sức khỏe cá nhân.

Kết quả thử nghiệm cho thấy mô hình đạt được độ chính xác tương đối tốt và có khả năng hỗ trợ đắc lực cho các chuyên gia y tế trong công tác sàng lọc ban đầu, nhất là tại các khu vực còn thiếu thiết bị y tế hiện đại. Tuy nhiên, để nâng cao hiệu quả thực tiễn, mô hình cần được tối ưu thêm về mặt dữ liệu, xử lý mất cân bằng lớp, và tích hợp với các phương pháp khác như cây rừng ngẫu nhiên (Random Forest) hay boosting.

Qua đề tài này, nhóm chúng em không chỉ củng cố kiến thức về trí tuệ nhân tạo và học máy, mà còn nhận thức sâu sắc hơn về vai trò của công nghệ trong lĩnh vực chăm sóc sức khỏe. Trong tương lai, chúng em hy vọng sẽ có thêm nhiều cơ hội nghiên cứu chuyên sâu, đóng góp vào việc xây dựng những giải pháp y tế thông minh, góp phần cải thiện chất lượng cuộc sống cho cộng đồng.

# TÀI LIỆU THAM KHẢO

1.Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81–106.  
 <https://link.springer.com/article/10.1007/BF00116251>

**2.**Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.  
<https://link.springer.com/article/10.1023/A:1010933404324>

3.UCI Machine Learning Repository – Pima Indians Diabetes Dataset.  
<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

4.Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O’Reilly Media. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/