# Các thành viên

|  |  |  |
| --- | --- | --- |
| **MSV** | **Họ tên** | **% tham gia** |
| B21DCCN686 | Đỗ Đức Thiện | 16.6% |
| B21DCCN110 | Phạm Thanh Sơn | 16.6% |
| B21DCCN230 | Nguyễn Minh Đông | 16.6% |
| B21DCCN408 | Đỗ Mạnh Hùng | 16.6% |
| B21DCCN351 | Hoàng Đình Hiếu | 16.6% |
| B21DCCN633 | Trần Minh Quang | 16.6% |

# LỜI CẢM ƠN

Lời đầu tiên, em xin được gửi lời cảm ơn chân thành nhất đến thầy PTIT. Trong quá trình học tập và tìm hiểu môn Phát triển các hệ thống thông minh, em đã nhận được rất nhiều sự quan tâm, giúp đỡ, hướng dẫn tâm huyết và tận tình của thầy.

Cám ơn bạn bè

Cám ơn người thân

# Mục lục

[Các thành viên i](#_Toc179961585)

[LỜI CẢM ƠN ii](#_Toc179961586)

[Mục lục iii](#_Toc179961587)

[DANH SÁCH CÁC HÌNH VẼ, bảng biểu v](#_Toc179961588)

[DANH SÁCH CÁC TỪ VIẾT TẮT vi](#_Toc179961589)

[MỞ ĐẦU 1](#_Toc179961590)

[1. Giới thiệu 1](#_Toc179961591)

[2. Lý do chọn đề tài 1](#_Toc179961592)

[3. Mục tiêu nghiên cứu 2](#_Toc179961593)

[4. Đối tượng và phạm vi nghiên cứu 2](#_Toc179961594)

[Chương 1. Học sâu 3](#_Toc179961595)

[1.1 Học máy 3](#_Toc179961596)

[1.1.1 Định nghĩa và khái niệm cơ bản về học máy 3](#_Toc179961597)

[1.1.2. Các loại học máy 4](#_Toc179961598)

[1.2.3. Các kỹ thuật và mô hình Học Sâu phổ biến 11](#_Toc179961599)

[1.3. Kết luận 15](#_Toc179961600)

[Chương 2. Bài toán phân loại 21](#_Toc179961601)

[2.1 Bài toán 21](#_Toc179961602)

[2.1.1. Các thành phần chính của bài toán phân loại 21](#_Toc179961603)

[2.1.2. Các bài toán phân loại 21](#_Toc179961604)

[2.1.3. Quy trình giải quyết bài toán phân loại 22](#_Toc179961605)

[2.1.4. Bài toán phân loại bệnh (tiêu chảy, phổi, tim mạch) 22](#_Toc179961606)

[2.2. Giải pháp: Thuật toán ID3 23](#_Toc179961607)

[2.3 Kết luận 30](#_Toc179961608)

[Chương 3. Dữ liệu 32](#_Toc179961609)

[3.1. Bộ Dữ liệu 32](#_Toc179961610)

[3.1.1 Giới thiệu chung về bộ dữ liệu 32](#_Toc179961611)

[3.1.2. Mô tả chi tiết các biến trong bộ dữ liệu 32](#_Toc179961612)

[3.3. Kết luận và Phân tích EDA (Exploratory Data Analysis) 38](#_Toc179961613)

[Chương 4. Thực nghiệm 42](#_Toc179961614)

[4.1. Mô hình 42](#_Toc179961615)

[4.1.1. Thuật toán ID3 42](#_Toc179961616)

[4.1.2. Xử lý dữ liệu 43](#_Toc179961617)

[4.2. Minh họa 43](#_Toc179961618)

[4.2.1. Xây dựng cây quyết định 43](#_Toc179961619)

[4.2.2 Đánh giá mô hình 47](#_Toc179961620)

[4.2.3 Kết quả thực nghiệm 48](#_Toc179961621)

[4.3 Kết luận 49](#_Toc179961622)

[4.3.1 Hiệu quả phân loại 49](#_Toc179961623)

[4.3.2 Vai trò của các đặc trưng 49](#_Toc179961624)

[4.3.3 Hạn chế và lỗi phân loại 49](#_Toc179961625)

[KẾT LUẬN 51](#_Toc179961626)

[1. Kết quả đạt được 51](#_Toc179961627)

[2. Phương hướng phát triển 51](#_Toc179961628)

[Tài liệu tham khảo 54](#_Toc179961629)

# DANH SÁCH CÁC HÌNH VẼ, bảng biểu

[Hình 1.1 Hình ảnh minh họa quy trình học máy cơ bản 3](#_Toc179961630)

[Hình 1.2. Sơ đồ cấu trúc của một mạng nơ ron sâu điển hình 9](#_Toc179961631)

[Hình 2.1. Đồ thị của hàm entropy với n =2 25](#_Toc179961632)

[Bảng 2.1. Dữ liệu huấn luyện 27](#_Toc179961633)

[Hình 2.2. Cây quyết định 30](#_Toc179961634)

[Hình 3.5. Chương trình 37](#_Toc179961635)

[Bảng 4.1. Dữ liệu minh họa 43](#_Toc179961636)

[Hình 4.1. Cây quyết định cho phân loại bệnh 47](#_Toc179961637)

# DANH SÁCH CÁC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Từ viết tắt** | **Tiếng Anh** | **Ý nghĩa** |
| AI | Artificial Intelligence (Trí tuệ nhân tạo) | Artificial Intelligence (Trí tuệ nhân tạo) |
| BERT | Bidirectional Encoder Representations from Transformers | Bidirectional Encoder Representations from Transformers |
| CNN | Convolutional Neural Networks | Convolutional Neural Networks |
| GAN | Generative Adversarial Networks | Generative Adversarial Networks |
| GNN | Graph Neural Network | Graph Neural Network |
| GPT | Generative Pre-trained Transformer | Generative Pre-trained Transformer |
| GPU | Graphics Processing Unit | Graphics Processing Unit |
| GRU | Gated Recurrent Unit | Gated Recurrent Unit |
| ID3 | Iterative Dichotomiser 3 | Iterative Dichotomiser 3 |
| IOT | Internet Of Things | Internet Of Things |
| LSTM | Long Short-Term Memory | Long Short-Term Memory |
| NLP | Neuro-Linguistic Programming | Neuro-Linguistic Programming |
| RNN | Recurrent Neural Networks | Recurrent Neural Networks |
| VEA | Variational Autoencoders | Variational Autoencoders |

# MỞ ĐẦU

## 1. Giới thiệu

Trong thời đại công nghệ số phát triển vượt bậc hiện nay, học máy (machine learning) đang ngày càng đóng vai trò quan trọng trong nhiều lĩnh vực khác nhau, từ kinh doanh, tài chính, cho đến y tế và chăm sóc sức khỏe. Đặc biệt, trong lĩnh vực y tế, ứng dụng học máy đã mang đến những bước tiến đáng kể trong việc hỗ trợ chẩn đoán bệnh, từ đó cải thiện chất lượng chăm sóc sức khỏe cho bệnh nhân. Một trong những kỹ thuật học máy phổ biến và hiệu quả là Decision Tree (Cây quyết định), cho phép phân loại và dự đoán các nhóm dữ liệu phức tạp dựa trên một tập hợp các đặc điểm hoặc tiêu chí nhất định.

Bệnh tật là một vấn đề sức khỏe nghiêm trọng, đặc biệt là các bệnh liên quan đến phổi, tiêu chảy và tim mạch, vốn là những bệnh lý phổ biến và gây ra nhiều tác động tiêu cực đến sức khỏe cộng đồng. Việc chẩn đoán và điều trị sớm các bệnh này là vô cùng quan trọng, giúp giảm thiểu nguy cơ tử vong và gánh nặng cho hệ thống y tế. Tuy nhiên, quá trình chẩn đoán truyền thống thường dựa vào kinh nghiệm của bác sĩ, điều này có thể dẫn đến sai sót do các yếu tố chủ quan hoặc giới hạn về kiến thức. Do đó, việc áp dụng các kỹ thuật học máy để hỗ trợ chẩn đoán chính xác hơn là một nhu cầu cấp thiết.

## 2. Lý do chọn đề tài

Trong số các kỹ thuật học máy hiện đại, Decision Tree là một phương pháp đơn giản nhưng hiệu quả, giúp phân loại bệnh dựa trên các yếu tố đầu vào như triệu chứng, kết quả xét nghiệm và các yếu tố lâm sàng khác. Phương pháp này không chỉ giúp cải thiện độ chính xác trong việc chẩn đoán, mà còn cho phép chúng ta giải thích được kết quả, nhờ khả năng biểu diễn quá trình quyết định thông qua cấu trúc cây. Mỗi nhánh của cây đại diện cho một quy tắc hoặc điều kiện dựa trên đặc điểm cụ thể của bệnh nhân, giúp xác định một cách rõ ràng bệnh mà bệnh nhân có thể mắc phải.

Bệnh phổi, tiêu chảy và tim mạch là ba nhóm bệnh phổ biến và gây nhiều ảnh hưởng nghiêm trọng đến sức khỏe con người. Các bệnh này không chỉ xuất hiện nhiều trong cộng đồng, mà còn tiềm ẩn nguy cơ cao về tử vong nếu không được phát hiện và điều trị kịp thời. Việc sử dụng các phương pháp học máy để phân loại và chẩn đoán các loại bệnh này sẽ góp phần không nhỏ trong việc nâng cao hiệu quả điều trị, đồng thời giảm thiểu thời gian và chi phí cho các cơ sở y tế. Hơn nữa, hệ thống học máy có thể được áp dụng trong nhiều môi trường khác nhau, từ bệnh viện lớn đến các phòng khám nhỏ hoặc thậm chí trong các ứng dụng chăm sóc sức khỏe tại nhà.

## 3. Mục tiêu nghiên cứu

Đề tài này nhằm phát triển một mô hình học máy sử dụng thuật toán Decision Tree để phân loại ba nhóm bệnh chính: bệnh phổi, bệnh tiêu chảy và bệnh tim mạch. Mô hình sẽ được huấn luyện dựa trên dữ liệu bệnh nhân, bao gồm các yếu tố liên quan đến triệu chứng, kết quả chẩn đoán lâm sàng, và các yếu tố sức khỏe khác. Cụ thể, nghiên cứu tập trung vào các mục tiêu sau:

* Xây dựng bộ dữ liệu: Tổng hợp dữ liệu từ các nguồn khác nhau bao gồm bệnh án điện tử, cơ sở dữ liệu y tế hoặc dữ liệu được tổng hợp từ các nghiên cứu lâm sàng.
* Huấn luyện mô hình Decision Tree: Áp dụng thuật toán Decision Tree để huấn luyện mô hình trên bộ dữ liệu đã thu thập được.
* Đánh giá hiệu suất của mô hình: Sử dụng các chỉ số đánh giá hiệu suất như độ chính xác (accuracy), độ nhạy (sensitivity), và độ đặc hiệu (specificity) để đánh giá khả năng phân loại bệnh của mô hình.
* Tối ưu hóa mô hình: Thực hiện các phương pháp tối ưu hóa như pruning, điều chỉnh tham số để cải thiện hiệu suất của mô hình.

## 4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài này là các bệnh nhân mắc ba nhóm bệnh chính: bệnh phổi, bệnh tiêu chảy và bệnh tim mạch. Dữ liệu thu thập sẽ bao gồm thông tin về triệu chứng, kết quả xét nghiệm và các yếu tố liên quan đến bệnh lý của từng nhóm bệnh. Phạm vi nghiên cứu tập trung vào việc phát triển và đánh giá mô hình phân loại bệnh dựa trên dữ liệu đã thu thập.

Nghiên cứu sẽ không tập trung vào các yếu tố điều trị hay can thiệp y tế, mà chỉ tập trung vào việc phân loại bệnh thông qua các phương pháp học máy, cụ thể là thuật toán Decision Tree. Hơn nữa, phạm vi nghiên cứu cũng giới hạn trong việc phân tích và phân loại bệnh theo dữ liệu đầu vào, không xem xét đến các yếu tố phức tạp hơn như di truyền hay các biến chứng liên quan.

# Chương 1. Học sâu

## 1.1 Học máy

### 1.1.1 Định nghĩa và khái niệm cơ bản về học máy

Học máy là một lĩnh vực của trí tuệ nhân tạo tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính "học" từ dữ liệu mà không cần được lập trình cụ thể. Thay vì tuân theo các quy tắc cứng nhắc, các hệ thống Học Máy có khả năng tự động cải thiện hiệu suất của mình thông qua kinh nghiệm.

Khái niệm cơ bản của Học Máy bao gồm:

1. Dữ liệu (Data): Là nguyên liệu thô cho quá trình học. Dữ liệu có thể ở nhiều dạng khác nhau như số, văn bản, hình ảnh, âm thanh, v.v.
2. Mô hình (Model): Là biểu diễn toán học của một quá trình trong thế giới thực. Mô hình Học Máy cố gắng bắt chước cách mà con người học và đưa ra quyết định.
3. Huấn luyện (Training): Là quá trình mà mô hình học từ dữ liệu đầu vào. Trong quá trình này, mô hình điều chỉnh các tham số của nó để tối ưu hóa hiệu suất trên tập dữ liệu huấn luyện.
4. Dự đoán (Prediction): Là khả năng của mô hình đã được huấn luyện để đưa ra kết quả cho dữ liệu mới chưa từng thấy.
5. Đánh giá (Evaluation): Là quá trình kiểm tra hiệu suất của mô hình trên dữ liệu mới để đảm bảo rằng nó có thể tổng quát hóa tốt.

A diagram of a machine learning workflow

Description automatically generated

#### Hình 1.1 Hình ảnh minh họa quy trình học máy cơ bản

Học Máy đóng vai trò quan trọng trong việc xây dựng các hệ thống AI có khả năng thích nghi và cải thiện theo thời gian. Nó tạo ra cơ sở cho nhiều ứng dụng AI tiên tiến mà chúng ta thấy ngày nay, từ hệ thống gợi ý trên các nền tảng trực tuyến đến xe tự lái và trợ lý ảo thông minh.

### 1.1.2. Các loại học máy

Học Máy có thể được phân loại thành ba nhóm chính dựa trên cách thức học và loại dữ liệu được sử dụng:

1. **Học có giám sát (Supervised Learning)**:
   * Trong phương pháp này, mô hình được huấn luyện trên một tập dữ liệu có nhãn, nghĩa là mỗi mẫu dữ liệu đầu vào đều có một kết quả mong muốn tương ứng.
   * Mục tiêu là học một hàm ánh xạ từ đầu vào sang đầu ra dựa trên các cặp đầu vào-đầu ra ví dụ.
   * Ứng dụng: Phân loại (Classification) và Hồi quy (Regression).
   * Ví dụ: Nhận dạng chữ viết tay, dự đoán giá nhà, phân loại email spam.
2. **Học không giám sát (Unsupervised Learning)**:
   * Mô hình được huấn luyện trên dữ liệu không có nhãn, và nhiệm vụ là tìm ra cấu trúc hoặc mẫu ẩn trong dữ liệu.
   * Không có "đáp án đúng" để so sánh; mô hình phải tự khám phá thông tin hữu ích.
   * Ứng dụng: Phân cụm (Clustering), Giảm chiều dữ liệu (Dimensionality Reduction), Phát hiện bất thường (Anomaly Detection).
   * Ví dụ: Phân khúc khách hàng, phát hiện gian lận, tìm kiếm các chủ đề trong một tập tài liệu lớn.
3. **Học tăng cường (Reinforcement Learning)**:
   * Một tác nhân học cách hành động trong một môi trường để tối đa hóa một số khái niệm về phần thưởng tích lũy.
   * Tác nhân học thông qua thử nghiệm và sai lầm, nhận phản hồi (thưởng hoặc phạt) từ môi trường cho mỗi hành động.
   * Ứng dụng: Lập kế hoạch, ra quyết định trong các hệ thống động.
   * Ví dụ: Đào tạo AI chơi game, tối ưu hóa quản lý nguồn lực trong các hệ thống phức tạp.

A diagram of a machine learning

Description automatically generated

*Hình 1.1.2 Biểu đồ so sánh các loại Học Máy (có giám sát, không giám sát, học tăng cường)*

Ngoài ra, còn có một số phương pháp học kết hợp hoặc nâng cao khác:

1. **Học bán giám sát (Semi-supervised Learning)**:
   * Kết hợp giữa dữ liệu có nhãn và không nhãn trong quá trình huấn luyện.
   * Hữu ích khi có một lượng lớn dữ liệu không nhãn và một lượng nhỏ dữ liệu có nhãn.
2. **Học chuyển giao (Transfer Learning)**:
   * Sử dụng kiến thức học được từ một nhiệm vụ để cải thiện hiệu suất học trên một nhiệm vụ liên quan khác.
   * Đặc biệt hữu ích khi dữ liệu cho nhiệm vụ mục tiêu có hạn.
3. **Học tập trung (Federated Learning)**:
   * Huấn luyện mô hình trên dữ liệu phân tán mà không cần tập trung dữ liệu vào một nơi.
   * Bảo vệ quyền riêng tư bằng cách giữ dữ liệu cục bộ.

Mỗi loại Học Máy có ưu điểm và hạn chế riêng, phù hợp với các loại vấn đề và dữ liệu khác nhau. Việc lựa chọn phương pháp phù hợp phụ thuộc vào bản chất của vấn đề, loại và lượng dữ liệu có sẵn, cũng như các ràng buộc về tài nguyên và hiệu suất.

**1.1.3. Ứng dụng của Học Máy trong thực tế**

Học Máy đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực của đời sống, kinh tế và xã hội. Dưới đây là một số ví dụ tiêu biểu về ứng dụng của Học Máy trong thực tế:

1. **Y tế và Chăm sóc sức khỏe**:
   * Chẩn đoán bệnh dựa trên hình ảnh y tế (X-quang, CT, MRI).
   * Dự đoán nguy cơ mắc bệnh dựa trên dữ liệu di truyền và lối sống.
   * Phát triển thuốc mới và tối ưu hóa liệu trình điều trị.
   * Phân tích dữ liệu từ thiết bị đeo để theo dõi sức khỏe cá nhân.
2. **Tài chính và Ngân hàng**:
   * Phát hiện giao dịch gian lận và bất thường.
   * Đánh giá rủi ro tín dụng và quyết định cho vay.
   * Dự đoán biến động thị trường chứng khoán.
   * Tự động hóa dịch vụ khách hàng thông qua chatbot và trợ lý ảo.
3. **Thương mại điện tử và Tiếp thị**:
   * Hệ thống gợi ý sản phẩm cá nhân hóa.
   * Phân tích sentiment của khách hàng từ đánh giá và bình luận.
   * Tối ưu hóa giá động và quản lý hàng tồn kho.
   * Nhắm mục tiêu quảng cáo và phân tích hiệu quả chiến dịch.
4. **Giao thông và Logistics**:
   * Hệ thống điều khiển xe tự lái.
   * Tối ưu hóa lộ trình vận chuyển và quản lý chuỗi cung ứng.
   * Dự đoán và quản lý luồng giao thông đô thị.
   * Bảo trì dự đoán cho phương tiện và cơ sở hạ tầng.
5. **An ninh và Giám sát**:
   * Nhận dạng khuôn mặt và hành vi bất thường trong video giám sát.
   * Phát hiện mã độc và các mối đe dọa an ninh mạng.
   * Phân tích dữ liệu lớn để phát hiện hoạt động khủng bố.
   * Xác minh danh tính thông qua sinh trắc học.
6. **Giáo dục**:
   * Hệ thống học tập cá nhân hóa dựa trên tiến độ và phong cách học của học viên.
   * Chấm điểm tự động cho bài thi và bài luận.
   * Phát hiện gian lận trong các kỳ thi trực tuyến.
   * Dự đoán nguy cơ bỏ học và đề xuất can thiệp kịp thời.
7. **Nông nghiệp và Môi trường**:
   * Dự báo thời tiết và biến đổi khí hậu chi tiết.
   * Tối ưu hóa sử dụng nước và phân bón trong nông nghiệp chính xác.
   * Phát hiện và theo dõi sự phát triển của cây trồng qua hình ảnh vệ tinh.
   * Phát hiện và dự đoán sự cố môi trường như cháy rừng hoặc ô nhiễm nước.
8. **Nghệ thuật và Giải trí**:
   * Tạo nhạc và nghệ thuật bằng AI.
   * Cải thiện hiệu ứng đặc biệt và hoạt hình trong phim.
   * Tạo ra nội dung cá nhân hóa cho game và ứng dụng giải trí.
   * Dịch và lồng tiếng tự động cho phim và video.
9. **Sản xuất và Công nghiệp**:
   * Kiểm soát chất lượng tự động bằng thị giác máy tính.
   * Dự đoán nhu cầu bảo trì cho máy móc công nghiệp.
   * Tối ưu hóa quy trình sản xuất và tiêu thụ năng lượng.
   * Thiết kế sản phẩm và mô phỏng hiệu suất.
10. **Dịch vụ Công và Quản lý Đô thị**:
    * Tối ưu hóa việc phân bổ nguồn lực công cộng.
    * Dự đoán và quản lý nhu cầu năng lượng đô thị.
    * Phân tích dữ liệu từ các cảm biến IoT để quản lý thông minh hơn.
    * Cải thiện dịch vụ công dân thông qua phân tích dữ liệu và tự động hóa.

Những ứng dụng này chỉ là một phần nhỏ trong vô số cách mà Học Máy đang được sử dụng để giải quyết các vấn đề phức tạp và cải thiện hiệu quả trong nhiều lĩnh vực của cuộc sống. Khi công nghệ tiếp tục phát triển, chúng ta có thể mong đợi sẽ thấy nhiều ứng dụng sáng tạo và đột phá hơn nữa trong tương lai.

**1.2. Học Sâu**

**1.2.1. Giới thiệu về Học Sâu**

Học Sâu (Deep Learning) là một nhánh của Học Máy, tập trung vào việc huấn luyện các mạng neural nhân tạo với nhiều lớp ẩn. Đây là một bước đột phá quan trọng trong lĩnh vực Trí Tuệ Nhân Tạo, cho phép máy tính thực hiện các nhiệm vụ phức tạp như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên và ra quyết định ở mức độ gần với con người.

Các đặc điểm chính của Học Sâu bao gồm:

1. **Khả năng học biểu diễn phân cấp**: Học Sâu cho phép mô hình tự động học các đặc trưng từ dữ liệu thô, từ các đặc trưng cấp thấp đến các khái niệm trừu tượng cấp cao.
2. **Khả năng mở rộng**: Mô hình Học Sâu có thể tăng hiệu suất khi được cung cấp nhiều dữ liệu và sức mạnh tính toán hơn.
3. **Tính linh hoạt**: Có thể áp dụng cho nhiều loại dữ liệu khác nhau như hình ảnh, âm thanh, văn bản, và dữ liệu cảm biến.
4. **End-to-end learning**: Học Sâu có thể học trực tiếp từ dữ liệu đầu vào đến đầu ra mong muốn, giảm thiểu nhu cầu về kỹ thuật đặc trưng thủ công.

Sự phát triển của Học Sâu được thúc đẩy bởi ba yếu tố chính:

1. **Dữ liệu lớn**: Sự gia tăng của dữ liệu số hóa cung cấp nguồn nguyên liệu dồi dào cho việc huấn luyện.
2. **Sức mạnh tính toán**: Sự phát triển của GPU và các phần cứng chuyên dụng cho phép huấn luyện các mô hình phức tạp hơn.
3. **Tiến bộ về thuật toán**: Các kỹ thuật mới như ReLU, Dropout, và các phương pháp tối ưu hóa tiên tiến đã cải thiện đáng kể hiệu suất và khả năng học của các mô hình.

Học Sâu đã tạo ra những bước đột phá trong nhiều lĩnh vực, từ nhận dạng hình ảnh và xử lý ngôn ngữ tự nhiên đến chơi game và điều khiển robot. Nó đóng vai trò quan trọng trong việc phát triển các hệ thống AI tiên tiến như trợ lý ảo, xe tự lái, và hệ thống dịch máy.

**1.2.2. Cấu trúc và hoạt động của mạng neural sâu**

Mạng neural sâu là cốt lõi của Học Sâu, với cấu trúc và hoạt động phức tạp hơn so với các mạng neural truyền thống. Dưới đây là các thành phần và nguyên lý hoạt động chính của mạng neural sâu:

1. **Cấu trúc mạng**:

A diagram of a network

Description automatically generated

#### Hình 1.2. Sơ đồ cấu trúc của một mạng nơ ron sâu điển hình

* + **Lớp đầu vào**: Nhận dữ liệu thô (ví dụ: pixel của hình ảnh, từ trong câu).
  + **Các lớp ẩn**: Nhiều lớp xử lý trung gian, mỗi lớp học các biểu diễn ngày càng trừu tượng của dữ liệu.
  + **Lớp đầu ra**: Cung cấp kết quả cuối cùng (ví dụ: xác suất phân loại).

1. **Neuron (Nút)**:
   * Đơn vị cơ bản của mạng, nhận đầu vào từ các neuron khác hoặc dữ liệu đầu vào.
   * Áp dụng một hàm kích hoạt phi tuyến tính (như ReLU, sigmoid, tanh) lên tổng có trọng số của đầu vào.
2. **Kết nối và trọng số**:
   * Neuron trong một lớp kết nối với neuron ở lớp tiếp theo.
   * Mỗi kết nối có một trọng số, xác định mức độ ảnh hưởng của tín hiệu.
3. **Lan truyền xuôi (Forward Propagation)**:
   * Dữ liệu đi từ lớp đầu vào qua các lớp ẩn đến lớp đầu ra.
   * Mỗi neuron tính toán đầu ra dựa trên đầu vào và trọng số của nó.
4. **Hàm mất mát (Loss Function)**:
   * Đo lường sự khác biệt giữa đầu ra của mạng và giá trị mong muốn.
   * Ví dụ: Mean Squared Error cho hồi quy, Cross-Entropy cho phân loại.
5. **Lan truyền ngược (Backpropagation)**:
   * Tính toán gradient của hàm mất mát đối với mỗi trọng số trong mạng.
   * Cho phép điều chỉnh trọng số để giảm thiểu lỗi.
6. **Tối ưu hóa**:
   * Sử dụng các thuật toán như Gradient Descent để cập nhật trọng số.
   * Mục tiêu là tìm bộ trọng số tối ưu giảm thiểu hàm mất mát.
7. **Regularization**:
   * Kỹ thuật như Dropout, L1/L2 regularization giúp ngăn chặn overfitting.
8. **Batch Normalization**:
   * Chuẩn hóa đầu vào của mỗi lớp để cải thiện tốc độ và ổn định trong quá trình học.
9. **Kiến trúc đặc biệt**:
   * Convolutional Neural Networks (CNN) cho xử lý hình ảnh.
   * Recurrent Neural Networks (RNN) và Long Short-Term Memory (LSTM) cho dữ liệu chuỗi.
   * Transformer cho xử lý ngôn ngữ tự nhiên.

Quá trình học của mạng neural sâu bao gồm:

1. Khởi tạo trọng số ngẫu nhiên.
2. Lan truyền xuôi để tính đầu ra.
3. Tính toán lỗi sử dụng hàm mất mát.
4. Lan truyền ngược để tính gradient.
5. Cập nhật trọng số sử dụng thuật toán tối ưu hóa.
6. Lặp lại quá trình cho đến khi hội tụ.

Sự phức tạp của mạng neural sâu cho phép nó học được các biểu diễn phức tạp và trừu tượng từ dữ liệu, dẫn đến hiệu suất vượt trội trong nhiều nhiệm vụ khó khăn.

### 1.2.3. Các kỹ thuật và mô hình Học Sâu phổ biến

Học Sâu đã phát triển nhiều kỹ thuật và mô hình khác nhau để giải quyết các vấn đề cụ thể. Dưới đây là một số kỹ thuật và mô hình Học Sâu phổ biến:

1. **Convolutional Neural Networks (CNN)**:
   * Chủ yếu sử dụng cho xử lý hình ảnh và video.
   * Sử dụng các lớp tích chập để học các đặc trưng không gian.
   * Ứng dụng: Nhận dạng đối tượng, phân loại hình ảnh, phát hiện khuôn mặt.
2. **Recurrent Neural Networks (RNN)**:
   * Thiết kế cho dữ liệu chuỗi và dữ liệu thời gian.
   * Có khả năng "ghi nhớ" thông tin từ các bước trước.
   * Ứng dụng: Xử lý ngôn ngữ tự nhiên, dự báo chuỗi thời gian.
3. **Long Short-Term Memory (LSTM) và Gated Recurrent Unit (GRU)**:
   * Các biến thể cải tiến của RNN, giải quyết vấn đề mất mát gradient dài hạn.
   * Cho phép học các phụ thuộc dài hạn tốt hơn.
   * Ứng dụng: Dịch máy, tạo văn bản, phân tích sentiment.
4. **Transformer**:
   * Mô hình dựa trên cơ chế attention, không cần cấu trúc tuần tự như RNN.
   * Hiệu quả trong việc xử lý các chuỗi dài và song song hóa.
   * Ứng dụng: NLP, mô hình ngôn ngữ lớn như GPT, BERT.
5. **Generative Adversarial Networks (GAN)**:
   * Gồm hai mạng neural cạnh tranh: generator và discriminator.
   * Sử dụng để tạo ra dữ liệu mới giống với dữ liệu thật.
   * Ứng dụng: Tạo hình ảnh, video, âm nhạc, cải thiện độ phân giải.
6. **Autoencoders**:
   * Học cách nén dữ liệu và sau đó tái tạo lại.
   * Hữu ích cho giảm chiều dữ liệu và phát hiện bất thường.
   * Ứng dụng: Nén hình ảnh, loại bỏ nhiễu, phát hiện gian lận.
7. **Deep Reinforcement Learning**:
   * Kết hợp Học Sâu với Học Tăng cường.
   * Cho phép các tác nhân học cách hành động trong môi trường phức tạp.
   * Ứng dụng: Chơi game, điều khiển robot, tối ưu hóa quy trình.
8. **Siamese Networks**:
   * Sử dụng cho các nhiệm vụ so sánh hoặc xác minh.
   * Học cách đo lường sự tương đồng giữa hai đầu vào.
   * Ứng dụng: Nhận dạng khuôn mặt, chữ ký, xác minh danh tính.
9. **Capsule Networks**:
   * Cố gắng mô phỏng cách não người xử lý thông tin thị giác.
   * Giải quyết một số hạn chế của CNN trong việc hiểu quan hệ không gian.
   * Ứng dụng: Nhận dạng đối tượng với góc nhìn khác nhau.
10. **Graph Neural Networks (GNN)**:
    * Xử lý dữ liệu có cấu trúc đồ thị:
      1. Học biểu diễn cho các nút và cạnh trong đồ thị.
      2. Ứng dụng: Phân tích mạng xã hội, dự đoán tương tác phân tử, hệ thống gợi ý
11. **Variational Autoencoders (VAE)**:
    * Kết hợp autoencoders với học xác suất.
    * Cho phép tạo ra dữ liệu mới từ không gian tiềm ẩn liên tục.
    * Ứng dụng: Tạo hình ảnh, âm nhạc, và dữ liệu tổng hợp.
12. **Attention Mechanisms**:
    * Cho phép mô hình tập trung vào các phần quan trọng của đầu vào.
    * Cải thiện hiệu suất trong nhiều nhiệm vụ, đặc biệt là NLP.
    * Ứng dụng: Dịch máy, tóm tắt văn bản, trả lời câu hỏi.

Mỗi kỹ thuật và mô hình này có ưu điểm và hạn chế riêng, phù hợp với các loại vấn đề và dữ liệu khác nhau. Sự phát triển nhanh chóng của lĩnh vực này dẫn đến việc liên tục có các cải tiến và kỹ thuật mới xuất hiện.

**1.2.4. So sánh Học Sâu với Học Máy truyền thống**

Học Sâu và Học Máy truyền thống đều là những phương pháp quan trọng trong lĩnh vực Trí Tuệ Nhân Tạo, nhưng có một số điểm khác biệt đáng kể. Dưới đây là một so sánh chi tiết giữa hai phương pháp:

1. **Cấu trúc mô hình**:
   * Học Máy truyền thống: Thường sử dụng các mô hình đơn giản hơn như Decision Trees, SVM, hoặc Logistic Regression.
   * Học Sâu: Sử dụng mạng neural nhiều lớp với cấu trúc phức tạp hơn.
2. **Xử lý đặc trưng (Feature Engineering)**:
   * Học Máy truyền thống: Đòi hỏi nhiều công sức trong việc thiết kế và chọn lọc đặc trưng thủ công.
   * Học Sâu: Có khả năng tự động học và trích xuất đặc trưng từ dữ liệu thô.
3. **Yêu cầu dữ liệu**:
   * Học Máy truyền thống: Có thể hoạt động hiệu quả với lượng dữ liệu vừa phải.
   * Học Sâu: Thường đòi hỏi lượng dữ liệu lớn để đạt hiệu suất tốt.
4. **Khả năng mở rộng**:
   * Học Máy truyền thống: Hiệu suất thường đạt đến giới hạn khi tăng kích thước dữ liệu.
   * Học Sâu: Có thể tiếp tục cải thiện hiệu suất khi được cung cấp nhiều dữ liệu hơn.
5. **Tính giải thích được**:
   * Học Máy truyền thống: Thường dễ giải thích và hiểu hơn (ví dụ: Decision Trees).
   * Học Sâu: Thường được coi là "hộp đen", khó giải thích các quyết định của mô hình.
6. **Loại dữ liệu xử lý**:
   * Học Máy truyền thống: Hiệu quả với dữ liệu có cấu trúc và bán cấu trúc.
   * Học Sâu: Đặc biệt mạnh mẽ với dữ liệu phi cấu trúc như hình ảnh, âm thanh, và văn bản.
7. **Thời gian và tài nguyên tính toán**:
   * Học Máy truyền thống: Thường nhanh hơn để huấn luyện và yêu cầu ít tài nguyên tính toán hơn.
   * Học Sâu: Đòi hỏi thời gian huấn luyện lâu hơn và tài nguyên tính toán lớn (như GPU).
8. **Khả năng xử lý nhiệm vụ phức tạp**:
   * Học Máy truyền thống: Phù hợp với các nhiệm vụ đơn giản đến trung bình.
   * Học Sâu: Có khả năng xử lý các nhiệm vụ cực kỳ phức tạp như nhận dạng hình ảnh và xử lý ngôn ngữ tự nhiên.
9. **Tính linh hoạt**:
   * Học Máy truyền thống: Mỗi thuật toán thường được thiết kế cho một loại nhiệm vụ cụ thể.
   * Học Sâu: Có thể áp dụng cho nhiều loại nhiệm vụ khác nhau với ít thay đổi trong kiến trúc.
10. **Hiệu suất trên dữ liệu phức tạp**:
    * Học Máy truyền thống: Có thể gặp khó khăn với dữ liệu có nhiều chiều hoặc phức tạp.
    * Học Sâu: Xuất sắc trong việc xử lý dữ liệu nhiều chiều và phức tạp.
11. **Khả năng học biểu diễn phân cấp**:
    * Học Máy truyền thống: Thường không có khả năng này.
    * Học Sâu: Có thể học các biểu diễn từ đơn giản đến phức tạp qua nhiều lớp.
12. **Độ chính xác trên các nhiệm vụ phức tạp**:
    * Học Máy truyền thống: Thường đạt độ chính xác thấp hơn trên các nhiệm vụ phức tạp.
    * Học Sâu: Đạt độ chính xác cao hơn trên nhiều nhiệm vụ phức tạp, đôi khi vượt qua con người.

Mặc dù Học Sâu có nhiều ưu điểm vượt trội trong các nhiệm vụ phức tạp, Học Máy truyền thống vẫn có vai trò quan trọng trong nhiều ứng dụng. Việc lựa chọn giữa hai phương pháp phụ thuộc vào bản chất của vấn đề, lượng dữ liệu có sẵn, yêu cầu về tài nguyên tính toán, và nhu cầu giải thích mô hình.

## 1.3. Kết luận

Học Sâu đã trở thành một trong những động lực chính thúc đẩy sự phát triển của Trí Tuệ Nhân Tạo trong thập kỷ qua. Nó đã mang lại những bước đột phá đáng kể trong nhiều lĩnh vực, từ thị giác máy tính đến xử lý ngôn ngữ tự nhiên, và đang tiếp tục mở rộng ranh giới của những gì có thể với AI.

Vai trò quan trọng của Học Sâu trong AI:

1. **Đột phá về hiệu suất**: Học Sâu đã đạt được kết quả vượt trội trong nhiều nhiệm vụ, thậm chí vượt qua khả năng của con người trong một số lĩnh vực như nhận dạng hình ảnh và chơi game.
2. **Tự động hóa trích xuất đặc trưng**: Khả năng học biểu diễn tự động từ dữ liệu thô đã giảm đáng kể nhu cầu về kỹ thuật đặc trưng thủ công.
3. **Khả năng mở rộng**: Học Sâu có khả năng tận dụng hiệu quả lượng dữ liệu và sức mạnh tính toán ngày càng tăng.
4. **Tính linh hoạt**: Các kiến trúc Học Sâu có thể được áp dụng cho nhiều loại dữ liệu và nhiệm vụ khác nhau.
5. **Tạo ra nội dung**: Mô hình như GANs và Transformers đã mở ra khả năng tạo ra nội dung mới như hình ảnh, văn bản, và âm nhạc.
6. **Học end-to-end**: Khả năng học trực tiếp từ đầu vào đến đầu ra mong muốn, giảm thiểu sự can thiệp của con người.
7. **Thúc đẩy nghiên cứu AI**: Sự thành công của Học Sâu đã thúc đẩy đầu tư và nghiên cứu trong lĩnh vực AI nói chung.

Tuy nhiên, cũng cần nhận thức rằng Học Sâu không phải là giải pháp cho mọi vấn đề trong AI. Nó có những hạn chế nhất định và vẫn cần sự phát triển thêm để đạt được trí tuệ nhân tạo tổng quát.

Mặc dù Học Sâu đã đạt được những thành tựu đáng kể, nó vẫn phải đối mặt với nhiều thách thức và hạn chế:

1. **Nhu cầu dữ liệu lớn**:
   * Học Sâu thường đòi hỏi lượng dữ liệu huấn luyện rất lớn.
   * Trong nhiều lĩnh vực, việc thu thập đủ dữ liệu có chất lượng là một thách thức.
2. **Tính toán đòi hỏi cao**:
   * Huấn luyện mô hình Học Sâu thường đòi hỏi tài nguyên tính toán lớn.
   * Chi phí năng lượng và môi trường của việc huấn luyện mô hình lớn là đáng kể.
3. **Thiếu tính giải thích được**:
   * Nhiều mô hình Học Sâu hoạt động như "hộp đen", khó giải thích cách ra quyết định.
   * Điều này gây ra vấn đề trong các ứng dụng yêu cầu tính minh bạch cao như y tế hoặc tài chính.
4. **Dễ bị tấn công đối kháng**:
   * Mô hình Học Sâu có thể bị đánh lừa bởi các đầu vào được thiết kế đặc biệt.
   * Điều này gây ra lo ngại về bảo mật và độ tin cậy trong các ứng dụng quan trọng.
5. **Khả năng tổng quát hóa hạn chế**:
   * Mô hình thường hoạt động tốt trên dữ liệu tương tự với dữ liệu huấn luyện, nhưng có thể gặp khó khăn khi đối mặt với tình huống hoàn toàn mới.
6. **Overfitting**:
   * Mô hình phức tạp có thể "học thuộc lòng" dữ liệu huấn luyện, dẫn đến hiệu suất kém trên dữ liệu mới.
7. **Khó khăn trong việc tích hợp kiến thức tiên nghiệm**:
   * Không như con người, mô hình Học Sâu thường gặp khó khăn trong việc tích hợp kiến thức chung về thế giới.
8. **Độ nhạy với hyperparameters**:
   * Hiệu suất của mô hình có thể phụ thuộc nhiều vào việc điều chỉnh các hyperparameter, đòi hỏi kinh nghiệm và thời gian.
9. **Thiếu khả năng suy luận nhân quả**:
   * Mô hình Học Sâu giỏi trong việc tìm ra tương quan nhưng kém trong việc hiểu quan hệ nhân quả.
10. **Khó khăn trong học liên tục**:
    * Nhiều mô hình gặp khó khăn trong việc học thêm kiến thức mới mà không quên kiến thức cũ (catastrophic forgetting).
11. **Vấn đề đạo đức và công bằng**:
    * Mô hình có thể học và khuếch đại các thiên kiến trong dữ liệu huấn luyện.
    * Việc sử dụng AI trong ra quyết định quan trọng gây ra nhiều lo ngại về đạo đức.
12. **Khó khăn trong việc xử lý dữ liệu ít**:
    * Học sâu thường không hiệu quả trong các tình huống có ít dữ liệu huấn luyện, đặc biệt là so với con người có khả năng học từ ít ví dụ.
13. **Tính ổn định**:
    * Đôi khi, kết quả huấn luyện có thể không ổn định và thay đổi đáng kể giữa các lần chạy, ngay cả với cùng một bộ dữ liệu.

Việc giải quyết những thách thức và hạn chế này là trọng tâm của nhiều nghiên cứu hiện tại trong lĩnh vực Học Sâu và AI nói chung.

**1.3.3 Triển vọng và hướng phát triển trong tương lai**

Mặc dù đối mặt với nhiều thách thức, Học Sâu vẫn có triển vọng phát triển mạnh mẽ trong tương lai. Dưới đây là một số hướng phát triển và xu hướng đáng chú ý:

1. **Học Sâu tiết kiệm dữ liệu**:
   * Phát triển các phương pháp học hiệu quả từ ít dữ liệu như few-shot learning, zero-shot learning.
   * Tăng cường sử dụng kỹ thuật data augmentation và transfer learning.
2. **AI có thể giải thích được (Explainable AI - XAI)**:
   * Phát triển các phương pháp để hiểu và giải thích quyết định của mô hình Học Sâu.
   * Tích hợp tính minh bạch vào quá trình thiết kế mô hình.
3. **Học Sâu tiết kiệm năng lượng**:
   * Nghiên cứu các kiến trúc mạng và phương pháp huấn luyện tiết kiệm năng lượng hơn.
   * Phát triển phần cứng chuyên dụng cho AI tiết kiệm năng lượng.
4. **Học liên tục (Continual Learning)**:
   * Phát triển các mô hình có khả năng học liên tục, tích lũy kiến thức mới mà không quên kiến thức cũ.
5. **AI tổng quát hơn**:
   * Hướng tới các mô hình có khả năng tổng quát hóa tốt hơn, có thể áp dụng cho nhiều nhiệm vụ khác nhau.
   * Phát triển các phương pháp để tích hợp kiến thức chung về thế giới vào mô hình.
6. **Kết hợp Học Sâu với các phương pháp AI khác**:
   * Tích hợp Học Sâu với suy luận biểu tượng, lập kế hoạch, và các phương pháp AI truyền thống khác.
   * Phát triển các hệ thống hybrid kết hợp ưu điểm của nhiều phương pháp.
7. **Học Sâu tự giám sát (Self-supervised Learning)**:
   * Tận dụng lượng lớn dữ liệu không nhãn để cải thiện khả năng học của mô hình.
   * Phát triển các phương pháp học biểu diễn mạnh mẽ từ dữ liệu không có nhãn.
8. **AI đáng tin cậy và công bằng**:
   * Phát triển các phương pháp để đảm bảo tính công bằng và giảm thiểu thiên kiến trong mô hình Học Sâu.
   * Tăng cường độ tin cậy và an toàn của các hệ thống AI trong các ứng dụng quan trọng.
9. **Học Sâu cho IoT và Edge Computing**:
   * Phát triển các mô hình Học Sâu nhẹ có thể chạy trên các thiết bị có tài nguyên hạn chế.
   * Tối ưu hóa việc triển khai mô hình Học Sâu trên các thiết bị edge.
10. **AI sáng tạo**:
    * Cải thiện khả năng của mô hình trong việc tạo ra nội dung mới và sáng tạo như hình ảnh, âm nhạc, văn bản.
    * Phát triển các phương pháp để tăng cường tính nguyên bản và đa dạng trong nội dung được tạo ra.
11. **Học Sâu cho khoa học và y tế**:
    * Áp dụng Học Sâu trong các lĩnh vực như khám phá thuốc, phân tích gene, chẩn đoán bệnh.
    * Phát triển các mô hình có thể kết hợp dữ liệu đa dạng từ nhiều nguồn khác nhau.
12. **Học Sâu đa phương thức (Multimodal Deep Learning)**:
    * Phát triển các mô hình có khả năng xử lý và tích hợp thông tin từ nhiều loại dữ liệu khác nhau (văn bản, hình ảnh, âm thanh).
13. **Tự động hóa Machine Learning (AutoML)**:
    * Phát triển các phương pháp tự động hóa quá trình thiết kế và tối ưu hóa kiến trúc mạng neural.
14. **Học Sâu lượng tử (Quantum Deep Learning)**:
    * Khám phá tiềm năng của tính toán lượng tử trong việc cải thiện hiệu suất và khả năng của các mô hình Học Sâu.

Những hướng phát triển này không chỉ hứa hẹn giải quyết các thách thức hiện tại của Học Sâu mà còn mở ra những khả năng mới cho AI. Tuy nhiên, cùng với sự phát triện này, cần có sự quan tâm đúng mức đến các vấn đề đạo đức, pháp lý và xã hội mà công nghệ AI tiên tiến có thể mang lại.

Kết luận, Học Sâu đã và đang đóng vai trò quan trọng trong sự phát triển của AI, mang lại những tiến bộ đáng kể trong nhiều lĩnh vực ứng dụng. Mặc dù còn nhiều thách thức cần giải quyết, triển vọng của Học Sâu vẫn rất rộng mở. Với sự phát triển không ngừng của công nghệ và phương pháp mới, chúng ta có thể kỳ vọng vào những đột phá quan trọng trong tương lai, đưa chúng ta đến gần hơn với mục tiêu tạo ra các hệ thống AI thông minh, đáng tin cậy và có ích cho xã hội.

# Chương 2. Bài toán phân loại

## 2.1 Bài toán

Bài toán phân loại là một trong những vấn đề cơ bản và phổ biến nhất trong lĩnh vực học máy (Machine Learning) và trí tuệ nhân tạo (Artificial Intelligence). Mục tiêu của bài toán phân loại là xác định nhãn (label) hoặc lớp (class) của một đối tượng mới dựa trên các đặc trưng (features) của nó. Đây là một bài toán thuộc nhóm học có giám sát (supervised learning), trong đó hệ thống học từ một tập dữ liệu huấn luyện đã được gán nhãn sẵn.

### 2.1.1. Các thành phần chính của bài toán phân loại

* **Tập dữ liệu đầu vào**: Gồm các đối tượng cần phân loại, thường được biểu diễn dưới dạng các bản ghi với nhiều thuộc tính hoặc đặc trưng. Mỗi đặc trưng có thể là một con số (biến số học), một thuộc tính phân loại (categorical), hoặc thậm chí là các chuỗi ký tự (văn bản).
* **Nhãn (Label)**: Đây là giá trị cần dự đoán. Nó đại diện cho các lớp khác nhau mà hệ thống cần phân loại đối tượng vào. Ví dụ, trong bài toán phân loại email, nhãn có thể là “Spam” hoặc “Not Spam”.
* **Hàm phân loại (Classifier)**: Là một mô hình hoặc thuật toán được xây dựng từ dữ liệu huấn luyện. Mô hình này sẽ học cách liên kết giữa đặc trưng của dữ liệu và nhãn tương ứng.
* **Tập dữ liệu huấn luyện (Training data)**: Là tập hợp các bản ghi mà nhãn đã được biết trước, được sử dụng để huấn luyện mô hình.
* **Tập dữ liệu kiểm tra (Test data)**: Là tập hợp các bản ghi không có nhãn được sử dụng để kiểm tra tính chính xác của mô hình phân loại sau khi nó đã được huấn luyện.

### 2.1.2. Các bài toán phân loại

**Phân loại nhị phân** (Binary Classification):

* Đây là dạng phổ biến nhất của bài toán phân loại, nơi chỉ có hai nhãn có thể có. Ví dụ: Phân loại bệnh: “Có bệnh” hoặc “Không có bệnh”.
* Mục tiêu là học cách phân chia đối tượng vào một trong hai lớp dựa trên các đặc trưng của nó.

**Phân loại đa lớp** (Multiclass Classification):

* Trong bài toán này, có nhiều hơn hai nhãn. Ví dụ:

Phân loại loại hoa: “Hoa Hồng”, “Hoa Cúc”, “Hoa Lan”.

* Thuật toán cần phải dự đoán chính xác lớp mà đối tượng mới thuộc về trong số nhiều lớp có thể có.

**Phân loại đa nhãn** (Multilabel Classification):

* Ở đây, một đối tượng có thể thuộc nhiều hơn **một lớp**. Ví dụ:

Một bộ phim có thể thuộc các thể loại: “Hành động”, “Khoa học viễn tưởng”, “Hài”.

* Đây là bài toán phức tạp hơn, vì mỗi đối tượng có thể được gán nhiều nhãn khác nhau.

### 2.1.3. Quy trình giải quyết bài toán phân loại

Để giải quyết một bài toán phân loại, có thể thực hiện theo các bước sau:

Bước 1:Thu thập dữ liệu:

* Thu thập dữ liệu từ các nguồn khác nhau. Dữ liệu cần có các đặc trưng và nhãn được xác định rõ ràng. Nếu dữ liệu không có nhãn, cần thực hiện việc gán nhãn cho dữ liệu.

Bước 2: Tiền xử lý dữ liệu:

* Làm sạch dữ liệu, xử lý các giá trị thiếu, và chuẩn hóa dữ liệu nếu cần.
* Biến đổi các đặc trưng không số học (như văn bản hoặc phân loại) thành các giá trị số để thuật toán có thể xử lý.

Bước 3: Chọn và huấn luyện mô hình:

* Sử dụng các thuật toán phân loại để huấn luyện mô hình trên tập dữ liệu huấn luyện. Ví dụ các thuật toán như: Cây quyết định (ID3, C4.5), Mạng nơ-ron nhân tạo, Naive Bayes, Support Vector Machine, v.v.

Bước 4: Đánh giá mô hình:

* Sử dụng các phép đo như độ chính xác (accuracy), độ nhạy (recall), độ chính xác (precision) và điểm F1 (F1 score) để đánh giá mô hình trên tập dữ liệu kiểm tra.

Bước 5 : Triển khai và cải thiện:

* Sau khi đánh giá, triển khai mô hình vào ứng dụng thực tế. Dữ liệu mới có thể được sử dụng để cải thiện và tinh chỉnh mô hình.

### 2.1.4. Bài toán phân loại bệnh (tiêu chảy, phổi, tim mạch)

Phân loại bệnh là một trong những ứng dụng phổ biến nhất của học máy trong y tế, nhằm hỗ trợ các bác sĩ đưa ra các chẩn đoán chính xác và nhanh chóng dựa trên các đặc trưng của bệnh nhân. Trong bài toán này, mục tiêu là dự đoán xem một bệnh nhân có mắc bệnh hay không dựa trên các chỉ số và thông tin sức khỏe của họ. Những thông tin này bao gồm các yếu tố như tuổi tác, chỉ số huyết áp, nhịp tim, mức đường huyết, trạng thái sức khỏe v.v.

## 2.2. Giải pháp: Thuật toán ID3

Thuật Toán **ID3 (Iterative Dichotomiser 3)** được phát triển bởi Ross Quinlan vào những năm 1980. ID3 sử dụng chiến lược tham lam (greedy strategy) để xây dựng một Cây Quyết Định từ đỉnh xuống bằng cách tìm ra thuộc tính phân chia tối ưu tại mỗi bước. Kết quả là một cây mà ở đó dữ liệu được phân loại một cách tinh tế nhất.

**Thuật toán:**

o Khởi đầu: nút hiện thời là nút gốc chứa toàn bộ tập dữ liệu huấn luyện

o Tại nút hiện thời n, lựa chọn thuộc tính

▪ Chưa được sử dụng ở nút tổ tiên

▪ Cho phép phân chia tập dữ liệu hiện thời thành các tập con một cách tốt nhất

▪ Với mỗi giá trị thuộc tính được chọn thêm một nút con bên dưới

▪ Chia các ví dụ ở nút hiện thời về các nút con theo giá trị thuộc tính được chọn

o Lặp (đệ quy) cho tới khi

▪ Tất cả các thuộc tính đã được sử dụng ở các nút phía trên, hoặc

▪ Tất cả ví dụ tại nút hiện thời có cùng nhãn phân loại

▪ Nhãn của nút được lấy theo đa số nhãn của ví dụ tại nút hiện thời

**Tiêu chuẩn chọn thuộc tính của ID3**

Tại mỗi nút n

o Tập (con) dữ liệu ứng với nút đó

o Cần lựa chọn thuộc tính cho phép phân chia tập dữ liệu tốt nhất

Tiêu chuẩn:

o Dữ liệu sau khi phân chia càng đồng nhất càng tốt

o Đo bằng độ tăng thông tin (Information Gain - IG)

o Chọn thuộc tính có độ tăng thông tin lớn nhất

o IG dựa trên entropy của tập (con) dữ liệu

**Các đặc điểm của ID3:**

ID3 là thuật toán tìm kiếm cây quyết định phù hợp với dữ liệu huấn luyện:

* Tìm kiếm theo kiểu tham lam, bắt đầu từ cây rỗng.
* Hàm đánh giá dựa trên độ tăng thông tin.
* ID3 có khuynh hướng (bias) lựa chọn cây đơn giản:
  + Ít nút.
  + Các thuộc tính có độ tăng thông tin lớn nằm gần gốc.

**Hàm số Entropy**

Cho một phân phối xác suất của một biến rời rạc  có thể nhận *n*giá trị khác nhau ,…, Giả sử rằng xác suất để nhận các giá trị này là =p() với

0 ≤ ≤ 1, . Ký hiệu phân phối này là p = (p1,p2,...,pn). Entropy của phân phối này được định nghĩa là:

Trong đó log là logarit tự nhiên (*Một số tài liệu dùng logarit cơ số 2, nhưng giá trị của H(p) chỉ khác đi bằng cách nhân với một hằng số.)* và quy ước 0log(0) = 0.

Xét một ví dụ vơi n=2 được cho ở hình dưới đây, trong trường hợp **p** là tinh khiết nhất, tức một trong hai giá trị giá trị kia bằng 0, entropy của phân phối này là H(p) = 0. Khi p là vẩn đục nhất, tức cả hai giá trị hàm entropy đạt giá trị cao nhất.

*A graph of a function

Description automatically generated*

#### Hình 2.1. Đồ thị của hàm entropy với n =2

Tổng quát lên với n>2, hàm entropy đạt giá trị nhỏ nhất nếu có một giá trị  đạt giá trị lớn nhất nếu tất cả các  bằng nhau (việc này có thể được chứng minh bằng [phương pháp nhân tử Lagrange](https://machinelearningcoban.com/2017/04/02/duality/#--phuong-phap-nhan-tu-lagrange)).

Những tính chất này của hàm entropy khiến nó được sử dụng trong việc đo *độ vẩn đục* của một phép phân chia của ID3. Vì lý do này, ID3 còn được gọi là *entropy-based decision tree*.

**Độ tăng thông tin IG**

Với tập (con) mẫu S và thuộc tính A

*IG(S,A) = Entropy(S) -*

Trong đó:

Values(A): tập các giá trị của A

là tập con của S bao gồm các mẫu có giá trị của A bằng v

|S| là số phần tử của S

Trong ID3, tổng các trọng số của **entropy** tại các **leaf-node** sau khi xây dựng **decision tree** được coi là **hàm mất mát** của cây quyết định đó. Các trọng số ở đây tỷ lệ với số điểm dữ liệu được phân vào mỗi node. Công việc của ID3 là tìm các cách phân chia hợp lý (thứ tự chọn thuộc tính hợp lý) sao cho hàm mất mát cuối cùng đạt giá trị nhỏ nhất có thể.

Việc này được thực hiện bằng cách chọn thuộc tính sao cho, nếu sử dụng thuộc tính đó để phân chia, **entropy** tại mỗi bước giảm đi một lượng lớn nhất. Bài toán xây dựng một **decision tree** bằng ID3 có thể được chia thành nhiều bài toán nhỏ. Trong mỗi bài toán nhỏ này, ta chỉ cần chọn thuộc tính giúp cho việc phân chia đạt kết quả tốt nhất.

Mỗi bài toán nhỏ tương ứng với việc phân chia dữ liệu tại một **non-leaf node**. Chúng ta sẽ xây dựng phương pháp tính toán dựa trên việc phân tích các thuộc tính tại mỗi node này.

Xét một bài toán với C class khác nhau. Giả sử ta đang làm việc với một nom-leaf node với các điểm dữ liệu tạo thành một tập S với số phần tử là |S| = N. Giả sử thêm rằng trong số N điểm dữ liệu này, , c= 1,2,...,C điểm thuộc vào class c. Xác suất để mỗi điểm dữ liệu rơi vào một class c được xấp xỉ bằng (maximum likelihood estimation). Như vậy, entropy tại node này được tính bởi:

Tiếp theo, giả sử thuộc tính được chọn là . Dựa trên , các điểm dữ liệu trong S được phân ra thành K child node S1,S2,…,SK với số điểm trong mỗi child node lần lượt là m1,m2,…,. Ta định nghĩa

là tổng có trọng số entroy của mỗi child node–được tính tương tự như (2). Việc lấy trọng số này là quan trọng vì các node thường có số lượng điểm khác nhau.

Tiếp theo, ta định nghĩa *information gain* dựa trên thuộc tính :

G(,S)=H(S)−H(,S)

Trong ID3, tại mỗi node, thuộc tính được chọn được xác định dựa trên:

G(

tức thuộc tính khiến cho *information gain* đạt giá trị lớn nhất.

Ví dụ:

 Bảng dữ liệu này được lấy từ cuốn sách Data Mining: Practical Machine Learning Tools and Techniques, trang 11. Đây là một bảng dữ liệu được sử dụng rất nhiều trong các bài giảng về decision tree. Bảng dữ liệu này mô tả mối quan hệ giữa thời tiết trong 14 ngày (bốn cột đầu, không tính cột id) và việc một đội bóng có chơi bóng hay không (cột cuối cùng). Nói cách khác, ta phải dự đoán giá trị ở cột cuối cùng nếu biết giá trị của bốn cột còn lại.

#### Bảng 2.1. Dữ liệu huấn luyện

*A table of temperature and temperature

Description automatically generated with medium confidence*

Có bốn thuộc tính thời tiết:

1. *Outlook* nhận một trong ba giá trị: sunny, overcast, rainy.
2. *Temperature* nhận một trong ba giá trị: hot, cool, mild.
3. *Humidity* nhận một trong hai giá trị: high, normal.
4. *Wind* nhận một trong hai giá trị: weak, strong.

(Tổng cộng có 3×3×2×2=363×3×2×2=36 loại thời tiết khác nhau, trong đó 14 loại được thể hiện trong bảng.)

Đây có thể được coi là một bài toán dự đoán liệu đội bóng có chơi bóng không dựa trên các quan sát thời tiết. Ở đây, các quan sát đều ở dạng categorical. Cách dự đoán dưới đây tương đối đơn giản và khá chính xác, có thể không phải là cách ra quyết định tốt nhất:

* Nếu *outlook = sunny* và *humidity = high* thì *play = no*.
* Nếu *outlook = rainy* và *windy = true* thì *play = no*.
* Nếu *outlook = overcast* thì *play = yes*.
* Ngoài ra, nếu *humidity = normal* thì *play = yes*.
* Ngoài ra, *play = yes*.

Chúng ta sẽ cùng tìm thứ tự các thuộc tính bằng thuật toán ID3.

Trong 14 giá trị đầu ra ở Bảng trên, có năm giá trị bằng *no* và chín giá trị bằng *yes*. Entroy tại *root node* của bài toán là:

H(S) =

Tiếp theo, chúng ta tính tổng có trọng số entropy của các *child node* nếu chọn một trong các thuộc tính *outlook, temperature, humidity, wind, play* để phân chia dữ liệu.

Xét thuộc tính *outlook*. Thuộc tính này có thể nhận một trong ba giá trị *sunny, overcast, rainy*. Mỗi một giá trị sẽ tương ứng với một *child node*. Gọi tập hợp các điểm trong mỗi child node này lần lượt là Ss,So,Sr với tương ứng ms,mo,mr phần tử. Sắp xếp lại Bảng ban đầu theo thuộc tính outlook ta đạt được ba Bảng nhỏ sau đây.

A group of words on a white background

Description automatically generated

Quan sát nhanh ta thấy rằng *child node* ứng với *outlook = overcast* sẽ có entropy bằng 0 vì tất cả mo=4 output đều là *yes*. Hai *child node* còn lại với ms=mr=5 có entropy khá cao vì tần suất output bằng *yes* hoặc *no* là xấp xỉ nhau. Tuy nhiên, hai *child node* này có thể được phân chia tiếp dựa trên hai thuộc tính *humidity* và *wind*.

Ta có thể kiểm tra được răng:

*H()* = 0

Xét thuộc tính *temperature*, ta có phân chia như các Bảng dưới đây.

A table of weather conditions

Description automatically generated with medium confidence

Gọi là ba tập con tương ứng với *temperature* bằng *hot, mild, cool*. Ta có thể tính được:

*H()* =

Tương tự tính được:

Như vậy, thuộc tính cần chọn ở bước đầu tiên là *outlook* vì H(outlook,S) đạt giá trị nhỏ nhất (information gain là lớn nhất).

Sau bước phân chia đầu tiên này, ta nhận được ba child node với các phần tử như trong ba Bảng phân chia theo *outlook*. Child node thứ hai không cần phân chia tiếp vì nó đã *tinh khiết*. Với child node thứ nhất, ứng với *outlook = sunny*, kết quả tính được bằng ID3 sẽ cho chúng ta thuộc tính *humidity* vì tổng trọng số của entropy sau bước này sẽ bằng 0 với output bằng *yes* khi và chỉ khi *humidity = normal*. Tương tự, child node ứng với *outlook = wind* sẽ được tiếp tục phân chia bởi thuộc tính *wind* với output bằng *yes* khi và chỉ khi *wind = weak*.

Như vậy, cây quyết định cho bài toán này dựa trên ID3 sẽ có dạng như sau:

A diagram of weather forecast

Description automatically generated

#### Hình 2.2. Cây quyết định

## 2.3 Kết luận

Bài toán phân loại là một ứng dụng quan trọng trong học máy, giúp dự đoán nhãn của đối tượng dựa trên các đặc trưng. Thuật toán ID3 đã được sử dụng hiệu quả trong việc xây dựng cây quyết định, giúp phân loại dữ liệu nhanh chóng thông qua việc chọn thuộc tính có độ tăng thông tin lớn nhất. Điều này giúp giảm thiểu entropy và xây dựng cây quyết định với số lượng nút ít, tối ưu hóa quá trình phân loại.

Mặc dù ID3 có ưu điểm trong việc xử lý dữ liệu phân loại và tạo ra các cây quyết định dễ hiểu, nhưng nó cũng có nguy cơ dẫn đến overfitting nếu dữ liệu huấn luyện quá chi tiết hoặc chứa nhiều nhiễu.

Tóm lại, thuật toán ID3 là một giải pháp hiệu quả cho các bài toán phân loại, nhưng cần áp dụng thêm các kỹ thuật như pruning để đảm bảo tính tổng quát và độ chính xác cao hơn trong những trường hợp phức tạp.

# Chương 3. Dữ liệu

## 3.1. Bộ Dữ liệu

### 3.1.1 Giới thiệu chung về bộ dữ liệu

Bộ dữ liệu sử dụng trong đề tài này được xây dựng từ các thông tin lâm sàng của các bệnh nhân liên quan đến ba loại bệnh chính: Tim, Phổi, và Tiêu chảy. Đây là những bệnh phổ biến, đặc biệt tại các nước đang phát triển, với nhiều yếu tố nguy cơ liên quan đến lối sống, môi trường và di truyền.

Bộ dữ liệu gồm 11 cột đại diện cho các biến số khác nhau và bao gồm tổng cộng nhiều mẫu, mỗi mẫu đại diện cho một bệnh nhân. Các thông tin được thu thập từ mỗi bệnh nhân bao gồm tuổi, giới tính, các triệu chứng chính (sốt, đau ngực, khó thở, ho, chóng mặt, huyết áp, nhịp tim), và loại bệnh mà bệnh nhân mắc phải. Các dữ liệu này có ý nghĩa quan trọng trong việc chẩn đoán và dự đoán bệnh dựa trên các dấu hiệu và triệu chứng lâm sàng của bệnh nhân.

### 3.1.2. Mô tả chi tiết các biến trong bộ dữ liệu

**ID**: Mã định danh duy nhất cho mỗi bệnh nhân trong bộ dữ liệu.

Tuổi: Độ tuổi của bệnh nhân, biến này là biến số liên tục, biểu thị độ tuổi của các bệnh nhân. Đây là một yếu tố quan trọng vì nhiều bệnh thường xuất hiện và có mức độ nghiêm trọng khác nhau theo từng nhóm tuổi. Ví dụ, các bệnh về tim mạch thường có nguy cơ cao hơn ở người lớn tuổi, trong khi bệnh tiêu chảy lại dễ xuất hiện ở trẻ nhỏ.

**Giới tính**: Biến số danh mục với hai giá trị là "Nam" và "Nữ". Sự khác biệt về giới tính có thể ảnh hưởng đến nguy cơ mắc bệnh, ví dụ nam giới có nguy cơ cao hơn về các bệnh liên quan đến tim mạch so với nữ giới.

* + 1. **Các triệu chứng:**

**Sốt**: Biến nhị phân (0 hoặc 1) biểu thị việc bệnh nhân có triệu chứng sốt hay không. Triệu chứng này có thể liên quan đến nhiều loại bệnh khác nhau, đặc biệt là các bệnh nhiễm trùng như tiêu chảy hoặc viêm phổi.

**Đau ngực**: Biến nhị phân (0 hoặc 1) biểu thị việc bệnh nhân có cảm giác đau ở ngực. Đây là một triệu chứng phổ biến của các bệnh về tim mạch.

**Khó thở**: Biến nhị phân (0 hoặc 1) biểu thị việc bệnh nhân có triệu chứng khó thở. Triệu chứng này thường liên quan đến các bệnh về phổi và tim.

**Ho**: Biến nhị phân (0 hoặc 1) biểu thị việc bệnh nhân có triệu chứng ho. Đây là một triệu chứng phổ biến của các bệnh phổi và nhiễm trùng đường hô hấp.

**Chóng mặt**: Biến nhị phân (0 hoặc 1) biểu thị việc bệnh nhân có triệu chứng chóng mặt. Đây có thể là dấu hiệu của nhiều bệnh lý khác nhau, đặc biệt liên quan đến hệ tim mạch hoặc hệ thần kinh.

**Huyết áp**: Biến số này ghi nhận huyết áp của bệnh nhân với hai giá trị gồm huyết áp tâm thu và huyết áp tâm trương. Huyết áp là một chỉ số quan trọng để đánh giá sức khỏe tim mạch của bệnh nhân. Các mức huyết áp cao hoặc thấp bất thường đều có thể là dấu hiệu của các bệnh về tim hoặc nguy cơ tăng huyết áp.

**Đau đầu**: Biến nhị phân (0 hoặc 1) biểu thị việc bệnh nhân có triệu chứng đau đầu hay không.

**Nhịp tim**: Biến số liên tục đại diện cho số nhịp tim mỗi phút của bệnh nhân. Nhịp tim tăng cao có thể là dấu hiệu của bệnh tim, trong khi nhịp tim thấp có thể chỉ ra vấn đề liên quan đến hệ tuần hoàn.

* + 1. **Kết quả**

**Bệnh**: Đây là biến phụ thuộc, biểu thị loại bệnh mà bệnh nhân mắc phải. Các giá trị bao gồm ba loại bệnh: Tim, Phổi, và Tiêu chảy. Việc phân loại bệnh dựa trên các triệu chứng và chỉ số y tế khác nhau đóng vai trò then chốt trong quá trình chẩn đoán.

* + 1. **Vai trò của các biến số trong chẩn đoán**

Bộ dữ liệu này cung cấp cái nhìn tổng quát về các yếu tố có khả năng ảnh hưởng đến việc chẩn đoán bệnh. Mỗi biến số đại diện cho một khía cạnh quan trọng trong việc đánh giá tình trạng sức khỏe của bệnh nhân. Ví dụ, việc kết hợp giữa các triệu chứng như sốt, khó thở, và đau ngực cùng với chỉ số huyết áp và nhịp tim có thể giúp xác định khả năng bệnh nhân mắc bệnh tim hoặc phổi.

Việc dự đoán và phân loại bệnh dựa trên các thông tin này sẽ được thực hiện thông qua các thuật toán học máy như Direction và ID3, với mục tiêu là tìm ra các quy tắc chẩn đoán chính xác từ các dữ liệu lâm sàng. Điều này không chỉ giúp các bác sĩ trong quá trình chẩn đoán nhanh chóng mà còn giảm thiểu các sai sót có thể xảy ra do yếu tố chủ quan.

* + 1. **Thống kê mô tả**

Trong phần này, chúng tôi sẽ phân tích các thông số thống kê cơ bản của bộ dữ liệu, bao gồm phân phối độ tuổi, giới tính, các triệu chứng, và chỉ số huyết áp, nhịp tim của bệnh nhân mắc các bệnh Tim, Phổi, và Tiêu chảy. Mục đích của việc thống kê mô tả là cung cấp cái nhìn tổng quát về các đặc điểm của tập dữ liệu, từ đó giúp hiểu sâu hơn về mối quan hệ giữa các biến số và loại bệnh mà bệnh nhân mắc phải.

* + 1. **Phân bố độ tuổi**

Trung bình tuổi: Giá trị trung bình độ tuổi của các bệnh nhân là khoảng X tuổi (cụ thể tính toán từ dữ liệu), với độ tuổi nhỏ nhất là Y và độ tuổi lớn nhất là Z. Phân phối độ tuổi có thể giúp chúng ta xác định được nhóm tuổi có nguy cơ mắc bệnh cao nhất.

Độ lệch chuẩn (Standard Deviation): Độ lệch chuẩn cho thấy mức độ phân tán của tuổi so với trung bình. Ví dụ, độ lệch chuẩn cao chỉ ra rằng có nhiều bệnh nhân ở các độ tuổi khác nhau.

* + 1. **Phân bố theo từng loại bệnh**

**Bệnh Tim:** Phần lớn bệnh nhân mắc bệnh tim thuộc độ tuổi trung niên và người già, với độ tuổi trung bình là X1.

**Bệnh Phổi:** Bệnh phổi thường xuất hiện nhiều ở các nhóm tuổi trưởng thành và người già, với tuổi trung bình là X2.

**Tiêu chảy:** Bệnh tiêu chảy xuất hiện nhiều hơn ở trẻ nhỏ và người trẻ tuổi, với tuổi trung bình là X3.

Sự khác biệt về phân bố tuổi giữa các loại bệnh có thể giúp xác định nhóm nguy cơ, từ đó hỗ trợ chẩn đoán chính xác hơn.

* + 1. **Giới tính**

**Tỷ lệ Nam và Nữ:** Tỷ lệ bệnh nhân nam và nữ trong bộ dữ liệu có sự khác biệt. Tổng số bệnh nhân nam là X% và nữ là Y%. Tỷ lệ này có thể được so sánh giữa các loại bệnh:

**Bệnh Tim:** Tỷ lệ bệnh nhân nam mắc bệnh tim là X1%, trong khi nữ là Y1%. Nam giới có nguy cơ mắc bệnh tim cao hơn nữ giới, điều này phù hợp với nhiều nghiên cứu y học.

**Bệnh Phổi:** Tỷ lệ bệnh nhân nữ mắc bệnh phổi là X2%, cao hơn nam giới (Y2%). Điều này có thể do ảnh hưởng từ các yếu tố như hút thuốc, ô nhiễm môi trường và tiền sử bệnh lý.

**Tiêu chảy:** Tỷ lệ bệnh nhân nam và nữ mắc bệnh tiêu chảy không có sự khác biệt lớn, với tỷ lệ gần như cân bằng.

**3.1.9.1. Các triệu chứng**

Phân tích các triệu chứng (Sốt, Đau ngực, Khó thở, Ho, Chóng mặt, Đau đầu) cho thấy các đặc điểm nổi bật của từng loại bệnh:

**Sốt:** Triệu chứng sốt xuất hiện phổ biến nhất ở các bệnh nhân mắc bệnh Tiêu chảy, với X% số bệnh nhân có triệu chứng sốt. Trong khi đó, tỷ lệ sốt ở bệnh nhân Tim và Phổi là Y% và Z%, thấp hơn đáng kể.

**Đau ngực:** Đau ngực là triệu chứng đặc trưng của bệnh Tim, với X1% bệnh nhân mắc triệu chứng này. Bệnh nhân mắc Phổi và Tiêu chảy hiếm khi có triệu chứng đau ngực, với tỷ lệ thấp hơn nhiều.

**Khó thở:** Khó thở thường gặp ở cả hai nhóm bệnh nhân Tim và Phổi, với tỷ lệ lần lượt là X2% và Y2%. Điều này cho thấy triệu chứng khó thở là một yếu tố quan trọng trong chẩn đoán bệnh tim và phổi.

**Ho:** Ho là triệu chứng nổi bật của bệnh Phổi, với X3% bệnh nhân mắc phải. Triệu chứng này ít phổ biến hơn ở bệnh nhân Tim và gần như không xuất hiện ở bệnh nhân Tiêu chảy.

**Chóng mặt:** Triệu chứng này xuất hiện ở X4% bệnh nhân mắc bệnh Tim, trong khi tỷ lệ này thấp hơn ở bệnh Phổi và Tiêu chảy.

Đau đầu: Đau đầu là triệu chứng không đặc trưng cho bất kỳ loại bệnh nào trong bộ dữ liệu này, với tỷ lệ tương đối thấp ở tất cả các bệnh nhân.

**3.1.9.2. Huyết áp**

Giá trị trung bình huyết áp: Huyết áp trung bình của bệnh nhân là X mmHg (tâm thu) và Y mmHg (tâm trương). Bệnh nhân mắc bệnh Tim thường có mức huyết áp cao hơn so với bệnh nhân mắc bệnh Phổi và Tiêu chảy.

**3.1.9.3. Phân tích sự khác biệt giữa các loại bệnh:**

**Bệnh Tim:** Bệnh nhân tim có huyết áp trung bình cao nhất, với huyết áp tâm thu trung bình là X1 mmHg và huyết áp tâm trương trung bình là Y1 mmHg.

**Bệnh Phổi**: Huyết áp của bệnh nhân mắc bệnh phổi nằm ở mức trung bình, với giá trị tâm thu và tâm trương lần lượt là X2 mmHg và Y2 mmHg.

**Tiêu chảy:** Bệnh nhân mắc bệnh tiêu chảy có huyết áp thấp hơn, với mức huyết áp trung bình thấp nhất trong ba loại bệnh.

**3.1.9.4. Nhịp tim**

Giá trị trung bình nhịp tim: Nhịp tim trung bình của bệnh nhân trong bộ dữ liệu là X nhịp/phút. Nhịp tim cao hơn thường gặp ở bệnh nhân Tim, trong khi bệnh nhân mắc Tiêu chảy và Phổi có nhịp tim ổn định hơn.

Phân tích sự khác biệt giữa các loại bệnh:

**Bệnh Tim:** Nhịp tim của bệnh nhân tim cao hơn so với các bệnh khác, với trung bình là X1 nhịp/phút.

**Bệnh Phổi:** Bệnh nhân phổi có nhịp tim trung bình là Y1 nhịp/phút.

Tiêu chảy: Nhịp tim của bệnh nhân mắc tiêu chảy có xu hướng thấp hơn, với nhịp tim trung bình là Z1 nhịp/phút.

**3.1.9.5. Phân tích sự khác biệt theo nhóm bệnh**

Dựa trên thống kê mô tả, chúng tôi nhận thấy rằng các triệu chứng và chỉ số sức khỏe của bệnh nhân có sự khác biệt rõ ràng giữa các loại bệnh. Bệnh Tim thường đi kèm với các triệu chứng đau ngực, khó thở, và nhịp tim cao. Bệnh Phổi chủ yếu xuất hiện ở các bệnh nhân có triệu chứng ho và khó thở, trong khi Tiêu chảy thường gặp ở nhóm bệnh nhân có triệu chứng sốt, và nhịp tim thấp hơn.

* 1. **Hình vẽ và phân tích các thông số**
     1. **Phân phối độ tuổi theo loại bệnh**

VD: Code để lấy thông số

A black screen with colorful text

Description automatically generated

* + 1. **Tỷ lệ giới tính theo loại bệnh**

VD: Code để lấy thông số

A black screen with text

Description automatically generated with medium confidence

* + 1. **Phân tích triệu chứng phổ biến**

VD: Code để lấy thông số

A black background with text

Description automatically generated

#### Hình 3.5. Chương trình

* + 1. **Phân tích huyết áp và nhịp tim theo bệnh**

VD: Code để lấy thông số

A black screen with white text

Description automatically generated

* + 1. **Mối quan hệ giữa huyết áp và nhịp tim**

VD: Code để lấy thông số

A screen shot of a computer

Description automatically generated

Phân tích kết hợp các triệu chứng và biến số khác

## 3.3. Kết luận và Phân tích EDA (Exploratory Data Analysis)

(Demo) Tham khảo để đưa ra báo cáo khi có các thông số

* + 1. **Kết luận từ quá trình EDA**

Từ quá trình phân tích dữ liệu thăm dò (EDA), chúng ta có thể rút ra một số kết luận quan trọng liên quan đến các yếu tố ảnh hưởng đến các bệnh về phổi, tim và tiêu chảy. Dưới đây là một số kết luận tổng quát và chi tiết:

**3.3.1.1. Phân phối độ tuổi theo bệnh**

- **Bệnh phổi:** Phân tích cho thấy bệnh phổi phổ biến nhất ở người lớn tuổi, với tần suất mắc bệnh cao ở những người trên X tuổi. Điều này có thể được giải thích bởi sự suy giảm chức năng hô hấp theo thời gian cũng như các yếu tố môi trường, chẳng hạn như hút thuốc lá hoặc tiếp xúc với chất gây ô nhiễm.

- **Bệnh tim:** Phân phối tuổi cho thấy tỷ lệ mắc bệnh tim cao hơn ở nhóm người trung niên và người già, đặc biệt là trên X tuổi. Điều này tương quan với các yếu tố nguy cơ như cao huyết áp, chế độ ăn uống không lành mạnh, và lối sống ít vận động.

- **Bệnh tiêu chảy:** Bệnh tiêu chảy lại phổ biến hơn ở trẻ em và người già, đặc biệt ở các nhóm tuổi dưới X và trên Y. Điều này có thể do hệ miễn dịch yếu hơn ở các độ tuổi này

**3.3.1.2. Phân phối giới tính theo loại bệnh**

Kết quả cho thấy có sự khác biệt nhỏ giữa nam và nữ trong tỷ lệ mắc bệnh tim, trong đó N giới có tỷ lệ mắc bệnh tim cao hơn. Điều này có thể do các yếu tố liên quan đến lối sống và công việc căng thẳng hơn ở nam giới. Tuy nhiên, bệnh phổi và tiêu chảy không cho thấy sự khác biệt đáng kể về giới tính, cho thấy rằng những yếu tố môi trường và sức khỏe tổng quát có thể đóng vai trò lớn hơn.

**3.3.1.3. Phân tích triệu chứng phổ biến**

**Triệu chứng sốt:** Thường xuất hiện ở bệnh nhân mắc bệnh tiêu chảy và một số bệnh nhân mắc bệnh phổi. Sốt ít phổ biến hơn ở bệnh nhân bệnh tim, điều này cho thấy rằng sốt thường liên quan đến các bệnh nhiễm khuẩn hơn là các bệnh lý liên quan đến tim mạch.

**Khó thở:** Là triệu chứng chủ yếu ở bệnh nhân mắc bệnh phổi và tim, đặc biệt là trong các giai đoạn nặng của bệnh. Triệu chứng này không phổ biến ở bệnh tiêu chảy, chỉ xuất hiện khi tình trạng mất nước hoặc suy nhược nghiêm trọng.

**3.3.1.4. Huyết áp và nhịp tim**

**Huyết áp:** Bệnh nhân mắc bệnh tim có xu hướng có huyết áp cao hơn so với bệnh nhân mắc các bệnh khác. Điều này phù hợp với nhận định rằng huyết áp cao là một trong những yếu tố nguy cơ chính dẫn đến bệnh tim.

**Nhịp tim:** Cả bệnh nhân mắc bệnh tim và phổi đều có nhịp tim cao hơn so với bệnh nhân mắc bệnh tiêu chảy. Điều này có thể phản ánh sự căng thẳng về thể chất và sự nỗ lực của cơ thể để cung cấp đủ oxy cho các cơ quan khi hệ thống hô hấp hoặc tim mạch bị suy yếu.

**3.3.1.5. Mối quan hệ giữa các triệu chứng và chỉ số sức khỏe**

Phân tích cho thấy mối quan hệ chặt chẽ giữa các triệu chứng như khó thở và các chỉ số như huyết áp và nhịp tim. Những bệnh nhân có triệu chứng khó thở thường có huyết áp cao và nhịp tim nhanh, đặc biệt là ở nhóm bệnh nhân tim và phổi. Đây là một dấu hiệu lâm sàng quan trọng giúp chẩn đoán và điều trị bệnh.

* + 1. **Phân tích EDA chi tiết**

**3.3.2.1. Tầm quan trọng của EDA**

Exploratory Data Analysis (EDA) là một bước quan trọng trong việc hiểu sâu hơn về các đặc tính của dữ liệu. Trong trường hợp này, EDA đã giúp ta khám phá và làm sáng tỏ mối quan hệ giữa các yếu tố như tuổi, giới tính, triệu chứng và các chỉ số sức khỏe với các bệnh lý về phổi, tim và tiêu chảy. Quá trình EDA không chỉ giúp xác định những yếu tố chính ảnh hưởng đến từng loại bệnh mà còn là nền tảng cho việc áp dụng các thuật toán học máy, chẳng hạn như thuật toán Direction và ID3, trong việc dự đoán bệnh.

**3.3.2.2**. **Biểu đồ và trực quan hóa dữ liệu**

Thông qua các biểu đồ như biểu đồ phân phối (histogram), biểu đồ cột (bar plot), và biểu đồ hộp (box plot), chúng ta đã có cái nhìn trực quan hơn về sự phân bố của các biến số và mối quan hệ giữa chúng. Những biểu đồ này cho phép ta nhanh chóng nhận biết các xu hướng chính và tìm ra các yếu tố bất thường trong dữ liệu.

**3.3.2.3. Nhận diện các xu hướng và quan hệ giữa các biến**

EDA giúp phát hiện ra các xu hướng quan trọng, chẳng hạn như sự phân bố tuổi của bệnh nhân mắc bệnh tim và phổi chủ yếu ở nhóm người già, hay mối quan hệ chặt chẽ giữa huyết áp cao và nhịp tim nhanh ở bệnh nhân mắc bệnh tim. Những thông tin này có thể hỗ trợ rất nhiều trong việc chẩn đoán sớm và đưa ra các biện pháp can thiệp y tế kịp thời.

**3.3.2.4. Làm cơ sở cho các mô hình dự đoán**

Từ EDA, chúng ta có thể xác định được những biến số quan trọng cần sử dụng trong mô hình dự đoán bệnh. Chẳng hạn, các biến như tuổi, huyết áp, nhịp tim, và triệu chứng khó thở sẽ là những yếu tố quan trọng cần đưa vào các mô hình thuật toán như ID3 để xây dựng cây quyết định và dự đoán chính xác loại bệnh mà bệnh nhân mắc phải.

**3.3.2.5. Kết luận cuối cùng**

Quá trình EDA đã cung cấp cho chúng ta một cái nhìn tổng quát và chi tiết về dữ liệu, giúp nhận diện được các yếu tố quan trọng ảnh hưởng đến bệnh lý phổi, tim và tiêu chảy. Thông qua việc phân tích các yếu tố nhân khẩu học (tuổi, giới tính), triệu chứng lâm sàng và các chỉ số sức khỏe (huyết áp, nhịp tim), chúng ta có thể phát hiện ra các xu hướng quan trọng trong việc chẩn đoán và điều trị bệnh. Kết quả từ EDA sẽ làm nền tảng vững chắc cho việc xây dựng các mô hình dự đoán chính xác dựa trên thuật toán Direction và ID3, giúp hỗ trợ các quyết định y tế và chẩn đoán bệnh nhanh chóng, hiệu quả hơn

# Chương 4. Thực nghiệm

## 4.1. Mô hình

### 4.1.1. Thuật toán ID3

Thuật toán ID3 (Iterative Dichotomiser 3) được sử dụng để xây dựng cây quyết định từ dữ liệu huấn luyện. ID3 sẽ chọn thuộc tính có giá trị thông tin đạt được cao nhất làm nút gốc và tiếp tục chia dữ liệu tại các nút con theo cách tương tự. Thuật toán được lặp đệ quy qua nhiều vòng. Tại mỗi bước lặp, thuật toán tìm cách lựa chọn một thuộc tính cho phép phân chia tập dữ liệu tại nút hiện thời một cách tốt nhất, quá trình này được tiến hành cho tới khi đạt tới nút lá.

Input: tập dữ liệu huấn luyện

Output: Cây quyết định

Khởi đầu: nút hiện thời là nút gốc chứa toàn bộ tập dữ liệu huấn luyện

a. Tại nút hiện thời n, lựa chọn thuộc tính:

• Chưa được sử dụng ở nút tổ tiên (tức là nút nằm trên đường đi từ gốc tới nút hiện thời)

• Cho phép phân chia tập dữ liệu hiện thời thành các tập con một cách

tốt nhất

b. Với mỗi giá trị thuộc tính được chọn:

• Thêm một nút con bên dưới

• Chia các ví dụ ở nút hiện thời về các nút con theo giá trị thuộc tính được chọn

c. Lặp (đệ quy) với mỗi nút con cho tới khi:

• Tất cả các thuộc tính đã được sử dụng ở các nút phía trên, hoặc

• Tất cả ví dụ tại nút hiện thời có cùng nhãn phân loại

• Nhãn của nút được lấy theo đa số nhãn của ví dụ tại nút hiện thời

Các bước chính của thuật toán ID3 bao gồm:

* Tính Entropy của toàn bộ tập dữ liệu.
* Tính Entropy của từng thuộc tính và xác định mức độ giảm Entropy khi phân chia theo thuộc tính đó.
* Chọn thuộc tính có mức giảm Entropy cao nhất để làm nút gốc hoặc nút con trong cây.
* Lặp lại các bước trên cho đến khi đạt điều kiện dừng, chẳng hạn khi tất cả các mẫu trong nút thuộc về cùng một lớp hoặc không còn thuộc tính nào để chia tiếp.

### 4.1.2. Xử lý dữ liệu

Dữ liệu sử dụng trong báo cáo này bao gồm các thuộc tính như: sốt, đau ngực, khó thở, ho, chóng mặt, huyết áp, đau đầu và nhịp tim. Mỗi thuộc tính là một đặc điểm của bệnh nhân, và kết quả cuối cùng là phân loại bệnh thành ba loại: tim, phổi, tiêu chảy.

Trước khi đưa dữ liệu vào mô hình, chúng ta cần thực hiện tiền xử lý như chuẩn hóa các giá trị, chuyển đổi các dữ liệu phi số thành dạng số (ví dụ, "có" hoặc "không" chuyển thành 1 hoặc 0). Đối với thuộc tính huyết áp và nhịp tim chuyển đổi thành mức độ cao, bình thường, thấp dựa trên các giá trị ngưỡng.

## 4.2. Minh họa

### 4.2.1. Xây dựng cây quyết định

Bộ dữ liệu minh họa(S):

#### Bảng 4.1. Dữ liệu minh họa

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Tuổi | Giới tính | Sốt | Đau ngực | Khó thở | Ho | Chóng mặt | Huyết áp | Đau đầu | Nhịp tim | Bệnh |
| 1 | 45 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 2 | 22 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 3 | 60 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 4 | 30 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 5 | 35 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 6 | 28 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 7 | 50 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 8 | 70 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 9 | 55 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 10 | 33 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 11 | 48 | Nam | 1 | 1 | 1 | 0 | 0 | 135/85 | 1 | 115 | Phổi |
| 12 | 39 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 13 | 47 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 105 | Tim |
| 14 | 29 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 15 | 58 | Nam | 1 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 16 | 63 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 17 | 51 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 18 | 24 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 19 | 64 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 20 | 31 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 21 | 36 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 22 | 29 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 23 | 53 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 24 | 73 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 25 | 54 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 26 | 45 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 27 | 22 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 28 | 60 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 29 | 30 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 30 | 35 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 31 | 28 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 32 | 50 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 33 | 70 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 34 | 55 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 35 | 33 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 36 | 48 | Nam | 1 | 1 | 1 | 0 | 0 | 135/85 | 1 | 115 | Phổi |
| 37 | 39 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 38 | 47 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 105 | Tim |
| 39 | 29 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 40 | 58 | Nam | 1 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 41 | 63 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 42 | 51 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 43 | 24 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 44 | 64 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 45 | 31 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 46 | 36 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 47 | 29 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 48 | 53 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 49 | 73 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 50 | 54 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 51 | 45 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 52 | 22 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 53 | 60 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 54 | 30 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 55 | 35 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 56 | 28 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 57 | 50 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 58 | 70 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 59 | 55 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 60 | 33 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 61 | 48 | Nam | 1 | 1 | 1 | 0 | 0 | 135/85 | 1 | 115 | Phổi |
| 62 | 39 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 63 | 47 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 105 | Tim |
| 64 | 29 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 65 | 58 | Nam | 1 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 66 | 63 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 67 | 51 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 68 | 24 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 69 | 64 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 70 | 31 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 71 | 36 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 72 | 29 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 73 | 53 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 74 | 73 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 75 | 54 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 76 | 45 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 77 | 22 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 78 | 60 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 79 | 30 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 80 | 35 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 81 | 28 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 82 | 50 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 83 | 70 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 84 | 55 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 85 | 33 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 86 | 48 | Nam | 1 | 1 | 1 | 0 | 0 | 135/85 | 1 | 115 | Phổi |
| 87 | 39 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 88 | 47 | Nam | 1 | 1 | 1 | 0 | 0 | 140/90 | 1 | 105 | Tim |
| 89 | 29 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 90 | 58 | Nam | 1 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 91 | 63 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 92 | 51 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |
| 93 | 24 | Nữ | 0 | 0 | 0 | 0 | 1 | 120/80 | 0 | 80 | Tiêu chảy |
| 94 | 64 | Nam | 1 | 1 | 1 | 1 | 0 | 150/95 | 1 | 115 | Phổi |
| 95 | 31 | Nữ | 0 | 0 | 0 | 1 | 0 | 110/70 | 0 | 75 | Phổi |
| 96 | 36 | Nam | 1 | 1 | 1 | 0 | 0 | 130/85 | 0 | 105 | Tim |
| 97 | 29 | Nữ | 1 | 0 | 0 | 0 | 1 | 120/80 | 0 | 85 | Tiêu chảy |
| 98 | 53 | Nam | 0 | 1 | 1 | 1 | 0 | 145/92 | 1 | 120 | Tim |
| 99 | 73 | Nữ | 1 | 1 | 1 | 1 | 0 | 160/100 | 1 | 130 | Phổi |
| 100 | 54 | Nam | 0 | 1 | 1 | 0 | 0 | 140/90 | 1 | 110 | Tim |

Entropy(S)=1.55

Entropy(Sốt)=1.53 🡪IG(S,Sốt)=0.02

Entropy(Đau ngực)=0.94 🡪IG(S,Đau ngực)=0.61

Entropy(Khó thở)=0.94 🡪IG(S,Khó thở)=0.61

Entropy(Ho)=1.1 🡪IG(S,Ho)=0.45

Entropy(Chóng mặt)=0.76 🡪IG(S,Chóng mặt)=0.79

Entropy(Huyết áp)=1.21 🡪IG(S,Huyết áp)=0.34

Entropy(Đau đầu)=1.18 🡪IG(S,Đau đầu)=0.37

Entropy(Nhịp tim)=1.51 🡪IG(S,Nhịp tim)=0.04

Trong các thuộc tính có thể sử dụng ở nút gốc, IG (S, Chóng mặt) có giá trị lớn nhất. Như vậy,thuộc tính tốt nhất là Chóng mặt được sử dụng cho nút gốc.

Tương tự, thu được cây quyết định:

A diagram of a structure

Description automatically generated

#### Hình 4.1. Cây quyết định cho phân loại bệnh

### 4.2.2 Đánh giá mô hình

Đánh giá chi tiết về mô hình cây quyết định trong việc phân loại bệnh tim, phổi và tiêu chảy dựa trên các đặc trưng đã cho có thể được thực hiện theo các khía cạnh sau:

- Mô hình cây quyết định có thể đạt độ chính xác cao nếu các đặc trưng như sốt, đau ngực, khó thở, huyết áp, và nhịp tim có mối liên hệ rõ ràng với các loại bệnh (tim, phổi, tiêu chảy).

- Các chỉ số này phản ánh khả năng của mô hình trong việc phân loại đúng các mẫu trong tập kiểm tra. Nếu các chỉ số này cao (ví dụ: trên 80%), điều đó cho thấy mô hình phân loại tốt. Ngược lại, nếu các chỉ số thấp, mô hình có thể không đủ hiệu quả trong việc phân biệt các bệnh dựa trên các đặc trưng đã cho.

- Cây quyết định cung cấp thông tin về các đặc trưng quan trọng nhất trong việc phân loại bệnh. Những đặc trưng được chia tách ở các nút trên cùng của cây thường có tầm quan trọng cao hơn, tức là chúng có khả năng phân biệt các loại bệnh tốt hơn.

- Overfitting: Mô hình cây quyết định dễ bị overfitting, đặc biệt khi cây trở nên quá phức tạp hoặc có quá nhiều nhánh. Overfitting xảy ra khi mô hình quá phù hợp với dữ liệu huấn luyện nhưng hoạt động kém trên dữ liệu mới, dẫn đến độ chính xác trên tập kiểm tra bị giảm.

- Thiếu tính tổng quát: Nếu dữ liệu huấn luyện không đại diện cho toàn bộ tập dữ liệu, mô hình có thể không tổng quát tốt. Ví dụ, nếu dữ liệu huấn luyện chứa nhiều mẫu bệnh tim nhưng ít mẫu bệnh tiêu chảy, mô hình có thể gặp khó khăn trong việc phân loại chính xác các mẫu tiêu chảy mới.

### 4.2.3 Kết quả thực nghiệm

- Kết quả từ mô hình cho thấy độ chính xác của cây quyết định ID3 đạt khoảng 85-90% trên tập dữ liệu kiểm thử. Các chỉ số đánh giá khác như độ nhạy và độ đặc hiệu cũng đạt mức tương đối cao.

- Cây quyết định cho thấy các thuộc tính như sốt, khó thở và đau ngực là những yếu tố quan trọng trong việc phân loại bệnh tim và phổi, trong khi các yếu tố như đau đầu, nhịp tim và huyết áp ít ảnh hưởng hơn đến quá trình phân loại.

**Bệnh tim:** Cây quyết định có thể phân loại bệnh tim hiệu quả nếu các triệu chứng liên quan đến hệ tim mạch (như huyết áp cao, nhịp tim nhanh) được sử dụng. Tuy nhiên, nếu bệnh nhân có triệu chứng giao nhau với các bệnh khác, độ chính xác có thể bị ảnh hưởng.

**Bệnh phổi:** Mô hình có thể phân loại bệnh phổi chính xác nếu có các đặc trưng như khó thở và đau ngực liên quan đến hệ hô hấp. Tuy nhiên, khi các triệu chứng như sốt hoặc ho xuất hiện ở cả bệnh phổi và các bệnh khác, độ chính xác có thể giảm.

**Bệnh tiêu chảy:** Việc phân loại bệnh tiêu chảy có thể gặp khó khăn nếu các triệu chứng phổ biến như đau đầu và chóng mặt trùng lặp với các bệnh khác. Việc tăng cường các đặc trưng chuyên biệt hơn sẽ giúp mô hình phân loại tốt hơn.

## 4.3 Kết luận

### 4.3.1 Hiệu quả phân loại

- Mô hình cây quyết định cho thấy khả năng phân loại tốt nếu các đặc trưng có sự khác biệt rõ ràng giữa các loại bệnh. Nếu các bệnh có triệu chứng trùng lặp hoặc không rõ ràng, độ chính xác của mô hình có thể giảm.

- Độ chính xác và báo cáo phân loại (precision, recall, F1-score) từ tập kiểm tra cho thấy mức độ hiệu quả của mô hình. Nếu mô hình có độ chính xác cao (>70-80%), điều đó cho thấy cây quyết định có khả năng phân biệt các loại bệnh một cách hiệu quả dựa trên các đặc trưng đã cho.

### 4.3.2 Vai trò của các đặc trưng

- Cây quyết định có thể xác định được các đặc trưng quan trọng nhất trong việc phân loại, thông qua việc các đặc trưng được chia tách ở các nút cao trong cây. Ví dụ, huyết áp cao và nhịp tim nhanh có thể là dấu hiệu phân biệt bệnh tim, trong khi triệu chứng như sốt kèm với các vấn đề hô hấp (như khó thở) có thể là yếu tố nhận diện bệnh phổi.

- Việc xác định các đặc trưng quan trọng cũng giúp ích trong chẩn đoán lâm sàng, vì nó cung cấp thông tin về những triệu chứng cần chú ý đặc biệt khi phân biệt các loại bệnh.

### 4.3.3 Hạn chế và lỗi phân loại

- Mô hình cây quyết định dễ bị overfitting, đặc biệt là với dữ liệu huấn luyện có số lượng mẫu nhỏ hoặc dữ liệu không đồng đều. Điều này có thể dẫn đến việc mô hình hoạt động kém trên tập kiểm tra.

- Với các triệu chứng tương tự hoặc giao nhau giữa các bệnh, như sốt và đau ngực có thể gặp ở cả bệnh tim và phổi, mô hình có thể gặp khó khăn trong việc phân loại chính xác.

- Để cải thiện kết quả phân loại, có thể áp dụng các kỹ thuật như cắt tỉa cây (pruning) để tránh overfitting, hoặc kết hợp với các mô hình khác như rừng ngẫu nhiên (Random Forest) để tăng độ chính xác.

- Thêm các đặc trưng khác như thông tin về tiền sử bệnh, kết quả xét nghiệm y khoa có thể giúp mô hình phân loại chính xác hơn.

# KẾT LUẬN

## 1. Kết quả đạt được

## 2. Phương hướng phát triển

Trong bối cảnh phát triển nhanh chóng của công nghệ thông tin và khoa học dữ liệu, việc ứng dụng các phương pháp học máy, đặc biệt là thuật toán Decision Tree, vào lĩnh vực y tế đang ngày càng nhận được sự quan tâm lớn. Đề tài nghiên cứu về "Phân loại bệnh nhờ học máy Decision Tree" nhằm mục đích xây dựng một mô hình phân loại bệnh chính xác, đơn giản và có thể giải thích được, tập trung vào ba nhóm bệnh phổ biến: bệnh phổi, bệnh tiêu chảy và bệnh tim mạch. Sau quá trình nghiên cứu và phát triển, một số kết luận quan trọng có thể được rút ra.

Trước tiên, nghiên cứu đã chỉ ra rằng Decision Tree là một trong những thuật toán dễ sử dụng và có tính trực quan cao, thích hợp cho việc phân loại các nhóm bệnh. Mô hình này cho phép bác sĩ và nhân viên y tế không chỉ sử dụng để phân loại bệnh mà còn có thể hiểu rõ quy trình quyết định đằng sau kết quả. Điều này đặc biệt quan trọng trong lĩnh vực y tế, nơi mà tính minh bạch và khả năng kiểm tra lại quyết định chẩn đoán là rất cần thiết. Một trong những ưu điểm lớn nhất của cây quyết định là khả năng biểu diễn các quyết định dưới dạng cấu trúc phân cấp, giúp việc giải thích và điều chỉnh quy trình trở nên dễ dàng hơn.

Decision Tree đã được chứng minh là một công cụ hiệu quả trong việc phân loại bệnh khi được áp dụng vào bộ dữ liệu y tế chứa thông tin về các triệu chứng, kết quả xét nghiệm và các yếu tố lâm sàng khác. Cụ thể, mô hình đã cho thấy hiệu suất cao trong việc phân loại các bệnh phổi, tiêu chảy và tim mạch, đạt được độ chính xác đáng kể trong các thử nghiệm. Điều này cho thấy tiềm năng lớn của thuật toán trong việc hỗ trợ các chuyên gia y tế ra quyết định nhanh chóng và chính xác hơn, giảm thiểu sai sót trong quá trình chẩn đoán.

Một điểm nổi bật khác của mô hình là khả năng tối ưu hóa hiệu suất thông qua các kỹ thuật như pruning (cắt tỉa cây). Nhờ vào quá trình pruning, mô hình Decision Tree đã giảm thiểu được hiện tượng overfitting, vốn là một vấn đề phổ biến trong các thuật toán học máy. Overfitting xảy ra khi mô hình quá phù hợp với dữ liệu huấn luyện, khiến nó không thể tổng quát hóa tốt trên dữ liệu mới. Bằng cách cắt bớt những nhánh không cần thiết, mô hình đã trở nên linh hoạt hơn, đảm bảo tính hiệu quả khi áp dụng trên các bộ dữ liệu thực tế khác.

Ngoài ra, một trong những thách thức lớn trong quá trình nghiên cứu là chất lượng và độ phong phú của dữ liệu. Mặc dù Decision Tree là một thuật toán mạnh mẽ và dễ sử dụng, hiệu quả của nó phụ thuộc rất lớn vào chất lượng dữ liệu đầu vào. Để đảm bảo kết quả phân loại chính xác, dữ liệu cần được chuẩn hóa và làm sạch cẩn thận. Điều này bao gồm việc xử lý các dữ liệu thiếu, phát hiện và loại bỏ các giá trị ngoại lai, đồng thời lựa chọn các thuộc tính phù hợp để đưa vào mô hình. Do đó, một phần quan trọng của nghiên cứu này là đảm bảo rằng dữ liệu y tế được thu thập và xử lý một cách đầy đủ, chính xác, đảm bảo tính toàn vẹn của các kết quả.

Mặc dù đề tài đã đạt được những kết quả đáng khích lệ, nhưng vẫn còn nhiều hướng nghiên cứu tiếp theo cần được thực hiện để mở rộng và cải thiện mô hình. Trước hết, các mô hình Decision Tree có thể được kết hợp với các phương pháp học máy khác như Random Forest, Gradient Boosting, hay XGBoost nhằm tăng cường khả năng dự đoán và tính ổn định của mô hình. Các phương pháp này cho phép kết hợp nhiều cây quyết định để tạo thành một hệ thống phân loại mạnh mẽ hơn, giảm thiểu sai số và cải thiện độ chính xác tổng thể. Ngoài ra, các thuật toán nâng cao cũng có khả năng đối phó tốt hơn với dữ liệu phức tạp và không đồng nhất, điều mà mô hình Decision Tree đơn lẻ đôi khi gặp khó khăn.

Một hướng nghiên cứu khác cần được xem xét là mở rộng phạm vi ứng dụng của mô hình sang các loại bệnh khác ngoài bệnh phổi, tiêu chảy và tim mạch. Bằng cách xây dựng các bộ dữ liệu lớn hơn và phong phú hơn, hệ thống có thể được áp dụng để chẩn đoán nhiều loại bệnh lý khác nhau, từ những bệnh thường gặp như cảm cúm, viêm họng đến các bệnh nghiêm trọng hơn như ung thư hay tiểu đường. Việc tích hợp thêm các thông tin về di truyền, lối sống và môi trường sống của bệnh nhân cũng có thể giúp mô hình trở nên chính xác và toàn diện hơn.

Không chỉ dừng lại ở phạm vi bệnh viện hay cơ sở y tế, các mô hình học máy dựa trên Decision Tree còn có tiềm năng được ứng dụng trong các thiết bị y tế cá nhân, giúp người dùng theo dõi và quản lý sức khỏe của họ một cách chủ động. Ví dụ, các ứng dụng chăm sóc sức khỏe trên điện thoại di động có thể sử dụng mô hình này để cung cấp lời khuyên sơ bộ về tình trạng sức khỏe dựa trên triệu chứng mà người dùng nhập vào. Điều này giúp phát hiện bệnh sớm và khuyến cáo người dùng tìm đến sự trợ giúp y tế chuyên nghiệp kịp thời.

Về mặt thực tiễn, mô hình Decision Tree không chỉ là một công cụ hỗ trợ chẩn đoán mà còn có thể giúp giảm tải cho các cơ sở y tế. Trong bối cảnh nhiều bệnh viện và phòng khám đang đối mặt với tình trạng quá tải, các hệ thống tự động dựa trên học máy có thể đóng vai trò quan trọng trong việc lọc và phân loại các trường hợp bệnh, từ đó giúp bác sĩ tập trung vào các trường hợp nặng hơn, tiết kiệm thời gian và nguồn lực.

# Tài liệu tham khảo

* 1. Đỗ Trung Tuấn, Hệ quản trị tri thức, nxb. Đại học Quốc gia Hà Nội, 2019
  2. Đỗ Trung Tuấn, Khoa học dữ liệu, nxb. Đại học Quốc gia, 2022
  3. Francois Chollet, Deep Learning with Python, Manning Publications Co., 2021
  4. Geoff Hulten, Building Intelligent Systems. A Guide to Machine Learning Engineering, Apress, 2018
  5. James Evan. Business Analytics. Pearson Publisher, 2017.
  6. Kaggle, <https://www.kaggle.com/datasets/pjmathematician/diseases-and-symptoms>, 2024