**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**---------------------------------------**

KHUẤT THỊ THANH HẢI – 2121051018

A blue circle with white text and a globe

Description automatically generated

|  |
| --- |
| **KHUẤT THỊ THANH HẢI - 2121051018** |
| **ĐỒ ÁN KHOA HỌC MÁY TÍNH** |
|  |
| TÊN ĐỀ TÀI: ÁP DỤNG THUẬT TOÁN ID3 KẾT HỢP VỚI PCA ĐỂ DỰ ĐOÁN KHẢ NĂNG BỊ BỆNH TIM |

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Hà Nội - Năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**---------------------------------------**

A blue circle with white text and a globe

Description automatically generated

**ĐỒ ÁN KHOA HỌC**

**MÁY TÍNH**

**CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH**

TÊN ĐỀ TÀI: ÁP DỤNG THUẬT TOÁN ID3 KẾT HỢP VỚI PCA ĐỂ DỰ ĐOÁN KHẢ NĂNG BỊ BỆNH TIM

|  |  |
| --- | --- |
| GIÁO VIÊN HƯỚNG DẪN | SINH VIÊN THỰC HIỆN |
| **Nguyễn Thị Phương Bắc** | **Khuất Thị Thanh Hải** |

**Hà Nội - Năm 2024**

# 

# LỜI NÓI ĐẦU

Trong bối cảnh y tế hiện đại, dự báo và chẩn đoán sớm các bệnh lý nguy hiểm, đặc biệt là bệnh tim, là một trong những ưu tiên hàng đầu. Bệnh tim mạch là nguyên nhân hàng đầu gây tử vong trên toàn cầu, và việc phát hiện sớm các yếu tố nguy cơ có thể giúp cứu sống hàng triệu người. Với sự phát triển mạnh mẽ của công nghệ thông tin và trí tuệ nhân tạo, các kỹ thuật học máy (Machine Learning) đang dần trở thành công cụ đắc lực trong việc hỗ trợ chẩn đoán y tế.

Bài tập lớn này tập trung vào việc ứng dụng Python và các kỹ thuật học máy để xây dựng một mô hình dự báo bệnh tim. Sử dụng các bộ dữ liệu y tế thực tế, chúng ta sẽ tiến hành các bước như tiền xử lý dữ liệu, lựa chọn đặc trưng, và triển khai các thuật toán học máy phổ biến để đưa ra dự báo về khả năng mắc bệnh tim của bệnh nhân. Mục tiêu của bài tập không chỉ là xây dựng mô hình chính xác mà còn là hiểu rõ quá trình triển khai và áp dụng các phương pháp học máy trong lĩnh vực y tế.

Thông qua bài tập này, chúng ta hy vọng sẽ có cái nhìn sâu sắc hơn về tiềm năng của học máy trong y học, đồng thời nắm bắt được các kỹ thuật và phương pháp cần thiết để phát triển các ứng dụng dự báo và hỗ trợ quyết định trong lĩnh vực chăm sóc sức khỏe.

# LỜI CẢM ƠN

Trước tiên, em xin gửi lời cảm ơn chân thành đến các thầy cô trong bộ môn Khoa học máy tính, đặc biệt là cô Nguyễn Thị Phương Bắc đã tận tình hướng dẫn, truyền đạt kiến thức và hỗ trợ em trong suốt quá trình thực hiện bài đồ án chuyên ngành này. Sự chỉ bảo tận tình của cô không chỉ giúp em hiểu sâu hơn về các khái niệm, kỹ thuật học máy mà còn giúp em định hướng và hoàn thiện ý tưởng cho đề tài.

Em cũng xin bày tỏ lòng biết ơn đến các thầy cô đã tạo điều kiện thuận lợi cho em tiếp cận với các tài liệu, công cụ cần thiết để hoàn thành bài tập. Sự động viên, góp ý quý báu từ cô là nguồn động lực lớn để em nỗ lực học hỏi và phát triển kỹ năng của mình.

Cuối cùng, em xin cảm ơn cô Nguyễn Thị Phương Bắc vì đã tạo ra một môi trường học tập đầy khuyến khích và thách thức, giúp em rèn luyện tư duy phân tích, kỹ năng giải quyết vấn đề và ứng dụng kiến thức vào thực tiễn. Những kinh nghiệm quý báu này sẽ là hành trang quan trọng cho em trên con đường học tập và sự nghiệp sau này.

Em xin kính chúc các thầy cô sức khỏe dồi dào và thành công trong sự nghiệp giáo dục.

# MỤC LỤC

Contents

[LỜI NÓI ĐẦU 3](#_Toc185796570)

[LỜI CẢM ƠN 4](#_Toc185796571)

[MỤC LỤC 5](#_Toc185796572)

[I. Chương 1: Giới Thiệu 7](#_Toc185796573)

[1.1 Tầm quan trọng của dự báo bệnh tim 7](#_Toc185796574)

[1.2 Vai trò của học máy trong y học 8](#_Toc185796575)

[1.3 Mục tiêu của đồ án 10](#_Toc185796576)

[1.3.1 Hiểu và phân tích dữ liệu y tế 10](#_Toc185796577)

[1.3.2 Áp dụng các thuật toán học máy để xây dựng mô hình dự báo 10](#_Toc185796578)

[1.3.3 Đánh giá và so sánh hiệu suất các mô hình 11](#_Toc185796579)

[1.3.4 Tối ưu hóa và cải thiện mô hình 11](#_Toc185796580)

[1.3.5 Rút ra kết luận và đề xuất hướng phát triển trong tương lai 11](#_Toc185796581)

[II. Chương 2: Tổng Quan Lý Thuyết 11](#_Toc185796582)

[2.1 Giới thiệu về bệnh tim và các yếu tố nguy cơ 11](#_Toc185796583)

[2.1.1 Giới thiệu về bệnh tim 12](#_Toc185796584)

[2.1.2 Các yếu tố nguy cơ của bệnh tim 12](#_Toc185796585)

[2.1.3 Tầm quan trọng của việc kiểm soát các yếu tố nguy cơ 13](#_Toc185796586)

[2.2 Tổng quan về các kỹ thuật học máy 14](#_Toc185796587)

[2.2.1 Học có giám sát (Supervised Learning) 14](#_Toc185796588)

[2.2.2 Học không giám sát (Unsupervised Learning) 15](#_Toc185796589)

[2.2.3 Học bán giám sát (Semi-Supervised Learning) 15](#_Toc185796590)

[2.2.4 Học tăng cường (Reinforcement Learning) 16](#_Toc185796591)

[2.2.5 Học sâu (Deep Learning) 16](#_Toc185796592)

[2.2.6 Kỹ thuật học máy tổng hợp (Ensemble Learning) 16](#_Toc185796593)

[2.3 Các mô hình học máy thường được sử dụng trong y học 17](#_Toc185796594)

[2.4 Các tiêu chuẩn đánh giá mô hình 19](#_Toc185796595)

[2.4.1 Độ chính xác (Accuracy) 20](#_Toc185796596)

[2.4.2 Độ nhạy (Recall) hay Tỉ lệ dương tính thực (True Positive Rate - TPR) 20](#_Toc185796597)

[2.4.3 Độ chính xác của dự đoán dương (Precision) 21](#_Toc185796598)

[2.4.4 Điểm F1 (F1 Score) 21](#_Toc185796599)

[III. Chương 3: Phương Pháp Nghiên Cứu 22](#_Toc185796600)

[3.1 Thu thập và xử lý dữ liệu 22](#_Toc185796601)

[3.1.1 Thu thập dữ liệu 22](#_Toc185796602)

[3.1.2 Xử lý dữ liệu(Data processing) 23](#_Toc185796603)

[3.1.3 Tiền xử lý dữ liệu(Data Preprocessing) 24](#_Toc185796604)

[3.2 Triển khai các mô hình học máy 25](#_Toc185796605)

[3.2.1. Lựa chọn mô hình học máy 25](#_Toc185796606)

[3.2.2 Lưu đồ thuật toán để huấn luyện mô hình 28](#_Toc185796607)

[3.2.3 Huấn luyện mô hình (Model Training) 30](#_Toc185796608)

[3.2.4 Đánh giá kết quả mô hình 30](#_Toc185796609)

[3.2.5 So sánh kết quả giữa các mô hình 30](#_Toc185796610)

[3.2.6. Kiểm thử với dữ liệu mới 31](#_Toc185796611)

[3.2.7 Thiết kế giao diện đồ họa người dùng 31](#_Toc185796612)

[IV. Chương 4: Kết Quả và Thảo Luận 32](#_Toc185796613)

[4.1 Kết quả phân tích dữ liệu 32](#_Toc185796614)

[4.1.1 Thống kê mô tả 32](#_Toc185796615)

[4.1.2 Phân tích mối quan hệ giữa các biến số 37](#_Toc185796616)

[4.2 Hiệu quả của các mô hình học máy 39](#_Toc185796617)

[4.2.1 Kết quả huấn luyện và kiểm thử 39](#_Toc185796618)

[4.2.2. Hiệu suất trên tập huấn luyện: 39](#_Toc185796619)

[4.2.3. Hiệu suất trên tập kiểm thử: 39](#_Toc185796620)

[4.3 Giao diện của mô hình 40](#_Toc185796621)

[4.3.1. Mô tả tổng quan giao diện: 40](#_Toc185796622)

[4.3.2. Cách hoạt động: 41](#_Toc185796623)

[4.3.3. Ưu điểm: 41](#_Toc185796624)

[4.4.4. Nhược điểm: 41](#_Toc185796625)

[4.4.5. Mô hình đề xuất 41](#_Toc185796626)

[4.4 Thảo luận về các yếu tố ảnh hưởng đến độ chính xác 42](#_Toc185796627)

[4.4.1. Độ chính xác của dữ liệu đầu vào: 43](#_Toc185796628)

[4.4.2. Tiền xử lý dữ liệu: 43](#_Toc185796629)

[4.4.3. Lựa chọn mô hình: 43](#_Toc185796630)

[4.4.4. Kỹ thuật huấn luyện: 43](#_Toc185796631)

[4.4.5. Đa dạng hóa dữ liệu: 44](#_Toc185796632)

[4.4.6. Khả năng giải thích của mô hình: 44](#_Toc185796633)

[4.4.7. Hạn chế của công nghệ và tính toán: 44](#_Toc185796634)

[4.4.8. Cải thiện chất lượng dữ liệu: 44](#_Toc185796635)

[V. Hạn chế của nghiên cứu và hướng phát triển 46](#_Toc185796636)

[5.1 Hạn chế của nghiên cứu 46](#_Toc185796637)

[5.1.1. Chất lượng và phạm vi của dữ liệu: 46](#_Toc185796638)

[5.1.2. Giới hạn của mô hình học máy: 46](#_Toc185796639)

[5.1.3. Thiếu sót trong đánh giá mô hình: 46](#_Toc185796640)

[5.1.4. Tính ứng dụng thực tế: 46](#_Toc185796641)

[5.2 Hướng phát triển 47](#_Toc185796642)

[5.2 1. Mở rộng và cải thiện dữ liệu: 47](#_Toc185796643)

[5.2.3. Tăng cường đánh giá mô hình: 47](#_Toc185796644)

[5.2.4. Nâng cao tính ứng dụng thực tế: 47](#_Toc185796645)

[5.2.5. Cải tiến khả năng giải thích mô hình: 48](#_Toc185796646)

[5.3 Kết luận chung 48](#_Toc185796647)

[VI. Danh mục tài liệu tham khảo 49](#_Toc185796648)

# I. Chương 1: Giới Thiệu

## 1.1 Tầm quan trọng của dự báo bệnh tim



Bệnh tim mạch là một trong những nguyên nhân hàng đầu gây tử vong trên toàn thế giới. Theo Tổ chức Y tế Thế giới (WHO), hàng năm có hơn 17 triệu người tử vong do các bệnh liên quan đến tim mạch, chiếm khoảng 31% tổng số ca tử vong toàn cầu. Trong số này, phần lớn là do nhồi máu cơ tim và đột quỵ, hai tình trạng cấp tính thường xuất hiện bất ngờ và gây hậu quả nghiêm trọng.

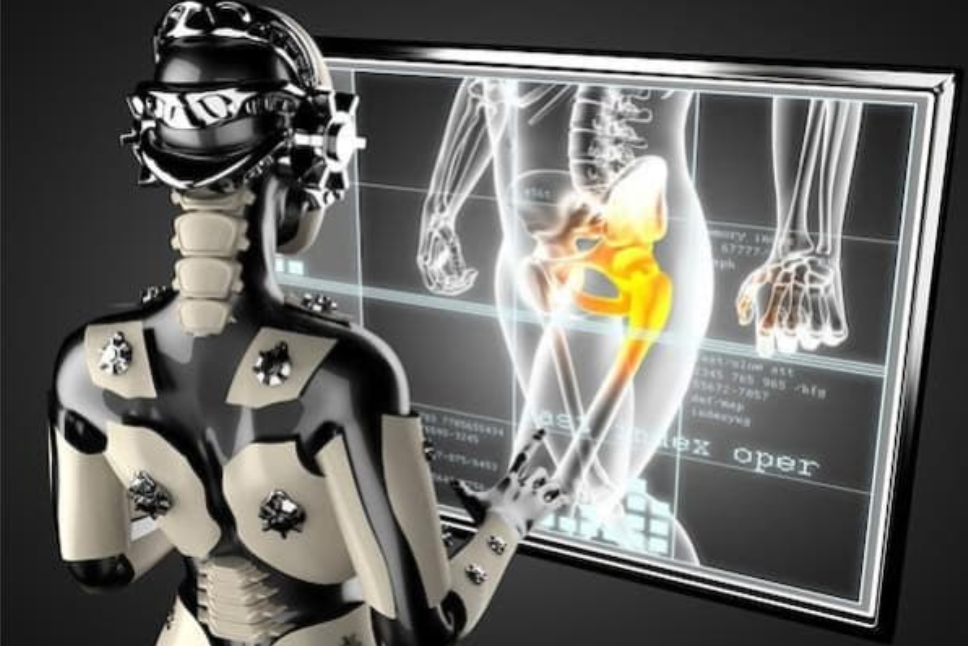
Dự báo sớm về nguy cơ mắc bệnh tim có vai trò vô cùng quan trọng trong việc ngăn chặn và giảm thiểu các biến chứng nguy hiểm. Việc phát hiện kịp thời các yếu tố nguy cơ giúp các bác sĩ có thể đưa ra các biện pháp phòng ngừa hoặc can thiệp sớm, từ đó giảm thiểu tỷ lệ mắc bệnh cũng như tăng cường chất lượng sống của bệnh nhân. Điều này đặc biệt quan trọng đối với các nhóm dân số có nguy cơ cao như người cao tuổi, người thừa cân, người có tiền sử bệnh lý tim mạch, hoặc những người có lối sống không lành mạnh.

Trong bối cảnh y học hiện đại, việc sử dụng các công nghệ tiên tiến như học máy để dự báo bệnh tim đang ngày càng trở nên phổ biến và hiệu quả. Các mô hình học máy có khả năng phân tích lượng dữ liệu lớn, nhận diện các mẫu liên quan đến bệnh lý mà các phương pháp truyền thống có thể bỏ qua. Điều này cho phép các chuyên gia y tế không chỉ dự báo nguy cơ mắc bệnh với độ chính xác cao mà còn cá nhân hóa các biện pháp điều trị và phòng ngừa cho từng bệnh nhân cụ thể.

Bên cạnh đó, dự báo bệnh tim còn giúp giảm bớt gánh nặng cho hệ thống y tế. Bằng cách tập trung nguồn lực vào những trường hợp có nguy cơ cao, các cơ sở y tế có thể tối ưu hóa việc sử dụng nhân lực và thiết bị, từ đó nâng cao hiệu quả chăm sóc và quản lý bệnh tật. Trong tương lai, với sự tiến bộ không ngừng của công nghệ, đặc biệt là trí tuệ nhân tạo và học máy, việc dự báo bệnh tim sẽ ngày càng chính xác và phổ biến, góp phần quan trọng vào việc bảo vệ sức khỏe cộng đồng.

Tóm lại, dự báo bệnh tim không chỉ là một công cụ y học mà còn là một chiến lược quan trọng trong việc bảo vệ và nâng cao sức khỏe cộng đồng. Sự phát triển của các phương pháp dự báo hiện đại, đặc biệt là dựa trên học máy, hứa hẹn mang lại những thay đổi tích cực trong việc phòng chống và điều trị bệnh tim mạch.

## 1.2 Vai trò của học máy trong y học



Trong thập kỷ qua, học máy (Machine Learning) đã nổi lên như một công cụ quan trọng trong lĩnh vực y học, mang lại những bước tiến đột phá trong chẩn đoán, điều trị và quản lý sức khỏe. Học máy, một nhánh của trí tuệ nhân tạo (AI), sử dụng các thuật toán và mô hình thống kê để phân tích dữ liệu y tế, tìm ra các mẫu và xu hướng mà con người có thể không dễ dàng nhận thấy. Điều này đã mở ra nhiều cơ hội mới trong việc cải thiện chất lượng chăm sóc sức khỏe và tăng cường hiệu quả của các phương pháp điều trị.

1.2.1 Tự động hóa và tối ưu hóa quá trình chẩn đoán

Một trong những vai trò nổi bật của học máy trong y học là khả năng tự động hóa và tối ưu hóa quá trình chẩn đoán. Các thuật toán học máy có thể phân tích hình ảnh y tế như X-quang, MRI, và CT scan với độ chính xác cao, thậm chí vượt qua khả năng của con người trong một số trường hợp. Ví dụ, các mô hình học máy đã được sử dụng để phát hiện ung thư ở giai đoạn sớm, nhận diện tổn thương mô, và dự đoán nguy cơ tái phát bệnh. Điều này không chỉ giúp rút ngắn thời gian chẩn đoán mà còn giảm thiểu sai sót, từ đó cải thiện kết quả điều trị cho bệnh nhân.

1.2.2 Cá nhân hóa điều trị

Học máy cũng đóng vai trò quan trọng trong việc cá nhân hóa điều trị. Bằng cách phân tích dữ liệu từ hồ sơ bệnh án, kết quả xét nghiệm, và dữ liệu di truyền, các mô hình học máy có thể dự đoán phản ứng của bệnh nhân với các liệu pháp điều trị khác nhau. Điều này cho phép các bác sĩ lựa chọn phương pháp điều trị tối ưu cho từng cá nhân, từ đó tăng cường hiệu quả điều trị và giảm thiểu tác dụng phụ không mong muốn. Sự cá nhân hóa trong y học không chỉ cải thiện chất lượng cuộc sống của bệnh nhân mà còn giúp giảm chi phí y tế.

1.2.3 Dự báo và quản lý bệnh tật

Ngoài chẩn đoán và điều trị, học máy còn được sử dụng rộng rãi trong việc dự báo và quản lý bệnh tật. Các mô hình dự báo dựa trên học máy có khả năng phân tích dữ liệu lịch sử của bệnh nhân để dự đoán nguy cơ phát triển các bệnh mãn tính như tiểu đường, cao huyết áp, và bệnh tim mạch. Việc này giúp các chuyên gia y tế có thể can thiệp sớm, áp dụng các biện pháp phòng ngừa thích hợp, từ đó giảm thiểu nguy cơ mắc bệnh và cải thiện sức khỏe cộng đồng.

1.2.4 Nâng cao nghiên cứu y học

Học máy cũng đang tạo ra những ảnh hưởng lớn đối với nghiên cứu y học. Các thuật toán học máy có thể xử lý và phân tích khối lượng dữ liệu khổng lồ từ các nghiên cứu lâm sàng, thử nghiệm thuốc, và dữ liệu gen để khám phá ra những phát hiện mới mà phương pháp truyền thống không thể đạt được. Điều này không chỉ giúp đẩy nhanh quá trình nghiên cứu mà còn mở ra những hướng đi mới trong việc phát triển các phương pháp điều trị và thuốc mới.

1.2.5 Hỗ trợ quyết định lâm sàng

Cuối cùng, học máy đóng vai trò quan trọng trong việc hỗ trợ quyết định lâm sàng. Bằng cách cung cấp các dự báo chính xác và phân tích chi tiết, học máy giúp các bác sĩ đưa ra quyết định nhanh chóng và chính xác hơn trong các tình huống phức tạp. Các hệ thống hỗ trợ quyết định lâm sàng dựa trên học máy có thể cung cấp các khuyến nghị về chẩn đoán, điều trị, và quản lý bệnh nhân, giúp cải thiện hiệu quả của quá trình chăm sóc sức khỏe.

## 1.3 Mục tiêu của đồ án

Đồ án này tập trung vào việc áp dụng các kỹ thuật học máy (Machine Learning) trong Python để dự báo bệnh tim dựa trên các yếu tố nguy cơ của bệnh nhân. Mục tiêu chính của bài tập là xây dựng và đánh giá các mô hình học máy nhằm xác định khả năng mắc bệnh tim của một cá nhân dựa trên các dữ liệu y tế có sẵn. Cụ thể, các mục tiêu chi tiết bao gồm:

### 1.3.1 Hiểu và phân tích dữ liệu y tế

Mục tiêu đầu tiên là hiểu và phân tích dữ liệu y tế liên quan đến bệnh tim. Điều này bao gồm việc thu thập dữ liệu, làm sạch dữ liệu, và tiền xử lý để đảm bảo rằng nó phù hợp cho các mô hình học máy. Việc hiểu rõ về dữ liệu và các yếu tố nguy cơ như tuổi tác, giới tính, mức cholesterol, huyết áp, và các yếu tố lối sống khác là bước đầu tiên quan trọng để xây dựng mô hình dự báo chính xác.

### 1.3.2 Áp dụng các thuật toán học máy để xây dựng mô hình dự báo

Sau khi dữ liệu đã được chuẩn bị, mục tiêu tiếp theo là áp dụng các thuật toán học máy phổ biến như Cây Quyết Định kết hợp với PCA để xây dựng mô hình dự báo bệnh tim. Mục tiêu của phần này là hiểu cách thức hoạt động của từng thuật toán, cách chúng xử lý dữ liệu và đưa ra dự báo, cũng như các ưu nhược điểm của từng phương pháp trong ngữ cảnh cụ thể này.

### 1.3.3 Đánh giá và so sánh hiệu suất các mô hình

Một mục tiêu quan trọng khác của bài tập lớn là đánh giá hiệu suất của các mô hình học máy đã xây dựng. Điều này được thực hiện bằng cách sử dụng các tiêu chuẩn đánh giá như độ chính xác, độ nhạy, độ đặc hiệu, và chỉ số F1. So sánh hiệu suất của các mô hình khác nhau sẽ giúp xác định mô hình nào hoạt động tốt nhất trong việc dự báo bệnh tim.

### 1.3.4 Tối ưu hóa và cải thiện mô hình

Dựa trên các kết quả đánh giá, mục tiêu tiếp theo là tối ưu hóa các mô hình để cải thiện độ chính xác và hiệu suất dự báo. Việc tối ưu hóa có thể bao gồm điều chỉnh các siêu tham số của mô hình, sử dụng các kỹ thuật lựa chọn đặc trưng, hoặc áp dụng các phương pháp nâng cao như Tăng cường Gradient hoặc Học sâu (Deep Learning).

### 1.3.5 Rút ra kết luận và đề xuất hướng phát triển trong tương lai

Cuối cùng, bài tập lớn hướng đến việc rút ra các kết luận quan trọng từ quá trình nghiên cứu và đề xuất các hướng phát triển trong tương lai. Các kết quả của bài tập không chỉ giúp hiểu rõ hơn về khả năng dự báo bệnh tim bằng học máy mà còn cung cấp những gợi ý cho các nghiên cứu tiếp theo hoặc áp dụng trong thực tiễn y tế.

# II. Chương 2: Tổng Quan Lý Thuyết

## 2.1 Giới thiệu về bệnh tim và các yếu tố nguy cơ



### 2.1.1 Giới thiệu về bệnh tim

Bệnh tim mạch là một nhóm các bệnh lý ảnh hưởng đến tim và mạch máu, bao gồm các tình trạng như bệnh động mạch vành, nhồi máu cơ tim, suy tim, và đột quỵ. Đây là nguyên nhân gây tử vong hàng đầu trên toàn thế giới, với hàng triệu người bị ảnh hưởng mỗi năm. Trong số các bệnh lý tim mạch, bệnh động mạch vành là phổ biến nhất, xảy ra khi các mạch máu cung cấp máu cho tim bị thu hẹp hoặc tắc nghẽn do sự tích tụ của mảng bám (plaque), gây ra thiếu máu cơ tim và có thể dẫn đến nhồi máu cơ tim.

Bệnh tim có thể phát triển dần dần qua nhiều năm mà không có triệu chứng rõ ràng, cho đến khi xuất hiện các biến chứng nghiêm trọng. Các triệu chứng của bệnh tim thường bao gồm đau ngực (đau thắt ngực), khó thở, mệt mỏi, nhịp tim không đều, và trong những trường hợp nghiêm trọng, đột quỵ hoặc nhồi máu cơ tim.

### 2.1.2 Các yếu tố nguy cơ của bệnh tim

Các yếu tố nguy cơ của bệnh tim là những điều kiện hoặc thói quen làm tăng khả năng phát triển bệnh tim mạch. Các yếu tố nguy cơ này có thể được chia thành hai nhóm chính: các yếu tố nguy cơ có thể thay đổi và các yếu tố nguy cơ không thể thay đổi.

a) Các yếu tố nguy cơ không thể thay đổi:

- Tuổi tác: Nguy cơ mắc bệnh tim tăng lên theo tuổi tác. Nam giới từ 45 tuổi trở lên và phụ nữ từ 55 tuổi trở lên có nguy cơ cao hơn.

- Giới tính: Nam giới thường có nguy cơ mắc bệnh tim mạch cao hơn phụ nữ ở tuổi trưởng thành. Tuy nhiên, sau khi mãn kinh, nguy cơ của phụ nữ cũng tăng lên và gần tương đương với nam giới.

- Tiền sử gia đình: Những người có tiền sử gia đình bị bệnh tim, đặc biệt là nếu người thân mắc bệnh sớm (trước 55 tuổi đối với nam và trước 65 tuổi đối với nữ), có nguy cơ cao hơn.

b) Các yếu tố nguy cơ có thể thay đổi:

- Hút thuốc: Hút thuốc lá là một trong những yếu tố nguy cơ chính của bệnh tim. Nicotine trong thuốc lá làm tăng nhịp tim và huyết áp, đồng thời các chất hóa học khác trong khói thuốc có thể làm tổn thương lớp lót bên trong động mạch.

- Cao huyết áp: Huyết áp cao làm tăng gánh nặng cho tim, gây ra sự cứng lại của động mạch và tăng nguy cơ mắc bệnh động mạch vành và nhồi máu cơ tim.

- Mức cholesterol cao: Mức cholesterol cao, đặc biệt là LDL (Low-Density Lipoprotein) cholesterol, có thể dẫn đến sự tích tụ mảng bám trong động mạch, gây hẹp mạch và dẫn đến bệnh tim.

- Tiểu đường: Tiểu đường làm tăng nguy cơ mắc bệnh tim do ảnh hưởng tiêu cực đến mạch máu và mức độ đường huyết cao có thể gây tổn thương tim và động mạch.

- Béo phì: Béo phì làm tăng nguy cơ mắc bệnh tim thông qua các yếu tố như cao huyết áp, mức cholesterol cao, và tiểu đường. Mỡ thừa, đặc biệt là mỡ bụng, có liên quan mật thiết đến các bệnh lý tim mạch.

- Lối sống ít vận động: Thiếu hoạt động thể chất làm tăng nguy cơ mắc bệnh tim, vì nó có thể dẫn đến béo phì, cao huyết áp, và mức cholesterol không lành mạnh.

- Chế độ ăn uống không lành mạnh: Chế độ ăn giàu chất béo bão hòa, muối, và đường có thể dẫn đến béo phì, cao huyết áp, và mức cholesterol cao, từ đó tăng nguy cơ mắc bệnh tim.

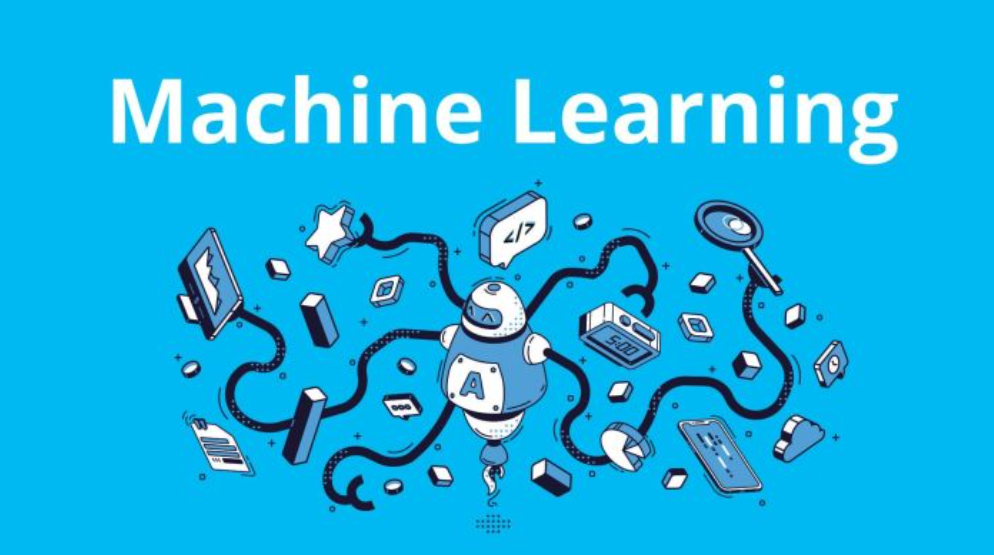
- Căng thẳng: Căng thẳng mãn tính có thể góp phần vào sự phát triển của bệnh tim bằng cách dẫn đến các thói quen không lành mạnh như hút thuốc, ăn uống không kiểm soát, và lối sống ít vận động.

### 2.1.3 Tầm quan trọng của việc kiểm soát các yếu tố nguy cơ

Việc kiểm soát các yếu tố nguy cơ có thể thay đổi là một phần quan trọng trong việc ngăn ngừa và quản lý bệnh tim. Thay đổi lối sống, như bỏ thuốc lá, ăn uống lành mạnh, tập thể dục thường xuyên, và kiểm soát căng thẳng, có thể giảm đáng kể nguy cơ mắc bệnh tim mạch. Ngoài ra, việc theo dõi thường xuyên các chỉ số sức khỏe như huyết áp, mức cholesterol, và đường huyết, cùng với việc sử dụng thuốc khi cần thiết, có thể giúp kiểm soát các yếu tố nguy cơ và ngăn ngừa sự tiến triển của bệnh tim.

Tóm lại, hiểu rõ các yếu tố nguy cơ của bệnh tim là điều cần thiết để có thể đưa ra các biện pháp phòng ngừa và can thiệp hiệu quả. Việc áp dụng các chiến lược dự phòng, cùng với các phương pháp điều trị tiên tiến như học máy trong y học, có thể giúp giảm thiểu tỷ lệ mắc bệnh tim và cải thiện chất lượng cuộc sống của người bệnh.

## 2.2 Tổng quan về các kỹ thuật học máy



Học máy (Machine Learning) là một nhánh của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng các hệ thống có khả năng học hỏi và cải thiện hiệu suất từ dữ liệu mà không cần phải lập trình một cách rõ ràng. Học máy đã trở thành một công cụ mạnh mẽ trong nhiều lĩnh vực, từ tài chính, thương mại điện tử, đến y học và khoa học. Các kỹ thuật học máy có thể được phân thành nhiều loại khác nhau, tùy thuộc vào cách thức học hỏi và ứng dụng của chúng. Trong phần này, chúng ta sẽ tìm hiểu các kỹ thuật học máy phổ biến và cách chúng được áp dụng trong các bài toán thực tiễn.

### 2.2.1 Học có giám sát (Supervised Learning)

Học có giám sát là một trong những kỹ thuật học máy phổ biến nhất. Trong học có giám sát, mô hình học từ một tập dữ liệu đã được gán nhãn, nghĩa là mỗi điểm dữ liệu đi kèm với một nhãn hoặc kết quả đầu ra mong muốn. Mục tiêu của mô hình là học mối quan hệ giữa các đặc trưng (features) đầu vào và nhãn đầu ra để có thể dự đoán chính xác nhãn của các điểm dữ liệu mới.

Các thuật toán học có giám sát phổ biến:

- Hồi quy tuyến tính (Linear Regression): Được sử dụng để dự đoán một giá trị liên tục dựa trên các đặc trưng đầu vào. Hồi quy tuyến tính cố gắng tìm ra một đường thẳng tốt nhất mô tả mối quan hệ giữa đầu vào và đầu ra.

- Hồi quy Logistic (Logistic Regression): Được sử dụng cho các bài toán phân loại nhị phân, như việc phân loại một bệnh nhân có mắc bệnh tim hay không. Hồi quy Logistic ước lượng xác suất của một biến phân loại nhị phân dựa trên một hoặc nhiều đặc trưng đầu vào.

- Cây Quyết Định (Decision Tree): Một thuật toán phân loại sử dụng một cấu trúc giống cây để đưa ra các quyết định dựa trên các đặc trưng của dữ liệu. Cây Quyết Định là dễ hiểu và có thể trực quan hóa, giúp giải thích các quyết định của mô hình.

- Rừng Ngẫu Nhiên (Random Forest): Một kỹ thuật học máy mạnh mẽ dựa trên việc kết hợp nhiều Cây Quyết Định. Rừng Ngẫu Nhiên sử dụng kỹ thuật Bagging để cải thiện độ chính xác và giảm thiểu nguy cơ quá khớp (overfitting).

- Máy Vector Hỗ Trợ (Support Vector Machine - SVM): Một thuật toán phân loại mạnh mẽ được sử dụng để tìm ra siêu phẳng (hyperplane) tốt nhất tách biệt các lớp trong không gian đặc trưng. SVM đặc biệt hữu ích cho các bài toán phân loại phức tạp và phi tuyến tính.

### 2.2.2 Học không giám sát (Unsupervised Learning)

Học không giám sát được sử dụng khi dữ liệu không có nhãn, và mục tiêu là khám phá ra các cấu trúc ẩn hoặc mô hình trong dữ liệu. Thay vì dự đoán kết quả cụ thể, học không giám sát tập trung vào việc nhóm hoặc phân cụm dữ liệu dựa trên sự tương đồng giữa các điểm dữ liệu.

Các thuật toán học không giám sát phổ biến:

- Phân cụm K-Means (K-Means Clustering): Là một thuật toán phân cụm phổ biến, K-Means chia dữ liệu thành k nhóm hoặc cụm sao cho các điểm dữ liệu trong cùng một cụm giống nhau hơn so với các điểm dữ liệu ở các cụm khác.

- Phân cụm Phân cấp (Hierarchical Clustering): Một kỹ thuật phân cụm khác sử dụng cấu trúc phân cấp để tạo ra một cây phân cụm, trong đó mỗi nút đại diện cho một cụm và các nhánh đại diện cho sự phân chia của cụm.

- Giảm chiều dữ liệu (Dimensionality Reduction): Các kỹ thuật như Phân tích thành phần chính (Principal Component Analysis - PCA) và Phân tích tương quan độc lập (Independent Component Analysis - ICA) được sử dụng để giảm số lượng đặc trưng trong dữ liệu, giúp đơn giản hóa mô hình và giảm độ phức tạp tính toán.

### 2.2.3 Học bán giám sát (Semi-Supervised Learning)

Học bán giám sát là sự kết hợp giữa học có giám sát và học không giám sát. Trong học bán giám sát, mô hình học từ một lượng nhỏ dữ liệu có nhãn và một lượng lớn dữ liệu không có nhãn. Kỹ thuật này hữu ích khi việc gán nhãn cho tất cả dữ liệu là tốn kém hoặc không khả thi, nhưng vẫn muốn tận dụng thông tin từ dữ liệu không có nhãn để cải thiện mô hình.

### 2.2.4 Học tăng cường (Reinforcement Learning)

Học tăng cường là một kỹ thuật trong đó một tác nhân (agent) học cách đưa ra các quyết định thông qua việc tương tác với môi trường. Tác nhân nhận được các phần thưởng (rewards) hoặc hình phạt (penalties) dựa trên hành động của mình, và mục tiêu là tối đa hóa tổng phần thưởng nhận được theo thời gian.

Ứng dụng của học tăng cường:

- Điều khiển tự động: Học tăng cường được sử dụng trong việc điều khiển các hệ thống phức tạp như robot, xe tự lái, và các trò chơi máy tính.

- Quản lý năng lượng: Các hệ thống quản lý năng lượng có thể sử dụng học tăng cường để tối ưu hóa việc sử dụng năng lượng trong các tòa nhà hoặc nhà máy.

### 2.2.5 Học sâu (Deep Learning)

Học sâu là một nhánh của học máy dựa trên mạng nơ-ron nhân tạo (Artificial Neural Networks). Học sâu sử dụng các mạng nơ-ron với nhiều lớp ẩn (hidden layers) để học các biểu diễn phức tạp của dữ liệu. Học sâu đã đạt được những thành tựu đáng kể trong các lĩnh vực như nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên, và chơi game.

Các loại mạng nơ-ron trong học sâu:

- Mạng Nơ-ron tích chập (Convolutional Neural Network - CNN): Được sử dụng chủ yếu trong các bài toán liên quan đến hình ảnh và video, CNN có khả năng nhận diện các mẫu phức tạp như khuôn mặt, vật thể, và chữ viết.

- Mạng Nơ-ron hồi quy (Recurrent Neural Network - RNN): Thường được sử dụng trong xử lý ngôn ngữ tự nhiên và chuỗi thời gian, RNN có khả năng ghi nhớ thông tin từ các bước trước đó trong chuỗi, giúp nó hiểu được ngữ cảnh.

- Mạng Nơ-ron biến đổi (Transformer): Được sử dụng chủ yếu trong xử lý ngôn ngữ tự nhiên, mạng Transformer đã cách mạng hóa các mô hình ngôn ngữ như GPT và BERT, nhờ khả năng xử lý đồng thời tất cả các vị trí trong chuỗi đầu vào.

### 2.2.6 Kỹ thuật học máy tổng hợp (Ensemble Learning)

Kỹ thuật học máy tổng hợp kết hợp nhiều mô hình học máy lại với nhau để cải thiện hiệu suất so với việc chỉ sử dụng một mô hình đơn lẻ. Các kỹ thuật học máy tổng hợp giúp giảm thiểu sai số và tăng độ ổn định của dự báo.

Các kỹ thuật học máy tổng hợp phổ biến:

- Bagging: Kỹ thuật này tạo ra nhiều phiên bản khác nhau của một mô hình học máy bằng cách lấy mẫu ngẫu nhiên dữ liệu với việc hoàn lại, sau đó kết hợp các kết quả dự báo của các mô hình này.

- Boosting: Boosting liên quan đến việc xây dựng các mô hình học máy theo tuần tự, trong đó mỗi mô hình mới học từ các lỗi của mô hình trước đó. Một ví dụ phổ biến là thuật toán AdaBoost.

- Stacking: Stacking kết hợp các mô hình học máy khác nhau bằng cách sử dụng một mô hình học máy khác (meta-model) để dự đoán dựa trên các kết quả đầu ra của các mô hình khác.

## 2.3 Các mô hình học máy thường được sử dụng trong y học

Học máy đã trở thành một công cụ quan trọng trong y học, giúp các nhà khoa học và bác sĩ dự đoán, chẩn đoán, và điều trị bệnh một cách chính xác và hiệu quả hơn. Nhiều mô hình học máy đã được áp dụng trong y học, từ các bài toán đơn giản như phân loại bệnh nhân, đến các ứng dụng phức tạp như dự đoán diễn tiến của bệnh hay hỗ trợ quyết định lâm sàng. Dưới đây là một số mô hình học máy phổ biến thường được sử dụng trong y học:

2.3.1 Hồi quy Logistic (Logistic Regression)

Hồi quy Logistic là một mô hình học có giám sát thường được sử dụng cho các bài toán phân loại nhị phân, chẳng hạn như phân loại một bệnh nhân có mắc bệnh tim hay không. Mô hình này dựa trên hàm logistic để ước lượng xác suất xảy ra của một sự kiện, giúp dự đoán khả năng mắc bệnh dựa trên các yếu tố nguy cơ.

- Ứng dụng trong y học:

+ Phân loại bệnh: Logistic Regression thường được sử dụng để phân loại bệnh nhân thành các nhóm có nguy cơ cao hoặc thấp đối với các bệnh như bệnh tim, tiểu đường, và ung thư.

+Dự đoán kết quả điều trị: Mô hình này cũng được sử dụng để dự đoán khả năng thành công của các phương pháp điều trị hoặc khả năng tái phát bệnh.

2.3.2 Cây Quyết Định (Decision Tree)

Cây Quyết Định là một mô hình dễ hiểu và trực quan, sử dụng cấu trúc giống cây để đưa ra các quyết định dựa trên các đặc trưng của dữ liệu. Mỗi nút trong cây đại diện cho một đặc trưng và mỗi nhánh đại diện cho kết quả của một quyết định dựa trên đặc trưng đó.

- Ứng dụng trong y học:

+ Chẩn đoán bệnh: Cây Quyết Định thường được sử dụng trong việc chẩn đoán bệnh, dựa trên các triệu chứng và dấu hiệu lâm sàng của bệnh nhân.

+ Hỗ trợ quyết định lâm sàng: Mô hình này có thể hỗ trợ bác sĩ trong việc đưa ra các quyết định lâm sàng bằng cách cung cấp các dự báo dựa trên dữ liệu bệnh nhân.

2.3.3 Rừng Ngẫu Nhiên (Random Forest)

Rừng Ngẫu Nhiên là một kỹ thuật học máy tổng hợp kết hợp nhiều Cây Quyết Định để tạo ra một mô hình mạnh mẽ và chính xác hơn. Bằng cách sử dụng kỹ thuật Bagging, Random Forest giúp giảm thiểu nguy cơ quá khớp và cải thiện độ chính xác của dự báo.

- Ứng dụng trong y học:

+ Phân loại bệnh: Random Forest thường được sử dụng để phân loại bệnh nhân trong các nghiên cứu về dịch tễ học, giúp xác định các yếu tố nguy cơ liên quan đến các bệnh lý khác nhau.

+ Phân tích yếu tố nguy cơ: Mô hình này cũng được sử dụng để phân tích và xác định các yếu tố nguy cơ quan trọng đối với sự phát triển của các bệnh mãn tính như bệnh tim và ung thư.

2.3.4 Máy Vector Hỗ Trợ (Support Vector Machine - SVM)

SVM là một mô hình phân loại mạnh mẽ được sử dụng để tìm ra siêu phẳng tốt nhất tách biệt các lớp trong không gian đặc trưng. Mô hình này đặc biệt hữu ích cho các bài toán phân loại phức tạp và phi tuyến tính.

- Ứng dụng trong y học:

+ Phân loại hình ảnh y tế: SVM thường được sử dụng trong các ứng dụng liên quan đến xử lý hình ảnh y tế, chẳng hạn như phân loại các khối u trong hình ảnh chụp X-quang, CT, hoặc MRI.

+ Dự đoán kết quả lâm sàng: SVM có thể được sử dụng để dự đoán các kết quả lâm sàng dựa trên dữ liệu bệnh nhân, chẳng hạn như khả năng sống sót sau một phẫu thuật hoặc điều trị.

2.3.5 Mạng Nơ-ron Nhân tạo (Artificial Neural Networks - ANN)

Mạng Nơ-ron Nhân tạo là một mô hình học sâu dựa trên cấu trúc của não bộ con người, với các nơ-ron kết nối với nhau trong nhiều lớp. ANN có khả năng học các biểu diễn phức tạp của dữ liệu và đã được sử dụng thành công trong nhiều lĩnh vực y học.

- Ứng dụng trong y học:

+ Chẩn đoán bệnh tự động: ANN có thể học cách phân loại và chẩn đoán bệnh tự động từ dữ liệu bệnh nhân, chẳng hạn như phân loại bệnh từ hình ảnh y tế hoặc dữ liệu sinh học.

+ Dự đoán diễn tiến bệnh: Mô hình này cũng được sử dụng để dự đoán diễn tiến của bệnh, chẳng hạn như tiến triển của ung thư hoặc mức độ nghiêm trọng của bệnh tim.

2.3.6 Mạng Nơ-ron Tích chập (Convolutional Neural Networks - CNN)

CNN là một loại mạng nơ-ron sâu đặc biệt mạnh mẽ trong việc xử lý dữ liệu hình ảnh và video. CNN có khả năng nhận diện các mẫu phức tạp trong hình ảnh, chẳng hạn như các đặc điểm của khối u hoặc tổn thương trong hình ảnh y tế.

- Ứng dụng trong y học:

+ Chẩn đoán hình ảnh y tế: CNN đã được sử dụng rộng rãi trong việc phân tích hình ảnh y tế, giúp phát hiện sớm các bệnh lý như ung thư, bệnh về mắt, và tổn thương nội tạng từ hình ảnh chụp X-quang, MRI, hoặc CT.

+ Phân loại mô bệnh học: CNN cũng được sử dụng trong việc phân loại các loại tế bào và mô bệnh học, giúp các nhà nghiên cứu và bác sĩ hiểu rõ hơn về cấu trúc và đặc điểm của các loại mô khác nhau.

2.3.7 Mạng Nơ-ron Hồi quy (Recurrent Neural Networks - RNN)

RNN là một loại mạng nơ-ron sâu được thiết kế để xử lý dữ liệu chuỗi thời gian, có khả năng ghi nhớ thông tin từ các bước trước đó trong chuỗi. Điều này làm cho RNN đặc biệt phù hợp cho các bài toán liên quan đến dữ liệu tuần tự, chẳng hạn như dữ liệu sinh học và chuỗi thời gian trong y học.

- Ứng dụng trong y học:

+ Phân tích chuỗi thời gian sinh học: RNN được sử dụng để phân tích các chuỗi thời gian sinh học như dữ liệu ECG (điện tâm đồ), giúp phát hiện các bất thường trong nhịp tim và chẩn đoán bệnh tim mạch.

+ Dự đoán diễn biến lâm sàng: RNN có thể dự đoán diễn biến lâm sàng của bệnh nhân dựa trên dữ liệu lâm sàng được thu thập theo thời gian, giúp bác sĩ dự báo các biến cố quan trọng như suy tim hoặc tái phát bệnh.

2.3.8 Mạng Biến đổi (Transformers)

Transformers là một kiến trúc mạng nơ-ron mạnh mẽ được phát triển ban đầu cho xử lý ngôn ngữ tự nhiên (NLP) nhưng đã được áp dụng rộng rãi trong các lĩnh vực khác, bao gồm y học. Transformers có khả năng xử lý đồng thời tất cả các vị trí trong chuỗi đầu vào, giúp cải thiện độ chính xác và hiệu suất của các mô hình học máy.

- Ứng dụng trong y học:

+ Phân tích và xử lý văn bản y khoa: Transformers được sử dụng để phân tích và xử lý các văn bản y khoa, chẳng hạn như báo cáo lâm sàng, ghi chú của bác sĩ, và tài liệu nghiên cứu. Điều này giúp trích xuất thông tin quan trọng và hỗ trợ quyết định lâm sàng.

+ Dự đoán kết quả dựa trên chuỗi dữ liệu: Transformers có thể dự đoán các kết quả y tế dựa trên chuỗi dữ liệu phức tạp, chẳng hạn như diễn tiến của bệnh dựa trên hồ sơ y tế điện tử của bệnh nhân.

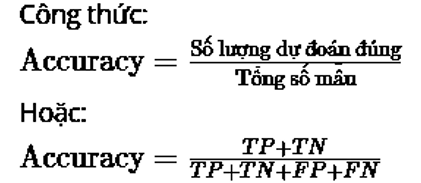
## 2.4 Các tiêu chuẩn đánh giá mô hình

Trong học máy, việc đánh giá hiệu suất của một mô hình là một bước quan trọng để đảm bảo rằng mô hình hoạt động tốt và đáp ứng được mục tiêu đặt ra. Đặc biệt trong y học, độ chính xác của mô hình là yếu tố then chốt, vì nó ảnh hưởng trực tiếp đến việc chẩn đoán và điều trị bệnh nhân. Dưới đây là các tiêu chuẩn phổ biến được sử dụng để đánh giá mô hình học máy:

### **2.4.1 Độ chính xác (Accuracy)**

**Độ chính xác** là tỷ lệ phần trăm của dự đoán đúng trên tổng số dự đoán. Đây là tiêu chuẩn cơ bản nhất để đánh giá mô hình, đặc biệt hữu ích khi các lớp trong dữ liệu có phân phối cân bằng.

**Công thức:**



Trong đó:

* **TP (True Positive)**: Số lượng dự đoán đúng về sự hiện diện của bệnh (mô hình dự đoán có bệnh và thực tế cũng có bệnh).
* **TN (True Negative)**: Số lượng dự đoán đúng về sự vắng mặt của bệnh (mô hình dự đoán không có bệnh và thực tế cũng không có bệnh).
* **FP (False Positive)**: Số lượng dự đoán sai về sự hiện diện của bệnh (mô hình dự đoán có bệnh nhưng thực tế không có bệnh).
* **FN (False Negative)**: Số lượng dự đoán sai về sự vắng mặt của bệnh (mô hình dự đoán không có bệnh nhưng thực tế có bệnh).

### 2.4.2 Độ nhạy (Recall) hay Tỉ lệ dương tính thực (True Positive Rate - TPR)

**Độ nhạy**, còn gọi là Recall đo lường tỷ lệ dự đoán dương chính xác trên tổng số các mẫu thực tế là dương. Recall quan trọng trong các tình huống cần phát hiện tất cả các trường hợp dương (ví dụ: phát hiện bệnh), ngay cả khi điều đó dẫn đến việc dự đoán một số trường hợp âm là dương (false negatives).

**Công thức:**

**A black and white math equation

Description automatically generated**

**Ứng dụng:**

* Trong y học, độ nhạy đặc biệt quan trọng trong các bài toán mà việc bỏ sót các trường hợp bệnh có thể dẫn đến hậu quả nghiêm trọng, chẳng hạn như trong phát hiện ung thư sớm.

### 2.4.3 Độ chính xác của dự đoán dương (Precision)

**Precision** đo lường tỷ lệ dự đoán dương chính xác trên tổng số dự đoán dương. Điều này đặc biệt quan trọng trong các ứng dụng mà dự đoán dương giả (false positives) có thể gây hậu quả nghiêm trọng, chẳng hạn như trong chẩn đoán bệnh..

**Công thức:**

A black text with a white background

Description automatically generated

**Ứng dụng:**

* Độ chính xác của dự đoán dương quan trọng trong các trường hợp việc chẩn đoán bệnh hiếm gặp, sàng lọc bệnh nguy hiểm, cảnh báo y tế tự động ,… Tránh làm bệnh nhân lo lắng do kết quả dương tính giả hoặc dẫn đến điều trị không cần thiết cho bệnh nhân.

### 2.4.4 Điểm F1 (F1 Score)

**Điểm F1** là trung bình điều hòa của Precision và Recall. Nó cung cấp một cách đánh giá cân bằng giữa độ chính xác của các dự đoán dương và khả năng phát hiện đầy đủ các mẫu dương. F1-Score đặc biệt hữu ích khi dữ liệu không cân bằng và có sự đánh đổi giữa Precision và Recall.

**Công thức:**

A mathematical equation with black text

Description automatically generated​

**Ứng dụng:**

* Điểm F1 được ưa chuộng trong các tình huống mà sự cân bằng giữa Precision và Sensitivity là quan trọng, chẳng hạn như trong phân loại bệnh hiếm gặp.

# III. Chương 3: Phương Pháp Nghiên Cứu

## 3.1 Thu thập và xử lý dữ liệu

Trong nghiên cứu sử dụng các kỹ thuật học máy để dự báo bệnh tim, việc thu thập và xử lý dữ liệu là một bước quan trọng và quyết định trực tiếp đến chất lượng của mô hình. Quá trình này bao gồm việc tìm kiếm, thu thập dữ liệu liên quan, làm sạch, và chuẩn hóa dữ liệu để chuẩn bị cho các bước phân tích và mô hình hóa sau này.

### 3.1.1 Thu thập dữ liệu

**Nguồn dữ liệu:** Dữ liệu sử dụng trong nghiên cứu này chủ yếu được lấy từ cơ sở dữ liệu UCI Machine Learning Repository, đặc biệt là tập dữ liệu về bệnh tim. Tập dữ liệu này bao gồm thông tin về nhiều yếu tố khác nhau như tuổi tác, giới tính, huyết áp, mức cholesterol, tình trạng đau ngực, và các chỉ số sinh học khác của bệnh nhân.

**Đặc điểm của dữ liệu:**

* **Số lượng mẫu:** Tập dữ liệu chứa các bản ghi từ 303 bệnh nhân với đầy đủ thông tin cần thiết để tiến hành phân tích.
* **Số lượng biến:** Mỗi mẫu dữ liệu bao gồm nhiều biến khác nhau, từ các biến định tính như giới tính, loại đau ngực đến các biến định lượng như mức cholesterol, huyết áp.

|  |  |
| --- | --- |
| Chiều thông tin | Nhãn lớp |
| Name | Tên bệnh nhân |
| age | Tuổi của bệnh nhân |
| sex | Giới Tính: 1=Nam,0=Nữ |
| cp | Loại đau ngực: 0 - Không có triệu chứng 1 - Đau thắt ngực không điển hình  2 - Không đau thắt ngực 3 - Đau thắt ngực điển hình |
| trestbps | Huyết áp lúc nghỉ: Tính bằng mm / Hg khi nhập viện |
| chol | Cholestorol: Tính bằng mg / dl |
| fbs | Lượng đường trong máu:  0: <120 mg / dl  1: > 120 mg / dl |
| restecg | Điện tâm đồ khi nghỉ ngơi:  0 - Bình thường 1- Có sóng ST-T bất thường 2 - Phì đại thất trái |
| thalach | Nhịp tim tối đa đạt được(bpm-nhịp mỗi phút) |
| exang | Tập thể dục gây ra đau thắt ngực:  1 - Có  0 - Không |
| oldpeak | Chênh lệch đoạn ST khi tập thể dục so với lúc nghỉ |
| slope | Độ dốc đoạn ST:   0 - Đi xuống 1 - Đi lên 2 - Cân bằng |
| ca | Số lượng các mạch chính:  0 -> 3 được tô màu bằng phương pháp soi huỳnh quang. |
| thal | Thalassemia:tam máu bẩm sinh 0 - Không 1 - Khuyết tật cố định 2 - Lưu lượng máu bình thường 3 - Khuyết tật có thể đảo ngược |
| target | 1- bị bệnh 0- không bị bệnh |

* **Mức độ đa dạng:** Tập dữ liệu có kích thước vừa phải với 303 bệnh nhân với các tình trạng sức khỏe khác nhau, đủ để huấn luyện và đánh giá các mô hình dự đoán.

A white grid with many numbers

Description automatically generated with medium confidence

##### Hình 3.1.1.1 : Các thông tin của dữ liệu

### 3.1.2 Xử lý dữ liệu(Data processing)

3.1.2.1. Làm sạch dữ liệu (Data Cleaning): Quá trình làm sạch dữ liệu nhằm loại bỏ các giá trị sai lệch hoặc bất thường có thể ảnh hưởng đến kết quả phân tích. Các bước chính bao gồm:

* **Xử lý giá trị trùng lặp (Duplicate Data):** Những bản ghi trùng lặp bằng cách xóa bớt bản ghi.
* **Xử lý giá trị thiếu (Missing Data):**  Bộ dữ liệu HeartAttackDataSet không có phát hiện dữ liệu thiếu ,nhưng nếu thay bằng một bộ dữ liệu khác thì những bản ghi thiếu dữ liệu ở một số biến có thể được loại bỏ hoặc xử lý bằng cách điền giá trị trung bình, giá trị phổ biến, hoặc sử dụng các kỹ thuật nội suy để dự đoán giá trị thiếu (tùy vào bài toán)

3.1.2.2. Mã hóa biến (Data Encoding): Bộ dữ liệu HeartAttackDataSet có dữ liệu đã được mã hóa về dạng số ,nhưng nếu thay bằng một bộ dữ liệu khác thì có thể xử lý các biến định tính bằng việc chuyển đổi chúng thành dạng số, giúp mô hình học máy có thể xử lý. Phương pháp mã hóa phổ biến bao gồm:

* **One-Hot Encoding:** Dùng cho các biến danh mục không thứ tự như giới tính, loại đau ngực, trạng thái thalassemia, chuyển chúng thành các biến nhị phân.
* **Label Encoding:** Dùng cho các biến danh mục có thứ tự như mức độ nghiêm trọng của bệnh.

### 3.1.3 Tiền xử lý dữ liệu(Data Preprocessing)

3.1.3.1. Loại bỏ ngoại lệ (Outliers): Những giá trị ngoại lệ có thể được xác định và loại bỏ hoặc xử lý tùy thuộc vào nguyên nhân và mức độ ảnh hưởng đến mô hình.

3.1.3.2 . Chuẩn hóa hóa dữ liệu (Normalization): Việc chuẩn hóa và tiêu chuẩn hóa là cần thiết để đảm bảo rằng các đặc trưng định lượng có thang đo tương tự nhau, giúp mô hình học máy hoạt động hiệu quả hơn**. Chuẩn hóa (Normalization):** Thường được sử dụng để đưa các giá trị về một khoảng nhất định (ví dụ từ 0 đến 1) nhằm loại bỏ sự khác biệt về đơn vị đo

3.1.3.3. Rời rạc hóa dữ liệu(Data Discretization):Rời rạc hóa là quá trình chuyển đổi dữ liệu liên tục (continuous data) thành các dữ liệu rời rạc (discrete data) bằng cách chia nhỏ khoảng giá trị liên tục thành các nhóm hoặc phân đoạn (bins). Mỗi nhóm sau đó sẽ được gán một nhãn hoặc giá trị cụ thể.

3.1.3.4. Giảm chiều dữ liệu (Dimensionality Reduction): Sử dụng kỹ thuật giảm chiều dữ liệu PCA áp dụng để giảm số lượng đặc trưng mà vẫn giữ lại phần lớn thông tin quan trọng.

3.1.3.5. Xử lý dữ liệu không cân bằng (Handling Imbalanced Data)**:** Bộ dữ liệu thực nghiệm tuy mất cân bằng nhẹ nhưng không cần xử lý.Dữ liệu không cân bằng là vấn đề phổ biến trong các bài toán y học, khi số lượng bệnh nhân mắc bệnh và không mắc bệnh chênh lệch đáng kể. Một số phương pháp xử lý bao gồm:

* **Tái lấy mẫu (Resampling):** Bao gồm kỹ thuật tăng cường dữ liệu (oversampling) cho lớp thiếu số hoặc giảm bớt dữ liệu (undersampling) cho lớp thừa số để cân bằng tập dữ liệu.
* **Sử dụng kỹ thuật tổng hợp (Synthetic Techniques):** Kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) tạo ra các mẫu mới cho lớp thiếu số bằng cách nội suy giữa các điểm dữ liệu gần nhau.

3.1.3.6. Phân chia tập dữ liệu (Data Splitting): Để đánh giá mô hình một cách khách quan, dữ liệu thường được chia thành các tập huấn luyện (training set) và kiểm thử (test set). Phân chia này giúp đảm bảo rằng mô hình được đánh giá trên dữ liệu mà nó chưa từng thấy trước đó, giảm thiểu nguy cơ overfitting.

* **Tập huấn luyện (Training Set):** Chiếm khoảng 70% tổng số dữ liệu và được sử dụng để huấn luyện mô hình.
* **Tập kiểm thử (Test Set):** Chiếm khoảng 30% còn lại và được sử dụng để đánh giá hiệu suất của mô hình trên dữ liệu mới.

## 3.2 Triển khai các mô hình học máy

Triển khai các mô hình học máy là bước quan trọng trong quy trình dự báo bệnh tim, khi các mô hình học máy khác nhau được lựa chọn, huấn luyện, và đánh giá hiệu suất. Mục tiêu là tìm ra mô hình có khả năng dự đoán tốt nhất trên tập dữ liệu thực tế, đồng thời đảm bảo rằng mô hình có thể được áp dụng trong các tình huống y tế để hỗ trợ các quyết định lâm sàng.

### 3.2.1. Lựa chọn mô hình học máy

#### 3.2.1.**1. Lý do lựa chọn mô hình:**

* ID3 là một thuật toán xây dựng cây quyết định, rất phù hợp để dự đoán và phân loại các loại bệnh dựa trên dữ liệu đầu vào. Một số lý do chính bao gồm:
* **Dễ hiểu và giải thích kết quả**:
  + ID3 tạo ra cây quyết định rõ ràng và dễ hiểu, giúp các bác sĩ hoặc chuyên gia y tế dễ dàng giải thích được các yếu tố quyết định trong dự đoán bệnh tim.
* **Phù hợp với dữ liệu phân loại**:
  + Bệnh tim thường được dự đoán dựa trên các biến phân loại hoặc nhị phân (ví dụ: có nguy cơ mắc bệnh hay không), và ID3 hoạt động hiệu quả trong việc xử lý các loại dữ liệu này.
* **Tập trung vào thông tin quan trọng**:
  + ID3 sử dụng **entropy** và **information gain** để chọn các thuộc tính quan trọng nhất, giúp giảm độ phức tạp và tăng độ chính xác của mô hình.
* **Khả năng xử lý dữ liệu thiếu**:
  + Trong nhiều trường hợp, dữ liệu y tế có thể không đầy đủ. ID3 có thể xử lý dữ liệu thiếu một cách tương đối dễ dàng.
* PCA là một kỹ thuật giảm số chiều được sử dụng để xử lý dữ liệu phức tạp, giúp cải thiện hiệu quả và độ chính xác của mô hình dự đoán. Các lý do chính bao gồm:
* **Giảm số chiều của dữ liệu**:
  + Dữ liệu y tế thường chứa nhiều biến số (ví dụ: tuổi, huyết áp, cholesterol, nhịp tim, v.v.), và không phải tất cả các biến này đều quan trọng như nhau. PCA giúp giảm số lượng biến đầu vào, giữ lại các thành phần chính có ý nghĩa nhất.
* **Loại bỏ sự dư thừa**:
  + Trong dữ liệu y tế, các biến thường có mối tương quan cao (ví dụ, huyết áp và nhịp tim). PCA loại bỏ các thông tin trùng lặp, giúp mô hình hoạt động nhanh hơn và chính xác hơn.
* **Giảm nhiễu**:
  + PCA giúp loại bỏ nhiễu từ các biến không quan trọng, giúp cải thiện hiệu suất dự đoán.
* **Tăng hiệu quả tính toán**:
  + Sau khi giảm số chiều, các thuật toán như ID3 có thể xử lý dữ liệu nhanh hơn mà không làm giảm đáng kể độ chính xác.
* **Kết hợp ID3 và PCA**
* PCA được sử dụng trước tiên để tiền xử lý dữ liệu, giảm số chiều và chuẩn bị dữ liệu sạch hơn, tập trung hơn cho ID3.
* ID3 sau đó được sử dụng để xây dựng cây quyết định, tận dụng dữ liệu đã được tinh giản và tối ưu từ PCA để tạo ra các dự đoán chính xác.
* Kết hợp cả hai giúp:

1. Tăng hiệu suất của mô hình.
2. Giảm thời gian tính toán.
3. Tạo ra một mô hình dễ hiểu và có khả năng ứng dụng cao trong thực tiễn y tế.

#### 3.2.1.**2. Các mô hình được sử dụng:**

* **ID3** **(Iterative Dichotomiser 3):**
* Cây quyết định là một mô hình có giám sát (supervised learning), có thể được áp dụng cho cả bài toán phân lớp (classification) và hồi quy (regression)
* Input: Tập dữ liệu huấn luyện
* Output: Cây quyết định được xây dựng từ tập dữ liệu huấn luyện
* Cách thực hiện:
* Cây quyết định được xây dựng theo cách chia để trị từ trên xuống (top-down)Tại vị trí khởi tạo, tất cả các mẫu thuộc nút gốc
* Các thuộc tính được phân loại (nếu giá trị của thuộc tính là liên tục, thì phải được rời rạc hoá trước)
* Các mẫu được phân chia đệ quy dựa vào các thuộc tính được chọn
* Thuộc tính kiểm tra được lựa chọn dựa vào kinh nghiệm (heuristic) hoặc độ đo thống kê (statistical measure) (vid dụ, information gain)
* Điều kiện dừng phân chia
  + Tất cả các mẫu của nút xem xét thuộc cùng một lớp
  + Không có thuộc tính nào để phân chia - biểu quyết đa số được sử dụng để gán nhãn phân loại cho lá
  + Không còn mẫu nào
* Hàm số Entropy

A mathematical equation with a square and a triangle

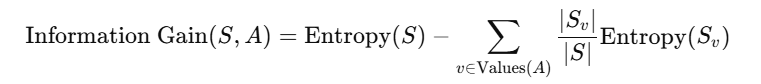
Description automatically generated with medium confidence

Trong đó:

pi​: Xác suất của từng lớp i.

c: Tổng số lớp.

* Information Gain đo lượng giảm của Entropy khi chia dữ liệu dựa trên một thuộc tính



Trong đó:

S: Tập dữ liệu hiện tại.

A: Thuộc tính cần xét.

Sv​: Tập con của S khi thuộc tính A có giá trị v.

* **Principal Component Analysis (PCA):**
* Dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân bố gần các đường/mặt đặc biệt nào đó.
* PCA xem xét một trường hợp đặc biệt khi các mặt đặc biệt đó có dạng tuyến tính là các không gian con (subspace).
* Input: Tập dữ liệu ban đầu
* Output: Tập dữ liệu mới tối ưu và chính xác
* Cách thực hiện:
* Bước 1: Tính vector kỳ vọng của toàn bộ dữ liệu:
* Bước 2: Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu:
* Bước 3: Tính ma trận hiệp phương sai S:
* Bước 4: Tính các trị riêng và vector riêng của S và sắp xếp chúng theo thứ tự giảm dần của trị riêng. (lưu ý về sự trực giao của các ui)

…

* Bước 5: Chọn 𝐾 vector riêng ứng với 𝐾 trị riêng lớn nhất để xây dựng ma trận 𝐔k có các cột tạo thành một hệ trực giao. 𝐾 vectors này là các thành phần chính
* Bước 6: Chiếu dữ liệu ban đầu đã chuẩn hoá xuống không gian con tìm được.
* Bước 7: Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới

### 3.2.2 Lưu đồ thuật toán để huấn luyện mô hình

**A diagram of a work flow

Description automatically generated**

##### Hình 3.2.1 : Lưu đồ thuật toán huấn luyện mô hình

Lưu đồ thuật toán thể hiện các bước cơ bản để huấn luyện mô hình dự đoán khả năng mắc bệnh tim mạch. Dưới đây là giải thích chi tiết cho từng bước trong lưu đồ:

**B1. Bắt đầu:**

Khởi động quá trình huấn luyện mô hình.

**B2. Nhập thư viện cần thiết:**

Import các thư viện cần thiết như numpy, pandas, matplotlib, seaborn, scikit-learn để xử lý dữ liệu, xây dựng và đánh giá mô hình.

**B3. Tải dữ liệu từ CSV:**

Đọc dữ liệu từ tệp CSV và lưu vào DataFrame.

**B4. Khám phá và tiền xử lý dữ liệu:**

Kiểm tra dữ liệu, làm sạch dữ liệu, tiền xử lý dữ liệu

**B5. Chia dữ liệu thành tập huấn luyện và kiểm tra:**

Sử dụng hàm train\_test\_split để chia dữ liệu thành tập huấn luyện và tập kiểm tra theo tỉ lệ thích hợp (70-30).

**B6. Chọn và huấn luyện mô hình:**

Lựa chọn thuật toán phù hợp và hiệu quả nhất để huấn luyện mô hình.

So sánh hiệu quả giữa các thuật toán ID3 khi có và không kết hợp với PCA.

**B7. Đánh giá mô hình:**

Đánh giá mô hình bằng cách sử dụng các phép đo như confusion matrix, precision, recall, và f1-score.Kiểm tra độ chính xác của mô hình. Nếu độ chính xác dưới 70%, điều chỉnh các thông số huấn luyện và lặp lại bước huấn luyện.

**B8.. Tối ưu hóa mô hình :**

Tinh chỉnh tham số , áp dụng nhiều kỹ thuật xử lý khác nhau, làm sạch dữ liệu

**B9. Lưu mô hình :**

Lưu trữ mô hình đã huấn luyện để sử dụng sau này.

**B10. Tạo hàm và ứng dụng của mô hình vào thực tế:**

Để đáp ứng cho nhu cầu tối ưu trong lĩnh vực y tế và bệnh của bệnh nhân việc tạo ra hàm PCA\_method sẽ giúp tiết kiệm nhiều thời gian cho việc tối ưu hóa các thuật toán khi kết hợp cùng PCA

**B11. Kết thúc:**

Khi độ chính xác của mô hình đạt trên 70%, dự đoán thử trên một vài mẫu có trong tập test và tạo báo cáo, đồ thị để so sánh với các thuật toán khác. Bằng cách tuân thủ các bước trong lưu đồ, mô hình có thể được huấn luyện và tối ưu hóa để đạt được hiệu suất tốt nhất trong việc dự đoán bệnh tim mạch.

### 3.2.3 Huấn luyện mô hình (Model Training)

3.2.3.1. Chia dữ liệu thành các tập huấn luyện và kiểm thử: Dữ liệu ban đầu được chia thành hai tập chính: tập huấn luyện (training set) để huấn luyện mô hình và tập kiểm thử (test set) để đánh giá hiệu suất của mô hình. Tỷ lệ chia thường là 70% dữ liệu cho huấn luyện và 30% cho kiểm thử( có thể tùy theo kích thước và đặc điểm của dữ liệu.)

3.2.3.2. Tinh chỉnh siêu tham số (Hyperparameter Tuning): Việc tinh chỉnh siêu tham số là quá trình thử nghiệm với nhiều giá trị khác nhau của các tham số như : Giới hạn độ sâu cây (max\_depth) và số lượng mẫu tối thiểu (min\_samples\_split) để tránh quá khớp.

### 3.2.4 Đánh giá kết quả mô hình

3.2.4.1. Sử dụng các tiêu chí đánh giá: Mỗi mô hình học máy sau khi huấn luyện được đánh giá trên tập dữ liệu kiểm thử sử dụng một số tiêu chí chính, bao gồm:

* **Độ chính xác (Accuracy):** Tỷ lệ dự đoán đúng trên tổng số mẫu kiểm thử. Đây là chỉ số tổng quát để đo lường hiệu suất của mô hình.
* **Độ nhạy (Recall):** Tỷ lệ mẫu dương tính (người mắc bệnh) được dự đoán đúng trên tổng số mẫu dương tính thực tế. Chỉ số này quan trọng trong y tế, nơi việc phát hiện bệnh nhân có nguy cơ cao là ưu tiên hàng đầu.
* **F1-Score:** Trung bình điều hòa của Precision và Recall, chỉ số này hữu ích khi cần cân bằng giữa việc phát hiện bệnh (Recall) và độ chính xác của dự đoán dương tính (Precision).

#### 3.2.4**.2. Sử dụng biểu đồ và ma trận đánh giá:**

* **Confusion Matrix:** Ma trận lỗi (confusion matrix) được sử dụng để cung cấp một cái nhìn chi tiết về kết quả dự đoán của mô hình, bao gồm số lượng dự đoán đúng, sai, dương tính giả (false positives), và âm tính giả (false negatives). Từ đó, có thể đánh giá chính xác những điểm mạnh và điểm yếu của mô hình.

### 3.2.5 So sánh kết quả giữa các mô hình

3.2.5.1. So sánh trên các tiêu chí đánh giá: Sau khi có kết quả đánh giá cho từng mô hình, tiến hành so sánh các chỉ số như Accuracy, Recall, Specificity, F1-Score. Mô hình nào có giá trị cao hơn trên các tiêu chí quan trọng sẽ được coi là tốt hơn.

3.2.5.2. Xem xét sự cân bằng giữa các tiêu chí: Không phải lúc nào một mô hình có Accuracy cao cũng là lựa chọn tốt nhất, đặc biệt trong y học. Ví dụ, mô hình có Recall cao nhưng Specificity thấp có thể dẫn đến nhiều cảnh báo giả (false positives), gây lo lắng không cần thiết cho bệnh nhân. Do đó, cần xem xét sự cân bằng giữa các tiêu chí để lựa chọn mô hình có hiệu suất toàn diện.

3.2.5.3. Lựa chọn mô hình tốt nhất: Cuối cùng, sau khi đánh giá và so sánh, mô hình có hiệu suất tổng thể tốt nhất trên các tiêu chí quan trọng sẽ được lựa chọn. Lựa chọn này không chỉ dựa trên kết quả của các chỉ số mà còn dựa trên khả năng áp dụng thực tế, tính dễ hiểu, và khả năng triển khai trong môi trường y tế.

### 3.2.6. Kiểm thử với dữ liệu mới

3.2.6.1. Đánh giá mô hình trên tập dữ liệu mới: Để đảm bảo rằng mô hình không chỉ hoạt động tốt trên tập kiểm thử ban đầu, mô hình sẽ được kiểm thử thêm trên tập dữ liệu mới (nếu có). Điều này giúp xác định khả năng tổng quát hóa của mô hình đối với dữ liệu chưa từng thấy trước đó.

3.2.6.2. So sánh với các kết quả trước đó: Kết quả từ tập dữ liệu mới được so sánh với các kết quả trước đó để kiểm tra tính ổn định và độ tin cậy của mô hình. Nếu mô hình hoạt động nhất quán trên cả hai tập dữ liệu, điều đó cho thấy mô hình có thể được áp dụng trong các tình huống thực tế với độ tin cậy cao.

### 3.2.7 Thiết kế giao diện đồ họa người dùng

* Mô tả giao diện: Giao diện gồm các thành phần
  + Khung nhập thông tin bệnh nhân gồm các trường dữ liệu đầu vào
  + Khung thông tin kế quả
* Hướng dẫn sử dụng
  + Bước 1: Nhập thông tin bệnh nhân vào khung “Thông tin bệnh nhân”
  + Bước 2: Bấm “Kết quả dự đoán” tại khung “Kết quả dự đoán”
* Hiển thị kết quả dự đoán dưới nút “Kết quả dự đoán”

A diagram of a work flow

Description automatically generated

##### Hình 3.1.2 :Biểu đồ sử dụng mô hình

# IV. Chương 4: Kết Quả và Thảo Luận

## 4.1 Kết quả phân tích dữ liệu

Phân tích dữ liệu là một bước quan trọng nhằm hiểu rõ hơn về các đặc trưng của tập dữ liệu và mối quan hệ giữa chúng, từ đó cung cấp cơ sở cho việc xây dựng và triển khai các mô hình học máy. Phần này sẽ trình bày chi tiết kết quả của các phân tích mô tả dữ liệu và khám phá những mẫu (pattern) tiềm ẩn trong dữ liệu.

### 4.1.1 Thống kê mô tả

#### **4.1.1.1. Tổng quan về dữ liệu sau khi làm sạch:**

* Tập dữ liệu gồm 302 mẫu (số lượng bệnh nhân), bao gồm 13 thuộc tính độc lập và 1 thuộc tính phụ thuộc (‘target’)

A screenshot of a computer program

Description automatically generated

* Sự phân bố giá trị của thuộc tính ‘target’:

A graph of a pie chart

Description automatically generated

##### Hình 4.1: Sự phân bố giá trị của thuộc tính ‘target’

* Dữ liệu cột 'target' mất cân bằng nhẹ nhưng không cần thiết phải cân bằng
* Phân phối độ tuổi của các bệnh nhân:

A graph with blue and brown bars

Description automatically generated

##### Hình 4.2: Phân phối độ tuổi theo thuộc tính ‘target’ của các bệnh nhân

- Nguy cơ mắc bệnh tim có xu hướng cao ở nhóm tuổi trung niên (40-60 tuổi).

- Nhóm tuổi cao (>70 tuổi) cũng có nguy cơ mắc bệnh tim tăng cao, mặc dù số lượng người ít hơn.

- Nhóm người trẻ tuối ít bị bệnh tim hơn.

* Phân phối giới tính: Tỷ lệ nam và nữ trong tập dữ liệu.

A graph with a bar and a rectangle

Description automatically generated with medium confidence

##### Hình 4.3: Phân phối giới tính theo thuộc tính ‘target’ của các bệnh nhân

- Nữ giới trong tập dữ liệu này có nguy cơ mắc bệnh cao hơn, mặc dù tổng số lượng nữ ít hơn nam.

- Nam giới chiếm đa số trong tập dữ liệu, với tỷ lệ không mắc bệnh cao hơn so với tỷ lệ mắc bệnh.

#### **4.1.1.2. Các biến số liên quan đến sức khỏe:**

* **Huyết áp (trestbps):** A graph with a line going up

  Description automatically generated

##### **Hình 4.**4: Biểu đồ phân phối thuộc tính **trestbps**

**- Có xu hướng lệch nhẹ về bên phải , với đa số các giá trị tập trung quanh khoảng 120-140.**

**-** Một số giá trị thấp hơn 110 mmHg hoặc cao hơn 150 mmHg xuất hiện với tần suất nhỏ hơn, biểu thị các trường hợp ngoại lệ hoặc ít phổ biến.

**- Đường trung bình (Mean) được hiển thị tại khoảng 130.07, nằm gần tâm của phân phối chính.**

* **Cholesterol (chol):**

A graph with a line

Description automatically generated

##### **Hình 4.**5: Biểu đồ phân phối thuộc tính **chol**

**- Cột ‘chol’ có phân phối gần như đối xứng, với tần xuất cao tập trung trong khoảng 200-280mg/dL**

- Giá trị trung bình (Mean) của chol, vào khoảng 241.56 mg/dL,chỉ ra xu hướng mức cholesterol hơi cao, có thể là một dấu hiệu cần theo dõi trong nghiên cứu sức khỏe tim mạch.

* **Nhịp tim tối đa (thalach):**

A graph with a line going up

Description automatically generated

##### **Hình 4.**6: Biểu đồ phân phối thuộc tính **thalach**

- Biểu đồ cho thấy thalach có phân phối gần giống hình chuông, khá cân đối với phần lớn giá trị tập trung trong khoảng từ 130 đến 170 nhịp/phút.

- Giá trị trung bình 149.83 nhịp/phút phù hợp với phân phối và cho thấy mức nhịp tim tối đa điển hình của các bệnh nhân trong tập dữ liệu.

#### **4.1.1.3. Phân tích các yếu tố nguy cơ:**

* **Tần suất các loại đau ngực (cp):** A graph with blue bars

  Description automatically generated

##### **Hình 4.**7: **Tần suất các loại đau ngực (cp)**

- Biểu đồ cho thấy phần lớn bệnh nhân có đau ngực điển hình (~47%) và không điển hình (~30%), phản ánh nguy cơ cao liên quan đến bệnh tim mạch

* **Đường huyết lúc đói (fbs):**

A graph with red squares

Description automatically generated with medium confidence

##### **Hình 4.**8: **Tỷ lệ bệnh nhân mắc bệnh có đường huyết>120mg/dL (fbs)**

- Biểu đồ cho thấy phần lớn bệnh nhân (hơn 80%) có đường huyết lúc đói ≤ 120 mg/dl, trong khi chỉ khoảng dưới 20% có mức đường huyết > 120 mg/dl. Điều này cho thấy đa số bệnh nhân không có nguy cơ cao về đường huyết, nhưng nhóm nhỏ có nguy cơ cao vẫn cần được theo dõi và kiểm soát cẩn thận.

* **Các bất thường điện tâm đồ (restecg):**

A graph with green bars

Description automatically generated

##### **Hình 4.**9: **Tỷ lệ bệnh nhân mắc bệnh có bất thường điện tâm đồ (restecg)**

- Biểu đồ cho thấy tỷ lệ bệnh nhân có kết quả điện tâm đồ bình thường và bất thường ST-T là gần tương đương (~50% mỗi loại), trong khi phì đại thất trái chiếm tỷ lệ rất thấp.

### 4.1.2 Phân tích mối quan hệ giữa các biến số

#### 4.1.2.**1. Tương quan giữa các biến số:**

* **Biểu đồ nhiệt (heatmap):** A diagram of different colored squares

  Description automatically generated

##### **Hình 4.**10: **Biểu đồ nhiệt**

- Biến cp (loại đau ngực) có tương quan dương mạnh với target (bệnh tim) (~0.43).

- exang (đau ngực do gắng sức) và oldpeak có tương quan âm mạnh với target (~-0.44 và -0.43).

- Các biến khác có tương quan yếu hoặc không đáng kể.

#### **4.1.2.2. Phân tích mối quan hệ giữa các yếu tố nguy cơ và bệnh tim:**

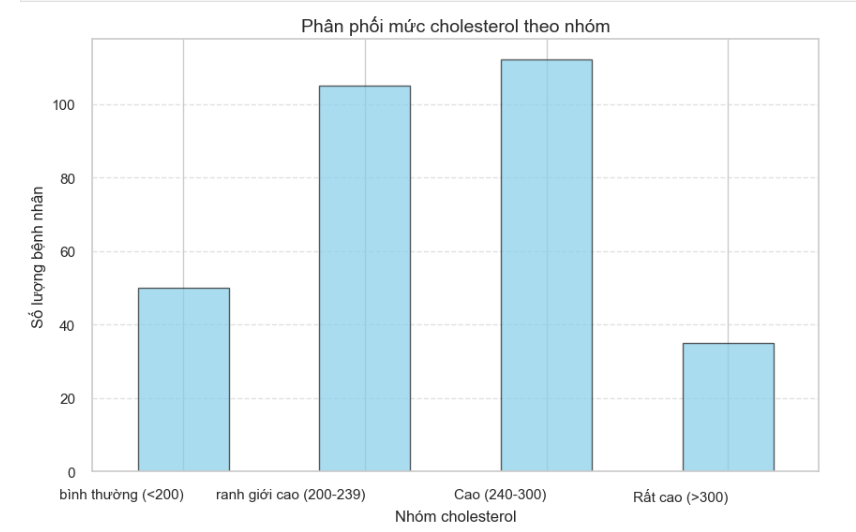
* **Đau ngực và bệnh tim:** A graph with blue bars

  Description automatically generated

##### **Hình 4.**11: **Tỷ lệ bệnh nhân mắc bệnh tim theo loại đau ngực**

- Biểu đồ cho thấy tỷ lệ mắc bệnh tim cao nhất ở nhóm "Đau thắt ngực không điển hình" (80%) và thấp nhất ở nhóm "Không có triệu chứng" (20%).

* **Cholesterol và bệnh tim:**



##### **Hình 4.**12: Phân phối mức cholesterol theo nhóm

- Biểu đồ cho thấy nhóm cholesterol cao (240-300) chiếm số lượng bệnh nhân lớn nhất, tiếp theo là nhóm ranh giới cao (200-239). Nhóm rất cao (>300) và bình thường (<200) có số lượng bệnh nhân thấp hơn đáng kể.

## 4.2 Hiệu quả của các mô hình học máy

Trong phần này, chúng ta sẽ đánh giá hiệu quả của các mô hình học máy được triển khai nhằm dự báo bệnh tim. Hiệu quả của mô hình không chỉ được đo lường qua các chỉ số thống kê, mà còn qua khả năng ứng dụng thực tế và độ tin cậy khi áp dụng cho các tập dữ liệu khác nhau.

### 4.2.1 Kết quả huấn luyện và kiểm thử

A white rectangular grid with black text

Description automatically generated

##### **Hình 4.**11: **So sánh hiệu xuất mô hình**

### 4.2.2**. Hiệu suất trên tập huấn luyện:**

* **Accuracy:**

- Mô hình không sử dụng PCA đạt độ chính xác khá cao (~94.31%), cho thấy khả năng học tốt từ dữ liệu.

- Mô hình sử dụng PCA đạt độ chính xác tuyệt đối (100%), có thể là dấu hiệu của overfitting nếu hiệu suất trên tập kiểm thử không tương đương.

* **Precision, Recall, F1-Score:**

- Mô hình sử dụng PCA cũng đạt giá trị tuyệt đối trên cả Precision, Recall và F1-Score, thể hiện rằng mô hình này hoạt động rất tốt trên dữ liệu huấn luyện.

### 4.2.3**. Hiệu suất trên tập kiểm thử:**

* **Accuracy:**

- Mô hình không sử dụng PCA đạt Accuracy 79.12%, trong khi mô hình sử dụng PCA cao hơn một chút (80.22%), thể hiện rằng PCA cải thiện khả năng tổng quát hóa một chút.

* **Recall và Precision:**

- Mô hình sử dụng PCA có Precision (82.98%) cao hơn so với không sử dụng PCA (85.71%), nhưng Recall (79.59%) lại tốt hơn, cho thấy PCA có khả năng cân bằng giữa việc phát hiện đúng và giảm sai sót.

* **F1-Score:**

- F1-Score của mô hình sử dụng PCA (81.25%) cao hơn so với không sử dụng PCA (79.12%), khẳng định rằng PCA giúp cải thiện khả năng cân bằng giữa Precision và Recall.

**Sự khác biệt giữa tập huấn luyện và kiểm thử:**

* **Overfitting (quá khớp):** Mô hình có độ chính xác rất cao trên tập huấn luyện nhưng lại có hiệu suất kém hơn trên tập kiểm thử có thể đang gặp phải vấn đề quá khớp, khi mô hình học thuộc tính cụ thể của tập huấn luyện thay vì học các đặc điểm chung.

## 4.3 Giao diện của mô hình

### 4.3.1. Mô tả tổng quan giao diện:

**A screenshot of a computer

Description automatically generated**

##### **Hình 4.**12: **Giao diện của mô hình dự đoán**

* **Mục đích:** Giao diện cho phép nhập các thông tin sức khỏe của bệnh nhân và sử dụng mô hình máy học để dự đoán khả năng mắc bệnh tim.
* **Thành phần giao diện chính:**
  + **Thông tin nhập của bệnh nhân:** Người dùng có thể nhập các đặc trưng sức khỏe như tuổi, giới tính, loại đau ngực, huyết áp, cholesterol, điện tâm đồ, và các chỉ số khác.
  + **Kết quả dự đoán:** Hiển thị kết quả dự đoán (có bị bệnh tim hay không) và các chỉ số hiệu suất của mô hình (độ chính xác, precision, recall, F1-score).
  + **Hai phương pháp dự đoán:**
    1. Không sử dụng PCA.
    2. Sử dụng PCA để giảm số chiều dữ liệu đầu vào.

### 4.3.2. Cách hoạt động:

* Người dùng nhập các thông tin sức khỏe vào các ô nhập liệu hoặc chọn từ các menu thả xuống.
* Giao diện cung cấp hai nút bấm:
  + **Dự đoán theo ID3:** Dự đoán dựa trên mô hình Decision Tree không sử dụng PCA.
  + **Dự đoán theo ID3\_PCA:** Dự đoán dựa trên mô hình Decision Tree kết hợp với PCA.
* Khi người dùng nhấn nút, chương trình kiểm tra dữ liệu nhập và hiển thị kết quả.

### 4.3.3. Ưu điểm:

* **Giao diện trực quan:** Các ô nhập liệu và nút bấm được bố trí gọn gàng, dễ sử dụng.
* **Hiển thị chi tiết:** Kết quả dự đoán kèm theo các chỉ số hiệu suất của mô hình, giúp đánh giá độ tin cậy.
* **Lựa chọn linh hoạt:** Cung cấp hai phương pháp dự đoán (với và không có PCA) để người dùng so sánh.
* **Kiểm tra dữ liệu nhập:** Chương trình có cơ chế kiểm tra thông tin nhập liệu, đảm bảo dữ liệu hợp lệ.

### 4.4.4. Nhược điểm:

* **Phụ thuộc dữ liệu nhập:** Nếu thông tin nhập thiếu hoặc không chính xác, kết quả dự đoán sẽ bị ảnh hưởng.
* **Giới hạn về hiển thị:** Kết quả chỉ hiển thị trong giao diện mà không lưu lại để phân tích sau.
* **Độ phức tạp:** Với người dùng không quen thuộc, việc nhập liệu có thể gây khó khăn nếu không có hướng dẫn chi tiết.

### 4.4.5. Mô hình đề xuất

**Mô hình được đề xuất:** Mô hình **Decision Tree kết hợp PCA** được lựa chọn để triển khai trong môi trường thực tế.

#### 4.4.5.1 Lý do đề xuất:

**Hiệu suất tốt hơn:** Kết quả so sánh chỉ ra rằng mô hình sử dụng PCA có độ chính xác (accuracy), F1-score cao hơn một chút so với mô hình không sử dụng PCA. Điều này thể hiện khả năng tổng quát hóa tốt hơn trên dữ liệu kiểm thử.

**Tối ưu dữ liệu:** PCA giúp giảm số chiều dữ liệu đầu vào mà vẫn giữ được các thông tin quan trọng, giảm thời gian xử lý và giúp mô hình hoạt động hiệu quả hơn.

**Khả năng giải thích:** Decision Tree là một mô hình dễ hiểu, trực quan, phù hợp để giải thích và trình bày với các chuyên gia y tế.

**Độ tin cậy:** F1-score cao đảm bảo rằng mô hình đạt được sự cân bằng giữa việc phát hiện bệnh nhân có nguy cơ và giảm thiểu chẩn đoán sai.

#### 4.4.5.2. Tính khả thi và hiệu quả trong thực tế

* **Cách triển khai:**
  + Mô hình có thể được tích hợp vào hệ thống quản lý thông tin y tế (HIS) hoặc ứng dụng hỗ trợ quyết định lâm sàng (CDSS).
  + Các bác sĩ hoặc nhân viên y tế nhập thông tin bệnh nhân (tuổi, giới tính, huyết áp, cholesterol, v.v.) qua giao diện hệ thống, và mô hình sẽ dự đoán khả năng mắc bệnh tim.
* **Khả năng mở rộng:**
  + Mô hình có thể được áp dụng tại các phòng khám đa khoa, bệnh viện lớn hoặc tích hợp vào các thiết bị IoT y tế.
  + Khi có thêm dữ liệu bệnh nhân, mô hình có thể được huấn luyện lại định kỳ để cải thiện độ chính xác và thích ứng với các thay đổi trong dữ liệu thực tế.
* **Tích hợp hệ thống:**
  + Các API có thể được xây dựng để kết nối mô hình với hệ thống HIS, cho phép tự động dự đoán và lưu trữ kết quả vào hồ sơ bệnh án điện tử.
  + Hệ thống có thể gửi cảnh báo tự động đến bác sĩ khi phát hiện nguy cơ cao, giúp tăng cường hiệu quả điều trị.

#### 4.4.5.3. Phân tích lợi ích và rủi ro

* **Lợi ích:**
  + **Phát hiện sớm:** Giúp xác định bệnh nhân có nguy cơ cao mắc bệnh tim từ giai đoạn sớm, tạo điều kiện cho các can thiệp kịp thời.
  + **Hỗ trợ quyết định:** Cung cấp thông tin hỗ trợ cho bác sĩ trong việc đưa ra quyết định điều trị, đặc biệt tại các cơ sở y tế thiếu chuyên gia tim mạch.
  + **Tối ưu hóa tài nguyên:** Hỗ trợ phân bổ tài nguyên y tế hợp lý, tập trung vào các bệnh nhân có nguy cơ cao.
  + **Dễ dàng sử dụng:** Mô hình có giao diện thân thiện, dễ dàng triển khai tại các cơ sở y tế.
* **Rủi ro:**
  + **Chẩn đoán sai (False Positives):** Dự đoán nhầm bệnh nhân không có bệnh là có bệnh, dẫn đến lo lắng không cần thiết hoặc các can thiệp y tế không cần thiết.
  + **Bỏ sót (False Negatives):** Không phát hiện đúng bệnh nhân có nguy cơ cao, dẫn đến việc bỏ lỡ cơ hội điều trị.
  + **Phụ thuộc vào dữ liệu:** Chất lượng và độ chính xác của dự đoán phụ thuộc vào dữ liệu đầu vào. Dữ liệu không đầy đủ hoặc không chính xác có thể ảnh hưởng đến hiệu suất mô hình.
  + **Khả năng bảo mật:** Cần đảm bảo an toàn dữ liệu bệnh nhân khi triển khai mô hình trong hệ thống y tế.

## 4.4 Thảo luận về các yếu tố ảnh hưởng đến độ chính xác

Độ chính xác của các mô hình học máy trong dự báo bệnh tim phụ thuộc vào nhiều yếu tố khác nhau, từ chất lượng dữ liệu, lựa chọn mô hình, phương pháp tiền xử lý, cho đến các yếu tố bên ngoài như cỡ mẫu và tính đa dạng của dữ liệu. Phần này sẽ thảo luận chi tiết về các yếu tố chính ảnh hưởng đến độ chính xác của các mô hình, từ đó rút ra những bài học và khuyến nghị cho việc xây dựng và triển khai mô hình học máy trong y học.

### **4.4.1. Độ chính xác của dữ liệu đầu vào:**

* **Dữ liệu bị thiếu hoặc sai lệch:** Khi dữ liệu đầu vào chứa nhiều giá trị bị thiếu hoặc không chính xác, mô hình có thể gặp khó khăn trong việc học các mẫu (patterns) chính xác. Ví dụ, nếu thông tin về các chỉ số sinh học quan trọng như cholesterol hoặc huyết áp không đầy đủ hoặc bị sai lệch, mô hình sẽ không thể dự đoán chính xác bệnh tim.
* **Sai số đo lường:** Sai số trong quá trình thu thập dữ liệu, chẳng hạn như sai số trong đo lường nhịp tim hoặc huyết áp, có thể dẫn đến việc mô hình học các mẫu không chính xác, từ đó làm giảm độ chính xác dự đoán.

### **4.4.2. Tiền xử lý dữ liệu:**

* **Chuẩn hóa và điều chỉnh dữ liệu:** Việc không chuẩn hóa hoặc điều chỉnh dữ liệu (như không chuẩn hóa các giá trị của các biến về cùng một thang đo) có thể làm cho mô hình học máy khó khăn trong việc học, đặc biệt là các mô hình tuyến tính như Logistic Regression. Điều này có thể dẫn đến việc mô hình hoạt động kém hiệu quả và không ổn định.
* **Xử lý dữ liệu bị thiếu:** Cách xử lý dữ liệu bị thiếu, như việc thay thế bằng giá trị trung bình hoặc loại bỏ các mẫu thiếu, cũng có thể ảnh hưởng lớn đến độ chính xác của mô hình. Phương pháp xử lý không phù hợp có thể làm mất đi thông tin quan trọng và gây ra sai lệch trong quá trình huấn luyện mô hình.

### **4.4.3. Lựa chọn mô hình:**

* **Mô hình đơn giản vs. mô hình phức tạp:** Mô hình quá đơn giản có thể không đủ khả năng để học các mối quan hệ phức tạp giữa các đặc trưng và bệnh tim (underfitting), trong khi mô hình quá phức tạp có thể học cả những nhiễu trong dữ liệu, dẫn đến quá khớp (overfitting). Cân nhắc chọn mô hình có độ phức tạp vừa phải, phù hợp với tập dữ liệu là yếu tố then chốt để đạt được độ chính xác cao.
* **Các giả định của mô hình:** Một số mô hình có các giả định nhất định về dữ liệu (như Linear Regression giả định mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc). Nếu dữ liệu không đáp ứng được các giả định này, mô hình có thể hoạt động kém và giảm độ chính xác.

### **4.4.4. Kỹ thuật huấn luyện:**

* **Kỹ thuật điều chỉnh mô hình:** Các kỹ thuật như điều chỉnh tham số (hyperparameter tuning), sử dụng các phương pháp chống quá khớp (như regularization), và phương pháp xác thực chéo (cross-validation) có thể ảnh hưởng lớn đến độ chính xác của mô hình. Việc tối ưu hóa không đúng cách có thể làm mô hình hoạt động kém hơn mong đợi.
* **Cách phân chia dữ liệu:** Cách chia dữ liệu thành các tập huấn luyện, kiểm thử và xác thực cũng ảnh hưởng đến độ chính xác của mô hình. Nếu dữ liệu không được chia đúng cách, mô hình có thể học quá nhiều từ tập huấn luyện và hoạt động kém trên tập kiểm thử.

### **4.4.5. Đa dạng hóa dữ liệu:**

* **Dữ liệu từ nhiều nguồn khác nhau:** Nếu dữ liệu được thu thập từ nhiều nguồn khác nhau với các phương pháp đo lường khác nhau, điều này có thể làm giảm độ chính xác của mô hình do sự không đồng nhất trong dữ liệu. Điều chỉnh và chuẩn hóa dữ liệu từ nhiều nguồn là cần thiết để duy trì độ chính xác của mô hình.
* **Tính cập nhật của dữ liệu:** Dữ liệu lỗi thời có thể không còn phù hợp với thực tế hiện tại, đặc biệt là trong các lĩnh vực như y tế, nơi các điều kiện và tiêu chuẩn có thể thay đổi nhanh chóng. Điều này yêu cầu mô hình phải được huấn luyện lại hoặc cập nhật thường xuyên để duy trì độ chính xác.

### **4.4.6. Khả năng giải thích của mô hình:**

* **Mô hình dễ hiểu vs. độ chính xác:** Các mô hình phức tạp như neural networks có thể đạt độ chính xác cao nhưng lại khó giải thích, gây khó khăn trong việc áp dụng trong y học. Điều này đòi hỏi sự cân bằng giữa khả năng giải thích và độ chính xác khi chọn mô hình cho các ứng dụng y tế.

### **4.4.7. Hạn chế của công nghệ và tính toán:**

* **Hạn chế về tài nguyên tính toán:** Việc huấn luyện các mô hình phức tạp có thể yêu cầu nhiều tài nguyên tính toán và thời gian, điều này có thể ảnh hưởng đến khả năng triển khai thực tế. Sự hạn chế này có thể làm giảm độ chính xác nếu mô hình không được huấn luyện đủ lâu hoặc không được tối ưu hóa kỹ càng.
* **Yếu tố nhân văn:** Sự can thiệp và diễn giải của con người trong quá trình phát triển và triển khai mô hình cũng có thể ảnh hưởng đến độ chính xác. Sự thiên lệch hoặc sai sót của con người có thể dẫn đến các quyết định sai lầm trong việc lựa chọn đặc trưng, mô hình, hoặc phương pháp xử lý dữ liệu.

### **4.4.8. Cải thiện chất lượng dữ liệu:**

* **Thu thập dữ liệu chính xác hơn:** Cải tiến quy trình thu thập dữ liệu để giảm sai số và tăng tính chính xác của dữ liệu.
* **Xử lý thiếu dữ liệu:** Áp dụng các phương pháp tiên tiến hơn để xử lý dữ liệu bị thiếu, như sử dụng mô hình dự đoán để thay thế các giá trị thiếu thay vì chỉ dùng giá trị trung bình.
* **Sử dụng kỹ thuật ensemble:** Kết hợp nhiều mô hình học máy khác nhau (ensemble methods) như Random Forests hoặc Gradient Boosting để tăng cường độ chính xác và tính ổn định của mô hình.
* **Cập nhật mô hình định kỳ:** Huấn luyện lại và cập nhật mô hình với dữ liệu mới để duy trì độ chính xác theo thời gian.
* **Sử dụng kỹ thuật hiện đại:** Ứng dụng các kỹ thuật hiện đại như deep learning hoặc reinforcement learning để cải thiện độ chính xác trong những bài toán phức tạp hơn.

# V. Hạn chế của nghiên cứu và hướng phát triển

Nghiên cứu này đã đạt được những kết quả đáng kể trong việc áp dụng các kỹ thuật học máy để dự báo bệnh tim, tuy nhiên, vẫn tồn tại một số hạn chế cần được xem xét và cải thiện trong các nghiên cứu tiếp theo. Phần này sẽ phân tích các hạn chế chính của nghiên cứu và đề xuất hướng phát triển nhằm khắc phục những hạn chế đó, từ đó nâng cao độ tin cậy và hiệu quả của các mô hình dự báo bệnh tim.

## 5.1 Hạn chế của nghiên cứu

### 5.1.**1. Chất lượng và phạm vi của dữ liệu:**

* **Dữ liệu hạn chế về quy mô và tính đa dạng:** Nghiên cứu chỉ sử dụng một tập dữ liệu cụ thể (ví dụ: từ cơ sở dữ liệu UCI Machine Learning Repository) với số lượng mẫu giới hạn và chủ yếu là từ một nguồn duy nhất. Điều này có thể không đủ để mô hình hóa tất cả các biến thể và đặc trưng của bệnh tim trong một dân số đa dạng hơn. Kết quả mô hình có thể không tổng quát hóa tốt cho các nhóm dân số khác hoặc không phản ánh chính xác tình trạng y tế tại các khu vực địa lý khác nhau.
* **Thiếu thông tin về các yếu tố nguy cơ khác:** Tập dữ liệu không bao gồm một số yếu tố nguy cơ quan trọng khác liên quan đến bệnh tim, chẳng hạn như tiền sử gia đình, chế độ ăn uống, mức độ hoạt động thể chất, hoặc căng thẳng tâm lý. Điều này có thể làm giảm khả năng mô hình dự đoán chính xác các trường hợp phức tạp hoặc không điển hình.

### 5.1**.2. Giới hạn của mô hình học máy:**

* **Khả năng giải thích của mô hình:** Một số mô hình phức tạp như các mô hình deep learning có thể đạt độ chính xác cao nhưng lại thiếu khả năng giải thích. Điều này làm cho việc áp dụng mô hình trong y tế trở nên khó khăn, vì các quyết định y tế yêu cầu phải có sự minh bạch và khả năng giải thích rõ ràng để bác sĩ và bệnh nhân có thể hiểu và tin tưởng.
* **Overfitting:** Mặc dù các biện pháp chống quá khớp (overfitting) đã được áp dụng, nhưng với một tập dữ liệu nhỏ và không đa dạng, vẫn có khả năng mô hình bị quá khớp với tập dữ liệu huấn luyện, dẫn đến giảm hiệu suất khi áp dụng cho các dữ liệu mới.

### 5.1**.3. Thiếu sót trong đánh giá mô hình:**

* **Hạn chế của các chỉ số đánh giá:** Nghiên cứu chủ yếu sử dụng các chỉ số đánh giá truyền thống như độ chính xác, độ nhạy, và độ đặc hiệu. Tuy nhiên, các chỉ số này có thể chưa đủ để đánh giá toàn diện hiệu quả của mô hình trong các tình huống phức tạp, chẳng hạn như dự đoán các trường hợp rủi ro cao.

### 5.1.**4. Tính ứng dụng thực tế:**

* **Thiếu sự kiểm nghiệm trên các tập dữ liệu thực tế khác:** Nghiên cứu chưa kiểm nghiệm mô hình trên các tập dữ liệu từ những nguồn khác nhau hoặc các hệ thống y tế thực tế khác. Điều này hạn chế khả năng mô hình được chấp nhận và triển khai rộng rãi trong các môi trường y tế khác nhau.
* **Hạn chế về tài nguyên tính toán:** Việc triển khai mô hình trên quy mô lớn hoặc trong môi trường thực tế có thể đòi hỏi nhiều tài nguyên tính toán hơn so với những gì đã được thử nghiệm trong nghiên cứu. Điều này có thể gây khó khăn cho việc triển khai mô hình trong thực tiễn, đặc biệt là tại các cơ sở y tế có hạn chế về hạ tầng công nghệ.

## 5.2 Hướng phát triển

### 5.2 **1. Mở rộng và cải thiện dữ liệu:**

* **Thu thập dữ liệu từ nhiều nguồn khác nhau:** Để cải thiện tính tổng quát và độ tin cậy của mô hình, nghiên cứu nên mở rộng phạm vi thu thập dữ liệu từ nhiều cơ sở y tế và dân số khác nhau, bao gồm các yếu tố nguy cơ bổ sung như lối sống, môi trường sống, và các chỉ số sinh học khác.
* **Tăng cường chất lượng dữ liệu:** Nâng cao chất lượng dữ liệu bằng cách sử dụng các kỹ thuật tiên tiến để xử lý dữ liệu bị thiếu, loại bỏ nhiễu, và chuẩn hóa các biến số. Điều này giúp cải thiện độ chính xác và tính nhất quán của các mô hình.

#### 5.2.2. **Phát triển và thử nghiệm các mô hình mới:**

* **Kết hợp các mô hình học máy:** Sử dụng các kỹ thuật ensemble để kết hợp nhiều mô hình học máy khác nhau, từ đó cải thiện độ chính xác và độ ổn định của dự báo. Các phương pháp như Random Forests, Gradient Boosting, và Stacking có thể mang lại hiệu quả tốt hơn so với việc chỉ sử dụng một mô hình đơn lẻ.
* **Phát triển mô hình deep learning:** Khám phá và triển khai các mô hình deep learning có khả năng tự động hóa quá trình lựa chọn đặc trưng và học các mối quan hệ phi tuyến tính phức tạp. Các mô hình này có thể được huấn luyện với dữ liệu lớn hơn và đa dạng hơn, từ đó cải thiện độ chính xác trong dự báo bệnh tim.

### 5.2.**3. Tăng cường đánh giá mô hình:**

* **Sử dụng các kỹ thuật đánh giá toàn diện:** Áp dụng các phương pháp đánh giá mới và toàn diện hơn, chẳng hạn như phân tích ROC-AUC, Precision-Recall curves, và các phương pháp kiểm định thống kê để hiểu rõ hơn về hiệu suất của mô hình trong các trường hợp khác nhau.
* **Thử nghiệm trên dữ liệu thực tế:** Kiểm nghiệm mô hình trên các tập dữ liệu thực tế từ nhiều nguồn khác nhau để đánh giá tính khả thi và hiệu quả khi áp dụng trong môi trường y tế thực tế. Điều này cũng giúp nhận diện và khắc phục những hạn chế mà mô hình có thể gặp phải trong thực tiễn.

### 5.2.**4. Nâng cao tính ứng dụng thực tế:**

* **Phát triển các công cụ hỗ trợ quyết định:** Xây dựng các hệ thống hỗ trợ quyết định dựa trên mô hình học máy, giúp bác sĩ và các chuyên gia y tế sử dụng kết quả dự báo một cách hiệu quả trong việc chẩn đoán và điều trị bệnh tim.
* **Đào tạo và nâng cao nhận thức:** Tăng cường đào tạo cho các chuyên gia y tế về cách sử dụng và hiểu rõ các mô hình học máy, từ đó giúp họ áp dụng công nghệ này một cách hiệu quả và an toàn trong thực hành y tế.

### 5.2.5. **Cải tiến khả năng giải thích mô hình:**

* **Phát triển các mô hình dễ giải thích:** Tập trung vào việc phát triển các mô hình dễ giải thích hơn, chẳng hạn như sử dụng các mô hình tuyến tính hoặc cây quyết định, hoặc kết hợp các mô hình phức tạp với các kỹ thuật giải thích như LIME (Local Interpretable Model-agnostic Explanations) hoặc SHAP (SHapley Additive exPlanations) để cung cấp sự minh bạch và giải thích rõ ràng cho các quyết định y tế.

## 5.3 Kết luận chung

Nghiên cứu này đã thể hiện tiềm năng của việc áp dụng các kỹ thuật học máy để dự đoán bệnh tim, cung cấp những kết quả khả quan và có giá trị cho việc cải thiện chẩn đoán và quản lý bệnh. Chúng tôi đã áp dụng và so sánh các mô hình học máy khác nhau, xác định những yếu tố quan trọng ảnh hưởng đến dự đoán, và chứng minh rằng học máy có thể nâng cao hiệu quả dự đoán bệnh tim so với các phương pháp truyền thống.

Tuy nhiên, nghiên cứu cũng gặp phải một số hạn chế, bao gồm quy mô dữ liệu hạn chế và khả năng giải thích mô hình. Việc khắc phục những hạn chế này và tiếp tục nghiên cứu sẽ góp phần nâng cao khả năng dự đoán và ứng dụng của học máy trong y tế. Các nghiên cứu trong tương lai cần tập trung vào việc mở rộng dữ liệu, khám phá các mô hình tiên tiến hơn, và phát triển các công cụ hỗ trợ quyết định để tận dụng tối đa tiềm năng của học máy trong việc cải thiện sức khỏe cộng đồng.

# VI. Danh mục tài liệu tham khảo

1. Phạm Huy Phong (2015), *Trí tuệ nhân tạo - Các thuật toán và ứng dụng*, Nhà xuất bản Khoa học và Kỹ thuật, Hà Nội.
2. Trần Văn Lâm (2020), *Học máy cơ bản và nâng cao*, Nhà xuất bản Bách Khoa, Thành phố Hồ Chí Minh.
3. Bộ Giáo dục và Đào tạo (2018), *Giáo trình học máy*, Nhà xuất bản Đại học Quốc gia, Hà Nội.
4. DES (2022), *Heart Disease Predictions*, Kaggle. Truy cập tại: https://www.kaggle.com/code/desalegngeb/heart-disease-predictions