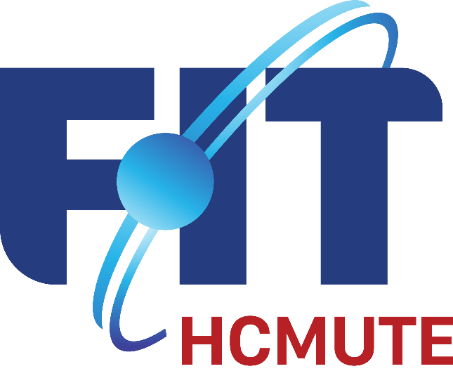
**Trường đại học sư phạm kỹ thuật TP.HCM**

**Khoa Công Nghệ Thông Tin**

**Bộ môn trí tuệ nhân tạo**



**BÁO CÁO CUỐI KÌ**

**xây dựng mô hình dự đoán tỉ lệ mắc bệnh tim thông qua các yếu tố sức khỏe**

môn học: nhập môn khoa học dữ liệu

**Giáo viên hướng dẫn**

**ThS. Lê Minh Tân**

**Nhóm sinh viên thực hiện**

|  |  |  |
| --- | --- | --- |
| STT | Họ và tên | MSSV |
| 1 | Lê Chánh Thành Tín | 21110930 |
| 2 | Lê Hoàng Phúc | 21110601 |
| 3 | Bùi Quang Thiện | 21110656 |

**TP. Hồ Chí Minh, Tháng 5 năm 2024**

lời cảm ơn

Thực tế luôn cho thấy, sự thành công nào cũng đều gắn liền với những sự hỗ trợ, giúp đỡ của những người xung quanh dù cho sự giúp đỡ đó là ít hay nhiều, trực tiếp hay gián tiếp. Trong suốt thời gian từ khi bắt đầu làm đồ án đến nay, nhóm 2 chúng em đã nhận được sự quan tâm, chỉ bảo, giúp đỡ rất nhiều từ thầy. Với tấm lòng biết ơn vô cùng sâu sắc, nhóm chúng em xin gửi lời cảm ơn chân thành đến thầy Lê Minh Tân vì đã dùng những tri thức và tâm huyết của mình để có thể truyền đạt cho chúng em những vốn kiến thức quý báu suốt thời gian học tập. Nhờ có những sự hướng dẫn, dạy bảo tận tình đó, báo cáo này của chúng em đã hoàn thành một cách hoàn chỉnh và đưa ra sản phẩm tốt nhất có thể. Một lần nữa, nhóm chúng em xin gửi lời cảm ơn chân thành nhất đến Thầy.

# MỤC LỤC

[MỤC LỤC 2](#_Toc165324483)

[DANH MỤC HÌNH ẢNH 5](#_Toc165324484)

[DANH MỤC BẢNG 7](#_Toc165324485)

[PHẦN 1: MỞ ĐẦU 8](#_Toc165324486)

[1. Tính cấp thiết của đề tài 8](#_Toc165324487)

[2. Đối tượng, mục tiêu nghiên cứu 8](#_Toc165324488)

[3. Phân tích phương pháp nghiên cứu 8](#_Toc165324489)

[4. Kết quả dự kiến 9](#_Toc165324490)

[PHẦN 2: NỘI DUNG 10](#_Toc165324491)

[CHƯƠNG 1: KHÁI QUÁT VỀ CÁC BỘ DỮ LIỆU CỦA ĐỀ TÀI 10](#_Toc165324492)

[1.1. Thu thập số liệu: P, A, S của 3 dataset 10](#_Toc165324493)

[1.2. Các yếu tố gây ảnh hưởng đến sự hữu ích của dữ liệu 10](#_Toc165324494)

[1.3 . Phương pháp thu thập dữ liệu 11](#_Toc165324495)

[1.4 . Quy tắc thu thập dữ liệu 11](#_Toc165324496)

[CHƯƠNG 2: CHI TIẾT VỀ CÁC BỘ DATASETS 12](#_Toc165324497)

[2.1. Dataset 1: Indicators of Heart Disease (2022 UPDATE) 12](#_Toc165324498)

[a. Khái quát về bộ dữ liệu 12](#_Toc165324499)

[b. Mô tả các thuộc tính 12](#_Toc165324500)

[c. Thống kê tổng quát 13](#_Toc165324501)

[d. Đánh giá chất lượng Dataset 16](#_Toc165324502)

[2.2. Dataset 2: UCI Heart Disease Data 17](#_Toc165324503)

[a. Khái quát về bộ dữ liệu 17](#_Toc165324504)

[b. Mô tả các thuộc tính 17](#_Toc165324505)

[c. Thống kê tổng quát 18](#_Toc165324506)

[d. Đánh giá chất lượng Dataset 20](#_Toc165324507)

[2.3. Dataset 3: Heart Disease Health Indicators Dataset 21](#_Toc165324508)

[a. Khái quát về bộ dữ liệu 21](#_Toc165324509)

[b. Mô tả các thuộc tính 21](#_Toc165324510)

[c. Thống kê tổng quát 22](#_Toc165324511)

[d. Đánh giá chất lượng Dataset 25](#_Toc165324512)

[CHƯƠNG 3: SO SÁNH VÀ KẾT LUẬN 26](#_Toc165324513)

[PHẦN 3: MỞ RỘNG VẤN ĐỀ 27](#_Toc165324514)

[CHƯƠNG 1: PHÂN TÍCH DỮ LIỆU 27](#_Toc165324515)

[1.1. Xử lý dữ liệu 27](#_Toc165324516)

[1.1.1. Chuyển đổi dữ liệu 27](#_Toc165324517)

[1.1.2. Chuẩn hóa dữ liệu 27](#_Toc165324518)

[1.1.3. Phân chia dữ liệu 27](#_Toc165324519)

[1.2. Tìm đặc trưng dữ liệu 27](#_Toc165324520)

[1.3. Phân tích dữ liệu thông qua biểu đồ, đồ thị 28](#_Toc165324521)

[1.3.1. Sự ảnh hưởng của BMI đến bệnh tim 28](#_Toc165324522)

[1.3.2. Số lượng người bị mắc các bệnh và không mắc các bệnh ở từng độ tuổi 30](#_Toc165324523)

[1.3.3. Giả thuyết về tỉ lệ mắc bệnh tim ở người trẻ tuổi 33](#_Toc165324524)

[CHƯƠNG 2: XÂY DỰNG MÔ HÌNH HỌC MÁY DỰ ĐOÁN 34](#_Toc165324525)

[2.1. Mô hình Logistic Regression 34](#_Toc165324526)

[2.1.1. Khái niệm 34](#_Toc165324527)

[2.1.2. Quy trình thực hiện 34](#_Toc165324528)

[2.2. Mô hình mạng nơ ron nhân tạo ANN 34](#_Toc165324529)

[2.2.1. Khái niệm 34](#_Toc165324530)

[2.2.2. Quy trình thực hiện 34](#_Toc165324531)

[2.3. Mô hình gradient boosting 35](#_Toc165324532)

[2.3.1. Khái niệm 35](#_Toc165324533)

[2.3.2. Quy trình thực hiện 36](#_Toc165324534)

[CHƯƠNG 3: ĐÁNH GIÁ MÔ HÌNH VÀ CẢI THIỆN 37](#_Toc165324535)

[3.1. Giải quyết vấn đề mất cân bằng dữ liệu 37](#_Toc165324536)

[3.2. Hướng cải thiện mô hình 39](#_Toc165324537)

[CHƯƠNG 4: KẾT QUẢ MÔ HÌNH 42](#_Toc165324538)

[CHƯƠNG 5: MỞ RỘNG ĐỀ TÀI 43](#_Toc165324539)

[5.1. Phát triển chiến lược phòng ngừa và can thiệp 43](#_Toc165324540)

[5.1.1 Chiến lược 43](#_Toc165324541)

[5.1.2. Chương trình phòng ngừa 43](#_Toc165324542)

[5.2. Đề xuất hướng nghiên cứu trong tương lai 43](#_Toc165324543)

[CHƯƠNG 6: KẾT LUẬN ĐỀ TÀI 45](#_Toc165324544)

[TÀI LIỆU THAM KHẢO 46](#_Toc165324545)

# DANH MỤC HÌNH ẢNH

[Hình 1: Số lượng người mắc bệnh tim 13](#_Toc165324184)

[Hình 2: Biểu đồ số lượng người mắc các loại bệnh 13](#_Toc165324185)

[Hình 3: Biểu đồ tình trạng thể chất là tinh thần của người bệnh 14](#_Toc165324186)

[Hình 4: Biểu đồ số lượng bệnh nhân mắc bệnh tim theo độ tuổi 14](#_Toc165324187)

[Hình 5: Biểu đồ về tỉ lệ mắc bệnh tim theo thời gian ngủ 15](#_Toc165324188)

[Hình 6: Biểu đồ về phân phối BMI của người mắc/ không mắc bệnh tim 15](#_Toc165324189)

[Hình 7: Biểu đồ về tỉ lệ tình trạng bệnh tiểu đường 16](#_Toc165324190)

[Hình 8: Dữ liệu bị thiếu trong mỗi cột của dataset 1 16](#_Toc165324191)

[Hình 9: Biểu đồ về số lượng người mắc bệnh/ không mắc bệnh tim 18](#_Toc165324192)

[Hình 10: Biểu đồ phân phối hàm lượng cholesterol của người mắc/không mắc bệnh tim 18](#_Toc165324193)

[Hình 11: Biểu đồ về số lượng bệnh nhân bị mắc/không mắc bệnh tim theo độ tuổi 19](#_Toc165324194)

[Hình 12: Biểu đồ về sự ảnh hưởng của huyết áp lúc nghỉ đối với bệnh tim 19](#_Toc165324195)

[Hình 13: Biểu đồ về tỷ lệ mắc các loại đau ngực khác nhau 20](#_Toc165324196)

[Hình 14: Dữ liệu bị thiếu trong mỗi cột của dataset 2 20](#_Toc165324197)

[Hình 15: Biểu đồ về Số lượng người mắc/không mắc bệnh tim 22](#_Toc165324198)

[Hình 16: Biểu đồ về tình trạng sức khỏe của bệnh nhân 23](#_Toc165324199)

[Hình 17: Biểu đồ về số lượng ca mắc bệnh tim theo độ tuổi 23](#_Toc165324200)

[Hình 18: Biểu đồ về phân phối BMI của người mắc và không mắc bệnh tim 23](#_Toc165324201)

[Hình 19: Biểu đồ về tỉ lệ mắc bệnh tim theo chế độ ăn 24](#_Toc165324202)

[Hình 20: Biểu đồ về Tỉ lệ mắc bệnh tim khi uống/không uống rượu bia 24](#_Toc165324203)

[Hình 21: Dữ liệu bị thiếu trong mỗi cột của dataset 3 25](#_Toc165324204)

[Hình 22: Biểu đồ tương quan giữa các thuộc tính 27](#_Toc165324205)

[Hình 23: Biểu đồ phân phối của chỉ số BMI theo HeartDisease 28](#_Toc165324206)

[Hình 24: Biểu đồ tỉ lệ mắc bệnh tim theo BMI 29](#_Toc165324207)

[Hình 25: Biểu đồ thể hiện tỉ lệ mắc bệnh tim theo độ tuổi 29](#_Toc165324208)

[Hình 26: Biểu đồ sức khỏe thể chất theo tình trạng bệnh tim 29](#_Toc165324209)

[Hình 27: Biểu đồ sức khỏe tâm thần theo tình trạng bệnh tim 30](#_Toc165324210)

[Hình 28: Biểu đồ thể hiện số lượng người hút thuốc mắc bệnh và không mắc bệnh ở từng độ tuổi 30](#_Toc165324211)

[Hình 29: Biểu đồ thể hiện số lượng người uống rượu mắc bệnh và không mắc bệnh ở từng độ tuổi 31](#_Toc165324212)

[Hình 30: Biểu đồ thể hiện số lượng người đột quỵ mắc bệnh và không mắc bệnh ở từng độ tuổi 31](#_Toc165324213)

[Hình 31: Biểu đồ thể hiện số lượng người khó khăn trọng việc đi lại mắc bệnh và không mắc bệnh ở từng độ tuổi 31](#_Toc165324214)

[Hình 32: Biểu đồ thể hiện số lượng người hoạt động thể chất mắc bệnh và không mắc bệnh ở từng độ tuổi 32](#_Toc165324215)

[Hình 33: Biểu đồ thể hiện số lượng người bị bệnh hen suyễn mắc bệnh và không mắc bệnh ở từng độ tuổi 32](#_Toc165324216)

[Hình 34: Biểu đồ thể hiện số lượng người bị bệnh thận mắc bệnh và không mắc bệnh ở từng độ tuổi 32](#_Toc165324217)

[Hình 35: Biểu đồ thể hiện số lượng người bị ung thư da mắc bệnh và không mắc bệnh ở từng độ tuổi 33](#_Toc165324218)

[Hình 36: Confusion matrix của mô hình gradient boosting ở tập dữ liệu cân bằng với ngưỡng 0.5 39](#_Toc165324219)

[Hình 37: Đường cong ROC và Confusion Matrix của mô hình với ngưỡng tốt nhất 40](#_Toc165324220)

[Hình 38: Biểu đồ thể hiện giữa Metrics và Threshold 40](#_Toc165324221)

[Hình 39: Kết quả của Confusion Matrix trên tập dữ liệu cuối cùng 42](#_Toc165324222)

# DANH MỤC BẢNG

[Bảng 1: Bảng mô tả các thuộc tính của dataset 1 13](#_Toc165618643)

[Bảng 2: Bảng mô tả các thuộc tính của dataset 2 18](#_Toc165618644)

[Bảng 3: Bảng mô tả các thuộc tính của dataset 3 22](#_Toc165618645)

[Bảng 4: Kết quả các chỉ số đánh giá khi sử dụng các mô hình 37](#_Toc165618646)

[Bảng 5: Kết quả các chỉ số đánh giá sau khi sử dụng phương pháp SMOTE 38](#_Toc165618647)

[Bảng 6: Bảng thể hiện các chỉ số metrics với các threshold khác nhau 41](#_Toc165618648)

[Bảng 7: Bảng thể hiện các chỉ số metrics với threshold là 0.39 42](#_Toc165618649)

PHẦN 1: MỞ ĐẦU

# Tính cấp thiết của đề tài

Bệnh tim là nguyên nhân hàng đầu gây tử vong trên toàn cầu, và việc dự đoán sớm nguy cơ mắc bệnh tim có thể giúp ngăn ngừa các biến cố tim mạch nghiêm trọng. Tuy nhiên, mỗi cá nhân có những yếu tố rủi ro khác nhau dựa trên lối sống, lịch sử y tế và các điều kiện sức khỏe khác. Hiện nay, việc phát hiện và phân tích các yếu tố rủi ro chính để dự đoán bệnh tim vẫn là một thách thức. Việc xác định một mô hình dự đoán chính xác sẽ giúp các nhà y tế triển khai các biện pháp can thiệp hiệu quả hơn và cá nhân hóa chiến lược phòng ngừa cho từng cá nhân.

Mô hình dự đoán chính xác và hiệu quả sẽ trở thành một công cụ quan trọng, giúp cá nhân hóa phương pháp điều trị và phòng ngừa bệnh tim, từ đó cải thiện chất lượng cuộc sống và giảm tải gánh nặng cho hệ thống y tế.

# Đối tượng, mục tiêu nghiên cứu

Đối tượng của đề tài: Đề tài sẽ tập trung các đối tượng là những người trưởng thành (từ 18 tuổi trở lên), đặc biệt tập trung vào những người thuộc độ tuổi trung niên và người lớn tuổi.

Mục tiêu nghiên cứu: Mục tiêu chính của bài báo cáo này là xây dựng một mô hình dự đoán tỷ lệ mắc bệnh tim dựa trên các yếu tố sức khỏe được ghi nhận trong bảng dữ liệu khảo sát. Các yếu tố bao gồm chỉ số khối cơ thể (BMI), hút thuốc, uống rượu, tiền sử đột quỵ, tình trạng sức khỏe thể chất và tinh thần, khó khăn trong việc đi lại, giới tính, độ tuổi, chủng tộc, tiền sử mắc bệnh tiểu đường, hoạt động thể chất, tổng quan sức khỏe, thời gian ngủ, hen suyễn, bệnh thận, và ung thư da.

# Phân tích phương pháp nghiên cứu

Phương pháp: Phân tích được thực hiện thông qua việc sử dụng phương pháp hồi quy logistic, Artificial Neural Network, Gradient Boosting so sánh để đưa ra đánh giá mô hình tốt nhất. Tiền xử lý dữ liệu bao gồm việc xử lý các giá trị và mã hóa các biến phân loại. Việc đánh giá mô hình dựa trên các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu, và diện tích ROC (AUC). Dựa vào đường cong ROC chúng ta sẽ thay đổi ngưỡng phù hợp sao cho tôi ưu hóa việc dự đoán người bệnh.

Kết quả: Các mô hình trên đã được phát triển có khả năng dự đoán chính xác nguy cơ mắc bệnh tim dựa trên các yếu tố sức khỏe cá nhân. Qua đó rút ra được mô hình nào sử dụng tối ưu nhất. Việc nhận biết sớm những người có nguy cơ cao sẽ giúp trong việc triển khai các biện pháp phòng ngừa và can thiệp kịp thời, góp phần vào việc giảm tỷ lệ mắc bệnh tim trong cộng đồng.

Đặt câu hỏi nghiên cứu và mục tiêu dự án: Những câu hỏi chúng ta cần làm rõ trong bài báo cáo:

1. Dữ liệu được phân bố như thế nào? Từ dữ liệu có thể rút ra điều gì?
2. Tương quan giữa các nhân tố với bệnh tim?
3. Nhóm rủi ro và nhân tố nguy cơ: Ước lượng nguy cơ mắc bệnh tim theo từng nhóm tuổi: Phân tích mức độ nguy cơ theo độ tuổi?
4. Dựa vào những phát hiện từ dữ liệu, đề xuất các biện pháp can thiệp để cải thiện sức khỏe và giảm thiểu nguy cơ mắc bệnh?
5. Dựa trên dữ liệu, làm thế nào để phát triển chiến lược phòng ngừa và nâng cao ý thức của người bệnh?
6. Có các điểm yếu nào về mô hình nghiên cứu không? Đề xuất hướng nghiên cứu tương lai.

Mục tiêu chung của việc nghiên cứu này là phát triển một mô hình phân tích dữ liệu sức khỏe để hiểu rõ và định lượng các yếu tố nguy cơ liên quan đến bệnh tim, nhằm nhận diện các nhân tố và nhóm dân số có nguy cơ cao. Qua đó, nghiên cứu này sẽ đề xuất các chiến lược can thiệp sức khỏe để cải thiện tình trạng tim mạch.

# Kết quả dự kiến

Các mô hình trên đã được phát triển có khả năng dự đoán chính xác nguy cơ mắc bệnh tim dựa trên các yếu tố sức khỏe cá nhân. Qua đó rút ra được mô hình nào sử dụng tối ưu nhất. Việc nhận biết sớm những người có nguy cơ cao sẽ giúp trong việc triển khai các biện pháp phòng ngừa và can thiệp kịp thời, góp phần vào việc giảm tỷ lệ mắc bệnh tim trong cộng đồng.

# PHẦN 2: NỘI DUNG

# CHƯƠNG 1: KHÁI QUÁT VỀ CÁC BỘ DỮ LIỆU CỦA ĐỀ TÀI

## 1.1. Thu thập số liệu: P, A, S của 3 dataset

1. Indicators of Heart Disease (2022 UPDATE)

P: Toàn bộ người dân trên nước Mỹ.

A: Danh sách đã và đang khám tại khoa tim mạch các bệnh viện trên nước Mỹ.

S: Sử dụng phương pháp SMOTE(Synthetic Minority Over-sampling) để giảm tình trạng mất cân bằng trong tập dữ liệu.

1. UCI Heart Disease Data

P: Người dân tại bốn vùng Hungarian,Cleveland,Switzerland và VA Long Beach.

A: Danh sách các bệnh nhân đến khám tại Viện Tim mạch Hungary,Bệnh viện Đại học Zurich,Bệnh viện Đại học, Basel,Trung tâm Y tế VA, Tổ chức Phòng khám Long Beach và Cleveland.

S: Sử dụng phương pháp SMOTE(Synthetic Minority Over-sampling) để giảm tình trạng mất cân bằng trong tập dữ liệu.

1. Heart Disease Health Indicators Dataset

P: Toàn bộ người dân trên nước Mỹ.

A: Danh sách người dân bên The Behavioral Risk Factor Surveillance System (BRFSS) cần gọi điện.

S: Sử dụng phương pháp SMOTE(Synthetic Minority Over-sampling) để giảm tình trạng mất cân bằng trong tập dữ liệu.

## 1.2. Các yếu tố gây ảnh hưởng đến sự hữu ích của dữ liệu

* Người hướng dẫn khảo sát có thể thiếu kinh nghiệm dẫn đến hướng dẫn sai.
* Người tham gia khảo sát có thể hiểu sai ý câu hỏi hoặc thiếu kiến thức về y tế.
* Độ trung thực của người tham gia khảo sát.
* Tính chính xác và mức độ liên quan của dữ liệu thu thập so với mục đích nghiên cứu.

## . Phương pháp thu thập dữ liệu

* Form điền khảo sát về các yếu tố liên quan đến bệnh tim: Nhằm thu thập các thông tin về lối sống, sinh hoạt, các chỉ số sức khỏe, thói quen, bệnh sử.
* Thiết bị đo huyết áp, nhịp tim: Nhằm thu thập các thông tin về huyết áp, nhịp tim.
* Khảo sát giấy hộ dân: Nhằm thu thập các thông tin về lối sống, sinh hoạt, các chỉ số sức khỏe, thói quen, bệnh sử.
* Thực hiện khảo sát thông qua điện thoại: Thu thập thông tin khảo sát trực tiếp qua điện thoại có thể đảm bảo phần nào tính trung thực của thông tin khảo sát.

## . Quy tắc thu thập dữ liệu

* Hướng dẫn người thực hiện trả lời các câu hỏi, ghi lại câu trả lời, giải thích nhằm đảm bảo thông tin chính xác và theo một cách thống nhất.
* Kiểm soát chất lượng của các thiết bị y tế để đảm bảo ít sai số và không bị ảnh hưởng bởi các yếu tố bên ngoài.
* Sử dụng các dữ liệu, thông tin của người dùng bằng cách phân loại và hệ thống lại dựa trên mức độ của các yếu tố khảo sát.

# CHƯƠNG 2: CHI TIẾT VỀ CÁC BỘ DATASETS

## 2.1. Dataset 1: Indicators of Heart Disease (2022 UPDATE)

Tác giả(s): Centers for Disease Control and Prevention(CDC)

Download link: [dataset 1](https://drive.google.com/uc?export=download&id=1UfF25fr1c-VQzUMi_ZiO5LGhDahM3dfF)

### Khái quát về bộ dữ liệu

Dataset Indicator of Heart Disease (2020) bao gồm khoảng 320.000 dòng với 18 cột. Bộ dữ liệu được lấy từ CDC (Centers for Disease Control and Prevention), thực hiện khảo sát qua điện thoại hàng năm để thu thập dữ liệu về tình trạng sức khỏe của người dân Hoa Kỳ. Đây là phiên bản nhỏ gọn, dễ sử dụng hơn phiên bản gốc. Mỗi dòng bao gồm các thông tin về tình trạng sức khỏe và các bệnh khác.

### Mô tả các thuộc tính

|  |  |
| --- | --- |
| Tên thuộc tính | Mô tả |
| HeartDisisease | Người tham gia từng báo cáo mắc bệnh tim mạch hoặc đau tim (CHD) hoặc trầm cảm cơ tim (MI) |
| BMI | Chỉ số khối cơ thể (BMI) |
| Smoking | Bạn đã hút ít nhất 100 điếu thuốc trong suốt cuộc đời của bạn không? [Lưu ý: 5 gói = 100 điếu thuốc] |
| AlcoholDrinking | Người uống rượu nhiều (đàn ông người lớn uống hơn 14 ly mỗi tuần và phụ nữ người lớn uống hơn 7 ly mỗi tuần) |
| Stroke | Bạn đã bao giờ mắc đột quỵ? |
| PhysicalHealth | Có bao nhiêu ngày trong 30 ngày qua bạn bị chấn thương bao gồm bệnh tật và thương tích thể chất? |
| MentalHealth | Có bao nhiêu ngày trong 30 ngày qua sức khỏe tinh thần của bạn không tốt? |
| DiffWalking | Bạn có khó khăn nghiêm trọng khi đi lại hoặc leo cầu thang không? |
| Sex | Bạn là nam giới hay nữ giới? |
| AgeCategory | Nhóm tuổi bao gồm mười bốn cấp độ |
| Race | Chủng tộc của bạn là gì? |
| Diabetic | Bạn đã mắc bệnh tiểu đường chưa? |
| PhysicalActivity | Bạn có hoạt động thể chất thường xuyên không |
| GenHealth | Bạn có thể nói rằng nói chung sức khỏe của bạn là. |
| SleepTime | Trung bình bạn ngủ bao nhiêu giờ trong một chu kỳ 24 giờ? |
| Asthma | Bạn có được thông báo đã mắc bệnh hen suyễn không  ? |
| KidneyDisease | Bạn có được báo là bạn đã mắc bệnh thận không? |
| SkinCancer | Bạn có được báo là bạn đã mắc bệnh ung thư da không? |

Bảng 1: Bảng mô tả các thuộc tính của dataset 1

### Thống kê tổng quát

*Số lượng người mắc bệnh tim của từng khảo sát*

A graph with red and blue squares

Description automatically generated

Hình 1: Số lượng người mắc bệnh tim

Xem xét số lượng người mắc và không mắc bệnh tim để phục vụ cho mô hình dự đoán, ta có thể thấy số lượng người không mắc bệnh tim chiếm rất cao so với số lượng người mắc bệnh (gấp 10 lần). Do đó dataset có thể gây mất cân bằng cho mô hình dự đoán.

*Tình trạng sức khỏe của người bệnh*

A graph of blue rectangular bars

Description automatically generated with medium confidence

Hình 2: Biểu đồ số lượng người mắc các loại bệnh

Quan sát đồ thị có thể thấy sự phân bố của các cột rất chênh lệch khả năng cao dẫn đến mất cân bằng cho mô hình dự đoán.

*Tình trạng về thể chất và tinh thần của người bệnh*

A graph of a number of different colored bars

Description automatically generated with medium confidence

Hình 3: Biểu đồ tình trạng thể chất là tinh thần của người bệnh

Tình trạng về thể chất và tinh thần bị sai lệch nghiêm trọng do có quá nhiều giá trị 0.

*Số lượng bệnh nhân theo độ tuổi và số ca mắc/không mắc bệnh tim theo tuổi*

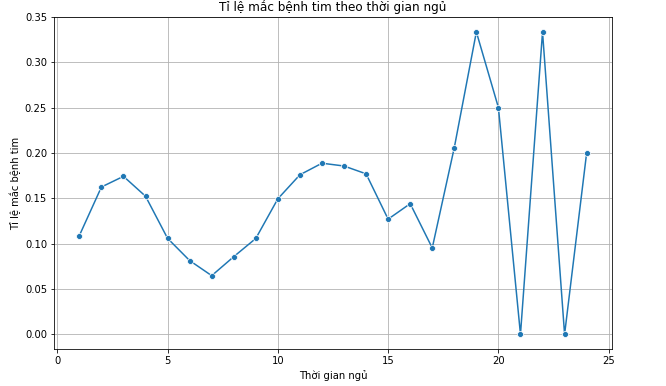
A graph and diagram of a graph

Description automatically generated with medium confidence

Hình 4: Biểu đồ số lượng bệnh nhân mắc bệnh tim theo độ tuổi

Biểu đồ về phân bố theo độ tuổi và bệnh tim khá đồng đều, biểu đồ cho thấy số ca mắc bệnh tim tỉ lệ thuận với độ tuổi của người bệnh, ngược lại ở các độ tuổi thấp số bệnh nhân mắc bệnh tim chiếm tỉ lệ rất thấp.

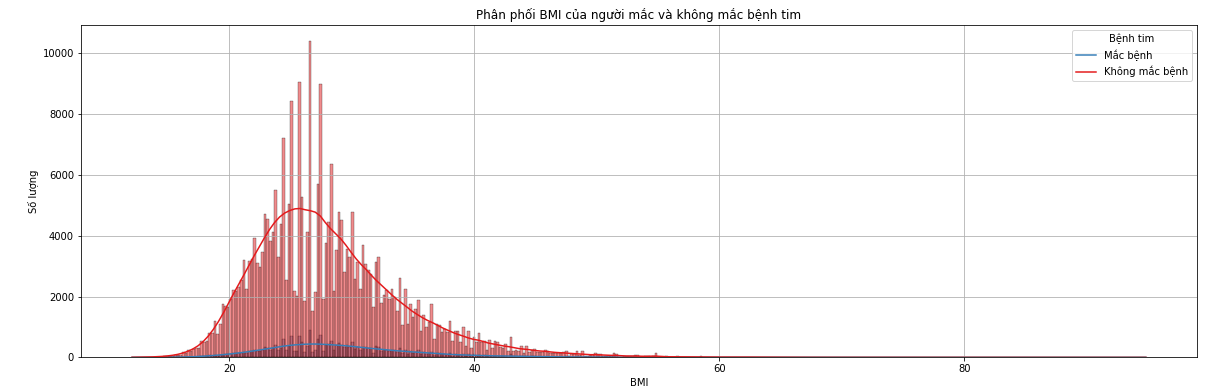
*Tỉ lệ mắc bệnh tim theo thời gian ngủ*



Hình 5: Biểu đồ về tỉ lệ mắc bệnh tim theo thời gian ngủ

Quan sát biểu đồ có thể thấy thông tin về thời gian ngủ của khảo sát bị sai lệch rất nhiều, làm cho biểu đồ không thể hiện được thông tin quan trọng, từ sau mốc 15 giờ, tỉ lệ bị thay đổi liên tục, thời gian ngủ cũng không hợp lí.

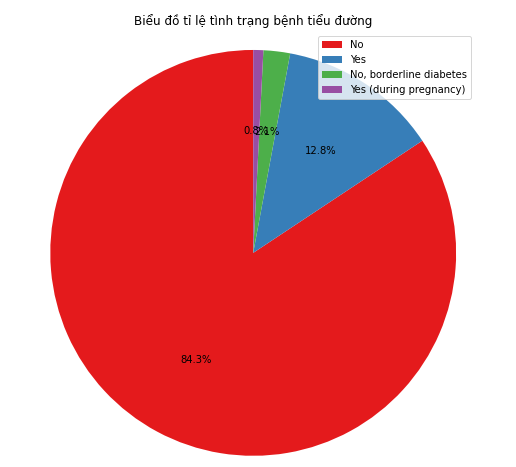
*Phân phối BMI của người mắc/không mắc bệnh tim*



Hình 6: Biểu đồ về phân phối BMI của người mắc/ không mắc bệnh tim

Tương tự như thời gian ngủ, thông tin về BMI cũng có nhiều thông tin sai lệch, gây nhiễu, phân bố không đều dễ gây tình trạng mất cân bằng.

*Tỉ lệ tình trạng bệnh tiểu đường*

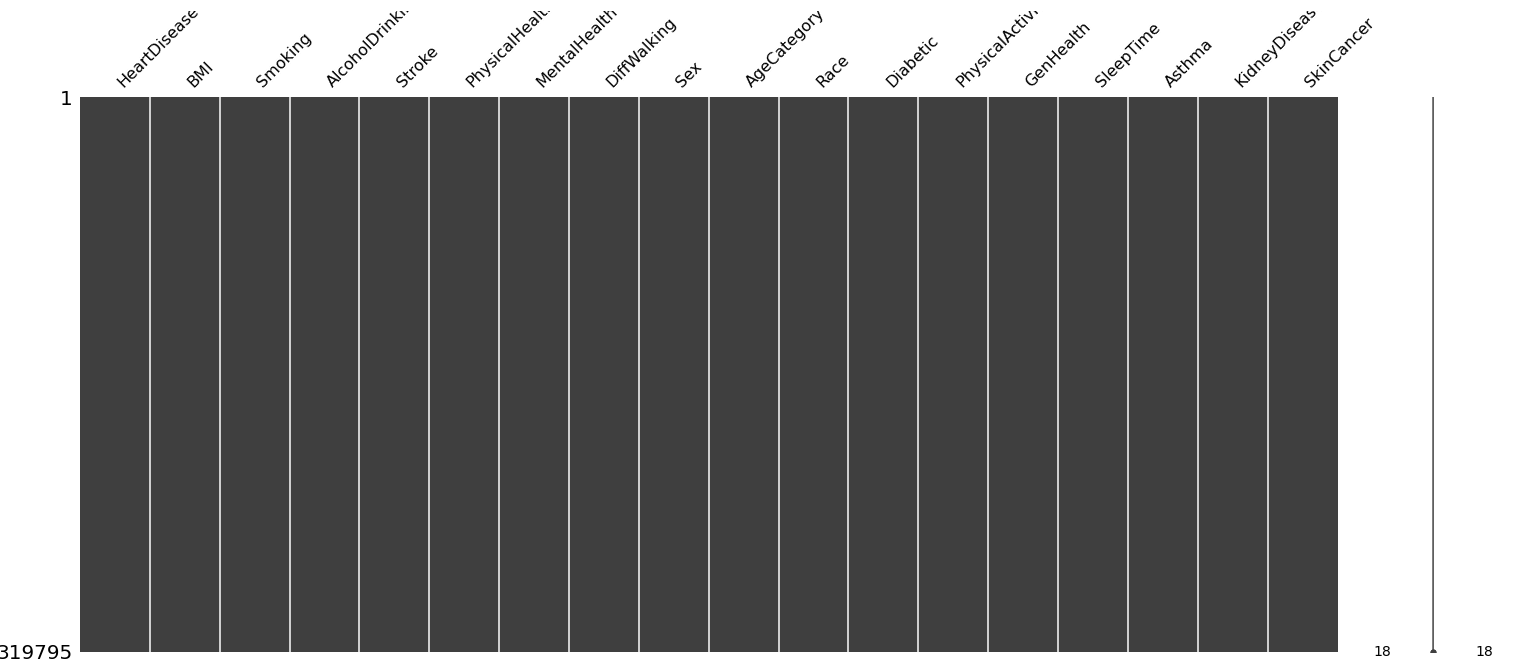


Hình 7: Biểu đồ về tỉ lệ tình trạng bệnh tiểu đường

Biểu đồ tỉ lệ tình trạng bệnh tiểu đường phân bố không đồng đều dễ gây mất cân bằng trong mô hình dự đoán.

### Đánh giá chất lượng Dataset

Thống kê số lượng giá trị null của từng dataset



Hình 8: Dữ liệu bị thiếu trong mỗi cột của dataset 1

Dataset đảm bảo đầy đủ dữ liệu. Không cần các thao tác xử lí việc thiếu dữ liệu.

Tuy nhiên qua thống kê về dữ liệu ta có thể nhận ra các điểm sau đây:

* + Phần lớn các cột về tình trạng sức khỏe là phân loại.
  + Dữ liệu về BMI bị sai lệch
  + SleepTime bị sai lệch
  + Thống kê về các tình trạng sức khỏe cho thấy dữ liệu mất cân bằng.
  + AgeCategory được phân bố đều.
  + PhysicalHealth và MentalHealth bị sai lệch nghiêm trọng
  + GenHealth không có vấn đề nghiêm trọng
  + Diabetic mất cân bằng.

## 2.2. Dataset 2: UCI Heart Disease Data

Các tác giả:

* Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
* University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
* University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
* V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Download link:  [dataset 2](https://drive.google.com/uc?export=download&id=1b1Qj-dBPOmefDWMoYvADfNo2ZZamRSO-)

### Khái quát về bộ dữ liệu

UCI Heart Disease Data bao gồm 920 dòng với 16 cột. Lấy từ dữ liệu của bệnh nhân từ các bệnh viện ở bốn vùng Hungary, Cleveland, Switzerland và VA Long Beach. Mỗi dòng bao gồm các chỉ số về huyết áp, hàm lượng cholesterol, các loại đau thắt ngực,… của bệnh nhân.

### Mô tả các thuộc tính

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Mô tả** |
| id | Chỉ số id của mỗi bệnh nhân |
| age | Tuổi của bệnh nhân |
| sex | Giới tính của bệnh nhân |
| dataset | Nơi thực hiện khảo sát bệnh nhân |
| cp(chest pain type) | Loại đau thắt ngực, gồm có 4 loại typical angina(đau thắt ngực điển hình), atypical angina( đau thắt ngực không điển hình), non-anginal(không đau thắc ngực), asymptomatic(không có triệu chứng). |
| trestbps(resting blood pressure) | Huyết áp lúc nghỉ, tính bằng mm Hg |
| chol | Cholesterol huyết thanh tính bằng mg/dl |
| fbs | Đường huyết lúc đói của bệnh nhân có > 120 mg/dl hay không? (TRUE là có, FALSE là không) |
| restecg | Điện tâm đồ lúc nghỉ, gồm có 3 lựa chọn normal(bình thường), st t abnormality(đoạn ST bất thường), lv hypertrophy(phì đại thất trái) |
| thalach | Nhịp tim tối đa bệnh nhân đạt được |
| exang | Bệnh nhân có bị đau thắt ngực do tập thể dục không? (TRUE là có, FALSE là không) |
| oldpeak | Độ chênh xuống của đoạn ST khi luyện tập so với khi nghỉ ngơi |
| slope | Độ dốc của đoạn ST lúc tập luyện |
| ca | Số lượng mạch máu tiêm cản quang |
| thal | Tình trạng bệnh Thalassemia(bệnh thiếu máu tán huyết di truyền hay thiếu máu tán huyết bẩm sinh) của bệnh nhân. |
| num | Cho thấy sự hiện diện bệnh tim của bệnh nhân. 0: Không có bệnh tim. 1: Nhẹ hoặc có biểu hiện bệnh tim. 2: Có bệnh tim ở mức độ vừa phải. 3: Có bệnh tim nặng. 4: Có bệnh tim rất nặng. |

Bảng 2: Bảng mô tả các thuộc tính của dataset 2

### Thống kê tổng quát

*Số lượng người mắc bệnh tim*

A red square and black square

Description automatically generated

Hình 9: Biểu đồ về số lượng người mắc bệnh/ không mắc bệnh tim

Xem xét số lượng người mắc và không mắc bệnh tim để phục vụ cho mô hình dự đoán, ta có thể thấy số lượng người không mắc bệnh và mắc bệnh tim là ngang nhau. Điều này sẽ không gây khó khăn trong việc phân tích.

*Phân phối hàm lượng cholesterol của người mắc/không mắc bệnh tim*

A graph with red and blue lines

Description automatically generated

Hình 10: Biểu đồ phân phối hàm lượng cholesterol của người mắc/không mắc bệnh tim

Nhìn vào biểu đồ trên, ta có thể thấy khi hàm lượng cholesterol càng cao thì tỉ lệ mắc bệnh tim càng tăng.

Nhìn chung dữ liệu khá ổn để phân tích, nhưng có một vấn đề trong tập dữ liệu cholesterol xuất hiện nhiều giá trị là 0. Điều này là vô lí trong y học. Điều này cũng gây phần nào khó khăn trong việc phân tích. Giải pháp ở đây sẽ là thay thế các giá trị bằng 0 thành giá trị NULL.

*Số lượng bệnh nhân bị mắc/không mắc bệnh tim theo độ tuổi*

A graph of different colored bars

Description automatically generated

Hình 11: Biểu đồ về số lượng bệnh nhân bị mắc/không mắc bệnh tim theo độ tuổi

Trong số các bệnh nhân bị mắc bệnh tim, bệnh nhân có độ tuổi từ 55-70 là nhiều nhất. Trong khi đó các bệnh nhân có độ tuổi từ 30-40 là độ tuổi có ít bệnh nhân mắc bệnh tim. Độ tuổi khoảng từ 30 trở xuống không có ai bị mắc bệnh tim.

Nhìn vào biểu đồ trên ta thấy khi tuổi càng cao sẽ càng có nguy cơ mắc bệnh tim.

*Sự ảnh hưởng của huyết áp lúc nghỉ đối với bệnh tim*

A graph with different colored bars

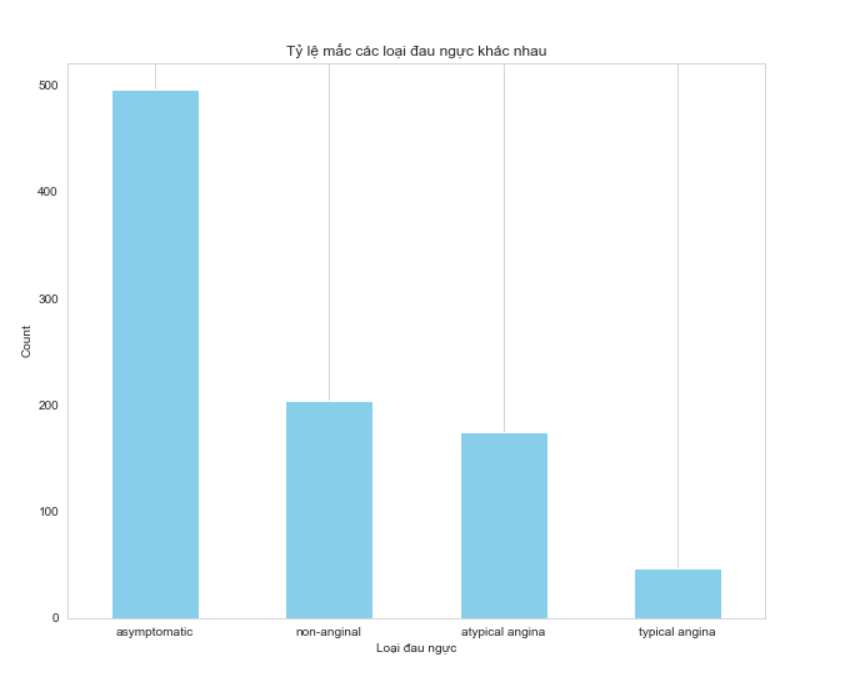
Description automatically generated

Hình 12: Biểu đồ về sự ảnh hưởng của huyết áp lúc nghỉ đối với bệnh tim

Trong biểu đồ trên, ta thấy khi huyết áp lúc nghỉ càng cao thì tỉ lệ mắc bệnh tim càng nhiều.

Các bệnh nhân có huyết áp cận 200 đều bị mắc bệnh tim. Trong khi đó các bệnh nhân có huyết áp thấp có tỉ lệ mắc bệnh tim thấp hơn.

*Tỉ lệ mắc các loại đau ngực khác nhau của bệnh nhân*



Hình 13: Biểu đồ về tỷ lệ mắc các loại đau ngực khác nhau

Nhìn vào biểu đồ trên ta thấy đa số các bệnh nhân không có triệu chứng(asymptomatic) về đau thắt ngực. Số bệnh nhân bị đau thắt ngực điển hình(typical angina) là thấp nhất.

### Đánh giá chất lượng Dataset

A black and white bar chart

Description automatically generated

Hình 14: Dữ liệu bị thiếu trong mỗi cột của dataset 2

Tập dữ liệu trên có sự thiếu sót đáng kể, đặc biệt là trong cột min\_major\_vessels và thalassemia.

Các giải pháp để xử lý dữ liệu bị thiếu như sau:

* + Điền các giá trị cho các cột resting\_blood\_pressure, thalch, st\_depression và serum\_cholesterol với median(giá trị trung bình) của mỗi cột.
  + Điền các giá trị cho các cột fasting\_blood\_sugar và resting\_ecg\_results bằng giá trị mode của mỗi cột.
  + Loại bỏ những cột có nhiều NULL hoặc không có giá trị trong phân tích.

Ngoài ra qua thống kê về dữ liệu ta có thể nhận ra các điểm sau đây:

* + Đa số các cột dữ liệu có phân bố đều, ngoại trừ các cột có nhiều giá trị bị thiếu.
  + Nhiều cột dữ liệu có nhiều giá trị bị thiếu, gây phần nào cản trở cho việc phân tích.
  + Dữ liệu này rất tốt bởi vì đa phần là dữ liệu con số. Điều này sẽ dễ dàng trong việc phân tích và làm ra mô hình chuẩn đoán bệnh tim.

## 2.3. Dataset 3: Heart Disease Health Indicators Dataset

Tác giả: Centers for Disease Control and Prevention(CDC)

Download link: [dataset 3](https://drive.google.com/uc?export=download&id=12lWPK8nlplpEUvWokzoYbcNI14lj_D8v)

### Khái quát về bộ dữ liệu

Dataset Indicator of Heart Disease bao gồm khoảng 320000 dòng với 22 cột. Bộ dữ liệu được lấy từ CDC (Centers for Disease Control and Prevention). Được thu thập từ bộ khảo sát BRFSS 2015. Bộ dữ liệu được dùng với mục đích nhấn mạnh tầm quan trọng của các biện pháp phòng ngừa và xét nghiệm, dự đoán các căn bệnh tim trong cộng đồng. Bộ dữ liệu này đã được làm sạch và thu gọn hơn. Các thuộc tính bao gồm tình trạng sức khỏe, thói quen sống và các thông tin khác về đời sống của bệnh nhân.

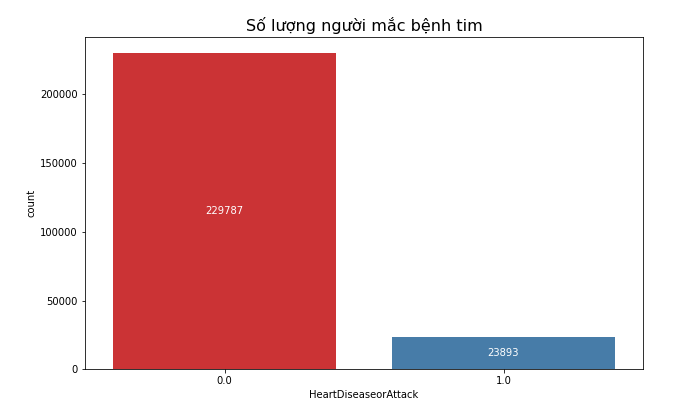
### Mô tả các thuộc tính

|  |  |
| --- | --- |
| Tên thuộc tính | Mô tả |
| HeartDiseaseorAttack | Bạn có từng bị bệnh tim mạch hoặc đau tim không? |
| HighBP | Bạn có huyết áp cao không? |
| HighChol | Bạn có mỡ máu cao không? |
| CholCheck | Bạn đã kiểm tra mỡ máu trong thời gian gần đây không? |
| BMI | Chỉ số BMI của bạn là bao nhiêu? |
| Smoker | Bạn có từng hút thuốc không? |
| Stroke | Bạn có từng bị đột quỵ không? |
| Diabetes | Bạn có bị bệnh tiểu đường không? |
| PhysActivity | Bạn có hoạt động thể chất không? |
| Fruits | Bạn ăn trái cây hàng ngày không? |
| Veggies | Bạn ăn rau cải hàng ngày không? |
| HvyAlcoholConsump | Bạn uống rượu nhiều hơn 14 ly mỗi ngày  không? |
| AnyHealthcare | Bạn có bảo hiểm y tế không? |
| NoDocbcCost | Bạn đã kiểm tra sức khỏe gần đây không? |
| GenHlth | Đánh giá sức khỏe tổng quát của mình như thế nào? (từ 1 đến 5) |
| MentHlth | Có bao nhiêu ngày trong 30 ngày qua sức khỏe tinh thần của bạn không tốt? |
| PhysHlth | Có bao nhiêu ngày trong 30 ngày qua bạn bị chấn thương bao gồm bệnh tật và thương tích thể chất? |
| DiffWalk | Bạn có gặp khó khăn khi đi lại không? |
| Sex | Giới tính của bạn là gì?(1 là nam/ 0 là nữ) |
| Age | Bạn thuộc nhóm tuổi nào? từ 1 (từ 18 đến 24 tuổi) đến 13 (80 tuổi trở lên), mỗi nhóm có khoảng cách 5 tuổi. |
| Education | Cho biết mức độ học vấn cao nhất mà người đó đã hoàn thành, từ 0 (chưa bao giờ đi học hoặc chỉ học mẫu giáo) đến 6 (đã hoàn thành 4 năm đại học hoặc nhiều hơn). |
| Income | Thu nhập của bạn là bao nhiêu trên 1 tháng ? |

Bảng 3: Bảng mô tả các thuộc tính của dataset 3

### Thống kê tổng quát

*Số lượng người mắc/không mắc bệnh tim*



Hình 15: Biểu đồ về Số lượng người mắc/không mắc bệnh tim

Quan sát biểu đồ cho thấy tỉ lệ giữa người mắc và không mắc bệnh tim chênh lệch rất nhiều khi đem vào mô hình dự đoán, dễ gây mất cân bằng.

*Tình trạng sức khỏe của bệnh nhân*

A graph with blue columns

Description automatically generated

Hình 16: Biểu đồ về tình trạng sức khỏe của bệnh nhân

Quan sát đồ thị có thể thấy sự phân bố của các cột rất chênh lệch khả năng cao dẫn đến mất cân bằng cho mô hình dự đoán.

*Số lượng bệnh nhân theo độ tuổi và số ca mắc/không mắc bệnh tim theo nhóm tuổi*

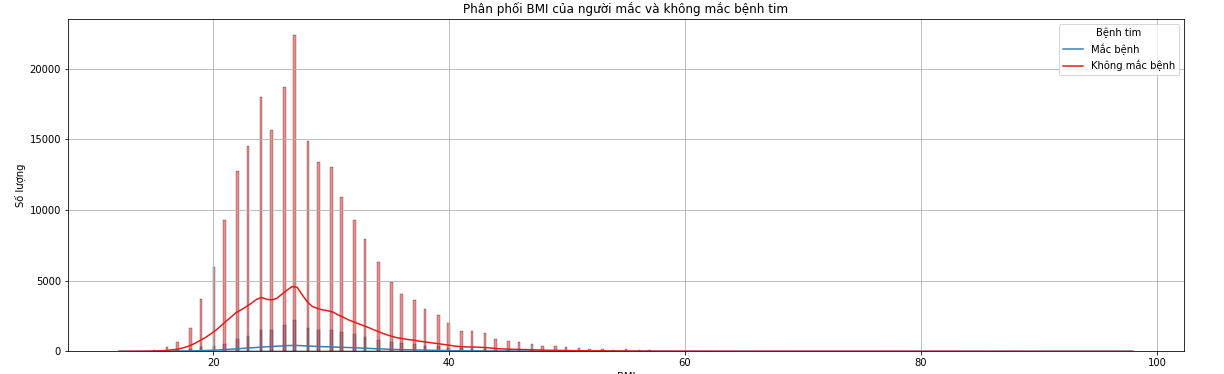
A graph and diagram of a graph

Description automatically generated

Hình 17: Biểu đồ về số lượng ca mắc bệnh tim theo độ tuổi

Biểu đồ về phân bố theo độ tuổi và bệnh tim khá đồng đều, biểu đồ cho thấy số ca mắc bệnh tim tỉ lệ thuận với độ tuổi của người bệnh, ngược lại ở các độ tuổi thấp số bệnh nhân mắc bệnh tim chiếm tỉ lệ rất thấp.

*Phân phối BMI của người mắc và không mắc bệnh tim*



Hình 18: Biểu đồ về phân phối BMI của người mắc và không mắc bệnh tim

Quan sát cho thấy biểu đồ phân phối về BMI ở mức tạm, tuy nhiên vẫn có vài dữ liệu sai lệch, cần xử lý để giảm tình trạng mất cân bằng

*Tỉ lệ mắc bệnh tim theo chế độ ăn*

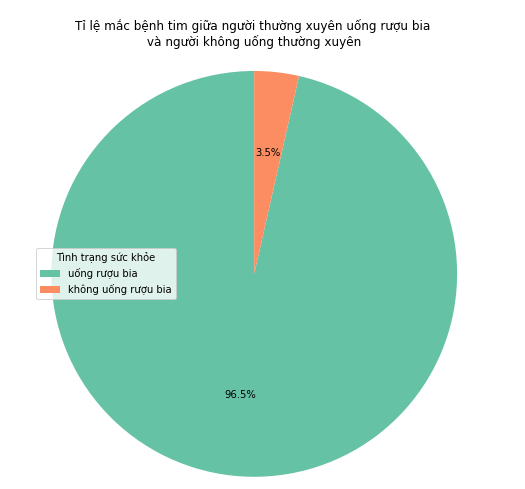
A close-up of a pie chart

Description automatically generated

Hình 19: Biểu đồ về tỉ lệ mắc bệnh tim theo chế độ ăn

Quan sát biểu đồ, ta có thể thấy, những người bệnh nhân không ăn rau củ, trái cây chiếm phần lớn tỉ lệ mắc bệnh tim. Cần bố trí lại dữ liệu để cân bằng hơn.

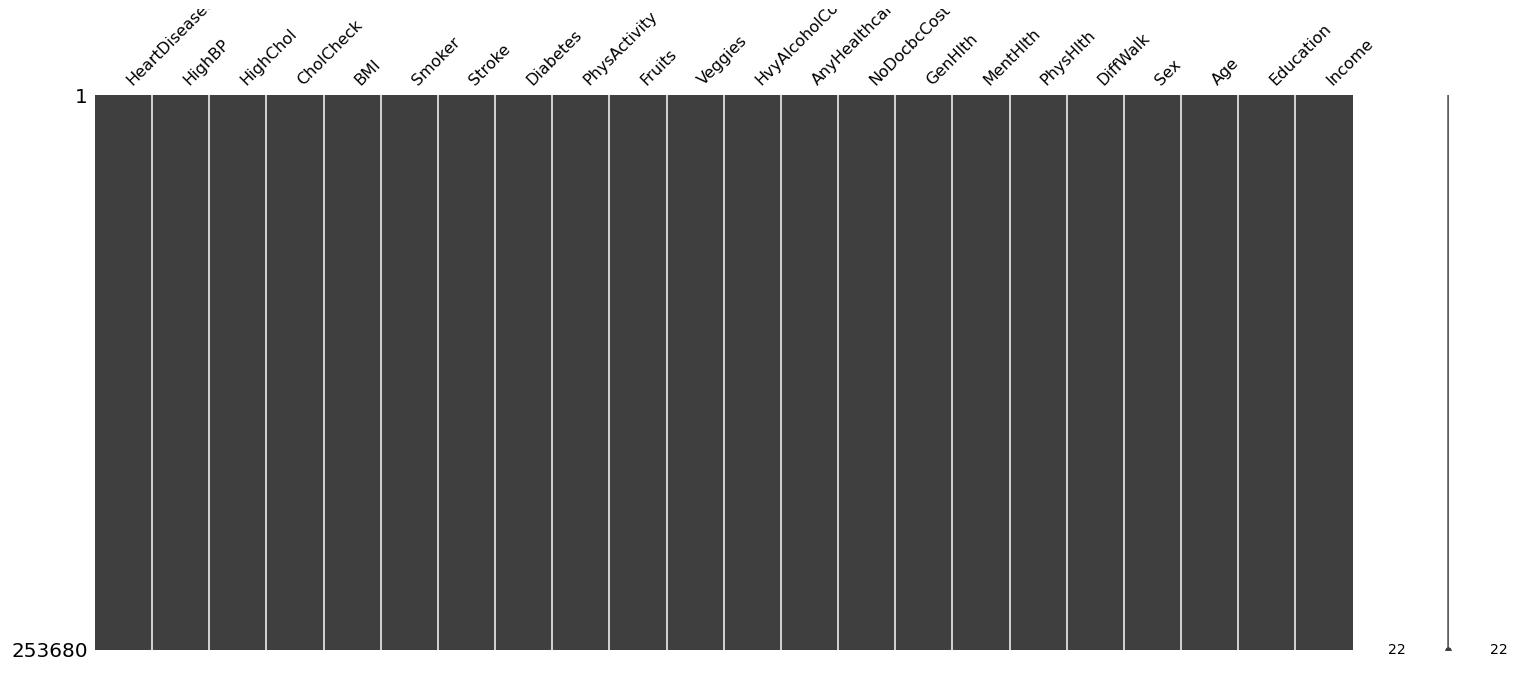
*Tỉ lệ mắc bệnh tim khi uống/không uống rượu bia*



Hình 20: Biểu đồ về Tỉ lệ mắc bệnh tim khi uống/không uống rượu bia

Biểu đồ cho biết, uống rượu bia có thể làm gia tăng nguy cơ mắc bệnh tim khi tỉ lệ người uống rượu bia mắc bệnh chiếm rất lớn (96.5%).

### Đánh giá chất lượng Dataset



Hình 21: Dữ liệu bị thiếu trong mỗi cột của dataset 3

Dataset đảm bảo đầy đủ dữ liệu. Không cần các thao tác xử lí việc thiếu dữ liệu.

Tuy nhiên qua thống kê về dữ liệu ta có thể nhận ra các điểm sau đây:

* + Phần lớn các cột về tình trạng sức khỏe là phân loại.
  + Dữ liệu về BMI bị sai lệch
  + Thống kê về các tình trạng sức khỏe cho thấy dữ liệu mất cân bằng.
  + AgeCategory được phân bố đều.
  + GenHealth không có vấn đề nghiêm trọng

# CHƯƠNG 3: SO SÁNH VÀ KẾT LUẬN

1. Indicators of Heart Disease (2022 UPDATE):

   - Dữ liệu này bao gồm thông tin về các chỉ số như BMI, hút thuốc, uống rượu, sức khỏe vật lý và sức khỏe tinh thần, giới tính, nhóm tuổi, chủng tộc, tiểu đường, hoạt động thể chất và nhiều yếu tố khác liên quan đến bệnh tim mạch.

   - Dữ liệu có tính đầy đủ, không chứa giá trị null, cung cấp thông tin về 20 thuộc tính cho mỗi mẫu.

2. UCI Heart Disease Data:

 - Dataset này chứa thông tin từ nhiều trường hợp bệnh nhân được đánh giá từ các bệnh viện khác nhau, bao gồm các yếu tố như tuổi, giới tính, các biểu hiện của bệnh tim mạch và các chỉ số y tế khác.

   - Dữ liệu trong dataset này cũng không đầy đủ, có các giá trị null, nhưng nó cung cấp một phần về thông tin về các yếu tố y tế cụ thể liên quan đến bệnh tim mạch.

3. HeartDiseaseorAttack Dataset:

   - Dataset này chứa thông tin về các yếu tố nguy cơ và biểu hiện của bệnh tim mạch, bao gồm các yếu tố như cao huyết áp, cholesterol cao, hút thuốc, tiểu đường, hoạt động thể chất, giới tính, nhóm tuổi và nhiều yếu tố khác.

   - Dữ liệu này không chứa giá trị null, nhưng nó tập trung vào các chỉ số sức khỏe và các biểu hiện của bệnh tim mạch từ một góc độ khác.

Cả ba dataset đều có phạm vi khảo sát lớn và cung cấp thông tin đa dạng về bệnh tim mạch. Tuy nhiên, mỗi dataset có phạm vi và góc nhìn khác nhau, từ việc thu thập dữ liệu từ các bệnh nhân đến việc khảo sát người dân chung.

Dataset 1 và dataset 3 đều đảm bảo rằng dữ liệu không chứa giá trị null, trong khi dataset 2 lại chứa nhiều giá trị null. Điều này có thể ảnh hưởng đến tính chính xác của các phân tích dữ liệu.

Cả ba dataset đều đo lường các yếu tố quan trọng liên quan đến bệnh tim mạch như hút thuốc, uống rượu, cao huyết áp, cholesterol, tiểu đường, và hoạt động thể chất. Tuy nhiên, mỗi dataset có thể chứa các biến số khác nhau hoặc đo lường các yếu tố khác nhau, tùy thuộc vào mục tiêu và phạm vi của nghiên cứu.

Tóm lại, dataset 1 là một lựa chọn lý tưởng cho phân tích vì nó cung cấp số lượng dữ liệu lớn với đa dạng thuộc tính, đồng thời không chứa giá trị null. Các thuộc tính có thể dễ dàng thu thập từ bệnh nhân mà không qua bất kỳ phương pháp xét nghiệm nào.

# PHẦN 3: MỞ RỘNG VẤN ĐỀ

# CHƯƠNG 1: PHÂN TÍCH DỮ LIỆU

## Xử lý dữ liệu

### Chuyển đổi dữ liệu

Mã hóa các biến phân loại thành số để có thể xử lý bằng các mô hình hồi quy. Ví dụ, biến "Giới tính" với các giá trị "Nam" và "Nữ" có thể được chuyển đổi thành 0 và 1.

### Chuẩn hóa dữ liệu

Chuẩn hóa các biến liên tục như BMI và tuổi để giảm thiểu ảnh hưởng của thang đo.

### Phân chia dữ liệu

Chia dữ liệu thành hai tập: tập huấn luyện (cho 80% dữ liệu) và tập kiểm thử (20%).

## Tìm đặc trưng dữ liệu

A red and black squares

Description automatically generated

Hình 22: Biểu đồ tương quan giữa các thuộc tính

Nhóm quy ước thang đo của độ tương quan: thấp: 0 → 0.4, trung bình: 0.4 → 0.8, cao: 0.8 → 1. Nhìn vào biểu đồ ta thấy đa số các thuộc tính có độ tương quan ở mức thấp và trung bình, do đa số các thuộc tính có tính chất phân loại, như trong trường hợp các thuộc tính có giá trị là “Yes” hoặc “No”. Điều này là hợp lý vì các thuộc tính phân loại thường không thể hiện mối quan hệ tuyến tính mạnh mẽ với nhau.

## Phân tích dữ liệu thông qua biểu đồ, đồ thị

### Sự ảnh hưởng của BMI đến bệnh tim

A graph of a number of patients

Description automatically generated

Hình 23: Biểu đồ phân phối của chỉ số BMI theo HeartDisease

Số lượng người có bệnh tim mạch cao hơn số lượng người không có bệnh tim mạch ở hầu hết các mức BMI. Điều này cho thấy rằng có mối liên hệ giữa chỉ số BMI và nguy cơ mắc bệnh tim mạch. Cụ thể, khi chỉ số BMI tăng, nguy cơ mắc bệnh tim mạch cũng tăng.

Có một sự khác biệt đáng kể về số lượng người có bệnh tim mạch và không có bệnh tim mạch ở các mức BMI cao. Điều này cho thấy rằng nguy cơ mắc bệnh tim mạch tăng cao hơn nhiều ở những người béo phì (BMI ≥ 30) so với những người thừa cân (25 ≤ BMI < 30) hoặc bình thường (18.5 ≤ BMI < 25).

A blue and red bars with white text

Description automatically generated

Hình 24: Biểu đồ tỉ lệ mắc bệnh tim theo BMI

* Tỷ lệ mắc bệnh tim mạch tăng dần theo mức độ BMI.
* Nhóm người béo phì và béo phì nặng có tỷ lệ mắc bệnh tim mạch cao nhất.
* Nhóm người có mức độ BMI thấp có tỷ lệ mắc bệnh tim mạch thấp nhất.
* Có sự khác biệt rõ ràng về tỷ lệ mắc bệnh tim mạch giữa các nhóm BMI khác nhau.

A graph and diagram of a graph

Description automatically generated with medium confidence

Hình 25: Biểu đồ thể hiện tỉ lệ mắc bệnh tim theo độ tuổi

* Tỷ lệ mắc bệnh tim mạch tăng dần theo độ tuổi.
* Nhóm người ở độ tuổi 65-69 và 70-74 có tỷ lệ mắc bệnh tim mạch cao nhất.
* Nhóm người ở độ tuổi 18-24 có tỷ lệ mắc bệnh tim mạch thấp nhất.
* Có sự khác biệt rõ ràng về tỷ lệ mắc bệnh tim mạch giữa các nhóm tuổi khác nhau.

A graph and diagram of a graph

Description automatically generated

Hình 26: Biểu đồ sức khỏe thể chất theo tình trạng bệnh tim

Nhìn chung, biểu đồ cho thấy một mối quan hệ tích cực giữa tỷ lệ bệnh tim và số ngày có vấn đề về sức khỏe thể chất. Điều này có nghĩa là khi số ngày có vấn đề về sức khỏe thể chất tăng lên, tỷ lệ bệnh tim cũng tăng lên.

A graph and a diagram

Description automatically generated with medium confidence

Hình 27: Biểu đồ sức khỏe tâm thần theo tình trạng bệnh tim

* Tỷ lệ bệnh tim có xu hướng tăng khi số ngày có vấn đề về sức khỏe tâm thần tăng.
* Mối quan hệ này không hoàn toàn tuyến tính, nghĩa là tỷ lệ bệnh tim không tăng đều theo số ngày có vấn đề về sức khỏe tâm thần.
* Có thể thấy một số điểm bùng nổ trong dữ liệu, cho thấy có sự thay đổi đột ngột trong tỷ lệ bệnh tim trong một số khoảng thời gian nhất định.

### Số lượng người bị mắc các bệnh và không mắc các bệnh ở từng độ tuổi

*Số lượng người hút thuốc mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of different sizes and colors

Description automatically generated with medium confidence

Hình 28: Biểu đồ thể hiện số lượng người hút thuốc mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người uống rượu mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of different colored bars

Description automatically generated with medium confidence

Hình 29: Biểu đồ thể hiện số lượng người uống rượu mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người đột quỵ mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of different sizes and colors

Description automatically generated with medium confidence

Hình 30: Biểu đồ thể hiện số lượng người đột quỵ mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người khó khăn trong việc đi lại mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of blue and red bars

Description automatically generated

Hình 31: Biểu đồ thể hiện số lượng người khó khăn trọng việc đi lại mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người hoạt động thể chất mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of a number of people

Description automatically generated with medium confidence

Hình 32: Biểu đồ thể hiện số lượng người hoạt động thể chất mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người bị bệnh hen suyễn mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of different colored bars

Description automatically generated with medium confidence

Hình 33: Biểu đồ thể hiện số lượng người bị bệnh hen suyễn mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người bị bệnh thận mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of different colored bars

Description automatically generated with medium confidence

Hình 34: Biểu đồ thể hiện số lượng người bị bệnh thận mắc bệnh và không mắc bệnh ở từng độ tuổi

*Số lượng người ung thư da mắc bệnh và không mắc bệnh ở từng độ tuổi*

A graph of different colored bars

Description automatically generated with medium confidence

Hình 35: Biểu đồ thể hiện số lượng người bị ung thư da mắc bệnh và không mắc bệnh ở từng độ tuổi

Qua các hình ảnh trên ta thấy được tỷ lệ người mắc bệnh ở các độ tuổi càng cao ứng với các trường khác nhau thì số lượng mắc bệnh cũng cao dần. => **Tuổi ảnh hưởng rất lớn đến việc mắc bệnh tim**

### Giả thuyết về tỉ lệ mắc bệnh tim ở người trẻ tuổi

Giả thuyết H0 trong nghiên cứu này giả định rằng không có sự khác biệt về tỉ lệ mắc bệnh tim giữa nhóm người trẻ tuổi có bệnh và không có bệnh. Ngược lại, giả thuyết H1 đặt ra là có sự khác biệt đáng kể về tỉ lệ này giữa hai nhóm. Nghiên cứu nhằm kiểm định hai giả thuyết này thông qua việc phân tích dữ liệu sức khỏe, qua đó giúp hiểu rõ hơn về mối quan hệ giữa các yếu tố sức khỏe và nguy cơ mắc bệnh tim ở người trẻ. Điều này được thực hiện bằng cách tính toán giá trị p-value, là xác suất của việc quan sát thấy kết quả hiện tại (hoặc cực đoan hơn) nếu giả thuyết null đúng. Nếu p-value nhỏ (thường dưới ngưỡng 0.05), ta có cơ sở để bác bỏ giả thuyết null, ngụ ý rằng có những bằng chứng thống kê đáng kể chống lại H0​ và ủng hộ giả thuyết đối (H1), biểu thị sự khác biệt hoặc tác động.

*Kết luận: “Không có sự khác biệt về tỉ lệ mắc bệnh tim giữa nhóm có bệnh và nhóm không có bệnh ở người trẻ tuổi là đúng “. Kết quả này chỉ ra rằng, trong bối cảnh của nghiên cứu này, không có bằng chứng đủ để cho thấy sức khỏe tổng thể và các yếu tố rủi ro khác ảnh hưởng đáng kể đến tỉ lệ mắc bệnh tim trong nhóm người trẻ tuổi.*

Giả thuyết này đã mang lại một cái nhìn mới về dữ liệu, từ đó có thể mở rộng phạm vi nghiên cứu bằng cách xem xét thêm các nguy cơ như: yếu tố di truyền, môi trường,…

# CHƯƠNG 2: XÂY DỰNG MÔ HÌNH HỌC MÁY DỰ ĐOÁN

## 2.1. Mô hình Logistic Regression

### 2.1.1. Khái niệm

Logistic Regression là 1 thuật toán phân loại được dùng để gán các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, ...). Một ví dụ điển hình là phân loại Email, gồm có email công việc, email gia đình, email spam, ... Giao dịch trực tuyến có là an toàn hay không an toàn, khối u lành tính hay ác tình.

### 2.1.2. Quy trình thực hiện

Xây dựng lớp mô hình Logistic Regression: lớp ModelLogisticRegression trong Python để quản lý việc huấn luyện và đánh giá mô hình Logistic Regression.

Cấu trúc của lớp ModelLogisticRegression

* Hàm khởi tạo \_\_init\_\_: lớp này lưu trữ dữ liệu huấn luyện (X\_train, y\_train), dữ liệu validation (X\_valid, y\_valid) và dữ liệu kiểm tra (X\_test, y\_test).
* Hàm train\_model: tạo một đối tượng LogisticRegression với tham số max\_iter=1000 để đảm bảo rằng thuật toán hội tụ đúng cách. Sau đó, gọi phương thức fit để huấn luyện mô hình trên dữ liệu huấn luyện.
* Hàm evaluate\_model: đánh giá mô hình trên tập validation và tập kiểm tra. Đánh giá được thực hiện bằng cách sử dụng các chỉ số đánh giá như accuracy, precision, recall và F1 score để đo lường khả năng dự đoán của mô hình.

Kết quả được trả về là một vector dự đoán tùy theo tập validation hoặc tập test, và các chỉ số đánh giá được in ra màn hình để đánh giá hiệu suất của mô hình trên tập dữ liệu validation hoặc test

## 2.2. Mô hình mạng nơ ron nhân tạo ANN

### 2.2.1. Khái niệm

ANN là viết tắt của Artificial Neural Networks. Về cơ bản, đây là một mô hình tính toán, chúng được xây dựng dựa trên cấu trúc và chức năng của mạng lưới nơ ron trong Sinh học (mặc dù cấu trúc của ANN sẽ bị ảnh hưởng bởi một luồng thông tin). Do đó, mạng nơ ron này sẽ thay đổi, chúng phụ thuộc vào đầu vào và đầu ra.

Chúng ta có thể coi ANN là dữ liệu thống kê phi tuyến. Điều này đồng nghĩa với một mối quan hệ phức tạp, được xác định giữa đầu vào và đầu ra. Kết quả làvsẽ có nhiều mẫu khác nhau.

### 2.2.2. Quy trình thực hiện

Xây dựng lớp mô hình NeuralNetwork: lớp Model NeuralNetwork trong Python để quản lý việc huấn luyện và đánh giá mô hình Neural Network.

Cấu trúc của lớp ModelNeuralNetwork

* Hàm khởi tạo \_\_init\_\_: lớp này lưu trữ dữ liệu huấn luyện (X\_train, y\_train), dữ liệu validation (X\_valid, y\_valid) và dữ liệu kiểm tra (X\_test, y\_test).
* Hàm build\_model: Mô hình được xây dựng bằng cách sử dụng lớp Sequential từ TensorFlow, cho phép chúng ta xây dựng một chuỗi các lớp một cách tuần tự. Trong trường hợp này, mô hình bao gồm ba lớp Dense:
  + Lớp 1: Lớp Dense với 32 nơ-ron và hàm kích hoạt 'relu'. Đây là lớp đầu tiên trong mạng nơ-ron và nhận đầu vào có kích thước là số lượng đặc trưng của dữ liệu (số lượng cột của ma trận đầu vào).
  + Lớp 2: Lớp Dense với 16 nơ-ron và hàm kích hoạt 'relu'. Đây là lớp ẩn thứ hai trong mạng nơ-ron.
  + Lớp 3: Lớp Dense với 1 nơ-ron và hàm kích hoạt 'sigmoid'. Đây là lớp đầu ra của mạng nơ-ron và cho ra dự đoán xác suất làm cho việc xảy ra một sự kiện (trong trường hợp này, xác suất mắc bệnh tim).

Sau khi định nghĩa mô hình, chúng ta sử dụng phương thức compile để biên dịch mô hình. Trong trường hợp này, chúng ta đã chọn optimizer là 'adam' và hàm mất mát là 'binary\_crossentropy' (thích hợp cho bài toán phân loại nhị phân). Chúng ta cũng đã chọn chỉ số đánh giá là 'accuracy' để theo dõi hiệu suất của mô hình trong quá trình huấn luyện.

* Hàm train\_model: sử dụng phương thức fit để huấn luyện mô hình trên tập dữ liệu huấn luyện. Trong trường hợp này, chúng ta huấn luyện mô hình trong 5 epochs và sử dụng dữ liệu validation để đánh giá hiệu suất mô hình sau mỗi epoch.
* Hàm evaluate\_model: đánh giá mô hình trên tập validation và tập kiểm tra. Đánh giá được thực hiện bằng cách sử dụng các chỉ số đánh giá như accuracy, precision, recall và F1 score dựa trên dự đoán của mô hình trên tập validation với ngưỡng quyết định tùy chọn.

Kết quả được trả về là một vector dự đoán tùy theo tập validation hoặc tập test, và các chỉ số đánh giá được in ra màn hình để đánh giá hiệu suất của mô hình trên tập dữ liệu validation hoặc test.

## 2.3. Mô hình gradient boosting

### 2.3.1. Khái niệm

Đây là một phương pháp tối ưu hóa học máy được sử dụng để tạo ra một mô hình dự đoán bằng cách kết hợp nhiều mô hình yếu lại với nhau. Thuật toán này tạo ra một mô hình dự đoán bằng cách lặp lại quá trình tối ưu hóa trên các mô hình yếu. Gradient Boosting thường được sử dụng trong các bài toán như dự đoán giá cổ phiếu, hoặc dự đoán thành công của một sản phẩm mới.

### 2.3.2. Quy trình thực hiện

Xây dựng lớp mô hình GradientBoosting: lớp ModelGradientBoosting trong Python để quản lý việc huấn luyện và đánh giá mô hình Gradient Boosting.

Cấu trúc của lớp ModelGradientBoosting:

* Hàm khởi tạo \_\_init\_\_: lớp này lưu trữ dữ liệu huấn luyện (X\_train, y\_train), dữ liệu validation (X\_valid, y\_valid) và dữ liệu kiểm tra (X\_test, y\_test).
* Hàm train\_model: chúng ta tạo một mô hình Gradient Boosting với tham số random\_state=42 để đảm bảo kết quả có thể tái lặp lại. Sau đó, chúng ta huấn luyện mô hình trên dữ liệu huấn luyện bằng cách sử dụng phương thức fit.
* Hàm evaluate\_model: chúng ta đánh giá mô hình trên tập validation hoặc tập kiểm tra, tùy thuộc vào giá trị của biến is\_test.
  + Nếu is\_test là False (mặc định), chúng ta đánh giá trên tập validation.
    - Đầu tiên, chúng ta sử dụng phương thức predict\_proba để dự đoán xác suất của các lớp trên tập validation.
    - Tiếp theo, chúng ta áp dụng ngưỡng threshold để quyết định lớp dự đoán, và tính các chỉ số đánh giá như accuracy, precision, recall và F1 score.
  + Nếu is\_test là True, chúng ta đánh giá trên tập test (tập kiểm tra)
    - Tương tự, chúng ta sử dụng predict\_proba để dự đoán xác suất trên tập kiểm tra và tính các chỉ số đánh giá, bao gồm accuracy, precision, recall và F1 score.

Kết quả được trả về là một vector dự đoán tùy theo tập validation hoặc tập test, và các chỉ số đánh giá được in ra màn hình để đánh giá hiệu suất của mô hình trên tập dữ liệu validation hoặc test.

# CHƯƠNG 3: ĐÁNH GIÁ MÔ HÌNH VÀ CẢI THIỆN

Nhóm đã sử dụng 4 metrics phổ biến cho bài toán dự đoán phân loại là: Accuracy, Precision, Recall, F1-Score.

**Accuracy (Độ chính xác):** Đo lường tỉ lệ các dự đoán đúng trên tổng số các dự đoán.

**Precision (Độ chính xác dương tính):** Đo lường tỷ lệ số lượng dự đoán đúng so với tổng số lượng dự đoán dương tính.

**Recall (Độ nhạy):** Đo lường tỷ lệ dự đoán đúng dương tính so với tổng số lượng dương tính thực tế. Chỉ số này đặc biệt quan trọng trong các bài chuẩn đoán bệnh.

**F1-Score:** Chỉ số trung bình điều hòa của precision và recall, là tổng hợp của hiệu suất mô hình.

Mô hình cho thấy sự kém hiệu quả khi các chỉ số metrics đều rất tệ ở mức threshold là 0.5:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 0.91 | 0.52 | 0.10 | 0.16 |
| Gradient Boosting | 0.92 | 0.55 | 0.08 | 0.14 |
| Neural Network | 0.92 | 0.59 | 0.06 | 0.11 |

Bảng 4: Kết quả các chỉ số đánh giá khi sử dụng các mô hình

Ở bước đầu sử dụng bộ dữ liệu gốc. Mô hình đạt độ chính xác rất cao (trên 90%), tuy nhiên có thể đây là do sự chênh lệch dữ liệu giữa số lượng người không mắc bệnh tim và mắc bệnh tim. Khi quan sát 2 chỉ số Precision và Recall, cả ba mô hình đều không đạt được mục tiêu nhóm đề ra - đặc biệt là Recall khi chỉ có 10% (bỏ sót những trường hợp dương tính), qua đó cũng khiến chỉ số F1-Score cực kì thấp.

Từ đó nhận thấy được điểm yếu của mô hình, tiến hành thực hiện một số cải thiện nhóm đề ra.

## 3.1. Giải quyết vấn đề mất cân bằng dữ liệu

Nhóm chọn sử dụng một phương pháp oversampling khá phổ biến là SMOTE (Synthetic Minority Over-sampling Technique). Phương pháp này sinh mẫu nhằm gia tăng kích thước cho nhóm thiểu số bằng cách lựa chọn k mẫu gần với nó và thực hiện tổ hợp tuyến tính để tạo ra mẫu giả lập. Điều này làm giảm nguy cơ overfiting và underfitting.

Kết quả khi áp dụng lên cả ba mô hình đều làm thay đổi đáng kể hiệu suất.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | | Precision | | Recall | | F1-Score | |
| Balanced | Imbalance | Balanced | Imbalance | Balanced | Imbalance | Balanced | Imbalance |
| Logistic Regression | 0.77 | 0.91 | 0.76 | 0.52 | 0.80 | 0.10 | 0.78 | 0.16 |
| **Gradient Boosting** | **0.86** | 0.92 | **0.86** | 0.55 | **0.87** | 0.08 | **0.87** | 0.14 |
| Neural Network | 0.78 | 0.92 | 0.75 | 0.59 | 0.84 | 0.06 | 0.79 | 0.11 |

Bảng 5: Kết quả các chỉ số đánh giá sau khi sử dụng phương pháp SMOTE

Dựa trên bảng so sánh giữa tập dữ liệu cân bằng và tập dữ liệu gốc, ta có thể nhận thấy sự cải thiện rõ rệt khi sử dụng phương pháp cân bằng dữ liệu. Dưới đây là một số điểm cải thiện:

**Accuracy:**

Đối với mọi mô hình, accuracy của tập dữ liệu cân bằng đều thấp hơn so với tập dữ liệu gốc. Mô hình Gradient Boosting chỉ giảm 6%, 2 mô hình còn lại đều giảm khoảng 14%  
Mặc dù vậy, accuracy trên tập dữ liệu cân bằng vẫn ở mức khá cao, chỉ thấp hơn một chút so với tập dữ liệu gốc. Điều này cho thấy mô hình vẫn giữ được khả năng dự đoán tổng quát tốt trên tập dữ liệu cân bằng.

**Precision:**

Khi áp dụng phương pháp cân bằng dữ liệu, chỉ số này tăng so với ban đầu.  
Mô hình Gradient Boosting đặc biệt cao khi tăng 30%, Mô hình Logistic chỉ tăng 24% thấp nhất trong 3 mô hình. Tuy nhiên mức tăng này là đáng kể.

**Recall:**  
Mô hình Gradient Boosting lại tiếp tục thể hiện sự vượt trội khi tăng thêm 79%. Hai mô hình cọn lại cũng tăng rất rõ rệt.

**F1 Score:**

F1 score, là sự kết hợp giữa precision và recall, thể hiện sự hiệu quả của mô hình.  
Cả ba mô hình điều cải thiện F1-Score rất lớn khi tăng khoảng hơn 70%. Trong đó mô hình Gradient Boosting tăng nhiều nhất.

* **Điều này chỉ ra rằng mô hình trên tập dữ liệu cân bằng có khả năng cân bằng giữa việc tối ưu hóa cả precision và recall cho cả hai lớp. Khi áp dụng phương pháp cân bằng dữ liệu, mô hình được cải tiến một cách rõ rệt. Đặc biệt là mô hình Gradient Boosting**

## 3.2. Hướng cải thiện mô hình

Thông thường khi xác định mức để phân loại. Ta thường chọn giá trị là 0.5 (mặc định) để phân loại nhị phân. Hoặc chọn giá trị cao để mô hình tốt tránh dự đoán nhằm. Việc chọn và xác định **ngưỡng** của mô hình có thể ảnh hưởng nghiêm trọng đến chuẩn đoán bệnh.

Với ngưỡng cao có thể dẫn đến tình trạng âm tính giả khi đó người bệnh có thể đã mắc bệnh tim rồi nhưng vẫn được chuẩn đoán là không mắc.

* Từ đó khiến người bệnh chủ quan, gây nguy hiểm cho sức khỏe và tính mạng.

Trong lĩnh vực y học, đường cong AUC-ROC và Confusion matrix là kỹ thuật quan trọng được sử dụng để đánh giá hiệu suất của các mô hình phân loại.

Đường cong AUC-ROC: Thể hiện giữa True Positive Rate(Tỉ lệ dương tính dự đoán đúng) và False Positive Rate (Tỉ lệ dự đoán sai ở trường hợp âm tính) ở các ngưỡng quyết định khác nhau.

Confusion matrix: Thể hiện tương quan số lượng giữa các phân loại dự đoán so với thực tế.

Confusion matrix của mô hình gradient boosting ở tập dữ liệu cân bằng (ngưỡng là 0.5):

A screenshot of a blue and white grid

Description automatically generated

Hình 36: Confusion matrix của mô hình gradient boosting ở tập dữ liệu cân bằng với ngưỡng 0.5

Mô hình đã khá hiệu quả khi dự đoán rất chính xác, số lượng các dự đoán sai chỉ chiếm tỉ lệ rất nhỏ (khoảng 15%). Tuy nhiên mục tiêu của nhóm đề ra là giảm thiểu các trường hợp âm tính giả mà vẫn giữ được sự ổn định của các trường hợp dương tính giả.

A graph with a line and a point

Description automatically generated with medium confidence A blue squares with white text

Description automatically generated

Hình 37: Đường cong ROC và Confusion Matrix của mô hình với ngưỡng tốt nhất

Đường cong ROC cho thấy mô hình đạt được hiệu quả tối đa, diện tích AUC = 0.95, điểm ngưỡng tốt nhất của đường cong ROC nằm ở vị trí mà ở đó tỉ lệ TPR và FPR đạt chênh lệch cao nhất.

Tuy nhiên có thể thấy mặc dù ngưỡng tốt nhất của đường cong thể hiện hiệu quả tối đa của mô hình, nhưng tỉ lệ âm tính giả lại tăng lên so với mô hình ban đầu. Điều này dẫn đến cần phải so sánh cụ thể các ngưỡng tuân theo đường cong ROC.

* **Lựa chọn ngưỡng phù hợp với mục đích**

A graph of different colored lines

Description automatically generated

Hình 38: Biểu đồ thể hiện giữa Metrics và Threshold

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Threshold = 0.30 | 0.83 | 0.76 | 0.95 | 0.85 |
| Threshold = 0.32 | 0.84 | 0.78 | 0.95 | 0.85 |
| Threshold = 0.34 | 0.84 | 0.79 | 0.94 | 0.86 |
| Threshold = 0.37 | 0.85 | 0.80 | 0.93 | 0.86 |
| Threshold = 0.39 | 0.85 | 0.81 | 0.92 | 0.86 |
| Threshold = 0.41 | 0.86 | 0.82 | 0.91 | 0.86 |
| Threshold = 0.43 | 0.86 | 0.83 | 0.90 | 0.86 |
| Threshold = 0.46 | 0.86 | 0.84 | 0.90 | 0.87 |
| Threshold = 0.48 | 0.86 | 0.85 | 0.89 | 0.87 |
| Threshold = 0.50 | 0.86 | 0.86 | 0.87 | 0.87 |
| Threshold = 0.57 (ROC) | 0.87 | 0.89 | 0.84 | 0.86 |

Bảng 6: Bảng thể hiện các chỉ số metrics với các threshold khác nhau

Quan sát cả đồ thị và bảng, các chỉ số đánh giá thay đổi không theo chiều hướng, ngưỡng càng tăng, precision và accuracy càng tăng, recall càng giảm. Khi ngưỡng ở mức thấp, recall đạt rất cao, đúng yêu cầu của nhóm, tuy nhiên lại phần nào ảnh hưởng đến precision và accuracy, có thể dẫn đến trường hợp dương tính giả khá nhiều. Để đảm bảo cho tất cả các chỉ số đều ở mức ổn định mà vẫn giảm thiểu được tình trạng âm tính giả, nhóm quyết định chọn ngưỡng là 0.39.

# CHƯƠNG 4: KẾT QUẢ MÔ HÌNH

Khi áp dụng ngưỡng nhóm đề ra có thể cố tình gây ra tình trạng dương tính giả, tuy nhiên lại không gây ảnh hưởng lớn, điều này còn có thể tăng ý thức bảo vệ sức khỏe của người bệnh khi họ vô tình bị dự đoán là dương tính. Nếu có thể áp dụng mô hình này với kinh nghiệm của các chuyên gia/ bác sĩ. Có thể xem xét qua các tình trạng bệnh nhân.

Kết quả có thể quan sát dựa trên tập dữ liệu test cuối cùng

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Mô hình Gradient Boosting với threshold = 0.39 | 0.85 | 0.81 | 0.92 | 0.86 |

Bảng 7: Bảng thể hiện các chỉ số metrics với threshold là 0.39

A blue squares with white numbers

Description automatically generated

Hình 39: Kết quả của Confusion Matrix trên tập dữ liệu cuối cùng

Tuy nhiên để đảm báo mô hình vận hành một cách hiệu quả nên kết hợp với kiến thức chuyên môn của các y bác sĩ.

# CHƯƠNG 5: MỞ RỘNG ĐỀ TÀI

Đề xuất các chiến lược và chương trình phòng ngừa bệnh tim dựa trên kết quả của mô hình dự đoán. Mô hình cho thấy sự hiệu quả trên cả tập test và tập valid. Có thể áp dụng vào thực tế, mô hình cho thấy mối quan hệ rõ rệt giữa các yếu tố và nguy cơ mắc bệnh tim. Thông qua so sánh về tỉ lệ giữa nhóm người sức khỏe tốt và sức khỏe kém ở người nhỏ tuổi, có thể thấy nhóm người sức khỏe kém nguy cơ mắc bệnh tim cao hơn.

## 5.1. Phát triển chiến lược phòng ngừa và can thiệp

### 5.1.1 Chiến lược

Xem xét các biện pháp can thiệp và cải thiện lối sống để giảm thiểu nguy cơ mắc bệnh tim, dựa trên những hiểu biết từ nghiên cứu:

+ Thúc đẩy chế độ ăn uống lành mạnh, giảm BMI, cholesterol.

+ Khuyến khích vận động thể chất đều đặn, bao gồm tập thể dục và hoạt động vận động hàng ngày.

+ Ngừng hút thuốc lá và hạn chế tiêu thụ cồn.

+ Cần chủ động phòng ngừa các bệnh như đột quỵ, suy thận, ung thư da.

+ Khám sức khỏe định kỳ để phát hiện được nguy cơ sớm

### 5.1.2. Chương trình phòng ngừa

+ Tổ chức các buổi hội thảo hoặc chiến dịch giáo dục cộng đồng để nâng cao nhận thức về bệnh tim và các biện pháp phòng ngừa.

+ Thiết lập hệ thống theo dõi và đánh giá hiệu quả của chương trình phòng ngừa, bao gồm việc theo dõi sự thay đổi trong chỉ số sức khỏe và số lượng người tham gia vào các hoạt động phòng ngừa.

+ Áp dụng mô hình dự đoán vào các chương trình có thể phát hiện sớm hoặc tăng ý thức đề phòng bệnh.

## 5.2. Đề xuất hướng nghiên cứu trong tương lai

Đề xuất các hướng nghiên cứu tiếp theo để cải thiện và mở rộng hiểu biết về bệnh tim, bao gồm cả việc sử dụng dữ liệu mới và phát triển các phương pháp phân tích tiên tiến hơn.

Rủi ro của mô hình là các thuộc tính dữ liệu là các bệnh khá nguy hiểm (suy thận, ung thư da, béo phì, hen suyễn).

Cần thu thập nhiều dữ liệu mới để tổng quát hóa mô hình bao gồm các dữ liệu phức tạp hơn: ảnh chụp, hồ sơ điện tử, đồng hồ theo dõi sức khỏe. Đồng thời kết hợp các thuộc tính mới khác bao gồm cả yếu tố di truyền, môi trường, và lối sống…

Cần xem xét thêm nhiều mô hình học sâu, học tăng cường, cải thiện mạng nơ ron nhân tạo. Sử dụng các phương pháp phân tích đa biến để phân tích dữ liệu phức tạp và tìm ra mối liên hệ giữa nhiều yếu tố khác nhau đối với bệnh tim.

# CHƯƠNG 6: KẾT LUẬN ĐỀ TÀI

Nhóm đã thực hiện một quá trình thu thập, tiền xử lý, phân tích kỹ lưỡng. Xây dựng nhiều mô hình, tiến hành so sánh, đánh giá và cải thiện mô hình qua từng phương pháp. Trong quá trình huấn luyện mô hình, nhóm đã áp dụng các phương pháp cân bằng dữ liệu, dùng đường cong AUC-ROC để lựa chọn ngưỡng phù hợp. Với mục tiêu dự đoán nhóm đã chọn ra các mô hình Gradient Boosting, Logistic Regression, Neural Network phù hợp với mô hình phân loại. Mô hình Gradient Boosting đạt hiệu quả cao nhất khi so sánh. Mô hình hình Logistic Regression và Neural Network tuy cho hiệu quả thấp so với Gradient Boosting nhưng vẫn đạt được hiệu suất rất cao, góp phần vào việc đánh giá toàn diện hiệu suất. Mô hình khi áp dụng ngưỡng đề ra mặc dù làm gia tăng tình trạng dương tính giả nhưng qua đó cũng có thể nâng cao ý thức bảo vệ sức khỏe của người bệnh, không ảnh hưởng quá lớn đến hiệu suất mô hình mà vẫn giứa tỉ lệ âm tính giả đạt mức thấp.

# TÀI LIỆU THAM KHẢO

[1]. Sách giáo trình [Learning Data Science Data Wrangling, Exploration, Visualization, and Modeling with Python - Sam LauFile](https://fhqx.hcmute.edu.vn/mod/resource/view.php?id=1543800).

[2]. UCI Heart Disease Data

<https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data>

[3]. Indicators of Heart Disease (2022 UPDATE)

<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

[4]. Heart Disease Health Indicators Dataset

<https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset>

[5]. Phạm Đình Khánh - Mất cân bằng dữ liệu (imbalanced dataset), đăng ngày 17/2/2020.

<https://phamdinhkhanh.github.io/2020/02/17/ImbalancedData.html#442-smote--adasyn>

[6]. AHMAD RAFIEE - Handling Imbalanced | 2-Resampling Techniques

<https://www.kaggle.com/code/ahmadrafiee/handling-imbalanced-2-resampling-techniques/notebook>

[7]. KIRAN HAYAT - EDA\_Heart\_Disease\_UCI

<https://www.kaggle.com/code/kirahhayatdata/eda-heart-disease-uci/notebook>

[8]. AHMAD RAFIEE - Handling imbalanced | 1-class-weight | halvinggrid

<https://www.kaggle.com/code/ahmadrafiee/handling-imbalanced-1-class-weight-halvinggrid>