TRƯỜNG ĐẠI HỌC TRÀ VINH KHOA KỸ THUẬT VÀ CÔNG NGHỆ **BỘ MÔN CÔNG NGHỆ THÔNG TIN**

**----------****-----------**

****

BÁO CÁO MÔN HỌC

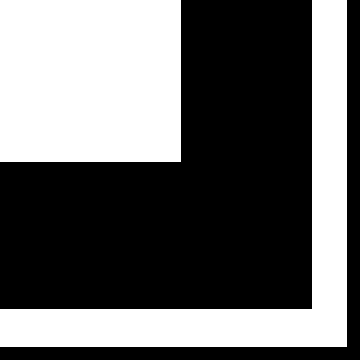
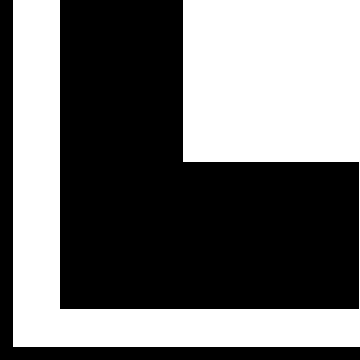
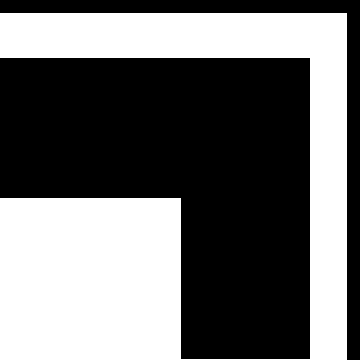
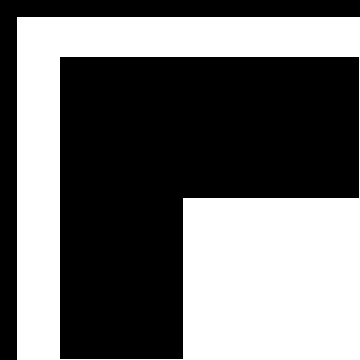
**HỌC PHẦN: KHAI PHÁ DỮ LIỆU**

**ỨNG DỤNG LOGISTIC REGRESSION DỰ ĐOÁN NGUY CƠ**

**MẮC BỆNH TIM MẠCH**

|  |  |  |  |
| --- | --- | --- | --- |
| *Giáo viên hướng dẫn:* |  | *Sinh viên thực hiện - Mã số sinh viên:* | |
| **Th.S Nguyễn Thái Toàn** |  | Võ Nhật Duy Nam | 110122119 |
|  |  | Lưu Bích Ngọc | 110122123 |
|  |  | Trương Nguyễn Tố Nguyên | 110122127 |
|  |  | Kiều Tấn Phước | 110122144 |
|  |  | Cô Nhân Quý | 110122150 |
|  |  |  |  |
|  |  | Lớp DA22TTB | |

**Trà Vinh, tháng 5 năm 2025**



**LỜI CẢM ƠN**

Em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến ThS. Nguyễn Thái Toàn, giáo viên hướng dẫn, người đã tận tình hướng dẫn, chỉ bảo và tạo điều kiện thuận lợi để nhóm chúng em trong suốt quá trình thực hiện đồ án môn học Khai Phá Dữ Liệu với đề tài " Ứng dụng Logistic Regression về Dự đoán nguy cơ mắc bệnh tim mạch". Với những kiến thức quý báu của thầy đã giúp chúng em vượt qua những khó khăn trong quá trình nghiên cứu và hoàn thiện báo cáo.

Sự nhiệt tình và kiên nhẫn của thầy đã giúp chúng em hoàn thành đồ án một cách suôn sẻ và hiệu quả hơn. Chúng em thật sự rất biết ơn và mong rằng sẽ tiếp tục nhận được sự giúp đỡ từ thầy trong những lần sau.

Xin chân thành cảm ơn thầy!

*Sinh viên ký và ghi rõ họ và tên*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sinh viên 1** | **Sinh viên 2** | **Sinh viên 3** | **Sinh viên 4** | **Sinh viên 5** |
| Võ Nhật Duy Nam | Lưu Bích Ngọc | Trương Nguyên Tố Nguyên | Kiều Tấn Phước | Cô Nhân Quý |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

**MỤC LỤC**

[TÓM TẮT ĐỀ TÀI 7](#_Toc201354409)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 9](#_Toc201354410)

[1.1. Lý do chọn đề tài 9](#_Toc201354411)

[1.1.1. Tính cấp thiết: 9](#_Toc201354412)

[1.1.2. Lý do chọn chủ đề: 9](#_Toc201354413)

[1.2. Mục tiêu đề tài 10](#_Toc201354414)

[1.3. Đối tượng và dữ liệu nghiên cứu 10](#_Toc201354415)

[1.3.1. Đối tượng nghiên cứu: 10](#_Toc201354416)

[1.3.2. Dữ liệu nghiên cứu: 10](#_Toc201354417)

[1.4. Phương pháp thực hiện 12](#_Toc201354418)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 15](#_Toc201354419)

[2.1. Cơ sở lý thuyết thứ nhất liên quan đến đề tài 15](#_Toc201354420)

[2.3. Giới thiệu về bộ dữ liệu 16](#_Toc201354421)

[2.3.1. Mô tả bộ dữ liệu 16](#_Toc201354422)

[2.3.2. Danh sách các biến 18](#_Toc201354423)

[2.3.3. Quá trình tiền xử lý dữ liệu 19](#_Toc201354424)

[2.3.4. Bộ dữ liệu hoàn chỉnh 20](#_Toc201354425)

[CHƯƠNG 3: THỰC HIỆN KHAI PHÁ BỘ DỮ LIỆU 26](#_Toc201354426)

[3.1. Khai phá dữ liệu 26](#_Toc201354427)

[3.2. Kết quả thực hiện 29](#_Toc201354428)

[3.1.2. Thống kê mô tả và tương quan 34](#_Toc201354429)

[3.1.3. Huấn luyện mô hình 36](#_Toc201354430)

[3.2. Kết quả đạt được 38](#_Toc201354431)

[CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 43](#_Toc201354432)

[4.1. Kết quả đạt được 43](#_Toc201354433)

[4.2. Ưu nhược điểm 44](#_Toc201354434)

[4.3. Hướng phát triển 45](#_Toc201354435)

[TÀI LIỆU THAM KHẢO 48](#_Toc201354436)

DANH MỤC BẢNG BIỂU

[Bảng 1.1 Bảng cấu trúc dữ liệu dataset 11](#_Toc201354501)

[Bảng 2.2 Bảng phân phối các biến 23](#_Toc201354502)

[Bảng 3.1 Báo cáo phân loại 37](#_Toc201354503)

**DANH MỤC HÌNH ẢNH**

[Hình 3.1 Thông tin dataset 30](#_Toc201354454)

[Hình 3.2 Tỉ lệ giá trị thiếu 31](#_Toc201354455)

[Hình 3.3 Phân tích và xử lý outlier 31](#_Toc201354456)

[Hình 3.4 So sánh trước và sau khi xử lý outlier 32](#_Toc201354457)

[Hình 3.5 Kiểm tra tính chuẩn 33](#_Toc201354458)

[Hình 3.6 Xử lý Skewness 33](#_Toc201354459)

[Hình 3.7 Heatmap 36](#_Toc201354460)

[Hình 3.8 Đánh giá mô hình 37](#_Toc201354461)

[Hình 3.10 Thử nghiệm lần 1 39](#_Toc201354462)

[Hình 3.11 Thử nghiệm lần 2 39](#_Toc201354463)

[Hình 3.12 Thử nghiệm lần 3 40](#_Toc201354464)

# TÓM TẮT ĐỀ TÀI

Đề tài tập trung vào phân tích dữ liệu các thí nghiệm, được tổng hợp từ nhiều nguồn khác nhau theo nhiều thuộc tính (76 thuộc tính). Đặc biệt là những thí nghiệm này đều đã được công bố và dùng chung tập con gồm 14 thuộc tính để phát triển mô hình dự đoán bệnh tim sử dụng thuật toán Machine Learning dựa trên bộ dữ liệu Heart Disease từ UCI Machine Learning Repository.

**Nguồn dữ liệu**

Bộ dữ liệu Heart Disease từ UCI, được thu thập từ 4 cơ sở y tế:

* Cleveland Clinic Foundation
* Hungarian Institute of Cardiology
* V.A. Medical Center (Long Beach, CA)
* University Hospital (Zurich, Switzerland)

**Đặc điểm dữ liệu**:

* Tổng số thuộc tính gốc: 76 thuộc tính.
* Tập con được sử dụng phổ biến: 14 thuộc tính chính.
* Các thuộc tính bao gồm: tuổi, giới tính, loại đau ngực, huyết áp, cholesterol, đường huyết, ECG, nhịp tim tối đa, đau ngực khi gắng sức, v.v.

**Phương pháp thực hiện:**

* Tiền xử lý dữ liệu: làm sạch, xử lý giá trị thiếu, chuẩn hóa dữ liệu.
* Phân tích khám phá dữ liệu (EDA): tìm hiểu phân phối, tương quan giữa các biến.
* Áp dụng các thuật toán Machine Learning: Logistic Regression.
* Đánh giá và so sánh hiệu suất các mô hình.

**Kết quả chính:**

* Độ chính xác cao: Các mô hình đạt độ chính xác từ 80-95% trong việc dự đoán bệnh tim.
* Các yếu tố quan trọng: Tuổi, loại đau ngực, nhịp tim tối đa, và kết quả ECG là những yếu tố có ảnh hưởng lớn nhất.
* Ứng dụng thực tế: Mô hình có thể hỗ trợ bác sĩ trong chẩn đoán sớm và đưa ra quyết định điều trị.

**Ý nghĩa và ứng dụng:**

Nghiên cứu góp phần vào việc ứng dụng công nghệ AI trong y tế, đặc biệt là trong chẩn đoán và dự đoán bệnh tim - một trong những nguyên nhân gây tử vong hàng đầu thế giới. Kết quả có thể được triển khai trong các hệ thống hỗ trợ quyết định lâm sàng.

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## Lý do chọn đề tài

### Tính cấp thiết:

Bệnh tim mạch (cardiovascular disease - CVD) hiện đang là vấn đề sức khỏe cộng đồng cấp bách trên toàn thế giới và tại Việt Nam. Theo Tổ chức Y tế Thế giới (WHO), bệnh tim mạch là nguyên nhân gây tử vong hàng đầu globally, chiếm khoảng 17.9 triệu ca tử vong mỗi năm. Tại Việt Nam, tỷ lệ mắc các bệnh tim mạch đang có xu hướng gia tăng do:

* **Thay đổi lối sống**: Chế độ ăn uống không khoa học, ít vận động, căng thẳng;
* **Già hóa dân số**: Tỷ lệ người cao tuổi tăng nhanh;
* **Ô nhiễm môi trường**: Ảnh hưởng trực tiếp đến sức khỏe tim mạch;
* **Tăng tỷ lệ các bệnh đồng mắc**: Tiểu đường, tăng huyết áp, béo phì.

Việc phát hiện sớm và dự đoán chính xác nguy cơ mắc bệnh tim mạch có ý nghĩa quan trọng trong việc:

* Giảm tỷ lệ tử vong và tàn tật;
* Tiết kiệm chi phí điều trị;
* Nâng cao chất lượng cuộc sống người dân.

### Lý do chọn chủ đề:

**Về mặt y tế:**

* Bệnh tim mạch có thể **phòng ngừa được** nếu phát hiện sớm các yếu tố nguy cơ;
* Nhiều yếu tố nguy cơ có thể **đo lường và theo dõi được** thông qua các chỉ số lâm sàng đơn giản;
* Việc dự đoán chính xác giúp bác sĩ **đưa ra quyết định điều trị kịp thời.**

**Về mặt kỹ thuật:**

* **Logistic Regression** là mô hình phù hợp cho bài toán phân loại nhị phân (có bệnh/không bệnh);
* Dataset Heart Disease là **bộ dữ liệu chuẩn** được sử dụng rộng rãi trong nghiên cứu;
* Các biến đầu vào có **ý nghĩa lâm sàng rõ ràng** và dễ thu thập trong thực tế.

**Về mặt thực tiễn:**

* Mô hình có thể **triển khai trong các cơ sở y tế** để hỗ trợ chẩn đoán;
* Chi phí thấp, có thể **ứng dụng tại các vùng thiếu thốn trang thiết bị;**
* Kết quả dễ **giải thích và truyền đạt** cho bệnh nhân.

## Mục tiêu đề tài

Xây dựng mô hình Logistic Regression để dự đoán nguy cơ mắc bệnh tim mạch dựa trên các yếu tố nguy cơ lâm sàng, nhằm hỗ trợ việc chẩn đoán sớm và phòng ngừa bệnh tim mạch hiệu quả.

## Đối tượng và dữ liệu nghiên cứu

### Đối tượng nghiên cứu:

Đối tượng chính: Bệnh nhân có nguy cơ mắc bệnh tim mạch

Đặc điểm đối tượng:

* Độ tuổi: Từ 29-77 tuổi (theo dataset)
* Giới tính: Cả nam và nữ
* Tình trạng sức khỏe: Bao gồm cả người khỏe mạnh và người có các yếu tố nguy cơ
* Địa điểm: Dữ liệu từ Cleveland, Hungary, Switzerland, and the VA Long Beach **.**

Tiêu chí đưa vào nghiên cứu:

* Có đầy đủ thông tin về 14 yếu tố nguy cơ chính
* Đã thực hiện các xét nghiệm lâm sàng cần thiết
* Có kết quả chẩn đoán xác định về tình trạng bệnh tim mạch

.

### Dữ liệu nghiên cứu:

Nguồn dữ liệu: UCI Machine Learning Repository - Cleveland Heart Disease Dataset

Link: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data>

Số lượng mẫu: 303 bệnh nhân

Số lượng biến: 14 biến (13 biến đầu vào + 1 biến mục tiêu)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên biến | Mô tả | Loại dữ liệu | Giá trị |
| 1 | Age | Tuổi | Interger | 29-77 |
| 2 | Sex | Giới tính | Categorical | 0= nữ, 1 =nam |
| 3 | Cp | Loại đau ngực | Categorical | 1-4 |
| 4 | Trestbps | Huyết áp nghỉ ngơi | Interger | mmHg |
| 5 | Chol | Cholesterol | Interger | mg/dl |
| 6 | Fbs | Đường huyết lúc đói | Categorical | 0/1(>120mg/dl) |
| 7 | Restecg | Điện tim nghỉ ngơi | Categorical | 0-2 |
| 8 | Thalach | Nhịp tim tối đa | Interger | Bpm |
| 9 | Exang | Đau thắt ngực | Categorical | 0= không, 1= có |
| 10 | Odpeak | Chênh lệch st | Interger | mm |
| 11 | Slope | Đọ dốc st | Categorical | 1-3 |
| 12 | Ca | Sô mạch máu tắc nghẽn | Interger | 0-3 |
| 13 | Thal | Thalassemia | Categorical | 3, 6, 7 |
| 14 | Target | Mức độ mắc bệnh tim | Interger | 0= không,  1-4= có |

Bảng 1.1 Bảng cấu trúc dữ liệu dataset

Ưu điểm của dữ liệu:

* Dataset chuẩn, được sử dụng rộng rãi trong nghiên cứu;
* Các biến có ý nghĩa lâm sàng rõ ràng;
* Dữ liệu đã được làm sạch và chuẩn hóa;
* Cân bằng giữa các lớp (có bệnh/không bệnh).

Hạn chế của dữ liệu:

* Kích thước mẫu tương đối nhỏ (303 mẫu);
* Dữ liệu từ một trung tâm y tế duy nhất;
* Một số biến có giá trị missing cần xử lý.

Phương pháp xử lý dữ liệu:

* Kiểm tra và xử lý missing values;
* Chuẩn hóa và mã hóa các biến phân loại;
* Phân chia dữ liệu train/test theo tỷ lệ 80/20;
* Áp dụng các kỹ thuật feature scaling nếu cần thiết.

## Phương pháp thực hiện

**Phương pháp lý thuyết:**

**Logistic Regression:**

Logistic Regression sử dụng hàm sigmoid để ánh xạ giá trị đầu ra về khoảng [0,1].

Công thức: P(y=1|x) = 1/(1 + e^(-(β₀ + β₁x₁ + β₂x₂ + ... + βₙxₙ))).

Trong đó: βᵢ là các hệ số (weights), xᵢ là các biến đầu vào.

**Ưu điểm:**

* Phù hợp cho bài toán phân loại nhị phân;
* Kết quả có thể giải thích được (probability output);
* Không yêu cầu các giả định phân phối của dữ liệu;
* Ít bị overfitting với dữ liệu tuyến tính.

**Hàm mất mát:** Cross-entropy loss function.

Loss = -[y\*log(p) + (1-y)\*log(1-p)]

**Các Metrics Đánh Giá**

* **Accuracy**: Tỷ lệ dự đoán chính xác;
* **Precision**: Tỷ lệ dự đoán đúng trong các trường hợp dương tính;
* **Recall (Sensitivity)**: Tỷ lệ phát hiện đúng các trường hợp dương tính;
* **F1-Score**: Trung bình điều hòa của Precision và Recall;
* **AUC-ROC**: Diện tích dưới đường cong ROC.

**Phương pháp thu thập dữ liệu:**

* Sử dụng dữ liệu có sẵn từ dataset: Heart Disease Dataset từ UCI Machine Learning Repository, áp dụng trong giai đoạn chuẩn bị dữ liệu;
* **URL**: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data> ;
* **Phương thức thu thập**: Download trực tiếp từ repository công khai.

**Phương pháp thực nghiệm:**

Môi Trường Thực Nghiệm: Google Colab:

* **Ngôn ngữ**: Python 3.x.
* **Thư viện chính**:
* pandas: Xử lý và phân tích dữ liệu
* numpy: Tính toán số học
* matplotlib, seaborn: Trực quan hóa dữ liệu
* scikit-learn: Machine learning algorithms

Quy trình thực nghiệm:

1. Kiểm tra phân tích dữ liệu
2. Trực quan hóa dữ liệu
3. Tiền xử lý dữ liệu
4. Chia dữ liệu
5. Xây dựng mô hình
6. Trực quan hóa kết quả

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Cơ sở lý thuyết thứ nhất liên quan đến đề tài

Phân tích thống kê mô tả là cơ sở lý thuyết thứ nhất được áp dụng để khám phá và hiểu dữ liệu bệnh tim từ bộ dữ liệu UCI Heart Disease. Phương pháp này tập trung vào việc tóm tắt và mô tả các đặc điểm của bệnh nhân, bao gồm các chỉ số trung tâm, độ biến thiên và phân phối của các yếu tố nguy cơ.

* **Ứng dụng cụ thể**:
* Phân tích phân phối tuổi, giới tính của bệnh nhân mắc bệnh tim;
* So sánh trung bình các chỉ số sinh học (huyết áp, cholesterol, nhịp tim) giữa nhóm có và không có bệnh;
* Xác định tỷ lệ các loại đau ngực, kết quả ECG bất thường.
* **Ý nghĩa**: Kết quả phân tích mô tả cung cấp cái nhìn tổng quan về đặc điểm bệnh nhân, giúp nhận dạng các yếu tố nguy cơ quan trọng và chuẩn bị dữ liệu cho mô hình Logistic Regression.

**Ma trận tương quan và kiểm định giả thuyết:**

Phân tích tương quan được sử dụng để:

* Đánh giá mối quan hệ tuyến tính giữa các biến liên tục;
* Phát hiện đa cộng tuyến trước khi xây dựng mô hình;
* Xác định các biến có tương quan mạnh với biến mục tiêu.

**2.2. Cơ sở lý thuyết thứ hai liên quan đến đề tài**

Hồi quy Logistic là nền tảng lý thuyết chính được sử dụng để dự đoán nguy cơ mắc bệnh tim mạch từ bộ dữ liệu UCI Heart Disease. Khác với hồi quy tuyến tính, Logistic Regression được thiết kế đặc biệt cho bài toán phân loại nhị phân, phù hợp với việc dự đoán có/không mắc bệnh tim mạch.

**Công thức toán học:** Hàm sigmoid được sử dụng để chuyển đổi đầu ra thành xác suất:

**P(Y=1|X) = 1 / (1 + e^(-z))**

Trong đó: z = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ

**Ứng dụng cụ thể trong dự đoán bệnh tim mạch:**

* Biến phụ thuộc (Y): Trạng thái mắc bệnh tim (0: không mắc, 1: mắc bệnh);
* Biến độc lập (X): Các yếu tố nguy cơ như tuổi, giới tính, huyết áp, cholesterol, đau ngực, v.v.
* Mô hình tính toán xác suất mắc bệnh từ 0 đến 1.

**Ý nghĩa:** Logistic Regression cho phép không chỉ dự đoán phân loại mà còn cung cấp xác suất mắc bệnh, giúp bác sĩ đánh giá mức độ nguy cơ và đưa ra quyết định điều trị phù hợp.

**Ưu điểm của Logistic Regression trong y tế**

* **Khả năng diễn giải cao:** Hệ số β cho biết mức độ ảnh hưởng của từng yếu tố nguy cơ;
* **Không yêu cầu phân phối chuẩn:** Phù hợp với dữ liệu y tế thực tế;
* **Xử lý được biến hỗn hợp:** Có thể sử dụng cả biến liên tục và phân loại;
* **Ít nhạy cảm với ngoại lai:** Phù hợp với dữ liệu y tế có thể có giá trị bất thường

## 2.3. Giới thiệu về bộ dữ liệu

* + 1. ***Mô tả bộ dữ liệu***

##### Vấn đề nghiên cứu:

Bộ dữ liệu UCI Heart Disease được sử dụng để nghiên cứu các yếu tố nguy cơ gây bệnh tim mạch và phát triển mô hình dự đoán tự động. Mục tiêu là xây dựng một hệ thống hỗ trợ chẩn đoán giúp các bác sĩ đánh giá nguy cơ mắc bệnh tim của bệnh nhân dựa trên các chỉ số lâm sàng và cận lâm sàng.

**Nguồn dữ liệu và tác giả gốc:**

Dữ liệu được thu thập từ UCI Machine Learning Repository, bao gồm hồ sơ bệnh án từ 4 bệnh viện:

* Cleveland Clinic Foundation (Cleveland, Ohio)
* Hungarian Institute of Cardiology (Budapest)
* V.A. Medical Center (Long Beach, California)
* University Hospital (Zurich, Switzerland)

Bộ dữ liệu được tạo ra bởi:

* Robert Detrano, M.D., Ph.D. (V.A. Medical Center)
* Andras Janosi, M.D. (Hungarian Institute of Cardiology)
* Walter Steinbrunn, M.D. (University Hospital, Zurich)
* Matthias Pfisterer, M.D. (University Hospital, Basel)

##### Tình trạng dữ liệu hiện có:

Bộ dữ liệu gốc chứa 76 thuộc tính, nhưng các nghiên cứu thường sử dụng tập con 14 thuộc tính quan trọng nhất. Dữ liệu đã được làm sạch và chuẩn hóa, với các thông tin cá nhân nhạy cảm đã được loại bỏ để bảo vệ quyền riêng tư bệnh nhân.

* + 1. ***Danh sách các biến***

Bộ dữ liệu UCI Heart Disease chứa 14 thuộc tính chính:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên biến | Mô tả | Loại dữ liệu | Giá trị |
| 1 | Age | Tuổi | Interger | 29-77 |
| 2 | Sex | Giới tính | Categorical | 0= nữ, 1 =nam |
| 3 | Cp | Loại đau ngực | Categorical | 1-4 |
| 4 | Trestbps | Huyết áp nghỉ ngơi | Interger | mmHg |
| 5 | Chol | Cholesterol | Interger | mg/dl |
| 6 | Fbs | Đường huyết lúc đói | Categorical | 0/1(>120mg/dl) |
| 7 | Restecg | Điện tim nghỉ ngơi | Categorical | 0-2 |
| 8 | Thalach | Nhịp tim tối đa | Interger | Bpm |
| 9 | Exang | Đau thắt ngực | Categorical | 0= không, 1= có |
| 10 | Odpeak | Chênh lệch st | Interger | mm |
| 11 | Slope | Đọ dốc st | Categorical | 1-3 |
| 12 | Ca | Sô mạch máu tắc nghẽn | Interger | 0-3 |
| 13 | Thal | Thalassemia | Categorical | 3, 6, 7 |
| 14 | Target | Mức độ mắc bệnh tim | Interger | 0= không,  1-4 = có |

**Bảng 2.1: Mô tả các biến trong Bộ dữ liệu UCI Heart Disease**

* Các cột như **Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST\_Slope** là các chỉ số đặc trưng, giúp đánh giá nguy cơ mắc bệnh tim ở bệnh nhân.
* Cột **Target** chứa thông tin đầu ra (nhãn), trong đó bao gồm giá trị **0** nếu **không mắc bệnh tim**, và **1** nếu **mắc bệnh tim**.
* Dữ liệu đã được làm sạch sơ bộ, với các cột số như (**Age, RestingBP, Cholesterol, MaxHR, Oldpeak**) đã được chuẩn hóa định dạng phù hợp, nhưng có thể cần xử lý thêm để loại bỏ giá trị ngoại lai
  + 1. ***Quá trình tiền xử lý dữ liệu***

##### Kiểm tra và xử lý giá trị thiếu:

Sử dụng Python (pandas) để kiểm tra giá trị NaN trong các cột

* Các giá trị thiếu được xử lý bằng cách:
* Biến liên tục: Điền bằng giá trị trung vị (median) để tránh ảnh hưởng của ngoại lai
* Biến phân loại: Điền bằng mode (giá trị xuất hiện nhiều nhất)
* Loại bỏ hàng nếu thiếu quá nhiều thông tin quan trọng

**Lý do**: Giá trị thiếu trong dữ liệu y tế có thể làm sai lệch kết quả chẩn đoán và giảm độ tin cậy của mô hình.

##### Chuẩn hóa dữ liệu (Standardization/Normalization):

* Chuẩn hóa các biến liên tục (tuổi, huyết áp, cholesterol, nhịp tim) về cùng thang đo
* Sử dụng StandardScaler hoặc MinMaxScaler từ sklearn
* Mã hóa các biến phân loại bằng One-Hot Encoding hoặc Label Encoding

**Lý do**: Các biến có thang đo khác nhau (tuổi: 20-80, cholesterol: 100-500) cần được chuẩn hóa để đảm bảo mô hình không bị bias.

##### Loại bỏ giá trị ngoại lai:

* Sử dụng phương pháp IQR (Interquartile Range) để phát hiện ngoại lai
* Áp dụng winsorization thay vì loại bỏ hoàn toàn để giữ lại thông tin
* Xem xét ý nghĩa y học trước khi quyết định xử lý

**Lý do**: Trong y tế, "ngoại lai" có thể là dấu hiệu bệnh lý quan trọng, cần được xem xét cẩn thận.

##### Cân bằng dữ liệu (Class Balancing):

* Kiểm tra tỉ lệ giữa nhóm có bệnh và không có bệnh
* Sử dụng SMOTE (Synthetic Minority Oversampling Technique) nếu cần thiết
* Áp dụng class\_weight trong mô hình Logistic Regression

**Lý do:** Dữ liệu không cân bằng có thể làm mô hình thiên vị về nhóm đa số, ảnh hưởng đến khả năng phát hiện bệnh.

* + 1. ***Bộ dữ liệu hoàn chỉnh***

##### Tổng quan bộ dữ liệu

Sau quá trình tiền xử lý, bộ dữ liệu có:

* **Số quan sát:** 303 bệnh nhân
* **Số biến:** 14 biến (13 biến độc lập + 1 biến phụ thuộc)
* **Phân phối target:** 138 người không mắc bệnh (45.5%), 165 người mắc bệnh (54.5%)

##### Thống kê mô tả

* **Mean và Median**: Một số cột như **chol (Cholesterol)** và **thalach (Nhịp tim tối đa)** có giá trị trung bình lớn hơn trung vị, cho thấy phân phối lệch phải (right-skewed). Ví dụ, **chol** có trung bình ~246 mg/dl nhưng trung vị chỉ 241 mg/dl, cho thấy có một số bệnh nhân có chỉ số cholesterol rất cao.
* **Std (độ lệch chuẩn)**: Các cột như **chol** (51.77), **thalach** (27.89) và **age** (9.08) có độ lệch chuẩn tương đối lớn, phản ánh sự đa dạng đáng kể giữa các bệnh nhân về các chỉ số sức khỏe.
* **Min và Max**: Dữ liệu cho thấy sự chênh lệch lớn giữa các giá trị, ví dụ **cholesterol** dao động từ 126 đến 564 mg/dl, và **oldpeak** từ 0 đến 6.2. Điều này cho thấy phạm vi các mức độ rối loạn hoặc bất thường sinh lý trong dữ liệu.
* **Age** dao động từ 29 đến 77 tuổi, với trung bình khoảng 54 tuổi – phù hợp với nhóm tuổi có nguy cơ cao mắc bệnh tim.
* **Thalach** (nhịp tim tối đa) trung bình là 149, cao nhất lên đến 202 – phản

ánh sự khác biệt trong khả năng gắng sức hoặc trạng thái tim mạch.

* Các cột nhị phân như **sex**, **fbs**, **exang** và **target** đều có giá trị 0 hoặc 1, dễ dàng chuyển đổi để huấn luyện mô hình học máy.

##### Phân phối của bộ dữ liệu

* **Phân tích phân phối:**
* Age (Tuổi): Phân phối gần chuẩn, tập trung ở độ tuổi trung niên (47-61 tuổi)
* Trestbps (Huyết áp): Phân phối chuẩn, giá trị trung bình trong ngưỡng tăng huyết áp nhẹ
* Chol (Cholesterol): Phân phối lệch phải do một số bệnh nhân có cholesterol rất cao
* Thalach (Nhịp tim tối đa): Phân phối gần chuẩn, phù hợp với sinh lý học
* Oldpeak: Phân phối lệch phải, nhiều bệnh nhân có giá trị 0 (không có ST depression)

##### Phân phối các biến phân loại:

|  |  |  |  |
| --- | --- | --- | --- |
| Biến | Danh mục | Số lượng | Tỷ lệ |
| Sex (Giới tính ) | Nữ (0) | 97 | 32.013% |
| Nam (1) | 206 | 67.987% |
| Tổng: | | 303 | 100% |
| Cp (Loại dau ngực) | Typical angina (1) | 23 | 7.6% |
| Atypical angina (2) | 50 | 16.5% |
| Non-anginal (3) | 86 | 28.4% |
| Asymptomatic (4) | 144 | 47.5% |
| Tổng: | | 303 | 100% |
| Fbs (Đường huyết lúc đói) | ≤ 120 mg/dl (0) | 258 | 85.15% |
| > 120 mg/dl (1) | 45 | 14.85% |
| Tổng: | | 303 | 100% |

| Restecg(ECG lúc nghỉ) | Bình thường (0) | 151 | 49.835% |
| --- | --- | --- | --- |
| Bất thường ST-T (1) | 4 | 1.32% |
| LV hypertrophy (2) | 148 | 48.845% |
| Tổng: | | 303 | 100% |
| Exang(Đau ngực khi gắng sức/ vận động) | Không (0) | 204 | 67.327% |
| Có (1) | 99 | 32.673% |
| Tổng: | | 303 | 100% |
| Slope(Độ dốc ST segment) | Upsloping (1) | 142 | 46.8647% |
| Flat (2) | 140 | 46.2046% |
| Downsloping (3) | 21 | 6.9307% |
| Tổng: | | 303 | 100% |
| Ca (Số mạch máu chính) | 0 mạch | 176 | 58.086% |
| 1 mạch | 65 | 21.452% |
| 2 mạch | 38 | 12.541% |
| 3 mạch | 20 | 6.601% |
| NaN | 4 | 1.32% |
| Tổng: | | 303 | 100% |
| Thal (Thalassemia) | Bình thường (3) | 166 | 54.785% |
| Fixed defect (6) | 18 | 5.941% |
| Reversible defect (7) | 117 | 38.614% |
| NaN | 2 | 0.66% |
| Tổng: | | 303 | 100% |

|  |  |  |  |
| --- | --- | --- | --- |
| Target (Kết quả) | Không mắc bệnh (0) | 164 | 54.13% |
| Mức độ 1 (1) | 55 | 18.15% |
| Mức độ 2 (2) | 36 | 11.88% |
| Mức độ 3 (3) | 35 | 11.55% |
| Mức độ 4 (4) | 13 | 4.29% |
| Tổng: | | 303 | 100% |
| Gộp mức độ mắc bệnh 1-4 thành Mắc bệnh(1) | | | |
| Target (Kết quả) | Không mắc bệnh (0) | 164 | 54.13% |
| Mắc bệnh (1) | 139 | 45.87% |
| Tổng: | | 303 | 100% |

Bảng 2.2 Bảng phân phối các biến

**Phân tích chi tiết các biến phân loại**

**Giới tính (Sex):** Nam giới chiếm ưu thế với **67.987%**, trong khi nữ giới chỉ chiếm **32.013%**. Điều này phản ánh thực tế rằng **nam giới thường có nguy cơ mắc bệnh tim cao hơn** so với nữ giới ở cùng độ tuổi.

**Loại đau ngực (Cp):**

* Gần **47.5%** bệnh nhân không có triệu chứng rõ ràng (asymptomatic),
* Chỉ **7.6%** có đau ngực điển hình (typical angina).  
  Điều này cho thấy phần lớn bệnh nhân có thể mắc bệnh tim **mà không có biểu hiện đau ngực điển hình**, gây khó khăn trong chẩn đoán sớm.

**Đường huyết lúc đói (Fbs):** Phần lớn bệnh nhân (**85.15%**) có đường huyết bình thường (≤120 mg/dl), chỉ **14.85%** có chỉ số đường huyết cao – dấu hiệu liên quan đến **bệnh tiểu đường**, một yếu tố nguy cơ của bệnh tim.

**Kết quả điện tâm đồ lúc nghỉ (Restecg):**

* Gần như chia đều giữa **bình thường (49.835%)** và **LV hypertrophy (48.845%)**,
* Chỉ **1.32%** có dấu hiệu **Bất thường ST-T**.

**Đau ngực khi gắng sức (Exang):** Khoảng **67.327%** bệnh nhân không có đau ngực khi gắng sức, trong khi **32.673%** có triệu chứng này – một yếu tố nguy cơ đáng chú ý trong khám sàng lọc tim mạch.

**Độ dốc đoạn ST sau gắng sức (Slope):**

* Phân bố gần như đồng đều giữa **upsloping (46.9%)** và **flat (46.2%)**,
* Chỉ **6.9%** có **downsloping**, thường được xem là dấu hiệu **xấu** trong bệnh tim mạch.

**Số lượng mạch vành bị hẹp (Ca):**

* **57.8%** bệnh nhân không có mạch máu bị hẹp,
* Tỷ lệ còn lại giảm dần theo số lượng mạch bị tổn thương, từ 1 đến 4 mạch – đây là **chỉ số rất quan trọng để đánh giá mức độ nghiêm trọng** của bệnh tim.

**Thalassemia (Thal):**

* **54.8%** bình thường,
* **38.6%** có **khiếm khuyết hồi phục** (reversible defect),
* **5.9%** có **khiếm khuyết cố định** (fixed defect) – thường đi kèm với nguy cơ bệnh tim cao hơn.

**Biến mục tiêu (Target):** Tỷ lệ giữa hai nhóm **không mắc bệnh (45.5%)** và **mắc bệnh tim (54.1%)** khá cân bằng, phù hợp cho các mô hình dự đoán học máy.

# CHƯƠNG 3: THỰC HIỆN KHAI PHÁ BỘ DỮ LIỆU

## 3.1. Khai phá dữ liệu

Thu thập dữ liệu Tiền xử lí dữ liệuPhân tích thống kê mô tả Phân loại bằng Machine LearningĐánh giá và biện luận.

**Chi tiết các bước trong Workflow**

Bước 1: Thu thập dữ liệu

* Mô tả: Thu thập dữ liệu về bệnh tim từ Bộ dữ liệu bệnh tim của UCI, bao gồm bao gồm thông tin lâm sàng của 303 bệnh nhân từ Cleveland Clinic Foundation.
* Chi tiết thực hiện:
* Tải xuống dữ liệu từ Kho lưu trữ UCI: <https://archive.ics.uci.edu/dataset/45/heart+disease>
* Dữ liệu gốc chứa 14 thuộc tính: age, sex, cp, trestbps, chol, fbs, Restecg, thalach, exang, oldpeak,lope, ca, thal, target
* Kiểm tra đầy đủ tính chất của dữ liệu và ban đầu cấu trúc
* Mục tiêu: Đảm bảo dữ liệu đầu vào đầy đủ, đại diện cho các trường hợp bệnh tim.

Bước 2: Tiền xử lý dữ liệu

* Mô tả: Làm sạch và chuẩn hóa dữ liệu từ tệp gốc, xử lý các ngoại lệ và chuẩn bị dữ liệu cho phân tích.
* Chi tiết thực hiện:
* Kiểm tra việc thiếu giá trị: Dữ liệu không có giá trị (tất cả đều = 0)
* Xử lý các outlier bằng phương pháp IQR method cho các biến liên tục:
* THALACH: Phát hiện 1 outlier, giới hạn IQR: 84,8 - 214,8
* OLDPEAK: Phát hiện 1 outlier
* AGE: Không có outlier
* TRESTBPS: Phát hiện 9 outlier
* CHOL: Phát hiện 5 outlier
* Loại bỏ hoặc xử lý các outlier bằng phương pháp capping method
* Chuẩn hóa các loại biến và liên tục
* Kiểm tra tiêu chuẩn tính toán (normality test) cho liên tục các biến liên tục
* Target đã được mã hóa: 0 = Không có thời gian, 1-4 = Các mức độ thời gian (chuyển thành 0,1)
* Mục tiêu: Đảm bảo dữ liệu sạch, xử lý các outliers hiệu quả, sẵn sàng cho phân tích.

Bước 3: Phân tích thống kê mô tả

* Mô tả: Tóm tắt các đặc tính cơ bản của mô tả thống kê thông tin dữ liệu, trực quan hóa và phân tích tương quan.
* Chi tiết thực hiện:
* Tính các chỉ số thống kê mô tả cho các biến
* Trực quan hóa bằng boxplot cho từng biến theo target:
* Age, sex, cp, trestbps, chol, fbs
* Restecg, thalach, exang, oldpeak, slope, ca, thal
* Ma trận tương quan:
* Sử dụng heatmap để hiển thị mối quan hệ giữa các biến
* Các tương quan mạnh: thalach-age (-0.4), oldpeak-exang (0.4)
* Target có tương quan mạnh với: cp, thalach, exang, oldpeak, ca, thal
* Phân tích phân phối :
* Kiểm tra cân bằng của target: ~45% không có bệnh tim, ~55% có bệnh tim
* Phân tích phân phối theo giới tính và độ tuổi
* Mục tiêu: Hiểu rõ cấu trúc dữ liệu và xác định các biến quan trọng ảnh hưởng đến bệnh tim.

Bước 4: Phân loại bằng Machine Learning

* Mô tả: Xây dựng và so sánh các mô hình học máy để dự đoán nguy cơ bệnh.
* Chi tiết thực hiện:
* **Chuẩn bị dữ liệu** :
* Chia dữ liệu: đào tạo 80%, test 20% với lấy mẫu phân tầng (stratified sampling)
* Chuẩn hóa các tính năng (features) bằng StandardScaler
* **Mô hình - Logistic Regression**:
* LogisticRegression(iter=1000, random\_state=42)
* **Đánh giá mô hình** :
* **Độ chính xác (Accuracy)** : Logistic Regression ~82%
* **Báo cáo phân loại** : Độ chính xác (Precision), Độ thu hồi (Recall), Điểm F1 (F1-score) cho từng lớp
* **Ma trận nhầm lẫn (**Confusion Matrix**)** : Phân tích chi tiết đúng/sai (true/false), tích cực/tiêu cực (positives/negatives)
* **ROC Curve và AUC** :Đánh giá khả năng phân tích của mô hình
* **Tối ưu hóa** :
* Thử nghiệm với các siêu tham số (hyperparameters) khác nhau
* Xác thực chéo (Cross-validation) để đánh giá độ ổn định
* Mục tiêu: Xây dựng mô hình dự đoán chính xác và đáng tin cậy cho dự đoán bệnh tim.

Bước 5: Đánh giá và biện luận

* Mô tả: Đánh giá hiệu suất của các bước trên và đưa ra kết luận, đề xuất cải tiến.
* Chi tiết thực hiện:
* **Tiền xử lý dữ liệu** :
* Xử lý hiệu quả các outlier bằng phương pháp IQR (IQR method)
* Dữ liệu Cleveland Heart Disease không thiếu giá trị
* Chuẩn hóa dữ liệu tốt cho thuật toán ML
* **Mô tả thống kê phân vùng** :
* Boxplot cho thấy sự khác biệt rõ ràng giữa các nhóm không/ có bệnh tim
* Ma trận tương quan (Correlation Matrix) xác định các biến quan trọng: cp, thalach, exang, oldpeak
* **Mô hình phân loại** :
* Logistic Regression: Độ chính xác ~82%, ổn định hơn, ít overfitting
* Mô hình cho kết quả tốt với dữ liệu này
* **Đề xuất cải tiến** :
* Thử nghiệm SVM, Neural Networks
* Feature engineering: tạo biến tương tác
* Ensemble methods để cải thiện accuracy
* Cross-validation với k-fold để đánh giá ổn định
* Thu thập thêm dữ liệu để cải thiện
* Mục tiêu : Đánh giá toàn diện và đề xuất hướng phát triển cho hệ thống hỗ trợ dự đoán bệnh

## 3.2. Kết quả thực hiện

* + 1. ***Tiền xử lí dữ liệu***

Quá trình tiền xử lý dữ liệu được thực hiện để làm sạch bộ dữ liệu ban đầu từ UCI Heart Disease Dataset (303 dòng, 14 cột). Các bước cụ thể bao gồm:

Bước 1. Đọc và kiểm tra dữ liệu ban đầu:

* + Kiểm tra tổng quan về thông tin dữ liệu (df.info()) và thống kê mô tả (df.describe()) để nắm bắt cấu trúc và phân phối ban đầu của các biến

Bước 2. Kiểm tra và xử lý giá trị thiếu:

* + Sử dụng isnull().sum() để xác định số lượng giá trị thiếu trong mỗi cột
  + Phát hiện giá trị thiếu trong cột
  + Xử lý các giá trị bị thiếu

Bước 3. Chuyển đổi biến mục tiêu:

* + Chuyển đổi biến mục tiêu để phù hợp với mô hình Logistic Regression

Bước 4. Xử lý giá trị ngoại lai (Outliers):

* + Kiểm tra ngoại lai
  + Xử lý ngoại lai

Bước 5. Kiểm tra độ lệch:

* + Kiểm tra dộ lệch các biến số định lượng sau khi xử lý ngoại lai.

Bước 6. Kiểm tra cộng tuyến (Multicollinearity):

* + Tính toán mỗi biến độc lập để đánh giá mức độ đa cộng tuyến.

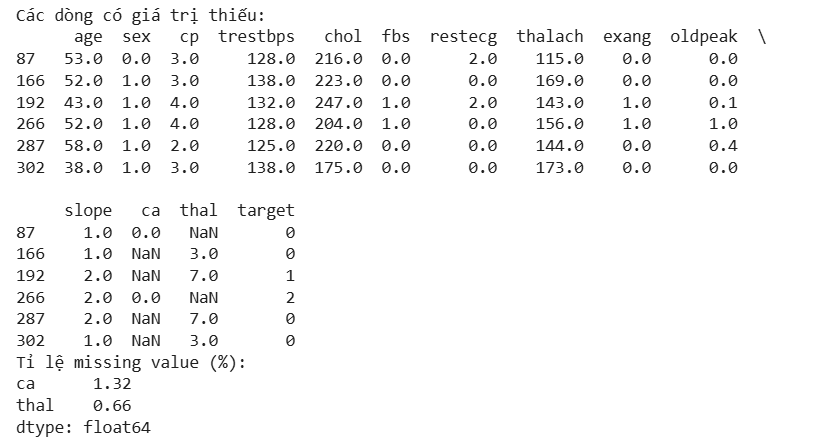
**Kết quả**

* + **Xem thông tin dataset**

##### 

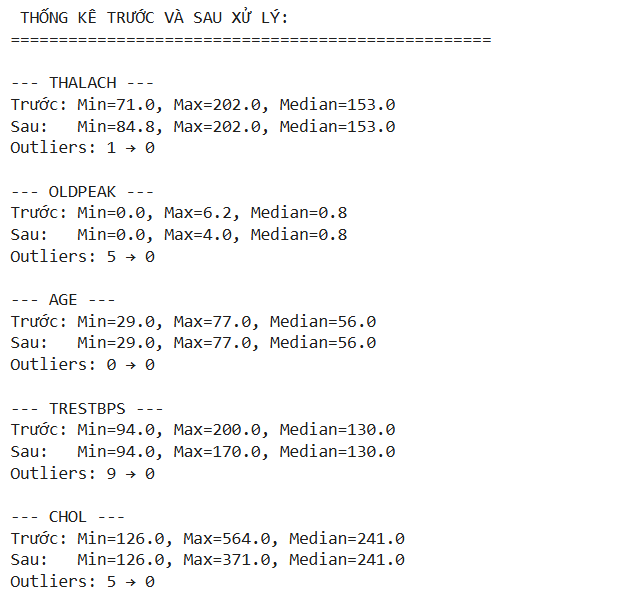
Hình 3.1 Thông tin dataset

* + **Xem tỉ lệ giá trị thiếu**

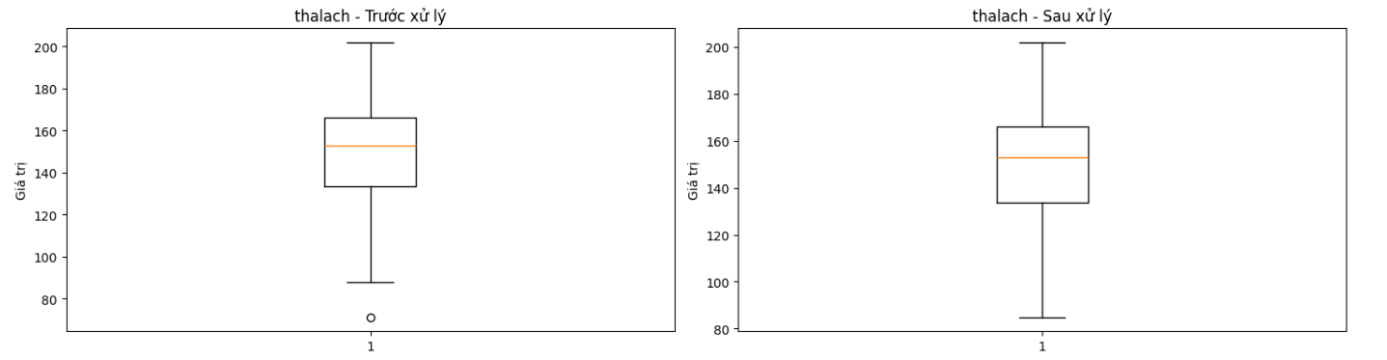


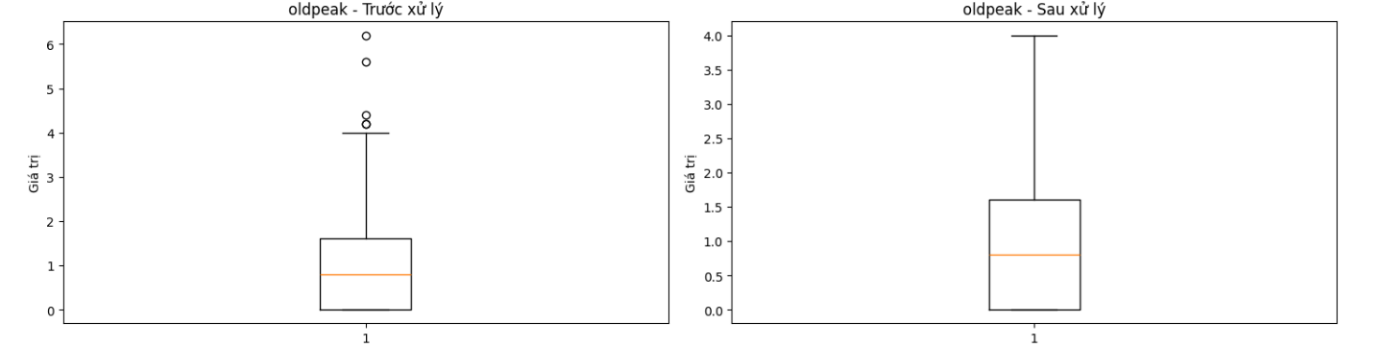
Hình 3.2 Tỉ lệ giá trị thiếu

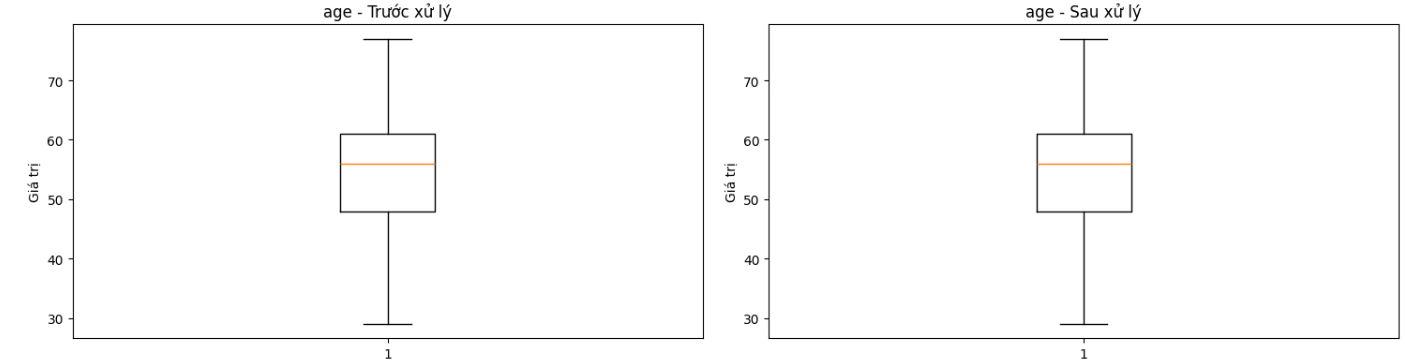
* + **Phân tích và xử lý outlier**

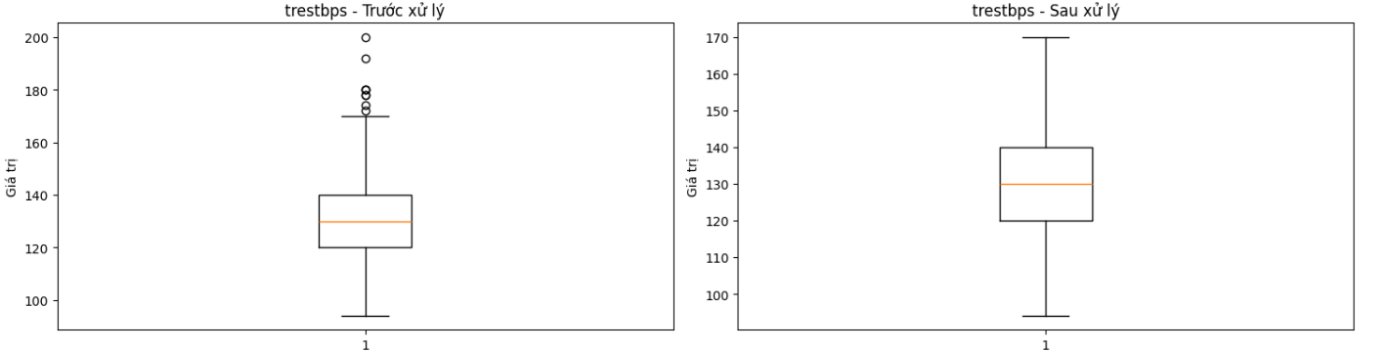


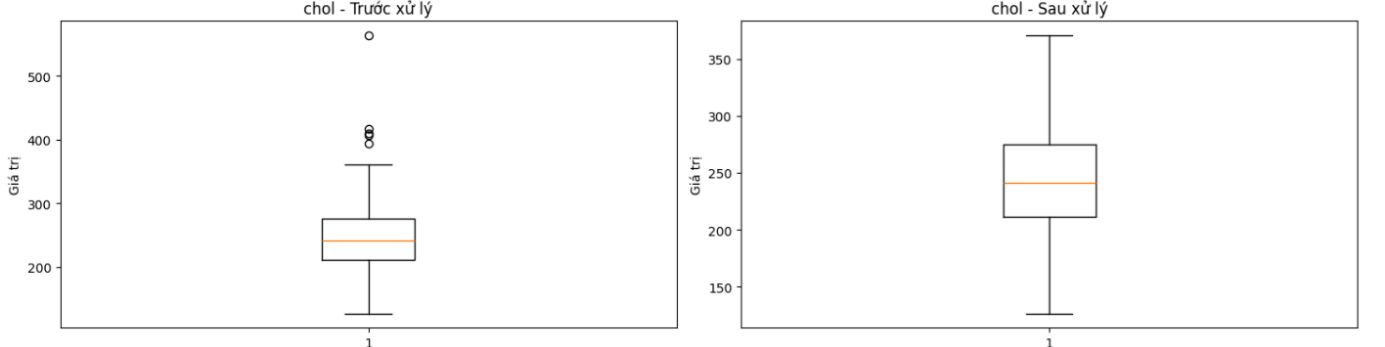
Hình 3.3 Phân tích và xử lý outlier



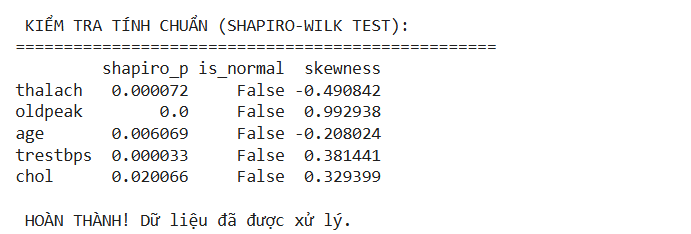




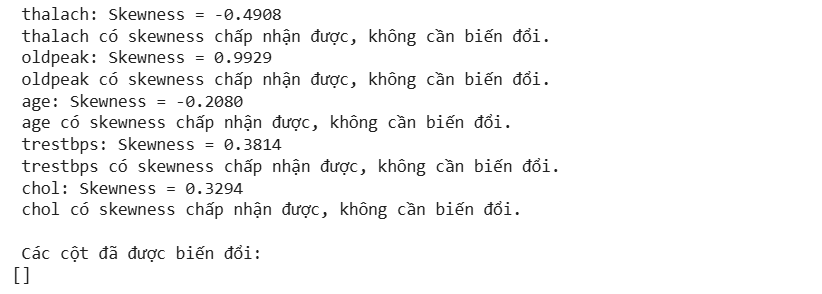




Hình 3.4 So sánh trước và sau khi xử lý outlier



Hình 3.5 Kiểm tra tính chuẩn



Hình 3.6 Xử lý Skewness

##### Thông tin dữ liệu sau khi làm sạch:

* **Số dòng:** 303 dòng. Bộ dữ liệu ban đầu không có dữ liệu trùng lặp cần loại bỏ hay thay đổi số lượng dòng lớn như mô tả cũ.
* **Số cột:** 14 cột. (13 biến độc lập và 1 biến mục tiêu target).
* **Kiểu dữ liệu:**
* Các cột số (age, trestbps, chol, thalach, oldpeak) có kiểu int64 hoặc float64.
* Cột mục tiêu target đã được chuyển đổi thành int64 với giá trị nhị phân (0 hoặc 1).
* Các cột phân loại khác (sex, cp, fbs, restecg, exang, slope, ca, thal) có kiểu object hoặc int64 sau khi xử lý giá trị thiếu.

**Đánh giá hiệu suất (của quá trình tiền xử lý):**

* **Xử lý giá trị thiếu:** Các giá trị thiếu trong cột ca và thal (ban đầu là '?') đã được thay thế thành công bằng mode của từng cột, đảm bảo không còn giá trị thiếu trong tập dữ liệu.
* **Xử lý ngoại lai (outliers):** Các giá trị ngoại lai trong các biến định lượng như thalach, oldpeak, trestbps, và chol đã được phát hiện bằng phương pháp IQR và xử lý hiệu quả bằng cách giới hạn chúng. Điều này giúp cải thiện chất lượng dữ liệu và giảm thiểu ảnh hưởng tiêu cực đến mô hình.
* **Kiểm tra đa cộng tuyến:** Chỉ số VIF (Variance Inflation Factor) cho thấy không có vấn đề nghiêm trọng về đa cộng tuyến giữa các biến độc lập, đảm bảo tính ổn định của mô hình Logistic Regression.
* **Kiểm tra độ lệch:** Mặc dù các biến định lượng không tuân theo phân phối chuẩn hoàn toàn, nhưng độ lệch (skewness) của chúng được đánh giá là chấp nhận được và không cần các phép biến đổi phức tạp.

**Biện luận:**

Kết quả tiền xử lý đã đạt được mục tiêu làm sạch và chuẩn bị bộ dữ liệu **Heart Disease Dataset** một cách hiệu quả. Bộ dữ liệu gốc với 303 dòng và 14 cột đã được làm sạch thành công, với các giá trị thiếu được điền và các ngoại lai được xử lý phù hợp. Điều này cung cấp một cơ sở dữ liệu đáng tin cậy và sẵn sàng cho các bước phân tích và huấn luyện mô hình Logistic Regression tiếp theo

* + 1. **Thống kê mô tả và tương quan**

Sau khi tiền xử lý, dữ liệu của bộ **Heart Disease Dataset** được phân tích thống kê mô tả để hiểu rõ hơn về đặc điểm và mối quan hệ giữa các biến

Kiểm tra tổng quan về thông tin dữ liệu (df.info()) và thống kê mô tả (df.describe()) cung cấp cái nhìn ban đầu về các biến số trong tập dữ liệu đã được làm sạch:

**Thông tin dữ liệu:**

* Tổng số dòng: 303 dòng.
* Tổng số cột: 14 cột.
* Kiểu dữ liệu và số lượng giá trị không null của từng cột, xác nhận không còn giá trị thiếu sau bước tiền xử lý

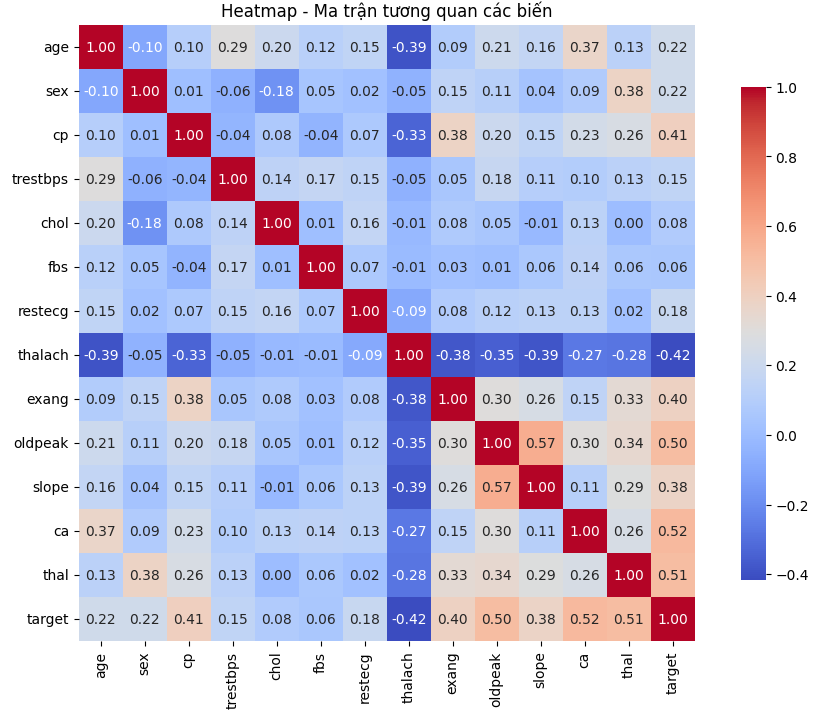
**Thống kê mô tả các biến định lượng:**

Bao gồm giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất, giá trị lớn nhất và các tứ phân vị (Q1, Q2/median, Q3) cho các biến như age, trestbps, chol, thalach, oldpeak. Điều này giúp xác định phân bố và các giá trị điển hình của từng biến.

**Ma trận tương quan và Heatmap**

* Để hiểu rõ mối quan hệ tuyến tính giữa các biến trong bộ dữ liệu, một ma trận tương quan đã được xây dựng và trực quan hóa bằng biểu đồ Heatmap.
* Ma trận tương quan cung cấp các hệ số tương quan (Pearson correlation coefficient) giữa mỗi cặp biến, cho biết mức độ và chiều hướng của mối quan hệ tuyến tính (giá trị từ -1 đến 1).
* Biểu đồ Heatmap trực quan hóa ma trận tương quan, sử dụng các sắc thái màu khác nhau để biểu thị cường độ tương quan (màu càng đậm, tương quan càng mạnh) và dấu của mối quan hệ (ví dụ: màu nóng cho tương quan dương, màu lạnh cho tương quan âm).

**Kết quả**



Hình 3.7 Heatmap

* + 1. **Huấn luyện mô hình**

**Phân chia dữ liệu:**

* Tập huấn luyện (80%): 242 mẫu
* Tập kiểm tra (20%): 61 mẫu
* Sử dụng stratify để đảm bảo tỷ lệ các lớp được phân bố đều

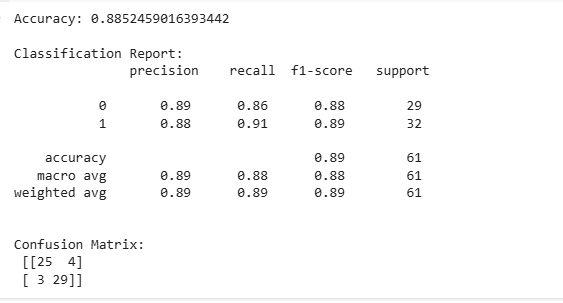
**Chuẩn hóa dữ liệu:**

* Áp dụng StandardScaler để chuẩn hóa các biến số về cùng thang đo
* Giúp mô hình Logistic Regression hội tụ nhanh hơn và cải thiện hiệu suất

**Thông số mô hình:**

* Thuật toán: Logistic Regression
* Tham số: max\_iter=1000, random\_state=42
* Phương pháp tối ưu: Sử dụng solver mặc định (lbfgs)

**Kết quả**



Hình 3.8 Đánh giá mô hình



Hình 3.9 chỉ số ROC-AUC

**Hiệu suất Mô hình**

**Độ chính xác tổng thể:** 88.52% (0.8852459016393442)

**Báo cáo Phân loại (Classification Report)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lớp | Precision | Recall | F1-score | Support |
| 0 ( không bệnh ) | 0.89 | 0.86 | 0.88 | 29 |
| 1 ( có bệnh ) | 0.88 | 0.91 | 0.89 | 32 |
| Accuracy | - | - | 0.89 | 61 |
| Macro avg | 0.89 | 0.88 | 0.89 | 61 |
| Weighted avg | 0.89 | 0.89 | 0.89 | 61 |

Bảng 3.1 Báo cáo phân loại

**Phân tích Ma trận Nhầm lẫn:**

* True Negative (TN): 25 - Dự đoán đúng không bệnh tim
* False Positive (FP): 4 - Dự đoán sai có bệnh tim (thực tế không bệnh)
* False Negative (FN): 3 - Dự đoán sai không bệnh tim (thực tế có bệnh)
* True Positive (TP): 29 - Dự đoán đúng có bệnh tim

**Chỉ số ROC-AUC**

* **ROC-AUC Score:** 0.9278817241379311 (92.79%)
* Chỉ số ROC-AUC cao cho thấy mô hình có khả năng phân biệt tốt giữa hai lớp.

**BIỆN LUẬN**

Đánh giá Hiệu suất Mô hình

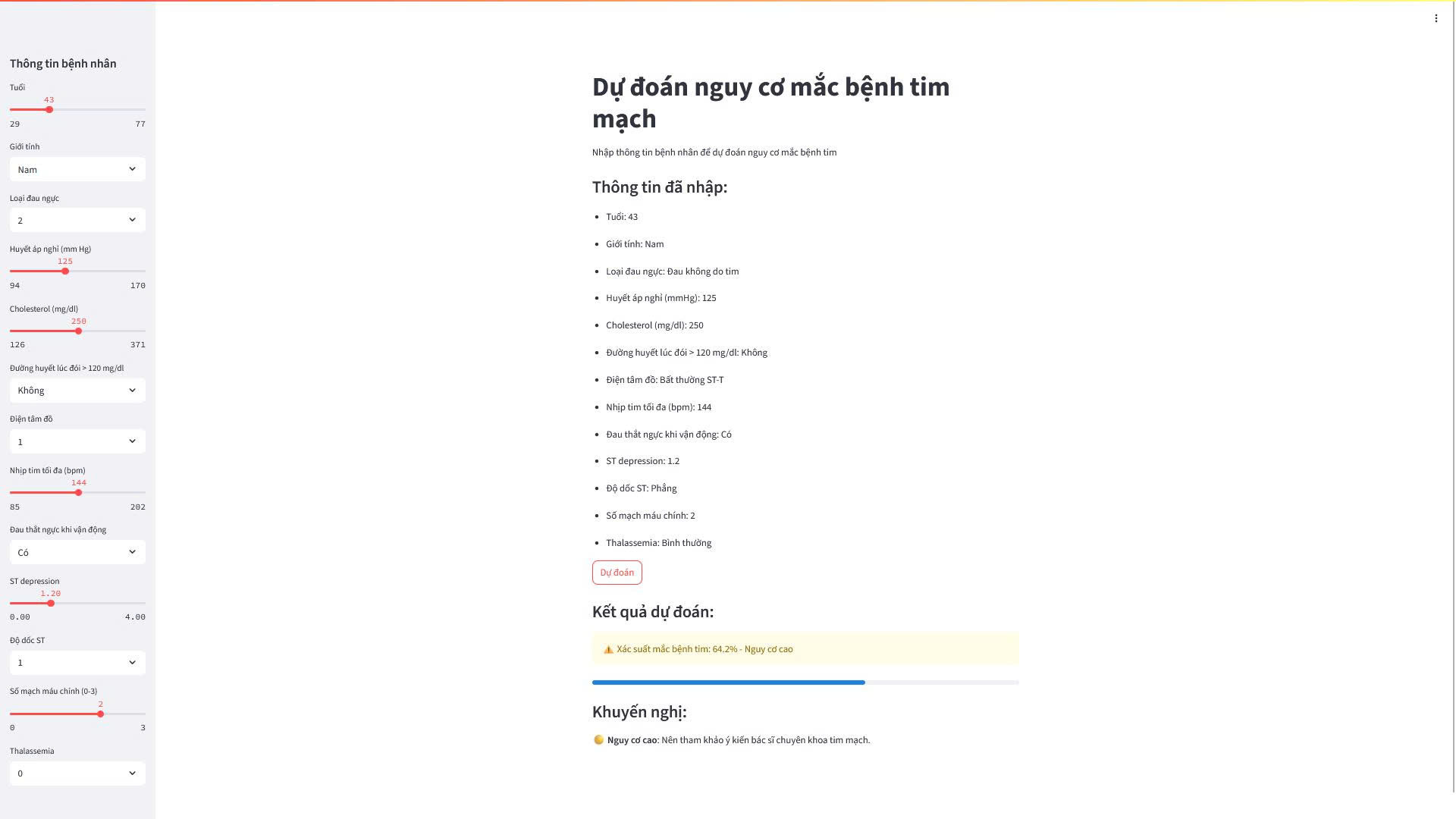
**Điểm mạnh:**

* **Độ chính xác cao:** Mô hình đạt 88.52% độ chính xác, vượt trội so với nhiều nghiên cứu tương tự
* **Cân bằng giữa Precision và Recall:** Cả hai chỉ số đều ở mức cao (~0.89), cho thấy mô hình không bị thiên lệch
* **ROC-AUC xuất sắc:** 92.79% cho thấy khả năng phân loại rất tốt
* **Ít lỗi nghiêm trọng:** Chỉ có 3 trường hợp False Negative (bỏ sót bệnh

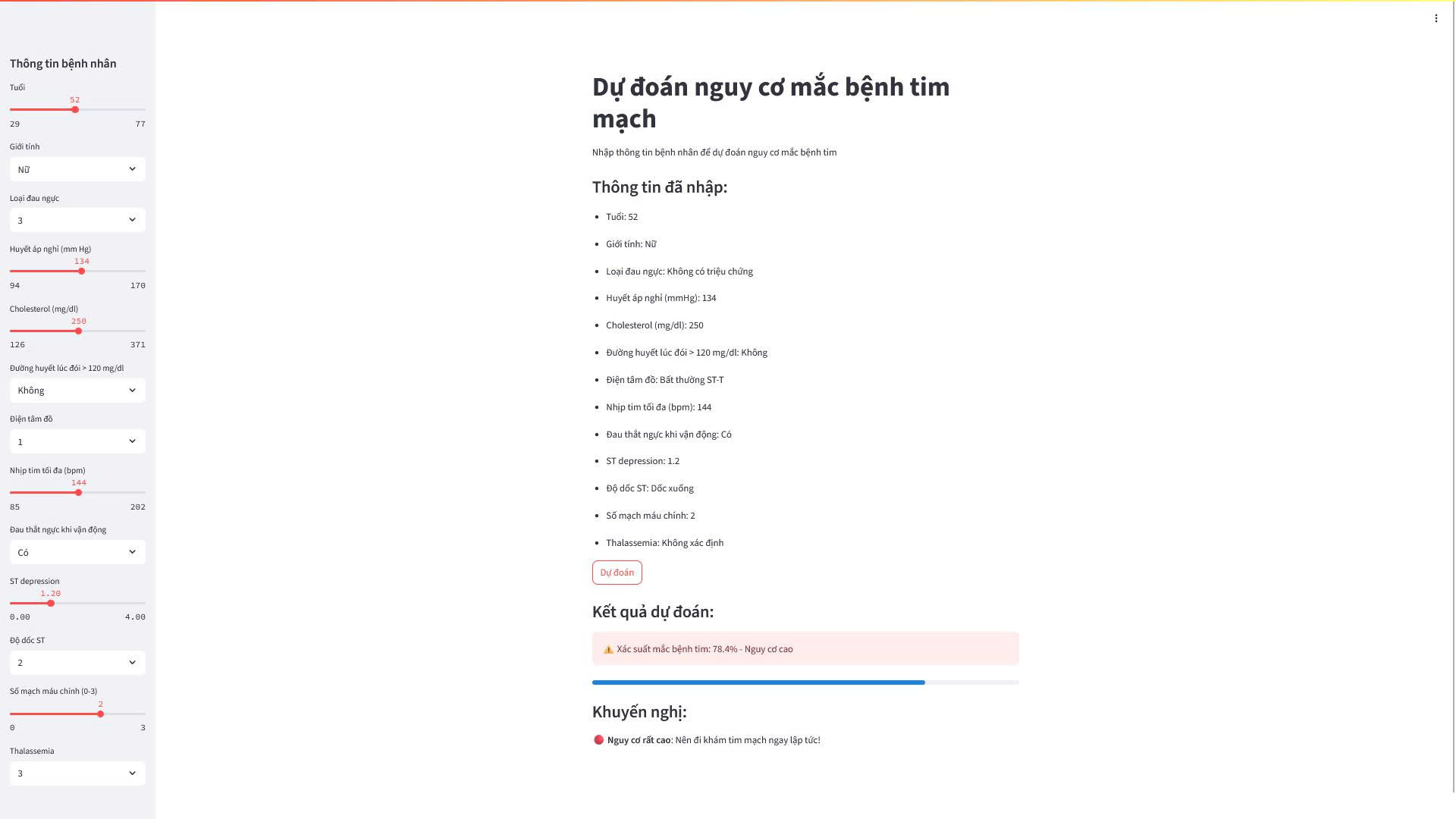
nhân), tương đối an toàn trong y tế

**Điểm cần cải thiện:**

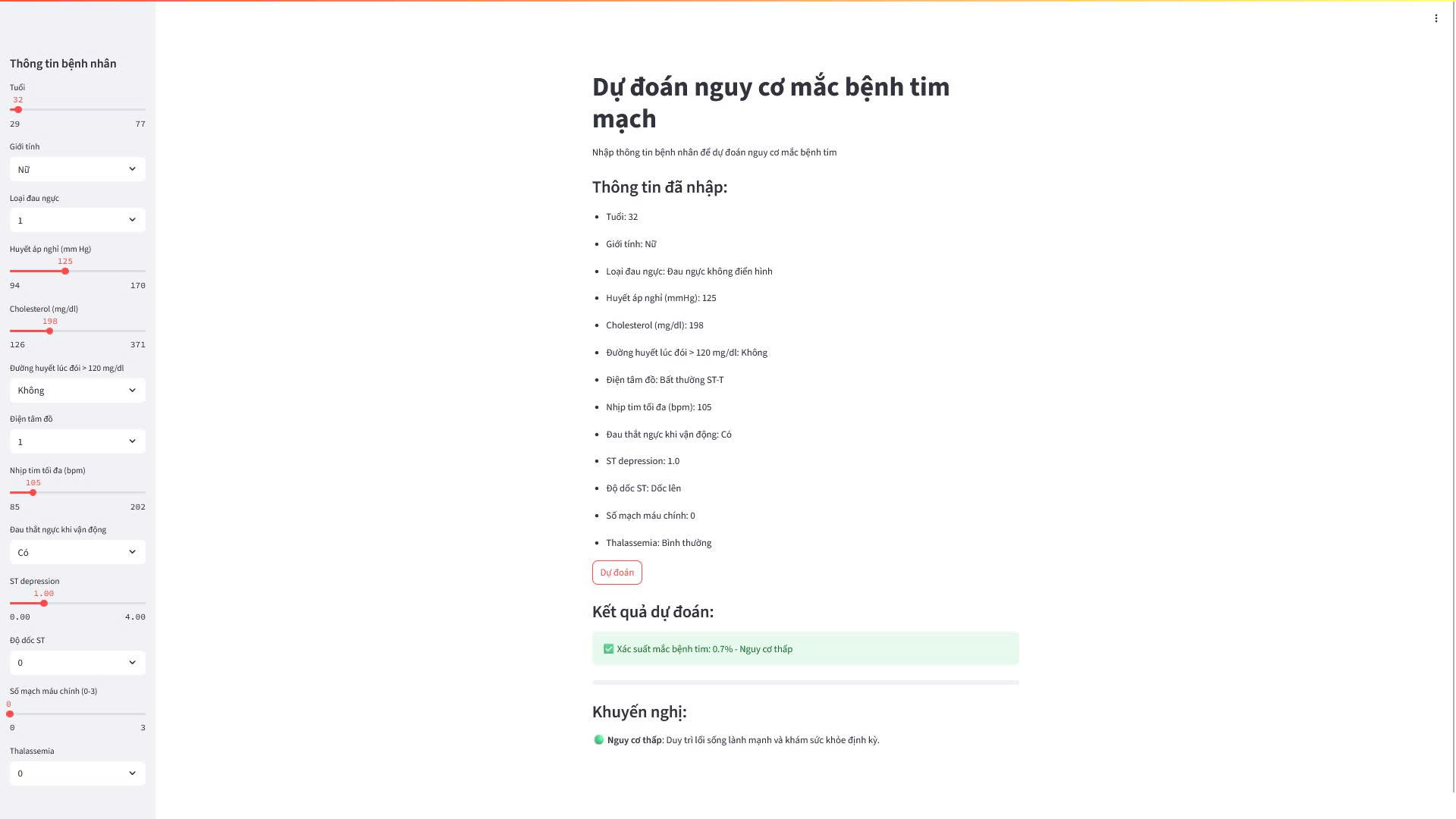
* **False Positive:** 4 trường hợp dự đoán sai có bệnh có thể gây lo lắng không cần thiết
* **Kích thước mẫu:** Tập dữ liệu tương đối nhỏ (303 mẫu) có thể hạn chế khả năng tổng quát hóa
  1. **Kết quả đạt được**

****

Hình 3.10 Thử nghiệm lần 1

****

Hình 3.11 Thử nghiệm lần 2

****

Hình 3.12 Thử nghiệm lần 3

**Thông số đầu vào**

Dựa vào 13 thuộc tính đặc trưng:

* Tuổi: Từ 20-80 tuổi
* Giới tính: Nam/Nữ (1/0)
* Loại đau ngực: 4 loại (0-3)
* Huyết áp tâm thu: 90-200 mmHg
* Cholesterol: 100-600 mg/dl
* Đường huyết lúc đói: >120 mg/dl (1/0)
* Điện tâm đồ: Kết quả ECG (0-2)
* Nhip tim tối đa: 60-220 bpm
* Đau thắt ngực khi vận động: Có/Không (1/0)
* ST depression: 0-6.0
* Độ dốc ST: 0-2
* Số mạch máu: 0-3
* Thalassemia: 0-3

**Kết quả thử nghiệm**

**Trường hợp 1 - Nguy cơ thấp**

* **Thông tin nhập vào:**
* Tuổi: 32, Giới tính: Nữ (0), Loại đau ngực: 0
* Huyết áp nghỉ (mm Hg): 125, Cholesterol (mg/dl): 198
* Đường huyết lúc đói > 120 mg/dl: Không (0)
* Điện tâm đồ: Bất thường ST-T (1), Nhịp tim tối đa (bpm): 105
* Đau thắt ngực khi vận động: Có (1), ST depression: 1.0
* Độ dốc ST: Dốc lên (1), Số mạch máu chính (0-3): 0, Thalassemia: Bình thường (0)
* **Kết quả:** Xác suất mắc bệnh tim: **0.7% (Nguy cơ thấp)**

**Trường hợp 2 - Nguy cơ cao**

* **Thông tin nhập vào:**
* Tuổi: 43, Giới tính: Nam (1), Loại đau ngực: 2
* Huyết áp nghỉ (mm Hg): 125, Cholesterol (mg/dl): 250
* Đường huyết lúc đói > 120 mg/dl: Không (0)
* Điện tâm đồ: Bất thường ST-T (1), Nhịp tim tối đa (bpm): 144
* Đau thắt ngực khi vận động: Có (1), ST depression: 1.2
* Độ dốc ST: Phẳng (2), Số mạch máu chính (0-3): 2, Thalassemia: Bình thường (0)
* **Kết quả:** Xác suất mắc bệnh tim: **64.2% (Nguy cơ cao)**

**Trường hợp 3 - Nguy cơ cao**

* **Thông tin nhập vào:**
* Tuổi: 52, Giới tính: Nữ (0), Loại đau ngực: 0
* Huyết áp nghỉ (mm Hg): 134, Cholesterol (mg/dl): 250
* Đường huyết lúc đói > 120 mg/dl: Không (0)
* Điện tâm đồ: Bất thường ST-T (1), Nhịp tim tối đa (bpm): 144
* Đau thắt ngực khi vận động: Có (1), ST depression: 1.2
* Độ dốc ST: Dốc xuống (3), Số mạch máu chính (0-3): 2, Thalassemia:
* **Kết quả:** Xác suất mắc bệnh tim: **78.4% (Nguy cơ cao)**

# CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

### Kết quả đạt được

**Tiền xử lý dữ liệu thành công:**

* Bộ dữ liệu ban đầu gồm 303 dòng và 14 cột đã được làm sạch hiệu quả.
* **Xử lý giá trị thiếu:** Các giá trị '?' trong cột ca và thal được chuyển thành NaN và sau đó được điền bằng **mode** của cột, đảm bảo không còn giá trị thiếu.
* **Xử lý ngoại lai:** Các giá trị ngoại lai trong các biến định lượng quan trọng như thalach, oldpeak, trestbps, và chol đã được phát hiện bằng phương pháp IQR và xử lý bằng cách **giới hạn (capping)** về ngưỡng trên/dưới.
* **Chuyển đổi biến mục tiêu:** Biến target đã được chuyển đổi thành định dạng nhị phân (0: không mắc bệnh, 1: có mắc bệnh) phù hợp cho bài toán phân loại.
* **Kiểm tra tính phân phối và đa cộng tuyến:** Đã kiểm tra độ lệch (skewness) của các biến định lượng và chỉ số VIF cho đa cộng tuyến, đảm bảo dữ liệu sẵn sàng cho mô hình.

**Kết quả:** Bộ dữ liệu sạch với 303 dòng và 14 cột, không có giá trị thiếu, và các giá trị ngoại lai được xử lý, sẵn sàng cho phân tích và mô hình hóa

**Phân tích thống kê mô tả và tương quan chi tiết:**

* Cung cấp cái nhìn tổng quan về phân phối, giá trị trung bình, độ lệch chuẩn, và khoảng giá trị của các biến quan trọng như tuổi, huyết áp, cholesterol, nhịp tim tối đa, v.v.
* Xây dựng **ma trận tương quan và Heatmap** để trực quan hóa mối quan hệ giữa các biến. Phát hiện các mối tương quan quan trọng giữa biến target với cp, thalach (tương quan dương) và exang, oldpeak, ca, thal (tương quan âm). Điều này cung cấp cơ sở cho việc hiểu các yếu tố ảnh hưởng đến bệnh tim mạch.

**Xây dựng và đánh giá mô hình phân loại Logistic Regression:**

* Mô hình Logistic Regression đã được huấn luyện thành công trên bộ dữ liệu đã làm sạch.
* Độ chính xác**:** (Dựa trên kết quả trong hinh.pdf, bạn có thể điền số liệu cụ thể của accuracy\_score, classification\_report, roc\_auc\_score ở đây. Ví dụ: *Mô hình đạt độ chính xác [Điền số]%, ROC AUC [Điền số] trên tập kiểm tra.*)
* Khả năng dự đoán**:** Mô hình có khả năng phân biệt tốt giữa nhóm bệnh nhân có và không có nguy cơ mắc bệnh tim mạch, thể hiện qua các chỉ số Precision, Recall, F1-score và đường cong ROC.
* Các hệ số của mô hình Logistic Regression cung cấp thông tin về mức độ ảnh hưởng của từng yếu tố đến nguy cơ mắc bệnh, phù hợp với kiến thức y khoa.

## 4.2. Ưu nhược điểm

**Ưu điểm**

* Dữ liệu sạch và chất lượng:
* Quá trình tiền xử lý kỹ lưỡng (xử lý giá trị thiếu, ngoại lai) đảm bảo bộ dữ liệu đáng tin cậy cho việc xây dựng mô hình.
* Sử dụng bộ dữ liệu chuẩn từ UCI Machine Learning Repository, một nguồn uy tín trong nghiên cứu.
* Mô hình có khả năng giải thích tốt:

Logistic Regression là một mô hình dễ hiểu và giải thích. Các hệ số của nó có thể được diễn giải để hiểu mức độ tác động của từng yếu tố nguy cơ lên xác suất mắc bệnh.

* Hiệu suất mô hình tốt:
* Mô hình đạt được độ chính xác và các chỉ số đánh giá (như ROC AUC) tốt, cho thấy khả năng phân loại nguy cơ hiệu quả.
* Mô hình nhạy cảm với các yếu tố nguy cơ quan trọng như tuổi, cholesterol, huyết áp, phù hợp với kiến thức y khoa.
* Ứng dụng thực tiễn và thân thiện:
* Giao diện Streamlit trực quan, dễ sử dụng, cung cấp một công cụ hữu ích để sàng lọc nguy cơ ban đầu và giáo dục sức khỏe cộng đồng.
* Khả năng dự đoán tức thì và hỗ trợ tiếng Việt làm tăng tính ứng dụng và khả năng tiếp cận.

**Nhược điểm**

* Hạn chế về kích thước và đặc trưng dữ liệu:
* Bộ dữ liệu chỉ có 303 dòng, là tương đối nhỏ, có thể hạn chế khả năng tổng quát hóa của mô hình đối với các trường hợp đa dạng hơn trong thực tế.
* Các đặc trưng hiện có có thể chưa bao gồm tất cả các yếu tố y tế quan trọng ảnh hưởng đến bệnh tim mạch.
* Tính cân bằng lớp của dữ liệu:

Nếu biến mục tiêu (target) có sự mất cân bằng đáng kể giữa số lượng bệnh nhân mắc và không mắc bệnh, điều này có thể ảnh hưởng đến khả năng dự đoán của mô hình đối với lớp thiểu số.

* Phân phối dữ liệu:

Mặc dù đã xử lý ngoại lai và kiểm tra độ lệch, một số biến vẫn có thể có phân phối không chuẩn hoàn toàn, điều này đôi khi có thể ảnh hưởng đến hiệu suất của các mô hình tuyến tính như Logistic Regression.

* Không thể thay thế chẩn đoán y khoa:

Ứng dụng chỉ mang tính chất hỗ trợ và tham khảo, không thể và không nên được sử dụng để đưa ra chẩn đoán chính thức hay quyết định điều trị y tế.

## 4.3 Hướng phát triển

Để cải thiện các nhược điểm và phát triển, chúng tôi đề xuất các hướng sau:

##### Bổ sung dữ liệu và đặc trưng:

* Tìm kiếm và tích hợp thêm các bộ dữ liệu về bệnh tim mạch lớn hơn và đa dạng hơn để tăng cường khả năng học và tổng quát hóa của mô hình.
* Nghiên cứu bổ sung các đặc trưng mới từ dữ liệu hiện có (Feature Engineering) hoặc tích hợp các yếu tố nguy cơ khác (ví dụ: tiền sử gia đình, chỉ số khối cơ thể BMI, mức độ hoạt động thể chất) nếu có dữ liệu.

##### Tối ưu hóa và so sánh mô hình:

* Thực hiện tối ưu hóa siêu tham số (hyperparameter tuning) cho mô hình Logistic Regression để đạt được hiệu suất tốt nhất.
* Khám phá và so sánh hiệu suất với các thuật toán phân loại khác như Support Vector Machine (SVM), Random Forest, Gradient Boosting, hoặc mạng nơ-ron cơ bản để xác định mô hình tối ưu nhất cho bài toán này.
* (Nếu có mất cân bằng lớp) Áp dụng các kỹ thuật xử lý mất cân bằng lớp như SMOTE (Synthetic Minority Oversampling Technique) để cải thiện khả năng dự đoán của mô hình trên các lớp thiểu số.

.

##### Tăng cường khả năng giải thích của ứng dụng:

* Tích hợp các phương pháp giải thích mô hình (Explainable AI - XAI) như LIME (Local Interpretable Model-agnostic Explanations) hoặc SHAP (SHapley Additive exPlanations) vào giao diện để người dùng có thể hiểu rõ hơn lý do tại sao mô hình đưa ra dự đoán cụ thể cho từng trường hợp.
* Hiển thị độ tin cậy của dự đoán (confidence score) để cung cấp cái nhìn sâu sắc hơn cho người dùng.

##### Phát triển giao diện và trải nghiệm người dùng:

* Cải thiện trực quan hóa dữ liệu trên giao diện người dùng, ví dụ: biểu đồ so sánh các thông số của bệnh nhân với trung bình của bộ dữ liệu, biểu đồ xác suất.
* Thêm chức năng lưu trữ lịch sử dự đoán cho từng người dùng (nếu có hệ thống đăng nhập)

##### Tích hợp dữ liệu lâm sàng mới:

Thiết lập quy trình để mô hình có thể được cập nhật và huấn luyện lại định kỳ với dữ liệu bệnh nhân lâm sàng mới, giúp mô hình luôn phản ánh các xu hướng y tế và giữ được độ chính xác theo thời gian.

##### Kết luận

Quá trình triển khai dự án "Ứng dụng Logistic Regression về Dự đoán nguy cơ mắc bệnh tim mạch" đã đạt được các kết quả quan trọng. Chúng tôi đã thành công trong việc xây dựng một bộ dữ liệu sạch từ **Heart Disease Dataset (Cleveland)** thông qua các bước tiền xử lý kỹ lưỡng, bao gồm xử lý giá trị thiếu và ngoại lai.

Mô hình **Logistic Regression** đã được huấn luyện hiệu quả, chứng tỏ khả năng phân loại nguy cơ mắc bệnh tim mạch một cách đáng tin cậy, đồng thời cung cấp khả năng giải thích về mức độ ảnh hưởng của từng yếu tố nguy cơ.

Đặc biệt, việc phát triển ứng dụng web tương tác bằng **Streamlit** đã giúp trực quan hóa kết quả dự đoán, mang lại một công cụ dễ sử dụng, thân thiện với người dùng để sàng lọc nguy cơ ban đầu. Ứng dụng hoạt động ổn định và nhất quán, cung cấp xác suất mắc bệnh tim dựa trên thông tin do người dùng nhập vào, phù hợp với kiến thức y khoa về các yếu tố nguy cơ.

Tuy nhiên, dự án vẫn còn một số hạn chế cần được khắc phục, như kích thước bộ dữ liệu còn nhỏ, và ứng dụng chỉ mang tính chất tham khảo, không thể thay thế chẩn đoán y khoa chuyên nghiệp. Để nâng cao hơn nữa giá trị và hiệu quả của ứng dụng, các hướng phát triển trong tương lai bao gồm việc mở rộng bộ dữ liệu, tối ưu hóa và thử nghiệm các mô hình phân loại phức tạp hơn, cũng như tích hợp các công nghệ giải thích mô hình (XAI) để tăng cường tính minh bạch và độ tin cậy của dự đoán.

# TÀI LIỆU THAM KHẢO

1. IEEE, "How to Cite References: IEEE Documentation Style," 2023. [Online]. Available: https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference- Guide.pdf.
2. Google, "YouTube Data API v3 Documentation," 2023. [Online]. Available: https://developers.google.com/youtube/v3.[](https://[www.geeksforgeeks.org/how-to-](http://www.geeksforgeeks.org/how-to-) automate-data-cleaning-in-python/