**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI: *Bài toán chuẩn đoán Bệnh tim***

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | : Vũ Trọng Sinh |
| **Lớp** | : 232IS54A02 |
| **Nhóm** | : 1 |

**Hà Nội – 2024**

**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI: *Bài toán chuẩn đoán Bệnh tim***

|  |  |  |  |
| --- | --- | --- | --- |
| **Thành viên** | **Mã sinh viên** | **Tỉ lệ đóng góp** | **Xác nhận** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Trần A Khang | 24A4040141 |  |  |

**Hà Nội – 2024**

# LỜI MỞ ĐẦU

Trong thế kỷ 21, trí tuệ nhân tạo đã trở thành một phần không thể thiếu trong các cuộc sống hiện đại, giúp cải thiện chất lượng cuộc sống và giải quyết nhiều vấn đề phức tạp. Môn học "Trí tuệ nhân tạo" là một phần quan trọng của chương trình đào tạo, đưa sinh viên vào thế giới phức tạp của công nghệ và quản lý thông tin. Trong sự phát triển không ngừng của trí tuệ nhân tạo, nhóm chúng em đã lựa chọn đề tài "Tìm hiểu và mô hình hóa bài toán chẩn đoán bệnh tim". Đây là một trong những vấn đề quan trọng trong ngành y tế, ảnh hưởng đến hàng triệu người trên thế giới. Chẩn đoán bệnh tim không chỉ phức tạp mà còn phản ánh một sự kết hợp hài hòa giữa các hoạt động y học, quản lý thông tin, và kỹ thuật trí tuệ nhân tạo.

Em xin gửi lời cảm ơn chân thành và sâu sắc đến các giảng viên đã dành thời gian và công sức để truyền đạt những kiến thức quan trọng cho chúng em, giúp chúng em tích lũy đủ kiến thức và kỹ năng để hoàn thành bài tập lớn này. Đặc biệt, chúng em muốn bày tỏ lòng biết ơn đến thầy Vũ Trọng Sinh, người đã hỗ trợ chúng em trong từng bước của dự án. Thầy không chỉ là người hướng dẫn tận tâm, mà còn giúp đỡ và hỗ trợ bọn em trong suốt quá trình làm việc. Trong quá trình học và làm việc cùng thầy, chúng em đã học được thái độ nghiêm túc trong công việc, cũng như sự phối hợp hiệu quả để hoàn thành công việc được tốt nhất.

Mặc dù chúng em đã cố gắng hoàn thiện bài tập trong phạm vi và khả năng của mình, tuy vậy bài của bọn em vẫn có thể sẽ xảy ra sai sót hoặc đôi khi diễn đạt chưa đủ ý, mong thầy có thể chỉ ra những thiếu sót để em có thể hoàn thiện và cải thiện dự án của mình.

# MỤC LỤC

[LỜI MỞ ĐẦU 3](#_Toc169046265)

[MỤC LỤC 4](#_Toc169046266)

[CHƯƠNG I: TỔNG QUAN ĐỀ TÀI 5](#_Toc169046267)

[1.1. Giới thiệu tổng quan dự án 5](#_Toc169046268)

[1.1.1. Giới thiệu dự án 5](#_Toc169046269)

[1.1.2. Phạm vi thực hiện dự án 5](#_Toc169046270)

[1.1.3. Ứng dụng 6](#_Toc169046271)

[1.2. Giới thiệu bài toán 6](#_Toc169046272)

[1.2.1. Phát biểu bài toán 6](#_Toc169046273)

[1.2.2 Giới thiệu về bộ dữ liệu 6](#_Toc169046274)

[1.2.3. Ứng dụng 8](#_Toc169046275)

[CHƯƠNG II: XÂY DỰNG BÀI TOÁN 9](#_Toc169046276)

[2.1. Thu thập dữ liệu 9](#_Toc169046277)

[2.1.1. Nguồn dữ liệu 9](#_Toc169046278)

[2.1.2. Mức độ liên quan đến bài toán dự đoán bệnh tim 9](#_Toc169046279)

[2.2. Tiền xử lý dữ liệu 9](#_Toc169046280)

[2.3. Visualization và trích chọn đặc trưng 13](#_Toc169046281)

[2.3.1. Visualization 13](#_Toc169046282)

[2.3.2. Trích chọn đặc trưng 20](#_Toc169046283)

[2.4. Xây dựng mô hình 23](#_Toc169046284)

[2.4.1. Chia tập dữ liệu 23](#_Toc169046285)

[2.4.2. Lựa chọn thuật toán 23](#_Toc169046286)

[2.5. Đánh giá mô hình 27](#_Toc169046287)

[2.5.1. Đánh giá kết quả các độ đo lần đầu 27](#_Toc169046288)

[2.5.2. Parameter tuning 28](#_Toc169046289)

[2.5.3. Phát triển thêm mô hình dự đoán bằng các thuật toán khác 29](#_Toc169046290)

[2.6. Triển khai mô hình 32](#_Toc169046291)

[2.6.1. Streamlit 32](#_Toc169046292)

[2.6.2. Áp dụng Streamlit vào việc triển khai mô hình 33](#_Toc169046293)

[KẾT LUẬN 35](#_Toc169046294)

# CHƯƠNG I: TỔNG QUAN ĐỀ TÀI

1.1. Giới thiệu tổng quan dự án**. (0.5đ)**

### 1.1.1. Giới thiệu dự án

**Bệnh tim** là thuật ngữ mô tả một loạt các tình trạng ảnh hưởng đến tim, bao gồm các bệnh về mạch máu như bệnh động mạch vành, các vấn đề về nhịp tim (rối loạn nhịp tim), và các dị tật tim bẩm sinh. Thuật ngữ "bệnh tim" thường được sử dụng đồng nghĩa với "bệnh tim mạch", mặc dù bệnh tim mạch thường cụ thể hơn đến các tình trạng liên quan đến sự thu hẹp hoặc tắc nghẽn của mạch máu, dẫn đến đau tim, đau thắt ngực hoặc đột quỵ. Ngoài ra, các bệnh tác động đến cơ tim, van tim hoặc nhịp tim cũng được coi là các dạng bệnh tim.

Bệnh tim là một trong những nguyên nhân hàng đầu gây bệnh tật và tử vong trên toàn cầu. Dự đoán bệnh tim mạch được xem là một lĩnh vực quan trọng trong phân tích dữ liệu lâm sàng. Ngành chăm sóc sức khỏe tạo ra lượng dữ liệu khổng lồ, và việc khai thác dữ liệu này nhằm chuyển đổi các bộ sưu tập dữ liệu lớn thành thông tin hữu ích có ý nghĩa rất quan trọng trong việc dự đoán và điều trị bệnh tim mạch.

Dự án này tập trung vào việc phát triển một mô hình trí tuệ nhân tạo để dự đoán khả năng mắc bệnh tim. Bệnh tim là một trong những nguyên nhân hàng đầu gây tử vong trên toàn cầu, do đó, việc dự đoán sớm và chính xác khả năng mắc bệnh có thể giúp cải thiện chăm sóc sức khỏe và cứu sống nhiều người. Sử dụng các kỹ thuật học máy và khai thác dữ liệu, chúng tôi sẽ xây dựng một mô hình có thể phân tích các yếu tố nguy cơ và dự đoán liệu một người có khả năng mắc bệnh tim hay không. Đồng thời, chúng tôi tiến hành so sánh các thuật toán học máy với nhau để tìm ra mô hình hiệu quả nhất

### 1.1.2. Phạm vi thực hiện dự án

Phạm vi thực hiện dự án bao gồm:

1. Thu thập và tiền xử lý dữ liệu y tế liên quan đến bệnh tim.
2. Khám phá và phân tích dữ liệu để xác định các yếu tố nguy cơ quan trọng.
3. Xây dựng và huấn luyện mô hình dự đoán sử dụng các thuật toán học máy.
4. Đánh giá hiệu suất của mô hình thông qua các chỉ số như độ chính xác,...
5. Triển khai ứng dụng demo hỗ trợ trong công tác chăm sóc sức khỏe và dự phòng bệnh tim.

### 1.1.3. Ứng dụng

Ứng dụng của dự án này bao gồm:

* Hỗ trợ bác sĩ và nhân viên y tế trong việc dự đoán và chẩn đoán sớm bệnh tim.
* Cung cấp công cụ tự đánh giá nguy cơ mắc bệnh tim cho người dùng cá nhân.
* Hỗ trợ trong việc thiết kế các chương trình phòng ngừa bệnh tim dựa trên dữ liệu thực tế.
* Đóng góp vào nghiên cứu y học và cải thiện các mô hình dự đoán bệnh tật trong tương lai.

## 1.2. Giới thiệu bài toán

### 1.2.1. Phát biểu bài toán

Bài toán đặt ra là xây dựng một mô hình dự đoán khả năng mắc bệnh tim dựa trên các thông tin y tế và lối sống của bệnh nhân. Mục tiêu là xác định liệu một người có nguy cơ cao mắc bệnh tim hay không, từ đó hỗ trợ trong việc chẩn đoán và điều trị sớm.

1.2.2 Giới thiệu về bộ dữ liệu***:***

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Vai trò trong dự đoán bệnh tim** |
| ***INPUT*** | | | |
| 1 | age | Tuổi của bệnh nhân | Tuổi là yếu tố nguy cơ quan trọng vì nguy cơ mắc bệnh tim thường tăng theo tuổi. |
| 2 | sex | Giới tính của bệnh nhân (1 = Nam, 0 = Nữ) | Giới tính ảnh hưởng đến nguy cơ mắc bệnh tim, với tỷ lệ mắc bệnh khác nhau giữa nam và nữ. |
| 3 | cp | Loại đau ngực (0: Đau thắt ngực điển hình, 1: Đau thắt ngực không điển hình, 2: Không đau thắt 3: Không có triệu chứng đau ngực) | Loại đau ngực có thể chỉ ra mức độ nghiêm trọng và loại bệnh tim. |
| 4 | trestbps | Huyết áp tâm thu khi nghỉ ngơi (mm Hg) | Huyết áp cao là yếu tố nguy cơ chính của bệnh tim. |
| 5 | chol | Mức độ cholesterol toàn phần trong máu (mg/dL) | Mức cholesterol cao có thể dẫn đến xơ vữa động mạch và bệnh tim. |
| 6 | fbs | Đường huyết lúc đói > 120 mg/dL (1 = Đúng, 0 = Sai) | Đường huyết cao khi đói có thể chỉ ra nguy cơ tiểu đường, một yếu tố nguy cơ của bệnh tim. |
| 7 | restecg | Kết quả điện tâm đồ lúc nghỉ ngơi (0: Bình thường, 1: Bất thường ST-T, 2: Phì đại thất trái hoặc bất thường khác) | Những bất thường này có thể liên quan đến nguy cơ mắc bệnh tim. |
| 8 | thalach | Nhịp tim tối đa đạt được khi gắng sức | Nhịp tim cao khi gắng sức có thể chỉ ra sức khỏe tim mạch tốt hơn. |
| 9 | exang | Đau thắt ngực do gắng sức (1 = Có, 0 = Không) | Đau ngực do gắng sức là dấu hiệu của bệnh tim. |
| 10 | oldpeak | Sự chênh lệch ST (đo trên ECG - điện tâm đồ) khi gắng sức so với lúc nghỉ ngơi | Thay đổi này có thể chỉ ra thiếu máu cục bộ của cơ tim. |
| 11 | slope | Độ dốc của đoạn ST khi gắng sức (0: Dốc lêm, 1: Bằng phẳng, 2: Dốc xuống) | Độ dốc này có thể liên quan đến nguy cơ mắc bệnh tim. |
| 12 | ca | Số lượng mạch máu chính bị hẹp (0-3) | Số lượng mạch máu bị hẹp cao hơn có thể chỉ ra nguy cơ cao hơn của bệnh tim. |
| 13 | thal | Tình trạng thalassemia (1: Khuyết tật cố đinh, 2: Bình thường, 3: Khuyết tật có thể điều chỉnh) | Thalassemia là rối loạn máu có thể ảnh hưởng đến sức khỏe tim mạch. |
| ***OUTPUT*** | | | |
| 14 | target | Biến mục tiêu, chỉ ra liệu bệnh nhân có mắc bệnh tim hay không (1 = Mắc bệnh tim, 0 = Không mắc bệnh tim). | Đây là biến đích mà mô hình cần dự đoán. |

### 1.2.3. Ứng dụng

Bài toán có thể áp dụng tại:

* Các cơ sở y tế và bệnh viện để hỗ trợ chẩn đoán.
* Các ứng dụng sức khỏe cá nhân giúp người dùng tự đánh giá nguy cơ bệnh tim.
* Các chương trình nghiên cứu và dự phòng bệnh tim trong cộng đồng.

# CHƯƠNG II: XÂY DỰNG BÀI TOÁN

2.1. Thu thập dữ liệu**. (1đ)**

### 2.1.1. Nguồn dữ liệu

Nguồn dữ liệu mà chúng tôi sẽ sử dụng lấy từ một trong các bộ dữ liệu phổ biến về bệnh tim trên Kaggle. Đó là bộ dữ liệu "Heart Disease UCI".Các bước để lấy dữ liệu như sau:

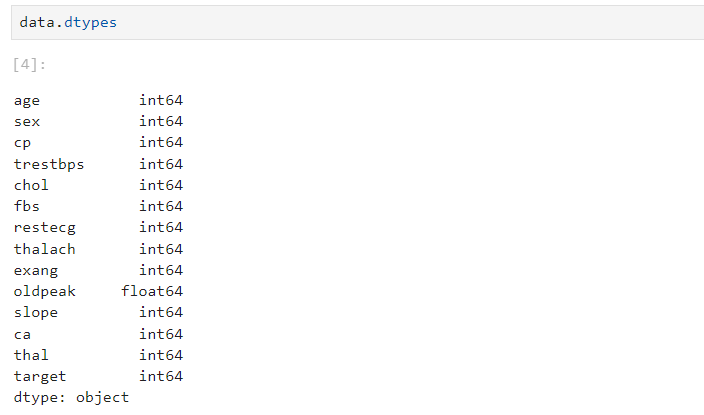
* Truy cập vào [Kaggle](https://www.kaggle.com/)
* Tìm kiếm bộ dữ liệu với từ khóa "Heart Disease"
* Chọn bộ dữ liệu phù hợp, ví dụ: "Heart Disease UCI"

Bộ dữ liệu này thường bao gồm các thông tin liên quan đến bệnh nhân như tuổi, giới tính, mức cholesterol, huyết áp, nhịp tim, và các thông số y tế khác.

### 2.1.2. Mức độ liên quan đến bài toán dự đoán bệnh tim

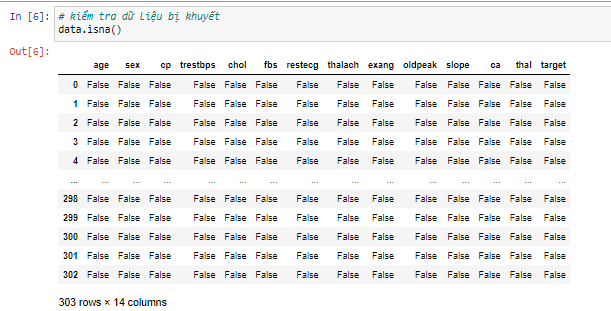
Bộ dữ liệu "Heart Disease UCI" là một trong những bộ dữ liệu phổ biến và được sử dụng rộng rãi trong nghiên cứu y tế, đặc biệt là trong việc dự đoán bệnh tim. Dữ liệu này chứa các thuộc tính (features) liên quan đến các yếu tố nguy cơ và triệu chứng của bệnh tim. Các thuộc tính này không chỉ có tính mô tả mà còn có tính dự báo cao, đã được chứng minh qua nhiều nghiên cứu lâm sàng và y học là có liên quan đến khả năng mắc bệnh tim. Những dữ liệu này rất phù hợp cho bài toán dự đoán bệnh tim

2.2. Tiền xử lý dữ liệu**. (1đ)**



Từ kiểu dữ liệu ta thấy tất cả thuộc tính đều là int64/float64. Hầu hết các dữ liệu này đều đã được mã hóa kể cả biến mục tiêu. Vì vậy để giải quyết được bài toán chúng ta cần đưa dữ liệu về giống bộ dữ liệu gốc

Bước 1: Kiếm tra dữ liệu khuyết thiếu

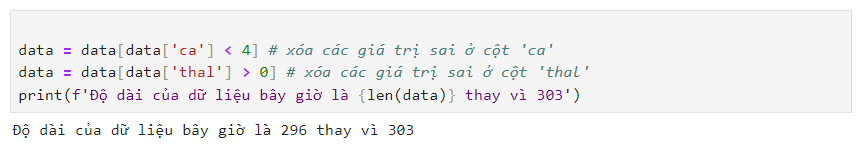


Bước 2: Xóa các dữ liệu lỗi

- Cột ‘ca’ có các giá trị =4, giá trị này không chính xác vì trong tập dữ liệu ban đầu của Cleveland chúng là NaN

- Cột ‘thal’ có thal = 0 cũng không chính xác vì chúng cũng là NaN trong tập dữ liệu gốc

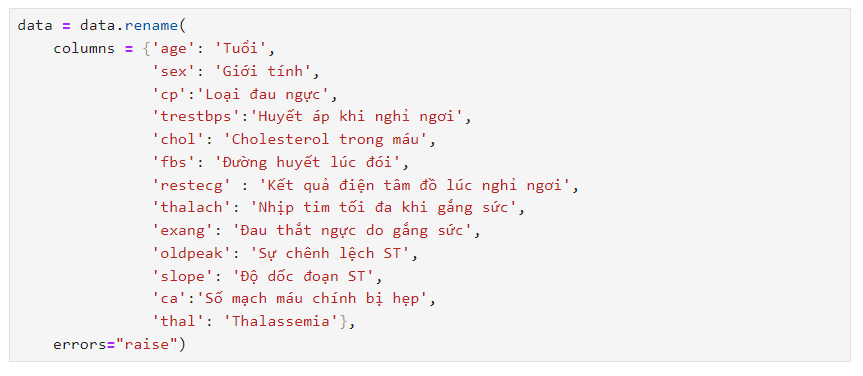
Do vậy ta sẽ xóa các dữ liệu có giá trị trên 4 ở cột ‘ca’ và giá trị = 0 ở cột ‘thal’



Bước 3: Đổi tên thuộc tính

Tên các thuộc tính trong tập dữ liệu được viết tắt và khó hiểu. Một cái tên y tế/ kỹ thuật đã đủ khó hiểu đối với hầu hết chúng ta chứ đừng nói đến dạng viết tắt của chúng. Vì vậy, để làm cho chúng dễ đọc hơn, nhóm sẽ thay đổi tên cột dữ liệu bằng cách sử dụng thông tin từ kho lưu trữ dữ liệu UCL và nhóm cũng sẽ thay thế các dữ liệu được mã hóa (0, 1, 2,..) thành ý nghĩa y tế của chúng (như đau thắt ngực không điển hình, đau thắt ngực điển hình,..)

*Hình: Thay đổi tên cột*



*Hình: Thay đổi dữ liệu được mã hóa*



Kết quả:





Bước 4: Phân nhóm dữ liệu

Phân thành 3 nhóm: nhóm thuộc tính số, nhóm thuộc tính phân loại binary chỉ có 2 giá trị và nhóm thuộc tính phân loại có nhiều hơn 2 giá trị



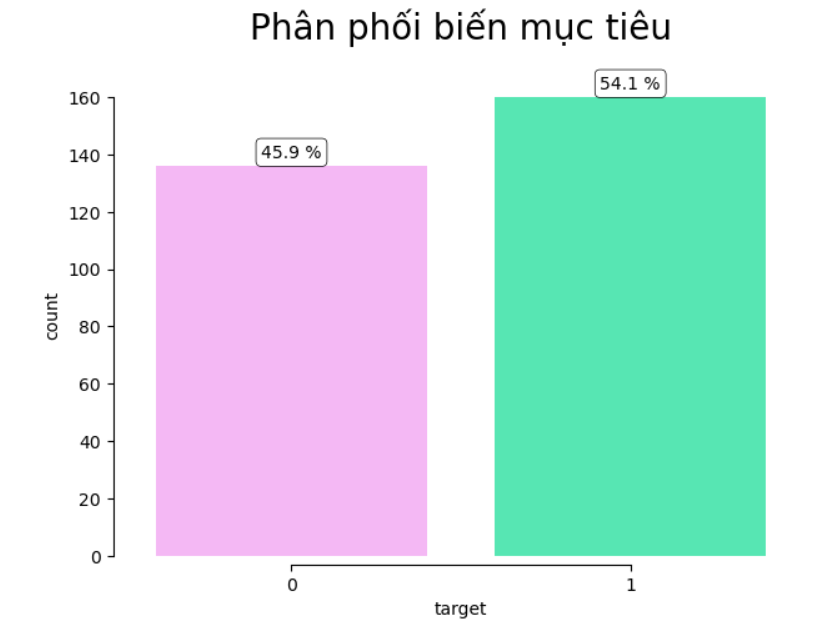
2.3. Visualization và trích chọn đặc trưng**.**

### 2.3.1. Visualization

*a. Phân tích sự phân bổ của biến mục tiêu*

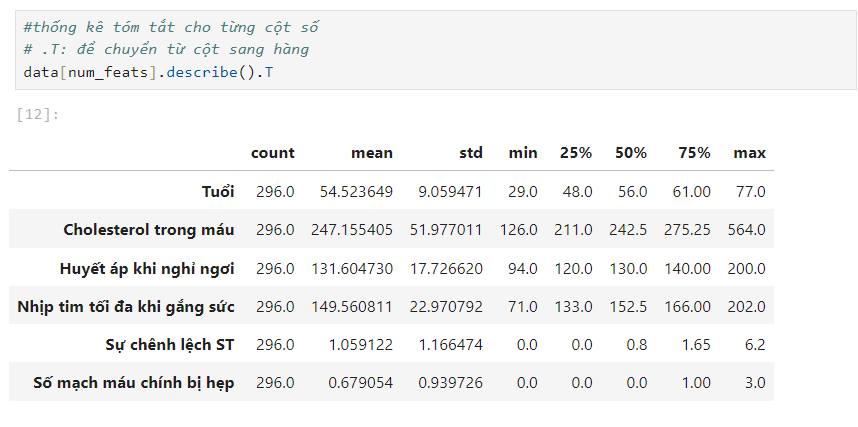


Chúng em quan sát thấy target khá cân bằng với 45.9% không mắc bệnh tim và 54,1% mắc bệnh tim. Vì vậy, không cần phải lo lắng về sự mất cân bằng mục tiêu.



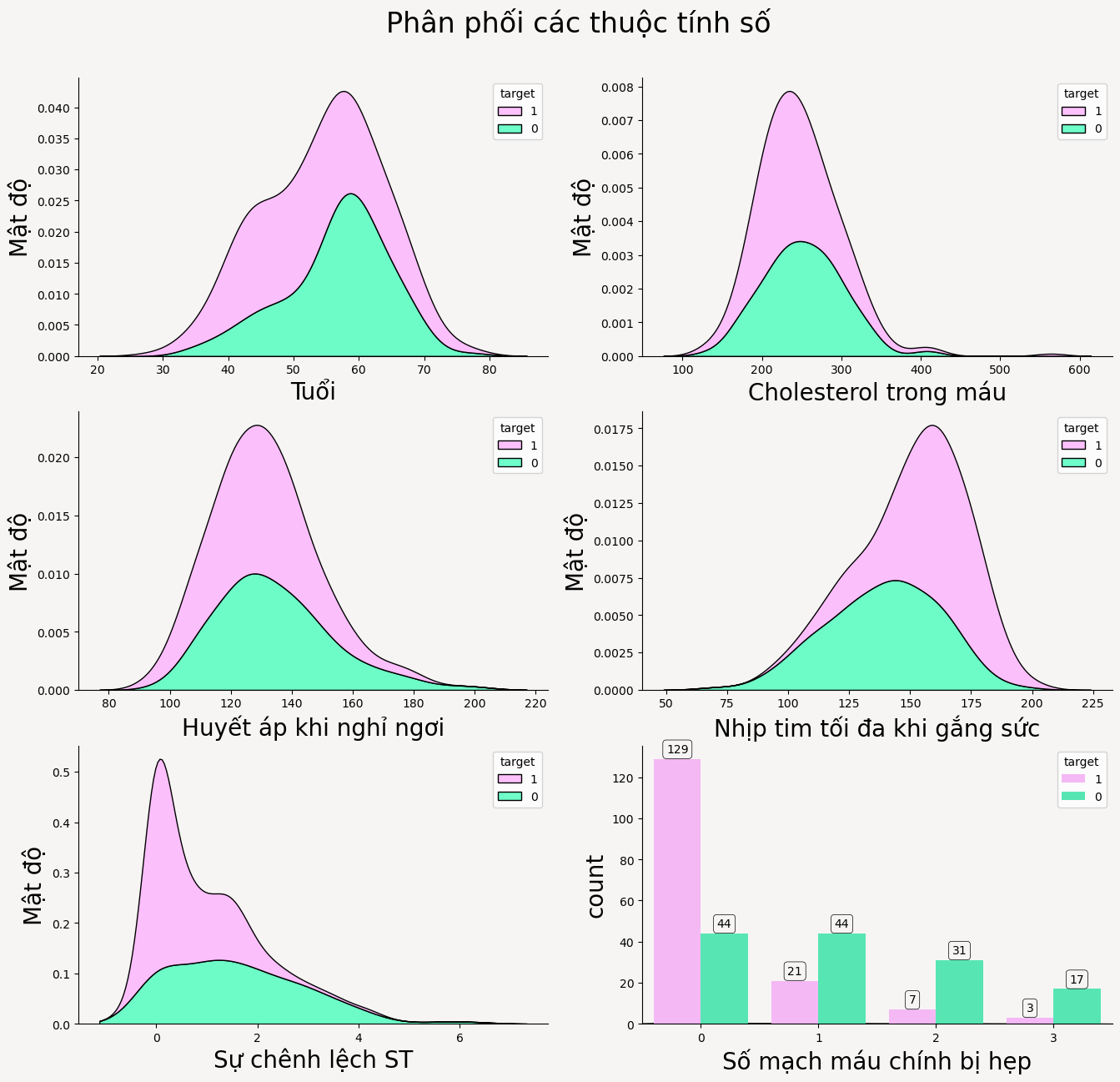
*b.  Phân tích thuộc tính số*

Đối với các thuộc tính số, chúng ta có thể áp dụng phương thức tiện dụng của pandas data.describe() và nhận được bản tóm tắt thống kê toàn cầu.



* Tuổi: Độ tuổi trung bình trong tập dữ liệu là 54,5 tuổi trong đó người lớn tuổi nhất là 77 tuổi và trẻ nhất là 29 tuổi
* Cholesterol trong máu: Mức cholesterol trung bình là 247,15 trong đó mức tối đa là 564 và tối thiểu là 126. Theo như nhóm tìm hiểu thì mức cholesterol khỏe mạnh là < 200mg/dl và thường mức cholesterol cao thì có liên quan đến bệnh tim.
* Huyết áp khi nghỉ ngơi: Trung bình là 131, lớn nhất là 200 và nhỏ nhất là 94
* Nhịp tim tối đa khi gắng sức: Nhịp tim tối đa trung bình là 149,5 bpm. Giá trị nhỏ nhất và lớn nhất lần lượt là 71 bpm và 202 bpm (bpm: số nhịp đập của tim trong 1 phút)
* Sự chênh lệch ST: Giá trị trung bình là 1,06. Giá trị lớn nhất là 6,2 và nhỏ nhất là 0
* Số lượng mạch máu chính bị hẹp: Tối đa là 3, tối thiểu là 0, trung bình là 0,68

*c. Biểu đồ phân tán*



**-** *Tuổi tác:* Nhìn chung, nguy cơ mắc bệnh tim tăng theo độ tuổi. Ở những người dưới 50 tuổi, tỷ lệ mắc bệnh tim tương đối thấp. Tuy nhiên, sau 50 tuổi, tỷ lệ mắc bệnh tim bắt đầu tăng nhanh chóng. Ở những người trên 70 tuổi, tỷ lệ mắc bệnh tim cao nhất.

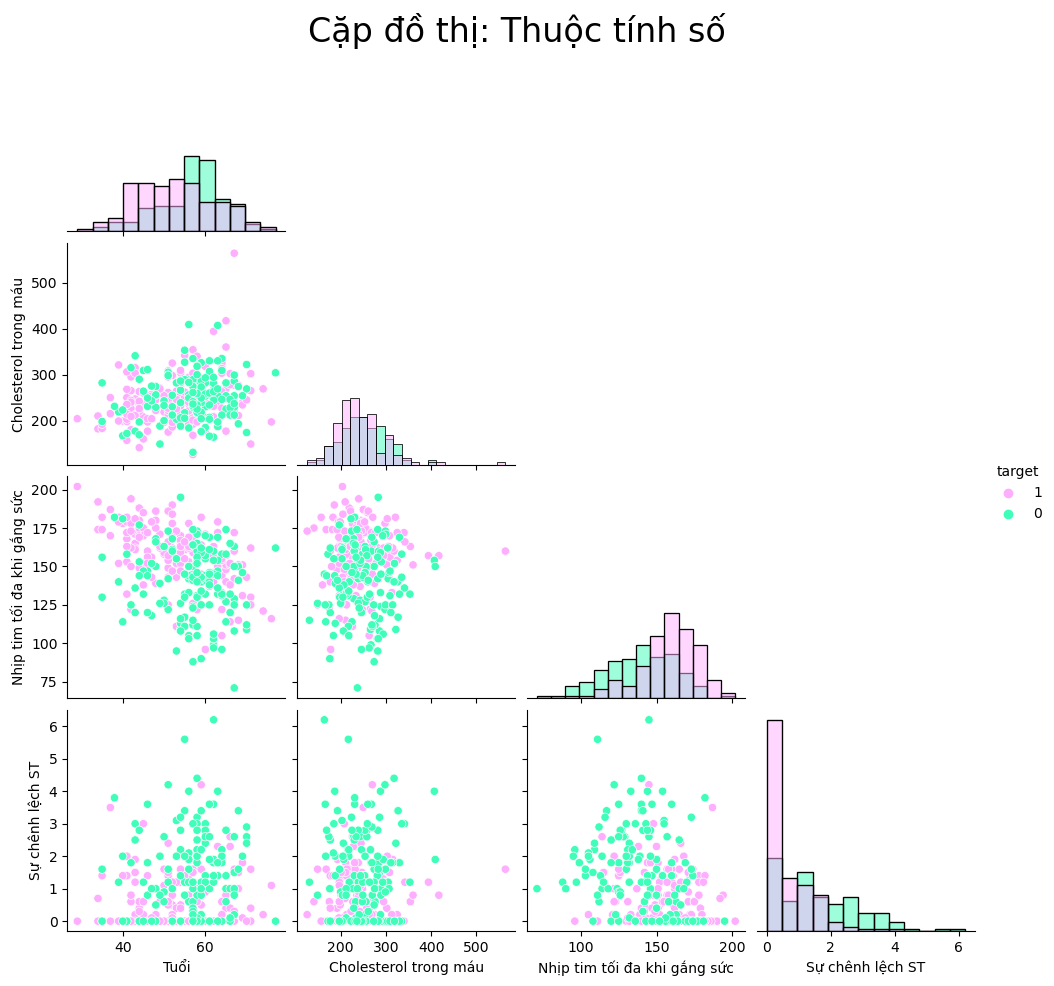
**-** *Mức cholesterol trong máu:*Những người có mức cholesterol trong máu cao (trên 200 mg/dL) có nguy cơ mắc bệnh tim cao hơn nhiều so với những người có mức cholesterol trong máu bình thường (dưới 200 mg/dL).

**-** *Huyết áp:*Huyết áp cao là một yếu tố nguy cơ quan trọng khác của bệnh tim. Ở những người có huyết áp cao, tỷ lệ mắc bệnh tim cao hơn đáng kể so với những người có huyết áp bình thường.

**-** *Nhịp tim tối đa khi gắng sức:*Ở những người có nhịp tim tối đa khi gắng sức thấp, tỷ lệ mắc bệnh tim cao hơn so với những người có nhịp tim tối đa khi gắng sức bình thường.

**-** *Sự chênh lệch ST***:** Những người có sự chênh lệch ST cao (trên 1 mV) có nguy cơ mắc bệnh tim cao hơn so với những người có sự chênh lệch ST bình thường (dưới 1 mV).

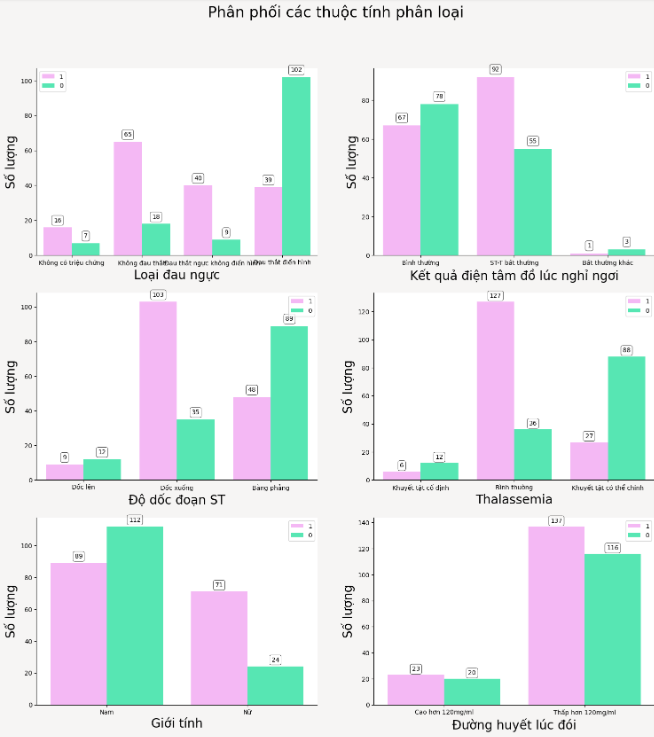
**-** *Số mạch máu chính bị hẹp***:** Ở những người có nhiều mạch máu chính bị hẹp, tỷ lệ mắc bệnh tim cao hơn so với những người có ít hoặc không có mạch máu chính bị hẹp.

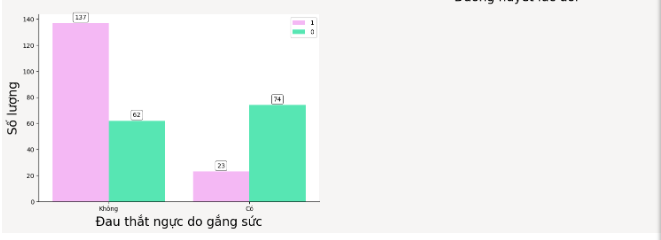


**-** *Mức độ phân bố của các điểm dữ liệu:*Các điểm dữ liệu màu xanh (không bị bệnh tim) tập trung chủ yếu ở khu vực có giá trị thấp của các yếu tố nguy cơ. Các điểm dữ liệu màu hồng (bị bệnh tim) tập trung chủ yếu ở khu vực có giá trị cao của các yếu tố nguy cơ.

**-** *Tác động của từng yếu tố nguy cơ***:** Trong biểu đồ này, độ dốc của đường xu hướng cho thấy rằng cholesterol trong máu và huyết áp là hai yếu tố nguy cơ có ảnh hưởng mạnh nhất đến nguy cơ mắc bệnh tim.

Tóm lại, có thể thấy rằng nguy cơ mắc bệnh tim phụ thuộc vào nhiều yếu tố khác nhau, bao gồm tuổi tác, mức cholesterol trong máu, huyết áp, nhịp tim tối đa khi gắng sức, sự chênh lệch ST và số mạch máu chính bị hẹp. Do đó, để dự đoán nguy cơ mắc bệnh tim một cách chính xác, cần phải xem xét tất cả các yếu tố này.





**-** *Giới tính:*Nhìn chung, tỷ lệ mắc bệnh tim ở nữ cao hơn nam giới

**-** *Kết quả điện tâm đồ lúc nghỉ ngơi:*Tỷ lệ mắc bệnh tim ở những người có kết quả điện tâm đồ lúc nghỉ ngơi bất thường cao hơn ở tất cả các nhóm tuổi.

**-** *Độ dốc đoạn ST:*Những người có độ dốc đoạn ST bất thường có nguy cơ mắc bệnh tim cao hơn những người có độ dốc đoạn ST bình thường.

**-** *Thalassemia:*Những người có Thalassemia bình thường có nguy cơ mắc bệnh tim cao nhất

**-** *Đường huyết lúc đói:*Tỷ lệ mắc bệnh tim khi đường huyết lúc đói nhỏ hơn 120mg/ml và lớn hơn 120 mg/ml không có nhiều sự chênh lệch

- *Đau thắt ngực do gắng sức:*Người bị đau thắt ngực do gắng sức có tỷ lệ mắc bệnh tim thấp hơn người không bị đau thắt ngực do gắng sức

**Kết luận:** Đường huyết lúc đói không có ảnh hưởng nhiều đến khả năng mắc bệnh tim

### 2.3.2. Trích chọn đặc trưng

*a. Biểu đồ tương quan*

- Tìm tương quan giữa các thuộc tính, đặc biệt xem xét mối quan hệ giữa các feature (đặc trưng) với target

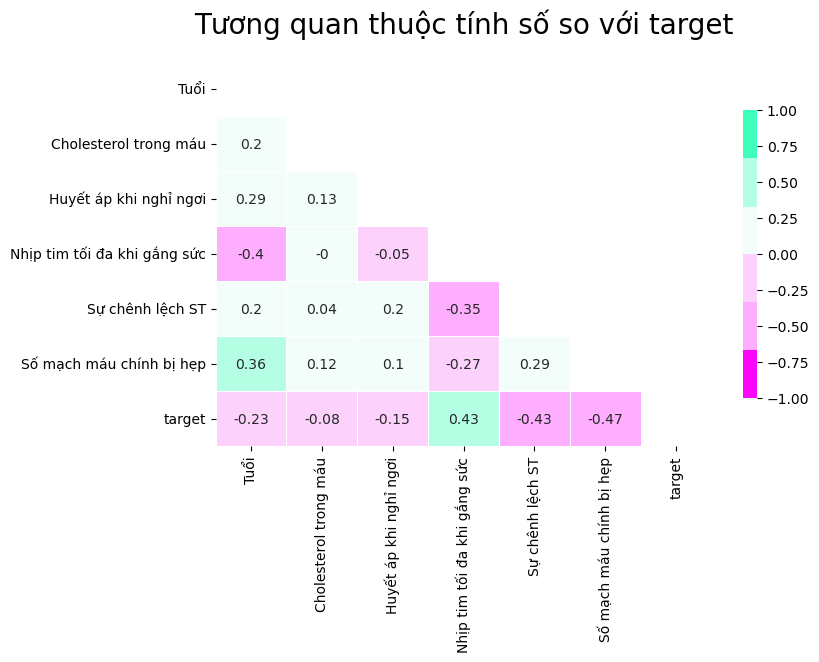
Mục đích của ma trận tương quan:

Tương quan đo lường mức độ phụ thuộc tuyến tính giữa hai biến. Giá trị tương quan có thể dao động từ -1 đến 1, với:

1: Tương quan hoàn toàn dương (khi một biến tăng, biến kia cũng tăng).

-1: Tương quan hoàn toàn âm (khi một biến tăng, biến kia giảm).

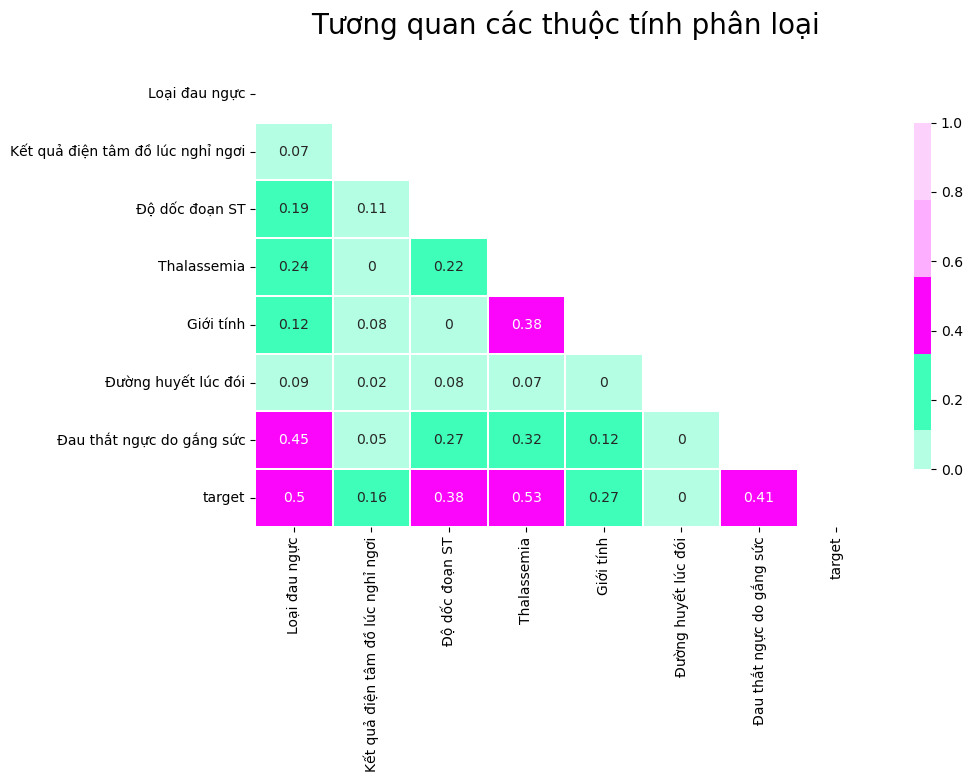
0: Không có tương quan (các biến không phụ thuộc lẫn nhau).



*b. Cramer's V-Biểu đồ  đánh giá mức độ mạnh mẽ của mối quan hệ giữa hai biến danh định*

Nó là một chỉ số thống kê được sử dụng phổ biến trong phân tích bảng chéo (contingency tables). Giá trị của Cramer's V nằm trong khoảng từ 0 đến 1, trong đó:

* 0 cho biết không có mối liên hệ giữa các biến.
* 1 cho biết có mối liên hệ mạnh mẽ hoàn hảo giữa các biến.



Cân nhắc bỏ đường huyết lúc đói vì tương quan với target bằng 0

*c. SelectKbest*

SelectKBest là một lớp trong thư viện scikit-learn được sử dụng để chọn ra các đặc trưng tốt nhất từ một tập dữ liệu dựa trên một hàm kiểm định được chỉ định.

SelectKBest sẽ tính toán điểm (score) của mỗi đặc trưng dựa trên hàm kiểm định được chỉ định và chọn ra k đặc trưng có điểm cao nhất. Điều này giúp giảm chiều dữ liệu và tăng hiệu suất của các mô hình học máy bằng cách chỉ sử dụng các đặc trưng quan trọng nhất.

Trong đó:

* score\_func là hàm kiểm định được sử dụng để tính điểm của các đặc trưng.
* k là số lượng đặc trưng tốt nhất sẽ được chọn.
* fit\_transform() tính toán điểm của các đặc trưng và chọn ra k đặc trưng tốt nhất.
* get\_support() trả về một mảng boolean hoặc chỉ số của các đặc trưng được chọn.



*c. Kết quả trích chọn đặc trưng*

Features: gồm 10 thuộc tính quan trọng nhất: ['Tuổi', 'Giới tính', 'Loại đau ngực', 'Cholesterol trong máu', 'Nhịp tim tối đa khi gắng sức', 'Đau thắt ngực do gắng sức', 'Sự chênh lệch ST', 'Độ dốc đoạn ST', 'Số mạch máu chính bị hẹp', 'Thalassemia']

Target: ‘target’

2.4. Xây dựng mô hình**.**

### 2.4.1. Chia tập dữ liệu

Chia bộ dữ liệu thành 2 tập train và test: 25% của dữ liệu sẽ được sử dụng cho bộ kiểm tra, còn lại (75%) sẽ được sử dụng cho việc huấn luyện.

- X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.25, random\_state=0)

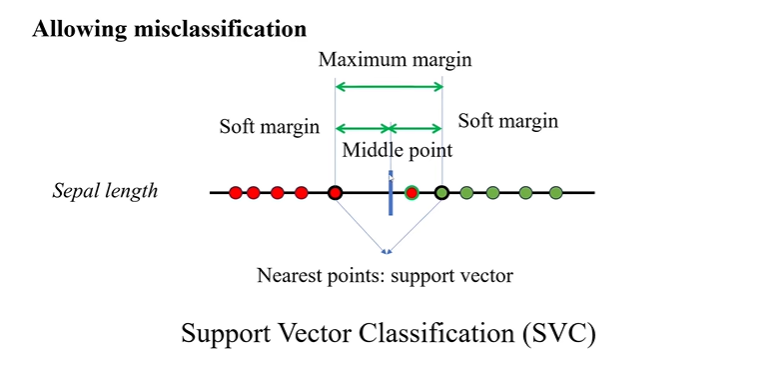
https://lh7-us.googleusercontent.com/docsz/AD_4nXceN_jEeEqpcWhs-Px9OqpPUyIlLwoMsKIGTmJL1gM7Zh1I7GcrUtvYsO0OzFHSYp4xg8MD480Au-u_vP7SpORoSr9PN_NFTAHTUJQzpWFaZXMBFqcMdzhZuIV92ZZx9ti1A7tIUXKIT-zUVsrzmsJooGE?key=QF3yEhjLoWCQNqT-kseIzQ

### 2.4.2. Lựa chọn thuật toán

Sau quá trình tiền xử lý dữ liệu, chúng ta chỉ có 297 dòng dữ liệu để có thể tiếp tục xây dựng mô hình dự đoán. Việc lựa chọn thuật toán là vô cùng quan trọng, hơn nữa với tập dữ liệu khá nhỏ trên, rất dễ xảy ra hiện tượng mất cân bằng dữ liệu, ảnh hưởng rất lớn tới hiệu suất dự đoán của mô hình.

Trước thực trạng trên, nhóm nghiên cứu xin đề xuất lựa chọn 3 thuật toán để tiến hành xây dựng mô hình và tiến hành dự đoán, 2 trong đó là 2 thuật toán rất quen thuộc, dễ biểu diễn và dễ hình dung trong các bài toán phân loại, là Logistic Regression và Decision Tree. Ngoài ra, nhóm có sử dụng thuật toán SVC, một loại của Support Vector Machines (SVM) - nhóm các phương pháp học máy giám sát mạnh mẽ và phổ biến.

*a. Lý thuyết thuật toán SVM*

**

Thuật toán Support Vector Machine (SVM) là một phương pháp học máy được sử dụng cho cả bài toán phân loại và hồi quy. Ý tưởng cơ bản của SVM là tìm ra ranh giới phân chia tốt nhất giữa các lớp dữ liệu.

Trong bài toán phân loại, SVM cố gắng tìm ranh giới phân chia tốt nhất (hay còn gọi là siêu phẳng) để phân tách các điểm dữ liệu thuộc các lớp khác nhau trong không gian đặc trưng. Cụ thể, SVM cố gắng tìm siêu phẳng có margin lớn nhất giữa các lớp. Margin là khoảng cách từ các điểm dữ liệu gần nhất tới siêu phẳng.

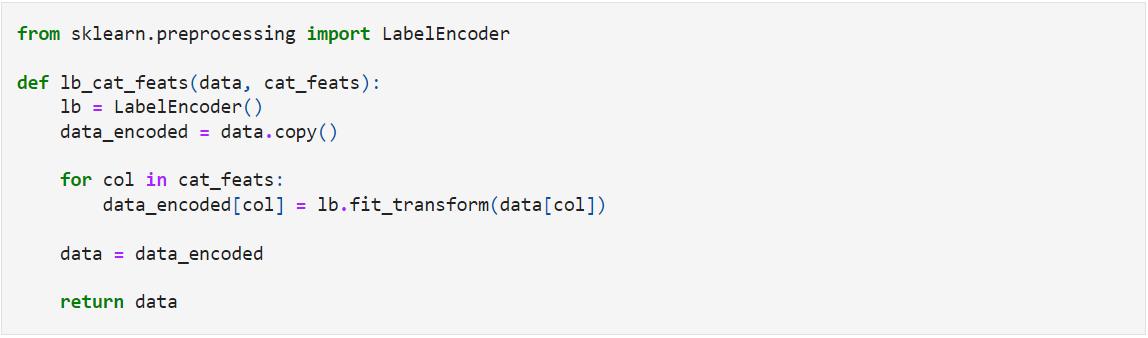
Một số đặc điểm của SVM:

* Hỗ trợ các siêu phẳng phi tuyến: SVM có thể được mở rộng để sử dụng các hàm nhân (kernel) để ánh xạ dữ liệu vào một không gian đặc trưng cao hơn, nơi các lớp dữ liệu có thể được phân loại bằng các siêu phẳng phi tuyến.
* Hạn chế overfitting: SVM có thể kiểm soát overfitting thông qua việc điều chỉnh siêu tham số, như siêu tham số regularization (C).
* Hiệu suất cao trong không gian chiều lớn: SVM vẫn hoạt động hiệu quả khi số chiều của dữ liệu lớn hơn số lượng mẫu.
* Phân loại đa lớp: SVM có thể được mở rộng để xử lý bài toán phân loại đa lớp bằng cách sử dụng các kỹ thuật như One-vs-One hoặc One-vs-All.

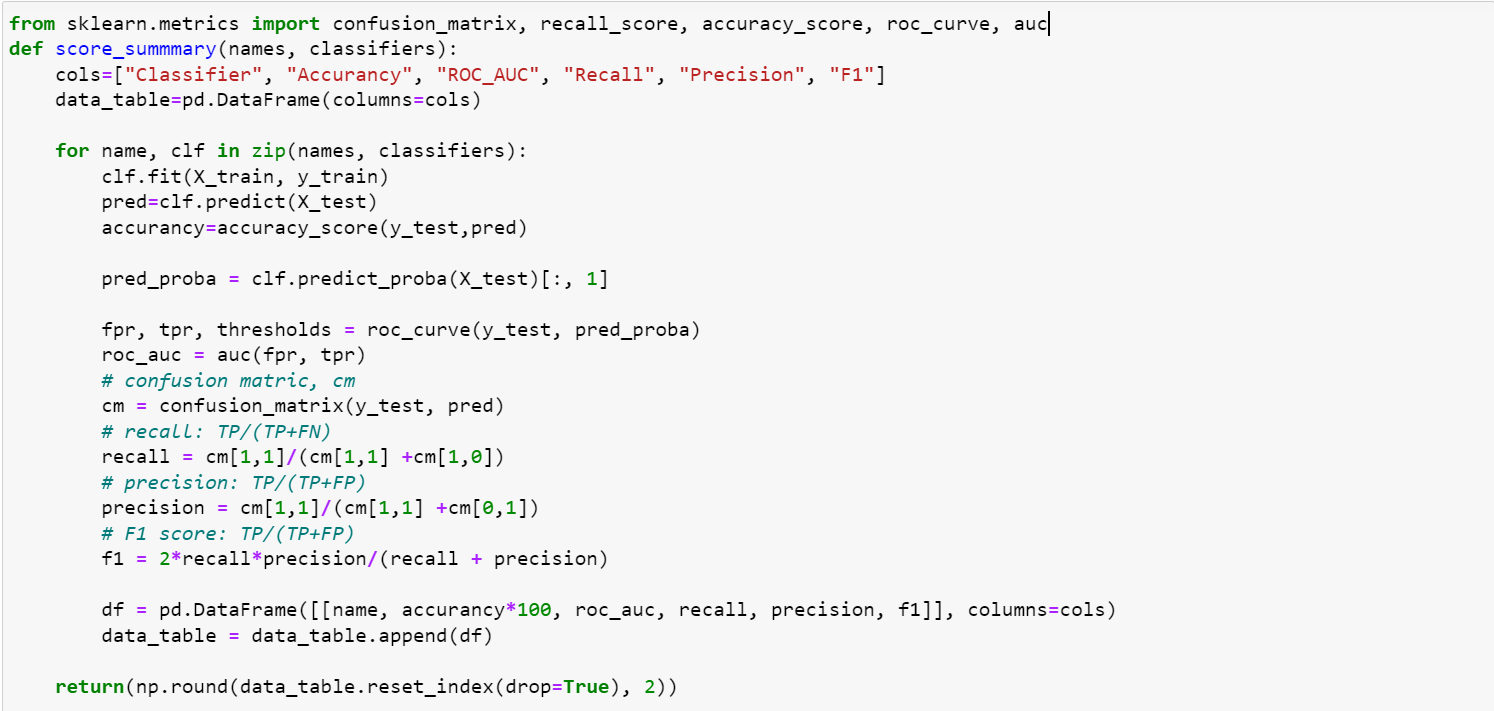
Tuy nhiên, SVM cũng có một số hạn chế, bao gồm đòi hỏi tính toán lớn khi có một lượng lớn các mẫu hoặc các đặc trưng, cũng như đòi hỏi lựa chọn siêu tham số tốt để đảm bảo hiệu suất tốt nhất của mô hình.

*b. Áp dụng các thuật toán*

Đầu tiên, sử dụng Label Encoder để mã hóa các thuộc tính đặc trưng đã được chọn.



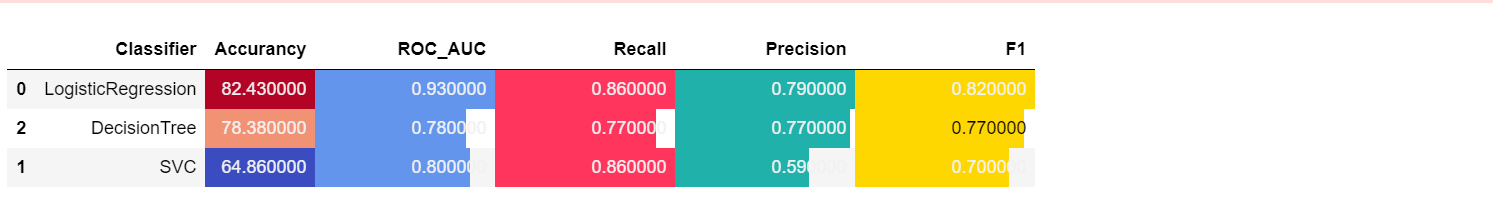
Nhóm khai báo một hàm dùng chung cho việc tính toán các độ đo của các thuật toán có tên “score\_summmary()”.



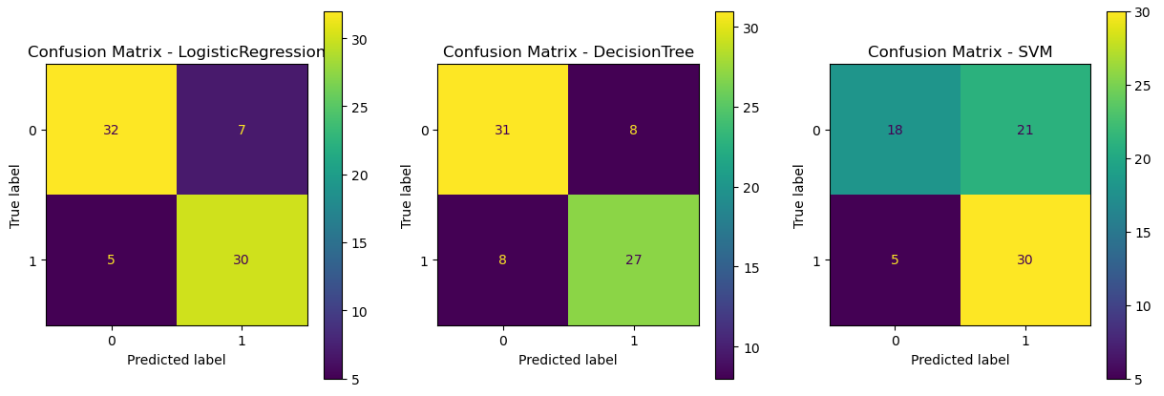
Áp dụng xây dựng các mô hình đã lựa chọn để cho ra kết quả các độ đo lần đầu tiên (sử dụng bộ tham số mặc định).



Kết quả:



Ma trận nhầm lẫn



## 2.5. Đánh giá mô hình

2.5.1. Đánh giá kết quả các độ đo lần đầu***.***

Với bài toán dự đoán khả năng mắc bệnh tim của một bệnh nhân, việc sử dụng các độ đo phù hợp cũng rất quan trọng. Bởi kết quả dự đoán của mô hình không chỉ phản ánh kết quả của một mô hình, mà còn ảnh hưởng tới cả quyết định và tâm lý con người. Với bài toán này, nhóm cân nhắc sử dụng cùng lúc 5 độ đo sau:

* Accuracy: Đo độ chính xác của tổng thể mô hình dự đoán.

Accuracy=TP+TNTP+TN+FP+FN

* Recall: Độ đo ảnh hưởng rất lớn tới yêu cầu của mô hình dự đoán - dự đoán chính xác số bệnh nhân mắc bệnh thực sự (True Positive).

Recall=TPTP+FN

* Precision: Độ đo giúp hạn chế tỷ lệ dự đoán sai một người không mắc bệnh là có bệnh (False Positive).

Precision=TPTP+FP

* F1: Đo sự cân bằng giữa Recall và Precision, đảm bảo khỏi sự mất cân bằng giữa các lớp dự đoán.

F1=2PrecisionRecallPrecision+Recall

* ROC\_AUC: Độ đo cung cấp cái nhìn tổng quát về hiệu suất của mô hình qua các ngưỡng dự đoán khác nhau.

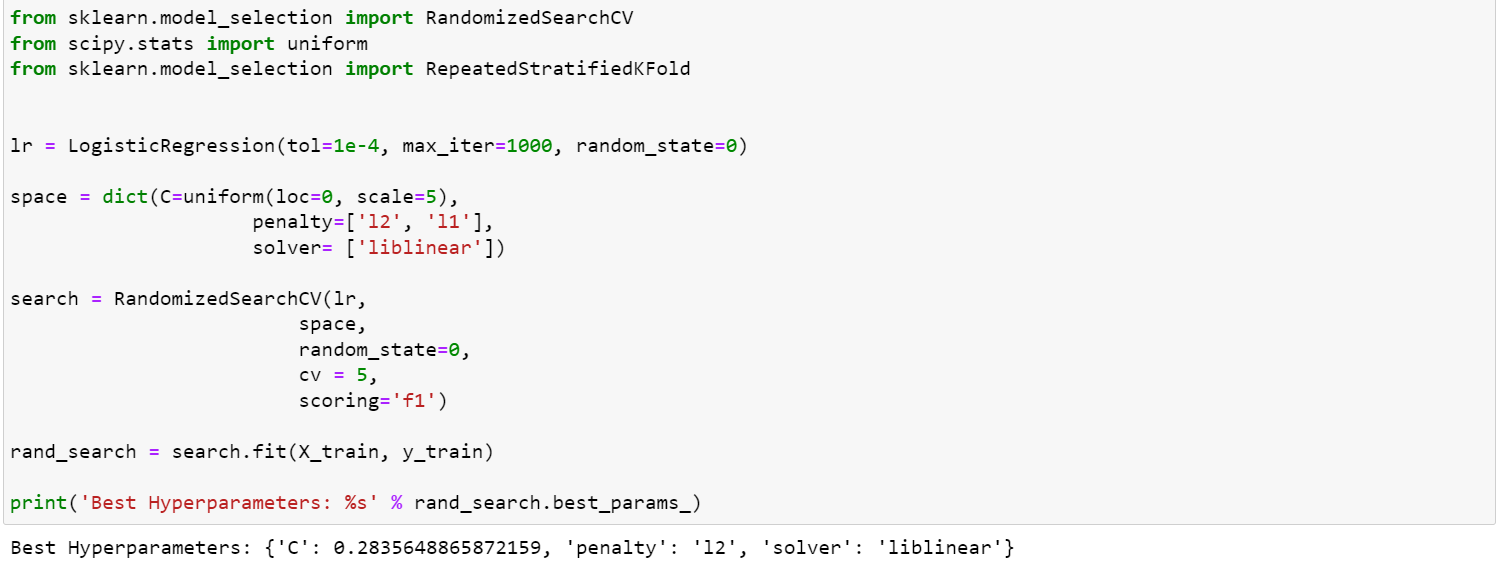
AUC=ROC

Với bài toán dự đoán khả năng mắc bệnh tim của một người, nhóm cân nhắc sử dụng các độ đo F1 Score, Recall, Precision và AUC để giảm thiểu ảnh hưởng của việc mất cân bằng dữ liệu, tất nhiên vẫn sử dụng thêm độ đo Accuracy để đánh giá độ chính xác tổng quát của mô hình.

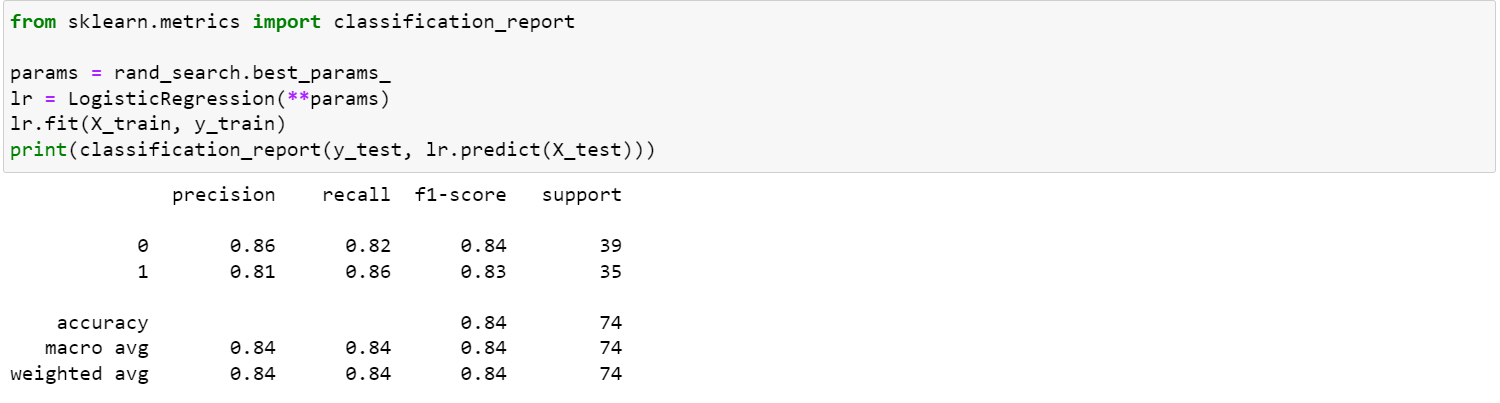
Đồng thời, ý nghĩa của bài toán dự đoán là giảm thiểu False Negatives, cho nên nhóm ưu tiên độ đo Recall. Tuy nhiên, trên kết quả chạy lần đầu, ta thấy 2 thuật toán Logistic Regression và SVC đều có độ đo Recall bằng nhau (cùng bằng 0.86). Đồng thời, nếu đây là một mô hình áp dụng cho 1 phòng khám, không ai dám mang danh tiếng của chính mình ra trước rủi ro, tức là giảm thiểu False Positive. Vì thế nhóm quyết định tin tưởng thuật toán cho ra kết quả F1 cao hơn, để cân bằng một cách tốt nhất Recall và Precision, và nhóm chọn thuật toán Logistic Regression với độ đo F1 cao nhất là 0.82 để tiếp tục cải thiện hiệu quả mô hình.

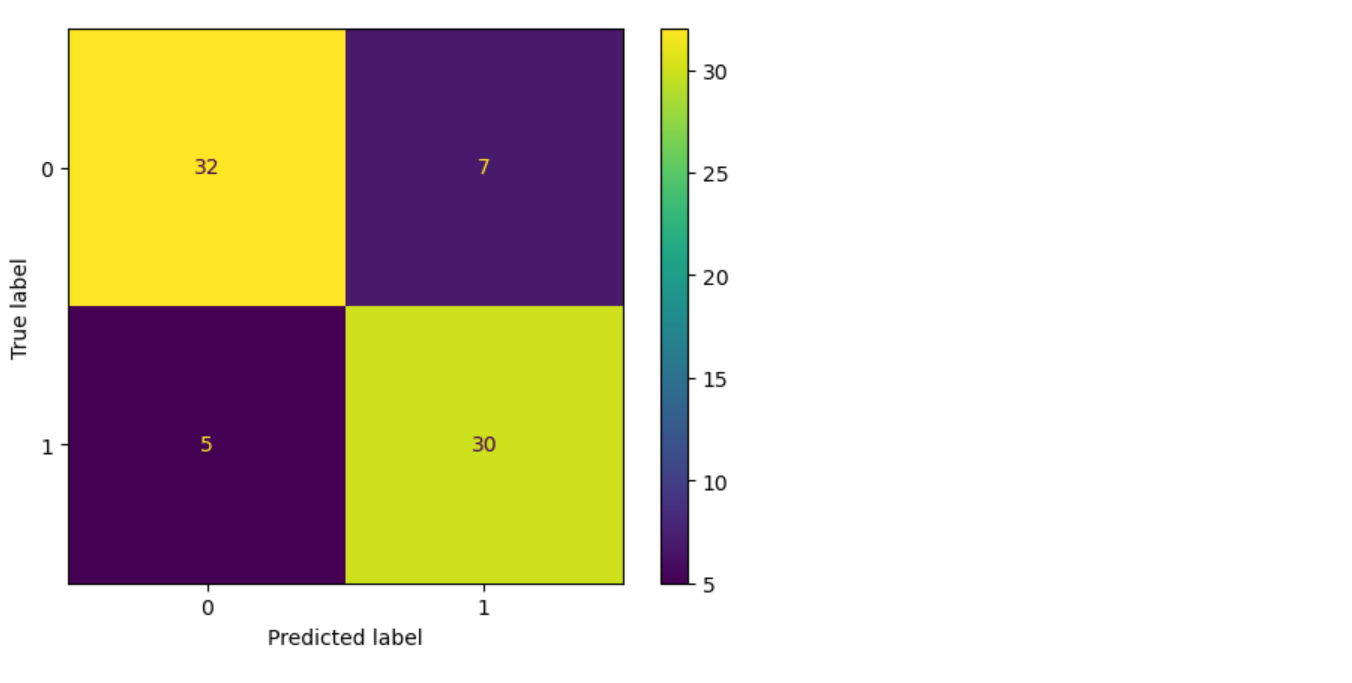
2.5.2. Parameter tuning***.***

Để tiếp tục cải thiện hiệu quả mô hình, nhóm tiến hành thay đổi các tham số cho thuật toán Logistic Regression, bằng cách sử dụng kĩ thuật Randomized Search để tiến hành chọn ra bộ tham số tốt nhất cho mô hình. Kết quả cho ra bộ tham số như sau:



Nhóm tiếp tục sử dụng bộ tham số đã tìm được ở trên để đo độ chính xác tốt nhất của mô hình Logistic Regression, đồng thời so sánh với kết quả chạy ban đầu xem độ đo đã được cải thiện như thế nào.





Có vẻ như việc thay đổi tham số không cải thiện được thêm hiệu suất của mô hình, chứng tỏ ngay từ lần chạy đầu tiên, chúng ta đã có được kết quả dự đoán đáng tin cậy

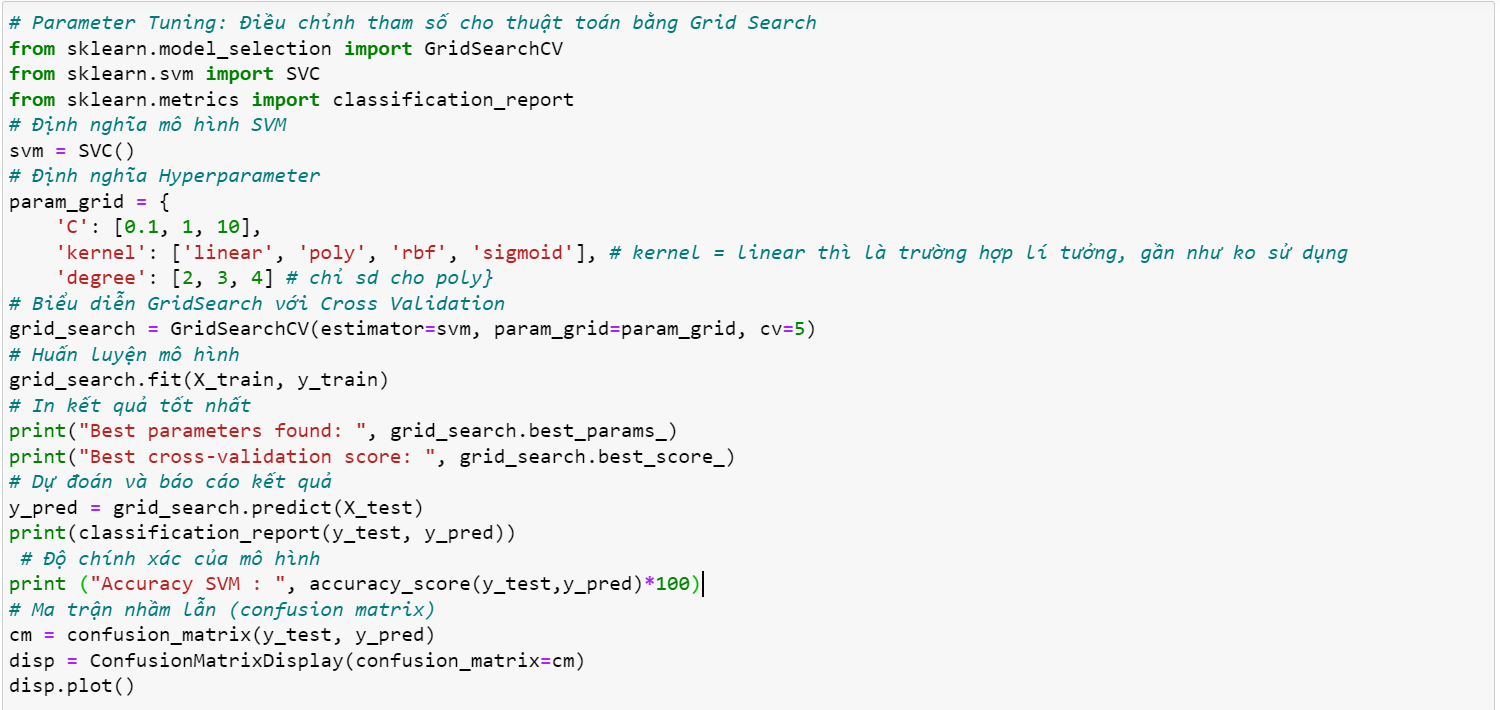
2.5.3. Phát triển thêm mô hình dự đoán bằng các thuật toán khác***.***

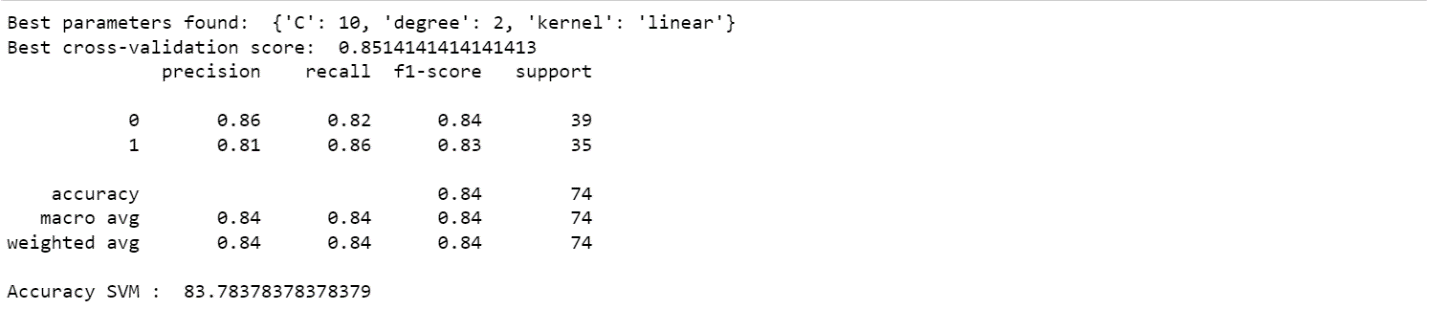
*a, Với thuật toán SVC.*

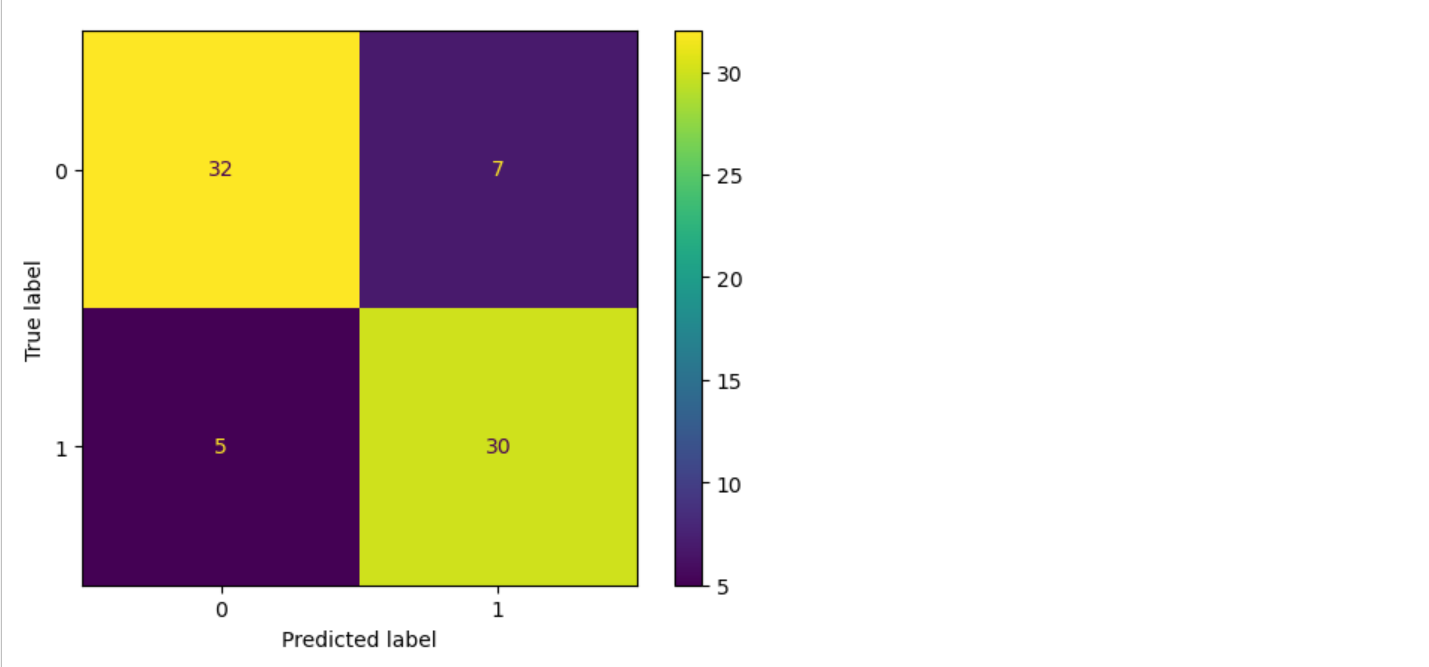
Trong phần đánh giá kết quả các độ đo trên, chúng ta đã lựa chọn thuật toán Logistic Regression để tiếp tục cải thiện hiệu suất mô hình do có độ Recall và F1 cao nhất. Giờ với thuật toán SVC, thuật toán này cũng cho độ đo Recall rất cao và ngang bằng Logistic Regression. Nhóm tiến hành tìm và chạy lại thuật toán này theo bộ tham số mới tốt nhất, xem có thể cải thiện hơn hiệu suất mô hình so với Logistic Regression đã cải tiến hay không. Lưu ý rằng SVC có một số siêu tham số quan trọng như siêu tham số C, kernel, và gamma. Việc lựa chọn các giá trị phù hợp cho các siêu tham số này có thể ảnh hưởng đến hiệu suất của mô hình. Nếu không điều chỉnh siêu tham số tốt, SVC có thể không hoạt động hiệu quả.

Đầu tiên, nhóm khai báo thuật toán SVC với hyperparameter có kernel=linear (trường hợp lý tưởng của SVC). Tiếp theo, nhóm sử dụng kĩ thuật Grid Search với Cross-Validation để truy tìm bộ tham số tốt nhất cho thuật toán này, đoạn code xử lý và kết quả thu được

như sau:







Đánh giá ban đầu cho thấy, sau khi thay đổi với bộ tham số tốt nhất, SVC vẫn chỉ đạt độ đo Recall và F1 bằng với Logistic Regression, tức là sau khi cải thiện cả 2 thuật toán trên, mô hình không thể cải thiện hiệu suất hơn nữa. Tuy nhiên, điều mà bộ tham số mới này làm được, là nâng độ chính xác dự đoán của mô hình SVC lên rất nhiều, từ 64,8% lên 83,78%.

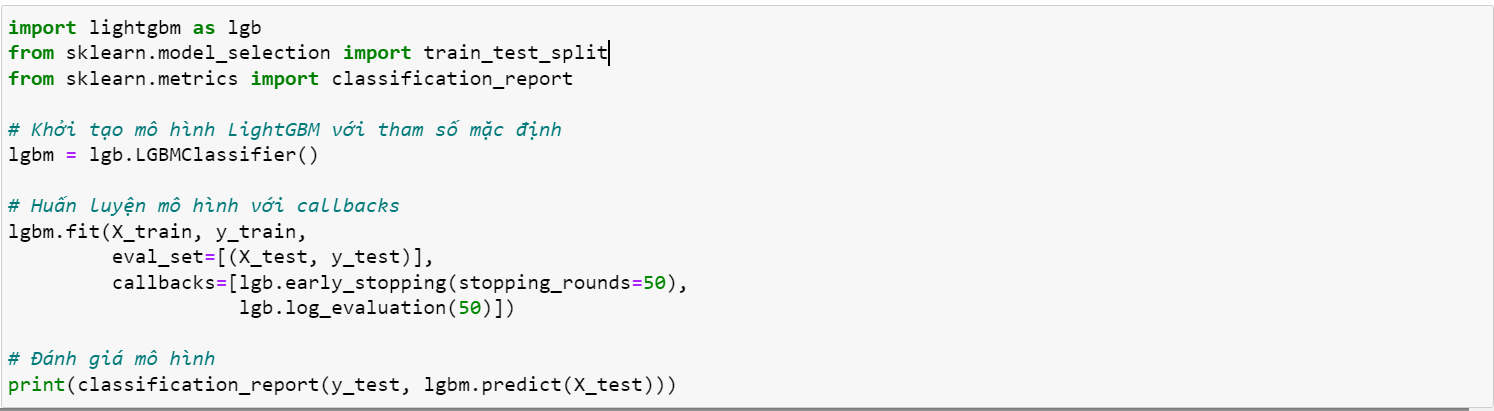
*b, Với thuật toán LightGBM.*

Ngoài các thuật toán phân lớp truyền thống trên, nhóm đề xuất thực hiện thêm 1 thuật toán ML là LightGBM để xem mô hình học máy có thể cải thiện hiệu suất hơn nữa không.

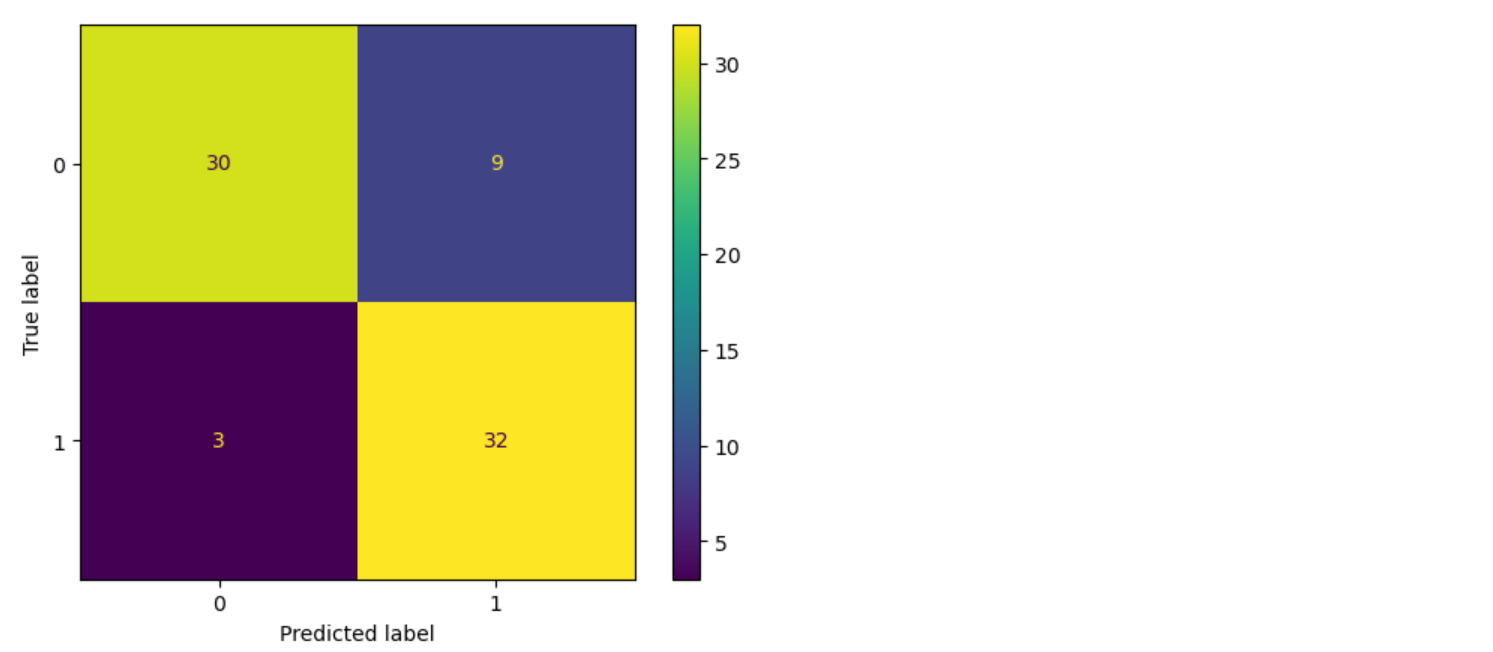
Một số thông tin về thuật toán LightGBM:

* Là một thuật toán học máy dựa trên Gradient Boosting, được thiết kế để cải thiện hiệu suất tính toán và khả năng mở rộng so với các thuật toán boosting truyền thống. Đây cũng là mô hình mạnh mẽ và có hiệu suất cao, phù hợp với dữ liệu đầu vào và ý nghĩa của bài toán.
* Sử dụng phương pháp boosting dựa trên gradient để xây dựng mô hình, nơi các cây quyết định được xây dựng tuần tự, và mỗi cây cố gắng sửa chữa các lỗi của cây trước đó. Một số kĩ thuật tiêu biểu như: Leaf-wise Growth, Histogram-based Decision Tree Algorithm, GOSS và EFB.
* Các tham số quan trọng của LightGBM: num\_leaves, max\_depth, learning\_rate, n\_estimators, boosting\_type.

Code xử lý việc khai báo và đánh giá độ đo của mô hình LightGBM:







Có thể thấy, với thuật toán LightGBM, ta có thể làm cải thiện đáng kể độ đo Recall lên 89%, đồng thời giảm số lượng False Negative xuống còn 4 trường hợp (so với 5 khi chạy mô hình Logistic Regression). Tuy nhiên lại vô tình làm giảm số lượng True Positive từ 32 xuống còn 31 trường hợp và độ chính xác Accuracy cũng thấp hơn Logistic Regression một chút.

## 2.6. Triển khai mô hình

2.6.1. Streamlit.

Streamlit là một framework mã nguồn mở được sử dụng để xây dựng và triển khai các ứng dụng web tương tác cho phân tích dữ liệu và học máy. Streamlit được thiết kế để dễ sử dụng, cho phép người dùng, đặc biệt là các Machine Learning Engineer, tạo ra các giao diện người dùng tùy chỉnh mà không cần phải có kiến thức sâu về việc sử dụng HTML, CSS hay JavaScript.

Streamlit được thành lập vào năm 2018 bởi Adrien Treuille, Thiago Teixeira và Amanda Kelly. Các nhà sáng lập của Streamlit có mục tiêu tạo ra một công cụ giúp các nhà khoa học dữ liệu và kỹ sư học máy dễ dàng chia sẻ kết quả nghiên cứu và mô hình của họ thông qua các ứng dụng web tương tác. Streamlit đã nhanh chóng thu hút được sự chú ý của cộng đồng nhờ tính đơn giản và linh hoạt của nó. Trong vòng một vài năm sau khi ra mắt, Streamlit đã trở thành một công cụ phổ biến trong lĩnh vực khoa học dữ liệu và học máy. Các nhà phát triển đã liên tục cải tiến và mở rộng các tính năng của Streamlit, giúp nó trở thành một lựa chọn hàng đầu cho việc tạo ra các dashboard và ứng dụng dữ liệu.

2.6.2. Áp dụng Streamlit vào việc triển khai mô hình

Nhóm của bọn em đã sử dụng Streamlit để tạo ra một ứng dụng cho phép người dùng nhập các thông số y tế của một bệnh nhân như tuổi, huyết áp, cholesterol, v.v., và nhận được dự đoán về nguy cơ mắc bệnh tim ngay lập tức. Sau đây là các bước để setup Streamlit cho bài toán trên:

Bước 1: Xuất mã từ Jupyter Notebook sang file Python và lưu trữ mô hình.

* Chúng ta cần lấy mã đã viết trong file ipynb, nơi mà mô hình đã được huấn luyện, để chuyển đổi nó thành file py để giúp ta dễ dàng lưu trữ mô hình sau khi được huấn, sử dụng thư viện pickle, và chạy file.
* Ta có thể sử dụng Jupytext để có thể dễ dàng chuyển đổi file ipynb thành file py.

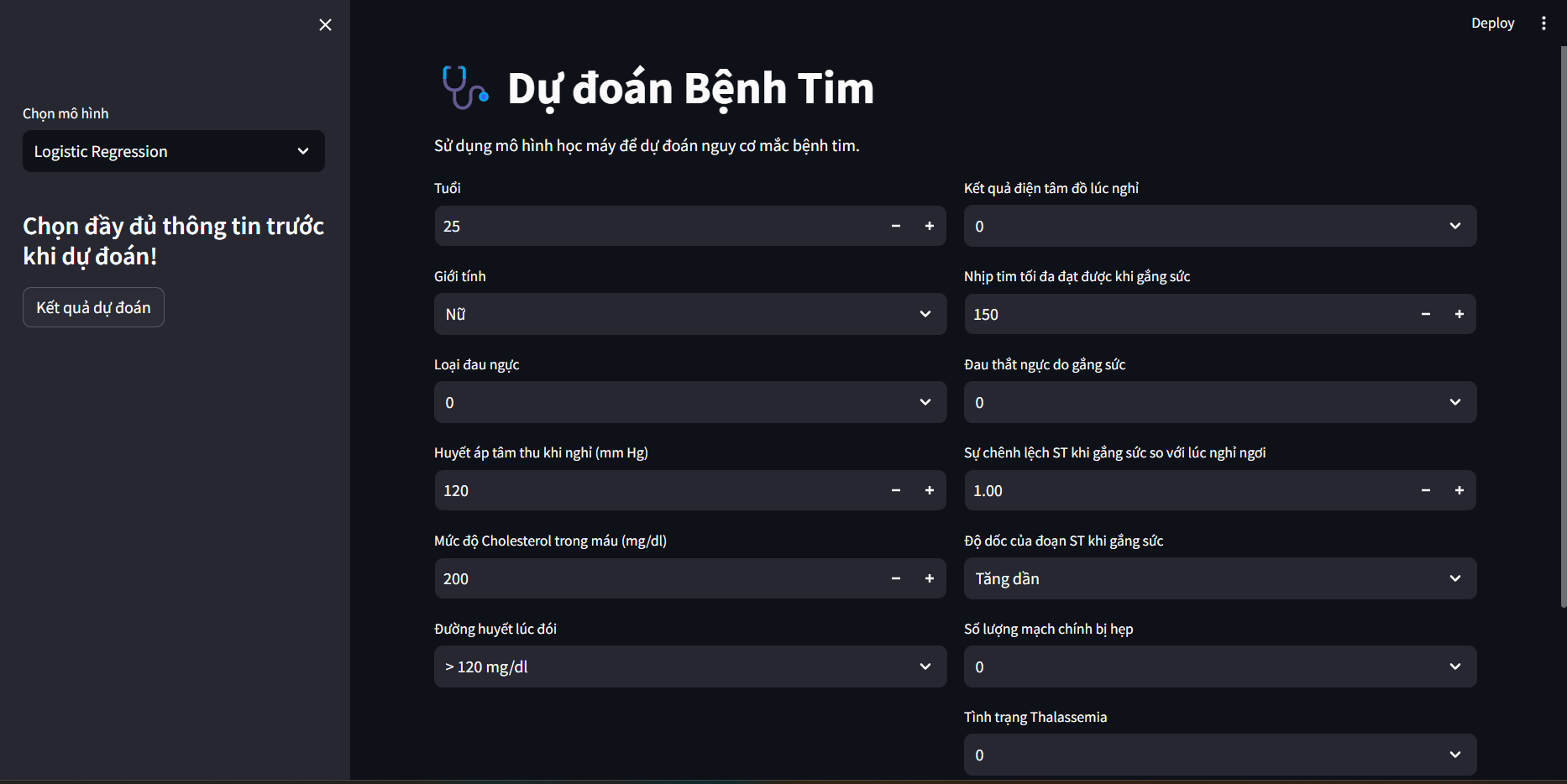
Bước 2: Tạo ứng dụng Streamlit để sử dụng mô hình đã lưu.

Viết code Python cho ứng dụng Streamlit để xây dựng giao diện người dùng cho phép nhập, chọn các thông số y tế và sử dụng mô hình đã lưu để dự đoán kết quả của bệnh tim.

Bước 3: Chạy ứng dụng Streamlit để dự đoán bệnh tim.

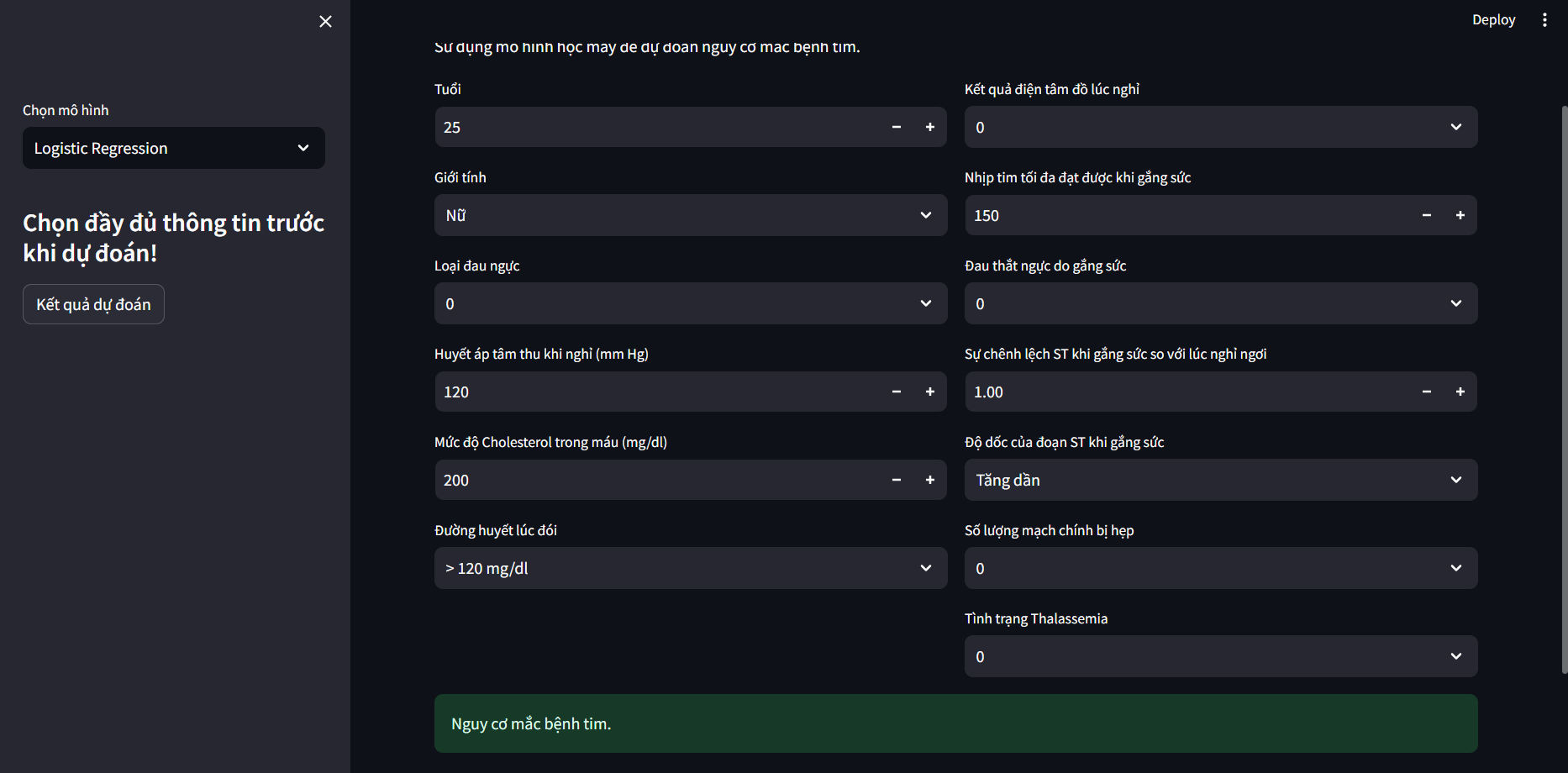
* Mở cmd và nhập lệnh streamlit run để bắt đầu chạy ứng dụng.

*Hình: Giao diện của ứng dụng Dự đoán bệnh tim.*



* Sau khi đã nhập đầy đủ các thông tin mà bài toán yêu cầu, nhấn Kết quả dự đoán để nhận về kết quả của bài toán.

*Hình: Ứng dụng trả về kết quả.*



# KẾT LUẬN

Qua bài nghiên cứu về việc xây dựng mô hình dự đoán khả năng mắc bệnh tim của một bệnh nhân, với nhiều kỹ thuật và phương pháp xử lý khác nhau, nhóm có thu được một số những điểm chính đáng khích lệ như:

* Sử dụng phương pháp Rename để giúp người sử dụng mô hình hiểu rõ hơn về các thuộc tính có trong tập dữ liệu, thay vì phải đọc những từ viết tắt và chuyên ngành của y học rất khó hiểu.
* Trích chọn được mẫu những đặc trưng quan trọng ảnh hưởng lớn nhất tới khả năng mắc bệnh tim của một bệnh nhân, hơn nữa có thể đánh giá mức độ tương quan giữa các thuộc tính đó.
* Kết hợp nhiều thuật toán và độ đo khác nhau để đánh giá hiệu suất của mô hình, có cân nhắc từ ý nghĩa nhân đạo của bài toán để ưu tiên độ đo Recall và F1 score, giúp cân bằng tập dữ liệu có khả năng mất cân bằng và giảm thiểu tối đa khả năng dự đoán sai 1 người không mắc bệnh thành có bệnh (False Positive), đặc biệt là nâng cao tỷ lệ dự đoán đúng chuẩn 1 bệnh nhân mắc bệnh thực sự (True Positive).

Tuy nhiên bên cạnh đó cũng còn nhiều vấn đề chưa được tối ưu nhất hoặc những hạn chế nhất định mà nhóm chưa cải thiện được:

* Tập dữ liệu được sử dụng quá nhỏ, có khả năng mất cân bằng dữ liệu, chưa thể ứng dụng vào các dự án nghiên cứu lớn trong thực tiễn.
* Các độ đo được lựa chọn đang bỏ qua các trường hợp một người mắc bệnh thực sự nhưng bị bỏ sót (False Negative).
* Độ chính xác của mô hình mới chỉ dừng lại ở mức đáng tin cậy (sau cải thiện tham số đạt 86%), cần tiếp tục nghiên cứu và phát triển sâu thêm.