**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC VIỆT NHẬT**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**BÁO CÁO CUỐI KỲ**

**DỰ ĐOÁN BỆNH TIM BẰNG MÔ HÌNH HỌC MÁY**

MÔN HỌC: HỌC MÁY

Năm 2024

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn :** |  |
| **Sinh viên:** | Nguyễn Duy Tùng |
| **Mã sinh viên** | 21110110 |
| **Lớp** | BCSE K1 |

*Hanoi, 6/2024*

**MỤC LỤC**

**[I. ĐẶT VẤN ĐỀ](#_gc702thxjlbl)** 2

[1. Thực trạng mắc bệnh tim hiện nay](#_jk7i1pv003e7) 2

[2. Mục tiêu nghiên cứu](#_2n5351qqmulj) 3

**[II. THUẬT TOÁN LOGISTICEREGRESSION](#_elwa7dxntiq3)** 4

[1. Giới thiệu](#_hykvjrv3amhn) 4

[2. LogisticRegression(Hồi quy Logistic)](#_a5cfv0jug9vk) 4

[a. Cách thức hoạt động](#_i8obm6hcbe3q) 4

[b. Lợi ích](#_km60p1mg4dfo) 7

[c. Thách thức](#_km60p1mg4dfo) 7

[d. Ứng dụng](#_km60p1mg4dfo) 8

**[III. KẾ HOẠCH THỰC HIỆN](#_h9plkzojyo9g)** 10

[1. Thu thập dữ liệu](#_ciko7i2ys1qi) 10

[2. Xử lý dữ liệu](#_p30g4fb4xoqe) 12

[3. Xây dựng mô hình](#_sqvb67rywp8z) 14

[4. Thuật toán code](#_sofmzxq93bvf) 15

**[IV. ĐÁNH GIÁ HIỆU SUẤT](#_vsznt9mpag0s)** 16

**[V. KẾT LUẬN](#_d3v5vgyoh9pm)** 19

**[TÀI LIỆU THAM KHẢO](#_l6jg4akfh3j9)** 20

# **ĐẶT VẤN ĐỀ**

### **Thực trạng măc bệnh tim hiện nay**

Bệnh tim đang là một trong những nguy cơ gây tử vong phổ biến nhất hiện nay ở người. Theo Tổ chức Y tế thế giới (WHO), bệnh tim đang có nguy cơ hàng đầu gây tử vong trên toàn thế giới với hơn 17 triệu người tử vong mỗi năm do các bệnh liên quan đến tim mạch, chiếm khoảng 31% tổng số ca tử vong trên toàn cầu.

Ở các nước phát triển, mặc dù tỷ lệ mắc bệnh tim cao nhưng hệ thống y tế tiên tiến giúp phát hiện và điều trị sớm. Các chiến dịch nâng cao nhận thức và thay đổi lối sống đã giúp giảm phần nào tỷ lệ tử vong. Tuy nhiên tại các **nước đang phát triển,** tỷ lệ mắc bệnh tim không có dấu hiệu ngừng gia tăng do lối sống thay đổi, gia tăng đô thị hóa và thiếu cơ sở hạ tầng y tế. Việc phát hiện và điều trị kém hiệu quả dẫn đến tỷ lệ tử vong cao.

Bệnh tim và các biến chứng có ảnh hưởng rất lớn đến kinh tế của người bệnh. Điều trị bệnh tim bao gồm các chi phí trực tiếp như khám bệnh, xét nghiệm, phẫu thuật, thuốc men và theo dõi điều trị. Những phương pháp này thường đòi hỏi công nghệ cao và chi phí lớn. Những chi phí gián tiếp liên quan như việc mất năng suất lao động do bệnh tật, nghỉ ốm, và giảm hiệu suất làm việc. Điều này cũng làm cho gia đình của người bệnh phải chi tiêu một phần lớn thu nhập vào các dịch vụ y tế và việc điều trị cho người bệnh. Áp lực kinh tế là rất lớn với gia đình của người bệnh khi trong một số trường hợp, chi phí điều trị cao có thể dẫn đến việc vay nợ hoặc bán tài sản, gây ra khó khăn tài chính lâu dài. Nó còn có thể ảnh hưởng đến kinh tế cả một quốc gia khi bệnh tim có thể làm giảm năng suất lao động, ảnh hưởng đến kinh tế tổng thể.

Bệnh tim cũng tác động tiêu cực đến đời sống xã hội của người bệnh và những người xung quanh. Nó làm chất lượng cuộc sống của người bệnh giảm sút khi người bệnh khó khăn trong các hoạt động hàng ngày, từ công việc giải trí đến các hoạt động xã hội. Không chỉ ảnh hưởng đến thể chất, bệnh tim còn gây lo lắng và căng thẳng cho người bệnh và người thân của họ. Bệnh tim cũng làm đảo lộn các vai trò trong gia đình khi người bệnh có thể không thực hiện được các trách nhiệm gia đình và gây áp lực này lên các thành viên khác. Nếu nặng hơn, gia đình và bạn bè của người bệnh còn cần giành thời gian và nguồn lực để chăm sóc người bệnh, gây ảnh hưởng đến cuộc sống của họ.

### **Mục tiêu nghiên cứu**

**Mục tiêu 1:** Cải thiện độ chính xác và hiệu quả trong việc chuẩn đoán khả năng mắc bệnh tim. Các mô hình học máy có thể phân tích được một lượng dữ liệu lớn và có độ chính xác cao mà con người có thể có sự thiếu sót và nhầm lẫn, từ đó nâng cao chất lượng chuản đoán. Ngoài ra các mô hình còn có thể giúp phát hiện khả năng mắc bệnh từ rất sớm khi các dữ liệu có thể chưa rõ ràng ở người bệnh giúp phát hiện và điều trị kịp thời.

**Mục tiêu 2:** Tạo ra một thiết bị có thể hỗ trợ các y bác sĩ trong công việc. Các mô hình học máy sẽ là thiết bị hỗ trợ các bác sĩ trong việc chuẩn đoán bệnh tim, vừa nâng cao độ chính xác vừa giúp các bác sĩ tự động hóa quát trình phân tích dữ liệu y tế, giúp giảm tải công việc để họ có thể tập trung vào việc điều trị và chăm sóc cho bệnh nhân.

**Mục tiêu 3:** Tăng cường khả năng tiếp cận với việc được chăm sóc sức khỏe đến tất cả mọi người. Khi mô hình máy học có hiệu quả đủ tốt và được phát triển rộng rãi, nó sẽ được mang tới những khu vực thiếu nguồn lực y tế, giúp nhiều người sẽ được tiếp cận với việc chuẩn đoán bệnh tim chất lượng cao. Điều này còn giúp giảm chi phí chuẩn đoán và chữa bệnh khi tối ưu hóa và sử dụng nguồn lực y tế hợp lý.

**Mục tiêu 4:** Cải thiện chất lượng cuộc sống và kéo dài tuổi thọ cho con người. Việc dự đoán khả năng mắc bệnh tim càng sớm càng giúp giảm biến chứng và thiệt hại do loại bệnh này gây ra khi đây là bệnh có tỉ lệ tử vong rất cao của loài người. Từ đó có thể cải thiện chất lượng cuộc sống và kéo dài được tuổi thọ của người bệnh.

**Mục tiêu 5:** Đóng góp vào phát triển Y học. Cung cấp cho Y học con người một công cụ hỗ trợ việc phân tích dữ liệu và chuẩn đoán bệnh tim giúp phát triển các phương pháp điều trị và phòng ngừa. Việc ứng dụng trí tuệ nhân tạo vào y tế sẽ thúc đẩy sự đổi mới và phát triển không ngừng, mang lại những tiến bộ trong dịch vụ chăm sóc sức khỏe con người.

# **THUẬT TOÁN** **LOGISTIC REGRESSION**

### **Giới thiệu**

****Logistic Regression**** Logistic Regression là một thuật toán học máy cơ bản và phổ biến, chủ yếu được sử dụng cho các bài toán phân loại nhị phân, mặc dù nó cũng có thể được mở rộng cho các bài toán phân loại đa lớp.

Mặc dù tên gọi chứa từ "regression" (hồi quy), logistic regression không phải là một thuật toán hồi quy trong nghĩa truyền thống. Thay vào đó, nó thuộc nhóm thuật toán phân loại, có nhiệm vụ dự đoán xác suất của một biến mục tiêu thuộc về một trong hai lớp.

Ví dụ: giả sử bạn muốn đoán xem khách truy cập trang web của bạn sẽ nhấp vào nút thanh toán trong giỏ hàng của họ hay không. Phân tích hồi quy logistic xem xét hành vi của khách truy cập trước đây, chẳng hạn như thời gian dành cho trang web và số lượng các mặt hàng trong giỏ hàng. Quá trình phân tích này xác định rằng, trước đây, nếu khách truy cập dành hơn năm phút trên trang web và thêm hơn ba mặt hàng vào giỏ hàng, họ sẽ nhấp vào nút thanh toán. Nhờ vào thông tin này, sau đó, hàm hồi quy logistic có thể dự đoán hành vi của một khách mới truy cập trang web.

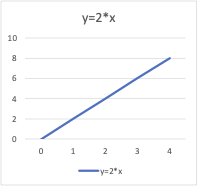
### **Logistic Regression (Hồi quy Logistic)**

1. **Cách hoạt động**

Để hiểu rõ về mô hình hồi quy logistic, trước tiên chúng ta phải hiểu các phương trình và biến.

### **\* Phương trình:**

Trong toán học, phương trình cho ta mối quan hệ giữa hai biến: x và y. Bạn có thể sử dụng các phương trình hoặc hàm này để vẽ đồ thị theo trục x và trục y bằng cách nhập các giá trị khác nhau của x và y. Ví dụ: nếu bạn vẽ đồ thị cho hàm y = 2\*x, bạn sẽ có một đường thẳng như hình dưới đây. Do đó hàm này còn được gọi là hàm tuyến tính.



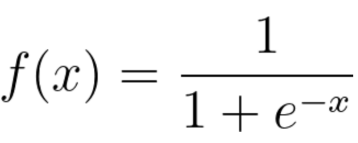
### **\* Biến:**

Trong thống kê, biến là các yếu tố dữ liệu hoặc thuộc tính có giá trị khác nhau. Bất kỳ phân tích nào cũng có một số biến nhất định là biến độc lập hoặc biến giải thích. Những thuộc tính này là nguyên nhân của một kết quả. Các biến khác là biến phụ thuộc hoặc biến đáp ứng; giá trị của chúng phụ thuộc vào các biến độc lập. Nhìn chung, hồi quy logistic khám phá cách các biến độc lập ảnh hưởng đến một biến phụ thuộc bằng cách xem xét các giá trị dữ liệu lịch sử của cả hai biến.

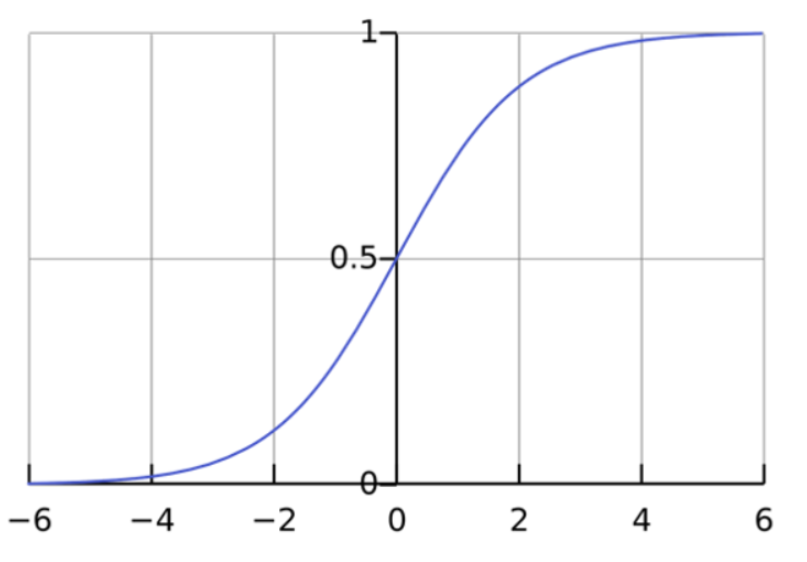
Trong ví dụ ở trên của chúng tôi, x được gọi là biến độc lập, biến dự đoán hoặc biến giải thích vì nó có một giá trị đã xác định. Y được gọi là biến phụ thuộc, biến kết quả hoặc biến đáp ứng vì giá trị của nó không xác định.

### **\* Hàm hồi quy logistic**

Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x.



Nếu vẽ phương trình hồi quy logistic này, bạn sẽ có một đường cong hình chữ S như hình dưới đây.



Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

### **\* Phân tích hồi quy logistic với nhiều biến độc lập**

Trong nhiều trường hợp, nhiều biến giải thích ảnh hưởng đến giá trị của biến phụ thuộc. Để lập mô hình các tập dữ liệu đầu vào như vậy, công thức hồi quy logistic phải giả định mối quan hệ tuyến tính giữa các biến độc lập khác nhau. Bạn có thể sửa đổi hàm sigmoid và tính toán biến đầu ra cuối cùng như sau

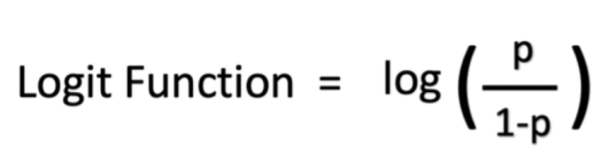
y = f(β0 + β1x1 + β2x2+… βnxn)

Ký hiệu β đại diện cho hệ số hồi quy. Mô hình logit có thể đảo ngược tính toán các giá trị hệ số này khi bạn cho nó một tập dữ liệu thực nghiệm đủ lớn có các giá trị đã xác định của cả hai biến phụ thuộc và biến độc lập.

### **\* Log của tỷ số odds**

Mô hình logit cũng có thể xác định tỷ số thành công trên thất bại hay log của tỷ số odds. Ví dụ: nếu bạn đang chơi poker với bạn bè và thắng bốn ván trên mười ván, tỷ số chiến thắng của bạn là bốn phần sáu, hoặc 4/6, và đó là tỷ số thành công trên thất bại của bạn. Mặt khác, xác suất thắng là 4/10.

Về mặt toán học, tỷ số odds về mặt xác suất của bạn là p/(1 - p) và log của tỷ số odds là log (p/(1 - p)). Bạn có thể biểu diễn hàm logistic bằng log của tỷ số odds như hình dưới đây:



1. **Lợi ích**

**Tính đơn giản:** Các mô hình hồi quy logistic ít phức tạp về mặt toán học hơn các phương pháp ML khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về ML.

**Tốc độ:** Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi quy logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án ML để đạt được một số thành tựu nhanh chóng.

**Sự linh hoạt:** Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu.

**Khả năng hiển thị:** Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

1. **Thách thức**

**Giới Hạn trong Phân Loại Tuyến Tính:** Logistic Regression mô hình hóa mối quan hệ tuyến tính giữa các đặc trưng và xác suất phân loại. Điều này có nghĩa là mô hình không thể xử lý các mối quan hệ phi tuyến tính giữa đặc trưng và mục tiêu, và có thể không phù hợp cho các bài toán mà mối quan hệ này phức tạp.

**Nhạy Cảm với Outliers:** Logistic Regression có thể nhạy cảm với dữ liệu ngoại lệ (outliers), điều này có thể làm giảm hiệu suất của mô hình. Các điểm dữ liệu ngoại lệ có thể ảnh hưởng đến ước lượng của các tham số, dẫn đến sự biến dạng của đường quyết định.

**Phân Loại Đa Lớp:** Logistic Regression ban đầu được thiết kế cho các bài toán phân loại nhị phân. Mặc dù có thể mở rộng cho phân loại đa lớp bằng cách sử dụng kỹ thuật One-vs-Rest (OvR) hoặc Softmax Regression, nhưng việc mở rộng này có thể không hiệu quả như các mô hình đa lớp khác như Random Forest hoặc Neural Networks.

**Khó Khăn trong Đối Phó với Sự Mất Cân Đối Dữ Liệu:** Nếu tập dữ liệu bị mất cân đối, có nghĩa là số lượng mẫu trong mỗi lớp không đồng đều, Logistic Regression có thể gặp khó khăn trong việc tạo ra một mô hình phân loại có hiệu suất cao. Các kỹ thuật như oversampling hoặc undersampling có thể cần được sử dụng để xử lý vấn đề này.

**Tính Tốt với Dữ Liệu Lớn:** Logistic Regression có thể không phù hợp với các tập dữ liệu lớn và phức tạp. Việc huấn luyện mô hình với dữ liệu lớn có thể tốn kém và đòi hỏi nhiều tài nguyên tính toán.

**Cần Chuẩn Hóa Dữ Liệu:** Logistic Regression yêu cầu các đặc trưng đầu vào được chuẩn hóa trước khi sử dụng. Nếu các đặc trưng có phạm vi hoặc độ biến động khác nhau, mô hình có thể không hoạt động hiệu quả.

1. **Ứng dụng**

### **Sản xuất:** Các công ty sản xuất áp dụng phân tích hồi quy logistic để ước tính xác suất xảy ra sự cố ở bộ phận trong máy móc. Sau đó, họ sẽ lên lịch bảo trì dựa trên xác suất đã ước tính này để giảm thiểu sự cố trong tương lai.

### **Y tế:** Các nhà nghiên cứu y khoa lên kế hoạch điều trị và chăm sóc dự phòng bằng cách dự đoán khả năng mắc bệnh ở bệnh nhân. Họ sử dụng các mô hình hồi quy logistic để so sánh tác động của tiền sử gia đình hoặc của bộ gen lên bệnh tật.

### **Tài chính:** Các công ty tài chính phải phân tích các giao dịch tài chính để đề phòng gian lận, xem xét các đơn xin vay và đơn bảo hiểm để đề phòng rủi ro. Những vấn đề này phù hợp với mô hình hồi quy logistic bởi chúng có kết quả cụ thể, chẳng hạn như rủi ro cao hoặc rủi ro thấp và gian lận hoặc không gian lận.

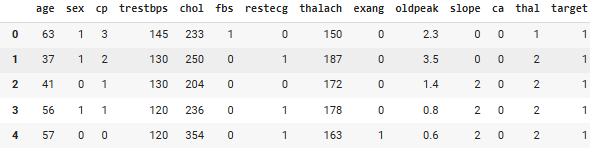
### **Bộ phận Tiếp thị:** Các công cụ quảng cáo trực tuyến sử dụng mô hình hồi quy logistic để dự đoán xem người dùng sẽ nhấp vào một quảng cáo hay không. Kết quả là, các nhà tiếp thị có thể phân tích phản ứng của người dùng đối với những từ ngữ và hình ảnh khác nhau, tạo ra các quảng cáo hiệu suất cao có khả năng thu hút khách hàng.

# **KẾ HOẠCH THỰC HIỆN**

### **Thu thập dữ liệu**

#### File dữ liệu: “heart.csv” được lấy từ trang Kaggle.com

* **Tổng quan hóa bộ dữ liệu:**



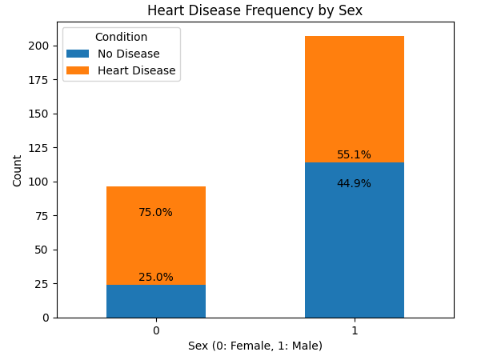
Hình 3.1: Bảng thống kê mô tả bộ dữ liệu

* Bộ dữ liệu gồm: 302 mẫu và 14 thuộc tính (trong đó thuộc tính target sẽ là nhãn của các bệnh nhân)
* **Dữ liệu gồm 14 thuộc tính:**

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Ý nghĩa |
| 1 | age | Tuổi của bệnh nhân |
| 2 | sex | Giới tính của bệnh nhân |
| 3 | cp | 4 loại đau ngực:   * 0: Đau thắt ngực điển hình: đau ngực liên quan đến giảm cung cấp máu cho tim * 1: Đau thắt ngực không điển hình: đau ngực không liên quan đến tim * 2: Đau không phải do thắt ngực: thường là co thắt thực quản (không liên quan đến tim) * 3: Không có triệu chứng: đau ngực không có dấu hiệu bệnh |
| 4 | trestbps | Huyết áp |
| 5 | chol | Lượng cholesterol trong máu (mg/dl) |
| 6 | fbs | Lượng đường trong máu khi đói |
| 7 | restecg | Kết quả điện tâm đồ khi nghỉ ngơi   * 0: Không có gì đáng chú ý * 1: Bất thường sóng ST-T   + Có thể từ triệu chứng nhẹ đến vấn đề nghiêm trọng   + Báo hiệu nhịp tim không bình thường * 2: Có thể hoặc chắc chắn phì đại tâm thất trái   + Buồng bơm chính của tim bị phì đại |
| 8 | thalach | Nhịp tim tối đa đạt được |
| 9 | exang | Đau thắt ngực khi tập thể dục (1 = có; 0 = không) |
| 10 | oldpeak | Mức độ trầm cảm ST do tập thể dục so với khi nghỉ ngơi   * Xem mức độ căng thẳng của tim khi tập thể dục * Tim không khỏe sẽ căng thẳng hơn |
| 11 | slope | Độ dốc của đoạn ST khi tập thể dục   * 0: Dốc lên: nhịp tim tốt hơn khi tập thể dục (hiếm gặp) * 1: Dốc ngang: thay đổi tối thiểu (tim khỏe mạnh điển hình) * 2: Dốc xuống: dấu hiệu của tim không khỏe |
| 12 | ca | Số lượng mạch chính (0-3) được hiển thị qua phương pháp chụp huỳnh quang   * Mạch máu có màu có nghĩa là bác sĩ có thể thấy máu chảy qua * Càng nhiều máu lưu thông càng tốt (không có cục máu đông) |
| 13 | thal | Mức độ khuyết tật của tim:   * 1: bình thường * 2: khuyết tật cố định: từng bị khuyết tật nhưng giờ đã ổn * 3: khuyết tật có thể hồi phục: không có sự di chuyển máu đúng cách khi tập thể dục |
| 14 | target | Có bệnh hay không (1=có, 0=không) |

### **Phân tích dữ liệu**

* **Tỉ lệ mắc bệnh theo giới tính:**

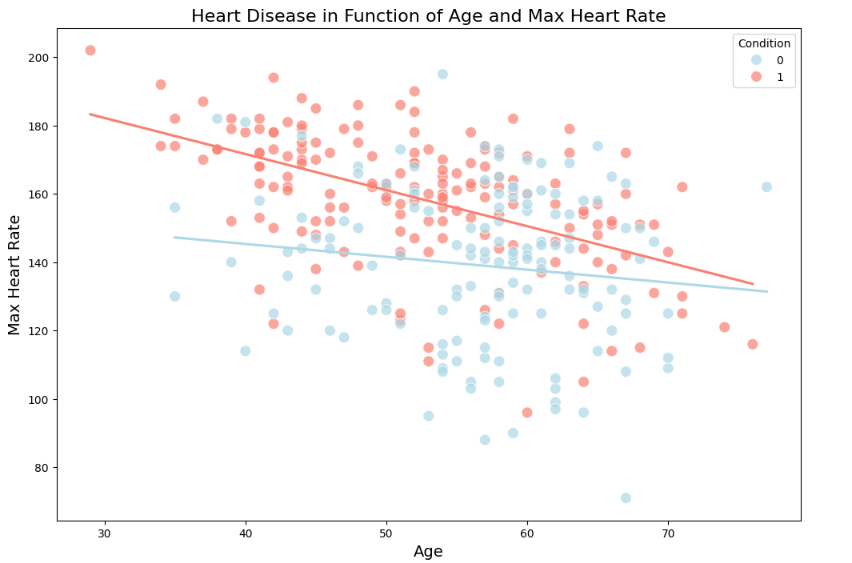


Hình 3.2a: Biểu đồ thể hiện tỷ lệ mắc bệnh tim theo giới tính

Nam có khoảng 200 người với tỉ lệ mắc bệnh khoảng 55.1%, nữ có số người ít hơn khoảng 100 người nhưng tỉ lệ mắc bệnh lại cao hơn rơi vào khoảng 75%.

**Kết luận:** Tỉ lệ mắc bệnh ở nam thấp hơn nhiều so với tỉ lệ mắc bệnh ở nữ với nam chỉ khoảng 55.1% còn nữ là khoảng 75%. Như vậy có thể thấy sự khác biệt về giới tính có ảnh hưởng lớn đến khả năng mắc bệnh tim của con người.

* **Tỷ lệ mắc bệnh tim theo mức độ của nhịp tim tối đa:**

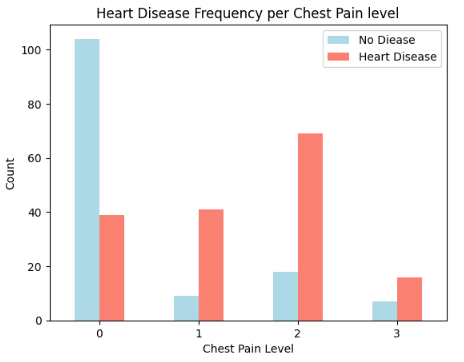


Hình 3.2c: Biểu đồ thể hiện tỷ lệ mắc bệnh tim theo nhịp tim tối đa

Người mắc bệnh tim thường có nhịp tim cao trên 160 và thường ở độ tuổi trên 40 tuổi. Biểu đồ chưa có sự trực quan khi số người ở trên 70 tuổi khá ít trong dữ liệu nên có cảm giác người trên 70 tuổi mắc bệnh ít hơn so với người ở khoảng 40-60 tuổi.

**Kết luận:** Khi nhịp tim tối đa tăng khả năng mắc bệnh về tim của bệnh nhân cũng tăng.

* **Tỷ lệ mắc bệnh theo mức độ đau tim:**



Hình 3.2b: Biểu đồ thể hiện tỷ lệ mắc bệnh tim theo kiểu đau tim

**Kết luận:** Những người có biểu hiện bị đau ở mức độ 0 có số lượng nhiều nhất tuy nhiên khả năng mắc bệnh tim khá thấp so với những người có biểu hiện đau mức độ 1, mức độ 2 và mức độ 3.

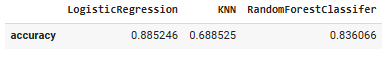
### **Xây dựng mô hình**

* Khai báo các thư viện: ****numpy** (**np**),  **pandas** (**pd**), **matplotlib** (**plt**), **seaborn** (**sns**)**
* Khai báo model sử dụng:

**- **LogisticRegression****

**- **KNeighborsClassifier** và **RandomForestClassifier** (đánh giá hiệu suât giữa 3 mô hình nhằm đưa ra mô hình tôi ưu nhất)**

* Khai báo các **Model Evaluations**
* Chia tập dữ liệu thành **X\_train,** **X\_test** với tỉ lệ 80/20 và **y\_train**, **y test**( nhãn của X) với tỉ lệ tương tự.
* Sử dụng cả 3 mô hình với bộ dữ liệu đã được chia

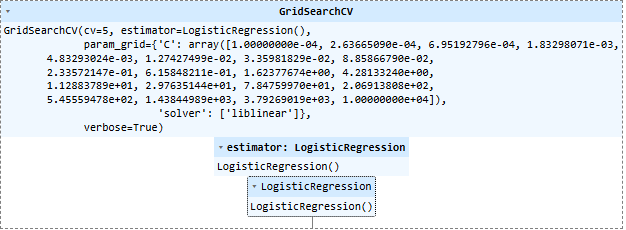


Có thể thấy mô hình **LogisticRegression** cho được độ chính xác tới 88.5% (accuracy = 0.885246). Đây là độ chính xác ở mức tốt và là tốt nhất trong 3 mô hình.

* Tìm siêu tham số tốt nhất cho mô hình:

- Tạo lưới tham số **“log\_reg\_grid”** với 20 giá trị **'C'** chạy từ -4 đến 4, Lựa chọn **‘solver': ['liblinear']** vì phù hợp với các tập dữ liệu nhỏ và bài toán phân loại nhị phân.

- Thiết lập tìm kiếm siêu tham số và huấn luyện mô hình với **X\_train** và **y\_train**.



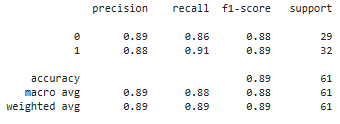
- Sử dụng lệnh **best\_params\_** để tìm ra thông số tối ưu nhất trong các thông số trên ta được:**{'C': 0.23357214690901212, 'solver': 'liblinear'}**

### **Thuật toán code**

Tham khảo tại Đường link liên kết GitHub:

# **IV . ĐÁNH GIÁ HIỆU SUẤT**

Sau khi sử dụng các tham số tối ưu, hiệu suất của mô hình có vẻ không thay đổi với accuracy =0.8852459016393442( 88.5%) vì vậy đây đang đang hiệu suất tốt nhất của mô hình. Cùng đi để kiểm tra và đánh giá các chỉ số khác của mô hình:



Hình 4a: Các chỉ số của mô hình

**\* Precision**:

* Precision là tỷ lệ của số lượng dự đoán đúng thuộc về một lớp cụ thể (true positives) so với tổng số lượng dự đoán thuộc về lớp đó (true positives và false positives).
* Precision = True Positives / (True Positives + False Positives)

Kết luận: Precision cao đồng nghĩa với tỉ lệ ít dự đoán sai dương tính

**\* Recall (Tỷ lệ phục hồi)**:

* Recall là tỷ lệ của số lượng dự đoán đúng thuộc về một lớp cụ thể (true positives) so với tổng số lượng mẫu thực sự thuộc về lớp đó (true positives và false negatives).
* Recall = True Positives / (True Positives + False Negatives)

Kết luận: Recall cao đồng nghĩa với khả năng phát hiện được nhiều mẫu thực sự thuộc về lớp dương tính.

**\* F1-score**:

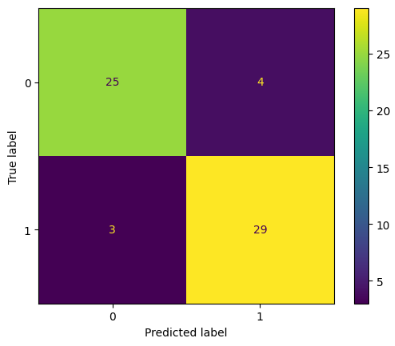
* F1-score là trung bình điều hòa giữa precision và recall. Nó cung cấp một phép đo tổng thể về hiệu suất của mô hình.
* F1-score = 2 \* (Precision \* Recall) / (Precision + Recall)

Kết luận: F1-score thường được sử dụng khi cần cân nhắc cả precision và recall, đặc biệt khi số lượng false negatives và false positives không đồng đều. Với việc F1-score cao tới 0.88 thể hiện rằng mô hình này khá hiệu quả.

**\* Support**:

* Support là số lượng mẫu thực sự thuộc về lớp trong tập dữ liệu kiểm tra.
* Support cung cấp thông tin về mức độ đại diện của mỗi lớp trong dữ liệu kiểm tra.

**\* Bảng confusion matrix:**



Hình 4b: Số lượng đúng sai mỗi lớp

* Từ bảng trên có thể thấy mô hình đã dự đoán đúng 25 trường hợp không mắc bệnh và 29 trường hợp mắc bệnh tim. Trong khi đó mô hình chỉ dự đoán sai 4 trường hợp không mắc bệnh và 3 trường hợp mắc bệnh.
* Mô hinh cho thấy tỉ lệ dự đoán đúng trường hợp không mắc bệnh là khoảng 86% và trường hợp mắc bệnh là 93%, đây là hiệu suất có thể chấp nhận được và được đánh giá là khá tốt.

\* **Ưu điểm:**

* Với việc mô hình có hiệu suất tốt, đây sẽ là công cụ tốt giúp con người có thể dự đoán được mình có mắc bệnh tim hay không một cách sớm nhất, điều này sẽ tăng hiệu quả điều trị trong tương lai.
* Mô hình cũng là công cụ hỗ trợ tốt cho các y bác sĩ trong việc xử lý dữ liệu và chuẩn đoán giúp giảm tải công việc của họ, tập trung hơn vào công cuộc chữa trị cho bệnh nhân.
* Là động lực thúc đẩy công nghệ học máy và trí tuệ nhân tạo vào y học làm cho ngành y học không ngừng phát triển để có thể giúp con người tránh khỏi bệnh tật.

\* **Nhược điểm:**

* Với dữ liệu đầu vào chưa phải quá nhiều, các chuẩn đoán của mô hình có thể thay đổi khi gặp số lượng dữ liệu lớn hơn, chủ yếu là theo chiều hướng tiêu cực.
* Logistic Regression giả định rằng quan hệ giữa các biến đầu vào và xác suất mắc bệnh tim là tuyến tính. Điều này có thể không phù hợp nếu có các quan hệ phi tuyến hoặc phức tạp giữa các biến dẫn đến giảm sự chính xác.
* Nếu dữ liệu không đồng đều giữa các lớp (ví dụ: có nhiều mẫu thuộc lớp một và ít mẫu thuộc lớp khác), Logistic Regression có thể không cho kết quả tốt như các thuật toán khác như Random Forest hoặc Gradient Boosting Machines.

# **KẾT LUẬN**

1. **Kết luận**

Mô hình Logistic Regression đã đạt được một mức độ hiệu suất khá tốt trong việc dự đoán bệnh tim dựa trên tập dữ liệu nhỏ, với precision, recall và F1-score đều ở mức khả quan. Điều này một phần là do dữ liệu đầu vào không quá nhiều nên mô hình vẫn có thể sử dụng được

Tuy nhiên, có giới hạn của mô hình Logistic Regression, bao gồm giả định về tuyến tính, nhạy cảm với outliers và hạn chế trong xử lý dữ liệu không đồng đều.

Để cải thiện hiệu suất của mô hình, ta có thể xem xét việc sử dụng các thuật toán phức tạp hơn như Random Forest hoặc Gradient Boosting Machines, có khả năng học được các mô hình phức tạp hơn và xử lý tốt hơn các đặc điểm phức tạp của dữ liệu.

### **Hướng Đi trong Tương Lai:**

Việc thu thập thêm dữ liệu có thể giúp cải thiện hiệu suất của mô hình, đặc biệt là trong việc xử lý các biến không đồng đều và các trường hợp hiếm gặp.

Nên thử nghiệm và so sánh hiệu suất của các thuật toán học máy khác nhau trên cùng một tập dữ liệu để tìm ra mô hình phù hợp nhất.

Tiến hành tối ưu hóa siêu tham số của các mô hình được chọn để đảm bảo rằng chúng hoạt động hiệu quả nhất trên dữ liệu thực tế.

Nghiên cứu sâu hơn về các đặc điểm của dữ liệu có thể giúp cải thiện hiệu suất của mô hình.

Tăng cường phân tích và đánh giá kết quả của mô hình để hiểu rõ hơn về các yếu điểm và điều chỉnh mô hình một cách linh hoạt.

# **TÀI LIỆU THAM KHẢO**

Tài liệu và sách:

1. Machine Learning with TensorFlow by Aurélien Géron

Video:

1. Machine Learning for Everybody - Full Course.

[JV YTB (youtube.com)](https://www.youtube.com/watch?v=i_LwzRVP7bg&t=82s)

The End