BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT

THÀNH PHỐ HỒ CHÍ MINH

TIỂU LUẬN

NHẬP MÔN LẬP TRÌNH PYTHON CHO PHÂN TÍCH

Giảng viên: Quách Đình Hoàng

Sinh viên thực hiện

STT	Họ và tên	MSSV
1	Trương Hoàng Anh Khôi	20133058
2	Phạm Minh Long	20133062
3	Nguyễn Văn Trường Tốt	20133098
4	Nguyễn Quốc Thắng	20133091

Thủ Đức, tháng 06 năm 2022

MỤC LỤC

1. Tóm tắt

Từ lâu bệnh tim đã được mọi người quan tâm, đã có rất nhiều cuộc nghiên cứu về bệnh tim được thực hiện. Xã hội ngày càng phát triển giúp chát lượng cuộc sống của con người được nâng cao. Thế nhưng mặt trái của sự phát triển này lại để lại hệ lụy môi trường và sức khỏe con người. Trong đó, các bệnh liên quan tới tim mạch ngày càng nhiều và để lại hẩu quả đáng tiếc. Tuy nhiên với sự tiến bộ của y học đã tìm ra nhiều phương pháp chẩn đoán, điều trị hiệu quả hơn, hiện đại hơn giúp phần nào giảm nỗi lo về bệnh tim.

Chính vì thế nhóm chúng em đã chọn lựa chọn đề tài "Dự đoán các khả năng mắc bệnh tim" thông qua tập dữ liệu cho trước để có thể hiểu sâu hơn và tổng hợp lại các triệu chứng và biểu hiện thường gặp của người mắc bệnh tim. Qua đó có thể phần nào phòng tránh cho chính mình và cho những người xung quanh.

Bài toán thuộc prediction nên nhóm sẽ dùng thuật toán hồi quy logictic để phân tích và đưa ra những dự đoán phù hợp để giải quyết các câu hỏi mà nhóm đã đặt ra.

2. Giới thiệu

Theo thống kê, trước đây bệnh tim thường chỉ xuất hiện ở người cao tuổi nhưng hiện nay độ tuổi mắc bệnh tim ngày càng trẻ hóa. Cùng với đó, có rất nhiều nguyên nhân gây ra khả năng mắc bệnh tim như tuổi tác, giới tính, huyết áp khống ổn định, lượng mỡ trong máu, ... Và người mắc bệnh tim có các biểu hiện khác nhau nên việc phát hiện sớm và ngăn chặn là điều vô cùng cần thiết. Điện tâm đồ là đồ thị ghi lại hoạt động của tim nên nó là cơ sở quan trọng để có thể phát hiện những yếu tố bất thường của tim có nguy cơ dẫn đến bệnh. Từ đó, nhóm đã đưa ra các câu hỏi cần phải được làm rõ như sau:

- Những yếu tố nào cho thấy rõ khả năng mắc bệnh tim?
- Độ tuổi có ảnh hưởng đến nguy cơ mắc bệnh tim hay không?
- Trạng thái điện tâm đồ của người mắc bệnh tim có gì bất thường?
- Có sự khác biệt về nguy cơ mắc bệnh tim giữa nam và nữ hay không?

Để trả lời các câu hỏi trên bài toán sử dụng tập các biến giải thích như: độ tuổi, giới tính, huyết áp, nhịp tim,... và thuật toán logistic regression để dự đoán khả năng mắc bệnh tim.

3. Dữ liệu

- Tập dữ liệu để phân tích được lấy từ <https://www.kaggle.com/> và có từ năm 1988. Bộ dữ liệu này có từ năm 1988 và bao gồm bốn cơ sở dữ liệu: Cleveland, Hungary, Thụy Sĩ và Long Beach V. Nó chứa 76 thuộc tính, bao gồm thuộc tính dự đoán, nhưng tất cả các thí nghiệm được công bố đều đề cập đến việc sử dụng một tập hợp con gồm 14 trong số đó. Trường "mục tiêu" đề cập đến sự hiện diện của bệnh tim ở bệnh nhân. Nó có giá trị số nguyên 0 = không có bệnh và 1 = bệnh.

- Trong quá trình nghiên cứu có thể xảy ra các trường hợp:

  - Số các ca dự đoán mắc bệnh và thực tế bị mắc bệnh thật.

  - Số các ca dự đoán không mắc bệnh và thực tế không mắc bệnh thật.

  - Số các ca dự đoán mắc bệnh và thực tế không bị mắc bệnh.

  - Số các ca dự đoán không mắc bệnh và thực tế bị mắc bệnh.

- Các thuộc tính:

  - age: tuổi tính theo năm

  - sex: giới tính

  - cp - 'check_pain_type': các loại đau thắt ngực

  - trestbps - 'resting_bool_pressure': huyết áp lúc nghỉ (mm Hg)

  - chol - 'cholesterol': cholestoral trong huyết thanh tính bằng (mg / dl)

  - fbs - 'fasting_blood_sugar': đường huyết lúc đói

  - restecg - 'rest_electrocardiographic': kết quả điện tâm đồ lúc nghỉ ngơi

  - thalach: nhịp tim tối đa đạt được (BPM: nhịp trên phút)

- exang - 'exercise_included_angina': đau thắt ngực do vận động (chỉ xảy ra do vận động)

- oldpeak: độ chênh lệch của đoạn ST gây ra do tập thể dục liên quan đến nghỉ ngơi (mm)

- slope - 'ST-slope': độ dốc của đoạn ST tập thể dục đỉnh cao

- ca: số lượng mạch chính quan sát được khi quang nội soi bằng huỳnh quang

- tha - 'thalassemia': bệnh tan máu bẩm sinh

- target: chẩn đoán bệnh tim

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

4. Trực quan hóa dữ liệu

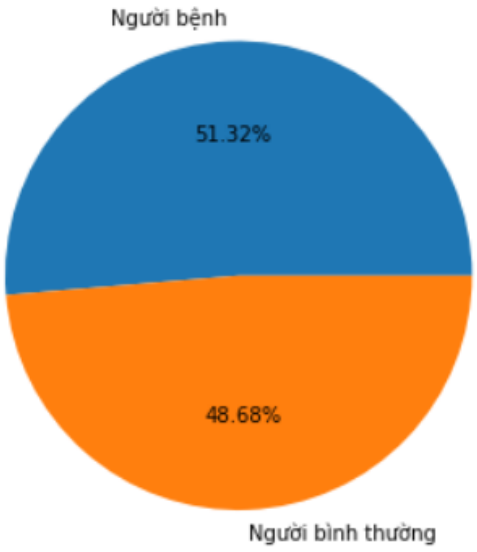
- Thông kê tóm tắt:

	mean	median	std	min	max
age	54.4341	56	9.0722	29	77
Chol	246	240	51.5925	126	564
Trestbps	131.611	130	17.5167	94	200
thalach	149.1141	152	23.0057	71	202

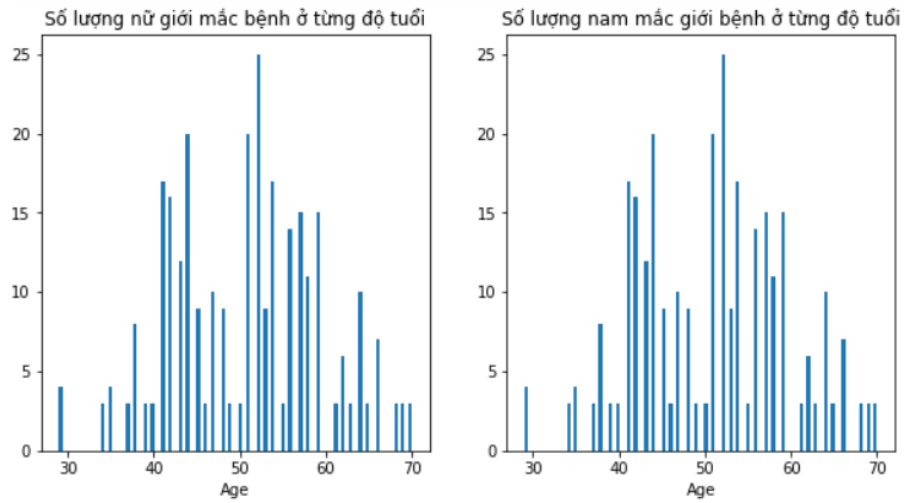
- Độ tuổi được khảo sát có trung bình là 54.4, trung vị là 56 và độ lệch chuẩn là 9.07 cho thấy độ tuổi có độ phân tán thấp và tập trung ở khoảng 45-65 tuổi, đồng thời cũng có một số giá trị ngoại vi có chênh lệch khá lớn như giá trị thấy nhất của độ tuổi là 29 và cao nhất 77.

- Cholesterol của tập dữ liệu có trung bình là 246mg/dl và lớn nhất là 564, trong đó, cholesterol của người khỏe mạnh là <200mg/dl và lớn hơn mức đó thường có nguy cơ mắc các bệnh về tim mạch.
- Huyết áp có trung bình là 131.6, trung vị là 130, giá trị thấp nhất là 94 và lớn nhất là 200.
- Nhịp tim tối đa có trung bình là 149.11, trung vị là 152, giá trị thấp nhất là 71 và lớn nhất là 202.

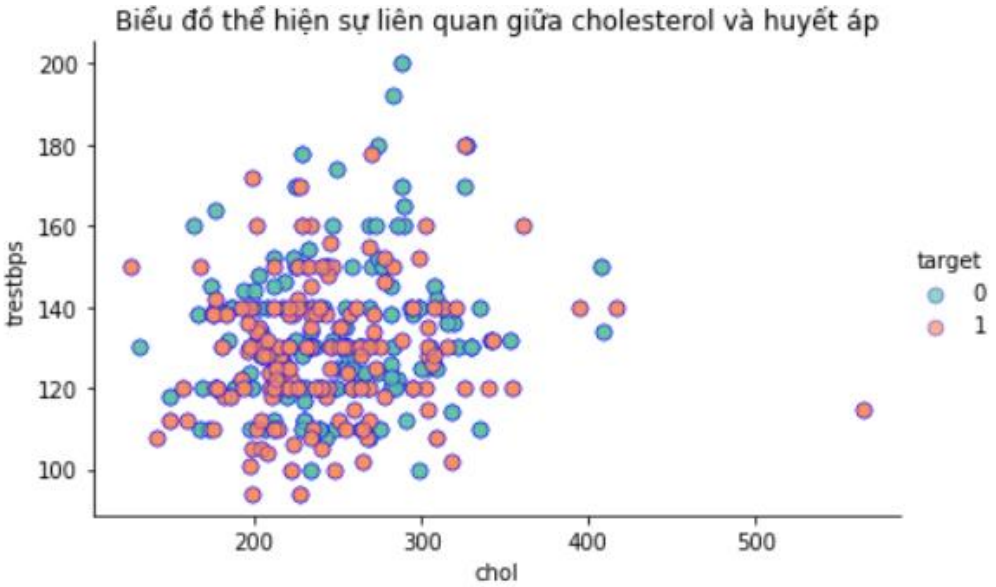
    - Biểu đồ cho thấy sự chênh lệch giữa người bệnh và người bình thường không cao từ đó việc phân tích và đánh giá sẽ đạt được độ tin cậy và chính xác cao hơn.



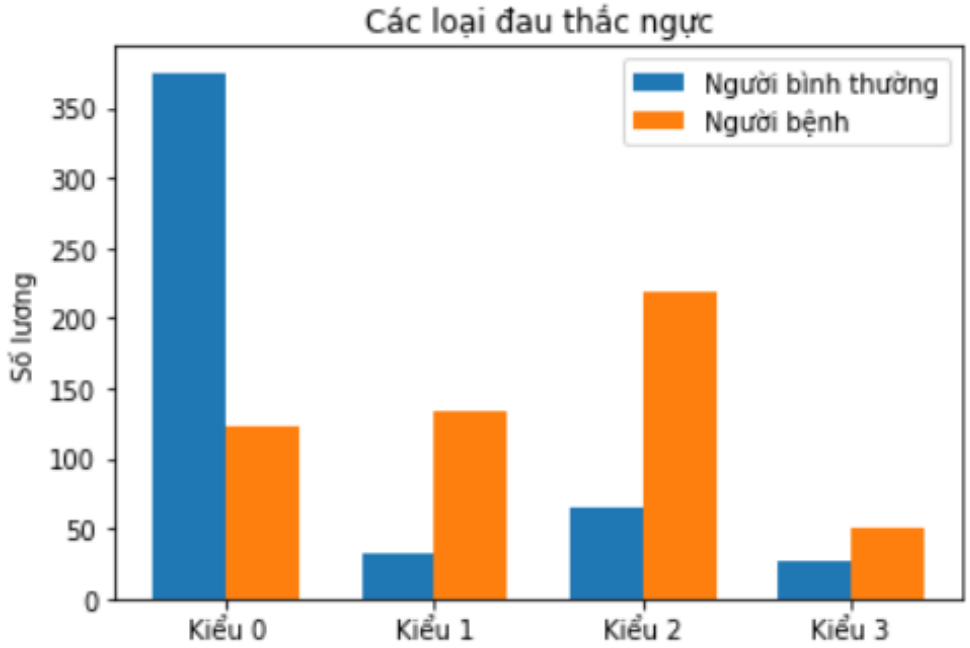
- Không có sự khác nhau về độ tuổi mắc bệnh giữa nam và nữ. Độ tuổi trung bình có khả năng mắc bệnh cao của cả nam và nữa là từ 41- 59 tuổi.



- Theo như khoa học thì cholesterol tỉ lệ thuận với huyết áp, khi cholesterol tăng sẽ ngăn cản sự lưu thông của máu làm cho tim phải tích cực đẩy máu từ đó làm tăng huyết áp và tăng nguy cơ mắc bệnh tim theo biểu đồ thì mức cholesterol ở bệnh nhân thường tập trung từ 200 - 300 mg/dl, từ đó huyết áp cũng ở mức cao từ 120 - 150 mmHg. Nên ta có thể thấy được mức cholesterol ảnh hưởng trực tiếp đến tim mạch và cũng là dấu hiệu rõ rệt cho người mắc bệnh.



- Những người đau ngực ở kiểu 0(đau điển hình) thường không mắc bệnh, so với số người không mắc bệnh ở các kiểu đau khác thì người không bệnh ở kiểu đau 0 cao vượt trội hơn rất nhiều cà khi so với người mắc bệnh đau kiểu 0 thì cũng có sự chênh lệch lớn. Tỉ lệ người đau kiểu 1(đau không điển hình) và kiểu 2(đau ngực không đặt hiệu) có khả năng mắc bệnh cao hơn rất nhiều với số người mắc bệnh, chiếm 80% ở kiểu 1 và 77,7% ở kiểu 2 trên tổng số của hai kiểu đau



5. Mô hình hóa dữ liệu

	coef	std err	z	P>|z|
const	3.6902	1.401	2.633	0.008
age	-0.0082	0.013	-0.650	0.516
sex	-1.8465	0.257	-7.197	0.000
cp	0.8546	0.100	8.516	0.000
trestbps	-0.0182	0.006	-3.245	0.001
chol	-0.0057	0.002	-2.757	0.006
fbs	-0.1012	0.285	-0.355	0.723
restecg	0.4132	0.189	2.187	0.029
thalach	0.0236	0.006	4.158	0.000
exang	-0.9908	0.224	-4.418	0.000
oldpeak	-0.5707	0.116	-4.920	0.000
slope	0.5341	0.189	2.831	0.005
ca	-0.7545	0.103	-7.321	0.000
thal	-0.8861	0.156	-5.693	0.000

Để xử lý và xây dựng mô hình dự đoán dữ liệu thì nhóm đã sử dụng thuật toán: hồi quy Logistic. Hồi quy Logistic là một kỹ thuật thống kê để xem xét mối liên hệ giữa các biến độc lập với biến phụ thuộc là biến nhị phân, ở tập dữ liệu này thì biến phụ thuộc là target với giá tri '1' tượng cho người mắc bệnh tim, và giá trị '0' tượng trung cho người không mắc bệnh tim.

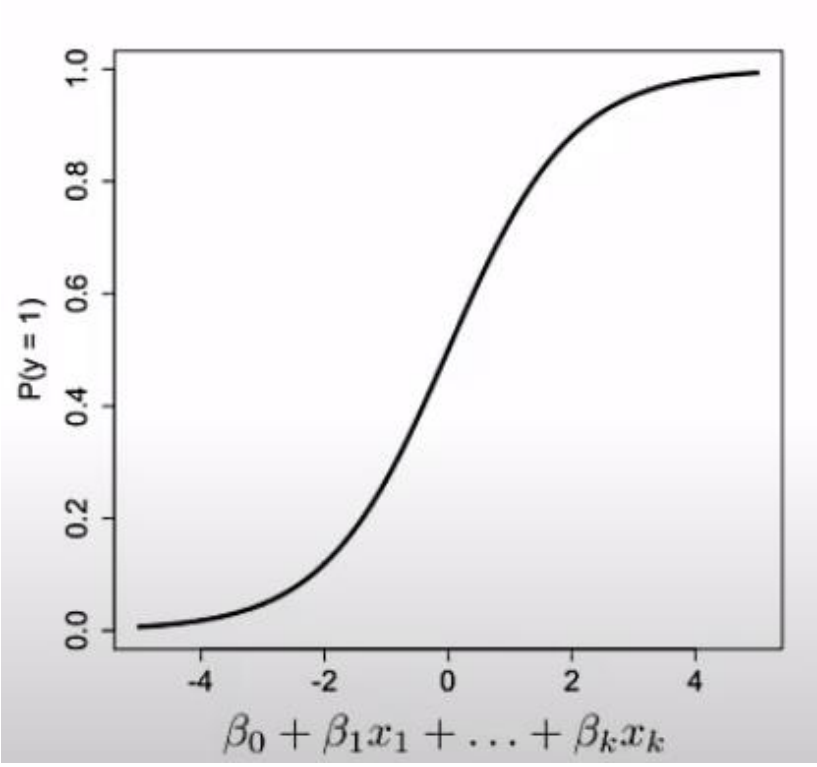
Phương trình hồi quy Logistic:

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}$$

Với β là các giá trị coef được tính ở trên của các thuộc tính trong dữ liệu và x là các giá trị thực tế của các thuộc tính.

Dựa vào thuật toán hồi quy logistic có phương trình như trên để suy ra P(y = 1) là khả năng mắc bệnh của một người và P(y = 0) = 1 – P(y = 1). Thông qua các biến độc lập là các thuộc tính đã cho trong tập dữ liệu ngoài trừ biến phụ thuộc target.

Trong đó, P(y = 1) tỉ lệ thuận với $\eta = \beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k$, η có giá trị càng lớn thì P(y = 1) càng tiến về 1 cho thấy khả năng mắc bệnh càng cao.



6. Thực nghiệm, kết quả và thảo luận

- Những yếu tố nào cho thấy rõ khả năng mắc bệnh tim?
- Độ tuổi có ảnh hưởng đến nguy cơ mắc bệnh tim hay không?
- Trạng thái điện tâm đồ của người mắc bệnh tim có gì bất thường?
- Có sự khác biệt về nguy cơ mắc bệnh tim giữa nam và nữ không?

- Câu hỏi 1 và 2 là kiểm tra mức độ ảnh hưởng của các biến độc lập tác động lên biến phụ thuộc là target như thế nào. Để làm được điều đó ta thực hiện kiểm định cho từng biến độc lập.

H_0: $\beta_i = 0$ H_A: $\beta_i \neq 0$

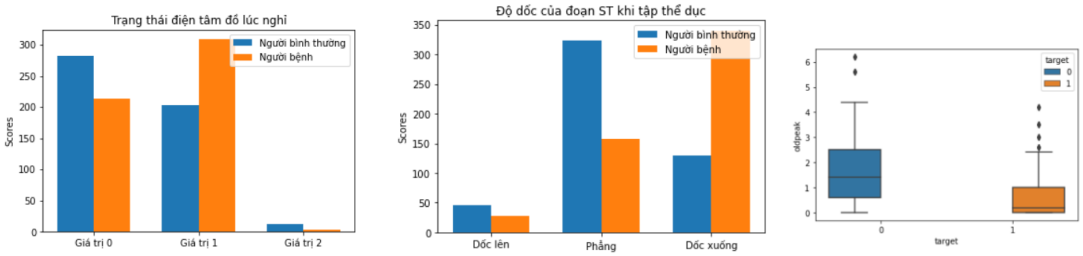
$$Z = \frac{\widehat{\beta_i} - \beta_i}{SE_i} = \text{m}$$

p-value = P(|Z| > m) = P(Z > m) + P(Z < -m)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
P(>|z|)	0.516	0	0	0.001	0.006	0.723	0.029	0	0	0	0.005	0	0

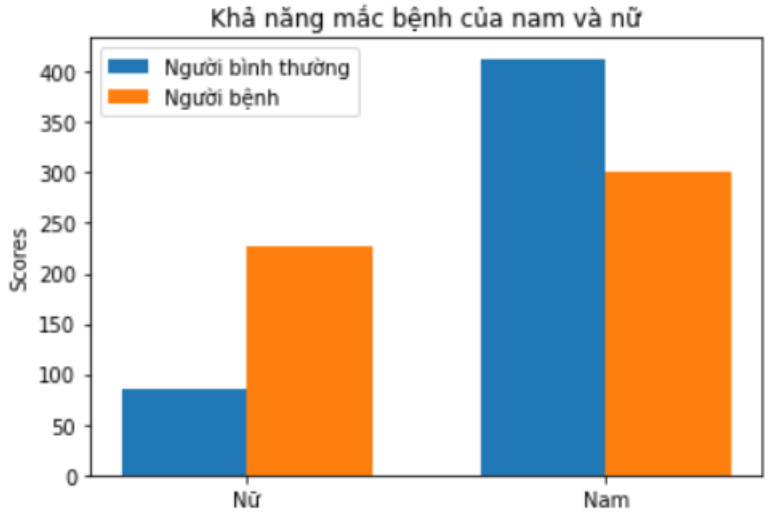
Với mức ý nghĩa 5% cùng với số liệu kiểm định đã tính được thì các biến độc lập > 0.05 không tác động đến biến phụ thuộc. Đồng nghĩa với độ tuổi, fbs (đường huyết lúc đói) không là yếu tố cho thấy một người có mắc bệnh tim hay không, ngược lại với các biến độc lập có kiểm định bằng 0 sẽ là các yếu tố có cơ sở để dự đoán một người có khả năng mắc bệnh hay không.

- Điện tâm đồ là một sơ đồ ghi lại hoạt động của tim nó có liên hệ mật thiết đến quá trình khám và phát hiện các vấn đề về tim. Để trả lời được câu hỏi điện tâm đồ của người bị bệnh tim có gì bất thường so với người bình thường thì ta sẽ dựa vào đồ thị trực quan hóa từ tập dữ liệu để phân tích.



Trạng thái điện tâm đồ lúc nghỉ của người bệnh thường tập trung ở giá trị 1 là có bất thường sóng ST ST-T (phắc phẫu sóng T và / hoặc độ cao hoặc chênh lệch của ST> 0,05 MV) cùng với đó độ dốc của đoạn ST khi tập thể dục của người bệnh thường dốc xuống và độ chênh lệch của độ dốc ST của người bệnh tập trung từ 0,1 đến 1.1.

- Theo như kiểm định đã tính ở trên thì với mức ý nghĩa 5%, giới tính cũng có ảnh hưởng đến khả năng mắc bệnh tim. Trong đó qua biểu đồ ta có thể thấy rõ khả năng mắc bệnh tim của nữ cao hơn nam với tỉ lệ mắc bệnh ở nữ là 72% và ở nam là 42%



- Khi xây dựng mô hình Logistic ta sẽ dùng đại lượng ma trận nhầm lẫn (Confusion matrix) để xác định thang dự báo của mô hình. Để biết có mắc bệnh hay không, ta sẽ so sánh các giá trị thực tế với giá trị của mô hình để xem khả dự báo của mô hình có tốt hay không.

Ở dữ liệu của ta, sau khi được phân tách và train thì mô hình dự đoán có kết quả như sau:

	Thực tế dương tính	Thực tế âm tính
Dự đoán dương tính	TP(82)	FP(23)
Dự đoán âm tính	FN(7)	TN(93)

- **TP**: Mô hình dự báo mắc bệnh và thực tế mắc bệnh
- **TN**: Mô hình dự báo không mắc bệnh và thực tế không mắc bệnh.
- **FP**: Mô hình dự báo mắc bệnh và thực tế không mắc bệnh
- **FN:** Mô hình dự báo không mắc bệnh và thực tế mắc bệnh

- Độ chính xác toàn thể của mô hình (Accuracy):

  (82 + 93) / (82 + 23 + 7 + 93) = 0,8534 -> 85,34 %

- Độ nhạy (Sensity): Trong số những người thực tế dự đoán mắc bệnh, bao nhiêu % người được dự đoán mắc bệnh?

  Độ nhạy = TP / (TP + FN) = 82 / (82 + 7) = 0,9213 -> 92,13%

- Độ đặc hiệu (Sencifycity): Trong số những người thực tế dự đoán không mắc bệnh, bao nhiêu % được dự đoán không mắc bệnh?

  Độ đặc hiệu = TN / (TN + FP) = 93 / (93 + 23) = 0.8017 -> 80,17%

Ở tập dữ liệu của ta cho thấy mô hình tốt vì có độ nhạy và độ đặc hiệu lớn, do đó có thể dùng mô hình để dự đoán.

7. Kết luận

Sau quá trình phân tích dữ liệu, nhóm đã tìm hiểu và biết cách dùng thuật toán hồi quy logictics để phân tích dữ liệu. Qua đó, biết được các thuộc tính ảnh hưởng đến khả năng mắc bệnh tim, đồng thời cũng tìm hiểu và biết được cách tính độ chính xác toàn thể, đồ nhạy, đồ đặc hiệu trong ma trận nhằm lẫn của một tập dữ liệu cho trước. Ngoài ra, nhóm còn học được cách nhận xét tập dữ liệu qua các biểu đồ trực quan của tập dữ liệu. Trong tương lai, nhóm sẽ tìm hiểu thêm về thuộc các thuộc tính ẩn có thể ảnh hưởng đến khả năng mắc bệnh tim, từ đó, giúp quá trình phân tích dữ liệu đạt kết quả chính xác hơn.

8. Đóng góp

Thứ tự	Họ tên	Phân chia việc	Tỉ lệ đóng góp
1	Trương Hoàng Anh Khôi	Tìm hiểu xây dựng mô hình hóa	100%
2	Nguyễn Văn Trường Tốt	Tìm hiểu xây dựng mô hình hóa	100%
3	Phạm Minh Long	Trực quan hóa dữ liệu, cung cấp thông tin	100%
4	Nguyễn Quốc Thắng	Trực quan hóa dữ liệu, cung cấp thông tin	100%

9. Tham khảo

Bài báo, sách:

- Mirko Stojiljkovic, Logistic Regression in Python, https://realpython.com/logistic-regression-python/?fbclid=IwAR0MrvJfdyk1kTlxIYvw6yO4AI8_Wyw9jZwlK_Iz27L0QYLFtasNp84cM6Q
- Avinash NavlanI, Understanding Logistic Regression in Python Tutorial, 16/12/2019, https://www.datacamp.com/tutorial/understanding-logistic-regression-python?fbclid=IwAR3lm6luGCR34VhBV6a2X1pVdpcUyh_s7ZJ66tMQ6UVMi6fZatg-06yXeA0
- Jake VanderPlas, Python Data Science Handbook, powered by Jupyter
- Hồi quy logistic là gì?, 23/09/2021, https://vietvuevent.vn/hoi-quy-logistic-la-gi/?fbclid=IwAR3optqKsJQ9_75hd5xJ66Ec95Fq0KiSfBigALd_b0NUnQ2UKv_RJeC17o8
- Code và thư viện
- David Lapp, Heart Disease Dataset, https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download&fbclid=IwAR1VUCSixdb0yvw2bGg3Nf_fiecNILEM6kYXLIjjv04o0rPDVuyYobxarno
- Sử dụng thư viện:pandas, numpy, matplotlib.pyplot, seaborn, statsmodels.api, sklearn.metrics, sklearn.linear_model