Đánh giá hiệu suất của ứng dụng định lí Bayes trong vấn đề phân loại bệnh nhân suy tim

Hồ Nguyễn Thiên Vũ – 22521689
KHDL2022
University of Information Technology - VNUHCM
22521689@gm.uit.edu.vn

Châu Nguyễn Tri Vũ – 22521687
KHDL2022
University of Information Technology - VNUHCM
22521687@gm.uit.edu.vn

TÓM TẮT

Suy tim là tình trạng tim không có khả năng bơm máu hiệu quả gây ra tình trạng không đáp ứng đủ nhu cầu Oxy và máu của cơ thể. Số ca tử vong hàng năm do bệnh lí này lên tới hàng trăm nghìn người trên toàn cầu, ảnh hưởng đến hàng triệu người. Bài báo này của chúng tôi trình bày việc sử dụng mô hình Machine Learning là thuật toán Naïve Bayes để phân loại tập dữ liệu gồm 299 dữ liệu của bệnh nhân suy tim thu được từ kho lưu trữ UCI vào năm 2015 dựa trên kết quả sống sót của họ trong quá trình theo dõi. Phương pháp của chúng tôi là sử dụng thuật toán Naïve Bayes tính toán xác suất cho từng danh mục dựa trên thuộc tính của bệnh nhân, chỉ định các xác suất như 0.68 cho những bệnh nhân có khả năng qua đời và 0.32 cho những bệnh nhân sống sót. Các tỷ lệ đào tạo – kiểm tra khác nhau bao gồm (60-40%), (70-30%) và (80-20%) được sử dụng để đánh giá trực quan về Naïve Bayes đối với sự phân loại này. Đáng chú ý, tỷ lệ (80-20%) luôn mang lại độ chính xác cao nhất, khẳng định tầm quan trọng của việc phân vùng dữ liệu tối ưu để phân loại bệnh nhân chính xác. Nghiên cứu này nêu bật sự thành công ứng dụng mô hình Naïve Bayes để phân loại kết quả sống sót của bệnh nhân suy tim. Nghiên cứu nhấn mạnh tầm quan trọng của việc phân vùng dữ liệu thích hợp và chứng minh tiềm năng của kỹ thuật Machine Learning trong nghiên cứu y học.

Từ khóa: Naïve Bayes, Machine Learning

1. GIỚI THIỆU

Suy tim, một tình trạng tim mạch phức tạp và phổ biến, ảnh hưởng đáng kể đến toàn cầu sức khỏe bằng cách làm suy giảm khả năng bơm máu đầy đủ của tim, dẫn đến thiếu oxy và cung cấp chất dinh dưỡng cho các cơ quan và mô. Việc phân loại và đưa ra các phỏng đoán chính xác về bệnh nhân suy tim là bắt buộc để điều chỉnh các phương pháp điều trị hiệu quả và nâng cao sức khỏe của bệnh nhân. Trong những năm gần đây, sự hội tụ của trang thiết bị y tế và Machine Learning đã mở đường cho những tiến bộ mang tính biến đổi trong chăm sóc sức khỏe. Các thuật toán Machine Learning mang lại nhiều hứa hẹn trong việc phân tích các bộ dữ liệu phức tạp và rộng lớn. Trong đó không thể không kể đến thuật toán Naïve Bayes nổi bật về hiệu quả trong các ứng dụng y tế khác nhau, bao gồm cả việc phân loại bệnh. Naïve Bayes – một thuật toán xác suất được thành lập dược trên định lí Bayes, đơn giản hóa quá trình lập mô hình bằng cách giả định sự độc lập có điều kiện giữa các tính năng, đạt được kết quả khả quan trong các tình huống thực tế đa dạng. Do tính chất phức tạp của bệnh suy tim và những lợi ích tiềm tàng của việc Machine Learning, việc áp dụng thuật toán Naïve Bayes để phân loại bệnh nhân suy tim có nhiều hứa hẹn. Việc phân loại chính xác bệnh nhân suy tim thành các phân nhóm riêng biệt dựa trên các đặc điểm lâm sàng, thông số chẩn đoán và các đặc điểm thích hợp khác là mấu chốt để điều chỉnh chiến lược điều trị, dự đoán kết quả của bệnh nhân và tối ưu hóa phân bổ nguồn lực chăm sóc sức khỏe. Nghiên cứu này tìm cách khám phá khả năng của thuật toán để phân biệt giữa các giai đoạn khác nhau. Những phát hiện của cuộc điều tra này có thể cung cấp cho các bác sĩ lâm sàng và người hành nghề chăm sóc sức khỏe những hiểu biết sâu sắc có giá trị về phân tầng bệnh nhân, cho phép các can thiệp y tế được cá nhân hóa và cứu chữa chính xác. Về bản chất, việc tích hợp các thuật toán Machine Learning, đặc biệt là Naïve Bayes có tiềm năng cách mạng hóa việc phân loại bệnh nhân suy tim. Khi tỷ lệ suy tim ngày càng gia tăng, các phương pháp tiếp cận sáng tạo để phân loại bệnh nhân có tầm quan trọng tối cao, giữ vai trò then chốt trong việc nâng cao khả năng ra quyết định lâm sàng và cuối cùng là nâng cao các tiêu chuẩn và kết quả chăm sóc bệnh nhân.

1.1. Lý thuyết về định lí Bayes

Định lí Bayes là phương pháp học tập thực tế nhất cho hầu hết các vấn đề học tập dựa trên việc đánh giá xác suất rõ ràng của các giả thuyết. Về việc phân loại

học tập, định lí Bayes cực kì cạnh tranh với các thuật toán học tập khác và trong nhiều trường hợp còn vượt trội hơn. Định lí Bayes cực kì quan trọng trong Machine Learning vì nó cung cấp góc nhìn độc đáo để hiểu nhiều thuật toán học tập mà không thao túng xác suất một cách rõ ràng.

Định lí Bayes phát biểu rằng:

Giả sử rằng $|A| \neq 0$ và $|B| \neq 0$, ta có thể phát biểu như sau:

$$P(\mathrm{A}|\mathrm{B}) = \frac{|A \cap B|}{|B|} = \frac{\frac{|AB|}{|\Omega|}}{|\frac{B}{\Omega}|} = \frac{P(A \cap B)}{P(B)} \quad (1)$$

Ta có điều tương tự ngược lại đối với P(B\A) (2)

Từ phương trình (1) và (2), ta có:

$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ (3)

Và do đó:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(B|A)P(A)}{P(B)} \qquad (4)$$

Đây là công thức đơn giản nhất của định lí Bayes. Nếu không gian mẫu Ω có thể được chia thành hữu hạn nhiều sự kiện loại trừ lẫn nhau $A_1, A_2, \dots, A_n$ và B là một biến cố với P(B) > 0, là tập con của hợp của tất cả A_i, thì với mỗi A_i, công thức tổng quát của Bayes là:

$$P(A_i|\mathrm{B}) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)} \qquad (5)$$

Có thể viết lại thành:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} \qquad (6)$$

Cả hai phương trình (5) và (6) đều theo sau phương trình (4) vì định lí tổng xác suất. Định lí Bayes có thể được sử dụng để suy ra các xác suất hậu nghiệm của một giả thuyết dựa trên dữ liệu quan sát được:

$$P(\text{Giả thuyết}|\text{Dữ liệu}) = \frac{P(Dữ\ liệu|Giả\ thuyết)P(Giả\ thuyết)}{P(Dữ\ liệu)} \qquad (7)$$

Hay $P(\mathrm{H}|\mathrm{D}) = \frac{P(D|H)P(H)}{P(D)}$

Trong đó:

- P(H): Xác suất của giả thuyết (H – Hypothesis)
- P(D): Xác suất của dữ liệu (D – Data)
- P(D|H): Xác suất của D khi cho trước H
- P(H|D): Xác suất của H khi cho trước D (Xác suất hậu nghiệm)

Trong trường hợp tổng quát, ta có k lớp loại trừ lẫn nhau và đầy đủ H_i với i = 1, 2, .., n. $P(D|H_i)$ là các xác suất nhìn thấy D là đầu vào khi viết nó thuộc lớp H_i. Các xác suất sau của lớp H_i có thể được tính như sau:

$$(H_i|\mathrm{D}) = \frac{P(D|H_i)P(H_i)}{\sum_{i=1}^{n} P(D|A_i)P(A_i)} \qquad (8)$$

Để lựa chọn giả thuyết tốt nhất trong số các giả thuyết được tạo ra, giả thuyết có xác suất tối đa được chọn là được gọi là giả thuyết hậu nghiệm tối đa (MAP) và nếu chúng ta giả sử rằng P(H) giống nhau cho tất cả các giả thuyết thì giả thuyết có thể xảy ra tối đa sẽ giảm xuống, giả thuyết khả năng sẽ tăng tối đa.

1.2. Trình phân loại Naïve Bayes

Naïve Bayes là một thuật toán xác suất được sử dụng để giải quyết các vấn đề phân loại. Nó hoạt động bằng cách tính xác suất của một điểm dữ liệu thuộc một lớp cụ thể dựa trên xác suất của các đặc điểm của điểm dữ liệu. Chúng tôi giả sử rằng một tập dữ liệu chứa n trường hợp x_i bao gồm p thuộc tính. Mỗi phiên bản được giả sử thuộc về một và chỉ một lớp y_i. Hầu hết các mô hình dự đoán trong Machine Learning đều tạo ra điểm số s cho mỗi trường hợp x_i. Điểm này định lượng mức độ thành viên lớp của một trường hợp nào đó trong lớp y_i. Nếu tập dữ liệu chỉ chứa các trường hợp âm và dương, y $\in\{0,1\}$ thì mô hình dự đoán có thể được sử dụng làm công cụ xếp hạng hoặc làm công cụ phân loại. Trình phân loại Naïve Bayes đơn giản sử dụng các xác suất này để gán một thể hiện cho một lớp. Áp dụng định lí Bayes (4) và đơn giản hóa ký hiệu một chút, chúng ta thu được:

$$P(y_j|x_i) = \frac{P(x_i \cap y_j)}{\mathrm{P}(x_i)} \qquad (9)$$

Lưu ý rằng tử số trong (9) kà xác suất chung của x_i và y_j. Do đó, tử số có thể được viết lại như sau (chúng ta sẽ chỉ sử dụng x, bỏ chỉ mục i để đơn giản:

$$\mathrm{P}(x|y_j)\mathrm{P}(y_j) = \mathrm{P}(x, y_j) = \mathrm{P}(x_1, x_2, \dots, x_p, y_j) \qquad (10)$$

$$= \mathrm{P}(x_1, |x_2, \dots, x_p, y_j)(x_2, x_3, \dots, x_p, y_j)$$

$$= \mathrm{P}(x_1, |x_2, \dots, x_p, y_j)\mathrm{P}(x_2, |x_3, \dots, x_p, y_j)..\mathrm{P}(x_p|y_j)P(y_j)$$

Giả sử rằng x_i có thể độc lập với nhau. Đây là một giả định mạnh mẽ, rõ ràng đã bị vi phạm trong hầu hết các ứng dụng thực tế và do đó nó được gọi là Naïve.

Giả định này suy ra rằng $\mathrm{P} = (x_2, |x_3,\ x_4 \dots, x_p, y_j) = \mathrm{P}(x_1, y_j)$ chẳng hạn. Như vậy xác suất cuối cùng của x và y_j là $\mathrm{P}=\mathrm{P}(x|y_j)\mathrm{P}(y_j) = \mathrm{P}((x_1|y_j)(x_2|y_j) \dots \mathrm{P}(x_p|y_j)P(y_j)$

$$\prod_{k=1}^{p} P(x_k|y_j)P(y_j) \qquad (11)$$

Cộng thêm vào (9), ta được:

$$\mathrm{P}(y_j|x) = \frac{\prod_{k=1}^{p} P=(x_k|y_j)P(y_j)}{P(x)} \qquad (12)$$

Lưu ý rằng P(x) không phụ thuộc vào lớp. P(x) hoạt động như một hệ số tỷ lệ và đảm bảo xác suất hậu nghiệm $\mathrm{P}(x|y_j)$ có tỷ lệ chính xác. Khi chúng ta quan tâm đến một quy tắc phân loại rõ ràng, tức là quy tắc gán mỗi thể hiện cho chính xác một lớp, thì chúng ta có thể chỉ cần tính giá trị của tử số cho mỗi lớp và chọn lớp đó mà giá trị này là lớn nhất.

Quy tắc này được gọi là quy tắc hậu nghiệm tối đa (13). Kết quả là lớp "chiến thắng' cũng là lớp MAP và được tính bằng $\hat{Y}$ cho trường hợp x như sau:

$$\hat{Y} = argmax \prod \prod_{k=1}^{p} P = (x_k|y_j)P(y_j) \qquad (13)$$

Một mô hình triển khai (11) gọi là bộ phân loại Naïve Bayes (đơn giản). Tuy nhiên việc phân loại rõ ràng thường không được như mong muốn.

Xác suất xếp sau của lớp ước tính là điểm xếp hạng tự nhiên. Áp dụng lại định lí tổng xác suất (13), chúng ta viết lại phương trình (12) như sau:

$$\mathrm{P}(y_j|x) = \frac{\prod_{k=1}^{p} P=(x_k|y_j)P(y_j)}{\prod_{k=1}^{p} P=(x_k|y_j)P(y_j) + \prod_{k=1}^{p} P=(x_k|y_j^c)P(y_j^c)} \qquad (14)$$

2. Tài liệu và mô tả của tập dữ liệu

Cơ sở dữ liệu Cleveland - một nguồn UCI đã cup cấp tập dữ liệu bao gồm 299 bệnh nhân suy tim được theo dõi, mỗi bệnh nhân có 13 đặc điểm lâm sàng. Mỗi hàng có một

hồ sơ bệnh nhân. Một trong 13 đặc điểm của hồ sơ là đặc điểm dự đoán được gọi là Y, giá trị của nó cho thấy loại suy tim (bệnh nhân tử vong trong thời gian theo dõi hay bệnh nhân không tử vong trong thời gian theo dõi). 12 thuộc tính cuối cùng được sử dụng trong giai đoạn dự đoán. Bảng bên dưới hiển thị tập dữ liệu được sử dụng trong bài báo cáo này.

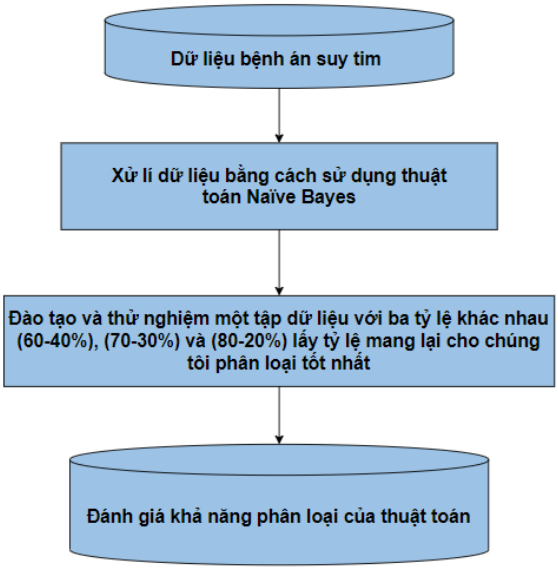
Bảng 1: Mô tả tập dữ liệu

STT	Tên thuộc tính	Mô tả thuộc tính	Giá trị
1	Tuổi	Tuổi của bệnh nhân (năm)	Không có phạm vi cụ thể
2	Giới tính	Nam hay nữ (binary)	Nữ = 0 Nam = 1
3	Thiếu máu	Giảm hồng cầu hoặc huyết sắc tố (boolean)	
4	Cao huyết áp	Nếu bệnh nhân bị tăng huyết áp (boolean)	Không có phạm vi cụ thể
5	CPK	Mức độ enzyme CPK trong máu (mcg/l)	Đường huyết lúc đói > 120 mg/dl Đúng =1 Sai = 0
6	Tiểu đường	Nếu bệnh nhân bị tiểu đường (boolean)	Bình thường = 0 Bất thường = 1
7	Phân suất tống máu	Lượng máu ra khỏi tim mỗi khi co thắt (%)	Không = 0 Có = 1
8	Tiểu cầu	Tiểu cầu trong máu (số lượng/mL)	Không = 0 Có = 1
9	Huyết thanh Creatinine	Lượng huyết thanh Creatinine trong máu (mg/dL)	Không = 0 Có = 1
10	Huyết thanh Sodium	Lượng huyết thanh sodium trong máu (mEq/L)	Không = 0 Có = 1
11	Hút thuốc	Bệnh nhân có hút thuốc hay không	Không = 0 Có = 1
12	Thời gian	Thời gian theo dõi (ngày)	Giá trị định nghĩa
13	[Mục tiêu]	Nếu bệnh nhân qua đời trong giai đoạn theo dõi (boolean) = A Nếu bệnh nhân không tử vong trong giai đoạn theo dõi (boolean) = B	A, $Y_i = +1$ B, $Y_i = -1$
	Sự kiện tử vong Y = Loại		

3. Thảo luận và kết quả

3.1. Các biện pháp đánh giá hiệu suất

Trong phần này, chúng ta sẽ đề cập đến chẩn đoán dự đoán suy tim dựa trên thuật toán Machine Learning là Naïve Bayes. Quá tình này được chia thành nhiều bước và được mô tả như sau:



3.2. Ma trận hỗn loạn (Confusion Matrix)

Ma trận hỗn loạn (Confusion Matrix) được sử dụng để đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Nó khớp với các giá trị thực tế và được dự đoán với bốn yếu tố:

- True Positive (TP): Dự đoán chính xác
- True Negative (TN): Dự đoán chính xác một cách gián tiếp
- False Positive (FP – Type 1 Error): Dự đoán sai lệch
- Falsee Negative (FN – Type 2 Error): Dự đoán sai lệch một cách gián tiếp

Ma trận hỗn loạn rất phù hợp để tính toán độ chính xác, thu hồi, điểm F1 và độ chính xác.

Bảng 2: Confusion matrix

Lớp thực tế	Lớp dự đoán	
	Bệnh nhân chưa tử vong	Bệnh nhân đã tử vong

Bệnh nhân chưa tử vong	TP	FN
Bệnh nhân đã tử vong	FP	TN

Ở phần này chúng ta sẽ kiểm ta tính hiệu quả của kỹ thuật được đề xuất bằng cách sử dụng độ chính xác, độ đặc hiệu, độ nhạy và giá trị trung bình. Dự đoán đúng tỷ lệ với tổng số dự đoán được đưa ra bởi bộ phân loại sẽ quyết định độ chính xác của nó được tính toán như sau:

$$\text{Độ chính xác} = \frac{(TP+TN)}{(TP+FP+FN+TN)} * 100\% \quad (15)$$

Bên cạnh đó, tỷ lệ lỗi được tính như sau:

$$\text{Tỷ lệ lỗi} = \frac{(FP+FN)}{(TP+FP+FN+TN)} * 100\% \quad (16)$$

Và cách tính tính đặc hiệu là:

$$\text{Tính đặc hiệu} = \frac{TN}{(FP+TN)} * 100\% \quad (17)$$

3.3. Phân tích dữ liệu bằng cách dùng Naïve Bayes

Trong phần này, chúng tôi đã mượn kết quả của một thực nghiệm thông qua thực hiện một số bước để phân loại dữ liệu tốt nhất bằng chương trình Weka bằng cách sử dụng mô hình Machine Learning là Naïve Bayes. Đĩnh nghĩa về các đặc điểm dự đoán cốt lõi được sử dụng để phân loại các triệu chứng suy tim (bệnh nhân tử vọng khi được theo dõi hoặc bệnh nhân sống sót sau thời gian theo dõi mà không tử vong).

Bảng 3: Độ chính xác của Naïve Bayes để đào tạo và kiểm tra tỷ lệ khác nhau cho tập dữ liệu

Độ chính xác của Naïve Bayes			
Mô hình	Đào tạo 60% - Thử 40 %	Đào tạo 70% - Thử 30%	Đào tạo 80% - Thử 20%
Hiệu suất	74.2%	74.4%	80%
Confushion Matrix	CM = $\begin{bmatrix} 20 & 26 \\ 5 & 69 \end{bmatrix}$	CM = $\begin{bmatrix} 15 & 19 \\ 4 & 52 \end{bmatrix}$	CM = $\begin{bmatrix} 12 & 10 \\ 2 & 36 \end{bmatrix}$

Bảng trên minh họa tỷ lệ chính xác của mô hình Naïve Bayes theo các tỷ lệ thử nghiệm khác nhau, nêu bật độ chính xác của mô hình này. Các ma trận hỗn loạn đi kèm cung cấp thông tin chi tiết về các dự đoán đúng, dự đoán đúng một cách gián tiếp, dự đoán sai lệch và dự đoán sai lệch một cách gián tiếp của mô hình trong từng điều kiện thử nghiệm.

Bảng 4: Ma trận hỗn loạn của Naïve Bayes đối với thử nghiệm đào tạo 80% - kiểm tra 20% của tập dữ liệu

	Lớp dự đoán		
Lớp thực tế	Bệnh nhân không tử vong	Bệnh nhân tử vong	Tổng
Bệnh nhân không tử vong	12 (TP)	2 (FN)	14
Bệnh nhân tử vong	10 (FP)	36 (TN)	46
Tổng	22	38	60

Ma trận hỗn loạn trên cung cấp nhiều chi tiết khác nhau về kết quả thu được từ phương pháp được đánh giá. Như chúng tôi đã đề cập trước đó, 80% tập dữ liệu được xử lý trước đã được sử dụng làm tập huấn luyện và 20% còn lại làm thử nghiệm. Có 60 trường hợp xét nghiệm hoặc thông tin từ 60 bệnh nhân đã dược sử dụng để kiểm nghiệm phương pháp đánh giá. 22 bệnh nhân được chẩn đoán thực sự theo đó bệnh nhân không tử vong trong thời gian theo dõi và 38 bệnh nhân còn lại được chẩn đoán là bệnh nhân tử vong trong thời gian theo dõi.

Bảng 5: Độ chính xác chi tiết theo lớp của Naïve Bayes đối với đào tạo 80% và kiểm tra 20%

Lớp	Bệnh nhân không tử vong	Bệnh nhân tử vong	Trọng số trung bình
Tỷ lệ TP	0.545	0.947	0.800
Tỷ lệ FP	0.053	0.455	0.307
Độ chính xác	0.857	0.783	0.810
Thu hồi (Độ nhạy)	0.545	0.947	0.800
Đo F	0.667	0.857	0.787
Khu vực ROC	0.829	0.829	0.829
Khu vực PRC	0.820	0.823	0.822

Trọng số trung bình của các tiêu chí của hai nhóm, có tính đến sự bổ sung của chúng, được trình bày ở bảng trên. Lấy một minh họa, tỷ lệ TP trung bình có trọng số là 0.850, độ chính xác có trọng số là 0.849,..v.v. Các thước đo này đưa ra một bức tranh toàn cảnh về hiệu suất của mô hình, có tính đến các yếu tối như độ nhạy, độ đặc hiệu, độ chính xác và sự cân bằng giữa chúng.

Bảng 6: Đánh giá tổng quát thuật toán Naïve bayes trong sự đo lường đào tạo 80% - kiểm tra 20%.

	Naïve bayes	
Các trường hợp được phân loại chính xác (Độ chính xác)	48	80%
Các trường hợp được phân loại không chính xác (Tỷ lệ lỗi)	12	20%
Tính đặc hiệu	78.26%	

- Độ chính xác = $\frac{12+36}{12+10+2+3} * 100 = 80\%$
- Tỷ lệ lỗi = $\frac{10+2}{60} * 100 = 20\%$
- Tính đặc hiệu = $\frac{36}{36+10} * 100 = 78.26\%$

4. Kết Luận

Trong bài luận này, chúng tôi đã tiến hành một đánh giá về việc sử dụng phương pháp thống kê Bayes trong phân loại bệnh nhân suy tim. Kết quả đánh giá cho thấy rằng phương pháp Bayes là một công cụ mạnh mẽ và hiệu quả cho việc phân loại bệnh nhân suy tim. Chúng tôi đã xây dựng một mô hình Bayes dựa trên dữ liệu thực tế và các đặc trưng quan trọng để tính toán xác suất mắc bệnh suy tim cho mỗi bệnh nhân. Mô hình đã được đánh giá bằng cách sử dụng các phương pháp đánh giá hiệu suất, bao gồm độ chính xác, độ nhạy, độ đặc hiệu và độ F1-score. Kết quả cho thấy rằng mô hình Bayes đạt được độ chính xác cao và khả năng phân loại tốt, vượt qua nhiều phương pháp khác trong việc phân loại bệnh nhân suy tim. Một điểm mạnh của phương pháp Bayes là khả năng tích hợp tri thức tiền định và dữ liệu thực tế để cung cấp dự đoán chính xác hơn. Việc sử dụng thông tin tiền định về tỷ lệ mắc bệnh suy tim trong dân số đã giúp cải thiện độ chính xác của mô hình. Đồng thời, phương pháp Bayes cũng cho phép chúng ta điều chỉnh mức độ tin tưởng vào thông tin tiền định và dữ liệu quan sát để tạo ra dự đoán linh hoạt và tin cậy. Tuy nhiên, cần lưu ý rằng phương pháp Bayes cũng có một số hạn chế. Việc ước lượng các tham số tiền định và tính toán xác suất có thể đòi hỏi sự hiểu biết sâu về lĩnh vực và dữ liệu được sử dụng. Đồng thời, việc xây dựng mô hình Bayes cần quan tâm đến các giả định và điều kiện tiên quyết, để đảm bảo tính chính xác và độ tin cậy của kết quả.

Tổng kết lại, phương pháp này là một công cụ hữu ích để phân loại bệnh nhân suy tim. Đánh giá của chúng tôi đã chứng minh khả năng phân loại chính xác và tin cậy của mô hình Machine Learning này. Tuy nhiên, cần tiếp tục nghiên cứu và phát triển phương pháp này để tối đa hóa hiệu quả và ứng dụng trong thực tế y tế.

Tài liệu tham khảo:

[1] Anitha, S., & Vanitha, M. (2022, February). Classification of VASA Dataset Using J48, Random Forest, and Naïve Bayes. In Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021) (pp. 283-291).

[2] .(2020) Classifying Patients with Myocardial Infarction and Heart Failure by Using SVM and KNN Learning Techniques .Journal of Administration and Economics.327-315 ,(126) .

[3] Brown, L. K., Williams, R. S., & Garcia, M. L. Application of Random Forest and Naïve Bayes algorithms in heart failure patient classification. *Cardiovascular Computing and Applications*.

[4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[5] Breiman, L. (2001). Random forests. Machine Learning 45 5–32.

[6] Biau, G. and Scornet, E. (2016). A random forest guided tour. Test 25 197–227.

[7] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29(2-3), 103-130.

[8] E. Alpaydin, "An Introduction to Machine Learning" The MIT press, Cambridge, Massachusetts, London, England, 2004.

[9] Faqe, M. Mohammad, S. & Hassan, K. "Using Random Forest algorithm to classify Iron anemia unspecified and coagulation defect unspecified diseases". A scientific Journal Issued by University of Sulaimani, Part (B- for Humanities) No. (68) (15 Feb.2022) .

[10] Lemons, K. (2020). A comparison between Naïve bayes and random forest to predict breast cancer. International Journal of Undergraduate Research and Creative Activities, 12(1).

[11] Lemons, K. (2020). A comparison between Naïve bayes and random forest to predict breast cancer. International Journal of Undergraduate Research and Creative Activities, 12(1).

[12] Pal, M., & Parija, S. (2021, March). Prediction of heart diseases using random forest. In Journal of Physics: Conference Series (Vol. 1817, No. 1, p. 012009). IOP Publishing.

[13] Rasul, Comparison between SVM and K-NN with application for diagnosis of heart disease patient. 2021.Master thesis, College of Administration & Economic, University of Sulaimani .

[14] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

[15] Runchuan Li, Shengya Shen, Xingjin Zhang, Runzhi Li, Shuhong Wang, Bing Zhou and Zongmin Wang, "Cardiovascular Disease Risk Prediction Based on Random Forest", Proceedings of the 2nd International Conference on Healthcare Science and Engineering, vol. 536, pp. 31-43, May 2019.

[16] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in aging neuroscience, 9, 3

BẢNG PHÂN CÔNG NHIỆM VỤ CHO BÁO CÁO CUỐI KÌ MÔN DS101				
STT	Công việc	Hạn	Người thực hiện	Mức độ hoàn thành
1	Phân công công việc	01/12/2023	Thiên Vũ	100%
2	Đưa ra ý tưởng	02/12/2023	Team	100%
3	Chốt ý tưởng	03/12/2023	Team	100%
4	Nộp đề tài	23/12/2023	Thiên Vũ	100%
5	Thảo luận	04/10/2023*	Team	100%
6	Tìm tài liệu tham khảo, dữ liệu	04/12/2023	Tri Vũ	98%
7	Phân tích	04/12/2023	Team	99%
8	Viết báo cáo	21/12/2023	Thiên Vũ	98%
9	Làm powerpoint	23/12/2023	Tri Vũ	100%
10	Sửa đổi, bổ sung báo cáo	24/12/2023	Tri Vũ	100%
11	Sửa đổi, bổ sung slide	24/12/2023	Thiên Vũ	100%
12	Rà soát, kiểm tra lại	24/12/2023	Team	100%
13	Thảo luận	25/12/2023	Team	100%
14	Tổng hợp	25/12/2023	Thiên Vũ	100%
15	Hoàn thành	25/12/2023	Team	100%
16	Thuyết trình (chia làm 2 phần)	28/12/2023	Team	

*Note: Đây là timeline khi làm lại và sửa đổi bài báo cáo mới do bài báo cáo đầu tiên không đạt yêu cầu khi nhận được kết quả của báo cáo tiến độ từ thầy