ỨNG DỤNG KỸ THUẬT MÁY HỌC VÀO PHÂN LOẠI BỆNH TIM

Trần Đình Toàn*, Dương Thị Mộng Thùy
Trường Đại học Công nghiệp Thực phẩm TP.HCM
*Email: *toantd@hufi.edu.vn*
Ngày nhận bài: 10/6/2022; Ngày chấp nhận đăng: 15/7/2022

TÓM TẮT

Trong nghiên cứu này, nhóm tác giả sử dụng kỹ thuật học máy vào phân loại bệnh tim dựa trên các triệu chứng và thông tin cận lâm sàng được ghi nhận trong tập dữ liệu của bệnh nhân. Thực nghiệm được tiến hành để phân loại có bệnh hay không có bệnh tim trên bộ dữ liệu công khai về bệnh tim lần lượt với thuật toán Naïve Bayes và mạng neuron nhân tạo (Artificial Neural Network - ANN). Kết quả thực nghiệm đạt được cho thấy việc ứng dụng kỹ thuật học máy vào phân loại bệnh tim đạt hiệu suất khá tốt với độ chính xác (Accuracy-Acc) lần lượt là 84% và 87%.

Từ khóa: Bệnh tim, Naïve Bayes, ANN, phân lớp.

1. GIỚI THIỆU

Bệnh tim mạch là do các rối loạn liên quan đến tim và mạch máu. Bệnh tim mạch bao gồm bệnh mạch vành (nhồi máu cơ tim), tai biến mạch máu não (đột quỵ), tăng huyết áp (cao huyết áp), bệnh động mạch ngoại biên, bệnh thấp tim, bệnh tim bẩm sinh và suy tim. Các nguyên nhân chính của bệnh tim mạch là do sử dụng thuốc lá, thiếu các hoạt động thể lực, chế độ ăn uống không lành mạnh và sử dụng rượu bia ở mức độ nguy hại. Theo các chuyên gia về y khoa, hầu hết các bệnh về tim mạch có thể phòng ngừa được bằng cách kiểm soát tốt các yếu tố nguy cơ dẫn đến bệnh này. Theo báo cáo của tổ chức y tế thế giới (WHO), bệnh tim mạch là nguyên nhân hàng đầu gây nên tử vong trên toàn cầu, chiếm tới 31% tổng số ca tử vong. Tại Việt Nam, cũng theo báo cáo bệnh tim mạch chiếm đến 31% tổng số ca tử vong trong năm 2016 tương đương với hơn 170.000 ca tử vong [1].

Trong những năm gần đây, máy học phát triển mạnh và được ứng dụng vào nhiều lĩnh vực của đời sống xã hội, trong đó có lĩnh vực y khoa do nhu cầu cao về phân tích dữ liệu để phát hiện các thông tin không xác định và có giá trị được hàm chứa trong các bộ dữ liệu y khoa. Trong số các kỹ thuật học máy được phát triển gần đây như khái quát hóa, đặc tính hóa, phân loại, phân cụm, kết hợp, so khớp mẫu, trực quan hóa dữ liệu, v.v. Việc học máy được áp dụng vào lĩnh vực y khoa cho một số lợi ích như phát hiện sớm các bệnh, giúp đưa ra giải pháp y tế cho bệnh nhân lựa chọn, phát hiện nguyên nhân bệnh, xác định và tư vấn các phương pháp y tế có thể sử dụng để điều trị bệnh. Học máy cũng giúp các nhà nghiên cứu chăm sóc sức khỏe thực hiện các chính sách chăm sóc sức khỏe hiệu quả, xây dựng hệ thống khuyến cáo thuốc, phát triển hồ sơ y tế của các cá nhân, ... Các kỹ thuật học máy được áp dụng trong các hệ thống chăm sóc sức khỏe cũng được sử dụng để phân tích các yếu tố khác nhau như loại thực phẩm, môi trường làm việc khác nhau, trình độ học vấn, điều kiện sống, nguồn nước sạch, dịch vụ chăm sóc sức khỏe, văn hóa và môi trường ảnh hưởng đối với các bệnh.

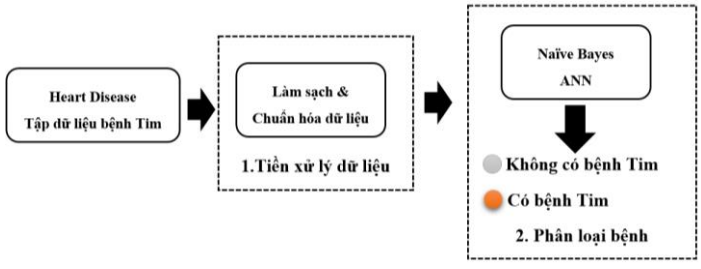
2. CÁC CÔNG TRÌNH LIÊN QUAN

Việc sử dụng các kỹ thuật máy học vào lĩnh vực y khoa tăng rất mạnh trong những năm gần đây với nhiều công trình đã được ghi nhận. Cụ thể, Victor Chang và cộng sự đã xây dựng hệ thống phát hiện bệnh tim dựa trên trí tuệ nhân tạo sử dụng các thuật toán học máy như logistic regression và random forest cho độ chính xác là 83% trên bộ dữ liệu huấn luyện [2]. Chithambaram T, Logesh Kannan N, and Gowsalya M đã đề xuất sử dụng 4 thuật toán là SVM, Decision tree, Random forest classifier and KNN vào phân loại bệnh tim kết quả đạt được khá tốt [3]. Tương tự, Likitha KN và cộng sự đã đề xuất sử dụng kỹ thuật Logistic Regression, KNN, Decision Tree, Naïve Bayes, Random Forest và SVM vào chẩn đoán bệnh tim [4]. Zaibunnisa L. H. Malik cùng cộng sự đã đề xuất sử dụng các thuật toán K Neighbours Classifier, SVM, Decision Tree, Random Forest vào xây dựng hệ thống chẩn đoán bệnh tim [5]. N. Deepika và cộng sự, đã áp dụng luật kết hợp vào phân loại bệnh nhân đau tim [6]. K. Srinivas và cộng sự đã sử dụng kỹ thuật khai phá dữ liệu vào dự đoán các cơn đau tim, trong đó cây quyết định đạt được hiệu suất tốt nhất [7]. Tương tự, A. Sudha và cộng sự sử dụng một số thuật toán phân loại để dự đoán bệnh đột quỵ, kết quả thực nghiệm cho thấy mạng nơron hoạt động tốt hơn nhiều so với các thuật toán còn lại [8]. Sujata Joshi và cộng sự đã tiến hành phân loại bệnh tim sử dụng kỹ thuật khai phá dữ liệu, kết quả với độ chính xác 84% [9]. Navdeep Singh và Sonika Jindal, đã sử dụng kỹ thuật chọn lọc đặc trưng và phân lớp dữ liệu bằng phương pháp Naïve Bayes để chẩn đoán bệnh tim [10]. H. Takci đã kết hợp giữa phương pháp học máy và phương pháp chọn lọc đặc trưng để chẩn đoán các cơn đau tim [11]. Hưng M.L. và nhóm nghiên cứu, đã chẩn đoán các bệnh về tim khác nhau dựa trên chọn lọc đặc trưng và kỹ thuật khai phá dữ liệu [12]. Trần Đình Toàn cùng nhóm nghiên cứu, đã áp dụng một số phương pháp học có giám sát để phân loại bệnh trên ba bộ dữ liệu cận lâm sàng là bệnh tim, bệnh thận mãn tính, và ung thư vú [13]. Devansh Shah cùng cộng sự sử dụng một số kỹ thuật học máy để hỗ trợ dự đoán bệnh tim như thuật toán KNN, Decision tree, Random Forest và Naïve Bayes kết quả đạt được tốt nhất với thuật toán KNN [14]. Saima Anwar Lashari cùng cộng sự đề xuất sử dụng một số kỹ thuật như SVM, DecisionTree, ANN, Bayesian Belief Network, KNN vào phân loại các bệnh trong y khoa [15].

Vẫn còn nhiều công trình nghiên cứu khác liên quan đến ứng dụng các kỹ thuật học máy và phân loại các bệnh dựa trên bộ dữ liệu được thu thập công khai. Kết quả đạt được nhìn chung vẫn còn khiêm tốn và nó cũng mở ra nhiều triển vọng cho các nhà nghiên cứu tìm hiểu cải tiến và phát triển để đạt hiệu suất phân loại tốt hơn, từ đó tiến tới xây dựng các hệ thống hỗ trợ y bác sỹ trong chẩn đoán bệnh và tư vấn điều trị cho bệnh nhân nhanh chóng kịp thời.

Nghiên cứu này đề xuất sử dụng 2 kỹ thuật trong học máy là Naïve Bayes (với kernel Gaussian và Bernoulli) và Mạng ANN vào phân loại bệnh tim (có nguy cơ và không có nguy cơ bệnh tim), trên bộ dữ liệu y khoa chứa thông tin cận lâm sàng của bệnh nhân. Phần còn lại của bài viết gồm: trình bày giản lược một số kỹ thuật đại diện cho phương pháp học máy được sử dụng trong bài viết ở phần 3; phần 4 trình bày kết quả thực nghiệm và bàn luận; phần 5 kết luận và hướng phát triển.

3. PHƯƠNG PHÁP NGHIÊN CỨU



Hình 1. Sơ đồ các bước thực hiện phân lớp bệnh tim

3.1. Naïve Bayes [16, 17]

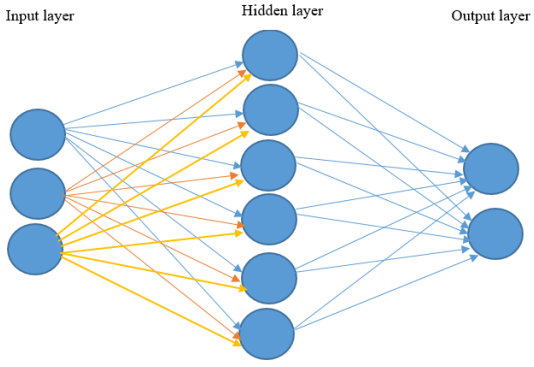
Phân lớp dựa trên lý thuyết Bayes được gọi là phân lớp Bayes. Định lý Bayes cung cấp cơ sở cho phân loại Naïve Bayes và Mạng Belief Bayes (BBN). Vấn đề chính với phân lớp Naïve Bayes là nó giả định rằng tất cả các thuộc tính là độc lập với nhau, trong khi thực tế các thuộc tính thuộc lĩnh vực y tế như triệu chứng bệnh và trạng thái sức khỏe như nhịp tim, huyết áp, đường huyết,… có mối tương quan với nhau. Mặc dù giả định các thuộc tính độc lập, phân lớp Naïve Bayes đã cho thấy hiệu quả về độ chính xác vì vậy nếu trong lĩnh vực y tế, các thuộc tính độc lập với nhau thì chúng ta có thể sử dụng phương pháp này. Định lý Bayes tập trung vào phân phối xác suất trước, sau và rời rạc của các mục dữ liệu (data items). Mạng Belief Bayes chủ yếu được sử dụng cho hệ thống phân loại bệnh nhân bị ung thư phổi. Và nó đã được sử dụng rộng rãi bởi nhiều nhà nghiên cứu trong lĩnh vực chăm sóc sức khỏe.

Mô hình Naïve Bayes có nhiều kernel khác nhau [17], tùy theo đặt tính dữ liệu và nhu cầu sử dụng sẽ có các sự kết hợp khác nhau. Trong nghiên cứu này, chúng tôi sử dụng mô hình Naïve Bayes với 2 kernel khác nhau là Gaussian và Bernoulli.

3.2. Mạng neuron nhân tạo (Artificial Neural Network - ANN) [16, 17]

ANN là một mô hình mô phỏng hệ thống thần kinh của con người. Hệ thống thần kinh của con người bao gồm các tế bào, được gọi là tế bào thần kinh. Các tế bào thần kinh sinh học được kết nối với nhau tại các điểm tiếp xúc, được gọi là khớp thần kinh. Với quá trình truyền thông tin từ tế bào thần kinh này qua tế bào thần kinh khác, các nhà khoa học đã tạo ra một mô hình máy tính có tính chất tương tự, nó có thể thực hiện tác vụ tính toán phức tạp nhanh hơn một hệ thống tính toán thông thường.

Một mạng neuron gồm có 3 layer như sau:



Hình 2. Cấu trúc đơn giản của mạng ANN

- Layer input: nhận giá trị đầu vào của các thuộc tính (feature) giải thích cho mỗi quan sát. Thông thường, số lượng node trong lớp đầu vào bằng số lượng thuộc tính (biến) giải thích.
- Layer ẩn: Các lớp ẩn áp dụng các phép biến đổi cho các giá trị đầu vào bên trong mạng. Trong lớp ẩn, quá trình xử lý thực tế được thực hiện thông qua một hệ thống các "kết nối" có trọng số. Có thể có một hoặc nhiều lớp ẩn, các lớp ẩn này thực hiện nhiều kiểu tính toán dạng toán học khác nhau trên dữ liệu đầu vào và nhận ra các mẫu là một phần công việc của nó.
- Layer output: Các lớp ẩn sau đó liên kết với "lớp đầu ra". Lớp đầu ra nhận các kết nối từ các lớp ẩn hoặc từ lớp đầu vào. Nó trả về một giá trị đầu ra tương ứng với dự đoán của biến phản hồi. Trong các bài toán phân loại, thường chỉ có một nút đầu ra. Các nút hoạt động của lớp đầu ra kết hợp và thay đổi dữ liệu để tạo ra các giá trị đầu ra.

Khả năng của mạng neuron trong việc cung cấp thao tác dữ liệu hữu ích nằm ở việc lựa chọn trọng số thích hợp. Điều này chính là điểm khác biệt với xử lý thông tin thông thường.

4. THỰC NGHIỆM VÀ THẢO LUẬN

4.1. Dữ liệu

Bộ dữ liệu bệnh tim được sử dụng trong nghiên cứu này là bộ dữ liệu công khai [18, 19] với 1025 mẫu, trong đó có 526 mẫu được gán nhãn thuộc lớp có nguy cơ mắc bệnh tim (target = 1), số mẫu còn lại 499 thuộc lớp không có nguy cơ (target = 0). Ở bộ dữ liệu gốc ban đầu, mỗi mẫu gồm 76 thuộc tính bao gồm các thuộc tính chẩn đoán khác nhau và thông tin y tế được thu thập từ mỗi bệnh nhân. Tuy nhiên, có nhiều thuộc tính do không có dữ liệu hoặc thiếu dữ liệu do đó đã được loại bỏ để còn lại 14 thuộc tính (13 thuộc tính chẩn đoán và 1 thuộc tính phân lớp).

Tiếp theo, chúng tôi đã tiến hành kiểm tra và loại bỏ các mẫu dữ liệu trùng lặp và kết quả còn lại là 302 mẫu được phân thành 2 lớp có nguy cơ mắc bệnh tim và không có nguy cơ mắc bệnh tim.

Bộ dữ liệu ghi nhận thông tin cận lâm sàng về bệnh tim của bệnh nhân nam và nữ, trưởng thành ở độ tuổi từ 29 đến 77 tuổi.

Bảng 1. Thuộc tính của bộ dữ liệu bệnh tim

STT	Thuộc tính	STT	Thuộc tính
1	age	8	thalach: maximum heart rate achieved
2	sex (1: male; 0: female)	9	exang: exercise induced angina (1 = yes; 0 = no)
3	cp: chest pain type	10	oldpeak = st depression induced by exercise relative to rest
4	trestbps: resting blood pressure	11	slope: the slope of the peak exercise st segment
5	chol: serum cholesterol in mg/dl	12	ca: number of major vessels (0-3) colored by fluoroscopy
6	fbs: (fasting blood sugar > 120 mg/dl)	13	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
7	restecg: resting electrocardiographic results	14	target: 1 = heart disease, 0 = No heart disease

Trong đó:

Age: Tuổi

Sex : giới tính (1 là nam, 0 là nữ)

Cp: loại đau ngực

1: cơn đau thắt ngực rõ ràng

2: đau thắt ngực không rõ ràng

3: không đau thắt ngực

4: không có triệu chứng

Trestbps: Huyết áp lúc nghỉ ngơi (đơn vị mmHg)

chol: cholestoral trong huyết thanh tính bằng mg/dl

fbs: (đường huyết lúc đói> 120 mg / dl): 1 = true; 0 = false

restecg: kết quả điện tâm đồ lúc nghỉ

thalach: nhịp tim tối đa

exang: đau thắt ngực do tập thể dục (1 = có; 0 = không)

oldpeak = ST ức chế do luyện tập so với khi nghỉ

slope: độ dốc của đoạn ST tập luyện đỉnh cao

1: dốc lên

2: bằng phẳng

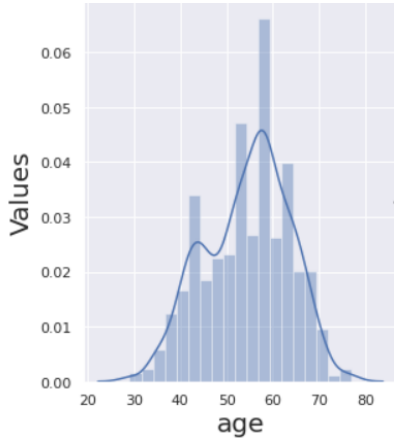
3: dốc xuống

ca: số lượng mạch chính (0-3) được tô màu bằng phương pháp soi huỳnh quang (nội soi)

thal: 3 = bình thường; 6 = khuyết tật cố định; 7 = khiếm khuyết có thể đảo ngược

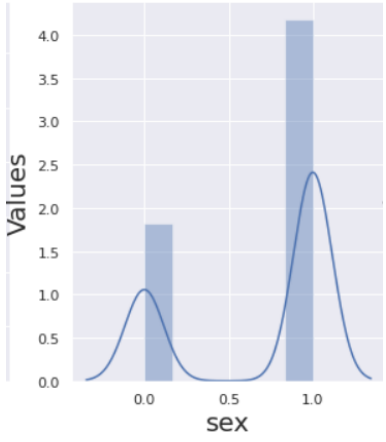
target: thuộc tính phân lớp (1 = Bệnh tim, 0 = Không có bệnh tim)

Tiến hành trực quan hóa để xem sự phân bố của bộ dữ liệu, kết quả cho thấy độ tuổi mắc bệnh tim nhiều nhất tập trung từ 45-65 tuổi, cao nhất ở khoảng 60 tuổi (Hình 3).



Hình 3. Biểu thị độ tuổi mắc bệnh tim trong tập dữ liệu

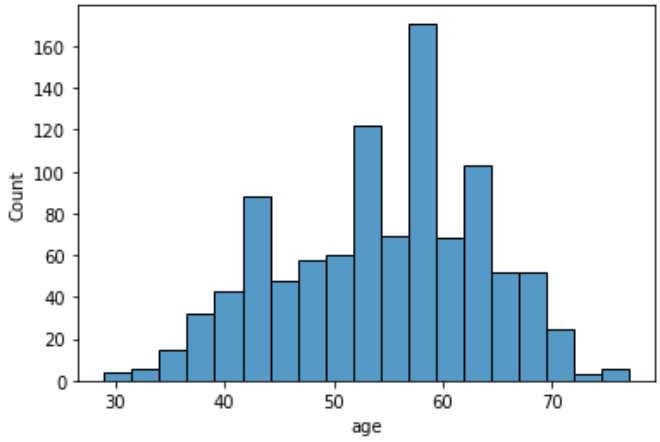
Từ tập dữ liệu cho thấy nam giới (sex = 1,0) có xu hướng mắc bệnh tim nhiều hơn nữ (sex = 0,0) (Hình 4)



Hình 4. Biểu thị nam mắc bệnh tim nhiều hơn nữ

Khi trực quan hóa tập dữ liệu này thay vì chỉ tính số lần xuất hiện của mỗi điểm dữ liệu là biểu đồ tần xuất, ở đây chúng tôi thực hiện phép biến đổi gaussian KDE (Kernel Density Estimation) dùng ước tính mật độ xác suất của biến age (tương tự cho các biến khác như sex,…). Giúp làm mịn dữ liệu cơ bản và làm biểu đồ smooth hơn cho chúng ta thấy tổng quan

hơn về thuộc tính này với một số lượng dữ liệu có hạn. Ưu điểm của phép biến đổi này chính là việc biến đổi phi tham số, thấy được hình dạng phân phối khác nhau ở các biểu đồ khác nhau và linh hoạt hơn trong việc khai phá dữ liệu (Hình 5).



Hình 5. Biểu đồ tần suất của biến Age

Nếu chỉ sử dụng biểu đồ tần suất, khi tập dữ liệu được bổ sung thêm 1 lượng dữ liệu nhưng không làm thay đổi phân phối thì khi trực quan bằng biểu đồ tần suất chúng ta sẽ khó phân biệt được phân phối trước và phân phối sau có giống nhau hay không.

Công thức chuyển đổi từ count sang value (các Hình 3, 4 và 5) như sau [20]:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Trong đó: $K(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, h là tham số bandwidth giúp phân phối 'mượt' hơn, n là số điểm dữ liệu

Chỉ số h trong công thức phải lớn hơn không (h>0) và có thể ước lượng theo công thức sau:

- $h = \left(\frac{4\sigma}{3n}\right)^{\frac{1}{5}}$
- $h = 0.9 \min\left(\sigma, \frac{IQR}{1.34}\right) n^{-\frac{1}{5}}$, trong đó IQR là khoảng giá trị từ tứ phân vị đầu tiên Q1 đến tứ phân vị cuối cùng Q3.

4.2. Thực nghiệm

Thực nghiệm được thực hiện trên tập dữ liệu gồm 302 mẫu, trong đó 162 mẫu của bệnh nhân có nguy cơ bệnh tim (target = 1), còn lại 140 mẫu của bệnh nhân không có nguy cơ bệnh tim (target = 0) và 14 thuộc tính được mô tả trong Bảng 1.

a. Đánh giá thực nghiệm

Dựa vào Độ chính xác, độ nhạy và độ đặc hiệu để đánh giá hiệu suất phân loại của mô hình. Độ chính xác, độ nhạy và độ đặc hiệu được xác định theo các công thức sau:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

Trong đó: TP, FP, TN và FN lần lượt là True Positive, False Positive, True Negative và False Negative.

b. Thực nghiệm

Trong phần này, chúng tôi tiến hành chia 80% tập dữ liệu ban đầu thành tập dữ liệu dùng để huấn luyện và 20% còn lại của tập dữ liệu được dùng để kiểm tra và tiến hành thực nghiệm phân lớp dữ liệu với các kỹ thuật nêu trên.

- Thực nghiệm 1: Sử dụng thuật toán Naïve Bayes lần lượt với nhân Gaussian và Bernoulli, kết quả đạt được về độ chính xác lần lượt là 83,82% và 83,40%, độ nhậy và độ đặc hiệu là như nhau lần lượt là 0,89 và 0,78, như trong Bảng 2:

Bảng 2. So sánh kết quả thực nghiệm của mô hình Naïve Bayes với 2 kernel

Methods	Accuracy (%)	Sensitivity	Specificity
Naïve Bayes (Gaussian)	83,82	0,89	0,78
Naïve Bayes (Bernoulli)	83,40	0,89	0,78

- Thực nghiệm 2: thực hiện trên mạng ANN lần lượt với các số epochs khác nhau là 50, 100, 150, 200 và 300, nhưng với epochs = 200, batch_size = 16, kết quả đạt được tốt nhất với độ chính xác là 87%, độ nhậy 0,91 và độ đặc hiệu 0,83.

Từ 2 kết quả thực nghiệm trên cho thấy, với bộ dữ liệu bệnh tim đang sử dụng trong nghiên cứu này thì khi dùng phương pháp ANN để phân loại có bệnh tim hay không có bệnh tim đạt kết quả tốt hơn phương pháp Naïve Bayes, như trình bày trong Bảng 3 và từ kết quả này cũng cho thấy phương pháp sử dụng trong nghiên cứu này đạt hiệu quả hơn một số công trình đã công bố trước đó [2-5].

Bảng 3. So sánh kết quả đạt được từ thực nghiệm của Naïve Bayes và ANN

Methods	Accuracy (%)	Sensitivity	Specificity
Naïve Bayes (Gaussian)	83,82	0,89	0,78
Naïve Bayes (Bernoulli)	83,40	0,89	0,78
ANN	87	0,91	0,83

5. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã tiến hành 2 thực nghiệm trên bộ dữ liệu bệnh tim. Thực nghiệm 1 sử dụng phương pháp Naïve Bayes với 2 kernel khác nhau để phân loại dữ liệu. Thực nghiệm 2 dùng kỹ thuật phân loại khác là mạng ANN cũng trên cùng bộ dữ liệu ở trên. Kết quả thực nghiệm của chúng tôi cho thấy rằng: (1) Dùng phương pháp Naïve Bayes nhưng với kernel Gaussian cho kết quả tốt hơn kernel còn lại, (2) Dùng mạng ANN với epochs=200 và batch_size = 16 cho kết quả tương đối tốt. Tuy nhiên tùy vào đặc tính khác nhau của dữ liệu mà các phương pháp khác nhau có thể cho kết quả hoàn toàn khác nhau. Nhìn chung trong nghiên cứu này, dùng phương pháp mạng ANN cho kết quả tốt hơn phương pháp Naïve Bayes.

Nhóm tác giả đã áp dụng thành công một số kỹ thuật máy học trong phân loại bệnh tim, từ kết quả đã đạt được cho thấy các kỹ thuật này có thể sử dụng vào các hệ thống hỗ trợ chăm sóc sức khỏe cộng đồng dựa trên dữ liệu cận lâm sàng của bệnh nhân.

Hướng tiếp theo, chúng tôi tiếp tục nghiên cứu cải tiến các mô hình đã thực hiện và các kỹ thuật phân loại dữ liệu khác, từ đó tìm ra được những kỹ thuật phân loại dữ liệu tốt nhất trên nhiều tập dữ liệu chăm sóc sức khỏe khác nhau.

TÀI LIỆU THAM KHẢO

1. https://www.who.int/vietnam/vi/health-topics/cardiovascular-disease
2. Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, and MA Hossain - An artificial intelligence model for heart disease detection using machine learning algorithms, Published by Elsevier Inc **2**, https://doi.org/10.1016/j.health.2022.100016, (2022) 1-17.
3. Chithambaram T., Logesh Kannan N., and Gowsalya M. - Heart disease detection using machine learning, Research Square (2020) 1-5. https://doi.org/10.21203/rs.3.rs-97004/v1
4. Likitha KN, Nethravathi R, Nithyashree K, Ritika Kumari, Sridhar N, and Venkateswaran K. - Heart disease detection using machine learning technique, DOI: 10.1109/ICESC51422.2021.9532705, IEEE, (2021) 1738-1743.
5. Zaibunnisa L. H. Malik, Momin Fatema, Nikam Pooja, and Gawandar Ankita - Prediction of Cardiovascular Disease Using Machine Learning Algorithms, IJERT **4** (4) (2021) 61-64.
6. Deepika N., Chandrashekar K. - Association rule for classification of heart attack patients, International Journal of Advanced Engineering Science and Technologies **11** (2) (2011) 253-257.
7. Srinivas K., Kavitha Rani B., and Govrdhan A. - Application of data mining techniques in healthcare and prediction of heart attacks, International Journal on Computer Science and Engineering **2** (2) (2011) 250-255.
8. Sudha A., Gayathiri P., and Jaisankar N. - Effective analysis and predictive model of stroke disease using classification methods, International Journal of Computer Applications **43** (14) (2012) 26-31.
9. Sujata Joshi, Mydhili K. Nair - Prediction of heart disease using classification based data mining techniques, Computational Intelligence in Data Mining **2**, Springer India, (2015) 502-511.
10. Navdeep Singh, Sonika Jindal - Heart disease prediction using classification and feature selection techniques, International Journal of Advance Research, Ideas and Innovations in Technology **4** (2) (2018) 1124-1127.
11. Takci H. - Improvement of heart attack prediction by the feature selection methods, Turkish Journal of Electrical Engineering & Computer Sciences **26** (1) (2018) 1-10.
12. Le Minh Hung, Tran Dinh Toan, Tran Van Lang - Automatic heart disease prediction using feature Selection and data mining technique, Journal of Computer Science and Cybernetics **34** (1) (2018) 33-47.
13. Trần Đình Toàn, Huỳnh Thị Châu Lan, Trần Văn Thọ, Hoàng Tùng, Lê Minh Hưng, Trần Văn Lăng - Data mining in healthcare system on patients clinical symptoms dataset, FAIR2019, Huế (2019) 92-101.

14. Devansh Shah, Samir Patel, Santosh Kumar Bharti - Heart Disease Prediction using Machine Learning Techniques, https://doi.org/10.1007/s42979-020-00365-y, Springer (2020) 1-6.
15. Saima Anwar Lashari, Rosziati Ibrahim, Norhalina Senan, and N. S.A. M Taujuddin - Application of data mining techniques for medical data classification: A Review, MATEC Web of Conferences **150**. EDP Sciences (2018) 1-6.
16. Charu C. Aggarwal, Data Mining, Springer, 2015.
17. https://scikit-learn.org
18. https://archive.ics.uci.edu/ml/datasets/heart+disease.
19. https://www.kaggle.com/datasets
20. En.wikipedia.org. 2022. Kernel density estimation - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Kernel_density_estimation> [Accessed 20 July 2022].

ABSTRACT

APPLICATION OF MACHINE LEARNING TECHNOLOGY TO CLASSIFICATION OF HEART DISEASES

Tran Dinh Toan*, Duong Thi Mong Thuy
Ho Chi Minh City University of Food Industry
*Email: *toantd@hufi.edu.vn*

In this study, we used machine learning to classify heart disease based on symptoms and laboratory information recorded in the patient dataset. Experiments were conducted to classify heart disease or not heart disease on the public data set of heart disease with the Naïve Bayes algorithm and Artificial Neural Network (ANN) respectively. The obtained experimental results show that the application of machine learning techniques to heart disease classification has quite good performance with accuracy (Accuracy-Acc) of 84% and 87%.

Keywords: Heart disease, Naïve Bayes, ANN, classification.