

GIẤY

Các mô hình học máy để phân loại và dự đoán trầm cảm ở sinh viên đại học

Orlando Iparraguirre-Villanueva¹(✉),
CleogePaulino-Moreno²,
AndrésEpifanía-Huerta¹,
CarmenTorres-Ceclén³Universidad
Tecnológica del Perú, Lima,
Perú²Universidad Católica
de Trujillo, Trujillo,
Perú³Universidad Católica
los Ángeles de Chimbote,
Chimbote,
Perúoiparraguirre@ieee.org

TÓM TẮT Mác áp lực là một tình trạng sức khỏe tâm thần ngày càng phổ biến trên toàn thế giới và bị ảnh hưởng bởi nhiều yếu tố khác nhau như lo lắng, thất vọng, béo phì, các vấn đề y tế, v.v. Trong trường hợp nghiêm trọng, nó thậm chí có thể dẫn đến tự tử. Nghiên cứu này nhằm mục đích sử dụng các mô hình máy học (ML) để phân loại và dự báo trầm cảm của học sinh. Nghiên cứu liên quan đến việc phân tích bộ dữ liệu của 787 sinh viên đại học thông qua một loạt các bước, bao gồm làm sạch, đào tạo mô hình và thử nghiệm bằng cách sử dụng các công nghệ để phân loại và dự đoán trầm cảm của sinh viên. Ba mô hình ML đã được sử dụng: logistic regression (LR), K-nearest neighbor (KNN) và cây quyết định (DT). Kết quả cho thấy mô hình LR đạt được độ chính xác cao nhất trong dự đoán, với tỷ lệ 77%, 70% nhớ lại và 72% điểm F1. Hơn nữa, nghiên cứu nhấn mạnh rằng hai trong số năm học sinh bị trầm cảm nhẹ, khoảng 90% học sinh trầm cảm không tìm cách điều trị, học sinh béo phì

Để bị trầm cảm gấp 2,5 lần, học sinh nam có nguy cơ béo phì cao gấp đôi và sinh viên nam thường có chỉ số khối cơ thể (BMI) cao hơn so với học sinh nữ. Nghiên cứu kết luận rằng việc tích hợp các mô hình ML vào các yếu tố kích hoạt dẫn đến trầm cảm ở học sinh.

Từ khóa phân loại, dự đoán, trầm cảm, học máy (ML), sinh viên

1 GIỚI THIỆU

Trầm cảm là một bệnh tâm thần ngày càng phổ biến trên toàn thế giới, ảnh hưởng đến khoảng 5,7% dân số trưởng thành [1]. Trên toàn thế giới, khoảng 280 triệu người bị trầm cảm [2]. Trầm cảm do nhiều yếu tố khác nhau gây ra, chẳng hạn như lo lắng, thất vọng và các vấn đề y tế [3], [4]. Bệnh này có thể trở thành một vấn đề sức khỏe nghiêm trọng, đặc biệt là khi nó tái phát và ở cường độ từ trung bình đến nặng, gây đau đớn cho cá nhân bị ảnh hưởng và làm gián đoạn cuộc sống hàng ngày của họ [5], [6]. Trong trường hợp xấu nhất, nó có thể dẫn đến tự tử [7]. Trong những năm gần đây, tự tử đã trở thành một trong những nguyên nhân gây tử vong hàng đầu, với tỷ lệ ngày càng tăng [8]. Hơn 700.000 người tự tử mỗi năm [9]. Những người trẻ tuổi ở các nước đang phát triển đặc biệt

Iparraguirre-Villanueva, O., Paulino-Moreno, C., Epifanía-Huerta, A., Torres-Ceclén, C. (2024). Các mô hình học máy để phân loại và dự đoán trầm cảm ở sinh viên đại học. Tạp chí Quốc tế về Công nghệ Di động Tương tác (IJIM), 18(14), trang 148–163. <https://doi.org/10.3991/ijim.v18i14.48669>Article đã gửi IJIM 2024-02-20. Bản sửa đổi đã tải lên ngày 18 tháng 4 năm 2024. Chấp nhận cuối cùng 2024-04-18. © 2024 bởi các tác giả của bài viết này. Xuất bản theo CC-BY.

để bị trầm cảm do những thách thức về xã hội, kinh tế, giáo dục và sức khỏe mà các quốc gia này phải đối mặt [10], [11]. Thế giới không ngừng thay đổi và sự phát triển công nghệ đóng vai trò quan trọng trong việc tạo ra những khả năng mới của con người [12]. Điều này đã khiến mọi người làm việc nhiều giờ hơn để theo kịp nhịp độ nhanh của thế giới, khiến họ phải chịu căng thẳng về thể chất và tinh thần ảnh hưởng đến sức khỏe của họ [13]. Theo Tổ chức Y tế Thế giới (WHO) [14], tỷ lệ trầm cảm đang tăng lên trên toàn cầu. Ví dụ, vào năm 2022, Ukraine báo cáo tỷ lệ suy thoái là 6,3%, tiếp theo là Hoa Kỳ là 5,9%, Úc là 5,9%, Phần Lan là 5,6%, Estonia là 5,9%, Hy Lạp là 5,7%, Brazil là 5,8%, Bồ Đào Nha là 5,7%, Belarus là 5,6% và Peru là 4,8%. Như minh họa trong Hình 1, các quốc gia có tỷ lệ lây lan thấp nhất trải rộng trên hầu hết các lục địa, với khu vực đảo Thái Bình Dương của Châu Đại Dương có tỷ lệ trầm cảm thấp nhất. Sinh viên đại học Peru phải đối mặt với nhiều thách thức cá nhân, học tập, kinh tế, sức khỏe, giao thông vận tải và các thách thức hàng ngày khác [15]. Một yếu tố quan trọng khác góp phần gây ra trầm cảm ở sinh viên đại học là việc quay trở lại các lớp học trực tiếp. Trong suốt đại dịch COVID-19, nhiều sinh viên đã trải qua căng thẳng tài chính nghiêm trọng đối với gia đình [16], khiến họ khó có thể chi trả các chi phí liên quan đến việc theo học đại học [17]. Ở Peru, khoảng 30% sinh viên đại học biểu hiện các triệu chứng trầm cảm, bao gồm buồn bã, cáu kỉnh, cảm giác trống rỗng, lo lắng, tội lỗi, lòng tự trọng thấp, thiếu hy vọng cho tương lai, suy nghĩ về cái chết hoặc tự tử và thiếu năng lượng [18]. Phương tiện truyền thông xã hội đã thay đổi cách chúng ta giao tiếp và trở thành một phần thiết yếu trong cuộc sống của nhiều người [19]. Những người bị trầm cảm thường có xu hướng tự tử và có thể chuyển sang mạng xã hội để thể hiện cảm xúc của họ [20], [21]. Tuy nhiên, việc sử dụng quá nhiều phương tiện truyền thông xã hội ở những người trẻ tuổi có liên quan đến tỷ lệ lo lắng, trầm cảm và rối loạn giấc ngủ cao hơn. Các mô hình và kỹ thuật học máy (ML) hiện đại ngày càng được sử dụng để dự đoán nguy cơ tự tử. Trong khi các phương pháp thống kê truyền thống vẫn được sử dụng, các mô hình ML kết hợp nhiều biến hơn để đánh giá rủi ro, dẫn đến cải thiện độ chính xác [22]. Các hệ thống dự đoán này tận dụng trí tuệ nhân tạo (AI) để đưa ra dự đoán chính xác bao cho các kịch bản mới [23]. Bộ dữ liệu cho nghiên cứu này bao gồm mười tám thuộc tính và bao gồm thông tin từ 787 sinh viên đại học tại Đại học Lahore vào năm 2021. Nghiên cứu nhằm mục đích phân loại và dự đoán trầm cảm của học sinh bằng cách sử dụng các kỹ thuật ML, sử dụng xác nhận chéo để xác định bộ phân loại hiệu quả nhất trong số các mô hình hồi quy logistic (LR), K-nearestneighbor (KNN) và cây quyết định (DT).

2 CÁC NGHIÊN CỨU TRƯỚC ĐÂY

Viện Sức khỏe Tâm thần Quốc gia (NIMH), Tổ chức Y tế Liên Mỹ (PAHO), các nhà nghiên cứu và học giả đã xuất bản công trình nghiên cứu liên quan đến sức khỏe tâm thần. Ví dụ, [24] đã phát triển một mô hình dựa trên DT để dự đoán nguy cơ rối loạn trầm cảm ở học sinh. Ngoài ra, các tác giả trong [25] đã làm việc trên mô hình aML để tìm ra một bộ phân loại tốt hơn và do đó đạt được kết quả cải thiện về độ chính xác của mô hình. Với mục đích này, họ đã sử dụng nhiều mô hình khác nhau như K-NN, RF, perceptron nhiều lớp (MLP), máy vectơ hỗ trợ (SVM) và logic mờ, đạt được mức hiệu suất lần lượt là 100%. Hơn nữa, trong bài báo [26], một mô hình ML đã được phát triển để phát hiện và dự đoán trầm cảm ở học sinh từ 4 đến 17 tuổi. Các yếu tố góp phần gây ra trầm cảm rất đa dạng, từ thiếu hỗ trợ xã hội đến

vấn đề tài chính và môi trường học tập. Tương tự, nghiên cứu [27] kiểm tra các vấn đề sức khỏe tâm thần ở học sinh. Họ cũng phân tích các mô hình ML để dự đoán sức khỏe tâm thần, với mô hình SVM nổi bật là phổ biến và chính xác nhất, đạt được độ chính xác từ 70% đến 96%. Tương tự như vậy, trong nghiên cứu [28], năm kỹ thuật ML đã được phân tích: RF, mạng nơ-ron, DT, SVM và Naïve Bayes (NB), với SVM và RF mang lại kết quả tốt nhất trong việc dự đoán trầm cảm. Hiện nay, ML được sử dụng rộng rãi để dự đoán cảm xúc và đặc điểm tâm lý, từ thiết kế đến triển khai. Trong bối cảnh này, nghiên cứu [29] đã thiết kế một mô hình dự đoán sử dụng các mô hình LR, NB, RF, DT và K-NN để xác định các rối loạn chính như lòng tự trọng chậm chạp, nghiện Internet và trầm cảm. Những tiến bộ công nghệ trong những năm gần đây đã đóng góp cho lĩnh vực y tế với nhiều kỹ thuật và công cụ khác nhau đã cải thiện kết quả và dự đoán đáng kể [29]. Do đó, các thuật toán ML khác nhau đã được phân tích để phân loại và phát hiện trầm cảm ở học sinh, với thuật toán SVM được xác định là hiệu quả nhất và mang lại kết quả tốt nhất để phát hiện trầm cảm. Phát hiện sớm trầm cảm là rất quan trọng, vì nó hỗ trợ phòng ngừa và đánh giá. Trong bối cảnh này, nghiên cứu [30] đã sử dụng các mô hình ML với 26 biến dự đoán để dự đoán nguy cơ tự tử ở thanh thiếu niên Hàn Quốc. Họ sử dụng các mô hình như LR, RF, SVM, ANN và tăng cường độ dốc cực cao. Kết quả cho thấy 12,4% thanh thiếu niên có tiền sử tự tử do trầm cảm. Mô hình tăng độ dốc cực cao hoạt động tốt nhất với 79%, tiếp theo là SVM ở mức 78,7%, LR ở mức 77,9%, RF ở mức 77,8% và ANN ở mức 77,5%. Các mô hình ML đã đóng góp đáng kể vào sự tiến bộ của các phương pháp chẩn đoán tự động trong các bệnh lý khác nhau. Ví dụ, trong một nghiên cứu được thực hiện với ML [31] dựa trên kết nối não, một tập hợp dữ liệu hình ảnh não đã được sử dụng để phát hiện trầm cảm. Kết quả được cung cấp bởi hầu hết các mô hình ML thiếu lời giải thích rõ ràng cho các cá nhân, khiến các dự đoán khó hiểu. Trong bài báo [32], một nghiên cứu đã được thực hiện để dự đoán các đặc điểm đa khía cạnh bằng cách sử dụng các mô hình học sâu. Kết quả chỉ ra rằng mô hình NB đạt được điểm F1 thấp nhất, trong khi mô hình học đa phương thức vượt trội hơn SVM, NB và BiGRU. Vì mô hình này được thiết kế đặc biệt để xác định người dùng trầm cảm, mô hình lai được đề xuất dựa trên MLP và CNN (MDHAN) đã chứng minh hiệu suất tốt nhất với điểm F1 89%, cho thấy rằng HAN với chiến lược đa khía cạnh đạt được độ chính xác đáng kể trong việc phát hiện trầm cảm. Ngoài ra, trong [33], họ đã phát triển một mô hình dựa trên ML để phân loại bệnh nhân dựa trên thông tin nhân khẩu học xã hội, đặc điểm tính cách và tâm trạng, kết luận rằng các mô hình LR và phương pháp Elastic Net mang lại kết quả với độ chính xác lần lượt là 84% và 80%. Để dự đoán trầm cảm bằng cách sử dụng các mô hình ML, nhiều thách thức phải được giải quyết, chẳng hạn như chất lượng dữ liệu, nhận dạng ký tự, khả năng diễn giải, mối quan tâm về đạo đức và quyền riêng tư, cũng như sự thay đổi cá nhân và thời gian. Khi sử dụng dữ liệu sức khỏe tâm thần để cải thiện độ chính xác và khả năng thích ứng, cần phải giải quyết những thách thức này để bảo vệ quyền riêng tư của dữ liệu.

3 PHƯƠNG PHÁP LUẬN

Phần này mô tả các mô hình ML sẽ được sử dụng, cũng như nghiên cứu điển hình sẽ được phát triển để giải quyết việc phân loại và dự đoán trầm cảm trong sinh viên đại học.

3.1 Hồi quy logistic

Mô hình học tăng cường (LR) được sử dụng để phân loại và phân tích dự đoán. Nó cũng thường được sử dụng để cố gắng tương quan xác suất xảy ra một sự kiện [34]. Khái niệm là mô hình LR ước tính xác suất mang lại bằng không khi không có sự kiện nào xảy ra hoặc một khi sự kiện xảy ra dựa trên giá trị của biến thể giải thích [35]. Trong LR, phép biến đổi logit được áp dụng cho xác suất, trong đó xác suất thành công được chia cho xác suất thất bại [36]. Điều này được gọi là xác suất logarit và được biểu diễn bằng phương trình (1).

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (1)$$

Phương trình (1) cho thấy biến độc lập với các tổ hợp $\beta_0 + \beta_1 X$, và biến phụ thuộc là xác suất ước tính $\pi(x)$. Mô hình LR trong ML thuộc danh mục học tập có giám sát. Trong bối cảnh này, mô hình LR sử dụng khả năng log âm làm hàm tổn thất và sử dụng quá trình đi xuống gradient để xác định mức tối đa toàn cầu và thu được các ước tính [36].

3.2 K-hàng xóm gần nhất

Thuật toán K-NN là một bộ phân loại học tập có giám sát phi tham số sử dụng khoảng cách để phân loại hoặc dự đoán [37]. Thuật toán lưu trữ các vector và nhãn attri-bute trong giai đoạn đào tạo để đào tạo lại [38]. Trong quá trình phân loại, K được định nghĩa là một biến do người dùng xác định và vector không được gắn nhãn được phân loại bằng cách đặt một nhãn trong số các thuộc tính đào tạo được coi là phù hợp nhất [39]. Khoảng cách Euclid được sử dụng cho các chỉ số khoảng cách cho các biến liên tục, giới hạn đối với các vector có giá trị thực. Phương trình (2) được sử dụng cho mục đích này, trong khi số liệu chồng lên được sử dụng để tạo các biến [40].

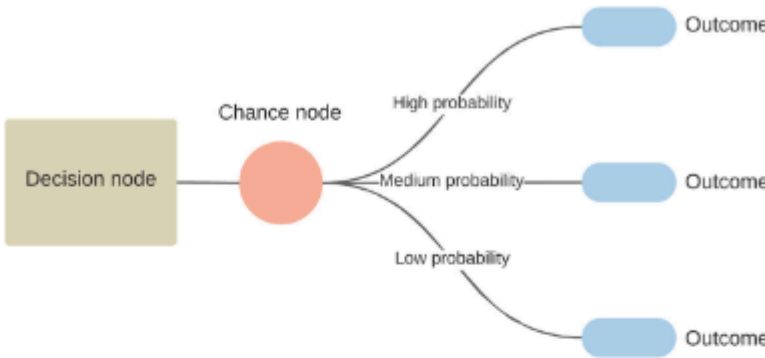
$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

Thuật toán K-NN được áp dụng trong ML cho nhiều mục đích khác nhau, chủ yếu là phân loại và dự đoán. Ví dụ: trong xử lý dữ liệu, thuật toán được sử dụng để ước tính các giá trị; trong các công cụ đề xuất, nó cung cấp các đề xuất tự động; trong tài chính được áp dụng vào dữ liệu tín dụng để đánh giá rủi ro và phân tích xu hướng kinh tế; Trong chăm sóc sức khỏe, nó đóng một vai trò quan trọng trong việc dự đoán nguy cơ đau tim và ung thư tuyến tiền liệt [41]. Ngoài ra, nó được sử dụng trong nhận dạng mẫu để hỗ trợ phân loại văn bản và hình ảnh.

3.3 Cây quyết định

Là một thuật toán phân loại và hồi quy phi tham số, cây quyết định (DT) được sử dụng để dự đoán các thuộc tính của các biến rời rạc và liên tục [42]. DTmodels được xây dựng tương tự như lưu đồ, trong đó mỗi nút đại diện cho một attri-bute và mỗi nhánh đại diện cho một đầu ra. Có ba loại nút trong DT [43]: nút quyết định, nút xác suất và nút cuối. Các nút quyết định thường là

được biểu thị bằng các hộp, các nút xác suất bằng hình tròn và các nút cuối bằng hình tam giác, như trong Hình 1.



Hình 1. Cây quyết định

3.4 Hiểu dữ liệu

Bộ dữ liệu được sử dụng để xử lý được lấy từ kho lưu trữ Kaggle và bao gồm 787 hồ sơ của sinh viên đại học. Một nghiên cứu đã được thực hiện bằng cách sử dụng bảng câu hỏi sức khỏe bệnh nhân (PHQ) để đánh giá mức độ nghiêm trọng của các vấn đề sức khỏe tâm thần (0–4: không tối thiểu hoặc bình thường; 5–9: nhẹ; 10–14: trung bình; 15–19: trung bình–nặng; 20–24: nặng). Các biến được phân tích bao gồm: php_score: điểm từ mức độ nghiêm trọng của PHQ questionnaire; ước tính bất nguồn từ trầm cảm php_score: cho biết liệu người tham gia có chẩn đoán trầm cảm hay không; cho biết liệu người tham gia có được chẩn đoán từ chuyên gia sức khỏe tâm thần hay không; liệu những người tham gia có đang điều trị trầm cảm hay không; xác suất của xu hướng tự tử từ điểm GAD: điểm từ đánh giá rối loạn lo âu mức độ nghiêm trọng của lo lắng; ước tính dựa trên chẩn đoán lo âu điểm GARD; chẩn đoán bởi chuyên gia sức khỏe tâm thần Điều trị lo âu: liệu người tham gia có đang được điều trị lo lắng hay không Điểm Epworth: điểm từ một bài kiểm tra đo độ buồn ngủ; khả năng buồn ngủ vào ban ngày BMI: chỉ số khối cơ thể who_bmp: Chỉ số khối cơ thể WHO tuổi: tuổi của người tham gia (từ 18 đến 31 tuổi) giới tính: giới tính của người tham gia; năm học, năm học. K-hàng xóm gần nhất.

3.5 Làm sạch dữ liệu

Phần này bắt đầu với việc làm sạch dữ liệu. Ví dụ: hàm 'dropna()' được sử dụng để loại bỏ các hàng có giá trị bị thiếu, trong khi các hàm 'Epworth()', 'loc()' và 'drop()' được sử dụng để xử lý dữ liệu ngoài phạm vi, xử lý dữ liệu không chính xác, ngoại lệ và loại bỏ dữ liệu trùng lặp. Sau đó, các biến được kiểm tra được hiển thị trong Bảng 1, hiển thị các biến có tổ chức mà kiểu dữ liệu phải được chuyển đổi thành các danh mục. Ngoài ra, số lượng thể loại đã được giảm và các chức năng đã được xác định để sắp xếp các danh mục. Cuối cùng, với

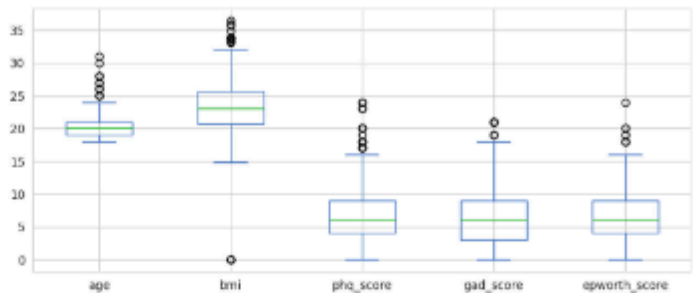
hàm 'reset_index()', một mẫu phân tầng được lấy từ mẫu origi -nal.

Bảng 1. Kiểm tra loại dữ liệu

#	Cột	Số lượng không rỗng	Loại D
0	school_year	757 [Không trống]	loại
1	tuổi	757 [Không trống]	Int64
2	giống	757 [Không trống]	loại
3	Bmi	757 [Không trống]	nôi64
4	who_bmi	757 [Không trống]	loại
5	phq_score	757 [Không trống]	Int6
6	depression_severity	757 [Không trống]	loại
7	trầm cảm	757 [Không trống]	đối tượng
8	Quyên	757 [Không trống]	đối tượng
9	depression_diagnosis	757 [Không trống]	đối tượng
10	depression_treatment	757 [Không trống]	đối tượng
11	gad_score	757 [Không trống]	Int64
12	anxiety_severity	757 [Không trống]	loại
13	lo lắng	757 [Không trống]	đối tượng
14	anxiety_diagnosis	757 [Không trống]	đối tượng
Dtypes: danh mục (5), float64 (2), int64 (3), đối tượng (8)			

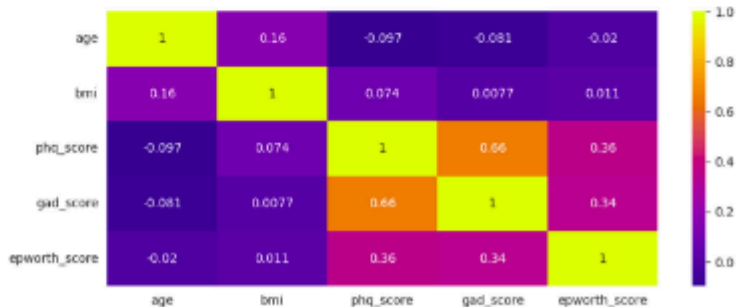
3.6 Phân tích dữ liệu thăm dò

Trong giai đoạn này, các thẻ đầu tiên được chuyển đổi thành định dạng số và sau đó chuyển đổi thành định dạng có thể hiểu được bởi thuật toán. Một giá trị số được gán cho mỗi giá trị phân loại. Các thuật toán ML có thể xác định tốt nhất cách xử lý thẻ. Bước này có tầm quan trọng tối quan trọng đối với việc xử lý dữ liệu và được thực hiện bằng cách sử dụng thư viện sklearn(), pandas() và NumPy(). Trích xuất ký tự được thực hiện bằng cách sử dụng kỹ thuật phân tích thành phần chính (PCA), sử dụng tất cả các tính năng do bộ dữ liệu cung cấp để dự đoán, phân tích, phân loại hoặc nhóm kết quả. Các đặc điểm có liên quan chặt chẽ đến các kích thước. Lựa chọn này xem xét một tập hợp con các đặc điểm quan trọng và loại bỏ những đặc điểm không đóng góp vào việc phân loại. Trích xuất tính năng cho phép tạo ra thông tin có giá trị từ dữ liệu thô bằng cách kết hợp và chuyển đổi các tính năng cốt lõi thành các tính năng mới cho đến khi một tập dữ liệu mới có thể được sử dụng bởi các mô hình ML. Sự phân bố của các biến số là một yếu tố quan trọng trong việc thực hiện các phép toán số học. Sử dụng hàm ggplot2(), chúng tôi đã quét để xác định các biến: tuổi, chỉ số khối cơ thể (BMI), điểm số để xác định mức độ nghiêm trọng của các vấn đề sức khỏe tâm thần (phq_score), tỷ lệ rối loạn lo âu tổng quát (gad_scor) và mức độ buồn ngủ ban ngày (Epworth_score).



Hình 2. Phân phối các biến số

Hình 2 chỉ ra rằng chỉ số BMI trung bình là 25, cho thấy hầu hết những người tham gia có cân nặng bình thường. Tuy nhiên, có một loạt các giá trị BMI, với một số người bị thừa cân và những người khác bị gầy. Ngoài ra, nó cho thấy rằng điểm PHQ là 8, cho thấy rằng phần lớn những người tham gia không biểu hiện các triệu chứng trầm cảm. Điểm GAD trung bình là 5, cho thấy hầu hết những người tham gia không lo lắng. Tuy nhiên, có một số người tham gia có điểm số cao có thể gây lo lắng. Tương tự, điểm trung bình của Epworth là 8, cho thấy hầu hết những người tham gia quần không bị buồn ngủ quá mức vào ban ngày. Phân tích thành phần chính (PCA) giúp xác định mối tương quan giữa các biến số và đặc điểm. Các biến trong Hình 3 mô tả các mối tương quan theo cặp trong tập hợp các biến. Ví dụ: các màu như đỏ, cam và yellow biểu thị mối tương quan dương, trong khi các màu như xanh lam, xanh lá cây và nâu đại diện cho mối tương quan tiêu cực. Cường độ của màu phản ánh sức mạnh của mối tương quan. Điều quan trọng là phải xác định rằng các giá trị trong ma trận tương quan nằm trong phạm vi từ -1 đến 1. Giá trị bằng không cho thấy không có mối tương quan giữa các biến. Giá trị 1 biểu thị mối tương quan dương hoàn hảo, trong khi -1 biểu thị tương quan âm hoàn hảo. Ví dụ, trong ma trận của Hình 3, rõ ràng là không có mối tương quan đáng kể giữa BMI, điểm PHQ và GAD. Có một mối tương quan yếu với điểm PHQ là 0,16, cho thấy rằng những học sinh có chỉ số BMI cao hơn có xu hướng có điểm PHQ cao hơn, chỉ ra các triệu chứng trầm cảm. Mối tương quan dương vừa phải với Epworth là 0,36 ngụ ý rằng những học sinh có điểm PHQ cao hơn biểu hiện nhiều triệu chứng trầm cảm hơn và có xu hướng buồn ngủ quá mức vào ban ngày. Ma trận cũng cho thấy mối tương quan tích cực mạnh mẽ với điểm số GAD, chỉ ra rằng những học sinh có điểm PHQ cao hơn có nhiều triệu chứng trầm cảm hơn và có xu hướng trải qua nhiều triệu chứng lo lắng hơn.



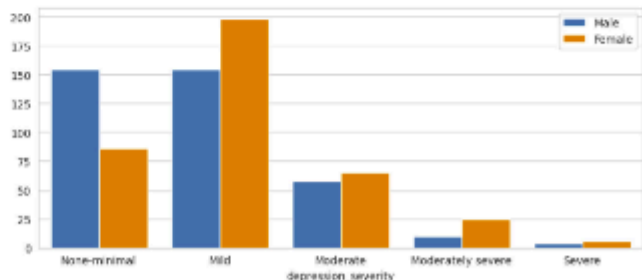
Hình 3. Ma trận tương quan giữa các biến số

Trong công trình này, tỷ lệ tính năng được thực hiện bằng cách sử dụng kỹ thuật tiêu chuẩn hóa, cho phép các tính năng được chia tỷ lệ sao cho các giá trị tập trung xung quanh chủ đề với độ lệch chuẩn đơn vị. Phương pháp này được sử dụng rộng rãi trong các thuật toán ML như SVM, LR và mạng nơ-ron. Tuy nhiên, các thuật toán như LR, K-NN và SVM yêu cầu các tính năng phải được chuẩn hóa. Bộ dữ liệu được sử dụng có các tính năng không đồng nhất ở mỗi thang đo do các thuộc tính khác nhau mà chúng đo lường. Dữ liệu được chuyển đổi sau đó được sử dụng để đào tạo với các hàm như StandarScaler(), loại bỏ giá trị trung bình và chia tỷ lệ từng tính năng thành phương sai đơn vị. Tuy nhiên, nó có thể bị ảnh hưởng bởi các ngoại lệ. Ngoài ra, có các hàm như scaler.fit_transform(x_train) và scaler.transform(x_test), như thể hiện trong Bảng 2.

Bảng 2. Mở rộng chức năng – tiêu chuẩn hóa

Ma trận		
(([[-1.25230037,	-0.51501498,	0.01682768],
[0.38909111,	-0.31446922,	0.65269925],
[0.93622161,	-0.2234338,	0.86465644],
..., [0.38909111, -1.11952747, -1.46687265],		
[-0.15803938, 0.27208495, -0.83100108], [1.4833521, -0.8437498, -0.4070867]		
]		
))		

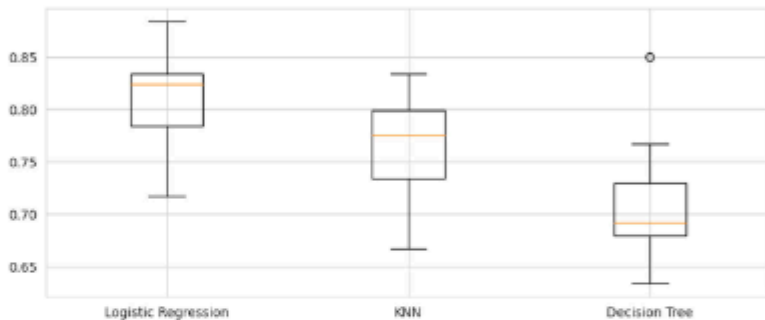
Sau khi tiến hành EDA, nó đã cho phép chúng tôi có được số liệu thống kê. Ví dụ, chỉ số khối cơ thể của một học sinh nam cao hơn của một học sinh nữ; khoảng 26% học sinh bị béo phì; Sinh viên nam có nguy cơ bị béo phì và bị trầm cảm nặng cao gấp đôi so với sinh viên nữ. Học sinh nữ có nguy cơ bị lo lắng nghiêm trọng cao gấp đôi so với nam sinh. Hai mươi lăm phần trăm sinh viên nữ có nhiều khả năng bị trầm cảm hơn học sinh nam. 25% học sinh trầm cảm được điều trị có ý định tự tử. Học sinh bị trầm cảm có chỉ số BMI cao hơn những học sinh không bị trầm cảm. Học sinh bị ám ảnh có nguy cơ bị trầm cảm cao gấp 2,5 lần so với học sinh bình thường. Khoảng 90% học sinh trầm cảm không được điều trị, và cuối cùng, 2 trong số 5 học sinh bị trầm cảm nhẹ. Hình 4 cho thấy tỷ lệ học sinh bị trầm cảm nặng. Theo phân tích, học sinh nữ có nhiều khả năng bị trầm cảm nặng; trong khi đó, học sinh nam có nhiều khả năng bị trầm cảm từ trung bình đến nặng; tương tự như vậy, trầm cảm vừa phải và nhẹ. Tuy nhiên, sinh viên nữ có nhiều khả năng không bị trầm cảm nào cả.



Hình 4. Tỷ lệ học sinh bị trầm cảm nặng

3.7 Đào tạo và thử nghiệm mô hình

Sau khi xử lý dữ liệu, trong phần này, đào tạo và thử nghiệm được thực hiện để có được bộ phân loại tốt nhất bằng cách sử dụng kỹ thuật xác thực chéo. Ngoài ra, khớp mô hình được thực hiện bằng cách sử dụng đường ống và cuối cùng, hiệu suất của mô hình được đánh giá. Mục đích của nghiên cứu này là xác định bộ phân loại hiệu quả nhất trong số các mô hình LR, K-NN và DT để dự đoán trầm cảm ở học sinh. Để đạt được điều này, kỹ thuật xác nhận chéo đã được sử dụng để đánh giá mô hình và hiệu suất của nó. Kỹ thuật này liên quan đến việc chia bộ dữ liệu thành hai phần: một phần để đào tạo (80%) và một phần để thử nghiệm (20%). Hơn nữa, kỹ thuật này liên quan đến việc đào tạo mô hình trên tập huấn luyện, xác thực nó trên tập thử nghiệm và lưu trữ kết quả xác thực. Các hàm tiếp theo đã được sử dụng để đào tạo: `logisticregression()`, `Kneighborsclassifier()`, `decisiontreeclassifier()`, `Kfold()`, `cross_val_score()` và `boxplot()`, cho phép thu thập các kết quả được mô tả trong Hình 5.



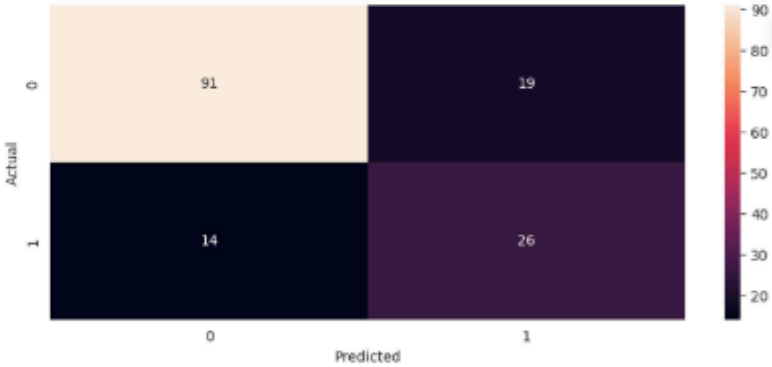
Hình 5. Xác nhận chéo để có được mô hình tốt nhất

4 KẾT QUẢ

Phần này trình bày kết quả của các giai đoạn phân loại, đào tạo và khám phá. Sau khi xử lý để xếp hạng mô hình tốt nhất để dự đoán, Hình 5 cho thấy thuật toán LR đạt được kết quả tốt nhất với tỷ lệ chính xác cao nhất, vượt quá 85%. Do đó, dự đoán trong nghiên cứu này được thực hiện bằng cách sử dụng mô hình này. Bảng 3 hiển thị chỉ số chính xác của ba mô hình. Ngoài ra, trong Hình 6, ma trận nhầm lẫn đã được sử dụng để đánh giá sự phân loại của mô hình dựa trên ML đã chọn. Các giá trị trên đường chéo thứ 91 và 26 đại diện cho các giá trị ước tính chính xác của mô hình, bao gồm dương tính thực và âm tính thực. Ngược lại, các giá trị trên đường chéo thứ hai cho biết sai số của mô hình, với 14 âm tính giả và 19 dương tính giả. Tỷ lệ chính xác đạt được với mô hình LR là 81%, nằm trong phạm vi chấp nhận được với độ chính xác là 71%. Do đó, có thể kết luận rằng mô hình chính xác hơn chính xác. Độ nhạy, là tỷ lệ các trường hợp dương tính được mô hình xác định trên tổng số trường hợp dương tính, được báo cáo là 48%, cho thấy khả năng phát hiện các trường hợp có liên quan của mô hình.

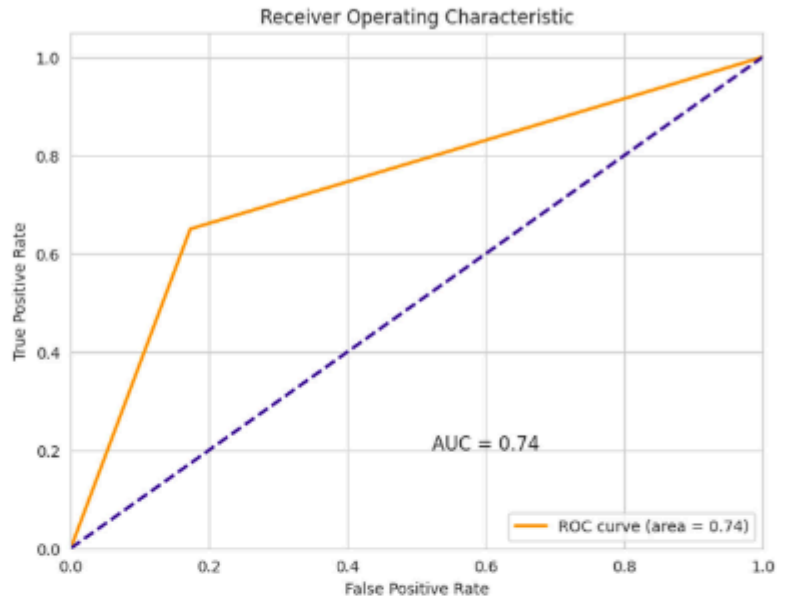
Bảng 3. Kiểm tra loại dữ liệu

Độ chính xácĐộ chính xác			Nhớ	Điểm F1	Hỗ trợ
LR	0	0.83	0.93	0.88	439
	1	0.71	0.48	0.57	159
	Chính xác			0.81	159
	Macro trung bình	0.77	0.70	0.72	159
	Trọng	0.80	0.81	0.80	159
KNN	0	0.81	0.90	0.85	439
	1	0.60	0.41	0.49	159
	Chính xác			0.77	598
	Macro trung bình	0.70	0.65	0.67	598
	Trọng	0.75	0.77	0.75	598
TIẾNG ĐỨC	0	0.80	0.79	0.79	439
	1	0.44	0.45	0.44	159
	Chính xác			0.70	598
	Macro trung bình	0.62	0.62	0.62	598
	Trọng	0.70	0.70	0.70	598



Hình 6. Ma trận nhầm lẫn

Đường cong ROC được sử dụng để đánh giá khả năng phân biệt giữa hai lớp của một mô hình. Ví dụ: Hình 7 trình bày đường cong ROC của mô hình đã chọn, cho thấy mối quan hệ giữa tỷ lệ dương tính giả và tỷ lệ vị trí thực. Diện tích dưới đường cong (AUC) là 0,74, cho thấy mô hình LR có khả năng phân biệt tốt giữa các trường hợp giảm giá tích cực và tiêu cực. Điều quan trọng cần lưu ý là đường cong ROC không phải là thước đo hoàn hảo về độ chính xác của mô hình. Hiệu suất của mô hình có thể thay đổi tùy thuộc vào kích thước của bộ dữ liệu được sử dụng.



Hình 7. Đường cong ROC của mô hình

5 THẢO LUẬN

Hiệu suất của các mô hình ML trong việc dự đoán các yếu tố gây trầm cảm ở học sinh đạt được bằng cách áp dụng ma trận chéo với các chỉ số sau: tính chính xác, độ chính xác, khả năng nhớ lại và điểm F1, trong đó 18 biến đầu vào đã được sử dụng (năm học, tuổi, giới tính, BMI, WHO_BMI, điểm PHQ, mức độ trầm cảm, trầm cảm, tự tử, chẩn đoán trầm cảm, điều trị trầm cảm, điểm GARD, buồn ngủ, Điểm Epworth, mức độ nghiêm trọng của lo lắng, chẩn đoán lo lắng, điều trị lo âu). Trong số các mô hình được đào tạo trong nghiên cứu này: LR, KNN và DT, mô hình LR mang lại kết quả đào tạo tốt nhất để dự đoán trầm cảm (LR: 77%; KNN: 70%; DT: 62%). Những phát hiện này phù hợp với các chỉ số được báo cáo trong một nghiên cứu trước đó [28], trong đó các mô hình ML SVM và RF được sử dụng để dự đoán trầm cảm và lo lắng, đạt được độ chính xác 92,5% đối với mô hình SVM và 76,4% đối với mô hình RF. Sự khác biệt trong các chỉ số này có thể là do nhiều yếu tố khác nhau, chủ yếu là kích thước của bộ dữ liệu. Cũng cần lưu ý khi so sánh những kết quả này với kết quả của một nghiên cứu khác [26], trong đó các mô hình DT và RF của ML được sử dụng để dự đoán trầm cảm ở trẻ em và thanh thiếu niên từ 4 đến 17 tuổi. Kết quả chỉ ra rằng mô hình RF đạt được độ chính xác cao nhất ở mức 95%. Điều này chứng minh rằng các mô hình có thể mang lại các kết quả khác nhau dựa trên quá trình đào tạo của chúng. Như đã chỉ ra trong tài liệu tham khảo [28] và [26], nghiên cứu của chúng tôi đã tiết lộ mối quan hệ chặt chẽ giữa béo phì, lo lắng và trầm cảm. Hơn nữa, các chỉ số dự đoán thu được có thể so sánh với tỷ lệ chính xác được báo cáo trong các nghiên cứu như [25], khám phá các rối loạn tâm thần. Nghiên cứu kết luận rằng các yếu tố có ảnh hưởng nhất đối với dự đoán lo lắng và trầm cảm là béo phì, tuổi tác, chỉ số BMI và lo lắng, khiến chúng trở thành bốn đặc điểm quan trọng nhất để dự đoán trầm cảm.

Đóng góp của nghiên cứu này nằm ở việc phân loại mô hình LR là yếu tố quan trọng nhất để dự đoán trầm cảm. Nó cũng cho phép xếp hạng những học sinh có nguy cơ béo phì, trầm cảm nhẹ, lo lắng nghiêm trọng, suy nghĩ tự tử và yếu tố BMI liên quan trực tiếp đến trầm cảm. Cuối cùng, chúng tôi nhận thấy rằng BMI có liên quan trực tiếp đến trầm cảm, cho thấy sự cần thiết phải có một cách tiếp cận toàn diện để điều trị sức khỏe thể chất và tinh thần. Những phát hiện này thể hiện một bước đột phá đáng kể trong lĩnh vực sức khỏe tâm thần. Do đó, chúng tôi có thể phát triển các chiến lược phòng ngừa và can thiệp sớm để nâng cao sức khỏe của học sinh.

6 KẾT THÚC

Bài báo đã đánh giá ba mô hình ML: LR, K-NN và DT, dựa trên học tập giám sát, để dự đoán các yếu tố liên quan đến trầm cảm ở sinh viên đại học. Theo kết quả, nghiên cứu kết luận rằng mô hình LR là hiệu quả nhất và thích hợp để dự đoán trầm cảm của học sinh, được chỉ ra bởi các chỉ số sau: độ chính xác (77%), nhớ lại: (70%) và điểm F1 (72%). Mô hình này có thể được tích hợp vào hệ thống thông tin của trường đại học để tự động truy cập thông tin của sinh viên dựa trên các biến chính. Ngoài ra, nghiên cứu điển hình đã phân loại thành công bộ dữ liệu. Ví dụ, khoảng hai trong số năm học sinh bị trầm cảm nhẹ; khoảng 90% học sinh bị trầm cảm không được điều trị; Học sinh béo phì có nguy cơ bị trầm cảm cao gấp 2,5 lần; học sinh nam có khả năng bị béo phì gấp đôi; và học sinh nam có chỉ số BMI cao hơn học sinh nữ. Thông tin này rất quan trọng cho việc ra quyết định. Các cơ quan quản lý trường đại học có thể sử dụng nó để giải quyết các vấn đề ảnh hưởng đến sức khỏe tinh thần của sinh viên. Hơn nữa, bằng cách sử dụng các kỹ thuật ML như mô hình LR, các chuyên gia y tế và những người ra quyết định có thể triển khai chúng trong các chương trình can thiệp để ngăn ngừa và tăng cường sức khỏe, giáo dục và hạnh phúc của học sinh. Mặc dù đã đạt được tiến bộ đáng kể trong lĩnh vực này, nhưng điều cần thiết là phải thừa nhận những hạn chế nhất định có thể ảnh hưởng đến nghiên cứu trong tương lai trong lĩnh vực này. Những hạn chế của nghiên cứu bao gồm kích thước mẫu, khả năng các biến bị bỏ qua và sự vắng mặt của xác nhận bên ngoài. Trong tương lai, nghiên cứu có thể nhấn mạnh các can thiệp phòng ngừa, đánh giá theo chiều dọc, kết quả thử nghiệm và nâng cao dự đoán thông qua các mô hình thống kê.

7 THAM KHẢO

- [1] WHO, "Rối loạn trầm cảm (trầm cảm)," Tổ chức Y tế Thế giới. Truy cập: Tháng Mười. 06, 2022. [Trực tuyến]. Có sẵn: <https://www.who.int/es/news-room/fact-sheets/detail/depression>[2] Sooma, "Điều trị trầm cảm không nên khó," soomamedical.com. Truy cập: 18 tháng 10, 2022. [Trực tuyến]. Có sẵn: <https://soomamedical.com/treating-depression-shouldnt-be-hard/> / # [3] F. S. Wicke và cộng sự, "Mối liên hệ giữa trầm cảm và tỷ lệ tử vong do mọi nguyên nhân: Giải thích các yếu tố và ảnh hưởng của giới tính," *J. Affect. Disord.*, tập 303, trang 315–322, 2022. <https://doi.org/10.1016/j.jad.2022.02.034>[4] E. N. Coman, H. Z. Wu và S. Assari, "Khám phá nguyên nhân gây trầm cảm và lo lắng sức khỏe chệnh lệch (HD) bằng cách kiểm tra sự khác biệt giữa các cá nhân phù hợp 1: 1," *Brain Sci.*, tập 8, số 12, tr. 207, 2018. <https://doi.org/10.3390/brainsci8120207>

- [5] N. F. Jamaludin, T. S. M. T. Wook, S. F. M. Noor và F. Qamar, "Thiết kế trò chơi hóa các yếu tố để tăng cường động lực thanh thiếu niên trong chẩn đoán trầm cảm," *Tạp chí Quốc tế về Công nghệ Di động Tương tác*, tập 15, số 10, trang 154–172, 2021.
<https://doi.org/10.3991/ijim.v15i10.21137>[6] M. A. Taheri và G. Rahimi, "Tự tử như được tiếp cận bởi Chủ nghĩa thần bí Halqeh," *Procedia – Khoa học Xã hội và Hành vi*, tập 114, trang 102–106, 2014.
<https://doi.org/10.1016/j.sbspro.2013.12.666>[7] R. S. Patel, T. Machado và W. E. Tankersley, "Rối loạn ăn uống và hành vi tự tử trong Thanh thiếu niên bị trầm cảm nặng: Thông tin chi tiết từ các bệnh viện Hoa Kỳ," *Khoa học Hành vi*, tập 11, số 5, tr. 78, 2021. <https://doi.org/10.3390/bs11050078>[8] Viện Sức khỏe Tâm thần Quốc gia, "Tự tử", [nimh.nih.gov](https://www.nimh.nih.gov), 2024. Truy cập: Tháng Mười Một 01, 2022. <https://www.nimh.nih.gov/health/statistics/suicide>[9] M. Chatterjee, P. Kumar, P. Samanta và D. Sarkar, "Phát hiện ý định tự tử từ phương tiện truyền thông xã hội trực tuyến: Kỹ thuật dựa trên tính năng đa phương thức," *Tạp chí Quốc tế về Thông tin Quản lý Dữ liệu Thông tin Thông tin* Tập 2, số 2, 100103, 2022. <https://doi.org/10.1016/j.ijime.2022.100103>[10] SE Murphy, LP Capitão, SLC Giles, PJ Cowen, A. Stringaris và CJ Harmer, "Những điều đã biết và chưa biết về điều trị SSRI ở những người trẻ tuổi bị trầm cảm và lo lắng: Hiệu quả, yếu tố dự đoán và cơ chế hoạt động," *Lancet Psychiatry*, tập 8, số 9, trang 824–835, 2021. [https://doi.org/10.1016/S2215-0366\(21\)00154-1](https://doi.org/10.1016/S2215-0366(21)00154-1)[11] JK Tan và cộng sự, "Tỷ lệ mắc bệnh và các yếu tố nguy cơ liên quan đến trầm cảm không được chẩn đoán giữa những người bị động kinh trong xã hội đa sắc tộc," *Epilepsy Res.*, tập 178, tr. 106772, 2021. <https://doi.org/10.1016/j.eplesyres.2021.106772>[12] L. Markauskaite và cộng sự, "Suy nghĩ lại về sự đan xen giữa trí tuệ nhân tạo và học tập của con người: Người học cần những khả năng nào cho một thế giới với AI?" *Máy tính và Giáo dục: Trí tuệ nhân tạo*, tập 3, tr. 100056, 2022.
<https://doi.org/10.1016/j.cacai.2022.100056>[13] C. Son, S. Hegde, A. Smith, X. Wang và F. Sasangohar, "Ảnh hưởng của COVID-19 đối với sức khỏe tâm thần của sinh viên đại học ở Hoa Kỳ: Nghiên cứu khảo sát phòng vắ", *J. Med. Internet Res.*, tập 22, số 9, tr. e21279, 2020. <https://doi.org/10.2196/21279>[14] "Trầm cảm và các rối loạn tâm thần phổ biến khác," WHO, 2022.
<https://www.who.int/publications/i/item/depression-global-health-estimates>[15] R. Alejandro-Salinas, AC Rojas-Cueva, F. J. Visconti-Lopez, M. L. Osorio-Martinez, và C. J. Toro-Huamanchumo, "Các yếu tố liên quan đến các triệu chứng rối loạn căng thẳng sau chấn thương trong bối cảnh sau cách ly của đại dịch COVID-19 ở sinh viên y khoa Peru," *Heliyon*, tập 8, số 5, trang e09446, 2022.
<https://doi.org/10.1016/j.heliyon.2022.e09446>[16] SM Islam và A. Habib, "Các công ty đầu tư tác động đang phản ứng như thế nào để duy trì và phát triển các doanh nghiệp kinh tế xã hội trong bối cảnh đại dịch COVID-19," *Tạp chí BusinessVenturing Insights*, tập 18, trang e00347, 2022.
<https://doi.org/10.1016/j.jbvi.2022.e00347>[17] RMR Parra, "Trầm cảm và ý nghĩa cuộc sống ở sinh viên đại học trong thời kỳ đại dịch," *Tạp chí Tâm lý Giáo dục Quốc tế*, tập 9, số 3, trang 223–242, 2020. <https://doi.org/10.17583/ijep.2020.6784>[18] J. Huarcaya-Victoria và cộng sự, "Các yếu tố liên quan đến sức khỏe tâm thần ở sinh viên y khoa Peru trong đại dịch COVID-19: Một nghiên cứu định lượng đa trung tâm," *Rev ColombPsiquiatr*, tập 52, số 3, trang 236–244, 2023.
<https://doi.org/10.1016/j.rcp.2021.06.002>[19] S. Kim và K. Lee, "Hiệu quả của việc dự đoán ý định tự tử thông qua các triệu chứng trầm cảm và cô lập xã hội bằng kỹ thuật học máy," *J. Pers. Med*, tập 12, số 4, tr. 516, 2022. <https://doi.org/10.3390/jpm12040516>[20] M. Garg, "Định lượng xu hướng tự tử trên mạng xã hội: Một cuộc khảo sát," *arXiv preprint arxiv: 2110.03663*, 2021.
<https://doi.org/10.48550/arxiv.2110.03663>

- [21] A. Stathopoulou và cộng sự, "Quy trình đánh giá di động về sức khỏe tâm thần và khả năng đọc viết kỹ năng trong giáo dục," *Tạp chí Quốc tế về Công nghệ Di động Tương tác*, tập 12, số 3, trang 21–37, 2018. <https://doi.org/10.3991/ijim.v12i3.8038>[22] Y. Huang và cộng sự, "So sánh ba mô hình học máy để dự đoán ý định tự tử-
ation và trầm cảm ở thanh thiếu niên Trung Quốc: Một nghiên cứu cắt ngang," *J. Affect. Disord.*, tập 319, trang 221–228, 2022. <https://doi.org/10.1016/j.jad.2022.08.123>[23] C. H. Espino-Salinas và cộng sự, "Mạng nơ-ron tích chập hai chiều cho depres-
phát hiện các tập sion trong thời gian thực bằng cách sử dụng chuỗi thời gian hoạt động vận động của bộ dữ liệu Depresjon," *Kỹ thuật sinh học*, tập 9, số 9, tr. 458, 2022. <https://doi.org/10.3390/bioengineering9090458>[24] Y. Y. Song và Y. Lu, "Phương pháp cây quyết định: Ứng dụng phân loại và dự đoán-
tion," *Thượng Hải Arch. Tâm thần học*, tập 27, số 2, tr. 130, 2015. <https://shanghaiarchivesof-psychiatry.org/en/215044.html>[25] T. Iliou và cộng sự, "Phương pháp tiền xử lý học máy ILIOU cho loại
trầm cảm
dự đoán," *Hệ thống phát triển*, tập 10, số 1, trang 29–39, 2017. <https://doi.org/10.1007/s12530-017-9205-9>[26] UM Haque, E. Kabir và R. Khanam, "Phát hiện trầm cảm ở trẻ em bằng máy
phương pháp học tập," *PLoS One*, tập 16, số 12, tr. e0261131, 2021. <https://doi.org/10.1371/journal.pone.0261131>[27] T. Zhang, Z. Zhong, W. Mao, Z. Zhang và Z. Li, "Một
lớp học máy mới
phương pháp dự đoán trung bình điểm cho sinh viên đại học kết hợp đánh giá tâm lý trong thời kỳ
hậu COVID-19," *Electronics (Thụy Sĩ)*, tập 13, số 10, tr. 1928, 2024. <https://doi.org/10.3390/electronics13101928>[28] R. Qasrawi, SP Vicuna Polo, D. Abu Al-Halawa, S. Hallaq và Z. Abdeen, "Đánh giá và dự đoán các yếu tố nguy cơ trầm cảm và lo lắng ở học sinh: Phân tích hiệu suất kỹ thuật học máy," *Biểu mẫu JMIR. Res.*, tập 6, số 8, tr. e32736, 2022. <https://doi.org/10.2196/32736>[29] S. Aleem, NU Huda, R. Amin, S. Khalid, SS Alshamrani và A. Alshehri, "Thuật toán học máy cho trầm cảm: Chẩn đoán, hiểu biết sâu sắc và hướng nghiên cứu," *Điện tử*, tập 11, số 7, tr. 1111, 2022. <https://doi.org/10.3390/electronics11071111>[30] JS Jung, SJ Park, EY Kim, KS Na, YJ Kim và KG Kim, "Các mô hình dự đoán nguy cơ tự tử cao ở thanh thiếu niên Hàn Quốc bằng kỹ thuật học máy," *PLoS One*, tập 14, số 6, trang e0217639, 2019. <https://doi.org/10.1371/journal.pone.0217639>[31] M. Mousavian, J. Chen, Z. Traylor và S. Greening, "Phát hiện trầm cảm từ hình ảnh sMRI và rs-fMRI bằng cách sử dụng học máy," *Tạp chí Hệ thống Thông tin Thông minh*, tập 57, số 2, trang 395–418, 2021. <https://doi.org/10.1007/s10844-021-00653-w>[32] H. Zogan, I. Razzak, X. Wang, S. Jameel và G. Xu, "Phát hiện trầm cảm có thể giải thích với các tính năng đa khía cạnh bằng cách sử dụng mô hình học sâu lai trên mạng xã hội," *WorldWide Web*, tập 25, số 1, trang 281–304, 2022. <https://doi.org/10.1007/s11280-021-00992-2>[33] PJC Suen, S. Goerigk, LB Razza, F. Padberg, IC Passos và AR Brunoni, "Phân loại trầm cảm đơn cực và lưỡng cực bằng kỹ thuật học máy," *Psychiatry Res.*, tập 295, tr. 113624, 2021. <https://doi.org/10.1016/j.psychres.2020.113624>[34] LH John, JA Kors, JM Reys, PB Ryan và PR Rijnbeek, "Mô hình hồi quy logistic để dự đoán cấp bệnh nhân dựa trên dữ liệu quan sát khổng lồ: Chúng ta có cần tất cả dữ liệu không?" *Int. J. Med. Thông tin.*, tập 163, tr. 104762, 2022. <https://doi.org/10.1016/j.ijmedinf.2022.104762>[35] KV Mhetre và AD Thube, "Phát triển các mô hình an toàn đường bộ bằng cách sử dụng các kỹ thuật mô hình hồi quy tuyến tính và logistic," trong *Xu hướng gần đây trong Công nghệ Xây dựng và Quản lý*, trong *Ghi chú Bài giảng về Kỹ thuật Xây dựng*, MS Ranadive, BB Das, YA Mehta và R. Gupta, Eds., Springer, Singapore, tập 260, 2023, trang 1205–1225. https://doi.org/10.1007/978-981-19-2145-2_89

- [36] Y. Amemiya, "Công cụ ước tính biến dụng cụ cho các sai số phi tuyến trong biến mô hình," J. Econom., tập 28, số 3, trang 273–289, 1985. [https://doi.org/10.1016/0304-4076\(85\)90001-6](https://doi.org/10.1016/0304-4076(85)90001-6)
- [37] S. Raschka, "STAT 479: Ghi chú bài giảng về học máy," 2018. <http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/> [38] E. Fix và JL Hodges, "Phân tích phân biệt đối xử, phân biệt đối xử phi tham số:

Thuộc tính nhất quán," Báo cáo kỹ thuật 4, Trường Y học Hàng không Không quân Hoa Kỳ, RandolphField, 1951. [39] TM Cover và PE Hart, "Phân loại mẫu hàng xóm gần nhất," IEEE Trans Inf.

Lý thuyết, tập 13, số 1, trang 21–27, 1967. [https://doi.org/10.1109/TIT.1967.1053964\[40\]](https://doi.org/10.1109/TIT.1967.1053964[40]) I. D. Wahyono và cộng sự, "Phát hiện cảm xúc dựa trên nhận xét cột trong tài liệu học trực tuyến bằng trí tuệ nhân tạo," Tạp chí Quốc tế về Công nghệ Di động Tương tác (IJIM), tập 16, số 3, trang 82–91, 2022. [https://doi.org/10.3991/ijim.v16i03.28963\[41\]](https://doi.org/10.3991/ijim.v16i03.28963[41]) F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch và J. B. O. Mitchell, "Tan chảy

Dự đoán điểm sử dụng thuật toán hàng xóm gần nhất K và tối ưu hóa tham số di truyền," J. Chem. Inf. Model, tập 46, số 6, trang 2412–2422, 2006. [https://doi.org/10.1021/ci060149f\[42\]](https://doi.org/10.1021/ci060149f[42]) O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente,

M. Cabanillas-Carbonell và R. Palma, "Phân tích cảm xúc của các tweet bằng cách sử dụng các kỹ thuật học tập không giám sát và thuật toán k-means," Tạp chí Quốc tế về Ứng dụng và Khoa học Máy tính Tiên tiến (IJACSA), tập 13, số 6, 2022. [https://doi.org/10.14569/IJACSA.2022.0130669\[43\]](https://doi.org/10.14569/IJACSA.2022.0130669[43]) B. Kamiński, M. Jakubczyk, và P. Szufel, "Một khuôn khổ để phân tích độ nhạy của deci-

cây sion," Cent. Eur. J. Oper. Res., tập 26, số 1, trang 135–159, 2018. <https://doi.org/10.1007/s10100-017-0479-6>

8 TÁC GIẢ

Orlando Iparraguirre-Villanueva Kỹ sư hệ thống với bằng thạc sĩ Quản lý Công nghệ Thông tin và Tiến sĩ Kỹ thuật Hệ thống của Đại học Quốc gia Federico Villarreal ở Peru. Chứng nhận ITIL®, đặc biệt về Quản lý Kinh doanh Liên tục (SBCM) và Chứng nhận SCRUM. Ông là diễn giả / tham luận viên trong nước và quốc tế đã trình bày tại Panama, Colombia, Ecuador, Venezuela, Mexico. Orlando có nhiều kinh nghiệm giảng dạy ở cả cấp đại học và sau đại học trong nhiều lĩnh vực khác nhau. Ông đã từng là cố vấn luận án và thành viên ban giám khảo tại các trường đại học khác nhau. Với nền tảng về các vị trí quản lý trong Công nghệ Thông tin, ông cũng là giáo sư nghiên cứu với các ấn phẩm trên các tạp chí được lập chỉ mục Scopus và WoS có tác động cao (Q1, Q2, Q3 và Q4). Các lĩnh vực quan tâm của ông bao gồm phần mềm mã nguồn mở, IoT, Thực tế tăng cường, Học máy, AI, CNN, khai thác văn bản, môi trường ảo, phương pháp nghiên cứu khoa học và luận án. Bạn có thể liên hệ với anh ấy tại (E-mail: oiparraguirre@ieec.org). Cleoge Paulino-Morenola một Kỹ sư Hệ thống chuyên nghiệp, có bằng Thạc sĩ về Máy tính Giáo dục và Công nghệ Thông tin. Ông chuyển về Quản lý Công cộng và được chứng nhận về Windows Server, Linux và WindowsServer. Anh ấy có trình độ tiếng Anh trung cấp. Cleoge có kinh nghiệm trong nghiên cứu khoa học và đã làm việc với tư cách là Điều phối viên Hỗ trợ Đổi mới và Công nghệ. Ông đã xuất bản các bài báo khoa học trên các tạp chí được bình duyệt, thể hiện sự cống hiến của mình cho nghiên cứu khoa học. Các kỹ năng của ông bao gồm khả năng nghiên cứu, đạo đức làm việc nhạy bén, giao tiếp hiệu quả, lãnh đạo nhóm và giải quyết vấn đề.

Andrés Epifanía-Huertacó bằng thạc sĩ về kỹ thuật hệ thống với chuyên ngành công nghệ thông tin và truyền thông. Ông hiện đang liên kết với Khoa Kỹ thuật tại Đại học Tecnológico del Perú (UTP). Ông giảng dạy về các chủ đề liên quan đến đổi mới công nghệ, công nghệ áp dụng cho giáo dục và quản lý công nghệ kinh doanh (E-mail: C25267@utp.edu.pe). Carmen Torres-Ceclénis là Kỹ sư Máy tính và Hệ thống với bằng Thạc sĩ về Máy tính và Hệ thống, chuyên về Công nghệ Thông tin và Truyền thông. Cô cũng có bằng Tiến sĩ Giáo dục, bằng tốt nghiệp giảng dạy đại học và được chứng nhận bởi EXIN về ITIL. Carmen được biết đến với ý thức mạnh mẽ về khả năng chịu trách nhiệm, sáng tạo, khả năng thích ứng với sự thay đổi, bản tính hòa đồng và cách tiếp cận chủ động. Cô có kiến thức sâu rộng trong lĩnh vực CNTT và cập nhật các công nghệ thông tin mới nhất.

