DỰ ĐOÁN TRẦM CẢM Ở SINH VIÊN BẰNG MÔ HÌNH HỌC MÁY TRÊN DỮ LIỆU ĐA ĐẶC TRƯNG

Trần Thị Thanh Nhàn
Khoa CNTT, Trường Đại học Đại Nam
Email: nhanttt@dainam.edu.vn

Tóm tắt: *Nghiên cứu ứng dụng các mô hình học máy để dự đoán nguy cơ trầm cảm ở sinh viên dựa trên các yếu tố như áp lực học tập, thói quen sinh hoạt và tiền sử bệnh lý tâm thần trong gia đình. Bộ dữ liệu được sử dụng là Student Depression Dataset từ Kaggle với 27.901 bản ghi, bao gồm nhiều đặc trưng về nhân khẩu học, hành vi và tâm lý. Quá trình nghiên cứu tập trung vào tiền xử lý dữ liệu, mã hóa biến phân loại và xây dựng các mô hình phân loại gồm Decision Tree, Extra Trees, Random Forest và LightGBM. Kết quả cho thấy tất cả mô hình đều đạt độ chính xác trên 80%, trong đó LightGBM đạt hiệu suất cao nhất với Accuracy 84,93%. Kết quả này khẳng định tiềm năng của học máy trong hỗ trợ sàng lọc sớm và giảm thiểu tác động tiêu cực của trầm cảm đối với sinh viên đại học.*

Từ khóa: *Phân loại, dự đoán, trầm cảm, học máy, sinh viên.*

PREDICTING DEPRESSION IN STUDENTS USING MACHINE LEARNING ON MULTI-FEATURE DATA

Abstract: *This study applies machine learning models to predict the risk of depression in students based on factors such as academic pressure, lifestyle habits and family history of mental illness. The dataset used is Student Depression Dataset from Kaggle with 27,901 records, including many demographic, behavioral and psychological features. The research process focuses on data preprocessing, encoding categorical variables and building classification models including Decision Tree, Extra Trees, Random Forest and LightGBM. The results show that all models achieve accuracy above 80%, in which LightGBM achieves the highest performance with Accuracy 84.93%. This result affirms the potential of machine learning in supporting early screening and minimizing the negative impact of depression on university students.*

Keyword: *Classification, prediction, depression, machine learning, students*

Nhận bài: 17/10/2025 *Phản biện: 20/11/2025* *Duyệt đăng: 25/11/2025*

I. ĐẶT VẤN ĐỀ

Trầm cảm đang trở thành một vấn đề sức khỏe tâm thần đáng lo ngại trong cộng đồng sinh viên đại học, khi các yếu tố như áp lực học tập, thay đổi môi trường sống, khó khăn tài chính và căng thẳng xã hội tác động mạnh đến sức khỏe tinh thần của họ. Nhiều nghiên cứu chỉ ra rằng trầm cảm ảnh hưởng tiêu cực đến kết quả học tập, khả năng thích nghi và chất lượng cuộc sống của sinh viên, đồng thời làm gia tăng nguy cơ tự hại. Việc phát hiện sớm trầm cảm vì vậy có ý nghĩa quan trọng trong việc hỗ trợ và can thiệp kịp thời.

Trong các nghiên cứu gần đây, các mô hình học máy được sử dụng rộng rãi nhằm dự đoán nguy cơ trầm cảm ở sinh viên. Iparraguirre-Villanueva và cộng sự (2024), sử dụng dữ liệu của 787 sinh viên để xây dựng mô hình Logistic Regression, KNN và Decision Tree trong phân loại trầm cảm. Tuy nhiên, quy mô dữ liệu nhỏ và bộ đặc trưng chưa phong phú khiến mô hình chưa đạt được độ chính xác cao và hạn chế về khả năng tổng quát hóa. Simarmata và Prasetyaningrum (2025) khai thác 2.028 phản hồi PHQ-9 và cải thiện hiệu năng SVM thông qua tối ưu siêu tham số, nhưng nghiên cứu vẫn còn phụ thuộc vào dữ liệu tự báo cáo và thiếu quy trình kiểm chứng chéo. Bên cạnh đó, nghiên cứu của Sonnadara et al. (2023) tại Sri Lanka áp dụng 13 mô hình khác nhau và xác định Gradient Boosting là hiệu quả nhất, nhưng cỡ mẫu 363 sinh viên hạn chế khả năng khái quát hóa.

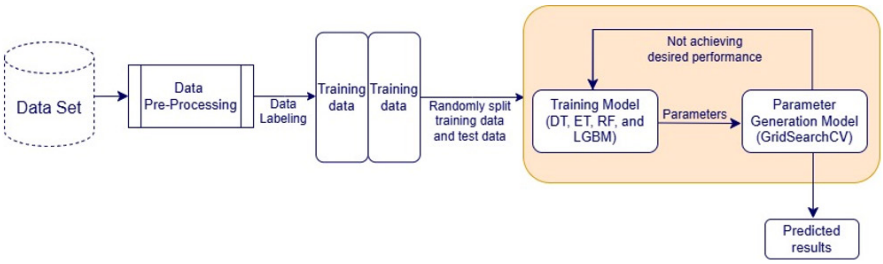
Các nghiên cứu trước nhìn chung còn hạn chế về cỡ mẫu, ít đặc trưng hành vi – tâm lý và chưa đánh giá toàn diện các mô hình mạnh như Extra Trees, Random Forest hay LightGBM trên bộ dữ liệu lớn và đa dạng. Xuất phát từ thực tiễn đó, nghiên cứu này sử dụng Student Depression Dataset gồm 27.901 bản ghi với 18 đặc trưng về nhân khẩu học, hành vi, tâm lý và học tập, giúp giảm nguy cơ overfitting và tăng độ tin cậy. 27.901 bản ghi với 18 đặc trưng về nhân khẩu học, hành vi, tâm lý và học tập, giúp giảm nguy cơ overfitting và tăng độ tin cậy.

II. NỘI DUNG NGHIÊN CỨU

2.1. Phương pháp nghiên cứu

Nghiên cứu triển khai và đánh giá hiệu năng của bốn thuật toán học máy gồm Decision Tree, Extra Trees, Random Forest và LightGBM nhằm xây dựng các mô hình phân loại phục vụ dự đoán

nguy cơ trầm cảm ở sinh viên. Các mô hình được đánh giá dựa trên các chỉ số phổ biến trong phân tích phân loại, bao gồm Accuracy, Recall và F1-score, nhằm đo lường mức độ chính xác, khả năng nhận diện đúng các trường hợp trầm cảm và hiệu suất tổng hợp của từng mô hình. Bên cạnh việc so sánh hiệu năng, nghiên cứu cũng hướng đến xác định những đặc trưng có ảnh hưởng mạnh nhất đến nguy cơ trầm cảm, từ đó góp phần xây dựng một hệ thống dự báo đáng tin cậy và có khả năng ứng dụng trong công tác sàng lọc sớm sức khỏe tâm thần cho sinh viên đại học.



Hình 1. Mô tả các thuật toán được thực hiện trên dữ liệu

2.1.1. Mô tả tập dữ liệu

Tập dữ liệu Student Depression Dataset ban đầu bao gồm 27.901 bản ghi với 18 đặc trưng phản ánh các thông tin nhân khẩu học, hành vi và yếu tố tâm lý của sinh viên. Các đặc trưng này mô tả toàn diện các yếu tố liên quan đến nguy cơ trầm cảm và được trình bày chi tiết trong Bảng 1.

Bảng 1. Bảng các đặc trưng của Dataset

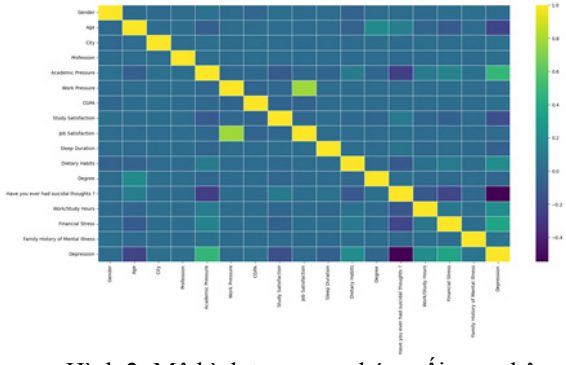
Đặc trưng	Kiểu dữ liệu	Diễn giải
ID	int64	Mã số sinh viên
Gender	object	Giới tính
Age	float64	Tuổi
City	object	Thành phố
Profession	object	Nghề nghiệp
Academic Pressure	float64	Áp lực học tập
Work Pressure	float64	Áp lực công việc
CGPA	float64	Thành tích học tập
Study Satisfaction	float64	Mức độ hài lòng với học tập
Job Satisfaction	float64	Mức độ hài lòng với công việc
Sleep Duration	object	Thời lượng ngủ
Dietary Habits	object	Thói quen ăn uống
Degree	object	Bằng cấp
Have you ever had suicidal thoughts?	object	Bạn đã bao giờ có ý định tự tử chưa?
Work/Study Hours	float64	Giờ làm việc/học tập
Financial Stress	float64	Căng thẳng về tài chính
Family History of Mental Illness	object	Tiền sử bệnh tâm thần trong gia đình
Depression	int64	Trầm cảm

2.1.2. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được thực hiện nhằm đảm bảo chất lượng và tính nhất quán của dữ liệu trước khi áp dụng các thuật toán học máy. Trước hết, các bản ghi chứa giá trị thiếu, đặc biệt tại biến Financial Stress, được xác định và loại bỏ để tránh ảnh hưởng đến hiệu năng mô hình. Biến ID được loại bỏ do không mang ý nghĩa dự báo và nhằm giảm độ phức tạp của dữ liệu. Các biến phân loại như Gender, City và Profession được mã hóa sang dạng số bằng phương pháp ánh xạ (mapping) nhằm phù hợp với yêu cầu đầu vào của các thuật toán học máy. Đối với biến Sleep Duration, các giá trị dạng chuỗi (ví dụ: "5–6 hours") được chuyển đổi sang dạng số tương ứng (như 5.5 giờ) nhằm chuẩn hóa dữ liệu và tăng hiệu quả mô hình hóa.

Sau khi hoàn tất các bước tiền xử lý, tập dữ liệu được kiểm tra lại bằng phương thức info() để xác

nhận không còn giá trị thiếu và tất cả các biến đã được chuyển đổi sang kiểu dữ liệu phù hợp. Kết quả thu được một tập dữ liệu gồm 27.898 bản ghi và 17 đặc trưng, đảm bảo điều kiện để triển khai các mô hình học máy như Decision Tree, Extra Trees, Random Forest và LightGBM trong nhiệm vụ dự đoán nguy cơ trầm cảm ở sinh viên.



Hình 2: Mô hình trực quan hóa mối quan hệ tương quan giữa các đặc trưng

Để phân tích mối quan hệ giữa các đặc trưng trong tập dữ liệu sau khi tiền xử lý, nghiên cứu sử dụng biểu đồ heatmap để trực quan hóa ma trận tương quan. Biểu đồ được biểu diễn theo thang màu viridis, cho phép nhận diện mức độ tương quan mạnh hoặc yếu giữa các cặp biến. Việc phân tích này đóng vai trò quan trọng trong quá trình lựa chọn và tối ưu hóa đặc trưng, góp phần nâng cao hiệu quả của các mô hình học máy được triển khai.

2.2. Kết quả nghiên cứu

Trong nghiên cứu này, mục tiêu trọng tâm là xây dựng các mô hình học máy nhằm phân loại sinh viên theo hai nhóm: có nguy cơ trầm cảm và không có nguy cơ trầm cảm. Để xác định mô hình tối ưu, nghiên cứu sử dụng phương pháp GridSearchCV nhằm tối ưu hóa siêu tham số cho từng thuật toán, qua đó lựa chọn được cấu hình phù hợp nhất. Các mô hình được huấn luyện và đánh giá dựa trên các chỉ số phổ biến trong bài toán phân loại gồm Accuracy, Recall và F1-Score.

Trước khi tiến hành huấn luyện, tập dữ liệu được tách thành hai phần: X (tập đặc trưng) và y (nhãn trầm cảm). Sau đó dữ liệu được chia thành hai tập: tập huấn luyện chiếm 80% và tập kiểm tra chiếm 20%, nhằm đảm bảo đánh giá khách quan hiệu năng mô hình. Mô hình Decision Tree được huấn luyện đầu tiên với các siêu tham số tối ưu gồm criterion, max_depth, min_samples_split và min_samples_leaf. Kết quả cho thấy mô hình đạt Accuracy 82.46%, Recall 82.46% và F1-Score 82.51%.

Tiếp theo, mô hình Random Forest được huấn luyện với các siêu tham số tốt nhất thu được từ GridSearchCV. Kết quả cho thấy Random Forest đạt Accuracy 83.64%, Recall 83.64% và F1-Score 83.52%, thể hiện hiệu suất vượt trội so với Decision Tree. Mô hình Extra Trees, với mức độ ngẫu nhiên hóa cao hơn, cũng được triển khai và tối ưu hóa; mô hình này đạt Accuracy 82.26%, Recall 82.26% và F1-Score 82.05%, cho thấy khả năng phân loại tốt nhưng vẫn thấp hơn Random Forest.

Cuối cùng, thuật toán Light Gradient Boosting Machine (LightGBM) - vốn nổi bật trong xử lý dữ liệu lớn và phức tạp, được triển khai với bộ siêu tham số tối ưu. Mô hình đạt hiệu năng cao nhất với Accuracy 84,93%, Recall 84,93% và F1-score 84,89%, vượt trội so với toàn bộ các mô hình còn lại.

Bảng 2: Đánh giá và so sánh mô hình

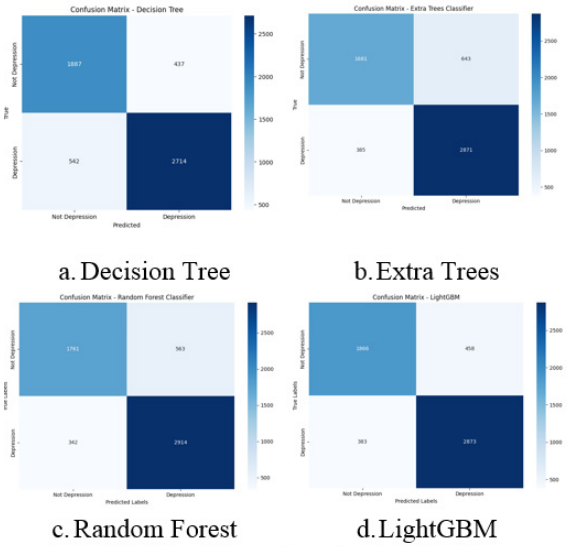
Mô Hình	Accuracy	Recall	F1 Score
Decision Tree	0.8246	0.8246	0.8251
Extra Trees	0.8226	0.8226	0.8205
Random Forest	0.8378	0.8378	0.8364
LightGBM	***0.8493***	***0.8493***	***0.8489***

Kết quả thực nghiệm cho thấy mô hình LightGBM đạt hiệu năng vượt trội, với các chỉ số Accuracy, Recall và F1-score đều vượt ngưỡng 84%, cao hơn đáng kể so với các mô hình còn lại như Decision Tree, Random Forest và Extra Trees. Điều này khẳng định khả năng phân loại mạnh mẽ của LightGBM trong việc nhận diện sinh viên có nguy cơ trầm cảm.

Bên cạnh các chỉ số tổng hợp, việc phân tích ma trận nhầm lẫn đóng vai trò quan trọng trong đánh giá chi tiết hiệu quả dự đoán của từng mô hình. Ma trận nhầm lẫn cho phép quan sát trực tiếp số lượng sinh viên được phân loại đúng và sai ở hai nhóm: có nguy cơ trầm cảm và không có nguy cơ, từ đó cung cấp cái nhìn sâu hơn về hiệu suất thực tế của mô hình đối với từng lớp.

Hình 2 thể hiện ma trận nhầm lẫn của bốn mô hình: Decision Tree, Extra Trees, Random Forest và LightGBM khi áp dụng trên bộ dữ liệu Student Depression Dataset. Việc phân tích các ma trận này cho phép đánh giá toàn diện hơn về hiệu quả phân loại và khả năng ứng dụng thực tiễn của từng thuật toán trong dự đoán trầm cảm ở sinh viên.

III. KẾT LUẬN



a. Decision Tree b. Extra Trees

c. Random Forest d. LightGBM

Hình 2. Ma trận nhầm lẫn trên dataset

III. KẾT LUẬN

Nghiên cứu đã xây dựng và tối ưu hóa bốn mô hình học máy: Decision Tree, Random Forest, Extra Trees và LightGBM nhằm dự đoán nguy cơ trầm cảm ở sinh viên dựa trên bộ dữ liệu quy mô lớn và đa dạng về đặc trưng. Hiệu năng của các mô hình được đánh giá thông qua ba chỉ số quan trọng gồm Accuracy, Recall và F1-score, phản ánh toàn diện khả năng phân loại và mức độ ổn định của từng thuật toán. Kết quả cho thấy tất cả các mô hình đều đạt độ chính xác trên 80%, khẳng định tiềm năng của học máy trong hỗ trợ sàng lọc sớm trầm cảm ở sinh viên. Trong số đó, mô hình LightGBM đạt hiệu suất cao nhất với Accuracy 84,93%, Recall 84,93% và F1-score 84,89%, thể hiện khả năng phân loại vượt trội, đặc biệt trong việc nhận diện các trường hợp trầm cảm tiềm ẩn - nhóm đối tượng cần được phát hiện sớm để triển khai các biện pháp can thiệp kịp thời.

TÀI LIỆU THAM KHẢO

A. Sudershan, S. Rehman, T. Manzoor, B. Shaban, S. Sultan, A. C. Pushap, S. Sudershan, M. Bashir, and S. A. Malik, "Depression, anxiety, and stress among college students: A Kashmir-based epidemiological study," *Frontiers in Psychiatry,* 2025, doi: 10.3389/fpsyt.2025.1633452.

R. Wei, L. Li, Y. Zhang, Y. Liu, R. Wang, S. Li, and Y. Wan, "Predicting depressive symptoms in college students: A nomogram integrating recent and early life events with social support," *Journal of Affective Disorders*, vol. 385, 2025, doi: 10.1016/j.jad.2025.119444.

O. Iparraguirre-Villanueva, C. Paulino-Moreno, A. Epifanía-Huerta, and C. Torres-Ceclén (2024). *Machine Learning Models to Classify and Predict Depression in College Students.* International Journal of Interactive Mobile Technologies. Vol.18 No.14, pp.148-163. DOI: 10.3991/ijim.v18i14.48669

P. W. Simarmata and P. T. Prasetyaningrum (2025). *Development of a Student Depression Prediction Model Based on Machine Learning with Algorithm Performance Evaluation.* Journal-of-Information Systems and Informatics. Vol.7, No.2. DOI:10.51519/journalisi.v7i2.1087

D.S. Sonnadara, S.D. Viswakula, M.D.T Attygalle (2023). *A machine learning approach for predicting the depression status among undergraduates: A case study from Sri Lanka.* DOI:10.21203/rs.3.rs-3555106/v1

S. Vidya, N. Swamy, and H. Lalitha (2024). *Leveraging Machine Learning Algorithms for Depression Prediction in College Students Using PHQ-9 Scores*. IEEE DOI:10.1109/PhDEDITS64207.2024.10838637