Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trang website: https://jstt.vn/index.php/vn

Machine Learning-Based Model for Predicting Students' Academic Performance

Vu Trong Hieu, Le Huyen Trang*, Nguyen Duc Dam, Nguyen Duc Son, Pham Quang Dung
University of Transport Technology, 54 Trieu Khuc Street, Thanh Liet Ward, Hanoi, Vietnam

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.vn.6.7.168-183

***Corresponding author:**
Email address:
lehuyentrang0500@gmail.com

Received: 04/06/2026
Received in Revised Form: 09/09/2026
Accepted: 02/10/2026

Abstract: This study focuses on predicting the second-year cumulative grade point average on a 4-point scale based on first-year academic data, aiming to provide additional quantitative evidence to support academic monitoring and early warning in educational management. The dataset consists of 1,433 students from cohort 75 of the University of Transport Technology, with input variables reflecting academic performance, accumulated credit progress, and several factors related to the learning process. Five machine learning models, including Ridge Regression, Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Adaptive Boosting (AdaBoost, AB), were developed and evaluated. In addition, hyperparameter optimization techniques, including Grid Search, Random Search, and Bayesian Optimization, were employed to identify suitable configurations for each model. The results show that tree-based and ensemble learning models achieved better predictive performance than the linear model. Among them, the GB model optimized using Random Search obtained the best result, with R^2 = 0.954 and a Root Mean Squared Error (RMSE) of 0.102 on the validation set. Shapley Additive Explanations (SHAP) and Partial Dependence Plots (PDP) analyses indicate that variables related to cumulative grade point average, total accumulated credits, and the number of repeated courses had a clear influence on the prediction results. The findings suggest that machine learning can support the early identification of students' academic trends and provide additional information for academic advising, monitoring, and educational management.

Keywords: Machine Learning; Academic Performance Prediction; Grade Point Average (GPA); Hyperparameter Optimization; Shapley Additive Explanations (SHAP); Partial Dependence Plots (PDP).

Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trang website: https://jstt.vn/index.php/vn

Nghiên cứu phát triển mô hình dự báo kết quả học tập của sinh viên bằng các mô hình học máy

Vũ Trọng Hiếu, Lê Huyền Trang*, Nguyễn Đức Đảm, Nguyễn Đức Sơn, Phạm Quang Dũng
Trường Đại học Công nghệ Giao thông vận tải, 54 Triều Khúc, Thanh Liệt, Hà Nội, Việt Nam

Thông tin bài viết
Dạng bài viết:
Bài báo nghiên cứu

DOI:
https://doi.org/10.58845/jstt.utt.2026.vn.6.7.168-183

***Tác giả liên hệ:**
Địa chỉ Email:
lehuyentrang0500@gmail.com

Ngày nộp bài: 04/06/2026
Ngày nộp bài sửa: 09/09/2026
Ngày chấp nhận: 02/10/2026

Tóm tắt: Nghiên cứu này tập trung dự báo điểm trung bình tích lũy hệ 4 của sinh viên năm thứ hai dựa trên dữ liệu học tập năm thứ nhất, nhằm cung cấp thêm cơ sở định lượng hỗ trợ công tác theo dõi và cảnh báo sớm trong quản lý đào tạo. Bộ dữ liệu gồm 1.433 sinh viên khóa 75 của Trường Đại học Công nghệ Giao thông vận tải, với các biến đầu vào phản ánh kết quả học tập, tiến độ tích lũy tín chỉ và một số thông tin liên quan đến quá trình học tập. Năm mô hình học máy gồm hồi quy Ridge, cây quyết định (DT), rừng ngẫu nhiên (RF), tăng cường độ dốc (GB) và tăng cường thích ứng AdaBoost (AB) được xây dựng và đánh giá. Đồng thời, các kỹ thuật tối ưu hóa siêu tham số gồm tìm kiếm lưới, tìm kiếm ngẫu nhiên và tối ưu hóa Bayes được sử dụng nhằm lựa chọn cấu hình phù hợp cho từng mô hình. Kết quả cho thấy các mô hình dựa trên cây và mô hình học tổ hợp cho hiệu quả dự báo tốt hơn mô hình tuyến tính. Trong đó, mô hình GB sau tối ưu hóa bằng tìm kiếm ngẫu nhiên đạt kết quả tốt nhất với R^2 = 0,954 và căn bậc hai của sai số bình phương trung bình (Root Mean Squared Error – RMSE) bằng 0,102 trên tập kiểm chứng. Phân tích giá trị Shapley cộng tính (Shapley Additive Explanations – SHAP) và biểu đồ phụ thuộc riêng phần (Partial Dependence Plots – PDP) cho thấy các biến liên quan đến điểm trung bình tích lũy, tổng số tín chỉ tích lũy và số học phần học lại có ảnh hưởng rõ rệt đến kết quả dự báo. Kết quả nghiên cứu cho thấy học máy có thể hỗ trợ nhận diện sớm xu hướng học tập của sinh viên và cung cấp thêm thông tin phục vụ tư vấn, theo dõi và quản lý đào tạo.

Từ khóa: Học máy; Dự báo kết quả học tập; Điểm trung bình tích lũy (GPA); Tối ưu hóa siêu tham số; Giá trị Shapley cộng tính (SHAP); Biểu đồ phụ thuộc riêng phần (PDP).

1. Giới thiệu

Trong bối cảnh chuyển đổi số trong giáo dục đại học, việc dự báo kết quả học tập của sinh viên ngày càng trở thành một vấn đề quan trọng, góp phần hỗ trợ công tác quản lý, tư vấn học tập và nâng cao chất lượng đào tạo. Điểm trung bình tích lũy (GPA) là một trong những chỉ số phản ánh trực tiếp quá trình học tập, mức độ tích lũy kiến thức và khả năng hoàn thành chương trình đào tạo của sinh viên. Tuy nhiên, tại nhiều cơ sở giáo dục đại học, việc đánh giá và cảnh báo kết quả học tập hiện nay vẫn chủ yếu dựa trên điểm thi kết thúc

học phần, ngưỡng điểm trung bình hoặc các quy tắc hành chính cố định [1].

Cách tiếp cận này có ưu điểm là đơn giản, dễ triển khai và thuận tiện trong quản lý, nhưng chưa hỗ trợ hiệu quả cho việc nhận diện sớm những sinh viên có nguy cơ suy giảm kết quả học tập. Bên cạnh đó, các phương pháp truyền thống thường chưa khai thác đầy đủ nguồn dữ liệu học tập đa chiều đang được tạo ra từ hệ thống quản lý đào tạo, bao gồm kết quả học phần, tiến độ tích lũy tín chỉ, tình trạng học lại, thi lại, điểm rèn luyện và thông tin đầu vào của sinh viên. Trong khi đó, các dữ liệu này có thể chứa nhiều thông tin có giá trị, phản ánh quá trình học tập và sự thay đổi của sinh viên theo thời gian.

Các nghiên cứu gần đây cho thấy đánh giá trong giáo dục đại học đang có xu hướng chuyển từ mô hình đo lường kết quả học tập đơn lẻ sang mô hình thu thập và phân tích chứng cứ học tập liên tục, đa nguồn và gắn với chuẩn đầu ra. Cách tiếp cận này không chỉ phục vụ việc xếp loại kết quả học tập, mà còn hướng đến ba mục tiêu quan trọng: cung cấp phản hồi kịp thời cho người học, hỗ trợ giảng viên điều chỉnh hoạt động dạy học và cung cấp căn cứ định lượng cho nhà trường trong quản lý đào tạo. Trong bối cảnh đó, việc ứng dụng các phương pháp phân tích dữ liệu và học máy có thể giúp chuyển từ cách đánh giá dựa trên quy tắc cố định sang cách tiếp cận dựa trên dữ liệu.

Trên cơ sở đó, học máy được xem là một hướng tiếp cận phù hợp để khai thác dữ liệu học tập phục vụ dự báo và hỗ trợ ra quyết định trong giáo dục. Trên thế giới, nhiều nghiên cứu đã cho thấy các mô hình như Decision Tree, Random Forest, Support Vector Regression, Gradient Boosting và các phương pháp học tổ hợp có khả năng khai thác dữ liệu lịch sử, nhận diện mối quan hệ tiềm ẩn giữa các yếu tố học tập và hỗ trợ dự báo kết quả học tập của người học. Cụ thể, nghiên cứu của Cortez và Silva đã sử dụng các thuật toán hồi quy và phân lớp để dự báo thành tích học tập của học sinh dựa trên dữ liệu học tập và nhân khẩu học, cho thấy các mô hình học máy có khả năng đạt độ chính xác cao trong dự báo giáo dục [1]. Nghiên cứu của Kotsiantis đã tổng hợp và đánh giá hiệu quả của nhiều thuật toán học máy trong môi trường giáo dục, khẳng định tiềm năng của các phương pháp ensemble trong việc cải thiện hiệu quả dự báo [2]. Bên cạnh đó, nghiên cứu của Asif và cộng sự đã xây dựng mô hình dự báo kết quả học tập dựa trên dữ liệu học tập theo từng học kỳ nhằm hỗ trợ cảnh báo sớm đối với sinh viên có nguy cơ học tập yếu [3]. Một số nghiên cứu gần đây cũng tập trung khai thác khả năng giải thích mô hình thông qua các kỹ thuật như SHAP hay PDP nhằm làm rõ mức độ ảnh hưởng của từng biến đầu vào đối với kết quả dự báo [4].

Tuy nhiên, trong bối cảnh Việt Nam, việc ứng dụng học máy vào quản lý đào tạo đại học vẫn còn nhiều dư địa nghiên cứu, đặc biệt là đối với các bài toán sử dụng dữ liệu học vụ thực tế được trích xuất trực tiếp từ hệ thống quản lý đào tạo của nhà trường. Một số nghiên cứu trong nước đã bước đầu áp dụng các thuật toán học máy để phân loại học lực, dự báo nguy cơ học vụ hoặc hỗ trợ cảnh báo sinh viên có kết quả học tập thấp. Các nghiên cứu này cho thấy dữ liệu học tập tích lũy, kết quả học phần, tiến độ tích lũy tín chỉ và các thông tin liên quan đến quá trình học tập có thể cung cấp cơ sở quan trọng cho việc hỗ trợ nhận diện sớm xu hướng học tập của sinh viên. Dù vậy, phần lớn các nghiên cứu trong nước vẫn chủ yếu dừng ở bài toán phân loại theo mức học lực hoặc cảnh báo theo ngưỡng, trong khi các nghiên cứu dự báo trực tiếp điểm trung bình tích lũy hệ 4 trên cơ sở dữ liệu năm học đầu tiên, kết hợp tối ưu hóa siêu tham số và phân tích khả năng diễn giải mô hình, còn tương đối hạn chế.

Xuất phát từ những vấn đề trên, nghiên cứu này tập trung xây dựng và đánh giá các mô hình học máy nhằm dự báo điểm trung bình tích lũy hệ 4 của sinh viên năm thứ hai dựa trên dữ liệu học tập năm thứ nhất tại Trường Đại học Công nghệ Giao thông vận tải. Nghiên cứu sử dụng bộ dữ liệu thực tế gồm 1.433 sinh viên khóa 75 hệ đại học chính quy đã hoàn thành năm học đầu tiên, với các biến phản ánh kết quả học tập, tiến độ tích lũy, kết quả rèn luyện và thông tin đầu vào. Trên cơ sở đó,

một số mô hình hồi quy tiêu biểu được xây dựng, so sánh và tối ưu hóa siêu tham số nhằm đánh giá mức độ phù hợp của từng thuật toán đối với bài toán dự báo.

Bên cạnh việc đánh giá hiệu suất thông qua các chỉ số định lượng, nghiên cứu còn kết hợp các phương pháp diễn giải mô hình để làm rõ vai trò của các biến đầu vào đối với kết quả dự báo. Qua đó, nghiên cứu hướng đến việc cung cấp thêm cơ sở định lượng phục vụ theo dõi xu hướng học tập, hỗ trợ cảnh báo sớm và nâng cao hiệu quả quản lý đào tạo trong giáo dục đại học.

2. Dữ liệu

2.1. Dữ liệu nghiên cứu

Dữ liệu sử dụng trong nghiên cứu được trích xuất từ hệ thống quản lý đào tạo của Trường Đại học Công nghệ Giao thông vận tải, gồm 1.433 sinh viên khóa 75 hệ đại học chính quy. Bài toán ban đầu được mô tả là dự báo điểm trung bình tích lũy hệ 4 của sinh viên năm thứ hai từ dữ liệu học tập năm thứ nhất. Tuy nhiên, các cột GPA đầu vào và mục tiêu trong tệp hiện có đều thuộc bảng điểm tổng hợp toàn khóa; tệp không cung cấp bản ghi độc lập về ngày chốt hoặc học kỳ tương ứng. Do đó, cách gán mốc năm học là giả định của thiết kế ban đầu, chưa phải một thuộc tính đã được xác minh của dữ liệu. Các kết luận về dự báo theo thời gian phải được đọc cùng giới hạn này.

Các biến đầu vào được ký hiệu từ X1 đến X9 (Bảng 1), bao gồm: tổng số tín chỉ đăng ký (tổng số ĐVHT/TC đăng ký) (X1), tổng số tín chỉ tích lũy (tổng số ĐVHT/TC tích lũy) (X2), số học phần học lại (X3), số học phần thi lại (X4), điểm trung bình toàn khóa theo hệ 10 (X5), điểm trung bình toàn khóa theo hệ 4 (X6), xếp loại rèn luyện (X7), thứ tự nguyện vọng trúng tuyển (X8) và điểm trúng tuyển đầu vào (X9). Trong đó, các biến X1, X2, X3, X4, X5, X6, X8, X9 được sử dụng dưới dạng định lượng; điểm trung bình hệ 10 được biểu diễn theo thang điểm 10, còn điểm trung bình hệ 4 được biểu diễn theo thang điểm 4. Riêng biến xếp loại rèn luyện (X7) là biến thứ bậc nên được mã hóa thành giá trị số theo mức xếp loại để phù hợp với yêu cầu đầu vào của các mô hình học máy. Biến mục tiêu của bài toán là điểm trung bình tích lũy hệ 4 của sinh viên năm thứ hai, ký hiệu là Y.

Bảng 1. Phân tích dữ liệu được sử dụng

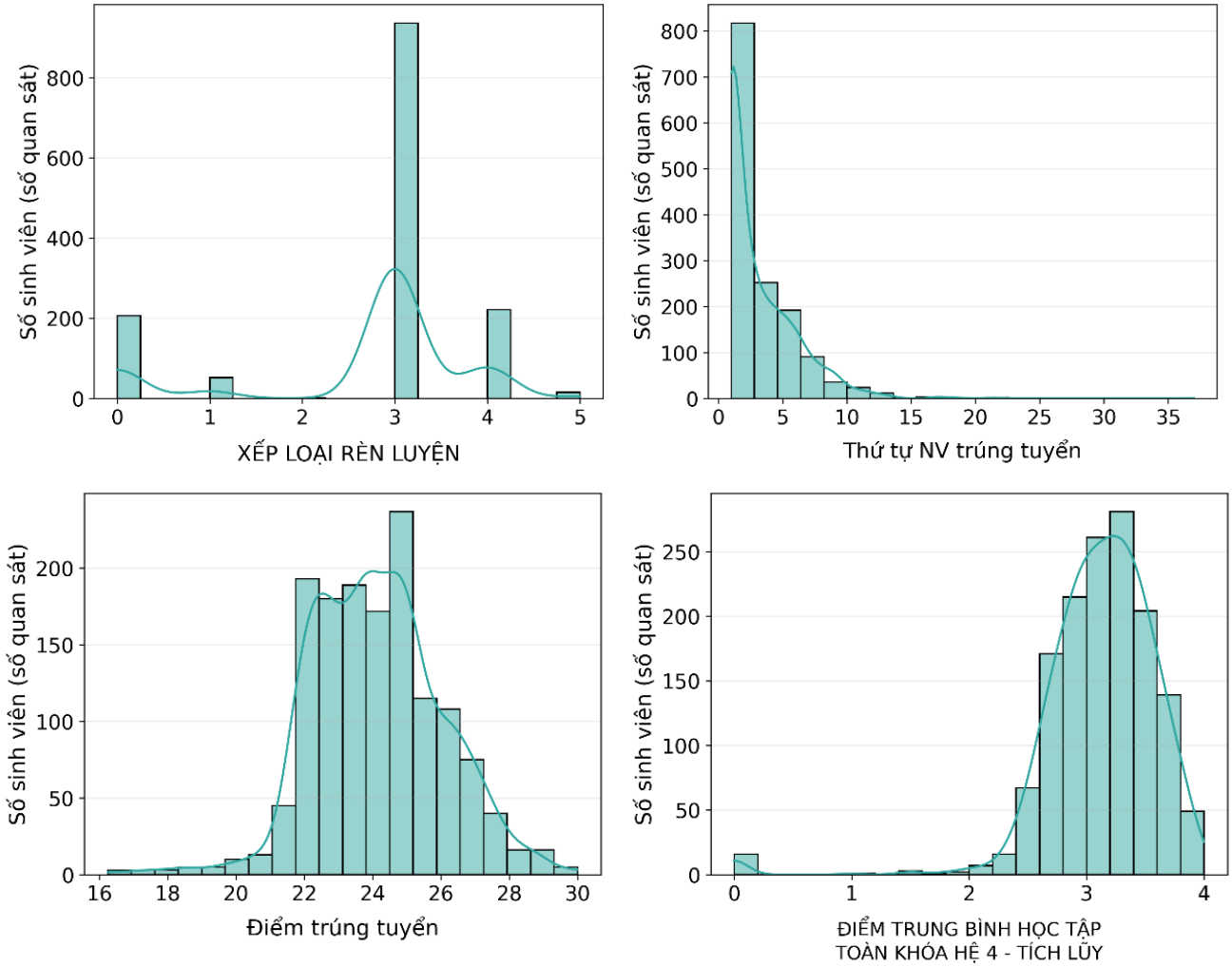
Ký hiệu	Trung bình	Độ lệch chuẩn	Nhỏ nhất	25%	Trung vị (50%)	75%	Lớn nhất
X1	52.86	9.42	0	54.00	55	55	57
X2	50.75	10.54	0	52.00	55	55	57
X3	0.96	2.18	0	0	0	1	21
X4	0	0.07	0	0	0	0	2
X5	7.29	1.53	0	6.98	7.60	8.11	9.38
X6	3.01	0.65	0	2.77	3.13	3.42	4
X7	2.67	1.24	0	3.00	3	3	5
X8	3.16	2.95	1	1	2	5	37
X9	24.13	1.90	16.25	22.70	24.01	25.24	30
Y	3.12	0.51	0	2.88	3.16	3.43	4

2.2. Phân tích dữ liệu nghiên cứu

Đối chiếu bộ dữ liệu dùng trong thí nghiệm cho thấy X5 có 27 giá trị bằng 0 (1.88%), X6 có 11 (0.77%) và Y có 16 (1.12%); các số lượng này có thể chồng lặp giữa các biến, không được cộng thành số sinh viên. Toàn bộ các quan sát bằng 0 được giữ lại, không điền thay và không loại bỏ, nên mẫu phân tích vẫn gồm 1,433 hồ sơ. Lý do là quy trình gốc xử lý chúng như giá trị được ghi nhận trong bảng điểm. Cả 16 hồ sơ có Y = 0 đều có X2 = 0, nhưng thông tin này không đủ để kết luận sinh viên nghỉ học hoặc xác nhận nguyên nhân của điểm 0. Bộ dữ liệu đã làm sạch không còn ô trống, song điều đó không chứng minh mọi giá trị 0 là điểm hợp lệ. Bảng ghép trước làm sạch có 13 hồ sơ thiếu thứ tự nguyện vọng và điểm trúng tuyển

(26 ô thiếu); các giá trị này được điền bằng trung vị từng cột trên toàn bộ mẫu trước khi chia tập. Những trường hợp thiếu này được phân biệt với các giá trị GPA bằng 0. Cách điền thiếu trước chia tập và mã hóa X7 theo nhãn được giữ đúng như thí nghiệm gốc và được thảo luận như hạn chế phương pháp, thay vì coi là tiền xử lý đã hoàn toàn tránh sử dụng thông tin của tập kiểm chứng.

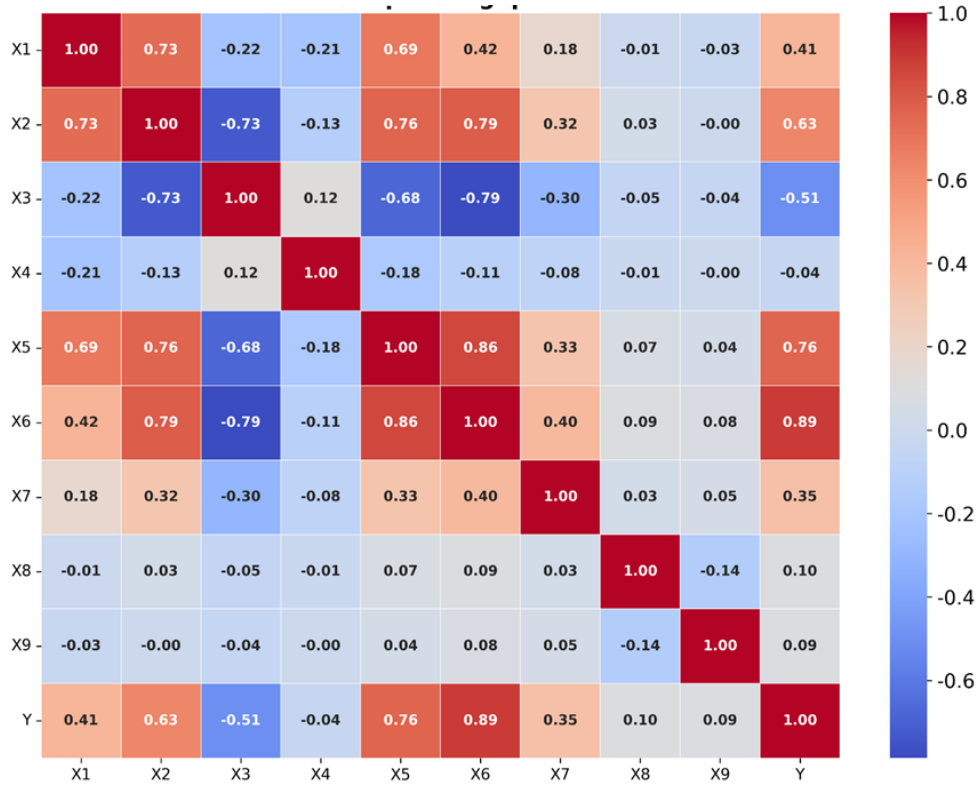
Phân bố của các biến sử dụng trong mô hình dự báo được thể hiện trong Hình 1. Nhìn chung, các biến có sự khác biệt về thang đo, mức độ phân tán và dạng phân bố, cho thấy dữ liệu có tính không đồng nhất nhất định. Trong đó, X1 và X2 có xu hướng tập trung ở vùng giá trị cao, phản ánh sự tương đồng về khối lượng học tập và mức độ tích lũy của phần lớn sinh viên trong bộ dữ liệu.



Hình 1. Phân phối dữ liệu của các biến được sử dụng trong mô hình

Ngược lại, X3 và X4 có phân bố lệch phải, với đa số quan sát tập trung ở giá trị thấp và chỉ một số ít trường hợp có giá trị cao hơn, cho thấy sự phân hóa giữa các nhóm sinh viên trong quá trình học tập. Các biến còn lại như X5, X6, X7, X8, X9 và Y thể hiện mức độ phân tán khác nhau, phản ánh sự đa dạng về kết quả học tập, rèn luyện và thông tin đầu vào. Ma trận tương quan giữa các biến trong bộ dữ liệu được thể hiện trong Hình 2. Kết quả cho thấy một số biến có mối liên hệ đáng chú ý với biến mục tiêu Y, trong đó X6 có tương quan cao nhất, với r = 0.89. Điều này cho thấy X6 là một trong những biến cần được xem xét trong quá trình xây dựng mô hình dự báo.

Bên cạnh đó, một số biến khác như X5 và X3 cũng thể hiện mối liên hệ nhất định với Y, nhưng ở mức thấp hơn. Ma trận tương quan cũng cho thấy giữa một số biến đầu vào có thể tồn tại sự liên hệ tương đối cao, đặc biệt trong nhóm biến liên quan đến kết quả học tập.



Hình 2. Ma trận tương quan của các biến được sử dụng trong mô hình

3. Phương pháp nghiên cứu

3.1. Mô hình học máy

3.1.1. Mô hình Cây quyết định (DT)

DT là mô hình hồi quy phi tuyến dựa trên việc phân chia không gian dữ liệu thành các vùng con thông qua một chuỗi điều kiện rẽ nhánh [5]. Tại mỗi nút, thuật toán lựa chọn biến đầu vào và ngưỡng phân tách nhằm giảm sai số sau khi chia. Mô hình có khả năng biểu diễn các quan hệ phi tuyến mà không cần giả định trước dạng hàm giữa đầu vào và đầu ra.

Độ phức tạp của cây được kiểm soát thông qua các tham số như độ sâu tối đa, số mẫu tối thiểu tại nút lá và số mẫu tối thiểu để tiếp tục phân chia. Sai số tại mỗi nút được đánh giá bằng MSE theo công thức (1):

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2 \quad (1)$$

Giá trị dự báo tại nút lá được xác định bằng giá trị trung bình của các mẫu thuộc nút đó theo công thức (2):

$$\hat{y}=\frac{1}{N_{leaf}}\sum_{i=1}^{N_{leaf}} y_i \quad (2)$$

trong đó y_i là giá trị thực tế của mẫu thứ i, $\bar{y}$ là giá trị trung bình của biến mục tiêu tại nút đang xét, N là số lượng mẫu tại nút và N_l là số lượng mẫu tại nút lá.

3.1.2. Mô hình Rừng ngẫu nhiên (RF)

RF là mô hình học tổ hợp được xây dựng từ nhiều cây quyết định hồi quy [6]. Mỗi cây được huấn luyện trên một tập mẫu bootstrap được lấy ngẫu nhiên từ tập huấn luyện ban đầu, đồng thời tại mỗi nút phân tách chỉ một tập con ngẫu nhiên của các biến đầu vào được xem xét. Cách xây dựng này tạo ra sự đa dạng giữa các cây thành phần, làm giảm sự phụ thuộc vào một cấu trúc cây đơn lẻ và giúp mô hình ổn định hơn so với DT. Dự báo cuối cùng được xác định bằng cách lấy trung bình kết quả dự báo của tất cả các cây trong rừng. Các tham số như số lượng cây, độ sâu tối đa, số biến được xét tại mỗi lần phân tách và số mẫu tối thiểu tại nút lá ảnh hưởng trực tiếp đến hiệu quả dự báo và chi phí tính toán. Khi số lượng cây tăng, mô hình thường ổn định hơn, nhưng khả năng diễn giải trực tiếp giảm và thời gian huấn luyện có thể tăng lên. Dự báo cuối cùng của RF được xác định như sau:

$$\hat{y}=\frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (3)$$

trong đó T là số lượng cây trong rừng và $f_t(x)$ là giá trị dự báo của cây thứ t.

3.1.3. Mô hình Tăng cường độ dốc (GB)

GB là mô hình học tổ hợp theo hướng tăng cường, trong đó các cây hồi quy được xây dựng tuần tự để cải thiện phần sai số còn lại của mô hình trước đó [7]. Ở mỗi bước lặp, một cây mới được thêm vào mô hình tổng thể nhằm mô tả phần dữ liệu mà các cây trước chưa dự báo tốt. Quá trình này tương ứng với việc cập nhật mô hình theo hướng giảm dần hàm mất mát, giúp mô hình từng bước nâng cao chất lượng dự báo. So với RF, GB không xây dựng các cây độc lập mà liên kết chúng theo trình tự học, do đó mô hình có thể biểu diễn tốt các quan hệ phi tuyến và các tương tác phức tạp trong dữ liệu. Các tham số như số lượng cây, tốc độ học, độ sâu cây và tỷ lệ mẫu sử dụng ở mỗi vòng lặp có ảnh hưởng lớn đến kết quả. Tốc độ học nhỏ giúp quá trình cập nhật thận trọng hơn nhưng thường cần nhiều cây hơn, trong khi tốc độ học lớn có thể rút ngắn thời gian huấn luyện nhưng làm tăng nguy cơ quá khớp.

Mô hình tổng thể tại bước lặp thứ m được biểu diễn như sau:

$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x) \quad (4)$$

Phần dư tại bước lặp có thể được xác định như sau:

$$r_{im}=y_i-F_{m-1}(x_i) \quad (5)$$

trong đó $F_{m-1}(x)$ là mô hình ở bước trước, $h_m(x)$ là cây hồi quy mới được bổ sung, γ_m là hệ số điều chỉnh mức đóng góp của cây mới và r_i là phần dư của mẫu thứ i.

3.1.4. Mô hình tăng cường thích ứng (AB)

AB là một phương pháp boosting trong đó các bộ học yếu được kết hợp tuần tự để hình thành mô hình dự báo tổng hợp [8]. Sau mỗi vòng lặp, các mẫu có sai số dự báo lớn được gán mức quan trọng cao hơn, giúp bộ học tiếp theo tập trung nhiều hơn vào những quan sát khó dự báo. Mô hình tổng thể được tạo thành từ tổ hợp có trọng số của các bộ học yếu, trong đó những bộ học có hiệu quả tốt hơn sẽ đóng góp nhiều hơn vào kết quả cuối cùng. Cách tiếp cận này giúp AB từng bước cải thiện chất lượng dự báo từ các mô hình đơn giản. Các tham số như số lượng bộ học yếu, tốc độ học và loại mô hình cơ sở có ảnh hưởng đáng kể đến quá trình cập nhật trọng số và kết quả dự báo. Mặc dù AB có thể cải thiện hiệu quả của các bộ học yếu, mô hình này thường nhạy với nhiễu và các điểm ngoại lai do các mẫu có sai số lớn được nhấn mạnh trong quá trình huấn luyện.

Dự báo tổng hợp của AB có thể được biểu diễn như sau:

$$F(x)=\sum_{m=1}^{M} \alpha_m h_m(x) \quad (6)$$

trong đó $h_m(x)$ là bộ học yếu thứ m, α_m là trọng số thể hiện mức độ đóng góp của bộ học yếu và M là số lượng bộ học yếu.

3.1.5. Mô hình hồi quy Ridge

Ridge là một dạng mở rộng của hồi quy tuyến tính, trong đó thành phần điều chuẩn L_2 được bổ sung vào hàm mục tiêu nhằm kiểm soát độ lớn của các hệ số hồi quy [9]. Khác với Linear Regression chỉ tập trung tối thiểu hóa sai số bình phương, Ridge đồng thời điều chỉnh độ lớn của hệ số thông qua tham số điều chuẩn λ. Khi λ tăng, các hệ số hồi quy có xu hướng bị co nhỏ lại, giúp hạn chế dao động của mô hình trong trường hợp các biến đầu vào có tương quan cao hoặc tồn tại hiện tượng đa cộng tuyến. Ridge vẫn giữ lại toàn bộ biến đầu vào trong mô hình, nhờ đó không làm mất thông tin ban đầu, đồng thời cải thiện tính ổn định của dự báo. Điểm cần lưu ý là mô hình này không thực hiện loại bỏ biến, do đó khả năng rút gọn tập đặc trưng và diễn giải vai trò của các biến ít quan trọng còn hạn chế.

Hàm mục tiêu của Ridge được viết như sau:

$$\min\left[\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right] \quad (7)$$

trong đó y_i là giá trị thực tế của mẫu thứ i, $\hat{y}_i$ là giá trị dự báo, β_j là hệ số hồi quy của biến thứ j, λ là tham số điều chuẩn, N là số lượng mẫu quan sát và p là số lượng biến đầu vào.

3.2. Tối ưu hóa siêu tham số

Để cải thiện hiệu suất của các mô hình dự báo, nghiên cứu tiến hành tối ưu hóa siêu tham số nhằm lựa chọn cấu hình phù hợp cho từng thuật toán. Việc lựa chọn siêu tham số có ảnh hưởng

đáng kể đến độ phức tạp của mô hình, khả năng dự báo trên dữ liệu kiểm tra và mức độ kiểm soát hiện tượng quá khớp. Trong quá trình này, ba kỹ thuật tối ưu hóa có thể được áp dụng bao gồm: Grid Search (GS), Random Search (RS) và Bayesian Optimization (BO). Mỗi kỹ thuật có cách thức tìm kiếm và mức độ hiệu quả khác nhau: Grid Search khảo sát có hệ thống các tổ hợp tham số được xác định trước, Random Search lựa chọn ngẫu nhiên các tổ hợp tham số trong không gian tìm kiếm, trong khi Bayesian Optimization định hướng quá trình tìm kiếm dựa trên kết quả đánh giá từ các lần thử trước đó. Quá trình tối ưu hóa được đánh giá thông qua các chỉ số R^2, RMSE và MAE. Cách tiếp cận này giúp mô hình khai thác tốt hơn cấu trúc dữ liệu, nâng cao độ chính xác dự báo và tạo cơ sở tin cậy hơn cho việc so sánh hiệu quả giữa các mô hình.

3.3. Đánh giá hiệu suất dự báo của mô hình

Để đánh giá mức độ chính xác và hiệu quả dự báo của các mô hình học máy, nghiên cứu sử dụng ba chỉ số gồm hệ số xác định R^2, căn bậc hai sai số bình phương trung bình (RMSE) và sai số tuyệt đối trung bình (MAE). Việc kết hợp các chỉ số này cho phép đánh giá mô hình dưới nhiều góc độ khác nhau, trong đó phản ánh khả năng giải thích sự biến thiên của biến mục tiêu, còn RMSE và MAE thể hiện mức độ sai lệch giữa giá trị dự báo và giá trị thực tế. Cụ thể, R^2 cho biết tỷ lệ phương sai của biến mục tiêu được mô hình giải thích thông qua các biến đầu vào; giá trị R^2 càng cao cho thấy mô hình mô tả tốt hơn mối quan hệ giữa đầu vào và đầu ra. Ngược lại, giá trị R^2 thấp hoặc âm cho thấy khả năng giải thích của mô hình còn hạn chế. RMSE được sử dụng để đo độ lệch trung bình giữa giá trị dự báo và giá trị thực tế, đồng thời nhạy hơn với các sai số lớn do sử dụng bình phương sai số. Trong khi đó, MAE phản ánh sai số tuyệt đối trung bình, ít bị ảnh hưởng bởi các sai lệch cực lớn hơn RMSE và có thể diễn giải trực tiếp hơn. Nhìn chung, mô hình có hiệu quả dự báo tốt khi đạt giá trị R^2 cao, đồng thời có RMSE và MAE thấp, thể hiện khả năng giải thích dữ liệu tốt và sai số dự báo nhỏ.

Các chỉ số này được xác định theo các công thức sau:

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \quad (8)$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \quad (9)$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (10)$$

Trong đó, y_i là giá trị thực tế của mẫu thứ i, $\hat{y}_i$ là giá trị dự đoán của mô hình, $\bar{y}$ là giá trị trung bình của tập dữ liệu thực tế, và N là số lượng mẫu quan sát.

3.4. Phân tích khả năng giải thích của các mô hình học máy

3.4.1. Phân tích đóng góp của biến bằng phương pháp SHAP

SHAP là phương pháp giải thích mô hình dựa trên giá trị Shapley trong lý thuyết trò chơi hợp tác, trong đó mỗi biến đầu vào được xem như một yếu tố đóng góp vào kết quả dự báo cuối cùng [4]. Giá trị SHAP cho biết mức độ một biến làm tăng hoặc làm giảm giá trị dự báo so với giá trị cơ sở của mô hình. Nhờ đó, phương pháp này không chỉ xác định mức độ quan trọng của từng biến mà còn cho phép phân tích hướng tác động của biến đó đối với đầu ra dự báo. Đây là điểm hữu ích khi cần giải thích các mô hình có cấu trúc phức tạp, bởi kết quả dự báo không chỉ được đánh giá bằng độ chính xác mà còn cần được kiểm tra về tính hợp lý trong mối quan hệ giữa biến đầu vào và biến mục tiêu. Tuy nhiên, chi phí tính toán của SHAP có thể tăng khi số lượng mẫu hoặc số lượng biến lớn, do đó trong một số trường hợp cần sử dụng các phương pháp xấp xỉ để rút ngắn thời gian xử lý.

Mô hình giải thích SHAP được biểu diễn như sau:

$$g(z')=\phi_0+\sum_{j=1}^{M}\phi_j z_j' \quad (11)$$

Giá trị Shapley của đặc trưng thứ j được xác định theo công thức:

$$\phi_j=\sum_{S\subseteq F\setminus\{j\}}\frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S\cup\{j\})-f(S)] \quad (12)$$

trong đó ϕ_0 là giá trị dự báo trung bình của mô hình, ϕ_j là giá trị SHAP của đặc trưng thứ j, F là tập hợp tất cả các đặc trưng, S là một tập con

không chứa đặc trưng j, f(S) là giá trị dự báo khi sử dụng tập đặc trưng S, và f(S∪{j}) là giá trị dự báo khi bổ sung đặc trưng j.

3.4.2. Phân tích ảnh hưởng của biến bằng biểu đồ phụ thuộc từng phần (PDP)

PDP là phương pháp được sử dụng để mô tả ảnh hưởng trung bình của một biến đầu vào đến giá trị dự báo của mô hình [7]. Phương pháp này thay đổi giá trị của biến đang xét trong khi các biến còn lại được lấy theo phân bố của dữ liệu, sau đó tính giá trị dự báo trung bình tương ứng. Nhờ đó, PDP giúp nhận diện xu hướng quan hệ giữa một biến đầu vào và biến mục tiêu, chẳng hạn biến đó làm tăng, làm giảm hoặc tạo ra sự thay đổi phi tuyến trong kết quả dự báo. So với SHAP, PDP tập trung nhiều hơn vào việc mô tả dạng quan hệ trung bình giữa biến và đầu ra của mô hình. Tuy nhiên, khi các biến đầu vào có tương quan mạnh, kết quả PDP cần được diễn giải thận trọng vì phương pháp này giả định rằng biến đang xét có thể thay đổi độc lập với các biến còn lại. Hàm phụ thuộc từng phần của biến x_s được xác định như sau:

$$PD(x_s)=\frac{1}{N}\sum_{i=1}^{N} f(x_s,x_c^{(i)}) \quad (13)$$

trong đó, x_s là biến cần phân tích, $x_c^{(i)}$ là các biến còn lại tại mẫu thứ i, và $f\left(x_s,x_c^{(i)}\right)$ là giá trị dự đoán khi thay thế giá trị của biến x_s trong mẫu.

3.5. Quy trình xây dựng và đánh giá mô hình

Nghiên cứu được triển khai bằng Python dưới dạng bài toán hồi quy, với chín biến đầu vào và GPA tích lũy hệ 4 là đầu ra. Quy trình dưới đây mô tả xử lý dữ liệu và so sánh nội bộ của thí nghiệm gốc. Việc gọi các cột là dữ liệu năm thứ nhất và GPA đến năm thứ hai thuộc mục tiêu nghiên cứu ban đầu; nghiên cứu chưa xác minh độc lập tính sẵn có của toàn bộ đầu vào tại thời điểm dự báo.

(1) Thu thập và chuẩn bị dữ liệu: Trích xuất các cột tổng hợp của bảng điểm, ghép thông tin tuyển sinh và thu được 1,433 hồ sơ. Đầu vào gồm tín chỉ, học lại, thi lại, GPA hệ 10 và hệ 4, rèn luyện, nguyện vọng và điểm trúng tuyển; đầu ra là cột GPA hệ 4 có hậu tố "- TÍCH LŨY". Tên cột, cách mã hóa và giới hạn xác định mốc thời gian được nêu ở mục 2.1.

(2) Xử lý và phân chia dữ liệu: Loại các cột định danh và cột GPA hệ 10 có hậu tố "- TÍCH LŨY" khỏi đầu vào. Các giá trị tuyển sinh thiếu được điền bằng trung vị trước khi chia tập; X7 được mã hóa theo nhãn như mục 2.1. Không loại hoặc thay thế các giá trị 0 ở X5, X6, Y. Dữ liệu được chia ngẫu nhiên với random_state = 42 thành 1.003 hồ sơ huấn luyện và 430 hồ sơ kiểm chứng, tương ứng xấp xỉ 70/30. StandardScaler được khớp trên tập huấn luyện để biến đổi đầu vào của hai tập. Việc điền thiếu trước khi chia tập sử dụng thống kê của toàn bộ mẫu là hạn chế tiền xử lý; không mô tả quy trình này như đã được cách ly hoàn toàn khỏi tập kiểm chứng.

(3) Xây dựng và huấn luyện mô hình: So sánh Ridge, DT, RF, GB và AB, sau đó báo cáo kết quả với Grid Search, Random Search và Bayesian Optimization. Bảng 3 trình bày các siêu tham số được chọn trong thí nghiệm gốc. Giới hạn tái lập quy trình tìm kiếm và đánh giá chéo được nêu ở mục 3.2; nghiên cứu không khẳng định tối ưu hóa đã loại bỏ nguy cơ quá khớp.

(4) Đánh giá và giải thích mô hình: Tính R^2, RMSE và MAE trên tập huấn luyện và tập kiểm chứng. Tập kiểm chứng đồng thời được sử dụng để so sánh và nêu bật cấu hình có kết quả cao; vì vậy, các chỉ số này chưa phải đánh giá cuối cùng trên một tập kiểm tra độc lập chưa được dùng cho lựa chọn. GB kết hợp Random Search được phân tích bằng SHAP và PDP như một cấu hình có kết quả cao trên lần chia đang xét. Các phân tích giải thích sự phụ thuộc của đầu ra mô hình vào dữ liệu, chưa xác lập quan hệ nhân quả hoặc hiệu quả cảnh báo sớm.

4. Kết quả và thảo luận

Bộ dữ liệu trong nghiên cứu được chia thành hai phần theo tỷ lệ 70/30, trong đó 70% dữ liệu được sử dụng cho quá trình đào tạo mô hình và 30% dữ liệu còn lại được sử dụng để kiểm chứng hiệu suất dự báo. Việc phân chia này nhằm bảo đảm rằng các mô hình không chỉ được đánh giá dựa trên khả năng học từ dữ liệu đào tạo mà còn được kiểm chứng trên các mẫu chưa tham gia vào

quá trình xây dựng mô hình. Do đó, kết quả trên tập kiểm chứng phản ánh khách quan hơn hiệu suất dự báo và mức độ phù hợp của từng thuật toán đối với bộ dữ liệu nghiên cứu.

Bảng 2. Kết quả của các mô hình học máy

Mô hình	Quá trình đào tạo			Quá trình kiểm chứng		
	R^2	RMSE	MAE	R^2	RMSE	MAE
DT	1	0	0	0.945	0.112	0.036
RF	0.994	0.041	0.013	0.951	0.105	0.029
GB	0.994	0.039	0.021	0.943	0.113	0.035
AB	0.947	0.120	0.084	0.907	0.145	0.087
Ridge	0.907	0.159	0.045	0.832	0.195	0.044

Bảng 3. Các siêu tham số tối ưu của các mô hình học máy

Phương pháp	Mô hình	Tham số tối ưu hóa
Grid Search	AB	{ "learning_rate": 0.01, "loss": "linear", "n_estimators": 50 }
Bayesian Optimization	AB	{ "n_estimators": 87, "learning_rate": 0.259, "loss": "exponential" }
Random Search	AB	{ "learning_rate": 0.113, "loss": "linear", "n_estimators": 162 }
Grid Search	DT	{ "criterion": "squared_error", "max_depth": 7, "min_samples_leaf": 1, "min_samples_split": 2 }
Random Search	DT	{ "criterion": "squared_error", "max_depth": 22, "min_samples_leaf": 4, "min_samples_split": 6 }
Bayesian Optimization	DT	{ "max_depth": 18, "min_samples_split": 5, "min_samples_leaf": 4, "criterion": "squared_error" }
Random Search	GB	{ "learning_rate": 0.243, "max_depth": 7, "n_estimators": 241, "subsample": 0.507 }
Bayesian Optimization	GB	{ "n_estimators": 218, "learning_rate": 0.164, "max_depth": 7, "subsample": 0.676 }
Grid Search	GB	{ "learning_rate": 0.1, "max_depth": 3, "n_estimators": 200, "subsample": 0.8 }
Grid Search	RF	{ "max_depth": 15, "min_samples_leaf": 1, "min_samples_split": 2, "n_estimators": 200 }
Bayesian Optimization	RF	{ "n_estimators": 141, "max_depth": 17, "min_samples_split": 2, "min_samples_leaf": 1 }
Random Search	RF	{ "max_depth": 18, "min_samples_leaf": 3, "min_samples_split": 6, "n_estimators": 135 }
Random Search	Ridge	{ "alpha": 0.553 }
Bayesian Optimization	Ridge	{ "alpha": 0.001 }
Grid Search	Ridge	{ "alpha": 0.001 }

Kết quả phân tích hiệu suất mô hình cho thấy, tất cả các mô hình đều đạt kết quả dự báo tốt với giá trị R^2 trên tập kiểm chứng lớn hơn 0.80. Trên tập đào tạo, các mô hình đều đạt R^2 cao, trong đó RF và GB lần lượt đạt R^2 = 0.994, cho thấy khả năng học dữ liệu rất tốt. Tuy nhiên, DT lại đạt R^2 = 1 điều này cho thấy mô hình này đã bị quá khớp. Trên tập kiểm chứng, RF là mô hình có hiệu suất

cao nhất với R^2 = 0.951 tiếp theo là DT với R^2 = 0.945 và GB với R^2 = 0.943, trong khi Ridge có giá trị thấp nhất với R^2 = 0.832. Kết quả đánh giá ban đầu cho thấy các mô hình đều đạt hiệu suất dự báo tương đối tốt, trong đó RF là mô hình có kết quả nổi bật nhất trên tập kiểm chứng. Tuy nhiên, hiệu suất của các mô hình học máy phụ thuộc đáng kể vào việc lựa chọn siêu tham số. Vì vậy, nghiên cứu tiếp tục thực hiện tối ưu hóa siêu tham số nhằm xác định cấu hình phù hợp hơn cho từng thuật toán, qua đó cải thiện độ chính xác dự báo, giảm sai số và hạn chế nguy cơ khớp quá mức trên tập đào tạo. Quá trình tối ưu hóa được thực hiện bằng ba phương pháp gồm Grid Search, Random Search và Bayesian Optimization. Các phương pháp này được sử dụng để khảo sát không gian tham số và lựa chọn tổ hợp siêu tham số phù hợp cho từng mô hình. Trong đó, Grid Search tìm kiếm có hệ thống trên các tổ hợp tham số đã xác định trước, Random Search lựa chọn ngẫu nhiên các tổ hợp trong không gian tìm kiếm, còn Bayesian Optimization định hướng quá trình tìm kiếm dựa trên kết quả của các lần thử trước đó. Trước khi đánh giá lại hiệu suất mô hình sau tối ưu hóa, các tham số tối ưu thu được từ từng phương pháp được tổng hợp và trình bày trong Bảng 3.

Bảng 4 cho thấy GB kết hợp Random Search đạt R^2 = 0.954 và RMSE = 0.102 trên tập kiểm chứng, thuộc nhóm kết quả cao nhất quan sát được ở lần chia này. DT kết hợp Grid Search cũng đạt R^2 = 0.954 và RMSE = 0.102 ở độ chính xác hiển thị của bảng; MAE của hai cấu hình lần lượt là 0.032 và 0.040. Vì các chênh lệch nhỏ và chưa có kiểm định trên dữ liệu độc lập, nghiên cứu không kết luận GB vượt trội một cách ổn định hoặc có ý nghĩa thống kê. RF duy trì các chỉ số kiểm chứng cao trong các cấu hình khảo sát. Sự khác biệt với Ridge là kết quả của tập dữ liệu và quy trình hiện tại, chưa đủ để khẳng định ưu thế phổ quát của mô hình phi tuyến.

Bảng 4. Kết quả tối ưu hóa mô hình

Phương pháp	Mô hình	Quá trình đào tạo			Quá trình kiểm chứng		
		R^2	RMSE	MAE	R^2	RMSE	MAE
Grid Search	DT	0.980	0.073	0.030	0.954	0.102	0.040
Grid Search	RF	0.993	0.042	0.013	0.952	0.104	0.029
Grid Search	GB	0.998	0.023	0.015	0.947	0.109	0.033
Grid Search	AB	0.930	0.138	0.080	0.927	0.129	0.079
Grid Search	Ridge	0.907	0.159	0.045	0.831	0.196	0.044
Random Search	GB	1	0	0	0.954	0.102	0.032
Random Search	RF	0.983	0.068	0.020	0.949	0.107	0.031
Random Search	DT	0.980	0.074	0.025	0.928	0.128	0.038
Random Search	AB	0.952	0.115	0.073	0.900	0.150	0.078
Random Search	Ridge	0.907	0.159	0.045	0.831	0.196	0.044
Bayesian Optimization	GB	1	0	0	0.953	0.103	0.029
Bayesian Optimization	RF	0.994	0.041	0.013	0.952	0.105	0.029
Bayesian Optimization	DT	0.980	0.074	0.025	0.928	0.128	0.038
Bayesian Optimization	AB	0.954	0.112	0.076	0.904	0.147	0.083
Bayesian Optimization	Ridge	0.907	0.159	0.045	0.831	0.196	0.044

4.1. Mức khớp huấn luyện và nguy cơ quá khớp

Ở Bảng 4, cả GB kết hợp Random Search và GB kết hợp Bayesian Optimization đều hiển thị R^2 huấn luyện = 1, RMSE = 0 và MAE = 0. Đối chiếu kết quả số lưu từ thí nghiệm gốc cho thấy đây là các giá trị sau làm tròn: với Random Search, $R^2 \approx 0.999999922$, RMSE ≈ 0.000145436 và MAE ≈ 0.000114096; với Bayesian Optimization, các giá trị tương ứng là 0.999999707; 0.000282256 và 0.000215224. Các sai số rất nhỏ nhưng không bằng 0 tuyệt đối. Chú thích này làm rõ độ chính xác hiển thị và không làm thay đổi các kết quả đã báo

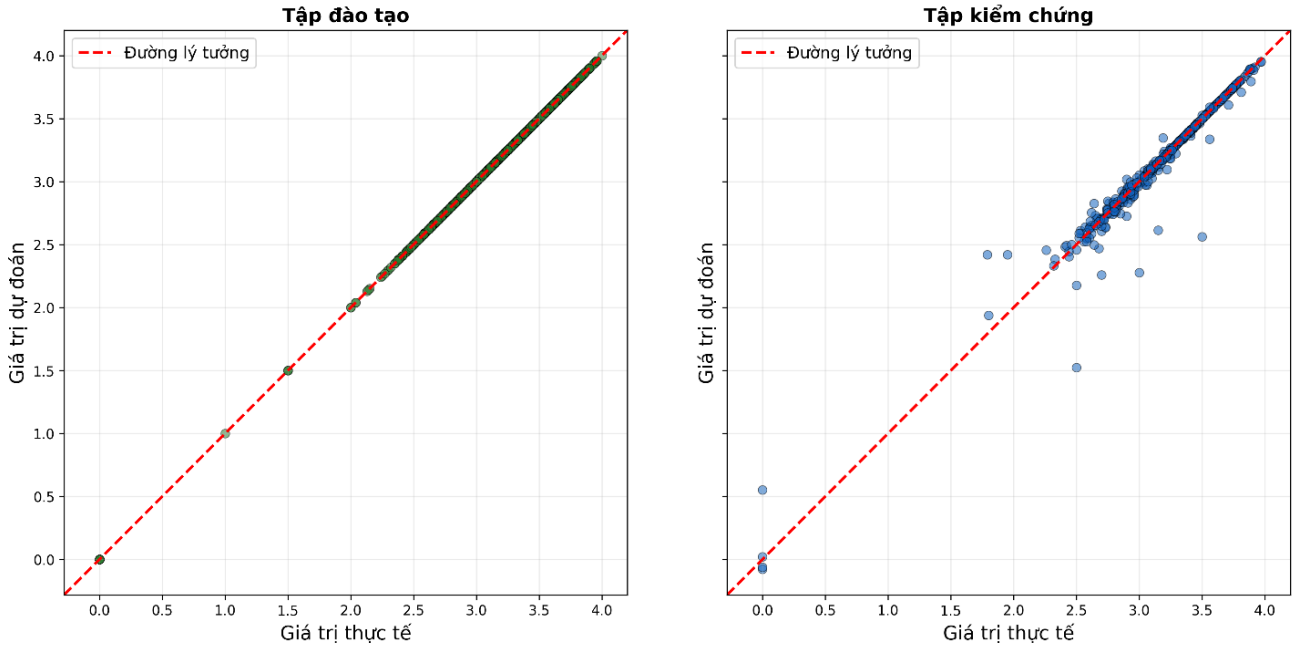
cáo.

Theo Bảng 3, hai cấu hình GB đều có max_depth = 7, với 241 và 218 cây, cùng learning_rate lần lượt khoảng 0.243 và 0.164. Độ phức tạp này cho phép mô hình khớp rất sát dữ liệu huấn luyện; việc giảm sai số huấn luyện đến gần 0 đặt ra nguy cơ học cả các đặc điểm riêng của mẫu. R^2 kiểm chứng lần lượt là 0,954 và 0,953, tạo chênh lệch huấn luyện - kiểm chứng khoảng 0,046 và 0,047 theo số liệu hiển thị. Một chênh lệch nhỏ hoặc thấp hơn một ngưỡng quy ước không đủ chứng minh mô hình không quá khớp.

Kết quả kiểm chứng vẫn có giá trị mô tả độ khớp trên các hồ sơ được giữ lại trong cùng lần chia, nhưng độ tin cậy chỉ giới hạn ở đánh giá nội bộ này. Tập kiểm chứng đã tham gia so sánh cấu hình, các biến GPA có thông tin chồng lặp với mục tiêu và tiền xử lý chưa hoàn toàn cách ly; vì vậy, R^2 cao chưa chứng minh khả năng khái quát hoặc năng lực dự báo phần kết quả học tập mới. Cách thảo luận rủi ro quá khớp được áp dụng nhất quán cho DT ở Bảng 2 và GB sau tối ưu hóa.

4.2. Độ phù hợp trên tập kiểm chứng nội bộ

Mức độ phù hợp của mô hình GB sau khi tối ưu hóa bằng Random Search được đánh giá thông qua mối quan hệ giữa giá trị thực tế và giá trị dự đoán. Kết quả cho thấy các điểm dữ liệu phân bố tương đối gần đường tham chiếu y=x, phản ánh xu hướng dự đoán bám sát giá trị quan sát. Sự phù hợp này thống nhất với kết quả đánh giá định lượng ở Bảng 4, trong đó mô hình GB kết hợp Random Search đạt giá trị R^2 cao và RMSE thấp, cho thấy hiệu quả dự báo tốt trên tập kiểm chứng (Hình 3).



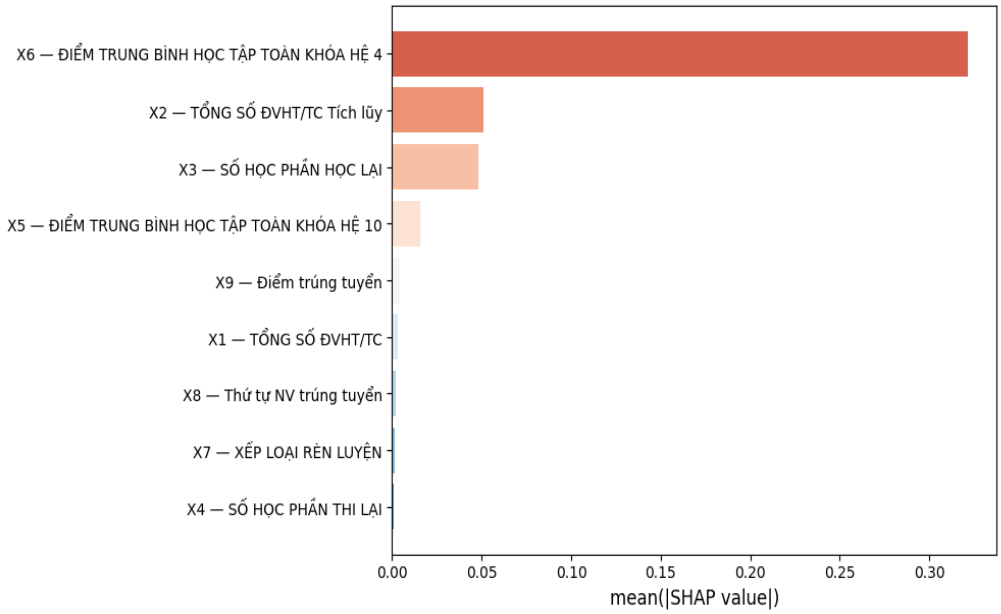
Hình 3. So sánh giá trị thực tế và giá trị dự đoán của mô hình GB tối ưu hóa Random Search

4.3. Diễn giải SHAP, PDP và giới hạn ứng dụng

Kết quả phân tích SHAP cho thấy mức độ đóng góp của các biến đầu vào đối với kết quả dự báo có sự khác biệt rõ rệt, như thể hiện trong Hình 4. Trong đó, biến X6, tương ứng với điểm trung bình học tập toàn khóa theo hệ 4, có giá trị SHAP lớn nhất, cho thấy đây là yếu tố có ảnh hưởng nổi bật nhất đến kết quả dự báo của mô hình. Tiếp theo là biến X2, tương ứng với tổng số đơn vị học trình/tín chỉ tích lũy, và biến X3, tương ứng với số học phần học lại; hai biến này cũng có đóng góp đáng kể nhưng thấp hơn so với X6. Trong khi đó, các biến X1, X4, X7, X8 và X9 có giá trị SHAP nhỏ hơn, phản ánh mức độ ảnh hưởng hạn chế hơn đến kết quả dự báo. Nhìn chung, kết quả SHAP cho thấy mô hình chủ yếu dựa vào các biến phản ánh trực tiếp kết quả và quá trình tích lũy học tập của sinh viên để hình thành dự báo, qua đó làm rõ vai trò tương đối của từng biến đầu vào trong mô hình.

Kết quả phân tích PDP cho thấy xu hướng ảnh hưởng của từng biến đầu vào đến giá trị dự báo của mô hình có sự khác biệt nhất định, như thể hiện trong Hình 5. Trong đó, biến X6, tương ứng với điểm trung bình học tập toàn khóa theo hệ 4, thể hiện mối quan hệ rõ rệt nhất với giá trị dự báo; khi X6 tăng, giá trị dự báo của mô hình cũng có xu hướng tăng tương ứng. Kết quả này phù hợp với phân tích SHAP, tiếp tục khẳng định vai trò nổi bật của X6 trong mô hình. Bên cạnh đó, biến X3, tương ứng với số học phần học lại, cũng cho thấy xu hướng tác động đáng kể đến đầu ra dự báo, trong khi X2, tương ứng với tổng số đơn vị học trình/tín chỉ tích lũy, có mức độ ảnh hưởng thấp hơn. Đối với biến X4, tương ứng với số học phần thi lại, đồ thị PDP gần như không hình thành đường xu hướng rõ ràng và thể hiện gần như trống. Nguyên nhân là X4 có mức độ biến thiên thấp trong tập dữ liệu nghiên cứu, với các quan sát tập trung tại một số ít giá trị. Khi số lượng giá trị phân biệt của biến hạn chế, quá trình tính PDP chỉ đánh giá được giá trị dự báo trong một khoảng rất hẹp, khiến đường biểu diễn bị thu hẹp và khó quan sát trên đồ thị. Hiện tượng này đồng thời cho thấy X4 cung cấp ít thông tin phân biệt cho mô hình trong bộ dữ liệu hiện tại, phù hợp với kết quả SHAP ở Hình 4 khi mức đóng góp của X4 tương đối thấp. Vì vậy, đồ thị gần như trống của X4 không phải là lỗi trực quan hóa mà chủ yếu phản ánh đặc điểm phân bố của biến trong dữ liệu nghiên cứu.

Các biến còn lại như X1, X7, X8 và X9 có đường PDP tương đối phẳng hoặc dao động không rõ rệt, cho thấy mức độ ảnh hưởng hạn chế hơn đến giá trị dự báo của mô hình.



Hình 4. Mức độ đóng góp của các biến đầu vào theo giá trị SHAP

So với nghiên cứu của Yağcı [10] trong đó điểm giữa kỳ được dùng để dự báo điểm cuối kỳ, nghiên cứu hiện tại có khác biệt quan trọng về cấu trúc mục tiêu: GPA tích lũy có thể chứa một phần kết quả đã có trong biến đầu vào. Do đó, không thể từ sự tương đồng về vai trò của điểm học tập suy ra mức độ cảnh báo sớm tương đương. Các kết quả SHAP và PDP ở đây chủ yếu giúp mô tả sự phụ thuộc của mô hình vào thông tin trong tập dữ liệu, đồng thời chỉ ra nhu cầu đánh giá riêng giá trị gia tăng so với GPA sẵn có.

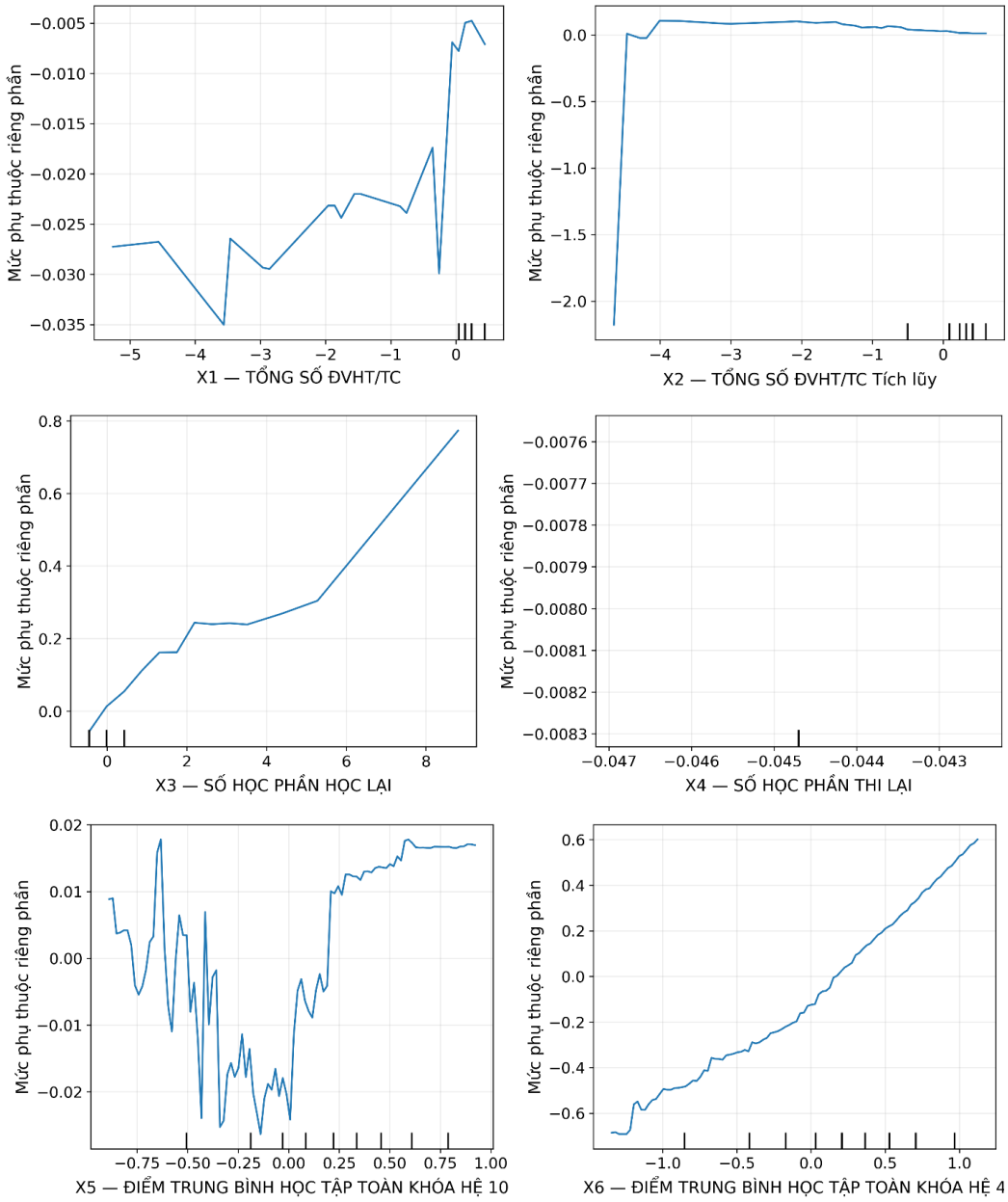
4.4. Hạn chế nghiên cứu và hướng phát triển

Thứ nhất, định nghĩa GPA tích lũy khiến đầu vào và mục tiêu có thể chồng lặp về mặt số học. Theo mục tiêu ban đầu, Y đã chứa phần kết quả của năm thứ nhất được phản ánh trong X6; X5 cũng mang thông tin GPA liên quan. Nghiên cứu chưa tách riêng thành tích học tập phát sinh sau thời điểm đầu vào. Ngoài ra, các cột được lấy từ cùng bảng điểm tổng hợp và chưa được đối chiếu với hai bản chốt điểm có mốc thời gian độc lập. Vì vậy, chưa xác nhận được toàn bộ biến đầu vào có

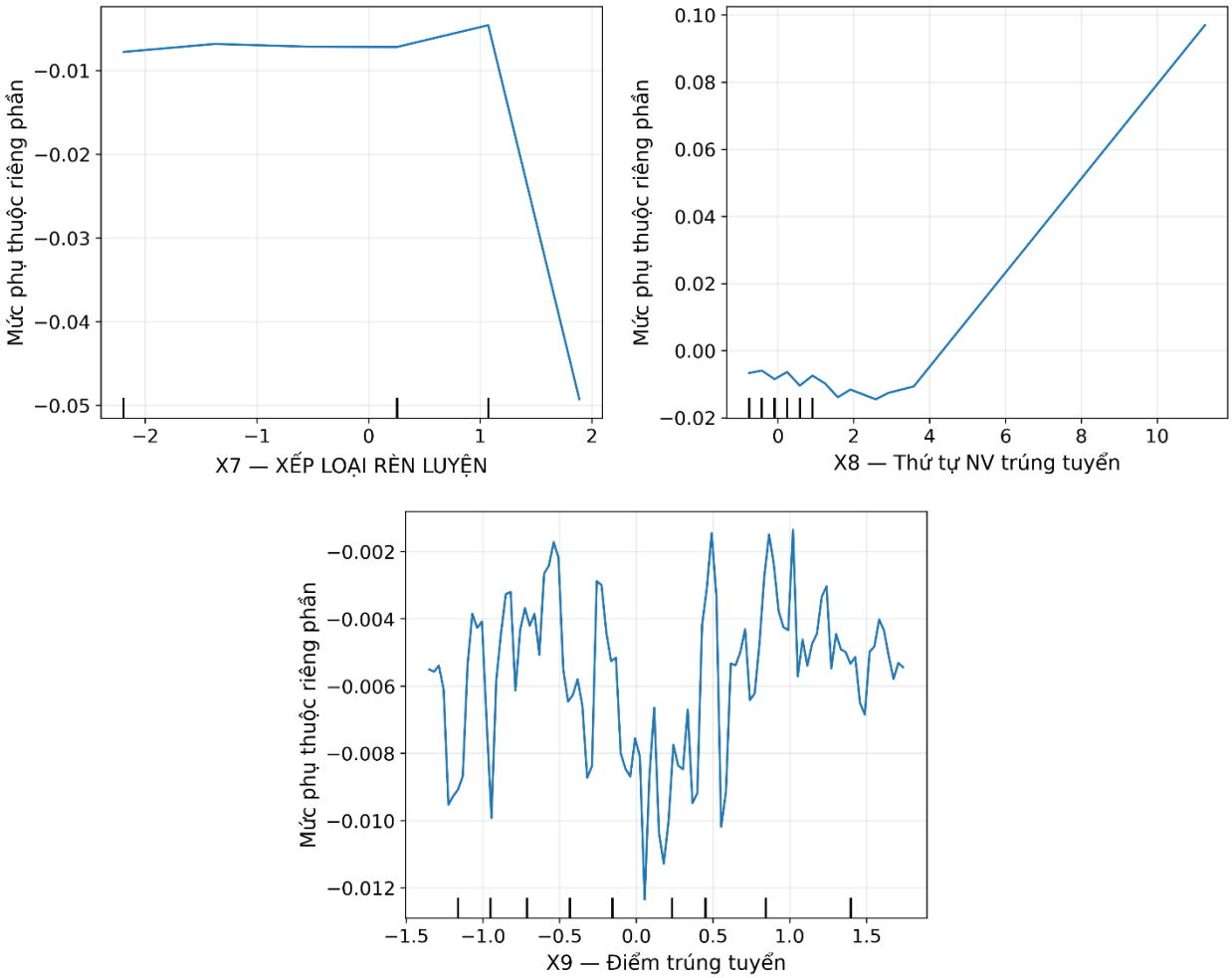
sẵn trước khi hình thành mục tiêu. Các giới hạn này ảnh hưởng trực tiếp đến cách gọi và diễn giải bài toán dự báo tương lai.

Thứ hai, Ridge trong bài là mô hình so sánh tuyến tính đa biến, chưa thay thế được các đối chứng ngây thơ như dự báo bằng trung bình GPA của tập huấn luyện, dùng trực tiếp GPA đã có hoặc hồi quy chỉ với X6. Khi chưa có các đối chứng và phân tích loại bỏ đồng thời X5 và X6, nghiên cứu chưa định lượng giá trị gia tăng của thuật toán phức tạp hoặc lượng thông tin mới ngoài nhóm GPA. Vì vậy, không khẳng định ưu thế của mô hình so với một quy tắc theo dõi học tập đơn giản.

Thứ ba, các kết quả chỉ dựa trên một lần chia ngẫu nhiên nội bộ; cùng tập kiểm chứng được dùng để so sánh cấu hình. Nghiên cứu chưa có tập kiểm tra cuối cùng độc lập chưa từng được dùng trong lựa chọn, chưa kiểm định theo thời gian hoặc trên khóa khác, và chưa đánh giá tại cơ sở đào tạo khác. Các cấu hình có điểm huấn luyện gần hoàn hảo cần được đánh giá thêm về quá khớp. Những chênh lệch nhỏ giữa các mô hình chưa được chứng minh có ý nghĩa thống kê; không suy ra mô hình tốt nhất ổn định từ một lần chia.



Hình 5. Ảnh hưởng của các biến đầu vào đến giá trị dự báo theo biểu đồ PDP của mô hình Random Search + GB.



Hình 5. (tiếp)

Thứ tư, việc điền thiếu dựa trên trung vị toàn bộ mẫu trước chia tập có nguy cơ đưa thông tin phân bố của tập kiểm chứng vào tiền xử lý. Mã số X7 không giữ đúng thứ bậc rèn luyện. Hồ sơ thí nghiệm chưa đủ để tái lập đầy đủ không gian tìm kiếm và xử lý trong từng nếp đánh giá chéo. Các giá trị GPA bằng 0 được giữ nguyên nhưng chưa có xác minh nghiệp vụ từng trường hợp hoặc phân tích độ nhạy. Những điểm này hạn chế tính tái lập và độ tin cậy của ước lượng; mức ảnh hưởng cụ thể chưa được đo trong bài.

Thứ năm, SHAP và PDP giải thích hành vi của mô hình trong bộ dữ liệu, không xác lập quan hệ nhân quả hoặc hiệu quả can thiệp. Nghiên cứu chưa đánh giá độ nhạy, độ đặc hiệu, tỷ lệ cảnh báo sai hay hiệu quả tư vấn đối với một quy trình cảnh báo học vụ. Ngay cả nếu xác nhận được đầu vào thuộc cuối năm thứ nhất, phạm vi ứng dụng cũng chỉ bắt đầu sau thời điểm đó, chưa phải cảnh báo từ đầu năm học hoặc trước khi hình thành GPA.

Hướng phát triển là xây dựng mục tiêu GPA của giai đoạn tương lai không chứa điểm đã dùng làm đầu vào; xác minh ngày chốt và tính sẵn có của từng biến; bổ sung đối chứng trung bình, GPA sẵn có và hồi quy một biến; kiểm tra khi loại nhóm GPA; xử lý thiếu, mã hóa và chuẩn hóa bên trong từng tập huấn luyện của quy trình đánh giá. Việc lựa chọn siêu tham số cần được tách khỏi đánh giá cuối cùng bằng quy trình đánh giá lồng hoặc một tập kiểm tra giữ riêng, rồi kiểm chứng trên khóa và cơ sở khác.

5. Kết luận

Nghiên cứu đã so sánh năm mô hình học máy và ba phương pháp tối ưu hóa siêu tham số trên 1,433 hồ sơ sinh viên khóa 75. Trong lần chia nội bộ 70/30, GB kết hợp Random Search đạt R^2 = 0.954, RMSE = 0.102 và MAE = 0.032 trên tập kiểm chứng; đây là một cấu hình thuộc nhóm kết

quả cao trong thí nghiệm, chưa phải bằng chứng về mô hình tối ưu cho triển khai. SHAP và PDP cho thấy đầu ra mô hình phụ thuộc nhiều vào X6, X2 và X3. Tuy nhiên, chồng lặp thông tin GPA, mốc thời gian chưa được xác minh độc lập, mức khớp huấn luyện gần hoàn hảo, cách tiền xử lý và mã hóa, giá trị 0 chưa được xác minh nghiệp vụ, cùng thiếu đối chứng ngây thơ và kiểm định độc lập giới hạn độ tin cậy và phạm vi ứng dụng. Kết quả được xem là bằng chứng khám phá trên dữ liệu nội bộ; chưa chứng minh năng lực dự báo phần kết quả học tập mới, hiệu quả cảnh báo sớm hay khả năng khái quát. Nghiên cứu tiếp theo cần thiết kế mục tiêu không chồng lặp, xác lập thời điểm sẵn có của từng biến, bổ sung đối chứng đơn giản và kiểm định độc lập trên khóa hoặc cơ sở đào tạo khác.

Tuyên bố về việc sử dụng AI tạo sinh

Trong quá trình chuẩn bị bài báo này, các tác giả đã không sử dụng công cụ AI để tạo ra nội dung cho bài báo.

Tài liệu tham khảo

[1] P. Cortez, A.M.G. Silva. (2008). Using data mining to predict secondary school student performance. In A. Brito, J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12. https://hdl.handle.net/1822/8024 (truy cập ngày 17/09/2026)

[2] S.B. Kotsiantis. (2012). Use of Machine Learning Techniques for Educational Proposes: A Decision Support System for Forecasting Students' Grades. *Artificial Intelligence Review*, 37(4), 331-344. https://doi.org/10.1007/s10462-011-9234-x

[3] R. Asif, A. Merceron, S.A. Ali, N.G. Haider. (2017). Analyzing Undergraduate Students' Performance Using Educational Data Mining. *Computers & Education,* 113, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

[4] S.M. Lundberg, S.I. Lee. (2017). A Unified Approach to Interpreting Model Predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

[5] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. (1984). Classification and Regression Trees. Wadsworth & Brooks/Cole. Monterey, CA, USA

[6] L. Breiman. (2001). Random forests. *Machine Learning,* 45(1). 5-32. https://doi.org/10.1023/A:1010933404324

[7] J.H. Friedman. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics,* 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[8] Y. Freund, R.E. Schapire. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences,* 55(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

[9] A.E. Hoerl, R. W. Kennard. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics,* 12(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

[10] M. Yağcı. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments,* 9(1), 11. https://doi.org/10.1186/s40561-022-00192-z