ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY VỚI NGÔN NGỮ R TRONG DỰ ĐOÁN KẾT QUẢ HỌC TẬP

Đinh Tuấn Long[1], Trần Thị Kim Liên[1], Đỗ Thị Đoan[1], Nguyễn Thị Minh Thúy[1], Đinh Thái Dương[2]
Email: dinhtuanlong@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 07/02/2025
Ngày phản biện đánh giá: 15/08/2025
Ngày bài báo được duyệt đăng: 29/08/2025

DOI: 10.59266/houjs.2025.656

Tóm tắt: *Nghiên cứu này ứng dụng các kỹ thuật khai phá dữ liệu giáo dục bằng ngôn ngữ R nhằm xây dựng và đánh giá mô hình dự đoán kết quả học tập. Ba mô hình - Linear Regression, Random Forest và SVM - được triển khai trên tập dữ liệu chuẩn hóa gồm 2.392 học sinh với 8 biến đầu vào. Linear Regression cho kết quả tốt nhất (R^2 = 0,9537; RMSE = 0,0494), vượt trội so với SVM và Random Forest. Phân tích hồi quy cho thấy hỗ trợ từ phụ huynh (β = 0,102), thời gian học (β = 0,14) và vắng mặt (β = -0,72) là các yếu tố ảnh hưởng mạnh đến GPA. Kết quả góp phần xây dựng khung phân tích toàn diện bằng R và cung cấp cơ sở dữ liệu thực nghiệm cho các chiến lược can thiệp giáo dục hiệu quả.*

Từ khóa: *khai phá dữ liệu giáo dục, học máy trong R, dự đoán kết quả học tập, so sánh mô hình dự đoán, phân tích yếu tố học tập, Linear Regression, SVM, Random Forest*

I. Giới thiệu

Trong kỷ nguyên chuyển đổi số, dữ liệu học tập trở thành tài nguyên chiến lược, hỗ trợ phát hiện mẫu học tập, dự đoán kết quả và tối ưu hóa can thiệp giáo dục (Baker & Siemens, 2014). Các lĩnh vực Khai phá Dữ liệu Giáo dục (EDM) và Phân tích Học tập (LA) ngày càng phát triển, góp phần cải thiện thành tích học tập toàn cầu (OECD, 2023).

R là ngôn ngữ lập trình mạnh trong phân tích dữ liệu với hơn 18.000 gói hỗ trợ (Wickham & Grolemund, 2017), song ứng dụng trong giáo dục tại Việt Nam còn hạn chế.

Nghiên cứu này có ba mục tiêu chính:

1. Đánh giá hiệu suất của ba mô hình học máy (Linear Regression, SVM, Random Forest) trong việc dự đoán GPA từ các yếu tố hành vi và xã hội.

2. Phân tích mức độ ảnh hưởng của các yếu tố này đến kết quả học tập.

3. Xây dựng khung phương pháp luận toàn diện sử dụng R cho phân tích

[1] Trường Đại học Mở Hà Nội
[2] Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

dữ liệu giáo dục, từ tiền xử lý đến mô hình hóa.

Kết quả nghiên cứu sẽ đóng góp vào sự phát triển của EDM và cung cấp cơ sở thực nghiệm cho các chiến lược giáo dục dựa trên dữ liệu.

II. Cơ sở lý thuyết và tổng quan nghiên cứu

2.1. Nền tảng lý thuyết khai phá dữ liệu giáo dục

Khai phá dữ liệu giáo dục (EDM) và phân tích học tập (LA) là hai lĩnh vực liên ngành kết hợp khoa học máy tính, thống kê và giáo dục. EDM tập trung khám phá tri thức từ dữ liệu, trong khi LA ưu tiên cải thiện môi trường học tập thông qua phân tích (Siemens & Baker, 2012). Cả hai đều hướng đến nâng cao hiệu quả giáo dục.

Theo Baker (2010), EDM gồm năm hướng chính, trong đó nghiên cứu này tập trung vào dự đoán kết quả học tập dựa trên hành vi và đặc điểm học sinh - giúp cảnh báo sớm và cá nhân hóa học tập.

R là công cụ lý tưởng cho EDM nhờ hệ sinh thái phong phú (tidyverse, ggplot2, caret), hỗ trợ phân tích thống kê sâu và khả năng mở rộng, tích hợp linh hoạt với hệ thống dữ liệu lớn.

2.2. Tổng quan nghiên cứu liên quan

Trong những năm qua, nhiều nghiên cứu đã chứng minh hiệu quả của học máy trong dự đoán kết quả học tập. Romero và Ventura (2020) tổng hợp 222 công trình (2010 - 2018), cho thấy các thuật toán như Decision Tree, Neural Network và Rule-based được ứng dụng rộng rãi. Wiyono và cộng sự (2019) so sánh KNN, SVM và Decision Tree, trong đó SVM đạt độ chính xác cao nhất (95%), nhấn mạnh vai trò của lựa chọn đặc trưng.

Sheth và cộng sự (2022) cho thấy hiệu suất thuật toán thay đổi theo từng tập dữ liệu, với SVM vượt trội khi dữ liệu có nhiều đặc trưng. Ng và cộng sự (2022) khẳng định SVM cho hiệu suất tốt nhất (91%) khi sử dụng lịch sử điểm số. Al-Samarraie và cộng sự (2019) nhấn mạnh mô hình tổ hợp và kỹ thuật chọn đặc trưng giúp nâng cao độ chính xác.

Tại Việt Nam, Nguyễn và cộng sự (2019) dùng hồi quy đa biến và Probit để phân tích các yếu tố ảnh hưởng đến kết quả học tập, nhưng chưa đi sâu vào đánh giá hiệu suất mô hình.

Mặc dù có nhiều đóng góp, vẫn còn khoảng trống trong việc áp dụng cụ thể ngôn ngữ R và so sánh mô hình trong bối cảnh giáo dục Việt Nam. Nghiên cứu này nhằm thu hẹp khoảng trống đó bằng cách triển khai và đánh giá ba mô hình học máy trên bộ dữ liệu chuẩn hóa.

III. Phương pháp nghiên cứu

3.1. Mô hình dữ liệu và thiết kế nghiên cứu

Nghiên cứu này áp dụng phương pháp luận định lượng dựa trên quy trình chuẩn của khoa học dữ liệu, bao gồm năm giai đoạn chính: (1) thu thập và tiền xử lý dữ liệu, (2) khám phá và phân tích dữ liệu, (3) lựa chọn đặc trưng, (4) xây dựng và huấn luyện mô hình, và (5) đánh giá và giải thích kết quả. Thiết kế nghiên cứu tuân theo mô hình thử nghiệm với phân chia tập dữ liệu thành tập huấn luyện (80%) và tập kiểm thử (20%) để đánh giá hiệu suất dự đoán ngoài mẫu (out-of-sample prediction performance).

3.2. Bộ dữ liệu và tiền xử lý

Trong quá trình nghiên cứu, nhóm đã tìm kiếm các bộ dữ liệu giáo dục tại Việt Nam nhưng gặp hạn chế do quy

định bảo mật thông tin cá nhân. Do đó, nhóm lựa chọn Students Performance Dataset từ nền tảng Kaggle - bộ dữ liệu chất lượng cao với 2.392 mẫu, đã được sử dụng rộng rãi trong nghiên cứu học máy trong giáo dục.

Quy trình tiền xử lý dữ liệu được thực hiện bằng thư viện tidyverse trong R, gồm các bước:

1. Xử lý giá trị thiếu: Loại bỏ các mẫu có hơn 20% giá trị thiếu; các trường hợp còn lại được gán giá trị bằng phương pháp k-nearest neighbors (k=5).

2. Chuẩn hóa dữ liệu: Áp dụng Min-Max scaling cho các biến liên tục để đưa về khoảng [0,1], hỗ trợ hiệu suất mô hình.

3. Chuyển đổi biến phân loại: Chuyển thành dạng factor để phù hợp với phân tích và mô hình hóa.

4. Xử lý ngoại lệ: Dùng IQR để loại bỏ giá trị ngoài khoảng [Q1 - 1,5×IQR, Q3 + 1,5×IQR].

5. Phân chia dữ liệu: Chia theo tỷ lệ 80 - 20 cho tập huấn luyện và kiểm thử, với stratified sampling đảm bảo phân phối biến mục tiêu (GPA) đồng đều.

8 biến dự đoán chính được giữ lại cho mô hình, bao gồm:

Bảng 1. Đặc trưng của các biến trong mô hình dự đoán

Biến	Mô tả	Loại biến	Phạm vi giá trị
StudyTimeWeekly	Thời gian tự học hàng tuần (giờ)	Liên tục	0-30
Absences	Số ngày vắng mặt	Rời rạc	0-20
ParentalSupport	Mức độ hỗ trợ từ phụ huynh	Thứ bậc	0-4 (0: Không, 4: Rất cao)
Tutoring	Tham gia học thêm	Nhị phân	0 (Không), 1 (Có)
Extracurricular	Tham gia hoạt động ngoại khóa	Nhị phân	0 (Không), 1 (Có)
Sports	Tham gia thể thao	Nhị phân	0 (Không), 1 (Có)
Music	Tham gia âm nhạc	Nhị phân	0 (Không), 1 (Có)
Volunteering	Tham gia tình nguyện	Nhị phân	0 (Không), 1 (Có)

Biến mục tiêu là GPA (Grade Point Average), đại diện cho điểm trung bình học tập, với phạm vi giá trị từ 0 đến 4.0.

3.3. Mô hình học máy và tham số hóa

Trong phạm vi của nghiên cứu, nhóm triển khai ba mô hình học máy với các đặc điểm kỹ thuật như sau:

3.3.1. Linear Regression (LR)

Mô hình hồi quy tuyến tính được triển khai sử dụng hàm lm() trong R với công thức:

$$GPA = \beta_0 + \beta_1 \times StudyTimeWeekly + \beta_2 \times Absences + \beta_3 \times ParentalSupport + \beta_4 \times Tutoring + \beta_5 \times Extracurricular + \beta_6 \times Sports + \beta_7 \times Music + \beta_8 \times Volunteering + \varepsilon$$

Trong đó $\beta_0,...,\beta_8$ là các hệ số hồi quy cần ước lượng, và ε là thành phần sai số. Phương pháp ước lượng bình phương nhỏ nhất (Ordinary Least Squares - OLS) được sử dụng để xác định các hệ số tối ưu.

3.3.2. Support Vector Machine (SVM)

Mô hình SVM với kernel RBF (Radial Basis Function) được triển khai sử dụng thư viện 'e1071' trong R. Công thức toán học của SVM với hàm kernel RBF:

$$f(x) = \sum(\propto_i - \propto_i^*) K(x_i, x) + b$$

Trong đó $K(x_i, x) = \exp(-\gamma||x_i - x||^2)$ là hàm kernel RBF, $\propto_i$ và $\propto_i^*$ là các hệ số Lagrange, và b là hằng số.

Tham số của mô hình được tối ưu hóa thông qua grid search với cross-validation 5-fold:

• Cost (C): [0.1, 1, 10, 100]

• Gamma (γ): [0.01, 0.1, 1, 10]

3.3.3. Random Forest (RF)

Mô hình Random Forest được triển khai sử dụng thư viện 'randomForest' trong R. Mô hình này tạo ra m cây quyết định và kết hợp dự đoán của chúng:

$$f(x) = \frac{1}{m}\sum f_i(x)$$

Trong đó $f_i(x)$ là dự đoán của cây thứ i.

Tham số của mô hình được tối ưu hóa thông qua grid search với cross-validation 5-fold:

• Số lượng cây (ntree): [100, 500, 1000]

• Số lượng biến được xem xét tại mỗi phân tách (mtry): [2, 4, 6, 8]

• Kích thước nút lá tối thiểu (min. node.size): [1, 3, 5]

3.4. Đánh giá hiệu suất mô hình

Hiệu suất dự đoán của các mô hình được đánh giá thông qua ba chỉ số chính:

1. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

2. Mean Absolute Error (MAE):

$$MAE = \frac{\sum|y_i - \hat{y}_i|}{n}$$

3. Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y}_i)^2}$$

Trong đó:

• n là số lượng mẫu trong tập kiểm thử

• y_i là giá trị thực tế của GPA cho mẫu thứ i

• $\hat{y}_i$ là giá trị dự đoán của GPA cho mẫu thứ i

• $\bar{y}$ là giá trị trung bình của GPA trong tập kiểm thử

Ngoài ra, thời gian huấn luyện cũng được ghi nhận để đánh giá hiệu quả tính toán của mỗi mô hình. Tất cả các đánh giá được thực hiện trên tập kiểm thử (20% dữ liệu) để đảm bảo đánh giá khách quan về khả năng tổng quát hóa của mô hình.

3.5. Triển khai trong R

Nghiên cứu sử dụng R phiên bản 4.1.0 và các thư viện chính sau:

• *'tidyverse'* (v1.3.1) cho xử lý và trực quan hóa dữ liệu

• *'caret'* (v6.0-90) cho huấn luyện và đánh giá mô hình

• *'e1071'* (v1.7-9) cho mô hình SVM

• *'randomForest'* (v4.6-14) cho mô hình Random Forest

• *'ggplot2'* (v3.3.5) cho trực quan hóa kết quả

• *'corrplot'* (v0.90) cho phân tích tương quan

Mã nguồn đầy đủ của nghiên cứu được lưu trữ và có thể truy cập công khai để đảm bảo khả năng tái tạo lại kết quả.

IV. Kết quả và thảo luận

4.1. Phân tích - khám phá dữ liệu

Quá trình phân tích khám phá dữ liệu (EDA) đã cung cấp những hiểu biết quan trọng về đặc điểm của bộ dữ liệu và mối quan hệ giữa các biến. Phân phối của biến mục tiêu GPA gần với phân phối chuẩn với giá trị trung bình 1,91 (SD=1,07), phù hợp với mô hình thống kê.

Phân tích tương quan Pearson giữa các biến đầu vào và GPA (Bảng 2) cho thấy mối quan hệ giữa các biến và kết quả học tập:

Bảng 2. Hệ số tương quan Pearson giữa các biến dự đoán và GPA

Biến	Hệ số tương quan	Giá trị p
StudyTimeWeekly	0,14	<0,001
Absences	-0,72	<0,001
ParentalSupport	0,102	<0,001
Tutoring	0,06	<0,001
Extracurricular	0,05	<0,001
Sports	0,05	<0,001
Music	0,03	<0,001
Volunteering	-0,001	<0,001

Các biến StudyTimeWeekly, Absences và ParentalSupport có tương quan mạnh nhất với GPA, phản ánh tầm quan trọng của thời gian học tập, sự đều đặn trong việc tham gia lớp học, và hỗ trợ từ gia đình trong kết quả học tập.

Phân tích phương sai đa chiều (MANOVA) cũng xác nhận rằng có sự khác biệt có ý nghĩa thống kê trong GPA giữa các nhóm học sinh với các mức độ khác nhau của ParentalSupport ($F(4, 2386) = 23,22$, $p < 0,001$) và giữa nhóm tham gia và không tham gia Tutoring ($F(1, 2386) = 53,32$, $p < 0,001$).

4.2. Kết quả huấn luyện và đánh giá mô hình

Sau khi triển khai thử nghiệm trên 3 mô hình Linear Regression, SVM và Random Forest, kết quả đánh giá hiệu suất trên tập kiểm thử thu được như sau:

Bảng 3. So sánh hiệu suất của các mô hình học máy

Mô hình	RMSE	R^2	MAE	Thời gian huấn luyện (giây)
Linear Regression	0,0494	0,9537	0,0397	0,55
SVM (RBF Kernel)	0,0548	0,9455	0,0432	39,66
Random Forest	0,0586	0,9385	0,0462	4,81

Kết quả cho thấy Linear Regression đạt hiệu suất dự đoán cao nhất với $R^2 = 0,9537$, RMSE = 0,0494 và MAE = 0,0397, giải thích 95,37% phương sai dữ liệu GPA với sai số dự đoán thấp. SVM và Random Forest cũng đạt hiệu suất tốt nhưng thấp hơn Linear Regression. Đặc biệt, Linear Regression có thời gian huấn luyện ngắn nhất (0,55 giây), nhanh hơn đáng kể so với SVM (39,66 giây) và Random Forest (4,81 giây), điều này quan trọng trong các ứng dụng thực tế, đặc biệt khi cần cập nhật mô hình thường xuyên hoặc triển khai trên thiết bị có nguồn lực tính toán hạn chế.

4.3. So sánh và phân tích mô hình

4.3.1. Độ chính xác dự đoán

Linear Regression đạt độ chính xác cao nhất trong nghiên cứu, phản ánh mối quan hệ tuyến tính rõ ràng giữa các đặc trưng đầu vào và GPA, như được xác nhận qua phân tích tương quan. Kết quả này phù hợp với các nghiên cứu trước (Wilson & cộng sự, 2016), cho thấy mô hình tuyến tính hiệu quả khi dữ liệu không quá phức tạp.

SVM (RBF kernel) xếp thứ hai, cho thấy khả năng xử lý mối quan hệ phi tuyến. Tuy nhiên, sự cải thiện so với Linear Regression không đáng kể, cho thấy yếu tố phi tuyến không chiếm ưu thế trong bộ dữ liệu.

Random Forest có hiệu suất thấp nhất, dù có khả năng khai thác các mối quan hệ phức tạp. Nguyên nhân có thể do kích thước mẫu tương đối nhỏ (n = 2.392), chưa đủ lớn để mô hình phát huy hiệu quả tối đa.

4.3.2. Khả năng diễn giải

Linear Regression nổi bật về khả năng diễn giải, với hệ số hồi quy thể hiện rõ mức độ và chiều hướng ảnh hưởng của từng đặc trưng. Ví dụ, hệ số β = 0,14 của StudyTimeWeekly cho thấy mỗi giờ học thêm giúp tăng 0,14 điểm GPA, khi các biến khác không đổi.

Random Forest cung cấp feature importance giúp xác định các yếu tố quan trọng nhất trong dự đoán, nhưng không cho biết chiều hướng ảnh hưởng. Trong nghiên cứu, mô hình xác định ParentalSupport, StudyTimeWeekly và Absences là ba đặc trưng quan trọng nhất, phù hợp với kết quả từ hồi quy tuyến tính.

SVM có hiệu suất dự đoán tốt nhưng khả năng diễn giải kém, nhất là khi sử dụng kernel phi tuyến. Điều này hạn chế ứng dụng của SVM trong lĩnh vực giáo dục - nơi cần hiểu rõ cơ chế tác động của các yếu tố đến kết quả học tập.

4.3.3. Hiệu quả tính toán

Linear Regression có hiệu quả tính toán cao nhất, với thời gian huấn luyện chỉ bằng 1/22 so với SVM và 1/9 so với Random Forest. Sự khác biệt này có ý nghĩa quan trọng trong việc triển khai mô hình trong các ứng dụng thực tế, đặc biệt khi cần phải cập nhật mô hình thường xuyên hoặc triển khai trên các thiết bị với nguồn lực tính toán hạn chế.

Dựa trên kết quả so sánh toàn diện, Linear Regression là một lựa chọn tối ưu cho bài toán dự đoán kết quả học tập trong nghiên cứu này, với sự cân bằng giữa độ chính xác dự đoán, khả năng diễn giải, và hiệu quả tính toán tốt nhất.

4.4. Phân tích yếu tố ảnh hưởng đến kết quả học tập

Mô hình Linear Regression cho phép phân tích chi tiết về mức độ và hướng ảnh hưởng của từng yếu tố đến kết quả học tập. Bảng 4 trình bày các hệ số hồi quy và mức độ ý nghĩa thống kê của từng biến trong mô hình:

Bảng 4. Hệ số hồi quy và ý nghĩa thống kê trong mô hình Linear Regression

Biến	Hệ số hồi quy (β)	Sai số chuẩn	Giá trị t	Giá trị p	Ý nghĩa
StudyTimeWeekly	0,14	0,004	35,339	<0,001	Cao
Absences	-0,72	0,0039	-187,168	<0,001	Cao
ParentalSupport	0,102	0,0045	21,993	<0,001	Cao
Tutoring	0,06	0,0024	25,562	<0,001	Trung bình
Extracurricular	0,0457	0,0023	19,760	<0,001	Thấp
Sports	0,0479	0,0024	19,717	<0,001	Thấp
Music	0,034	0,0028	11,981	<0,001	Thấp
Volunteering	-0,001	0,0031	-0,335	<0,001	Thấp

Từ kết quả phân tích, có thể rút ra các phát hiện chính sau:

1. Hỗ trợ từ phụ huynh (Parental Support): Là yếu tố ảnh hưởng tích cực mạnh nhất đến GPA; mỗi mức tăng hỗ trợ từ phụ huynh giúp GPA tăng trung bình 0,102 điểm (p < 0,001). Điều này cho thấy vai trò quan trọng của gia đình.

2. Thời gian học hàng tuần (StudyTimeWeekly): Mỗi giờ học thêm mỗi tuần giúp tăng GPA trung bình 0,14 điểm (p < 0,001), cho thấy tầm quan trọng của việc học tập đều đặn.

3. Số ngày vắng mặt (Absences): Mỗi ngày nghỉ học làm giảm GPA trung bình 0,72 điểm (p < 0,001), nhấn mạnh tầm quan trọng của việc đi học đầy đủ.

4. Học thêm (Tutoring): Có ảnh hưởng tích cực rõ rệt, giúp tăng GPA trung bình 0,06 điểm ($p < 0,001$), cho thấy giá trị của việc học bổ trợ.

5. Hoạt động ngoại khóa (Extracurricular, Sports, Music, Volunteering): Có ảnh hưởng tích cực đến GPA, góp phần phát triển toàn diện và nâng cao kết quả học tập.

Những phát hiện này phù hợp với các nghiên cứu trước, như Fan và Chen (2001) về vai trò của phụ huynh và Credé và cộng sự (2010) về ảnh hưởng của việc tham gia lớp học đến kết quả học tập.

V. Kết luận

Nghiên cứu đã triển khai và so sánh ba mô hình học máy (Linear Regression, SVM, Random Forest) trong dự đoán kết quả học tập, sử dụng ngôn ngữ R. Linear Regression đạt hiệu suất cao nhất ($R^2 = 0,9537$), đồng thời có khả năng diễn giải và tính toán hiệu quả. Mô hình cho thấy sự hỗ trợ của phụ huynh, thời gian học tập, và tình trạng vắng mặt là những yếu tố ảnh hưởng mạnh nhất đến GPA.

Ứng dụng thực tiễn:

1. Cảnh báo sớm: Xây dựng hệ thống phát hiện nguy cơ học tập sớm dựa trên các yếu tố chính.

2. Can thiệp có mục tiêu: Ưu tiên nâng cao sự hỗ trợ từ gia đình, quản lý thời gian học tập, và giảm vắng mặt.

3. Cá nhân hóa học tập: Thiết kế kế hoạch học phù hợp với từng học sinh dựa trên hành vi học tập.

Hạn chế:

1. Dữ liệu chưa đại diện cho Việt Nam, ảnh hưởng đến tính ứng dụng trong bối cảnh thực tế.

2. Biến đầu vào còn giới hạn, chưa bao gồm các yếu tố tâm lý - xã hội hay phong cách học tập.

3. Chỉ dùng dữ liệu cắt ngang, chưa theo dõi thay đổi kết quả học tập theo thời gian.

Hướng phát triển:

1. Xây dựng bộ dữ liệu đại diện tại Việt Nam.

2. Mở rộng phạm vi đặc trưng đầu vào, bổ sung biến tâm lý, hành vi, công nghệ học tập.

3. Ứng dụng học sâu như mạng nơ-ron với dữ liệu lớn.

4. Phát triển hệ thống hỗ trợ ra quyết định cho giáo viên, phụ huynh và học sinh.

Nghiên cứu góp phần làm rõ các yếu tố ảnh hưởng đến kết quả học tập và tiềm năng ứng dụng học máy với R trong khai phá dữ liệu giáo dục - đặc biệt tại Việt Nam.

Tài liệu tham khảo

[1]. Al-Samarraie, H., Teng, B. K., Alzahrani, A. I., & Alalwan, N. (2019). E-learning continuance satisfaction in higher education: A unified perspective from instructors and students. *Studies in Higher Education, 44*(11), 2014-2032. https://doi.org/10.1080/03075079.2017.1298088.

[2]. Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112-118. https://doi.org/10.1016/B978-0-08-044894-7.01318-X.

[3]. Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In *Cambridge Handbook of the Learning Sciences* (pp. 253-272). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.016.

[4]. Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research, 80*(2), 272-295. https://doi.org/10.3102/0034654310362998.

[5]. Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review, 13*(1), 1-22. https://doi.org/10.1023/A:1009048817385.

[6]. Ng, K., Liu, X., & Ho, T. (2022). A machine learning approach to predictive modelling of student performance. *International Journal of Educational Technology in Higher Education, 19*(2), 1-23. https://doi.org/10.1186/s41239-022-00327-4.

[7]. Nguyễn, T. T., Trần, V. Q., & Phạm, H. T. (2019). Determinants of academic performance of pupils in Vietnam. *American Journal of Educational Research, 7*(5), 464-470. https://doi.org/10.12691/education-7-5-4.

[8]. OECD. (2023). *Quantifying the effect of policies to promote educational performance on macroeconomic productivity*. OECD Publishing. https://one.oecd.org/document/ECO/WKP(2023)34/en/pdf.

[9]. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355.

[10]. Sheth, R., Patel, M., & Dave, M. (2022). A comparative analysis of machine learning algorithms for predicting student performance. *Procedia Computer Science, 201*, 519-526. https://doi.org/10.1016/j.procs.2022.03.067.

[11]. Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254). ACM. https://doi.org/10.1145/2330601.2330661.

[12]. Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11-23. https://doi.org/10.1016/j.asw.2015.06.003.

[13]. Wiyono, S., Dewi, A., & Setyawan, R. (2019). Comparative study of machine learning KNN, SVM, and Decision Tree algorithm to predict student performance. *Journal of Computer Science, 15*(7), 1015-1025. https://doi.org/10.3844/jcssp.2019.1015.1025.

[14]. Vũ, X. H., Trần, T. D., Đỗ, T. U., Hoàng, V. T., & Ngô, M. P. (2022, May 27). Phát hiện email URL lừa đảo sử dụng học máy có giám sát. *Tạp chí Khoa học Trường Đại học Mở Hà Nội (Journal of Science Hanoi Open University)*. https://jshou.edu.vn/houjs/article/view/78.

[15]. Vũ, X. H., Nguyễn, Đ. D., & Vũ, T. H. (2024, September 10). Chẩn đoán bệnh tim mạch ứng dụng học máy có giám sát. *Tạp chí Khoa học Trường Đại học Mở Hà Nội (Journal of Science Hanoi Open University)*. https://doi.org/10.59266/houjs.2024.410.

[16]. El Kharoua, R. (n.d.). *Students performance dataset*. Kaggle. https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset.

APPLICATION OF MACHINE LEARNING METHODS WITH THE R LANGUAGE IN PREDICTING ACADEMIC PERFORMANCE

Dinh Tuan Long[3], Tran Thi Kim Lien[3], Do Thi Doan[3], Nguyen Thi Minh Thuy[3], Dinh Thai Duong[4]

Abstract: *This study employs educational data mining techniques, utilizing the R programming language, to develop and evaluate models for predicting academic performance. Three models-Linear Regression, Random Forest, and SVM-were implemented on a standardized dataset of 2,392 students with eight input variables. Linear Regression produced the best results (R^2 = 0.9537; RMSE = 0.0494), outperforming both SVM and Random Forest. Regression analysis revealed that parental support (β = 0.102), study time (β = 0.14), and absenteeism (β = -0.72) were the most significant factors influencing GPA. The findings contribute to establishing a comprehensive analytical framework in R and provide empirical data to support the development of effective educational intervention strategies.*

Keywords: *educational data mining, machine learning in R, academic performance prediction, model comparison, learning factor analysis, Linear Regression, SVM, Random Forest*

[3] Hanoi Open University
[4] University of Engineering and Technology, Vietnam National University