HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1059.2023-0006
Natural Sciences 2023, Volume 68, Issue 1, pp. 63-76
This paper is available online at http://stdb.hnue.edu.vn

SỬ DỤNG HỌC MÁY ĐỂ XÁC ĐỊNH NHÂN TỐ ẢNH HƯỞNG VÀ DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN TRONG MÔ HÌNH HỌC TẬP KẾT HỢP

Nguyễn Thị Hồng, Trần Hải Long và Đỗ Trung Kiên
Khoa Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội

Tóm tắt. Nghiên cứu này nhằm mục đích xác định các nhân tố ảnh hưởng đến kết quả học tập và sử dụng chúng để xây dựng mô hình dự đoán kết quả học tập của sinh viên nhằm hỗ trợ nâng cao chất lượng đào tạo. Trong các nghiên cứu trước đây, việc chọn và đánh giá các nhân tố chỉ được thực hiện trên dữ liệu học tập trực tuyến. Trong nghiên cứu này, chúng tôi đề xuất sử dụng tập thuộc tính được lựa chọn từ dữ liệu thực nghiệm thu thập được cả trên lớp học trực tiếp và trên hệ thống học tập trực tuyến tại Trường Đại học Sư phạm Hà Nội. Để xây dựng mô hình dự đoán kết quả học tập, chúng tôi đã thực hiện hai phương pháp lựa chọn biến: một là chọn các biến có mức độ tương quan cao; hai là sử dụng phương pháp phân tích hồi quy tuyến tính Stepwise. Ngoài ra, hai thuật toán học máy được sử dụng để xây dựng mô hình dự đoán là hồi quy tuyến tính và hồi quy véc tơ hỗ trợ. Kết quả thực nghiệm cho thấy mô hình hồi quy véc tơ hỗ trợ với hàm nhân poly được xây dựng dựa trên các biến lựa chọn bằng phương pháp Stepwise có hiệu quả cao nhất.

Từ khóa: dự đoán kết quả học tập, học tập kết hợp, học máy.

1. Mở đầu

Những năm gần đây, mô hình học tập kết hợp đang được triển khai rộng rãi ở nhiều cấp học trên thế giới. Học tập kết hợp là sự kết hợp giữa phương thức học tập trực tuyến (E-learning) và học tập truyền thống (học tập giáp mặt) nhằm nâng cao chất lượng dạy và học. Việc dạy học theo phương thức này giúp khắc phục được những hạn chế và phát huy những ưu điểm của hai phương pháp trên [1]. Thực tiễn cho thấy, nhiều cơ sở giáo dục đã áp dụng mô hình dạy học kết hợp đạt được hiệu quả tích cực [2, 3]. Đặc biệt, trong giai đoạn đại dịch Covid 19 diễn biến phức tạp, dạy học kết hợp là sự lựa chọn phù hợp nhất cho việc tổ chức dạy học an toàn, bảo đảm chương trình và mục tiêu chất lượng giáo dục, đào tạo.

Việc đổi mới phương pháp dạy học và thay đổi mô hình học tập nhằm mục đích nâng cao chất lượng dạy và học. Một trong những thước đo đánh giá chất lượng dạy và học

Ngày nhận bài: 10/1/2023. Ngày sửa bài: 30/1/2023. Ngày nhận đăng: 6/2/2023.
Tác giả liên hệ: Nguyễn Thị Hồng. Địa chỉ e-mail: nguyenhong@hnue.edu.vn

chính là kết quả học tập của người học. Vì vậy, dự đoán kết quả học tập của người học là bài toán quan trọng và được nhiều nhà khoa học quan tâm. Trước đây, bài toán này thường được giải quyết theo hướng tiếp cận của lĩnh vực lí luận và phương pháp dạy học. Tuy nhiên, những năm gần đây, khai phá dữ liệu cũng là một cách tiếp cận hiệu quả để giải quyết các bài toán trong lĩnh vực giáo dục.

Để giải quyết bài toán dự đoán kết quả học tập của người học dựa trên các phương pháp khai phá dữ liệu, các nhà khoa học cần phải nghiên cứu để xác định được các yếu tố ảnh hưởng đến kết quả học tập của người học và lựa chọn thuật toán học máy phù hợp với dữ liệu thu thập được từ quá trình dạy học kết hợp. Vì vậy, trong bài báo này, nhóm tác giả tập trung trả lời hai câu hỏi như sau: *Câu hỏi 1*: Nhân tố nào ảnh hưởng đến kết quả học tập của người học trong mô hình học tập kết hợp? *Câu hỏi 2*: Mô hình học máy nào phù hợp để dự đoán kết quả học tập trong dạy học kết hợp?

Để thống kê các yếu tố ảnh hưởng đến kết quả học tập, Amjed Abu Saa và đồng nghiệp [4] đã thực hiện khảo sát các công trình giải quyết bài toán dự đoán kết quả học tập. Nghiên cứu chỉ ra rằng có 26% công trình sử dụng các nhân tố về thành tích học tập trong quá khứ, 25% các công trình sử dụng các nhân tố về hành vi học tập của sinh viên, 23% công trình sử dụng thông tin nhân khẩu (giới tính, quê quán), 12% số công trình sử dụng thông tin về kinh tế xã hội (khu vực sinh sống, điều kiện kinh tế).

Để dự đoán kết quả học tập của sinh viên trong một khóa học kết hợp, Nick Z. Zacharis [5] đã sử dụng phân tích hồi quy Stepwise để lựa chọn các thuộc tính đưa vào mô hình phân lớp. Mô hình dự đoán đề xuất đạt khoảng 81%. Kiran Fahd và cộng sự [6] cũng thực hiện một nghiên cứu tương tự nhưng sử dụng mức độ tương quan để lựa chọn các thuộc tính dữ liệu đưa vào huấn luyện xây dựng mô hình. Cả hai nghiên cứu này đều sử dụng dữ liệu hành vi được ghi lại từ hệ thống quản lí học tập. Trong các nghiên cứu [7- 9], bên cạnh dữ liệu trực tuyến, các tác giả đã sử dụng thêm các dữ lieu được thu thập từ lớp học trực tiếp. Tuy nhiên, các dữ liệu trên lớp học trực tiếp chỉ được thu thập thông qua việc khảo sát từ người học.

Trong nghiên cứu này, nhóm tác giả thực hiện thu thập dữ liệu từ cả môi trường học tập trực tuyến và học tập trên lớp. Để xác định các biến ảnh hưởng đến kết quả học tập, chúng tôi sử dụng hai phương pháp phân tích Stepwise và đánh giá mức độ tương quan. Hai tập biến được lựa chọn từ hai phương pháp này sẽ được đưa vào để huấn luyện xây dựng các mô hình hồi quy dự đoán kết quả học tập và đánh giá để lựa chọn ra mô hình tốt nhất.

2. Nội dung nghiên cứu

2.1. Mô hình dạy học kết hợp tại Trường Đại học Sư phạm Hà Nội

Ứng dụng công nghệ thông tin trong dạy học đã được triển khai nhiều năm ở trường Đại học Sư phạm Hà Nội. Khi dịch Covid-19 diễn biến phức tạp với hàng loạt các chính sách giãn cách và vệ sinh phòng dịch nghiêm ngặt đòi hỏi các cơ sở giáo dục cần phải áp dụng các hình thức dạy học linh hoạt mà vẫn đảm bảo chất lượng. Chính vì vậy, Trường Đại học Sư phạm đã nhanh chóng triển khai mô hình dạy học kết hợp cho tất cả môn học của trường để thích ứng với tình hình dịch bệnh. Hiện nay, nhận thấy hiệu quả của mô

hình dạy học này nên trường vẫn tiếp tục triển khai dạy học kết hợp cho nhiều môn chung và môn chuyên ngành ở các khoa.

Khi triển khai dạy học kết hợp, trong thời gian học tập trực tiếp, các hoạt động của giảng viên (GV) và sinh viên (SV) sẽ diễn ra tại giảng đường. Ngoài ra, sinh viên sẽ có thời gian tự học trên hệ thống học tập trực tuyến Moodle https://cst.hnue.edu.vn. Các học liệu như bài giảng, câu hỏi trắc nghiệm, bài tập được đưa lên hệ thống để sinh viên có thể học mọi lúc mọi nơi.

Các hoạt động của GV và SV trong môi trường học tập kết hợp sẽ diễn ra theo mô hình lớp học đảo ngược bao gồm:

- *Trước giờ học (trực tuyến):* SV tự học trên hệ thống học tập trực tuyến của trường: xem trước bài giảng, làm bài tập và chuẩn bị các câu hỏi cần giải đáp.

- *Trong giờ học (trực tiếp*): GV sẽ tổng hợp kiến thức trọng tâm và giải đáp các thắc mắc; SV thảo luận và trao đổi theo nhóm.

- *Sau giờ học (trực tuyến):* Để củng cố kiến thức đã được học trên lớp, SV phải hoàn thiện bài tập về nhà và nộp lên hệ thống học tập trực tuyến, trả lời các câu hỏi trắc nghiệm ôn tập kiến thức.

2.2. Đề xuất quy trình xây dựng mô hình dự đoán kết quả học tập của sinh viên trong dạy học kết hợp

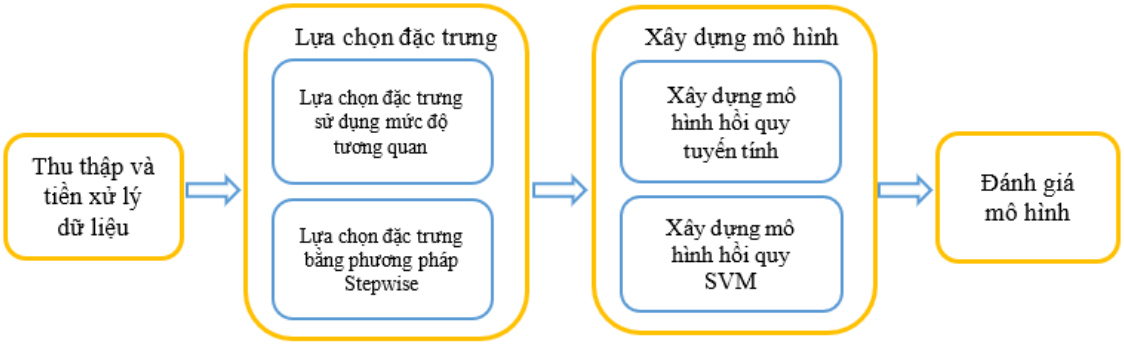
Trong phần này, nhóm tác giả đề xuất việc lựa chọn các biến và xây dựng mô hình dự đoán kết quả học tập của sinh viên trong môi trường học tập kết hợp với dữ liệu thu thập từ lớp học trực tiếp và hệ thống quản lí học tập Moodle. Quy trình xây dựng mô hình dự đoán kết quả học tập trong dạy học kết hợp bao gồm 4 bước như trong Hình 1:

(1) Thu thập dữ liệu từ một số nguồn và tiến hành tiền xử lí dữ liệu.

(2) Lựa chọn các thuộc tính dữ liệu ảnh hưởng tới kết quả học tập.

(3) Xây dựng mô hình hồi quy dự đoán kết quả học tập của sinh viên.

(4) Đánh giá hiệu quả các mô hình.

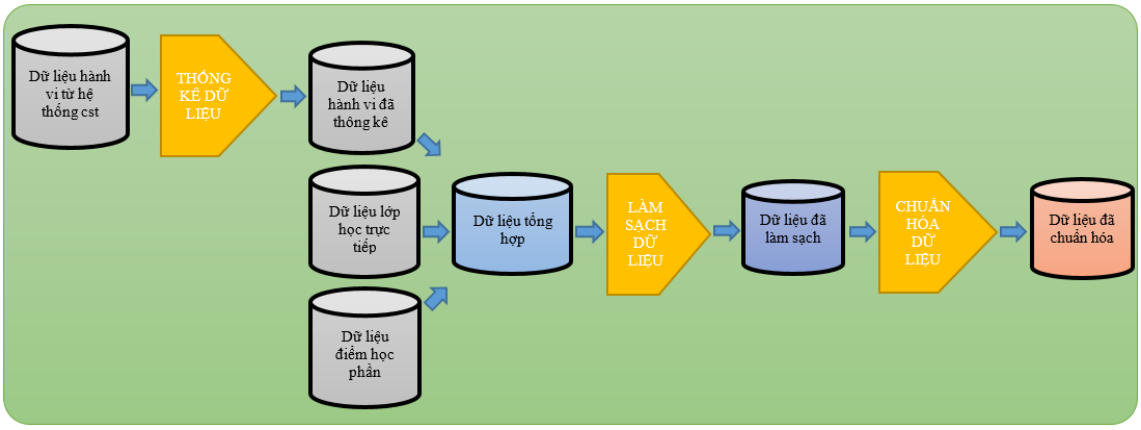


Hình 1. Quy trình xây dựng mô hình dự đoán điểm học tập trong môi trường học tập kết hợp

2.2.1. Dữ liệu

Trong nghiên cứu này, nhóm tác giả sử dụng dữ liệu về hành vi và dữ liệu điểm thi học phần của người học để xây dựng mô hình. Nghiên cứu được thực hiện trên 392 sinh viên năm nhất đến từ các khoa của Trường Đại học Sư phạm Hà Nội. Các sinh viên này tham gia học phần "*Tin học Đại cương*" với hình thức học tập kết hợp diễn ra trong 10 tuần của học kì hè từ tháng 5 đến tháng 7 năm 2022 do cùng một giáo viên giảng dạy.

Quy trình thu thập và tiền xử lí được mình họa trong Hình 2.



Hình 2. Quy trình thu thập và tiền xử lí dữ liệu

**** Thu thập dữ liệu***

Dữ liệu thực nghiệm được thu thập từ các nguồn sau: hệ thống học tập trực tuyến, lớp học trực tiếp và hệ thống quản lí điểm của nhà trường:

Môi trường học tập trực tuyến: Hệ thống học tập trực tuyến https://cst.hnue.edu.vn có thể ghi lại nhật kí hành vi của người sử dụng trong mỗi khóa học như: xem tệp, xem video, làm trắc nghiệm, nộp bài tập. Các hành vi này thể hiện thái độ học tập của người học vì vậy đây là những yếu tố có thể ảnh hưởng đến kết quả học tập. Hình 3 minh họa dữ liệu hành vi được trích xuất từ hệ thống học tập trực tuyến.

	A	B	C	D	E	F	G	H	
1	Thời gian	Tên đầy đủ	người dùn	Bối cảnh của sự kiện	thành phần	Tên sự kiện	Mô tả	Nguyên thi	Địa chỉ
2	7/08/2022 17:04	Phạm Thị Ngọc Hoa -K71 Toán	-	Khoá học: COMP 103 - Tin họ	Hệ thống	Đã xem khóa họ	The user with id '23701' view	web	171.255
3	7/08/2022 07:31	Phạm Thị Ngọc Hoa -K71 Toán	-	Khoá học: COMP 103 - Tin họ	Hệ thống	Đã xem khóa họ	The user with id '23701' view	web	27.67.3
4	7/08/2022 02:32	Phạm Thị Ngọc Hoa -K71 Toán	-	Khoá học: COMP 103 - Tin họ	Hệ thống	Đã xem khóa họ	The user with id '23701' view	web	171.255
5	7/08/2022 00:53	Phạm Thị Ngọc Anh -K71 Toán	-	Khoá học: COMP 103 - Tin họ	Hệ thống	Đã xem khóa họ	The user with id '23614' view	web	100.88.
6	6/08/2022 19:39	Dương Ngọc Anh -K71 Toán	-	Thư mục: Project 1	Thư mục	Mô-đun khóa họ	The user with id '23603' view	web	27.71.4
7	6/08/2022 15:03	Trần Thị Linh Chi -K71A SH	-	Bài tập: Nộp bài thi Ca 3	Bài tập	Trạng thái của bi	The user with id '22992' has v	web	10.14.1
8	6/08/2022 15:03	Trần Thị Linh Chi -K71A SH	-	Bài tập: Nộp bài thi Ca 3	Bài tập	Mô-đun khóa họ	The user with id '22992' view	web	10.14.1
9	6/08/2022 15:03	Trần Thị Linh Chi -K71A SH	-	Khoá học: COMP 103 - Tin họ	Hệ thống	Đã xem khóa họ	The user with id '22992' view	web	10.14.1
10	6/08/2022 14:53	Trần Thị Linh Chi -K71A SH	-	Bài tập: Nộp bài thi Ca 3	Bài tập	Trạng thái của bi	The user with id '22992' has v	web	10.14.1
11	6/08/2022 14:53	Trần Thị Linh Chi -K71A SH	-	Bài tập: Nộp bài thi Ca 3	Bài tập	Mô-đun khóa họ	The user with id '22992' view	web	10.14.1
12	6/08/2022 14:53	Trần Thị Linh Chi -K71A SH	-	Bài tập: Nộp bài thi Ca 3	Bài tập	Một bài làm đã c	The user with id '22992' has s	web	10.14.1

Hình 3. Dữ liệu nhật kí hành vi xuất ra từ hệ thống https://cst.hnue.edu.vn

Môi trường học tập trực tiếp: Sự chuyên cần góp mặt trong các tiết học trực tiếp thể hiện thái độ học tập của người học, vì vậy, chúng tôi tiến hành thu thập dữ liệu chuyên

cần bằng hình thức điểm danh trực tiếp. Ngoài ra, điểm bài tập nhóm và điểm kiểm tra trên lớp cũng thể hiện mức độ hiểu bài và một phần hiệu quả học tập của người học. Do đó, chúng tôi thu thập hai điểm này từ giáo viên và trợ giảng.

Điểm thi học phần: Kết thúc học kì, sinh viên phải làm một bài thi kết thúc học phần trên máy tính. Dữ liệu điểm thi có thể được tải trực tiếp từ hệ thống https://daotao.hnue.edu.vn. Điểm này được dùng làm biến phụ thuộc trong việc dự đoán kết quả học tập của người học.

**** Tiền xử lí dữ liệu***

Quá trình tiền xử lí dữ liệu được tiến hành qua bốn giai đoạn:

Giai đoạn 1 - Thống kê dữ liệu: Dữ liệu hành vi trích xuất từ hệ thống học tập trực tuyến chưa thể sử dụng ngay để dự đoán kết quả học tập. Để đánh giá mức độ tích cực của người học, từ dữ liệu hành vi, nhóm tác giả thực hiện thống kê các thông tin: số lần xem khóa học, số lần xem tệp, số lần xem video, số lần làm bài trắc nghiệm, điểm trung bình các lần trả lời trắc nghiệm, số lần nộp bài tập. Việc thống kê này thực hiện bằng một chương trình được lập trình trên ngôn ngữ Python. Tương tự như vậy, dữ liệu chuyên cần trên lớp học trực tiếp cũng được thống kê nhưng việc thống kê này khá đơn giản được thực hiện bằng phần mềm Excel.

Giai đoạn 2 - Tích hợp dữ liệu: Dữ liệu từ 3 nguồn (dữ liệu hành vi trên hệ thống trực tuyến được thống kê trong giai đoạn 1, dữ liệu lớp học trực tiếp, dữ liệu điểm học phần) sẽ được lưu trữ và tổng hợp trên Excel dựa trên mã sinh viên của người học.

Giai đoạn 3 - Làm sạch dữ liệu: Để mô hình dự đoán hoạt động hiệu quả thì bộ dữ liệu đầu vào không được chứa các dữ liệu khuyết, dữ liệu nhiễu. Quá trình loại bỏ các dữ liệu khuyết, dữ liệu nhiễu được gọi là làm sạch dữ liệu. Một số sinh viên tham gia đầy đủ trong cả quá trình học nhưng vì một lí do nào đó mà không đến dự thi và đạt điểm thi học phần là 0 hoặc một số sinh viên có lí do chính đáng được Nhà trường cho phép lùi lịch thi sang học kì tiếp theo thì không có điểm thi học phần. Đây chính là các dữ liệu nhiễu và khuyết cần phải loại bỏ. Dữ liệu sau khi làm sạch có thể làm đầu vào xây dựng mô hình hồi quy tuyến tính.

Sau khi thực hiện xong các thao tác làm sạch dữ liệu, trường Mã sinh viên sẽ được loại bỏ và bộ dữ liệu thu được còn 365 bản ghi. Hình 4 minh họa bộ dữ liệu sau khi được làm sạch và Bảng 1 mô tả chi tiết ý nghĩa từng trường dữ liệu của bộ dữ liệu. Trong đó trường *Final_exam* là biến phụ thuộc và các trường còn lại là biến độc lập.

	A	B	C	D	E	F	G	H	I	J	
1	num_video_view	num_quiz_attemp	quiz_avg	num_course_view	num_assignment_submitted	num_file_view	num_behavior	num_attend_F2F	F2F_quiz_mark	group_mark	Fina
2	136	109	93.33333333	259	5	55	4318	13	100	7	
3	30	48	91.66666667	156	6	41	1517	18	76.67	7	
4	1	10	33.666	46	0	38	247	8	0	7	
5	67	113	100	309	6	37	3463	17	96.67	7	
6	62	62	99.16666667	253	6	42	2111	15	90	7	
7	99	127	100	348	6	62	3351	14	93.33	7	
8	100	88	93.61166667	236	6	59	3376	15	70	7	
9	50	13	37.77833333	211	6	54	1820	14	73.33	7	
10	69	84	100	197	6	39	3558	18	93.33	7	

Hình 4. Dữ liệu sau khi đã làm sạch

Giai đoạn 4 - Chuẩn hóa dữ liệu: Với thuật toán hồi quy véc-tơ hỗ trợ, dữ liệu đầu vào cần được chuẩn hóa về các giá trị trong khoảng từ 0 đến 1. Việc chuẩn hóa được thực hiện bằng cách sử dụng hàm MinMaxScaler trong thư viện Scikit-learn của Python.

Bảng 1. Mô tả các trường dữ liệu sau quá trình tiền xử lí

Nguồn dữ liệu	Tên trường	Ý nghĩa trường	Khoảng giá trị
Dữ liệu hành vi trên hệ thống trực tuyến	*num_video_view*	Số lượt xem video	≥ 0
	num_quiz_attemp	Số lượt làm trắc nghiệm	≥ 0
	quiz_avg	Điểm trung bình các lần làm trắc nghiệm	0…100
	num_course_view	Số lần xem khóa học	≥ 0
	num_assignment_submitted	Số lần nộp bài tập	0…6
	num_file_view	Số lần xem tệp	≥ 0
	num_behavior	Tổng số hành vi trong nhật ký khóa học	≥ 0
Dữ liệu hành vi trên lớp học trực tiếp	*num_attend_F2F*	Số tiết học tham dự trực tiếp	0…20
	F2F_quiz_mark	Điểm trắc nghiệm lớp học trực tiếp	0…100
	group_mark	Điểm bài tập nhóm	0…10
Dữ liệu điểm học phần	*Final_exam*	Điểm thi học phần	0…10

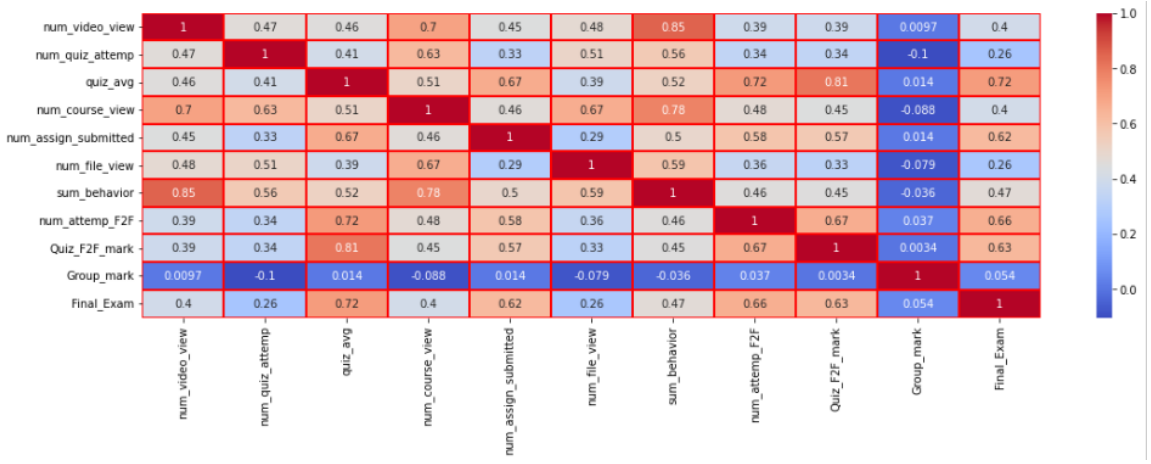
2.2.2. Lựa chọn các biến để xây dựng mô hình dự đoán kết quả học tập

Để xây dựng được mô hình dự đoán hiệu quả, chúng ta cần phải lựa chọn các biến có ảnh hưởng đến kết quả học tập của người học. Trong phần này, nhóm tác giả tiến hành lựa chọn các biến dùng để xây dựng mô hình bằng hai cách. Cách thứ nhất là sử dụng hệ số tương quan giữa biến *Final_exam* và các biến còn lại. Cách thứ hai là sử dụng phương pháp phân tích hồi quy tuyến tính Stepwise trong SPSS 20.

**** Lựa chọn các biến dựa trên mức độ tương quan***

Để xác định mức độ tương quan giữa các biến độc lập với biến phụ thuộc *Final_exam*, nhóm tác giả đã sử dụng thư viện matplotlib của Python để vẽ biểu đồ heatmap (Hình 5) và tính toán giá trị tương quan cho từng cặp biến. Mỗi ô trong biểu đồ được tô một màu tương ứng với giá trị tương quan của hai biến tại hàng và cột tương ứng. Giá trị trong mỗi ô càng gần 1 hoặc màu sắc của ô càng gần màu đỏ thì hai biến càng có độ tương quan lớn. Biến *quiz_avg* là biến có độ tương quan lớn nhất với biến *Final_exam* (0.72). Ngoài ra, các biến *num_assign_submitted*, *num_attend_F2F* và *Quiz_F2F_mark* cũng có tương quan mạnh với *Final_exam* với giá trị tương quan lần lượt là 0.62, 0.66 và 0.63. Bốn biến

có mức độ tương quan lớn nhất này (> 0,5) và biến *Final_exam* được lựa chọn để làm đầu vào cho việc xây dựng mô hình hồi quy.



Hình 5. Biểu đồ Heatmap biểu diễn mức độ tương quan giữa các cặp biến

**** Lựa chọn các biến dựa trên phương pháp Stepwise***

Trong cách tiếp cận thứ hai, để xác định các biến có ảnh hưởng đến mô hình hồi quy, nhóm tác giả tiến hành phân tích hồi quy tuyến tính theo phương pháp Stepwise [10]. Với phương pháp này, các biến sẽ lần lượt được đưa vào mô hình, biến nào có ảnh hưởng sẽ làm cho hiệu suất của mô hình tăng lên được thể hiện bởi chỉ số R-Square (R^2), và Std.Error of the Estimate. Trong đó, R-Square thể hiện mức độ phù hợp của mô hình và Std. Error of Estimate thể hiện mức độ lỗi của mô hình với các biến được đưa vào.

- *Mô hình 1:* Biến quiz_avg được đưa vào đầu tiên và là biến có ảnh hưởng lớn nhất đến mô hình. Chỉ số R2 của mô hình là 0.503 hay biến này giải thích 50% sự thay đổi của biến Final_exam.

- *Mô hình 2:* Biến num_attend_F2F được xác định là biến thứ hai có ảnh hưởng đến mô hình. Chỉ số R2 của mô hình hai biến tăng lên 0,546.

- *Mô hình 3:* Biến num_assignment_submitted được xác định là biến thứ ba ảnh hưởng đến mô hình. Chỉ số R2 của mô hình ba biến tăng lên 0,571.

- *Mô hình 4:* Biến num_quiz_attemp được xác định là biến thứ tư ảnh hưởng đến mô hình. Chỉ số R2 của mô hình bốn biến tăng lên 0,576.

- *Mô hình 5*: Biến num_behavior được xác định là biến thứ năm ảnh hưởng đến mô hình. Chỉ số R^2 của mô hình bốn biến tăng lên 0,583.

Bảng 2 minh họa kết quả khi đưa các biến *quiz_avg*, *num_attend_F2F*, *num_assignment_submitted*, *num_quizz_attemp* và *num_behavior* vào mô hình. Mỗi biến được đưa vào mô hình làm cho chỉ số R^2 tăng lên và lỗi giảm đi. Như vậy, chỉ số R^2 của mô hình 5 biến là 0.583 nghĩa là 5 biến trên có ảnh hưởng và giải thích 58% sự biến thiên của biến *Final_exam*. Còn lại, 42% sự thay đổi còn lại của biến *Final_exam* do các biến khác tác động hoặc do các yếu tố khách quan tác động.

Bảng 2. Minh họa các chỉ số R-Squared và mức độ lỗi cho năm mô hình khi thực hiện Stepwise

Tóm tắt mô hình[f]			
Mô hình	**R**	**R-Square**	**Lỗi ước tính**
1	0,709[a]	0,503	1,5640
2	0,739[b]	0,546	1,4977
3	0,755[c]	0,571	1,4575
4	0,759[d]	0,576	1,4510
5	0,764[e]	0,583	1,4406

a. Predictors: (Constant), quiz_avg

b. Predictors: (Constant), quiz_avg, num_attend_F2F

c. Predictors: (Constant), quiz_avg, num_attend_F2F, num_assignment_ submitted

d. Predictors: (Constant), quiz_avg, num_attend_F2F, num_assignment_ submitted, num_quiz_attemp

e. Predictors: (Constant), quiz_avg, num_attend_F2F, num_assignment_ submitted, num_quiz_attemp, num_behavior

f. Dependent Variable: Final_exam

2.2.3. Mô hình dự đoán kết quả học tập

Để dự đoán kết quả học tập của sinh viên trong mô hình học tập kết hợp, chúng tôi xây dựng các mô hình hồi quy. Các biến được lựa chọn từ hai phương pháp trên là các biến độc lập và biến *Final_exam* là biến phụ thuộc (biến dự đoán).

Để đánh giá mức độ ảnh hưởng của các biến với sự biến thiên của biến phụ thuộc, dữ liệu được đưa vào SPSS để xây dựng mô hình hồi quy tuyến tính, xây dựng biểu đồ phần dư và đánh giá mức ý nghĩa của các hệ số trong mô hình xây dựng được.

Ngoài ra, để đánh giá hiệu quả các mô hình xây dựng từ hai tập biến, nhóm tác giả thực hiện xây dựng các mô hình hồi quy tuyến tính và hồi quy SVR [11] bằng cách sử dụng thư viện Scikit-learn trong Python với phương pháp thực nghiệm 10fold cross-validation [12]. Các độ đo hệ số xác định (R-Square - R^2), trung bình trị tuyệt đối giữa giá trị thực tế và giá trị dự đoán (Mean Absolute Error - MAE) và độ lệch chuẩn các phần dư (Root Mean Square Error - RMSE) [13] được sử dụng để so sánh các mô hình hồi quy. Các mô hình hồi quy SVR được xây dựng với 2 loại hàm nhân *linear* và *poly*. Bảng sau mô tả các tham số được sử dụng khi xây dựng mô hình.

Bảng 3. Tham số được sử dụng khi xây dựng mô hình

kernel	**C**	**gamma**	**degree**	**epsilon**
linear	1000	auto	-	-
poly	1000	0,1	6	0,1

2.3. Kết quả thực nghiệm và đánh giá

Dữ liệu sau khi thực hiện quá trình tiền xử lí thu được còn 365 bản ghi. Từ đây, hai bộ dữ liệu tương ứng với hai nhóm thuộc tính được lựa chọn theo hai phương pháp ở trên được trích xuất. Bảng 4 minh họa danh sách các biến được lựa chọn theo hai phương pháp.

Bảng 4. Các biến được lựa chọn dựa trên mức độ tương quan và phân tích Stepwise

Biến	Các biến lựa chọn dựa trên mức độ tương quan	Các biến lựa chọn dựa trên phân tích Stepwise
Biến độc lập	*quiz_avg* *num_assignment_submitted* *num_attend_F2F* *F2F_quiz_mark*	*num_quiz_attemp* *quiz_avg* *num_assignment_submitted* *num_behavior* *num_attend_F2F*
Biến phụ thuộc	*Final_exam*	*Final_exam*

2.3.1. Các mô hình hồi quy xây dựng bằng SPSS

**** Mô hình hồi quy tuyến tính dựa trên năm biến được chọn bằng phương pháp Stepwise***

Năm biến lựa chọn từ phương pháp phân tích Stepwise được đưa vào huấn luyện để xây dựng mô hình hồi quy tuyến tính trên SPSS. Bảng 5 minh họa mô hình hồi quy tuyến tính được xây dựng từ năm biến độc lập đã xác định có ảnh hưởng đến biến phụ thuộc *Final_exam* với hai bộ hệ số chưa chuẩn hóa (Unstandardized Coefficients) và đã chuẩn hóa (Standardized Coefficients). Mô hình hồi quy với bộ hệ số được chuẩn hóa cho biết điểm thi học phần của sinh viên có thể được tính theo công thức sau:

$$\begin{aligned} Final\_exam = &-0.124 \times num\_quiz\_attemp + 0.385 \times quiz\_avg + \\ &0.201 \times num\_assignment\_submitted + \\ &0.015 \times num\_behavior + 0.25 \times num\_attend\_F2F \end{aligned} \quad (1)$$

Hệ số của các biến trong công thức này cũng cho thấy mức độ tác động của các biến đến biến *Final_exam*. Trong đó, điểm trung bình các lần làm trắc nghiệm (*quiz_avg*) của sinh viên có tác động lớn nhất và biến thiên cùng chiều với kết quả học tập (*Final_exam*). Điều này cho biết rằng sinh viên có điểm trung bình các lần làm trắc nghiệm càng cao thì đạt kết quả học tập càng tốt. Tương tự như vậy, các biến *num_assignment_submitted*, *num_behavior* và *num_attend_F2F* cũng biến thiên cùng chiều với *Final_exam*. Trong khi đó, biến *num_quiz_attemp* có tác động biến thiên ngược chiều với *Final_exam* hay người học có số lần làm trắc nghiệm ít thì vẫn có thể đạt thành tích học tập cao.

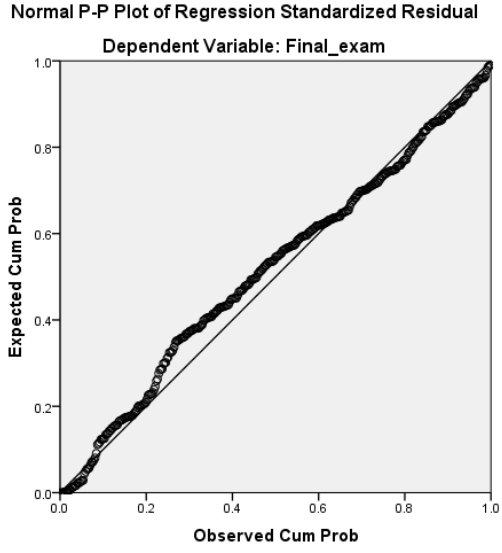
Cột t thể hiện các giá trị kiểm định t cho các hệ số của các biến trong mô hình hồi quy. Cột Sig. cho biết các biến này đều có ý nghĩa thống kê (< 0,05) và có tác động đáng kể đến biến phụ thuộc *Final_exam*.

Bảng 5. Mô hình hồi quy tuyến tính dự đoán kết quả thi học phần

Mô hình	Bộ hệ số chưa chuẩn hóa		Bộ hệ số chuẩn hóa	Kiểm định *t*	Mức ý nghĩa
	B	Std. Error	Beta		
(Constant)	-2,164	0,524		-4,129	0,000
num_quiz_attemp	-0,006	0,002	-0,124	-2,973	0,003
quiz_avg	0,048	0,007	0,385	6,829	0,000
num_assignment_ submitted	0,279	0,067	0,201	4,181	0,000
num_behavior	0,000	0,000	0,115	2,495	0,013
num_attend_F2F	0,222	0,045	0,250	4,951	0,000

a. Biến phụ thuộc: Final_exam

Hình 7 là biểu đồ P-P Plot mô tả phân bố phần dư của các điểm dữ liệu so với đường hồi quy của mô hình xây dựng. Các điểm dữ liệu trong phân phối của phần dư trên biểu đồ bám khá sát đường chéo. Điều này thể hiện phân phối của phần dư là một phân phối chuẩn. Như vậy, bộ dữ liệu nghiên cứu là bộ dữ liệu tốt và mô hình lựa chọn là phù hợp.



Hình 7. Biểu đồ P-P Plot biểu diễn phần dư của các điểm dữ liệu

**** Mô hình hồi quy tuyến tính dựa trên 4 biến có mức độ tương quan cao***

Các biến có mức độ tương quan cao cũng được đưa vào để xây dựng mô hình hồi quy tuyến tính dự đoán kết quả học tập. Kết quả (Bảng 6) cho thấy mô hình có hệ số R^2 là 0,573. Giá trị này thấp hơn giá trị R^2 của mô hình năm biến hay mô hình năm biến phù hợp với bộ dữ liệu hơn so với mô hình bốn biến này.

Bảng 6. Các chỉ số đánh giá cho mô hình hồi quy tuyến tính xây dựng từ 4 biến có độ tương quan cao với biến Final_exam

Model	R	R-Square	Std. Error of the Estimate
1	0,757[a]	0,573	1,4558

a. Biến phụ thuộc: Final_exam

Bảng 7 minh họa mô hình hồi quy tuyến tính được xây dựng từ bốn biến nêu trên. Tuy nhiên, giá trị trong cột Sig. của biến *F2F_quiz_mark* là 0,175 (> 0,05) cho thấy biến này không có ý nghĩa thống kê trong mô hình xây dựng được.

Bảng 7. Mô hình hồi quy tuyến tính xây dựng từ bốn biến có mức độ tương quan cao

Mô hình	Hệ số chưa chuẩn hóa		Hệ số chuẩn hóa	Kiểm định *t*	Mức ý nghĩa
	B	*Std. Error*	*Beta*		
(Constant)	-2,197	0,526		-4,173	0,000
quiz_avg	0,040	0,009	0,325	4,686	0,000
num_assignment_ submitted	0,302	0,066	0,217	4,587	0,000
num_attend_F2F	0,212	0,046	0,239	4,621	0,000
F2F_quiz_mark	0,011	0,008	0,082	1,363	0,174

2.3.2. Các mô hình hồi quy xây dựng bằng Scikit-learn

**** Mô hình hồi quy tuyến tính***

Kết quả đánh giá các mô hình hồi quy tuyến tính sử dụng thư viện Scikit-learn trong Python với phương pháp thực nghiệm 10-fold cross-validation được minh họa trong Bảng 8. Mô hình hồi quy tuyến tính được xây dựng từ tập biến lựa chọn bằng phương pháp Stepwise có hệ số xác định R^2 cao hơn và mức độ lỗi (MAE, RMSE) thấp hơn so với mô hình được xây dựng từ tập biến lựa chọn bằng mức độ tương quan. Độ đo R^2 càng cao thì mô hình càng phù hợp với bộ dữ liệu và các độ đo MAE và RMSE càng thấp thể hiện mức độ lỗi khi áp dụng mô hình với dữ liệu kiểm thử càng thấp.

Bảng 8. Các độ đo MAE, RMSE khi thực hiện xây dựng mô hình hồi quy tuyến tính sử dụng Scikit-learn

Mô hình	R^2	MAE	RMSE
Mô hình sử dụng biến xác định dựa trên Stepwise	0,571	1,082	1,435
Mô hình sử dụng biến có mức độ tương quan cao	0,559	1,111	1,454

**** Mô hình hồi quy Suport Vector Regression (SVR)***

Bảng 9 minh họa các độ đo R^2, MAE và RMSE của các mô hình hồi quy SVR sau khi chạy thực nghiệm với phương pháp 10-fold cross-validation. Kết quả này cũng cho thấy các mô hình được xây dựng từ tập biến lựa chọn bằng phương pháp Stepwise có hiệu quả tốt hơn trong đó mô hình hồi quy SVR với hàm nhân *poly* là mô hình tốt nhất.

Bảng 9. Các độ đo MAE, RMSE khi thực hiện xây dựng mô hình hồi quy SVR

Mô hình	Kernel	R^2	MAE	RMSE
Mô hình sử dụng biến xác định dựa trên Stepwise	linear	0,5573	1,0761	1,4575
	poly	**0,583**	**1,0394**	**1,4131**
Mô hình sử dụng biến có mức độ tương quan cao	linear	0,5217	1,1310	1,5150
	poly	0,5158	1,1335	1,5243

2.4. Một số đánh giá

Khi thực hiện cả hai phương pháp lựa chọn biến, biến *num_attend_F2F* (số lần tham gia lớp học trực tiếp) đều được đánh giá là biến quan trọng. Khi đánh giá bằng mức độ tương quan, biến này có mức độ tương quan với biến phụ thuộc *Final_exam* lớn thứ hai (0.66). Khi phân tích hồi quy Stepwise, biến này cũng là biến thứ hai được xác định là có ảnh hưởng đến mô hình hồi quy. Hệ số của biến này trong phương trình hồi quy cũng thể hiện mức độ ảnh hưởng của biến này đến sự biến thiên của biến phụ thuộc. Như vậy, chúng ta có thể khẳng định rằng, việc sử dụng thêm dữ liệu thu thập từ môi trường học tập trực tiếp góp phần nâng cao hiệu quả dự đoán kết quả học tập của sinh viên.

Đối với việc xây dựng mô hình hồi quy tuyến tính, trong các biến lựa chọn dựa trên mức độ tương quan thì biến *F2F_quiz_mark* không có ý nghĩa với mô hình. Bên cạnh đó, các mô hình hồi quy được xây dựng sử dụng thư viên Scikit-learn được huấn luyện từ dữ liệu gồm các biến lựa chọn từ phương pháp phân tích Stepwise có mức độ lỗi thấp hơn. Như vậy, tập biến lựa chọn từ phương pháp Stepwise phù hợp để xây dựng các mô hình hồi quy hơn.

3. Kết luận

Trong bài báo này, nhóm tác giả đã xác định được các yếu tố ảnh hưởng đến kết quả học tập của sinh viên trong mô hình học tập và đề xuất một số mô hình dự đoán từ các yếu tố này với dữ liệu thu thập từ cả hai môi trường học tập. Kết quả thực nghiệm cho thấy rằng việc lựa chọn các biến dữ liệu bằng phương pháp Stepwise phù hợp để xây dựng các mô hình hồi quy hơn và việc sử dụng thêm các biến thu thập từ lớp học trực tiếp cũng góp phần nâng cao hiệu quả dự đoán. Mặc dù quy mô nghiên cứu còn nhỏ nhưng kết quả thu được là đáng tin cậy và phương pháp này có thể áp dụng để xác định các nhân tố ảnh hưởng đến kết quả học tập của các học phần khác. Từ việc xác định được các nhân tố ảnh hưởng và xây dựng mô hình dự đoán kết quả học tập, các nhà giáo dục có thể biết được những hành vi nào trong môi trường trực tiếp và trực tuyến có ảnh hưởng đến kết quả học tập và có phương án tác động để nâng cao hiệu quả học tập của người học.

Trong các nghiên cứu sau này, nhóm tác giả sẽ hướng đến sử dụng hệ thống camera lớp học để thu thập hành vi người học trong lớp học trực tiếp một cách tự động. Điều này có thể giảm thiểu công sức thu thập và tiền xử lí dữ liệu cũng như góp phần nâng cao hiệu quả dự đoán. Ngoài ra, nhóm tác giả cũng sẽ tiến hành nghiên cứu bài toán này với các thuật toán học máy khác như phân lớp, phân cụm cũng như giải quyết vấn đề mất cân bằng dữ liệu.

Lời cảm ơn. Nghiên cứu được hoàn thành dưới sự tài trợ của đề tài nghiên cứu khoa học cấp Trường Đại học Sư phạm Hà Nội, mã số SPHN22-08.

TÀI LIỆU THAM KHẢO

[1] K. Do Trung and H. Nguyen Thi, 2021. Application of B-learning in teaching data structures and algorithms at Hanoi National University of Education. *Journal of Science Educational Science*, Vol. 66, No. 3, pp. 229-241, doi: 10.18173/2354-1075.2021-0129.

[2] Lê Thị Thu Hiền, 2013. Áp dụng mô hình dạy học hỗn hợp trong dạy học vật lí ở trường Trung học phổ thông. *Tạp chí khoa học Giáo dục*, Vol. 98, pp. 23-25.

[3] T. Yigit, A. Koyun, A. S. Yuksel, and I. A. Cankaya, 2014. Evaluation of Blended Learning Approach in Computer Engineering Education. *Procedia Soc Behav Sci*, Vol. 141, pp. 807-812, doi: 10.1016/j.sbspro.2014.05.140.

[4] A. Abu Saa, M. Al-Emran, and K. Shaalan, 2019. Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. *Technology, Knowledge and Learning*, Vol. 24, No. 4, pp. 567-598, doi: 10.1007/s10758-019-09408-7.

[5] N. Z. Zacharis, 2015. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *Internet and Higher Education*, Vol. 27, pp. 44-53, doi: 10.1016/j.iheduc.2015.05.002.

[6] K. Fahd, S. J. Miah, and K. Ahmed, 2021. Predicting student performance in a blended learning environment using learning management system interaction data. *Applied Computing and Informatics*, doi: 10.1108/ACI-06-2021-0150.

[7] J. H. Zhang, L. cong Zou, J. jia Miao, Y. X. Zhang, G. J. Hwang, and Y. Zhu, 2020. An individualized intervention approach to improving university students' learning performance and interactive behaviors in a blended learning environment. *Interactive Learning Environments*, Vol. 28, No. 2, pp. 231-245, doi: 10.1080/10494820.2019.1636078.

[8] H. Sokout, T. Usagawa, and S. Mukhtar, 2020. Learning analytics: Analyzing various aspects of learners' performance in blended courses. The case of Kabul Polytechnic University, Afghanistan. *International Journal of Emerging Technologies in Learning*, Vol. 15, No. 12, pp. 168-190, doi: 10.3991/ijet.v15i12. 13473.

[9] W. Chango, R. Cerezo, and C. Romero, 2021. Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Computers and Electrical Engineering*, Vol. 89, doi: 10.1016/j.compeleceng.2020. 106908.

[10] Johnsson, 1992. A procedure for stepwise regression analysis Johnsson. *Statistical Papers Statistische Hefte*, pp. 21-29.

[11] Mariette Awad and Rahul Khanna, 2015. *Support Vector Regression*, Apress, Berkeley, CA. Springer.

[12] J. D. Rodríguez, A. Pérez, and J. A. Lozano, 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Trans Pattern Anal Mach Intell*, Vol. 32, No. 3, pp. 569-575, doi: 10.1109/TPAMI.2009.187.

[13] T. Chai and R. R. Draxler, 2014. Root mean square error (RMSE) or mean absolute error (MAE)? *Geosci. Model Dev. Discuss*, Vol. 7, pp. 1525-1534, doi: 10.5194/gmdd-7-1525-2014.

ABSTRACT

Using machine learning to identify influential factors and predict student academic performance in blended learning

Nguyen Thi Hong, Tran Hai Long and Do Trung Kien

Faculty of Information Technology, Hanoi National University of Education

This study aims to identify the factors that influence academic performance and use them to develop a predictive model for student academic achievement, in order to support the improvement of education quality. In previous studies, the selection and evaluation of factors were only conducted on online learning data. In this study, we propose using a selected set of attributes from experimental data collected both in face-to-face classes and on the online learning system at Hanoi National University of Education. To build the predictive model for academic performance, we employed two variable selection methods: one is to choose highly correlated variables, and the other is to use the Stepwise linear regression analysis. Furthermore, two machine learning algorithms, linear regression, and support vector regression were used to construct the predictive model. The experimental results show that the support vector regression model with a polynomial kernel function built from the Stepwise-selected variables is the most effective.

Keywords: predicting performance, blended learning, machine learning.