APPLYING MACHINE LEARNING FOR PREDICTING THE DROPOUT OF STUDENTS

Nong Thi Hoa
Thuy Loi University

ARTICLE INFO

Received: 04/3/2025
Revised: 11/6/2025
Published: 25/6/2025

KEYWORDS

Neural network
Random Forest
Support Vector Machine
Machine Learning
Prediction

ABSTRACT

Currently, the number of students dropping out of some universities is increasing due to many factors affecting students. Predicting the possibility of students dropping out will help to provide the supports for students in time. In this paper, the most new effective machine learning models were applied on the benchmark dataset to predict students dropping out. The benchmark dataset has 36 features about the learning results in the first two years and social factors. Important features were analyzed to improve the classification performance of machine learning models. The dataset was preprocessed to meet the input of each machine learning model. Neural network, Random Forest, Support Vector Machine were applied in this study. Parameters of each machine learning model were adjusted to get the highest classification accuracy. Experimental results show that Random Forest is the best machine learning model for the problem. Its accuracy reaches 91.33%.

ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY VÀO DỰ ĐOÁN TÌNH TRẠNG BỎ HỌC CỦA SINH VIÊN

Nông Thị Hoa
Trường Đại học Thủy Lợi

THÔNG TIN BÀI BÁO

Ngày nhận bài: 04/3/2025
Ngày hoàn thiện: 11/6/2025
Ngày đăng: 25/6/2025

TỪ KHÓA

Neural network
Random Forest
Support Vector Machine
Học máy
Dự đoán

TÓM TẮT

Ngày nay, số sinh viên nghỉ học ở các trường đại học ngày càng tăng do nhiều yếu tố bởi nhiều yếu tố ảnh hưởng đến sinh viên. Từ kết quả dự đoán sinh viên bỏ học, các trường học đưa ra các giải pháp hỗ trợ để các sinh viên đó kịp thời. Trong bài báo này, các mô hình học máy mới và hiệu quả nhất được áp dụng trên tập dữ liệu chuẩn để dự đoán các sinh viên bỏ học. Tập dữ liệu chuẩn quốc tế có 36 đặc trưng về kết quả học tập hai năm học đầu tiên và các yếu tố về xã hội. Các đặc trưng quan trọng đã được phân tích để cải thiện hiệu quả phân lớp của các mô hình học máy. Tập dữ liệu được tiền xử lý để phù hợp với các dữ liệu vào của từng mô hình học máy. Neural network, Random Forest, Support Vector Machine là các mô hình học máy được ứng dụng trong nghiên cứu này. Từng mô hình học máy được điều chỉnh bộ tham số để thu được kết quả phân lớp có độ chính xác cao nhất. Kết quả thực nghiệm cho thấy Random Forest là mô hình học máy phù hợp nhất cho bài toán với độ chính xác là 91,33%.

DOI: https://doi.org/10.34238/tnu-jst.12201

Email: nongthihoa@tlu.edu.vn

1. Giới thiệu

Hiện nay, số lượng sinh viên ở các trường đại học bỏ học ngày càng tăng. Để cải thiện tình trạng này, việc dự đoán sinh viên bỏ học giúp nhà trường đưa ra các giải pháp hỗ trợ phù hợp đến các sinh viên đó. Sinh viên bỏ học do nhiều yếu tố khác nhau như kết quả học tập thấp, lo lắng cho chi phí học tập và ăn ở, chuyên ngành đang theo học, và tác động của các yếu tố kinh tế xã hội của quốc gia. Vì vậy, một ứng dụng thông minh tích hợp tri thức về các yếu tố ảnh hưởng đến kết quả học tập của sinh viên là hết sức cần thiết và hữu ích cho các trường đại học.

Hiện nay, một số nghiên cứu đã được thực hiện trên các sinh viên ở Mỹ, Phần Lan, Tây Ban Nha. Matti Vaarma [1] dùng cây quyết định (CatBoost), mạng nơ-ron và hồi quy để dự đoán các sinh viên bỏ học trên tập dữ liệu về sinh viên ở Phần Lan và các thử nghiệm cho thấy độ chính xác đạt 81%. Tương tự, Achmad Ridwana [2] dùng cây quyết định (XGBoost) trên tập dữ liệu chuẩn của UCI Machine Learning Repository (UCI) [5] và độ chính xác đạt 88%. Alice Villar [3] dùng cây quyết định (LightGBM, CatBoost) thử nghiệm trên tập dữ liệu chuẩn của UCI và có độ chính xác là 87%. Divvyam Arora [4] dùng kỹ thuật Stacking Classifier trên tập dữ liệu chuẩn của UCI và đạt độ chính xác 89%. Tuti Purwoningsih [6] dùng Random Forest trên tập dữ liệu của Đại học Terbuka, Indonesia. Trong nước, các nghiên cứu còn rất ít và tập trung vào dự đoán điểm học tập của sinh viên. Huỳnh Lê Uyên Minh [7] dùng cây quyết định để dự đoán khả năng tốt nghiệp của sinh viên năm 4 với tập dữ liệu của Đại học Đồng Tháp. Lưu Hoài Sang [8] dùng mạng nơ-ron đa tầng với kỹ thuật học sâu để dự đoán điểm một môn học của sinh viên dựa vào điểm thi đầu vào, điểm tích luỹ học tập, ngành, khóa học. Huỳnh Lý Thanh Nhàn [9] dùng giải thuật Biased Matrix Factorization để dự đoán điểm các môn chưa học dựa vào điểm của các môn học trước. Các nghiên cứu trên chưa có dự đoán tình trạng bỏ học của sinh viên ở Việt Nam. Trong bài báo này, ba kỹ thuật học máy tốt nhất cho bài toán phân loại (Neural network, Random Forest, Support Vector Machine) được dùng để dự đoán tình trạng bỏ học của sinh viên. Tập dữ liệu dùng cho thực nghiệm là tập dữ liệu chuẩn của UCI. Tập dữ liệu được tiền xử lý để tránh sự mất cân đối của số lượng mẫu giữa các lớp và chuẩn hóa dữ liệu để trở thành các dữ liệu vào phù hợp cho từng kỹ thuật học máy. Hơn nữa, tính quan trọng của từng đặc trưng của tập dữ liệu cũng được xem xét để cải thiện kết quả phân lớp. Kết quả thực nghiệm cho thấy Random Forest là kỹ thuật tốt nhất cho việc dự đoán sinh viên bỏ học.

Bài báo gồm các phần: giới thiệu vấn đề, cách giải quyết, kết quả thực nghiệm và kết luận. Phần 2 mô tả tập dữ liệu chuẩn và các kỹ thuật học máy áp dụng trong nghiên cứu này. Trong phần 3, các kết quả thực nghiệm được so sánh, giải thích. Các kết luận được nêu ra trong phần 4.

2. Phương pháp nghiên cứu

2.1. Tập dữ liệu chuẩn

Tập dữ liệu chuẩn của UCI [5] được tạo ra từ một cơ sở giáo dục đại học liên quan đến sinh viên theo học các chuyên ngành khác nhau ở Tây Ban Nha. Đây là bộ dữ liệu chuẩn của quốc tế nên tập dữ liệu sẽ thể hiện đúng, đủ các trường hợp đang có của sinh viên và các thông tin trong các mẫu dữ liệu có tính chính xác. Hơn nữa việc dùng tập dữ liệu chuẩn sẽ đánh giá tốt nhất hiệu quả của từng kỹ thuật học máy.

Tập dữ liệu có 4424 mẫu dữ liệu, mỗi mẫu có dữ liệu 36 đặc trưng. Danh sách đặc trưng gồm trạng thái hôn nhân, chế độ nhập học (mới tốt nghiệp, đã tốt nghiệp, đã đi làm,...), thứ tự chọn trường, chuyên ngành, thời gian học (ngày/đêm), bằng cấp đã có, điểm học tập của bằng cấp đã có, quốc gia, bằng cấp của mẹ, bằng cấp của cha, điểm đầu vào, nơi ở có cùng với nơi có trường đại học, yêu cầu đặc biệt về giáo dục, nợ tiền ngân hàng (có/không), trường có thay đổi học phí, giới tính, học bổng (có/không), tuổi nhập học, sinh viên nước ngoài (có/không), số tín chỉ đã đăng ký ở kỳ 1, số tín chỉ đã học ở kỳ 1, số tín chỉ đã thi ở kỳ 1, số tín chỉ đã đạt ở kỳ 1, điểm học tập kỳ 1, số tín chỉ đã đăng ký ở kỳ 2, số tín chỉ đã học ở kỳ 2, số tín chỉ đã thi ở kỳ 2, số tín chỉ đã đạt ở kỳ 2, số tín chỉ chưa thi ở kỳ 2, điểm học tập kỳ 2, tỷ lệ thất nghiệp, tỷ lệ lạm phát. Tập dữ liệu không có

các mẫu dữ liệu bất thường và các đặc trưng không bị thiếu dữ liệu nhưng có sự mất cân bằng về số mẫu dữ liệu giữa hai phân loại. Tập dữ liệu được gắn hai nhãn lớp gồm Dropout (bỏ học) và Graduate (tốt nghiệp). Hình 1 thể hiện một số mẫu dữ liệu đầu tiên trong tập dữ liệu.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1	Ma	Ap	Ap	Co	Da	Pre	Pre	Na	Ma	Fat	Ma	Fat	Ad	Dis	Edu	De	Tui	Ge	Sch	Ag	Int	Cu	Cu	Cu	Cu	Cu	Cu	Cu	Cu	Cu	Cu	Cu	Cu	Un	Inf	GDP	Target
2	1	17	5	171	1	1	122	1	19	12	5	9	127	1	0	0	1	1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	10.8	1.4	1.74	Dropouts
3	1	15	1	9254	1	1	160	1	1	3	3	3	143	1	0	0	0	1	0	19	0	0	6	6	6	14	0	0	6	6	6	13.7	0	13.9	-0.3	0.79	Graduate
4	1	1	5	9070	1	1	122	1	37	37	9	9	125	1	0	0	0	1	0	19	0	0	6	0	0	0	0	0	6	0	0	0	0	10.8	1.4	1.74	Dropouts
5	1	17	2	9773	1	1	122	1	38	37	5	3	120	1	0	0	1	0	0	20	0	0	6	8	6	13.4	0	0	6	10	5	12.4	0	9.4	-0.8	-3.12	Graduate
6	2	39	1	8014	0	1	100	1	37	38	9	9	142	0	0	0	1	0	0	45	0	0	6	9	5	12.3	0	0	6	6	6	13	0	13.9	-0.3	0.79	Graduate
7	2	39	1	9991	0	19	133	1	37	37	9	7	115	0	0	1	1	1	0	50	0	0	5	10	5	11.9	0	0	5	17	5	11.5	5	16.2	0.3	-0.92	Graduate
8	1	1	1	9500	1	1	142	1	19	38	7	10	128	1	0	0	1	0	1	18	0	0	7	9	7	13.3	0	0	8	8	8	14.3	0	15.5	2.8	-4.06	Graduate
9	1	18	4	9254	1	1	119	1	37	37	9	9	113	1	0	0	0	1	0	22	0	0	5	5	0	0	0	0	5	5	0	0	0	15.5	2.8	-4.06	Dropouts
10	1	1	3	9238	1	1	137	62	1	1	9	9	129	0	0	0	1	0	1	21	1	0	6	8	6	13.9	0	0	6	7	6	14.1	0	16.2	0.3	-0.92	Graduate
11	1	1	1	9238	1	1	138	1	1	19	4	7	123	1	0	1	0	0	0	18	0	0	6	9	5	11.4	0	0	6	14	2	13.5	0	8.9	1.4	3.51	Dropouts
12	1	1	1	9670	1	1	139	1	38	19	5	7	131	1	0	0	1	0	0	18	0	0	6	6	6	12.3	0	0	6	7	5	14.2	0	13.9	-0.3	0.79	Graduate
13	1	1	1	9500	1	1	136	1	19	38	9	9	119	1	0	0	1	0	1	18	0	0	8	8	7	13.2	0	0	8	8	7	13.2	0	12.7	3.7	-1.7	Graduate
14	1	1	2	9853	1	1	133	1	19	37	4	9	130	1	0	0	1	0	0	19	0	0	6	6	0	0	0	0	6	0	0	0	0	12.7	3.7	-1.7	Dropouts
15	1	53	1	9254	1	42	110	1	1	1	4	7	112	1	0	0	1	0	1	21	0	0	6	7	6	10.6	0	0	6	8	5	11	0	8.9	1.4	3.51	Graduate
16	1	1	1	9085	1	1	149	1	38	37	5	5	137	1	0	0	1	0	1	18	0	0	5	7	4	13.3	0	0	5	5	5	12	0	10.8	1.4	1.74	Graduate
17	1	1	1	9773	1	1	127	1	19	37	9	3	121	1	0	0	1	0	0	20	0	0	6	6	5	13.2	0	0	6	7	0	0	0	15.5	2.8	-4.06	Dropouts
18	1	18	1	9238	1	1	137	1	19	38	5	8	137	1	0	0	1	0	0	18	0	0	6	10	1	12	0	0	6	14	2	11	0	10.8	1.4	1.74	Enrolled
19	1	17	2	9500	1	1	135	1	19	1	5	4	127	1	0	0	1	0	0	18	0	0	7	8	7	13.3	0	0	8	8	8	14.5	0	15.5	2.8	-4.06	Graduate
20	1	1	1	9130	1	1	137	1	3	19	3	5	136	1	0	0	1	0	0	20	0	0	5	8	4	12.5	1	0	5	8	4	12.3	2	10.8	1.4	1.74	Graduate
21	1	1	1	9853	1	1	140	1	19	19	7	7	125	1	0	0	1	0	0	18	0	0	7	7	6	11.7	0	0	7	8	6	13.5	0	16.2	0.3	-0.92	Enrolled
22	1	1	3	171	1	1	122	1	1	1	9	8	120	0	0	0	1	0	1	21	0	0	0	0	0	0	0	0	0	0	0	0	0	11.1	0.6	2.02	Graduate
23	1	18	4	9556	1	1	127	1	1	38	4	7	122	1	0	0	1	0	0	20	0	0	7	14	7	11.4	0	0	8	9	8	11.4	0	12.7	3.7	-1.7	Enrolled
24	1	1	4	9500	1	1	142	1	19	19	1	1	126	1	0	0	1	0	0	18	0	0	8	12	7	12.9	0	0	8	12	7	12.9	0	12.7	3.7	-1.7	Graduate
25	1	1	4	9670	1	1	125	1	1	38	4	7	115	1	0	0	1	0	1	19	0	0	6	8	6	13.4	0	0	6	7	6	12.3	0	11.1	0.6	2.02	Graduate
26	1	1	1	9500	1	1	126	1	19	19	3	7	124	0	0	0	1	0	0	19	0	0	7	8	6	13.3	0	0	8	9	7	14.1	0	11.1	0.6	2.02	Graduate

Hình 1. *Một số mẫu dữ liệu từ tập dữ liệu chuẩn của UCI*

Nếu bộ dữ liệu này áp dụng cho sinh viên Việt Nam thì cần đánh trọng số thêm cho các đặc trưng do quan điểm coi trọng việc học tập, tinh thần vượt qua khó khăn của người Việt Nam khác so với các nước châu Âu. Hơn nữa, cần bổ sung thêm các mẫu dữ liệu của sinh viên Việt Nam để cải thiện kết quả dự đoán.

2.2. Ứng dụng học máy để giải quyết bài toán

Bài toán dự đoán tình trạng bỏ học của sinh viên là bài toán phân loại. Để giải bài toán phân loại, chúng ta có thể sử dụng các kỹ thuật Neural Network, Random Forest, Suport Vector Machine (SVM), Bayes Clacsifier, Logitic Regresion, Decision Tree, K-nearest neighbors của học máy. Ba kỹ thuật học máy tốt nhất được chọn gồm Neural network, Random Forest, SVM. Neural network, SVM có thể được huấn luyện cho đến khi lỗi rất nhỏ (lỗi nhỏ hơn một ngưỡng cho trước) và Random Forest là phiên bản cải tiến của Decision Tree để giải quyết tốt các trường hợp thiếu thông tin để phân loại.

Các bước xử lý của hệ thống gồm: tiền xử lý dữ liệu, trích chọn các đặc trưng quan trọng và gắn trọng số cho các đặc trưng, huấn luyện các mô hình/thuật toán học máy và dùng kết quả huấn luyện để phân loại cho các mẫu dữ liệu mới.

Tiền xử lý gồm các bước sau: chuyển các nhãn lớp từ văn bản sang số (Dropout (bỏ học) => 0, Graduate (tốt nghiệp), Enrolled (đã nhập học) => 1); chuẩn hóa giá trị của một số thuộc tính để trở thành tập dữ liệu chuẩn cung cấp cho mô hình/thuật toán học máy; giảm sự mất cân bằng dữ liệu giữa các lớp trong tập dữ liệu (Dropout có 1421 mẫu, Graduate và Enrolled có 3003 mẫu) bằng cách nhân 2 số mẫu có nhãn là Dropout trong tập dữ liệu để thu được (Dropout có 2840 mẫu, Graduate và Enrolled có 3003 mẫu).

Trích chọn các đặc trưng quan trọng được làm như sau: đo sự ảnh hưởng của mỗi đặc trưng đến kết quả phân loại; loại bỏ các đặc trưng ít quan trọng; gắn trọng số cho các đặc trưng quan trọng.

Huấn luyện các mô hình/thuật toán học máy gồm các bước sau: chia ngẫu nhiên tập dữ liệu theo tỷ lệ 80% số mẫu cho huấn luyện và 20% số mẫu cho kiểm tra; điều chỉnh các tham số của mỗi mô hình để thu được kết quả phân lớp tốt nhất; huấn luyện mô hình/thuật toán học máy với tập mẫu huấn luyện ứng với từng bộ tham số. Kết quả huấn luyện sẽ khác nhau đối với mỗi mô hình/thuật toán học máy. Với SVM, tập các hệ số gắn với mỗi đặc trưng được đưa ra. Với Random Forest, cây phân loại được hình thành. Với Neural Network, các trọng số thể hiện liên kết giữa các nơ-ron được trình bày.

Phân loại cho các mẫu dữ liệu mới được dùng cho hai trường hợp. Đầu tiên, dùng các mẫu dữ liệu trong tập kiểm tra để mô hình/thuật toán học máy phân loại. Kết quả thu được dùng để đánh giá hiệu quả của mô hình/thuật toán học máy. Vì vậy, trường hợp này giúp chọn được một mô

hình/thuật toán học máy tốt nhất cho bài toán. Trường hợp thứ hai là dự đoán nhãn lớp cho các mẫu dữ liệu mới mà có thể chưa có trong tập kiểm tra bằng mô hình/thuật toán học máy tốt nhất đã chọn được từ tập dữ liệu.

Để các mô hình/thuật toán học máy hoạt động, các tham số cần được thiết lập cho cả bước huấn luyện và kiểm tra. Các giá trị phù hợp của các tham số này được lựa chọn để đưa vào các thử nghiệm.

3. Kết quả thực nghiệm

Các thực nghiệm được viết bằng ngôn ngữ Python. Với Neural Network, một mạng nơ-ron nhiều tầng có lan truyền ngược và một mạng nơ-ron với deep learning được dùng. Tham số cho các mạng là số nơ-ron ở mỗi tầng, số lượng tầng trong mạng, tốc độ học, dạng hàm chuyển. Với Random Forest, dùng tham số là số cây trong rừng. Với SVM, thiết lập hàm nhân và hệ số C để thử nghiệm. Kết quả thực nghiệm cho từng mô hình/thuật toán học máy được làm với các bộ tham số khác nhau. Tỷ lệ chia dữ liệu là 80:20 trên tập dữ liệu đã cân bằng. Số mẫu huấn luyện là 4676 và số mẫu kiểm tra là 1169.

3.1. Kết quả phân loại của từng mô hình học máy

3.1.1. Kết quả của Neural Network

Kết quả thử nghiệm với Neural Network được thể hiện trong Bảng 1. Số nơ-ron ở tầng vào là 36 ứng với 36 đặc trưng. Số nơ-ron ở tầng ra là 2 ứng với hai nhãn lớp Dropout và lớp Graduate. Mạng nhiều tầng có lan truyền ngược và mạng học sâu đều có 2 lớp ẩn, mỗi lớp có 100 nơ-ron, dùng hàm chuyển là hàm relu và có số lần lặp là 200. Các trọng số của mạng được khởi tạo là các số ngẫu nhiên với tham số là 42.

Bảng 1. *Kết quả thử nghiệm với mạng nơ-ron*

Loại mạng	Nhiều tầng có lan truyền ngược	Dùng học sâu v1	Dùng học sâu v2
Tham số trong kiểm tra	solver = 'adam', alpha = 0,0001	solver = 'adam', alpha = 0,001	solver = 'adam', alpha = 0,00000001
Độ chính xác	**0,7793**	**0,7767**	**0,8178**
Ma trận confusion	[519 53] [205 392]	[348 224] [37 560]	[429 143] [70 527]

Bảng 1 cho thấy mạng học sâu version 2 với tham số alpha = 0,00000001 cho kết quả phân lớp tốt nhất đạt 0,8178. Theo ma trận confusion, mạng nhiều tầng có lan truyền ngược phân lớp tốt với mẫu thuộc lớp Dropout và phân lớp kém với lớp còn lại. Mạng với học sâu version 1 nhận dạng tốt lớp Graduate. Mạng với học sâu version 1 nhận dạng các mẫu ở lớp Graduate tốt hơn mạng nhiều tầng có lan truyền ngược nhưng nhận dạng các mẫu ở lớp Dropout kém hơn mạng nhiều tầng có lan truyền ngược. Tổng số mẫu bị nhận dạng sai của mạng học sâu version 2 là ít nhất.

3.1.2. Kết quả của Random Forest

Với Random Forest, số lượng cây con trong rừng được thay đổi gồm 100 cây và 200 cây. Cách lấy mẫu để xây dựng cây thay đổi (chọn hai giá trị gồm 42 và 0). Thuật toán có chọn đặc trưng quan trọng cho việc phân lớp. Vì vậy, độ chính xác của Random Forest cao hơn so với mạng nơ-ron. Kết quả phân lớp của Random Forest với ba version được thể hiện trong Bảng 2.

Bảng 2. *Kết quả thử nghiệm với Random Forest*

Các version	Random Forest v1	Random Forest v2	Random Forest v3
Tham số trong kiểm tra	n_estimators = 100, random_state = 0	n_estimators = 200, random_state = 42	n_estimators = 200, random_state = 0
Độ chính xác	**0,9133**	**0,9133**	**0,9133**
Ma trận confusion	[773 60] [92 829]	[773 60] [92 829]	[773 60] [92 829]

Do Random Forest đã tối ưu hóa việc chọn đặc trưng tốt nhất và bầu chọn theo số đông cho các mẫu chưa đủ thông tin phân lớp nên các version đều cho độ chính xác là 0,9133. Dữ liệu của ma trận confusion cho thấy lớp Dropout nhận diện đúng 773 mẫu/833 mẫu và lớp Graduate nhận diện đúng 829 mẫu/921 mẫu.

3.1.3. Kết quả của SVM

Đối với SVM, thực hiện chọn hai chế độ gồm SVM tuyến tính và SVM có dùng hàm nhân. Các mẫu trong tập dữ liệu được chuẩn hóa để phù hợp với các dữ liệu vào của SVM. Kết quả phân lớp của SVM tuyến tính với ba version được thể hiện trong Bảng 3. Bảng 4 trình bày kết quả phân lớp của SVM với hai nhân poly và signmoid.

Với SVM tuyến tính, Bảng 3 cho thấy version 3 có tham số C = 100 cho độ chính xác cao nhất là 0,874 với khả năng phân lớp tốt và cân đối cho cả hai lớp. Theo ma trận confusion, lớp Dropout nhận diện đúng 753/853 mẫu và lớp Graduate nhận diện đúng 780/900 mẫu.

Bảng 3. *Kết quả thử nghiệm với SVM tuyến tính*

Các version	SVM tuyến tính v1	SVM tuyến tính v2	SVM tuyến tính v3
Tham số trong kiểm tra	C = 10,0	C = 1,0	C = 100,0
Độ chính xác	**0,8706**	**0,8592**	**0,8740**
Ma trận confusion	[728 126] [101 799]	[699 155] [92 808]	[753 101] [120 780]

Bảng 4. *Kết quả thử nghiệm với SVM có dùng hàm nhân*

Các version	SVM có nhân v1	SVM có nhân v2	SVM có nhân v3
Tham số trong kiểm tra	kernel = 'poly', C = 100,0	kernel = 'poly', C = 10,0	kernel = 'sigmoid', C = 100,0
Độ chính xác	**0,8626**	**0,8683**	**0,7537**
Ma trận confusion	[748 106] [135 765]	[716 138] [93 807]	[637 217] [215 685]

Bảng 4 thể hiện SVM có dùng hàm nhân ở version 2 cho độ chính xác cao nhất 0,8683. Dữ liệu của ma trận confusion cho thấy lớp Dropout nhận diện đúng 716/854 mẫu và lớp Graduate nhận diện đúng 807/900 mẫu.

3.2. So sánh kết quả của các kỹ thuật học máy

Kết quả phân loại tốt nhất của từng kỹ thuật học máy được tổng hợp trong Bảng 5. Dữ liệu từ Bảng 5 cho thấy mô hình Random Forest là tốt nhất cho bài toán. Độ chính xác của Random Forest cao hơn mô hình tốt thứ hai (SVM tuyến tính v3) là 0,0393. Nghĩa là, tỷ lệ phân lớp đúng cao hơn khoảng 4% so với SVM tuyến tính v3.

Bảng 5. *So sánh kết quả phân loại của các mô hình học máy*

Các mô hình	SVM tuyến tính v3	SVM có nhân v2	Random Forest v1	Mạng nơ-ron học sâu v2
Tham số trong kiểm tra	C = 100,0	kernel = 'poly', C = 10,0	n_estimators = 100, random_state = 0	solver = 'adam', alpha = 0,00000001
Độ chính xác	**0,8740**	**0,8683**	**0,9133**	**0,8178**

Bảng 6 thể hiện độ chính xác của các mô hình/ thuật toán học máy dùng cho tập dữ chuẩn của UCI. Số liệu được lấy từ kết quả nghiên cứu trong [2] - [4] và Random Forest trong Phần 3.1. Dữ liệu từ Bảng 6 cho thấy độ chính xác của Random Forest cao nhất.

Bảng 6. *So sánh kết quả phân loại với các nghiên cứu đã công bố*

Các mô hình	Random Forest	XGBoost [2]	LightGBM, CatBoost [3]	Stacking Classifier [4]
Độ chính xác	91%	88%	87%	89%

Bảng 5 và Bảng 6 cho thấy Random Forest là kỹ thuật tốt nhất cho bài toán dự đoán sinh viên bỏ học. Random Forest cho kết quả phân lớp cao nhất do Random Forest đã giảm số đặc trưng

không quan trọng khi đưa vào phân lớp. Hơn nữa, Random Forest đã tổng hợp kết quả phân lớp của các cây quyết định con theo nguyên tắc bầu cử theo số đông.

4. Kết luận

Bài báo này trình bày việc ứng dụng các mô hình/thuật toán học máy vào dự đoán tình trạng bỏ học của sinh viên đại học. Đầu tiên, các mô tả chi tiết về tập dữ liệu chuẩn dùng cho bài toán được mô tả. Các đặc trưng được giải thích chi tiết. Các bước tiền xử lý dữ liệu phù hợp được áp dụng để dữ liệu đúng với định dạng của các dữ liệu vào của các mô hình phân lớp. Các mô hình học máy tốt nhất gồm Neural network, Random Forest, SVM đã được áp dụng cho bài toán. Các tham số quan trọng của mỗi mô hình đã được điều chỉnh để tìm ra bộ tham số phù hợp giúp cho kết quả phân lớp đạt độ chính xác cao. Kết quả thực nghiệm trên tập dữ liệu chuẩn cho thấy mô hình Random Forest là thích hợp nhất cho bài toán.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education, " *Technology in Society*, vol. 76, pp. 1-10, 2024.

[2] A. Ridwana and A. M. Priyatnob, "Predict Students' Dropout and Academic Success with XGBoost," *Journal of Education and Computer Applications*, vol. 1, no. 2, pp. 1-8, 2024

[3] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 2, pp. 1-24, 2024.

[4] D. Arora, "Predicting Students Academic Success and Dropout Using Supervised Machine Learning," *International Journal of Scientific Study*, vol. 11, no. 6, pp. 72-78, 2023.

[5] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 146, pp. 1-17, 2022.

[6] T. Purwoningsih, H. B. Santoso, K. A. Puspitasari, and Z. A. Hasibuan, "Early Prediction of Students' Academic Achievement: Categorical Data from Fully Online Learning on Machine-Learning Classification Algorithms," *Journal of Hunan University* (Natural Sciences), vol. 48, no. 9, pp. 131-141, 2021.

[7] L. U. M. Huynh, T. T. Pham, and V. N. Nguyen, "Predicting students' ability to graduate on time: a case study at Dong Thap University," (in Vietnamese), *Vietnam Journal of Education*, vol. 24, no. 1, pp. 48-53, 2024.

[8] H. S. Luu, T. D. Tran, T. H. Nguyen, and T. N. Nguyen, "Predicting learning outcomes using deep learning techniques with multi-layered neural networks," (in Vietnamese), *Journal of Science, Can Tho University*, vol. 56, no. 3A, pp. 20-28, 2020.

[9] L. T. N. Huynh and T. N. Nguyen, "Student learning outcome prediction system using open source recommender system library MYMEDIALITE," (in Vietnamese), *National Conference of Information Technology*, 2013, pp. 20-28.