VJE Tạp chí Giáo dục (2024), 24(16), 53-58 ISSN: 2354-0753

KHAI PHÁ DỮ LIỆU NGƯỜI HỌC HỖ TRỢ CÔNG TÁC QUẢN LÍ ĐÀO TẠO VÀ TƯ VẤN: NGHIÊN CỨU TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

Đặng Ngọc Hùng+, Hoàng Anh, Mai Văn Thanh, Vũ Quý Trung, Bùi Thanh Huyền

Trung tâm Đảm bảo chất lượng - Trường Đại học Công nghiệp Hà Nội
+Tác giả liên hệ ● Email: dangngochung@haui.edu.vn

Article history
Received: 31/5/2024
Accepted: 11/6/2024
Published: 20/8/2024

Keywords
On-time graduation, graduation classification, data mining, Hanoi University of Industry

ABSTRACT
Hanoi University of Industry is using an electronic university system in management and training, with a huge source of data collected and synthesized in recent times; however, the analysis and mining of this data has not been done yet. Therefore, conducting research on learner data at Hanoi University of Industry to support training management and student consulting is meaningful and necessary. In this study, the first research model incorporates demographic data and 4-semester academic results, predicting an on-time graduation rate of 88% and a graduation classification prediction of 89%. The second model includes learning outcome data of 6 semesters, resulting in an increased prediction rate. It projected an on-time graduation rate of 90% and a graduation classification prediction of 95%. The research findings proposed several recommendations aimed at improving on-time graduation rates and achieving good and excellent graduation results for the University and its students; providing students with personalized learning that takes into account students' skills, goals and interests to create their ideal learning journey.

1. Mở đầu

Trong những năm gần đây, việc ứng dụng các kĩ thuật khai phá dữ liệu và phát hiện tri thức trong các lĩnh vực giáo dục, tài chính, ngân hàng, marketing… rất được quan tâm nghiên cứu. Đối với GD-ĐT sinh viên trong các trường đại học theo hình thức đào tạo tín chỉ đòi hỏi sinh viên phải có sự chủ động cao. Sinh viên trong quá trình học tập phải tự mình lựa chọn, phân bổ các môn học cho từng kì sao cho tích lũy đủ số tín chỉ theo quy chế đào tạo, sinh viên hoàn toàn có thể ra trường sớm hoặc đúng hạn.

Khai phá dữ liệu là trích xuất và khai thác những thông tin hữu ích, tiềm ẩn của dữ liệu. Công việc này giải quyết các vấn đề bằng cách phân tích lượng dữ liệu lớn hiện có để khám phá ra các xu hướng và các quy tắc có ý nghĩa (Baradwaj & Pal, 2011). Theo Nguyễn Trí Thành và cộng sự (2013), khai phá dữ liệu là một bước trong quy trình khai phá tri thức trong cơ sở dữ liệu được tiến hành qua 6 giai đoạn sau: *Gom dữ liệu; Trích lọc dữ liệu; Làm sạch, tiền xử lí dữ liệu; Chuyển đổi dữ liệu; Khai phá dữ liệu; Đánh giá luật và biểu diễn tri thức.*

Nghiên cứu tổng quan chung về phân tích dữ liệu trong giáo dục đại học của Mai (2018) cho rằng, phân tích dữ liệu giáo dục được xem là một giải pháp khá hữu hiệu. Bài viết được thực hiện nhằm đưa ra cái nhìn bao quát về dữ liệu lớn trong ngữ cảnh của giáo dục đại học, đặc biệt tại Việt Nam như là một mối quan tâm mới xuất hiện gần đây trong các nghiên cứu liên ngành. Bài báo xác định rõ các loại dữ liệu cần được thu thập, lưu trữ và phân tích tại các cơ sở giáo dục đại học, qua đó nêu lên những vấn đề tồn tại và gợi ý những định hướng ban đầu cho việc ứng dụng lĩnh vực nghiên cứu này vào thực tiễn giáo dục đại học.

Một số nghiên cứu sử dụng các ứng dụng khai phá dữ liệu để hỗ trợ cho công tác quản lí trong các cơ sở đào tạo, như nghiên cứu của Đinh Chung Dũng (2017), Romero & Ventura (2020) đã tổng quan các nghiên cứu với tiêu đề "khai thác dữ liệu trong giáo dục", về cách áp dụng Phân tích học tập và Khai thác dữ liệu giáo dục trên dữ liệu giáo dục. Yağcı (2022) đề xuất một mô hình mới dựa trên thuật toán học máy như Random Forests, Nearest Neighbour, Support Vector Machines, Logistic Regression, Naïve Bayes và K-Nearest Neighbour Algorithms để dự đoán điểm thi cuối kì của sinh viên đại học, lấy điểm thi giữa kì làm dữ liệu nguồn. Mai Thu Giang (2020) đã ứng dụng khai phá dữ liệu trong dự báo kết quả học tập tại Trường Đại học Kinh tế - Đại học Huế bằng việc sử dụng kĩ thuật trích chọn thuộc tính và kĩ thuật phân lớp dựa trên các thuật toán Cây quyết định (Decision Tree) trong phần mềm WEKA

(Waikato Environment for Knowledge Analysis) để xây dựng các mô hình dự báo kết quả cuối khóa sau khi kết thúc từng học kì. Các nghiên cứu Nguyễn Xuân Hải (2016), Nguyễn Văn Thủy (2023), Huỳnh Lê Uyên Minh và cộng sự (2024), đã sử dụng các công cụ máy học trong nghiên cứu kết quả tốt nghiệp, nâng cao chất lượng học tập và hiệu quả đào tạo của nhà trường.

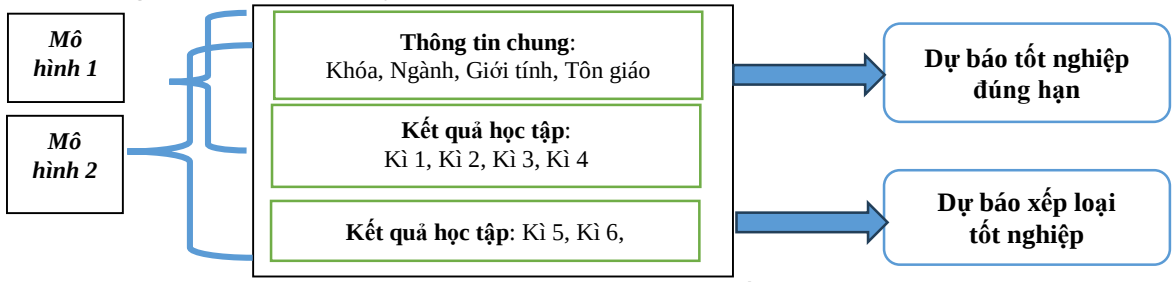
Như vậy trên cơ sở nghiên cứu trong và ngoài nước, chúng tôi nhận thấy khai phá dữ liệu trong lĩnh vực GD-ĐT đang rất được quan tâm và mang lại nhiều ý nghĩa. Trường Đại học Công nghiệp Hà Nội đang sử dụng hệ thống đại học điện tử trong công tác quản lí và đào tạo, với nguồn dữ liệu rất lớn được thu thập và tổng hợp trong thời gian qua, tuy nhiên việc phân tích và khai phá dữ liệu này chưa được thực hiện nhiều. Do đó mục tiêu của nghiên cứu này là khai phá dữ liệu người học tại Trường nhằm hỗ trợ cho công tác quản lí đào tạo và tư vấn sinh viên, tập trung vào dự báo tốt nghiệp đúng hạn và xếp loại tốt nghiệp.

2. Kết quả nghiên cứu

2.1. Phương pháp nghiên cứu

Học máy (Machine Learning) là một phương tiện trong trí tuệ nhân tạo, sử dụng các thuật toán cho phép máy tính có thể tự học từ dữ liệu để giải quyết những vấn đề cụ thể như làm cho máy tính có khả năng nhận thức cơ bản của con người (nghe, nhìn, hiểu, giải toán,…) và hỗ trợ con người xử lí một lượng thông tin khổng lồ phải đối diện hằng ngày. Học máy đóng một vai trò quan trọng trong nhiều ngành khoa học và các ứng dụng của nó là một phần trong cuộc sống hằng ngày. Học máy được sử dụng để lọc thư rác điện tử, để dự đoán thời tiết, trong chẩn đoán y tế, khuyến cáo sản phẩm, nhận diện khuôn mặt, phát hiện gian lận thẻ tín dụng, dự báo kiệt quệ tài chính hay phá sản doanh nghiệp. Trong nghiên cứu này chúng tôi sử dụng một số thuật toán sử dụng phổ biến để dự báo kiệt quệ tài chính như: Rừng ngẫu nhiên (Random Forrest); Máy vec-tơ hỗ trợ (Support Vector Machine); K láng giềng gần nhất (K-nearest neighbor); Hồi quy Logistic (Logistic Regression) và Cây quyết định (Decision tree).

Mô hình nghiên cứu được trình bày tại hình 1.



Hình 1. Mô hình nghiên cứu dự báo tốt nghiệp

Hầu hết các bộ dữ liệu đều khó đạt được trạng thái cân bằng mà luôn có sự khác biệt về tỉ lệ giữa 2 classes. Đối với những trường hợp dữ liệu mất cân bằng nhẹ như tỉ lệ 60:40 thì sẽ không ảnh hưởng đáng kể tới khả năng dự báo của mô hình, do đó chúng tôi sẽ sử dụng thước đo độ chính xác (Accuracy). Tuy nhiên nếu hiện tượng mất cân bằng nghiêm trọng xảy ra, chẳng hạn như tỉ lệ 90:10 sẽ thường dẫn tới ngộ nhận chất lượng mô hình. Khi đó thước đo đánh giá mô hình là độ chính xác (Accuracy) có thể đạt được rất cao mà không cần tới mô hình. Ví dụ, một dự báo ngẫu nhiên đưa ra tất cả đều là nhóm đa số thì độ chính xác đã đạt được là 90%. Do đó việc sử dụng độ chính xác làm thước đo đánh giá mô hình thường không hiệu quả bởi hầu hết chúng đều đạt độ chính xác rất cao.. Một mô hình ngẫu nhiên dự báo toàn bộ là nhãn thuộc nhóm đa số cũng sẽ mang lại kết quả gần bằng 100%. Khi đó chúng tôi sẽ tiếp tục cân nhắc tới một số chỉ số thay thế như Precision, Recall, F1-score,… Các chỉ số này sẽ không quá lớn để dẫn tới ngộ nhận độ chính xác, đồng thời chúng tập trung hơn vào việc đánh giá độ chính xác trên nhóm thiểu số, nhóm mà chúng ta muốn dự báo chính xác hơn so với nhóm đa số. Một mô hình có các chỉ số trên đều cao thì mô hình đó có chất lượng dự báo càng tốt. Trong bài này, chúng tôi sẽ sử dụng chỉ số Accuracy, Precision, Recall, F1 - score và AUC là thước đo đánh giá mô hình.

Trên cơ sở dữ liệu nghiên cứu, tiếp theo để huấn luyện, lựa chọn và kiểm tra kết quả của mô hình chúng ta sẽ phân chia một cách ngẫu nhiên, không trùng lặp bộ dữ liệu thành các tập train/test. Các bộ dữ liệu này có ý nghĩa và vai trò như sau: + Tập train: Dựa trên các biến input và target của tập train, ta sẽ huấn luyện mô hình phân loại tốt nghiệp. Mô hình thu được sẽ được đánh giá ở những tập dữ liệu độc lập khác như tập test; + Tập test: Đây cũng là tập dữ liệu có các trường giống tập train được coi như những quan sát mới hoàn toàn. Tập test nên có phân phối giống nhất với dữ liệu thực tế mà người dùng sẽ tạo ra để đánh giá khả năng áp dụng mô hình vào thực tiễn.

2.2. Dữ liệu nghiên cứu

Nghiên cứu này sử dụng dữ liệu người học được thu thập từ hệ thống Đại học điện tử của nhà trường, đối tượng là sinh viên chính quy tốt nghiệp giai đoạn từ 2017-2023 (tương ứng từ Đại học khóa 9 đến Đại học khóa 15), mốc thời gian lấy dữ liệu là ngày 31/12/2023.

Dữ liệu sử dụng trong nghiên cứu, được trích xuất từ hệ thống Đại học điện tử của Trường Đại học Công nghiệp Hà Nội, các trường dữ liệu được trích xuất ngoài các thông tin nhân khẩu học của sinh viên, các trường dữ liệu trong mô hình nghiên cứu đước trích xuất ra Excel, loại bỏ những sinh viên thôi học, chỉ xem xét những sinh viên đã tốt nghiệp trong giai đoạn 2017-2023. Nhóm tác giả sử dụng ngôn ngữ lập trình Python, để xử lí và chuyển đổi dữ liệu sang dạng định tính và định lượng thông qua hàm chuyển đổi, sau đó sử dụng các thư viện của Python như sklearn.preprocessing, sklearn.impute, sklearn.compose chuyển đổi dữ liệu. Sau đó chia dữ liệu thành 2 phần với tỉ lệ 85%:15%, trong đó 85% là để train và 15% là dữ liệu test. Tiếp theo các tác giả sử dụng thư viện sklearn, với các thuật toán Random Forrest; Máy vec-tơ hỗ trợ (Support Vector Machine); K láng giềng gần nhất (K-nearest neighbor); Hồi quy Logistic (Logistic Regression) và Cây quyết định (Decision tree) để thực hiện hồi quy và dự báo. Cuối cùng sử dụng sklearn.metrics để lập bảng dự báo về thôi học và tốt nghiệp.

Trong thuật toán Cây quyết định (Decision Tree), khi xây dựng cây quyết định nếu để độ sâu tùy ý thì cây sẽ phân loại đúng hết các dữ liệu trong tập training dẫn đến mô hình có thể dự đoán tệ trên tập train/test. Thuật toán Random Forest gồm nhiều cây quyết định, mỗi cây quyết định đều có những yếu tố ngẫu nhiên: Lấy ngẫu nhiên dữ liệu để xây dựng cây quyết định; Lấy ngẫu nhiên các thuộc tính để xây dựng cây quyết định. Do mỗi cây quyết định trong thuật toán Random Forest không dùng tất cả dữ liệu training, cũng như không dùng tất cả các thuộc tính của dữ liệu để xây dựng cây nên mỗi cây có thể sẽ dự đoán không tốt, khi đó mỗi mô hình cây quyết định không bị overfitting mà có thể bị underfitting, hay nói cách khác là mô hình có high bias. Tuy nhiên, kết quả cuối cùng của thuật toán Random Forest lại tổng hợp từ nhiều cây quyết định, thế nên thông tin từ các cây sẽ bổ sung thông tin cho nhau, dẫn đến mô hình có low bias và low variance, hay mô hình có kết quả dự đoán tốt.

Bảng 1 trình bày dữ liệu bao gồm 34.400 sinh viên đã tốt nghiệp (trong đó Nam là 18.898, chiếm tỉ lệ 54,94%, Nữ là 15.502, chiếm tỉ lệ 45,06%), tỉ lệ sinh viên tốt nghiệp giỏi và xuất sắc là 19,06%, tỉ lệ sinh viên chiếm tỉ lệ cao nhất là sinh viên tốt nghiệp loại khá là 69,8%.

Trong số sinh viên tốt nghiệp có 29.385 tốt nghiệp đúng hạn, chiếm tỉ lệ 85,42%, còn lại là tốt nghiệp chậm, chiếm tỉ lệ 14,58%. Trong dữ liệu của Đại học khóa 15, có 290 sinh viên thuộc diện tốt nghiệp sớm.

Bảng 2 trình bày dữ liệu sinh viên tốt nghiệp theo tôn giáo, trong đó phần lớn là không theo tôn giáo chiếm tỉ lệ 96,51%, kế đến là theo phật giáo, công giáo và những tôn giáo khác. Bảng 3 trình bày tổng hợp thống kê mô tả kết quả học tập của 6 kì của những sinh viên đã tốt nghiệp, với thang điểm 4, kết quả học tập của từng kì có xu hướng tăng lên.

Bảng 1. Dữ liệu nghiên cứu

Khóa	Xếp loại tốt nghiệp				Tổng cộng
	Trung bình	Khá	Giỏi	Xuất sắc	
ĐH K9	1.144	4.263	490	22	***5.919***
	19,33	72,02	8,28	0,37	***100***
ĐH K10	1.152	4.229	794	40	***6.215***
	18,54	68,05	12,78	0,64	***100***
ĐH K11	711	3.537	890	91	***5.229***
	13,6	67,64	17,02	1,74	***100***
ĐH K12	505	3.810	773	48	***5.136***
	9,83	74,18	15,05	0,93	***100***
ĐH K13	229	4.383	1.313	71	***5.996***
	3,82	73,1	21,9	1,18	***100***
ĐH K14	87	3.601	1.761	166	***5.615***
	1,55	64,13	31,36	2,96	***100***
ĐH K15	4	188	91	7	***290***
	1,38	64,83	31,38	2,41	***100***
Tổng cộng	***3.832***	***24.011***	***6.112***	***445***	***34.400***
Tỉ lệ (%)	***11,14***	***69,8***	***17,77***	***1,29***	***100***

Bảng 2. Dữ liệu nghiên cứu sinh viên tốt nghiệp phân loại theo tôn giáo

Tôn giáo	Số lượng	Tỉ lệ (%)
Không tôn giáo	33.200	96,51
Công giáo	477	1,39
Phật giáo	713	2,07
Tôn giáo khác	10	0,03

Bảng 3. Tổng hợp thống kê mô tả kết quả học tập 6 kì

Kì	Số quan sát	Giá trị trung bình	Độ lệch chuẩn	Thấp nhất	Cao nhất
KQHTTLKY1	34400	2,30	0,72	0,00	4,00
KQHTTLKY2	34400	2,41	0,67	0,00	4,00
KQHTTLKY3	34400	2,40	0,65	0,00	4,00
KQHTTLKY4	34400	2,60	0,65	0,00	4,00
KQHTTLKY5	34400	2,74	0,64	0,00	4,00
KQHTTLKY6	34400	2,69	0,55	0,00	4,00

2.3. Kết quả phân tích và thảo luận

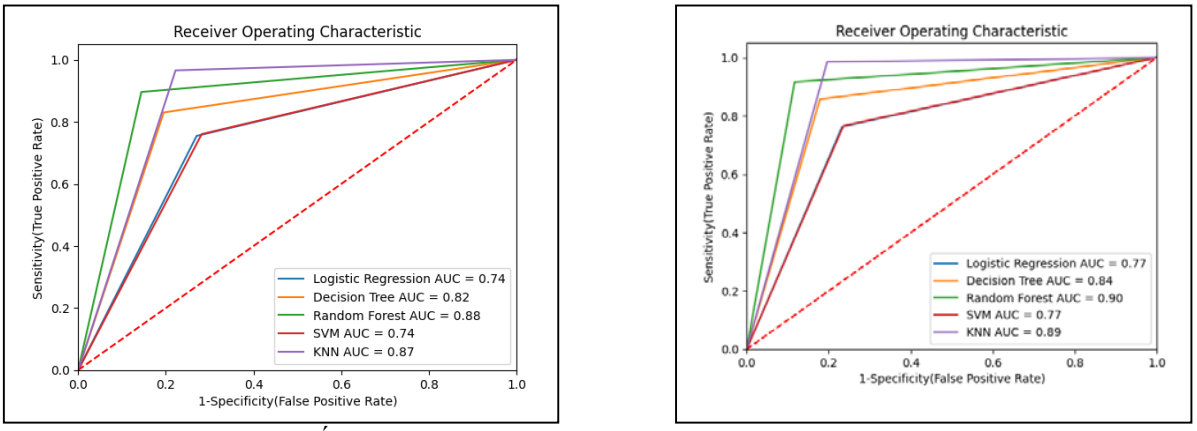
2.3.1. Dự báo tốt nghiệp đúng hạn

Trong dữ liệu nghiên cứu (có 85,42% về tốt nghiệp đúng hạn), để xử lí dữ liệu mất cân bằng về tỉ lệ tốt nghiệp đúng hạn và tỉ lệ tốt nghiệp chậm, chúng tôi sử dụng các kĩ thuật SMOTE (Synthetic Minority Over-sampling). Tại bảng 5 cho thấy kết quả xem xét độ chính xác của các mô hình dự báo (Accuracy), các thuật toán Logistic Regression, SVM, Decision Tree, Random Forest, KNN, Bayes ở mô hình 1 (gồm dữ liệu chung và dữ liệu kết quả học tập của 4 kì), trong 5 thuật toán được sử dụng để dự báo sinh viên tốt nghiệp đúng hạn thì thuật toán Random Forest đạt độ chính xác cao nhất với tỉ lệ 88%. Bên cạnh đo lường độ chính xác của các mô hình (Accuracy), chúng tôi sử dụng các thước đo khác để xem xét một cách toàn diện và đầy đủ hơn như theo các thước đo Precision, Recall, F1 - score (bảng 4).

Bảng 4. Kết quả dự báo chính xác của từng thuật toán theo mô hình 1

STT	Method	Accuracy	Precision	Recall	F1-score
1	Logistic Regression	0,74	0,74	0,74	0,74
2	SVM	0,74	0,74	0,74	0,74
3	Decision Tree	0,82	0,82	0,82	0,82
4	Random Forest	**0,88**	0,88	0,88	0,88
5	KNN	0,87	0,89	0,87	0,87

Đối với mô hình 2, (gồm dữ liệu chung và dữ liệu kết quả học tập của 6 kì), Bảng 5 trình bày kết quả dự báo có độ chính xác cao hơn, tuy nhiên mức tăng không đáng kể, mức cao nhất là thuật toán Random Forest, với tỉ lệ dự báo chính xác là 90%.



Hình 2. Kết quả dự báo theo đường cong ROC theo mô hình 1 và mô hình 2

Chỉ số AUC (Area Under Curve) đo lường phần diện tích nằm dưới đường cong ROC cho biết khả năng phân loại của nhóm Tốt nghiệp đúng hạn/Tốt nghiệp chậm của các thuật toán đã trình bày ở trên là mạnh hay yếu. AUC ∈ [0, 1], giá trị của nó càng lớn thì mô hình càng tốt. Theo hình 3 thuật toán Random Forest đạt tỉ lệ dự báo chính xác cao, AUC = 0.9. Như vậy cho thấy khả năng dự báo của mô hình tốt và có thể áp dụng mô hình vào thực tiễn. Kết quả nghiên cứu tương đồng và có độ chính xác hơn so với nghiên cứu của Mai (2018), Nguyễn Văn Thủy (2023) và Huỳnh Lê Uyên Minh và cộng sự (2024).

Bảng 5. Kết quả dự báo chính xác của từng thuật toán theo mô hình 2

STT	Method	Accuracy	Precision	Recall	F1-score
1	Logistic Regression	0,77	0,77	0,77	0,77
2	SVM	0,77	0,77	0,77	0,77
3	Decision Tree	0,84	0,84	0,84	0,84
4	Random Forest	**0,90**	0,90	0,90	0,90
5	KNN	0,89	0,91	0,89	0,89

2.3.2. Dự báo xếp loại tốt nghiệp

Tại bảng 2, dữ liệu nghiên cứu đã trình bày kết quả xếp loại tốt nghiệp của sinh viên, trong đó tỉ lệ tốt nghiệp xếp loại khá chiếm tỉ lệ cao nhất là 69,8%, kế tiếp là loại giỏi chiếm tỉ lệ là 17,77%, loại trung bình là 11,14% và thấp nhất là tốt nghiệp xuất sắc với tỉ lệ là 1,29%. Như vậy dữ liệu về xếp loại tốt nghiệp được chia thành 4 mức, với tỉ lệ không cân bằng, để làm rõ hơn về mức độ chính xác của dự báo đối với dữ liệu không cân bằng chúng tôi tiến hành

sử dụng 3 thuật toán có hiệu quả cao hơn đã được sử dụng ở nội dung trên là các thuật toán Decision Tree, Random Forest và KNN.

Bảng 6 trình bày kết quả dự báo cho thấy thuật toán Random Forest có mức độ dự báo chính xác trung bình đạt 85%, tuy nhiên kết quả của dự báo theo các tiêu chí Precision, Recall, F1-score đối với mức xếp loại tốt nghiệp trung bình, giỏi và xuất sắc thì có tỉ lệ dự báo đạt mức thấp trong khi dự báo mức xếp loại tốt nghiệp loại khá đạt tỉ lệ cao.

Bảng 7 trình bày kết quả dự báo sau khi sử dụng kĩ thuật SMOTE, xử lí dữ liệu mất cân bằng thuật toán Random Forest có mức độ dự báo chính xác (Accuracy) trung bình đạt 94%, đồng thời dự báo theo các tiêu chí Precision, Recall, F1-score đạt tỉ lệ cũng ở mức cao, thậm chí tỉ lệ dự báo sinh viên tốt nghiệp xuất sắc đạt 100% theo tiêu chí Recall.

Bảng 6. Kết quả dự báo phân loại tốt nghiệp theo mô hình 1 chưa xử lí mất cân bằng

Method	Xếp loại tốt nghiệp	Accuracy	Precision	Recall	F1-score
Decision Tree	Trung bình	***0,79***	0,54	0,58	0,56
	Khá		0,86	0,85	0,85
	Giỏi		0,68	0,7	0,69
	Xuất sắc		0,58	0,57	0,58
Random Forest	Trung bình	***0,85***	0,74	0,57	0,65
	Khá		0,87	0,93	0,9
	Giỏi		0,79	0,73	0,76
	Xuất sắc		0,73	0,4	0,52
KNN	Trung bình	***0,83***	0,67	0,63	0,65
	Khá		0,87	0,9	0,89
	Giỏi		0,76	0,69	0,72
	Xuất sắc		0,68	0,37	0,48

Bảng 7. Kết quả dự báo phân loại tốt nghiệp theo mô hình 1 đã xử lí mất cân bằng

Method	Xếp loại tốt nghiệp	Accuracy	Precision	Recall	F1-score
Decision Tree	Trung bình	***0,89***	0,89	0,91	0,9
	Khá		0,82	0,79	0,8
	Giỏi		0,88	0,89	0,89
	Xuất sắc		0,98	0,98	0,98
Random Forest	Trung bình	***0,94***	0,92	0,96	0,94
	Khá		0,9	0,84	0,87
	Giỏi		0,93	0,9	0,94
	Xuất sắc		0,99	1	0,99
KNN	Trung bình	***0,93***	0,88	0,99	0,93
	Khá		0,96	0,76	0,85
	Giỏi		0,9	0,97	0,94
	Xuất sắc		0,99	1	0,99

Đối với mô hình 2, khi bổ sung kết quả học tập của kì 5 và kì 6, kết quả dự báo loại tốt nghiệp có được cải thiện, nhưng không đáng kể, mức độ dự báo chính xác tổng thể đạt 95%, kết quả được trình bày chi tiết tại bảng 8 và bảng 9. Kết quả nghiên cứu này cũng phù hợp với kết quả nghiên cứu của Lê Quốc Tiến và Đặng Hoàng Anh (2019).

Bảng 8. Kết quả dự báo phân loại tốt nghiệp theo mô hình 2 chưa xử lí mất cân bằng

Method	Xếp loại tốt nghiệp	Accuracy	Precision	Recall	F1-score
Decision Tree	Trung bình	***0,84***	0,63	0,65	0,64
	Khá		0,89	0,89	0,89
	Giỏi		0,77	0,77	0,77
	Xuất sắc		0,64	0,63	0,63
Random Forest	Trung bình	***0,89***	0,81	0,67	0,73
	Khá		0,91	0,95	0,93
	Giỏi		0,86	0,82	0,84
	Xuất sắc		0,88	0,52	0,65
KNN	Trung bình	***0,78***	0,54	0,57	0,55
	Khá		0,86	0,84	0,85
	Giỏi		0,67	0,68	0,68
	Xuất sắc		0,53	0,57	0,55

Bảng 9. Kết quả dự báo phân loại tốt nghiệp theo mô hình 2 đã xử lí mất cân bằng

Method	Xếp loại tốt nghiệp	Accuracy	Precision	Recall	F1-score
Decision Tree	Trung bình	***0,92***	0,91	0,93	0,92
	Khá		0,87	0,84	0,85
	Giỏi		0,92	0,93	0,93
	Xuất sắc		0,98	0,99	0,99
Random Forest	Trung bình	***0,95***	0,93	0,97	0,95
	Khá		0,93	0,89	0,91
	Giỏi		0,95	0,96	0,96
	Xuất sắc		0,99	1	1
KNN	Trung bình	***0,95***	0,9	1	0,95
	Khá		0,98	0,81	0,89
	Giỏi		0,92	0,98	0,95
	Xuất sắc		1	1	1

3. Kết luận

Nghiên cứu này khai phá dữ liệu người học để dự báo khả năng tốt nghiệp đúng hạn và dự báo xếp loại tốt nghiệp của sinh viên Trường Đại học Công nghiệp Hà Nội trong giai đoạn 2017-2023. Kết quả nghiên cứu cho thấy, sau khi áp dụng phương pháp trích chọn thuộc tính trên tập dữ liệu đã được thu thập, đối với dự báo khả năng tốt nghiệp đúng hạn, với tệp dữ liệu nhân khẩu học và kết quả học tập 4 kì thì tỉ lệ dự báo là 88%, và kết quả học tập 6 kì kết quả đạt 90%. Đối với dự báo xếp loại tốt nghiệp kết quả dự báo tương ứng đạt là 89% và 95%. Mô hình dự báo phân lớp dựa trên giải thuật Random Forest là các tài liệu hữu ích không chỉ giúp cho sinh viên mà còn giúp ích cho các

nhà quản lí giáo dục trong việc ra quyết định và hỗ trợ sinh viên trong định hướng cho toàn bộ quá trình học tập của sinh viên. Để nâng cao được tỉ lệ sinh viên tốt nghiệp đúng hạn, dựa trên kết quả phân tích trên, một số khuyến nghị được nghiên cứu đề xuất với các cơ sở giáo dục đại học là: (1) Cơ sở giáo dục cần quan tâm xây dựng các công cụ và phương pháp dự báo sớm kết quả học tập của sinh viên. Ứng dụng phân tích dữ liệu trí tuệ nhân tạo để xây dựng các bộ công cụ và phương pháp có thể dự báo sớm chính xác kết quả học tập của sinh viên theo từng học kì, năm học để từ đó có các căn cứ và các biện pháp tác động tới cá nhân từng sinh viên trong quá trình đào tạo tại trường; (2) Cơ sở giáo dục cần chú trọng tăng cường các hệ thống hỗ trợ sinh viên bao gồm hệ thống cố vấn học tập có khả năng nắm bắt đầy đủ thông tin và kết quả dự báo sớm tình hình học tập của sinh viên. Đội ngũ cố vấn học tập với thông tin toàn diện về sinh viên sẽ có khả năng tư vấn tốt nhất tới sinh viên trong việc giải quyết các vấn đề liên quan đến học tập và cuộc sống có thể giúp tăng khả năng sinh viên tốt nghiệp đúng hạn; (3) Sinh viên dựa trên năng lực của bản thân, mục tiêu học tập và công cụ hỗ trợ của nhà trường, chủ động xây dựng kế hoạch học tập có khả năng tốt nghiệp đúng hạn cao nhất và kết quả xếp loại tốt nghiệp cao.

Lời cảm ơn: *Nghiên cứu này được tài trợ bởi Trường Đại học Công nghiệp Hà Nội, trong đề tài mã số: HĐ 28-2024-RD/HĐ-ĐHCN.*

Tài liệu tham khảo

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *IJACSA, 2*(6).

Đinh Chung Dũng (2017). *Nghiên cứu và áp dụng kĩ thuật khai phá dữ liệu trên bộ dữ liệu sinh viên đại học phục vụ công tác cố vấn học tập*. Luận văn thạc sĩ Công nghệ thông tin, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội.

Huỳnh Lê Uyên Minh, Phạm Trường Trinh, Nguyễn Văn Nhựt (2024). Dự báo khả năng tốt nghiệp đúng hạn của sinh viên: nghiên cứu trường hợp tại Trường Đại học Đồng Tháp. *Tạp chí Giáo dục, 24*(1), 48-53.

Lê Quốc Tiến, Đặng Hoàng Anh (2019). Khai phá dữ liệu: phân tích xếp loại tốt nghiệp và cơ hội việc làm của sinh viên sử dụng kĩ thuật phân lớp. *Tạp chí Khoa học Công nghệ Hàng hải, 59*(8), 125-129.

Mai Thu Giang (2020). Khai phá cơ sở dữ liệu trong hệ thống quản lí đào tạo của Trường Đại học Kinh tế - Đại học Huế. *Tạp chí Khoa học Đại học Huế: Kinh tế và Phát triển, 129*(5B), 123-137.

Mai, A. T. (2018). An Overview of Big Data Analytics in Higher Education Sector. *Journal of Technical Education Science, HCMC University of Technology and Education, 50*(11), 96-104.

Nguyễn Văn Thủy (2023). Sử dụng các mô hình Machine Learning dự đoán tình trạng sinh viên tốt nghiệp đúng hạn. *Tạp chí Khoa học & Đào tạo Ngân hàng, 225*(8), 52-64.

Nguyễn Xuân Hải (2016). *Khai phá dữ liệu và ứng dụng trong dự báo tiến trình học tập của sinh viên Trường Đại học Thủy lợi*. Luận văn thạc sĩ ngành Khoa học máy tính, Học viện Công nghệ Bưu chính viễn thông.

Nguyễn Trí Thành, Nguyễn Hà Nam, Hà Quang Thụy (2013). *Giáo trình Khai phá dữ liệu*. NXB Đại học Quốc gia Hà Nội.

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.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11.