HỌC MÁY DỰ ĐOÁN KẾT QUẢ XẾP LOẠI TỐT NGHIỆP SINH VIÊN

Nguyễn Bá Duy[1], Trần Lê Duy Anh[1] và Diệp Bình Nguyên[1]

[1]Trường Đại học Kỹ thuật - Công nghệ Cần Thơ
Email: nbduy@ctuet.edu.vn

Thông tin chung
Ngày nhận bài: 27/01/2025
Ngày nhận bài sửa: 12/5/2025
Ngày duyệt đăng: 14/5/2025

Từ khóa: *Dự đoán kết quả tốt nghiệp, máy học, phân loại tốt nghiệp*

TÓM TẮT

Khả năng dự báo thành tích học tập tại thời điểm tốt nghiệp có tầm quan trọng sâu sắc đối với các trường đại học, đặc biệt là trong việc phân biệt các yếu tố ảnh hưởng đến xếp loại tốt nghiệp sẽ góp phần vào việc nâng cao hiệu quả xếp loại tốt nghiệp sinh viên. Nghiên cứu này sử dụng nhiều thuật toán học máy bao gồm K-Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine và Recurrent Neural Network để dự đoán kết quả tốt nghiệp của 1.817 sinh viên đại học chính quy thuộc trường Đại học Kỹ thuật - Công nghệ Cần Thơ bao gồm hệ kỹ sư và cử nhân từ năm 2022 đến năm 2024. Các kết quả cho thấy mô hình Decision Tree đưa ra các dự đoán đáng tin cậy nhất và thời gian huấn luyện nhanh. Các yếu tố tác động đến xếp loại tốt nghiệp bao gồm: điểm trung bình tích lũy, tuổi, ngành, giới tính... Dựa trên các phát hiện thực nghiệm, các yếu tố này được xếp hạng để xác định tác động của chúng đến việc phân loại tốt nghiệp của sinh viên.

1. GIỚI THIỆU

Việc sinh viên tốt nghiệp đúng hạn không chỉ là mối quan tâm của chính họ mà còn là ưu tiên hàng đầu của các trường đại học. Một tỷ lệ tốt nghiệp cao kết hợp với chất lượng đào tạo tốt sẽ giúp thu hút nhiều sinh viên tiềm năng đăng ký học tại Trường trong tương lai. Nó củng cố vai trò quan trọng và cải thiện thứ hạng của Nhà trường (Nick, 2016).

Quá trình học tập và tốt nghiệp của sinh viên đại học chịu ảnh hưởng từ nhiều yếu tố, bao gồm sự hỗ trợ từ các tổ chức, công tác kiểm tra giám sát, kỹ năng tự học và động lực học tập. Tuy nhiên, để có được đánh giá chính xác và đáng tin cậy, các nghiên cứu cần được thực hiện trên tập dữ liệu lớn và đủ dài. Bên cạnh đó, việc lựa chọn chuyên ngành cũng đóng vai trò quan trọng trong kết quả tốt nghiệp (Alsayed, 2021).

Học máy được xem là một công cụ hiệu quả trong việc dự báo kết quả học tập của sinh viên (Sekeroglu, 2021). Kỹ thuật này hoạt động bằng cách lựa chọn các tham số phù hợp và phân tích dữ liệu một cách toàn diện. Các thuật toán học máy, chẳng hạn như hồi quy logistic, mạng nơ-ron và nhiều thuật toán khác thường được áp dụng để so sánh và tìm ra phương pháp tối ưu nhằm dự đoán kết quả học tập của sinh viên trong môi trường giáo dục đại học (Yakubu, 2022).

Một số nghiên cứu liên quan

Phương pháp Khai thác Dữ liệu Giáo dục (Educational Data Mining) đã được nhiều nghiên cứu đề cập đến như một công cụ hữu ích trong việc phân tích dữ liệu giáo dục (Wook, 2017). Kỹ thuật này hỗ trợ dự đoán thành tích học tập, cung cấp hướng dẫn và đưa ra quyết định cho sinh viên dựa trên dữ liệu thu thập được. Mustafa (2022) cũng đã áp dụng Educational Data Mining (EDM) kết hợp với các thuật toán học máy để dự đoán kết quả học tập của sinh viên.

Việc dự đoán kết quả học tập của sinh viên đã được nghiên cứu rộng rãi nhằm hỗ trợ quá trình ra quyết định ở các cấp quản lý giáo dục. Ibrahim và AlBarwani (1993) đã sử dụng dữ liệu từ 1.511 sinh viên để dự đoán điểm trung bình của sinh viên năm nhất, cung cấp thông tin hữu ích cho việc cải thiện chương trình đào tạo.

Ngoài ra, Al-Alawi và cộng sự (2023) đã áp dụng các thuật toán học máy để xác định các yếu tố ảnh hưởng đến thời gian học tập, bao gồm giới tính, năm tốt nghiệp dự kiến và nhóm sinh viên.

Khiêm và cộng sự (2023) đã thực hiện dự đoán điểm tốt nghiệp của 7.837 sinh viên thuộc trường Đại học Cần Thơ trong năm 2022 bằng các thuật toán học máy. Kết quả nghiên cứu cho thấy mô hình Random Forest đã dự đoán tốt nhất với độ chính xác khi xác thực chéo là 96,2%.

Nghiên cứu mang lại hai đóng góp chính. Một là, áp dụng các thuật toán học máy để dự đoán xếp loại tốt nghiệp của sinh viên dựa trên các yếu tố như giới tính, tuổi, chuyên ngành, số tín chỉ tích luỹ, điểm trung bình tích luỹ... Bộ dữ liệu sử dụng trong nghiên cứu bao gồm 1.817 sinh viên các ngành đã tốt nghiệp hệ kỹ sư và cử nhân từ năm 2022 đến 2024 thuộc Đại học Kỹ thuật - Công nghệ Cần Thơ. Hai là, phân tích các yếu tố ảnh hưởng đến xếp loại tốt nghiệp của sinh viên nhằm làm rõ vai trò và tác động của chúng đến kết quả dự đoán.

2. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP

2.1. Bộ dữ liệu

Bộ dữ liệu được thu thập tại bộ phận in văn bằng, chứng chỉ từ năm 2022 đến 2024. Kể từ tháng 9 năm 2022, Nhà trường thực hiện cấp bằng tốt nghiệp đại học theo quy định mới của Bộ Giáo dục và Đào tạo về cấp văn bằng tốt nghiệp kỹ sư, cử nhân theo tổng số tín chỉ tích lũy của sinh viên.

Nhóm tác giả đã thực hiện thu thập các tập tin Excel về danh sách viên tốt nghiệp kỹ sư, cử nhân từ năm 2022 đến năm 2024 của Nhà trường. Thực hiện tổng hợp lại thành một tập tin duy nhất để làm bộ dữ liệu huấn luyện và đánh giá các mô hình máy học.

Bảng 1. Số lượng sinh viên tốt nghiệp theo ngành

STT	Ngành	Số sinh viên tốt nghiệp
1	Công nghệ thực phẩm	207
2	Kỹ thuật phần mềm	179
3	Công nghệ kỹ thuật cơ điện tử	133
4	Quản lý xây dựng	86
5	Công nghệ kỹ thuật công trình xây dựng	179
6	Khoa học máy tính	110
7	Kỹ thuật hệ thống công nghiệp	99
8	Hệ thống thông tin	103
9	Công nghệ kỹ thuật điều khiển và tự động hóa	93
10	Quản lý công nghiệp	188
11	Công nghệ sinh học	67
12	Công nghệ thông tin	53
13	Công nghệ kỹ thuật điện, điện tử	197

STT	Ngành	Số sinh viên tốt nghiệp
14	Logistics và quản lý chuỗi cung ứng	86
15	Khoa học dữ liệu	37
	Tổng cộng	1.817

Nguồn: Trung tâm Ngoại ngữ - Tin học, (2025)

2.2. Tối ưu hoá dữ liệu

Để đạt được kết quả dự đoán chính xác nhất, nghiên cứu đã triển khai một loạt các bước xử lý dữ liệu chi tiết như sau:

- Ngày sinh của mỗi sinh viên được chuyển đổi thành một biến mới gọi là "tuoi".

- Nơi sinh được mã hóa về số nguyên theo mã đơn vị địa phương do Bộ Giáo dục và Đào tạo chỉ định.

- Tốt nghiệp kỹ sư gán giá trị 0 và tốt nghiệp cử nhân gán giá trị 1.

- Đánh số từ 0 đến 3 tương ứng với các xếp loại Trung bình, Khá, Giỏi, Xuất sắc.

Sau khi tiền xử lý, nhóm tác giả chuyển thành tập tin định dạng .csv để đưa vào các mô hình máy học.

mssv	ho	ten	tuoi	noi_sinh	gioi_tinh	dan_toc	quoc_tich	nganh	thang_tn	loai_tn	tong_tctl	diem_tbtl	xeploai_tn
			24	86	0	0	Viá»‡t Nam	4	9	1	150	2.6	1
			24	89	1	0	Viá»‡t Nam	4	9	1	150	2.71	1
			24	38	1	0	Viá»‡t Nam	4	9	1	150	2.25	0
			26	92	0	0	Viá»‡t Nam	4	9	1	150	2.7	1
			24	94	1	0	Viá»‡t Nam	4	9	1	150	2.26	0
			24	94	0	0	Viá»‡t Nam	4	9	1	150	2.67	1
			24	91	1	0	Viá»‡t Nam	4	9	1	150	2.62	1
			24	95	0	0	Viá»‡t Nam	4	9	1	150	2.91	1
			24	92	1	0	Viá»‡t Nam	4	9	1	150	2.95	1
			24	91	1	0	Viá»‡t Nam	4	9	1	150	2.64	1
			26	96	1	0	Viá»‡t Nam	3	9	0	134	2.35	0
			23	92	1	0	Viá»‡t Nam	3	9	0	133	2.45	0
			25	92	1	0	Viá»‡t Nam	3	9	0	137	2.04	0
			24	84	1	0	Viá»‡t Nam	3	9	0	133	3.43	2
			22	87	1	0	Viá»‡t Nam	3	9	0	134	2.79	1
			24	89	1	0	Viá»‡t Nam	3	9	0	133	2.11	0
			24	89	1	0	Viá»‡t Nam	3	9	0	133	2.49	0
			25	89	1	0	Viá»‡t Nam	3	9	0	133	2.77	1
			23	87	1	0	Viá»‡t Nam	3	9	0	133	2.25	0
			25	87	1	0	Viá»‡t Nam	3	9	0	133	2.01	0
			26	89	1	0	Viá»‡t Nam	3	9	0	134	2.78	1
			25	92	1	0	Viá»‡t Nam	3	9	0	133	3.06	1

Hình 1. Tập tin csv sau khi tối ưu hóa

Các biến không có ý nghĩa đáng kể đối với việc dự đoán điểm tốt nghiệp đã bị loại bỏ, bao gồm:

- Mã số sinh viên: mã định danh duy nhất, không đóng góp vào việc dự đoán.

- Họ và tên: là một phần của hồ sơ cá nhân, không ảnh hưởng đến kết quả tốt nghiệp.

- Năm tốt nghiệp: xác định thời gian tốt nghiệp của sinh viên, không liên quan đến việc dự đoán.

- Quốc tịch: biến này không có ảnh hưởng đáng kể đến xếp loại tốt nghiệp.

Sau quá trình tối ưu hóa dữ liệu, từ 16 biến ban đầu, bộ dữ liệu cuối cùng bao gồm 9 biến và tổng số mẫu (N) = 1.817.

2.3. Xác thực kết quả

Hai phương pháp phổ biến được sử dụng để phân chia tập dữ liệu trong dự đoán học máy là:

- Phương pháp giữ lại đơn giản: Dữ liệu được tách thành hai tập riêng biệt gồm tập đào tạo và tập đánh giá. Trong nghiên cứu này, phương pháp giữ lại được áp dụng với tỷ lệ 3:1, tương ứng với 75% mẫu dành cho

đào tạo và 25% mẫu trong tập dữ liệu dành cho đánh giá.

- Phương pháp xác thực chéo k-fold: Dữ liệu được chia thành k tập con bằng nhau để đào tạo và đánh giá theo từng lần lặp. Nghiên cứu sử dụng giá trị k = 10. Trong xác thực chéo k-fold, giá trị của k phụ thuộc vào thuật toán và số lần lặp. Đáng chú ý, mỗi mẫu trong dữ liệu sẽ xuất hiện một lần trong tập đánh giá và có thể xuất hiện trong cả tập đào tạo và đánh giá ở các lần lặp khác nhau, đảm bảo tính toàn diện và độ tin cậy của quá trình dự đoán.

2.4. Đánh giá độ chính xác

Đánh giá độ chính xác và độ tin cậy của các thuật toán học máy có tầm quan trọng đáng kể trong các nhiệm vụ dự đoán. Tuy nhiên, do số lượng xếp loại tốt nghiệp khác nhau và mức độ chênh lệch đáng kể, các lớp có thể được coi là không cân bằng.

Bảng 2. Bảng thống kê xếp loại tốt nghiệp theo lớp

	Tập dữ liệu gốc	Tập dữ liệu cân bằng
Lớp 0	472	885
Lớp 1	1158	885
Lớp 2	175	885
Lớp 3	12	885

Nghiên cứu đã thực hiện Oversampling nhằm nhân số lượng dữ liệu bị chênh lệch lên nhiều lần để có một bộ dữ liệu cân xứng hơn. Thư viện SMOTE được áp dụng nhằm tạo ra các mẫu mới bằng cách nội suy giữa các mẫu thiểu số gần nhau.

Trong nghiên cứu này, các số liệu được sử dụng để đánh giá các mô hình là Precision, Recall, F-measure (F1-Score) và Accuracy.

2.5. Thuật toán học máy

2.5.1. Logistic Regresstion

Logistic Regression là một mô hình học máy được sử dụng cho bài toán phân loại. Mặc dù có chữ regression trong tên nhưng nó thực sự được sử dụng để dự đoán xác suất thuộc một lớp cụ thể. Sử dụng hàm logistic (hoặc sigmoid function) để chuyển đổi đầu ra thành giá trị xác suất nằm trong khoảng từ 0 đến 1.

2.5.2. Decision Tree

Mô hình cây quyết định (Decision Tree) là một thuật toán học máy sử dụng cấu trúc cây để đưa ra quyết định hoặc phân loại dữ liệu. Cây quyết định bao gồm các nút (nodes) và các nhánh (branches). Mỗi nút trong cây đại diện cho một thuộc tính trong tập dữ liệu và mỗi nhánh biểu thị một giá trị cụ thể của thuộc tính đó. Quá trình đưa ra quyết định bắt đầu từ nút gốc (root node) và tiếp tục qua các nhánh cho đến khi đạt đến nút lá (leaf node), nơi chứa nhãn quyết định hoặc giá trị dự đoán.

2.5.3. Random Forest

Mô hình Random Forest Classifier là một thuật toán máy học được sử dụng trong bài toán phân loại. Một hệ thống rừng ngẫu nhiên dựa vào nhiều cây quyết định khác nhau. Mỗi cây quyết định bao gồm các nút quyết định, nút lá và nút gốc. Nút lá của mỗi cây là đầu ra cuối cùng được tạo ra bởi cây quyết định cụ thể đó. Việc lựa chọn kết quả cuối cùng tuân theo hệ thống bỏ phiếu đa số. Trong trường hợp này, đầu ra được đa số cây quyết định lựa chọn sẽ trở thành đầu ra cuối cùng của hệ thống rừng ngẫu nhiên.

2.5.4. K-Nearest Neighbor

Mô hình K-Nearest Neighbors Classifier (KNN) là một mô hình phân loại không gian vector được sử dụng trong lĩnh vực học máy. Nó dựa trên nguyên tắc rằng các điểm dữ liệu có thuộc tính tương tự cũng có khả năng thuộc

vào cùng một lớp. Nguyên tắc hoạt động của KNN rất đơn giản. Khi có một điểm dữ liệu mới cần phân loại, mô hình KNN tìm K điểm dữ liệu huấn luyện gần nhất (có khoảng cách nhỏ nhất) với điểm dữ liệu mới đó trong không gian đặc trưng.

2.5.5. Support Vector Machine

Support Vector Machine (SVM) là một thuật toán học máy mạnh mẽ, được sử dụng để phân loại và hồi quy. SVM tìm ra một siêu phẳng (hyperplane) tối ưu trong không gian đặc trưng để phân chia các lớp dữ liệu sao cho khoảng cách giữa các điểm dữ liệu gần nhất (các vector hỗ trợ - support vectors) và siêu phẳng này là lớn nhất.

2.5.6. Recurrent Neural Network

Mạng nơ-ron hồi quy (RNN) là một họ mô hình kết nối mạnh mẽ. RNN hoạt động bằng cách nhận dữ liệu tuần tự và duyệt qua từng phần tử của chuỗi. Tại mỗi bước, RNN cập nhật trạng thái ẩn bằng cách kết hợp thông tin từ trạng thái ẩn trước đó và đầu vào hiện tại. Điều này cho phép RNN học được các mẫu và quan hệ phức tạp trong dữ liệu tuần tự. RNN là một công cụ mạnh mẽ cho các bài toán phân lớp với dữ liệu tuần tự, cho phép mô hình ghi nhớ và sử dụng thông tin từ các bước trước đó để đưa ra dự đoán chính xác.

3. KẾT QUẢ THỰC NGHIỆM

Các mô hình được thực nghiệm trên máy tính có cấu hình phần cứng: Core i5-14000F, RAM 16GB và phần mềm: Windows 10 Pro, PyCharm Community Edition 2022. Thêm vào đó, các mô hình được khởi tạo và chia dữ liệu thành 10 phần (k = 10) để đào tạo và đánh giá, sau đó sử dụng GridSearchCV để tìm kiếm các siêu tham số tốt nhất cho mô hình máy học.

Logistic Regression

Kết quả thử nghiệm trên mô hình cho thấy bằng cách sử dụng tập siêu tham số 'C': 100, 'penalty': 'l1', 'solver': 'saga', mô hình này đạt được hiệu suất tốt nhất với độ chính xác tốt nhất là 92,25% trên tập đào tạo với thời gian thực hiện là 48,23 giây và độ chính xác trên tập đánh giá là 87,25%.

Bảng 3. Kết quả Classification Report của mô hình Logistic Regression

Class	Precision	Recall	F1-Score	Support
0	0,98	1,00	0,99	125
1	0,97	0,82	0,89	273
2	0,49	0,85	0,63	54
3	0,67	0,67	0,67	3

K-Nearest Neighbor

Kết quả thử nghiệm trên mô hình cho thấy khi sử dụng bộ siêu tham số với giá trị 'n_neightbors': 5, 'weights': 'distance' và 'metric': 'manhattan', mô hình này đạt hiệu suất tốt nhất với độ chính xác tốt nhất là 97% trên tập đào tạo với thời gian thực hiện là 2,9 giây và độ chính xác trên tập đánh giá là 90,1%.

Bảng 4. Kết quả Classification Report của mô hình K-Nearest Neighbor

Class	Precision	Recall	F1-Score	Support
0	0,84	0,93	0,88	125
1	0,95	0,90	0,92	273
2	0,88	0,85	0,87	54
3	0,33	0,67	0,44	3

Decision Tree

Kết quả thử nghiệm trên mô hình cho thấy bằng cách sử dụng tập siêu tham số 'criterion': 'gini', 'max_depth': None, 'min_samples_split': 2, 'min_samples_leaf':

1, mô hình này đạt được hiệu suất tốt nhất với độ chính xác tốt nhất là 100% trên tập đào tạo với thời gian thực hiện là 4,62 giây và độ chính xác trên tập đánh giá là 99,78%.

Bảng 5. Kết quả Classification Report của mô hình Decision Tree

Class	Precision	Recall	F1-Score	Support
0	1,00	1,00	1,00	125
1	1,00	1,00	1,00	273
2	1,00	0,98	0,99	54
3	0,75	1,00	0,86	3

Random Forest

Kết quả thử nghiệm trên mô hình cho thấy bằng cách sử dụng tập siêu tham số 'criterion': 'gini', 'max_depth': None, 'min_samples_split': 5, 'min_samples_leaf': 5 và 'n_estimators': 20, mô hình này đạt được hiệu suất tốt nhất với độ chính xác tốt nhất là 100% trên tập đào tạo với thời gian thực hiện là 178,71 giây và độ chính xác trên tập đánh giá là 99,34%.

Bảng 6. Kết quả Classification Report của mô hình Random Forest

Class	Precision	Recall	F1-Score	Support
0	1,00	0,99	1,00	125
1	1,00	1,00	1,00	273
2	0,96	1,00	0,98	54
3	1,00	0,33	0,50	3

Support Vector Machine

Kết quả thử nghiệm trên mô hình cho thấy bằng cách sử dụng tập siêu tham số 'C': 100, 'degree': 2, 'gamma': 'scale' và 'kernel': 'poly', mô hình này đạt được hiệu suất tốt nhất với độ chính xác tốt nhất là 99,58% trên tập đào tạo với thời gian thực hiện là 53,34 giây và độ chính xác trên tập đánh giá là 97,36%.

Bảng 7. Kết quả Classification Report của mô hình SVM

Class	Precision	Recall	F1-Score	Support
0	0,99	0,98	0,99	125
1	0,99	0,98	0,98	273
2	0,89	0,94	0,92	54
3	0,67	0,67	0,67	3

Recurrent Neural Network

Kết quả thử nghiệm trên mô hình cho thấy bằng cách sử dụng 'epochs': 100, 'batch_size': 4, 'dropout': 0,2, 'optimizer': 'Adam', mô hình này đạt được hiệu suất tốt nhất với độ chính xác trên tập đánh giá là 90,1% trong thời gian thực hiện là 82,19 giây.

Bảng 8. Kết quả Classification Report của mô hình RNN

Class	Precision	Recall	F1-Score	Support
0	0,93	0,98	0,96	125
1	0,90	0,97	0,93	273
2	0,90	0,50	0,64	54
3	0,00	0,00	0,00	3

Các biến quan trọng được xác định bằng cách sử dụng feature_importances_ của mô hình Random Forest. Kết quả như sau:

Bảng 9. Bảng xếp hạng các yếu tố quan trọng

STT	Biến	Tỷ lệ
1	diem_tbtl	77%
2	gioi_tinh	6%
3	noi_sinh	4%
4	thang_tn	4%
5	nganh	3,6%
6	tuoi	2,15%
7	tong_tctl	2%
8	loai_tn	1,2%
9	dan_toc	0,05%

Như kết quả trình bày trong Bảng 9, biến diem_tbtl chiếm ưu thế vượt trội với 77%, đây là biến quan trọng nhất trong việc dự đoán xếp loại tốt nghiệp. Điều này phù hợp với thực tế học thuật vì điểm trung bình tích lũy là yếu tố cốt lõi trong việc phân loại bằng tốt nghiệp. Các biến gioi_tinh, noi_sinh, tuoi chỉ chiếm tỷ trọng nhỏ cho thấy kết quả tốt nghiệp chủ yếu dựa vào năng lực học thuật, không chịu ảnh hưởng quá lớn từ thông tin cá nhân. Biến nganh, tong_tctl và thời điểm tốt nghiệp (thang_tn) cũng có ảnh hưởng nhất định, có thể phản ánh sự khác biệt về độ khó hoặc khối lượng học tập theo ngành hoặc theo đợt tốt nghiệp đúng hạn, trễ hạn. Các biến gần như không ảnh hưởng (dan_toc, loai_tn) với tỷ lệ 1,2% trở xuống có thể bị loại bỏ khỏi mô hình để giảm nhiễu và tăng hiệu quả tính toán.

4. KẾT LUẬN

Nghiên cứu này đã áp dụng các kỹ thuật học máy để dự đoán xếp loại tốt nghiệp của 1.817 sinh viên thuộc các ngành đào tạo tại Trường Đại học Kỹ thuật - Công nghệ Cần Thơ. Các thuật toán được sử dụng bao gồm K-Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine và Recurrent Neural Network với tỷ lệ chính xác đạt trên 87%. Đáng chú ý, các thuật toán dựa trên cây như Decision Tree, Random Forest cho kết quả vượt trội với tỷ lệ chính xác cao hơn 99%. Trong số đó, Decision Tree đạt hiệu suất tốt nhất với độ chính xác 99,78% trong kiểm tra xác thực chéo (cross-validation) với k = 10 và thời gian huấn luyện nhanh với 4,62 giây. Nghiên cứu cũng đóng góp vào việc dự đoán xếp loại tốt nghiệp của sinh viên dựa trên các tiêu chí quan trọng. Ngoài điểm trung bình tích luỹ là biến quan trọng nhất, các thông tin khác cũng đóng vai trò quan trọng trong quá trình dự đoán, bao gồm thông tin cá nhân của sinh viên (giới tính, nơi sinh…) và thông tin liên quan đến chương trình học (ngành, tổng số tín chỉ tích luỹ,…).

Trong tương lai, để cải thiện kết quả dự đoán và tăng cường độ tin cậy của mô hình, chúng tôi dự định bổ sung số lượng sinh viên tốt nghiệp qua các năm và tăng số lượng các đặc trưng. Có nhiều yếu tố ảnh hưởng đến kết quả xếp loại của sinh viên, chẳng hạn như điều kiện sống, học đúng chuyên ngành và ảnh hưởng quan hệ tình cảm. Những yếu tố này sẽ được thu thập và tích hợp vào mô hình nhằm tăng độ chính xác dự đoán. Độ chính xác của các mô hình được thực hiện trên 02 tập dữ liệu là đào tạo và đánh giá, do đó trong tương lai chúng tôi sẽ chia tập dữ liệu làm 03 tập là đào tạo, tập đánh giá, tập kiểm tra để đánh giá kết quả toàn diện hơn.

Tài liệu tham khảo

Al-Alawi, L., Al Shaqsi, J. and Tarhini, A., (2023), "Using machine learning to predict factors affecting academic performance: the case of college students on academic probation", Educ. Inf. Technol.

Alsayed, A.O., Rahim, M.S.M., AlBidewi, I., Hussain, M., Jabeen, S.H., Alromema, N., Hussain, S. and Jibril, M. L., (2021), "Selection of the right undergraduate major by students using supervised learning techniques", Appl. Sci., 11.

Ibrahim, A. and Al-Barwani, T. A., (1993), "A study of Omani secondary school Certificate Examination as a predictor of academic performance of Sultan Qaboos University", Research in College Teaching Practicum Research in Sultan Qaboos University, 1(1), 1–29.

Khiêm, N.M, Tú, H.V. và Dũng, N.H., (2023), "Predicting graduation grades using Machine Learning: A case study of Can Tho University students", CTU Journal of Innovation and Sustainable Development. Vol. 15, Special issue on ISDS (2023): 83-92.

Mustafa, Y., (2022), "Educational data mining: prediction of students' academic performance using machine learning

algorithms", Smart Learn. Environ. 9(11). https://doi.org/10.1186/s40561-022-00192-z.

Nick, D., (2016), "The effect of university attended on graduates' labour market prospects: A field study of Great Britain", Economics of Education Review, 52. DOI: https://doi.org/10.1016/j.econedurev.2016.03.001.

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M. and Idoko, J. B., (2021), "Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies", Applied Sciences, 11(22), 10907.

Wook, M., Yusof, Z. M. and Nazri, M. Z. A., (2017), "Educational data mining acceptance among undergraduate students", Educ. Inf. Technol., 22, 1195–1216.

Yakubu, M. N. and Abubakar, A. M., (2022), "Applying machine learning approach to predict students'performance in higher educational institutions", Kybernetes, 51(2), 916–934. https://doi.org/10.1108/K-12-2020-0865.

Danh sách kèm theo quyết định tốt nghiệp chính quy. Liên kết: https://mail.google.com/mail/u/0/#search/cmthanh%40ctuet.edu.vn/FMfcgzQZSjmKWwDtrptNrPJVBTHmwnfn. Ngày truy cập: 10/04/2025.

PREDICTING STUDENT GRADUATION CLASSIFICATIONS USING MACHINE LEARNING TECHNIQUES

ABSTRACT

The ability to predict academic performance at the time of graduation is of profound importance to universities, especially in distinguishing factors affecting graduation ranking will contribute to improving the efficiency of student graduation ranking. This research uses many machine learning algorithms including K-nearest neighbor, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine and Recurrent Neural Network to predict the graduation results of 1,817 full-time university students at Can Tho University of Technology including engineering and bachelor's programs from 2022 to 2024. The results showed that the Decision Tree model provided the most reliable predictions and fast training time. The factors affecting graduation classification included: GPA, age, major, gender, etc. Based on the experimental results, these factors were ranked to determine their impact on the graduation classification of students.

Keywords: *Graduation classification, graduation outcome prediction, machine learning*