ỨNG DỤNG MACHINE LEARNING DỰ BÁO SINH VIÊN DIỆN CẢNH BÁO HỌC TẬP TẠI TRƯỜNG ĐẠI HỌC KINH TẾ HUẾ

Trần Bá Thuấn

Tóm tắt. Máy học đang trở thành ứng dụng cao cấp hỗ trợ con người tìm kiếm những bí ẩn bên trong dữ liệu lớn. Các mô hình máy học được huấn luyện sẽ có khả năng tự phân tích dữ liệu hiện tại để dự đoán xu hướng tương lai. Bài báo này có mục đích xây dựng thuật toán, ứng dụng thực tế, tính khả thi của mô hình máy học có giám sát Random Forest dự báo sinh viên cảnh báo học tập tại Trường Đại học Kinh tế Huế. Tập dữ liệu huấn luyện là điểm thi học kì 1 của 2239 sinh viên năm thứ nhất của các khoá K51 và K52. Dữ liệu huấn luyện có đặc điểm mất cân bằng giữa sinh viên diện cảnh báo và diện không cảnh báo. Trực quan cho thấy dữ liệu phân cụm phi tuyến. Lớp các mô hình máy học phân lớp như k-Nearest Neighbors (kNN), Decision Tree, Perceptron (PLA), Navie Bayes, Logistics Regression, Random Forest và Multip Layers Perceptron (MLP) được đưa vào huấn luyện. Hiệu suất, độ chính xác và hiệu quả trong phân tích dự báo được so sánh để chọn mô hình tốt. Qua huấn luyện cho thấy mô hình máy học Random Forest dự báo hiệu quả nhất. Kết quả dự báo giúp sinh viên điều chỉnh việc học tập, giúp nhà trường quản lý tốt hơn và nâng cao chất lượng đào tạo tại Trường Đại học Kinh tế Huế. Nghiên cứu này cho thấy tính khả thi và hướng nghiên cứu mới tiếp theo.

Từ khoá: Mô hình máy học phân lớp; Sinh viên diện cảnh báo học tập; Mất cân bằng dữ liệu; Mô hình máy học có giám sát; Mô hình máy học không giám sát.

1. Mở đầu

Theo số liệu thống kê sinh viên thuộc diện cảnh báo, năm học 2019-2020 nhà trường có 499 và chỉ riêng học kì 1 năm học 2020-2021 nhà trường có tổng số 565 sinh viên thuộc diện cảnh báo . Số lượng sinh viên thuộc diện cảnh báo tăng trong khi tiêu chí đánh giá sinh viên diện cảnh báo đã giảm bớt khắt khe (Quy chế Đào tạo, 2020). Trước thực trạng cấp thiết đó, làm thế nào nhà trường cảnh báo sớm với sinh viên kết quả học tập để sinh viên nhanh chóng điều chỉnh việc học tập của mình tránh tình trạng "sự việc đã rồi". Do đó, cần có một nghiên cứu mới đầy đủ để tạo ra ứng dụng với độ chính xác cao, tin cậy để phục vụ việc dự báo sinh viên thuộc diện cảnh báo trong những học kì của năm học tới. Gần đây, những ứng dụng của trí tuệ nhân tạo cụ thể là phương pháp máy học đã và đang đóng góp có ý nghĩa vào cải thiện chất lượng giáo dục thông qua học phân tích dự báo và huấn luyện khám phá dữ liệu. Trong bài báo này, mô hình máy học không có giám sát kNN được đưa vào để xử lý dữ liệu trống và sử dụng mô hình máy học có giám sát Random Forest để dự báo. Phần 2 của bài báo sẽ trình bày tổng quan các kết quả đạt được

trong nghiên cứu dự báo sinh viên thuộc diện cảnh báo học tập, phần 3 là phương pháp nghiên cứu và phân tích dữ liệu nghiên cứu và phần 4 là kết quả nghiên cứu và thảo luận.

2. Tổng quan nghiên cứu

Máy học là một ứng dụng của trí tuệ nhân tạo được phân thành học không giám sát, học có giám sát của con người, học bán giám sát và học tăng cường. Phương pháp học không giám sát và có giám sát được sử dụng phổ biến hiện nay. Máy học có giám sát được thực hiện khi dữ liệu có thể gán nhãn đầy đủ giúp con người phân loại, dự đoán, gợi ý,…Trong khi đó, học không giám sát được sử dụng khi dữ liệu không gán nhãn giúp tìm ra mối quan hệ và khuôn mẫu của dữ liệu đầu vào. Phương pháp học bán giám sát và học tăng cường ứng dụng cho các bài toán phân loại và phát hiện gian lận. Các thuật toán máy học được lập trình bằng ngôn ngữ Python (Wes McKinney, 2017) hướng đối tượng hỗ trợ bởi hệ thống các thư viện Pandas, Numpy, Matplotblib, Seaborn và Scikit-learn (Andreas C.Muller, Sarah Guido, 2016) là nguồn tài nguyên vô giá trong nghiên cứu phân tích dữ liệu. Hai ứng dụng Google Colab và Kaggle đã giúp tạo dataset và xử lý mạnh mẽ dữ liệu phục vụ cho phân tích sâu. Thời gian qua, trên thế giới đã có nhiều nghiên cứu dự báo kết quả học tập của sinh viên ứng dụng phương pháp máy học. Năm 2021, một công bố trên tạp chí IEEE Access, nhóm tác giả (Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, 2021) đã sử dụng hệ thống mô hình máy học phân lớp để dự báo kết quả học tập của sinh viên. Nghiên cứu sử dụng kỹ thuật Smote tạo sự cân bằng dữ liệu giúp các mô hình dữ liệu có tính khái quát hơn, tránh bị học lệch hoặc không học. Kĩ thuật Smote không trùng dữ liệu nhưng sử dụng dữ liệu nhân tạo. Mô hình chỉ hoạt động tốt nếu các dữ liệu sinh viên yếu kém sinh ra dữ liệu nhân tạo là giống nhau. Nếu dữ liệu sinh ra và dữ liệu gốc không quá giống nhau có thể gặp vấn đề tạo nhiễu vì dữ liệu được sinh ra chưa chắc là dữ liệu điểm sinh viên yếu kém. Nhóm tác giả (M. Hussain, W.Zhu, W.Zhang, S.M.R Abidi và S. Ali, 2020) đã dự báo những khó khăn sinh viên gặp phải trong một học kì bằng cách sử dụng các mô hình SVM, ANN và Decision Tree. Hiện nay trong nước, xu hướng ứng dụng máy học đang được quan tâm đúng mức. Nhóm tác giả (Đào Đức Anh, Nguyễn Tu Trung và Vũ Văn Thoả, 2020, 48) trong nghiên cứu của mình, đăng trên tạp chí Khoa học Trường Đại học Thuỷ Lợi đã "Ứng dụng của thuật toán Bayes trong vấn đề dự báo học lực của học sinh phổ thông", nhóm tác giả đã sử dụng phương pháp đo độ chính xác trên tập dữ liệu kiểm tra "ta thấy độ chính xác trên dữ liệu test của phương án 1-Kĩ thuật 1 là nhỏ nhất, phương án 1-Kỹ thuật 2 là lớn nhất. Độ chính xác chỉ đạt được như vậy có thể do tập dữ liệu huấn luyện chưa đủ lớn và bao quát miền dữ liệu." Chỉ sử dụng phương pháp đo độ chính xác Accuracy chưa hoàn toàn phát hiện ra những bất thường. Độ chính xác của mô hình chưa cao cho thấy mô hình Navie Bayes dự đoán

đúng học sinh khá giỏi nhưng không thể dự đoán hoặc dự đoán không chính xác đối với diện học sinh yếu chiếm tỉ lệ thấp trong tập dữ liệu và như vậy không loại trừ hết khả năng mô hình học lệch. Tại trường Đại học Cần thơ, nhóm tác giả (Lưu Hoài Sang, Trần Thanh Điện, Nguyễn Thanh Hải và Nguyễn Thái Nghe, 2020) đã xây dựng mạng Neural Multip Layers Perceptron để dự báo kết quả học tập của sinh viên tuy nhiên mô hình chưa đánh giá được độ chính xác và "sử dụng kỹ thuật Dropout giảm tham số ngẫu nhiên để hy vọng mô hình không bị học lệch" như vậy chưa loại trừ khả năng mô hình học lệch (Overfitting) rất lớn trong thực tế và các chỉ số đánh giá MSE của mô hình trong kiểm tra lần lượt là 0.7888 và 0.5708. Bài báo cũng chưa đánh giá đúng mức hiệu suất cũng như hiệu quả và đo độ chính xác của mô hình. Trong bài báo này, nghiên cứu mới tập trung phân tích dự báo sinh viên diện cảnh báo học tập trong tập dữ liệu điểm mất cân bằng điều mà trước đây chưa có bài báo nào đặt làm trọng tâm nghiên cứu. Hơn nữa, trong tập dữ liệu mất cân bằng đó, dữ liệu trống sử dụng kỹ thuật KNNImputer với k=5 để đảm bảo tối ưu. Kết quả nghiên cứu có mục tiêu tìm ra giải pháp phù hợp nhất cho sinh viên diện cảnh báo học tập tại Trường Đại học Kinh tế Huế.

3. Phương pháp và dữ liệu nghiên cứu

3.1 Phát biểu bài toán

Sau khi sinh viên thi một số học phần thuộc học kỳ 1 của năm thứ nhất, nhà trường sẽ dự báo điểm trung bình học kỳ 1 của sinh viên có thuộc diện cảnh báo học tập hay không? Hiện nay, theo quy định đào tạo Đại học hệ chính quy theo hệ thống tín chỉ tại Trường Đại học Kinh tế Huế, theo điều 18, cảnh báo học tập được thực hiện theo từng học kỳ, điều kiện cảnh báo là sinh viên có điểm trung bình chung học kỳ 1 đạt dưới 0,80 và dưới 1,00 đối với các học kì tiếp theo.

3.2 Dữ liệu nghiên cứu

3.2.1 Nguồn dữ liệu và cấu trúc dữ liệu gốc

Dữ liệu được thu thập và chọn lọc từ kho lưu trữ điểm của Trường Đại học Kinh tế Huế. Điểm của sinh viên lưu trữ dưới dạng bảng tính Excel. Mỗi học phần được sắp xếp theo cột gồm điểm chuyên cần, điểm quá trình và điểm kết thúc học phần. Theo quy chế đào tạo của Trường Đại học Kinh tế Huế, sinh viên năm 1 của học kỳ 1 các khoá K51, K52, K53 và K54 học bốn học phần bắt buộc là Toán ứng dụng trong kinh tế, Tin học ứng dụng, Pháp luật đại cương và Triết học Mác Lênin cùng với hai học phần tự chọn là Khoa học môi trường và Địa lí kinh tế.

3.2.2 Thiết kế mô hình dữ liệu

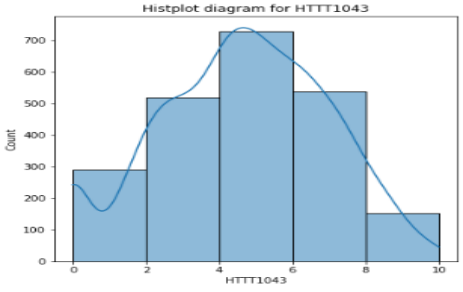
Bảng 1: Dữ liệu huấn luyện

MNA	HKI	HTTT 1043	HTTT 1053	LUAT 1062	KTCT 1022	KTPT 1052	KTPT 1012	LAB
734045	1	9.3	6.8	6.1	5.7	8.1	6.9	0
734045	1	7.9	7.7	5.5	7.6	7.4		0
734045	1	8.3	6.9	6.7	8.2	8.7		0
…	…	…	…	…	…	…	…	…
								2239 sinh viên

Nguồn: Tác giả

Trường dữ liệu đầu vào và đầu ra được thành lập gồm đặc trưng 1 (feature 1) mã ngành (MNA) là dữ liệu điểm sinh viên bao gồm các chuyên ngành Hệ thống Thông tin Quản lý 7340405, Thống kê Kinh tế 7310107,... Đặc trưng 2 (feature 2) KH1 là học kỳ 1 (HK1) của năm thứ. Đặc trưng 3 (feature 3) học phần HTTT1043_Toán ứng dụng trong kinh tế (3): điểm trung bình là 4.65 và độ lệch chuẩn là 2.32.

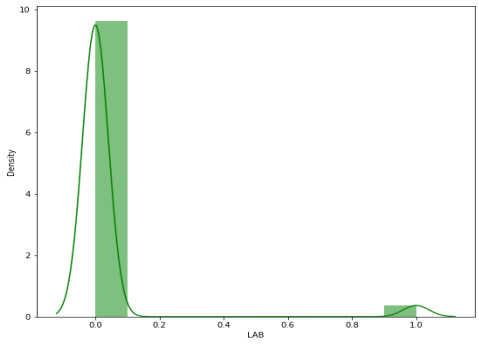
Hình 1: Mẫu dữ liệu điểm Toán ứng dụng trong kinh tế



Nguồn: Tác giả

Đặc trưng 4 (feature 4) HTTT1053_Tin học ứng dụng (3): điểm trung bình 6.32 và độ lệch chuẩn là 1.77. Đặc trưng 5 (feature 5) LUAT1062_Pháp luật đại cương (2): điểm trung bình 6.13 và độ lệch chuẩn là 1.68. Đặc trưng 6 (feature 6) KTCT1022_Những nguyên lý cơ bản của Chủ nghĩa Mác-Lênin 1 (2): điểm trung bình 6.13 và độ lệch chuẩn là 1.68. Đặc trưng 7 (feature 7) KTPT1052_Khoa học môi trường (2): điểm trung bình 5.74 và độ lệch chuẩn là 2.07. Đặc trưng 8 (feature 8) KTPT10_Địa lý kinh tế (2): điểm trung bình 7.18 và độ lệch chuẩn là 2.1. Đặc trưng 9 (feature 9) đầu ra được gán nhãn (Label) LAB nhận hai giá trị mục tiêu (Target) là lớp 0 và 1.

Hình 2: Tỉ lệ giữa lớp 0 và 1



Nguồn: Tác giả

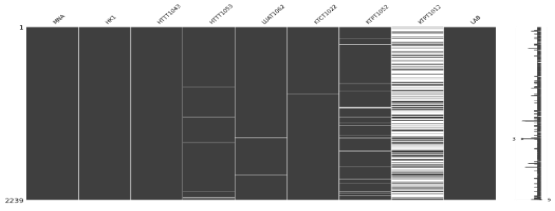
3.2.3 Làm sạch dữ liệu

Tập dữ liệu được bỏ đi dòng trống do sinh viên thôi học hoặc không đăng kí thi tất cả các môn trong học kì 1 năm thứ nhất và giữ lại dòng dữ liệu nhận giá trị 0 đây là đặc điểm nổi bật của bộ dữ liệu thực phi tuyến mất cân bằng này.

3.2.4 Tiền xử lý dữ liệu

Số hoá dữ liệu được thực hiện trên thuộc tính LAB. Nhãn nhận giá trị mục tiêu (Target) 0 là sinh viên không thuộc diện cảnh báo và 1 là sinh viên thuộc diện cảnh báo. NaN (Missing value) là dữ liệu trống do sinh viên chưa thi hoặc không chọn môn học đó mà sẽ chọn môn thay thế. Máy sẽ không học được từ dữ liệu trống chưa qua xử lý. Để có thể dự báo kết quả điểm trung bình học kì 1 chúng ta sẽ phải chuẩn hoá bộ dữ liệu huấn luyện bằng cách đưa ra phương pháp tính phù hợp thay thế vị trí dữ liệu trống bằng dữ liệu điểm thích hợp.

Hình 3: Dữ liệu trống



Nguồn: Tác giả

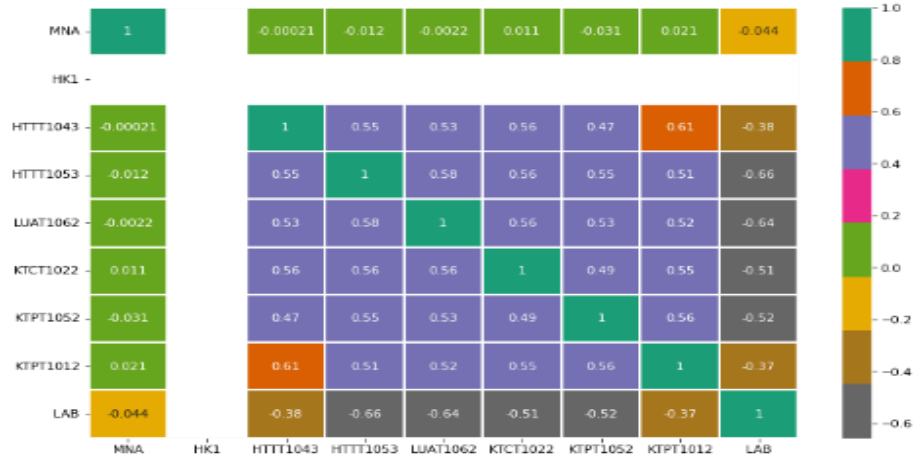
Các đặc trưng học phần tự chọn KTPT1052(2) 6.52% và dữ liệu trống xuất hiện nhiều nhất ở học phần KTPT1012(2) gồm 1341 dữ liệu trống mức 59.89% trên tổng số dữ liệu điểm của học phần.

3.2.5 Phân chia dữ liệu huấn luyện và dữ liệu kiểm tra

Tập dữ liệu huấn luyện (Training dataset) nhận từ 50% đến 70% trong dataset điểm học kỳ 1 năm 1 của 2239 sinh viên khoá K51 và K52. Dữ liệu kiểm tra (Data test) nhận từ 50% đến 30% trong dataset. Tập dữ liệu kiểm tra và tập dữ liệu đánh giá được chọn cùng phân phối chuẩn.

3.2.6 Trực quan hoá dữ liệu

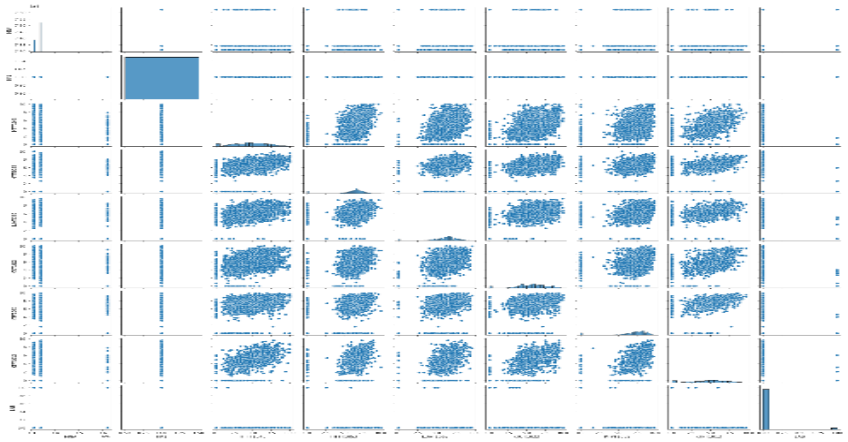
Hình 4: Tương quan dữ liệu



Nguồn: Tác giả

Tương quan Pearson giữa các thuộc tính thể hiện qua các màu sắc khác nhau. Màu xanh ngọc [0.8,1.0] thể hiện tương quan thuận chặt chẽ hay màu xanh lá cây 0.0 không có tương quan hay màu xám thể hiện tương quan ngược chiều chặt chẽ. Trong đó hệ số tương quan từ -1.0 đến bé hơn 0.0 thể hiện mối tương quan ngược chiều. Hệ số tương quan 0.0 đến 1.0 đại diện cho mối tương quan thuận. Chẳng hạn, khi điểm kết thúc học phần của KTPT1012 càng cao (càng thấp) tác động giảm (tăng) nguy cơ cảnh báo của sinh viên với mức tương quan không chặt chẽ 0.37.

Hình 5: Phân cụm dữ liệu



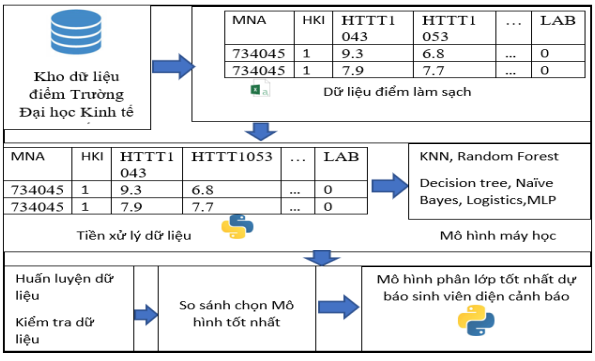
Nguồn: Tác giả

Dữ liệu phân cụm sẽ tương ứng với những mô hình phân lớp dạng phi tuyến. Các mô hình hồi quy tuyến tính và mô hình phân lớp SVM (Support Vector Machine) không phù hợp với dạng dữ liệu này. Bài toán sẽ được phân tích trên mô hình máy học phân lớp phi tuyến phù hợp nhất.

3.3 Mô hình máy học phân lớp

Thống kê dữ liệu điểm của sinh viên khoá K51 và K52 ở học kì 1 năm thứ nhất ta thấy dữ liệu mất cân bằng. Số sinh viên bị cảnh báo là 83 trên tổng số 2239 sinh viên được chọn, tỷ lệ $\frac{83}{2239} = 0,0370701$ bé hơn 4% . Mô hình phân lớp phi tuyến có thể phù hợp với mô hình dữ liệu gồm k-Nearest Neighbors (kNN), Naïve Bayes, Logistic Regression, Perceptron (PLA), Multip Layers Perceptron (MLP), Random Forest và Decision Tree.

Hình 6. Lựa chọn mô hình dự báo tốt nhất



Nguồn: Tác giả

3.3.1 Chọn mô hình qua dữ liệu kiểm tra

Ma trận nhầm lẫn $Confusion\_matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$ là một kỹ thuật đo hiệu suất các mô hình phân loại. True Positive (TP) là số sinh viên không thuộc diện cảnh báo thật. True Negative (TN) số sinh viên thuộc diện diện cảnh báo thật. False Positive (FP) số sinh viên không thuộc diện cảnh báo giả. False Negative (FN) số sinh viên thuộc diện cảnh báo giả. Độ chính xác của mô hình được tính theo các phương pháp sau:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \qquad \text{Re}\,call = \frac{TP}{TP + TN}$$

$$\text{Pr}\,ecision = \frac{TP}{TP + FP} \qquad F1\_score = 2.\frac{\text{Pr}\,ecision \times Recall}{\text{Pr}\,ecision + Recall}$$

Bảng 2: Ma trận nhầm lẫn và độ chính xác

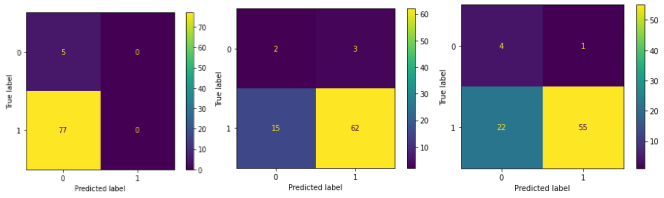
Mô hình	Ma trận nhầm lẫn	Độ chính xác
Logistic Regression	$\begin{bmatrix} 646 & 0 \\ 26 & 0 \end{bmatrix}$	0.9613095238095238
Naïve Bayes	$\begin{bmatrix} 646 & 0 \\ 26 & 0 \end{bmatrix}$	0.9613095238095238
Perceptron (PLA)	$\begin{bmatrix} 646 & 0 \\ 26 & 0 \end{bmatrix}$	0.9613095238095238
kNN	$\begin{bmatrix} 645 & 1 \\ 3 & 23 \end{bmatrix}$	0.9940476190476191
Random Forest	$\begin{bmatrix} 645 & 1 \\ 2 & 24 \end{bmatrix}$	0.9955357142857143
Decision Tree	$\begin{bmatrix} 643 & 3 \\ 1 & 25 \end{bmatrix}$	0.9940476190476191
Multip-Layers Perceptron (MLP)	$\begin{bmatrix} 646 & 0 \\ 26 & 0 \end{bmatrix}$	0.9613095238095238

Nguồn: Tác giả

Đánh giá hiệu suất của các mô hình trên tập dữ liệu kiểm tra 30 %, những mô hình phân lớp học lệch (Overfitting) là Perceptron (PLA), Multip Layers Perceptron (MLP), Navie

Bayes và Logistic Regression. Mô hình Random Forest tạo ra nhiều cây quyết định ngẫ u nhiên có sức mạnh vượt trội với độ chính xác 99.55% , mô hình Decision Tree sử dụn g một cây quyết định dự báo với độ chính xác 99.40%, mô hình máy học đơn giản nhất k-Nearest Neighbors (kNN) dự báo có độ chính xác 99.40%. Để phát hiện bất thường, một tập dữ liệu 82 điểm nhân tạo có nhiễu mạnh được đưa vào kiểm tra đánh giá ba mô hình kNN, Random Forest và Decision Tree. Hai mô hình Random Forest và Decision Tree cho kết quả khả quan. Ngược lại, mô hình kNN dự báo lệch 77 về diện sinh viên th uộc diện cảnh báo giả mà không dự đoán được sinh viên nào thuộc diện cảnh báo thật.

Hình 7: Ma trận nhầm lẫn của kNN, Random Forest và Decision Tree



Nguồn: Tác giả

Mô hình Random Forest cho kết quả tốt nhất với dự báo 62 sinh viên thuộc diện cảnh báo thật, 2 sinh viên không thuộc diện cảnh báo thật, 15 sinh viên thuộc diện cảnh báo giả và 3 sinh viên không thuộc diện cảnh báo giả.

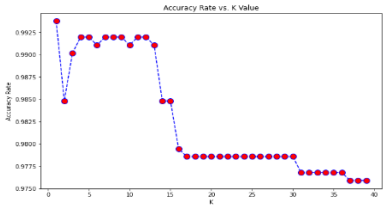
Bảng 3: Độ chính xác của Random Forest và Decision Tree

Model	Predict	Precision	Recall	f1-score	Support
Random	0	0.15	0.80	0.26	5
Forest	1	0.96	0.84	0.90	77
Decision	0	0.12	0.40	0.18	5
Tree	1	0.96	0.71	0.82	77

Nguồn: Tác giả

Qua đánh giá hiệu suất và đo độ chính xác, mô hình Random Forest dự báo có độ chính xác và hiệu suất cao nhất. Đây là mô hình được chọn để dự báo sinh viên thuộc diện cản h báo.

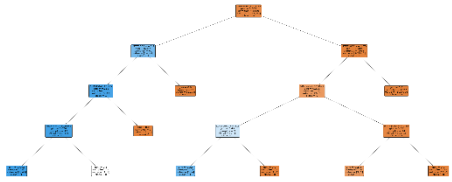
Hình 8: Tìm k tối ưu trong mô hình kNN



Nguồn: Tác giả

Trong xử lý dữ liệu trống, phương pháp Elbow chọn k bằng 5 lân cận gần nhất là tối ưu. Bằng cách sử dụng thuật toán KNNImputer, giá trị trung bình của 5 điểm dữ liệu gần nhất thay thế vào giá trị NaN của dữ liệu điểm trống. Đầu ra của mô hình kNN do đó làm đầu vào trong mô hình Random Forest. Một cây trong mô hình Decision Tree:

Hình 9: Cây phân tách



Nguồn: Tác giả

3.3.2 Hoạt động của mô hình máy học có giám sát Random Forest dự báo sinh viên diện cảnh báo tại Trường Đại học Kinh tế Huế

Chọn ngẫu nhiên $D_i = (MNA, HK1, ..., KTCT1022, KTPT1052, KTPT10)$, $i = \overline{1,n}$ một dữ liệu (sample) từ tập dữ liệu huấn luyện với kĩ thuật bootstrapping (Loe Breiman, 2001) và tiếp tục tạo mẫu dữ liệu sample để tạo ra đủ n dữ liệu nTraining sample. Kỹ thuật boostrapping tạo ra những dữ liệu không hoàn toàn khác nhau. Bộ dữ liệu mới gồm n dữ liệu và mỗi dữ liệu có k thuộc tính ($1 \le k \le 8$). Dùng thuật toán Decision Tree cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi cây quyết định. nDecision Tree, mỗi cây nhận giá trị mục tiêu 0 hoặc 1. Kết quả dự đoán của thuật toán Random Forest sẽ được tổng hợp, bỏ phiếu theo số đông Majority voting cho mỗi kết quả dự đoán. Random Forest chọn kết quả dự đoán nhiều nhất là dự đoán cuối cùng Final-class.

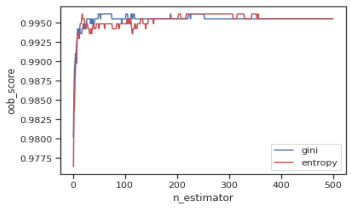
```
Data.fillna(0,inplace=True)
X=np.array(Data.iloc[:,:8]).reshape(-1,8)
Y=np.array(Data.iloc[:,-1]).reshape(-1,1)
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3
,random_state=0)

clf=RandomForestClassifier(n_estimators=500)

clf.fit(X_train,Y_train)

Y_pred=clf.predict(X_test)
```

Hình 10. Biểu đồ Oob_score



Nguồn: Tác giả

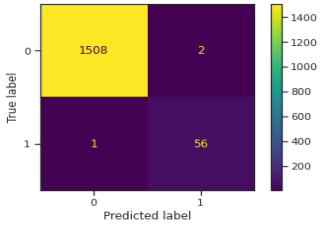
Out-of-Bag error (OOB_error) (Leo Breiman, 2001) dùng để kiểm tra sai số tạo ra từ việc kết hợp kết quả phân loại riêng lẻ sau đó được tổng hợp lại trong mô hình Random Forest.

$$OOB\_error = 1 - OOB\_score$$

Geni đo lường sự suy giảm bình quân tại mỗi nút, làm giảm xác suất phạm sai lầm khi gán nhãn. Entropy thể hiện độ nhiễu của dữ liệu. Độ giảm của Gini được gọi là Geni Gain.

```
from sklearn.metrics import classification_report, plot_confusio
n_matrix,plot_roc_curve
print (classification_report(Y_train,clf.predict(X_train)))
plot_confusion_matrix(clf,X_train, Y_train)
plot_roc_curve(clf,X_train, Y_train)
plt.show()
```

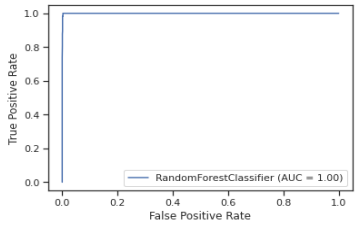
Hình 11. Hiệu suất của mô hình Random Forest



Nguồn: Tác giả

Trong số 1567 điểm dữ liệu huấn luyện chiếm (70%) dataset, mô hình dự đoán 1508 sinh viên không cảnh báo thật, dự đoán 56 sinh viên thuộc diện cảnh báo thật, 1 sinh viên cảnh báo giả và 2 sinh viên không cảnh báo giả với độ chính xác Precision 0.97, Recall 0.98, F1-Score 0.97 và Accuracy 1.00.

Hình 12: Biểu đồ ROC



Nguồn: Tác giả

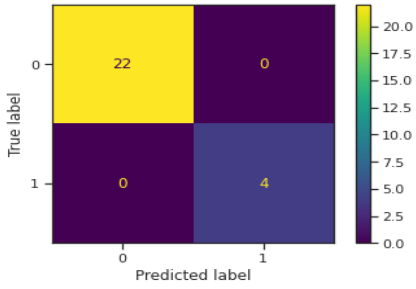
Biểu đồ ROC (Hanley JA, 1982) có trục tung là tỉ lệ sinh viên không cảnh báo thật và trục hoành là tỉ lệ sinh viên cảnh báo giả. Cả hai tỉ lệ có giá trị dao động trong đoạn [0;1]. Hai tỉ lệ này được ước tính trên dữ liệu huấn luyện. Mô hình Random Forest dự báo tốt vì có những giá trị tham chiếu tập trung vào góc vuông có toạ độ (0;1), tức là những điểm ở góc trái thuộc phía trên của biểu đồ. Những điểm này cho chúng ta biết mô hình dự báo hiệu quả và cảnh báo giả rất thấp.

3.3.3 Kiểm tra mô hình

Mô hình Random Forest được kiểm tra qua bộ dữ liệu điểm thực tế là điểm học kì 1 năm 1 của 26 sinh viên ngành Tin học kinh tế của Khoa HTTT Kinh tế, Trường Đại học Kinh tế Huế.

```
X_t=np.array(Data_test.iloc[:,:8]).reshape(-1,8)
Y_t=np.array(Data_test.LAB).reshape(-1,1)
imputer = KNNImputer(n_neighbors=5, weights="uniform")
imputer.fit_transform(X)
X_t=imputer.transform(X_t)
```

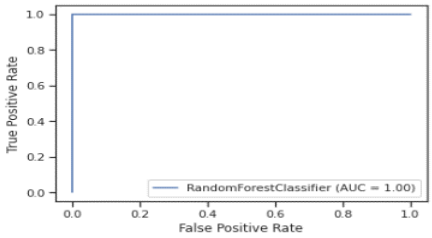
Hình 13: Hiệu suất dự đoán



Nguồn: Tác giả

Kết quả dự báo và kết quả thực tế hoàn toàn trùng khớp. Mô hình Random Forest dự báo đúng 22 sinh viên không thuộc diện cảnh báo thật, 4 sinh viên thuộc diện cảnh báo thật và không có dự báo giả.

Hình 14: Biểu đồ ROC



Nguồn: Tác giả

Biểu đồ ROC (Receiver Operating Characteristic) đo hiệu quả của mô hình Random Forest trên tập dữ liệu kiểm tra vuông góc ở góc (0;1). Mô hình Random Forest vì vậy dự báo tốt nhất với chỉ số dung hoà ước tính diện tích dưới đường biểu diễn ROC hay còn gọi là Area under the curve (Michael J.Pencina, Ralph B. D'Agostino, Ralph B. D'Agostino Jr, Ramachandran S.Vasan, 2007), AUC=1.00. Đây là mô hình tốt hiệu quả nhất để dự báo sinh viên thuộc diện cảnh báo với độ chính xác 100%.

3.3.4 Ứng dụng dự báo của mô hình Random Forest

Trên tập dữ liệu thực điểm học kỳ 1 năm 1 của 11 sinh viên ngành Thống kê kinh doanh khoá K54, mô hình Random Forest cho kết quả dự báo:

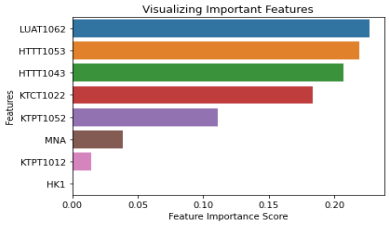
```
X_predict=np.array(Data_predict.iloc[:,:8]).reshape(-1,8)
imputer = KNNImputer(n_neighbors=5, weights="uniform")
imputer.fit_transform(X)
X_predict=imputer.transform(X_predict)
X_predict=np.array([Data_predict.iloc[:,:8]]).reshape(-1,8)
X_predict=imputer.transform(X_predict)
clf.predict(X_predict)
array([0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0])
from collections import Counter
Counter(clf.predict(X_predict))
Counter({0: 9, 1: 2})
```

Trong số 11 sinh viên, 2 sinh viên thuộc diện cảnh báo và 9 sinh viên không thuộc diện cảnh báo.

3.3.5 Xếp hạng mức độ quan trọng của các đặc trưng

Mô hình Random Forest được sử dụng để lựa chọn đặc trưng quan trọng. Thuật toán Random Forest cho phép xếp hạng mức độ quan trọng của các đặc trưng bằng cách gắn thuật toán vào dữ liệu huấn luyện trong quá trình huấn luyện.

Hình 15. Biểu đồ nhân tố quan trọng



Nguồn: Tác giả

Trong quá trình huấn luyện sai số Oob_error(out-of-bag error) cho mỗi điểm dữ liệu được ghi lại và tính trung bình trên toàn bộ khu rừng. Tầm quan trọng của đặc trưng thứ k được tính bằng cách lấy trung bình sai số Oob (out-of-bag error) xuất hiện trước và sau khi hoán vị trên tất cả các cây. Trong 8 đặc trưng là đầu vào của mô hình, đặc trưng LUAT1062 là quan trọng nhất (feature-scores: 0.226479) và đặc trưng HK1 không quan trọng (feature-scores: 0.000000) với mô hình Random Forest dự báo sinh viên cảnh báo.

4. Kết quả nghiên cứu và thảo luận

4.1 Kết quả nghiên cứu

Mô hình máy học Random Forest phi tuyến đã giải thành công bài toán dự báo sinh viên diện cảnh báo tại Trường Đại học Kinh tế Huế. Những phát hiện mới: (i) sinh viên diện cảnh báo trong học kỳ 1 năm 1; (ii) có thể xây dựng mô hình Random Forest dự báo sinh viên diện cảnh báo cho các học kỳ; (iii) các mô hình SVM, Navie Bayes, Logistic Regression, PLA và MLP không sử dụng được cho mô hình dữ liệu mất cân bằng này; (iv) mô hình kNN hỗ trợ xử lý dữ liệu đầu vào không tham gia vào quá trình dự báo.

Mô hình Random Forest mang tính khách quan hơn mô hình Decision Tree vì tạo được nhiều cây ngẫu nhiên trên các tập con ngẫu nhiên và kết quả thu được bằng cách bỏ phiếu.

Thuật toán Random Forest giúp phát hiện hai nhân tố quan trọng nhất là LUAT1062 và HTTT1053. Trong khi đó, nhân tố HK1 không ảnh hưởng đến mô hình do đó có thể xây dựng các mô hình dự báo sinh viên thuộc diện cảnh báo học tập cho mỗi học kỳ.

4.2 Thảo luận

Ưu điểm của mô hình Random Forest là độ chính xác cao khắc phục được hạn chế học lệch hoặc không học của các mô hình kNN khi dữ liệu lệch hẳn về phía diện không cảnh báo. Nhược điểm của mô hình là hoạt động mất nhiều thời gian hơn mô hình Decision Tree vì phải tạo ra nhiều cây hơn.

Nhà trường nên xây dựng một ứng dụng hiệu quả trên cơ sở kết quả nghiên cứu này để dự báo sinh viên diện cảnh báo. Dự báo sinh viên diện cảnh báo tại trường Đại học Kinh tế Huế giúp Nhà trường quản lý giáo dục ngăn chặn sinh viên bị đuổi học hoặc bỏ học trước khi hoàn thành khoá học. Nghiên cứu này giúp hiểu rõ hơn những sinh viên cần sự giúp đỡ.

Mô hình tương tự như mô hình Random Forest là Adaboost được huấn luyện thử nghiệm. Trong tương lai mô hình máy học K-mean và Bdscan phân cụm trên nền ngôn ngữ lập trình Python cùng các thư viện hỗ trợ sẽ nghiên cứu hành vi học tập của sinh viên khi tương tác với các môn học Toán ứng dụng, Tin học ứng dụng, Kinh tế lượng hay Lập trình nâng cao.

Tài liệu tham khảo

Andreas C.Muller & Sarah Guido (2016), 'Introduction to Machine Learning with Python', O'Reilly Media, USA.

Các quy chế đào tạo, quản lý sinh viên và các văn bản về chế độ chính sách của sinh viên (2020), Trường Đại học Kinh tế Huế, Đại học Huế.

Đào Đức Anh, Nguyễn Tu Trung, Vũ Văn Thoả (2020), 'Ứng dụng thuật toán Navie Bayes trong vấn đề dự báo học lực của học sinh phổ thông', Tạp chí khoa học công nghệ thông tin và truyền thông, 1(CS.01).

Hanley JA, McNeil BJ, 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', Radiology, 1982, 143(1), 29-36.

Lưu Hoài Sang, Trần Thanh Điện, Nguyễn Thanh Hải và Nguyễn Thái Nghe (2020), 'Dự báo kết quả học tập bằng kĩ thuật học sâu với mạng Neural đa tầng', Tạp chí khoa học trường Đại học Cần Thơ, 56, 3A, 20-28.

Leo Breiman, 'Random Forest' (2001), Statistics Departement, University of California Berkeley, CA 94720.

Michael J.Pencina, Ralph B. D'Agostino, Ralph B. D'Agostino Jr, Ramachandran S.Vasan, 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond', Statistics in Medicine, Research Article,15-3-2007, Wiley.

M. Hussain, W.Zhu, W.Zhang, S.M.R Abidi và S. Ali (2019), 'Using machine learning to predict student difficulties from learning session data,' Artiff Intell. Rev, Vol 52, No.1, 381-407, Spinger.

Sklearn.impute.KNNImputer, truy cập ngày 27 tháng 5 năm 2021, từ https://scikit-learn.org.

Siti Dianah Adul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-Viedma, Hamido Fujita và Nor Azura Md. Ghanni (2021), 'Multiclass Prediction Model for Student Grade Prediction Using Machine Learning', IEEE Access, 9, 95608-95621.

Wes McKinney (2017), 'Python for Data Analysis', O'Reilly Media, USA.

USING MACHINE LEARNING TO PREDICT THE STUDENTS ON ACADEMIC WARNING STATUS AT HUE UNIVERSITY OF ECONOMICS

Tran Ba Thuan

Abstract. Machine learning is becoming a high-end application to help people find mysteries in big data. Trained machine learning models will self-analyze current data to predict future trends. This article aims to build a classifier machine learning model that predicts the students on academic warning status at Hue University of Economics. The training dataset is the scores of 2239 first year students of K51 and K52 . The visualization shows that data is imbalanced and nonlinear clustering. Multiclass machine learning classifier models k-Nearest Neighbors (kNN), Decision Tree, Perceptron (PLA), Navie Bayes, Logistics Regression, Random Forest and Multip Layers Perceptron (MLP) are included in the training. By using confusion matrix, accuracy of model and Roc-curve achieved in predictive analysis dataset and datatest, we compare results to select the best final model. The best model is Random Forest used to predict the students on academic warning status. The resutls of research will help students to adjust their studies and help the university administrative staffs better manage and improve the training quality at Hue University of Economics. The study also shows the feasibility and directions for future research.

Keywords: Machine learning classifier model; Students on academic warning status; Imbalanced data; Supervised learning; Unsupervised learning.