ỨNG DỤNG CÁC GIẢI THUẬT HỌC MÁY TRONG DỰ ĐOÁN HIỆU SUẤT HỌC TẬP CỦA SINH VIÊN QUA BÀI KIỂM TRA THỰC HÀNH CUỐI KÌ

LÊ QUỐC HUY, NGUYỄN HỮU QUANG, NGUYỄN THÀNH TRỌNG, PHẠM GIA KHÁNH, NGUYỄN MINH PHÚC, PHẠM THỊ THIẾT*

[1]*Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh*
22636191.huy@student.iuh.edu.vn, nguyenhuuquang@iuh.edu.vn,
22642481.trong@student.iuh.edu.vn, 22724051.khanh@student.iuh.edu.vn,
22637001.phuc@student.iuh.edu.vn, *phamthithiet@iuh.edu.vn*

Tóm tắt. Nghiên cứu này sử dụng các mô hình học máy để dự đoán khả năng đạt hay không đạt trong bài kiểm tra thực hành cuối kỳ của sinh viên trong khóa học kỹ thuật lập trình, dựa trên dữ liệu từ các bài tập thực hành. Bộ dữ liệu gồm 1430 sinh viên với các biến như điểm số, số lần thử và thời gian làm bài cho các bài tập Prelab, Inlab và Postlab. Sinh viên được phân loại thành hai nhóm: đạt yêu cầu và không đạt yêu cầu, dựa trên điểm số bài kiểm tra cuối kỳ. Sáu thuật toán học máy (Random Forest, SVM, Hồi quy Logistic, K-Nearest Neighbors, Naive Bayes và Cây quyết định) được áp dụng và đánh giá qua độ chính xác (Accuracy), ma trận nhầm lẫn (Confusion Matrix), độ chính xác (Precision), độ nhạy (Recall), và F1-score. Kết quả cho thấy các mô hình học máy có khả năng dự đoán chính xác việc sinh viên đạt hay không đạt bài kiểm tra cuối kỳ, từ đó hỗ trợ quá trình cá nhân hóa học tập và nâng cao chất lượng đào tạo.
Từ khóa: Dự đoán kết quả, Giải thuật phân lớp, Mô hình dự đoán, Kết quả học tập….

APPLYING MACHINE LEARNING ALGORITHMS TO PREDICT STUDENTS' ACADEMIC PERFORMANCE THROUGH FINAL PRACTICAL EXAMS

Abstract. This study utilizes machine learning models to predict students' ability to pass or fail the final practical exam in a programming techniques course, based on data from practice exercises. The dataset consists of 1430 students with variables such as scores, number of attempts, and time spent on Prelab, Inlab, and Postlab exercises. Students are classified into two groups: pass and fail, based on their final exam scores. Six machine learning algorithms (Random Forest, SVM, Logistic Regression, K-Nearest Neighbors, Naive Bayes, and Decision Tree) were applied and evaluated using accuracy, confusion matrix, precision, recall, and F1-score. The results indicate that machine learning models can accurately predict students' performance on the final exam, thereby supporting personalized learning and improving training quality.
Keywords: Predicting results, Classification algorithms, Predictive models, Learning outcomes….

1 GIỚI THIỆU

Trong bối cảnh giáo dục hiện đại, việc ứng dụng công nghệ và phân tích dữ liệu vào quá trình đánh giá và cải thiện hiệu suất học tập của sinh viên đang ngày càng trở nên quan trọng. Dự đoán kết quả học tập không chỉ giúp các giảng viên và nhà quản lý đánh giá chính xác năng lực của sinh viên mà còn hỗ trợ cá nhân hóa quá trình học, từ đó nâng cao chất lượng đào tạo và cải thiện các phương pháp giảng dạy [1]. Mặt khác, dự đoán này còn đóng vai trò quan trọng trong việc nhận diện sớm những sinh viên có nguy cơ không đạt yêu cầu, giúp đưa ra các can thiệp kịp thời và hiệu quả.

Một trong những yếu tố quan trọng trong quá trình học tập là sự tham gia của sinh viên vào các bài tập thực hành. Các bài thực hành trong các môn học như lập trình, với sự ghi nhận điểm số, số lần thử và thời

gian làm bài, có thể cung cấp những dữ liệu hữu ích để dự đoán khả năng hoàn thành các mục tiêu học tập, như việc vượt qua bài kiểm tra cuối khóa. Dữ liệu từ các bài thực hành trước, trong và sau giờ học trên lớp có thể phản ánh mức độ tham gia và nỗ lực của sinh viên, từ đó giúp xây dựng các mô hình dự đoán kết quả học tập của họ.

Nghiên cứu này nhằm mục đích sử dụng dữ liệu thu thập từ các bài thực hành để phát triển mô hình học máy dự đoán khả năng đạt được mục tiêu học tập của sinh viên, cụ thể là kết quả thi cuối khóa. Bằng cách áp dụng các thuật toán học máy như Random Forest, Support Vector Machines (SVM), Hồi quy Logistic, K-Nearest Neighbors, Naive Bayes và Cây quyết định, nghiên cứu sẽ đánh giá hiệu quả của từng mô hình trong việc phân loại sinh viên vào các nhóm "đạt yêu cầu" hoặc "không đạt yêu cầu" dựa trên điểm số cuối khóa.

Kết quả từ nghiên cứu này không chỉ giúp đánh giá hiệu quả của các mô hình học máy trong việc dự đoán kết quả học tập, mà còn cung cấp thông tin hữu ích cho các giảng viên và nhà quản lý trong việc cải thiện phương pháp giảng dạy và tối ưu hóa quá trình học tập.

2 CÁC NGHIÊN CỨU LIÊN QUAN

Kumar et al. [2] đã tiến hành một nghiên cứu trong lĩnh vực Khai thác Dữ liệu Giáo dục, tập trung vào việc trích xuất thông tin từ dữ liệu giáo dục để nâng cao chất lượng giảng dạy. Các tác giả phát hiện rằng giáo viên có thể dự đoán xu hướng hiệu suất của học sinh bằng cách sử dụng các kỹ thuật khai thác dữ liệu như Decision Trees, Random Forests, and Naïve Bayes. Các thử nghiệm trên dữ liệu môn toán và tiếng Bồ Đào Nha cho thấy độ chính xác hợp lý trong việc dự đoán điểm thi cuối kỳ. Các kỹ thuật lựa chọn và phân loại đặc trưng đã giúp cải thiện độ chính xác của dự đoán.

Thien-Wan AU et al. [3] đã nghiên cứu các kỹ thuật học máy để dự đoán hiệu suất của học sinh trong các khóa học lập trình, sử dụng dữ liệu từ 71 sinh viên. Họ đã phát triển các mô hình dự đoán sử dụng Naïve Bayes, C4.5 Decision Tree, Random Forest, and K-Nearest Neighbour, trong đó Random Forest đạt được độ chính xác cao nhất. Nghiên cứu này gợi ý mở rộng các dự đoán qua các khoảng thời gian khóa học khác nhau để nâng cao kết quả giáo dục.

Yu-Sheng Su et al. [4] đã nghiên cứu các công nghệ học máy để phân tích hành vi học tập của sinh viên và dự đoán kết quả học tập. Họ đã áp dụng các phương pháp phân loại như Random Forest and Neural Networks, trong đó Neural Network đạt được độ chính xác dự đoán cao nhất. Nghiên cứu nhấn mạnh tiềm năng của các mô hình học máy được điều chỉnh tốt trong việc dự đoán kết quả giáo dục.

Al-Alawi and Alsubaiee [5] đã xác định các phương pháp khai thác dữ liệu phổ biến được sử dụng để nâng cao hiệu suất học tập của sinh viên, đồng thời ghi nhận sự xuất hiện của các yếu tố ảnh hưởng mới trong giáo dục. Hầu hết các nghiên cứu sử dụng Random Forest và Naïve Bayes, với phương pháp kết hợp LMT đạt độ chính xác cao nhất. Các tác giả kêu gọi việc áp dụng rộng rãi các thuật toán này, đặc biệt là tại khu vực Vịnh Ả Rập, để dự đoán hiệu suất học tập của sinh viên một cách hiệu quả.

Nguyen Dinh Van et al. [6] nhấn mạnh tầm quan trọng của việc dự đoán sớm hiệu suất học tập của sinh viên để xác định những sinh viên có nguy cơ. Phân tích hồ sơ của gần 400 sinh viên, họ đã sử dụng mô hình Neural Network sâu để dự đoán hiệu suất của sinh viên năm thứ tư, đạt độ chính xác 77%. Họ phát hiện rằng mặc dù KNN có độ chính xác tốt nhất, nhưng tỷ lệ sai sót cao của nó đã hạn chế khả năng tổng quát.

Mustafa Yağcı [7] đã áp dụng các kỹ thuật học máy khác nhau để dự đoán điểm thi cuối kỳ của sinh viên đại học từ kết quả thi giữa kỳ. Sử dụng dữ liệu từ 1.854 sinh viên, tác giả báo cáo độ chính xác dự đoán trong khoảng từ 70–75%, đồng thời nhấn mạnh rằng điểm thi giữa kỳ là yếu tố dự đoán mạnh mẽ cho kết quả cuối kỳ.

Sivasakthi M. and Pandiyan M. [8] đã khảo sát các thuật toán học máy trong các khóa học lập trình, phát hiện Naïve Bayes đạt độ chính xác cao nhất (91,02%), nhấn mạnh tầm quan trọng của việc dự đoán sớm để giúp sinh viên đạt được thành công.

Jose Llanos et al. [9] đã giải quyết vấn đề tỷ lệ thất bại cao trong các khóa học lập trình cơ bản bằng cách phát triển một mô hình dự đoán dựa trên các đánh giá sớm. Mô hình cho thấy bộ phân loại gradient boosting đạt hiệu suất tốt nhất trong việc dự đoán sớm, từ đó giúp can thiệp kịp thời.

Shan Chen and Yuanzhao Ding [10] đã sử dụng học máy để dự đoán hiệu suất học tập dựa trên dữ liệu giáo dục và xã hội học, với Neural Network đạt độ chính xác cao nhất (60%). Kết quả nghiên cứu chỉ ra ảnh hưởng đáng kể của các yếu tố xã hội học đến kết quả học tập.

The authors [11] đã đề xuất một mô hình dự đoán điểm số của sinh viên sử dụng RFE_RF cho việc lựa chọn đặc trưng, đạt được độ chính xác 84,38%. Họ đã chỉ ra rằng các yếu tố hành vi, như tham gia lớp học và tham gia các hoạt động, có ảnh hưởng đáng kể đến hiệu suất học tập, cho thấy rằng những yếu tố này quan trọng hơn các yếu tố dân số học trong việc dự đoán thành công học tập.

3 PHƯƠNG PHÁP NGHIÊN CỨU

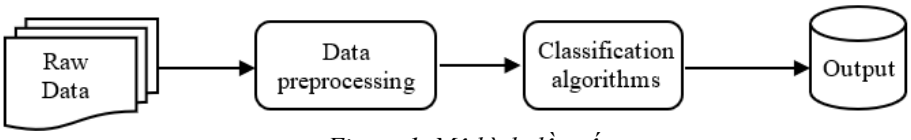

Figure 1. Mô hình đề xuất

Hình trên minh họa trên mô tả quy trình phân tích dữ liệu trong bài toán dự đoán kết quả học tập của sinh viên thông qua các bài tập thực hành. Quá trình bắt đầu với Raw Data (dữ liệu thô), bao gồm các thông tin về kết quả thực hiện bài tập của sinh viên. Dữ liệu này sau đó được trải qua bước Data preprocessing (tiền xử lý dữ liệu), trong đó các giá trị thiếu và sai lệch được xử lý để chuẩn bị cho các bước tiếp theo. Tiếp theo, Classification algorithm (thuật toán phân loại) được áp dụng nhằm huấn luyện mô hình học máy dựa trên các đặc trưng đã được tiền xử lý, với mục tiêu dự đoán kết quả kỳ thi cuối kỳ của sinh viên. Cuối cùng, kết quả dự đoán sẽ được đưa ra trong phần Output, cung cấp thông tin về khả năng đạt hay không đạt của sinh viên trong bài kiểm tra thực hành, từ đó hỗ trợ việc đánh giá hiệu quả học tập và dự báo kết quả học tập tương lai.

3.1 Mô tả dữ liệu

Nghiên cứu này sử dụng một tập dữ liệu nhằm dự đoán kết quả cuối cùng của các bài thi thực hành của sinh viên trong môn Kỹ thuật Lập trình vào học kỳ mùa xuân năm 2023. Môn học này bao gồm hai phần chính: lý thuyết và thực hành. Nghiên cứu tập trung phân tích các chỉ số về hiệu suất học tập của sinh viên trong các buổi thực hành, được tổ chức thành các bài tập chính gọi là Labs. Có tất cả 03 bài Labs lớn, mỗi lab sẽ có nhiều chủ đề, mỗi được chia thành ba phần: bài tập chuẩn bị trước buổi thực hành (Prelab), bài tập tại phòng lab (Inlab), và bài tập nâng cao sau buổi thực hành (Postlab), như mô tả chi tiết trong Bảng 01.

Table 1. Summary of Lab Exercises and Corresponding Topics

No.	Type	Topic
1	Lab 1	C-String
2	Lab 1	Class string
3	Lab 1	Multi-dimensional Array
4	Lab 1	File IO
5	Lab 1	Function
6	Lab 2	Recursion
7	Lab 2	Pointer Basic

8	Lab 2	Pointer 2
9	Lab 2	Struct
10	Lab 3	Linked List
11	Lab3	OOP

Kết thúc các buổi thực hành, sinh viên sẽ tham gia một bài kiểm tra cuối cùng nhằm đánh giá mức độ thành thạo của mình. Nghiên cứu sử dụng tổng hợp các kết quả từ tất cả các labs để làm biến dự đoán cho kết quả bài thi cuối cùng. Trong từng chủ đề, sinh viên có thể thực hiện các bài tập nhiều lần để cải thiện thành tích học tập.

Tập dữ liệu bao gồm: Điểm số của từng bài tập trong Prelab, Inlab, và Postlab; Số lần thử bài tập (Attempts) đối với từng phần Prelab, Inlab, và Postlab; Thời gian thực hiện bài tập (thời điểm bắt đầu và kết thúc); Điểm số cuối cùng (labFinalScore) – biến mục tiêu trong nghiên cứu. Những biến số này được thiết kế nhằm phân tích chi tiết khả năng học tập của sinh viên, giúp dự đoán hiệu quả kết quả bài thi cuối cùng.

3.2 Tiền xử lý dữ liệu:

Vì sinh viên có quyền thực hiện lại mỗi bài tập thực hành nhiều lần để đạt được điểm số mong muốn và hệ thống ghi nhận tất cả các lần thực hiện, nhóm nghiên cứu chỉ sử dụng các biến liên quan như: Điểm số cao nhất mà sinh viên đạt được cho từng chủ đề; Số lần thực hiện bài lab mà sinh viên đã tiến hành cho chủ đề đó; Tổng thời gian sinh viên đã sử dụng để hoàn thành bài lab.

Sau khi tổng hợp, tập dữ liệu bao gồm kết quả học tập của 1,430 sinh viên. Trong trường hợp sinh viên không tham gia hoặc bỏ qua một chủ đề cụ thể, hệ thống ghi nhận giá trị "N/A" cho các mục dữ liệu tương ứng. Theo quy định học thuật của trường, những trường hợp này được tính là 0 điểm. Do đó, nhóm nghiên cứu đã chuyển đổi toàn bộ các giá trị "N/A" thành 0 điểm, đảm bảo tuân thủ các quy định học tập.

Điểm số cuối cùng (labFinalScore) được sử dụng làm biến mục tiêu để dự đoán, được phân loại thành hai nhãn: 1 (Đạt): Sinh viên đạt yêu cầu với kết quả từ 4 điểm trở lên; 0 (Không đạt): Sinh viên không đạt yêu cầu với kết quả dưới 4 điểm. Phân bố dữ liệu cho hai lớp này trong tập dữ liệu gồm 942 sinh viên thuộc lớp 1 (đạt) và 488 sinh viên thuộc lớp 0 (không đạt).

3.3 Thực nghiệm

Thực nghiệm sử dụng bộ dữ liệu gồm 1.430 mẫu, được chia thành hai tập dữ liệu: 80% cho huấn luyện và 20% cho kiểm tra. Dữ liệu đã được xử lý và chuẩn bị cho quá trình phân tích, bao gồm việc xử lý các giá trị thiếu, chuẩn hóa dữ liệu và lựa chọn các đặc trưng quan trọng. Các mô hình học máy được lựa chọn dựa trên thành công đã được chứng minh trong các nghiên cứu trước. Quá trình huấn luyện và kiểm tra các mô hình này được thực hiện trên nền tảng Google Colab, tận dụng khả năng tính toán nhanh chóng nhờ GPU.

Nghiên cứu này tập trung vào các nhiệm vụ phân loại, sử dụng ba giải thuật nổi bật: Random Forest, Support Vector Machines (SVM), và hồi quy logistic.

Random Forest [12] Là phương pháp học máy theo kiểu tập hợp, sử dụng nhiều cây quyết định kết hợp với kỹ thuật "bagging" để giảm thiểu overfitting và nâng cao độ chính xác. Phương pháp này hoạt động bằng cách tạo ra nhiều cây quyết định từ các tập con ngẫu nhiên của dữ liệu huấn luyện, với dự đoán cuối cùng được thực hiện thông qua bỏ phiếu đa số từ các cây này.

Phương pháp Support Vector Machines (SVM) [13] được sử dụng để phân loại dữ liệu có thể phân chia được tuyến tính hoặc phi tuyến tính. Bằng cách chiếu dữ liệu vào không gian có chiều cao hơn, SVM tìm ra mặt phẳng phân chia tối ưu, giúp phân loại chính xác trong các bộ dữ liệu phức tạp.

Logistic Regression [14] là phương pháp phân loại nhị phân, giúp dự đoán xác suất của các sự kiện bằng cách sử dụng các biến độc lập. Hồi quy logistic có khả năng xử lý các biến liên tục và phân loại, là công cụ mạnh mẽ trong các bài toán dự đoán nhị phân.

K-Nearest Neighbors (KNN) [3]: Là một thuật toán phân loại dựa trên khoảng cách giữa các điểm dữ liệu. Mô hình sẽ phân loại một điểm dữ liệu dựa trên số điểm gần nhất (k điểm gần nhất) trong không gian đặc trưng.

Naive Bayes (NB) [5]: Là mô hình phân loại dựa trên định lý Bayes, giả định các biến độc lập với nhau. Mô hình này tính toán xác suất có điều kiện của các lớp và sử dụng chúng để phân loại dữ liệu.

Decision Tree (Cây quyết định) [2]: Là mô hình học máy sử dụng các cây quyết định để phân loại, dựa trên các câu hỏi có giá trị nhị phân. Cây quyết định chia dữ liệu thành các nhánh theo từng đặc trưng và đưa ra quyết định dựa trên các tiêu chí đã được xác định.

4 KẾT QUẢ THỰC NGHIỆM

Table 2. Kết quả thực nghiệm

Model	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
Logistic Regression	0.67	0.61 / 0.68	0.14 / 0.95	0.23 / 0.79
Support Vector Machine	0.69	0.58 / 0.72	0.34 / 0.87	0.43 / 0.79
K-Nearest Neighbors	0.69	0.56 / 0.73	0.41 / 0.83	0.47 / 0.78
Random Forest	**0.73**	0.65 / 0.75	0.44 / 0.88	0.52 / 0.81
Gradient Boosting	0.7	0.58 / 0.75	0.46 / 0.82	0.51 / 0.78
Neural Network (MLP)	0.69	0.55 / 0.74	0.45 / 0.81	0.49 / 0.77

Các mô hình được đánh giá qua các chỉ số hiệu năng: độ chính xác (Accuracy), ma trận nhầm lẫn (Confusion Matrix), độ chính xác (Precision), độ nhạy (Recall), và F1-score. Kết quả thực nghiệm được thể hiện trong bảng 2.

Random Forest là mô hình có hiệu năng cao nhất, đạt độ chính xác 73% và F1-score cho lớp đạt (lớp 1) là 0.81. Mô hình này cho thấy khả năng phân loại chính xác cao cả hai lớp. Cụ thể, độ nhạy của lớp đạt là 88%, vượt trội so với các mô hình khác, cho thấy khả năng dự đoán tốt cho sinh viên đạt điểm đạt chuẩn. Tuy nhiên, lớp không đạt (lớp 0) vẫn gặp khó khăn với độ nhạy chỉ đạt 44%, phản ánh sự mất cân bằng trong dự đoán giữa hai lớp. Gradient Boosting đạt độ chính xác 70%, với F1-score cho lớp đạt là 0.78. Mô hình này có độ nhạy tốt cho lớp đạt (82%) nhưng độ chính xác cho lớp không đạt chỉ đạt 46%. Điều này chỉ ra rằng mô hình thiên về phân loại các sinh viên đạt điểm đạt, trong khi khả năng phát hiện sinh viên không đạt còn hạn chế. SVC và KNN đều đạt độ chính xác 69%. Trong đó, SVC có F1-score cho lớp đạt là 0.79 và độ nhạy là 87%, cao hơn so với KNN, cho thấy khả năng phân loại của SVC ổn định hơn trong môi trường dữ liệu mất cân bằng. Tuy nhiên, cả hai mô hình đều có hiệu suất kém khi phân loại lớp không đạt, với độ nhạy lần lượt là 34% và 41%. MLP đạt độ chính xác 69%, với F1-score lớp đạt là 0.77. Mặc dù độ nhạy lớp đạt đạt 81%, hiệu suất dự đoán lớp không đạt chỉ đạt 45%, phản ánh sự hạn chế của mô hình trong việc nhận diện các trường hợp thất bại. Logistic Regression đạt độ chính xác thấp nhất, chỉ 67%, với F1-score lớp đạt là 0.79 nhưng độ nhạy lớp không đạt chỉ đạt 14%, điều này cho thấy mô hình gặp khó khăn đáng kể trong việc dự đoán chính xác các trường hợp không đạt điểm chuẩn.

Kết quả cho thấy, mặc dù các mô hình như Random Forest, Gradient Boosting, và SVC đạt hiệu năng cao, chúng vẫn gặp khó khăn khi phân loại lớp không đạt (lớp 0), do sự mất cân bằng trong phân phối dữ liệu (số lượng sinh viên đạt điểm đạt vượt trội so với không đạt). Điều này được minh họa qua độ nhạy và độ chính xác thấp ở lớp không đạt trong các mô hình. Ngược lại, mô hình Logistic Regression, dù đơn giản, không thể xử lý tốt các đặc điểm phức tạp của dữ liệu. Kết quả này cũng nhấn mạnh rằng việc sử dụng các giải thuật máy học tiên tiến như Random Forest và Gradient Boosting là lựa chọn tối ưu để dự đoán kết quả học tập. Đồng thời, việc xử lý sự mất cân bằng trong dữ liệu thông qua các kỹ thuật như resampling hoặc điều chỉnh trọng số có thể cải thiện hiệu quả phân loại ở cả hai lớp.

Hình 02 cho thấy sơ đồ ma trận nhầm lẫn (Confusion Matrix) của các giải thuật đã sử dụng. Mỗi biểu đồ thể hiện: Trục hoành: Nhãn dự đoán (Predicted Labels); Trục tung: Nhãn thực tế (True Labels). Các ô trong ma trận: Class 0 (Not Pass): Số lượng dự đoán không đạt đúng và sai; Class 1 (Pass): Số lượng dự đoán đạt đúng và sai. Sơ đồ cho thấy hiệu suất của từng giải thuật thông qua các giá trị dự đoán chính xác và nhầm lẫn.

Kết quả từ các ma trận nhầm lẫn của các mô hình học máy cho thấy những khác biệt rõ rệt trong khả năng phân loại giữa các lớp điểm đạt (Class 1) và điểm không đạt (Class 0). Đối với mô hình Logistic Regression, mặc dù độ chính xác tổng thể đạt 67%, mô hình này gặp khó khăn trong việc phân loại chính xác các trường hợp không đạt (Class 0), với tỷ lệ dự đoán sai khá cao (84 trường hợp được dự đoán là đạt

nhưng thực tế là không đạt). Tuy nhiên, mô hình này thể hiện khả năng phân loại Class 1 tương đối tốt, với 179 trường hợp đạt được dự đoán đúng. Tương tự, mô hình SVC đạt độ chính xác tổng thể là 69%, nhưng hiệu suất đối với Class 0 chỉ đạt 34% recall, cho thấy khả năng nhận diện các trường hợp không đạt vẫn còn hạn chế, mặc dù mô hình vẫn giữ được độ chính xác cao với Class 1 (87% recall). Mô hình K-Nearest Neighbors (KNN) cũng có kết quả tương tự với SVC, với độ chính xác tổng thể đạt 69%, nhưng hiệu suất phân loại Class 0 vẫn chưa được cải thiện rõ rệt. Mặt khác, Random Forest thể hiện ưu thế vượt trội với độ chính xác tổng thể lên tới 73%, nhờ vào khả năng phân loại tốt cả hai lớp, đặc biệt là Class 0 (recall đạt 44%), tuy vẫn còn một số nhầm lẫn. Mô hình Gradient Boosting đạt độ chính xác là 70%, với kết quả tốt ở Class 1 (82% recall), tuy nhiên, hiệu suất nhận diện Class 0 vẫn chưa cải thiện đáng kể so với Random Forest. Cuối cùng, MLP Classifier (Neural Network) cũng có độ chính xác tổng thể tương tự Gradient Boosting (69%), nhưng recall cho Class 0 vẫn còn thấp, cho thấy mô hình này cũng gặp khó khăn trong việc phân loại các trường hợp không đạt. Nhìn chung, Random Forest cho thấy hiệu suất tốt nhất trong việc phân loại cả hai lớp điểm đạt và không đạt, trong khi các mô hình khác như Logistic Regression, SVC, và KNN gặp nhiều khó khăn trong việc nhận diện chính xác Class 0. Những kết quả này chỉ ra rằng việc tối ưu hóa các mô hình để cải thiện độ chính xác đối với Class 0 sẽ là một hướng quan trọng trong các nghiên cứu tiếp theo.

Figure 2. Confusion Matrices

5 TỔNG KẾT

Bài toán dự đoán kết quả labFinalScore của sinh viên là một bài toán phân loại nhị phân, với mục tiêu xác định xem sinh viên có đạt điểm trong kỳ thi cuối kỳ (điểm >=5) hay không (điểm <5). Với dữ liệu thu thập từ các hoạt động thực hành (Prelab, Inlab, Postlab), mỗi sinh viên có thể tham gia nhiều lần vào các bài tập để cải thiện điểm số. Do đó, chúng ta chỉ chọn điểm cao nhất của sinh viên cho mỗi chủ đề, kết hợp với số lần tham gia và tổng thời gian dành cho mỗi bài tập, để làm các đặc trưng trong mô hình dự đoán.

Các mô hình học máy được áp dụng trong nghiên cứu bao gồm Logistic Regression, SVC, KNeighborsClassifier, Random Forest, Gradient Boosting, và MLP Classifier. Kết quả thực nghiệm cho thấy mô hình Random Forest đạt độ chính xác cao nhất (73%) và hiệu suất tốt trong việc phân loại sinh viên đạt (Class 1) với recall lên đến 88%. Tuy nhiên, việc phân loại sinh viên không đạt (Class 0) vẫn là một thách thức lớn, với các mô hình như Logistic Regression và SVC có recall cho Class 0 khá thấp (14% và 34% tương ứng).

Một trong những yếu tố quan trọng ảnh hưởng đến kết quả dự đoán là sự phân bố không đều của các lớp trong tập dữ liệu. Lớp sinh viên đạt (Class 1) chiếm tỷ lệ cao hơn nhiều so với lớp không đạt (Class 0), dẫn đến các mô hình thiên về dự đoán lớp đạt. Điều này phản ánh trong việc các mô hình đạt recall cao đối với Class 1 nhưng lại gặp khó khăn trong việc phân loại đúng nhóm sinh viên không đạt.

Mặc dù các mô hình học máy hiện tại đã cho kết quả khả quan, tuy nhiên, vẫn còn nhiều không gian để cải thiện, đặc biệt là trong việc phân loại sinh viên không đạt. Việc điều chỉnh các tham số của mô hình,

kết hợp với việc bổ sung các yếu tố khác như thái độ học tập hoặc đặc điểm học tập của sinh viên, có thể cải thiện hiệu suất của các mô hình. Thêm vào đó, việc xử lý mất mát dữ liệu và điều chỉnh cân bằng giữa các lớp (ví dụ, sử dụng phương pháp sampling hoặc điều chỉnh trọng số lớp) sẽ giúp cải thiện hiệu quả phân loại cho cả hai lớp đạt và không đạt hoặc có thể dùng các phương pháp học sâu như Gan để có thể sinh ra dữ liệu giả để train để tăng độ chính xác của tập test. Mô hình Gan giúp cho dữ liêu được cân bằng cho tập train giúp dữ liệu được cân bằng cải thiện độ chính xác của lớp thiểu số.

Kết quả của bài toán dự đoán này có thể ứng dụng trong việc hỗ trợ giảng viên trong việc đánh giá và can thiệp kịp thời đối với sinh viên, giúp nâng cao chất lượng giảng dạy và học tập. Bài toán này cũng có thể mở rộng để dự đoán kết quả của các kỳ thi khác trong quá trình học tập của sinh viên, từ đó cung cấp thông tin dự báo hữu ích cho công tác quản lý giáo dục.

REFERNCES

[1] Brazhkin, Vitaly and Strakos, Joshua K., "Student Preferences for Multiple Attempts and Feedback on Online Quantitative Assessments," *Intersection: A Journal at the Intersection of Assessment and Learning,* vol. 4, no. 2, 2023.

[2] M. Kumar, C. Sharma, S. Sharma, N. Nidhi and N. Islam, "Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance," *2022 International Conference on Decision Aid Sciences and Applications (DASA),* pp. 1013-1017, 2022.

[3] Au, T.-W. and Salihin, R. and Saiful, O., "Performance Prediction of Learning Programming - Machine Learning Approach," in *30th International Conference on Computers in Education Conference, ICCE 2022 - Proceedings*, 2022.

[4] Su, Y. S., Lin, Y. D., & Liu, T. Q, "Applying machine learning technologies to explore students' learning features and performance prediction," *Frontiers in neuroscience,* 2022.

[5] A. I. Al-Alawi and N. M. A. Alsubaiee, "Predicting Student's Academic Performance Using Data Mining Methods: Review Paper," *2023 International Conference On Cyber Management And Engineering (CyMaEn),* pp. 18-23, 2023.

[6] V. D. Nguyen, T. V. Nguyen and P. V. Ha, "Early educational performance prediction a deep learning approach," *Journal of Science and Technology - Hanoi University of Industry,* vol. 58, pp. 37-41, 2022.

[7] Yağcı, M., "Educational data mining: prediction of students' academic performance using machine learning algorithms.," *Smart Learning Environments,* vol. 9, no. 11, 2022.

[8] Sivasakthi M and Pandiyan M, "Machine Learning Algorithms to Predict Students' Programming Performance: A comparative Study," *Journal of University of Shanghai for Science and Technology ,* vol. 24, no. 6, 2022.

[9] Llanos Mosquera, José & Bucheli, Victor & Restrepo-Calle, Felipe., "Early prediction of student performance in CS1 programming courses.," *PeerJ Computer Science,* vol. 9, 2023.

[10] Chen S, Ding Y., " Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools.," *Social Sciences.,* vol. 12, no. 3, 2023.

[11] Niu, Yajing & Zhou, Tao & Li, Zhigang & Liu, Haochen., "Student Grade Prediction Model Based on RFE_RF and Integrated Learning Voting Algorithm.," *Proceedings of the 2023 2nd International Conference on Educational Innovation and Multimedia Technology (EIMT 2023),* 2023.

[12] Sakshi, Gautam, S., Sharma, C., Kukreja, V., "Handwritten Mathematical Symbols Classification Using WEKA," *Applications of Artificial Intelligence and Machine Learning,* vol. 778, 2021.

[13] Al-Shehri, Huda & Al-Qarni, Amani & Al-Saati, Leena & Batoaq, Arwa & Badukhen, Haifa & Alrashed, Saleh & Alhiyafi, Jamal & Olatunji, Sunday, "Student performance prediction using Support Vector Machine and K-Nearest Neighbor.," *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE),* 2017.

[14] David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, Inc., 2013.