Dự đoán kết quả học tập của học sinh dựa trên mô hình máy học có giám sát

Nguyễn Nhứt Lam*

**Trường Đại học Trà Vinh*

Received: 5/6/2023; Accepted: 12/6/2023; Published: 19/6/2023

Abstract: *Predicting student learning outcomes is an important area of research in the field of education. The purpose of this paper is to propose a learning machine model that can reliably predict students' learning outcomes. The model is based on a set of attributes related to students such as family, school, and previous learning outcomes. The experimental results of the paper show that the Random Forest algorithm is very effective in predicting the learning outcomes of with a Mean Absolute Error (MAE) of 1.13. Besides, attributes such as family size, student age, school and reasons for choosing a school are important factors affecting the prediction results. The proposed predictive model could help educators identify students at risk of low academic achievement and take appropriate remedial measures to improve their academic performance. Furthermore, predicting academic performance also helps students and parents make the necessary adjustments to improve academic achievement.*

Keywords: *Prediction, learning outcomes, students, supervised learning machine model*

1. Giới thiệu

Ngày nay với sự phát triển mạnh mẽ của trí tuệ nhân tạo, ứng dụng của lĩnh vực này, cụ thể là máy học, ngày càng được mở rộng trong nhiều lĩnh vực khác nhau như giáo dục, y tế, giao thông thông minh. Trong lĩnh vực giáo dục, dự đoán kết quả học tập của HS là một trong những vấn đề rất được quan tâm nghiên cứu của các học giả từ trước đến nay. Dự đoán kết quả học tập của HS giúp GV và nhà trường có những điều chỉnh thích hợp nhằm phát huy hiệu quả năng lực của HS cũng như có những hỗ trợ phù hợp và đúng thời điểm. HS dựa trên kết quả dự đoán cũng tự đánh giá được năng lực của bản thân. Từ đó, HS sẽ có những điều chỉnh phù hợp nhằm phát huy tối đa năng lực của bản thân, khắc phục những nhược điểm để hoàn thành mục tiêu học tập của mình.

Nghiên cứu về các yếu tố ảnh hưởng đến kết quả học tập của HS cho thấy có nhiều yếu tố ảnh hưởng đến thành tích học tập của HS. Các yếu tố liên quan đến gia đình như quy mô gia đình, kiểu gia đình có ảnh hưởng lớn đến thành tích học tập của HS [1]. Tác giả bài báo này cho rằng đây là các thuộc tính quan trọng đối với mô hình dự đoán. Ngoài ra, các yếu tố như thái độ học tập của HS, phương pháp giảng dạy, phương tiện học tập, môi trường học tập cũng ảnh hưởng đến kết quả học tập của HS.

Trong bài báo này, tác giả đề xuất một mô hình dự đoán kết quả học tập của HS. Tác giả sử dụng các thuộc tính như yếu tố gia đình, môi trường học tập, tuổi HS và một số thuộc tính khác để huấn luyện mô hình dự đoán. Thực nghiệm về tầm quan trọng của các thuộc tính được sử dụng cũng được thực hiện để xác định các yếu tố quan trọng ảnh hưởng đến kết quả học tập của HS.

2. Nội dung nghiên cứu

2.1. Nghiên cứu liên quan

Dự đoán kết quả học tập của HS đã là một chủ đề được nghiên cứu rộng rãi trong lĩnh vực giáo dục. Nhiều nhà nghiên cứu đã khám phá các kỹ thuật và phương pháp khác nhau để dự đoán kết quả học tập của HS thông qua các yếu tố khác nhau. Bài báo [2] đề xuất mô hình khai phá dữ liệu để dự đoán kết quả học tập của sinh viên tại các trường đại học ở Bulgaria. Kết quả nghiên cứu cho thấy mô hình Random Forest (RF) và mô hình mạng thần kinh đạt hiệu quả tốt khi dự đoán kết quả học tập.

Nghiên cứu [3] khai thác các yếu tố liên quan đến HS như nhân khẩu học, đặc điểm hành vi và thực hiện huấn luyện cá mô hình máy học có giám sát khác nhau như Suport Vector Machine (SVC), K-nearest neighbor (KNN), Decision Tree (DT), Logistic Regression (LR). Dữ liệu huấn luyện mô hình dự đoán được thu thập trong hai năm học liên tiếp tại Trường Đại học Basra, Irad. Kết quả thực nghiệm chỉ ra rằng LR đáng tin cậy nhất.

Nghiên cứu [4] ứng dụng các mô hình máy học khác nhau để dự đoán kết quả học tập của HS dựa trên tập dữ liệu thu thập từ hoạt động của HS trên hệ

thống Moodle. Kết quả so sánh hiệu quả của bảy mô hình máy học khác nhau cho thấy RF đạt hiệu quả dự đoán tốt nhất.

2.2. Giải pháp đề xuất dự đoán kết quả học tập của học sinh dựa trên mô hình máy học có giám sát

2.2.1. Cơ sở dữ liệu

Trong bài báo này, tác giả thu thập dữ liệu từ UCI Machine Learning Reposity [1]. Dữ liệu được thu thập là dữ liệu về thành tích học tập của HS trung học của hai trường học ở Bồ Đào Nha. Dữ liệu được trích xuất từ các báo cáo và bảng hỏi liên quan đến kết quả học tập của HS đối với hai môn học là Toán và Tiếng Bồ Đào Nha. Trong bài báo này, tác giả chỉ sử dụng dữ liệu về môn Toán. Bộ dữ liệu bao gồm 33 thuộc tính và 395 mẫu tin.

2.2.2. Mô hình máy học.

Các mô hình máy học có giám sát sau đây được sử dụng để huấn luyện các mô hình dự đoán kết quả học tập của HS.

1) K-nearest Neighbors: KNN là một thuật toán máy học có giám sát phi tham số được sử dụng cho các bài toán phân loại và hồi quy. Thuật toán này thường được gọi là "lazy learning" bởi vì nó không học gì từ dữ liệu huấn luyện mà đơn giản là lưu trữ dữ liệu huấn luyện này. Việc tính toán được thực hiện khi có dữ liệu mới cần dự đoán. Trong bài báo này, kết quả dự đoán của một mẫu dữ liệu mới là giá trị trung bình của các láng giềng gần nhất.

2) Random Forest: RF là một thuật toán máy học kết hợp sử dụng cây quyết định để cải thiện độ chính xác. RF kết hợp các kết quả từ nhiều cây quyết định được đào tạo bằng cách sử dụng các tập hợp con khác nhau của dữ liệu đào tạo. Trong bài báo này, kết quả dự đoán của một mẫu dữ liệu mới là giá trị trung bình của các kết quả dự đoán từ các cây quyết định này.

3) Support Vector Machine: SVM là một trong những thuật toán máy học được sử dụng phổ biến. SVM có thể được sử dụng cho cả bài toán phân loại và hồi quy. Thuật toán sử dụng tập huấn luyện để tìm ra các "hyperplanes" phù hợp nhất với các điểm dữ liệu.

2.2.3. Đánh giá mô hình

Trong bài báo này, tác giả sử dụng các tiêu chí sau để đánh giá hiệu quả của mô hình dự đoán.

1) Mean Squared Error (MSE): Đây là một phương pháp đánh giá mô hình hồi quy được sử dụng rất phổ biến. MSE là giá trị trung bình của bình phương sai số giữa giá trị thực tế và giá trị dự đoán. Giá trị MSE được tính theo công thức sau:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y'_i)^2$$

Miền giá giá trị của MSE là [0,+∞]. Giá trị của MSE càng nhỏ thì mô hình càng hiệu quả trong việc dự đoán.

2) Mean Absolute Error (MAE): MAE là phướng pháp đánh giá mô hình hồi quy dựa trên giá trị trung bình của giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán. Giá trị MAE được tính toán theo công thức sau:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y'_i|$$

Tương tự như MSE, giá trị của MAE thuộc miền giá trị [0,+∞]. Giá trị của MAE càng nhỏ thì mô hình hồi quy càng có độ chính xác cao.

2.3. Kết quả và thảo luận

Để đánh hiệu quả của các mô hình dự đoán đối với tập dữ liệu đã thu thập, tập dữ liệu được chia thành hai phần: tập huấn luyện 70% và tập dữ liệu dùng để kiểm thử và đánh giá độ chính xác của mô hình 30%. Tập huấn luyện là tập dữ liệu dùng để huấn luyện mô hình máy học hay nói cách khác là máy sẽ học từ dữ tập dữ liệu này. Tập dữ liệu huấn luyện thông thường có kích thước lớn hơn tập học. Sau khi mô hình máy học đã được huấn luyện, chúng ta cần phải đánh giá hiệu quả hay độ chính xác của mô hình để kết luận mô hình máy học tương ứng với một bài toán cụ thể có đáng tin cậy hay không. Tập dữ liệu này được gọi là tập kiểm thử. Tập kiểm thử được đưa vào mô hình máy học sau huấn luyện như là đầu vào. Dựa vào kết quả dự đoán và giá trị thực tế của của các mẫu được kiểm thử chúng ta đánh giá độ tin cậy, so sánh hiệu quả của các mô hình khác nhau và từ đó chọn mô hình thích hợp cho bài toán. Tập dữ liệu kiểm thử thường có kích thước nhỏ hơn tập dữ liệu huấn luyện.

Để cài đặt các mô hình huấn luyện được đề xuất trong bài báo này, tác giả sử dụng ngôn ngữ lập trình Python. Python là ngôn ngữ lập trình được sử dụng phổ biến hiện nay để giải quyết các bài toán trong lĩnh vực trí tuệ nhân tạo và khoa học dữ liệu. Các thuật toán máy học sử dụng trong bài báo được nhập từ thư viện Scikit-learn[1]. Scikit-learn là thư viện mã nguồn mở chứa tập các mô hình máy học được cài đặt bằng ngôn ngữ lập trình Python. Scikit-learn cung cấp các mô hình máy học phân loại, hồi quy và phân cụm phổ biến như KNN, RF và SVM. Việc cài đặt và huấn luyện các mô hình máy học đề xuất trong bài

1. https://scikit-learn.org

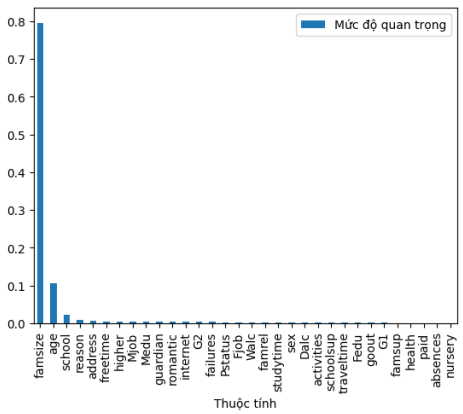
báo, tác giả sử dụng công cụ Google Colab[2]. Colab cho phép lập trình Python trên trình duyệt web mà không đòi hỏi máy tính có cấu hình cao. Colab cho phép truy cập miễn phí vào GPU (bộ xử lý đồ hoạ) và dễ dàng chia sẽ mã nguồn (code).

Tập đặc trưng sử dụng cho mô hình dự đoán bao gồm tất cả 33 thuộc tính của bộ dữ liệu. Yêu cầu của các mô hình là giúp dự đoán điểm môn Toán cuối khoá của HS. Điểm môn Toán của HS có miền giá trị từ [0, 20]. Kết quả thực nghiệm trên tập kiểm thử được trình bày ở bảng 2.1. Như trình bày trong bảng bên dưới, kết quả dự đoán của cả ba mô hình đều tương đối hiệu quả. Đặc biệt, mô hình RF cho kết quả vượt trội so với hai mô hình còn lại với độ lỗi MAE là 1.13 và MSE là 3.45. Kết quả này chứng minh rằng mô hình máy học RF rất đáng tin cậy trong việc dự báo điểm môn Toán của HS trung học trong tập dữ liệu thực nghiệm.

Bảng 2.1. Kết quả dự đoán của các mô hình máy học

Mô hình	MAE	MSE
RF	1.13	3.45
SVM	1.97	10.13
KNN	2.49	11.85

Mức độ quan trọng (sự ảnh hưởng) của các thuộc tính đến việc dự đoán điểm môn Toán của HS cũng được đánh giá dự trên thuật toán RF. Kết quả đánh giá được thể hiện trên hình 2.1. Hình 2.1 cho thấy thuộc tính familysize (quy mô gia đình) có mức độ quan trọng cao nhất (0.80), tiếp theo là tuổi của HS (mức độ quan trọng là 0.1). Trường học của HS của ảnh hưởng đến kết quả học tập của sinh viên. Giá trị của mức độ quan trọng của thuộc tính này là 0.02. Nguyên nhân chọn trường học như nhà gần trường hay độ tin cậy của trường cũng là một yếu tố quan trọng ảnh hưởng kết quả dự đoán.



Hình 2.1. Tầm quan trọng cuả thuộc tính

2. https://colab.research.google.com/

3. Kết luận

Bài báo đề xuất một mô hình máy học có giám sát để dự đoán kết quả học tập của HS. Mô hình được huấn luyện dựa trên tập các thuộc tính liên quan đến HS như yếu tố gia đình, trường học, kết quả học tập trước đó. Kết quả thực nghiệm chứng minh thuật toán RF rất hiệu quả trong việc dự đoán kết quả học tập của HS với độ lỗi MAE là 1.13 và MSE là 3.45. Các thuộc tính như quy mô gia đình, độ tuổi của HS, trường học của HS và lí do chọn trường học là các yếu tố quan trọng ảnh hưởng kết quả dự đoán. Mô hình dự đoán đề xuất có thể giúp các nhà giáo dục xác định những HS có nguy cơ đạt thành tích học tập thấp và đưa ra các biện pháp khắc phục phù hợp để cải thiện kết quả học tập.

Tài liệu tham khảo

[1] Marks, Gary N. "*Family size, family type and student achievement: Cross-national differences and the role of socioeconomic and school factors." Journal of comparative family studies* 37.1 (2006): 1-24.

[2] Kabakchieva, Dorina. "Student performance prediction by using data mining classification algorithms." *International journal of computer science and management research* 1.4 (2012): 686-690.

[3] Hashim, Ali Salah, Wid Akeel Awadh, and Alaa Khalaf Hamoud. "Student performance prediction model based on supervised machine learning algorithms." *IOP Conference Series: Materials Science and Engineering*. Vol. 928. No. 3. IOP Publishing, 2020.

[4] Tamada, Mariela Mizota, Rafael Giusti, and José Francisco de Magalhães Netto. "Predicting student performance based on logs in moodle LMS." *2021 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2021.

[5]. Okubo, Fumiya, Takayoshi Yamashita, Atsushi Shimada, and Hiroaki Ogata. (2017), A neural network approach for students' performance prediction. Proceedings of the seventh international learning analytics & knowledge conference.

[6]. Amra, Ihsan A. Abu, and Ashraf YA Maghari (2017), Students performance prediction using KNN and Naïve Bayesian. 2017 8th international conference on information technology (ICIT). IEEE.

[7]. Mohammadi, Mehdi, et al (2019), Comparative study of supervised learning algorithms for student performance prediction. 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE.