So sánh J48 và Naive Bayes trong phân tích dữ liệu giáo dục

Đỗ Quỳnh Anh*

**ThS. Công nghệ thông tin, Trường Đại học Đồng Tháp*

Received: 10/01/2024; Accepted: 18/01/2024; Published: 22/01/2024

***Abstract**. In the present, data mining can be applied in various fields. One of them is the field of education. By applying data mining in the field of education, education providers can make an analysis of students in their schools. Schools can predict student achievement, make assessments of students more thoroughly, and can also predict students' interests and talents. This study will discuss the prediction of student learning habits and the prediction of student achievement in order to find out the right steps to take afterwards. In this study, two (2) classification algorithms were used, namely J48 and NaiveBayes. This is done to find the best results from each algorithm.*

***Keywords:** J48 and Naive Bayes, educational data analysis*

1. Đặt vấn đề

Trong thời đại này, mọi lĩnh vực đều bắt buộc phải tiến hành phân tích, dự đoán một cách nhanh chóng và chính xác. Một trong số các lĩnh vực này là lĩnh vực giáo dục. Nâng cao chất lượng giáo dục là một trong những vấn đề quan trọng nhất được nhiều bên quan tâm vì nó có vai trò quan trọng đối với sự tiến bộ trong tương lai [1]. Để có được điều tốt nhất phân tích, kỹ thuật khai thác dữ liệu được sử dụng. Khai thác dữ liệu là một kỹ thuật phân tích được sử dụng để có thể tìm hiểu sâu hơn về dữ liệu thô [2] [3]. Khai thác dữ liệu giáo dục (EDM) là một môn khoa học sử dụng các kỹ thuật khai thác dữ liệu trong giáo dục [4]. EDM trong giáo dục rất hữu ích trong quá trình dự đoán thành tích của học sinh, đánh giá học sinh kỹ lưỡng hơn, dự đoán sở thích và tài năng và nhiều phân tích khác [5] [6] [7]. Khai thác dữ liệu được sử dụng rộng rãi vì nó rất hữu ích trong kiểm tra dữ liệu bằng nhiều cách tiếp cận khác nhau và có những đặc điểm riêng. Dữ liệu khai thác cũng được sử dụng để đơn giản hóa dữ liệu thành thông tin chức năng. Các phương pháp khai phá dữ liệu được sử dụng rộng rãi được sử dụng trong EDM là k-hàng xóm gần nhất, cây quyết định, mạng lưới thần kinh, Bayes ngây thơ, v.v. [8] [9] đều có thể thực hiện phân tích, nhiều công cụ nguồn mở có thể được sử dụng để triển khai khai thác dữ liệu. Những công cụ bao gồm WEKA. Những công cụ này được thiết kế để có thể thực hiện điều tra dữ liệu và nhận được các mô hình hoặc cấu trúc có thể hữu ích trong tương lai [10].

Trong các nghiên cứu trước đây, nhiều người đã tiến hành nghiên cứu về khai phá dữ liệu với mục tiêu chính là dự đoán thành tích của học sinh. Trong nghiên cứu, thuật toán k-gần nhất có vai trò hiệu quả nhất với độ chính xác phân loại. Trong nghiên cứu về dự đoán kết quả học tập của học sinh trong trường và sử dụng một số thông số như điểm danh và giá trị bài tập. Nghiên cứu này sử dụng Thuật toán Naïve Bayes và cho ra độ chính xác cao nhất so với các thuật toán phân loại khác. Sau đó, trong nghiên cứu phương pháp Cây quyết định và Mạng lưới thần kinh được tham khảo để dự đoán thành tích của học sinh vì họ có điểm chính xác cao, hiệu suất được đánh giá bằng phương pháp Cây quyết định. Kết quả là mô hình này chỉ có thể sản xuất giá trị chính xác là 60%. Trong nghiên cứu đo lường thành tích học sinh bằng cách sử dụng phương pháp Cây quyết định và Mạng lưới thần kinh. Kết quả là, nghiên cứu này cho thấy hiệu quả của áp dụng phương pháp trong EDM cao hơn. Trong nghiên cứu tương đối nhằm phân tích một số Tạp chí Vật lý: Chuỗi hội nghị 1933 (2021) 012062 IOP Publishing doi:10.1088/1742-6596/1933/1/012062 2 Phương pháp Cây quyết định và tác dụng của chúng đối với các bộ dữ liệu giáo dục đã được thực hiện. Kết quả cho thấy các phương pháp phân tích hồi quy và phân loại là sự kết hợp tốt nhất vì chúng có mức độ tương thích cao để cho ra kết quả tốt hơn.

2. Nội dung nghiên cứu

2.1. Mô tả tập dữ liệu

Trong nghiên cứu này, tập dữ liệu này được lấy từ hệ thống quản lý học tập (LMS) của một trường đại học tư thục ở Jakarta, Indonesia. Bộ dữ liệu bao gồm 340 hàng dữ liệu và 10 thuộc tính. 10 thuộc tính được sử dụng bao gồm ba loại, đó là:

a) Nhân khẩu học, cụ thể là các thuộc tính về giới

tính, quốc tịch và nơi cư trú;

b) Trình độ học vấn, cụ thể là điểm trung bình khi vào đại học, điểm trung bình cuối năm ở trường đại học, học kỳ, chuyên ngành;

c) Thói quen như tham gia khảo sát phụ huynh, thói quen mở tài liệu, thói quen trả lời diễn đàn;

2.2. Kỹ thuật phân loại

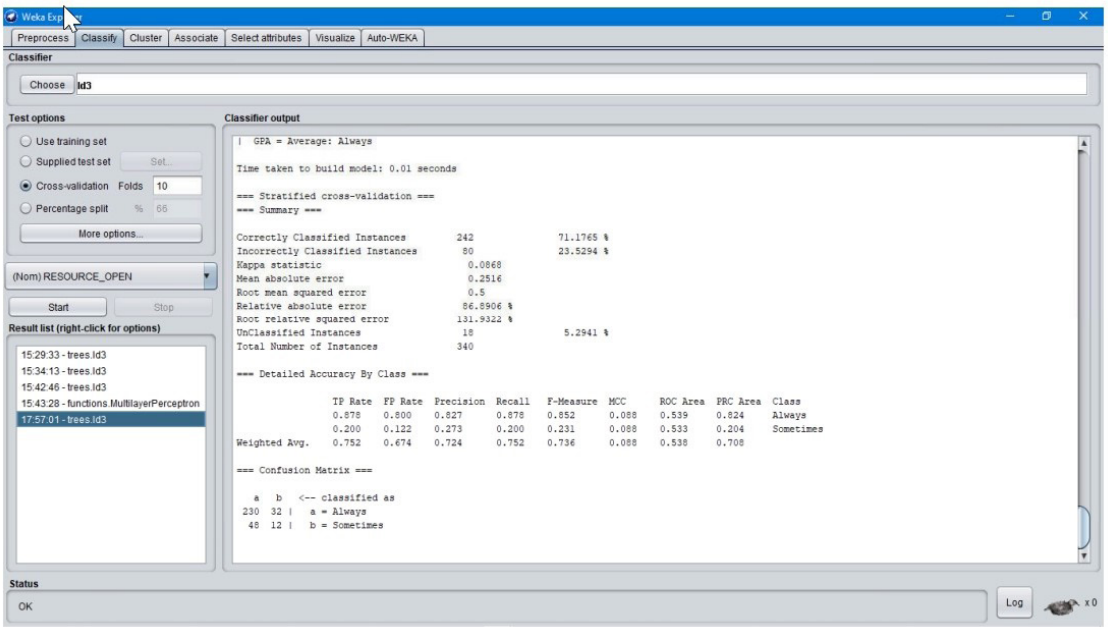
Phân loại là một kỹ thuật khai thác dữ liệu được sử dụng rộng rãi vì nó khá đơn giản. Có hai giai đoạn trong kỹ thuật phân loại, đó là phát triển mô hình để huấn luyện và đánh giá mô hình bằng dữ liệu huấn luyện. Phân loại cũng có nhiều phương pháp, ví dụ như thuật toán thống kê, phân tích tương quan, phân tích hồi quy, mô hình Bayes, thuật toán dựa trên khoảng cách, cách tiếp cận đơn giản, k-láng giềng gần nhất, cây quyết định, mạng lưới thần kinh và thuật toán dựa trên quy tắc. Trong nghiên cứu này, các phương pháp được sử dụng là mô hình Bayesian, cây quyết định và mạng lưới thần kinh.

a) Mô hình Bayes là phương pháp sử dụng kỹ thuật ứng dụng xác suất vào dữ liệuhoặc nó còn được gọi là kỹ thuật thường xuyên. Các tính toán sử dụng phương pháp này mang lại giá trị trực tiếp cho xác suất giả thuyết.

b) Cây quyết định là một phương pháp có điều kiện xây dựng giống như một cây. Để có thể sử dụng phương pháp này cần có hai (2) bước là xây dựng Cây quyết định và sau đó triển khai vào cơ sở dữ liệu.

2.3. Kết quả và thảo luận

Sau khi xử lý dữ liệu trên hai (2) thuật toán khác nhau trong các phương pháp phân loại (J48 và Naive Bayes), mỗi thuật toán đều có những đặc điểm riêng. Trong nghiên cứu này, các giá trị trọng tâm là CC (Các trường hợp được phân loại chính xác), IC (Các trường hợp được phân loại không chính xác), các giá trị Độ chính xác, Thu hồi và FMeasure.



Hình 2.1. So sánh kết quả thuật toán J48 và ID3

2.3.1. Thuật toán J48 hoặc ID3

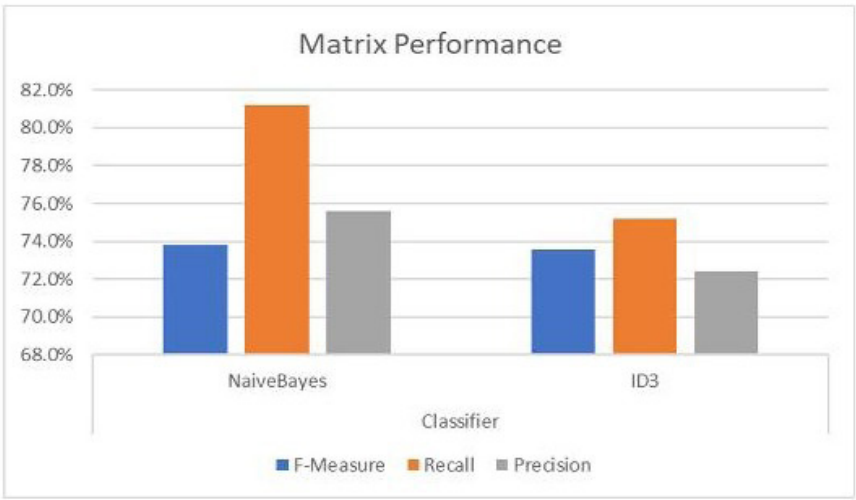
Thuật toán J48 là sự phát triển của thuật toán ID3, là một thuật toán thông thường. Thuật toán này có thể phân loại dữ liệu số và dữ liệu rời rạc bằng phương pháp Cây quyết định. Bằng cách sử dụng thuật toán này, giá trị Phiên bản được phân loại chính xác là 71,17% và độ chính xác hoặc F-Measure là 73,6%.

2.3.2. Thuật toán Naïve Bayes

Thuật toán Naïve Bayes là một thuật toán trong phương pháp phân loại. Thuật toán này sử dụng xác suất và tính toán thống kê. Bằng cách sử dụng thuật toán này, chúng tôi nhận được giá trị Trường hợp được phân loại chính xác là 81,17% và độ chính xác hoặc F-Measure là 73,8%.3.6 Kết quả thực hiện Kết quả xử lý dữ liệu bằng năm (5) thuật toán khác nhau này như sau.

Có thể thấy năm thuật toán này có tỷ lệ chính xác khá tốt (trên 70%). Giá trị chính xác cao nhất được tìm thấy trong thuật toán NaïveBayes (73,8%), sau đó là ID3 (73,6%). Trong khi đó, đối với CC hoặc Các trường hợp được phân loại chính xác, giá trị cao nhất cũng được tìm thấy ở thuật toán NaïveBayes

(276 dữ liệu), sau đó là ID3 (242 dữ liệu). Nhìn chung, thuật toán Naive Bayes có giá trị độ chính xác dựa trên F-Measure, Recall và Precision cao nhất so với các thuật toán khác.



Hình 2.2. So sánh thuật toán NaiveBayes và thuật toán ID3

3. Kết luận

Trong thế giới giáo dục, việc khai thác dữ liệu là cần thiết để thực hiện phân tích dự đoán về tập dữ liệu của học sinh. Một trong những thông tin hoặc kiến thức có thể được tạo ra từ kỹ thuật khai thác dữ liệu là dự đoán về kết quả học tập và thành tích của học sinh, sau đó sẽ được sử dụng để quyết định các bước tiếp theo cho những học sinh này. Trong nghiên cứu này, tập dữ liệu được sử dụng là dữ liệu được lấy từ hệ thống quản lý học tập (LMS) của trường đại học. Tập dữ liệu này có 340 hàng dữ liệu và 10 thuộc tính. Để xử lý, hai (2) thuật toán phân loại được sử dụng là J48 và NaïveBayes. Hai thuật toán này có những đặc điểm khác nhau. Kết quả, hai thuật toán đều cho độ chính xác tốt (>70%). Tuy nhiên, thuật toán Naive Bayes có giá trị độ chính xác cao nhất (dựa trên F-Measure, Recall và Precision) và giá trị Phiên bản được phân loại chính xác cao nhất so với các thuật toán khác. Để nghiên cứu sâu hơn, để sử dụng nhiều dòng dữ liệu hơn và cũng thử sử dụng các kỹ thuật khai thác dữ liệu khác.

Tài liệu tham khảo

[1] Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In 2017 8th International Conference on Information Technology (ICIT) (pp. 909-913). IEEE.

[2] Ramaphosa, K. I. M., Zuva, T., & Kwuimi, R. (2018, August). Educational data mining to improve learner performance in Gauteng primary schools. In 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-6). IEEE.

[3] Kumar, A. D., Selvam, R. P., & Kumar, K. S. (2018). Review on prediction algorithms in educational data mining. International Journal of Pure and Applied Mathematics, 118(8), 531-537.

[4] Triayudi, A., Sumiati, S., Nurhadiyan, T., & Rosalina, V. (2020). Data Mining Implementation to Predict Sales Using Time Series Method. Proceeding of the Electrical Engineering Computer Science and Informatics, 7(2), 1-6.

[5] Jalota, C., & Agrawal, R. (2019, February). Analysis of educational data mining using classification. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 243-247). IEEE.

[6] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. Ieee Access, 5, 15991-16005.

[7] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.

[8] Fitri I, Triayudi A, Iksal, Muttaqin Z, Sumiati. Visualization of Data Mining Distribution of COVID-19 in Indonesia Using Self-Organizing Maps Algorithm. Icic Express Letters. 2021, Vol. 15 (3), pp. 241-248.

[9] Rawat, K. S., & Malhan, I. V. (2019). A hybrid classification method based on machine learning classifiers to predict performance in educational data mining. In Proceedings of 2nd International Conference on Communication, Computing and Networking (pp. 677-684). Springer, Singapore.

[10] Hegde, V., & Prageeth, P. P. (2018, January). Higher education student dropout prediction and analysis through educational data mining. In 2018 2nd International Conference on Inventive Systems and Control (ICISC) (pp. 694-699). IEEE.