Equipment with new general education program, Volume 2, Issue 313 (May 2024)
ISSN 1859 - 0810

Ứng dụng thuật toán phân loại J48 trong dự đoán kết quả học tập của học sinh, sinh viên

Trịnh Anh Tuấn*

**Trường Đại Học Y Dược Hải Phòng*
Received: 26/4/2023; Accepted: 6/5/2023; Published: 15/5/2023

Abstract: *This paper sets out to use J48 classification algorithm to predict students' academic performance towards the end of the semester in the Data Structure course under the Computer Science Program. This algorithm aimed to help faculty in forecasting who among the students would likely to fail and who would make it until the end of the semester. In this way, the faculty could make remedial measures to help those struggling students pass the subject and advance to the next level, thus, increasing students' success rate and retention in a Higher Education Institutions (HEI). This research employed a descriptive correlational design using Exploratory Data Analysis (EDA) for Data Mining in testing and verifying data to generate new information. Data mining is part of the Knowledge Discovery in Databases (KDD) process where it follows six steps: data selection, data pre-processing, data transformation, data mining, interpretation, and knowledge discovery.*

Keywords**:** *Academic Performance, Classification Algorithm, Educational Data Mining, J48 Decision Tree*

1. Đặt vấn đề

Khai thác dữ liệu là việc trích xuất các mẫu hoặc kiến thức thú vị (không tầm thường, tiềm ẩn, chưa biết trước đây và có khả năng hữu ích) từ lượng dữ liệu khổng lồ. Sự sẵn có của lượng dữ liệu lớn này đã thúc đẩy nhà nghiên cứu chuyển đổi dữ liệu thô thành thông tin hữu ích mà các nhà giáo dục có thể sử dụng để đưa ra các quyết định sáng suốt và hành động phù hợp bằng cách sử dụng EDM.

Tác giả đã áp dụng một trong những kỹ thuật khai thác dữ liệu có tên là phân loại để dự đoán kết quả học tập của SV ở một trong những môn học khó nhất như Cấu trúc dữ liệu trong chương trình Khoa học Máy tính. Phân loại là một trong những kỹ thuật học có giám sát nhằm xây dựng mô hình phân loại một mục dữ liệu thành nhãn lớp được xác định trước. Mục đích là dự đoán sản lượng trong tương lai dựa trên dữ liệu có sẵn. Cấu trúc dữ liệu được nhà nghiên cứu coi là một chủ đề đang được nghiên cứu vì tỷ lệ thất bại cao. Dữ liệu được thu thập thông qua việc lấy hồ sơ lớp học từ một giảng viên phụ trách bộ môn trong học kỳ 2 của năm học 2014-2015 & 2015-2016 với 103 SV đăng ký khóa học. Do hầu hết các nghiên cứu tập trung nhiều hơn vào việc phân biệt và đánh giá các thuật toán phân loại khác nhau trong các lĩnh vực quan tâm khác nhau như y tế và giáo dục, nên các tác giả khác đã sử dụng hồ sơ nhân khẩu học xã hội của học sinh để dự đoán kết quả học tập của học sinh [5], và những tác giả khác vẫn sử dụng dữ liệu cho nữ SV, đóng góp của bài viết này là sử dụng thuật toán phân loại J48 (dựa trên so sánh hiệu suất được thực hiện bởi một số tác giả) trong việc tạo ra mô hình dự đoán để dự báo sự thành công hay thất bại của SV trong khóa học bằng cách sử dụng cả hồ sơ nam và nữ trong giáo dục đại học. Hơn nữa, nhà nghiên cứu phát hiện ra rằng chỉ có một số ít tài liệu về lĩnh vực này ở Philippines.

Được hướng dẫn bởi các khái niệm và phát hiện của nhiều nghiên cứu khác nhau, nghiên cứu này nhằm mục đích sử dụng thuật toán phân loại để xây dựng một mô hình giúp dự đoán kết quả học tập của học sinh trong môn Cấu trúc dữ liệu và cuối cùng sẽ cung cấp những hiểu biết sâu sắc hữu ích cho các nhà giáo dục trong việc thiết kế lại phương pháp sư phạm cho những học sinh gặp khó khăn.

2. Nội dung nghiên cứu

Nghiên cứu này nhằm tạo ra mô hình cây quyết định dự đoán sẽ dự báo các yếu tố có ảnh hưởng lớn đến kết quả học tập của SV trong khóa học Cấu trúc dữ liệu bằng Thuật toán phân loại J48.

Khung nghiên cứu và phạm vi và giới hạn của nghiên cứu

Thuật toán được sử dụng trong nghiên cứu này chỉ có khả năng dự đoán các giá trị phân loại. Độ chính xác của dự đoán chỉ được đánh giá ở mức độ phần trăm của các thuộc tính/giá trị được phân loại. Hơn nữa, mô hình này chỉ mang tính gợi ý, nghĩa là các tiêu chí hoặc thuộc tính khác phải được đưa vào và đo lường cẩn thận để đạt được kết quả chính xác hơn.

2.1 Phương pháp

Thiết kế nghiên cứu: Nghiên cứu này sử dụng thiết kế tương quan mô tả bằng cách sử dụng Phân tích dữ liệu khám phá (EDA) để khai thác dữ liệu để kiểm tra

và xác minh dữ liệu nhằm tạo ra thông tin mới và xây dựng mô hình phân loại.

Xây dựng mô hình: Mô hình nghiên cứu mô tả cách chuẩn bị và trình bày dữ liệu. Hồ sơ của SV cho khóa học Cấu trúc dữ liệu được thu thập từ các giảng viên phụ trách môn học nói trên tại Trường Cao đẳng Công nghệ Thông tin & Truyền thông thuộc Cơ sở ESSU Salcedo từ Năm học 2014-2015 đến 2015-2016. Tổng cộng có 188 hồ sơ đã được thu thập. Hồ sơ cho thấy mỗi học sinh đều có mã số học sinh, tên học sinh, điểm trong các bài kiểm tra được chia thành 4 học kỳ (bài kiểm tra 1, câu hỏi 2 đến câu hỏi 11), các kỳ thi chuyên ngành (thi sơ bộ, giữa kỳ, thi trước và cuối kỳ), điểm thực hành và yêu cầu và điểm thi. tổng điểm trung bình. Học sinh phải đạt ít nhất 3.0 để vượt qua khóa học. Điều đó có nghĩa là, 3,1 đến 3,5 được coi là "Điểm có điều kiện" trong khi 3,6 đến 5,0 là "Không đạt".

2.2. Chuẩn bị dữ liệu

3 bước để chuẩn bị dữ liệu:

- Loại bỏ hồ sơ học sinh đã rút khỏi lớp học vì có thể thiếu một số giá trị liên quan.

- Phân chia thuộc tính điểm cuối cùng thành 7 loại: Xuất sắc, Xuất sắc, Rất tốt, Tốt, Khá, Có điều kiện, Không đạt

- Sự rời rạc hóa dữ liệu có nghĩa là người ta sẽ sử dụng một tập hợp các khoảng được xác định trước và nhóm các giá trị trong tương lai theo khoảng đó. Dữ liệu sau khi tiền xử lý được tải lên phần mềm WEKA để áp dụng thuật toán phân loại.

2.3. Phân loại dữ liệu: Mục tiêu của nghiên cứu này là dự đoán kết quả học tập của học sinh đối với môn Cấu trúc dữ liệu. Phân loại được sử dụng ở đây vì đây là quá trình đặt một đối tượng vào một lớp hoặc danh mục. Cây quyết định được sử dụng để phân loại các mẫu chưa biết. Một trong những mục tiêu của phân loại là xây dựng cây quyết định có cấu trúc cây giống như sơ đồ, với các nút bên trong và nút lá. Tất cả các nút bên trong đều có hai hoặc nhiều nút con biểu thị việc kiểm tra một thuộc tính. Các nút lá đại diện cho nhãn lớp hoặc phân bố lớp. Việc tạo cây quyết định bao gồm việc xây dựng cây và cắt tỉa cây. Tất cả các ví dụ huấn luyện ở phần đầu đều nằm ở gốc của cây. Việc phân vùng được thực hiện đệ quy dựa trên các thuộc tính đã chọn. Việc cắt tỉa cây xác định và loại bỏ các nhánh phản ánh tiếng ồn và phân vùng ngoại lệ.

2.4. Cây quyết định: ID3 do J. R. Quinlan phát triển là thuật toán trung tâm trong việc xây dựng cây quyết định. Nó sử dụng tìm kiếm tham lam, từ trên xuống trong không gian của các nhánh có thể có mà không cần quay lại. Cây quyết định mô tả các quy tắc phân chia dữ liệu thành các nhóm. J48 (trong công cụ WEKA) xây dựng cây quyết định từ một tập hợp dữ liệu huấn luyện theo cách tương tự như ID3 bằng cách sử dụng entropy và thu được thông tin. Thuật toán ID3 sử dụng entropy để tính toán tính đồng nhất của mẫu. Nếu mẫu hoàn toàn đồng nhất thì entropy bằng 0 và nếu mẫu được chia đều thì nó có entropy bằng 1.

Tại mỗi nút của cây, J48 chọn một thuộc tính của dữ liệu để phân chia tập mẫu của nó thành các tập con được làm giàu trong lớp này hay lớp khác một cách hiệu quả nhất. Tiêu chí của nó là mức tăng thông tin được chuẩn hóa (sự khác biệt về entropy) là kết quả của việc chọn một thuộc tính để phân tách dữ liệu. Thuộc tính có mức tăng thông tin chuẩn hóa cao nhất được chọn để đưa ra quyết định. Thuật toán J48 sau đó sẽ lặp lại trên các danh sách con nhỏ hơn.

Sau đây là các điều kiện để dừng phân vùng:

- Tất cả các mẫu cho một nút nhất định đều thuộc cùng một lớp.

- Không còn thuộc tính nào để phân chia tiếp - biểu quyết đa số được sử dụng để phân loại lá.

- Không còn mẫu nào (Apte & Weiss, 1997) (Fayyad, 1994 theo trích dẫn của [9]).

2.5. Trích xuất các quy tắc từ cây

Từ nghiên cứu của [9], dưới đây là một số quy tắc trích xuất từ cây quyết định:

- Tri thức được biểu diễn dưới dạng luật IF-THEN.

- Một quy tắc được tạo cho mỗi đường dẫn từ gốc tới cây.

- Mỗi cặp thuộc tính-giá trị dọc theo một đường dẫn tạo thành một kết hợp.

- Các nút lá chứa dự đoán lớp. 5. Con người dễ hiểu các quy tắc hơn (Kamber, et.al., 1997)

3. Kết luận

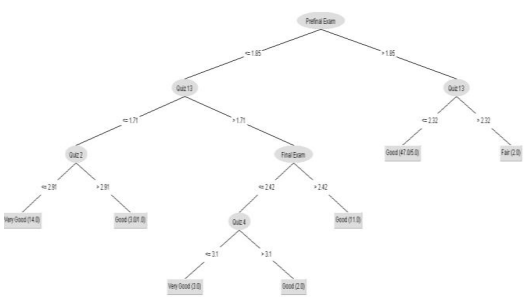
Đối với thí nghiệm này, người nghiên cứu đã sử dụng tất cả các thuộc tính (ở dạng số): câu hỏi 1 đến câu hỏi 13, điểm thực hành cuối kỳ, điểm trong các kỳ thi chính (sơ tuyển, giữa kỳ, sơ tuyển, cuối kỳ) và điểm trung bình cuối cùng đã được phân loại như (xuất sắc, xuất sắc, rất tốt, tốt, trung bình hoặc có điều kiện). Dữ liệu được mã hóa vào ứng dụng MS Excel với định dạng tệp CSV (giá trị được phân tách bằng dấu phẩy). Tập dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm tra (20%). Tập huấn luyện được sử dụng để tạo mô hình dự đoán trong khi tập kiểm tra được sử dụng để xác thực tính chính xác của mô hình. Điều đáng chú ý từ cây là học sinh phải thực hiện tốt hơn trong các kỳ thi lớn để có điểm trung bình cuối kỳ được cải thiện. Hình 3 cho thấy mã giả của mô hình đã phát triển trong đó các thuộc tính khác nhau như Kỳ thi sơ tuyển, Bài kiểm tra 13, Bài kiểm tra 2 và Bài kiểm tra cuối kỳ được mô tả . Điều này có nghĩa là nếu một học sinh đạt điểm từ 1,85 trở lên, tương

tự như vậy trong bài kiểm tra 13 và bài kiểm tra 2 thì học sinh đó có thể sẽ nhận được điểm trung bình cuối kỳ "rất tốt" ở môn học đó. Tuy nhiên, nếu một học sinh nhận được điểm thấp hơn 1,85 trong Bài kiểm tra trước cuối kỳ và bài kiểm tra 13 thấp hơn 2,32, thì rất có thể học sinh đó sẽ chỉ nhận được điểm "Khá" trong môn học đó.

Để đo lường độ chính xác của việc phân loại, Bảng II đưa ra hiệu suất dự đoán được thực hiện bởi công cụ đánh giá của WEKA. Tập hợp các phép đo được lấy từ dữ liệu huấn luyện. Nó hiển thị số liệu thống kê về dự đoán chính xác của bộ phân loại để phân loại đúng lớp. Rõ ràng là 92,68% \là độ chính xác của dự đoán, nghĩa là chỉ có khoảng 7% trường hợp được phân loại sai. Sai số tuyệt đối trung bình của xác suất là 0,08 và sai số bình phương trung bình gốc là 0,20, nghĩa là căn bậc hai của tổn thất bậc hai. Tổng số tập dữ liệu huấn luyện sau khi phân tách là 82. Lỗi không phải là 1 hoặc 0 vì nó không phân loại chính xác tất cả các mẫu huấn luyện. Hình 4 là mô hình dự đoán của thuật toán phân loại. Mô hình mô tả yếu tố dự đoán mạnh nhất hoặc biến số có ảnh hưởng lớn đến kết quả học tập của học sinh trong môn Cấu trúc dữ liệu. Nó cho thấy rằng học sinh phải thể hiện tốt nhất trong các kỳ thi chuyên ngành để đạt được ít nhất điểm trung bình cuối kỳ rất tốt.

Hai kỳ thi chính – Kỳ thi sơ tuyển và Kỳ thi cuối kỳ được hiển thị trong cây quyết định với thuộc tính được dự đoán mạnh nhất là kỳ thi trước. Khi một học sinh đạt điểm trung bình cuối cùng nhỏ hơn hoặc bằng 1,85, một thuộc tính khác Câu đố 13 (tức là trong học kỳ cuối cùng của học kỳ) sẽ được đánh giá. Nếu học sinh đạt 1,71 trở lên và bài kiểm tra 2 (đầu học kỳ) cũng cao hơn 2,91 thì rất có thể điểm trung bình cuối kỳ của học sinh sẽ là "Rất tốt" nếu không là "Tốt".

Cây quyết định cũng cho thấy rằng một học sinh cụ thể sẽ chỉ được đánh giá "tốt" hoặc "khá" nếu học sinh đó không làm tốt bài kiểm tra cuối kỳ, bài kiểm tra 4 và bài kiểm tra 13 nếu không sẽ đạt điểm "Rất tốt". " Xếp hạng.



Chủ đề cấu trúc – đó là Bài kiểm tra trước, Bài kiểm tra cuối kỳ, Câu đố 13, Câu đố 2 và Câu hỏi 4, trong đó Bài kiểm tra trước khi kết thúc được gọi là nút gốc. Các con số dưới mỗi nút lá là điểm đánh giá cuối cùng của SV trong khóa học. Các ô vuông thể hiện sự phân loại là Vergy Good, Good hoặc Fair. Cây cho thấy rằng học sinh cần phải nắm vững các bài học trong giai đoạn sơ cấp và giữa kỳ vì các bài kiểm tra ở học kỳ trước và cuối kỳ hầu hết được lấy từ các bài học trước đó (tiền sơ bộ và giữa kỳ), đó là lý do tại sao yếu tố dự báo mạnh nhất từ thí nghiệm này là bài kiểm tra trước cuối kỳ. tiết lộ thêm rằng các khái niệm nền tảng vững chắc là rất quan trọng để vượt qua môn học.

Một kỹ thuật khai thác dữ liệu có tên là phân loại đã được sử dụng để dự đoán kết quả học tập của học sinh. Cụ thể, nghiên cứu này sử dụng thuật toán phân loại J48 để tạo ra mô hình dự đoán hiệu quả dựa trên dữ liệu lịch sử của học sinh để phân loại các thuộc tính vào từng lớp/loại. Từ mô hình dự đoán, học sinh cần có kiến thức nền tảng vững chắc về các lĩnh vực học tập của môn học để có thể đạt kết quả tốt nhất trong kỳ thi chuyên ngành nhằm đạt điểm đậu trong khóa học. Các kỳ thi chính (Pre-final & Final) và các bài kiểm tra đầu học kỳ được coi là yếu tố góp phần nâng cao kết quả học tập của SV. Nếu không, học sinh sẽ có khả năng trượt hoặc bị điểm trung bình kém nếu 2 trong số các kỳ thi chính có điểm thấp hơn 3.0. Kết quả đạt tỷ lệ dự đoán chính xác 92,68% sau khi áp dụng thuật toán phân loại.

Tài liệu tham khảo

1. International Data Mining Society. Available at educationaldatamining.org.

2. M. Al-Barrak & M. Al-Razgan "Predicting Students' Performance through Classification: A Case Study", Journal of Theoretical and

3. Abu-Oda, & A.M. El-Halees, "Data Mining in Higher Education: University Dropout CaseStudy", International Journal of Data Mining and Knowledge Discovery process, Vol. 5, No. 1, January 2015.

4.Behrouz, et.al., "Predicting Student Performance: An Application of

5. S. Aher & L.,L.M.R.J., "Data Mining in Educational System using WEKA", International Conference on Emerging Technology Trends, International Journal of Computer Applications, pp. 20-25., 2011

6. S.K. Yadav, S. Pal, & B.K. Bhardwaj, "Mining Data to Predict

7. Students' Retention: A Comparative Study", www.researchgate.net/2012

8. A. Goyal & M. Rajni "Performance Comparison of Naïve Bayes and J48 Classification Algorithms "International Journal of Applied Engineering