Tạp chí Khoa học Đại học Huế: Kinh tế và Phát triển
pISSN: 2588-1205; eISSN: 2615-9716
Tập 129, Số 5B, 2020, Tr. 123–137; 10.26459/hueunijed.v129i5B.5743

KHAI PHÁ CƠ SỞ DỮ LIỆU TRONG HỆ THỐNG QUẢN LÝ ĐÀO TẠO CỦA TRƯỜNG ĐẠI HỌC KINH TẾ, ĐẠI HỌC HUẾ

Mai Thu Giang*

Trường Đại học Kinh tế, Đại học Huế, 99 Hồ Đắc Di, Huế, Việt Nam

Tóm tắt: Dự báo kết quả học tập và tìm ra các yếu tố có ảnh hưởng đến kết quả đó có ý nghĩa vô cùng quan trọng đối với không chỉ các nhà quản lý giáo dục mà cả đối với sinh viên. Tuy nhiên, các nghiên cứu về ứng dụng khai phá dữ liệu trong dự báo kết quả học tập tại Trường Đại học kinh tế, Đại học Huế còn chưa được khai thác tương xứng với tiềm năng của dữ liệu được lưu trữ. Nghiên cứu này sử dụng kỹ thuật trích chọn thuộc tính và kỹ thuật phân lớp dựa trên các giải thuật cây quyết định được trong phần mềm WEKA (Waikato Environment for Knowledge Analysis) để xây dựng các mô hình dự báo kết quả cuối khóa sau khi kết thúc từng kỳ học. Kết quả cho thấy các thuộc tính bao gồm: giới tính, số tín chỉ tích lũy ngành và điểm trung bình chung của mỗi học kỳ là các thuộc tính được giữ lại ở hầu hết trong các tập dữ liệu con sau khi trích chọn. Đặc biệt, J48 là giải thuật phù hợp nhất trong xây dựng mô hình cây quyết định dự báo kết quả cuối khóa của sinh viên.

Từ khóa: cây quyết định, dự báo, khai phá dữ liệu, phân lớp, trích chọn thuộc tính

1 Đặt vấn đề

Khai phá dữ liệu là trích xuất và khai thác những thông tin hữu ích, tiềm ẩn của dữ liệu. Công việc này giải quyết các vấn đề bằng cách phân tích lượng dữ liệu lớn hiện có để khám phá ra các xu hướng và các quy tắc có ý nghĩa [1]. Rõ ràng, các trường đại học luôn lưu trữ một cơ sở dữ liệu lớn của sinh viên. Cùng với sự phát triển của nhà trường thì cơ sở dữ liệu này ngày càng lớn về quy mô cũng như về số lượng. Tuy nhiên, vấn đề không chỉ ở việc lưu trữ, mà hơn nữa là việc khám phá và trích xuất ra được các mô hình có ý nghĩa và khai phá được tri thức tiềm ẩn trong cơ sở dữ liệu khổng lồ đó [2]. Triển khai công cụ khai phá dữ liệu là một cách để phân tích và quản lý khối lượng lớn dữ liệu sao cho có thể khám phá được các mô hình hữu ích cho giải quyết vấn đề và hỗ trợ ra quyết định [3]. Đây là thách thức của các trường đại học nói chung và của Trường Đại học Kinh tế, Đại học Huế nói riêng.

Kết quả học tập của sinh viên chịu ảnh hưởng của nhiều yếu tố như các đặc điểm riêng của từng cá nhân, đặc điểm kinh tế xã hội và các yếu tố liên quan đến môi trường sống [4]. Biết rõ những yếu tố này và ảnh hưởng của chúng đến quá trình và kết quả học tập của sinh viên có thể giúp cho không chỉ sinh viên mà cả các nhà quản lý giáo dục triển khai công tác đào tạo một cách hiệu quả.

* *Liên hệ:* mtgiang@hce.edu.vn
Nhận bài: 27-3-2020; Hoàn thành phản biện: 16-7-2020; Ngày nhận đăng: 21-7-2020

Hiện nay, rất nhiều nghiên cứu về khai phá dữ liệu trong giáo dục được các nhà nghiên cứu quan tâm. Khai phá dữ liệu giáo dục là công cụ nghiên cứu được thiết kế để tự động chiết xuất ngữ nghĩa từ hoạt động học tập của người học trong môi trường giáo dục [5]. Dự báo kết quả học tập của sinh viên càng sớm càng trở nên quan trọng đối với người học và cả các nhà quản lý giáo dục trong mục tiêu nâng cao chất lượng đầu ra. Tuy nhiên, việc dự báo trở nên khó khăn hơn do lượng lớn cơ sở dữ liệu giáo dục đã lưu trữ càng ngày càng lớn. Bên cạnh đó, sinh viên và các nhà quản lý giáo dục đều mong muốn xác định những yếu tố ảnh hưởng đến kết quả học tập của sinh viên để có hành động cụ thể và kịp thời và hỗ trợ cho việc cải thiện kết quả học tập.

Nghiên cứu này sử dụng các kỹ thuật trích chọn thuộc tính để tìm ra các yếu tố ảnh hưởng đến kết quả học tập của từng học kỳ của sinh viên khóa 2014–2018 của Trường Đại học Kinh tế, Đại học Huế. Đồng thời, một mô hình dự báo phân lớp áp dụng các giải thuật Cây quyết định trong WEKA (Waikato Environment for Knowledge Analysis) được xây dựng để dự báo kết quả. Việc dự báo cho phép phát hiện kịp thời những sinh viên có khả năng nằm trong diện đạt kết quả thấp hoặc không đủ điều kiện ra trường. Từ đó, các nhà quản lý giáo dục có biện pháp tư vấn, hỗ trợ kịp thời đối với sinh viên, đồng thời sinh viên sẽ có kế hoạch tốt hơn cho việc học của mình.

2 Tổng quan

Khai phá dữ liệu trong giáo dục đại học là một lĩnh vực còn mới và lĩnh vực này được gọi Khai phá dữ liệu giáo dục. Đã có nhiều nghiên cứu trong lĩnh vực này bởi vì khả năng tiềm ẩn của nó đối với sự phát triển của các tổ chức giáo dục, đặc biệt trong lĩnh vực đào tạo.

Từ một cuộc khảo sát về khai phá dữ liệu giáo dục từ năm 1995 đến 2005, Romero và Sebastian đã kết luận rằng khai phá dữ liệu giáo dục là một lĩnh vực nghiên cứu đầy hứa hẹn [6]. Trong một nghiên cứu khác về sử dụng khai phá dữ liệu áp dụng giải thuật 'Cây quyết định' để chỉ ra hành vi của những học sinh thuộc diện cảnh báo để từ đó cảnh báo nguy cơ ngừng học trước kỳ thi cuối học kỳ, Merceron và Ycef đã giúp sinh viên có ý thức học tập tốt hơn để chuẩn bị cho kỳ thi và cải thiện kết quả học tập [7]. Bayer và cs. đã kết hợp công cụ phân tích mạng xã hội với kỹ thuật khai phá dữ liệu bao gồm Cây quyết định và Naïve Bayes để dự báo khả năng sinh viên đạt và không đạt kết quả tốt ngay từ đầu khóa học với mục đích cải thiện độ chính xác của mô hình phân lớp đối với dữ liệu giáo dục và cho thấy mô hình xây dựng trên giải thuật Cây quyết định (J48) mang lại tỷ lệ phân lớp chính xác hơn Naïve Bayes [8]. Đặc biệt, Kapoor và cs. công bố giải thuật J48 là một trong những giải thuật tốt nhất trong việc xây dựng mô hình dự báo phân lớp [9]. Ngoài ra, Sharma đã so sánh kết quả dự báo phân lớp của các mô hình dựa trên các giải thuật cây tìm kiếm sử dụng trong WEKA và kết luận rằng J48 là giải thuật có kết quả phân lớp tương đối tốt với thời gian thực ít nhất [10]. Để đánh giá tỷ lệ phân lớp một cách chính xác,

Kohavi cho thấy đánh giá chéo 10 lần là phương pháp tốt nhất khi xây dựng mô hình phân lớp mặc dù cần phải thực hiện các thao tác tính toán nhiều hơn [11].

Phương pháp tìm kiếm và đánh giá thuộc tính cũng là một vấn đề nghiên cứu được nhiều tác giả quan tâm. Trong đó, giải thuật và đánh giá thuộc tính BestFirst–CfsSubsetEval của WEKA được nhiều tác giả sử dụng. Điển hình, Lei và Pingfan đã chứng minh rằng phương pháp lựa chọn thuộc tính theo BestFirst là một giải thuật lựa chọn thuộc tính tối ưu, cho ra tập thuộc tính ít hơn nhiều so với các phương pháp tìm kiếm khác [12]. Aggarwal và cs. đã sử dụng CfsSubsetEval làm công cụ đánh giá một tập hợp con của các thuộc tính bằng cách xem xét khả năng riêng của từng thuộc tính cùng với mức độ dư thừa của chúng và đưa ra tập thuộc tính của mô hình với độ phân lớp chính xác tới 99,95% [13].

3 Phương pháp

3.1 Thu thập số liệu và chuẩn hóa

Thu thập số liệu

Dữ liệu được thu thập từ các tác vụ khác nhau từ phần mềm quản lý đào tạo của trường, bao gồm Quản lý sinh viên, Quản lý đào tạo và Đánh giá và phân loại xếp hạng của sinh viên. Trong đó, dữ liệu thu được từ tác vụ Quản lý sinh viên trên hệ thống quản lý đào tạo gồm các bảng dữ liệu Quản lý hồ sơ sinh viên với ba nhóm thông tin: Thông tin người học, Thông tin học tập và rèn luyện và Thông tin tuyển sinh. Thông tin hồ sơ người học bao gồm họ tên, giới tính, quê quán và dân tộc. Nhóm thông tin học tập và rèn luyện bao gồm điểm xếp loại rèn luyện năm 1, 2, 3 và 4; điểm xếp loại học tập năm 1, 2, 3 và 4; tổng số tín chỉ đã học, điểm trung bình chung hệ số 4 và xếp loại học tập và rèn luyện toàn khóa học. Nhóm thông tin tuyển sinh bao gồm điểm tuyển sinh đầu vào của ba môn, điểm thưởng, khối thi, ngành thi, xếp loại THPT và xếp loại hạnh kiểm. Tác vụ Quản lý hồ sơ sinh viên có Bảng kiểm tra hoàn thành chương trình học với các trường dữ liệu bao gồm thông tin về số tín chỉ đã hoàn thành đối với từng khối kiến thức yêu cầu như: kiến thức giáo dục đại cương, lý luận chính trị, ngoại ngữ, khoa học xã hội – nhân văn – nghệ thuật, khối kiến thức giáo dục đại cương tự chọn, kiến thức giáo dục chuyên nghiệp, kiến thức chung của ngành, kiến thức chuyên sâu của ngành, kiến thức cơ sở, kiến thức bổ trợ, thực tập nghề, thực tập tốt nghiệp và khóa luận và kiến thức giáo dục chuyên nghiệp.

Dữ liệu Quản lý đào tạo bao gồm Xếp loại học tập toàn khóa và Quản lý điểm. Trong đó, Quản lý xếp loại học tập toàn khóa bao gồm các trường dữ liệu Xếp loại học lực, Điểm trung bình hệ số 10 và Điểm trung bình hệ số 4. Thông tin về điểm của sinh viên được trích xuất theo từng học kỳ và từng năm học.

Bảng 1 trình bày tổng số thuộc tính đã được thu thập và sử dụng trong cơ sở dữ liệu cùng với viết tắt và diễn giải.

Bảng 1. Tổng số thuộc tính được lưu trữ, viết tắt và diễn giải

STT	Viết tắt	Diễn giải thuộc tính	STT	Viết tắt	Diễn giải thuộc tính
1	NS	Năm sinh	29	STCTLN_K3	Số tín chỉ tích lũy ngành kỳ 3
2	GT	Giới tính	30	SMKD_K3	Số môn không đạt kỳ 3
3	NoiSinh	Nơi sinh	31	DTBC_K3	Điểm trung bình chung kỳ 3
4	TG	Tôn giáo	32	TongTC_K4	Tổng tín chỉ đăng ký kỳ 4
5	KV	Khu vực	33	STCTLN_K4	Số tín chỉ tích lũy ngành kỳ 4
6	KQTSM1	Kết quả tuyển sinh môn 1	34	DTBCQD_K4	Điểm trung bình chung quy đổi kỳ 4
7	KQTSM2	Kết quả tuyển sinh môn 2	35	SMKD_K4	Số môn không đạt kỳ 4
8	KQTSM3	Kết quả tuyển sinh môn 3	36	STCKD_K4	Số tín chỉ không đạt kỳ 4
9	Khoa	Khoa theo học	37	DTBCQD_K4	Điểm trung bình chung quy đổi kỳ 4
10	TongTC_K1	Tổng số tín chỉ đăng ký kỳ 1	38	TongTC_K5	Tổng tín chỉ đăng ký kỳ 5
11	STCTLN_K1	Số tín chỉ tích lũy ngành kỳ 1	39	STCTLN_K5	Số tín chỉ tích lũy ngành kỳ 5
12	SMKD_K1	Số môn không đạt kỳ 1	40	DTBCQD_K5	Điểm trung bình chung quy đổi kỳ 5
13	STCKD_K1	Số tín chỉ không đạt kỳ 1	41	SMKD_K5	Số môn không đạt kỳ 5
14	MacLenin	Điểm học phần Mác–Lênin	42	STCKD_K5	Số tín chỉ không đạt kỳ 5
15	PLDC	Điểm học phần Pháp luật đại cương	43	DTBCQD_K5	Điểm trung bình chung quy đổi kỳ 5
16	THDC	Điểm học phần Tin học đại cương	44	TongTC_K6	Tổng tín chỉ kỳ 6
17	TCC1	Điểm toán cao cấp 1	45	STCTLN_K6	Số tín chỉ tích lũy ngành kỳ 6
18	TCC2	Điểm toán cao cấp 2	46	DTBCQD_K6	Điểm trung bình chung quy đổi kỳ 6
19	DTBCQD_K1	Điểm trung bình chung quy đổi kỳ 1	47	SMKD_K6	Số môn không đạt kỳ 6
20	TongTC_K2	Tổng số tín chỉ đăng ký kỳ 2	48	STCKD_K6	Số tín chỉ không đạt kỳ 6
21	STCTLN_K2	Số tín chỉ tích lũy ngành kỳ 2	49	DTBCQD_K6	Điểm trung bình chung quy đổi kỳ 6
22	DTBCQD_K2	Điểm trung bình chung quy đổi kỳ 2	50	TongTC_K7	Tổng số tín chỉ đăng ký kỳ 7
23	SMKD_K2	Số môn không đạt kỳ 2	51	STCTLN_K7	Số tín chỉ tích lũy ngành kỳ 7
24	STCKD_K2	Số tín chỉ không đạt kỳ 2	52	DTBCQD_K7	Điểm trung bình chung quy đổi kỳ 7
25	XSTK	Xác suất thống kê	53	SMKD_K7	Số môn không đạt kỳ 7
26	MacLenin2	Điểm học phần Mác–Lenin 2	54	STCKD_K7	Số tín chỉ không đạt kỳ 7
27	DTBCQD_K2	Điểm trung bình chung quy đổi kỳ 2	55	DCTK_K7	Điểm chữ tổng kết kỳ 7
28	TongTC_K3	Tổng số tín chỉ kỳ 3			

Bảng 2. Số lượng từng nhãn lớp tương ứng với số bản ghi

STT	Nhãn lớp	Số bản ghi	Tỷ lệ (%)
1	Xuất sắc	24	1,55%
2	Giỏi	145	9,35%
3	Khá	627	40,43%
4	Trung bình	329	21,21%
5	Yếu	326	21,02%
6	Chưa xếp hạng	100	6,45%

Chuẩn hóa số liệu

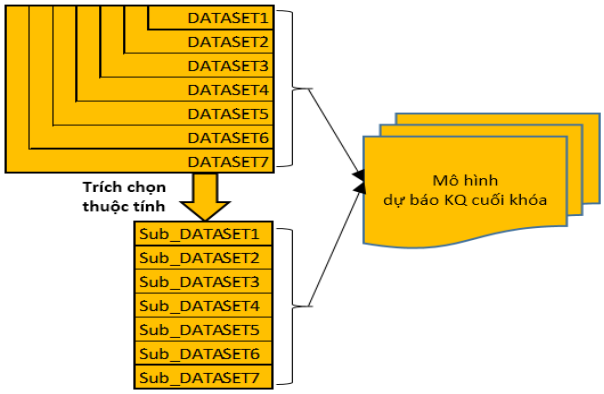
Dữ liệu trích xuất từ tác vụ quản lý sinh viên và tác vụ quản lý điểm được kết nối với nhau dựa vào trường dữ liệu khóa (Mã sinh viên). Dữ liệu được thu thập và lưu trữ dưới dạng file Excel với tổng số bản ghi là 1881. Những bản ghi thiếu thông tin được loại bỏ. Tổng số bản ghi cuối cùng được sử dụng trong cơ sở dữ liệu là 1551, tương ứng với 1551 sinh viên.

Điểm cuối khóa được quy đổi từ điểm hệ số 10 sang các nhãn lớp Xếp loại gồm Xuất sắc, Giỏi, Khá, Trung bình, Yếu và Chưa xếp hạng và đây được gọi là nhãn lớp trong cơ sở dữ liệu. Trong đó, nhãn "Chưa xếp hạng" là nhãn được gán cho những sinh viên chưa hoàn thành chương trình học (Bảng 2).

Cuối cùng, dữ liệu file Excel đã lưu trữ được chuyển đổi để đưa về định dạng file ARFF (Atrribute Relation File Format) để thực hiện các bước trích chọn thuộc tính, xây dựng mô hình và kiểm thử trong WEKA.

3.2 Phương pháp

Ứng dụng phần mềm mở WEKA để tiến hành nghiên cứu. Giải thuật BestFirst–CfsSubsetEval được sử dụng để trích chọn thuộc tính. Giải thuật Cây quyết định được sử dụng để xây dựng mô hình dự báo phân lớp; ứng dụng phân lớp theo các giải thuật đã được xây dựng để dự báo kết quả học tập cuối khóa ngay sau mỗi kỳ học để có được dự báo sớm nhất có thể. Các giải thuật này bao gồm J48, Decision Stump, HoeffdingTree, LMT, RandomForest, RandomTree và REPTree. Mô hình dự báo được xây dựng đồng thời trên các tập dữ liệu trước và sau khi trích chọn thuộc tính. Cuối cùng, so sánh tỷ lệ dự báo phân lớp chính xác của các mô hình để từ đó lựa chọn mô hình cho ra kết quả dự báo phân lớp với tỷ lệ chính xác cao nhất, dựa trên phương pháp đánh giá chéo 10 lần [11] (Hình 1).



Hình 1. Các tập dữ liệu tham gia vào quá trình xây dựng mô hình dự báo

Cơ sở dữ liệu được thu thập sau mỗi học kỳ từ học kỳ 1 đến học kỳ 7 được lưu trữ trong 7 tập thuộc tính DATASET1, DATASET2, DATASET3, DATASET4, DATASET5, DATASET6 và DATASET7. Bước đầu, các tập dữ liệu được sử dụng để xây dựng mô hình phân lớp trước khi trích chọn và kiểm tra độ chính xác của phân lớp. Sau đó, áp dụng kỹ thuật trích chọn thuộc tính lên các tập dữ liệu đã thu được ở trên để có được các tập dữ liệu con tương ứng với tên Sub_DATASET1, Sub_DATASET2, Sub_DATASET3, Sub_DATASET4, Sub_DATASET5, Sub_DATASET6 và Sub_DATASET7.

Nghiên cứu không thực hiện cho học kỳ 8 vì kết quả cuối khóa được ghi nhận tại kỳ thứ 7. Chi tiết các tập cơ sở dữ liệu được thu thập và sử dụng để xây dựng mô hình dự báo được mô tả như sau:

DATASET1 bao gồm các trường dữ liệu điểm tuyển sinh đầu vào (3 môn), khoa, và các thuộc tính liên quan đến lý lịch trích ngang của sinh viên như: năm sinh, nơi sinh, giới tính, dân tộc, tôn giáo, khu vực, số tín chỉ đăng ký học kỳ 1, số tín chỉ tích lũy ngành, số môn không đạt học kỳ 1, số tín chỉ không đạt học kỳ 1, điểm trung bình chung học kỳ 1, điểm chữ tổng kết học kỳ 1.

DATASET2 bao gồm các thuộc tính từ tập dữ liệu DATASET1 và được bổ sung thêm thuộc tính sau khi đăng ký tín chỉ học kỳ 2 như tổng số tín chỉ đăng ký, số tín chỉ tích lũy ngành, nhóm thuộc tính gồm điểm số của 5 môn học bắt buộc trong học kỳ (Những nguyên lý cơ bản của chủ nghĩa Mác–Lê nin, Pháp luật đại cương, Tin học đại cương, Toán cao cấp 1, Toán cao cấp 2), số môn không đạt kỳ 2, số tín chỉ không đạt học kỳ 2, điểm trung bình học kỳ 2, điểm chữ tổng kết học kỳ 2.

DATASET3, DATASET4, DATASET5, DATASET6 và DATASET7 là các tập dữ liệu lần lượt kế thừa các tập dữ liệu của học kỳ trước đó và bổ sung thêm sáu thuộc tính bao gồm tổng số

tín chỉ đăng ký, số tín chỉ tích lũy ngành, số môn không đạt, số tín chỉ không đạt, điểm trung bình chung và điểm chữ tổng kết của từng học kỳ.

4 Kết quả và thảo luận

Kết quả từ Bảng 3 cho thấy, đối với trước khi thực hiện trích chọn thuộc tính và sau mỗi học kỳ kết thúc, tập thuộc tính dùng để dự báo cho kết quả cuối khóa được bổ sung thêm đáng kể số trường tham gia vào quá trình xây dựng mô hình. Cụ thể, sau kết thúc học kỳ 1, mô hình dự báo kết quả cuối khóa được xây dựng dựa trên 18 trường dữ liệu. Tuy nhiên, đến cuối học kỳ 2, cơ sở dữ liệu tăng lên đến 27 trường. Kết thúc học kỳ 2, các môn học đại cương là chung cho tất cả các ngành đã hoàn thành, do đó mỗi kỳ tiếp theo sau chỉ bổ sung thêm so với kỳ trước sáu trường, bao gồm các trường liên quan đến Tổng số tín chỉ đăng ký học, Số tín lũy tích lũy ngành, Điểm trung bình chung học kỳ, Số môn không đạt, Số tín chỉ không đạt và Điểm chữ tổng kết. Đến cuối học kỳ 7, tổng số lượng thuộc tính tham gia vào xây dựng mô hình dự báo là 55. Ngược lại, đối với trường hợp sau trích chọn thuộc tính, tổng số trường được giữ lại để tham gia vào xây dựng mô hình dự báo nhỏ hơn nhiều so với tập dữ liệu ban đầu, chỉ dao động từ 5 đến 10 thuộc tính. Đặc biệt, trong hầu hết các tập thuộc tính con nhận được sau khi trích chọn, thuộc tính giới tính, số tín chỉ tích lũy ngành và điểm trung bình chung được giữ lại ở hầu hết các tập thuộc tính kết quả.

Kết quả sau trích chọn thuộc tính cho thấy cả sự tương đồng lẫn khác biệt đối với một số nghiên cứu trước đó. Các thuộc tính về đặc điểm riêng của từng cá nhân, đặc điểm kinh tế xã hội và các yếu tố liên quan đến môi trường sống được thể hiện qua các trường bao gồm năm sinh, giới tính, nơi sinh, tôn giáo và khu vực (Bảng 1). Trong đó, đặc điểm về cá nhân có thuộc tính giới tính được giữ lại ở hầu hết các tập dữ liệu con sau khi trích chọn, còn các thuộc tính về đặc điểm

Bảng 3. Tổng hợp thuộc tính trước và sau khi áp dụng biện pháp trích chọn thuộc tính

DATASET sử dụng	Tổng số thuộc tính trước trích chọn	Thuộc tính được giữ lại sau trích chọn	Tổng số thuộc tính sau trích chọn
DATASET1	18	GT, STCKD_K1, MacLenin, THDC, TCC1, TCC2, TBC_K1	7
DATASET2	27	MacLenin, THDC, TCC1, TCC2, TBC_K1, STCTLN_K2, STCKD_K2, XSTK, MacLenin2, DTBC_K2	10
DATASET3	33	DTBC_K1, DTBC_K2, TongTC_K3, STCTLN_K3, DTBC_K3	5
DATASET4	37	GT, DTBC_K1, DTBCQD_K2, TongTC_K3, DTBC_K3, STCKD_K4, DTBC_K4	7
DATASET5	43	GT, DTBC_K1, DTBCQD_K2, TongTC_K3, DTBC_K3, DTBC_K4, DTBC_K5	7
DATASET6	49	GT, DTBC_K1, DTBCQD_K2, TongTC_K3, DTBC_K3, DTBC_K4, DTBC_K5, DTBC_K6	7
DATASET7	55	DTBC_K1, DTBC_K2, TongTC_K3, DTBC_K3, STCKD_K4, DTBC_K4, DTBC_K5, DTBC_K6, DCTK_K7	9

kinh tế xã hội và môi trường sống đều không được giữ lại trong kết quả sau trích chọn. Điều này xuất phát từ bộ cơ sở dữ liệu đầu vào khác nhau giữa các nghiên cứu.

Kết quả tỷ lệ dự báo phân lớp chính xác của mô hình dựa báo dựa trên các giải thuật cây quyết định được xây dựng trong WEKA, với các tập dữ liệu đầu vào là các tập dữ liệu được thu thập ngay sau mỗi học kỳ và các tập con sau khi được trích chọn được trình bày trong Bảng 4.

Bảng 4. Tỷ lệ dự báo phân lớp chính xác của các mô hình dựa trên các giải thuật cây quyết định trong WEKA (%)

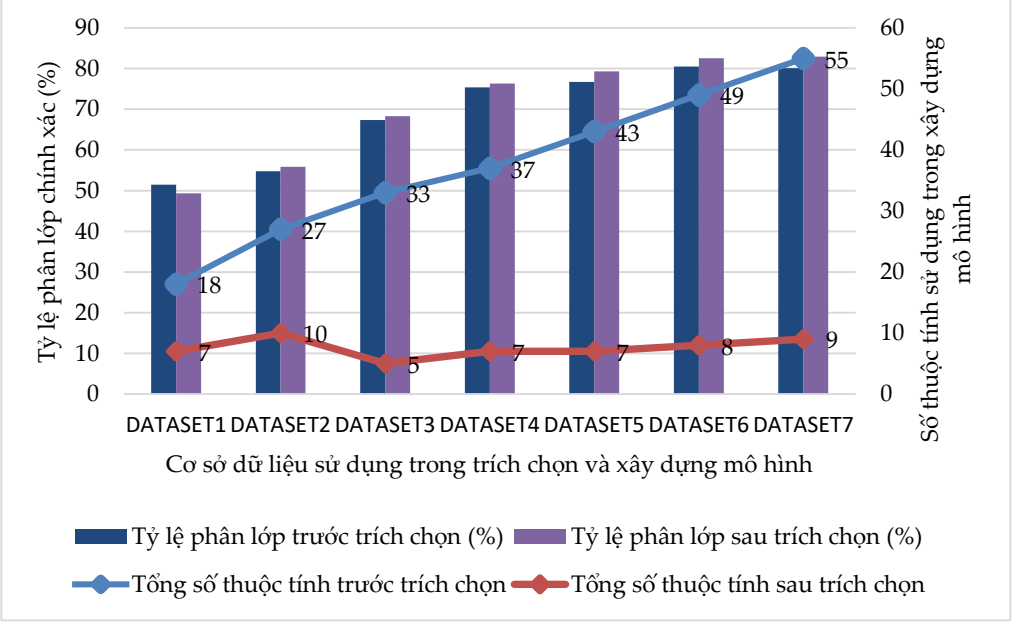
DATASET sử dụng		DATAST1	DATAST2	DATAST3	DATAST4	DATAST5	DATAST6	DATAST7
Trước trích chọn	J48	49,32	54,73	67,37	75,37	76,72	80,46	80,14
	Decision Stump	50,68	53,25	56,76	67,72	68,72	76,72	79,72
	Hoeffding_Tree	44,35	46,83	60,53	64,33	65,33	73,33	76,33
	LMT	58,74	61,37	58,84	66,47	67,47	75,47	78,47
	Random_Forest	58,89	61,50	56,36	65,93	66,93	74,93	77,93
	Random_Tree	47,96	50,03	60,53	63,50	64,50	72,50	75,50
	REPTree	56,03	58,54	59,71	64,47	65,47	73,47	76,47
Sau trích chọn	J48	51,45	55,83	68,34	76,33	79,30	82,52	82,97
	Decision Stump	45,45	53,25	66,42	74,35	77,74	80,87	81,55
	Hoeffding_Tree	48,08	53,88	66,51	75,03	77,86	81,41	81,20
	LMT	44,40	52,17	65,02	72,90	75,76	79,10	80,02
	Random_Forest	47,08	52,35	64,68	73,39	76,44	80,36	79,38
	Random_Tree	42,52	50,87	63,96	71,00	74,61	77,36	78,21
	REPTree	45,91	50,43	63,54	71,85	75,26	79,12	77,62

Bảng 4 cho thấy tỷ lệ phân lớp chính xác của các mô hình dự báo đối với các tập dữ liệu sau khi được trích chọn thường có xu hướng cao hơn so với trước trích chọn trên hầu hết tất cả các giải thuật được áp dụng. Đặc biệt, giải thuật cây quyết định J48 cho ra tỷ lệ dự báo phân lớp chính xác cao nhất với 51,45 % sau khi kết thúc học kỳ 1, tăng lên trên 75% sau khi kết thúc học kỳ 4 và đạt trên 82% sau khi kết thúc học kỳ 6 và học kỳ 7.

Hình 2 cho thấy mối quan hệ giữa tổng số thuộc tính trước và sau khi trích chọn với tỷ lệ phân lớp chính xác của các mô hình dự báo cuối khóa được xây dựng với các tập thuộc tính tương

ứng dựa trên giải thuật J48. Đối với trước khi thực hiện trích chọn, khi các thuộc tính được bổ sung vào tập thuộc tính sau mỗi kỳ học, mô hình phân lớp xây dựng cho kỳ học sau cho tỷ lệ phân lớp chính xác cao hơn so với mô hình xây dựng dựa trên kỳ học trước. Cụ thể, khi xây dựng mô hình dự báo kết quả cuối khóa từ ngay sau khi kết thúc học kỳ 1 và học kỳ 2, kết quả phân lớp chính xác của các mô hình dự báo rất thấp (49–55%). Tuy nhiên, tỷ lệ phân lớp cải thiện và tăng nhanh bắt đầu từ mô hình phân lớp sau khi kết thúc học kỳ 3, với tỷ lệ phân lớp chính xác đạt trên 67%. Tỷ lệ này tăng lên dần qua mô hình xây dựng ở các học kỳ sau đó và đạt cao nhất ở mô hình được xây dựng sau khi kết thúc học kỳ 7 với trên 80%. Sau khi thực hiện trích chọn thuộc tính, hầu hết các mô hình dự báo đều có kết quả phân lớp cao hơn mô hình dự báo trước khi thực hiện trích chọn từ 1,1 đến 2,83%, trong khi số lượng các thuộc tính cần để xây dựng mô hình ít hơn nhiều so với dữ liệu trước trích chọn từ 11 đến 46 thuộc tính. Trong đó, mô hình dự báo kết quả cuối khóa có tỷ lệ phân lớp chính xác cao là các mô hình được xây dựng sau khi kết thúc học kỳ 4 với tỷ lệ phân lớp đạt trên 76% đến gần 83% sau khi kết thúc học kỳ thứ 7.

Kết quả các mô hình dự báo phân lớp theo J48 trên dữ liệu sau khi thực hiện trích chọn được lưu lại nhằm hỗ trợ cho sinh viên và người quản lý dự báo kết quả cuối khóa bằng hai cách. Thứ nhất, có thể ứng dụng mô hình trên tập dữ liệu đầu vào cụ thể với nhãn lớp chưa được xác định để cho ra ngay kết quả dự báo nhãn lớp. Thứ hai, người dùng có thể quan sát trực quan cây



Hình 2. Tổng số thuộc tính và tỷ lệ phân lớp chính xác của các mô hình dự báo phân lớp dựa trên giải thuật J48

quyết định hoặc tập luật sinh ra từ cây quyết định để hiểu được luật khi rẽ nhánh trong cây đối với điều kiện cụ thể để đưa đến kết quả dự báo.

Do giới hạn về không gian trình bày của bài báo nên tác giả chỉ trình bày kết quả của một ví dụ về các tập luật được rút ra dựa trên giải thuật cây quyết định cho mô hình dự báo phân lớp kết quả cuối khóa sau khi kết thúc học kỳ 4 với tỷ lệ dự báo phân lớp chính xác đạt 76,33% (Bảng 5).

Bảng 5. Các tập luật trên cây quyết định của mô hình dự báo kết quả cuối khóa sau khi kết thúc học kỳ 4

DTBC_K4 ≤ 6,08	**DTBC_K4 > 6,08**
| TongTC_K3 ≤ 0: Chua xep hang	| DTBC_K4 ≤ 7,62
| TongTC_K3 > 0	| | DTBC_K3 ≤ 6,83
| | DTBC_K4 ≤ 5,02	| | | DTBC_K4 ≤ 6,71
| | | DTBC_K1 ≤ 7,02: Yeu	| | | | DTBC_K1 ≤ 5,61
| | | DTBC_K1 > 7,02	| | | | | DTBCQD_K2 ≤ 2,09: Trungbinh
| | | | DTBC_K4 ≤ 1,53	| | | | | DTBCQD_K2 > 2,09: Kha
| | | | | DTBCQD_K2 ≤ 2,36: Chua xep hang	| | | | DTBC_K1 > 5,61
| | | | | DTBCQD_K2 > 2,36: Kha	| | | | | TongTC_K3 ≤ 17
| | | | DTBC_K4 > 1,53: Trungbinh	| | | | | | DTBC_K4 ≤ 6,31
| | DTBC_K4 > 5,02	| | | | | | | TongTC_K3 ≤ 15: Trungbinh
| | | DTBC_K3 ≤ 6,76	| | | | | | | TongTC_K3 > 15
| | | | DTBC_K3 ≤ 5,03	| | | | | | | | TongTC_K3 ≤ 16: Kha
| | | | | Gioitinh = Nu	| | | | | | | | TongTC_K3 > 16
| | | | | SoTCKhongDat_K4 ≤ 8	| | | | | | | | SoTCKhongDat_K4 ≤ 0: Trungbinh
| | | | | | TongTC_K3 ≤ 17	| | | | | | | | SoTCKhongDat_K4 > 0
| | | | | | | TongTC_K3 ≤ 16	| | | | | | | | | Gioitinh = Nu
| | | | | | | | SoTCKhongDat_K4 ≤ 5	| | | | | | | | | DTBC_K4 ≤ 6,18: Trungbinh
| | | | | | | | DTBC_K1 ≤ 4,46: Yeu	| | | | | | | | | DTBC_K4 > 6,18
| | | | | | | | DTBC_K1 > 4,46	| | | | | | | | | | DTBCQD_K2 ≤ 2,12
| | | | | | | | DTBC_K4 ≤ 5,85: Trungbinh	| | | | | | | | | | DTBC_K4 ≤ 6,27: Trungbinh
| | | | | | | | DTBC_K4 > 5,85	| | | | | | | | | | DTBC_K4 > 6,27: Kha
| | | | | | | | SoTCKhongDat_K4 ≤ 0: Trungbinh	| | | | | | | | | DTBCQD_K2 > 2,12: Kha
| | | | | | | | SoTCKhongDat_K4 > 0: Yeu	| | | | | | | | Gioitinh = Nam: Kha
| | | | | | | SoTCKhongDat_K4 > 5: Yeu	| | | | | DTBC_K4 > 6,31

DTBC_K4 ≤ 6,08	DTBC_K4 > 6,08
| | | | | | | | TongTC_K3 > 16: Yeu	| | | | | | | SoTCKhongDat_K4 ≤ 2
| | | | | | | TongTC_K3 > 17	| | | | | | | | DTBC_K3 ≤ 6,26
| | | | | | | | DTBC_K4 ≤ 5,64	| | | | | | | | | DTBC_K4 ≤ 6,63: Trungbinh
| | | | | | | | | DTBC_K4 ≤ 5,44: Trungbinh	| | | | | | | | | DTBC_K4 > 6,63: Kha
| | | | | | | | | DTBC_K4 > 5,44: Yeu	| | | | | | | | DTBC_K3 > 6,26: Kha
| | | | | | | | DTBC_K4 > 5,64: Trungbinh	| | | | | | | SoTCKhongDat_K4 > 2: Kha
| | | | | SoTCKhongDat_K4 > 8: Yeu	| | | | | TongTC_K3 > 17: Trungbinh
| | | | | Gioitinh = Nam: Yeu	| | | DTBC_K4 > 6,71
| | | | DTBC_K3 > 5,03	| | | | DTBCQD_K2 ≤ 0,82
| | | | | DTBC_K1 ≤ 4,65	| | | | | Gioitinh = Nu
| | | | | Gioitinh = Nu: Yeu	| | | | | | TongTC_K3 ≤ 17: Trungbinh
| | | | | Gioitinh = Nam	| | | | | | TongTC_K3 > 17: Kha
| | | | | | DTBCQD_K2 ≤ 0,92	| | | | | Gioitinh = Nam: Trungbinh
| | | | | | | TongTC_K3 ≤ 16: Trungbinh	| | | | DTBCQD_K2 > 0,82
| | | | | | | | TongTC_K3 > 16: Yeu	| | | | | DTBC_K4 ≤ 7,01
| | | | | | | DTBCQD_K2 > 0,92: Trungbinh	| | | | | | SoTCKhongDat_K4 ≤ 0
| | | | | DTBC_K1 > 4,65	| | | | | | | DTBCQD_K2 ≤ 1,59: Trungbinh
| | | | | Gioitinh = Nu	| | | | | | | DTBCQD_K2 > 1,59: Kha
| | | | | | | TongTC_K3 ≤ 17	| | | | | | SoTCKhongDat_K4 > 0: Kha
| | | | | | | | DTBC_K4 ≤ 5,67	| | | | | DTBC_K4 > 7,01: Kha
| | | | | | | | TongTC_K3 ≤ 15: Trungbinh	| | DTBC_K3 > 6,83
| | | | | | | | | TongTC_K3 > 15	| | | DTBC_K3 ≤ 8,14: Kha
| | | | | | | | | DTBC_K3 ≤ 6,26	| | | DTBC_K3 > 8,14
| | | | | | | | | SoTCKhongDat_K4 ≤ 4: Yeu	| | | | DTBC_K4 ≤ 7,27: Kha
| | | | | | | | | SoTCKhongDat_K4 > 4	| | | | DTBC_K4 > 7,27: Gioi
| | | | | | | | | DTBC_K1 ≤ 6,38	| DTBC_K4 > 7,62
| | | | | | | | | SoTCKhongDat_K4 ≤ 5: Trungbinh	| | DTBC_K3 ≤ 7,97
| | | | | | | | | SoTCKhongDat_K4 > 5	| | | DTBC_K3 ≤ 7: Kha
| | | | | | | | | SoTCKhongDat_K4 ≤ 7: Kha	| | | DTBC_K3 > 7
| | | | | | | | | SoTCKhongDat_K4 > 7: Trungbinh	| | | | DTBCQD_K2 ≤ 3,21
| | | | | | | | | DTBC_K1 > 6,38: Yeu	| | | | DTBC_K4 ≤ 8: Kha
| | | | | | | | DTBC_K3 > 6,26	| | | | DTBC_K4 > 8

DTBC_K4 ≤ 6,08	DTBC_K4 > 6,08
| | | | | | | | | | | DTBC_K1 ≤ 6,84: Kha	| | | | | | DTBCQD_K2 ≤ 2,47: Kha
| | | | | | | | | | | DTBC_K1 > 6,84: Trungbinh	| | | | | | DTBCQD_K2 > 2,47: Gioi
| | | | | | | | DTBC_K4 > 5,67: Trungbinh	| | | | DTBCQD_K2 > 3,21: Gioi
| | | | | | | TongTC_K3 > 17: Trungbinh	| | DTBC_K3 > 7,97
| | | | | | Gioitinh = Nam	| | | DTBC_K4 ≤ 8,28
| | | | | | SoTCKhongDat_K4 ≤ 2	| | | | TongTC_K3 ≤ 15: Kha
| | | | | | | | DTBC_K4 ≤ 5,72: Yeu	| | | | TongTC_K3 > 15
| | | | | | | | DTBC_K4 > 5,72	| | | | | TongTC_K3 ≤ 18: Gioi
| | | | | | | | DTBC_K1 ≤ 6,05: Trungbinh	| | | | | TongTC_K3 > 18
| | | | | | | | | DTBC_K1 > 6,05: Yeu	| | | | | | DTBC_K4 ≤ 7,97: Gioi
| | | | | | SoTCKhongDat_K4 > 2	| | | | | | DTBC_K4 > 7,97: Kha
| | | | | | | | DTBC_K4 ≤ 5,61	| | | DTBC_K4 > 8,28
| | | | | | | | SoTCKhongDat_K4 ≤ 4: Trungbinh	| | | | DTBCQD_K2 ≤ 3,69
| | | | | | | | SoTCKhongDat_K4 > 4	| | | | | DTBC_K3 ≤ 8,74: Gioi
| | | | | | | | | DTBC_K3 ≤ 5,86: Yeu	| | | | | DTBC_K3 > 8,74
| | | | | | | | | DTBC_K3 > 5,86: Trungbinh	| | | | | | DTBCQD_K2 ≤ 3,26: Gioi
| | | | | | | | DTBC_K4 > 5,61: Trungbinh	| | | | | | DTBCQD_K2 > 3,26: Xuatsac
| | | DTBC_K3 > 6,76	| | | | DTBCQD_K2 > 3,69: Xuatsac
| | | | DTBCQD_K2 ≤ 1,95	
| | | | | DTBC_K4 ≤ 5,83	
| | | | | SoTCKhongDat_K4 ≤ 4: Trungbinh	
| | | | | | SoTCKhongDat_K4 > 4	
| | | | | | DTBC_K1 ≤ 6,58: Trungbinh	
| | | | | | DTBC_K1 > 6,58: Kha	
| | | | | DTBC_K4 > 5,83: Kha	
| | | | DTBCQD_K2 > 1,95: Kha	

Gốc của cây quyết định cho mô hình dự báo phân lớp được xây dựng sau học kỳ 4 dựa trên các thuộc tính sau khi thực hiện trích chọn là ĐTBC_K4 (Điểm trung bình chung học kỳ 4), Trong đó, phía bên phải của Bảng 5 thể hiện cho nhánh cây con phải với ĐTBC_K4 > 6,08 cho kết quả phân lớp cuối khóa không có kết quả Yếu hoặc Chưa xếp hạng. Kết quả dự báo là tất cả đều từ Trung Bình trở lên. Tuy nhiên, phía bên trái của Bảng 5 tương ứng với nhánh cây con trái khi ĐTBC_K4 < 6,08 cho thấy luật cho kết quả dự báo cuối khóa đạt loại Khá rất ít, tức là khả năng để đạt xếp loại Khá là khó khi sinh viên có các điều kiện thỏa mãn nhánh cây con trái. Kết quả

dự báo xếp loại hay giá trị nút lá trong cây quyết định ở nhánh cây con này chủ yếu là Xếp loại Trung bình, Yếu, hoặc Chưa xếp loại. Các nhãn lớp Yếu hoặc Chưa xếp loại được đánh dấu trong bảng.

Có thể hiểu một số tập luật đầu tiên của cây quyết định như sau: Nút gốc là ĐTBC_K4; nếu ĐTBC_K4 ≤ 6,08 thì đi về phía cây con trái của cây quyết định. Luật sẽ đi xuống nút con là TongTC_K3 (Tổng tín chỉ kỳ 3) để kiểm tra. Nếu TongTC_K3 ≤ 0 thì dự báo kết quả cuối khóa là "Chưa xếp hạng". Ngược lại, nếu TongTC_K3 > 0, cây quyết định đi xuống nhánh con trái và kiểm tra quan hệ ĐTBC_K4 ≤ 5,02. Nếu đúng thì đi xuống nhánh con trái tiếp theo là nút ĐTBC_K1 để kiểm tra quan hệ ĐTBC_K1 < 7,02. Nếu đúng thì dự báo kết quả xếp loại cuối khóa "Yếu",

Cây quyết định hay tập luật tạo ra từ xây dựng mô hình cây quyết định là một cách trực quan và dễ hiểu nhất để sinh viên và người quản lý có thể dự báo được kết quả học tập cuối khóa dựa trên các giả định hoặc tình huống cụ thể trong quá trình học. Người học và cả người quản lý sẽ có được định hướng tốt nhất có thể và tránh được những trường hợp đáng tiếc như khi dự báo kết quả cuối khóa đưa về xếp loại Yếu hoặc Chưa xếp loại. Tuy nhiên, trong trường hợp người dùng chỉ cần có kết quả dự báo cuối cùng mà không cần quan tâm đến luật sinh ra kết quả đó, người dùng chỉ cần cung cấp dữ liệu đầu vào với nhãn lớp để trống và thực hiện lệnh gọi mô hình đã được lưu trữ trước đó. WEKA sẽ cho ra kết quả dự báo với tỷ lệ dự báo phân lớp chính xác đã đề cập trong Bảng 4.

5 Kết luận

Kết quả nghiên cứu cho thấy sau khi áp dụng phương pháp trích chọn thuộc tính trên tập dữ liệu đã được thu thập ngay sau kết thúc mỗi học kỳ để xây dựng mô hình dự báo kết quả học tập toàn khóa, độ chính xác phân lớp của các mô hình đạt tỷ lệ phân lớp chính xác cao hơn so với trước trích chọn thuộc tính. Mô hình dự báo cuối khóa đạt kết quả cao nhất khi sử dụng kết quả trích chọn thuộc tính sau khi kết thúc học kỳ 6 và học kỳ 7, đạt gần 83%. Để cải thiện kết quả mô hình dự báo cuối khóa sau khi kết thúc học kỳ 1 và học kỳ 2, và kể cả các mô hình cho các học kỳ sau, nghiên cứu cần được bổ sung và phối hợp nhiều thuộc tính khác nhau thể hiện việc theo dõi lộ trình học như quản lý chuyên cần, quá trình và kể cả nền tảng học tập của người học từ các bậc học trước đó cũng như truyền thống học tập của gia đình sinh viên vào cơ sở dữ liệu. Các tập thuộc tính sau trích chọn của mỗi kỳ học để phục vụ cho dự báo kết quả cuối khóa, mô hình dự báo phân lớp dựa trên giải thuật cây quyết định J48, cây quyết định và tập luật tương ứng đều là các tài liệu hữu ích không chỉ giúp cho sinh viên mà còn giúp ích cho các nhà quản lý giáo dục trong việc ra quyết định và hỗ trợ sinh viên trong định hướng cho toàn bộ quá trình học tập của sinh viên.

Tài liệu tham khảo

1. Brijesh B. (2011), Mining Educational Data to Analyze Students' Performance, *International Journal of Advanced Computer Science and Applications*, 5(7), 65–75.
2. Dekker G. and Pechenizkiy M. (2009), *Predicting students drop out: A case study*, in International Conference on Educational Data Mining, 41–50, The Netherlands.
3. Yehuala M. A. (2015), Application Of Data Mining Techniques For Student Success And Failure Prediction, *International Journal of Scientific & Technology research*, 4(4), 342–250.
4. Baradwaj B., Pal S. (2012), Mining educational data to analyze students' performance, *IJACSA 2*, 4(4), 63–69.
5. Nithya P., Umamaheswari B., Umadevi A. (2016), A survey on educational data mining in field of education, *Journal Computer Science Software Development*, 7(8), 1–6.
6. Romero C., Sebastian V. (2007), Educational data mining: A survey from 1995 to 2005, *Expert Systems with Applications*, 3(5), 135–146.
7. Merceron A., Ycef K. (2005), *Educational Data mining: A case study*, *in International Conference on Artificial Intelligence in Education*, The Netherlands.
8. Bayer J., Bydzovska H., G´eryk J. (2012), *Predicting drop-out from social behaviour of students*, in the 5th International Conference on Educational Data Mining, Czech Republic.
9. Kapoor P.. Reena R. (2015), Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning, *International Journal of Engineering Research and General Science*, 5(7), 67–90.
10. R. Kohavi (1995), *A study of cross-validation and bootstrap for accuracy*, in International Joint Conference on Artificial Intelligence, Quebec, Canada.
11. Sharma P. (2014), Comparative Analysis of Various Decision Tree Classification Algorithms using WEKA, *International Journal on Recent and Innovation Trends in Computing and Communication*, 3(2), 684–690.
12. Lei X., Pingfan Y. (1988), *Best first strategy for feature selection*, the 9th International Conference in Pattern Recognition, Roma.
13. Aggarwal M. (2013), Performance Analysis Of Different Feature Selection Methods In Intrusion Detection, *International journal of scientific & technology research*, 2(6), 225–235.
14. Sembiring S., Hartama D. (2011), *Prediction of Student Academic Performance*, in International Conference on Management and Artificial Intelligence, Bali.
15. Delavari N., Beikzadeh M. R. (2005), *Application of Enhanced Analysis Model for*, in Juan Dolio, Dominican Republic.

16. Sarker, F., Thanassis T., Davis H. C. (2013), *Student's performance prediction by using institutional internal and external open data sources*, in 5th International Conference on Computer Supported Education. 6–8 May, Aachen Germany.

17. Do Q. H., Chen J. F. (2013), A Neuro-Fuzzy Approach in the Classification of Students' Academic Performance, *Computational Intelligence and Neuroscience*, 4(6), 60–67.

18. Kiranmai A., Jaya L. (2018), Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy, *Protection and Control of Modern Power System*, 8(6), 470–482.

MINING DATABASE OF THE TRAINING MANAGEMENT SYSTEM AT UNIVERSITY OF ECONOMICS, HUE UNIVERSITY

Mai Thu Giang*

University of Economics, Hue University, 99 Ho Dac Di St., Hue, Vietnam

Abstract: The prediction of the learning outcome and finding the factors that influence the outcome are extremely important for not only educational managers but also students. However, research on data mining applications in predicting learning outcomes at University of Economics, Hue University, has not been adequately exploited with the stored data. The purpose of this study is to apply the attribute selection technique and the classification technique with the decision tree algorithm, supported by the Waikato Environment for Knowledge Analysis (WEKA) software to build prediction models at the end of each semester. The results show that attributes including gender, cumulative major credits, and the average score of each semester are frequently retained in almost subsets results. Especially, the J48 algorithm returns the best model in predicting final results with the highest accuracy.

Keywords: attribute selection, J48, classification, data mining, decision tree