VJE Tạp chí Giáo dục (2024), 24(1), 48-53 ISSN: 2354-0753

DỰ BÁO KHẢ NĂNG TỐT NGHIỆP ĐÚNG HẠN CỦA SINH VIÊN: NGHIÊN CỨU TRƯỜNG HỢP TẠI TRƯỜNG ĐẠI HỌC ĐỒNG THÁP

Huỳnh Lê Uyên Minh[1,+],
Phạm Trường Trinh[2],
Nguyễn Văn Nhựt[2]

[1]Trường Đại học Đồng Tháp;
[2]Sinh viên K20, Trường Đại học Đồng Tháp
+ Tác giả liên hệ ● Email: hluminh@dthu.edu.vn

Article history
Received: 28/6/2023
Accepted: 26/7/2023
Published: 05/01/2024

Keywords
Data mining, on time graduation, computer science, students

ABSTRACT
Currently, data mining and knowledge discovery are areas of interest to scientists. The applications of data mining are implemented in many different fields such as: education, healthcare, finance, banking, trading, etc. Particularly in the field of education, the application of data mining has obtained multiple practical results. This study presents a 6-step data mining process and applies this process to predict the likelihood of on-time graduation of Computer Science students at Dong Thap University. The results obtained will help students timely determine their possibility to graduate on time so that they can adjust their study plans accordingly, thereby increasing the proportion of students graduating on time, improving the learning quality and training efficiency of the school.

1. Mở đầu

Trong những năm gần đây, khai phá dữ liệu (KPDL) và phát hiện tri thức đang là lĩnh vực được các nhà khoa học quan tâm nghiên cứu. Các ứng dụng KPDL được thực hiện trong nhiều lĩnh vực khác nhau như: giáo dục, y tế, tài chính, ngân hàng, kinh doanh,… Đặc biệt trong lĩnh vực giáo dục, ứng dụng KPDL đã thu được nhiều kết quả mang tính thực tiễn như: Nghiên cứu của Nguyễn Thị Uyên và Nguyễn Minh Tâm (2019) đã sử dụng phương pháp hai thuật toán KPDL là Logistic Regressio và Naive Bayes để tìm ra mô hình tốt nhất cho việc dự báo tình trạng học tập của SV. Nhóm tác giả Đỗ Thanh Nghị và cộng sự (2014) đã sử dụng giải thuật rừng ngẫu nhiên từ dữ liệu để rút trích các môn học quan trọng trong chương trình đào tạo ngành Công nghệ thông tin. Lưu Hoài Sang và cộng sự (2020) đã đề xuất một phương pháp dự báo kết quả học tập của sinh viên (SV) bằng kĩ thuật học sâu nhằm khai thác cơ sở dữ liệu trong hệ thống quản lí SV tại các trường đại học,…

Hiện nay, đa số các trường đại học và cao đẳng đang đào tạo theo học chế tín chỉ. Đối với hình thức đào tạo này, yêu cầu SV phải có sự chủ động cao, có sự lựa chọn mềm dẻo các môn học trong chuyên ngành đào tạo. SV cần tự phân bổ các môn học cho từng kì sao cho đủ số tín chỉ theo quy chế đào tạo, các em có thể học nhanh để tốt nghiệp sớm hoặc đúng thời hạn với số điểm cao, bên cạnh đó SV cũng cần hoàn thành các nội dung khác để đủ điều kiện xét tốt nghiệp như: Tin học, Ngoại ngữ, Giáo dục quốc phòng - an ninh,… Trên thực tế, đã có rất nhiều trường hợp thời gian học đã hết nhưng SV vẫn chưa hoàn thành đủ số tín chỉ, còn nợ môn chuyên ngành, các chứng chỉ và điều kiện liên quan. SV chưa quen và gặp nhiều khó khăn trong việc định hướng học tập, làm ảnh hưởng đến quá trình học tập của mình cũng như kết quả đào tạo của nhà trường.

Do vậy, việc phân loại và dự báo khả năng tốt nghiệp đúng hạn cho SV có vai trò quan trọng đối với hình thức đào tạo theo tín chỉ. Bài báo đề cập vấn đề ứng dụng kĩ thuật KPDL để dự báo khả năng tốt nghiệp đúng hạn của SV ngành Khoa học máy tính, Trường Đại học Đồng Tháp dựa trên công nghệ khám phá tri thức và KPDL. Kết quả thu được nhằm giúp SV nhận biết kịp thời khả năng tốt nghiệp đúng hạn để có sự điều chỉnh kế hoạch học tập; qua đó góp phần hỗ trợ cho công tác quản lí cũng như cố vấn học tập đạt hiệu quả cao hơn; giúp gia tăng tỉ lệ SV tốt nghiệp đúng hạn, nâng cao hiệu quả đào tạo của nhà trường.

2. Kết quả nghiên cứu

2.1. Một số vấn đề lí luận

2.1.1. Cây quyết định

Trong những năm qua, nhiều mô hình phân lớp dữ liệu đã được các nhà khoa học đề xuất như mô hình thống kê tuyến tính bậc 2, cây quyết định, di truyền,… Trong số các mô hình đó, cây quyết định với những ưu điểm riêng, được đánh giá là một công cụ mạnh, phổ biến và đặc biệt thích hợp cho KPDL nói chung và phân lớp dữ liệu nói riêng (Shafer et al., 1996; Pal & Pal, 2013). Ý tưởng cơ bản của cây quyết định có thể tóm tắt như sau: Cây quyết

định là dạng cấu trúc dữ liệu cây, mỗi nút trong có từ 2 đến nhiều nút con. Nút lá chỉ chứa dữ liệu thuộc về một lớp, nút thuần khiết (purity). Quá trình xây dựng cây được thực hiện theo luật trên xuống (top-down) (bắt đầu từ nút gốc, tất cả các dữ liệu học ở nút gốc): - Tại mỗi nút t trong cây, nếu tất cả phần tử dữ liệu thuộc về chỉ một lớp c thì gán nhãn nút t là c và return t là nút lá; - Ngược lại, tìm ra khả năng cắt dữ liệu (split) s* trong tất cả các khả năng s; - Tạo k nút con của t tương ứng với sự chia cắt s* với tập dữ liệu trong t. Gán nhãn cạnh nối từ t xuống các nút con dựa trên sự phân chia của s*; đồng thời phân hoạch các phần tử dữ liệu từ t xuống các nút con tương ứng; - Đệ quy bước 1 cho mỗi nút con (1…k) của t.

Dữ liệu mới được phân loại theo đường dẫn từ gốc đến lá của cây quyết định. Luật sinh ra dựa vào các nhãn của cạnh (IF - THEN), kết quả phân lớp dựa vào nhãn của nút lá mà dữ liệu mới xếp vào.

2.2.2. Rừng ngẫu nhiên các cây quyết định

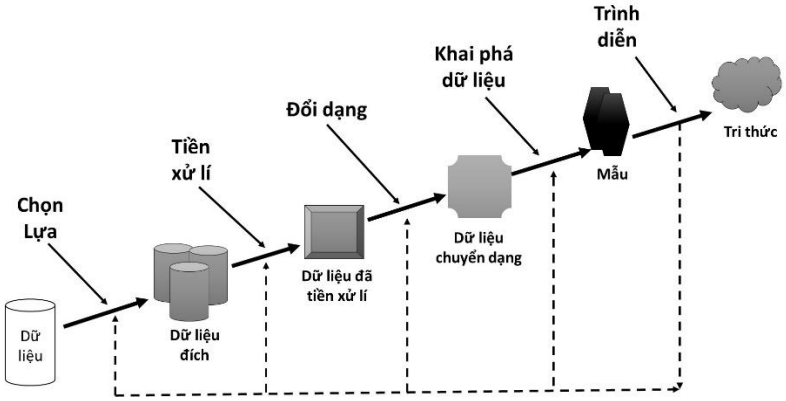
Tiếp cận rừng ngẫu nhiên do Breiman đưa ra năm 2001 là một trong những phương pháp tập hợp mô hình thành công nhất. Giải thuật rừng ngẫu nhiên tạo ra một tập hợp các cây quyết định không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap (lấy mẫu có hoàn lại từ tập học), tại mỗi nút, phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính (Quinlan, 1993; Bukralia et al., 2012). Lỗi tổng quát của rừng phụ thuộc vào độ chính xác của từng cây thành viên trong rừng và sự phụ thuộc lẫn nhau giữa các cây thành viên. Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho thành phần lỗi bias thấp (thành phần lỗi bias là thành phần lỗi của giải thuật học, độc lập với tập dữ liệu học) và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng. Tiếp cận rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán có giám sát hiện nay. Rừng ngẫu nhiên được đánh giá là giải thuật học nhanh, chịu đựng nhiễu tốt và tránh bị tình trạng học vẹt (Breiman, 2001). Giải thuật rừng ngẫu nhiên sinh ra mô hình có độ chính xác cao, đáp ứng được yêu cầu thực tiễn cho vấn đề phân loại, hồi quy.

Giải thuật máy học rừng ngẫu nhiên có thể được trình bày ngắn gọn như sau: - Từ tập dữ liệu học LS có m phần tử và n biến (thuộc tính), xây dựng T cây quyết định một cách độc lập với nhau; - Mô hình cây quyết định thứ t được xây dựng trên tập mẫu Bootstrap thứ t (lấy mẫu m phần tử có hoàn lại từ tập học LS); - Tại nút trong, chọn ngẫu nhiên n' biến (n'<< n) và tính toán phân hoạch tốt nhất dựa trên n' biến này.

Cây được xây dựng đến độ sâu tối đa không cắt nhánh. Kết thúc quá trình xây dựng T mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử mới đến X. Lặp lại thao tác trên k lần, tạo ra một rừng gồm k cây thành viên để thấy rõ quá trình xây dựng và áp dụng rừng ngẫu nhiên để phân lớp.

2.2. Quy trình khai phá dữ liệu

KPDL là một kĩ thuật phổ biến, được sử dụng để trích xuất thông tin hữu ích từ dữ liệu đã có, từ đó đưa ra các quyết định có lợi cho tương lai. Nghiên cứu của Fayyad và cộng sự (1996) đã tóm lược quá trình KPDL được tiến hành qua 6 giai đoạn như sau (xem hình 1):



Hình 1. Quy trình KPDL

- Bước 1: Xác định mục tiêu và hiểu về dữ liệu. Nhiệm vụ cơ bản trong giai đoạn này là: Xác định mục tiêu của quá trình KPDL, nắm rõ mục đích và yêu cầu của quá trình KPDL; Thu thập và khám phá dữ liệu từ các nguồn khác nhau, đảm bảo có đủ dữ liệu để thực hiện quá trình KPDL; Hiểu về cấu trúc và đặc điểm của dữ liệu, nhận biết các đặc trưng quan trọng và tiềm năng trong dữ liệu.

- *Bước 2: Chuẩn bị và tiền xử lí dữ liệu.* Xử lí dữ liệu thiếu bằng cách điền giá trị hoặc loại bỏ các mẫu dữ liệu không đầy đủ, xử lí dữ liệu nhiễu bằng các phương pháp như: làm sạch dữ liệu, lọc nhiễu hoặc sử dụng các kĩ thuật khác như smoothing hay interpolation.

- *Bước 3: Chuyển đổi dữ liệu.* Chọn và rút trích những đặc trưng quan trọng từ dữ liệu để tạo ra dữ liệu thuần và cung cấp thông tin cần thiết cho các thuật toán KPDL.

- *Bước 4: Áp dụng các thuật toán để KPDL.* Lựa chọn thuật toán phù hợp với mục tiêu và yêu cầu của dự án, tiến hành áp dụng và kiểm thử mô hình sử dụng các thuật toán KPDL, tìm ra mô hình tốt nhất phù hợp với dữ liệu và mục tiêu cụ thể.

- *Bước 5: Đánh giá và phân tích kết quả.* Đánh giá độ chính xác và hiệu suất của mô hình KPDL thông qua các phương pháp đánh giá như cross-validation, confusion matrix, precision, recall, F1-score và ROC curve. Phân tích và diễn giải kết quả để hiểu rõ hơn về thông tin và thông điệp được khám phá từ dữ liệu.

- *Bước 6: Triển khai và quản lí mô hình.* Triển khai mô hình KPDL vào môi trường thực tế để tận dụng các thông tin và lợi ích từ quá trình KPDL.

2.3. Minh họa quá trình khai phá dữ liệu để dự báo khả năng tốt nghiệp đúng hạn của sinh viên ngành Khoa học máy tính, Trường Đại học Đồng Tháp

Nghiên cứu của chúng tôi bao gồm bước nghiên cứu cơ bản và bước vận dụng mô hình để xây dựng ứng dụng. Ở bước nghiên cứu cơ bản được tiến hành sưu tầm, đọc, tra cứu tài liệu, sách báo có liên quan đến vấn đề KPDL, nghiên cứu các kĩ thuật và công cụ cho phép phân lớp trong KPDL; tìm hiểu về quy chế đào tạo đại học và cao đẳng hệ chính quy theo hệ thống tín chỉ, được ban hành kèm theo Quyết định số 1465/QĐ-ĐHĐT của Trường Đại học Đồng Tháp (2018). Sau đó, thực hiện bước tiền xử lí dữ liệu để trích chọn, làm sạch và biến đổi dữ liệu về dạng cấu trúc bảng cho phù hợp, từ đó áp dụng giải thuật rừng ngẫu nhiên của Breiman (2001) trên tập dữ liệu đã được tiền xử lí và đưa ra mô hình dự báo phù hợp cho khả năng tốt nghiệp đúng thời hạn của SV ngành Khoa học máy tính, Trường Đại học Đồng Tháp.

Ở bước vận dụng mô hình để xây dựng hệ thống, chúng tôi tập trung vào hai mô hình dự báo cho SV thuộc hai mốc thời gian như sau: thứ nhất là chức năng hỗ trợ dự đoán cho SV năm thứ 3 - đã có kết quả học tập thuộc các học kì 1, 2, 3, 4, 5; thứ hai là chức năng hỗ trợ dự đoán cho SV đầu năm thứ 4 - đã có kết quả học tập thuộc các học kì 1, 2, 3, 4, 5, 6. Ứng với mỗi mốc thời gian, ứng dụng sẽ đưa ra dự đoán khả năng tốt nghiệp đúng hạn của SV, từ đó SV sẽ có sự tự điều chỉnh kế hoạch học tập, cố gắng để có thể tốt nghiệp đúng hạn cũng như phấn đấu đạt được kết quả học tập tốt nhất. Vận dụng các bước KPDL của Fayyad và cộng sự (1996), chúng tôi đã triển khai quá trình KPDL như sau:

- *Bước 1: Xác định mục tiêu và hiểu về dữ liệu.* Trong nghiên cứu này, chúng tôi nghiên cứu về khả năng tốt nghiệp đúng hạn của SV ngành Khoa học máy tính Trường Đại học Đồng Tháp, thông tin có được dựa trên kết quả khai thác từ các thông tin cơ bản và kết quả học tập của SV tương ứng với các học kì trước đó (họ tên, độ tuổi, giới tính, số tín chỉ đạt, số tín chỉ không đạt, điểm trung bình theo học kì,…) nhằm giúp các em có thể tự đánh giá được khả năng tốt nghiệp đúng hạn của mình, từ đó có sự điều chỉnh và sắp xếp kế hoạch học tập của cá nhân cho phù hợp hơn, đồng thời giúp cho các giảng viên và cố vấn học tập theo dõi được khả năng tốt nghiệp đúng hạn của SV.

- *Bước 2: Chuẩn bị và tiền xử lí dữ liệu.* Để có được tập dữ liệu cho mô hình dự đoán, chúng tôi đã tìm hiểu và thu thập dữ liệu từ hệ thống quản lí đào tạo của Trường Đại học Đồng Tháp, bao gồm các bảng thống kê về kết quả học tập của SV đã tốt nghiệp ở các khóa đào tạo 2015, 2017, 2018 thuộc ngành Khoa học máy tính. Dữ liệu thu thập được có dạng bảng, bao gồm các tập tin ứng với kết quả học tập theo từng học kì của SV. Thông tin bao gồm họ tên, ngành học, lớp học, ngày sinh, giới tính,… các kết quả học phần liên quan đến điểm tương ứng với mỗi học kì, điểm trung bình hệ 4, điểm trung bình hệ 10, xếp loại học tập trong học kì, xếp loại kết quả tốt nghiệp. Dữ liệu thu được ở mỗi học kì được trình bày như sau (xem bảng 1).

Sau đó, dữ liệu được tính toán, xử lí và biến đổi về dạng tổng hợp, thống kê theo thông tin chung của SV, số tín chỉ đạt, số tín chỉ không đạt, số môn không đạt, điểm trung bình học kì và xếp loại học tập theo mỗi học kì như ở bảng 2.

Tiếp theo, chúng tôi tiến hành tiền xử lí dữ liệu. Dữ liệu thu thập sẽ được tổng hợp và chuyển về một bảng duy nhất, mỗi cột (field) của bảng biểu diễn một thuộc tính đặc thù riêng của từng người học, mỗi dòng (record) mô tả đầy đủ về các thông tin liên quan đến kết quả học tập của đối tượng người học tương ứng. Để làm được điều này, chúng tôi tiến hành xử lí dữ liệu thu được từ thông tin kết quả học tập ở từng học kì của SV, kết quả đưa về bảng dữ liệu gồm 56 cột, bao gồm mã số SV, giới tính và kết quả điểm tương ứng với các học phần, kết quả xếp loại tốt nghiệp của SV. Tổng hợp dữ liệu từ các học kì về một bảng duy nhất, xóa bỏ các trường dữ liệu không hợp lệ, các

thuộc tính không quan trọng. Ở bước này ta thu được một bảng đã bỏ qua các thuộc tính không quan trọng, ví dụ như mã số, họ tên, ngày sinh. Tiếp tục tổng hợp dữ liệu đã xử lí từ các lớp SV đã tốt nghiệp để đưa về một bảng dữ liệu duy nhất, trong đó giữ lại các trường dữ liệu giao nhau giữa các dữ liệu từ các lớp, cột cuối cùng là kết quả xếp loại tốt nghiệp ứng với mỗi SV.

Bảng 1. Bảng dữ liệu thu được ở mỗi học kì

STT	Mã số	Họ và tên		Ngày sinh	Giới tính	Toán cao cấp 1	Tiếng Anh 1	Điện tử căn bản	....	Kĩ thuật số	Toán rời rạc 1	Tin học căn bản	ĐIỂM TBC Hệ 4	ĐIỂM TBC Hệ 10	Xếp loại học tập
						2	3	2	....	3	3	2			
1	15411490	Mai Quốc	Việt	16/03/1997	Nam	6.1		9.1	...	9.2	6.7	8.5	3.43	8.3	Giỏi
2	15411500	Nguyễn Thành	Được	23/10/1997	Nam	7	9.4	9.1	...	7.2	5.6	8.7	3.36	8.13	Giỏi
3	15411514	Lưu Sĩ	Quí	03/05/1997	Nam	8.4		9.4	...	8.6	6.3	8.3	3.51	8.46	Giỏi
4	15411935	Nguyễn Quốc	Cường	17/11/1997	Nam	7.7	8.3	9.4	...	9.2	6.5	6.9	3	7.98	Khá
5	15411973	Nguyễn Phúc	Anh	22/09/1996	Nam	2.7	7.1	9.1	...	5.6	5.5	7.2	2.11	6.27	Trung bình
6	15412014	Võ Khánh	An	28/04/1997	Nam	3.7	7.5	7.2	...	7.6	6.2	4.9	1.98	6.22	Yếu
7	15412095	Nguyễn Văn Dương	Linh	23/11/1997	Nam	8.2	7.5	9.1	...	6.6	6.5	5.7	3.28	7.96	Giỏi
8	15412124	Trần Trung	Chánh	22/09/1997	Nam	4.9	7	8.5	...	7.6	6.4	7.5	2.65	7.09	Khá
9	15412136	Đỗ Nguyễn Anh	Huy	20/02/1997	Nam	4.6	7.1	9.1	...	7.4	5.7	7.2	2.56	6.9	Khá
10	15412161	Huỳnh Thị Diễm	Trinh	14/07/1997	Nữ	7	6.2	8.4	...	7.6	5.6	6.2	2.31	6.62	Trung bình
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Bảng 2. Bảng dữ liệu xử lí tổng hợp ở mỗi học kì

STT	Mã số	Họ và tên		Ngày sinh	Giới tính	Tín chỉ Đạt	Tín chỉ Không đạt	Số môn Không đạt	Điểm trung bình	Xếp loại học tập
1	15411490	Mai Quốc	Việt	16/03/1997	Nam	18	0	0	7.43	Giỏi
2	15411500	Nguyễn Thành	Được	23/10/1997	Nam	19	0	0	7.65	Giỏi
3	15411514	Lưu Sĩ	Quí	03/05/1997	Nam	18	0	0	7.54	Giỏi
4	15411935	Nguyễn Quốc	Cường	17/11/1997	Nam	17	0	0	7.95	Khá
5	15411973	Nguyễn Phúc	Anh	22/09/1996	Nam	17	2	1	6.64	Trung bình
6	15412014	Võ Khánh	An	28/04/1997	Nam	15	2	1	6.57	Yếu
7	15412095	Nguyễn Văn Dương	Linh	23/11/1997	Nam	19	0	0	7.27	Giỏi
8	15412124	Trần Trung	Chánh	22/09/1997	Nam	19	0	0	7.12	Khá
9	15412136	Đỗ Nguyễn Anh	Huy	20/02/1997	Nam	17	0	0	6.71	Khá
10	15412161	Huỳnh Thị Diễm	Trinh	14/07/1997	Nữ	17	0	0	6.6	Trung bình
...	...	...	...	...	...	...	...	...	...	...

- *Bước 3: Chuyển đổi dữ liệu.* Sử dụng các kĩ thuật trong xử lí dữ liệu để loại bỏ nhiễu và các dữ liệu bị khuyết, không cần thiết, trích chọn các thuộc tính có giá trị cho việc phân lớp, chuyển đổi dữ liệu về dạng thích hợp để sử dụng một cách hiệu quả, phù hợp cho quá trình xử lí. Kết quả thu được ở 2 bảng dữ liệu: + Tập dữ liệu bao gồm 16 thuộc tính độc lập, được sử dụng nhằm dự đoán kết quả tốt nghiệp đúng hạn dành cho SV đã có kết quả học tập ứng với các học kì 1, 2, 3, 4, 5 (SV năm thứ 3); + Tập dữ liệu bao gồm 19 thuộc tính độc lập, được sử dụng cho mục đích dự đoán kết quả tốt nghiệp đúng hạn cho SV đã có kết quả học tập ứng với các học kì 1, 2, 3, 4, 5, 6 (SV năm thứ 4).

- *Bước 4: Áp dụng thuật toán KPDL để xây dựng mô hình dự báo.* Chương trình xử lí được thực hiện dựa trên 02 tập dữ liệu đầu vào từ kết quả trên, được sử dụng tương ứng cho 2 mô hình dự báo. Quá trình thực nghiệm được tiến hành với ngôn ngữ lập trình Python trong môi trường Anaconda. Đầu tiên, chúng tôi sử dụng giao thức kiểm tra chéo (k-folds cross-validation) để phân chia tập dữ liệu ra làm hai phần, gồm: tập dữ liệu đào tạo (training dataset) và tập dữ liệu kiểm tra (testing dataset). Sử dụng hàm train_test_split của sklearn.model selection với thông số phân chia test size = 0,3; random state = 150 nhằm mang đến kết quả đánh giá tốt nhất cho mô hình. Tiếp theo, nghiên cứu đã sử dụng gói chương trình rừng ngẫu nhiên các cây quyết định (RandomForestClassifier) được cung cấp trong sklearn.ensemble, tiến hành xây dựng mô hình thực nghiệm rừng ngẫu nhiên với 100 cây quyết định và lưu mô hình dự báo bằng cách tuần tự hóa với hàm pickle.dump() và tên Randomforestmodel.pkl.

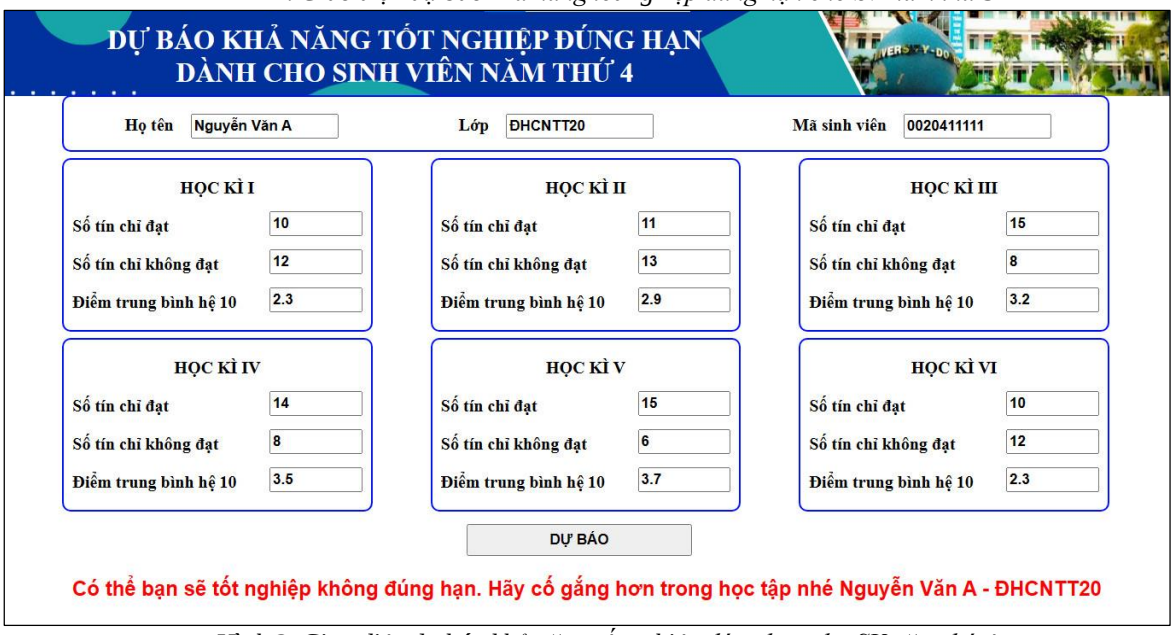
- *Bước 5: Đánh giá và phân tích kết quả.* Kết quả quá trình kiểm tra mang lại độ chính xác cao (chiếm 92,45%), được thể hiện qua ma trận trực giao (confusion matrix) nhằm lưu trữ kết quả phân lớp - dự đoán ở giai đoạn kiểm tra. Bên cạnh đó, chúng tôi cũng đánh giá mức độ quan trọng, độ ảnh hưởng của các thuộc tính độc lập đến kết quả của quá trình phân lớp.

- *Bước 6: Triển khai mô hình.* Tiếp tục sử dụng mô hình Randomforestmodel.pkl được đánh giá và lưu trữ ở bước 5, vận dụng cho việc xây dựng ứng dụng web để dự báo khả năng tốt nghiệp đúng hạn cho SV ngành Khoa học máy

tính, Trường Đại học Đồng Tháp. Để thực hiện, chúng tôi sử dụng các hàm Flask, Pandas, Scikit-learn, Numpy, Pickle, Sklearn, render_template, request, redirect, url for từ Flask để nhúng mô hình học máy đã có vào ứng dụng web. Ở bước này, chương trình tạo một đối tượng ứng dụng thuộc lớp Flask (name) của module flask. Đối tượng này sẽ đại diện cho ứng dụng web, cho phép định nghĩa các đường dẫn và các hàm xử lí cho từng đường dẫn. Tiếp theo, chương trình đưa mô hình từ file nhị phân Randomforestmodel.pkl bằng cách sử dụng module pickle.

Chúng tôi đã định nghĩa một chức năng định tuyến (router) cho ứng dụng web bằng cách sử dụng lệnh @app.route('/'). Định tuyến này sẽ trả về trang chủ của ứng dụng web khi người dùng truy cập vào địa chỉ http://127.0.0.1:5000/. Trang web được hiển thị (render) từ file HTML có tên index.html bằng cách sử dụng hàm render template () của module flask. File HTML này chứa một giao diện để người dùng nhập các thông tin đầu vào và một nút để gửi yêu cầu dự đoán. Giao diện ứng dụng tương ứng được trình bày như trong hình 2 và hình 3 để dự báo kết quả tốt nghiệp đúng hạn dành cho SV (năm thứ 3 và năm thứ 4 khoá K20) ngành Khoa học máy tính đang học tập tại Trường:

Hình 2. Giao diện dự báo khả năng tốt nghiệp đúng hạn cho SV năm thứ 3



Hình 3. Giao diện dự báo khả năng tốt nghiệp đúng hạn cho SV năm thứ 4

Kết quả nghiên cứu nhằm cung cấp thông tin hữu ích, giúp SV ngành Khoa học máy tính, Trường Đại học Đồng Tháp có thể đánh giá khả năng của mình trong việc tốt nghiệp đúng hạn dựa trên học lực hiện tại. Thông qua giao diện ở hình 2 sẽ hỗ trợ cho SV đang học ở học kì 6 (năm thứ 3) và giao diện hình 4 sẽ hỗ trợ cho SV đang học ở học kì 7 (năm thứ 4) có thể tự thực hiện để nhận được dự báo về tình trạng học tập cá nhân. Bên cạnh đó, để tiết kiệm thời gian cho người dùng, chương trình chỉ cho phép người dùng nhập vào dữ liệu kiểu số tương ứng trong thang điểm quy định.

SV sử dụng các ứng dụng và cung cấp đầy đủ thông tin cần thiết về kết quả học tập của cá nhân tương ứng với các học kì, sau đó ấn vào chức năng dự báo, hệ thống sẽ đưa ra kết quả. Kết quả thu được từ nghiên cứu có thể ứng dụng cho các trường đại học để nhận biết kịp thời khả năng tốt nghiệp đúng hạn của SV, giúp các em có những điều chỉnh cho tiến trình học tập của bản thân, từ đó gia tăng tỉ lệ tốt nghiệp đúng hạn và hiệu quả đào tạo của nhà trường.

3. Kết luận

Kết quả nghiên cứu đã trình bày một tiếp cận KPDL giáo dục để dự báo khả năng tốt nghiệp đúng hạn của SV ngành Khoa học máy tính, Trường Đại học Đồng Tháp. Các bước thực hiện bao gồm thu thập dữ liệu của SV đã tốt nghiệp thuộc các khóa đào tạo khác nhau ngành Khoa học máy tính, sau đó thực hiện bước tiền xử lí và chuyển đổi dữ liệu để có thể sử dụng mô hình rừng ngẫu nhiên, đánh giá mô hình thu được và triển khai xây dựng ứng dụng dự báo dành cho SV đang học năm thứ 3 và năm thứ 4. Trong những nghiên cứu tiếp theo, chúng tôi sẽ tiếp tục thu thập dữ liệu nhiều hơn nữa nhằm xây dựng mô hình và ứng dụng một cách đầy đủ hơn, giúp SV cũng như các nhà quản lí có đầy đủ thông tin khi cần tư vấn về tiến độ học tập của SV tại trường.

Lời cảm ơn: *Nhóm tác giả cảm ơn sự tài trợ của Trường Đại học Đồng Tháp qua đề tài nghiên cứu khoa học của sinh viên: "Áp dụng kĩ thuật khai phá dữ liệu để dự báo khả năng tốt nghiệp đúng hạn của sinh viên ngành Khoa học máy tính Trường Đại học Đồng Tháp" với mã số: SPD2022.02.16.*

Tài liệu tham khảo

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Bukralia, R., A-V., Deokar, A-V., Sarnikar, S., & Hawkes, M. (2012). *Using Machine Learning Techniques in Student Dropout Prediction*. Chapter 7 in Cases on Institutional Research Systems, Hansel Burley Eds., IGI Global.

Đỗ Thanh Nghị, Phạm Nguyên Khang, Nguyễn Minh Trung, Trịnh Trung Hưng (2014). Phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập sinh viên ngành Công nghệ thông tin. *Tạp chí Khoa học, Trường Đại học Cần Thơ*, *33*(1), 49-57.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37-54.

Lưu Hoài Sang, Trần Thanh Điện, Nguyễn Thanh Hải, Nguyễn Thái Nghe (2020). Dự báo kết quả học tập bằng kĩ thuật học sâu với mạng nơ-ron đa tầng. *Tạp chí Khoa học, Trường Đại học Cần Thơ*, *56*(3A), 20-28.

Nguyễn Thị Uyên, Nguyễn Minh Tâm (2019). Dự đoán kết quả học tập của sinh viên bằng kĩ thuật khai phá dữ liệu. *Tạp chí Khoa học, Trường Đại học Vinh*, *48*(3A), 68-73.

Pal, A. K. & Pal, S. (2013). Analysis and Mining of Educational Data for Predicting the Performance of Students. *In International Journal of Electronics Communication and Computer Engineering*, *4*(5), 2278-4209.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Shafer, J., Agrawal, R., & Mehta, M. (1996). *SPRINT: a scalable parallel classifier for data mining*. In Proceedings of 22nd International Conference on Very Large Data Bases.

Trường Đại học Đồng Tháp (2018). *Quy chế đào tạo đại học và cao đẳng hệ chính quy theo hệ thống tín chỉ* (ban hành kèm theo Quyết định số 1465/QĐ-ĐHĐT ngày 23/10/2018 của Hiệu trưởng Trường Đại học Đồng Tháp).