ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRONG GIÁO DỤC: NGHIÊN CỨU TRƯỜNG HỢP VỀ DỰ BÁO KẾT QUẢ HỌC TẬP CỦA SINH VIÊN ĐẠI HỌC NĂM THỨ NHẤT

PGS.TS. Nguyễn Thị Thu Thủy[1*], ThS. Nguyễn Thị Liệu[2]

[1]*Trường Đại học Thương Mại*
[2]*Trường Đại học Công nghệ Đồng Nai*
*Tác giả liên hệ: *Nguyễn Thị Thu Thuỷ, nguyenTthuthuy@gmail.com*

THÔNG TIN CHUNG

Ngày nhận bài: 27/07/2023

Ngày nhận bài sửa: 27/08/2023

Ngày duyệt đăng: 21/09/2023

TỪ KHOÁ

Khai phá dữ liệu;
Dự báo;
Học máy;
Giáo dục.

TÓM TẮT

Thành tích học tập của sinh viên đại học sau một học kỳ hoặc một năm học trực tiếp thể hiện quá trình phấn đấu, rèn luyện của mỗi sinh viên trong trường đại học. Các kết quả này sẽ có tác động đến những chuỗi các hoạt động giáo dục của bản thân các em những kỳ kế tiếp hay xa hơn là cơ hội phát triển bản thân trong việc tìm kiếm việc làm và học tập ở bậc cao hơn. Trên thực tế, các em sinh viên năm thứ nhất là có sự thay đổi một cách đáng kể nhất về thành tích học tập so với các em sinh viên khóa trên. Do có nhiều yếu tố tác động đến thành tích học tập của sinh viên như nơi ở, gia đình, mối quan hệ khác giới,...và quan trọng nhất là sự chuyển đổi từ giáo dục phổ thông lên bậc Đại học. Nghiên cứu này thu thập dữ liệu với các em sinh viên năm thứ nhất nhằm phân tích tích dữ liệu để xác định những yếu tố chủ yếu tác động đến kết quả học tập của các em. Sau đó, nghiên cứu sẽ sử dụng mô hình cây quyết định trong khai phá dữ liệu để dự đoán thành tích học tập của các em với mục đích giúp cho các em nhận thức được việc học và có chiến lược học trong những năm học kế tiếp. Thêm vào đó, nghiên cứu hi vọng có thể trợ giúp cho các nhà quản lý trong công tác quản lý sinh viên ở bậc đại học.

ABSTRACT

The academic results of university students after a semester or a year directly reflect their efforts and learning progress in higher education environment. These results

will have an impact on the chains of educational activities of the students in subsequent semesters or the future, such as opportunities for personal development in seeking employment and higher education. In fact, first-year students show the most significant changes in academic performance compared to students in other years. This proves that there are many factors that affect students' academic performance, such as their place of residence, family, and relationships, etc. This study collected data from first-year university students to analyze the data and determine the main factors that affect their academic performance. At the same time, the study employed a decision tree model in data mining to predict the academic performance of students with the aim of helping them understand the learning process and develop a learning strategy in the following years. Additionally, the study hopes to assist university administrators in student managements at university level.

1. GIỚI THIỆU

Tình trạng nhiều sinh viên rơi vào trường hợp bị cảnh báo học tập trước khi buộc thôi học đang diễn ra ngày càng phổ biến. Nhiều gia đình không biết tình hình học tập của sinh viên cho đến khi có thông báo về tình trạng cảnh báo học tập đến gia đình. Một trong những nguyên nhân dẫn đến tình trạng này là ý thức học tập của các sinh viên trong quá trình học tập và rèn luyện dẫn đến kết quả học tập không hiệu quả. Vì vậy, việc khai phá các dữ liệu liên quan để dự đoán kết quả học tập và tìm xem các nhân tố ảnh hưởng đến kết quả học tập của sinh viên là vấn đề được các trường đại học quan tâm. Đối với các em sinh viên năm thứ nhất, sự thay đổi môi trường học tập, thay đổi mối quan hệ bạn bè,… được coi là các nhân tố quan trọng cần xem xét hơn so với các sinh viên các khóa sau. Chính vì vậy, nghiên cứu muốn tìm hiểu và khai thác các khía cạnh này ảnh hưởng thế nào đến kết quả học tập của các em sinh viên. Để làm được điều đó, nghiên cứu đề xuất sử dụng phương pháp cây quyết định để làm công cụ dự đoán kết quả học tập của các em. Việc áp dụng công cụ này dựa trên các ưu điểm của phương pháp về tính dễ giải thích, phù hợp với các dữ liệu rời rạc,... (Witten và Frank, 2005).

Khai phá dữ liệu, còn được hiểu là sự tìm tòi, phát hiện các tri thức mới trong cơ sở dữ liệu cho trước. Nói cách khác đây là sự khám phá thông tin mới lạ và rút trích các thông tin hữu ích cho người dùng từ một tập dữ liệu cho trước. Khai phá dữ liệu đã được áp dụng trong rất nhiều lĩnh vực khác nhau trong đời sống xã hội như:

kinh doanh, phân tích thói quen mua sắm của các khách hàng,… Trong những năm gần đây, khai phá dữ liệu ngày càng được quan tâm với sự bùng nổ của dữ liệu. Ví dụ như, nó có thể được sử dụng trong việc khai thác các dữ liệu điều tra xã hội học và một trong số đó có lĩnh vực giáo dục. Việc khai phá dữ liệu trong giáo dục thường được xem xét ở nhiều cấp độ khác nhau và thông thường việc này được tích hợp với các phương pháp khác như thống kê, phân tích, để đạt được một mục tiêu nào đó trong giáo dục (Baker, 2010, Baker & Yacef, 2009). Ví dụ, khai phá dữ liệu có thể được sử dụng trong trường về phần mềm sẽ được học đối với học phần Tin học Đại cương. Dữ liệu có thể phân tích ở góc độ cách sinh viên đã tiếp cận đến máy tính thế nào? Cách sinh viên đã được trang bị các kiến thức nền tảng ra sao để có thể quyết định dạy cho sinh viên về các kỹ năng tin học văn phòng hay dạy cho họ cả những tính toán về cơ sở dữ liệu. Tất cả những yếu tố này sẽ đóng vai trò quan trọng trong việc nghiên cứu dữ liệu giáo dục.

Tóm lại theo Baker (2010), khai phá dữ liệu trong lĩnh vực giáo dục có thể hiểu là việc khám phá và phát hiện tri thức từ dữ liệu trong lĩnh vực giáo dục và sử dụng các phương pháp, công cụ thống kê hay công cụ khác để hiểu rõ hơn về đối tượng học sinh, sinh viên với các vấn đề liên quan đến họ trong lĩnh vực này.

Có thể phân chia khai phá dữ liệu thành các loại sau (Baker, 2010):

- Dự đoán (Prediction): Dự đoán sẽ bao gồm các phương pháp như: Phân loại; hồi quy; Ước tính mật độ;…
- Phân cụm: Các phương pháp được sử dụng trong phân cụm sẽ bao gồm khai phá mối quan hệ; khai phá luật kết hợp; khai thác tương quan; khai thác mẫu tuần tự; khai thác dữ liệu nhân quả;…
- Ngoài ra còn có chắt lọc (Distillation) và khai phá dữ liệu với các mô hình (Discovery with models).

Trong bài toán dự đoán, mục tiêu của bài toán là xây dựng và phát triển thuật toán hoặc mô hình mà từ đó có thể tính toán được hay suy ra một giá trị duy nhất (output) (kết quả của dự đoán) từ một số các kết hợp của các yếu tố khác của dữ liệu (các biến dự đoán). Sự dự đoán này cần yêu cầu phải có nhãn cho trước biến đầu ra, trong đó nhãn được coi là sự đại diện cho các kết quả tin cậy về giá trị của biến đầu ra trong các trường hợp cụ thể.

Ứng dụng của phương pháp dự đoán có hai nhân tố nền tảng là: Ứng dụng việc dự đoán trong nghiên cứu để tìm ra các yếu tố quan trọng mà các thông tin đưa ra đang trong giai đoạn thu thập và xây dựng cơ sở dữ liệu. Các nghiên cứu liên quan đến phương pháp này có thể xem ở các công bố trước đây như: dự đoán kết quả trượt hoặc đỗ trong học tập (Romero et al, 2007, Dekker et al, 2009) hay quản lý học tập của sinh viên (Lin, 2012). Yếu tố cốt lõi thứ hai là các phương pháp dự đoán được sử dụng để dự đoán giá trị đầu ra sẽ là bao nhiêu của tập dữ liệu mà không quan tâm đến nhãn cho cấu trúc đó. Ví dụ như các tập dữ liệu dự đoán kết quả điểm số của sinh viên dựa vào các điểm số của họ trước đó mà không cần gán nhãn cho biến kết quả đầu ra.

Từ các mô hình dự đoán ban đầu với một tập dữ liệu nhỏ, có thể phát triển mô hình dự đoán bằng cách sử dụng dữ liệu được thu thập tự động từ tương tác giữa sinh viên và phần mềm cho các biến dự đoán. Sau đó mô hình sẽ tính toán độ chính xác để khái quát hóa cho các sinh viên và ngữ cảnh bổ sung (Baker et al, 2009). Trong bài toán dự đoán, với phương pháp phân loại, biến dự đoán được dùng là biến nhị phân hoặc phân loại trong khi hồi quy thì biến dự đoán là biến liên tục và ước tính mật độ thì biến dự báo sẽ là hàm tính xác xuất mật độ.

Ở bài toán phân cụm (clustering), mục đích là đi tìm các nhóm tương đồng và phân chia tập dữ liệu thành các nhóm khác nhau. Ví dụ, nhóm các trường có chất lượng tương đồng hoặc nhóm sinh viên có lực học tương tự (Amershi & Conati, 2006; Beal, Qu, & Lee, 2006).

Trong bài toán khai phá mối quan hệ, mục đích là đi tìm mối quan hệ giữa các biến trong tập dữ liệu có nhiều biến. Nhiệm vụ chủ yếu của loại bài toán này là tìm hoặc đánh giá mức quan trọng của các biến, hay đi tìm mối tương quan giữa hai biến trong tập dữ liệu. Chắt lọc dữ liệu với mô hình được sử dụng thông qua các phương pháp kể trên là dự đoán, phân lớp, hoặc lọc dữ liệu. Mô hình được coi là một thành phần trong hệ thống phân tích dữ liệu để dự đoán dữ liệu hoặc khai phá mối quan hệ của các nhân tố trong tập dữ liệu của hệ thống mà ở đó dữ liệu đã được lọc và hiển thị ở dạng "dễ nhận biết" đối với người dùng.

Trong nghiên cứu này, phương pháp phân loại (classification) được áp dụng để dự đoán mức độ hiệu quả của việc học tập cho sinh viên năm thứ nhất của đại học. Nghiên cứu được tổ chức như sau: Phần 2 trình bày các tài liệu liên quan đến khai phá dữ liệu trong giáo dục, Phần 3 trình bày một nghiên cứu trường hợp về việc ứng dụng khai phá dữ liệu trong việc dự báo kết quả học tập của sinh viên năm thứ nhất của đại học. Các hàm ý quản lý và kết luận được trình bày ở phần thứ 4 của nghiên cứu.

2. TỔNG QUAN VỀ DỰ BÁO KẾT QUẢ HỌC TẬP

2.1. Khái niệm và các nhân tố chính của việc dự báo kết quả học tập

Trong khoa học giáo dục, kết quả học tập là chỉ mức độ đạt được về mặt kiến thức, kỹ năng hay nhận thức của người học trong một lĩnh vực nào đó thông qua điểm của các môn học (học phần giảng dạy) và chương trình giảng dạy trong trường Đại học. Để đo lường nó, người ta sử dụng các tiêu chí khác nhau (thi cử, điểm số,…). Trong các trường Đại học, kết quả học tập có thể đo thông qua thang điểm 10 hoặc quy đổi sang thang điểm 4 theo thông lệ quốc tế. Cụ thể thì với bài báo này, kết quả học tập mà bài báo muốn dự đoán được thực hiên thông qua chỉ số GPA (Grade Point Average) của sinh viên trong kỳ học hay năm học. Đây là hệ thống đánh giá theo hệ thống giáo dục của Mỹ. Trong đó điểm GPA của một sinh viên được tính thông quan tổng điểm trung bình của các học phần khác nhau trong kỳ và sau đó chia đều để có GPA cuối cùng trên thang điểm 4.

Những nhân tố ảnh hưởng tới kết quả học tập

- Mối quan hệ khác giới: Trong tâm lý học, mối quan hệ khác giới đối với sinh viên là việc phát triển tâm sinh lý thông thường. Tuy nhiên, nó có thể tồn tại thành hai thái cực tốt và xấu. Mặt tốt là, nó có thể giúp cho nhau cùng tiến bộ, có thể đồng cảm và cùng giúp đỡ nhau cố gắng trong học tập. Ngược lại, nó có thể làm sao nhãng trong học tập, có những biểu hiện tiêu cực, ảnh hưởng đến việc học tập.
- Gia đình: Yếu tố gia đình góp phần định hướng sự phát triển của mỗi em sinh viên. Gia đình chính là yếu tố quan trọng trong sự phát triển về nhân cách, lối sống của thanh niên. Chính vì vậy, yếu tố gia đình cũng có thể hiểu là nhân tố quan trọng, có ảnh hưởng trực tiếp đến kết quả học tập của các em. Trong yếu tố gia đình, sự hỗ trợ về mặt tài chính cho việc học tập của sinh viên được coi là yếu tố cần thiết để sinh viên có thể tập trung vào việc học.
- Nơi ở: Việc được ở trong ký túc xá của trường cũng giúp cho sinh viên trong việc đi lại, học tập một cách dễ dàng. Sinh viên có thể dễ dàng trao đổi việc học, gần gũi và giao tiếp với nhiều người hơn. Mở rộng mối quan hệ, tự tin giao tiếp và rèn được kỹ năng sống chung với một tập thể cũng là ưu điểm của sinh viên. Sự lựa chọn thứ hai đối với sinh viên Đại học là việc ở trọ. Đối với việc này, các sinh viên thích thú với việc lựa chọn này vì nó thỏa mãn sự tự do của tuổi trẻ. Bên cạnh sự tự do được làm điều mình thích, việc ở trọ cũng rèn cho sinh viên được tính cách độc lập, tự làm chủ cuộc sống và tính tự lập cá nhân. Chính vì các lựa chọn khác nhau nên cũng có thể hiểu đây cũng là yếu tố tác động đến kết quả học tập của sinh viên, nhất là sinh viên năm thứ nhất. Lý do đối với sinh viên năm nhất là do các em phải rời xa gia đình đến một nơi ở khác khi vào học đại học.
- Ngoài ra còn có các yếu tố khác cũng ảnh hưởng đến việc học tập đó là quan điểm cá nhân và phương pháp học tập. Quan điểm cá nhân về ngành học mình lựa chọn cũng như chương trình mình đang học ở mức độ nào. Phương pháp học tập liên quan đến cách thức học như: học khi đến mùa thi, học trên thư viện, phương pháp tự học,…
- Các hoạt động khác bao gồm các hoạt động tham gia câu lạc bộ, học thêm,… cũng ảnh hưởng đến kết quả học tập của sinh viên.

2.2. Một số kỹ thuật khai phá dữ liệu

Trong khai phá dữ liệu, một số kỹ thuật sử dụng được phân chia thành các phương pháp như sau (Lee and Siau, 2001):

- Phương pháp thống kê: phương pháp này thường được sử dụng để phát hiện các ngoại lai, xử lý các dữ liệu khuyết thiếu trong tập dữ liệu trước khi đưa vào các mô hình thực nghiệm tiếp theo.
- Phương pháp luật kết hợp: Phương pháp này thường được sử dụng để khai phá các dữ liệu giao dịch hoặc cơ sở dữ liệu quan hệ vì phương pháp này có thể quét các giao dịch nhiều lần để tìm ra mối quan hệ giữa chúng.
- Phương pháp mô hình hóa dữ liệu (visualization): Các kỹ thuật được sử dụng như: ma trận phân tán satterplot matrices, ma trận giải phẫu (prosection matrices), phép

chiếu ma trận (projection matrices),… để trực quan hóa dữ liệu nhiều chiều.

- Các phương pháp sử dụng trí tuệ nhân tạo (AI): Các kỹ thuật như nhận dạng mẫu (pattern recognition), học máy (machine learning) và mạng nơ ron (neural networks) được sử dụng để phân loại (classification), phân cụm (clustering),… Trong các kỹ thuật này, cây quyết định được coi là kỹ thuật cơ bản để xây dựng các luật theo cấu trúc hình cây và dựa vào đó có thể giải thích một cách rõ ràng kết quả quyết định của mô hình. Chính vì lý do này nên nghiên cứu lựa chọn mô hình cây quyết định là mô hình thực nghiệm cho bài toán dự báo kết quả học tập của sinh viên đại học.

2.3. Mô hình cây quyết định trong bài toán dự báo của khai phá dữ liệu

Mô hình cây quyết định (Decision Tree) (Witten và Frank, 2005) là một trong các phương pháp dự đoán cơ bản nhất của lớp bài toán phân nhãn hoặc hồi quy. Cây quyết định được thể hiện như một cấu trúc cây nhị phân mà ở đó các nhánh được mọc ra từ gốc và phát triển thành các nhánh con cho đến tận lá cây. Việc phân loại các đối tượng sẽ được thực hiện dựa vào dãy các luật duyệt trên cây. Luật chính là yếu tố cơ bản để xây dựng gốc và các nút nhánh trên cây. Như vậy, gốc cây và các nút nhánh chính là thể hiện của các thuộc tính phản ánh về dữ liệu. Việc quyết định đi theo nhánh nào diễn ra ở các nút và quyết định cuối cùng sẽ nằm ở các nút lá. Tập các nút bao gồm nút gốc, nút nhánh (nút quyết định, và nút lá) tạo thành tập các luật để đưa ra quyết định cuối cùng cho đối tượng cần xem xét.

Các tiêu chí đánh giá cây

Chỉ số Gini: Chỉ số Gini (Gini Index) hay còn gọi là hệ số Gini. Chúng thường được biểu diễn ở dạng tỷ lệ phần trăm. Hệ số này cho phép đánh giá mức độ phân nhánh (purity) của từng nút của cây quyết định.

Cách tính chỉ số Gini của nút trên cây (NODE t):

$$Gini(t) = 1 - \sum_{j=1}^{c} p(j|t)^2 \ (1)$$

Trong đó $p(j|t)$ là tần suất của lớp j trong nút t

Giá trị gini là lớn nhất (bằng 1) khi các quan sát phân bố đều trên các lớp, và giá trị thấp nhất (bằng 0) khi các quan sát chỉ thuộc về một lớp.

Độ lợi thông tin (Information Gain-IG): đây là phép đo sự thay đổi thông tin trong Entropy sau khi phân nhánh tập dữ liệu đối với một thuộc tính A nào đó. Độ lợi thông tin này sẽ tính toán lượng thông tin có lợi mà một thuộc tính sẽ cung cấp cho người dùng về lớp mà cây đang cần phân loại. Có 2 cách tính độ lợi thông tin như sau:

Cách 1: Độ lợi thông tin được đo bằng lượng Entropy của nút nhánh (nút cha mẹ) trừ đi trung bình Entropy của các nút con của nhánh đó.

Cách 2: Gọi S là số mẫu huấn luyện của tập dữ liệu ban đầu; S_i: số lượng mẫu của tập S có trong lớp C_i với i = {1, …, m}. Giả sử gọi A là thuộc tính của tập dữ liệu cần phân chia thành n tập con $\{S_1, S_2, …, S_n\}$, gọi S_{ij} là số lượng mẫu của lớp C_i trong tập S_j ($A=a_j$), thì:

$I(s_1, s_2, \dots, s_m)$ là thông tin cần biết để phân lớp một mẫu, và được tính bằng:

$$I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} log_2 \frac{s_i}{s} \quad (2)$$

Khi đó Entropy của thuộc tính A:

$$E(A) = \sum_{j=1}^{n} \frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) \quad (3)$$

Và độ lợi thông tin khi cây sẽ phân chia thành các nhánh khác nhau dựa trên thuộc tính A sẽ là:

$$IG(A) = I(s_1, s_2, \dots, s_m) - E(A) \quad (4)$$

Tại mỗi nhánh, việc chia nhánh sẽ được thực hiện nếu nhánh nào đó có độ lợi lớn nhất so với các nhánh còn lại.

Các bước xây dựng cây quyết định

Bước 1. Phân chia dữ liệu ban đầu thành các tập dữ liệu con khác nhau một cách đệ quy.

Bước 2. Xác định thuộc tính của dữ liệu cho mỗi nút và tìm xem luật liên kết với nút sao cho việc phân tách các giá trị trên nút là tốt nhất.

Bước 3. Sử dụng luật vừa tìm được tại nút để tách nhánh của cây.

Bước 4. Lặp lại các bước nói trên đối với các nút con khác trên cây.

Bước 5. Lặp lại quá trình phân tách nút cho đến khi thỏa mã điều kiện dừng của cây.

Bước 6. Gán nhãn quyết định tại các nút lá căn cứ vào tập nhãn đã cho.

Các chỉ số đánh giá hiệu quả của phương pháp dự đoán.

Một số tiêu chí để đánh giá cho tất cả các kỹ thuật phân loại là độ chính xác dự báo được tính toán dựa trên ma trận nhầm lẫn (Confusion matrix) hay đường cong ROC. Trong bài báo này, ma trận nhầm lẫn được tính toán để xác định độ chính xác của mô hình.

3. NGHIÊN CỨU THỰC NGHIỆM ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRONG DỰ BÁO KẾT QUẢ HỌC TẬP CỦA SINH VIÊN ĐẠI HỌC

3.1. Phân tích dữ liệu

Để có được dữ liệu về sinh viên, nghiên cứu đã thực hiện khảo sát đối với 150 sinh viên của trường Đại học Thương Mại thông qua khảo sát thực tế và online. Nhóm nghiên cứu thực hiện khảo sát với sinh viên năm thứ nhất. Lý do chọn các em sinh viên năm thứ nhất là vì các em thay đổi hẳn môi trường sinh hoạt và học tập cũng như các mối quan hệ xã hội, từ các học sinh cấp 3 chuyển sang môi trường giáo dục chuyên nghiệp bậc cao và sống cuộc sống tự lập. Vì vậy các nhân tố ảnh hưởng kể trên có thể sẽ có tác động nhiều nhất đến các sinh viên.

Phương pháp thu thập dữ liệu được tiến hành như sau: Nhóm đề xuất phiếu khảo sát, gửi phiếu khảo sát cho một số em sinh viên của năm thứ nhất khoa Hệ thống thông tin kinh tế và Thương mại điện tử của trường Đại học Thương Mại làm thử và thăm dò các ý kiến các em về câu hỏi. Chỉnh sửa câu hỏi về nội dung, từ ngữ,… cho phù hợp với đối tượng sinh viên và nghiên cứu sử dụng phương pháp lấy mẫu thuận tiện để thực hiện khảo sát với các sinh viên của Trường Đại học Thương Mại.

Các thông tin thu thập bao gồm các thông tin về nhân khẩu học như họ tên, giới tính và các

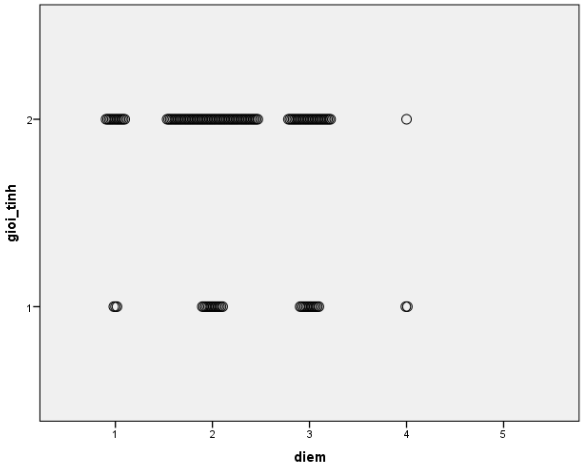
thông tin hành vi của sinh viên trong quá trình học như nơi ở, hoạt động lên thư viện (thời điểm, tần suất…), hoạt động làm thêm, tham gia CLB, học thêm, các mối quan hệ,… các quan điểm về học tập (phương pháp học, thời gian học). Sau khi thu thập được dữ liệu thô ban đầu, nghiên cứu đã tiền xử lý dữ liệu và lọc ra được 134 sinh viên đạt đủ các yêu cầu để tiến hành xử lý các bước tiếp theo. Các bản ghi bị loại bỏ do ghi sai thông tin, thiếu thông tin,...

Sau khi dữ liệu thu thập được sẽ được xử lý thông qua việc biến đổi về dạng số. Ví dụ: ở cột giới tính, sẽ được mã hóa thành 0 và 1. Cột nơi ở sẽ được mã hóa thành 3 giá trị (1: ký túc xá; 2: Nhà trọ; 3: Nhà riêng), Mối quan hệ tình yêu (0: không; 1: có),… Thống kê mô tả về dữ liệu thu thập được miêu tả ở Bảng 1 dưới đây. Trong đó người tham gia nghiên cứu chủ yếu là nữ (mode=2), nơi ở là nhà trọ (mode=2),…

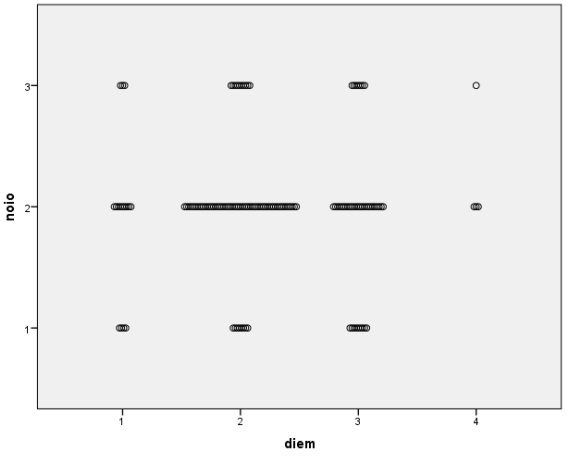
Bảng 1. Thống kê mô tả các biến của tập dữ liệu

	N	Minimum	Maximum	Mode	Std. Deviation
gioi_tinh	134	1	2	2	.432
noio	134	1	3	2	.562
mqhe	134	1	2	2	.494
pponthi	134	1	3	2	.563
thuvien	134	1	4	1	.946
tuhoc	134	1	3	2	.590
hocthem	134	1	4	4	1.273
dilamthem	134	1	2	2	.474
ngoaikhoa	134	1	2	1	.487
diem	134	1	5	2	.698

Theo Hình 1, phân bố của giới tính là nữ (giá trị bằng 2) có mức điểm chủ yếu phân bố vào điểm 2 và 3 (điểm trung bình chung tích lũy đạt điểm B, C theo thang điểm 4)

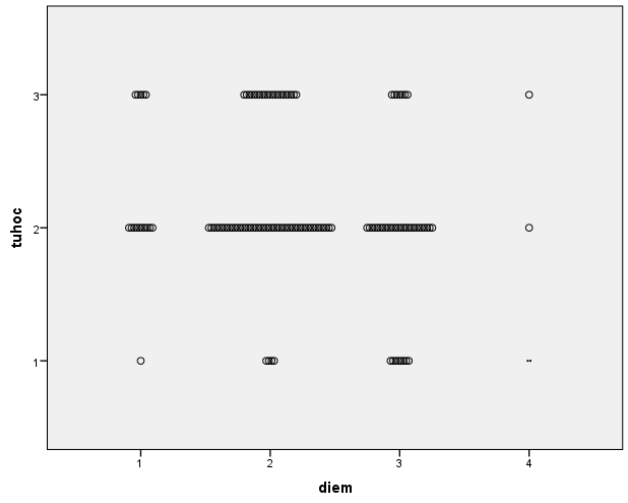


Hình 1. Mối quan hệ giữa giới tính và điểm số của các sinh viên trong khảo sát



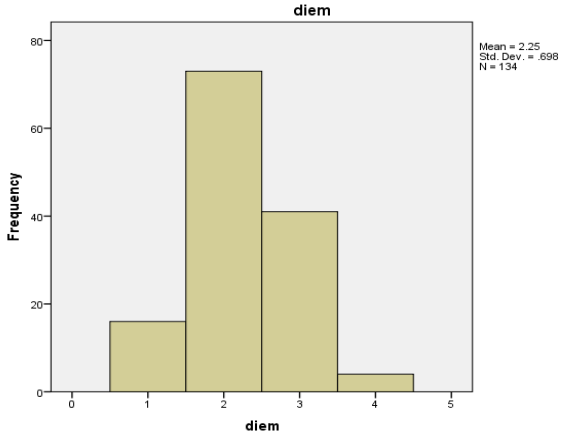
Hình 2. Mối quan hệ giữa nơi ở và điểm số của các sinh viên trong khảo sát.

Nhìn vào Hình 2, rõ ràng sinh viên sống trong các nhà trọ là chiếm đa số và sự tương quan giữa các em sống ở ký túc xá (1); nhà trọ (2) và nhà riêng (3) với điểm số là khá đồng đều và không có sự khác biệt.



Hình 3. Mối quan hệ giữa sự tự học và điểm số của các sinh viên trong khảo sát

Trong các giá trị biểu diễn việc tự học của sinh viên (1: tự học mỗi ngày dưới 1 giờ hay còn gọi là 0 giờ; 2: tự học mỗi ngày dưới 2 giờ; và 3: tự học mỗi ngày trên 2 giờ), ta có thể thấy hầu hết sinh viên được khảo sát đều dành thời gian từ 1-2 giờ tự học trở lên và cũng đạt học lực (điểm) là 2,3,4 nhiều hơn số sinh viên dành ít thời gian tự học (<1 giờ).



Hình 4. Tần suất phân bố điểm của các sinh viên khảo sát

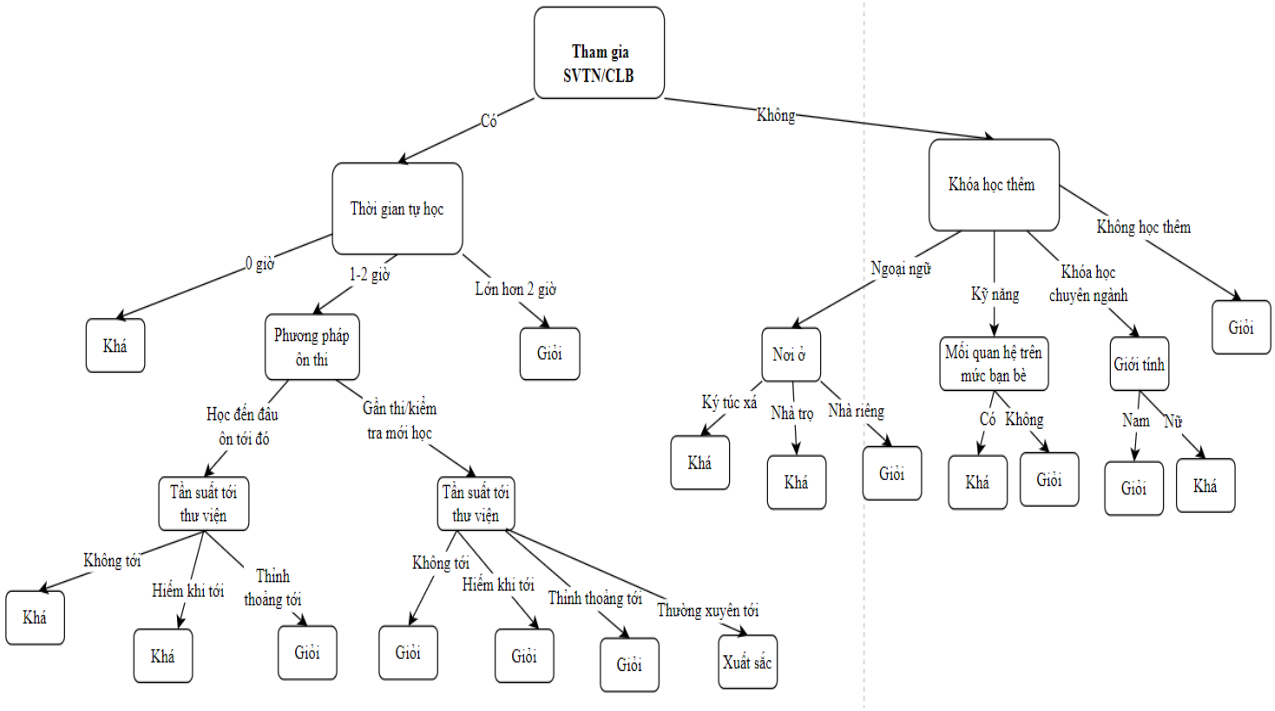
Hình 4 cho thấy trong số 134 sinh viên khảo sát, số lượng các em đạt điểm 2, 3 (điểm trung bình chung tích lũy đạt điểm B, C) cũng chiếm phổ biến. Số lượng các em đạt điểm xuất sắc (giá trị 4) chiếm ít nhất (khoảng 4 em). Vì vậy, nhu cầu hỗ trợ các em để dự báo tình hình học tập trong các kỳ tiếp theo là cấp thiết đối với mỗi sinh viên.

3.2. Kết quả thực nghiệm

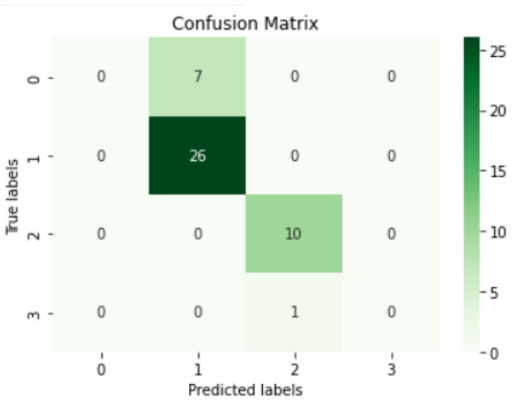
Dữ liệu được thực hiện với thuật toán cây quyết định cho 2/3 dữ liệu ban đầu. Nghĩa là 90/134 trường hợp được sử dụng để xây dựng cây và phần còn lại, 44 trường hợp, được sử dụng để test cây. Chỉ số Gini (Gini Index) được tính toán tại mỗi nút để xác định gốc cây và các nhánh của cây. Các bước của thuật toán xây dựng cây quyết định được miêu tả chi tiết ở mục 2. Sử dụng chương trình trong Python, áp dụng bộ công cụ xây dựng cây quyết định với các tham số như sau: độ sâu tối đa của cây (max_depth=25; số lá tối thiểu của cây là 10 (có 10 thuộc tính).

Kết quả sau khi thực hiện chương trình cây dự báo được xây dựng ở Hình 5.

Kiểm tra độ chính xác của cây như sau: Sử dụng tập test với 44 trường hợp còn lại trong mô hình. Ma trận confusion matrix vẽ với 44 dòng dữ liệu cho thấy các nhãn được phân vào 2 lớp chính là lớp 1 và lớp 2. 7 trường hợp của lớp 0 được phân lớp vào lớp 1, và 1 trường hợp của lớp 3 (điểm =4) được phân vào lớp 2. Rõ ràng về mầu sắc ta nhận thấy 7 trường hợp của lớp 0 mặc dù được phân lớp vào lớp 1 nhưng màu sắc của chúng nhạt hơn rất nhiều so với 26 trường hợp phân lớp đúng vào lớp 1. Điều này có nghĩa là chúng thuộc về lớp 0, nhưng do đầu ra của cây quyết định là bài toán nhị phân (phân về 2 lớp) nên chúng được gán về lớp 1. Tương tự như vậy đối với 01 trường hợp của lớp 3 được gán về lớp 2.



Hình 5. Kết quả dự báo của cây quyết định



Hình 6. Bảng ma trận hỗn loạn của cây quyết định sau khi test.

Kết quả test còn được tính toán độ chính xác (accuracy rate) và các chỉ số dự báo khác như sau: Độ chính xác của cây dự báo là 0.82, còn các chỉ số như Precision, Recall, và F1-Score được xem ở Bảng 2.

Bảng 2: Các chỉ số hỗ trợ các mẫu trong tập Test

Chỉ số/lớp	Precision	Recall	F1-Score	Support (số mẫu)
1	0.00	0.00	0.00	1
2	0.91	1.00	0.95	10
3	0.79	1.00	0.88	26
4	0.00	0.00	0.00	7

Nhìn vào Bảng 2, số mẫu được hỗ trợ nhiều nhất là ở lớp điểm 2 và 3 (10 và 26 mẫu) sẽ có chỉ số cân bằng dự báo F1-Score cao nhất là 95 % và 88%. Như đã phân tích mầu sắc ở ma trận Hình 6, chỉ số cân bằng của lớp dự báo 1 và 4 (hỗ trợ cho các mẫu này là 1 và 7 mẫu) là 0% vì nhược điểm

của bài toán dự báo cây quyết định thu về bài toán nhị phân.

4. KẾT LUẬN

Như đã phân tích ở trên, có nhiều yếu tố có thể ảnh hưởng đến kết quả học tập của sinh viên. Về cơ bản các yếu tố này có thể được phân loại thành ba nhóm chính: yếu tố của bản thân cá nhân, yếu tố đến từ gia đình và yếu tố môi trường sống. Các yếu tố cá nhân ở đây thường bao gồm năng lực bẩm sinh, động lực và thái độ của học sinh đối với học tập. Trong bài báo này, yếu tố cá nhân được thể hiện thông qua các câu hỏi liên quan như: Phương pháp ôn thi, tần suất đến thư viện, thời gian tự học, tham gia các khóa học thêm,… Như phân tích ở Hình 3, sinh viên dành thời gian tự học nhiều hơn thì có kết quả tốt hơn.

Yếu tố gia đình được quan tâm trong nghiên cứu này chính là yếu tố tài chính. Sinh viên phải đi làm thêm từ năm thứ nhất tương đương với việc các em không có nhiều sự trợ giúp từ gia đình về tài chính. Điều này dẫn đến có sự ảnh hưởng đến việc học tập của các em.

Các yếu tố môi trường bao gồm chất lượng giảng dạy của trường, môi trường sống,…có khả năng thành công hơn trong học tập. Chính vì vậy, trong nghiên cứu này, yếu tố nơi ở được đặt ra trong câu hỏi đối với sinh viên năm thứ nhất. Ngoài ra, đối với sinh viên năm thứ nhất việc thay đổi môi trường sống (chủ yếu) từ gia đình lên một môi trường sống khác như ở trọ hay ở ký túc xá là một bước ngoặt khá lớn. Và điều này cũng ảnh hưởng đến việc học của sinh viên. Tuy nhiên theo kết quả khảo sát ở Hình 2, sinh viên chủ yếu ở trọ và các kết quả học tập không bị ảnh hưởng bởi môi trường sống (nơi ở). Yếu tố tâm lý của sinh viên cũng là một trong những nhân tố được đề cập đến ở nghiên cứu. Quan hệ khác giới đối với sinh viên năm thứ nhất cũng được đưa vào bảng hỏi để đánh giá mức độ ảnh hưởng của chúng đối với kết quả học tập của sinh viên. Ngoài ra, các yếu tố khác như sức khỏe, sự phát triển tâm lý,… không được đề cập đến nghiên cứu này.

Cây quyết định được xây dựng theo cấu trúc phân cấp xuất phát từ nút gốc và các nút nhánh liên kết với nhau. Phương pháp Gini được sử dụng để đo mức độ phản ánh thông tin của các biến cần xem xét đối với biến quyết định mà cụ thể là điểm số của các sinh viên. Để xác định nút gốc và sau đó là các nút tiếp theo, đối với mỗi thuộc tính trong tập dữ liệu xem xét, các câu hỏi (luật) được xây dựng và tính toán dựa trên ngưỡng để phân loại đối tượng cần xem xét. Các nhánh được liên kết với nút cha mẹ dựa trên các giá trị phân loại của nút thuộc tính. Các nút lá cuối cùng phản ánh thông tin quyết định sau khi thỏa mãn một loạt các luật của cây trước đó. Như vậy, để dự báo điểm số tương lai cho một sinh viên với một loạt các câu hỏi đưa ra cho sinh viên đó lựa chọn bao gồm nhóm câu hỏi về phương pháp học tập của cá nhân (tự học, tần suất lên thư viện,…), về hỗ trợ tài chính của gia đình, và môi trường sống,… Mô hình sẽ căn cứ vào bộ dữ liệu đã có để đưa ra quyết định cho sinh viên đó với điểm số tương ứng là bao nhiêu.

Nghiên cứu lựa chọn cây quyết định làm công cụ trong việc dự báo kết quả học tập của sinh viên vì ưu điểm của công cụ trong khai phá dữ liệu là tính "dễ giải thích và suy diễn" của mô hình cây quyết định.

Mục tiêu của bài báo này là mong muốn xây dựng một công cụ hỗ trợ, giúp cho sinh viên nhận thức được các yếu tố cơ bản của bản thân, môi trường,… sẽ tác động đến kết quả học tập của bản thân. Khi xác định được những yếu tố ảnh hưởng đến công việc học tập của sinh viên, các yếu tố đó sẽ được đưa vào mô hình với sự trợ giúp của máy tính (cây quyết định), họ sẽ biết được kết quả dự báo cho những kỳ tiếp theo. Từ đó, các em có thể thay đổi lại chiến lược học tập và biết những yếu tố nào mình cần thay đổi và nên thay đổi để đạt được kết quả học tập tốt hơn. Đồng thời nghiên cứu cũng hỗ trợ cho các cố vấn học tập, người phụ trách các em sinh viên có thể giải thích kết quả học tập của sinh viên dựa vào các yếu tố ảnh hưởng, đồng thời cảnh báo cho sinh viên trong những năm kế tiếp.

Liệu các yếu tố ảnh hưởng đến kết quả học tập có ổn định qua các năm trong cùng một trường đại học không? Đối với các nhóm sinh viên khác nhau cũng có thể có ảnh hưởng đến kết quả học tập khác nhau. Việc mở rộng nghiên cứu đối với các nhóm khác nhau và trải rộng nghiên cứu ở các trường Đại học khác nhau với số lượng mẫu nhiều lên cũng có thể củng cố thêm về kỹ thuật lựa chọn. Hơn nữa cần phải có việc kết hợp các phương pháp dự đoán khác nhau hoặc sử dụng các phương pháp mới như học sâu (deep learning) để có thể có các kết quả tốt hơn chẳng hạn. Tuy nhiên, các vấn đề này cần có các nghiên cứu sâu hơn ở tương lai.

TÀI LIỆU THAM KHẢO

Amershi, S., and Conati, C. (2006). Automatic Recognition of Learner Groups in Exploratory Learning Environments. Proceedings of ITS 2006. 8th International Conference on Intelligent Tutoring Systems.

Baker, R.S.J.D. (2010). Mining Data for Student Models. In: Advances in Intelligent Tutoring Systems. Studies in Computational Intelligence. Vol 3.08. Springer.

Baker, R.S.J.D., and Yacef, K. (2009). The state of educational data mining: A review and future vision. Journal of Educational Data Mining. Vol. 1. No. 1. Pp: 3-17.

Beal, C., Qu, L., and Lee, H. (2007). Classifying learner engagement through integration of multiple data sources. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2007).

Bharadwaj, B.K., and Pal, S. (2011). Mining Educational Data to Analyze Student's Performance. International Journal of Advance Computer Science and Applications (IJACSA). Vol. 2. No. 6. Pp: 63-69.

Dekker, G., Pechenizkiy M., and Vleeshouwers, J. (2009). Predicting students drop out: A case study. Proceedings of the 2nd International Conference on Educational Data Mining. Pp:41-50.

Lee, S. J. and Siau. K. (2001). A review of data mining techniques. Industrial Management & Data Systems. Vol. 101/1. Pp: 41-46

Lin, S. H. (2012). Data Mining for student retention management. ACM journal of ComputingSciences in Colleges. Vol.27. No.4. Pp: 92-99.

Romero, C., and Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications. Vol. 33. No. 1,. Pp: 135-146.

Witten, I.H., and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann publisher.