MÔ HÌNH TỐI ƯU CHO BÀI TOÁN DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN TRƯỜNG ĐẠI HỌC ĐỒNG THÁP

• Lê Quang Minh(*)

Tóm tắt

Mục tiêu của nghiên cứu này là vận dụng phương pháp hồi quy Naïve Bayes, cây quyết định và mạng nơ-ron để xây dựng, đánh giá và tìm ra mô hình tối ưu trên tập dữ liệu thực tế tại Trường Đại học Đồng Tháp. Bài báo giới thiệu phương pháp hồi quy Naïve Bayes là mô hình tối ưu cho bài toán dự đoán kết quả học tập của sinh viên Trường Đại học Đồng Tháp. Từ đó, giúp cho sinh viên xác định mục tiêu và lập kế hoạch học tập phù hợp cho cả khóa học, cho từng học kỳ để mang lại kết quả học tập như mong muốn.

Từ khóa: Phương pháp phân lớp, Naïve Bayes, cây quyết định, mạng nơ-ron.

1. Đặt vấn đề

Bài toán dự đoán kết quả học tập của sinh viên đã và đang thu hút được nhiều sự quan tâm nghiên cứu của các nhà khoa học. Có một số mô hình đã được các nhà nghiên cứu đề xuất với mục tiêu cuối cùng là nâng cao độ chính xác của kết quả dự đoán. Bài toán dự đoán kết quả học tập của sinh viên chủ yếu được tiếp cận dưới ba dạng sau: dự đoán kết quả cho sinh viên lựa chọn các môn học cho học kỳ sau dựa trên kết quả của các học kỳ trước, dự đoán kết quả cuối khóa của sinh viên dựa trên điểm của các học kỳ trước đó, bài toán dự đoán kết quả học tập cuối khóa của sinh viên dựa trên lộ trình học.

Một trong những hướng tiếp cận phổ biến hiện nay để giải quyết bài toán dự đoán kết quả học tập cho sinh viên là xây dựng các mô hình đánh giá dựa trên tập dữ liệu đầu vào, phân tích đánh giá các mô hình để lựa chọn mô hình tối ưu nhất cho bài toán đặt ra. Một số thuật toán khai phá dữ liệu hiệu quả được nhóm tác giả J. Ross Quinlan, X. Wu, V. Kumar nghiên cứu và công bố trong [6]. Theo hướng tiếp cận này nhiều tác giả đã nghiên cứu đề xuất và ứng dụng phương pháp hồi quy Naïve Bayes cho các bài toán phân lớp [8].

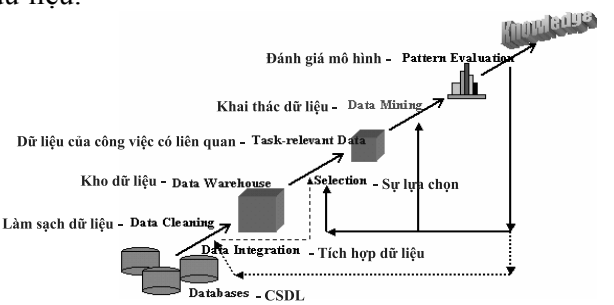
Trong bài báo này, tác giả dẫn xuất từ kết quả thực nghiệm trên cơ sở xây dựng mô hình bằng ba thuật toán là Naïve Bayes, cây quyết định và mạng nơ-ron, đánh giá các mô hình khai phá dữ liệu đã xây dựng từ đó lựa chọn mô hình tối ưu nhất để áp dụng cho bài toán dự đoán kết quả học tập của sinh viên Trường Đại học Đồng Tháp.

(*) Trường Đại học Đồng Tháp.

2. Nội dung nghiên cứu

2.1. Tổng quan về khai phá dữ liệu

Khám phá tri thức là quá trình tìm ra những tri thức, đó là những mẫu tìm ẩn, trước đó chưa biết và là thông tin hữu ích đáng tin cậy. "Khai phá dữ liệu là một quá trình khám phá, chắt lọc các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có của một công ty, đơn vị, tổ chức nào đó, từ đó giúp cho chúng ta có được quyết định sáng suốt" [1]. Nói một cách khác, mục đích của khám phá tri thức và khai phá dữ liệu chính là tìm ra các mẫu hoặc mô hình đang tồn tại trong các cơ sở dữ liệu (CSDL) nhưng vẫn còn bị che khuất bởi hàng núi dữ liệu.



Hình 1. Các bước của quá trình khai phá dữ liệu

2.2. Một số thuật toán về khai phá dữ liệu

2.2.1. Hồi quy Naïve Bayes

a. Định lý Bayes

- Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là $P(A \mid B)$, và đọc là "xác suất của A với điều kiện B xảy ra". Đại lượng này được gọi là xác suất có điều kiện vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

- Theo định lý Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

+ $P(A)$: Xác suất xảy ra A của riêng nó, không quan tâm đến bất kỳ thông tin nào về B.

+ $P(B)$: Xác suất xảy ra B của riêng nó, không quan tâm đến A. Đại lượng này còn gọi là hằng số chuẩn hóa (normalising constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.

+ $P(A \mid B)$: Xác suất xảy ra B khi biết A xảy ra và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra.

+ Khi biết ba đại lượng trên, xác suất của A khi biết B cho bởi công thức:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \quad (1)$$

Từ đó dẫn đến:

$$P(A \mid B)P(B) = P(A \cap B) = P(B \mid A)P(A).$$

+ Khi có n giả thuyết thì

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}.$$

b. Mô hình phân lớp Naïve Bayes (NBC)

Cho $\{C_1, C_2 ... C_n\}$ là phân hoạch của không gian mẫu C (được xem là các lớp C_i). Không gian thể hiện X bao gồm tất cả các thể hiện được mô tả trên tập thuộc tính $(a_1, a_2 ... a_n)$ hàm đích $f(x)$ có thể nhận bất kỳ giá trị nào trong $C\ (f(x) = C_i \mid i = 1...n)$. Không gian thể hiện X được xem là các ví dụ học. Khi có một thể hiện mới với bộ giá trị $< a_1, a_2 ... a_n >$ bộ phân lớp sẽ dự đoán giá trị hàm đích $f(x)$ hoặc lớp cho thể hiện mới này $f(x) \in \{C_1, C_2 ... C_n\}$.

Cách tiếp cận Bayes để phân lớp một thể hiện mới là lấy giá trị đích có xác suất cao nhất $C_{\max}$ của thể hiện này. Hay nói cách khác, định lý Bayes được sử dụng để chọn giả thuyết có xác suất cao nhất từ tập mẫu huấn luyện, giả thuyết này được gọi là giả thuyết cực đại xác suất hậu nghiệm MAP-Maximum A Posterior:

$$C_{MAP} = \max_{C_i \in C} P(C_i \mid a_1, a_2 ... a_n). \quad (2)$$

Sử dụng định lý Bayes - Công thức (1) - Áp dụng vào (2), ta có:

$$C_{MAP} = \max_{C_i \in C} \frac{P(C_i)P(a_1, a_2 ... a_n \mid C_i)}{P(a_1, a_2 ... a_n)} = \max_{C_i \in C} P(C_i)P(a_1, a_2 ... a_n \mid C_i). \quad (3)$$

(Cùng mẫu số $P(a_1, a_2 ... a_n)$ nên ta bỏ qua so sánh mẫu)

Các $P(C_i)$ được tính bằng cách đếm số lần có mặt của giá trị đích C_i trong tập dữ liệu học. Tuy nhiên để tính $P(a_1, a_2 ... a_n)$ bộ phân lớp Naïve Bayes dựa trên việc đơn giản hóa các giả định ban đầu là các giá trị thuộc tính độc lập điều kiện với giá trị đích cho trước.

Nói cách khác, xác suất của một thể hiện quan sát được $< a_1, a_2 ... a_n >$ trên mỗi lớp C_i sẽ là tích của các khả năng của từng thuộc tính riêng biệt trên C_i

$$P(a_1, a_2 ... a_n \mid C_i) \prod_{i=1}^{n} P(a_i \mid C_i).$$

Công thức (3) được viết lại:

$$C_{NB} = \max_{C_i \in C} P(C_i) \prod_{i=1}^{n} P(a_i \mid C_i). \quad (4)$$

Bộ phân lớp Naïve Bayes liên quan đến một bước học mà trong đó $P(C_i)$ và $P(a_1, a_2 ... a_n)$ được ước đoán dựa trên tần số xuất hiện của chúng trên toàn bộ tập dữ liệu học. Tập dự đoán này tương ứng với kết luận học được, kết quả của bộ phân lớp trong công thức (4) được sử dụng để phân lớp thể hiện mới này.

c. Các bước thực hiện thuật toán Naïve Bayes

+ Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính $P(C_i)$ và $P(a_i \mid C_i)$

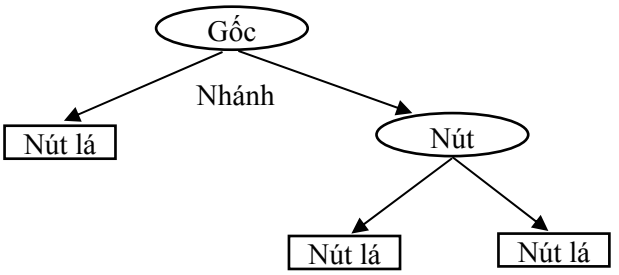
+ Bước 2: Phân lớp $A^{new} = (a_1, a_2 ... a_n)$, ta cần tính xác suất thuộc từng phân lớp khi đã biết trước A^{new}. A^{new} được gán vào lớp có xác suất lớn nhất theo công thức:

$$\max_{C_i \in C} P(C_i) \prod_{i=1}^{n} P(a_i \mid C_i).$$

2.2.2. Cây quyết định

Cây quyết định là một cấu trúc biểu diễn dưới dạng cây. Trong đó, mỗi nút trong (Internal node) biểu diễn một thuộc tính, mỗi nhánh

(branch) biểu diễn giá trị có thể có của thuộc tính, mỗi nút lá (Leaf node) biểu diễn các lớp quyết định và đỉnh trên cùng của cây gọi là nút gốc (Root).



Hình 2. Biểu diễn cây quyết định cơ bản

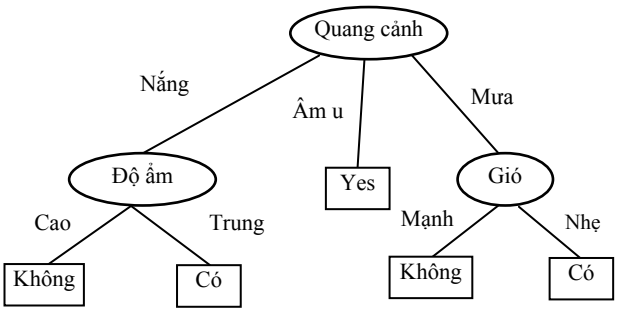
Việc phân lớp một phần tử mới bắt đầu từ nút gốc của cây. Nếu nút này có nút con thì quá trình sẽ được lặp lại sử dụng cây con thích hợp. Khi đến nút lá, nhãn của nút lá sẽ cho ta lớp dự đoán của phần tử này.

Một số điều kiện cần thiết khi làm việc với cây quyết định:

- Mô tả giá trị thuộc tính: các phần tử phải được thể hiện dưới dạng một tập hợp cố định các thuộc tính.
- Các lớp gán cho các phần tử phải được định nghĩa trước.
- Các lớp phải rời rạc: một phần tử phải thuộc hoặc không thuộc về một lớp và số phần tử phải nhiều hơn số lớp.

Bảng 1. Ví dụ dữ liệu minh họa cây quyết định

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
Nắng	Nóng	Cao	Nhẹ	Không
Nắng	Nóng	Cao	Mạnh	Không
Âm u	Nóng	Cao	Nhẹ	Có
Mưa	Ấm áp	Cao	Nhẹ	Có
Mưa	Mát	Trung bình	Nhẹ	Có
Mưa	Mát	Trung bình	Mạnh	Không
Âm u	Mát	Trung bình	Mạnh	Có
Nắng	Ấm áp	Cao	Nhẹ	Không
Nắng	Mát	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Trung bình	Nhẹ	Có
Nắng	Ấm áp	Trung bình	Mạnh	Có
Âm u	Ấm áp	Cao	Mạnh	Có
Âm u	Nóng	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Cao	Mạnh	Không



Hình 3. Mô hình cây quyết định minh họa

Việc xây dựng cây quyết định được tiến hành một cách đệ quy, lần lượt từ nút gốc xuống tới tận các nút lá. Tại mỗi nút hiện hành đang xét, nếu kiểm tra thấy thoả điều kiện dừng: thuật toán sẽ tạo nút lá. Nút này được gán một giá trị của nhãn lớp tùy điều kiện dừng được thoả. Ngược lại, thuật toán tiến hành chọn điểm chia tốt nhất theo một tiêu chí cho trước, phân chia dữ liệu hiện hành theo điều kiện chia này. Lưu ý dữ liệu hiện hành không phải hoàn toàn là tập dữ liệu ngay khi bắt đầu thuật toán, có thể là tập dữ liệu đã được phân chia theo điều kiện chia của nút liền trước đó (nút cha).

2.2.3. Mạng nơ-ron

Mạng nơ-ron là một trong những kỹ thuật khai phá dữ liệu được ứng dụng phổ biến hiện nay. Kỹ thuật này phát triển dựa trên một nền tảng toán học vững vàng, khả năng huấn luyện trong kỹ thuật này dựa trên mô hình thần kinh trung ương của con người.

Mạng nơ-ron nhân tạo là hệ thống máy tính song song bao gồm nhiều nơ-ron đơn giản, kết nối với nhau theo một cấu trúc nào đó để thực hiện một nhiệm vụ cụ thể. Mặc dù được đơn giản hóa về mặt mô phỏng nhưng mạng nơ-ron nhân tạo vẫn không mất đi các tính chất đặc trưng của bộ não thật [4].

Ưu điểm lớn nhất của mạng nơ-ron nhân tạo là tính hoàn toàn song song, thêm vào đó mạng nơ-ron nhân tạo có thể học từ các dữ liệu huấn luyện và khái quát hóa tình huống mới nên nó không yêu cầu nhiều về kỹ thuật lập trình [4], có khả năng chịu lỗi tốt.

Khả năng học của mạng nơ-ron nhân tạo là học từ một tập dữ liệu huấn luyện và khái quát hóa những tình huống mới. Phương pháp hoạt động của chúng là thay đổi trọng số các kết nối

giữa các nơ-ron sao cho kết quả ra cuối cùng là chính xác. Có ba loại học tập được xem xét trong mạng nơ-ron nhân tạo:

- Học có giám sát;
- Học dựa trên thông tin phản hồi;
- Học không giám sát.

Mạng nơ-ron nhân tạo được ứng dụng nhiều như: mô phỏng hành vi giúp chúng ta hiểu được cách thức hoạt động của bộ não, các mô hình giải phẫu não, dự đoán chứng khoán, cổ phiếu, dự đoán khí hậu, thời tiết, tiếp thị hàng không, nhận dạng giọng nói, nhận dạng chữ viết tay, khai phá dữ liệu, cấu trúc protein thứ cấp.

2.3. Xây dựng, phân tích và đánh giá mô hình dự đoán kết quả học tập

2.3.1. Thu thập dữ liệu

a. Về điểm trung bình chung tích lũy

Bảng 2. Điểm trung bình chung tích lũy

Trường	Dữ liệu ban đầu	Dữ liệu sau khi xử lý
ĐTBC	Từ 3,6 đến 4	Xuất sắc
ĐTBC	Từ 3,2 đến 3,59	Giỏi
ĐTBC	Từ 2,5 đến 3,19	Khá
ĐTBC	Từ 2 đến 2,49	Trung bình

Ghi chú: ĐTBC: Điểm trung bình chung.

b. Về thông tin sinh viên: Thu được 3.402 bản ghi của các sinh viên đã tốt nghiệp tại Trường Đại học Đồng Tháp, gồm 11 thuộc tính sau:

Bảng 3. Thuộc tính sinh viên

STT	Thuộc tính	Mô tả	Giá trị
1	MaSV	Mã sinh viên (thuộc tính khóa)	Text
2	Ketqua	Kết quả cuối khóa (thuộc tính dự đoán)	XS: Xuất sắc; G: Giỏi; K: Khá TB: Trung bình
3	Gioitinh	Giới tính	0: Nam; 1: Nữ
4	Diemthits	Điểm thi tuyển sinh	Real
5	Manganh	Mã ngành dự thi	Integer (rời rạc, theo mã)
6	Malotrinh	Mã lộ trình học của sinh viên	Integer (rời rạc, theo mã)
7	MaTP	Mã tỉnh, thành phố của sinh viên	Text
8	MaQuanHuyen	Mã quận, huyện của sinh viên	Text
9	Khoithi	Khối sinh viên dự thi đầu vào	A: Khối A; D1: Khối D1
10	Doituong	Sinh viên thuộc đối tượng dự thi nào	01: đối tượng 01; 02: đối tượng 02; 03: đối tượng 03; 04: đối tượng 04; 05: đối tượng 05; 06: đối tượng 06
11	Khuvuc	Sinh viên thuộc khu vực dự thi nào	1: khu vực 1; 2: khu vực 2 2NT: khu vực 2 nông thôn; 3: khu vực 3

c. Về lộ trình học: Thu được 233.510 bản ghi về điểm tổng kết các học phần và dữ liệu cá nhân, tuyển sinh và điểm tổng kết cuối khóa cho 3.402 sinh viên thuộc 21 ngành học với 840 học phần của các ngành học đối với hệ đào tạo chính quy bậc đại học tại Trường Đại học Đồng Tháp.

2.3.2. Xây dựng mô hình

Để xây dựng được một hệ thống dự đoán kết quả học tập với độ chính xác cao và ổn định, tiến hành triển khai các mô hình dự báo kết quả học tập dựa trên ba thuật toán như: Naïve Bayes, mạng nơ-ron, cây quyết định. Các mô hình được xây dựng dựa trên công cụ BIDS của Microsoft. CSDL để xây dựng các mô hình gồm 3.402 bản ghi và chia theo tỷ lệ 85% cho training và 15% cho testing bằng cách lựa chọn ngẫu nhiên 2.892 bản ghi cho việc training và 510 bản ghi cho việc testing.

Sau khi xây dựng mô hình, tiến hành kiểm tra sự phụ thuộc của kết quả thuộc tính dự đoán vào các thuộc tính khác thông qua tab Dependency Network. Từ kết quả kiểm tra cho thấy thuộc tính dự đoán chỉ phụ thuộc vào 5 thuộc tính sau: Giới tính, ngành học, lộ trình học, khối thi và điểm thi tuyển sinh. Đầu vào của hệ thống dự đoán chỉ yêu cầu sinh viên nhập vào thông tin các thuộc tính như: giới tính, ngành học, khối thi, điểm thi tuyển sinh các thuộc tính khác như: đối tượng dự thi, tỉnh/thành phố, quận/huyện, khu vực sẽ không được xét đến trong mô hình.

Hiệu quả của các mô hình DMM đã xây dựng sẽ được đánh giá thông qua hai phương pháp: Lift Chart và Classification Matrix. Mục tiêu của nghiên cứu là xác định xem mô hình nào cho phần trăm dự đoán chính xác cao nhất trong việc dự đoán kết quả học tập của sinh viên.

2.3.3. Đánh giá mô hình

a. Với Lift Chart có xác định giá trị thuộc tính dự đoán

Trong ba mô hình đã xây dựng thì mô hình Naïve Bayes cho kết quả tốt nhất, sau đó đến mạng nơron, cây quyết định, kết quả cụ thể được thể hiện ở Bảng 4.

Bảng 4. Kết quả đánh giá với Lift Chart

STT	Series, Model	Score
1	Naïve Bayes	0,93
2	Mạng nơron	0,92
3	Cây quyết định	0,90

b. Với Classification Matrix

Kết quả dự báo cụ thể của từng mô hình được thể hiện ở Bảng 5.

Bảng 5. Kết quả đánh giá với Classification Matrix

Mô hình	Kết quả dự đoán	K (Khá)	TB (Trung bình)	G (Giỏi)	XS (Xuất sắc)
Naïve Bayes	K	***278***	29	9	0
	TB	48	***116***	0	0
	G	9	1	***19***	1
	XS	0	0	0	***0***
Mạng nơ-ron	K	***292***	47	16	0
	TB	29	***98***	0	0
	G	14	1	***12***	*1*
	XS	0	0	0	**0**
Cây quyết định	K	***317***	78	21	0
	TB	9	***77***	0	0
	G	0	0	***7***	1
	XS	0	0	0	***0***

Từ kết quả Bảng 5 cho ta thấy tỷ lệ chính xác trung bình các mô hình lần lượt là 80,98% (Naïve Bayes), 78,82% (mạng nơ-ron) và cuối cùng là 78,62% (cây quyết định). Tất cả các mô hình đều không dự báo chính xác cho sinh viên đạt loại xuất sắc (độ chính xác là 0%), vì bộ dữ liệu phục vụ xây dựng mô hình cũng chỉ có 02 sinh viên đạt xếp loại tổng kết xuất sắc, số lượng quá ít như vậy sẽ không có ý nghĩa trong việc khai phá dữ liệu.

3. Kết luận

Nghiên cứu này đã tìm hiểu cơ sở lý thuyết về khai phá dữ liệu, tập trung vào các kỹ thuật khai phá dữ liệu như: Naïve Baye, cây quyết định, mạng nơ-ron. Xây dựng thành công ba mô hình dự đoán kết quả học tập trên các thuật toán đã đề xuất bằng công cụ BIDS của Microsoft. Với kết quả nghiên cứu này tác giả đề xuất mô hình tối ưu để dự đoán kết quả học tập của sinh viên Trường Đại học Đồng Tháp là sử dụng thuật toán Naïve Bayes./.

Tài liệu tham khảo

[1]. Jiawei Han and Micheline Kamber (2006), *Data Mining Concepts and Techniques*, Second Edition, Published by Elsevier Inc.

[2]. Brian Knight, Devin Knight, Adam Jorgensen, Patrick LeBlanc, Mike Davis (2010), *Knight's Microsoft Business Intelligence 24-Hour Trainer*, Published by Wiley Publishing, Inc.

[3]. Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat (2008), *Data Mining with Microsoft SQL Server 2008*, Published by Wiley Publishing, Inc., Indianapolis, Indiana.

[4]. Võ Viết Minh Nhật (2013), *Mạng Nơ-ron nhân tạo và ứng dụng*, NXB Giáo dục Việt Nam.

[5]. Lê Văn Phùng và Quách Xuân Trưởng (2012), *Khai phá dữ liệu*, NXB Thông tin và Truyền thông.

[6]. J.Ross Quinlan, X. Wu, V. Kumar (2009), *Top 10 Algorithms in Data Mining*, Chapman & Hall/ CRC, ©Taylor & Francis Group, LLC.

[7]. Hà Quang Thụy (2009), *Giáo trình khai phá dữ liệu Web*, NXB Giáo dục Việt Nam.

[8]. Nguyễn Thị Thanh Thủy (2012), "Ứng dụng khai phá dữ liệu xây dựng công cụ dự đoán kết quả học tập của sinh viên", *Kỷ yếu Hội nghị Sinh viên Nghiên cứu Khoa học lần thứ 8 Đại học Đà Nẵng,* tr. 15-21.

OPTIMAL MODEL FOR PREDICTING STUDENTS' LEARNING RESULTS AT DONG THAP UNIVERSITY

Summary

This article applies the Naive Bayesian classifier, decision tree and neural nets, to set up, evaluate and come up with an optimal model, based on database at Dong Thap University. It recommends that the Naïve Bayesian is the optimal model for predicting students' learning results at Dong Thap University. Thereby, it helps students to set learning objectives and make plans for their entire traning programs and each semester, as such to obtain results as expected.

Keywords: Classification methods, Naive Bayesian, decision tree, neural nets.

Ngày nhận bài: 28/3/2017; Ngày nhận lại: 13/5/2017; Ngày duyệt đăng: 03/7/2017.