# DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN BẰNG KỸ THUẬT KHAI PHÁ DỮ LIỆU

**Nguyễn Thị Uyên, Nguyễn Minh Tâm**
*Trường Đại học Vinh*
Ngày nhận bài 22/5/2019, ngày nhận đăng 12/9/2019

**Tóm tắt:** Hiện nay, tình trạng sinh viên bị buộc ngừng học đang diễn ra rất phổ biến tại các trường đại học ở Việt Nam. Bài báo này đề xuất phương pháp cho phép dự đoán được khả năng bị buộc ngừng học dựa vào phân tích dữ liệu từ điểm thi đầu vào, điểm thi các môn của ba học kỳ đầu và tình trạng hiện thời (tiếp tục học hoặc ngừng học) của hơn 555 sinh viên khóa 54, 55, 56 ngành Công nghệ thông tin, Trường Đại học Vinh. Từ dữ liệu đã có, hai thuật toán khai phá dữ liệu Logistic Regression, Naive Bayes đã được áp dụng để tìm ra mô hình tốt nhất cho việc dự báo tình trạng học tập cho sinh viên các khóa tiếp theo. Việc nghiên cứu này sẽ giúp cho Nhà trường đưa ra được những cảnh báo sớm và có phương án hỗ trợ để giảm tỷ lệ bị buộc thôi học cho các sinh viên khóa sau.

**Từ khóa:** Khai phá dữ liệu giáo dục; cảnh báo ngừng học.

## 1. Giới thiệu

Trong những năm qua, công tác tuyển sinh ngày càng khó khăn, nhưng số lượng sinh viên bị buộc thôi học, cảnh báo thôi học lại ngày càng có xu hướng gia tăng. Theo thống kê chưa chính thức tại Trường Đại học Vinh, mỗi năm có tới hàng trăm sinh viên rơi vào tình trạng bị buộc thôi học, chủ yếu tập trung vào các sinh viên học năm thứ 3, hoặc năm thứ 4, khi các em đã gần tốt nghiệp. Vì vậy, việc phát hiện sớm các sinh viên có khả năng bị buộc ngừng học nhằm giúp họ lập kế hoạch học tập sao cho phù hợp là một nhu cầu rất cần thiết của nhà trường hiện nay.

Khai phá dữ liệu giáo dục là một lĩnh vực nghiên cứu đã và đang được nhiều nhà khoa học quan tâm. Các thuật toán khai phá dữ liệu như Logistic Regression, Naive Bayes đã được áp dụng nhiều trong các bài toán thực tế như dự báo chứng khoán, dự báo dữ liệu y tế, phân tích dữ liệu giáo dục [1] - [4]. Các thực nghiệm cho thấy việc xây dựng các mô hình dự đoán hay phân lớp bằng các thuật toán này cho kết quả khá tốt, hỗ trợ được cho việc ra các quyết định tiếp theo.

Trong bài báo này, chúng tôi thu thập dữ liệu về điểm thi đầu vào đại học, điểm thi các môn của ba học kỳ đầu và tình trạng cảnh báo (đang học hoặc ngừng học) của sinh viên khóa 54, 55, 56 ngành Công nghệ thông tin làm dữ liệu huấn luyện để xây dựng mô hình dự đoán. Sau khi xây dựng được mô hình, dựa vào dữ liệu đầu vào bao gồm điểm thi đầu vào và điểm thi các môn của ba học kỳ đầu ta có thể dự đoán được sinh viên nào đó trong tương lai có thể bị buộc ngừng học.

Trên cơ sở trình bày nhận thức chung về khai phá dữ liệu trong giáo dục cùng các công trình nghiên cứu ứng dụng kĩ thuật này, bài viết tập trung mô tả quá trình xây dựng mô hình dự toán tình trạng ngừng học tại Trường Đại học Vinh. Quá trình này bao gồm các bước: lựa chọn và chuẩn hóa dữ liệu, áp dụng thuật toán khai phá dữ liệu, kết quả thực nghiệm. Từ kết quả đạt được, chúng tôi rút ra các kết luận và đề xuất nhằm hạn chế tình trạng sinh viên bị buộc ngừng học tại Trường Đại học Vinh.

Email: uyendhv@gmail.com (N. T. Uyên)

## 2. Khai phá dữ liệu trong giáo dục

Khai phá dữ liệu là lĩnh vực nghiên cứu để trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Quá trình khai phá dữ liệu là quá trình khám phá kiến thức có trong cơ sở dữ liệu [5]. Khai phá dữ liệu giáo dục là lĩnh vực nghiên cứu có sự kết hợp của các phương pháp tính toán và phương pháp tâm lý nhằm mục đích hiểu thêm về hành vi học tập của người học [6]. Mục tiêu của việc khai phá dữ liệu giáo dục là: (1) dự đoán hành vi học tập trong tương lai bằng cách tạo ra mô hình dựa trên sự kết hợp các thông tin như kiến thức, thái độ, động lực, nhận thức của người học; (2) xác định được các nội dung quan trọng cần học và tối ưu hóa được trình tự giảng dạy; (3) nghiên cứu sự ảnh hưởng của các hình thức giảng dạy đến quá trình học tập của người học; và (4) thúc đẩy được các nghiên cứu khoa học về quá trình học tập thông qua việc xây dựng các mô hình tính toán dựa trên các dữ liệu giáo dục [7].

Việc nghiên cứu khai phá dữ liệu giáo dục cho phép trả lời được một số câu hỏi dạng như sau:

- Sinh viên sẽ có kết quả học tập như thế nào trong tương lai?
- Sinh viên nên học theo tiến trình nào để đạt được hiệu quả tốt nhất?
- Những hành vi nào của sinh viên có liên quan đến việc học tiếp lên các bậc học cao hơn (ví dụ: Thạc sỹ, Tiến sỹ)?
- Những hành vi nào của sinh viên cho thấy sự hài lòng, chủ động tham gia để hoàn thành tiến độ học tập?
- Môi trường học tập trực tuyến cần có được các chức năng nào để giúp cho việc học tập trực tuyến đạt hiệu quả tốt hơn?
- Yếu tố nào có thể cho phép dự đoán được mức độ thành công của người học trong tương lai.

Khai phá dữ liệu trong giáo dục đã và đang được nhiều nhà nghiên cứu quan tâm. Superby và cộng sự [3] sử dụng bảng câu hỏi để thu thập dữ liệu bao gồm thông tin cá nhân, các hành vi và nhận thức học tập của sinh viên. Các tác giả áp dụng các cách tiếp cận khác nhau như cây quyết định (decision tree), rừng ngẫu nhiên (random forest), mạng lưới thần kinh (neural network) và phân tích phân biệt tuyến tính (linear discriminant) để phân tích và dự đoán các yếu tố ảnh hưởng đến việc học tập của sinh viên. Tuy nhiên, có thể vì số lượng thông tin thu thập còn ít nên độ chính xác dự đoán chưa cao. Ashby và cộng sự [4] thu thập dữ liệu để nghiên cứu các yếu tố ảnh hưởng đến kết quả học tập của sinh viên khi tham gia các khóa học trực tuyến từ xa. Ayesha và cộng sự [3] áp dụng thuật toán K-means để dự đoán hành vi học tập của sinh viên. Những thông tin thu được có thể giúp cho giáo viên có những điều chỉnh kịp thời trong quá trình giảng dạy. Bharadwaj và cộng sự [9], Yadav và cộng sự [10] thu thập thông tin về tính chuyên cần, điểm thi, các hoạt động ngoại khóa của sinh viên để dự đoán kết quả học tập vào cuối học kỳ. Các thuật toán khai phá dữ liệu được các tác giả sử dụng là ID3, C4.5 and CART. Marie Bienkowski và cộng sự [11] nghiên cứu ứng dụng khai phá dữ liệu giáo dục để xây dựng chương trình học cá thể hóa. Lin [12] nghiên cứu xây dựng mô hình cho phép dự đoán được những sinh viên nào sẽ gặp khó khăn trong việc học, để từ đó có giải pháp hỗ trợ kịp thời. Dekker và cộng sự [13] sử dụng thuật toán khai phá dữ liệu Cây quyết định để xây dựng mô hình dự đoán tỷ lệ sinh viên có thể bị ngừng học sau học kỳ đầu tiên.

## 3. Xây dựng mô hình dự đoán

### *3.1. Thu thập và chuẩn hóa dữ liệu*

Các thông tin cần lấy thu thập để thực hiện xây dựng mô hình là: mã sinh viên, họ và tên, ngày sinh, nơi sinh, giới tính, điểm đầu vào, điểm các môn học trong 3 kỳ đầu của mỗi sinh viên. Những dữ liệu này được thu thập từ Phòng Công tác chính trị và Học sinh, sinh viên, Phòng Đào tạo và Trung tâm Công nghệ thông tin của Trường Đại học Vinh. Vì vậy, dữ liệu có độ tin cậy và chính xác cao, phản ánh đúng thông tin của sinh viên. Chúng tôi đã thu thập được thông tin của 555 sinh viên khóa 54, 55 và 56 ngành Công nghệ thông tin.

### *3.2. Tính độ ảnh hưởng của các thuộc tính*

Trích chọn các thuộc tính là việc lựa chọn các thuộc tính có ảnh hưởng đến kết quả dự đoán, các thuộc tính khác sẽ bị loại ra. Để xác định được thuộc tính nào có ảnh hưởng đến mô hình dự đoán, chúng tôi đã dùng phương pháp tính Độ lợi thông tin (Information Gain). Thực nghiệm phương pháp tính độ lợi thông tin bằng phần mềm WEKA, chúng tôi đã tính được trọng số ảnh hưởng và xếp hạng được các thuộc tính như Bảng 1.

**Bảng 1:** *Trọng số ảnh hưởng của từng thuộc tính*

| STT | Thuộc tính | Trọng số |
|---|---|---|
| **Nhóm thông tin chung** | | |
| 1 | Quê quán | 0,06326 |
| 2 | Thành phần gia đình | 0,02431 |
| 3 | Tôn giáo | 0,01945 |
| 4 | Giới tính | 0,01199 |
| **Nhóm thông tin điểm các môn** | | |
| 1 | Điểm đầu vào | 0,08135 |
| 2 | Ngôn ngữ Lập trình C | 0,04894 |
| 3 | Toán A2 - Giải tích I | 0,04866 |
| 4 | Tư tưởng Hồ Chí Minh | 0,03499 |
| 5 | Vật lý đại cương A1 | 0,03141 |
| 6 | Lý thuyết tối ưu | 0,02855 |
| 7 | Những nguyên lý cơ bản của Chủ nghĩa Mác Lênin II | 0,02786 |
| 8 | Kỹ thuật điện tử | 0,02603 |
| 9 | Toán A1 - Đại số tuyến tính | 0,02586 |
| 10 | Toán cao cấp nâng cao | 0,02384 |
| 11 | Ngoại ngữ 1 - Tiếng Anh | 0,02196 |
| 12 | Những nguyên lý cơ bản của Chủ nghĩa Mác Lênin I | 0,02149 |
| 13 | Giáo dục quốc phòng 1 | 0,01778 |
| 14 | Giáo dục quốc phòng 2 | 0,01489 |
| 15 | Ngoại ngữ 2 - Tiếng Anh | 0,01058 |
| 16 | Giáo dục quốc phòng 3 | 0,00574 |

### *3.3. Áp dụng thuật toán khai phá dữ liệu*

Chúng tôi sẽ tiến hành áp dụng thuật toán Naïve Bayes và Logistic Regression cho các tập thuộc tính như sau:

Trường hợp 1: Chạy thuật toán với tất cả 20 thuộc tính đầu vào được cho ở Bảng 2. Thuộc tính dự đoán là tình trạng cảnh báo Ngừng học (Có/Không).

Trường hợp 2: Chạy thuật toán với việc loại bỏ 2 thuộc tính có độ ảnh hưởng thấp nhất (GDQP 3 và Ngoại ngữ 2).

Trường hợp 3: Chạy thuật toán với việc loại bỏ 4 thuộc tính có độ ảnh hưởng thấp nhất (GDQP 3, Ngoại ngữ 2, Giới tính, GDQP 2).

Trường hợp 4: Chạy thuật toán với việc loại bỏ 6 thuộc tính có độ ảnh hưởng thấp nhất (GDQP 3, Ngoại ngữ 2, Giới tính, GDQP 2, GDQP 1, Tôn giáo).

Kết quả huấn luyện để xây dựng mô hình dự đoán với hai thuật toán khai phá dữ liệu Naïve Bayes và Logistic Regression cho cả 4 trường hợp được cho ở Bảng 2.

**Bảng 2:** *Độ chính xác của mô hình dự đoán so với dữ liệu thực tế*

| **Phương pháp** | **Độ chính xác** | | | |
|---|---|---|---|---|
| | **Trường hợp 1** | **Trường hợp 2** | **Trường hợp 3** | **Trường hợp 4** |
| Naive Bayes | 62% | 62% | 68% | 68% |
| Logistic Regression | 88% | 88% | 88% | 88% |

Như vậy, thuật toán Logistic Regression cho kết quả dự đoán cao hơn so với thuật toán Naive Bayes.

### *3.4. Kết quả và phân tích*

Qua thực nghiệm với sinh viên ngành Công nghệ thông tin, có thể thấy các yếu tố ảnh hưởng nhiều đến tình trạng ngừng học là: điểm đầu vào, quê quán, môn Ngôn ngữ Lập trình C, môn Toán A2 (Giải tích I), môn Tư tưởng Hồ Chí Minh. Chi tiết các yếu tố ảnh hưởng đã được trình bày ở Bảng 1. Những sinh viên có điểm thấp ở các môn học Ngôn ngữ Lập trình C, Toán A2 (Giải tích I), Tư tưởng Hồ Chí Minh và có điểm thấp khi thi đầu vào đại học thì có xu thế bị buộc ngừng học. Ngoài ra yếu tố quê quán cũng ảnh hưởng cao đến tình trạng ngừng học của sinh viên. Những sinh viên cùng quê thường có xu hướng đạt kết quả học tập tương tự nhau.

## 4. Kết luận

Hiện nay, vấn đề dự báo tình trạng bị buộc ngừng học là khá cấp thiết. Tại Trường Đại học Vinh, việc này đang được thực hiện một cách cơ học thông qua tính điểm tích lũy theo từng kỳ. Trong bài báo này, chúng tôi đề xuất phương pháp dự đoán tình trạng bị buộc ngừng học bằng sử dụng kỹ thuật khai phá dữ liệu Naïve Bayes và Logistic Regression. Bằng phương pháp này, các nhân tố ảnh hưởng đến tình trạng ngừng học của sinh viên có thể được phát hiện sớm để nhà trường có biện pháp hỗ trợ sinh viên trong việc học tập ở các kỳ tiếp theo. Việc thực nghiệm với dữ liệu sinh viên ngành Công nghệ thông tin đã chứng minh được tính khả thi của phương pháp. Trong tương lai, chúng tôi sẽ thực nghiệm với dữ liệu sinh viên các ngành khác, để có thể đề xuất được một mô hình dự đoán kết quả học tập ở nhiều mức khác nhau như: xuất sắc, giỏi, khá, trung bình, yếu, ngừng học,...

## TÀI LIỆU THAM KHẢO

[1] Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis", *Depok*: *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 147-154, 2015.

[2] Kharya Shweta, Shika Agrawal and Sunita Soni, "Naive Bayes classifiers: A probabilistic detection model for breast cancer", *International Journal of Computer Applications 92.10: 0975-8887*, 2014.

[3] Superby J. F., Vandamme J. P. and Meskens N., *Determination of factors influencing the achievement of the first-year university students using data mining methods*, Workshop on Education, 2006.

[4] Ashby A., *Monitoring Student Retention in the Open University: Detritions, measurement, interpretation and action*, Open Learning, 19(1), pp. 65-78, 2004.

[5] Hand David J., *Data Mining*, Encyclopedia of Environmetrics 2, 2006.

[6] Romero Cristobal, Ventura Sebastian, "Data mining in education", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. 12-27, 2013.

[7] Baker Ryan S. J. D. and Yacef Kalina, "The state of educational data mining in 2009: A review and future visions", *Journal of Educational Data Mining*, Vol. 1, No. 1, pp. 3-17, 2009.

[8] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar and M. Inayat Khan , "Data mining model for higher education system", *European Journal of Scientific Research*, Vol. 43, No. 1, pp. 24-29, 2010.

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

[10] S. K. Yadav, B. K. Bharadwaj and S. Pal, *Data Mining Applications: A Comparative Study for Predicting Student's Performance*, International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19, 2011.

[11] Marie Bienkowski, Mingyu Feng and Barbara Means, *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics*, Washington D. C. : U. S. Department of Education, 2012.

[12] Lin S. H., "Data mining for student retention management", *ACM Journal of Computing Sciences in Colleges*, Vol. 27, No. 4, pp. 92-99, 2012.

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

## SUMMARY

## PREDICTING STUDENT'S ACADEMIC PERFORMANCE BY APPLYING DATA MINING TECHNIQUE

The situation of students being forced to stop their studies is currently very popular at universities in Vietnam. This paper proposes a method for predicting students' dropout based on the analysis of data from the university entrance scores, paper scores of subjects in the first three semesters and the current learning status of more than 555 students majoring in IT at Vinh University. Through these data, the Logistic Regression and Naïve Bayes data mining algorithms were applied to find a suitable model for predicting students' dropout in the next courses. This study will help the university to give early warnings and supports to reduce the rate of students' dropout in the next courses.

**Key words:** Education data mining(EDM); Dropout prediction.