BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC SAO ĐỎ

Địa chỉ:
- **Số 1:** Số 76, Nguyễn Thị Duệ, Thái Học 2, phường Sao Đỏ, thành phố Chí Linh, tỉnh Hải Dương.
- **Số 2:** Số 72, đường Nguyễn Thái Học, phường Thái Học, thành phố Chí Linh, tỉnh Hải Dương.
- **Điện thoại:** (0220) 3882 269 **Fax:** (0220) 3882 921 **Website:** http://saodo.edu.vn **Email:** info@saodo.edu.vn

P. ISSN 1859-4190
E. ISSN 2815-553X

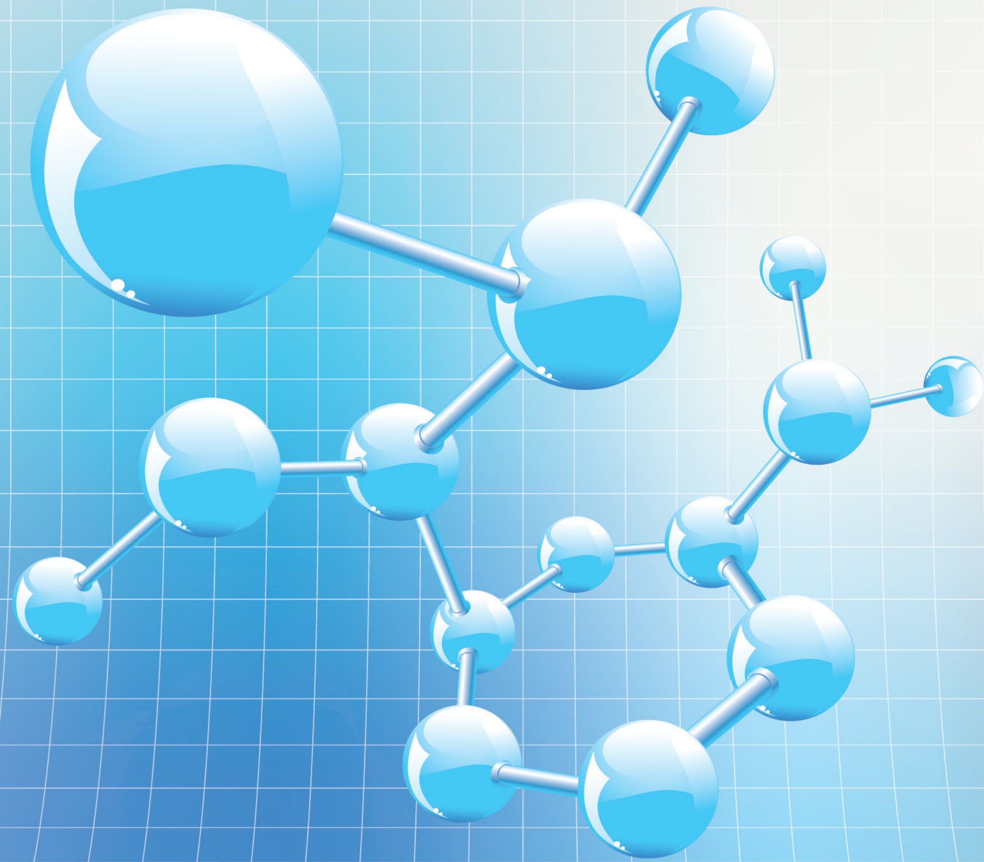
Số 2 (81)
2023

SỐ 2 (81) 2023
TẠP CHÍ NGHIÊN CỨU KHOA HỌC
P.ISSN 1859-4190 - E.ISSN 2815-553X

Tạp chí NGHIÊN CỨU KHOA HỌC ĐẠI HỌC SAO ĐỎ

SCIENTIFIC JOURNAL - SAO DO UNIVERSITY

P. ISSN 1859-4190
E. ISSN 2815-553X



Địa chỉ Tòa soạn:
Trường Đại học Sao Đỏ.
Số 76, Nguyễn Thị Duệ, Thái Học 2, phường Sao Đỏ, thành phố Chí Linh, tỉnh Hải Dương.
Điện thoại: (0220) 3587213, Fax: (0220) 3882 921, Hotline: 0912 107858/0936 847980.
Website: http://tapchikhcn.saodo.edu.vn/Email: tapchikhcn@saodo.edu.vn.

Giấy phép xuất bản số: 620/GP-BTTTT ngày 17/9/2021 của Bộ Thông tin và Truyền thông.
In 2.000 bản, khổ 21 × 29,7cm, tại Công ty TNHH in Tre Xanh, cấp ngày 17/02/2011.

Số 2 (81)
2023

P. ISSN 1859-4190
E. ISSN 2815-553X

Địa chỉ Tòa soạn:
Trường Đại học Sao Đỏ.
Số 76, Nguyễn Thị Duệ, Thái Học 2, phường Sao Đỏ, thành phố Chí Linh, tỉnh Hải Dương.
Điện thoại: (0220) 3587213, Fax: (0220) 3882 921, Hotline: 0912 107858/0936 847980.
Website: http://tapchikhcn.saodo.edu.vn/Email: tapchikhcn@saodo.edu.vn.

Giấy phép xuất bản số: 620/GP-BTTTT ngày 17/9/2021 của Bộ Thông tin và Truyền thông.
In 2.000 bản, khổ 21 × 29,7cm, tại Công ty TNHH in Tre Xanh, cấp ngày 17/02/2011.

THỂ LỆ GỬI BÀI
TẠP CHÍ NGHIÊN CỨU KHOA HỌC, TRƯỜNG ĐẠI HỌC SAO ĐỎ

Tạp chí Nghiên cứu khoa học, Trường Đại học Sao Đỏ (P. ISSN 1859-4190, E. ISSN 2815-553X), thường xuyên công bố kết quả, công trình nghiên cứu khoa học và công nghệ của các nhà khoa học, cán bộ, giảng viên, nghiên cứu sinh, học viên cao học, sinh viên ở trong và ngoài nước.

1. Tạp chí xuất bản 01 số/quý bằng hai ngôn ngữ tiếng Việt và tiếng Anh. Tạp chí nhận đăng các bài báo khoa học thuộc các lĩnh vực: Điện - Điện tử - Tự động hóa; Cơ khí - Động lực; Kinh tế; Triết học - Xã hội học - Chính trị học; Các lĩnh vực khác gồm: Công nghệ thông tin; Hóa học - Công nghệ thực phẩm; Ngôn ngữ học; Toán học; Vật lý; Văn hóa - Nghệ thuật - Thể dục thể thao...
2. Bài nhận đăng là những công trình nghiên cứu khoa học chưa công bố trong bất kỳ ấn phẩm khoa học nào.
3. Tòa soạn chỉ nhận bài báo gửi online trên website http://tapchikhcn.saodo.edu.vn. Bài báo gửi về toà soạn dưới dạng file điện tử (*.doc *.docx và *.pdf); cuối bài báo, tác giả ghi rõ thông tin địa chỉ liên hệ, số điện thoại, email và cập nhật thông tin trên website. Bài báo phải được trình bày đúng định dạng, rõ ràng; Trường hợp bài báo phải chỉnh sửa theo thể lệ hoặc theo yêu cầu của Phản biện thì tác giả sẽ cập nhật trên website. Người phản biện sẽ do toà soạn mời. Toà soạn không gửi lại bài nếu không được đăng.
4. Các công trình thuộc đề tài nghiên cứu có Cơ quan quản lý cần kèm theo giấy phép cho công bố của cơ quan (Tên đề tài, mã số, tên chủ nhiệm đề tài, cấp quản lý,...).
5. Tên bài báo trình bày bằng hai ngôn ngữ (tiếng Việt và tiếng Anh), font Arial, cỡ chữ 14, in đậm, căn giữa.
6. Tên tác giả (không ghi học hàm, học vị), font Arial, cỡ chữ 10, in đậm, căn lề phải; cơ quan công tác của các tác giả, font Arial, cỡ chữ 9, in nghiêng, căn lề phải.
7. Chữ "Tóm tắt" in đậm, font Arial, cỡ chữ 10; Nội dung tóm tắt của bài báo không quá 10 dòng, trình bày bằng hai ngôn ngữ (tiếng Việt và tiếng Anh), font Arial, cỡ chữ 10, in thường.
8. Chữ "Từ khóa" in đậm, nghiêng, font Arial, cỡ chữ 10; Có từ 03÷05 từ khóa, font Arial, cỡ chữ 10, in nghiêng, ngăn cách nhau bởi dấu chấm phẩy, cuối cùng là dấu chấm.
9. Nội dung bài báo viết bằng tiếng Việt hoặc tiếng Anh; Nếu là bài báo viết bằng tiếng Việt: Tiêu đề tiếng Việt trước, tiếng Anh sau; Tóm tắt tiếng Việt trước, tiếng Anh sau; Từ khóa tiếng Việt trước, tiếng Anh sau; Nếu là bài báo viết bằng tiếng Anh: Tiêu đề tiếng Anh trước, tiếng Việt sau; Tóm tắt tiếng Anh trước, tiếng Việt sau; Từ khóa tiếng Anh trước, tiếng Việt sau.
10. Bài báo được đánh máy trên khổ giấy A4 (21 × 29,7cm) có độ dài không quá 8 trang, font Arial, cỡ chữ 10, giãn dòng At least 12pt, Before 3pt, After 3pt; căn lề trên 2.5cm, dưới 2.5cm, trái 3cm, phải 2cm; hình vẽ phải rõ ràng, đủ nét và được định dạng dưới dạng file ảnh (*.jpg); Phương trình, công thức phải soạn thảo bằng Mathtype hoặc Equation; Phần nội dung bài báo được chia thành 02 cột, khoảng cách cột là 1cm; Trong trường hợp hình vẽ, hình ảnh có kích thước lớn, bảng biểu có độ rộng lớn hoặc công thức, phương trình dài thì cho phép trình bày dưới dạng 01 cột.
11. Tài liệu tham khảo được sắp xếp theo thứ tự tài liệu được trích dẫn trong bài báo.
    - Nếu là sách/luận án: Tên tác giả (năm), Tên sách/luận án/luận văn, Nhà xuất bản/Trường/Viện, lần xuất bản/tái bản.
    - Nếu là bài báo/báo cáo khoa học: Tên tác giả (năm), Tên bài báo/báo cáo, Tạp chí/Hội nghị/Hội thảo, Tập/Kỳ yếu, số, trang.
    - Nếu là trang web: Phải trích dẫn đầy đủ tên website và đường link, ngày cập nhật.
12. Định dạng mẫu bài báo tham khảo tại địa chỉ http://tapchikhcn.saodo.edu.vn/news/detail/198/format_paper

Bài báo sau khi xuất bản sẽ được công bố trên http://tapchikhcn.saodo.edu.vn.

THÔNG TIN LIÊN HỆ:

Ban Biên tập Tạp chí Nghiên cứu khoa học, Trường Đại học Sao Đỏ
Phòng 203, Tầng 2, Nhà B1, Trường Đại học Sao Đỏ.
Địa chỉ: Số 76, Nguyễn Thị Duệ, Thái Học 2, phường Sao Đỏ, thành phố Chí Linh, tỉnh Hải Dương.
Điện thoại: (0220) 3587213, Fax: (0220) 3882921, Hotline: 0912 107858/0936 847980.
Website: http://tapchikhcn.saodo.edu.vn
Email: tapchikhcn@saodo.edu.vn

TẠP CHÍ
NGHIÊN CỨU KHOA HỌC
ĐẠI HỌC SAO ĐỎ

TRONG SỐ NÀY

Số 2(81) 2023

TẠP CHÍ
NGHIÊN CỨU KHOA HỌC
ĐẠI HỌC SAO ĐỎ

TRONG SỐ NÀY

Số 2(81) 2023

SCIENTIFIC JOURNAL
SAO DO UNIVERSITY

CONTENTS

No 2(81) 2023

TITLE FOR ELECTRICITY - ELECTRONICS - AUTOMATION

TITLE FOR INFORMATION TECHNOLOGY

TITLE FOR MECHANICAL AND DRIVING POWER ENGINEERING

SCIENTIFIC JOURNAL
SAO DO UNIVERSITY

CONTENTS

No 2(82) 2023

Phân lớp người dùng tiềm năng của hệ thống học trực tuyến vuihoc

Classify potential users of online learning system vuihoc

Hoàng Thị Ngọc Diệp[1*], Trần Duy Khánh[1], Phạm Huy Hoàng[2], Trần Đình Khang[2]

**Email của tác giả liên hệ: hoangdiepdtth@gmail.com*

[1]Trường Đại học Sao Đỏ

[2]Đại học Bách khoa Hà Nội

Ngày nhận bài: 22/5/2023

Ngày nhận bài sửa sau phản biện: 02/3/2023

Ngày chấp nhận đăng: 30/6/2023

Tóm tắt

Hệ thống phân lớp người dùng tiềm năng của trang học trực tuyến *Vuihoc* có nhiệm vụ phân tích từ dữ liệu của người học trong quá khứ để đưa ra dự đoán những người dùng nào có khả năng sẽ mua khóa học kế tiếp trong tương lai. Vấn đề ở đây là cần trích chọn được những thông tin quan trọng của người dùng và một phương pháp phù hợp để mô hình phân lớp đạt hiệu quả. Để giải quyết vấn đề này, phương pháp học có giám sát được ứng dụng vào hệ thống để giúp phân lớp và dự đoán từ tập người dùng của trang học trực tuyến *Vuihoc*. Cụ thể, bài báo sử dụng phương pháp phân loại Support Vector Machine và kỹ thuật cân bằng dữ liệu SMOTE, dựa trên dữ liệu về thông tin cá nhân cũng như lịch sử hoạt động của người dùng trên trang web (ID của người dùng, họ tên, năm sinh, tổng tiền đã mua khóa học, số lượng khóa học đã mua, thời gian tương tác với hệ thống thông qua chức năng ôn luyện và làm bài kiểm tra,điểm ôn tập và kiểm tra...), chia họ thành hai lớp: Lớp những người dùng không mua khóa tiếp theo và lớp người dùng mua khóa học tiếp theo.

Từ khóa: Phương pháp phân loại Support Vector Machine; kỹ thuật cân bằng dữ liệu SMOTE; hệ thống phân lớp; học có giám sát.

Abstract

The potential user classification system of the Vuihoc online learning site has the task of analyzing data from past learners to predict which users are likely to buy the next course in the future. The problem here is to extract important user information and a suitable method for the classification model to be effective. To solve this problem, the supervised learning method is applied to the system to help classify and predict from the user set of the Vuihoc online learning site. Specifically, the paper uses the Support Vector Machine classification method and the SMOTE data balancing technique, based on data about personal information as well as the user's activity history on the website, dividing them into two classes: The class of users who do not purchase the next course and the class of users who buy the next course.

Keywords: Classification method Support Vector Machine; data balancing technique SMOTE; classification system; Supervised learning.

1. ĐẶT VẤN ĐỀ

Với sự phát triển của công nghệ, của Internet như hiện nay, có thể thấy càng ngày càng có nhiều người quan tâm đến việc học trực tuyến bởi lợi ích mà nó mang lại: Tiết kiệm thời gian, tiết kiệm chi phí, phù hợp với mọi lứa tuổi,.... Trong thời kỳ dịch bệnh Covid-19, nền giáo dục vẫn được duy trì khắp nơi, nhờ các hệ thống học trực tuyến. Trong các hệ thống đó, lưu trữ nguồn dữ liệu về nội dung các khóa học, về người học và về quá trình học tập, quá trình tương tác của từng người học. Từ các dữ liệu đó, có thể trích rút được các thông tin về thái độ, năng lực, hành vi,… của người học (những người tương tác thường xuyên hơn gần đây dựa vào điểm số các bài luyện tập và bài kiểm tra, có thể người dùng đang thích thú với các bài học), giúp dự đoán về kết quả học tập, trợ giúp nâng cao hiệu quả học tập hoặc giới thiệu các khóa học tiếp theo.

Việc huấn luyện mô hình học máy với tập dữ liệu mất cân bằng thường làm cho mô hình bị nghiêng về lớp đa số, gây ra lỗi cao hơn. Một trong số những phương pháp hiệu quả để giải quyết vấn đề này là phương pháp lấy lại mẫu (resampling): Phương pháp Oversampling (lấy vượt mẫu) là lựa chọn tốt khi bài toán không có

Người phản biện: 1. GS.TSKH. Thân Ngọc Hoàn
2. TS. Nguyễn Phúc Hậu

lượng dữ liệu thực sự lớn. Nhược điểm của phương pháp này là việc sao chép chính xác những phần tử của lớp thiểu số có thể gây ra “overfitting” (mô hình quá khớp với dữ liệu huấn luyện, gây việc dự đoán nhầm, chất lượng mô hình không tốt với dữ liệu test). Đồng thời, cách này cũng làm cho số lượng mẫu dữ liệu cho huấn luyện tăng lên, làm tăng thời gian huấn luyện [1]. Phương pháp Undersampling (lấy giảm mẫu) lớp đa số sẽ được lấy lại mẫu bằng cách giảm bớt những mẫu bên trong nó để cân bằng với lớp thiểu số. Phương pháp này tốt khi lượng dữ liệu là quá lớn. Tuy nhiên, có thể khi giảm mẫu sẽ làm mất những thông tin quan trọng của tập dữ liệu, gây ra "underfitting" (mô hình quá đơn giản, không mô tả được xu hướng của dữ liệu) [2]. Máy vector hỗ trợ (SVM - Support Vector Machine hay Support-Vector Networks [4]) xây dựng cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. Với bài toán phân loại 2 lớp SMV tìm một siêu phẳng để phân tác các điểm dữ liệu. Siêu phẳng sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu. Trường hợp phân tách tuyến tính thì dùng siêu phẳng phân tách có lề lớn nhất [5]. Trường hợp phân tách không tuyến tính dùng phương pháp kernel trick [6]. Trong quá trình huấn luyện có thể cân nhắc việc sử dụng kỹ thuật SMOTE [3] để cân bằng dữ liệu giữa hai lớp. Các công trình nghiên cứu trên đều chỉ rõ các ưu nhược điểm của từng phương pháp. Việc kết hợp các phương pháp lấy mẫu và phân loại mẫu có sử dụng SMC và kỹ thuật SMOTE đã chọn được mô hình phân lớp phù hợp, lưu lại mô hình đó để phục vụ cho bước dự đoán tiếp theo.

Nội dung bài báo này đề cập đến việc khai phá dữ liệu về thông tin người dùng, dữ liệu về hành vi của người dùng tương tác với hệ thống học trực tuyến để phân loại nhóm người dùng tiềm năng có thể sẽ mua tiếp khóa học giúp cho doanh nghiệp có chiến lược marketing, chăm sóc khách hàng hiệu quả. Từ đó, đưa hệ thống học trực tuyến phát triển hơn, không chỉ trong cung cấp những khóa học chất lượng, mà còn là dịch vụ thân thiện với người học, góp phần tăng doanh thu cho doanh nghiệp. Dữ liệu thử nghiệm trong bài báo này lấy từ trang học trực tuyến VUIHOC là một trang học mới hoạt động từ đầu năm 2019, nhưng sau gần một năm đã có trên 15,000 người học, với trung bình khoảng 2,000 học viên sử dụng hệ thống mỗi ngày [7].

Bài báo được trình bày trong năm phần, ở phần 2 trình bày chung về bài toán phân lớp người học trong hệ thống học trực tuyến, trong phần 3 trình bày các kỹ thuật giải quyết bài toán, phần 4 trình bày thực nghiệm với các dữ liệu từ hệ thống học trực tuyến VUIHỌC, phần 5 nêu kết luận.

2. BÀI TOÁN PHÂN LỚP NGƯỜI HỌC TRONG HỆ THỐNG HỌC TRỰC TUYẾN

Bài toán phân lớp nhóm người dùng tiềm năng mua khóa học tiếp theo của hệ thống học trực tuyến là việc xây dựng mô hình phân tích, huấn luyện từ dữ liệu về thông tin cá nhân, hành vi tương tác người dùng trang web và lịch sử mua khóa học để phân lớp, dự đoán khách hàng tiềm năng có thể mua khóa học tiếp theo của trang web.

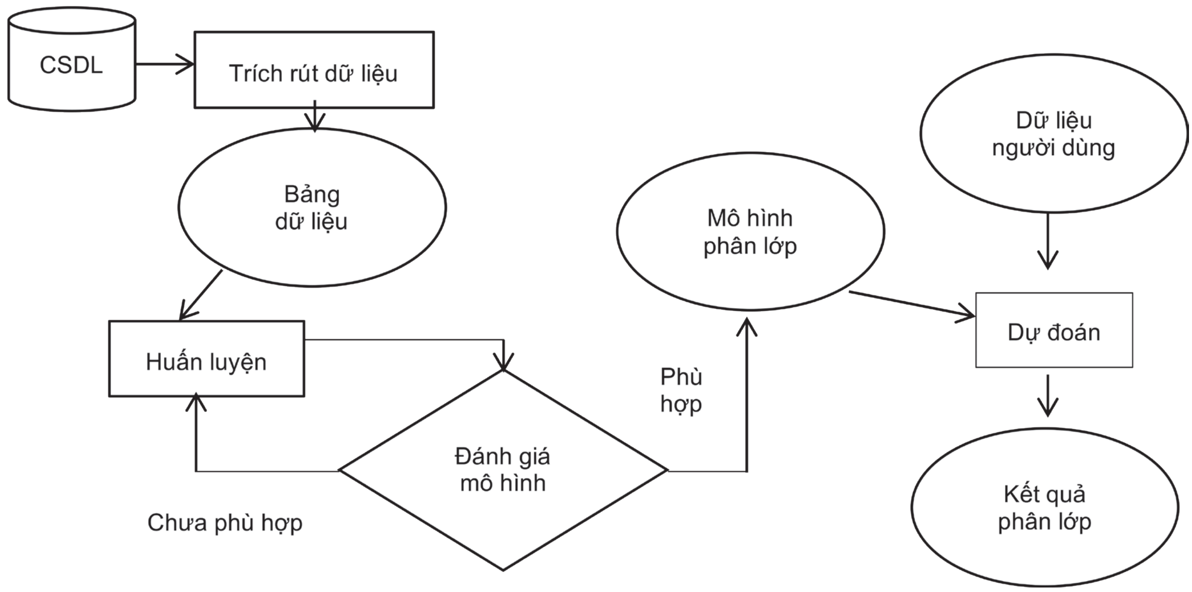
Các vấn đề cần giải quyết:

Nhập dữ liệu của hệ thống học trực tuyến.

Trích rút dữ liệu ra bảng dữ liệu, sử dụng cho bài toán phân lớp.

Tiến hành huấn luyện để tạo ra mô hình phân lớp.

Dự đoán cho tập dữ liệu khách hàng tiềm năng trong tương lai.



Hình 1. *Mô hình tổng quát bài toán phân lớp người dùng tiềm năng cho hệ thống học trực tuyến*

3. PHƯƠNG PHÁP

3.1. Trích rút dữ liệu từ hệ thống học trực tuyến

Cơ sở dữ liệu của một hệ thống học trưc tuyến gồm rất nhiều bảng, trong bài toán này sẽ quan tâm đến các bảng dữ liệu về thông tin cá nhân cũng như về hành vi, hoạt động tương tác của người học trên hệ thống. Ví dụ, ở hệ thống VUIHOC có các bảng:

Bảng "*core_user*": Bảng dữ liệu có những thông tin cơ bản về người dùng (học sinh).

Bảng "*core_user_profile*": Thông tin chi tiết về người dùng là học sinh.

Bảng "*bus_order*": Thông tin về khóa học mà phụ huynh mua cho con.

Bảng "*bus_order_detail*": Thông tin chi tiết ứng với từng đơn hang mua khóa học.

Bảng "*core_log_activity_beta*": Bảng về các hành vi của người dùng trên hệ thống.

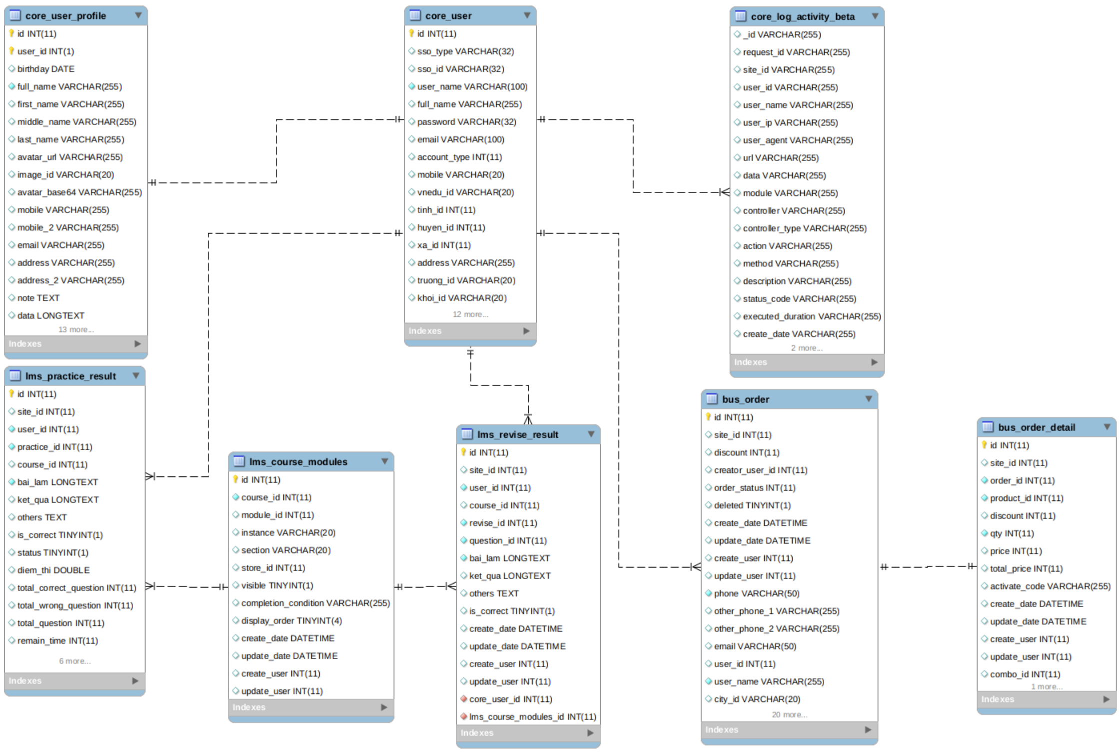
Bảng "*lms_course_modules*": Bảng về các câu hỏi, bài tập và bài kiểm tra của từng khóa học.

Bảng "*lms_revise_result*": Bảng về kết quả làm các câu hỏi trên hệ thống.

Bảng "*lms_practice_result*": Bảng về kết quả của các bài kiểm tra trên hệ thống.

Hình 2 là biểu đồ liên kết giữa các bảng:

Tiếp theo là "Trích rút dữ liệu", ta sẽ xử lý, trích rút dữ liệu để lấy ra Bảng dữ liệu phục vụ cho bài toán phân lớp. Cụ thể từ dữ liệu có trong những bảng trên, ta cần nối các bảng với nhau thông qua các khóa bằng các truy vấn SQL, từ đó sẽ lấy ra được các thuộc tính hay đặc trưng ứng với từng người dùng. Khi đó ta sẽ được bảng đặc trưng gồm các cột, mỗi cột sẽ là một đặc trưng của người dùng. Sau đó, ta sẽ tiền xử lý dữ liệu như chuẩn hóa hoặc thay thế các giá trị bị thiếu cũng như loại bỏ những điểm ngoại lai. Do tập dữ liệu chưa có nhãn nên ta sẽ tiến hành gán đối với dữ liệu trong quá khứ. Sẽ có hai nhãn, nhãn ứng với những người dùng không mua khóa học tiếp theo và nhãn ứng với những người dùng mua khóa học sau đó. Cuối cùng ta sẽ thu được một Bảng dữ liệu gồm các đặc trưng và nhãn ứng với từng người dùng để sử dụng cho việc huấn luyện.



Hình 2. *Sơ đồ thực thể - liên kết từ hệ thống VUIHOC*

3.2. Huấn luyện mô hình phân lớp

Trong bước này, ta sẽ tiến hành huấn luyện từ Bảng dữ liệu lấy được trước đó sử dụng phương pháp Support Vector Machine (SVM). Trong quá trình huấn luyện có thể cân nhắc việc sử dụng kỹ thuật SMOTE để cân bằng dữ liệu giữa hai lớp. Các tham số của phương pháp SVM, cũng như tỷ lệ chọn để cân bằng dữ liệu bằng SMOTE có thể được điều chỉnh bằng cách tối ưu hóa hiệu năng trên một tập con của tập huấn luyện (validation set). Khi đã chọn được mô hình phân lớp phù hợp, lưu lại mô hình đó để phục vụ cho bước dự đoán tiếp theo.

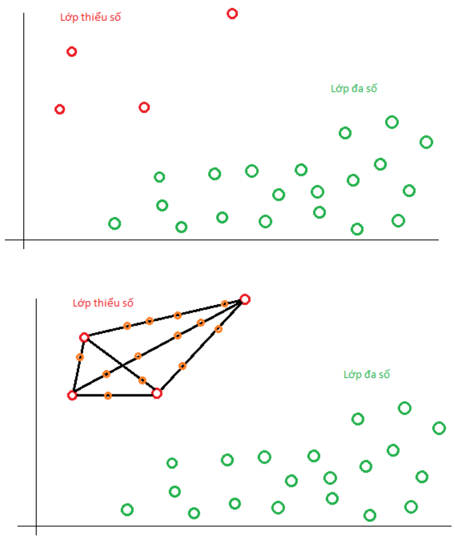
Tập dữ liệu mất cân bằng khi mà sự chênh lệch giữa số phần tử của các lớp là khác nhau rõ rệt. Việc mất cân bằng giữa các lớp dữ liệu là một vấn đề khá phổ biến của các bài toán phân loại trong học máy, đặc biệt là các bài toán thực tế như: Chẩn đoán y khoa, phát hiện gian lận... Đối với bài toán có 2 lớp, lớp thiểu số (minority class) sẽ là lớp có ít phần tử hơn và lớp đa số (majority class) sẽ là lớp có nhiều phần tử hơn.

Việc huấn luyện mô hình học máy với tập dữ liệu mất cân bằng thường làm cho mô hình bị nghiêng về lớp đa số, gây ra lỗi cao hơn. Một trong số những phương pháp hiệu quả để giải quyết vấn đề này là phương pháp lấy lại mẫu (resampling):

Oversampling - lấy vượt mẫu: Lớp thiểu số sẽ được thêm các phần tử mới bằng cách sao chép những phần tử bên trong nó. Đây là một trong những phương pháp được công bố sớm nhất và được chứng minh là mạnh mẽ [1]. Phương pháp là lựa chọn tốt khi bài toán không có lượng dữ liệu thực sự lớn. Nhược điểm của phương pháp này là việc sao chép chính xác những phần tử của lớp thiểu số có thể gây ra “overfitting” (mô hình quá khớp với dữ liệu huấn luyện, gây việc dự đoán nhầm, chất lượng mô hình không tốt với dữ liệu test). Đồng thời, cách này cũng làm cho số lượng mẫu dữ liệu cho huấn luyện tăng lên, làm tăng thời gian huấn luyện.

Undersampling - lấy giảm mẫu: Lớp đa số sẽ được lấy lại mẫu bằng cách giảm bớt những mẫu bên trong nó để cân bằng với lớp thiểu số. Phương pháp này tốt khi lượng dữ liệu là quá lớn. Tuy nhiên, có thể khi giảm mẫu sẽ làm mất những thông tin quan trọng của tập dữ liệu, gây ra “underfitting” (mô hình quá đơn giản, không mô tả được xu hướng của dữ liệu [2].

Một trong các kỹ thuật phát triển từ Oversampling là Synthetic Minority Oversampling Technique (SMOTE). SMOTE sẽ sử dụng giải thuật hàng xóm gần nhất (nearest neighbors) để tạo những mẫu dữ liệu mới cho mô hình huấn luyện. Kỹ thuật SMOTE sẽ tạo dữ liệu mới bằng cách kết hợp các đặc trưng của điểm dữ liệu của lớp thiểu số với các hàng xóm của nó trong lớp này. Cách tiếp cận này giúp mẫu dữ liệu của lớp thiểu số tổng quát hơn.



Hình 3. *Các điểm dữ liệu được tạo bởi SMOTE*

Máy vector hỗ trợ (SVM - Support Vector Machine hay Support-Vector Networks [4]) là một khái niệm trong thống kê và khoa học máy tính do Vapnik và Chervonenkis xây dựng cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó, SVM là một thuật toán phân loại nhị phân. Với một tập huấn luyện thuộc hai lớp cho trước, thuật toán huấn luyện SVM xây dựng một mô hình để phân loại các mẫu khác vào hai lớp đó. Một mô hình SVM là một cách biểu diễn các điểm trong không gian và lựa chọn ranh giới giữa hai lớp sao cho khoảng cách từ các mẫu huấn luyện tới ranh giới là xa nhất có thể. Các mẫu mới cũng được biểu diễn trong cùng một không gian và được thuật toán dự đoán thuộc một trong hai lớp tùy vào mẫu đó nằm ở phía nào của ranh giới.

Ý tưởng của SVM là tìm một siêu phẳng (hyper lane) để phân tách các điểm dữ liệu. Siêu phẳng này sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu. Siêu phẳng được biểu diễn bằng hàm số $w^Tx = b$. Nhưng vấn đề là có rất nhiều siêu phẳng, chúng ta phải chọn cái nào tối ưu nhất. Chúng ta cần một đường phân chia sao cho khoảng cách từ điểm gần nhất của mỗi lớp tới đường phân chia (hay còn gọi là lề - margin) như nhau và khoảng cách đó là lớn nhất. Việc margin rộng hơn sẽ mang lại hiệu ứng phân lớp tốt hơn vì sự phân chia giữa hai lớp là rạch ròi hơn. Cụ thể, chúng ta giả sử phải phân loại tập dữ liệu các lớp dương (màu xanh) nhãn là 1 và các dữ liệu lớp âm (màu đỏ) là -1. Siêu phẳng phân tách giữa hai lớp tạo ra hai nửa không gian dữ liệu. Không gian các dữ liệu lớp âm thỏa mãn và không gian dữ liệu lớp dương thỏa mãn .

Tiếp theo, ta chọn hai siêu phẳng lề H_1: $w^Tx + b = -1$ đi qua điểm thuộc lớp âm, H_2: $w^Tx + b = 1$ đi qua điểm thuộc lớp dương và đều song song với H_0. Khoảng cách từ H_1 đến H_0 là d_ và từ H_2 đến H_0 là d+. Margin sẽ là m = d + d_+. Siêu phẳng tối ưu mà chúng ta cần chọn là siêu phẳng phân tách có lề lớn nhất [5].

Khoảng cách từ một điểm x_i nằm trên H_1 đến siêu phẳng H_0 là:

$$d_- = \frac{|w^T x_i + b|}{\|w\|} = \frac{1}{\|w\|} \qquad (1)$$

Khoảng cách từ một điểm x_j nằm trên H_2 đến siêu phẳng H_0 là:

$$d_+ = \frac{|w^T x_j + b|}{\|w\|} = \frac{1}{\|w\|} \qquad (2)$$

Trong đó:

||w|| là độ dài của vector w: $\|w\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$.

Từ đó ta có thể tính được mức lề.

$$m = d_- + d_+ = \frac{2}{\|w\|} \qquad (3)$$

Vì vậy, mà các điểm nằm trên hai siêu phẳng H_1 và H_2 là các Support Vector.

Vậy việc huấn luyện trong SVM tương đương với giải quyết bài toán cực tiểu có ràng buộc sau đây:

Cực tiểu hóa:

$$\frac{\|w\|}{2} \tag{4}$$

Với điều kiện:

$$y_i(w^T x + b) \geq 1, \forall i = 1..n \tag{5}$$

Việc xác định siêu phẳng H_0 được giả sử trong điều kiện lý tưởng tập dữ liệu có thể phân tách tuyến tính, tìm được hai siêu phẳng lề H_1 và H_2 mà không có điểm dữ liệu nào nằm giữa chúng. Vậy trong trường hợp tập dữ liệu có nhiều điểm gây nhiễu, các điểm này không thỏa mãn điều kiện (5), bài toán không tìm được lời giải. Đối với các trường này, chúng ta cần nới lỏng các điều kiện lề bằng việc sử dụng các biến . Bài toán trong trường hợp này trở thành:

Cực tiểu hóa:

$$\frac{\|w\|}{2} + c\sum_{i=1}^{n} \epsilon_i \tag{6}$$

Với điều kiện:

$$\begin{cases} y_i(w^T x + b) \geq 1 - \epsilon_i, \forall i = 1..n \\ \epsilon_i \geq 0, \forall i = 1..n \end{cases} \tag{7}$$

Trong trường hợp phân tách không tuyến tính, SVM dựa trên các hàm nhân để ánh xạ các điểm dữ liệu sang một không gian khác (thường có số chiều lớn hơn). Trong không gian mới bằng cách áp dụng một hàm ánh xạ phi tuyến ϕ:

$$\begin{aligned} \emptyset &: X \to F \\ x &\mapsto \emptyset(x) \end{aligned} \tag{8}$$

Các điểm dữ liệu sẽ khả tách tuyến tính hoặc tìm được mặt phân tách với ít lỗi hơn so với không gian ban đầu. Nhưng trở ngại là trong không gian này, do số chiều của dữ liệu tăng lên rất nhiều so với không gian ban đầu làm cho chi phí tính toán vô cùng tốn kém. Vì vậy, trong SVM người ta tìm ra cách không cần phải tính trực tiếp hàm $\phi()$, mà ta chỉ cần tính tích vô hướng của hai vector bất kỳ trong không gian mới: $k(x,z) = \phi(x)^T\phi(z)$. Phương pháp này gọi là kernel trick [6]. Các hàm nhân thường dùng: Polynomial, Gaussian RBF, Sigmoidal.

SVM được xây dựng cho bài toán phân loại 2 lớp, khi áp dụng vào bài toán phân lớp nhóm người dùng tiềm năng mua khóa học tiếp theo của hệ thống Vuihoc, thực hiện huấn luyện trên tập dữ liệu về thông tin và hành vi của người học trên hệ thống học trực tuyến này.

3.3. Dự đoán khách hàng tiềm năng

Cuối cùng là pha "Dự đoán", sau khi huấn luyện với các tham số, hệ thống sẽ lưu mô hình học (là mô hình phân lớp phù hợp được chọn sau khi đánh giá) và có thể đưa ra dự đoán đối với một tập kiểm tra độc lập với tập huấn luyện. Dữ liệu người dùng với những thông số ứng với các đặc trưng ở trên khi đi qua mô hình đã được huấn luyện ở bước trước đó sẽ cho kết quả phân lớp người dùng đó thuộc lớp nào: Lớp người dùng có tiềm năng mua khóa học tiếp theo hoặc lớp người dùng không có tiềm năng mua khóa học tiếp theo.

Sau khi thu được mô hình học, ta sẽ áp dụng với dữ liệu người dùng mà không trùng với dữ liệu đã huấn luyện ở bước trước đó. Dữ liệu người dùng sẽ có các đặc trưng giống với dữ liệu huấn luyện. Dữ liệu cho việc dự đoán có thể lấy ra tương tự như việc lấy bảng dữ liệu ở bước "Trích rút dữ liệu" nhưng sẽ không đánh nhãn cho những người dùng trong dữ liệu này.

Sau đó các dữ liệu của khách hàng sẽ được đưa qua mô hình đã được huấn luyện với phương pháp SVM dưới dạng các vector đặc trưng, cùng với các Support Vectors lưu trong mô hình, nhãn của dữ liệu người dùng này sẽ được tính ra. Do chỉ cần thay mẫu dữ liệu vào phương trình siêu phẳng và kiểm tra xem lớn hơn 0 hay nhỏ hơn 0 thì ta sẽ có kết quả nhãn nên việc dự đoán của SVM là khá nhanh. Với bài toán phân lớp nhóm người dùng tiềm năng mua khóa học tiếp theo của hệ thống học trực tuyến VUIHOC, kết quả dữ liệu dự đoán sẽ là một trong hai nhãn tương ứng với hai lớp: "Người dùng có mua tiếp khóa học khác" hoặc "Người dùng không mua tiếp khóa học khác".

Khi có danh sách những người dùng có tiềm năng mua khóa học tiếp theo, chúng ta có thể kiểm chứng hiệu quả của mô hình phân lớp nhóm người dùng tiềm năng này bằng cách xem thực sự có bao nhiêu người trong danh sách dự đoán này có mua tiếp khóa học sau. Nếu kết quả của mô hình này tốt, thì có thể sử dụng trong hoạt động kinh doanh.

4. THỰC NGHIỆM

4.1. Dữ liệu thực nghiệm

Hệ thống học trực tuyến Vuihoc mới hoạt động từ đầu năm 2019, nên trong phần thử nghiệm dữ liệu huấn luyện bảng dữ liệu ứng với mốc thời gian: 2019-08-01 và dữ liệu cho việc dự đoán với mốc thời gian: 2019-10-01. Chúng tôi sẽ thử nghiệm với T_1 là khoảng thời gian 6 tháng trước mốc thời gian t, và T_2 là khoảng thời gian trong vòng 1 tháng sau mốc thời gian t. Các bảng này được trích rút từ cơ sở dữ liệu về người dùng và các khóa học của hệ thống học trực tuyến Vuihoc

Bảng 1. *Dữ liệu người dùng với mốc thời gian tháng 8 năm 2019*

Mốc thời gian	Số mẫu dữ liệu	Số mẫu nhãn 0	Số mẫu nhãn 1
2019-08-01	6508	6407	101

Từ bộ dữ liệu huấn luyện, ta sẽ học ra mô hình phân lớp nhờ những đặc trưng của người dùng, sau đó, ta thực hiện dự đoán dựa trên mô hình đó với dữ liệu tháng 10 này để dự đoán xem những người dùng nào sẽ có tiềm năng mua khóa học tiếp theo trong tháng đó.

4.2. Thử nghiệm và đánh giá

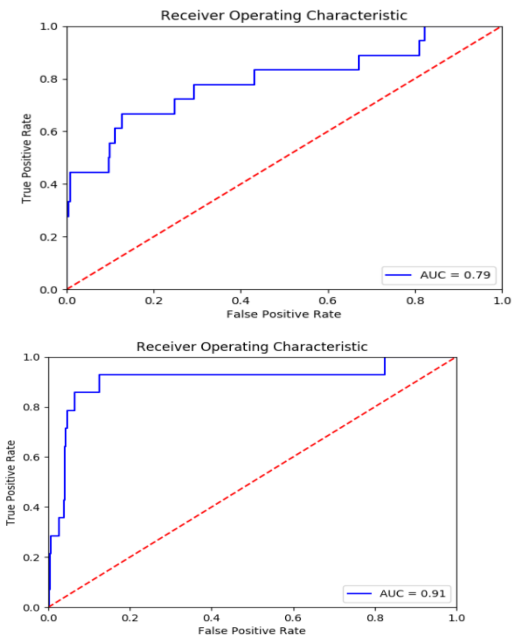
Bước 1: Tiến hành huấn luyện với tập dữ liệu người dùng tại mốc thời gian tháng 8 năm 2019 lấy được ở trên. Chúng ta sẽ thử nghiệm trên tham số smote để xem độ ảnh hưởng của nó đến kết quả của mô hình

Bước 2: Tiến hành thực hiện huấn luyện với các cặp tham số khác nhau để đánh giá kết quả. Ở đây, ta quy ước lớp C_0 là lớp người dùng không mua khóa học tiếp theo, lớp C_1 là lớp người dùng có mua tiếp khóa học tiếp theo.

Bước 3: Xem xét chung sự ảnh hưởng của tỉ lệ cân bằng smote của kỹ thuật cân bằng dữ liệu SMOTE đến mô hình huấn luyện phân lớp này. Ta sẽ xét lần lượt với smote = 0 (giữ nguyên bộ dữ liệu huấn luyện, không cân bằng lại), và tăng dần smote đến smote = 1 (cân bằng hai lớp của tỉ lệ bằng nhau). Các tham số còn lại được cố định là: C: 1.0, gamma: 0.1, kernel: rbf.

Bảng 2. *Giá trị các độ đo hiệu năng với smote = 0.0 và 0.1*

C: 1.0, gamma: 0.1, kernel: rbf						
Smote	**Lớp**	**Average accuracy**	**G-mean**	**Precision**	**Recall**	**F-score**
0.0	C_0	0.99	0.00	0.99	1.00	0.99
	C_1			0.00	0.00	0.00
0.1	C_0	0.98	0.51	0.98	1.00	0.99
	C_1			0.25	0.29	0.27

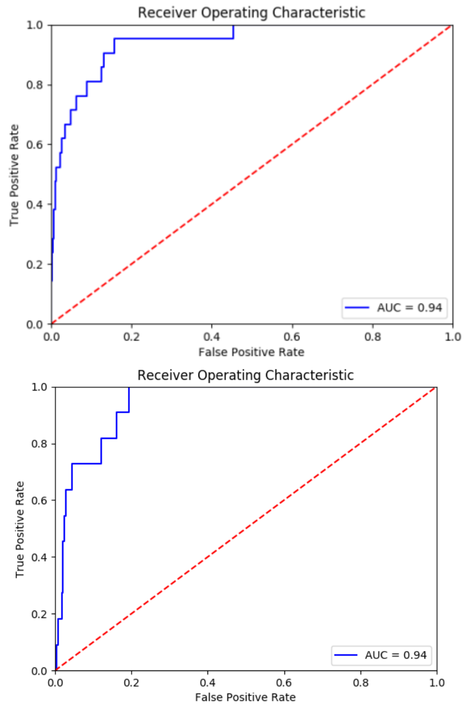


Hình 4. *Biểu đồ AUC-ROC với smote = 0.0 (trái) và = 0.1 (phải)*

Cụ thể, với smote = 0, tức là giữ nguyên dữ liệu huấn luyện ban đầu (tỷ lệ khoảng 1:67), không cân bằng dữ liệu hai lớp, mặc dù accurary có điểm số cao, nhưng mô hình phân lớp này không hiệu quả. Có thể dễ thấy điểm số này cao là do sự chênh lệch của lớp C_0 hay là lớp người dùng không mua tiếp khóa tiếp theo. Các điểm số Precision, Recall, F-score của lớp C_0 đều rất cao, Recall của lớp này là 1.0 tức là tất cả các mẫu dữ liệu lớp C_0 đều được dự đoán nhãn chính xác. Tuy nhiên, G-mean ở trường hợp này có giá trị bằng 0.0, chứng tỏ độ chính xác phân lớp của một trong hai lớp C0 hoặc C_1 là bằng 0 (ở đây là lớp C_1). Đồng thời, Recall và các độ đo khác của lớp C_1 đều bằng 0, nghĩa là không có mẫu dữ liệu nào trong lớp này được dự đoán nhãn đúng. Giá trị AUC trong biểu đồ ở trường hợp này cũng rất thấp.

Chúng ta tăng giá trị smote = 0.1, lúc này tỉ lệ giữa hai lớp được cân bằng thành 1:10. Có thể thấy, điểm số G-mean đã tăng rõ rệt, các chỉ số của lớp C_1 đã có cải thiện hơn một chút, mô hình phân lớp đã có thể dự đoán được một số dữ liệu của lớp này. Giá trị AUC ở trường hợp này cũng tăng lên, đường cong ROC cũng hướng lên góc trên bên, gần với 1 hơn, chứng tỏ mô hình đã được tốt hơn.

Ở những thử nghiệm dưới, chúng ta sẽ không xét độ đo Accuracy để tránh bị hiểu lầm về hiệu năng của mô hình.



Hình 5. *Biểu đồ AUC-ROC với smote = 0.2 (trái) và = 0.3 (phải)*

Bảng 3. *Giá trị các độ đo hiệu năng với smote = 0.2 và 0.3*

C: 1.0, gamma: 0.1, kernel: rbf					
Smote	**Lớp**	**G-mean**	**Precision**	**Recall**	**F-score**
0.2	C_0	0.71	0.98	0.99	0.98
	C_1		0.41	0.52	0.46
0.3	C_0	0.78	1.00	0.97	0.98
	C_1		0.16	0.64	0.27

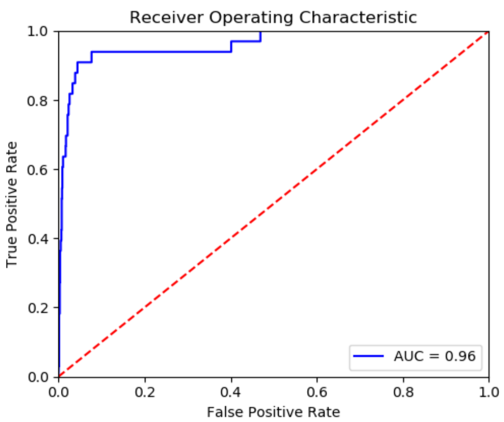
Bảng 4. *Giá trị các độ đo hiệu năng với smote = 0,4 và 0,5*

	C: 1.0, gamma: 0.1, kernel: rbf				
Smote	Lớp	G-mean	Precision	Recall	F-score
0.4	C_0	0.80	0.96	0.99	0.98
	C_1		0.26	0.72	0.38
0.5	C_0	0.84	1.00	0.95	0.97
	C_1		0.18	0.78	0.30

Tiếp tục tăng smote lên các giá trị từ 0.2 đến 0.5, qua biểu đồ AUC-ROC và các giá trị các độ đo, ta nhận thấy rằng kết quả tăng lên khá đáng kể đối với phân lớp của nhãn C_1 (lên đến 78% với smote = 0.5), mà phân lớp nhãn C_1 gần như vẫn giữ nguyên được độ hiệu quả, chỉ có giảm một chút. Độ đo G-mean trong những trường hợp này đều tăng lên khi smote tăng. Giá trị AUC cũng tăng lên, mặc dù chỉ tăng một chút, cho thấy hiệu quả của mô hình phân lớp đang dần được cải thiện, có thể phân lớp được những điểm dữ liệu của lớp C_1 (lớp thiểu số).

Bảng 5. *Giá trị các độ đo hiệu năng với smote = 0.6 và 0.7*

	C: 1.0, gamma: 0.1, kernel: rbf				
Smote	Lớp	G-mean	Precision	Recall	F-score
0.6	C0	0.85	1.00	0.93	0.96
	C1		0.18	0.79	0.30
0.7	C0	0.88	1.00	0.93	0.96
	C1		0.23	0.89	0.28



Hình 6. *Biểu đồ AUC-ROC với smote = 1.0*

Bảng 6. *Giá trị các độ đo hiệu năng với smote = 1.0*

Smote: 1.0, C: 1.0, gamma: 0.1, kernel: rbf				
Lớp	G-mean	Precision	Recall	F-score
C0	0.90	1.00	0.90	0.95
C1		0.10	0.88	0.18

Với các giá trị smote từ 0.6 đến 1.0, hiệu quả phân lớp vẫn tiếp tục tăng, tuy nhiên kết quả phân lớp của lớp C_0 cũng giảm tương đối. Độ đo AUC ở những trường hợp này cũng có hơi giảm một chút. Nguyên nhân là do giá trị smote lớn, rất nhiều các mẫu dữ liệu trong lớp C_1 được tạo ra, dẫn đến mô hình phân lớp có thể gặp hiện tượng overfitting.

Như vậy, có thể thấy việc cân bằng dữ liệu đã thực sự giúp cải thiện kết quả của mô hình huấn luyện. Chúng ta sẽ tiến hành thử nghiệm dự đoán trên tập mẫu dữ liệu người dùng của Vuihoc với mốc thời gian là tháng 10. Dữ liệu vẫn lấy theo bước "Trích rút dữ liệu" như ở trên để lấy ra các đặc trưng của từng dữ liệu người dùng trước tháng 10, nhưng sẽ không thực hiện gán nhãn cho chúng. Bộ dữ liệu này có 9576 bản ghi tương ứng với 9576 người dùng của hệ thống Vuihoc, có cả người dùng cũ và những người dùng mới đăng kí vào hệ thống. Đầu tiên, ta sẽ chọn một mô hình huấn luyện phù hợp để phân lớp. Từ thử nghiệm các mô hình phân lớp SVM với các tham số ở trên, chúng ta chọn mô hình với smote = 0.8 là trường hợp có kết quả tốt hơn so với các trường hợp khác. Thực hiện dự đoán trên bộ dữ liệu này với mô hình huấn luyện, ta được kết quả: Mô hình phân lớp dự đoán có khoảng 2860 người/9576 người có tiềm năng mua khóa học tiếp theo của hệ thống Vuihoc trong tháng 10.

Trên thực tế kiểm tra, trong tháng 10, có 37 người dùng cũ có mua thêm khóa học tiếp theo. Số lượng người mua tiếp có thể hơi ít vì tháng 10 đang là giữa kỳ học. Trong đó, với kết quả dự đoán trên, đã đoán đúng được 26 người trong số 37 người dùng kia trong khi chỉ cần xét 2860 người trên 9576 người dùng. Có thể thấy, kết quả này chưa phải là tốt nhất nhưng cũng đã nâng hiệu quả lên so với thực tế.

5. KẾT LUẬN

Nhóm tác giả đã trình bày các bước phân lớp người dùng tiềm năng của hệ thống học trực tuyến vuihoc. Kết quả chỉ ra rằng khi kết hợp phương pháp SVM và kỹ thuật cân bằng lại mẫu dữ liệu bằng SMOTE thì kết quả phân lớp người dùng tiềm năng của hệ thống học trực tuyến sẽ hiệu quả hơn nhiều so với thực tế.

TÀI LIỆU THAM KHẢO

[1]. Ling, Charles X., and Chenghui Li (1998), *Data mining for direct marketing: Problems and solutions,* Kdd. Vol. 98.

[2]. Chawla, Nitesh V.; Herrera, Francisco; Garcia, Salvador; Fernandez, Alberto, *SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary,* Journal of Artificial Intelligence Research. 61: 863-905.

[3]. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer (2002), *SMOTE: Synthetic Minority Over-sampling Technique,* arXiv preprint 1106.1813, Available at: https://arxiv.org/pdf/1106.1813.pdf.

[4]. Cortes, Corinna, Vapnik, Vladimir N.(1995), *Support-vector networks(PDF).,* Machine Learning. 20 (3): 273-297, CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018.

[5]. Boser, Bernhard E.; Guyon, Isabelle M.; Vapnik, Vladimir N., *A training algorithm for optimal margin classifiers (1992),* Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. p. 144., CiteSeerX 10.1.1.21.3818. doi:10.1145/130385.130401. ISBN 978-0897914970..

[6]. Aizerman, Mark A.; Braverman, Emmanuel M. & Rozonoer, Lev I (1964), *Theoretical foundations of the potential function method in pattern recognition learning,* Automation and Remote Control. 25: 821–837.

[7]. https://vuihoc.vn/

AUTHORS INFORMATION

Hoang Thi Ngoc Diep[1], Tran Duy Khanh[1],
Pham Huy Hoang[2], Tran Dinh Khang[2]

**Corresponding Author: hoangdiepdtth@gmail.com*

[1] Sao Do University;

[2] Hanoi University of Science and Technology.