ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÀI THỰC HÀNH SỐ 2
MÔN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

THUYẾT MINH ĐỀ TÀI ĐỒ ÁN
DỰ ĐOÁN SINH VIÊN BỊ CẢNH CÁO HỌC VỤ HAY KHÔNG

GVHD: ThS. Nguyễn Thị Anh Thư

Nhóm sinh viên thực hiện: Nhóm 8

1. Bùi Duy Anh Đức	MSSV: 20520047
2. Nguyễn Phúc Khang	MSSV: 20520569
3. Võ Trung Kiên	MSSV: 20521492
4. Chu Kim Chí	MSSV: 20521129

□□ Tp. Hồ Chí Minh, 05/2023 □□

MỤC LỤC

BẢNG PHÂN CÔNG CÔNG VIỆC

Bảng 1: Bảng phân công, đánh giá thành viên

Họ và tên	MSSV	Nội dung tìm hiểu	Đánh giá
Bùi Duy Anh Đức	20520047	- Tiền xử lý dữ liệu - Tìm hiểu các phương pháp dự đoán	Hoàn thành tốt công việc được giao
Nguyễn Phúc Khang	20520569	- Tiền xử lý dữ liệu - Tìm hiểu các phương pháp dự đoán	Hoàn thành tốt công việc được giao
Võ Trung Kiên	20521492	- Tìm hiểu nội dung đề tài, các phương pháp thực hiện	Hoàn thành tốt công việc được giao
Chu Kim Chí	20521129	- Tìm hiểu các phương pháp dự đoán	Hoàn thành tốt công việc được giao

CHƯƠNG I. GIỚI THIỆU ĐỀ TÀI

I. Tên đề tài, thời gian thực hiện

- Tên đề tài: Dự đoán sinh viên có bị cảnh cáo học vụ hay không dựa trên các mô hình phân lớp.
- Thời gian thực hiện: Từ lúc nhận được bộ dữ liệu đến nay, nhóm đã tìm hiểu về dữ liệu khoảng một tháng và quyết định chọn đề tài để tìm hiểu trong gần hai tuần.

II. Mô tả đề tài

1. Lĩnh vực bài toán

Bài toán mà nhóm thực hiện áp dụng cho lĩnh vực giáo dục, kết hợp ứng dụng học máy để xây dựng mô hình dựa trên các yếu tố liên quan để đánh giá kết quả học tập của sinh viên.

2. Khả năng ứng dụng

Kết quả mô hình nếu đạt độ chính xác tốt sẽ có thể áp dụng tại các trường đại học, cũng như các bậc học khác.

3. Các vấn đề liên quan cần tìm hiểu

Đối với đề tài này, nhóm có các nội dung chính cần tìm hiểu:

- Mô hình phân lớp: Hiện nay có nhiều mô hình, cần tìm hiểu để áp dụng và so sánh các mô hình.
- Cách xử lý dữ liệu: Tìm hiểu về cách biến đổi các giá trị trong dữ liệu cho phù hợp với mô hình.
- Các yếu tố ảnh hưởng đến kết quả học tập của sinh viên: Đây là một đề tài đã có khá nhiều công trình nghiên cứu, nhóm đã tham khảo và chọn lọc cho phù hợp với dữ liệu được cung cấp.

III. Tổng quan

1. Ý tưởng

Ý tưởng chính mà nhóm thực hiện đó là áp dụng các mô hình máy học để phân loại sinh viên vào hai lớp: bị cảnh cáo học vụ hoặc không bị cảnh cáo học vụ và chọn ra mô hình tốt nhất. Tuy nhiên, trước đó một điều quan trọng là phải tìm hiểu các yếu tố có thể dùng cho bài toán.

2. Tính cấp thiết, tính mới

Việc dự đoán một sinh viên có bị cảnh cáo học vụ hay không là một vấn đề có tính cấp thiết, không chỉ giúp sinh viên được hỗ trợ kịp thời, mà còn giúp nhà trường cải thiện chất lượng giáo dục, nâng cao thành tích. Đây không phải là một chủ đề mới, vì đã có khá nhiều công trình nghiên cứu. Tuy nhiên các yếu tố áp dụng vào bài toán đã được nhóm tìm hiểu cho phù hợp với dữ liệu.

3. Giới thiệu bài toán

- Input: Giới tính, nơi ở, khóa học, khoa, học kỳ và năm học hiện tại, xếp loại dựa trên điểm trúng tuyển, xếp loại điểm anh văn đầu vào, xếp loại điểm trung bình và điểm rèn luyện của học kỳ trước, số tín chỉ đăng ký trong học kỳ hiện tại.
- Output: Bị cảnh cáo học vụ hay không (trong dữ liệu huấn luyện, bị cảnh cáo học vụ khi có điểm trung bình học kỳ dưới 3).

IV. Mục tiêu đề tài

Dự đoán kết quả với độ chính xác cao, càng nhiều thông tin thì mô hình dự đoán càng tốt.

V. Phạm vi thực hiện

- Lý thuyết liên quan, kỹ thuật cần dùng: Kiến thức về bài toán phân lớp trong máy học, cách đánh giá kết quả mô hình.

- Bộ dữ liệu: Nhóm thực hiện đề tài môn học trên bộ dữ liệu được giảng viên và trường Đại học Công nghệ thông tin – ĐHQG TPHCM cung cấp.

CHƯƠNG II. NỘI DUNG VÀ PHƯƠNG PHÁP THỰC HIỆN

I. Khảo sát tổng quan

1. Các công trình nghiên cứu, lý thuyết liên quan

Các công trình nghiên cứu liên quan mà nhóm tìm hiểu: "*Dự đoán kết quả học tập của sinh viên bằng kỹ thuật khai phá dữ liệu (Trường Đại học Vinh, 2019)"*, "*Các yếu tố ảnh hưởng đến kết quả học tập của sinh viên hệ chính quy tại trường Đại học Kinh tế, Đại học Huế*". Từ các nguồn tham khảo, có thể thấy các yếu tố thường được dùng để đánh giá là giới tính, nơi thường trú của sinh viên, điểm trúng tuyển và điểm trung bình của học kỳ trước, khoa học, khóa học.

2. Mô hình giải bài toán phân lớp

- Tiền xử lý dữ liệu.
- Chia dữ liệu train, test.
- Chọn và huấn luyện mô hình.
- Đánh giá mô hình.
- Tinh chỉnh mô hình.

II. Mô tả bài toán

- Input:
    - Các thông tin chung của sinh viên:
        - Giới tính.
        - Nơi thường trú (nhóm đang xây dựng dữ liệu theo 7 vùng kinh tế của Việt Nam hoặc là giá trị nhị phân chứa thông tin thuộc địa phận thành phố Hồ Chí Minh hoặc không).
        - Xếp loại trúng tuyển (dựa trên điểm và phương thức trúng tuyển).
        - Xếp loại anh văn đầu vào (dựa vào các mức xếp lớp anh văn của trường).
    - Các thông tin về học tập của sinh viên tại trường:

- Học kỳ hiện tại.
- Số tín chỉ đăng ký.
- Xếp loại học lực ở học kì trước.
- Xếp loại điểm rèn luyện ở học kì trước.

- Output: Cho biết sinh viên có bị cảnh cáo học vụ không.

Trong đó, các thông tin về điểm học tập, điểm rèn luyện được quy đổi về các mức đánh giá xếp loại (0, 1, 2, 3, 4,…).

III. Cài đặt thực nghiệm

1. Dataset được cung cấp

Nhóm đã xử lý bộ dữ liệu được cung cấp để chỉ giữ lại những thông tin cần thiết cho đề tài, cũng như biến đổi các giá trị cho phù hợp. Tuy nhiên, do số lượng mẫu bị cảnh cáo thấp hơn nhiều so với mẫu không bị cảnh cáo, nên cần phải áp dụng thêm các kỹ thuật cân bằng dữ liệu trước khi đưa vào mô hình phân lớp.

	gioitinh	xl_tt	khuvuc_1.0	khuvuc_2.0	khuvuc_3.0	khuvuc_4.0	khuvuc_5.0	khuvuc_6.0	khuvuc_7.0	xl_av	sotchk	hocky_sx	xl_drlhk_truoc	xlhk_truoc
22017	1.0	0.0	0	0	0	1	0	0	0	0.0	6	9	5.0	3
9047	1.0	2.0	0	0	0	0	0	0	1	1.0	23	5	3.0	3
5729	1.0	2.0	0	0	0	0	0	1	0	0.0	10	9	4.0	4
9779	1.0	3.0	0	0	1	0	0	0	0	1.0	20	3	3.0	3
1185	1.0	1.0	0	0	0	0	0	1	0	0.0	17	7	3.0	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
31496	1.0	1.0	0	0	0	1	0	0	0	0.0	14	6	3.0	2
31537	1.0	2.0	0	0	0	0	0	0	1	0.0	21	6	4.0	2
31600	1.0	2.0	0	0	1	0	0	0	0	0.0	6	4	3.0	2
31627	1.0	3.0	0	0	0	1	0	0	0	0.0	14	7	3.0	3
31636	1.0	1.0	0	0	1	0	0	0	0	0.0	16	7	2.0	1

3102 rows × 14 columns

2. Các phương pháp đánh giá

- Độ đo đánh giá: accuracy, precision, recall, F1-score
- So sánh các mô hình.

IV. Kết luận và hướng phát triển

Bên cạnh các yếu tố về số liệu, kết quả học tập của sinh viên còn phù thuộc vào nhiều yếu tố khác như sức khỏe, tâm lý,… không thể đo đạc bằng số liệu.

CHƯƠNG III. TÀI LIỆU THAM KHẢO

1. Nguyễn Thị Uyên, Nguyễn Minh Tâm, "DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN BẰNG KỸ THUẬT KHAI PHÁ DỮ LIỆU", 12/9/2019, [Trực tuyến]. Địa chỉ: https://vjol.info.vn/index.php/vinhuni/article/download/47819/38792/.
2. Nguyễn Mạnh Hùng*, Hoàng Thị Kim Thoa, Nguyễn Thanh Thiện, Phan Thị Bích Hạnh, "CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN HỆ CHÍNH QUY TẠI TRƯỜNG ĐẠI HỌC KINH TẾ, ĐẠI HỌC HUẾ", 2020, [Trực tuyến]. Địa chỉ: View of PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN HỆ CHÍNH QUY TẠI TRƯỜNG ĐẠI HỌC KINH TẾ, ĐẠI HỌC HUẾ | Hue University Journal of Science: Social Sciences and Humanities.