ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

ĐỀ TÀI

CÁC YẾU TỐ ẢNH HƯỞNG TỚI KẾT QUẢ TIẾNG ANH QUÁ TRÌNH CỦA SINH VIÊN

Giảng viên hướng dẫn: ThS. Nguyễn Thị Anh Thư

Nhóm 7 - Sinh viên thực hiện:

Mai Duy Ngọc	20520654
Trần Đăng Khoa	20520589
Đào Danh Đăng Phụng	20520699
Đặng Phước Sang	21521377

TP. Hồ Chí Minh, tháng 5 năm 2023

MỤC LỤC

CHƯƠNG 1: PHẦN MỞ ĐẦU

1.1 Sơ lược đề tài

Tên đề tài: "Các yếu tố ảnh hưởng tới kết quả tiếng Anh quá trình của sinh viên".

Thời gian thực hiện: 3 tháng

1.2 Mô tả đề tài

Trong thời đại toàn cầu hóa ngày nay, tiếng Anh đã trở thành một yêu cầu tối thiểu đối với nhiều ngành nghề nói chung và ngành liên quan tới công nghệ thông tin nói riêng. Tuy nhiên, không phải tất cả sinh viên đều đạt được kết quả tiếng Anh chuẩn quá trình đúng như mong đợi. Điều này có thể phụ thuộc vào nhiều yếu tố ảnh hưởng khác nhau.

Ngoại trừ các yếu tố chủ quan như sự kiên trì, cố gắng, khả năng tiếp thu, … của một sinh viên thì các yếu tố khác cũng sẽ góp phần không nhỏ vào việc đánh giá trình độ tiếng Anh đầu ra của sinh viên như giới tính, năm sinh, … hoặc những yếu tố trước khi nhập học như trường THPT đã học, hình thức xét tuyển vào trường, … và các yếu tố sau khi nhập học bao gồm ngành xét tuyển, hệ đào tạo, điểm tiếng Anh các học kì, … Thông qua đó, chúng ta có thể kiểm tra sinh viên có đáp ứng yêu cầu tiếng Anh quá trình. Từ đó có thể đưa lộ trình hoặc nhắc nhở sinh viên từ sớm.

1.3 Ý tưởng

Ý tưởng của nhóm là sẽ thông qua các dữ liệu có sẵn của sinh viên UIT để lọc ra những yếu tố mà nhóm cho rằng là có thể để đánh giá tiếng Anh của sinh viên có đạt đủ chuẩn quá trình. Từ đó, nhóm sẽ xây dựng một mô hình máy học phân lớp để phân định các trình độ tiếng Anh của sinh viên vào năm cuối xem có đáp ứng yêu cầu hay chưa.

Với việc nhiều sinh viên ỷ lại trong quá trình rèn luyện tiếng Anh, cho rằng năm 1, năm 2 còn quá sớm để cải thiện khả năng ngoại ngữ của bản thân thì việc có thể đánh giá sớm khả năng tiếng Anh của sinh viên góp phần răn đe cũng như nhắc nhở sinh viên trong công cuộc xây dựng vốn ngoại ngữ, từ đó giúp nhà trường đưa ra lộ trình

hợp lí và giúp sinh viên có thời gian nhìn lại bản thân để cải thiện tốt hơn để tránh bị cảnh cáo học vụ.

1.4 Mục tiêu

Mục tiêu của đề tài sẽ tập trung vào việc xác định sinh viên có đạt đủ yêu cầu tiếng Anh quá trình thông qua việc phân lớp sinh viên vào hai lớp "đạt chuẩn" và "không đạt chuẩn".

Các yếu tố mà nhóm sẽ sử dụng bao gồm: giới tính, năm sinh, hình thức xét tuyển, trường THPT, khoa ứng tuyển, hệ đào tạo, xếp loại AV đầu vào và xếp loại AV gần nhất. Các yếu tố sẽ được khai thác trong bộ dữ liệu được giảng viên cung cấp sẵn trong các bảng "sinh viên", "sinh viên và chứng chỉ", "xếp loại av" và "thí sinh" và "điểm". Các bảng sẽ được xử lí để cung cấp dữ liệu cho mô hình máy học để phân lớp sinh viên đạt chuẩn tiếng Anh ở mức tốt nhất có thể.

CHƯƠNG 2: NỘI DUNG THỰC HIỆN

2.1 Tổng quát

Đề tài nhóm thực hiện "Các yếu tố ảnh hưởng tới kết quả tiếng Anh quá trình của sinh viên" sẽ là một bài toán phân lớp nhị phân với input đầu vào là thông tin của sinh viên về: giới tính, năm sinh, hình thức xét tuyển, trường THPT, ngành ứng tuyển, hệ đào tạo, xếp loại AV đầu vào và xếp loại AV gần nhất. Output đầu ra sẽ là gán nhãn cho sinh viên đó "đạt tiêu chuẩn" hoặc "không đạt tiêu chuẩn".

Đối tượng được đánh giá sẽ là sinh viên đang theo học tại UIT, mục tiêu là đánh giá sinh viên trước khi tốt nghiệp có đủ điều kiện tiếng Anh quá trình hay không. Thách thức bài toán đề ra là việc nhóm phải xử lí dữ liệu tốt do từng năm học khác nhau, trường sẽ có các xếp loại AV khác nhau. Việc có nhiều thông tin input đòi hỏi nhóm phải lựa chọn hướng giải quyết cũng như phương pháp phù hợp để có thể tối ưu lượng thông tin đồng thời tối ưu khả năng đánh giá của mô hình.

2.2 Mô hình giải bài toán

2.2.1 Tiền xử lí dữ liệu

Để có dữ liệu đúng với mong muốn, nhóm đã tìm hiểu dữ liệu và lựa chọn ra các bảng dữ liệu phù hợp với mục đích của nhóm là bảng "sinh viên", "sinh viên và chứng chỉ", "xếp loại av" và "thí sinh" và "điểm".

Sau đó sẽ tiến hành bước làm sạch dữ liệu các bảng để có thể sử dụng dữ liệu.

2.2.1.1 Sinh viên

Xử lý dữ liệu trùng lắp: không có dữ liệu trùng.

Đưa dữ liệu về cùng một định dạng thống nhất:

- Cột id, năm sinh, giới tính, lớp sinh hoạt, hệ đào tạo, khoa, khóa học, chuyên ngành 2, tình trạng: đã cùng định dạng.
- Cột nơi sinh: các điểm dữ liệu chưa thống nhất về định dạng (các giá trị viết hoa, viết thường, có dấu ngoặc kép, viết tắt, …).

```
df[' noisinh'].unique()
```

```
array([' TP. Hồ Chí Minh', ' Đồng Tháp', ' Hà Nam Ninh',
       ' Thành phố Hồ Chí Minh', ' Hà Tĩnh', ' Thanh Hóa', ' Quảng Ngãi',
       ' Khánh Hoà', ' Cần Thơ', ' Gia Lai', ' Tiền Giang', ' Vĩnh Long',
       ' Sông Bé', ' Kiên Giang', ' Lâm Đồng', ' Hải Dương', ' Trà Vinh',
       ' Nam Định', ' Quảng Nam', ' Bình Thuận', ' Quảng Bình',
       ' Bến Tre', ' Ninh Thuận', ' Bình Định', ' An Giang',
       ' Bà Rịa - Vũng Tàu', ' Thừa Thiên Huế', ' Kon Tum', ' Tây Ninh',
       ' Hà Nội', ' Đồng Nai', ' Thanh Hoá', ' Nghệ An', ' NULL',
       ' Phú Yên', ' Đăk Lăk', ' Tp Hồ Chí Minh', ' Đắc Nông',
       ' Bình Dương', ' Khánh Hòa', ' Long An', ' Thái Bình',
       ' Bình Phước', ' Hà Sơn Bình', ' Đà Nẵng', ' Ninh Bình', ' Cà Mau',
       ' Quảng Trị', ' Hải Hưng', ' Hưng Yên', ' Hà Tây',
       ' Thành phố Cần Thơ', ' Đắk Lắk', ' Tỉnh Đồng Nai', ' Đắk Nông',
       ' Bắc Giang', ' Thành phố Đà Nẵng', ' Hải Phòng',
       ' Tỉnh Bình Dương', ' Minh Hải', ' Sóc Trăng', ' Hà Nam',
       ' Hậu Giang', " 'Tây Ninh'", " 'Tiền Giang'", " 'Đồng Nai'",
       " 'Bình Định'", " 'Đồng Tháp'", " 'Phú Yên'", " 'TP. Hồ Chí Minh'",
       " 'Quảng Ngãi'", " 'Quảng Nam'", " 'Bình Dương'", " 'Đăk Lăk'",
       " 'Gia Lai'", " 'Bến Tre'", " 'Nghệ An'", " 'Bà Rịa - Vũng Tàu'",
       " 'Ninh Bình'", " 'An Giang'", " 'Lâm Đồng'", " 'Khánh Hoà'",
       " 'Thành phố Hồ Chí Minh'", " 'Quảng Bình'", " 'Thừa Thiên Huế'",
       " 'Bình Thuận'", " 'Long An'", " 'Ninh Thuận'", " 'Bắc Cạn'",
```

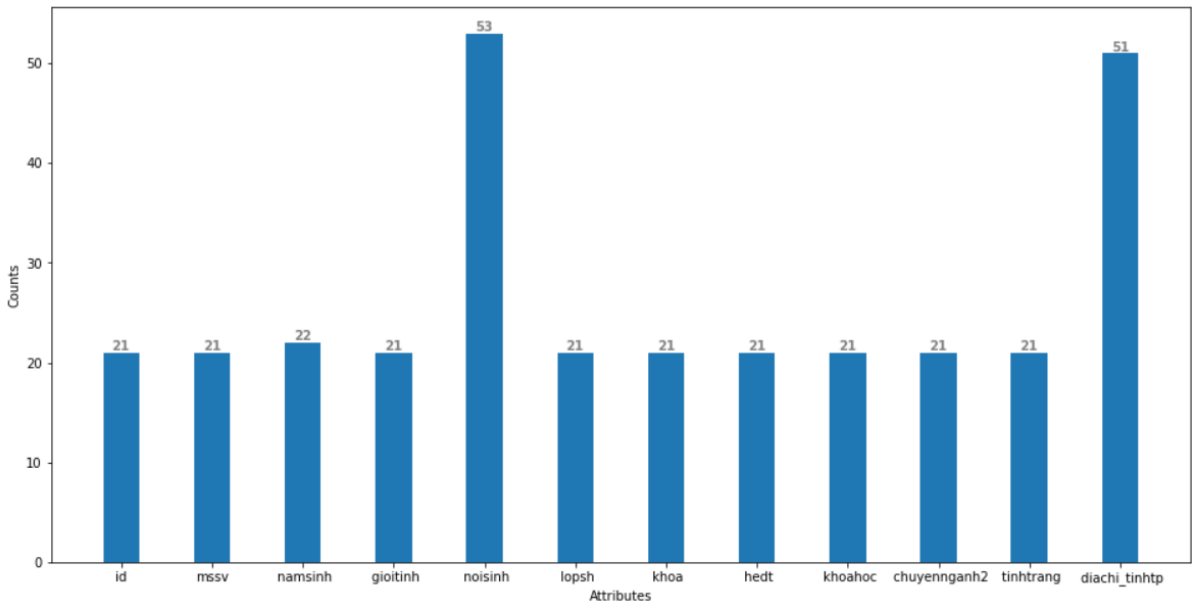
Nhóm nghiên cứu sẽ đưa tất cả dữ liệu về cùng một định dạng: viết hoa chữ cái đầu, không viết tắt, không dấu ngoặc và chỉ gồm tên tỉnh/thành phố.

```
In [56]: df_final[' noisinh'].unique()
Out[56]: array([' Hồ Chí Minh', ' Đồng Tháp', ' Hà Nam Ninh', ' Hà Tĩnh',
       ' Thanh Hóa', ' Quảng Ngãi', ' Khánh Hoà', ' Cần Thơ', ' Gia Lai',
       ' Tiền Giang', ' Vĩnh Long', ' Sông Bé', ' Kiên Giang',
       ' Lâm Đồng', ' Hải Dương', ' Trà Vinh', ' Nam Định', ' Quảng Nam',
       ' Bình Thuận', ' Quảng Bình', ' Bến Tre', ' Ninh Thuận',
       ' Bình Định', ' An Giang', ' Bà Rịa - Vũng Tàu', ' Thừa Thiên Huế',
       ' Kon Tum', ' Tây Ninh', ' Hà Nội', ' Đồng Nai', ' Nghệ An', ' ',
       ' Phú Yên', ' Đắk Lắk', ' Đắk Nông', ' Bình Dương', ' Long An',
       ' Thái Bình', ' Bình Phước', ' Hà Sơn Bình', ' Đà Nẵng',
       ' Ninh Bình', ' Cà Mau', ' Quảng Trị', ' Hải Hưng', ' Hưng Yên',
       ' Hà Tây', '  Đồng Nai', ' Bắc Giang', ' Hải Phòng',
       '  Bình Dương', ' Minh Hải', ' Sóc Trăng', ' Hà Nam', ' Hậu Giang',
       ' Bắc Cạn', ' Vĩnh Phúc', ' Bạc Liêu', '  Tây Ninh', ' Campuchia',
       ' Bắc Ninh', ' Liên Bang Nga', ' Nha Trang', 'Nam Định',
       ' Quảng Ninh', ' Phú Thọ', ' Lạng Sơn', ' Cộng hoà Séc',
       ' Yên Bái', '  Đắk Lắk', ' Gia lai', ' Đà Lạt', '  Ninh Thuận',
       ' Bình Phước ', '  Thanh Hóa', ' Cao Bằng', ' Trà vinh',
       '  Nghệ An', ' Ninh Thuận ', ' Vũng Tàu', '', ' Hà Nam ',
       ' Australia', ' Quảng BÌnh', ' Hòa Bình', ' Bình phước ',
       ' Lai Châu', ' Bình phước', ' An giang', ' Buôn Ma Thuột',
       ' Tuyên Quang', ' Thái Nguyên', 'Long Hồ'], dtype=object)
```

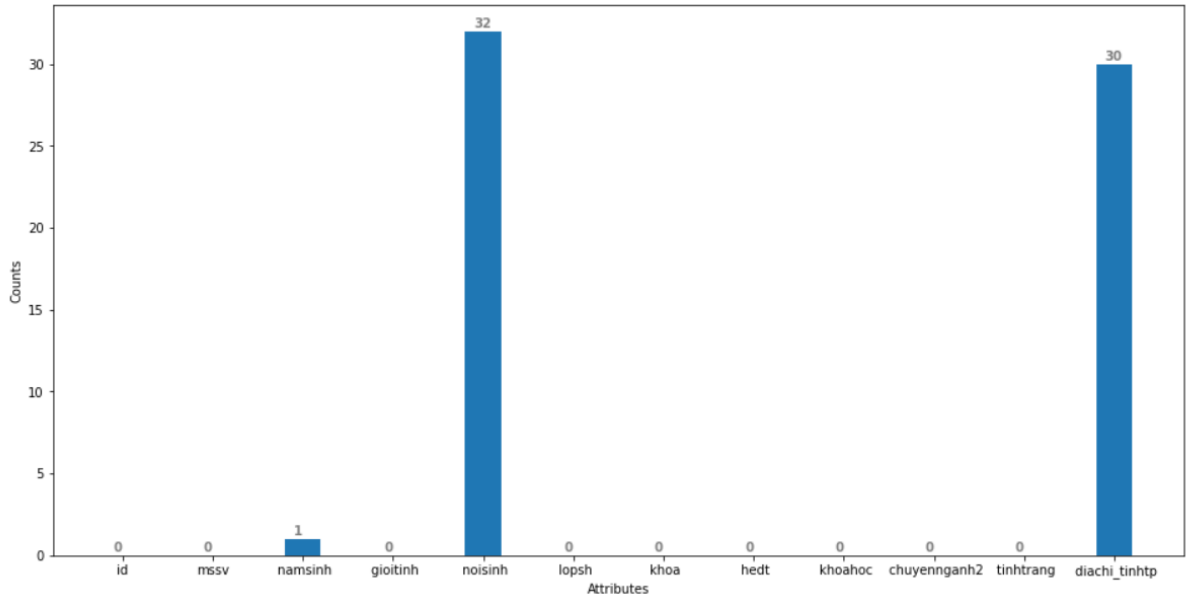
- Cột địa chỉ tỉnh-thành phố gồm các dữ liệu chưa thống nhất về định dạng. Thực hiện biến đổi đối với cột này bằng cách viết hoa chữ cái đầu của mỗi từ, xóa các dấu ngoặc.

```
In [72]: df_final[' diachi_tinhtp'].unique()
Out[72]: array([' Thành Phố Hồ Chí Minh', ' Huyện Hóc Môn', ' Tỉnh Hà Nam',
       ' Tỉnh Hà Tĩnh', ' Tỉnh Thanh Hóa', ' Tỉnh Quảng Ngãi', ' Quận 12',
       ' Tỉnh Khánh Hoà', ' Thành Phố Cần Thơ', ' Thị Xã An Khê',
       ' Tỉnh Tiền Giang', ' Huyện Bình Tân', ' Quận Thủ Đức',
       ' Phường Phú Cường', ' Tỉnh Lâm Đồng', ' Huyện Di Linh',
       ' Tỉnh Kiên Giang', ' Tỉnh Trà Vinh', ' Huyện Đức Hoà',
       ' Thành Phố Hải Phòng', ' Tỉnh Quảng Nam', ' Huyện Kinh Môn',
       ' Tỉnh Bình Thuận', ' Tỉnh Quảng Bình', ' Tỉnh Quảng Ngãi',
       ' Thành Phố Tây Ninh', ' Thành Phố Sa Đéc', ' Tỉnh Ninh Thuận',
       ' Tỉnh Gia Lai', ' Tỉnh Vĩnh Phúc', ' Tỉnh An Giang',
       ' Thành Phố Bà Rịa', ' Tỉnh Thừa Thiên Huế', ' Huyện Kiên Lương',
       ' Tỉnh Kon Tum', ' Tỉnh Vĩnh Long', ' Tỉnh Tây Ninh',
```

Xử lý dữ liệu trống: nhóm thực hiện thống kê số lượng dữ liệu trống của tất cả các thuộc tính.



Do mã số sinh viên thuộc tính rất quan trọng đóng vai trò khóa chính, vì thế, ta sẽ xóa tất cả các dòng dữ liệu bị thiếu thuộc tính này.



Với giá trị năm sinh bị rỗng, ta sẽ điền vào đó là giá trị năm sinh trung bình của các nhóm đối tượng trong chung lớp sinh hoạt MTLK2015 là 1990.

id	mssv	namsinh	gioitinh	noisinh	lopsh	khoa	hedt	khoahoc	chuyennganh2	tinhtrang
10561	FEDA6C86XPvAibaEXe/TvTjLTyzflhOWf9+W8he6	1995	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10562	88A832B4XPvAibaEXe/TvTjLTyzfljbL/KBUI/Dl	1995	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10563	A123BBE3XPvAibaEXe/TvTjLTyzflj6RjUEvimtW	1982	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10564	9363FC62XPvAibaEXe/TvTjLTyzflrcciunSlsad	1979	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10565	B2E95A9FXPvAibaEXe/TvTjLTyzflg1BIejC6JWj	1994	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10566	F53C9805XPvAibaEXe/TvTjLTyzfljpqpBx0oeCr	1979	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10567	C0F22F51XPvAibaEXe/TvTjLTyzflmKWmieH/MIJ	1991	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10568	BB032AC5XPvAibaEXe/TvTjLTyzfllUW7coWUPgk	1993	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10569	86B4040FXPvAibaEXe/TvTjLTyzflvYZvkl6yT4t	1993	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10570	C407DC7EXPvAibaEXe9z+h7aUGr+1B/Sn7nH+3Jd	1991	0		MTLK2015	MMT&TT	CQUI	10	D480299	11
10571	AB1E292EXPvAibaEXe9z+h7aUGr+1Pa1vxccdu6I	1992	1		MTLK2015	MMT&TT	CQUI	10	D480299	11
10572	5EE0C76EXPvAibaEXe9z+h7aUGr+1HViw3dtd1q		1		MTLK2015	MMT&TT	CQUI	10	D480299	11

Với đối tượng nơi sinh và địa chỉ tỉnh/thành phố, ta sẽ điền giá trị xuất hiện nhiều nhất đó là Hồ Chí Minh.

```
df_final[' noisinh'].value_counts()
```

```
Hồ Chí Minh     1633
Đồng Nai         525
Đăk Lăk          504
Bình Định        422
Lâm Đồng         354
                ...
Cộng hoà Séc       1
Hải Hưng           1
Nha Trang          1
Campuchia          1
Long Hồ            1
```

2.2.1.1 Sinh viên và chứng chỉ

Xử lý dữ liệu trùng lắp: 24 điểm dữ liệu trùng. Thực hiện xóa các điểm dữ liệu này.

Đưa dữ liệu về cùng một định dạng thống nhất:

- Cột ngày thi: đưa kiểu dữ liệu về dạng date (dd/mm/yy).
- Thực hiện xóa các dấu nháy đơn bị lỗi ở các cột.
- Bắt đầu từ cột url1 trở về sau, giá trị dữ liệu bị điền nhầm lẫn giữa các cột, giá trị dữ liệu đúng là cột url1 -> listening, loaixn2 -> speaking, listening -> reading, speaking -> writing… Nhóm sẽ thực hiện đưa dữ liệu về đúng cột đồng thời, xóa cột url1 và loaixn2 vì không chứa bất kỳ thông tin nào.
- Riêng cột listening, speaking, reading, writing nhóm sẽ đưa về kiểu dữ liệu số. Đối với cột tổng điểm, chưa có sự thống nhất về mặt ý nghĩa khi với loại chứng chỉ tiếng Nhật có điểm dữ liệu số có điểm dữ liệu chữ (N3, N4), nhóm sẽ quy đổi N3 thành 95, N4 thành 90. Ngoài ra, có một điểm dữ liệu vô lí đó chính là tổng điểm chứng chỉ IELTS = 29.5. Nhóm sẽ thực hiện xóa giá trị này.
- Ngoài ra, còn một số trường hợp lỗi ở cột lý do dẫn đến dữ liệu bị đẩy sang cột khác.

E	F	G	H	I	J	K	L	M	N
loaixn	listening	speakin	reading	writing	tongdien	lydo	trangthai	ngayxl	Column1
VNU-EPT						Miễn AV2	3 và Đạt chuẩn đầu ra	1	2/2/2018 9:46
IELTS					6	Sinh viên vui lòng up lại chứng chỉ	hiện tại hình chụp không rõ	-1	7/31/2020 14:50
TOEIC_LR	270		230		500	Chứng chỉ đã làm lại kết quả điểm	điểm chính xác là 240.	-1	7/7/2020 9:22
TOEIC_LR	185		160		345	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	2/24/2021 9:53
TOEIC_LR	235		185		420	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	2/24/2021 9:53
TOEIC_LR	190		130		320	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	2/24/2021 9:54
TOEIC_LR	285		185		470	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	2/24/2021 9:54
TOEIC_LR	230		200		430	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	2/24/2021 9:56
TOEIC_LR	170		165		335	Chứng chỉ Đánh giá năng lực	sv được miễn AVSC	-1	3/11/2021 8:20
TOEIC_LR	200		220		420	Chứng chỉ ĐGNL	không phảiTOEIC	SV đã được miễn AV1	
VNU-EPT						Miễn AV2	3 và Đạt chuẩn đầu ra	1	
IELTS					6	Sinh viên vui lòng up lại chứng chỉ	hiện tại hình chụp không rõ	-1	
TOEIC_LR	270		230		500	Chứng chỉ đã làm lại kết quả điểm	điểm chính xác là 240.	-1	
TOEIC_LR	185		160		345	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	
TOEIC_LR	235		185		420	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	
TOEIC_LR	190		130		320	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	
TOEIC_LR	285		185		470	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	
TOEIC_LR	230		200		430	Chứng chỉ này là chứng chỉ Đánh giá năng lực	sinh viên vui lòng chọn lại lc	-1	
TOEIC_LR	170		165		335	Chứng chỉ Đánh giá năng lực	sv được miễn AVSC	-1	
TOEIC_LR	200		220		420	Chứng chỉ ĐGNL	không phảiTOEIC	SV đã được miễn AV1	

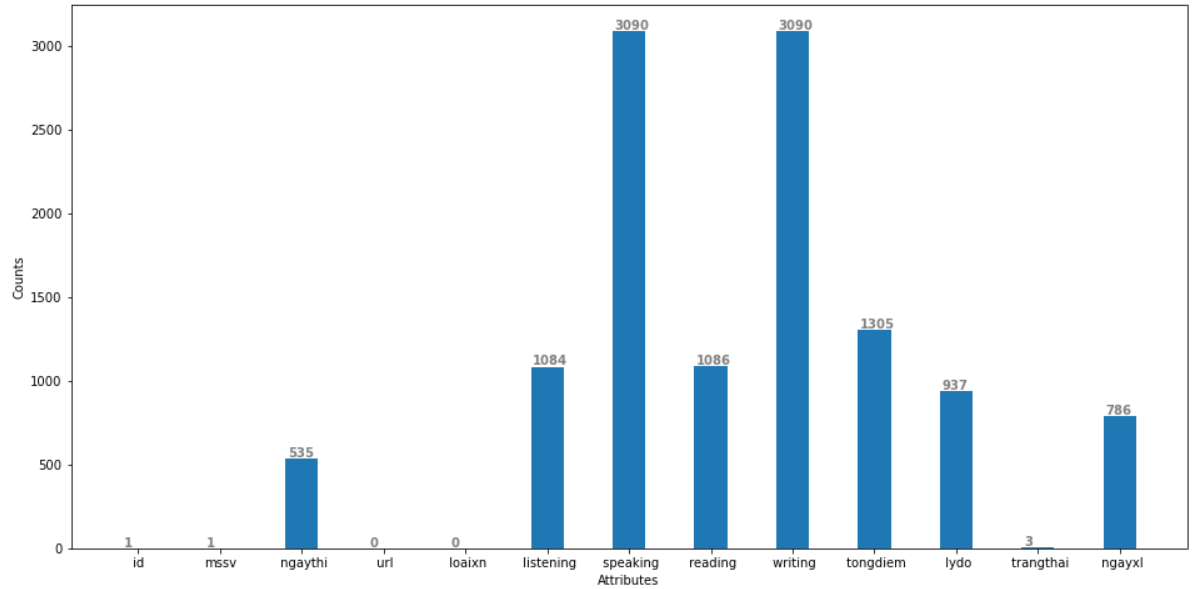
Trước khi chỉnh sửa

E	F	G	H	I	J	K	L	M	N
loaixn	listening	speaking	reading	writing	tongdiem	lydo	trangthai	ngayxl	Column1
VNU-EPT						Miễn AV2, 3 và Đạt chuẩn đầu ra	1	2/2/2018 9:46	
IELTS					6	Sinh viên vui lòng up lại chứng chỉ, hiện tại hình chụp không rõ nộ	-1	7/31/2020 14:50	
TOEIC_LR	270		230		500	Chứng chỉ đã làm lại kết quả điểm, điểm chính xác là 240.	-1	7/7/2020 9:22	
TOEIC_LR	185		160		345	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1	2/24/2021 9:53	
TOEIC_LR	235		185		420	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1	2/24/2021 9:53	
TOEIC_LR	190		130		320	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1	2/24/2021 9:54	
TOEIC_LR	285		185		470	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1	2/24/2021 9:54	
TOEIC_LR	230		200		430	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1	2/24/2021 9:56	
TOEIC_LR	170		165		335	Chứng chỉ Đánh giá năng lực, sv được miễn AVSC	-1	3/11/2021 8:20	
TOEIC_LR	200		220		420	Chứng chỉ ĐGNL, không phảiTOEIC, SV đã được miễn AV1			
VNU-EPT						Miễn AV2, 3 và Đạt chuẩn đầu ra	1		
IELTS					6	Sinh viên vui lòng up lại chứng chỉ, hiện tại hình chụp không rõ n	-1		
TOEIC_LR	270		230		500	Chứng chỉ đã làm lại kết quả điểm, điểm chính xác là 240.	-1		
TOEIC_LR	185		160		345	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1		
TOEIC_LR	235		185		420	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1		
TOEIC_LR	190		130		320	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1		
TOEIC_LR	285		185		470	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1		
TOEIC_LR	230		200		430	Chứng chỉ này là chứng chỉ Đánh giá năng lực, sinh viên vui lòng	-1		
TOEIC_LR	170		165		335	Chứng chỉ Đánh giá năng lực, sv được miễn AVSC	-1		
TOEIC_LR	200		220		420	Chứng chỉ ĐGNL, không phảiTOEIC, SV đã được miễn AV1			

Sau khi chỉnh sửa

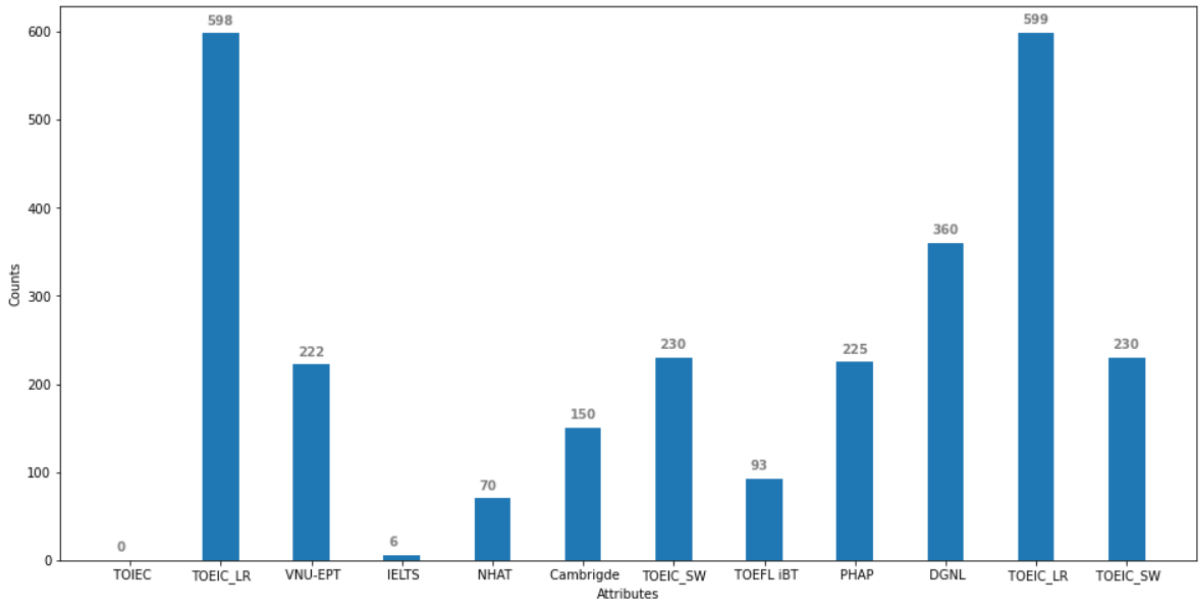
- Sau khi đã đưa đúng dữ liệu về các cột, ta sẽ xóa cột Column1.
- Cột ngày xử lý đưa về dạng dữ liệu date (dd/mm/yy hh:mm:ss)

<u>Xử lý dữ liệu trống:</u> nhóm thực hiện thống kê số lượng dữ liệu trống của tất cả các thuộc tính.



- Thực hiện xóa các điểm dữ liệu thiếu mã số sinh viên.
- Với cột tổng điểm, có 2 trường hợp: thiếu điểm thành phần và có điểm thành phần
    - Với trường hợp có điểm thành phần, ta sẽ cộng các giá trị này lại thành tổng điểm.

- Với trường hợp thiếu điểm, ta sẽ lấy trung bình tổng điểm của từng loại chứng chỉ.
- Riêng chứng chỉ TOEIC không có điểm trung bình, vậy có nghĩa là không có điểm dữ liệu nào có giá trị điểm. Tuy nhiên, nhóm chứng chỉ này có một điểm chung đó là ngày xử lý đều vào năm 2017 và trạng thái là 1 (đã được duyệt). Điều này thể hiện khả năng chứng chỉ TOEIC = TOEIC_LR (ở những khóa trước chưa có TOEIC_SW). Do đó, nhóm sẽ điền bằng giá trị TOEIC_LR.



Tổng điểm trung bình

- Đối với 4 cột listening, speaking, reading, writing ta sẽ dựa vào loại chứng chỉ để điền vào chỗ trống:
  - Cambrigde: speaking = writing = reading = listening = tổng điểm
  - TOEIC_LR, TOEIC, DGNL: speaking = 0, writing = 0, listening = reading = tổng điểm / 2.
  - TOEIC_SW: listening = 0, reading = 0, speaking = writing = tổng điểm / 2.
  - Đối với NHAT, speaking = writing = 0, reading = 2/3 tổng điểm.
  - Đối với PHAP, speaking = 0, writing = listening = reading = 1/3 tổng điểm
  - Đối với các cột còn lại, speaking = writing = reading = listening = 1/4 tổng điểm.

- Đối với cột lý do, nếu dữ liệu đó trống thì sẽ được điền là không lý do.
- Đối với cột trạng thái, ta sẽ kiểm tra xem cột lý do để điền vào chỗ trống.
- Đối với cột ngày thi và ngày xử lý, nhóm sẽ điền một giá trị không có thực đó là 0 (tương đương 0/1/1990).

2.2.1.3 Xếp loại anh văn

Xử lý dữ liệu trùng lắp: 5 dữ liệu trùng. Xóa các dòng dữ liệu trùng.

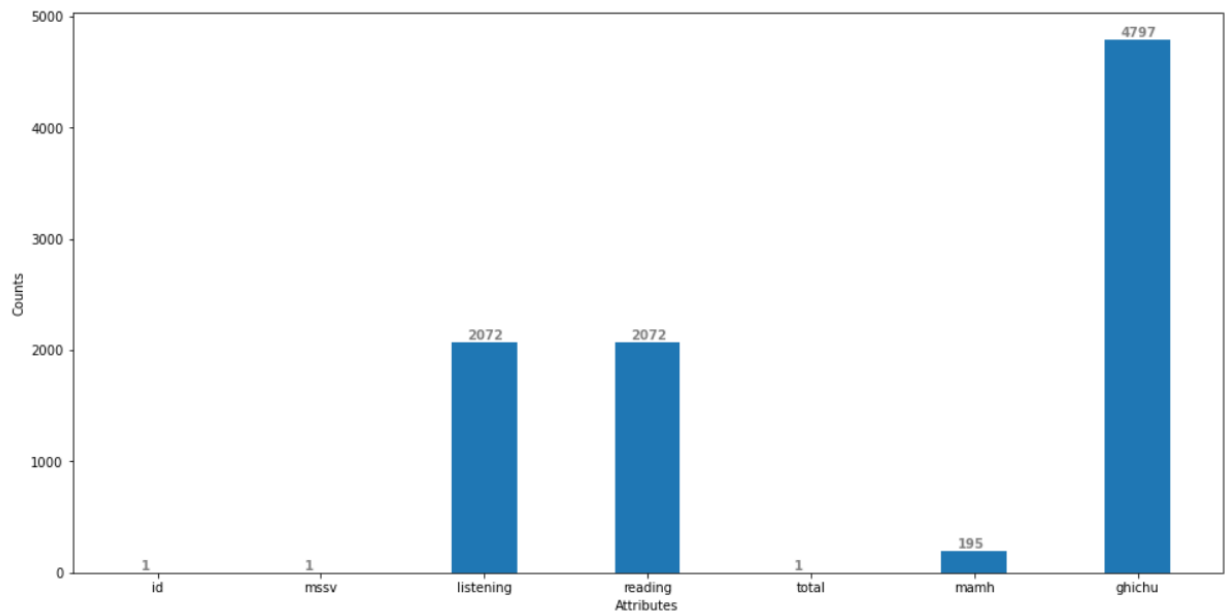
Đưa dữ liệu về cùng một định dạng thống nhất:

- Thay đổi kiểu dữ liệu của listening và reading về dạng số.
- Xóa các dấu nháy đơn dư thừa ở cột ghi chú.
- Xét điểm của AVSC, nhóm đưa AVSC về AVSC1 nếu điểm <50 và đưa về AVSC2 nếu điểm >50.
- Trong cột mamh của bảng có giá trị miễn ENG3 với điểm TOEIC tương ứng, nhóm quyết định xét theo điều kiện xếp lớp của lớp CLC để chuyển hết các giá trị "miễn ENG3" sang "ENG4" và "ENG5"

Bảng 2. Xếp lớp đầu khóa

<table>
<tr><th>Điểm kiểm tra (dạng thức TOEIC Nghe và Đọc)</th><th>CTĐTr và CTTN</th><th>CTCLC</th><th>CTTT</th></tr>
<tr><td><300</td><td colspan="2">Không được học ENG01,
học AVSC ngoài CTĐT chính quy
do TTNN phụ trách (không bắt buộc)</td><td rowspan="2">HọcENG01</td></tr>
<tr><td>300 – 345</td><td colspan="2">Học ENG01</td></tr>
<tr><td>350 – 395</td><td colspan="3">Miễn ENG01;
Học ENG02</td></tr>
<tr><td>400 – 445</td><td colspan="2">Miễn ENG01, ENG02;
HọcENG03</td><td>Miễn ENG01, ENG02;
Học ENG03</td></tr>
<tr><td>450 – 495</td><td rowspan="2">Miễn
ENG01→ENG03</td><td>Miễn ENG01 → ENG03;
Học ENG04</td><td>Miễn ENG01 → ENG03;
Học ENG04</td></tr>
<tr><td>≥500</td><td>Miễn ENG01 → ENG04;
HọcENG05</td><td>Miễn ENG01 → ENG04;
HọcENG05</td></tr>
</table>

Xử lý dữ liệu trống: nhóm thực hiện thống kê số lượng dữ liệu trống của tất cả các thuộc tính.



- Thực hiện xóa các điểm dữ liệu thiếu mã số sinh viên.
- Đối với cột ghi chú, nếu dữ liệu đó trống thì sẽ được điền là không lý do.
- Đối với cột listening và reading, ta sẽ lấy giá trị là total / 2.
- Với cột mã môn học, ta sẽ xét theo điều kiện total < 50 thì mamh = AVSC1, còn lại thì mamh = Miễn ENG03.

2.2.1.4 Thí sinh

Xử lí dữ liệu trùng lắp: Không có dữ liệu trùng.

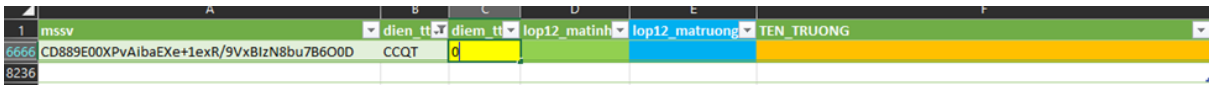
Xử lí dữ liệu bị thiếu trống:

Với mỗi hình thức xét tuyển sẽ có một mốc điểm riêng tương ứng. Nhóm sẽ xử lí những điểm dữ liệu điểm bị thiếu dựa trên hình thức xét tuyển.

- Hình thức THPT: Nhóm tính được AVG điểm là 23.88 nên sẽ làm tròn xuống 23.85 để điền vào các ô thiếu.

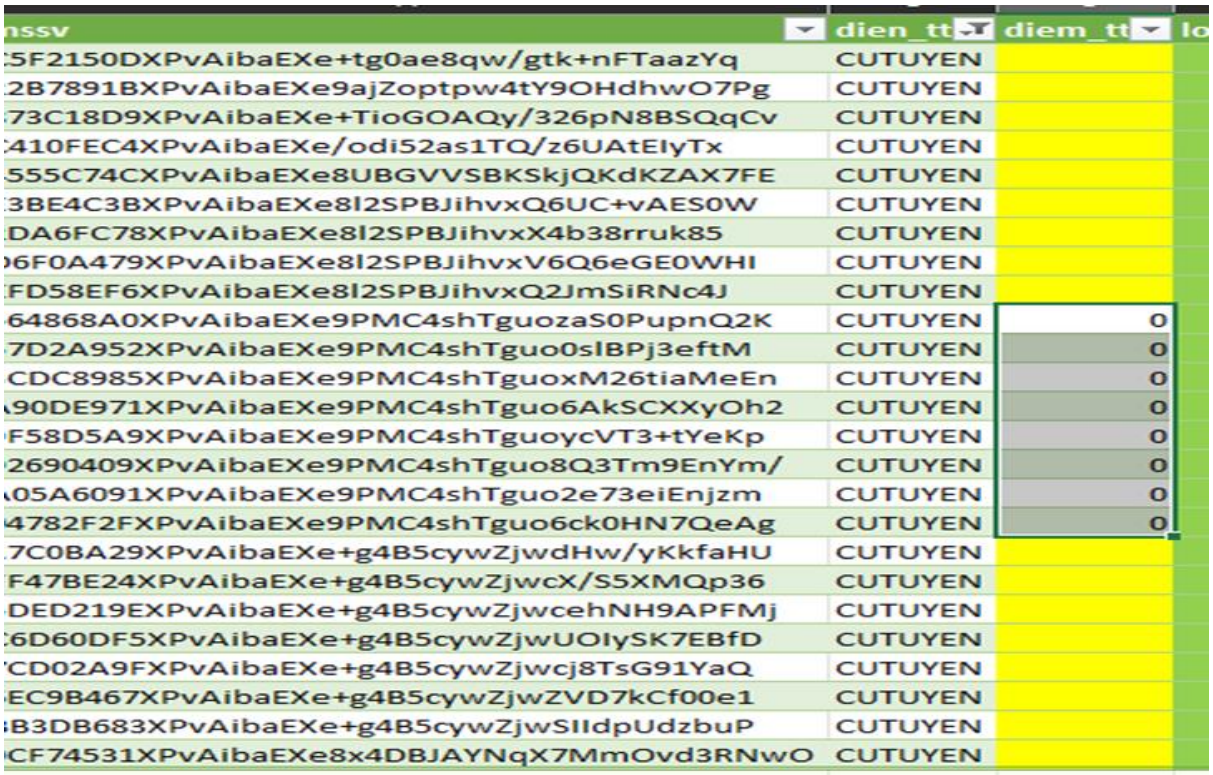
8946F2XPvAibaEXe8Wb350a8ibnm5kQzFjIM2p	THPT	23.95	37	1 Quốc
2E756EXPvAibaEXe8Wb350a8ibnvs4VyUWS40K	THPT	21.5	2	1 THPT
3431338XPvAibaEXe8xMCTZ03/BexWzDCyWVsT3	THPT	24.7	47	25 THPT
AD7B4AXPvAibaEXe8xMCTZ03/BewoGrWSt0ZXM	THPT	24.95	42	2 THPT
		23.889727		

- Hình thức CCQT chỉ có 1 dòng dữ liệu nên không đáng kể.



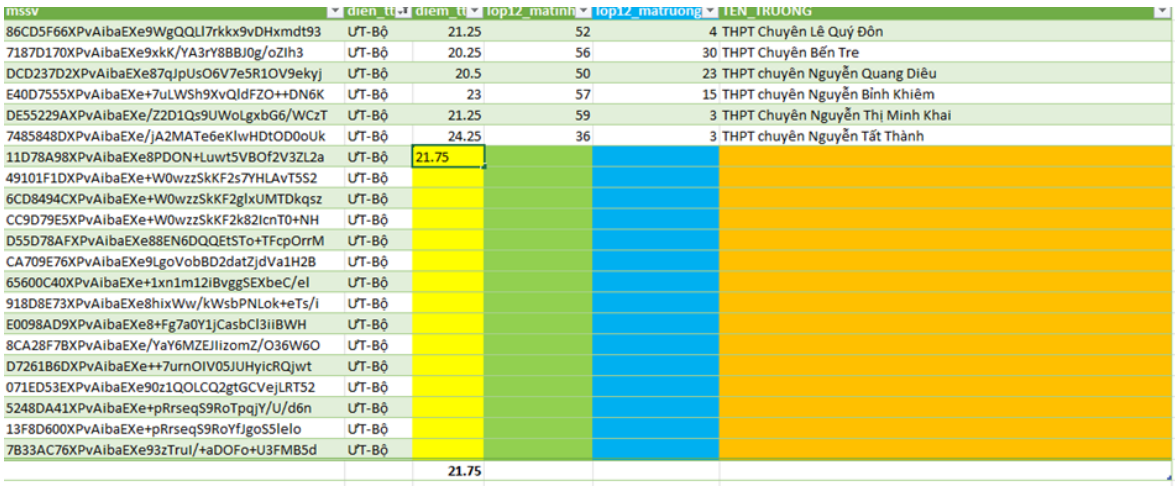
	A	B	C	D	E	F
1	mssv	dien_tt	diem_tt	lop12_matinh	lop12_matruong	TEN_TRUONG
6666	CD889E00XPvAibaEXe+1exR/9VxBIzN8bu7B6O0D	CCQT	0			
8236						

- Hình thức CUTUYEN là dạng tuyển thẳng và nhóm nhận thấy có một số dòng có giá điểm để là 0 nên nhóm lấy giá trị 0 điền cho các ô bị thiếu ở cột điểm thuộc hình thức CUTUYEN. Cách này cũng được áp dụng cho hình thức TTBộ



mssv	dien_tt	diem_tt	lo
5F2150DXPvAibaEXe+tg0ae8qw/gtk+nFTaazYq	CUTUYEN		
2B7891BXPvAibaEXe9ajZoptpw4tY9OHdhwO7Pg	CUTUYEN		
73C18D9XPvAibaEXe+TioGOAQy/326pN8BSQqCv	CUTUYEN		
410FEC4XPvAibaEXe/odi52as1TQ/z6UAtEIyTx	CUTUYEN		
555C74CXPvAibaEXe8UBGVVSBKSkjQKdKZAX7FE	CUTUYEN		
3BE4C3BXPvAibaEXe8l2SPBJihvxQ6UC+vAES0W	CUTUYEN		
DA6FC78XPvAibaEXe8l2SPBJihvxX4b38rruk85	CUTUYEN		
6F0A479XPvAibaEXe8l2SPBJihvxV6Q6eGE0WHI	CUTUYEN		
FD58EF6XPvAibaEXe8l2SPBJihvxQ2JmSiRNc4J	CUTUYEN		
64868A0XPvAibaEXe9PMC4shTguozaS0PupnQ2K	CUTUYEN	0	
7D2A952XPvAibaEXe9PMC4shTguo0slBPj3eftM	CUTUYEN	0	
CDC8985XPvAibaEXe9PMC4shTguoxM26tiaMeEn	CUTUYEN	0	
90DE971XPvAibaEXe9PMC4shTguo6AkSCXXyOh2	CUTUYEN	0	
F58D5A9XPvAibaEXe9PMC4shTguoycVT3+tYeKp	CUTUYEN	0	
2690409XPvAibaEXe9PMC4shTguo8Q3Tm9EnYm/	CUTUYEN	0	
05A6091XPvAibaEXe9PMC4shTguo2e73eiEnjzm	CUTUYEN	0	
4782F2FXPvAibaEXe9PMC4shTguo6ck0HN7QeAg	CUTUYEN	0	
7C0BA29XPvAibaEXe+g4B5cywZjwdHw/yKkfaHU	CUTUYEN		
F47BE24XPvAibaEXe+g4B5cywZjwcX/S5XMQp36	CUTUYEN		
DED219EXPvAibaEXe+g4B5cywZjwcehNH9APFMj	CUTUYEN		
6D60DF5XPvAibaEXe+g4B5cywZjwUOIySK7EBfD	CUTUYEN		
CD02A9FXPvAibaEXe+g4B5cywZjwcj8TsG91YaQ	CUTUYEN		
EC9B467XPvAibaEXe+g4B5cywZjwZVD7kCf00e1	CUTUYEN		
B3DB683XPvAibaEXe+g4B5cywZjwSIIdpUdzbuP	CUTUYEN		
CF74531XPvAibaEXe8x4DBJAYNqX7MmOvd3RNwO	CUTUYEN		

- Hình thức ƯT-Bộ: Nhóm tính được AVG điểm là 21.75 nên sẽ điền vào các ô thiếu.

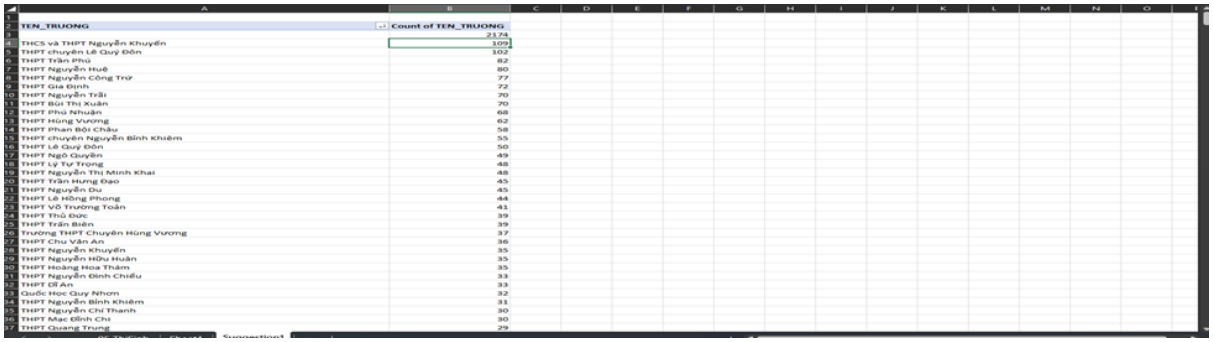


mssv	dien_tt	diem_tt	lop12_matinh	lop12_matruong	TEN_TRUONG
86CD5F66XPvAibaEXe9WgQQLl7rkkx9vDHxmdt93	ƯT-Bộ	21.25	52	4	THPT Chuyên Lê Quý Đôn
7187D170XPvAibaEXe9xkK/YA3rY8BBJ0g/oZIh3	ƯT-Bộ	20.25	56	30	THPT Chuyên Bến Tre
DCD237D2XPvAibaEXe87qJpUsO6V7e5R1OV9ekyj	ƯT-Bộ	20.5	50	23	THPT chuyên Nguyễn Quang Diêu
E40D7555XPvAibaEXe+7uLWSh9XvQIdFZO++DN6K	ƯT-Bộ	23	57	15	THPT chuyên Nguyễn Bỉnh Khiêm
DE55229AXPvAibaEXe/Z2D1Qs9UWoLgxbG6/WCzT	ƯT-Bộ	21.25	59	3	THPT Chuyên Nguyễn Thị Minh Khai
7485848DXPvAibaEXe/jA2MATe6eKIwHDtOD0oUk	ƯT-Bộ	24.25	36	3	THPT chuyên Nguyễn Tất Thành
11D78A98XPvAibaEXe8PDON+Luwt5VBOf2V3ZL2a	ƯT-Bộ	21.75			
49101F1DXPvAibaEXe+W0wzzSkKF2s7YHLAvT5S2	ƯT-Bộ				
6CD8494CXPvAibaEXe+W0wzzSkKF2glxUMTDkqsz	ƯT-Bộ				
CC9D79E5XPvAibaEXe+W0wzzSkKF2k82IcnT0+NH	ƯT-Bộ				
D55D78AFXPvAibaEXe88EN6DQQEtSTo+TFcpOrrM	ƯT-Bộ				
CA709E76XPvAibaEXe9LgoVobBD2datZjdVa1H2B	ƯT-Bộ				
65600C40XPvAibaEXe+1xn1m12iBvggSEXbeC/el	ƯT-Bộ				
918D8E73XPvAibaEXe8hixWw/kWsbPNLok+eTs/i	ƯT-Bộ				
E0098AD9XPvAibaEXe8+Fg7a0Y1jCasbCl3iiBWH	ƯT-Bộ				
8CA28F7BXPvAibaEXe/YaY6MZEJIizomZ/O36W6O	ƯT-Bộ				
D7261B6DXPvAibaEXe++7urnOIV05JUHyicRQjwt	ƯT-Bộ				
071ED53EXPvAibaEXe90z1QOLCQ2gtGCVejLRT52	ƯT-Bộ				
5248DA41XPvAibaEXe+pRrseqS9RoTpqjY/U/d6n	ƯT-Bộ				
13F8D600XPvAibaEXe+pRrseqS9RoYfJgoS5lelo	ƯT-Bộ				
7B33AC76XPvAibaEXe93zTrul/+aDOFo+U3FMB5d	ƯT-Bộ				
		21.75			

- Hình thức ƯT-ĐHQG: Nhóm tính được AVG điểm là 25.8 nên điền vào các ô thiếu.

86DF147XPvAibaEXe/q/d6S7LoUw4tVaLlsfE7p	ƯT-ĐHQG	24.1	57	15	THPT chuyên Nguyễn Bỉnh Khiêm
FE3F325XPvAibaEXe/q/d6S7LoUwzxPRBYXTiEh	ƯT-ĐHQG	27.2	34	7	THPT chuyên Nguyễn Bỉnh Khiêm
4EECABBXPvAibaEXe/q/d6S7LoUw9AWKspJi03t	ƯT-ĐHQG	25.9	63	37	THPT chuyên Nguyễn Chí Thanh
B43D033XPvAibaEXe/q/d6S7LoUwzBJQEyK9w6Z	ƯT-ĐHQG	26.2	50	23	THPT chuyên Nguyễn Quang Diêu
1261CCAXPvAibaEXe/LvKmPOy7w5zZfHv6Q5Cro	ƯT-ĐHQG	25.7	45	17	THPT chuyên Lê Quí Đôn
D671B38XPvAibaEXe/LvKmPOy7w5/y4iP4ff15b	ƯT-ĐHQG	25.9	35	13	Trường THPT chuyên Lê Khiết
A92BD38XPvAibaEXe/pu9fEdzG2a8JcIvh5MmvZ	ƯT-ĐHQG	25.1	51	2	THPT Chuyên Thoại Ngọc Hầu
A6A920DXPvAibaEXe/pu9fEdzG2ay0K718vt9kt	ƯT-ĐHQG	27.3	2	70	THPT Nguyễn Hữu Huân
F6B7C0EXPvAibaEXe9lUBdG2blqkqC9pVR6qzQN	ƯT-ĐHQG	25.4	60	9	THPT Chuyên Bạc Liêu
FE284AAXPvAibaEXe9lUBdG2blqkoKR4vqB7+UN	ƯT-ĐHQG	26	63	37	THPT Chuyên Nguyễn Chí Thanh
25DE12BXPvAibaEXe9lUBdG2blqkvl/b/nkUvmT	ƯT-ĐHQG	26.2	29	7	THPT Chuyên - Đại học Vinh
D3F195DXPvAibaEXe9lUBdG2blqkl/kK90e0GkM	ƯT-ĐHQG	25.5	51	2	THPT Chuyên Thoại Ngọc Hầu
3918911XPvAibaEXe8uDPDBBTbYrBR+erGSTT5q	ƯT-ĐHQG	26.2	2	53	THPT Nguyễn Công Trứ
A50F4FEXPvAibaEXe8uDPDBBTbYrCmcAvdq0oUX	ƯT-ĐHQG	25.3	39	5	THPT Chuyên Lương Văn Chánh
F8A419FXPvAibaEXe8uDPDBBTbYrMZOW6f+ONC4	ƯT-ĐHQG	26.3	34	7	THPT chuyên Nguyễn Bỉnh Khiêm
ADE2105XPvAibaEXe8X5f+vLkYyLm5FxXtBkamp	ƯT-ĐHQG	28	2	41	THPT Nguyễn Du
9A18589XPvAibaEXe+Bn/Kbq8h84OcmHXv/YLAD	ƯT-ĐHQG	26.4	34	10	THPT chuyên Lê Thánh Tông
C198FB7XPvAibaEXe+Bn/Kbq8h84Nh4Gk40SFXB	ƯT-ĐHQG	27.5	53	15	THPT Nguyễn Đình Chiểu
0C7AB81XPvAibaEXe+Bn/Kbq8h84PR7K1SVKA8p	ƯT-ĐHQG	25.3	64	39	THPT chuyên Vị Thanh
E8E6269XPvAibaEXe+Bn/Kbq8h84OsrwaWP9SkQ	ƯT-ĐHQG	25.8	2	70	THPT Nguyễn Hữu Huân
1E88E8EXPvAibaEXe+1exR/9VxBIzE4VWy1ppy/	ƯT-ĐHQG	26	42	26	THPT Bảo Lộc
1F82710XPvAibaEXe+1exR/9VxBI9LfEtB0cVb8	ƯT-ĐHQG	24.4	51	8	THPT Chuyên Thủ Khoa Nghĩa
		25.800913			

- Phần TEN_TRUONG trống nhiều điểm dữ liệu nên nhóm đã phân tích giá trị xuất hiện nhiều nhất để gán cho các ô bị thiếu. Trường hợp này là THCS và THPT Nguyễn Khuyến.



TEN_TRUONG	Count of TEN_TRUONG
	2174
THCS và THPT Nguyễn Khuyến	109
THPT chuyên Lê Quý Đôn	102
THPT Trần Phú	82
THPT Nguyễn Huệ	80
THPT Nguyễn Công Trứ	77
THPT Gia Định	72
THPT Nguyễn Trãi	70
THPT Bùi Thị Xuân	70
THPT Phú Nhuận	68
THPT Hùng Vương	62
THPT Phan Bội Châu	58
THPT chuyên Nguyễn Bỉnh Khiêm	55
THPT Lê Quý Đôn	50
THPT Ngô Quyền	49
THPT Lý Tự Trọng	48
THPT Nguyễn Thị Minh Khai	48
THPT Trần Hưng Đạo	45
THPT Nguyễn Du	45
THPT Lê Hồng Phong	44
THPT Võ Trường Toản	41
THPT Thủ Đức	39
THPT Trần Biên	39
Trường THPT Chuyên Hùng Vương	37
THPT Chu Văn An	36
THPT Nguyễn Khuyến	35
THPT Nguyễn Hữu Huân	35
THPT Hoàng Hoa Thám	35
THPT Nguyễn Đình Chiểu	33
THPT Dĩ An	33
Quốc Học Quy Nhơn	32
THPT Nguyễn Bỉnh Khiêm	31
THPT Nguyễn Chí Thanh	30
THPT Mạc Đĩnh Chi	30
THPT Quang Trung	29

- Các dòng dữ liệu có sẵn mã trường hoặc mã tỉnh thì nhóm ưu tiên tìm kiếm thông tin trên mạng khớp với dữ liệu đã cho, nếu mã trường và mã tỉnh không khớp thì nhóm ưu tiên lấy mã trường để điền. Ngoài ra với các sai sót về mã trường, nhóm sẽ lấy mã trường phổ biến nhất của tỉnh tương ứng để điền vào, và sai sót về mã tỉnh thì sẽ dựa trên thông tin bằng cách tìm kiếm mã trường để điền mã tỉnh phù hợp.

2.2.1.5 Điểm (diem_Thu)

- Nhóm loại bỏ những dòng có môn học khác với tiếng Anh. Từ cột điểm học phần và điểm môn, nhóm thêm một cột “mamh_tiep”để phù hợp hơn với bài toán, đồng thời giúp nhóm dễ dàng đánh giá các sinh viên năm trước.

- Với các mã môn học tương ứng với hiện nay như ENG01, ENG02, … thì nhóm sẽ xét điểm học phần.
    - Điểm học phần lớn hơn bằng 5 sẽ nâng thêm 1 bậc (ENG01 với điểm học phần 6.0 sẽ là ENG02)
    - Điểm học phần bé hơn 5 sẽ giữ nguyên bậc.
- Với các mã môn học cũ thì sẽ dựa vào danh sách môn học của trường để tìm ra mức độ Anh văn tương ứng của môn học đó, và cũng sẽ áp dụng điều kiện đổi như trên.

Sau khi đã xử lí xong các bảng dữ liệu cần thiết, nhóm sẽ tiến hành kết hợp các bản và lọc ra các yếu tố mà nhóm cần để tạo ra một file data mới phù hợp hơn là data_final.

2.2.1.6 Data_final

Bảng data_final được merge lại từ bảng sinhvien, xeploaiav, thisinh đã qua xử lý (khóa chính là mssv). Nhóm sẽ giữ lại các thuộc tính mà nhóm dự đoán sẽ ảnh hưởng đến chuẩn quá trình anh văn của sinh viên: namsinh, gioitinh, khoa, hedt, khoahoc, mamh, dien_tt. Ngoài ra, nhóm thêm vào bảng data_final 3 thuộc tính mới:

- khu_vuc: khu vực ưu tiên trong thi trung học phổ thông quốc gia với mỗi khu vực sẽ có điều kiện học tập ngoại ngữ khác nhau. Được gom nhóm từ file thisinh với 3 thuộc tính gom nhóm là lop12_matruong, TEN_TRUONG, lop12_matinh. Dựa vào quy định của Bộ Giáo Dục Và Đào Tạo, nhóm gom thành 4 nhóm: khu vực 1, khu vực 2, khu vực 2NT và khu vực 3.
- chuanav_1: thông tin xếp loại anh văn gần nhất của sinh viên kể từ khi thi anh văn đầu vào. Được trích xuất từ file diem_Thu với cấu trúc chung là nếu điểm anh văn cuối kì học phần gần nhất kể từ lúc thi anh văn đầu vào lớn hơn 5 thì chuẩn anh văn của sinh viên sẽ được thêm 1 bậc, ví dụ: sinh viên học lớp ENG01 đạt 6 điểm thì chuanav_1 sẽ ENG02. Trường hợp sinh viên đó không có thông tin anh văn trong file diem_Thu tức là sinh viên đó không tham gia học tại trung tâm ngoại ngữ của trường, nhóm sẽ xem như là không có thông tin và điền giá trị thuộc tính này bằng với xếp loại anh văn đầu vào.
- Label: gán nhãn để lấy dữ liệu với qui ước như sau: 0 (sinh viên chưa đạt đủ chuẩn quá trình) – 1 (sinh viên đã đạt chuẩn quá trình).
    - Các tiêu chí để gán nhãn 1:
        - Sinh viên có mã số sinh viên trong file tốt nghiệp (tức là đã đạt chuẩn quá trình thì mới được tốt nghiệp).
        - Sinh viên nộp chứng chỉ hợp lệ (trong file sinhvien_chungchi_final) đủ điểm để qua ngưỡng anh văn.
        - Sinh viên thi xếp lớp anh văn đầu vào đủ điểm (trong file xeploaiav_final).

        - Điểm học anh văn mới nhất của sinh viên đủ đạt chuẩn anh văn.
    - Các điểm dữ liệu còn lại sẽ được gán nhãn 0.

Vậy, bảng data_final sẽ bao gồm 11 cột tương ứng là mssv, namsinh (năm sinh), gioitinh (giới tính), khoa (ngành đào tạo), hedt (hệ đào tạo), khoahoc (khóa), mamh (xếp loại tiếng Anh đầu vào), dien_tt (diện xét tuyển), khu_vuc (khu vực địa lí trường THPT), chuanav_1 (xếp loại tiếng Anh gần nhất), Label (Nhãn).

Sau khi kết bảng và lọc dữ liệu, bảng data_final có tổng cộng 4916 điểm dữ liệu tương ứng với 4916 sinh viên với 11 thuộc tính.

Bảng mô tả dữ liệu data_final

STT	Tên cột	Thuộc tính	Ý nghĩa	Kiểu dữ liệu	Ghi chú
1	mssv	Mã số sinh viên	Mã số của sinh viên được mã hóa	string	
2	namsinh	Năm sinh	Năm sinh của sinh viên	int	Giá trị từ 1988 tới 2001
3	gioitinh	Giới tính	Giới tính của sinh viên	int	0 là nữ 1 là nam
4	khoa	Khoa	Chuyên ngành sinh viên theo học	string	Có các ngành: CNPM, HTTT, KHMT, KTMT, KTTT, MMT&TT
5	hedt	Hệ đào tạo	Hệ đào tạo của sinh viên	string	Có các hệ CLC, CQUI, KSTN, CNTT, CTTT
6	khoahoc	Khóa học	Khóa mà sinh viên vô trường	int	Giá trị từ 9 đến 14
7	mamh	Xếp loại AV đầu vào	Xếp loại của sinh viên sau kì kiểm tra tiếng Anh đầu vào	string	Có các mức: AVSC1, AVSC2, ENG01, ENG02, ENG03, ENG04, ENG05

8	dien_tt	Diện tuyển	Cách sinh viên xét tuyển vào trường	string	Có các dạng: THPT, 30A, CCQT, CUTUYEN, ĐGNL, TT-Bộ, ƯT-Bộ, ƯT-ĐHQG
9	khu_vuc	Khu vực	Khu vực trường THPT mà sinh viên theo học	string	Có các khu: 1, 2, 3, 2NT
10	chuanav_1	Xếp loại AV gần nhất	Xếp loại Anh văn của sinh viên gần nhất kể từ khi thi đầu vào	string	Có các mức: AVSC1, AVSC2, ENG01, ENG02, ENG03, ENG04, ENG05, ENG06
11	Label	Nhãn của điểm dữ liệu	Nhãn kiểm tra điều kiện đạt chuẩn quá trình của sinh viên	int	0 là không đạt chuẩn 1 là đạt chuẩn

2.2.2 Thuộc tính sử dụng

Thuộc tính được sử dụng bao gồm:

Thuộc tính	Kiểu dữ liệu	Ý nghĩa
gioitinh	int	Giới tính của sinh viên
namsinh	int	Năm sinh của sinh viên
dien_tt	string	Hình thức được xét tuyển vào trường
khoahoc	int	Khóa học
khu_vuc	string	Khu vực của trường THPT đã theo học
khoa	string	Ngành sinh viên ứng tuyển

hedt	string	Hệ đào tạo của sinh viên
mamh	string	Xếp loại Anh văn đầu vào
chuanav_1	string	Xếp loại Anh văn của sinh viên gần nhất kể từ khi thi đầu vào

2.2.3 Phương pháp đề xuất

Vì đây là một bài toán phân lớp nhị phân nên nhóm sẽ thực hiện dựa trên một quá trình để xây dựng các mô hình phân lớp dữ liệu.

Đầu tiên là chuẩn bị dataset: Sau khi tiến hành xử lí, làm sạch, nhóm đã xây dựng được file dữ liệu là data_final. Từ đó nhóm sẽ xây dựng các trường hợp để cho ra các mô hình đánh giá khác nhau.

Thứ hai là xây dựng mô hình phân lớp: Để xây dựng mô hình cho bài toán phân lớp cần sử dụng các thuật toán học có giám sát (supervised learning) nên nhóm đã chọn ba thuật toán để tiến hành xây dựng mô hình và đánh giá là Logistic Regression, Support Vector Machine, Adaboost và thêm cả Multi-layer Perceptron.

Thứ ba là kiểm tra dữ liệu với mô hình và cuối cùng là đánh giá mô hình phân lớp. Nhóm sử dụng bốn thang đo đặc trưng cơ bản là Precision, Recall, F1-score và Accuracy.

2.3 Cài đặt thực nghiệm

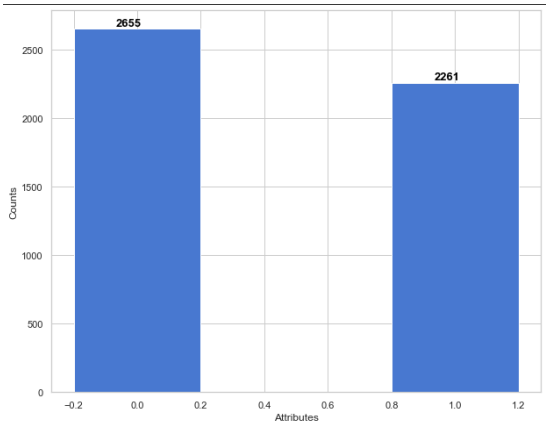
2.3.1 Dataset

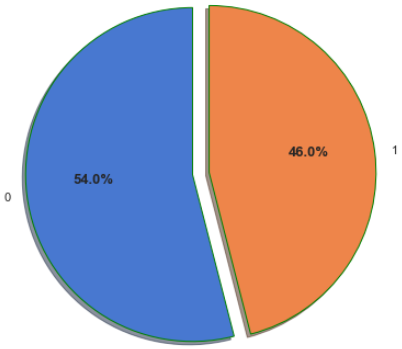
Nhóm sẽ phân tích dữ liệu của data_final. Ngoài hai cột là mssv dùng để kết bản và cột chuanav_1 là dữ liệu dùng để học tăng cường, nhóm sẽ phân tích các thuộc tính còn lại.

```
---  ------    --------------  -----
 0   namsinh   4916 non-null   object
 1   gioitinh  4916 non-null   object
 2   khoa      4916 non-null   object
 3   hedt      4916 non-null   object
 4   khoahoc   4916 non-null   object
 5   mamh      4916 non-null   object
 6   dien_tt   4916 non-null   object
 7   khu_vuc   4916 non-null   object
 8   Label     4916 non-null   object
```

Đầu tiên là thống kê Label.

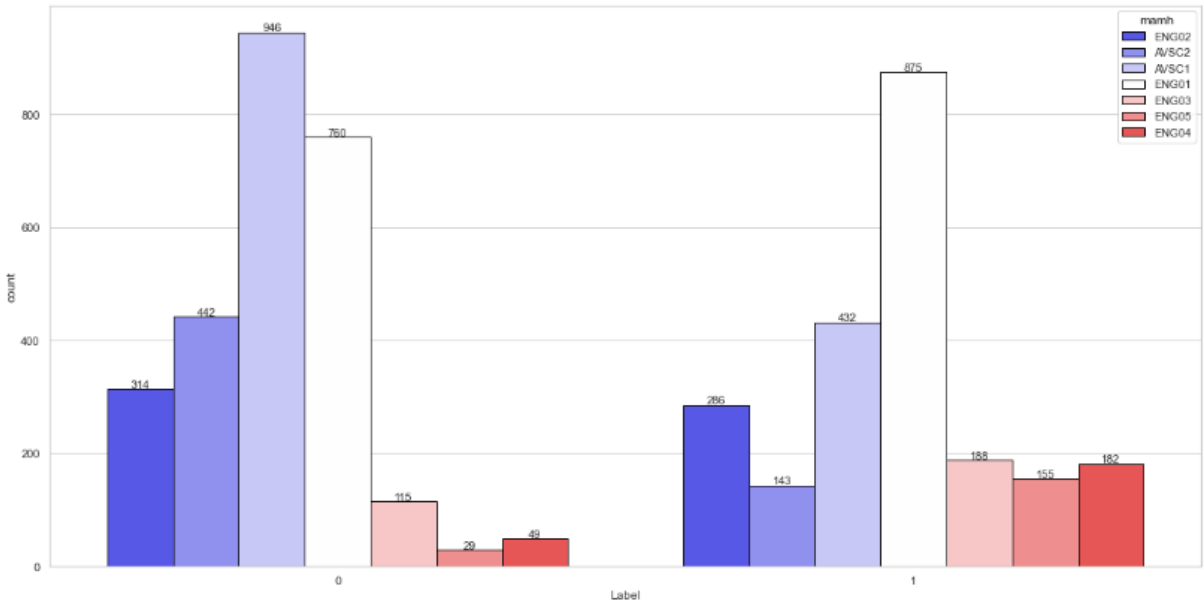
Dựa qua biểu đồ bên dưới, nhóm có thể đưa ra nhận định:





Dữ liệu giữa hai label là 1 và 0 khá cân bằng. Vì vậy nhóm không cần sử dụng các kĩ thuật cân bằng dữ liệu.

Thứ hai sự ảnh hưởng của Anh văn đầu vào (thuộc tính mamh).



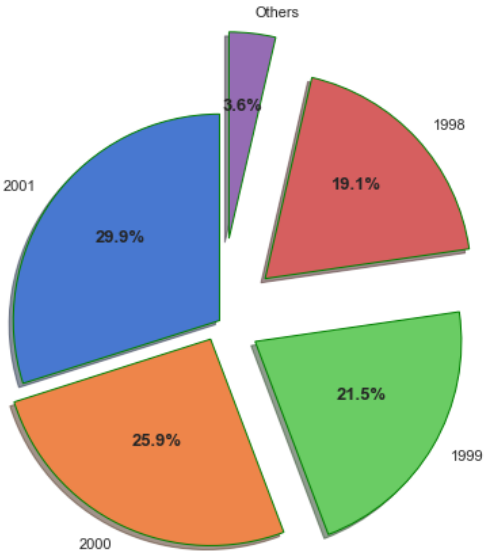
Qua thống kê của biểu đồ, nhóm có tỉ lệ:

- AVSC1[0] / AVSC1[1] = 2.2
- AVSC2[0] / AVSC2[1] = 3
- ENG01[0] / ENG01[1] = 0.87
- ENG02[0] / ENG02[1] = 1.1
- ENG03[0] / ENG03[1] = 0.61
- Tỉ lệ ENG04, ENG05 đều bé hơn 1

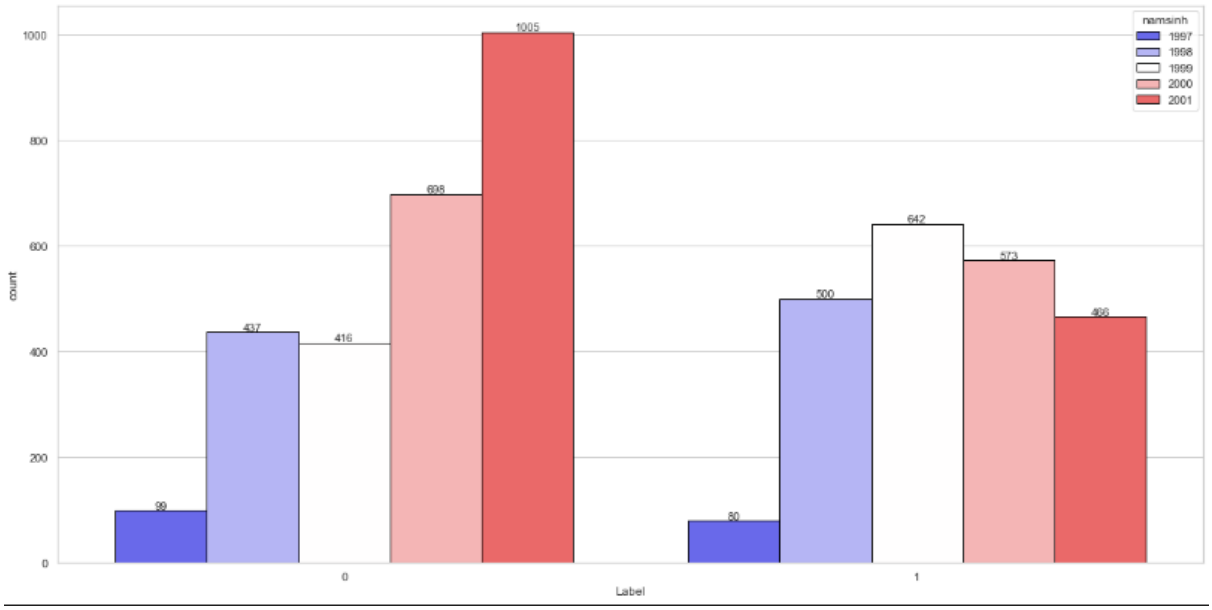
Từ đây nhóm thấy rằng sinh viên khi đầu vào được AVSC1 hoặc AVSC2 thường sẽ không đạt chuẩn quá trình lớn hơn các mức Anh văn khác.

Thứ ba là sự ảnh hưởng của năm sinh (thuộc tính namsinh)

2001	1471
2000	1271
1999	1058
1998	937
1997	124
1996	33
1995	15
1992	3
1993	1
1994	1
1991	1
1988	1



Với bản thông tin trên, các giá trị từ 1988 tới 1997 có số lượng rất ít nên nhóm sẽ gom nhóm các thành Others cho thuận tiện việc visualize.

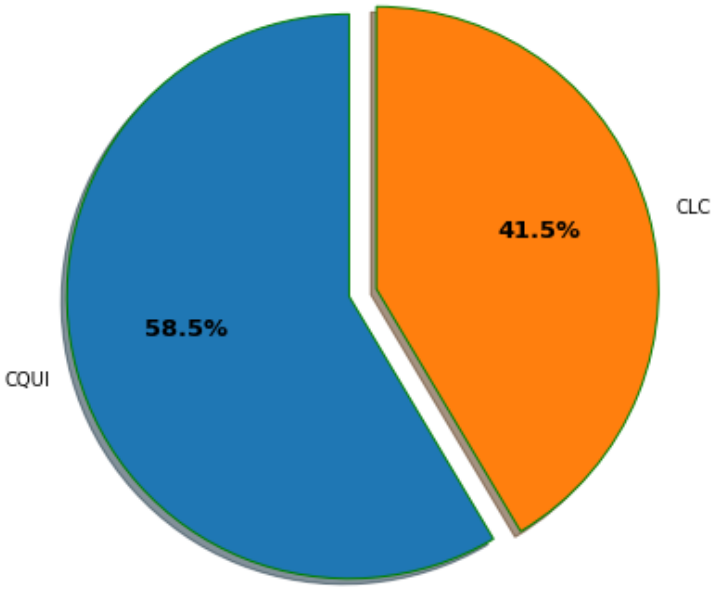


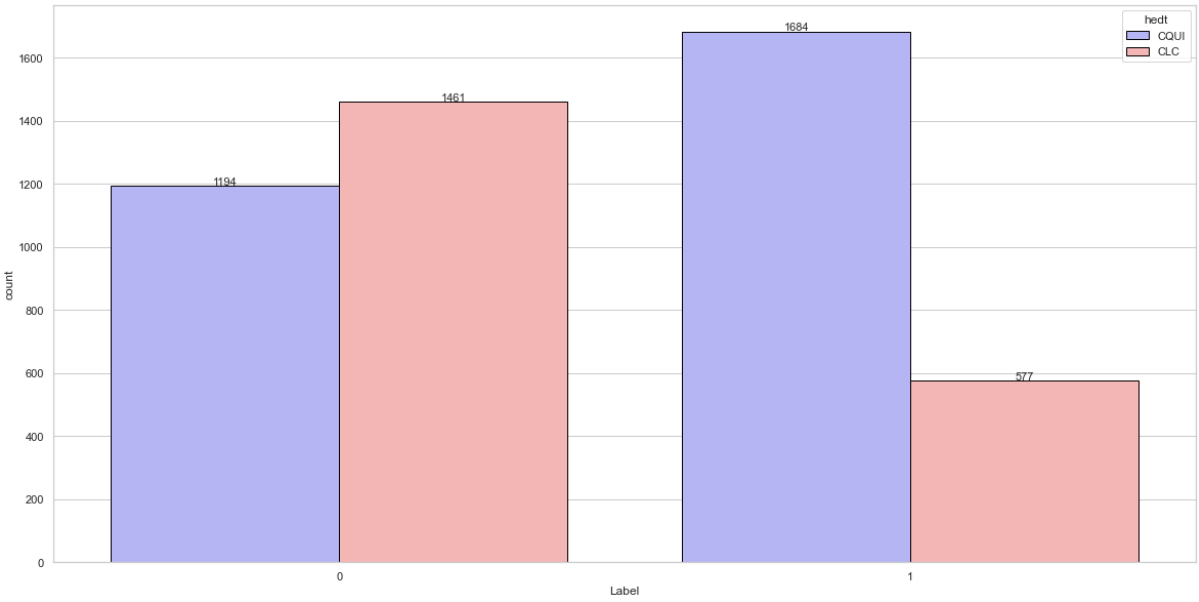
Những sinh viên có số tuổi nhỏ (2001-2000) có số lượng lớn chưa đủ quá trình, lí do có thể giải thích cho trường hợp này là dữ liệu được lấy khi sinh viên đang còn ở năm 1,2 nên chưa cần xét chuẩn quá trình.

Thứ tư là sự ảnh hưởng của hệ đào tạo (thuộc tính hedt)

Ở đây có 5 hệ đào tạo khác nhau. Tuy nhiên, dựa theo quy định mà nhóm có thể gom CQUI, KSTN, CNTN thành 1 nhóm do có cùng chuẩn AV như nhau và CLC, CTTT thành một nhóm. Từ đây, khi đề cập đến hệ CQUI sẽ bao gồm KSTN và CNTN. Tương tự với hệ CLC.

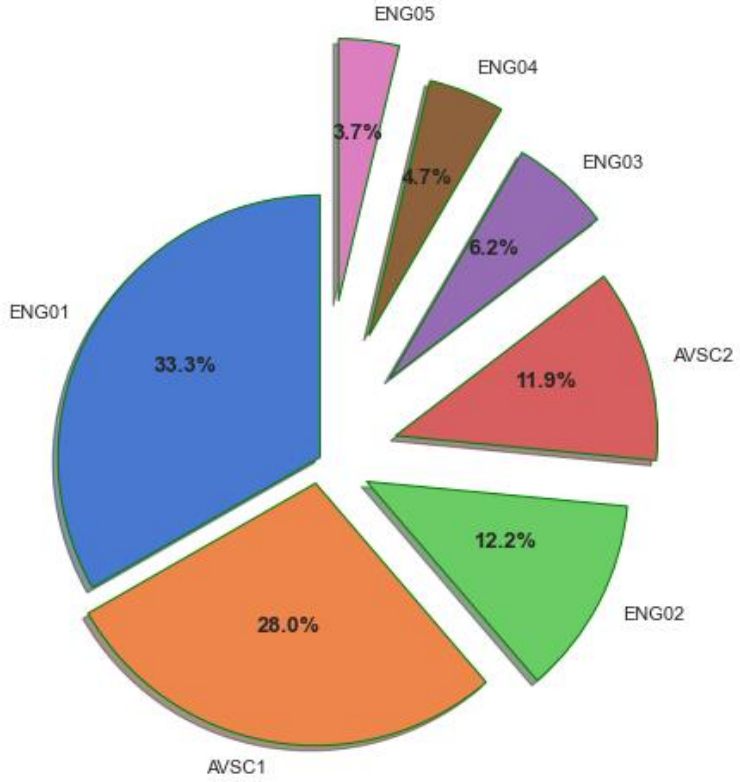
CQUI	2683
CLC	1858
CTTT	180
KSTN	101
CNTN	94

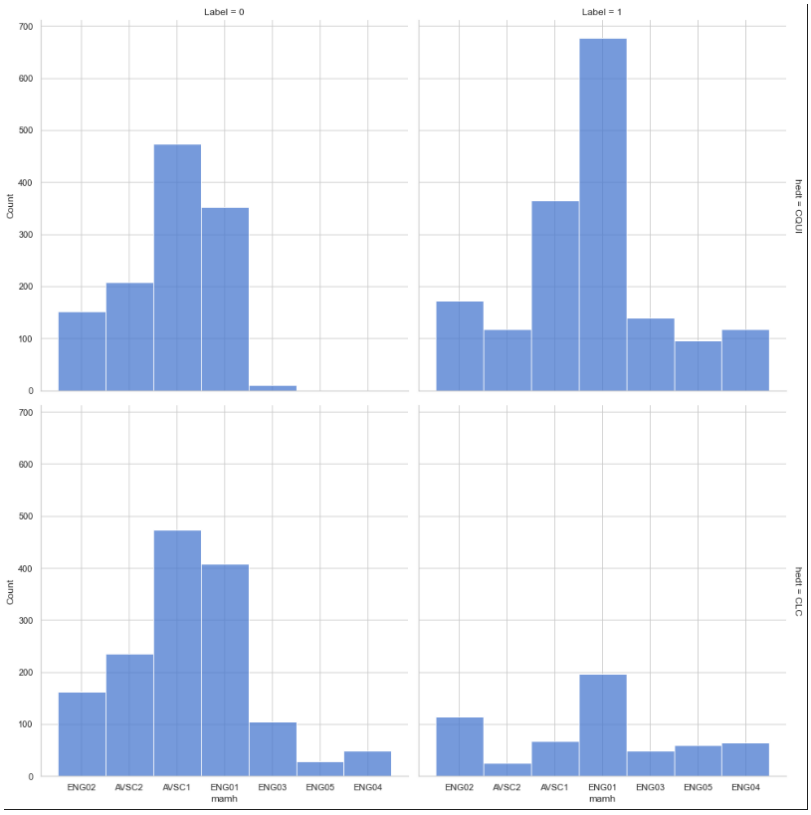




Hệ CLC có tỉ lệ sinh viên không đạt chuẩn nhiều gấp đôi so với hệ CQUI do ngưỡng AV để đạt chuẩn cao hơn hệ CQUI.

Thông qua quan sát biểu đồ trái dưới có thể thấy: Phần lớn sinh viên có mức tiếng Anh đầu vào ở mức trung bình và yếu.





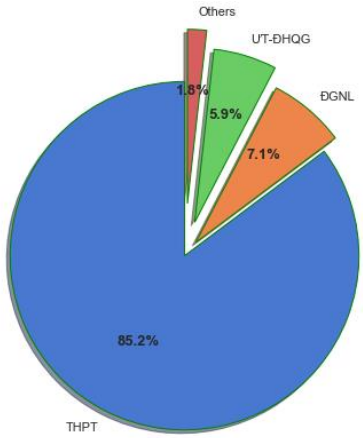
Dựa vào biểu đồ trên có thể nói: Đối với hệ chính quy, khi thi Anh văn đầu vào đạt mức ENG04 và ENG05 thì chắc chắn đạt chuẩn quá trình, ENG03 cũng rất cao sinh

viên thi vừa chạm ngưỡng để xét. Đối với hệ CLC, sinh viên khi thi AVSC1, AVSC2 đa số sẽ không thể đạt chuẩn quá trình.

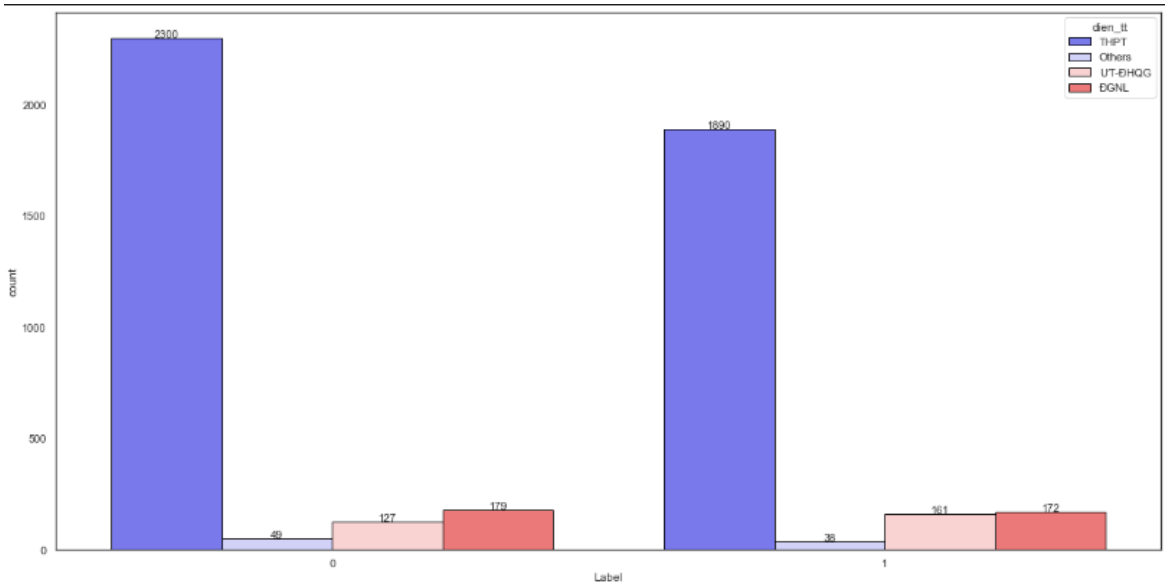
Thứ tư là sự ảnh hưởng của diện xét tuyển vào trường.

Thuộc tính của diện xét tuyển gòm nhiều giá trị khác nhau nên nhóm sẽ gom lại thành 4 cụm chính là THPT, ĐGNL, ƯT-ĐHQG, Others.

THPT	4190
ĐGNL	351
ƯT-ĐHQG	288
TT-BỘ	42
CUTUYEN	22
ƯT-BỘ	20
30A	2
CCQT	1

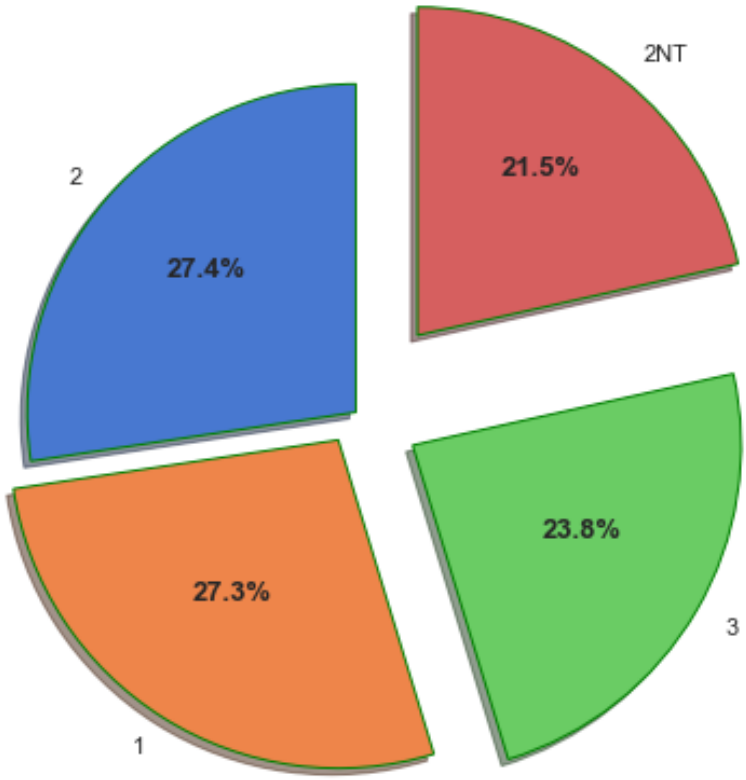


Phần lớn sinh viên xét tuyển vào trường bằng hình thức thi THPT.

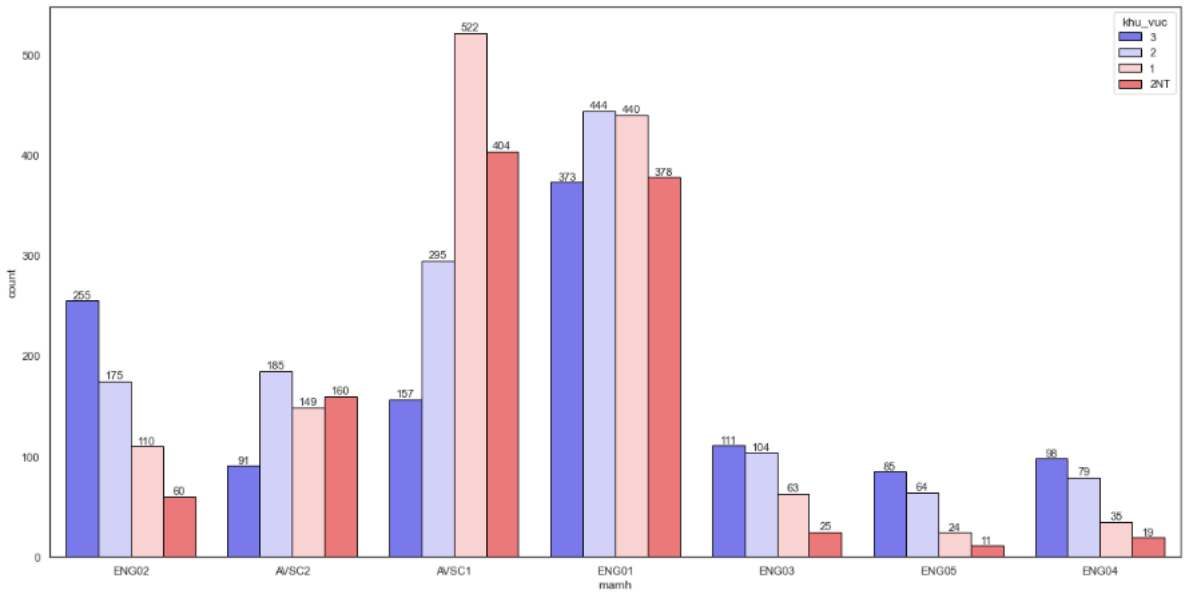


Nhìn vào đồ thị trên có thể thấy rõ không có sự chênh lệch giữa các diện xét tuyển, tức tỉ lệ thuộc mỗi class là như nhau. Điều này có thể dẫn đến diện xét tuyển không ảnh hưởng nhiều tới chuẩn quá trình của sinh viên.

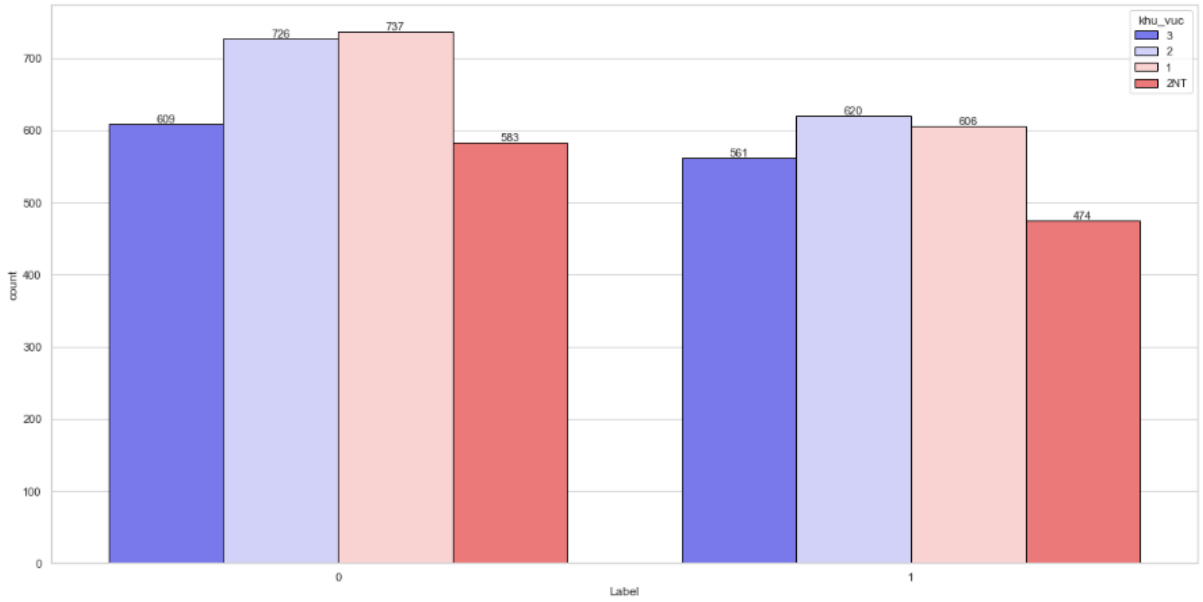
Thứ sáu là khu vực.



Tỉ lệ các khu vực cân bằng với nhau.



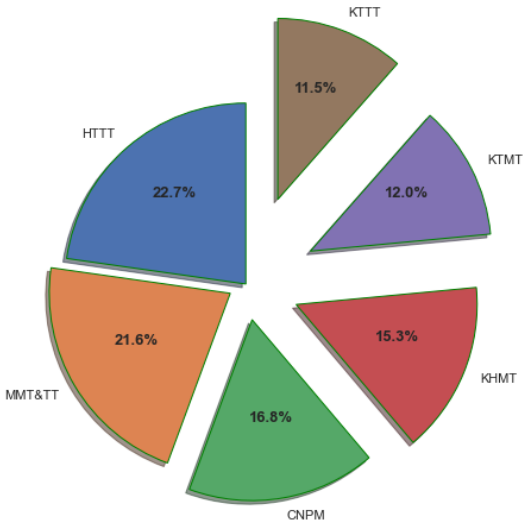
Thuộc tính khu vực có ảnh hưởng đến kết quả thi anh văn đầu vào của sinh viên (mamh). Khi các sinh viên thuộc khu vực 3 (không được cộng điểm ưu tiên) rất giỏi tiếng anh khi phần lớn đều đạt ENG01 trở lên. Ngược lại các bạn khu vực 1 (cộng nhiều điểm ưu tiên nhất) thường học anh văn không tốt bằng các khu vực khác



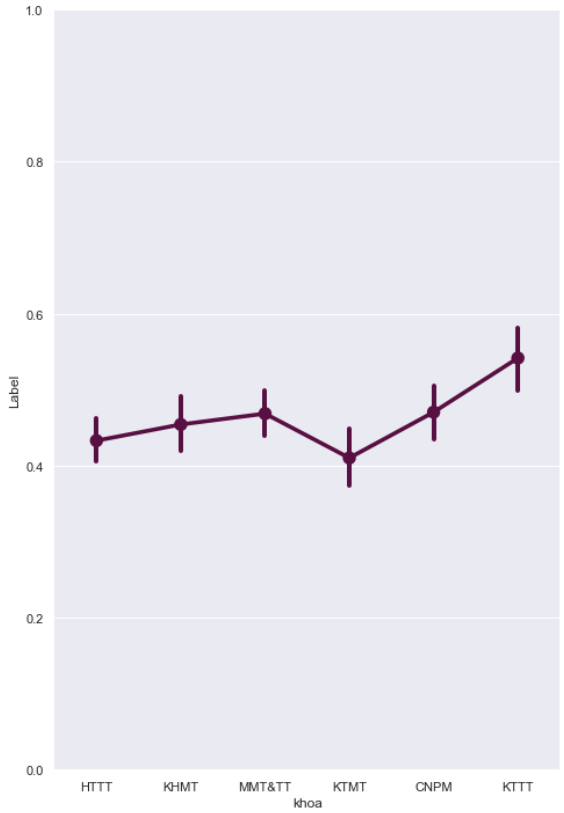
Tuy khu vực ảnh hưởng đến kết quả thi anh văn đầu vào, nhưng dường như thuộc tính này lại không ảnh hưởng đến việc các bạn có đạt chuẩn quá trình hay không khi tỉ lệ giữa các khu vực đều tương tự nhau (~50%). Lí do có thể là khi các bạn đều học cùng một nơi thì không có sự khác biệt về điều kiện học anh văn nên việc các bạn có đạt hay không đạt thì xác suất đều như nhau.

Thứ bảy là mối quan hệ của thuộc tính khoa đến nhãn.

HTTT	1118
MMT&TT	1063
CNPM	827
KHMT	753
KTMT	590
KTTT	565



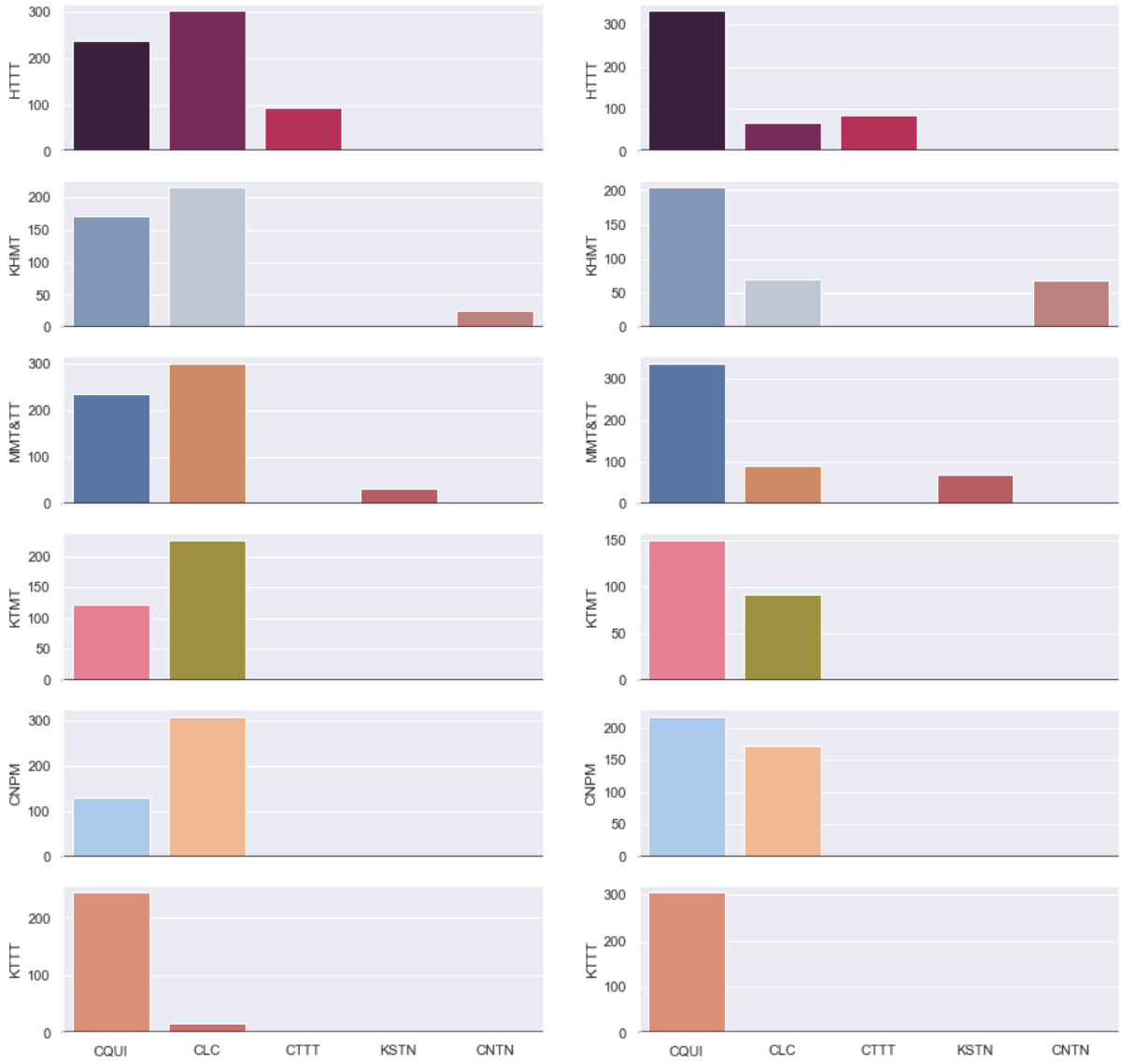
Có tổng cộng 6 khoa, với số lượng sinh viên nhiều nhất thuộc khoa HTTT và ít nhất là KTTT (chênh lệch gấp đôi).



Đồ thị trên thể hiện xác suất có đạt chuẩn quá trình anh văn với biến khoa.

Đối với mỗi khoa, tỉ lệ đều rơi vào cùng một khoảng xác suất xác định (40%-60%). Nếu chỉ tính riêng mỗi thuộc tính khoa thì sẽ không xuất hiện một xu hướng đặc biệt nào để có thể phân loại các nhãn. Ví dụ như có một khoa tỉ lệ đạt chỉ 10%, hay cao tới 80%.

Câu hỏi đặt ra là? Nếu khoa kết hợp với hệ đào tạo thì có thể xảy ra xu hướng nào cho dữ liệu không? Vì có những khoa có hệ đào tạo đặc biệt như tài năng mà những khoa khác không có.

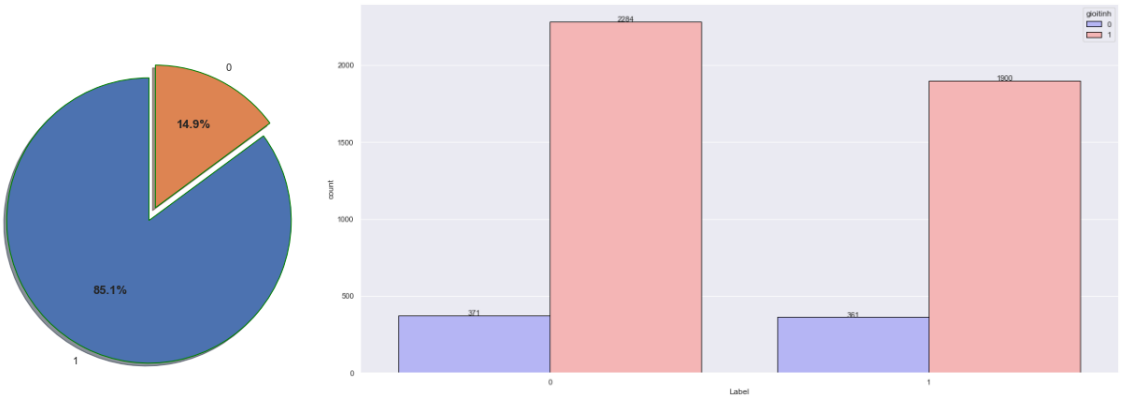


Đồ thị trên thống kê số lượng các sinh viên của các hệ đào tạo thuộc các khoa. Phía trên trái thuộc về nhãn 0 và bên phải thuộc về nhãn 1.

Dựa vào đồ thị nhóm nhìn thấy có một điểm chung giữa các khoa, đó là hệ CQUI sẽ có tỉ lệ đạt chuẩn quá trình AV cao hơn. Trong khi đó hệ CLC tỉ lệ này lại ngược lại. Vậy điều này cho thấy mối quan hệ của khoa và hệ đào tạo không ảnh hưởng gì đến output. Thuộc tính ảnh hưởng trong trường hợp này là hệ đào tạo.

Tuy nhiên có một ngoại lệ, đó chính là khoa KTTT toàn hệ CLC đều nhãn 0. Nhưng số lượng này rất nhỏ (chưa đến 50 sinh viên) nên ta có thể bỏ qua trường hợp này.

Thứ tám là sự ảnh hưởng của giới tính.



Mặc dù nam chiếm số lượng lớn do tính chất trường công nghệ thông tin nhưng tỉ lệ giới tính chia đều cho các nhãn.

Cuối cùng, để thống kế toàn bộ các thuộc tính, nhóm có một bảng thể hiện mối tương quan giữa các biến.

Từ đồ thị tương quan, ta thấy rằng các biến sau đây có mối quan hệ mạnh với Label: namsinh, hedt, khoahoc, mamh.

Còn lại còn biến: gioitinh, khoa, dien_tt, khu_vuc lại dường như rất ít ảnh hưởng với Label.

Dựa vào những phân tích trên và đồ thị tương quan, ta sẽ xem xét loại bỏ các biến gioitinh, khoa, dien_tt, khu_vuc khi xây dựng mô hình để quan sát mô hình có tăng hay giảm độ chính xác.

2.3.2 Phương pháp thực nghiệm

2.3.2.1 Chuẩn bị dữ liệu cho các mô hình học máy

Đầu tiên, trước khi xây dựng mô hình, ta cần tiến hành mã hóa các thuộc tính về dạng số. Với các thuộc tính hedt, dien_tt, khoa không mang tính thứ tự nên ta sẽ dùng kĩ thuật biến đổi các biến này thành các biến nhị phân giả, tức là mỗi biến sẽ được chuyển đổi thành nhiều biến nhị phân.

	hedt_CLC	hedt_CNTN	hedt_CQUI	hedt_CTTT	hedt_KSTN
1	0	0	1	0	0
2	0	0	1	0	0
3	0	0	1	0	0
4	0	0	1	0	0
5	1	0	0	0	0
...	...	...	...	...	...
4912	1	0	0	0	0
4913	0	0	1	0	0
4914	1	0	0	0	0
4915	0	0	1	0	0
4916	1	0	0	0	0

Còn với các biến mamh, khu_vuc, chuanav_1 mang tính thứ tự (ENG01 > AVSC1, khu vực 3 > khu vực 2, …) ta sẽ biến đổi thành dạng số 0, 1, 2, 3, 4, … tương ứng từ thấp đến cao vì như vậy giúp thuật toán sẽ hiểu rằng giữa biến 0 và 4 có khoảng cách xa nhau hơn là biến 0 và 1 (từ AVSC1 đến AVSC2 sẽ khác AVSC1 đến ENG03).

Sau khi đã mã hóa, nhóm sẽ phân chia dữ liệu thực nghiệm. Nhóm lựa chọn Holdout để chia bộ dữ liệu thành 2 tập train và test (8/2). Tiếp đến, nhóm sử dụng GridSearch để tìm bộ tham số thích hợp cho các mô hình. Khi đã chọn được bộ tham số phù hợp, dùng Kfold với mặc định chia thành 20 nhóm nhỏ để đánh giá mô hình trong quá trình học (do bộ dữ liệu khá nhỏ).

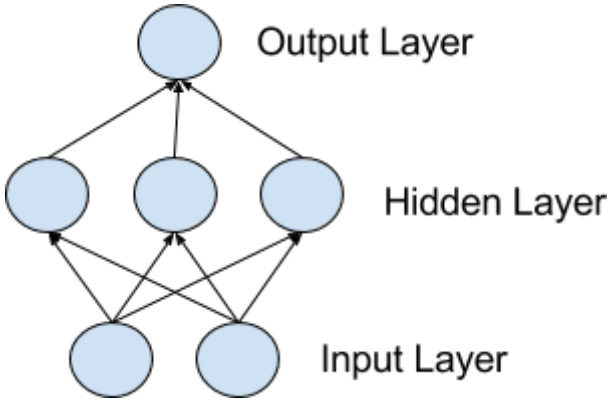
2.3.2.2 Xây dựng mô hình

Để có thể đánh giá và đưa ra kết luận chính xác nhất, nhóm lựa chọn 4 thuật toán phân lớp khác nhau để thực hiện huấn luyện.

- Logistic Regression: là thuật toán cơ bản nhất để giải quyết bài toán phân lớp, dùng hàm sigmoid để đưa ra đánh giá theo xác suất. [1]
- Support Vector Machine: là kĩ thuật cố gắng tối ưu các đường phân chia giữa các lớp, được hỗ trợ bởi nhiều kernel giúp thuật toán linh hoạt hơn. [2]
- AdaBoost: là thuật toán họ Boosting trong Ensemble learning và cũng là thuật toán giành chiến thắng nhiều cuộc thi trên Kaggle. Ý tưởng chính của thuật

toán này đánh trọng số cho các điểm dữ liệu từ đó tối ưu hóa các điểm sai do có trọng số lớn. [3]

- Multi-layer Perceptron (MLP): là thuật toán Neural Network cổ điển. Có thể gồm 1 hay nhiều layers. MLP thích hợp cho việc dự đoán với inputs được gán nhãn với bộ dữ liệu dạng bảng hơn là các thuật toán khác cùng họ (CNN – thích hợp cho dự đoán dữ liệu dạng hình ảnh với các điểm dữ liệu có mối quan hệ không gian, …) [4] [5]



Nhóm sẽ lần lượt tiến hành xây dựng các mô hình dự đoán với 2 giai đoạn:

- Giai đoạn 1: sinh viên chỉ mới thi anh văn đầu vào (bỏ thuộc tính chuanav_1). Giai đoạn này có thể tỉ lệ dự đoán đúng chưa cao do vẫn chưa đủ thông tin.
    - Sử dụng tất cả thuộc tính.
    - Sử dụng thuộc tính được phân tích là có liên quan đến nhãn.

- Giai đoạn 2: đã có kết quả tiến trình học tập ngoại ngữ của sinh viên (thêm thuộc tính chuanav_1). Nhóm kì vọng giai đoạn này tỉ lệ đúng sẽ cao hơn giai đoạn 1.
    - Sử dụng tất cả thuộc tính.
    - Sử dụng thuộc tính được phân tích là có liên quan đến nhãn.

Dưới đây là bộ tham số tối ưu nhất sau kho nhóm sử dụng GridSearch.

	GĐ1 - 1	**GĐ1 - 2**	**GĐ2 - 1**	**GĐ2 - 2**
Logistic Regression	C=0.1, class_weight=balanced, max_iter=3000, multi_class=ovr	C=0.1, class_weight=balanced, max_iter=3000, multi_class=ovr	C=0.01, class_weight=balanced, max_iter=3000, multi_class=ovr	C=0.01, class_weight=balanced, max_iter=3000, multi_class=ovr

SVM	C=0.1, gamma=0.001, kernel=linear, probability=True	C=0.1, gamma=0.1, kernel=rbf, probability=True	C=0.1, gamma=0.1, kernel=rbf, probability=True	C=0.1, gamma=0.001, kernel=linear, probability=True
AdaBoost	DecisionTreeClassifier(criterion=gini), learning_rate=0.1, n_estimators=500	DecisionTreeClassifier (criterion=entropy), learning_rate=0.001	DecisionTreeClassifier (criterion=entropy), n_estimators=10 learning_rate=0.01	DecisionTreeClassifier(criterion=gini), learning_rate=0.001
MLP	activation=identity, early_stopping=True, max_iter=800, solver=lbfgs	activation=identity, early_stopping=True, max_iter=800, solver=lbfgs	activation=identity, early_stopping=True, max_iter=800. solver=lbfgs	activation=identity, early_stopping=True, max_iter=800. solver=lbfgs

1) Giai đoạn 1

Kết quả trên tập test

	Models Used	Accuracy	Precision	Recall	F1
0	LogisticRegression	69.817073	72.160356	65.322581	68.571429
1	Support Vector Machine	70.731707	66.815145	68.337130	67.567568
2	AdaBoost	69.918699	68.596882	66.522678	67.543860
3	MLP	70.731707	66.146993	68.591224	67.346939

Toàn bộ thuộc tính

	Models Used	Accuracy	Precision	Recall	F1
0	LogisticRegression	72.154472	73.719376	67.967146	70.726496
1	Support Vector Machine	71.443089	66.146993	69.718310	67.885714
2	AdaBoost	72.764228	66.815145	71.599045	69.124424
3	MLP	71.951220	68.151448	69.703872	68.918919

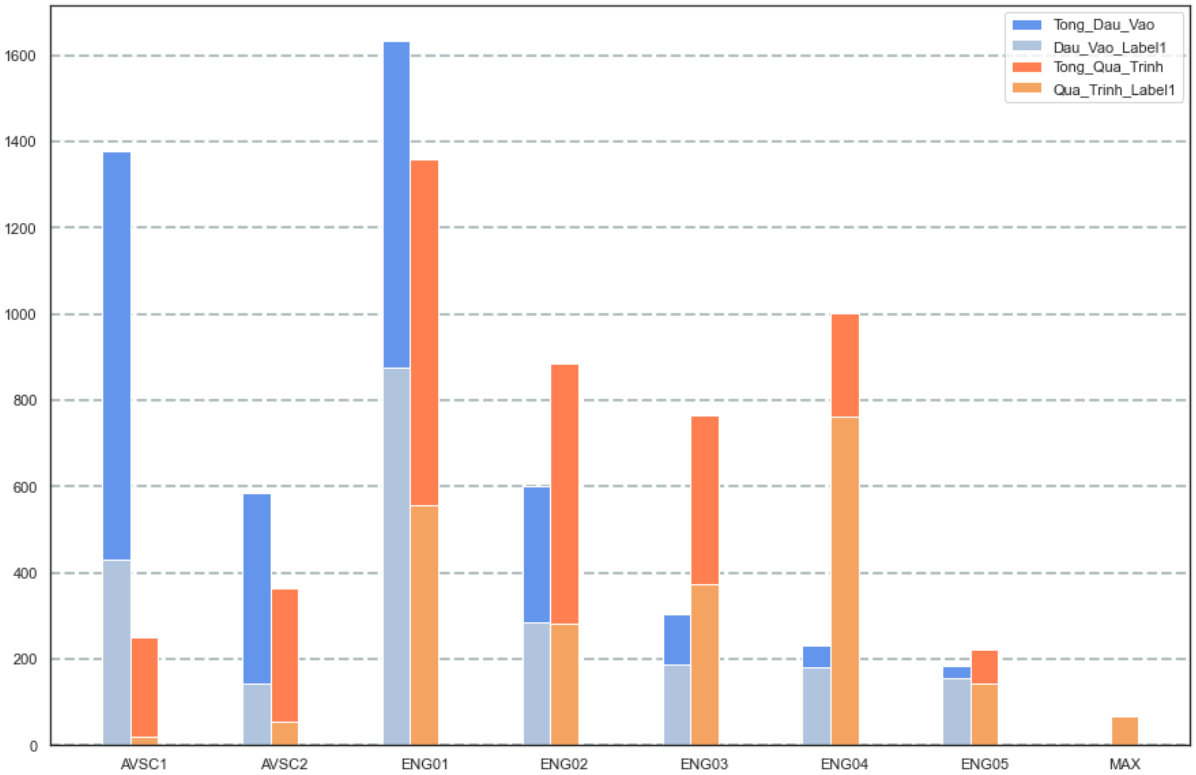
Các thuộc tính có liên quan

Khi so sánh 2 bảng kết quả chạy trên tập test trên, nhóm thấy rằng khi sử dụng đúng một số thuộc tính có nhiều ảnh hưởng lên label sẽ làm kết quả mô hình cải thiện. Điều này chứng tỏ phân tích của nhóm là chính xác về sự liên quan của các thuộc tính.

2) Giai đoạn 2

Thêm vào dữ liệu thuộc tính xếp loại AV đầu vào để tiến hành xây dựng mô hình giai đoạn 2.

Khi thêm thuộc tính xếp loại AV gần nhất vào, nhóm sẽ tiến hành một bước nhỏ phân tích dữ liệu đã thêm vào.



Sự phân bố tiếng anh có sự thay đổi đáng kể. Có sự chuyển đổi rõ rệt khi các bạn thi đầu vào AVSC1 đã tăng bậc và chuyển thành cột ENG01. Tương tự khi các bạn thi đầu vào được ENG01 đã phân bố lại thành ENG02, ENG03, ENG04. Điều này cho thấy sự hiệu quả trong quá trình học anh văn của các bạn sinh viên và chắc chắn sẽ ảnh hưởng đến nhãn cuối cùng của bài toán.

Ngoài ra, khi nhìn vào tỉ lệ label của các cột quá trình. Ta thấy rằng, khi các bạn đã học anh văn rồi (hoặc không học anh văn) nhưng vẫn ở mức AVSC1, AVSC2, ENG01, ENG02 thì có tỉ lệ không đạt chuẩn rất cao (AVSC1 gần như là tuyệt đối).

Đồng thời qua heat map giữ các thuộc tính, nhóm thấy rằng thuộc tính xếp loại AV gần nhất (chuanav_1) có mối quan hệ mạnh mẽ với Label. Từ đó suy ra dữ liệu khi có thêm thuộc tính chuanav_1 sẽ cải thiện tỉ lệ chính xác của mô hình.

Kết quả trên tập test

	Models Used	Accuracy	Precision	Recall	F1
0	LogisticRegression	72.662602	78.396437	67.175573	72.353546
1	Support Vector Machine	73.272358	77.282851	68.307087	72.518286
2	AdaBoost	74.898374	67.483296	75.000000	71.043376
3	MLP	75.203252	59.242762	81.345566	68.556701

Toàn bộ thuộc tính

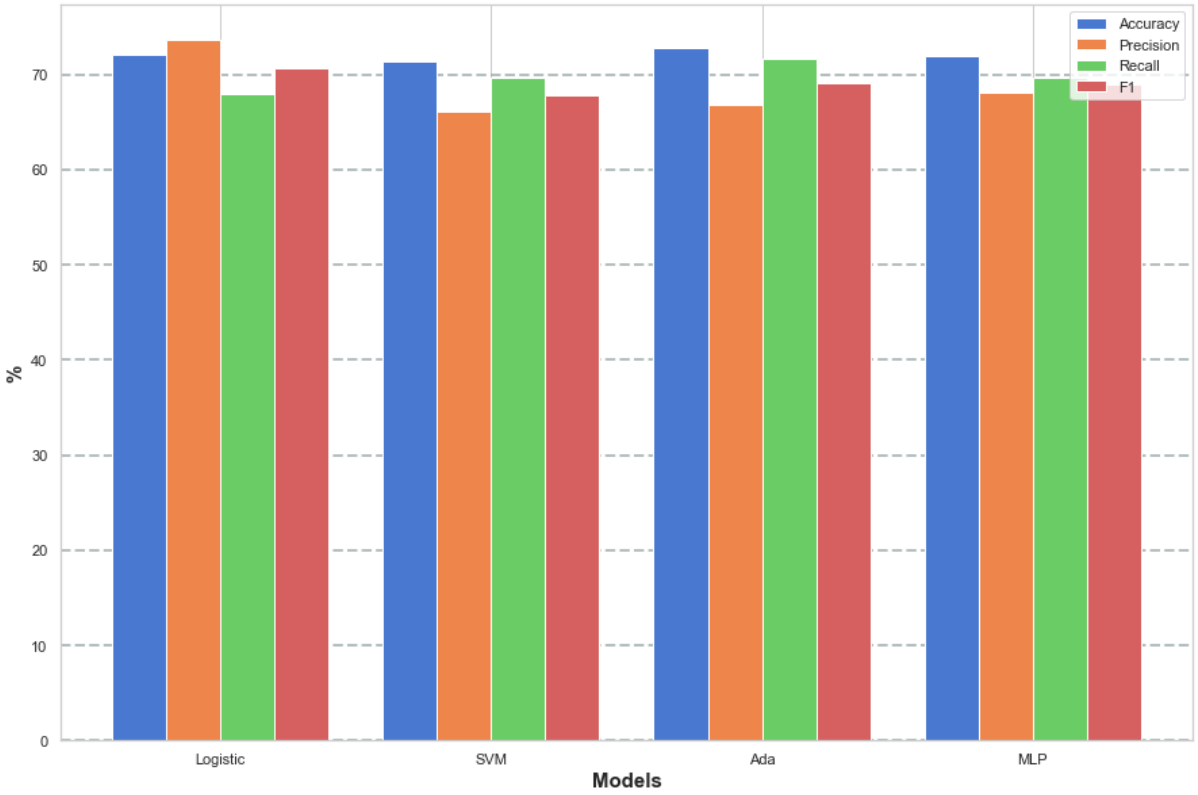
	Models Used	Accuracy	Precision	Recall	F1
0	LogisticRegression	73.272358	76.614699	68.525896	72.344900
1	Support Vector Machine	74.085366	71.937639	71.460177	71.698113
2	AdaBoost	77.845528	65.701559	82.172702	73.019802
3	MLP	73.780488	75.055679	69.772257	72.317597

Các thuộc tính có liên quan

2.3.2.3 Đánh giá mô hình

Để đánh giá mô hình cũng như dữ liệu của nhóm, nhóm sẽ đánh giá dựa trên thang đo Accuracy, Precision, Recall và F1 như đã trình bày phía trên.

Với mục tiêu chính của nghiên cứu này sẽ giúp phòng đào tạo cảnh báo những trường hợp sinh viên sẽ không đạt chuẩn quá trình ngoại ngữ (label 0). So với label 1, ta sẽ ưu tiên việc dự đoán chính xác label 0, ngoài tỉ lệ dự đoán đúng lớp 0, ta cần quan tâm đến việc giảm thiểu trường hợp label là 0 nhưng ta lại dự đoán là 1 (phòng đào tạo sẽ không cảnh báo). Vì vậy, khi cần thiết, ta cần ưu tiên các mô hình có độ precision cao.



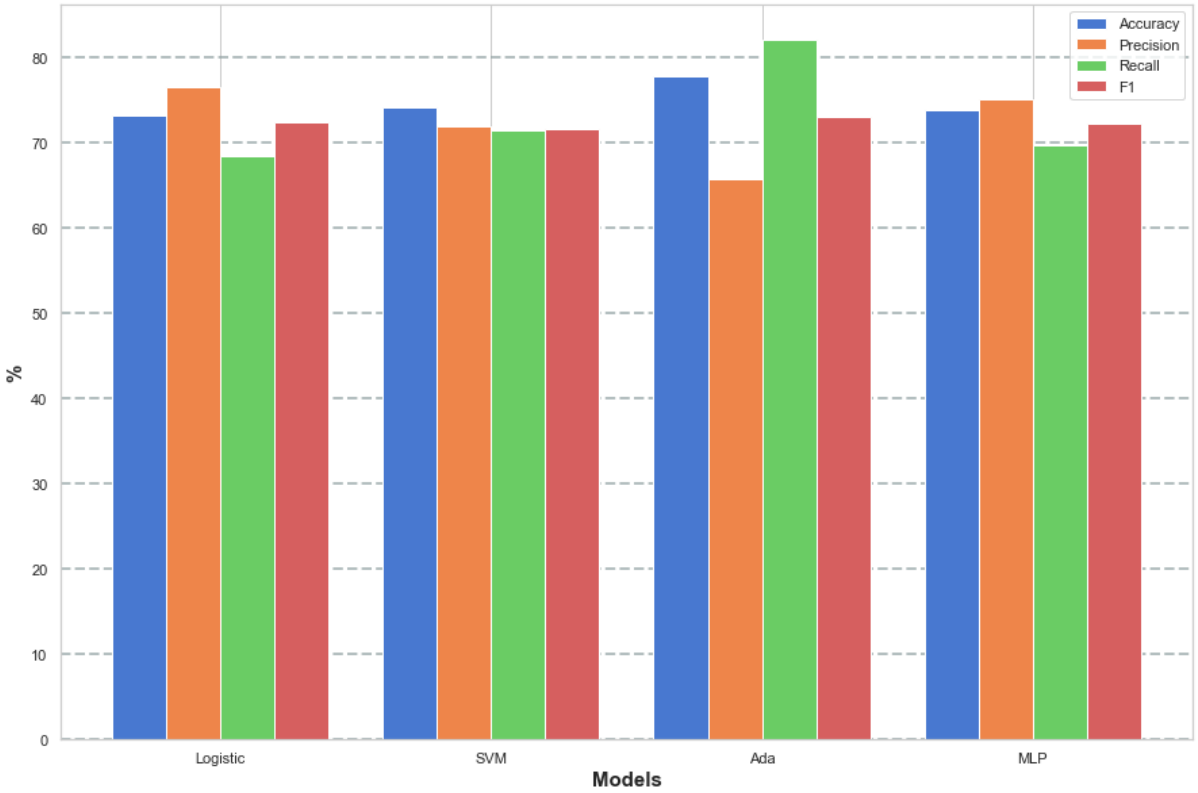
Giai đoạn 1 với thuộc tính có liên quan

Accuracy: cao nhất là mô hình AdaBoost, tuy nhiên nhìn chung không có sự chênh lệnh rõ rệt giữa các mô hình (~1%)

Precision: nhìn chung các mô hình đều thấp ở thang đo này ngoại trừ Logistic.

Recall: có sự đánh đổi với precision, các mô hình thấp ở precision đều đạt recall cao.

F1: chỉ có Logistic đạt ngưỡng 70%.



Giai đoạn 1 với thuộc tính có liên quan

Accuracy: có sự tăng lên rõ rệt khi Ada ~ 78% (tăng gần 5% so với ban đầu).

Precision: phần lớn các mô hình đêu tăng precision, tuy nhiên Ada lại giảm.

Recall: Ada đạt hơn 80%.

F1: các mô hình nhìn chung không chêch lệnh với nhau.

Qua đó, thuộc tính chuẩn anh văn 1 có ảnh hưởng đến hiệu quả của các mô hình. Hầu hết các chỉ số mô hình đều tăng đáng kể (~3% -> 5%). Dựa vào nhận định của nhóm, mô hình Logistic sẽ đạt hiệu quả cao nhất so với các mô hình còn lại, vì các chỉ số đều đạt mức ổn không thấp hơn quá nhiều và đạt được sự ổn định ở thang đo precision. Riêng với trường hợp Ada, do tỉ lệ của precision quá thấp (~65%) nên mô hình này có xu hướng sẽ dự đoán các sinh viên chưa đạt chuẩn thành đạt chuẩn cao.

2.4 Demo

Nhóm tiến hành kiểm tra phân lớp một số mẫu dữ liệu mà nhóm đã thu thập với 2 mô hình: Logistic và AdaBoost.

	namsinh	khoahoc	mamh	chuanav_1	hedt_CLC	hedt_CNTN	hedt_CQUI	hedt_CTTT	hedt_KSTN
1	2002	15	0	4	0	0	1	0	0
2	2002	15	4	4	0	0	1	0	0
3	2002	15	4	5	0	0	0	1	0
4	2000	13	0	3	1	0	0	0	0
5	2002	16	0	3	0	0	1	0	0
6	2001	15	3	6	0	0	0	1	0

2.4.1. Giai đoạn 1

Cả 2 mô hình đều cho kết quả dự đoán giống nhau, đó là chỉ có sinh viên thứ 2 đạt chuẩn quá trình, còn lại đều không đạt. Điều này khá hợp lí vì chỉ có sinh viên 2 thi chạm ngưỡng quá trình (ENG03) còn lại cách khá xa ngưỡng.

2.4.2. Giai đoạn 2

2 mô hình cũng cho kết quả dự đoán giống nhau, đó là sinh viên 2, 3 và 6 sẽ đạt chuẩn quá trình, còn lại đều không đạt. Qua kết quả kiểm tra, nhóm thấy rằng trong cùng một điều kiện, mô hình sẽ có xu hướng thiên vị cho các điểm dữ liệu có khoảng cách nhỏ hơn giữa mamh đến chuanav_1 (sinh viên 1 và sinh viên 2). Giải thích cho vấn đề này là do sự thiếu dữ liệu dẫn đến việc biểu diễn dữ liệu sai, tức là việc biểu diễn sinh viên 2 học từ ENG03 đến ENG03 thì không có nghĩa rằng sinh viên đó không có sự phát triển trong học ngoại ngữ (sinh viên học bên ngoài và không nhập thông tin vào cơ sở dữ liệu). Nếu xét cùng một khoảng thời gian, sinh viên 1 học từ AVSC1 đến ENG03 và sinh viên 2 sẽ học từ ENG03 đến ENG0x ($x \geq 3$).

KẾT LUẬN

Qua các phân tích trên và vô số lần đánh giá, nhóm đã đút kết rằng: Nếu ngoại trừ các yếu tố chủ quan như sự cần cù, nỗ lực, … của sinh viên thì các yếu tố khách quan cũng sẽ giúp đánh giá trình độ tiếng Anh quá trình của sinh viên liệu có đạt chuẩn tương đối chính xác. Các yếu tố này có ảnh hưởng tới kết quả output và output có thể thay đổi khi sinh viên học tiếng Anh tại trường hay cập nhật thường xuyên trình độ tiếng Anh của bản thân.

Nhóm đánh giá mô hình đã hoạt động tương đối tốt so với dự kiến ban đầu. Tuy nhiên các mô hình còn gặp nhiều hạn chế như điểm chuẩn quá trình anh văn của các sinh viên không cùng một mốc thời gian với nhau, thiếu thông tin do sinh viên học ngoại ngữ ở bên ngoài. Để giải quyết vấn đề này, nhóm đề xuất sẽ thu thập thêm dữ liệu như quá trình học tiếng Anh trong và ngoài trường, quá trình học tiếng Anh của sinh viên từng các kì, … hoặc áp dụng các thuật toán phân lớp đánh giá tốt hơn, từ đó có thể đánh giá đúng mức trình độ Anh văn thay vì chỉ đánh giá có hay không sinh viên đạt chuẩn.

BẢNG PHÂN CÔNG CÔNG VIỆC

Thành viên	MSSV	Công việc	Mức độ hoàn thành
Mai Duy Ngọc	20520654	Phân tích và làm sạch bảng dữ liệu: diem, sinhvien; Lập bảng diem_av; Trực quan hóa dữ liệu và phân tích tương quan 2 thuộc tính; Xây dựng mô hình 4 thuật toán Logistic, SVM, AdaBoost và đánh giá (chính);	100%
Trần Đăng Khoa	20520589	Phân tích và làm sạch bảng dữ liệu: diem_Thu; Chương 2 thuyết minh; Trực quan hóa dữ liệu và phân tích tương quan 2 thuộc tính; Slide thuyết trình (chính), Trình bày đồ án	100%
Đào Danh Đăng Phụng	20520699	Phân tích và làm sạch bảng dữ liệu: sinhvien_chungchi; Xây dựng mô hình 4 thuật toán Logistic, SVM, AdaBoost và đánh giá (phụ), Kiểm tra và quản lí kế hoạch	100%
Đặng Phước Sang	21521377	Phân tích và làm sạch bảng dữ liệu: thisinh, xeploai; Chương 1, Tổng kết thuyết minh; Trực quan hóa dữ liệu và phân tích tương quan 4 thuộc tính; Slide thuyết trình (phụ)	100%

TÀI LIỆU THAM KHẢO

[1] scikit-learn developers, “Linear Models”, 2023. [Trực tuyến]. Địa chỉ: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression [Truy cập 23/5/2023]

[2] scikit-learn developers, “Support Vector Machines”, 2023. [Trực tuyến]. Địa chỉ: https://scikit-learn.org/stable/modules/svm.html [Truy cập 23/5/2023]

[3] scikit-learn developers, “Support Vector Machines”, 2023. [Trực tuyến]. Địa chỉ: https://scikit-learn.org/stable/modules/svm.html [Truy cập 23/5/2023]

[4] scikit-learn developers, “AdaBoostClassifier”, 2023. [Trực tuyến]. Địa chỉ: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html#sklearn.ensemble.AdaBoostClassifier [Truy cập 23/5/2023]

[5] Jason Brownlee, “When to Use MLP, CNN, and RNN Neural Networks”,2022. [Trực tuyến]. Địa chỉ: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/?fbclid=IwAR3zlWV1CaP7CnqR53PVQYu6U8gDYsw4uTmmxF8WgIYfJbW7LtBVgvyrQTw [Truy cập 23/5/2023]