

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KINH TẾ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH

UEH UNIVERSITY

ĐỒ ÁN MÔN HỌC
BIỂU DIỄN TRỰC QUAN DỮ LIỆU

**Đề tài: Phân tích và biểu diễn trực quan dữ liệu về tình hình
học tập của sinh viên.**

Thành viên: Nguyễn Đình Toàn
Nguyễn Thị Phương Thảo
Lý Gia Thuận
Phạm Quốc Thuận
Lưu Trọng Tốt

Chuyên ngành: Khoa học dữ liệu

Lớp: DS001

Giảng viên: TS. Nguyễn An Tê

Thành phố Hồ Chí Minh , ngày 23 tháng 11 năm 2023

MỤC LỤC

LỜI CẢM ƠN.....	1
CHƯƠNG I : TỔNG QUAN ĐỀ TÀI.....	2
1.1 Giới thiệu về bộ dữ liệu.....	2
1.2 Mục tiêu nghiên cứu.....	2
1.3 Phương pháp nghiên cứu.....	2
1.4 Tài nguyên sử dụng.....	3
CHƯƠNG II: TỔNG QUAN BỘ DỮ LIỆU.....	4
2.1 Tổng quan bộ dữ liệu thu thập.....	4
2.2 Các thuộc tính của bộ dữ liệu.....	4
CHƯƠNG III: TIỀN XỬ LÝ DỮ LIỆU.....	9
3.1 Exploratory Data Analysis (EDA).....	9
3.1.1 Chính dạng bộ dữ liệu.....	12
a, Thay đổi tên cột.....	12
b, Gộp, xóa cột.....	12
3.1.2 Phân tích sơ bộ bộ dữ liệu.....	13
3.2 Xử lý dữ liệu bị thiếu (Missing Values).....	15
3.3 Phân loại dữ liệu.....	19
3.4 Xử lý outliers.....	22
3.5 Biểu diễn lại dữ liệu sau khi xử lý.....	23
CHƯƠNG IV: BIỂU DIỄN TRỰC QUAN DỮ LIỆU.....	26
4.1 Biểu diễn theo nhân khẩu học.....	26
4.2 Biểu diễn theo tình hình kinh tế - xã hội.....	32
4.3 Biểu diễn theo tình hình học tập.....	39
CHƯƠNG V: KIỂM ĐỊNH.....	47
5.1 Nhân khẩu học.....	47
5.2 Tình hình kinh tế - xã hội.....	52
5.3 Tình hình học tập.....	53
CHƯƠNG VI: XÂY DỰNG MÔ HÌNH DỰ BÁO - PCA.....	57
6.1 Xây dựng mô hình:.....	57
6.1.2 Chạy thử với mô hình.....	58
6.1.3 Sử dụng thư viện `GridSearchCV` để tối ưu hóa mô hình.....	59
6.2 Chạy các thuật toán dự báo.....	61
6.2.1 Đánh giá với mục tiêu dự báo.....	62
6.3 Giảm chiều dữ liệu.....	63
6.3.1 Tính lại hiệu suất các mô hình sau khi giảm chiều.....	64
6.3.2 Áp dụng vào bài toán dự báo học sinh nghỉ học.....	66
6.4 Kết luận.....	67
PHỤ LỤC.....	69

BẢNG PHÂN CÔNG NHIỆM VỤ

STT	Họ và tên	Công việc phụ trách	Đánh giá
1.	Nguyễn Thị Phương Thảo	Tổng quan bộ dữ liệu, kiểm định, xây dựng mô hình dự báo, điều chỉnh định dạng Word, PPT.	100%
2.	Lý Gia Thuận	Tiền xử lý, phân tích tổng quan dữ liệu, biểu diễn trực quan, tổng hợp source code.	100%
3.	Phạm Quốc Thuận	Biểu diễn trực quan, nhận xét, điều chỉnh định dạng Word.	100%
4.	Nguyễn Đình Toàn	Tiền xử lý, phân tích tổng quan dữ liệu, kiểm định, xây dựng mô hình dự báo, điều chỉnh định dạng Word, PPT.	100%
5.	Lưu Trọng Tốt	Tổng quan đề tài, nhận xét một số biểu đồ.	40%

Lời cảm ơn

Chúng em xin gửi lời cảm ơn chân thành nhất đến thầy Nguyễn An Tế về sự hỗ trợ và hướng dẫn tận tình trong suốt quá trình học môn "Biểu diễn trực quan dữ liệu". Dưới sự dẫn dắt của thầy, chúng em đã có cơ hội tiếp xúc và ứng dụng những kiến thức mới mẻ và kỹ năng liên quan đến biểu diễn trực quan dữ liệu.

Qua môn học này, chúng em đã có cơ hội nắm bắt những khái niệm cơ bản và nâng cao về việc biểu diễn dữ liệu một cách sinh động và dễ hiểu. Nhờ sự hỗ trợ nhiệt tình và những lời khuyên hữu ích của thầy, tụi em đã có thể hoàn thành được báo cáo Đồ án kết thúc môn học này.

Trong quá trình làm đồ án vẫn còn các hạn chế, sai sót, chưa tối ưu về mặt kiến thức và kỹ năng. Nhóm chúng em mong nhận được sự phản hồi, nhận xét của thầy để có thể cải thiện được báo cáo tốt hơn nữa.

CHƯƠNG I : TỔNG QUAN ĐỀ TÀI

1.1 Giới thiệu về bộ dữ liệu

Bộ dữ liệu "Predict students' dropout and academic success" cung cấp cái nhìn toàn diện về sinh viên theo học các chương trình đại học khác nhau được cung cấp tại một cơ sở giáo dục. Nó bao gồm dữ liệu nhân khẩu học, các yếu tố kinh tế xã hội và thông tin về kết quả học tập có thể được sử dụng để phân tích các yếu tố dự đoán liên quan đến tình trạng học sinh bỏ học và thành công trong học tập. Tập dữ liệu bao gồm nhiều cơ sở dữ liệu riêng biệt chứa thông tin liên quan có sẵn tại thời điểm đăng ký, chẳng hạn như trạng thái đăng ký, tình trạng hôn nhân, các khóa học đã chọn, v.v. Ngoài ra, dữ liệu này có thể được sử dụng để ước tính kết quả học tập tổng thể của sinh viên vào cuối mỗi học kỳ bằng cách đánh giá các đơn vị học tập về tín chỉ/ghi danh/danh giá/phê duyệt, cùng với điểm số tương ứng của chúng. Cuối cùng, tập dữ liệu bao gồm tỉ lệ thất nghiệp, tỉ lệ lạm phát và GDP trong khu vực, có thể giúp hiểu được các yếu tố kinh tế ảnh hưởng như thế nào đến tỉ lệ bỏ học hoặc thành công trong học tập của sinh viên. Công cụ phân tích mạnh mẽ này cung cấp những hiểu biết có giá trị về động cơ thúc đẩy sinh viên ở lại trường hoặc bỏ học, theo đuổi các lĩnh vực nghiên cứu khác nhau như nông nghiệp, thiết kế, giáo dục, điều dưỡng, quản lý báo chí, dịch vụ xã hội hoặc công nghệ.

1.2 Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của bộ dữ liệu "Predict students' dropout and academic success" là tạo điều kiện hiểu biết toàn diện về các yếu tố khác nhau ảnh hưởng đến việc giữ chân học sinh và thành tích học tập trong giáo dục đại học. Bằng cách cung cấp cái nhìn chi tiết về nhân khẩu học, hoàn cảnh kinh tế xã hội và kết quả học tập của sinh viên, bộ dữ liệu này nhằm mục đích cho phép các tổ chức giáo dục và các nhà hoạch định chính sách xác định các yếu tố có thể dự báo được tỉ lệ nghỉ học của sinh viên. Với cách phân tích thông tin này, bộ dữ liệu tìm cách hỗ trợ phát triển các chiến lược và biện pháp can thiệp hiệu quả nhằm nâng cao tỷ lệ giữ chân học sinh, kết quả học tập và kết quả giáo dục tổng thể. Ngoài ra, tập dữ liệu này nhằm mục đích làm sáng tỏ tác động của các yếu tố kinh tế đối với lộ trình học tập của sinh viên, từ đó cho phép các tổ chức thực hiện các biện pháp hỗ trợ có mục tiêu nhằm cải thiện về việc học của sinh viên, thành tích của sinh viên trong công việc học tập trong giáo dục đại học và giảm tỷ lệ bỏ học.

1.3 Phương pháp nghiên cứu

- EDA: Sử dụng các biểu đồ nhằm tương quan cũng như làm rõ mục đích nghiên cứu đề tài, sự liên kết với nhau giữa các biến.
- Trực quan hóa dữ liệu: Sử dụng các loại biểu đồ chuyên dụng và phù hợp với mục đích trực quan hóa các dữ liệu, giúp người đọc báo cáo dễ dàng quan sát và đánh giá kết quả phân tích.
- Kiểm định Chi-Square: Kiểm định tính độc lập giữa 2 biến phân loại, xác định xem liệu có mối liên hệ giữa 2 biến phân loại hay không.
- Kiểm định T-test: Kiểm định giá trị trung bình của một biến liên tục giữa hai nhóm độc lập.
- Kiểm định Anova: Kiểm định giá trị trung bình của một biến liên tục giữa ba hoặc nhiều nhóm độc lập.
- PCA: Giảm chiều dữ liệu, thông qua việc phép chiếu dữ liệu từ không gian đa chiều ban đầu xuống một không gian chiều thấp hơn, từ đó giúp bảo toàn thông tin quan trọng trong dữ liệu, đồng thời loại bỏ hoặc giảm thiểu những chiều dữ liệu ít quan trọng.

1.4 Tài nguyên sử dụng

- Công cụ: Google Colab
- Ngôn ngữ lập trình: Python.
- Bộ dữ liệu “Predict students' dropout and academic success” từ nền tảng Kaggle.

CHƯƠNG II: TỔNG QUAN BỘ DỮ LIỆU

2.1 Tổng quan bộ dữ liệu thu thập

Bộ dữ liệu “Predict students' dropout and academic success” chứa dữ liệu về sinh viên của một tổ chức giáo dục, bao gồm dữ liệu về nhân khẩu học, kinh tế xã hội, và kết quả học tập.

Bộ dữ liệu bao gồm 35 thuộc tính và 4424 dòng.

2.2 Các thuộc tính của bộ dữ liệu

STT	Tên thuộc tính	Ý nghĩa	Mô tả
1	Marital status (Categorical)	Tình trạng hôn nhân của sinh viên	Bao gồm 6 tình trạng được đánh số theo thứ tự từ 1 đến 6: 1 = Độc thân, 2 = Đã kết hôn, 3 = Góa phụ, 4 = Ly hôn, 5 = Sống thử, 6 = Ly thân.
2	Application mode (Categorical)	Phương thức nhập học của sinh viên	Bao gồm 18 phương thức nhập học của sinh viên được đánh số theo thứ tự từ 1 đến 18.
3	Application order (Numerical)	Thứ tự đăng ký của sinh viên	Bao gồm 6 thứ tự đăng ký được đánh số theo thứ tự từ 1 đến 6.
4	Course (Categorical)	Ngành học của sinh viên	Bao gồm 17 ngành học, được đánh số theo thứ tự từ 1 đến 17: 1= Công nghệ sản xuất nguyên liệu sinh học, 2 = Thiết kế hoạt hình và đa phương tiện, 3 = Dịch vụ xã hội (học buổi tối), 4 = Nông học, 5 = Thiết kế truyền thông, 6 = Thú y, 7 = Kỹ thuật máy tính, 8 = Chăn nuôi ngựa, 9 = Quản trị, 10 = Công tác xã hội, 11 = Du lịch, 12 = Điều dưỡng, 13 = Vệ sinh nha khoa, 14 = Quảng trị quảng cáo và tiếp thị, 15 = Báo chí truyền thông, 16 = Giáo dục căn bản, 17 = Quản trị (học buổi tối).
5	Daytime/evening attendance (Categorical)	Sinh viên học ban ngày hoặc buổi tối	Bao gồm 2 khoảng thời gian được đánh số 0, 1: Buổi tối, Ban ngày.
6	Previous qualification (Categorical)	Trình độ trước đó	Bao gồm 7 trình độ được đánh số theo thứ tự từ 1 đến 7: 1 = Trung học phổ thông, 2 = Đại học chính quy, 3 = Bằng cấp đại học, 4 =

			Thạc sĩ, 5 = Tiến sĩ, 6 = Tỷ lệ giáo dục đại học, 7 = Lớp 12 - chưa tốt nghiệp
7	Nationality (Categorical)	Quốc tịch của sinh viên	Bao gồm 21 quốc gia được đánh số theo thứ tự từ 1 đến 21: 1 = Portuguese, 2 = German, 3 = Spanish, 4 = Italian, 5 = Dutch, 6 = English, 7 = Lithuanian, 8 = Angolan, 9 = Cape Verdean, 10 = Guinea, 11 = Mozambican, 12 = Santomean, 13 = Turkish, 14 = Brazilian, 15 = Romanian, 16 = Moldovan, 17 = Mexican, 18 = Ukrainian, 19 = Russian, 20 = Cuban, 21 = Colombian.
8	Mother's qualification (Categorical)	Nghề nghiệp của mẹ sinh viên	Bao gồm 46 ngành nghề được đánh số thứ tự từ 1 đến 46.
9	Father's qualification (Categorical)	Nghề nghiệp của bố sinh viên	Bao gồm 46 ngành nghề được đánh số thứ tự từ 1 đến 46.
10	Displaced (Categorical)	Sinh viên có phải là người di dời hay không	Bao gồm 2 tình trạng có hoặc không được đánh số 1, 0: 1 = có, 0 = không
11	Educational special needs (Categorical)	Sinh viên có nhu cầu giáo dục đặc biệt nào không	Bao gồm 2 tình trạng có hoặc không được đánh số 1, 0: 1 = có, 0 = không
12	Debtor (Categorical)	Sinh viên có nợ không	Bao gồm 2 tình trạng có hoặc không được đánh số 1, 0: 1 = có, 0 = không
13	Tuition fees up to date (Categorical)	Học phí của sinh viên có được cập nhật hay không.	Bao gồm 2 tình trạng có hoặc không được đánh số 1, 0: 1 = có, 0 = không
14	Gender (Categorical)	Giới tính của sinh viên.	Bao gồm 2 giới tính: 1= Nam, 0= Nữ

15	Scholarship holder (Categorical)	Sinh viên có nhận học bổng hay không.	Bao gồm 2 tình trạng có hoặc không được đánh số 1, 0: 1 = có, 0 = không
16	Age at enrollment (Numerical)	Tuổi của sinh viên vào thời điểm nhập học.	Bao gồm độ tuổi của các sinh viên
17	International (Categorical)	Sinh viên có phải là sinh viên quốc tế hay không.	Bao gồm 2 tình trạng có hoặc không được đánh số 1, 0: 1 = có, 0 = không
18	Curricular units 1st sem (credited) (Numerical)	Các đơn vị học phần học kỳ 1 (đã được tính tín chỉ)	Số lượng đơn vị học phần
19	Curricular units 1st sem (enrolled) (Numerical)	Các đơn vị học phần học kỳ 1 (đã đăng ký)	Số lượng đơn vị học phần
20	Curricular units 1st sem (evaluations) (Numerical)	Các đơn vị học phần học kỳ 1 (đã đánh giá)	Số lượng đơn vị học phần
21	Curricular units 1st sem (approved) (Numerical)	Các đơn vị học phần học kỳ 1 (đã được chấp thuận)	Số lượng đơn vị học phần
22	Curricular units 1st sem (grade) (Numerical)	Điểm các đơn vị học phần học kỳ 1	Điểm số
23	Curricular units 1st sem (without)	Các đơn vị học phần học	Số lượng đơn vị học phần

	evaluations) (Numerical)	kỳ 1 (chưa đánh giá)	
24	Curricular units 2nd sem (credited) (Numerical)	Các đơn vị học phần học kỳ 2 (đã được tính tín chỉ)	Số lượng đơn vị học phần
25	Curricular units 2nd sem (enrolled) (Numerical)	Các đơn vị học phần học kỳ 2 (đã đăng ký)	Số lượng đơn vị học phần
26	Curricular units 2nd sem (evaluations) (Numerical)	Các đơn vị học phần học kỳ 2 (đã đánh giá)	Số lượng đơn vị học phần
27	Curricular units 2nd sem (approved) (Numerical)	Các đơn vị học phần học kỳ 2 (đã được chấp thuận)	Số lượng đơn vị học phần
28	Curricular units 2nd sem (grade) (Numerical)	Điểm các đơn vị học phần học kỳ 2	Điểm số
29	Curricular units 2nd sem (without evaluations) (Numerical)	Các đơn vị học phần học kỳ 2 (chưa đánh giá)	Số lượng đơn vị học phần
30	Unemployment rate (Numerical)	Tỷ lệ thất nghiệp	Bao gồm các giá trị trong khoảng [7.6; 16.2]
31	Inflation rate (Numerical)	Tỷ lệ lạm phát	Bao gồm các giá trị trong khoảng [-0.8; 3.7]

32	GDP (Numerical)	GDP	Bao gồm các giá trị trong khoảng [-4.06; 3.51]
33	Target (Categorical)	Mục tiêu	Bao gồm 3 trạng thái: Bỏ học, Tốt nghiệp, Đang học

Bảng 1: Các thuộc tính của bộ dữ liệu

CHƯƠNG III: TIỀN XỬ LÝ DỮ LIỆU

3.1 Exploratory Data Analysis (EDA)

Để thăm dò bộ dữ liệu, ta cần biết được tổng quan các thông tin về: số dòng, số cột, có tồn tại giá trị bị thiếu hay không, nếu có thì ở dòng nào, thuộc cột nào và chiếm bao nhiêu phần trăm của bộ dữ liệu.

Xem số dòng, số cột hiện có của bộ dữ liệu nguyên bản để nắm được các thông tin sơ lược trước khi tiến hành tiền xử lý:

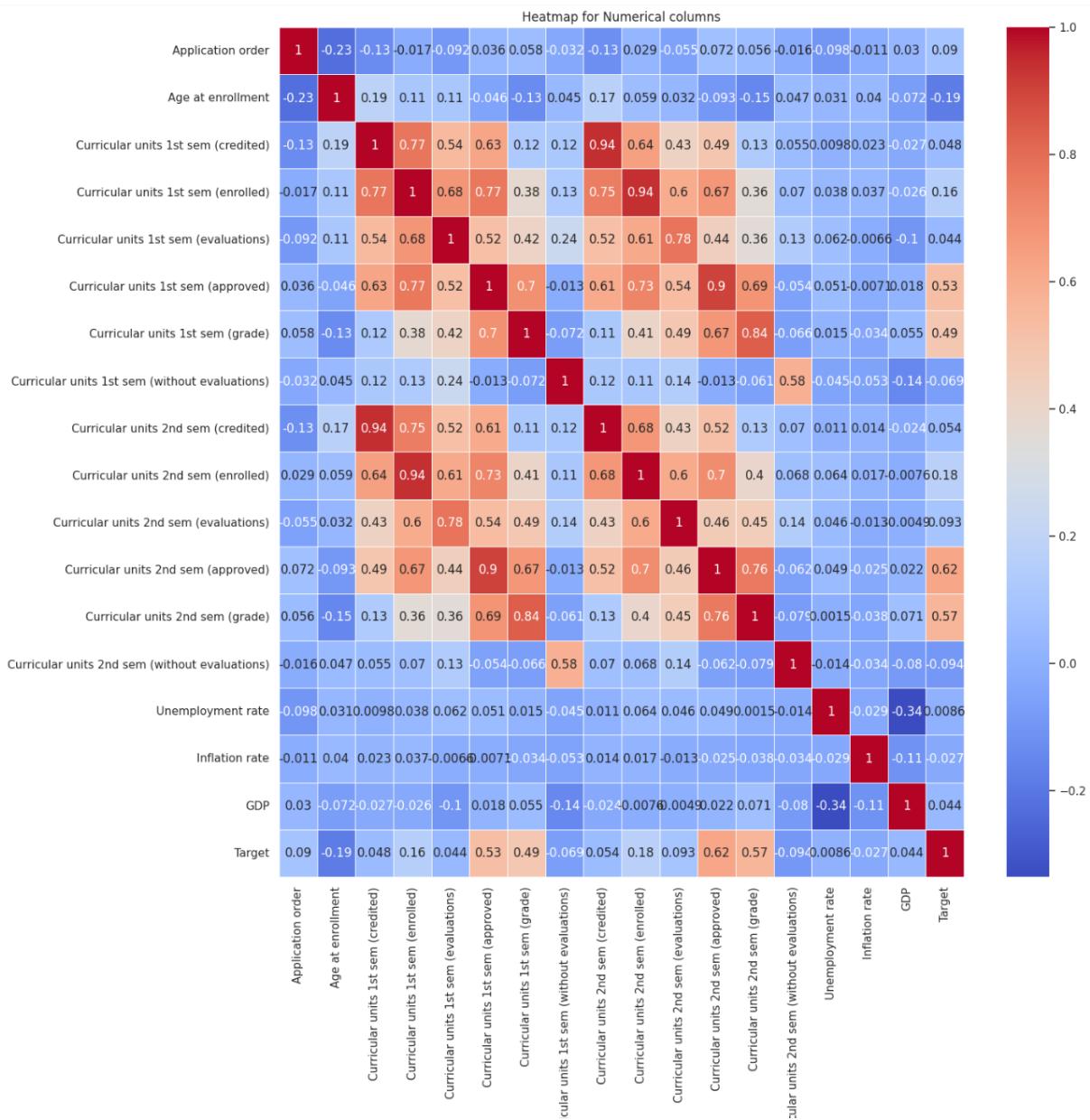
	Datasets columns name:
1.	Marital status
2.	Application mode
3.	Application order
4.	Course
5.	Daytime/evening attendance
6.	Previous qualification
7.	Nationality
8.	Mother's qualification
9.	Father's qualification
10.	Mother's occupation
11.	Father's occupation
12.	Displaced
13.	Educational special needs
14.	Debtor
15.	Tuition fees up to date
16.	Gender
17.	Scholarship holder
18.	Age at enrollment
19.	International
20.	Curricular units 1st sem (credited)
21.	Curricular units 1st sem (enrolled)
22.	Curricular units 1st sem (evaluations)
23.	Curricular units 1st sem (approved)
24.	Curricular units 1st sem (grade)
25.	Curricular units 1st sem (without evaluations)
26.	Curricular units 2nd sem (credited)
27.	Curricular units 2nd sem (enrolled)
28.	Curricular units 2nd sem (evaluations)
29.	Curricular units 2nd sem (approved)
30.	Curricular units 2nd sem (grade)
31.	Curricular units 2nd sem (without evaluations)
32.	Unemployment rate
33.	Inflation rate
34.	GDP
35.	Target
Data shape: (4424, 35)	

Tiếp theo là kiểm tra sự toàn vẹn của bộ dữ liệu:

↳ Marital status	0.587703
Application mode	3.074141
Application order	0.000000
Course	0.000000
Daytime/evening attendance	0.000000
Previous qualification	0.000000
Nationality	0.000000
Mother's qualification	0.000000
Father's qualification	0.000000
Mother's occupation	0.000000
Father's occupation	0.000000
Displaced	0.000000
Educational special needs	0.000000
Debtors	0.000000
Tuition fees up to date	0.000000
Gender	0.000000
Scholarship holder	0.000000
Age at enrollment	1.966546
International	0.000000
Curricular units 1st sem (credited)	0.000000
Curricular units 1st sem (enrolled)	0.000000
Curricular units 1st sem (evaluations)	0.000000
Curricular units 1st sem (approved)	0.000000
Curricular units 1st sem (grade)	0.000000
Curricular units 1st sem (without evaluations)	0.000000
Curricular units 2nd sem (credited)	0.000000
Curricular units 2nd sem (enrolled)	0.000000
Curricular units 2nd sem (evaluations)	0.000000
Curricular units 2nd sem (approved)	0.000000
Curricular units 2nd sem (grade)	0.000000
Curricular units 2nd sem (without evaluations)	0.000000
Unemployment rate	0.000000
Inflation rate	0.000000
GDP	0.000000
Target	0.000000
dtype:	float64

↳ Marital status has: 26 values
 Application mode has: 136 values
 Age at enrollment has: 87 values

- **Nhận xét:**
- + Có thể thấy được qua tổng quan bộ dữ liệu thu thập được rằng chúng ta có 35 cột với 4425 dòng dữ liệu. Trong đó có tồn tại missing values ở 3 cột “Marital status” với 26 giá trị, “Application mode” với 136 giá trị và “Age at enrollment” với 87 giá trị bị thiếu. Việc xử lý vấn đề này sẽ nằm ở khâu tiền xử lý phía sau.



Hình 1: Heatmap biểu diễn tương quan giữa các cột dữ liệu

- Nhận xét:**
- Qua biểu đồ heatmap ta thấy được có nhiều sự tương quan lẫn nhau giữa các cột. Điều này là một tín hiệu tốt cho thấy có những sự phụ thuộc vào nhau giữa các biến. Sẽ giúp mô hình dự báo có hiệu suất cao hơn sau này.
- Hơn nữa những cột về số tín chỉ ở cả kì 1 và kì 2 đều là tâm điểm của các sự tương quan. Vì vậy, chúng ta có thể gộp 2 kì lại thành một mà không ảnh hưởng nhiều tới tương quan các biến, từ đó giúp giảm số cột phải xử lý cũng như có cái nhìn tổng quan hơn về cả năm học.

3.1.1 Chính dạng bộ dữ liệu

a, Thay đổi tên cột

```
# Chuyển về định dạng lowercase
df.columns = [col.lower() for col in df.columns]
```

Việc chuyển đổi tên cột giúp thuận tiện cho việc nhập tên các thuộc tính.

Ở phần trước, chúng ta đã in danh sách các tiêu đề cột trong bộ dữ liệu và qua quan sát thấy được rằng các tiêu đề có chứa ký tự đặc biệt như "/" hoặc "s" sẽ có độ dài tên khá lớn. Vậy nên để tiện cho việc đọc dữ liệu thì ta sẽ đổi lại tên cho các cột này. Trước tiên hãy cùng kiểm tra lại những tiêu đề trên:

```
[ ] # Liệt kê ra các cột có tên đặc biệt (chứa 's ; () )
special_columns_name = []
for col in df.columns:
    if re.search(r"('s[ ])|(\s\()", col):
        special_columns_name.append(col)
    cols_to_rename = [val for val in df.columns if val not in special_columns_name]
renamed = [val.replace(" ", "_") for val in cols_to_rename]
dictr = {key: value for key, value in zip(cols_to_rename, renamed)}
df.rename(columns=dictr, inplace=True)
special_columns_name

["mother's qualification",
 "father's qualification",
 "mother's occupation",
 "father's occupation",
 'curricular units 1st sem (credited)',
 'curricular units 1st sem (enrolled)',
 'curricular units 1st sem (evaluations)',
 'curricular units 1st sem (approved)',
 'curricular units 1st sem (grade)',
 'curricular units 1st sem (without evaluations)',
 'curricular units 2nd sem (credited)',
 'curricular units 2nd sem (enrolled)',
 'curricular units 2nd sem (evaluations)',
 'curricular units 2nd sem (approved)',
 'curricular units 2nd sem (grade)',
 'curricular units 2nd sem (without evaluations)']
```

b, Gộp, xóa cột

Có thể thấy được một số tên cột khá dài, cùng với đó là những cột như “curricular units 1st sem(credited)” và “curricular units 2nd(credited)” cùng là số tín chỉ trong kỳ 1 và kỳ 2 của năm học, vậy nên ta sẽ gộp lại thành một cột tổng về số tín chỉ trong một năm học luôn để tiện tính toán sau này (tương tự với những cột còn lại):

```

# Gộp những cột Curricular_units 1st vs 2nd vì nó có sự tương quan với nhau, và tổng lại thì nó là số tín chỉ trong 1 năm học
column_indexes = list(range(4,10))
column_suffixes = ['curricular_units_credited', 'curricular_units_enrolled', 'curricular_units_evals',
                   'curricular_units_grade', 'curricular_units_grade', 'curricular_units_without_evals']

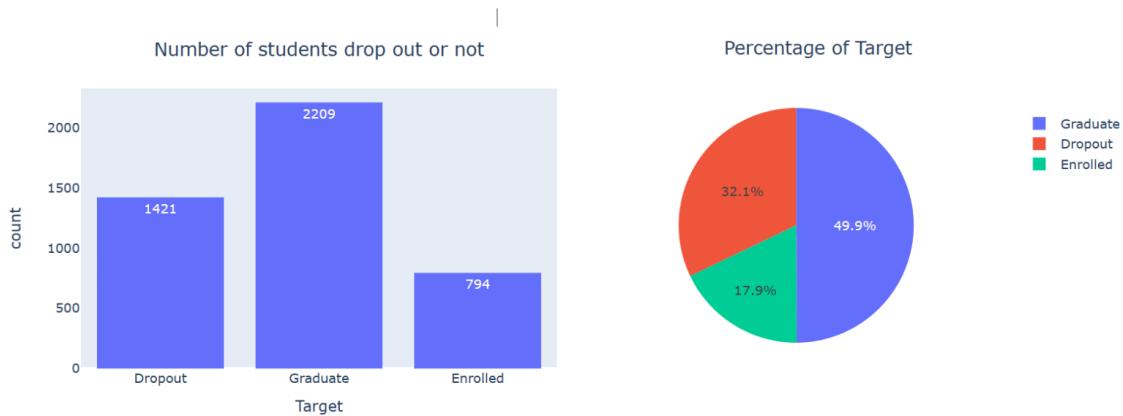
# Tương tự với Parent_occupation
column_indexes2 = list(range(0,4))
column_suffixes2 = ['parent_occupation', 'parent_qualification']

for i, suffix in zip(column_indexes, column_suffixes):
    df[suffix] = df[special_columns_name[i]] + df[special_columns_name[i + 6]]
for i, suffix in zip(column_indexes2, column_suffixes2):
    df[suffix] = df[special_columns_name[i]] + df[special_columns_name[i + 2]]
df.drop(columns=special_columns_name,inplace=True)

```

3.1.2 Phân tích sơ bộ dữ liệu

Ở đây với việc biến phụ thuộc của bộ dữ liệu này là “Target”, vậy trước tiên hãy xem qua thông số của biến này.

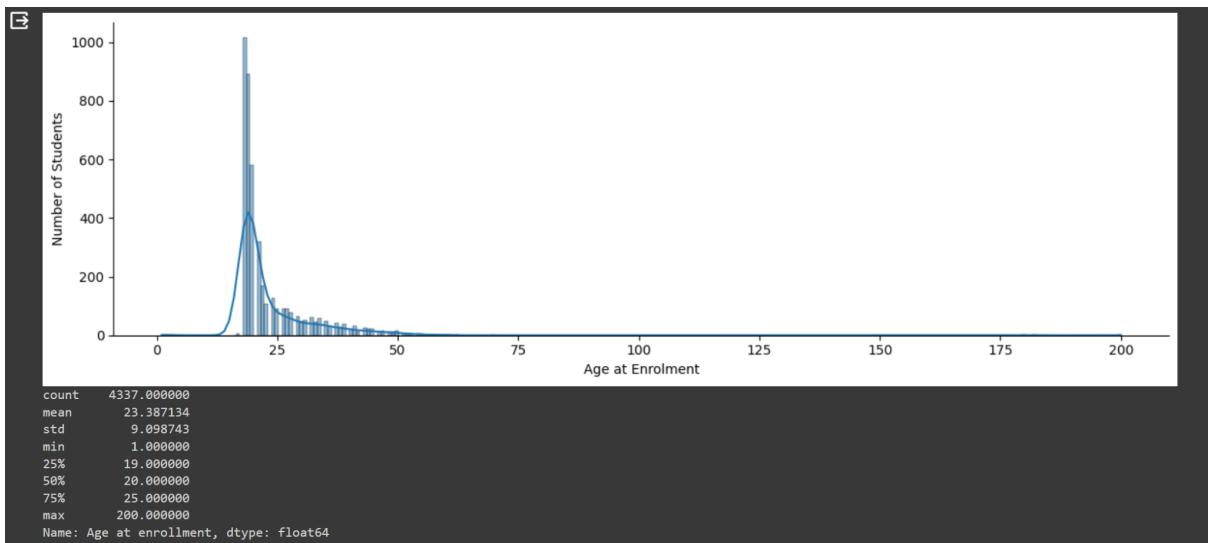


Hình 2: Thông số của biến “Target”

- **Nhận Xét:**
- + Phân Bố Phàn Trăm:
 - + "Graduate" chiếm 49.9% của biểu đồ, đây là nhóm có tỉ lệ cao nhất, cho thấy rằng một nửa sinh viên của nhóm mục tiêu đạt được tình trạng tốt nghiệp.
 - + "Dropout" chiếm 32.1%, chỉ ra rằng khoảng một phần ba không hoàn thành chương trình.
 - + "Enrolled" chiếm 17.9%, có nghĩa là ít hơn một phần năm của nhóm mục tiêu đang trong quá trình học.
- + Ý Nghĩa Giáo Dục:
 - + Số lượng người tốt nghiệp cao gợi ý rằng có tỉ lệ thành công khá tốt trong nhóm mục tiêu này.
 - + Tỉ lệ bỏ học cũng đáng kể, đặt ra câu hỏi về những thách thức hoặc rào cản mà sinh viên có thể gặp phải.
 - + Tỉ lệ đang nhập học cho thấy có một lượng sinh viên đáng kể đang trong quá trình hoàn thành các yêu cầu để tốt nghiệp.
- + Quản Lý và Chính Sách:
 - + Đối với các quản lý giáo dục hoặc nhà hoạch định chính sách, tỉ lệ bỏ học có thể là một chỉ số quan trọng để xem xét các can thiệp nhằm cải thiện.

- + Sự cân nhắc về nguồn lực và sự hỗ trợ có thể cần được tập trung vào những sinh viên đang tiếp tục học để đảm bảo rằng họ có khả năng tốt nghiệp.

Vậy điều gì đã làm cho tỉ lệ nghỉ học của sinh viên ở bộ dữ liệu này đạt ngưỡng như vậy, cùng khám phá xem một vài biến có thể ảnh hưởng tới kết quả này. Hãy xem qua với độ tuổi nhập học:

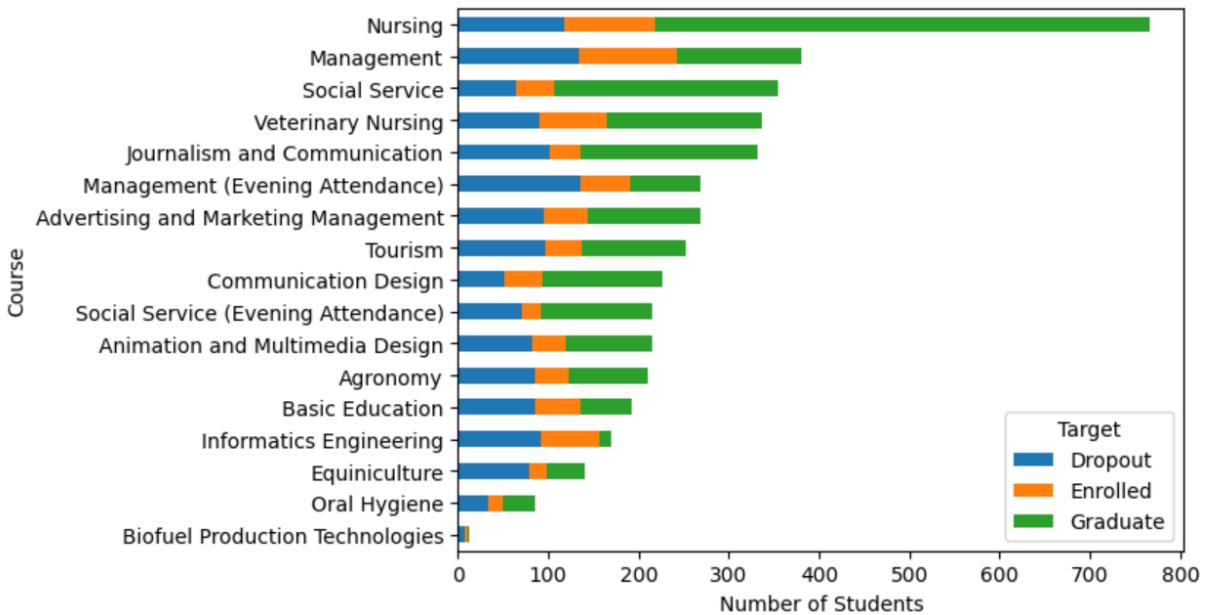


Hình 3: Số lượng sinh viên nghỉ học theo độ tuổi (Trước xử lý)

- Nhận xét:

- + Thông qua biểu đồ, chúng ta có thể nhận thấy một số điểm quan trọng đáng chú ý về sự phân bố độ tuổi của sinh viên. Điều này mang lại cái nhìn sâu hơn về sự đa dạng của độ tuổi trong cộng đồng học viên và những ảnh hưởng quan trọng mà nó có đối với kết quả học tập. Có thể thấy, độ tuổi trung bình của sinh viên khi đăng ký là khoảng 23 tuổi, trong đó độ tuổi thường gặp nhất là từ 19 đến 25 tuổi. Điều này có thể phản ánh một xu hướng chung của sinh viên là nhiều người quyết định tham gia học đại học sau khi tốt nghiệp trung học. Điểm đáng chú ý là phân bố độ tuổi lệch dương, cho thấy sự tập trung cao ở các độ tuổi trẻ hơn. Ngoài ra, độ tuổi của sinh viên còn kéo dài hơn cho tới hơn 50 tuổi cho thấy sự đa dạng trong môi trường học tập này.
- + Thông qua thống kê mô tả cho thấy được có tồn tại giá trị nhỏ nhất là 1 và lớn nhất là 200, trong đó độ lệch chuẩn chỉ ở mức 9, vậy thì đây có vẻ là ngoại lệ hoặc sai sót trong quá trình thu thập hoặc nhập liệu dữ liệu. Chúng sẽ được xem xét và thực hiện xử lý dữ liệu trong bước tiền xử lý outliers sau này.

Thực trạng sinh viên nghỉ học do chọn sai ngành học không phù hợp cũng rất phổ biến, vậy chúng ta kiểm tra trong bộ dữ liệu này có như vậy không:



Hình 4: Số lượng sinh viên của mỗi ngành học

- Nhận xét:

- + Có thể thấy rằng mặc dù có tỷ lệ tốt nghiệp cao nhưng Nursing vẫn có số lượng sinh viên bỏ học là khá lớn. Chứng tỏ ngành học có thể khó là khó và vẫn là thách thức với nhiều học sinh để đi tới hết chương trình.
- + Cùng với đó là một số ngành học có tỷ lệ sinh viên bỏ học cao như Management, Tourism, Informatics Engineering. Đây đều là những nhóm ngành đầu vào không quá cao nhưng tỷ lệ chọi đầu ra là rất lớn. Đây cũng có thể là nguyên nhân khiến nhiều sinh viên quyết định bỏ học.

3.2 Xử lý dữ liệu bị thiếu (Missing Values)

Tiến hành kiểm tra số lượng dòng chứa giá trị bị thiếu của bộ dữ liệu:

```

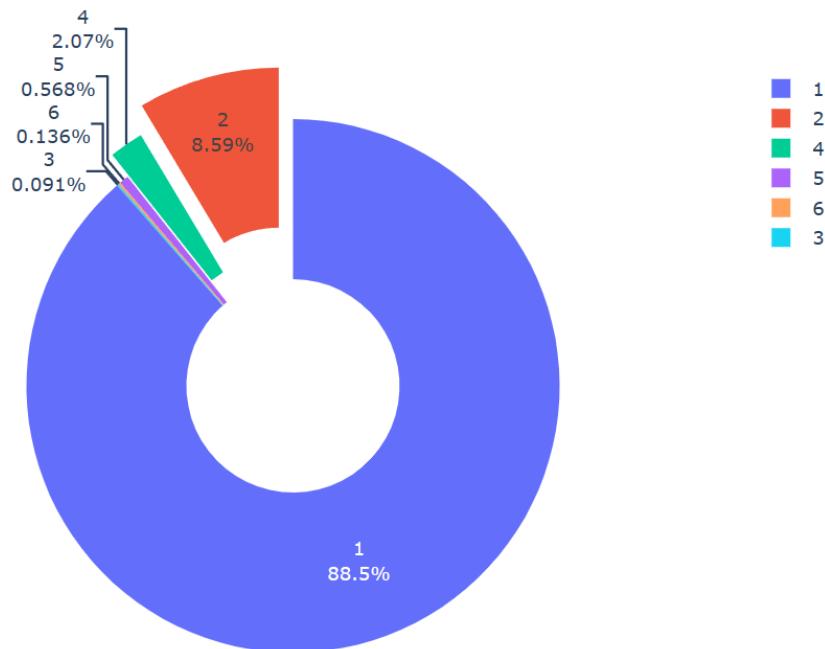
[32] missing=df.isnull().sum()/ len(df)* 100
      print(missing)
      for column, percentage in missing.items():
          if percentage > 0:
              missing_count = df[column].isnull().sum()
              print(f'{column} has: {missing_count} values')

      Marital status has: 26 values
      Application mode has: 136 values
      Age at enrollment has: 87 values
  
```

Column	Percentage (%)
Marital status	0.587703
Application mode	3.074141
Application order	0.000000
Course	0.000000
Daytime/evening attendance	0.000000
Previous qualification	0.000000
Nacionality	0.000000
Mother's qualification	0.000000
Father's qualification	0.000000
Mother's occupation	0.000000
Father's occupation	0.000000
Displaced	0.000000
Educational special needs	0.000000
Debtor	0.000000
Tuition fees up to date	0.000000
Gender	0.000000
Scholarship holder	0.000000
Age at enrollment	1.966546
International	0.000000
Curricular units 1st sem (credited)	0.000000
Curricular units 1st sem (enrolled)	0.000000
Curricular units 1st sem (evaluations)	0.000000
Curricular units 1st sem (approved)	0.000000
Curricular units 1st sem (grade)	0.000000
Curricular units 1st sem (without evaluations)	0.000000
Curricular units 2nd sem (credited)	0.000000
Curricular units 2nd sem (enrolled)	0.000000
Curricular units 2nd sem (evaluations)	0.000000
Curricular units 2nd sem (approved)	0.000000
Curricular units 2nd sem (grade)	0.000000
Curricular units 2nd sem (without evaluations)	0.000000
Unemployment rate	0.000000
Inflation rate	0.000000
GDP	0.000000
Target	0.000000

- a. Qua kết quả thấy được có 3 biến có dữ liệu bị thiếu, ta sẽ bắt đầu xử lý với “Marital status”. Đầu tiên hãy kiểm tra qua các giá trị của biến này:

Biểu đồ tròn thể hiện tỉ lệ các giá trị trong biến Marital Status

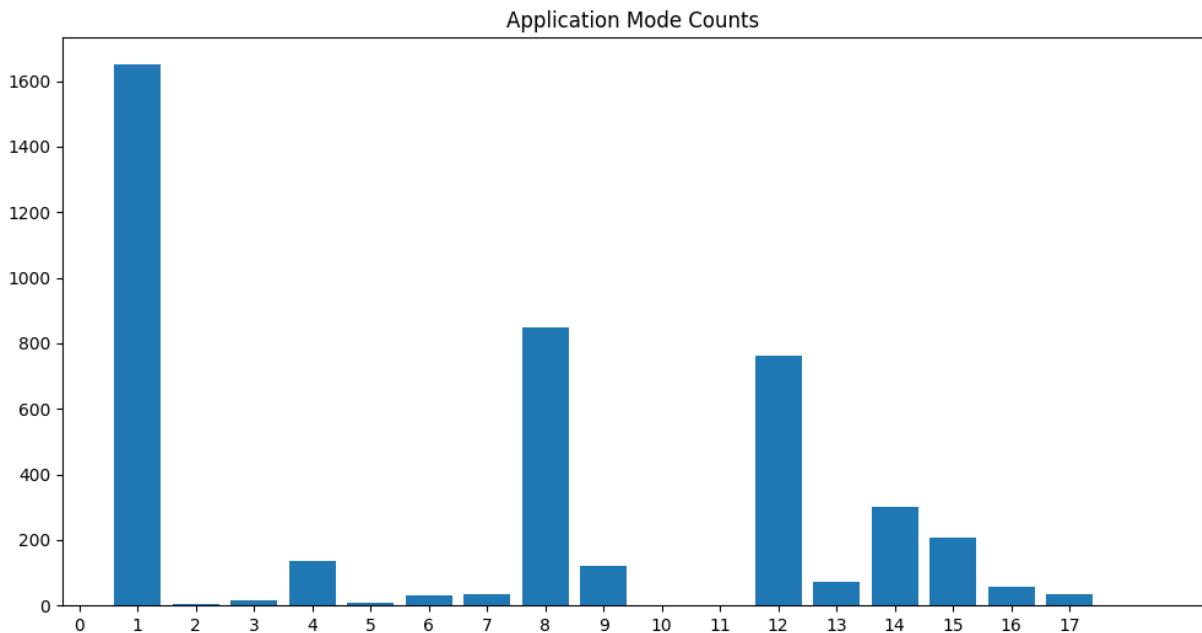


Hình 5: Tỉ lệ Các giá trị trong biến “Marital Status”

Có thể thấy được rằng các biến phân loại được phân bố từ 1 đến 6, với giá trị 1 chiếm phần lớn (88.5%), điều này cho thấy rằng hơn % số lượng sinh viên được khảo sát là đang trong tình trạng độc thân, điều này cũng hợp lý. Vì vậy chúng ta sẽ bổ sung những giá trị bị thiếu bằng mode (giá trị lặp lại nhiều nhất):

```
df['marital_status'] = df['marital_status'].fillna(df['marital_status'].mode()[0])
```

- b. Tiếp đến là cột “application mode”, hãy nhìn sơ qua phân bố dữ liệu của cột này:

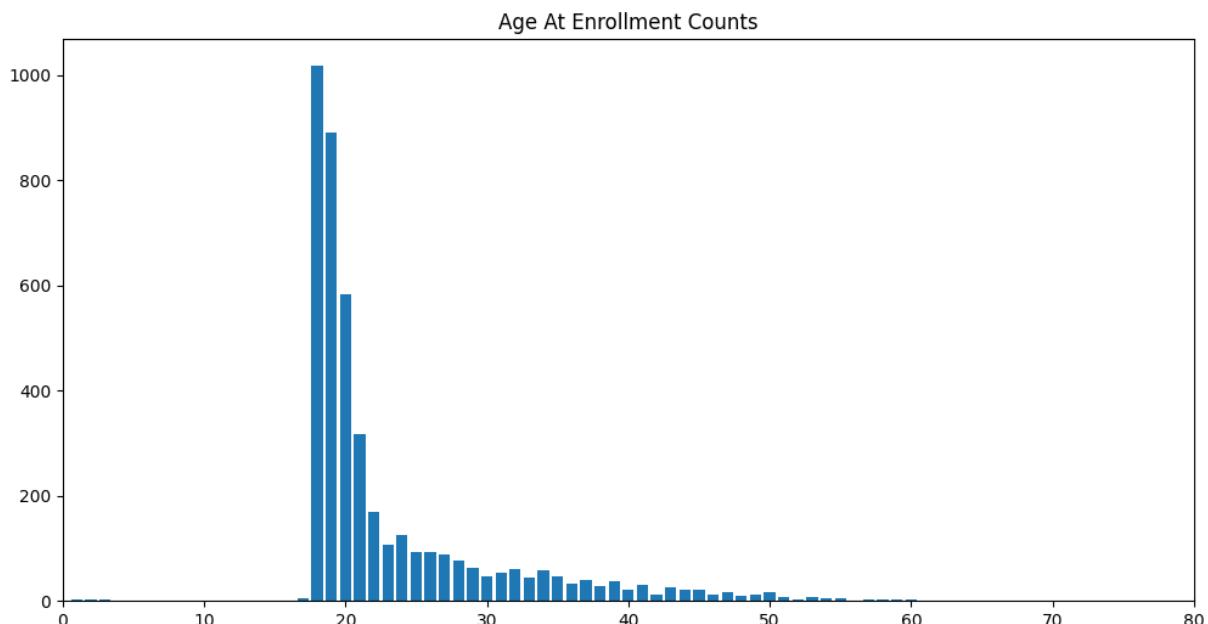


Hình 6: Số lượng các giá trị của biến “Application Mode”

Thông qua biểu đồ có thể thấy được rằng biến có 18 giá trị phân loại và dữ liệu được phân bổ không đồng đều, phương thức 1 và 8 là 2 phương thức được áp dụng nhiều nhất. Vì vậy chúng ta sẽ bổ sung những giá trị bị thiếu bằng mode (giá trị lặp lại nhiều nhất):

```
df['application_mode'] = df['application_mode'].fillna(df['application_mode'].mode()[0])
```

- c. Với biến có dữ liệu bị thiếu cuối cùng là “Age at enrollment”, trước tiên cũng xem qua dữ liệu của biến này:



Hình 7: Độ tuổi tại thời điểm nhập học của sinh viên

Với bộ dữ liệu có phân bố lệch phải như vậy, chúng ta có thể điền các giá trị bị thiếu bằng trung vị của độ tuổi lúc ghi danh. Trong trường hợp này, sử dụng trung vị tốt hơn giá trị

trung bình vì trung vị không bị ảnh hưởng bởi các giá trị ngoại lai trong dữ liệu. Ngoài ra việc thay thế bằng trung vị còn giúp giảm thiểu sự biến động do các giá trị bất thường gây ra và tạo ra một phép đo trung bình ổn định hơn.

```
1 age_mean_value = round(df['age_at_enrollment'].mean())
2 print(f"Trung bình của tuổi khi nhập học: {age_mean_value}")
3 age_median_value = df['age_at_enrollment'].median()
4 print(f"Trung vị của tuổi khi nhập học: {age_median_value}")

Trung bình của tuổi khi nhập học: 23
Trung vị của tuổi khi nhập học: 20.0

1 # Thay thế missing values bằng giá trị trung bình
2 df['age_at_enrollment'] = df['age_at_enrollment'].fillna(round(age_median_value))
```

- Tiến hành kiểm tra lại:

```
df.isnull().sum()

marital_status          0
application_mode         0
application_order        0
course                   0
daytime/evening_attendance 0
previous_qualification   0
nacionality              0
mother's qualification    0
father's qualification     0
mother's occupation       0
father's occupation       0
displaced                 0
educational_special_needs 0
debtor                    0
tuition_fees_up_to_date   0
gender                    0
scholarship_holder        0
age_at_enrollment         0
international              0
curricular units 1st sem (credited) 0
curricular units 1st sem (enrolled) 0
curricular units 1st sem (evaluations) 0
curricular units 1st sem (approved) 0
curricular units 1st sem (grade) 0
curricular units 1st sem (without evaluations) 0
curricular units 2nd sem (credited) 0
curricular units 2nd sem (enrolled) 0
curricular units 2nd sem (evaluations) 0
curricular units 2nd sem (approved) 0
curricular units 2nd sem (grade) 0
curricular units 2nd sem (without evaluations) 0
unemployment_rate          0
inflation_rate             0
gdp                       0
target                     0
dtype: int64
```

3.3 Phân loại dữ liệu

```
[233] for col in df:  
    if df[col].dtype == 'float64' and col not in ['unemployment_rate', 'inflation_rate', 'gdp']:  
        df[col] = df[col].astype('int64')
```

Đầu tiên chúng ta cần phải chuyển hóa các cột có giá trị float (ngoài unemployment, inflation rate và gdp là các biến định lượng liên tục) về lại int cho đồng bộ dữ liệu và phục vụ cho việc phân loại dữ liệu

```
# Xác định các cột dữ liệu số:  
numerical_features = [col for col in df.columns if col in df.select_dtypes(include='float64')]  
numerical_features.extend(['age_at_enrollment', 'curicular_units_credited',  
    'curicular_units_enrolled', 'curicular_units_evals',  
    'curicular_units_grade', 'curicular_units_without_evals'])  
print(f"{len(numerical_features)} features: {numerical_features}")
```

Sau đó khởi tạo một danh sách gồm các cột có dữ liệu là float và thêm vào một số cột.

Lưu ý: Các giá trị đều là biến phân loại và được gán nhãn theo số nguyên **trừ cột age_at_enrollment, curicular_units_credited, curicular_units_enrolled, curicular_units_evals, curicular_units_grade, curicular_units_without_evals** nên các cột vừa được nhắc đến sẽ được thêm vào danh sách biến định lượng

- Tạo một list chứa các biến định lượng với:

9 features: ['unemployment_rate', 'inflation_rate', 'gdp', 'age_at_enrollment', 'curicular_units_credited', 'curicular_units_enrolled', 'curicular_units_evals', 'curicular_units_grade', 'curicular_units_without_evals']

```
# Xác định các cột dữ liệu Nhị phân:  
binary_features = list()  
for col in df:  
    if all(value in range(2) for value in df[col].values):  
        binary_features.append(col)  
print(f"{len(binary_features)} features: {binary_features}")
```

Xác định các thuộc tính nhị phân (chỉ có hai giá trị duy nhất trong thuộc tính) đó để phục vụ cho việc biểu diễn các biểu đồ được tối ưu và phục vụ cho kiểm định.

- Các thuộc tính nhị phân:

8 features: ['daytime/evening_attendance', 'displaced', 'educational_special_needs', 'debtor', 'tuition_fees_up_to_date', 'gender', 'scholarship_holder', 'international']

```
# Xác định các cột dữ liệu được gán nhãn:  
labeled_features = [col for col in df.columns  
                    if (col in df.select_dtypes(include='int').columns and col not in binary_features  
                        and col not in numerical_features)]  
print(f"{len(labeled_features)} features: {labeled_features}")
```

Xác định các thuộc tính được gán nhãn còn lại với điều kiện các biến đều số nguyên (Labeled), không nằm trong danh sách các thuộc tính nhị phân và danh sách các thuộc tính định lượng đã khởi tạo bên trên.

Tương tự việc phân loại các dữ liệu được gán nhãn này cũng sẽ phục vụ cho việc biểu diễn trực quan được tối ưu (Xem xét những thuộc tính nào có quá nhiều giá trị độc lập - unique thì sẽ áp dụng những phương pháp plot khác nhau).

- Các thuộc tính được gán nhãn:

9 features: ['marital_status', 'application_mode', 'application_order', 'course', 'previous_qualification', 'nationality', 'parent_occupation', 'parent_qualification', 'target_labeled']

- Tiến hành kiểm tra lại:

```
bool(len(df.columns) - (len(binary_features) + len(labeled_features)) == 1)  
True
```

Kết quả trả về “True” sau khi lấy tổng độ dài các biến trừ đi độ dài các biến đã được gán nhãn là **binary_features** và **labeled_features** trả về kết quả bằng 1 (Ở đây là biến **Target**) thì tất cả các biến đã được gán nhãn đầy đủ.

- Tiến hành kiểm tra lại:

```
] df.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 4424 entries, 0 to 4423  
Data columns (total 27 columns):  
 #   Column           Non-Null Count  Dtype     
---  --  
 0   marital_status    4424 non-null   int64    
 1   application_mode  4424 non-null   int64    
 2   application_order 4424 non-null   int64    
 3   course            4424 non-null   int64    
 4   daytime/evening_attendance 4424 non-null   int64    
 5   previous_qualification 4424 non-null   int64    
 6   nationality        4424 non-null   int64    
 7   displaced          4424 non-null   int64    
 8   educational_special_needs 4424 non-null   int64    
 9   debtor             4424 non-null   int64    
 10  tuition_fees_up_to_date 4424 non-null   int64    
 11  gender              4424 non-null   int64    
 12  scholarship_holder 4424 non-null   int64    
 13  age_at_enrollment 4424 non-null   int64    
 14  international       4424 non-null   int64    
 15  unemployment_rate 4424 non-null   float64   
 16  inflation_rate     4424 non-null   float64   
 17  gdp                 4424 non-null   float64   
 18  target               4424 non-null   object    
 19  curicular_units_credited 4424 non-null   int64    
 20  curicular_units_enrolled 4424 non-null   int64    
 21  curicular_units_evals 4424 non-null   int64    
 22  curicular_units_grade 4424 non-null   int64    
 23  curicular_units_without_evals 4424 non-null   int64    
 24  parent_occupation    4424 non-null   int64    
 25  parent_qualification 4424 non-null   int64    
 26  target_labeled      4424 non-null   int64    
 dtypes: float64(3), int64(23), object(1)  
memory usage: 933.3+ KB
```

TỔNG QUAN ĐỘ TƯƠNG QUAN CỦA CÁC THUỘC TÍNH VỚI TARGET:

- Tạo ra một thuộc tính target_labeled: tương tự như target nhưng được gán nhãn để đồng bộ hóa đơn vị (int) và sẽ sử dụng target_labeled thay cho thuộc tính target gốc (object):

```
df['target_labeled'] = df['target'].map({  
    'Dropout':0,  
    'Enrolled':1,  
    'Graduate':2  
})
```

- Tổng quan về Rank của các biến so với target_labeled với một số biến được chọn:

```
[223] selected_columns = ['application_mode', 'age_at_enrollment', 'curicular_units_credited', 'curicular_units_enrolled',  
    'curicular_units_evals', 'curicular_units_grade', 'nacionality', 'marital_status',  
    'curicular_units_without_evals', 'gender', 'debtor', 'international', 'daytime/evening_attendance',  
    'unemployment_rate', 'inflation_rate', 'gdp', 'target_labeled']  
selected_df = df[selected_columns]  
correlation_matrix = selected_df.corr().sort_values(by='target_labeled', ascending=False)  
plt.figure(figsize=(10, 5))  
sns.heatmap(correlation_matrix[['target_labeled']], annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)  
plt.title('Features Ranker')  
plt.show()
```



Hình 8: Tổng quan độ tương quan của các thuộc tính với “Target”

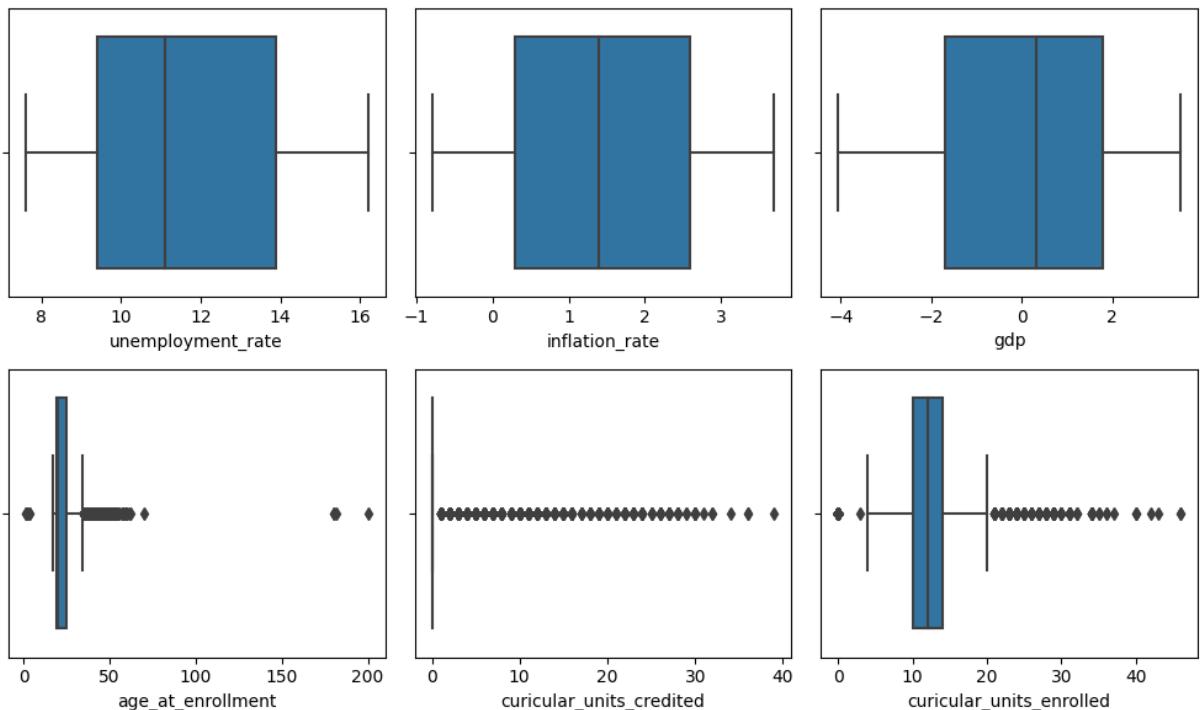
3.4 Xử lý outliers

```

cols = [element for element in numerical_features]
n_rows=2
n_cols=3
fig, ax = plt.subplots(n_rows,n_cols,figsize=(n_rows*5,n_cols*2))

for r in range(0,n_rows):
    for c in range(0,n_cols):
        i = r*n_cols + c
        if i < len(cols):
            ax_i = ax[r,c]
            sns.boxplot(data=df,x=cols[i],ax=ax_i)
plt.tight_layout()
ax.flat[-1].set_visible(False)
ax.flat[-2].set_visible(False)

```



Hình 9: Box plot cho biết độ dàn trải của dữ liệu (Trước xử lý)

- **Nhận xét:**
- + Các số liệu về tỷ lệ lạm phát hoặc tỷ lệ thất nghiệp, gdp đều không chứa ngoại lai. Các số lượng đơn vị học phần là các chỉ số quan trọng vì thế không thể loại các giá trị ngoại lai. Chỉ có thuộc tính tuổi chứa nhiều giá trị không hợp lý và chúng ta nhận định đây là ngoại lai nên sẽ **chỉ loại bỏ ngoại lai cho cột age_at_enrollment**.
- + Các biến khác đều là biến phân loại được gán nhãn nên sẽ không được coi là ngoại lai
- Xử lý ngoại lai:

```

] Q1 = df['age_at_enrollment'].quantile(0.25)
Q3 = df['age_at_enrollment'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - (1.5 * IQR)
upper_bound = Q3 + (4 * IQR)

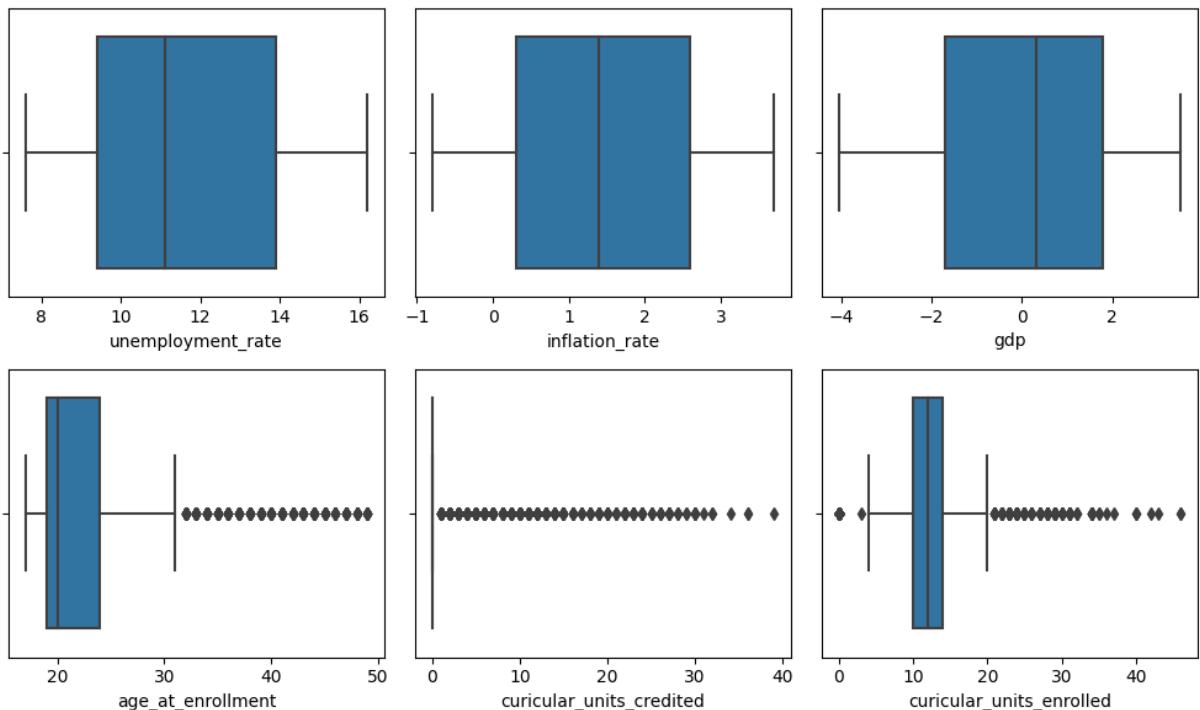
df = df[~((df['age_at_enrollment'] < lower_bound) | (df['age_at_enrollment'] > upper_bound))]

```

- Ở đây, ta sẽ xử lý những giá trị ngoại lai bằng phương pháp Interquartile range với khoảng dưới = $Q1 - (1.5 * IQR)$ vì sẽ chỉ loại những giá trị cực gần 0 nên số nhân sẽ nhỏ, với khoảng trên = $Q3 + (4 * IQR)$ vì sẽ loại những giá trị có độ biến thiên lớn nên chúng ta sẽ tùy chỉnh số nhân = 4 để loại những khoảng siêu ngoại lai.

3.5 Biểu diễn lại dữ liệu sau khi xử lý

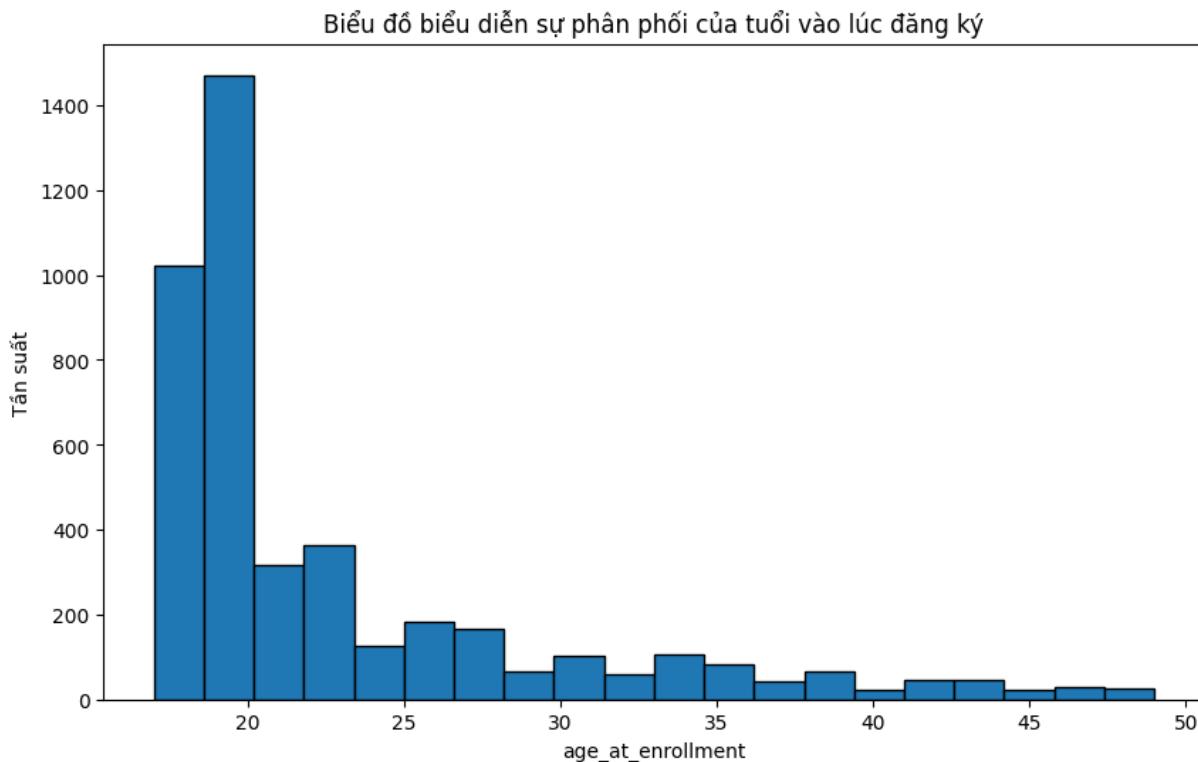
BIỂU ĐỒ HỘP SAU KHI XỬ LÝ:



Hình 10: Boxplot cho biết độ dàn trải của dữ liệu (Sau xử lý)

- **Nhận xét:**
- + Các số liệu không hợp lệ về tuổi đã được xóa bỏ.

BIỂU ĐỒ PHÂN BỐ TUỔI:

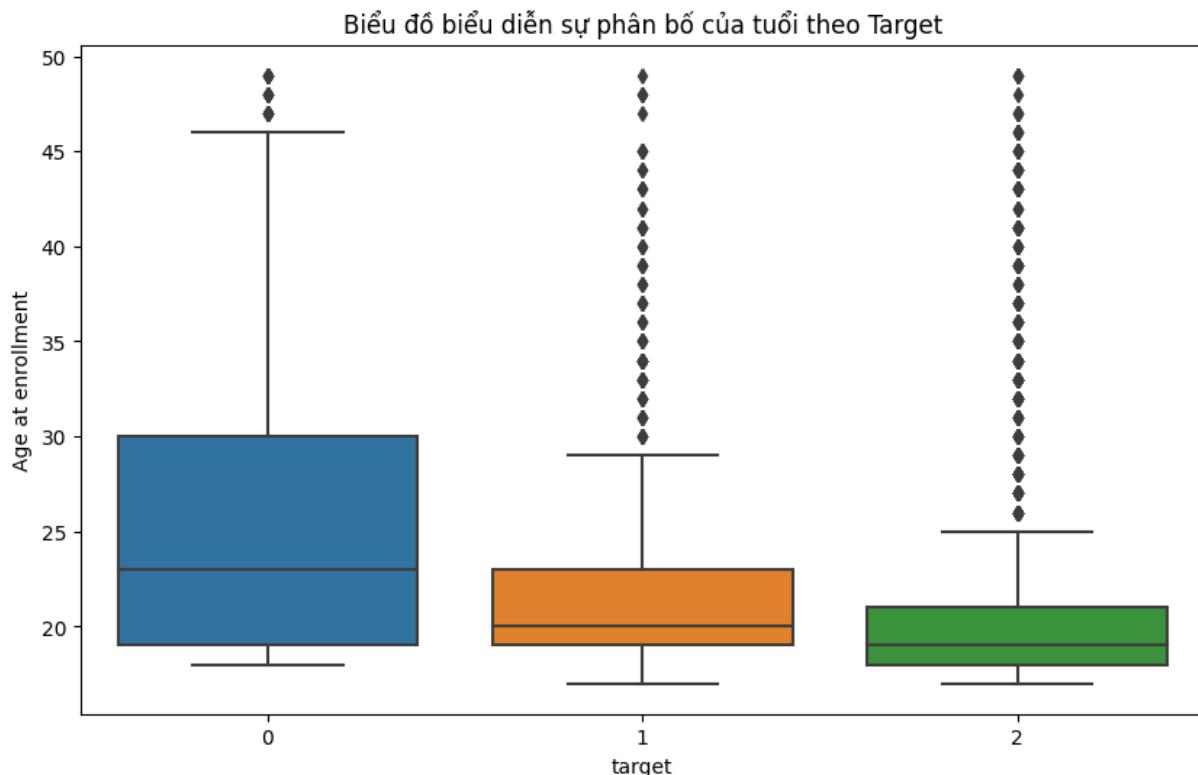


Hình 11: Biểu đồ phân bố tuổi

- Nhận xét:

- + Biểu đồ mô tả phân phối độ tuổi của sinh viên tại thời điểm nhập học, được gọi là `age_at_enrollment`. Quan sát đầu tiên từ biểu đồ là có một lượng lớn sinh viên nhập học ở lứa tuổi từ 18 đến 20, điều này phản ánh xu hướng truyền thống của việc bắt đầu giáo dục đại học ngay sau khi hoàn thành trung học. Điểm cao nhất của biểu đồ cho thấy độ tuổi phổ biến nhất để bắt đầu học là khoảng 18 tuổi, với số lượng giảm dần khi độ tuổi tăng lên. Điều này cũng cho thấy một mẫu hình giảm dần rõ rệt: càng lớn tuổi, sinh viên càng ít nhập học vào đại học hoặc các chương trình sau trung học.
- + Mặc dù có một số lượng đáng kể sinh viên nhập học ở độ tuổi từ 15 đến 25, biểu đồ cũng ghi nhận sự hiện diện của những sinh viên ở độ tuổi từ 30 trở lên, mặc dù số lượng này không lớn. Điều này có thể cho thấy một số người chọn tiếp tục giáo dục ở một độ tuổi muộn hơn, có thể vì lý do nâng cao trình độ, thay đổi nghề nghiệp, hoặc phát triển cá nhân. Sự phân bố này cung cấp một cái nhìn thú vị về đa dạng độ tuổi trong giáo dục đại học và cho thấy giáo dục không chỉ giới hạn ở lứa tuổi thanh niên.

BIỂU ĐỒ PHÂN BỐ TUỔI THEO TARGET (WHISKER):



Hình 12: Biểu đồ phân bố tuổi theo biến "Target" (Whisker)

- **Nhận xét:**

- + Dropout (0): Phân bố độ tuổi phổ biến của nhóm này có phạm vi từ khoảng dưới 20 đến hơn 30 tuổi. Trung vị (median) của nhóm sinh viên bỏ học ở khoảng 23 tuổi, cho thấy nửa số lượng sinh viên bỏ học là dưới 23 tuổi và nửa còn lại là trên 23 tuổi. Biểu đồ hộp này có dải phân phối khá rộng, cho thấy có sự phân tán lớn về độ tuổi trong nhóm này. Có một lượng ít sinh viên ở độ tuổi trên 45 quyết định bỏ học, còn lại, có thể nói đa số số lượng sinh viên quyết định bỏ học nằm trong độ tuổi phổ biến của học vấn.
- + Enrolled (1): Độ tuổi của nhóm đang theo học phân bố trong phạm vi hẹp hơn so với nhóm Dropout (bỏ học), từ khoảng 18 đến 23, với trung vị ở khoảng 20 tuổi. Điều này cho thấy phần lớn sinh viên đang theo học nằm trong độ tuổi này. Đây biểu hiện rằng nhóm sinh viên hiện đang theo học nằm ở độ tuổi trẻ.
- + Graduate (2): Phân bố độ tuổi trong nhóm này là hẹp nhất so với hai nhóm còn lại (từ 17 đến 22), Trung vị nằm ở khoảng 18 tuổi và có thể nói số lượng sinh viên đã tốt nghiệp trải dài ở đa số các độ tuổi.

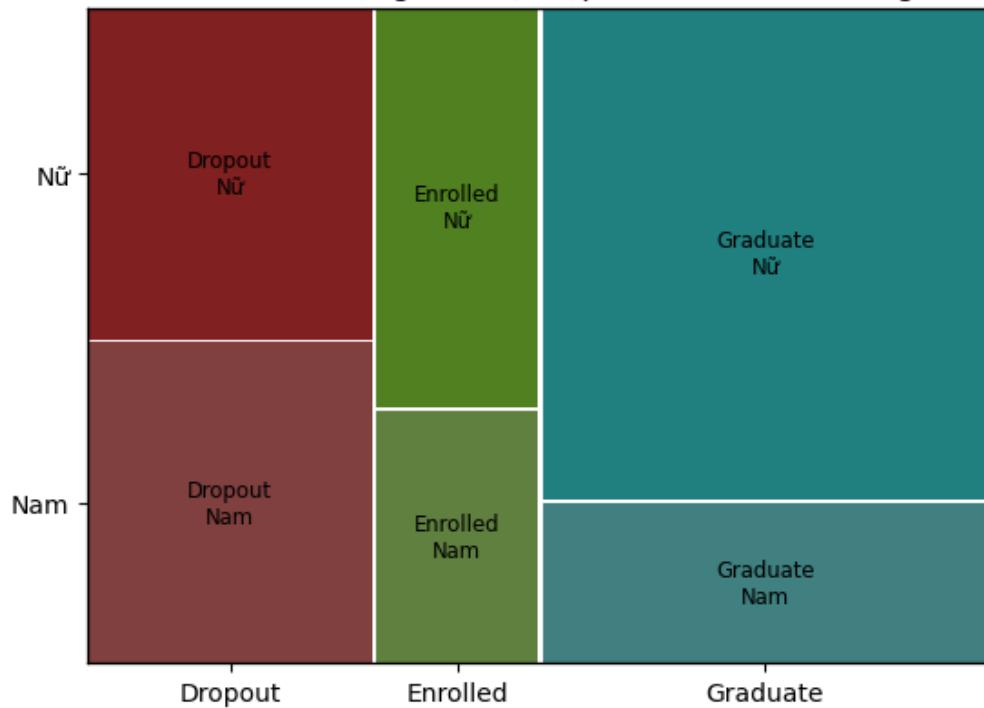
CHƯƠNG IV: BIỂU ĐỒ BIỂU DIỄN TRỰC QUAN DỮ LIỆU

(Trạng thái học tập của sinh viên: Bỏ học, đang học, tốt nghiệp.)

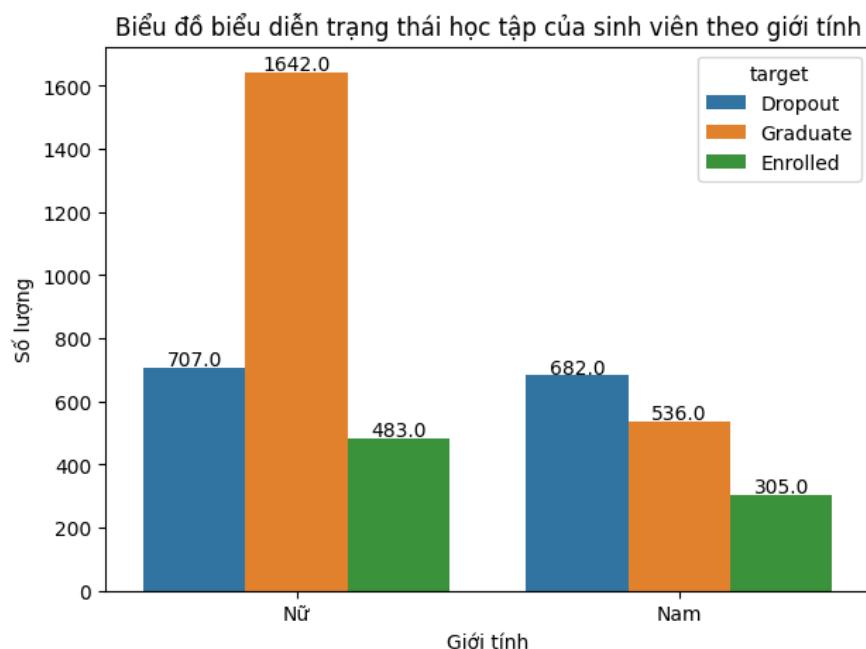
4.1 Biểu diễn theo nhân khẩu học

4.1.1 Biểu đồ biểu diễn trạng thái học tập của sinh viên theo giới tính:

Biểu đồ biểu diễn trạng thái học tập của sinh viên theo giới tính



Hình 13: Biểu đồ biểu diễn trạng thái học tập của sinh viên theo giới tính (1)

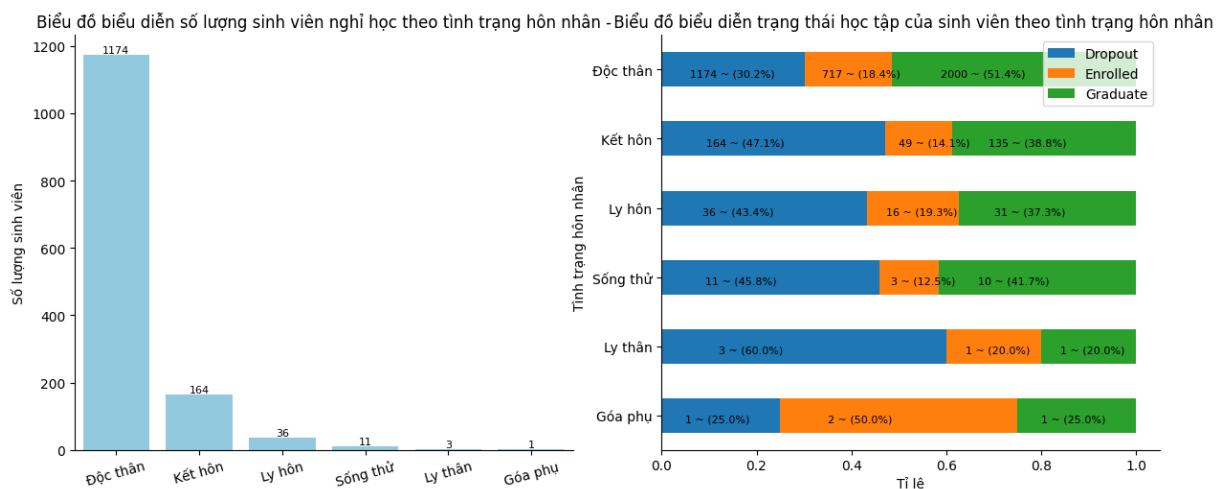


Hình 14: Biểu đồ biểu diễn trạng thái học tập của sinh viên theo giới tính (2)

- **Nhận xét:**

- + Thông qua biểu đồ, có thể thấy được tỉ lệ nghỉ học (Dropout) có sự đồng đều giữa 2 giới. Tuy nhiên, điều đáng chú ý là tỉ lệ đang theo học và tốt nghiệp của nữ giới cao hơn so với nam giới, đặc biệt là tỉ lệ tốt nghiệp xấp xỉ gấp đôi.
- + Điều này cho thấy được rằng mặc dù số lượng nghỉ học ở cả 2 giới là như nhau, nhưng nữ giới lại thể hiện sự ưu thế với tỉ lệ tốt nghiệp cao hơn hẳn. Điều này có thể là một tín hiệu tích cực về sự tiến triển và thành công học vụ của nữ giới.

4.1.2 Biểu đồ biểu diễn trạng thái học tập của sinh viên theo tình trạng hôn nhân:



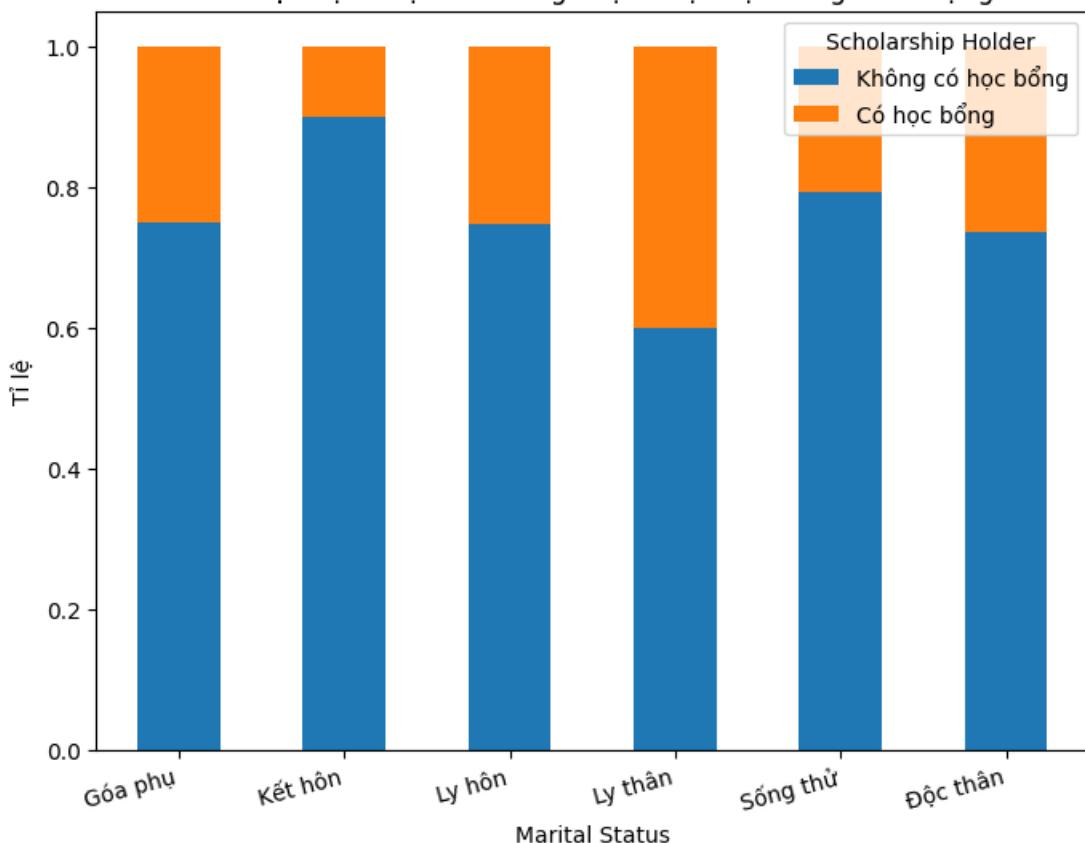
Hình 15: Biểu đồ biểu diễn trạng thái học tập của sinh viên theo tình trạng hôn nhân

- **Nhận xét:**

- + Từ biểu đồ “Biểu đồ biểu diễn số lượng sinh viên nghỉ học theo tình trạng hôn nhân”, chúng ta có thể quan sát được rằng nhóm sinh viên độc thân có số lượng sinh viên nghỉ học cao hơn đáng kể so với sinh các nhóm sinh viên còn lại, cao hơn xấp xỉ 7 lần so với nhóm sinh viên có số lượng sinh viên nghỉ học đứng thứ hai - nhóm đã kết hôn.
- + Tuy nhiên, khi xem xét trên một góc độ khác thông qua biểu đồ “Biểu đồ biểu diễn tình trạng hôn nhân của sinh viên và tình trạng học” thì trong tổng thể toàn bộ sinh viên đang độc thân, tỉ lệ sinh viên bỏ học chỉ chiếm khoảng 30.2 %, thấp hơn so với tỉ lệ sinh viên đã tốt nghiệp (khoảng 20%). Trong khi đó, ở các nhóm sinh viên còn lại, tỉ lệ bỏ học dao động trong khoảng 40%, xấp xỉ với tỉ lệ tốt nghiệp. Ngoại trừ nhóm sinh viên Góa phụ, có tỉ lệ bỏ học và tốt nghiệp bằng nhau và đều ở mức 25% và nhóm sinh viên đã ly thân có tỉ lệ bỏ học cao hơn hẳn (khoảng 60%)
- + Nếu xét trung bình trên tổng thể, tỉ lệ giữa các tình trạng học tập của sinh viên có tham gia hoạt động hôn nhân và không tham gia khá tương đương nhau. Tuy nhiên, khi xem xét chi tiết từng nhóm nhỏ, ta thấy rằng đối với nhóm sinh viên có tình trạng hôn nhân là ly thân, nguy cơ bỏ học cao hơn so với các nhóm khác. Điều này đặt ra câu hỏi về tác động của tình trạng hôn nhân đến hiệu suất học tập của sinh viên và cần sự nghiên cứu sâu hơn để hiểu rõ về các yếu tố ảnh hưởng.

4.1.3 Biểu đồ biểu diễn tỷ lệ nhận học bổng theo trạng thái hôn nhân.

Biểu đồ biểu diễn tỉ lệ nhận được và không nhận được học bổng theo trạng thái hôn nhân

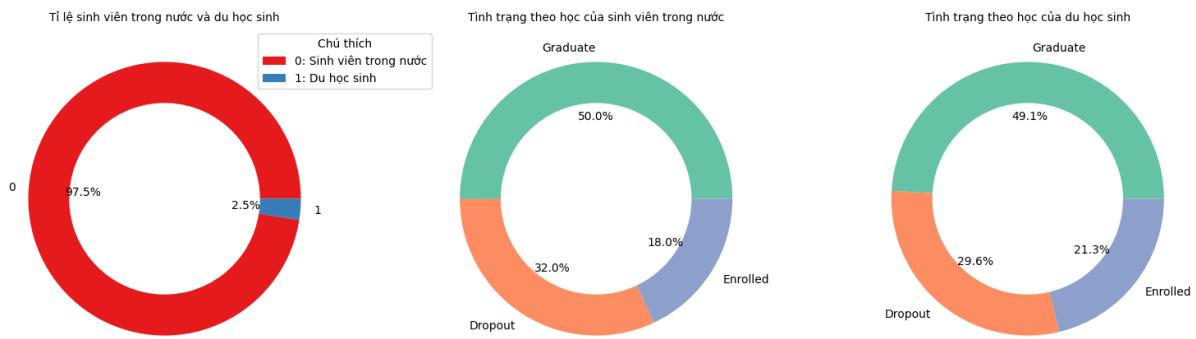


Hình 16: Biểu đồ biểu diễn tỷ lệ nhận học bổng theo trạng thái hôn nhân

- **Nhận xét:**

- + Nhận xét chung: chỉ một nhóm nhỏ trong từng nhóm đối tượng trên được có học bổng
- + Độc thân - Ly hôn - Ly thân - Góa phụ: các nhóm này có tỉ lệ nhận học bổng cao hơn các nhóm khác vì không có hoặc đã qua thời kỳ dành nhiều thời gian cho hôn nhân hay những mối quan hệ tình cảm. Vì đó sẽ có nhiều thời gian hơn và tập trung cho việc học dẫn đến tỉ lệ nhận học bổng cao hơn.
- + Sống thử - Kết hôn: Nhóm này có nhiều vấn đề quan tâm hơn trong chuyện tình cảm và gia đình. Do việc phải phân bổ thời gian và sự quan tâm cho cả việc học và cuộc sống cá nhân nên nhóm này có xu hướng ít nhận được học bổng hơn

4.1.4 Biểu đồ biểu diễn số lượng sinh viên trong nước và quốc tế theo trạng thái học tập của sinh viên:

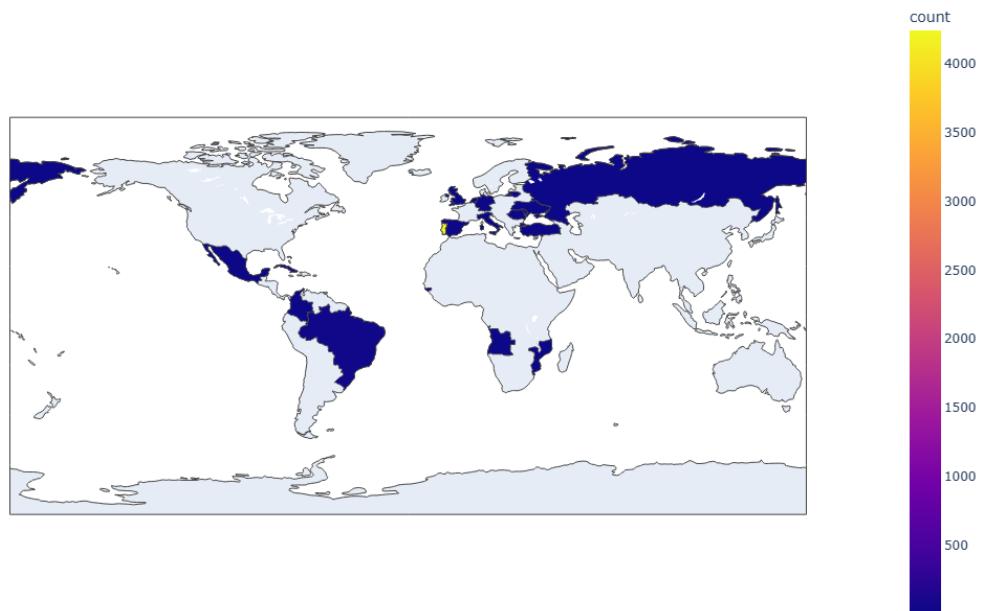


Hình 17: Biểu đồ biểu diễn số lượng sinh viên trong nước và quốc tế theo trạng thái học tập của sinh viên

- **Nhận xét:**

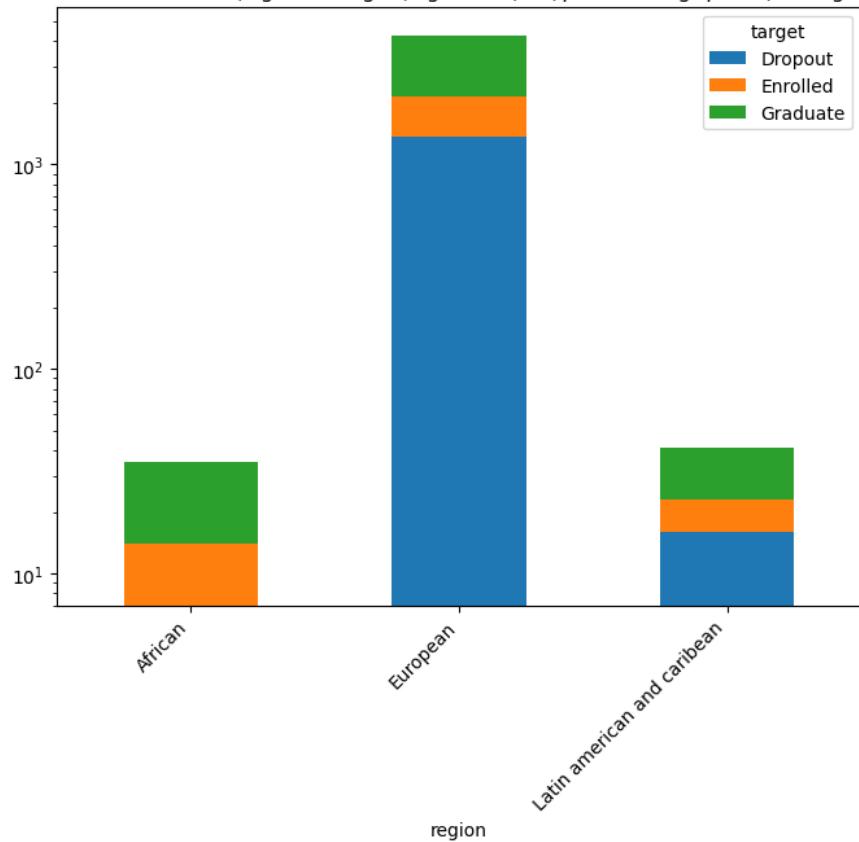
- + Qua biểu đồ đầu tiên chúng ta có thể thấy được một sự chênh lệch đáng kể trong số lượng giữa 2 nhóm sinh viên. Với nhiều lý do, có thể đây không phải là một điểm đến của những du học sinh, hoặc đơn vị giáo dục này chưa có kế hoạch nhắm tới nhóm học sinh này.
- + Tuy vậy, tỉ lệ trạng thái học tập giữa sinh viên trong nước và sinh viên quốc tế có vẻ tương đồng, hơn nữa cho thấy được rằng tỉ lệ tốt nghiệp của sinh viên theo 2 nhóm này vẫn ở mức cao. Đây là một tín hiệu tích cực rằng sinh viên ngoại quốc không bị ảnh hưởng bởi các yếu tố bên ngoài nhiều dẫn tới áp lực và nghỉ học.
- + Tín hiệu tích cực này có thể là kết quả của môi trường học tập tích cực, hỗ trợ sinh viên ngoại quốc để vượt qua thách thức và duy trì cam kết học tập. Điều này cũng có thể chỉ ra rằng hệ thống giáo dục đang thành công trong việc tạo ra một môi trường học tập tích cực và thân thiện với sinh viên quốc tế, khuyến khích họ tiếp tục đào tạo và tốt nghiệp một cách hiệu quả.

4.1.5. Biểu đồ biểu diễn số lượng sinh viên theo khu vực.



Hình 18: Biểu đồ biểu diễn số lượng sinh viên theo khu vực

Biểu đồ biểu diễn số lượng của từng trạng thái học tập theo vùng quốc tịch (log-scaled)



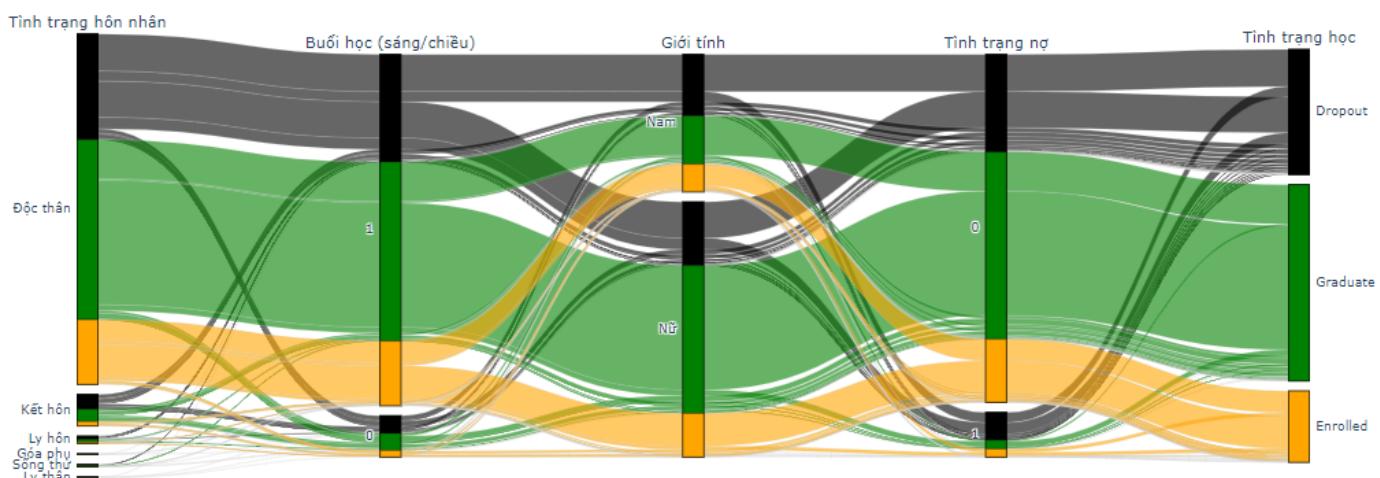
Hình 19: Biểu đồ biểu diễn số lượng sinh viên theo vùng quốc tịch

Thay đổi biến quốc tịch (nationality) thành biến vùng (region) vì có quá nhiều quốc gia:

- **Nhận xét:**
- + European có lượng sinh viên đông đảo, gấp nhiều lần so với African và Latin American & Caribbean. Sự chênh lệch lớn có thể phản ánh sự phát triển kinh tế và hệ thống giáo dục mạnh mẽ trong khu vực này.
- + Tuy nhiên European cũng là nhóm sinh viên Dropout nhiều nhất trong cả 3 nhóm sinh viên. Hiện tượng này có thể giải thích bằng một hoặc nhiều yếu tố sau đây:
 - + *Áp lực học tập:* Môi trường học tập ở các quốc gia châu Âu thường rất cạnh tranh và yêu cầu cao. Áp lực này có thể gây stress cho sinh viên và dẫn đến quyết định bỏ học nếu họ cảm thấy không thể đáp ứng được yêu cầu.
 - + *Áp lực từ xã hội và tâm lý lựa chọn ngành:* Có thể có áp lực hoặc kỳ vọng từ gia đình hoặc xã hội dẫn đến lựa chọn ngành nghề của sinh viên không đúng với mong muốn thực sự của chính bản thân người chọn. Nếu sinh viên không cảm thấy họ đang theo đuổi sự nghiệp mà họ thực sự quan tâm, họ có thể quyết định bỏ học để tìm kiếm các lựa chọn khác.
 - + *Thiếu sự hỗ trợ về tài chính:* Mặc dù quyết định theo học một ngành nghề nào đó có thể đúng như mong muốn của sinh viên nhưng việc gánh vác những gánh nặng tài chính như học phí, chi phí sinh hoạt đắt đỏ cũng trở thành một trở ngại lớn đối với người học. Việc thiếu đi các hỗ trợ tài chính như học bổng, tài trợ có thể khiến gánh nặng đó trở nên lớn hơn dẫn đến lý do bỏ học của sinh viên.
 - + *Mất cân bằng giữa học tập - công việc - đời sống:* Công việc làm thêm có thể giúp sinh viên trang trải một phần gánh nặng tài chính nhưng đối với những sinh viên không thể cân bằng giữa 3 yếu tố trên có thể dẫn đến những vấn đề phát sinh và phải bỏ học

4.1.6 Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Buổi học, Tình trạng nợ và Tình trạng học

Biểu đồ biểu diễn số lượng sinh viên nghỉ học theo Giới tính, Buổi học, Tình trạng nợ và Tình trạng học



Hình 20: Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Buổi học, Tình trạng nợ và Tình trạng học

- **Nhận xét:**

- + Tình Trạng Hôn Nhân: Phần lớn sinh viên là "Single" (Độc thân), tiếp theo là "Married" (Đã kết hôn) và một số rất ít trong nhóm "Other" (Khác). Có thể thấy luồng từ "Single" chiếm ưu thế, điều này có thể phản ánh đặc điểm đối tượng sinh viên chủ yếu chưa kết hôn.
- + Buổi Học: Các luồng từ "Buổi học (sáng/chiều)" cho thấy xu hướng học ban ngày vẫn chiếm đa số và những sinh viên học buổi tối có xu hướng có nợ nhiều hơn so với nhóm còn lại.
- + Giới Tính: Tỉ lệ sinh viên nam và nữ được biểu diễn có sự chênh lệch, trong đó thì sinh viên nữ chiếm ưu thế về số lượng và tỷ lệ sinh viên nữ còn độc thân, tham gia học vào ban ngày, không có nợ và đã tốt nghiệp chiếm đa số.
- + Tình Trạng Nợ: Đa số sinh viên không có nợ ("No"), và một tỷ lệ nhỏ hơn có nợ ("Yes"). Ngoài ra, có thể thấy rằng sinh viên có nợ có xu hướng nghỉ học nhiều hơn so với sinh viên không mang nợ.

❖ **Tổng kết:** Qua một số phân tích về các yếu tố nhân khẩu học tác động tới sinh viên, chúng ta rút ra được một số thông tin như sau:

- + Tuy có sự chênh lệch giữa số lượng sinh viên giữa hai giới, nhưng tỉ lệ bỏ học của sinh viên là ngang nhau. Tuy vậy tỉ lệ tốt nghiệp thì nữ giới vẫn cao hơn rất nhiều so với nam giới. Đây là vấn đề quan sát thêm để cân bằng lại.
- + Có sự tác động của tình trạng hôn nhân đối với việc sinh viên nghỉ học khi những sinh viên tham gia hôn nhân có xu hướng nghỉ học nhiều hơn.
- + Có sự tương đồng giữa các thông số tỉ lệ nghỉ học và tốt nghiệp giữa sinh viên trong nước và du học sinh. Qua đó giúp nhận thấy rằng đây là môi trường học tập tích cực, các du học sinh không bị ảnh hưởng đáng kể hoặc có sự trợ giúp từ đơn vị giáo dục trước các yếu tố mới từ môi trường học tập tới sinh hoạt, qua đó làm giảm tỉ lệ nghỉ học của nhóm sinh viên này.
- + Tuy nhiên vẫn còn nhiều vấn đề như:
 - + Các yếu tố như hôn nhân, quốc tịch có thể đặt ra thách thức về cả khả năng đạt học bổng cũng như quyết định nghỉ học của sinh viên.

⇒ Cần tìm kiếm giải pháp hỗ trợ và giúp đỡ các sinh viên. Đồng thời qua đó có thể thu thập thêm những nhận xét của sinh viên và giúp tăng cường môi trường học tập tích cực, cung cấp hỗ trợ đặc biệt cho sinh viên có tình trạng hôn nhân, cũng như phát triển các chương trình hỗ trợ đặc biệt cho sinh viên quốc tế, có thể là những bước quan trọng trong việc giảm tỉ lệ nghỉ học và tăng cường thành công học tập. Đồng thời, nghiên cứu và triển khai các chính sách học bổng linh hoạt có thể giúp đỡ các sinh viên đang đối mặt với những khó khăn tài chính, giúp họ duy trì cam kết và tiếp tục con đường học vụ của mình.

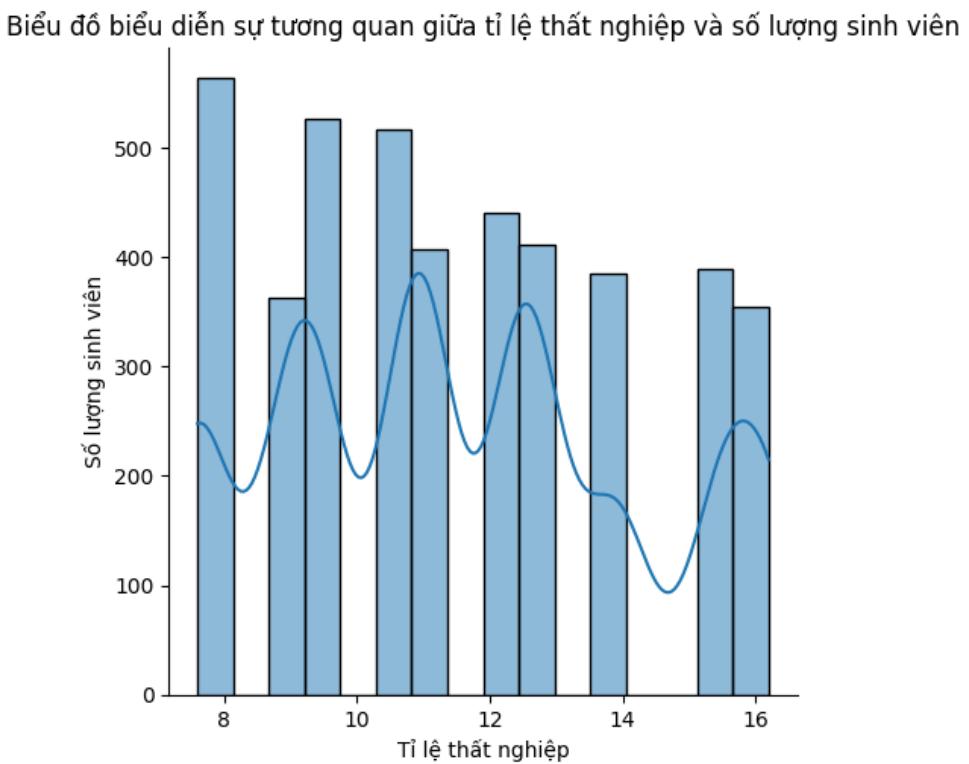
4.2 Biểu diễn theo tình hình kinh tế - xã hội

4.2.1 Biểu đồ biểu diễn sự tương quan giữa tình trạng học tập và tỷ lệ thất nghiệp



Hình 21: Biểu đồ biểu diễn sự tương quan giữa tình trạng học tập và tỉ lệ thất nghiệp (1)

- **Nhận xét:**
- + Qua biểu đồ trên chúng ta không thấy có sự ảnh hưởng gì của tỉ lệ thất nghiệp tới tình trạng sinh viên nghỉ học hay tốt nghiệp. Nhưng hình như có sự tương quan giữa số lượng sinh viên và tỉ lệ thất nghiệp khi các biểu đồ có xu hướng thấp hơn về cuối.

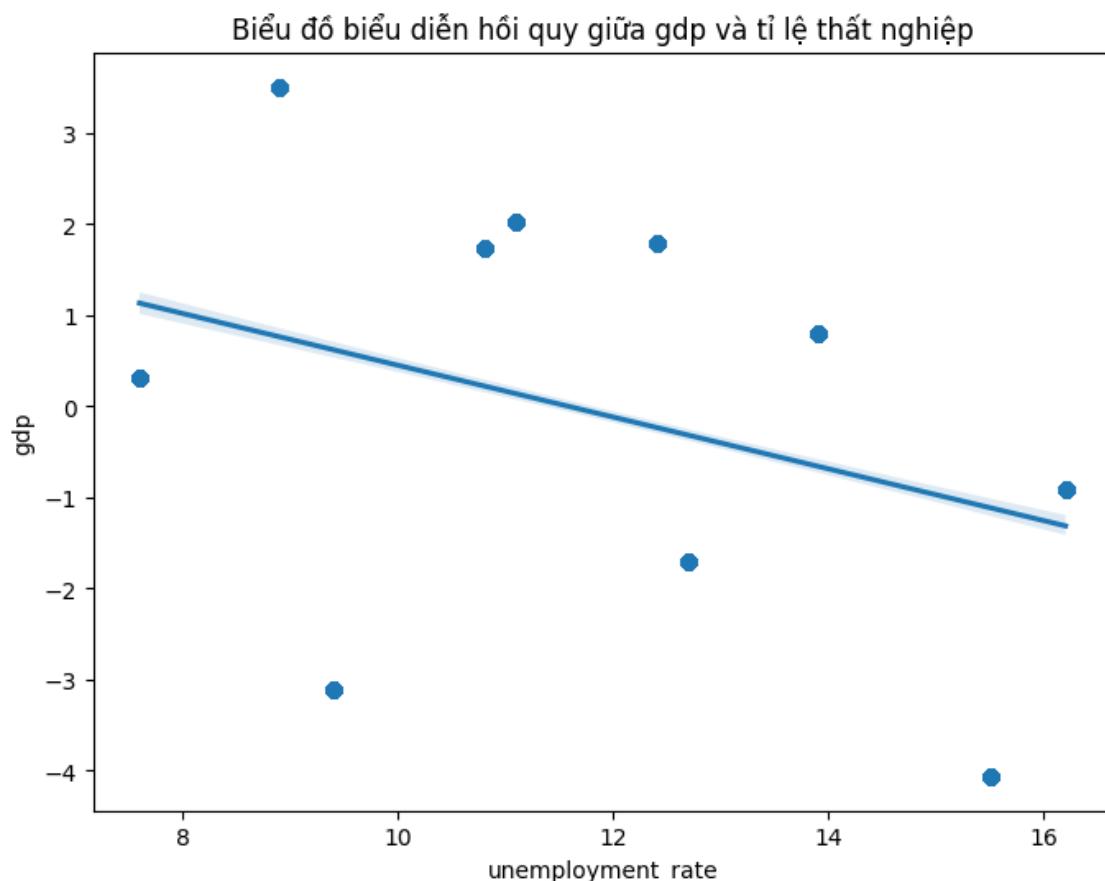


Hình 22: Biểu đồ biểu diễn sự tương quan giữa tình trạng học tập và tỉ lệ thất nghiệp (2)

- **Nhận xét:**

- + Có thể thấy được rõ ràng hơn ở biểu đồ này rằng có sự giảm của số lượng sinh viên khi tỷ lệ thất nghiệp tăng.
- + Làm nổi bật mối liên kết mạnh mẽ giữa sức khỏe của nền kinh tế và khả năng tiếp tục học của sinh viên. Sự giảm hỗ trợ tài chính, có thể xuất phát từ nguồn tài trợ gia đình giảm bớt hoặc mất đi, đặt ra áp lực không nhỏ đối với sinh viên. Các sinh viên vừa học vừa làm để chi trả học phí cũng trở nên tỏ ra yếu đuối trước sự biến động trong thị trường lao động.

4.2.2 Biểu đồ biểu diễn hồi quy giữa gdp và tỷ lệ thất nghiệp



Hình 23: Biểu đồ biểu diễn hồi quy giữa gdp và tỷ lệ thất nghiệp

- **Nhận xét:**

- + Quan hệ Giữa GDP và Tỷ Lệ Thất Nghiệp: Có vẻ như tỷ lệ thất nghiệp có mối quan hệ nghịch với GDP; khi tỷ lệ thất nghiệp tăng thì GDP giảm. Điều này được biểu thị bởi đường hồi quy có xu hướng giảm.
- + Phân Bó Dữ Liệu: Các điểm dữ liệu phân bố rộng trên biểu đồ, cho thấy mức độ biến động của GDP ở các mức tỷ lệ thất nghiệp khác nhau.
- + Có thể thấy được có mối tương quan âm mặc dù không rõ ràng giữa GDP và tỷ lệ thất nghiệp. Và ở biểu đồ “hình 22” ở trên chúng ta thấy được khi tỷ lệ thất nghiệp tăng sẽ làm số lượng sinh viên đăng ký học giảm. Điều này thể hiện rằng cũng có thể có mối liên quan giữa tình hình kinh tế chung giữa và số lượng sinh viên.

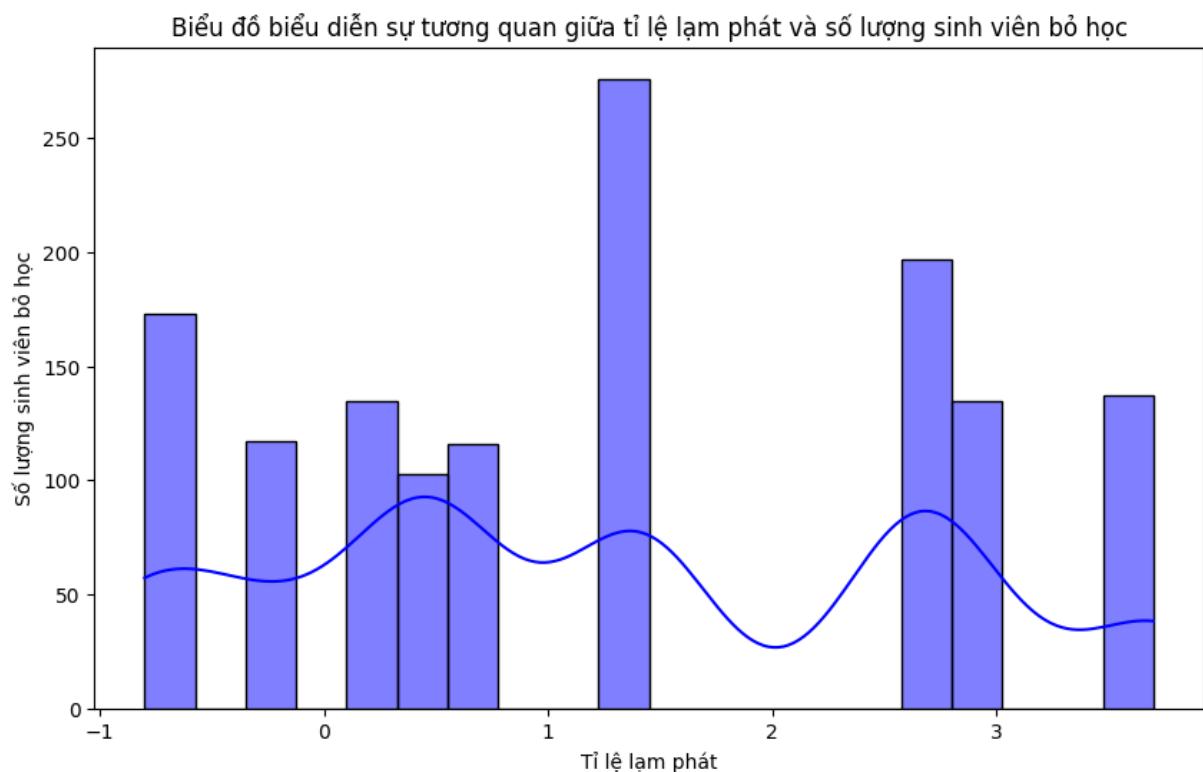
4.2.3 Biểu đồ biểu diễn hồi quy giữa tỷ lệ lạm phát và tỷ lệ thất nghiệp:



Hình 24: Biểu đồ biểu diễn hồi quy giữa tỷ lệ lạm phát và tỷ lệ thất nghiệp

- **Nhận xét:**
- + Quan Hệ Giữa Hai Biến: Đường hồi quy tuyến tính trên biểu đồ cho thấy không có mối quan hệ rõ ràng hoặc mạnh mẽ giữa tỷ lệ lạm phát và tỷ lệ thất nghiệp. Đường hồi quy có vẻ khá ngang, cho thấy rằng không có sự thay đổi đáng kể của tỷ lệ lạm phát khi tỷ lệ thất nghiệp thay đổi.
- + Không có sự liên hệ giữa tỷ lệ lạm phát và tỷ lệ thất nghiệp trong bộ dữ liệu này.

4.2.4 Biểu đồ biểu diễn sự tương quan giữa tỷ lệ lạm phát và số lượng sinh viên bỏ học:



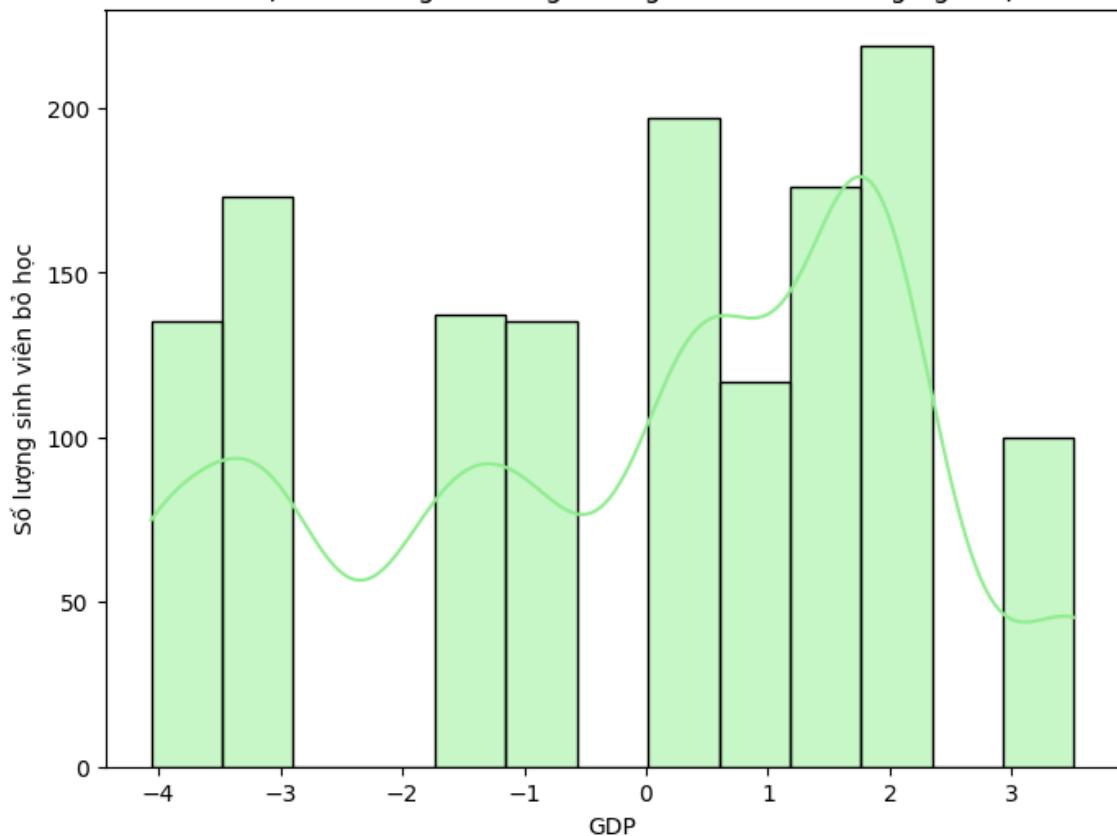
Hình 25: Biểu đồ biểu diễn sự tương quan giữa tỷ lệ lạm phát và số lượng sinh viên bỏ học

- **Nhận xét:**

- + Biểu đồ cho thấy ảnh hưởng của tỷ lệ lạm phát đến khả năng bồi học của sinh viên mặc dù có nhưng không nhiều. Tại mức lạm phát 1.4%, sự tăng đột ngột ở đó có thể là dấu hiệu của một mức lạm phát khiến sinh viên gặp khó khăn.
- + Tỷ lệ lạm phát khi từ âm qua dương dường như đã ảnh hưởng khá nhiều tới số lượng sinh viên bỏ học, nhưng về sau thì ổn định dần hơn. Có thể là do sinh viên chưa kịp thích nghi với sự biến động của nền kinh tế, vật giá leo thang làm họ bị áp lực về tài chính và đi tới quyết định nghỉ học. Tuy nhiên sau đó đã có sự giảm đáng kể, có thể là do chính phủ và các đơn vị giáo dục đã có những hỗ trợ kịp thời để sinh viên ổn định trạng thái của mình.
- + Tuy vậy, để đánh giá rõ hơn về mối quan hệ này, cần kết hợp với các phương pháp thống kê và phân tích chi tiết khác cũng như thu thập số lượng dữ liệu nhiều hơn để có cái nhìn tổng thể về vấn đề này.

4.2.5 Biểu đồ biểu diễn sự ảnh hưởng của tăng trưởng GDP tới khả năng nghỉ học của học sinh

Biểu đồ biểu diễn sự ảnh hưởng của tăng trưởng GDP tới khả năng nghỉ học của sinh viên

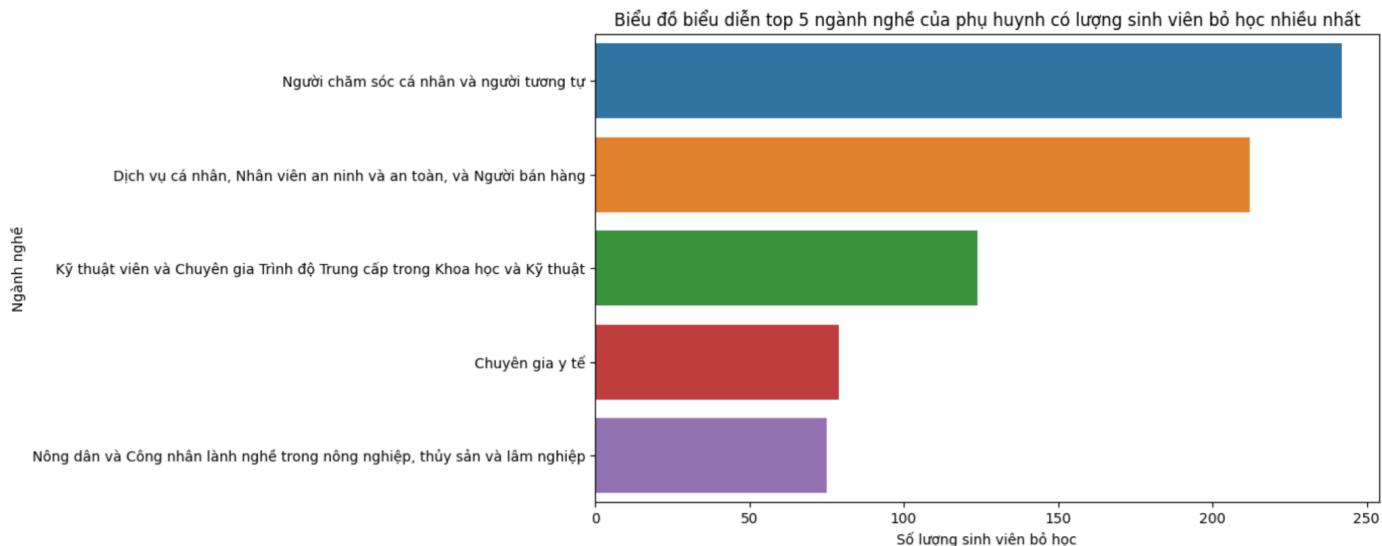


Hình 26: Biểu đồ biểu diễn sự ảnh hưởng của tăng trưởng GDP tới khả năng nghỉ học của học sinh

- **Nhận xét:**

- + Biểu đồ trên cho thấy có một xu hướng tăng dần về số lượng sinh viên bỏ học khi tỷ lệ tăng trưởng GDP tăng đỉnh điểm là ở khoảng từ 1.5 - 2%. Việc GDP tăng có thể phản ánh một số mặt tích cực của tăng trưởng kinh tế nhưng nghịch lý là nó kéo theo số lượng sinh viên nghỉ học tăng lên. Có thể đây là do tăng trưởng không đồng đều, có thể tạo ra áp lực tài chính. Có thể do giáo dục trở nên đắt đỏ hơn do tăng giá học phí, chi phí sinh hoạt, và các chi phí khác. Điều này có thể tạo ra áp lực tài chính đặc biệt đối với sinh viên và gia đình, dẫn đến quyết định bỏ học.
- + Điều đáng chú ý là sau sự tăng mạnh, ta thấy một sự giảm mạnh ngay sau đó. Có thể là do sinh viên nhanh chóng thích nghi hoặc có những biện pháp hỗ trợ được triển khai để giảm áp lực. Điều này có thể bao gồm các biện pháp tài chính, hỗ trợ học thuật, hoặc chính sách hỗ trợ sinh viên.
- + Điều này cho thấy tỷ lệ tăng trưởng GDP không ảnh hưởng đến quyết định nghỉ học của sinh viên. Nhưng để chắc chắn ta cần phân tích sâu hơn để biết được sự liên hệ đằng sau đó.

4.2.6 Biểu đồ biểu diễn top 5 ngành nghề của phụ huynh có lượng sinh viên bỏ học nhiều nhất



Hình 27: Biểu đồ biểu diễn top 5 ngành nghề của phụ huynh có lượng sinh viên bỏ học nhiều nhất

- Nhận xét:

- + *Ảnh Hưởng Của Yếu Tố Tài Chính và An Sinh Xã Hội:* Các ngành nghề như chăm sóc cá nhân, an ninh và nông nghiệp có thể chỉ ra rằng yếu tố gánh nặng tài chính và môi trường làm việc phải làm việc nhiều giờ hoặc có những yêu cầu công việc căng thẳng, có thể là những yếu tố quan trọng khiến sinh viên quyết định bỏ học. Các ngành nghề này thường đối mặt với áp lực tài chính và điều kiện làm việc khó khăn.
- + *Tác Động Của Yếu Tố Tâm Lý:* Các ngành nghề y tế và kỹ thuật thường đòi hỏi kiến thức và kỹ năng cao cũng có thể liên quan đến áp lực gia đình và kỳ vọng về sự thành công trong ngành này, đặc biệt nếu sinh viên không đáp ứng được kỳ vọng..
- + *Sự Biến Động Môi Trường Của Nông Nghiệp:* Số sinh viên bỏ học có phụ huynh làm trong lĩnh vực nông, ngư, lâm nghiệp cao có thể liên quan đến khả năng tài chính gia đình, chi phí đào tạo, hoặc trong việc kinh doanh.

❖ **Tổng kết:** Qua việc phân tích các yếu tố kinh tế ảnh hưởng đến quyết định nghỉ học của sinh viên, chúng ta nhận thức được rằng mối liên kết giữa sức khỏe của nền kinh tế và khả năng tiếp tục học của sinh viên là rất rõ ràng.

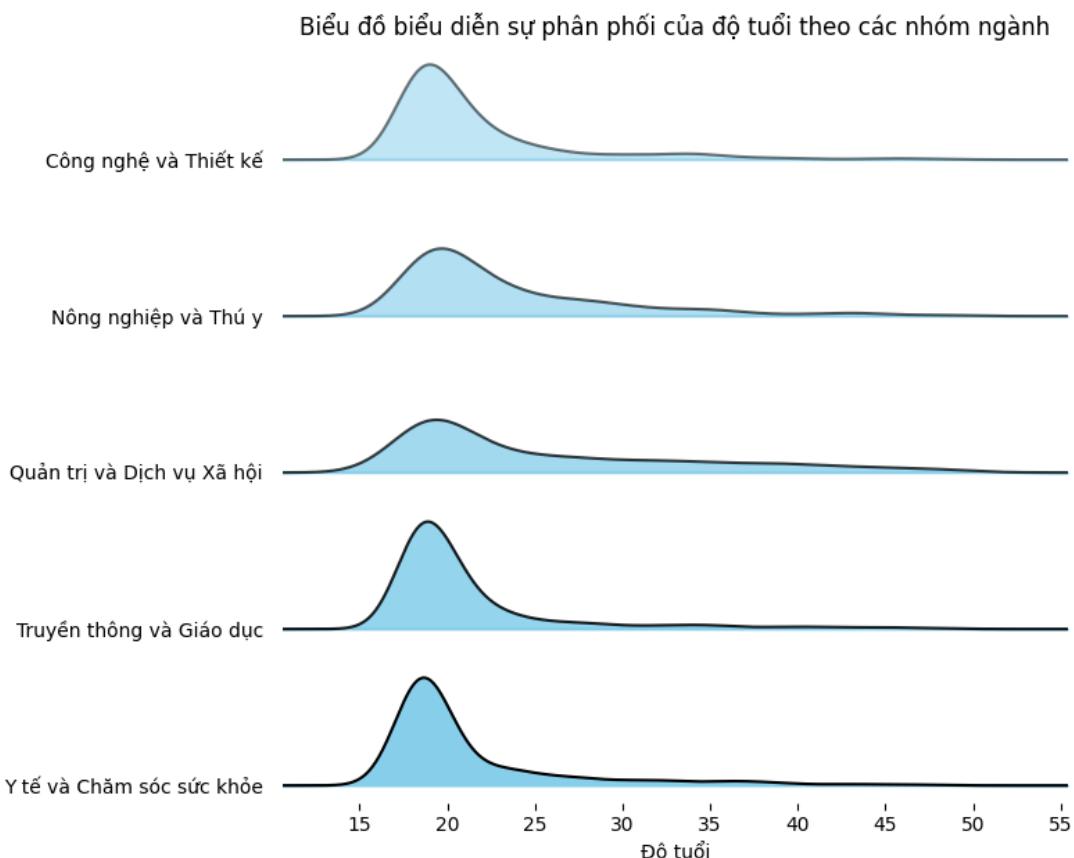
- + *Sự giảm hỗ trợ tài chính và biến động trong thị trường lao động* có thể đặt ra áp lực lớn đối với sinh viên, đặc biệt là nhóm sinh viên đang làm việc để chi trả học phí.
- + *Mối quan hệ giữa GDP và tỷ lệ thất nghiệp* có vẻ như là yếu tố quan trọng, và sự biến động của GDP ảnh hưởng đáng kể đến số lượng sinh viên. Tuy nhiên, cần thêm nghiên cứu và phân tích chi tiết để hiểu rõ hơn về mối liên hệ này và các yếu tố nguyên nhân khác có thể ảnh hưởng đến quyết định nghỉ học.
- + Ngoài ra, *sự tác động của phụ huynh* khi làm việc ở những ngành nghề đặc thù có thể đưa ra những kỳ vọng lớn mà sinh viên không đáp ứng được dẫn tới áp lực, cũng có thể là sự truyền con nối khiến sinh viên bỏ học và về làm với gia đình, hoặc ngành nghề bị phụ

thuộc vào tình hình kinh tế khiến họ không thể trợ cấp học phí cho sinh viên, đây cũng có thể là những tác nhân làm tăng số lượng sinh viên nghỉ học.

⇒ **Kết luận:** Sau khi hiểu rõ những yếu tố này có thể giúp chúng ta phát triển các chính sách và biện pháp hỗ trợ phù hợp để giảm tỉ lệ nghỉ học cũng như tạo điều kiện cho việc học tập ổn định của sinh viên trong bối cảnh biến động của nền kinh tế và xã hội.

4.2 Biểu diễn theo tình hình học tập

4.2.1 Biểu đồ biểu diễn sự phân phối của độ tuổi theo các nhóm ngành:

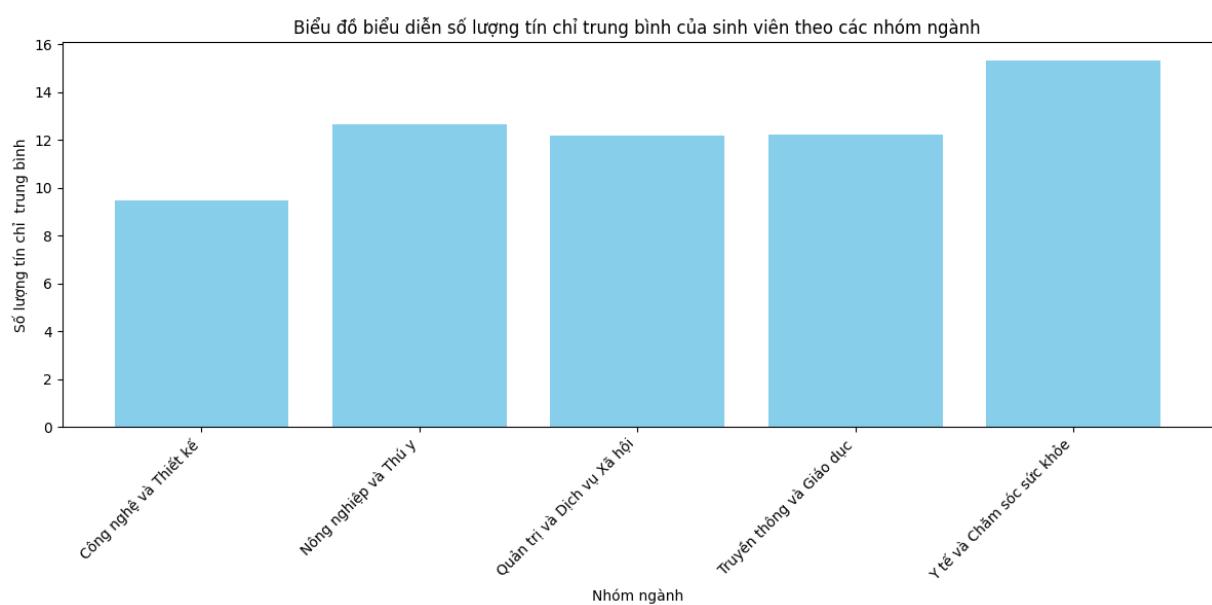


Hình 28: Biểu đồ biểu diễn sự phân phối của độ tuổi theo các nhóm ngành

- **Nhận xét:**
- + Qua biểu đồ trên, chúng ta có thể thấy được sự phân phối độ tuổi chia thành 2 nhóm, một nhóm có độ tuổi trẻ, và một nhóm có độ tuổi trung bình cao hơn, trải dài tới trên 25-30 tuổi;
- + Nhóm ngành Công nghệ và Thiết kế, Truyền thông và Giáo dục, Y tế và Chăm sóc sức khỏe:
 - + Có sự tương đồng trong phân phối độ tuổi, với sự tập trung chủ yếu vào đối tượng 18-20 tuổi.
 - + Cho thấy đây là những ngành có được nhiều sự quan tâm đặc biệt của giới trẻ, kết hợp với thời gian đào tạo linh hoạt, chủ yếu là thông qua tích lũy kinh nghiệm trong cả quá trình học và làm việc thực tế. Điều này giúp các ngành này đáp ứng nhanh chóng và linh hoạt với nhu cầu thị trường ngày càng tăng.
- + Nhóm ngành Quản trị và Dịch vụ xã hội, Nông nghiệp và Thú y:

- + Phân bố độ tuổi có vẻ trải rộng hơn, có thể chỉ ra sự linh hoạt trong thời gian đào tạo và yêu cầu về những kiến thức kinh nghiệm liên quan trong lĩnh vực này.
- + Ngoài ra, các ngành này ít có sự cạnh tranh giữa những nhóm người có độ tuổi khác nhau. Vậy nên những người ngoài độ tuổi trung bình của sinh viên vẫn có thể đăng ký theo học, nhằm nâng cao trình độ chuyên môn hoặc cũng có thể là sự thay đổi trong nghề nghiệp hiện tại.
- + Tuy biểu đồ không được mô tả chi tiết, nhưng những nhận định trên có thể giúp ta hiểu được xu hướng tổng quan về độ tuổi và mức độ học vấn trong các ngành nghề khác nhau. Đồng thời, cũng có thể suy luận về sự linh hoạt của thời gian đào tạo và đặc tính của kiến thức trong từng lĩnh vực nghề nghiệp.

4.1.4 Biểu đồ biểu diễn số lượng tín chỉ trung bình của sinh viên theo các nhóm ngành

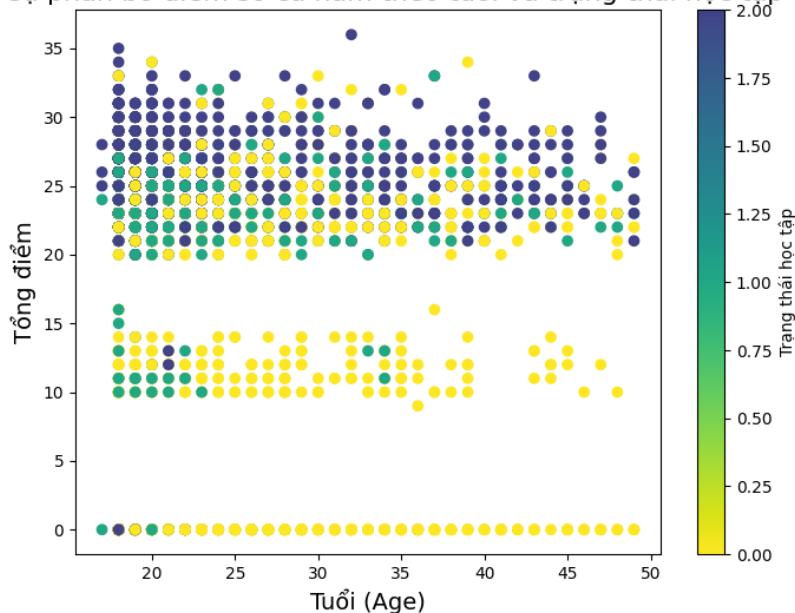


Hình 29: Biểu đồ biểu diễn số lượng tín chỉ trung bình của sinh viên theo các nhóm ngành

- **Nhận xét:**
- + Số lượng tín chỉ trung bình mà một sinh viên đăng ký trong năm học cao nhất nằm ở nhóm ngành Y tế và chăm sóc sức khỏe. Điều này là do khối lượng kiến thức mà nhóm ngành này yêu cầu lớn so với các nhóm ngành khác. Đặc thù công việc yêu cầu sự hiểu biết sâu rộng và chính xác trong việc xác định bệnh tình cũng là một nguyên nhân dẫn đến lượng tín chỉ đăng ký mỗi kỳ cao.
- + Với đặc thù là nhóm ngành phải thường xuyên thay đổi để thích nghi với sự tiến bộ chóng mặt của khoa học kỹ thuật, nhóm ngành công nghệ và thiết kế thường có lượng tín chỉ mỗi năm thấp hơn các ngành được trình bày ở trên vì những nguyên nhân như: các giờ học lý thuyết được giảm bớt để sinh viên có thêm thời gian tìm hiểu, nghiên cứu và ứng dụng các công nghệ mới để giải quyết các bài toán thực tế, thời gian cho những nghiên cứu và thử nghiệm công nghệ mới, các dự án đội nhóm/cá nhân,...

4.3.3 Biểu đồ biểu diễn sự phân bố điểm số cả năm theo tuổi và trạng thái học tập của sinh viên:

Biểu đồ biểu diễn sự phân bố điểm số cả năm theo tuổi và trạng thái học tập của sinh viên



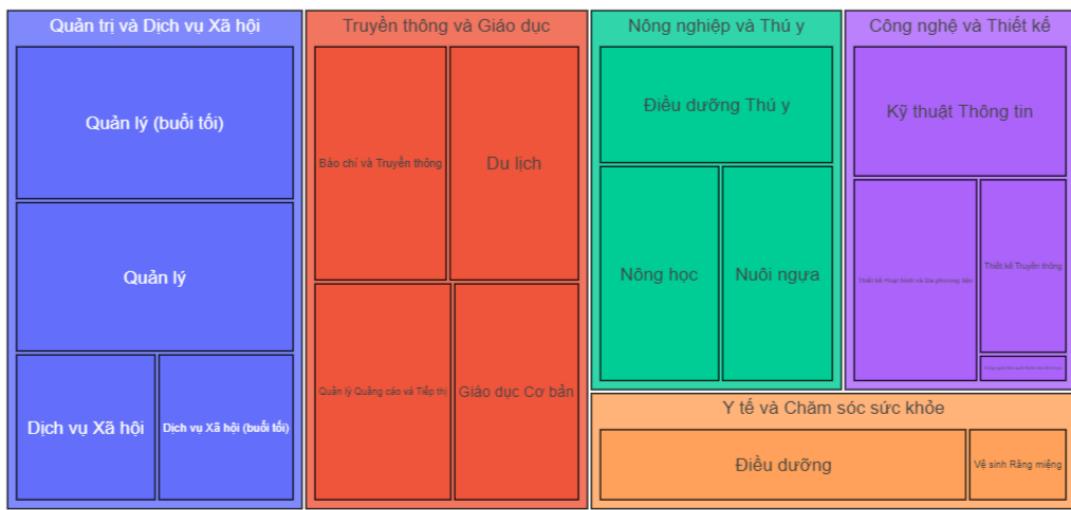
Hình 30: Biểu đồ biểu diễn sự phân bố điểm số cả năm theo tuổi và trạng thái học tập của sinh viên

- **Nhận xét:**

- + Biểu đồ được chia làm 2 phần rõ rệt tại mức điểm 18 - 19, dưới mức này là những sinh viên có học lực dưới trung bình và hầu hết trong số đó đều đã bỏ học tại thời điểm thu thập số liệu với nhiều lý do đã phân tích ở các phần trước.
- + Phần trên của biểu đồ, là những sinh viên có học lực khá giỏi thì trong số đó hầu hết các sinh viên đều tốt nghiệp ở độ tuổi 20 đến 25, là một độ tuổi lý tưởng và thường thấy.
- + Dù vậy với điểm số và học lực ở mức này vẫn có không ít sinh viên lựa chọn bỏ học. Việc này có thể do những vấn đề chúng ta khám phá ra trước đó như hôn nhân, gia đình, tình hình kinh tế xã hội,... dẫn tới việc nghỉ học của sinh viên.

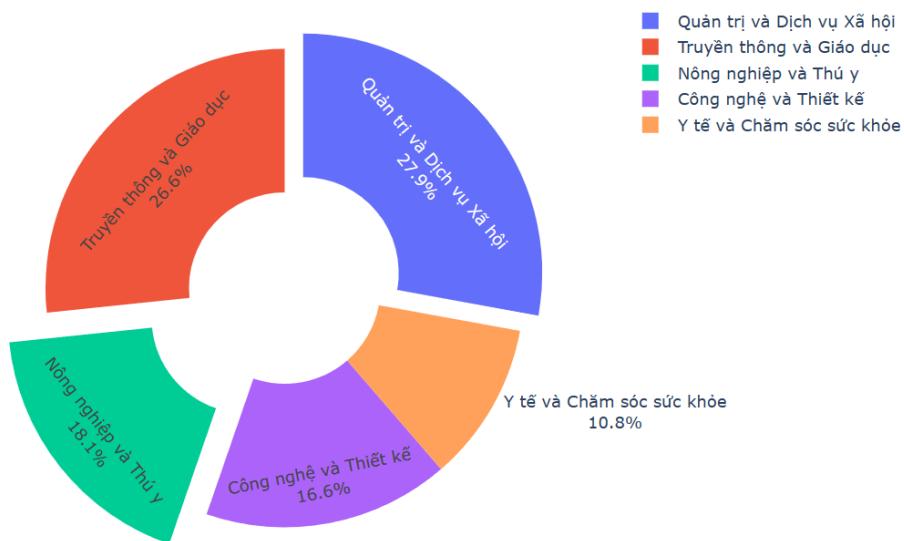
4.3.4. Biểu đồ biểu diễn tỉ lệ bỏ học của các nhóm ngành học.

Biểu đồ biểu diễn số lượng sinh viên nghỉ học theo các ngành, nhóm ngành



Hình 31: Biểu đồ biểu diễn số lượng sinh viên bỏ học của các nhóm ngành học

Biểu đồ thể hiện phần trăm sinh viên bỏ học theo nhóm ngành học

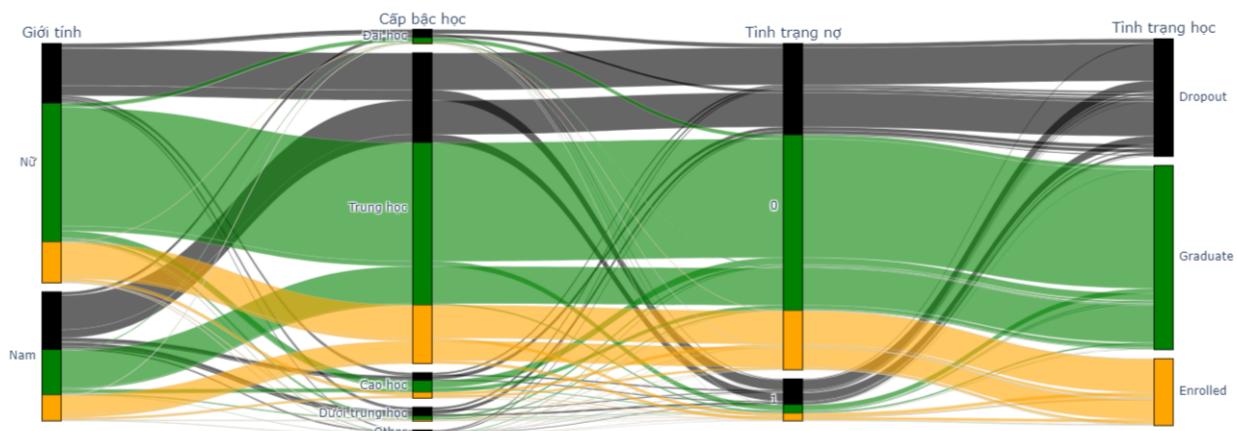


Hình 32: Biểu đồ biểu diễn tỉ lệ bỏ học của các nhóm ngành học

- **Nhận xét:**
- + Nhìn vào 2 biểu đồ, chúng ta thấy được rằng Quản trị và Dịch vụ xã hội cũng như Truyền thông và Giáo dục có số lượng và tỉ lệ sinh viên nghỉ học là cao nhất, tới hơn 25%:
 - + Điều này có thể là dấu hiệu của một vấn đề trong chương trình đào tạo hoặc phương pháp giáo dục có thể chưa phù hợp với sinh viên.
 - + Cũng như đây là 2 ngành liên quan tới xã hội và phục vụ công chúng, từ đó sản sinh ra những yêu cầu về kỹ năng xã hội và tư duy công đồng cao. Sinh viên không cảm thấy hài lòng hoặc không thích nghi được với những khía cạnh này có thể dẫn đến quyết định nghỉ học.
- + Sau ngành truyền thông và giáo dục là ngành nông nghiệp và thú y với tỉ lệ là 18.1% (chiếm gần 20% tỷ lệ sinh viên) và ngành công nghệ thiết kế với tỉ lệ là 16.6% và ngành học có tỉ lệ sinh viên bỏ học ít nhất đó là ngành y tế và chăm sóc sức khỏe với 10.8% sinh viên bỏ học chiếm 1/10 tỉ lệ sinh viên.
- + Ngành ‘Nông nghiệp và Thú y’ cùng ‘Công nghệ và thiết kế’ cũng đạt gần $\frac{1}{2}$ tỉ lệ sinh viên bỏ học. Là một nhóm ngành yêu cầu tích lũy kinh nghiệm thực tiễn cao, cùng với đó là những nhóm ngành đang xu thế khi dịch vụ thú y cũng nghe công nghệ thông tin ngày càng phát triển. Kéo theo đó là lượng sinh viên đăng ký theo số đông, dần dần khi học sẽ mất đi sự yêu thích ban đầu và nghỉ học.
- + Y tế và chăm sóc sức khỏe là nhóm ngành được quan tâm hàng đầu ở các quốc gia, từ đó đầu vào nhóm ngành này luôn đứng top đầu và được đào tạo bài bản. Từ đó giúp tỉ lệ bỏ học ngành này luôn không quá cao so với các ngành khác.

4.3.5. Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Cấp bậc, Tình trạng nợ, Tình trạng học:

Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Cấp bậc, Tình trạng nợ, Tình trạng học

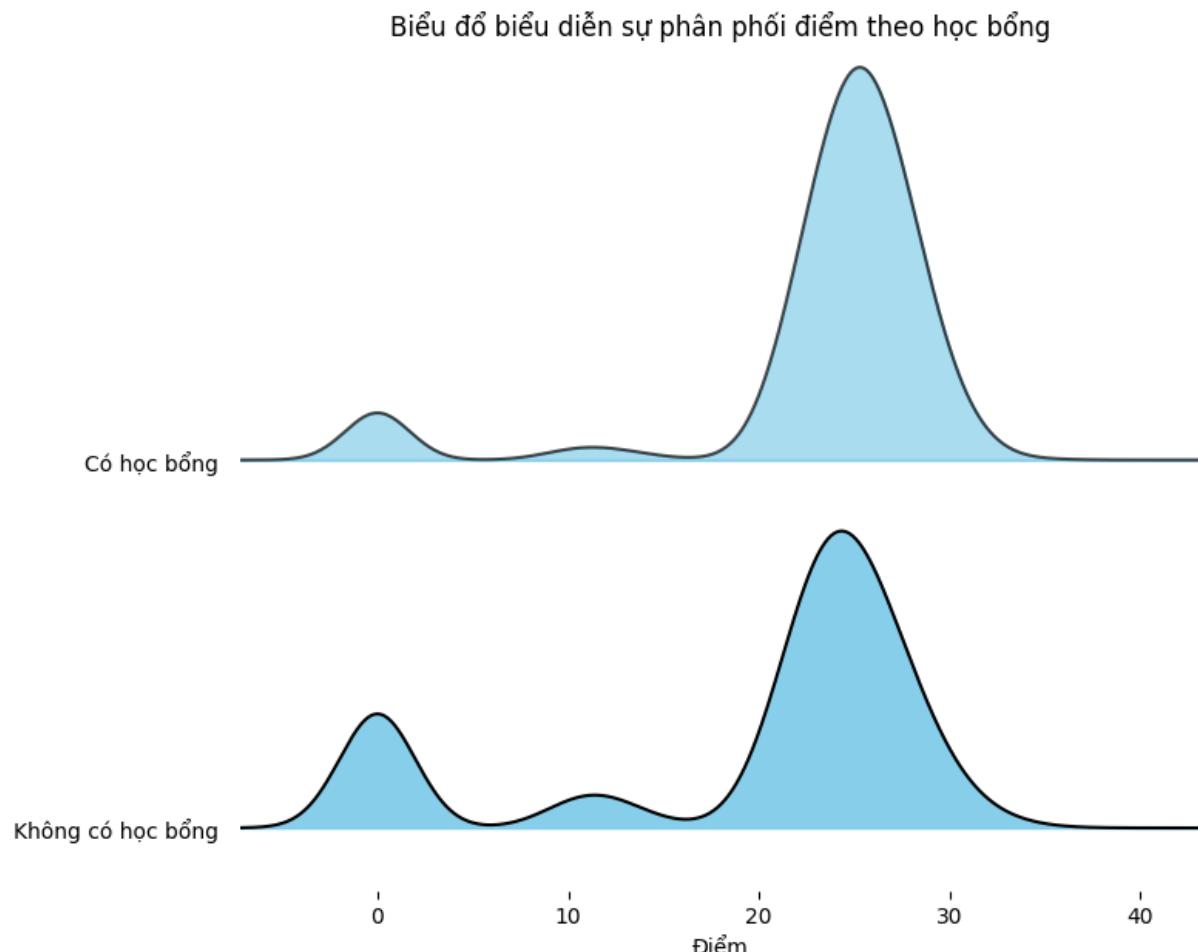


Hình 33: Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Cấp bậc, Tình trạng nợ, Tình trạng học

- **Nhận xét:**

- + Cấp bậc học : Người có bằng Trung học nhiều nhất, đại học với dưới trung học ngang nhau, có rất ít người đang học có bằng đại học, dưới trung học, other.
- + Có thể thấy được số lượng sinh viên đang theo học chủ yếu đến từ những sinh viên có trình độ trước khi vào học là “Trung học” và không mang nợ. Tuy vậy những sinh viên này cũng là nhóm sinh viên có số lượng mang nợ và bỏ học nhiều nhất. Có vẻ họ chưa thể đi làm để có thể tự chủ về mặt tài chính cho việc chi trả học phí.

4.3.6. Biểu đồ biểu diễn sự phân phối của điểm số theo học bổng:



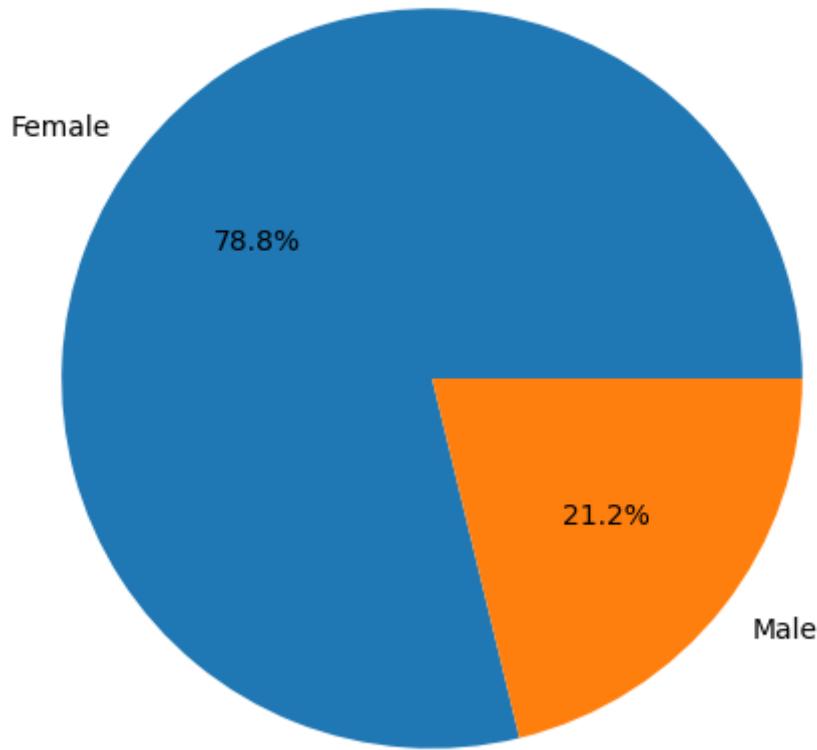
Hình 34: Biểu đồ biểu diễn sự phân phối của điểm số theo học bổng

- Nhận xét:

- + Chúng ta sẽ chú ý vào phổ điểm từ 20 đến 30 điểm vì đây là mức điểm phổ biến nhất của sinh viên cũng như được thể hiện trên biểu đồ bằng 1 đỉnh dữ liệu ở 2 biểu đồ.
- + Nguyên nhân dẫn đến việc có hoặc không có học bổng trong khi có cùng 1 mức điểm đến từ việc học bổng đó có thể dành cho một đối tượng được hưởng đãi ngộ đặc biệt, hoàn cảnh gia đình, đánh giá rèn luyện của sinh viên, các học bổng cho các trường hợp cụ thể do cá nhân/ tổ chức trao học bổng quy định.
- + Vì đó là phổ điểm thường thấy để nhận học bổng nhưng với những đỉnh dữ liệu nhỏ hơn trên 2 biểu đồ đặc biệt là tại điểm 0 ta có thể giải thích nó với các học bổng đầu vào. Vì tại thời điểm lấy dữ liệu chưa có dữ liệu về điểm cho những sinh viên mới.

4.3.7. Biểu đồ biểu diễn tỉ lệ đạt được học bổng giữa hai giới:

Biểu đồ biểu diễn tỉ lệ đạt được học bổng theo giới tính



Hình 35: Biểu đồ biểu diễn tỉ lệ đạt được học bổng giữa hai giới

- **Nhận xét:**
- + Biểu đồ thể hiện một sự chênh lệch rõ ràng giữa nam và nữ về tỉ lệ được hỗ trợ bằng học bổng.
- + Một phần như đã phân tích trước đó số lượng nữ sinh ở đây nhiều hơn nam sinh và các ngành học chủ đạo là về chăm sóc và y tế. Vậy nên số lượng nữ giới có thể đạt học bổng sẽ nhiều hơn nam giới. Và tỉ lệ này cũng cho thấy tín hiệu tích cực rằng nữ giới tại đơn vị giáo dục này có sự phân đều trong học tập nhiều hơn nam giới.
- + Tuy nhiên với tỉ lệ chênh lệch nhiều như vậy thì cũng sẽ phải xem lại rằng chính sách phân phối và điều kiện nhận học bổng của đơn vị giáo dục này đã phù hợp hay chưa để có sự phân bổ hợp lý hơn trong tương lai.

❖ **Tổng kết:**

- + Thông qua việc phân tích các biểu đồ và nhận xét về độ tuổi, tín chỉ, điểm số, và học bổng, chúng ta nhận thấy sự đa dạng của các yếu tố ảnh hưởng đến quá trình học tập của sinh viên.
- + Sự phân bố độ tuổi trong từng ngành nghề, cũng như số tín chỉ và điểm số, cho thấy tính đặc thù của từng ngành, nhóm ngành học mà sinh viên đã đăng ký.

- + Mối liên quan giữa số lượng sinh viên nghỉ học và các yếu tố như môi trường học tập, chương trình đào tạo, ngành nghề theo học là phức tạp và đòi hỏi sự quan sát sâu rộng để có thể đưa ra được những biện pháp hỗ trợ hợp lý. Các ngành như Quản trị và Dịch vụ xã hội, Truyền thông và Giáo dục có tỷ lệ nghỉ học cao có thể đặt ra câu hỏi về chất lượng chương trình và khả năng hỗ trợ sinh viên. Ngược lại, ngành Y tế và Chăm sóc sức khỏe có tỷ lệ thấp có thể là kết quả của sự chú trọng vào chất lượng giáo dục và cơ hội nghề nghiệp sau này.
- + Trong khi nhận học bổng có thể phản ánh sự cố gắng và tập trung của sinh viên đối với giáo dục cũng như sự quan tâm của cơ sở giáo dục đối với những hoàn cảnh, trường hợp cần hỗ trợ thì sự chênh lệch rõ ràng giữa tỉ lệ có học bổng giữa nam và nữ ở trường hợp này là cần xem xét. Điều này có thể là dấu hiệu của một vấn đề trong chính sách phân phối học bổng hoặc cần thiết lập lại các tiêu chí để đảm bảo sự công bằng giới tính trong quá trình trao học bổng.

⇒ Cuối cùng, để hiểu rõ hơn về các yếu tố ảnh hưởng đến quyết định nghỉ học, cần thêm nghiên cứu sâu rộng và sự hợp tác giữa các bên liên quan, từ các tổ chức giáo dục đến doanh nghiệp và chính phủ. Điều này sẽ giúp xây dựng các chiến lược hỗ trợ sinh viên và tối ưu hóa chất lượng giáo dục, từ đó giảm thiểu tỷ lệ nghỉ học và thúc đẩy sự thành công học tập của sinh viên.

CHƯƠNG V: KIỂM ĐỊNH

5.1 Nhân khẩu học

5.1.1 Kiểm định giả thuyết 1: Sinh viên khi tham gia vào hoạt động hôn nhân không làm ảnh hưởng tới quyết định nghỉ học

Các bước thực hiện:

Bước 1: Xác định giả thuyết kiểm định:

Ho: Sinh viên khi tham gia vào hoạt động hôn nhân không làm ảnh hưởng tới quyết định nghỉ học.

H1: Sinh viên khi tham gia vào hoạt động hôn nhân có làm ảnh hưởng tới quyết định nghỉ học của sinh viên.

Bước 2: Phân loại sinh viên ‘Độc thân’ và nhóm còn lại ‘Đã tham gia hôn nhân’.

Bước 3: Tạo bảng tần số (Crosstab) giữa biến Marital status và Target.

Bước 4: Tiến hành kiểm định Chi-square và kết luận theo phương pháp p-value (trị số p).

Với alpha = 0.05 và độ tin cậy = 95%. Điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

Code mẫu:

- Xem qua crosstab giữa 2 biến:

```
[15] df['Marital status'] = df['Marital status'].replace({
    1.0: 'Độc thân',
    2.0: 'Tham gia hôn nhân',
    3.0: 'Tham gia hôn nhân',
    4.0: 'Tham gia hôn nhân',
    5.0: 'Tham gia hôn nhân',
    6.0: 'Tham gia hôn nhân'
})

[17] Marital=df[['Target','Marital status']]
crosstab = pd.crosstab(Marital["Target"], Marital["Marital status"])
crosstab
```

Target	Marital status	Tham gia hôn nhân	Độc thân
Dropout		236	1185
Enrolled		74	720
Graduate		194	2015

- Ở đây chúng ta sẽ dùng kiểm định Chi-square:

```

[18] stat, p, dof, expected = chi2_contingency(crosstab)
     print('Degrees of freedom = %d' % dof)

Degrees of freedom = 2

[19] contingency_table = pd.crosstab(df['Marital status'], df['Target'])
     chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

     print(f"Chi-squared statistic: {chi2_stat}")
     print(f"P-value: {p_value}")

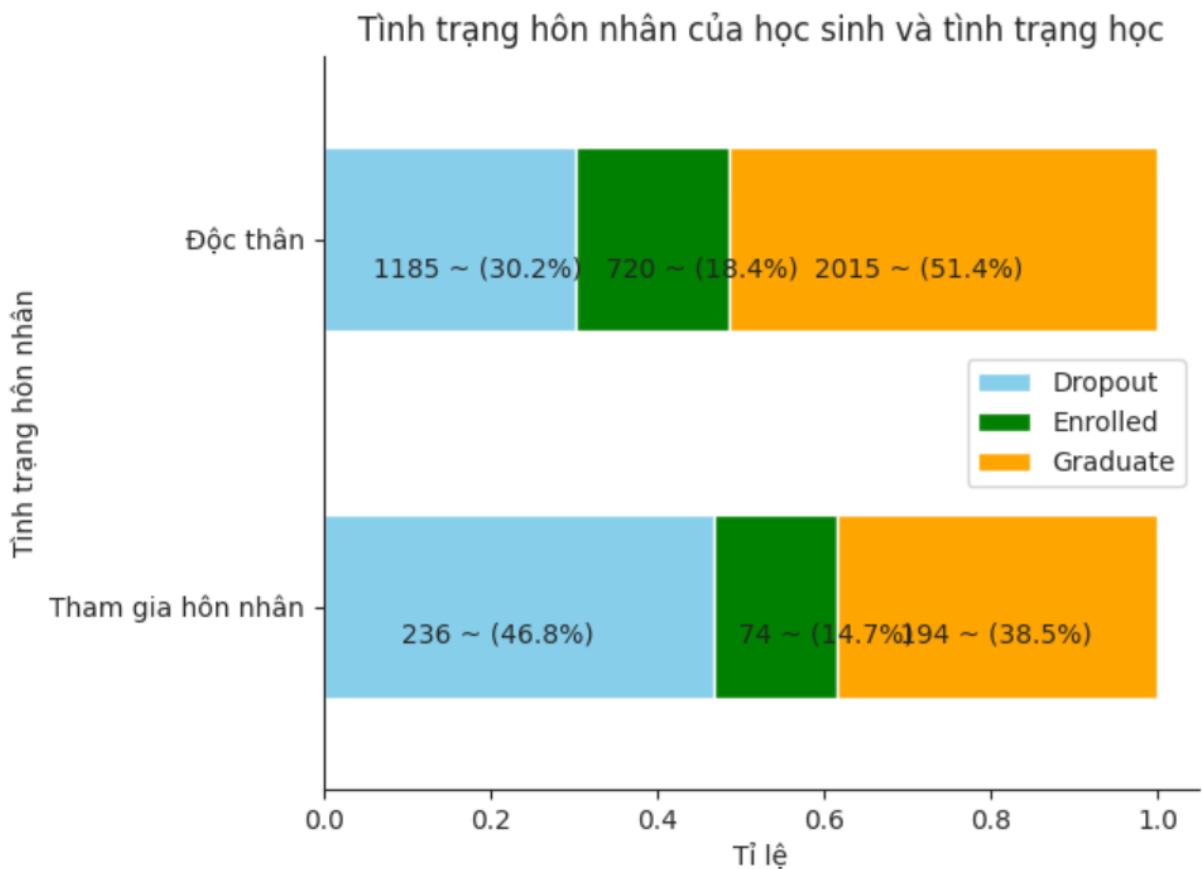
     alpha = 0.05
     if p_value < alpha:
         print("Có đủ bằng chứng để bác bỏ H0.")
     else:
         print("Không đủ bằng chứng để bác bỏ H0.")

Chi-squared statistic: 56.580218798426046
P-value: 5.173227713685819e-13
Có đủ bằng chứng để bác bỏ H0.

```

Nhận xét: Dựa vào kết quả kiểm định Chi-square, ta thấy được giá trị Chi-squared statistic khá lớn đồng thời p-value nhỏ hơn ngưỡng alpha đặt ra (0.05). Kết luận từ những số liệu này là có đủ bằng chứng để bác bỏ giả thuyết không (H0). Do đó, chúng ta có cơ sở để kết luận rằng sinh viên khi tham gia vào hoạt động hòn nhân có làm ảnh hưởng tới quyết định nghỉ học.

Biểu đồ trực quan:



Hình 36: Biểu đồ tình trạng hôn nhân của sinh viên và tình trạng học

- **Nhận xét biểu đồ:** Biểu đồ Bar plot cung cấp một cái nhìn trực quan hơn về mối liên hệ giữa tình trạng hôn nhân và quyết định nghỉ học của sinh viên. Với các sinh viên còn độc thân, tỉ lệ nghỉ học chỉ chiếm khoảng 30% trên tổng số sinh viên, còn đối với nhóm sinh viên đã tham gia vào hôn nhân thì số lượng sinh viên nghỉ học chiếm đến khoảng 47%, gần một nửa số lượng sinh viên ở tình trạng hôn nhân này đã nghỉ học. Kết quả này hoàn toàn hợp lý và càng khẳng định hơn về tính chính xác của kiểm định ở trên.

5.1.2 Kiểm định giả thuyết 2: Không có sự khác biệt giữa tỷ lệ nghỉ học giữa sinh viên quốc tế và sinh viên trong nước

Các bước thực hiện:

Bước 1: Xác định giả thuyết kiểm định:

Ho: Không có sự khác biệt giữa tỷ lệ nghỉ học giữa sinh viên quốc tế và sinh viên trong nước

H1: Có sự khác biệt giữa tỷ lệ nghỉ học giữa sinh viên quốc tế và sinh viên trong nước

Bước 2: Chia các giá trị trong biến Target thành 2 giá trị là Dropout và Not Dropout (bao gồm Enrolled và Graduate)

Bước 3: Tạo bảng tần số (Crosstab) giữa biến International và Target

Bước 4: Tiến hành kiểm định Chi-square và kết luận theo phương pháp p-value (trị số p).

Với alpha = 0.05 và độ tin cậy = 95%. Điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

Code mẫu:

- Xem qua crosstab giữa 2 biến:

```
[100] df['International'].value_counts()

0    4314
1     110
Name: International, dtype: int64

[101] df['Target'] = df['Target'].replace({
    'Dropout': 'Dropout',
    'Enrolled': 'Not Dropout',
    'Graduate': 'Not Dropout'
})

[102] cols = ['International']
      for col in cols:
          df = df.replace({col: {0: 'No', 1: 'Yes'}})

[103] inter=df[['Target','International']]
      crosstab = pd.crosstab(inter["Target"], inter["International"])
      crosstab
```

International	No	Yes
Target		
Dropout	1389	32
Not Dropout	2925	78

- Ở đây chúng ta sẽ dùng kiểm định Chi-square:

```
[104] stat, p, dof, expected = chi2_contingency(crosstab)
print('Degrees of freedom = %d' % dof)

Degrees of freedom = 1

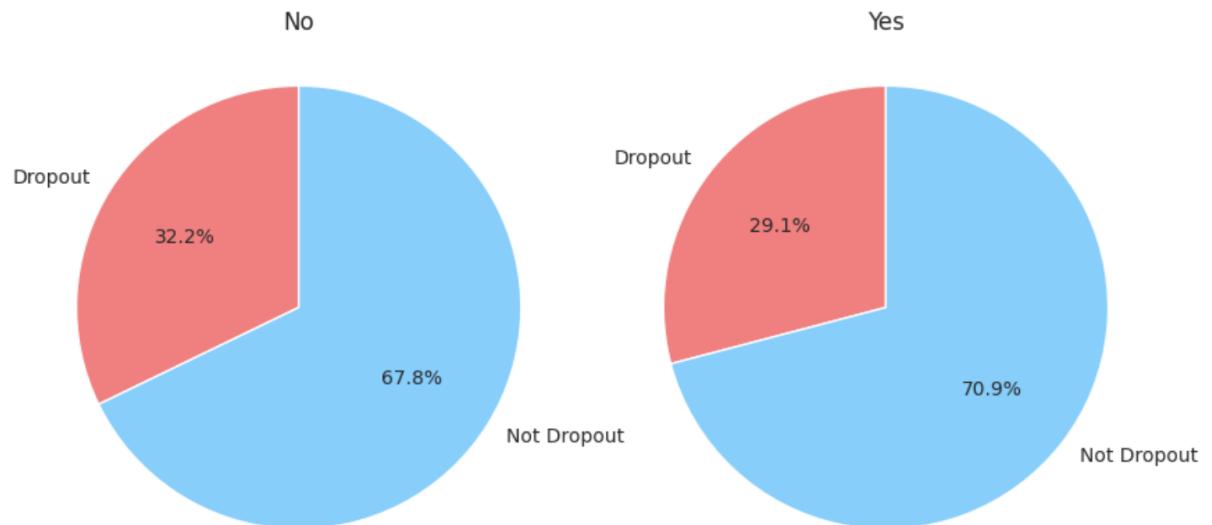
[105] contingency_table = pd.crosstab(df['International'], df['Target'])

print(f"Chi-squared statistic: {chi2_stat}")
print(f"P-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0.")
else:
    print("Không đủ bằng chứng để bác bỏ H0.")

Chi-squared statistic: 0.3430007141187338
P-value: 0.5581022527918915
Không đủ bằng chứng để bác bỏ H0.
```

- **Nhận xét:** Dựa vào kết quả kiểm định Chi-square, ta thấy được giá trị Chi-squared statistic khá nhỏ đồng thời p-value lớn hơn ngưỡng alpha đặt ra (0.05). Kết luận từ những số liệu này là không đủ bằng chứng để bác bỏ giả thuyết không (H0) về nhận định không có sự khác biệt giữa tỷ lệ nghỉ học giữa sinh viên quốc tế và sinh viên trong nước



Hình 37: Biểu đồ tỉ lệ nghỉ học giữa sinh viên trong nước và du học sinh

- **Nhận xét biểu đồ:** Dựa vào quan sát biểu đồ, chúng ta có thể nhận thấy rằng tỷ lệ nghỉ học giữa hai nhóm sinh viên, bao gồm sinh viên quốc tế và sinh viên trong nước, có vẻ gần nhau nhau, càng khẳng định cho kết quả chấp nhận giả thuyết Ho về nhận định không có sự khác biệt giữa tỷ lệ nghỉ học giữa sinh viên quốc tế và sinh viên trong nước là hợp lý.

5.2 Tình hình kinh tế - xã hội

Giả thuyết : Mức độ tăng giảm của tỉ lệ lạm phát của nền kinh tế không ảnh hưởng tới quyết định nghỉ học của sinh.

Các bước thực hiện:

Bước 1: Xác định giả thuyết kiểm định:

Ho: Mức độ tăng giảm tỉ lệ lạm phát của nền kinh tế không ảnh hưởng tới quyết định nghỉ học của sinh viên.

H1: Mức độ tăng giảm tỉ lệ lạm phát của nền kinh tế có ảnh hưởng tới quyết định nghỉ học của sinh viên.

Bước 2: Chia các giá trị trong biến Target thành 2 giá trị là Bỏ học là Dropout và Có học bao gồm Enrolled và Graduate.

Bước 3: Ở đây với số mẫu lớn, chúng ta sẽ thực hiện kiểm định Z-test và và kết luận theo phương pháp p-value (trị số p).

Với alpha = 0.05 và độ tin cậy = 95%. Điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

Code mẫu:

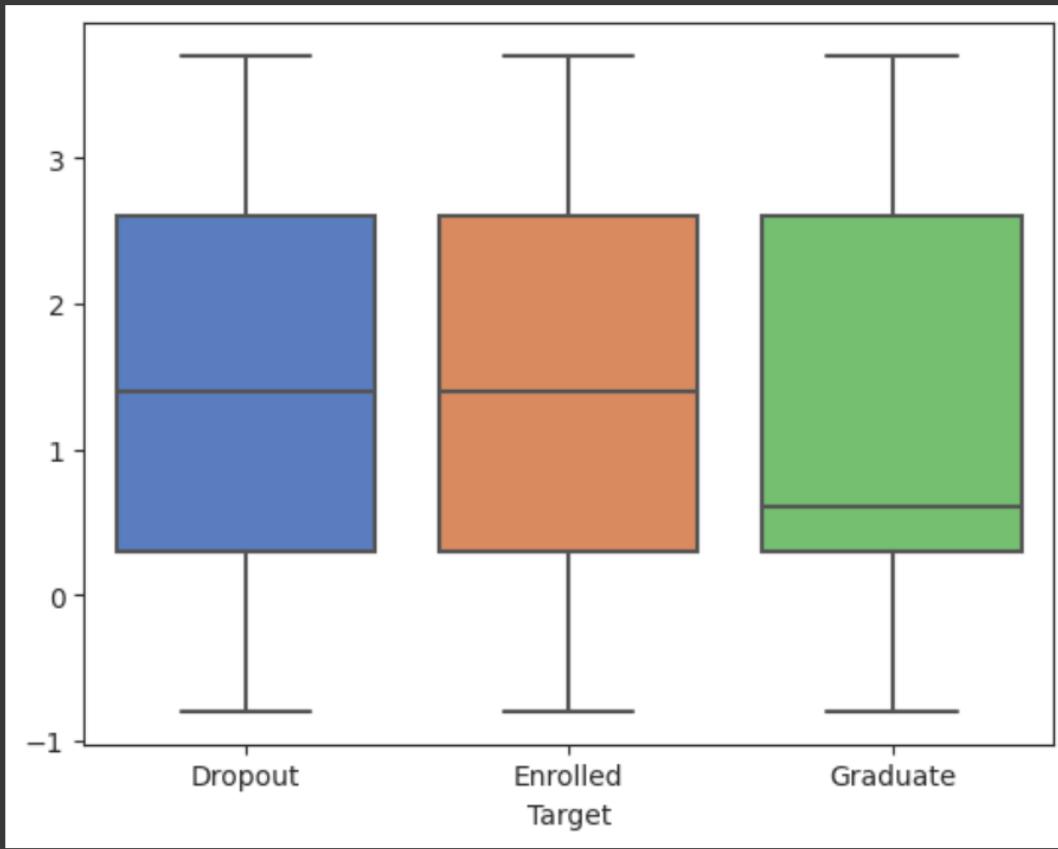
- Ở đây chúng ta sẽ dùng kiểm định T-test:

```
[ ] df2 = df[['target', 'inflation_rate']]  
# Chia dữ liệu thành 2 nhóm: Bỏ học (Dropout) và Có học (Graduate, Enrolled)  
dropout_group = df2[df2["target"] == 'Dropout'][["inflation_rate"].dropna()  
enrolled_graduate_group = df2[df2["target"].isin(["Enrolled", "Graduate"])]["inflation_rate"].dropna()  
# Đối với z-test, cần biết độ lệch chuẩn của cả hai nhóm  
std_dropout = dropout_group.std()  
std_enrolled_graduate = enrolled_graduate_group.std()  
  
# Thực hiện kiểm định z-test  
z_statistic, p_value_z = ztest(dropout_group, enrolled_graduate_group, value=0, alternative='two-sided')  
  
print("Z-statistic:", z_statistic)  
print("Giá trị p (z-test):", p_value_z)  
  
alpha = 0.05  
if p_value_z < alpha:  
    print("Có đủ bằng chứng để bác bỏ H0.")  
else:  
    print("Không đủ bằng chứng để bác bỏ H0.")
```

Z-statistic: 1.611581534998867
Giá trị p (z-test): 0.10705303147005597
Không đủ bằng chứng để bác bỏ H0.

- **Nhận xét:** Dựa vào kết quả kiểm định T-test, ta thấy được giá trị p-value lớn hơn ngưỡng alpha đặt ra (0.05). Kết luận từ những số liệu này là không đủ bằng chứng để bác bỏ giả thuyết không (H0) về nhận định rằng mức độ tăng giảm GDP của nền kinh tế không ảnh hưởng tới quyết định nghỉ học của sinh.

```
[211] df1 = df.pivot(columns = "Target", values = "Inflation_rate")
      sns.boxplot(data = df1)
      plt.show()
```



Hình 38: Box plot biểu diễn phân phối của các giá trị trong biến "Target" theo tỉ lệ lạm phát

- **Nhận xét biểu đồ:** Quan sát biểu đồ boxplot ta thấy được sự chênh lệch không đáng kể giữa phân phối của các giá trị trong biến Target theo tỷ lệ lạm phát, càng khẳng định cho kết quả rằng không đủ bằng chứng để bác bỏ Ho (Mức độ tăng giảm GDP của nền kinh tế không ảnh hưởng tới quyết định nghỉ học của sinh).

5.3 Tình hình học tập:

Giả thuyết: Không có sự khác biệt của số tín chỉ đăng ký theo nhóm ngành học.

Bước 1: Xác định giả thuyết kiểm định:

Ho: Không có sự khác biệt của số tín chỉ đăng ký theo nhóm ngành học.

H1: Có sự phụ thuộc của số tín chỉ đăng ký theo nhóm ngành học.

Bước 2: Chia 17 ngành học có trong biến Course thành 5 nhóm ngành chính.

Bước 3: Tạo bảng tần số (Crosstab) giữa biến International và Target

Bước 4: Thực hiện kiểm định Anova và và kết luận theo phương pháp p-value (trị số p).

Với alpha = 0.05 và độ tin cậy = 95%. Điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

Bước 5: Vì có đủ bằng chứng để bác bỏ giả thuyết Ho nên tiếp tục thực hiện Hậu kiểm Tukey HSD để kiểm tra sự khác biệt giữa các nhóm.

Code mẫu:

```
] field_mapping = {'Công nghệ và Thiết kế': [1, 2, 5, 7],  
                  'Quản trị và Dịch vụ Xã hội': [3, 9, 10, 17],  
                  'Nông nghiệp và Thú y': [4, 6, 8],  
                  'Truyền thông và Giáo dục': [11, 14, 15, 16],  
                  'Y tế và Chăm sóc sức khỏe': [12,13]}  
  
def map_to_field(index):  
    for field, indices in field_mapping.items():  
        if index in indices:  
            return field  
    return 'Other'  
  
df['field'] = df['course'].apply(map_to_field)
```

```
df3=df[['curricular_units_credited','field']]  
crossstab_units_field = pd.crosstab(course["curricular_units_credited"], df3["field"])  
crossstab_units_field.head()  
  
field Công nghệ và Thiết kế Nông nghiệp và Thú y Quản trị và Dịch vụ Xã hội Truyền thông và Giáo dục Y tế và Chăm sóc sức khỏe  
curricular_units_credited  
0 504 522 964 980 802  
1 15 18 14 4 7  
2 6 32 13 3 8  
3 4 25 5 3 5  
4 7 4 10 7 1
```

- Kiểm định Anova

```
field_data = {}  
for field in df['field'].unique():  
    field_data[field] = df[df['field'] == field]['curricular_units_enrolled']  
anova_result = f_oneway(*field_data.values())  
print("ANOVA Result:")  
print(anova_result)  
alpha = 0.05  
if anova_result.pvalue < alpha:  
    print("Có đủ bằng chứng để bác bỏ giả thuyết H0.")  
else:  
    print("Không đủ bằng chứng để bác bỏ giả thuyết H0.")  
  
ANOVA Result:  
F_onewayResult(statistic=174.3747718740743, pvalue=1.0061466881517475e-138)  
Có đủ bằng chứng để bác bỏ giả thuyết H0.
```

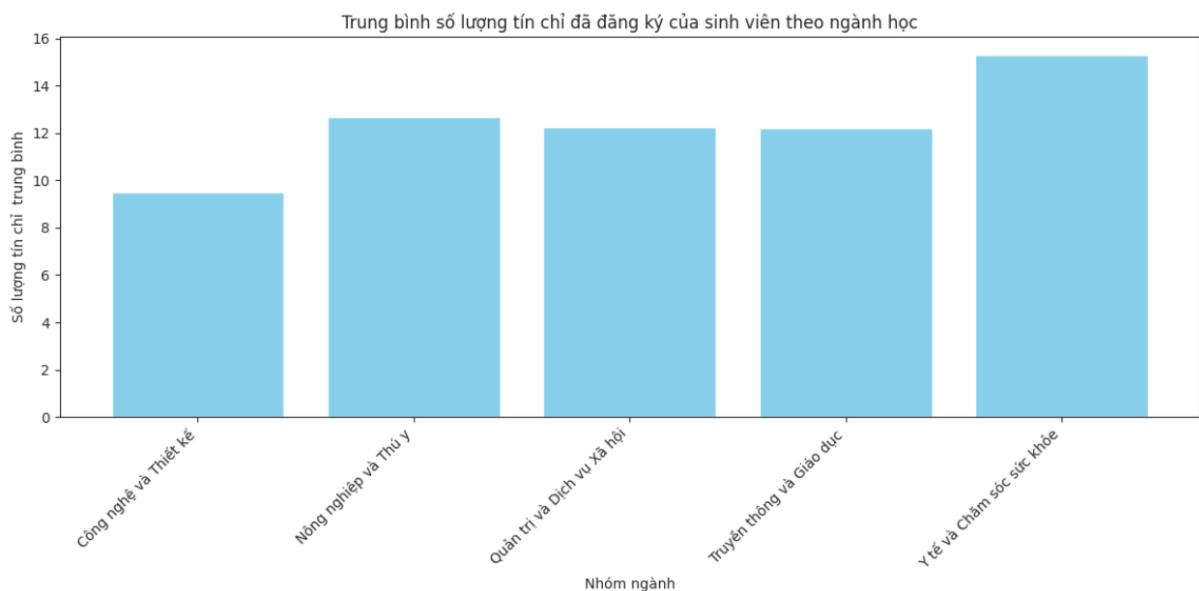
- Hậu kiểm Tukey HSD

# Thực hiện kiểm định Tukey HSD						
tukey_results = pairwise_tukeyhsd(df['curricular_units_enrolled'], df['field'])						
print("\nTukey's HSD Results:")						
print(tukey_results)						
Tukey's HSD Results:						
Multiple Comparison of Means - Tukey HSD, FWER=0.05						
=====						
group1	group2	meandiff	p-adj	lower	upper	reject
Công nghệ và Thiết kế	Nông nghiệp và Thú y	3.1749	0.0	2.5309	3.8189	True
Công nghệ và Thiết kế	Quản trị và Dịch vụ Xã hội	2.6853	0.0	2.11	3.2606	True
Công nghệ và Thiết kế	Truyền thông và Giáo dục	2.7306	0.0	2.1409	3.3204	True
Công nghệ và Thiết kế	Y tế và Chăm sóc sức khỏe	5.8324	0.0	5.2196	6.4452	True
Nông nghiệp và Thú y	Quản trị và Dịch vụ Xã hội	-0.4896	0.1171	-1.0477	0.0685	False
Nông nghiệp và Thú y	Truyền thông và Giáo dục	-0.4443	0.2134	-1.0173	0.1287	False
Nông nghiệp và Thú y	Y tế và Chăm sóc sức khỏe	2.6575	0.0	2.0608	3.2542	True
Quản trị và Dịch vụ Xã hội	Truyền thông và Giáo dục	0.0453	0.9991	-0.4492	0.5399	False
Quản trị và Dịch vụ Xã hội	Y tế và Chăm sóc sức khỏe	3.1471	0.0	2.6252	3.669	True
Truyền thông và Giáo dục	Y tế và Chăm sóc sức khỏe	3.1018	0.0	2.564	3.6395	True

- Nhận xét:** Dựa vào giá trị p nhỏ hơn mức ý nghĩa (alpha) 0.05, chúng ta có đủ chứng cứ để phủ nhận giả thuyết không có sự khác biệt về giá trị trung bình của số lượng tín chỉ đăng ký giữa các khóa học. Vì vậy, chúng ta có thể suy luận rằng có sự khác biệt đáng kể về mặt thống kê giữa số tín chỉ đăng ký, ít nhất ở một cặp khóa học nào đó.

Sau khi tiến hành Hậu kiểm Tukey HSD, chúng ta rút ra được nhận định rằng các cặp giá trị có reject là “True” có giá trị khác biệt đáng kể về mặt thống kê, còn 3 cặp có kết quả “False” là không có sự khác biệt đáng kể. Để kiểm chứng lại kết quả này, chúng ta tiến hành trực quan hóa bằng biểu đồ.

- Biểu đồ minh họa:**



Hình 39: Biểu đồ biểu diễn số lượng tín chỉ trung bình của sinh viên theo các nhóm ngành

- Nhận xét biểu đồ:** Thông qua biểu đồ ta thấy được 3 nhóm ngành bao gồm “Nông nghiệp và Thú ý”, “Quan trị và Dịch vụ xã hội”, “Truyền thông và Giáo dục” có số lượng

tín chỉ trung bình xấp xỉ nhau, đối chiếu với kết quả trong hậu kiểm Turkey HSD có reject mang giá trị False có thể kết luận được rằng kiểm định là hợp lý.

CHƯƠNG VI: XÂY DỰNG MÔ HÌNH DỰ BÁO - PCA

- Chúng ta sẽ sử dụng 4 loại mô hình phân lớp phổ biến đã được học trong khuôn khổ chương trình là : k-NN, Decision Trees, Random Forest, Naive Bayes Classification
- **Các bước xây dựng mô hình:**
 - Bước 1:** Chia tập Train-Test và thử nghiệm các mô hình
 - Bước 2:** Tối ưu hóa hiệu suất mô hình bằng việc tìm kiếm các siêu tham số (hyperparameter tuning) thông qua `GridSearchCV` của thư viện sklearn.model_selection
 - Bước 3:** In kết quả ra và so sánh hiệu suất của mỗi mô hình
 - Bước 4:** Giảm chiều dữ liệu bằng PCA
 - Bước 5:** Tính lại điểm số các mô hình sau khi giảm chiều dữ liệu.

6.1 Xây dựng mô hình:

- Ở đây chúng ta sẽ chỉ dự đoán những học sinh nào có nghỉ học hay không theo tập dữ liệu đã thu thập, nhằm giúp nhà trường và các đơn vị giáo dục dự báo được rằng tệp những sinh viên có thể nghỉ học trong tương lai và tìm hướng để xử lý.

```
[313] df['target'] = df['target'].replace({'Dropout':0, 'Enrolled':1, 'Graduate':1})
      df['target']

      0      0
      1      1
      2      0
      3      1
      4      1
      ..
  4419      1
  4420      0
  4421      0
  4422      1
  4423      1
Name: target, Length: 4424, dtype: int64
```

- Chia tập Train_Test bằng KFold và cross-validation:

```
[21] target = df['target']
      features = df.drop(['target'], axis=1)
      target.shape, features.shape

      ((4424,), (4424, 25))

[22] from sklearn.model_selection import train_test_split, KFold

      kf = KFold(n_splits = 5, shuffle = True, random_state = 2304)
      for tr_idx, te_idx in kf.split(features):
          X_train, X_test = features.iloc[tr_idx], features.iloc[te_idx]
          y_train, y_test = target.iloc[tr_idx], target.iloc[te_idx]

          X_train.shape, X_test.shape, y_train.shape, y_test.shape

      ((3540, 25), (884, 25), (3540,), (884,))
```

6.1.2 Chạy thử với mô hình

- Cài đặt các mô hình vào colab:

```
[84] from sklearn.ensemble import RandomForestClassifier
     from sklearn.svm import SVC, LinearSVC
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.naive_bayes import GaussianNB
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.ensemble import AdaBoostClassifier, GradientBoostingClassifier
Classifiers=[
    ["Random_Forest",RandomForestClassifier()],
    ["KNN",KNeighborsClassifier()],
    ["Naive_Bayes",GaussianNB()],
    ["Decision_Tree",DecisionTreeClassifier()]]
```



```
[85] from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score
name_list = []
recall_list = []
precision_list = []
f1_list = []
acc_list = []
for name, classifier in Classifiers:
    classifier = classifier
    classifier.fit(X_train, y_train)
    pred = classifier.predict(X_test)
    globals()['pred_result_' + name] = pd.DataFrame({'true_label': y_test.squeeze(), "predict": pred})

    recall = recall_score(y_test, pred, average='macro')
    precision = precision_score(y_test, pred, average='macro')
    f1 = f1_score(y_test, pred, average='macro')
    accuracy = accuracy_score(y_test, pred)
    name_list.append(name)
    recall_list.append(recall)
    precision_list.append(precision)
    f1_list.append(f1)
    acc_list.append(accuracy)
```

- In kết quả chạy mô hình:

```
[86] performance = pd.DataFrame({"name":name_list, "recall":recall_list, "precision":precision_list,
                                 "f1_score":f1_list, "accuracy":acc_list })
performance
```

	name	recall	precision	f1_score	accuracy
0	Random_Forest	0.810641	0.859038	0.828638	0.861991
1	KNN	0.727331	0.775000	0.742624	0.796380
2	Naive_Bayes	0.775470	0.765593	0.769991	0.798643
3	Decision_Tree	0.756364	0.752465	0.754328	0.787330

- **Nhận xét:** Dựa vào bảng kết quả trên, ta có thể đưa ra các nhận định rằng:
 - + **Random_Forest:** Với các chỉ số lần lượt là recall: 0.8106, precision: 0.8590, f1_score: 0.8286, accuracy: 0.8619, có thể thấy được rằng mô hình Random_Forest hoạt động khá

tốt, nó có thể dự đoán đúng chính xác đến 86.19% các trường hợp có trong bộ dữ liệu. Sự kết hợp tốt giữa precision và recall, được thể hiện qua f1_score, chỉ ra mô hình đạt được sự cân bằng giữa việc tránh dự đoán sai positive (precision cao) và bỏ sót ít positive thực tế (recall cao).

- + **KNN:** Mô hình có các chỉ số recall: 0.7273, precision: 0.7750, f1_score: 0.7426, accuracy: 0.7963. So với Random_Forest, KNN có các chỉ số thấp hơn, nhưng vẫn giữ được độ chính xác trên 80%. Mặc dù precision và recall không đạt mức cao như Random_Forest, nhưng mô hình vẫn đảm bảo một khả năng dự đoán chính xác và cân bằng tương đối giữa việc tránh dự đoán sai positive và bỏ sót ít positive thực tế.
- + **Naive_Bayes:** Tương tự ở trên, mô Naive_Bayes có các chỉ số lần lượt là recall: 0.7754, precision: 0.7655, f1_score: 0.7699, accuracy: 0.7986. Đây là một mô hình có hiệu suất trung bình, nhưng ổn định. Mô hình có thể dự đoán đúng khoảng 79.86% trên tổng số các dự đoán. Mặc dù không đạt được chỉ số cao như Random Forest hay KNN, nhưng Naive_Bayes là một lựa chọn hợp lý nếu cần yêu cầu về độ chính xác toàn cục và một hiệu suất ổn định trên cả precision và recall.
- + **Decision_Tree:** Mô hình có các chỉ số recall: 0.7563, precision: 0.7524, f1_score: 0.7543, accuracy: 0.7873, có thể thấy trong 4 mô hình được ứng dụng, mô hình Decision_Tree có các chỉ số thấp hơn hẳn, mô hình cho mức độ dự đoán chính xác xấp xỉ 78.73%. Tổng thể, Decision_Tree không đạt được hiệu quả cao như các mô hình khác trong bài toán này.
- **Kết luận:** Thông qua các nhận xét, chúng ta có thể đúc kết lại được những kết luận quan trọng rằng mô hình Random_Forest có hiệu suất cao nhất trong bốn mô hình được đánh giá, với các chỉ số recall, precision, f1_score, accuracy đều ở mức cao. Mô hình này có khả năng dự đoán chính xác cao và đồng thời duy trì được sự cân bằng giữa việc tránh dự đoán sai positive và bỏ sót positive thực tế. Tiếp theo đó là mô hình KNN cũng có hiệu suất khá ổn với độ chính khá cao.

6.1.3 Sử dụng thư viện `GridSearchCV` để tối ưu hóa mô hình

▼ k-NN

```
[23] from sklearn.model_selection import GridSearchCV
      from sklearn.neighbors import KNeighborsClassifier
      knn = KNeighborsClassifier()
      knn_param_grid = {
          'n_neighbors': [1, 3, 5, 7],
          'weights': ['uniform', 'distance'],
          'p': [1, 2] }

      knn_grid_search = GridSearchCV(knn, knn_param_grid, cv=10)
      knn_grid_search.fit(X_train, y_train)
      print("K-NN Best Parameters:", knn_grid_search.best_params_)

K-NN Best Parameters: {'n_neighbors': 7, 'p': 2, 'weights': 'distance'}
```

- **Mục đích:**
- + `n_neighbors`: Số lượng hàng xóm gần nhất mà mô hình sẽ xem xét khi đưa ra dự đoán.

- + `weights`: Cách mà mô hình sẽ đánh giá trọng số cho các điểm dữ liệu láng giềng, 'uniform' có nghĩa là tất cả các điểm láng giềng có cùng trọng số, 'distance' có nghĩa là điểm láng giềng gần hơn có trọng số lớn hơn.
- + `p`: Tham số p trong công thức khoảng cách (mặc định là 2 cho khoảng cách Euclidean). `p=1` tương ứng với khoảng cách Manhattan.
- + Grid Search: Sử dụng kỹ thuật cross-validation để đánh giá hiệu suất của mô hình với từng bộ giá trị siêu tham số. Nó tạo ra tất cả các kết hợp có thể của các giá trị trong lưới và đánh giá chúng. Ở đây chúng ta sử dụng `cv=10` chỉ định rằng bạn đang sử dụng 10 folds trong cross-validation để đánh giá mô hình.

▼ Decision Tree

```
4s [26] from sklearn.model_selection import GridSearchCV
      from sklearn.tree import DecisionTreeClassifier
      dtree = DecisionTreeClassifier(class_weight='balanced')
      param_grid = [
          'max_depth': [3, 4, 5],
          'min_samples_split': [2, 3],
          'min_samples_leaf': [1, 2],
          'random_state': [0, 30]
      ]
      grid_search = GridSearchCV(dtree, param_grid, cv=10)
      grid_search.fit(X_train, y_train)
      print(grid_search.best_params_)

      {'max_depth': 4, 'min_samples_leaf': 2, 'min_samples_split': 2, 'random_state': 0}
```

- Giải thích:

- + `max_depth` (độ sâu tối đa của cây)
- + `min_samples_split` (số mẫu tối thiểu để phân chia một nút)
- + `min_samples_leaf` (số mẫu tối thiểu trên mỗi lá)
- + Grid Search: Tương tự như KNN

▼ Random Forest

```
41s [28] from sklearn.ensemble import RandomForestClassifier
      from sklearn.model_selection import GridSearchCV
      rfc = RandomForestClassifier(class_weight='balanced')
      param_grid = [
          'n_estimators': [10, 20, 50],
          'max_depth': [None, 2, 4],
          'random_state': [0, 20]}
      grid_search = GridSearchCV(rfc, param_grid, cv=10)
      grid_search.fit(X_train, y_train)
      print(grid_search.best_params_)

      {'max_depth': None, 'n_estimators': 50, 'random_state': 0}
```

- Giải thích:

- + `n_estimators`: Số cây trong rừng.

- + `max_depth`: Độ sâu tối đa của mỗi cây..
- + Grid Search: Tương tự như KNN

6.2 Chạy các thuật toán dự báo

- Cài đặt các mô hình với các tham số ở bước trước:

```
[35] from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC, LinearSVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
Classifiers:[
    ["Random_Forest",RandomForestClassifier(random_state=0, n_estimators=50, max_depth=None, class_weight='balanced')],
    ["KNN",KNeighborsClassifier(n_neighbors= 7, p= 2, weights='distance')],
    ["Naive_Bayes",GaussianNB()],
    ["Decision_Tree",DecisionTreeClassifier(random_state=0, max_depth=4, min_samples_leaf=2, min_samples_split=2,class_weight='balanced')]]
```



```
[366] from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score
name_list = []
recall_list = []
precision_list = []
f1_list = []
acc_list = []
for name, classifier in Classifiers:
    classifier = classifier
    classifier.fit(X_train, y_train)
    pred = classifier.predict(X_test)
    globals()[f'pred_result_{name}'] = pd.DataFrame({"true_label": y_test.squeeze(), "predict": pred})

    recall = recall_score(y_test, pred, average='macro')
    precision = precision_score(y_test, pred, average='macro')
    f1 = f1_score(y_test, pred, average='macro')
    accuracy = accuracy_score(y_test, pred)
    name_list.append(name)
    recall_list.append(recall)
    precision_list.append(precision)
    f1_list.append(f1)
    acc_list.append(accuracy)
```

- In kết quả:

```
[86] performance = pd.DataFrame({"name":name_list, "recall":recall_list, "precision":precision_list,
                                "f1_score":f1_list, "accuracy":acc_list })
performance
```

	name	recall	precision	f1_score	accuracy
0	Random_Forest	0.816409	0.856017	0.831896	0.863122
1	KNN	0.736044	0.803900	0.755670	0.811086
2	Naive_Bayes	0.775470	0.765593	0.769991	0.798643
3	Decision_Tree	0.798865	0.777138	0.785210	0.807692

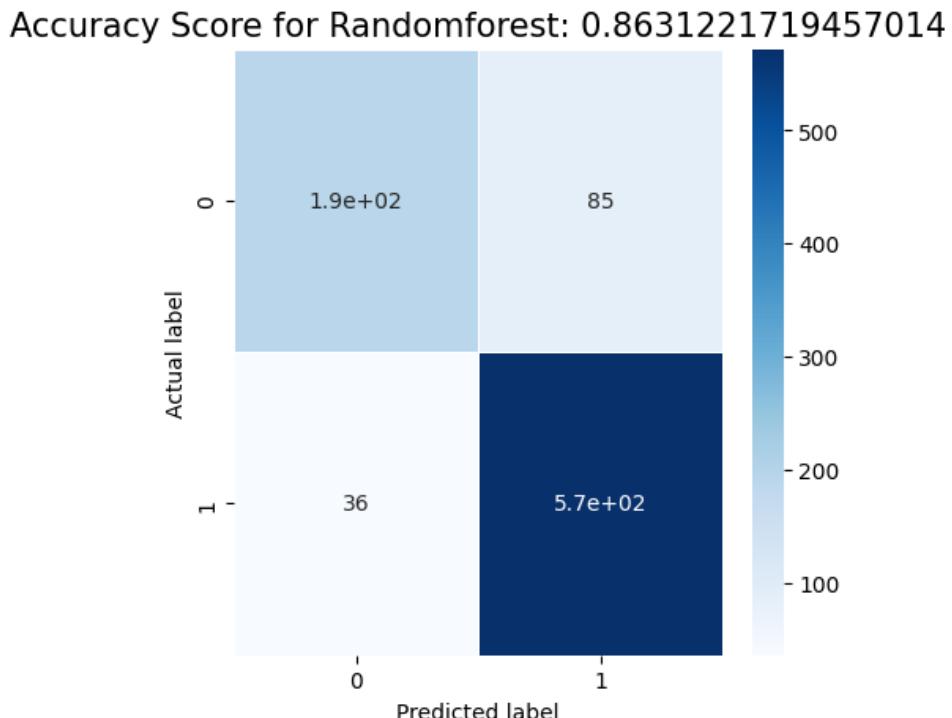
- So sánh với trước khi chưa thêm tham số:

	name	recall	precision	f1_score	accuracy
0	Random_Forest	0.810641	0.859038	0.828638	0.861991
1	KNN	0.727331	0.775000	0.742624	0.796380
2	Naive_Bayes	0.775470	0.765593	0.769991	0.798643
3	Decision_Tree	0.756364	0.752465	0.754328	0.787330

→ Thông qua việc đánh giá các chỉ số, có thể đi đến kết luận rằng sau khi thêm vào các siêu tham số, các mô hình đã cho kết quả tích cực, đa số các chỉ trong đánh giá trong các mô hình đều có chiều hướng tăng. Nếu chúng ta ưu tiên mức độ chính xác cao và khả năng nhận diện positive instances thì Random_Forest và Decision_Tree là những lựa chọn hợp lý, còn nếu ưu tiên độ ổn định có thể cân nhắc việc chọn Naive_Bayes. KNN cần phải được điều chỉnh để cải thiện được sự cân bằng giữa precision, recall và nâng cao độ chính xác toàn cục.

6.2.1 Đánh giá với mục tiêu dự báo

Với việc có hiệu suất và dự báo điểm số cao nhất cả trước và sau khi hiệu chỉnh tham số, Random Forest là một thuật toán có vẻ sẽ phù hợp cho mô hình dữ liệu này. Hãy cùng phân tích qua kết quả của nó trong dự báo học sinh nghỉ học.



Hình 40: Ma trận nhầm lẫn của Random Forest

	Dự đoán nghỉ học	Dự đoán không nghỉ học

Thực tế nghỉ học	190 (TP)	85 (FP)
Thực tế không nghỉ học	36 (FN)	570 (TN)

Bảng 2: Ma trận nhầm lẫn của Random Forest

- Trong ma trận nhầm lẫn này, trong số **275 mẫu thực tế nghỉ học**, hệ thống đã đánh giá rằng có **190/275 mẫu khớp với thực tế** (85 mẫu dự báo sai thực tế) đạt 69% và trong số **606 mẫu thực tế không nghỉ học** thì có tới 570/606 **mẫu khớp với thực tế đưa ra** (36 mẫu dự báo sai thực tế) đạt 94%.
- Điều này cho thấy rằng việc dự báo sinh viên có nghỉ học hay không thực sự là một vấn đề khó và nan giải cho mỗi trường học để có kế hoạch xử lý vấn đề này. Khi tỉ lệ dự báo một sinh viên học tiếp lên tới hơn 90% trong khi nếu có 10 học sinh nghỉ học thì mô hình chỉ dự báo được khoảng 7/10 người.

6.3 Giảm chiều dữ liệu

Thông qua quan sát, với bộ dữ liệu đầu vào tới 26 cột tuy nhiên tỉ lệ dữ báo vẫn chỉ đạt hiệu suất chưa tới 70%. Từ đó có thể thấy được rằng việc dự báo sinh viên nghỉ học là rất khó so với nhóm còn lại. Tuy nhiên, để kiểm tra rằng việc giảm số lượng biến đầu vào thêm nữa ảnh hưởng nhiều như thế nào tới hiệu suất. Nhóm quyết định áp dụng PCA để tìm được câu trả lời:

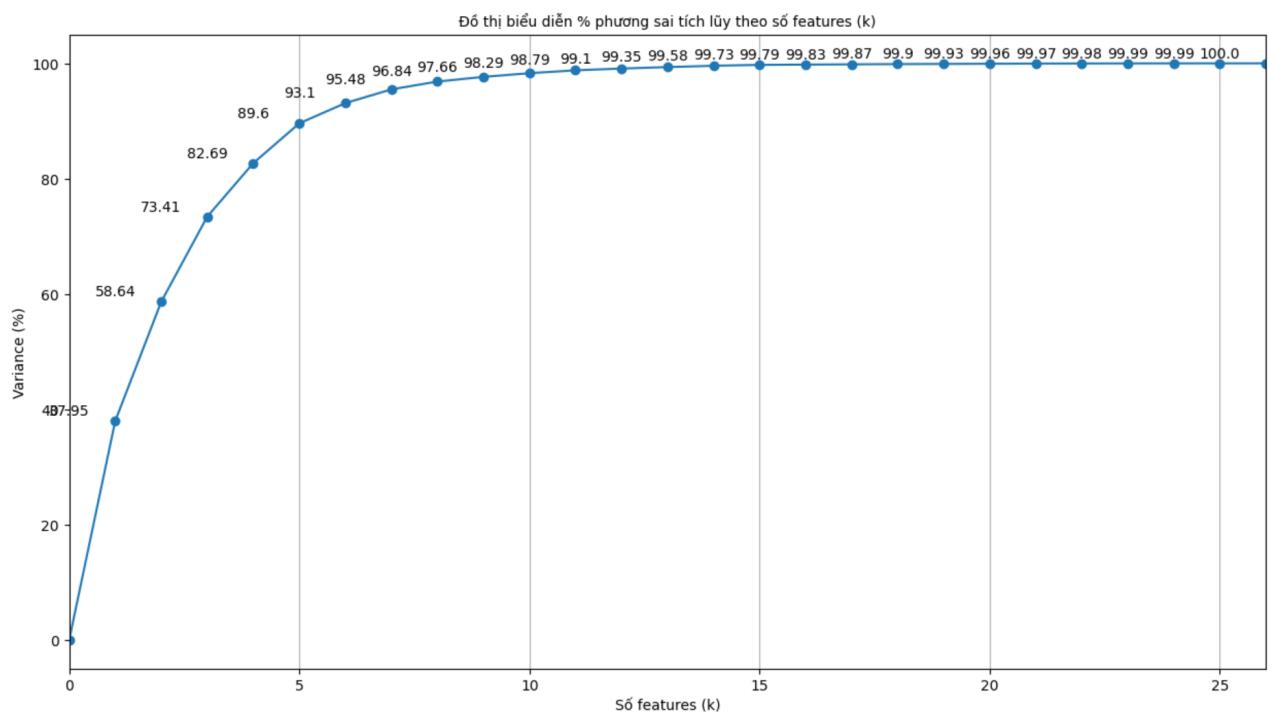
a. Cài đặt PCA và loại biến Target ra khỏi bộ dữ liệu:

```
[368] from sklearn.decomposition import PCA
      pca = PCA().fit(df)

[369] target = 'target'
      print('* Biến phân lớp:', target)

      ## Danh sách các features
      nb_features = df.shape[1] - 1
      features    = df.columns[:nb_features]
      print('* Số lượng features = %2d' %nb_features)
      print('  Các features:', ', '.join(features))
```

b. Vẽ biểu đồ phân tích phương sai tích lũy theo features:



Hình 41: biểu đồ phân tích phương sai tích lũy theo features

Ở đây chúng ta thấy từ k=8 trở đi thì sự biến thiên của phương sai là không nhiều. Vậy chúng ta sẽ lấy k = 8.

6.3.1 Tính lại hiệu suất các mô hình sau khi giảm chiều

Áp dụng PCA vào tập huấn luyện.

```
[ ] from sklearn.decomposition import PCA
k = 8
pca = PCA(n_components=k)
X_train_pca = pca.fit_transform(X_train)
X_test_pca = pca.transform(X_test)

[ ] from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score

name_list = []
recall_list = []
precision_list = []
f1_list = []
acc_list = []

for name, classifier in Classifiers:
    classifier.fit(X_train_pca, y_train)
    pred = classifier.predict(X_test_pca)
    globals()['pred_result_{}'.format(name)] = pd.DataFrame({"true_label": y_test.squeeze(), "predict": pred})

    recall = recall_score(y_test, pred, average='macro')
    precision = precision_score(y_test, pred, average='macro')
    f1 = f1_score(y_test, pred, average='macro')
    accuracy = accuracy_score(y_test, pred)

    name_list.append(name)
    recall_list.append(recall)
    precision_list.append(precision)
    f1_list.append(f1)
    acc_list.append(accuracy)
```

- Xem hiệu suất:

	name	recall	precision	f1_score	accuracy
0	Random_Forest	0.749225	0.802525	0.766627	0.815611
1	KNN	0.746424	0.802281	0.764264	0.814480
2	Naive_Bayes	0.727176	0.763384	0.739824	0.790724
3	Decision_Tree	0.757020	0.755026	0.756001	0.789593

- **Nhận xét:** Sau quá trình giảm chiều dữ liệu từ 25 chiều xuống còn 8 chiều bằng phương pháp PCA, tiến hành áp dụng các thuật toán phân loại, bao gồm Random_Forest, KNN, Naive_Bayes, Decision_Tree, để xây dựng mô hình trên bộ dữ liệu đã giảm chiều. Mục tiêu để đánh giá hiệu suất của các mô hình này và xem xét liệu việc giảm chiều dữ liệu có mang lại lợi ích không thông qua việc đánh giá trên các chỉ số như recall, precision, f1_score và accuracy.
- + **Random_Forest:** Với các chỉ số recall: 0.7492, precision: 0.8025, f1_score: 0.7666, accuracy: 0.8156, ta thấy được rằng mặc dù chỉ số recall và precision giảm so với dữ liệu ban đầu, nhưng mô hình vẫn duy trì được hiệu suất tốt với các chỉ số f1_score và accuracy đều ở mức ổn định.
- + **KNN:** Tương tự ở trên, các chỉ số của KNN là recall: 0.7362, precision: 0.7985, f1_score: 0.7542, accuracy: 0.8057. KNN có sự giảm nhẹ ở các chỉ số so với trước khi giảm chiều dữ liệu, tuy nhiên vẫn duy trì được hiệu suất khá tốt, đặc biệt là độ chính xác vẫn đạt trên 80%.
- + **Naive_Bayes và Decision Tree:** Hầu như các chỉ số của 2 mô hình này đều ở dưới mức 80%. Mặc dù không có sự thuyên giảm đáng kể so với trước khi PCA nhưng hiệu suất so với 2 mô hình bên trên thì vẫn thấp hơn.

Kết luận: Dựa vào việc đánh giá các chỉ số trong các mô hình trước và sau khi giảm chiều dữ liệu, ta có thể đi đến các kết luận rằng:

- Tổng thể, hiệu suất của các mô hình học máy sau khi giảm chiều dữ liệu giảm nhẹ so với trước khi giảm chiều có thể vì mất đi một số lượng thông tin.
- Mức độ giảm hiệu suất ở các mô hình là khác nhau. Mô hình Random Forest là mô hình giảm hiệu suất nhiều nhất khoảng 5% về độ chính xác.

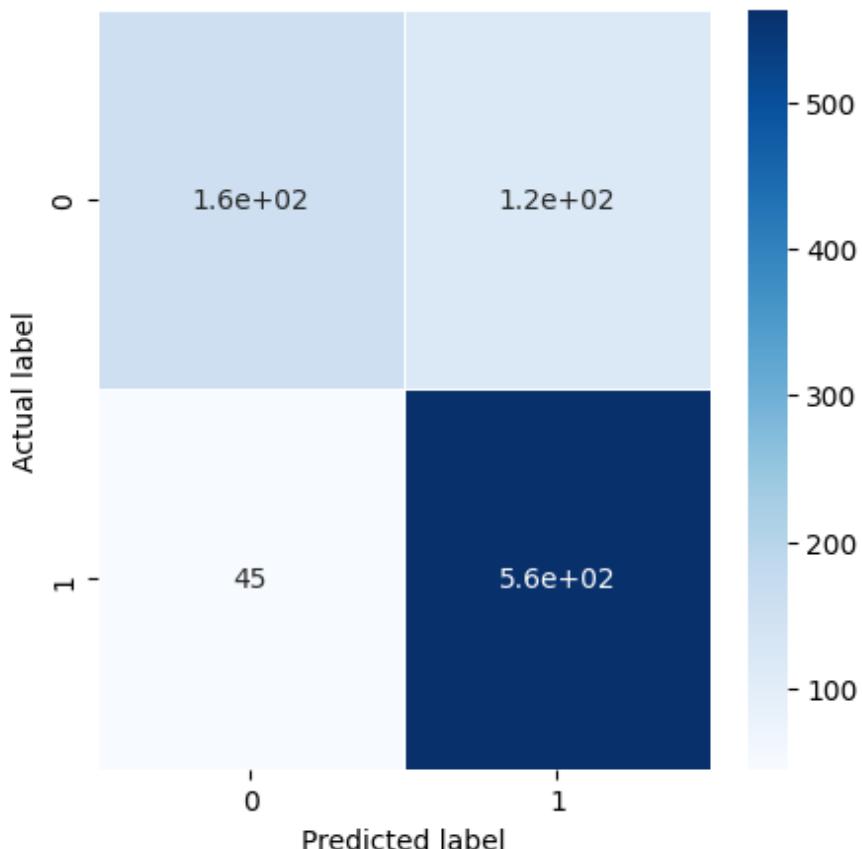
=> Từ kết quả trên, có thể nhận thấy rằng quá trình giảm chiều dữ liệu từ 25 chiều xuống còn 8 chiều dẫn đến việc mất đi một số lượng thông tin, đồng thời giảm đi hiệu suất của các mô hình, nhưng sự sụt giảm ở đây khá nhỏ, và ưu điểm của PCA đó là sử dụng tất cả các biến đầu vào nên phương pháp này không bỏ sót những biến quan trọng. Ngoài ra, nó còn xây dựng những trục tọa độ mới, thay vì giữ lại các trục của không gian cũ, nhưng lại có khả năng biểu diễn dữ liệu tốt tương đương, đảm bảo độ biến thiên của dữ liệu trên mỗi chiều mới, và vẫn giữ được khả năng dự đoán chính xác khá cao, trong trường hợp này $\frac{3}{4}$ thuật toán cho độ chính xác trên 80%. Quyết định giảm chiều dữ liệu từ 25 chiều xuống còn 8 chiều là một quyết định hợp lý và có tính khả thi đối với bộ dữ liệu này.

⇒ Tuy nhiên cần phải xem xét tiếp rằng mức dự báo sinh viên thực sự nghỉ học có bị ảnh hưởng nhiều không.

6.3.2 Áp dụng vào bài toán dự báo học sinh nghỉ học

Qua những lần chạy dự báo, chúng ta thấy được mô hình Random Forest luôn có sự ổn định cũng như hiệu suất đứng đầu trong 4 mô hình. Và việc dự báo về việc học sinh nghỉ học thì chúng ta sẽ thiên về điểm số accuracy cũng như recall. Vừa hay Random Forest đáp ứng được cả 2 tiêu chí khi cả 2 thông số này nó đều đứng đầu.

Accuracy Score for RandomForest: 0.8156108597285068



Hình 42: Ma trận nhầm lẫn của Random Forest (Sau giảm chiều dữ liệu)

	Dự đoán nghỉ học	Dự đoán không nghỉ học
Thực tế nghỉ học	160 (TP)	120 (FP)
Thực tế không nghỉ học	45 (FN)	560 N)

Bảng 43: Ma trận nhầm lẫn của Random Forest (Sau giảm chiều dữ liệu)

Trong ma trận nhầm lẫn này, trong số **280 mẫu thực tế nghỉ học**, hệ thống đã đánh giá rằng có **160/280 mẫu khớp với thực tế** (110 mẫu dự báo sai thực tế) đạt **57%** và trong số **604**

mẫu thực tế không nghỉ học thì có tới 560/605 **mẫu khớp với thực tế đưa ra** (45 mẫu dự báo sai thực tế) đạt **92%**

Mặc dù việc **giảm** chiều dữ liệu đã giảm kích thước của đầu vào cho mô hình dự báo, nhưng nó cũng đi kèm với một sự đánh đổi. Tỷ lệ dự đoán đúng học sinh nghỉ học giảm 12%, trong khi tỷ lệ dự đoán đúng học sinh không nghỉ học giảm chỉ 2%. Điều này có thể được hiểu là việc dự đoán học sinh nghỉ học đòi hỏi nhiều yếu tố hơn so với dự đoán học sinh không nghỉ học. Trên thực tế, việc xác định một học sinh có tiếp tục học hay không dễ dàng hơn rất nhiều so với việc xác định một học sinh có nghỉ học hay không.

Do đó, nhóm khuyến khích không giảm chiều dữ liệu khi dự đoán một học sinh có nghỉ học, mà hơn nữa là cần tăng cường thu thập dữ liệu và tìm hiểu thêm về các yếu tố ảnh hưởng đến quá trình học tập của học sinh. Việc này có thể bao gồm việc thu thập thông tin về sức khỏe, môi trường gia đình, và các yếu tố xã hội khác. Nhằm có thể đưa ra dự báo tối ưu nhất về việc một sinh viên thực sự nghỉ học từ đó sẽ đưa ra được những phương án xử lý kịp thời và hiệu quả để bảo vệ, giữ chân sinh viên hoặc có hướng tư vấn hợp lý để sinh viên đưa ra quyết định chuẩn xác hơn trong các trường hợp sau này.

6.4 Kết luận

Qua những mô hình và kết quả dự báo, việc dự đoán liệu một học sinh có đi đến quyết định nghỉ học đặt ra những thách thức đáng kể, điều này yêu cầu sự kết hợp đa dạng của nhiều yếu tố, sẽ tăng cường thu thập thêm dữ liệu và tìm hiểu sâu rộng về các nhân tố ảnh hưởng đến quá trình học tập, sinh hoạt của sinh viên. Từ đó giúp tạo ra một mô hình dự báo đa chiều, chứa đựng thông tin đa dạng về sức khỏe, môi trường gia đình, và các yếu tố xã hội, để có thể cung cấp cái nhìn toàn diện về tình hình của sinh viên.

Việc sử dụng mô hình không chỉ nhằm mục đích dự đoán việc sinh viên nghỉ học mà còn hướng đến mục tiêu quan trọng hơn là đưa ra các biện pháp can thiệp và hỗ trợ phù hợp. Qua đó cũng có thể giúp cơ sở quản lý và giáo dục có những thay đổi về chương trình đào tạo, cơ sở vật chất, học phí, địa điểm cũng như thời gian học cũng như là tư vấn cho những sinh viên có ý định đăng ký sau này về đặc điểm của những ngành học một cách hợp lý.

Tóm lại, dự báo nghỉ học không chỉ là về việc xác định xác suất mà còn là về việc xây dựng cơ sở cho các biện pháp hiệu quả để hỗ trợ học sinh và duy trì môi trường học tốt nhất có thể.

TÀI LIỆU THAM KHẢO

- [0] Giáo trình Biểu diễn Trực quan Dữ liệu, TS. Nguyễn An Té, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.
- [1] Giáo trình Máy học, TS. Nguyễn An Té, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.
- [2] Giáo trình Lập trình Phân tích Dữ liệu, TS. Nguyễn An Té, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.
- [3] Giáo trình Khai phá Dữ liệu, TS. Nguyễn An Té, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.
- [4] Predict students' dropout and academic success (Kaggle) .
- [5] Predicting Student Dropout and Academic Success - Của nhóm nhà nghiên cứu Valentim Realinho, Jorge Machado, Luís Baptista và Mónica V. Martin.
- [6] Predict students' dropout by ML.
- [7] Tuning the hyper-parameters of an estimator by Sklearn.
- [8] Data Visualization with Python.

PHỤ LỤC

DANH MỤC HÌNH ẢNH

Hình 1: Heatmap biểu diễn tương quan giữa các cột dữ liệu	11
Hình 2: Thông số của biến “Target”	13
Hình 3: Số lượng sinh viên nghỉ học theo độ tuổi (Trước xử lý)	14
Hình 4: Số lượng sinh viên của mỗi ngành học	15
Hình 5: Tỉ lệ Các giá trị trong biến “Marital Status”	16
Hình 6: Số lượng các giá trị của biến “Application Mode”	17
Hình 7: Độ tuổi tại thời điểm nhập học của sinh viên	18
Hình 8: Tổng quan độ tương quan của các thuộc tính với “Target”	22
Hình 9: Box plot cho biết độ dàn trải của dữ liệu (Trước xử lý)	23
Hình 10: Boxplot cho biết độ dàn trải của dữ liệu (Sau xử lý)	24
Hình 11: Biểu đồ phân bố tuổi	24
Hình 12: Biểu đồ phân bố tuổi theo biến "Target" (Whisker)	25
Hình 13: Biểu đồ biểu diễn trạng thái học tập của sinh viên theo giới tính (1)	27
Hình 14: Biểu đồ biểu diễn trạng thái học tập của sinh viên theo giới tính (2)	27
Hình 15: Biểu đồ biểu diễn trạng thái học tập của sinh viên theo tình trạng hôn nhân	28
Hình 16: Biểu đồ biểu diễn tỷ lệ nhận học bổng theo trạng thái hôn nhân	29
Hình 17: Biểu đồ biểu diễn số lượng sinh viên trong nước và quốc tế theo trạng thái học tập của sinh viên	30
Hình 18: Biểu đồ biểu diễn số lượng sinh viên theo khu vực	31
Hình 19: Biểu đồ biểu diễn số lượng sinh viên theo vùng quốc tịch	31
Hình 20: Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Buổi học, Tình trạng nợ và Tình trạng học	33
Hình 21: Biểu đồ biểu diễn sự tương quan giữa tình trạng học tập và tỷ lệ thất nghiệp (1)	34
Hình 22: Biểu đồ biểu diễn sự tương quan giữa tình trạng học tập và tỷ lệ thất nghiệp (2)	35
Hình 23: Biểu đồ biểu diễn hồi quy giữa gdp và tỷ lệ thất nghiệp	36
Hình 24: Biểu đồ biểu diễn hồi quy giữa tỷ lệ lạm phát và tỷ lệ thất nghiệp	37
Hình 25: Biểu đồ biểu diễn sự tương quan giữa tỷ lệ lạm phát và số lượng sinh viên bỏ học	38
Hình 26: Biểu đồ biểu diễn sự ảnh hưởng của tăng trưởng GDP tới khả năng nghỉ học của học sinh	39
Hình 27: Biểu đồ biểu diễn top 5 ngành nghề của phụ huynh có lượng sinh viên bỏ học nhiều nhất	41
Hình 28: Biểu đồ biểu diễn sự phân phối của độ tuổi theo các nhóm ngành	42
Hình 29: Biểu đồ biểu diễn số lượng tín chỉ trung bình của sinh viên theo các nhóm ngành	43
Hình 30: Biểu đồ biểu diễn sự phân bố điểm số cả năm theo tuổi và trạng thái học tập của sinh viên	44
Hình 31: Biểu đồ biểu diễn số lượng sinh viên bỏ học của các nhóm ngành học	45
Hình 32: Biểu đồ biểu diễn tỉ lệ bỏ học của các nhóm ngành học	45
Hình 33: Biểu đồ biểu diễn số lượng sinh viên theo Giới tính, Cấp bậc, Tình trạng nợ, Tình trạng học	48
Hình 34: Biểu đồ biểu diễn sự phân phối của điểm số theo học bổng	49
Hình 35: Biểu đồ biểu diễn tỉ lệ đạt được học bổng giữa hai giới	50
Hình 36: Biểu đồ tình trạng hôn nhân của sinh viên và tình trạng học	53

Hình 37: Biểu đồ tỉ lệ nghĩ học giữa sinh viên trong nước và du học sinh	55
Hình 38: Box plot biểu diễn phân phối của các giá trị trong biến "Target" theo tỉ lệ lạm phát	57
Hình 39: Biểu đồ biểu diễn số lượng tín chỉ trung bình của sinh viên theo các nhóm ngành	59
Hình 40: Ma trận nhầm lẫn của Random Forest	67
Hình 41: biểu đồ phân tích phương sai tích lũy theo features	68
Hình 42: Ma trận nhầm lẫn của Random Forest (Sau giảm chiều dữ liệu)	71

DANH MỤC BẢNG BIỂU:

Bảng 1: Các thuộc tính của bộ dữ liệu	8
Bảng 2: Ma trận nhầm lẫn của Random Forest	67
Bảng 3: Ma trận nhầm lẫn của Random Forest (Sau giảm chiều dữ liệu)	71