

TRƯỜNG ĐẠI HỌC MỞ
THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN



THÀNH VIÊN THỰC HIỆN
2251050038 - Nguyễn Vĩ Khang
2251050025 - Thạch Nhựt Hào
2251050008 - Trương Thái Bảo

MÔN HỌC: KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: KHAI PHÁ DỮ LIỆU KHẢO SÁT
SỨC KHỎE TÂM THẦN SINH VIÊN
ĐỂ PHÂN TÍCH NGUY CƠ TRẦM CẢM

TP. Hồ Chí Minh, Năm 2025

MỤC LỤC

Mục lục	i
1 Tổng quan	1
1.1 Giới thiệu bài toán	1
1.2 Dữ liệu sử dụng	2
1.3 Ý nghĩa của một số thuộc tính chính	3
2 Tiền xử lý dữ liệu	4
2.1 Làm sạch dữ liệu	4
2.1.1 Chuẩn hóa tên cột	5
2.1.2 Xử lý giá trị thiếu và không hợp lệ	6
2.1.3 Gom nhóm giá trị danh mục không phổ biến	6
2.1.4 Mã hóa nhị phân cho biến Yes/No	7
2.1.5 Xử lý giá trị bất thường	7
2.1.6 Loại bỏ các cột không có giá trị phân biệt	8
2.1.7 Review dữ liệu	8
2.2 Biến đổi dữ liệu	12
2.2.1 Mã hóa các biến phân loại	12
2.2.2 Chuẩn hóa các biến số liên tục	13

2.2.3	Tạo mới các đặc trưng có ý nghĩa	13
2.3	Đánh giá dữ liệu sau xử lý	14
2.3.1	Không còn giá trị thiếu	14
2.3.2	Các biến phân loại được mã hóa đồng nhất	14
2.3.3	Biến định lượng được chuẩn hóa	15
2.3.4	Các đặc trưng mới giúp tăng giá trị học máy	15
3	Khám phá và trực quan hóa dữ liệu	16
3.1	Phân phối biến mục tiêu	16
3.2	Phân tích một số biến phân loại	18
3.2.1	Giới tính	18
3.2.2	Độ tuổi	19
3.2.3	Học lực (CGPA)	21
3.2.4	Suy nghĩ tiêu cực	22
3.3	Thói quen ăn uống	23
3.4	Sự phân tán các biến số dạng số	24
3.5	Phân tích ma trận tương quan	25
4	Áp dụng các mô hình khai phá dữ liệu	27
4.1	Logistic Regression	27
4.1.1	Cơ sở lý thuyết	27
4.1.2	Lý do lựa chọn	28
4.1.3	Quá trình thực thi	28
4.2	Decision Tree	28
4.2.1	Cơ sở lý thuyết	28
4.2.2	Lý do lựa chọn	29
4.3	Random Forest	30
4.3.1	Cơ sở lý thuyết	30

4.3.2	Lý do lựa chọn	30
4.3.3	Quá trình thực thi	30
4.4	XGBoost	31
4.4.1	Cơ sở lý thuyết	31
4.4.2	Lý do lựa chọn	31
4.4.3	Quá trình thực thi	32
4.5	Naive Bayes	33
4.5.1	Cơ sở lý thuyết	33
4.5.2	Lý do lựa chọn	33
4.6	Apriori – Luật kết hợp	34
4.6.1	Cơ sở lý thuyết	34
4.6.2	Lý do lựa chọn	35
4.6.3	Tiền xử lý dữ liệu	35
4.7	Phân cụm – KMeans & Agglomerative Clustering	36
4.7.1	KMeans	36
4.7.2	Agglomerative Clustering	37
5	Kết quả và phân tích	39
5.1	Logistic Regression	39
5.1.1	Hiệu suất mô hình	39
5.1.2	Phân tích và diễn giải	40
5.2	Decision Tree	42
5.2.1	Trực quan hóa cây quyết định	42
5.2.2	Hiệu suất mô hình	43
5.2.3	Phân tích và nhận xét	44
5.3	Random Forest	45
5.3.1	Hiệu suất mô hình	45
5.3.2	Phân tích và diễn giải	46

5.4	XGBoost	48
5.4.1	Hiệu suất mô hình	48
5.4.2	Phân tích và diễn giải kết quả	49
5.5	Naive Bayes	50
5.5.1	Hiệu suất mô hình	50
5.5.2	Phân tích và diễn giải kết quả	50
5.6	Apriori – Khai phá luật kết hợp	52
5.6.1	Các tập mục phổ biến	52
5.6.2	Các luật dẫn đến trầm cảm	52
5.6.3	Các luật dẫn đến suy nghĩ tự tử	54
5.6.4	Các luật bảo vệ ($Lift < 1$)	55
5.6.5	Diễn giải tổng thể	55
5.7	Phân cụm – KMeans và Agglomerative Clustering	56
5.7.1	KMeans	56
5.7.2	Agglomerative Clustering	59
5.8	Tổng kết hiệu suất các mô hình	60
5.9	Phân tích mối quan hệ giữa các biến	62
5.10	Kiểm định giả thuyết thống kê	67
6	Kết luận và Hướng phát triển	69
6.1	Tổng kết nội dung nghiên cứu	69
6.2	Hạn chế của đề tài	70
6.3	Hướng phát triển	71
	Phân công công việc nhóm	73

Chương 1

TỔNG QUAN

1.1 Giới thiệu bài toán

Trầm cảm là một trong những vấn đề sức khỏe tâm thần nghiêm trọng đang ảnh hưởng đến nhiều đối tượng trong xã hội, đặc biệt là sinh viên, người trẻ. Nhóm này thường xuyên phải đối mặt với áp lực học tập, tài chính và các thay đổi trong cuộc sống cá nhân. Việc phát hiện sớm các yếu tố nguy cơ gây trầm cảm có thể giúp nhà trường, gia đình và bản thân sinh viên có biện pháp can thiệp phù hợp, từ đó góp phần cải thiện sức khỏe tâm lý và chất lượng cuộc sống.

Dự án này hướng tới việc khai thác dữ liệu khảo sát sinh viên để dự đoán nguy cơ trầm cảm dựa trên các yếu tố cá nhân, học tập và xã hội. Thông qua việc áp dụng các kỹ thuật khai phá dữ liệu như phân tích mối quan hệ giữa các biến, xây dựng mô hình dự đoán và tìm luật kết hợp, nhóm nghiên cứu mong muốn tìm ra những đặc điểm nổi bật có thể đóng vai trò quan trọng trong việc sàng lọc và phòng ngừa trầm cảm trong cộng đồng sinh viên.

1.2 Dữ liệu sử dụng

- **Tên tập dữ liệu:** Student Depression Dataset
- **Nguồn dữ liệu:** <https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset>
- **Chủ đề:** Khảo sát sức khỏe tâm thần của sinh viên, tập trung vào các yếu tố học tập, cá nhân và hành vi để phân tích nguy cơ trầm cảm.
- **Quy mô dữ liệu:**
 - Số dòng: khoảng 27.000 sinh viên
 - Số cột: 18 cột, bao gồm cả thuộc tính và nhãn mục tiêu
- **Các nhóm thuộc tính chính:**
 - *Demographics:* Age, Gender, City
 - *Academic Indicators:* CGPA, Academic Pressure, Study Satisfaction
 - *Lifestyle & Wellbeing:* Sleep Duration, Dietary Habits, Work Pressure, Job Satisfaction, Work/Study Hours
 - *Additional Factors:* Profession, Degree, Financial Stress, Family History of Mental Illness, Suicidal Thoughts
 - *Target Variable:* **Depression_Status** – nhãn nhị phân (Yes/No hoặc 0/1)
- **Mục tiêu khai phá dữ liệu:**
 - Phân tích các yếu tố hành vi và tâm lý có liên quan đến nguy cơ trầm cảm ở sinh viên.

- Phân cụm sinh viên theo mức độ stress, áp lực học tập và chất lượng cuộc sống.
- Dự đoán nguy cơ trầm cảm bằng các mô hình học máy.
- Khai phá các luật kết hợp để nhận diện nhóm nguy cơ cao.
- Gợi ý xây dựng hệ thống sàng lọc sớm trầm cảm trong môi trường giáo dục.

1.3 Ý nghĩa của một số thuộc tính chính

- **Depression_Status**: Biến mục tiêu, phản ánh tình trạng trầm cảm của sinh viên (Yes/No hoặc 1/0).
- **Suicidal_Thoughts**: Sinh viên có từng có ý định tự tử hay không.
- **Academic_Pressure**: Mức độ áp lực học tập của sinh viên.
- **CGPA**: Điểm trung bình tích lũy.
- **Financial_Stress**: Mức độ căng thẳng tài chính.
- **Dietary_Habits**: Thói quen ăn uống (Unhealthy, Moderate, Healthy).
- **Study_Satisfaction**: Mức độ hài lòng với việc học.
- **Sleep_Duration**: Thời lượng giấc ngủ mỗi ngày.
- **Work/Study_Hours**: Tổng thời gian học và làm việc mỗi ngày.
- **Family_Mental_History**: Tiền sử bệnh tâm thần trong gia đình.

Các thuộc tính này đóng vai trò quan trọng trong mô hình hóa dữ liệu, là đầu vào cho các kỹ thuật phân tích và học máy được sử dụng trong nghiên cứu.

Chương 2

TIỀN XỬ LÝ DỮ LIỆU

Tiền xử lý dữ liệu là bước quan trọng trong toàn bộ quy trình khai phá dữ liệu. Dữ liệu thực tế thường không hoàn hảo – có thể chứa lỗi, thiếu giá trị, hoặc không đồng nhất về định dạng. Do đó, chất lượng của mô hình học máy phụ thuộc lớn vào độ sạch và sự chuẩn hóa của dữ liệu đầu vào. Trong chương này, các bước tiền xử lý dữ liệu sẽ được trình bày cụ thể, bao gồm làm sạch, mã hóa, chuẩn hóa, và tạo đặc trưng mới.

2.1 Làm sạch dữ liệu

Làm sạch dữ liệu là một bước quan trọng và không thể thiếu trong quá trình tiền xử lý dữ liệu, nhằm đảm bảo đầu vào chất lượng cao cho các mô hình máy học (ML) và phân tích nghiệp vụ thông minh (BI). Dữ liệu thu thập từ thực tế thường tồn tại các vấn đề như lỗi chính tả, định dạng không thống nhất, dữ liệu thiếu, không hợp lệ hoặc không liên quan đến mục tiêu phân tích. Những sai lệch này nếu không được xử lý có thể gây ảnh hưởng tiêu cực đến hiệu suất của mô hình và độ tin cậy của kết quả. Các bước làm sạch dữ liệu trên bao gồm:

2.1.1 Chuẩn hóa tên cột

Trong hệ thống phân tích dữ liệu, chuẩn hóa tên cột là bước tiền xử lý quan trọng giúp đảm bảo tính nhất quán và tránh lỗi khi thao tác với dữ liệu, đặc biệt trong môi trường như Python/Pandas. Các thao tác bao gồm: loại bỏ khoảng trắng, thay thế bằng dấu gạch dưới, xóa ký tự đặc biệt và chuyển toàn bộ tên cột về chữ thường. Việc này giúp mã lệnh đơn giản, dễ truy xuất và phù hợp với quy tắc đặt tên biến trong lập trình.

```
[ ] # Chuẩn hóa tên cột
def clean_column_names(df):
    df.columns = (
        df.columns
        .str.strip()
        .str.replace(" ", "_")
        .str.replace("?", "")
        .str.lower()
    )
    return df
```

Hình 2.1: Chuẩn hóa tên cột

Ngoài ra, một số tên cột dài hoặc dư thừa được rút gọn có chủ đích, chẳng hạn như `have_you_ever_had_suicidal_thoughts` được đơn giản thành `suicidal_thoughts`, vừa đảm bảo ngắn gọn, vừa giữ nguyên ý nghĩa ban đầu.

```
[ ] df_student.columns

Index(['id', 'Gender', 'Age', 'City', 'Profession', 'Academic Pressure',
      'Work Pressure', 'CGPA', 'Study Satisfaction', 'Job Satisfaction',
      'Sleep Duration', 'Dietary Habits', 'Degree',
      'Have you ever had suicidal thoughts ?', 'Work/Study Hours',
      'Financial Stress', 'Family History of Mental Illness', 'Depression'],
      dtype='object')
```

Hình 2.2: Trước khi chuẩn hóa

```
[22] df_cleaned.columns  
  
Index(['id', 'gender', 'age', 'city', 'profession', 'academic_pressure',  
       'work_pressure', 'cgpa', 'study_satisfaction', 'job_satisfaction',  
       'sleep_duration', 'dietary_habits', 'degree', 'suicidal_thoughts',  
       'work/study_hours', 'financial_stress', 'family_mental_history',  
       'depression'],  
      dtype='object')
```

Hình 2.3: Sau khi chuẩn hóa

2.1.2 Xử lý giá trị thiếu và không hợp lệ

Cột `financial_stress` xuất hiện một số giá trị không hợp lệ (ký tự '?'). Các giá trị này được chuyển thành NaN để phù hợp với định dạng dữ liệu thiếu mà pandas hỗ trợ.

Để xử lý, phương pháp điền trung vị được áp dụng. Trung vị được tính từ các giá trị hợp lệ còn lại và dùng để thay thế toàn bộ giá trị thiếu.

Lý do chọn trung vị thay vì trung bình là vì trung vị: ít nhạy cảm với giá trị ngoại lai và phản ánh xu hướng trung tâm tốt hơn trong các phân phối lệch. Cách làm này giúp giữ nguyên số lượng mẫu, hạn chế mất mát thông tin và đảm bảo dữ liệu đồng nhất cho các bước phân tích tiếp theo.

2.1.3 Gom nhóm giá trị danh mục không phổ biến

Cột `city` trong tập dữ liệu bao gồm nhiều giá trị phân loại, trong đó một số thành phố xuất hiện với tần suất rất thấp. Để xử lý, các thành phố ít gặp được gom vào nhóm chung mang nhãn `Others`, dựa trên ngưỡng xuất hiện tối thiểu.

```
[ ] # City
valid_cities = [
    'Visakhapatnam', 'Bangalore', 'Srinagar', 'Varanasi', 'Jaipur', 'Pune', 'Thane',
    'Chennai', 'Nagpur', 'Nashik', 'Vadodara', 'Kalyan', 'Rajkot', 'Ahmedabad',
    'Kolkata', 'Mumbai', 'Lucknow', 'Indore', 'Surat', 'Ludhiana', 'Bhopal',
    'Meerut', 'Agra', 'Ghaziabad', 'Hyderabad', 'Vasai-Virar', 'Kanpur', 'Patna',
    'Faridabad', 'Delhi'
]

# Những thành phố có tên không hợp lệ sẽ được gán là 'Others'
df_cleaned['city'] = df_cleaned['city'].apply(lambda x: x if x in valid_cities else 'Others')
```

Hình 2.4: Gom nhóm giá trị city không phổ biến

Việc rút gọn này giúp giảm số chiều khi mã hóa one-hot, tiết kiệm tài nguyên và tránh làm loãng dữ liệu. Đồng thời, nó hạn chế overfitting do loại bỏ các nhãn hiếm dễ gây nhiễu, và tăng khả năng khái quát khi mô hình gặp các giá trị mới không xuất hiện trong tập huấn luyện. Đây là kỹ thuật tiền xử lý hiệu quả đối với các biến phân loại có nhiều nhãn và phân phối không đồng đều.

2.1.4 Mã hóa nhị phân cho biến Yes/No

Các biến phân loại dạng Yes/No như `suicidal_thoughts` và `family_mental_history` được ánh xạ thành giá trị nhị phân (Yes \rightarrow 1, No \rightarrow 0), giúp chuyển đổi dữ liệu ký tự sang dạng số – định dạng cần thiết cho các thuật toán học máy. Việc mã hóa này không chỉ giữ nguyên ý nghĩa ban đầu mà còn giúp tối ưu hiệu suất mô hình. Việc ánh xạ được thực hiện như sau:

```
[ ] # Chuyển yes/no về 1/0
df_cleaned['suicidal_thoughts'] = df_cleaned['suicidal_thoughts'].map({'No': 0, 'Yes': 1})
df_cleaned['family_mental_history'] = df_cleaned['family_mental_history'].map({'No': 0, 'Yes': 1})
```

Hình 2.5: Mã hóa nhị phân cho dữ liệu Yes/No

2.1.5 Xử lý giá trị bất thường

Trong dữ liệu, cột `cgpa` xuất hiện các giá trị bằng 0 – điều này không hợp lệ trong hệ thống chấm điểm học thuật. Do đó, các bản ghi có `cgpa = 0` được

loại bỏ hoàn toàn. Sau đó, tiến hành làm tròn giá trị `cgpa` đến hai chữ số thập phân để thống nhất định dạng và thuận tiện khi hiển thị.


2.1.6 Loại bỏ các cột không có giá trị phân biệt

Các cột không có giá trị phân biệt hoặc thể hiện sự biến thiên thấp đã được loại bỏ khỏi tập dữ liệu để tối ưu hóa quá trình huấn luyện mô hình. Cụ thể, biến `profession` chỉ chứa một giá trị duy nhất là `Student`, nên không mang lại thông tin phân loại. Biến `work_pressure` không cho thấy mối liên hệ rõ ràng với biến mục tiêu, trong khi `job_satisfaction` có phân phối lệch nghiêm trọng, hầu hết các giá trị đều bằng 0.

Việc loại bỏ các đặc trưng như vậy giúp giảm nhiễu, rút gọn số chiều, và nâng cao khả năng tổng quát của mô hình bằng cách tập trung vào các biến thực sự có ý nghĩa trong phân tích.

2.1.7 Review dữ liệu


Dữ liệu trước khi được làm sạch




	id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction
0	2	Male	33.0	Visakhapatnam	Student	5.0	0.0	8.97	2.0	0.0
1	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5.0	0.0
2	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0
3	30	Female	28.0	Varanasi	Student	3.0	0.0	5.59	2.0	0.0
4	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3.0	0.0

Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
5-6 hours	Healthy	B.Pharm	Yes	3.0	1.0	No	1
5-6 hours	Moderate	BSc	No	3.0	2.0	Yes	0
Less than 5 hours	Healthy	BA	No	9.0	1.0	Yes	0
7-8 hours	Moderate	BCA	Yes	4.0	5.0	Yes	1
5-6 hours	Moderate	M.Tech	Yes	1.0	1.0	No	0

Hình 2.6: 5 dòng dữ liệu đầu



```
[55] df_student.shape
```

```
(27870, 18)
```

Hình 2.7: Số dòng dữ liệu

```
[51] df_student.describe()
```

	id	Age	Academic Pressure	Work Pressure	CGPA
count	27870.000000	27870.000000	27870.000000	27870.000000	27870.000000
mean	70439.182634	25.821134	3.140617	0.000431	7.656028
std	40633.159539	4.906574	1.381640	0.044016	1.470835
min	2.000000	18.000000	0.000000	0.000000	0.000000
25%	35055.000000	21.000000	2.000000	0.000000	6.290000
50%	70654.500000	25.000000	3.000000	0.000000	7.770000
75%	105813.000000	30.000000	4.000000	0.000000	8.920000
max	140699.000000	59.000000	5.000000	5.000000	10.000000

Study Satisfaction	Job Satisfaction	Work/Study Hours	Financial Stress	Depression
27870.000000	27870.000000	27870.000000	27867.000000	27870.000000
2.943559	0.000682	7.157948	3.139520	0.585145
1.361209	0.044419	3.707180	1.437289	0.492706
0.000000	0.000000	0.000000	1.000000	0.000000
2.000000	0.000000	4.000000	2.000000	0.000000
3.000000	0.000000	8.000000	3.000000	1.000000
4.000000	0.000000	10.000000	4.000000	1.000000
5.000000	4.000000	12.000000	5.000000	1.000000

Hình 2.8: Mô tả dữ liệu

Dữ liệu sau khi được làm sạch

```
df_cleaned.head()
```

	id	gender	age	city	academic_pressure	cgpa	study_satisfaction	sleep_duration	dietary_habits
0	2	Male	33.0	Visakhapatnam	5.0	8.97	2.0	5-6 hours	Healthy
1	8	Female	24.0	Bangalore	2.0	5.90	5.0	5-6 hours	Moderate
2	26	Male	31.0	Srinagar	3.0	7.03	5.0	Less than 5 hours	Healthy
3	30	Female	28.0	Varanasi	3.0	5.59	2.0	7-8 hours	Moderate
4	32	Female	25.0	Jaipur	4.0	8.13	3.0	5-6 hours	Moderate

degree	suicidal_thoughts	work/study_hours	financial_stress	family_mental_history	depression
B.Pharm	1	3.0	1.0	0	1
BSc	0	3.0	2.0	1	0
BA	0	9.0	1.0	1	0
BCA	1	4.0	5.0	1	1
M.Tech	1	1.0	1.0	0	0

Hình 2.9: 5 dòng dữ liệu đầu

```
[58] df_cleaned.shape
```

```
(27870, 18)
```

Hình 2.10: Số dòng dữ liệu

```
df_cleaned.describe()
```

	id	age	academic_pressure	work_pressure	cgpa
count	27870.000000	27870.000000	27870.000000	27870.000000	27870.000000
mean	70439.182634	25.821134	3.140617	0.000431	7.656028
std	40633.159539	4.906574	1.381640	0.044016	1.470835
min	2.000000	18.000000	0.000000	0.000000	0.000000
25%	35055.000000	21.000000	2.000000	0.000000	6.290000
50%	70654.500000	25.000000	3.000000	0.000000	7.770000
75%	105813.000000	30.000000	4.000000	0.000000	8.920000
max	140699.000000	59.000000	5.000000	5.000000	10.000000

study_satisfaction	job_satisfaction	work/study_hours	financial_stress	depression
27870.000000	27870.000000	27870.000000	27870.000000	27870.000000
2.943559	0.000682	7.157948	3.139505	0.585145
1.361209	0.044419	3.707180	1.437212	0.492706
0.000000	0.000000	0.000000	1.000000	0.000000
2.000000	0.000000	4.000000	2.000000	0.000000
3.000000	0.000000	8.000000	3.000000	1.000000
4.000000	0.000000	10.000000	4.000000	1.000000
5.000000	4.000000	12.000000	5.000000	1.000000

Hình 2.11: Mô tả dữ liệu

Nhận xét: Quá trình làm sạch dữ liệu trên đã góp phần quan trọng trong việc chuẩn bị dữ liệu đầu vào và chất lượng cho các bước phân tích và mô hình hóa.

- Tăng tính nhất quán và dễ thao tác trong quá trình xử lý, nhờ việc chuẩn

hóa cú pháp tên biến và chuẩn định dạng dữ liệu.

- Giảm nhiễu và rút gọn số chiều, bằng cách loại bỏ các cột dư thừa, ít biến thiên hoặc không liên quan đến mục tiêu phân loại.

2.2 Biến đổi dữ liệu

Sau khi làm sạch dữ liệu, bước tiếp theo là biến đổi dữ liệu để mô hình học máy có thể xử lý hiệu quả. Việc biến đổi bao gồm: mã hóa các biến phân loại, chuẩn hóa các biến liên tục, và tạo ra các đặc trưng mới có ý nghĩa.

2.2.1 Mã hóa các biến phân loại

Các biến định tính (categorical features) không thể được đưa trực tiếp vào mô hình học máy. Vì vậy, cần mã hóa các biến này thành định dạng số. Dữ liệu sử dụng các kỹ thuật sau:

- **Label Encoding:** Áp dụng cho biến có thứ tự tự nhiên. Ví dụ:
 - `gender`: Male = 0, Female = 1
 - `sleep_duration`: được ánh xạ từ ít ngủ đến nhiều ngủ với các giá trị từ 1 đến 4
 - `dietary_habits`: Unhealthy = 1, Moderate = 2, Healthy = 3
- **One-Hot Encoding:** Áp dụng cho biến không có thứ tự như `city`, `degree`. Mỗi giá trị duy nhất được chuyển thành một cột nhị phân riêng biệt.

Việc chọn đúng kỹ thuật mã hóa là quan trọng vì nó ảnh hưởng đến khả năng học của mô hình. Với mô hình tuyến tính như Logistic Regression,

Label Encoding phù hợp cho biến có thứ tự, trong khi One-Hot giúp tránh mô hình hiểu sai mối quan hệ giữa các nhãn rời rạc.

2.2.2 Chuẩn hóa các biến số liên tục

Các biến định lượng như `age`, `cgpa`, `academic_pressure`, `study_satisfaction`, `work/study_hours`, `financial_stress` có thang đo khác nhau. Để đảm bảo tất cả biến có trọng số tương đương khi huấn luyện, dữ liệu được chuẩn hóa bằng:

- **StandardScaler**: Đưa giá trị về phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1:

$$z = \frac{x - \mu}{\sigma}$$

Chuẩn hóa đặc biệt quan trọng với các mô hình nhạy cảm với khoảng cách như KNN hoặc sử dụng Gradient Descent như Logistic Regression.

2.2.3 Tạo mới các đặc trưng có ý nghĩa

Ngoài các biến gốc, một số đặc trưng được tạo thêm để tăng khả năng phân biệt cho mô hình:

- **sleep_adequate**: Biến nhị phân, gán 1 nếu sinh viên ngủ từ 7 giờ trở lên (tương ứng `sleep_duration` ≥ 3), ngược lại là 0.
- **high_academic_pressure**: Gán 1 nếu điểm chuẩn hóa của `academic_pressure` lớn hơn trung bình (0).
- **stress_interaction**: Biến tương tác giữa `financial_stress` và `academic_pressure`, được tính bằng:

$$\text{stress_interaction} = \text{financial_stress} \times \text{academic_pressure}$$

Biến này phản ánh ảnh hưởng kết hợp giữa áp lực tài chính và học tập. Hai yếu tố này được giả thuyết là có tác động cộng hưởng lên tình trạng tâm lý.

Việc tạo đặc trưng mới giúp mô hình học được mối liên hệ phi tuyến giữa các yếu tố, đồng thời phản ánh tri thức chuyên môn trong quá trình phân tích.

2.3 Đánh giá dữ liệu sau xử lý

Sau khi hoàn tất các bước làm sạch và biến đổi dữ liệu, cần đánh giá lại toàn bộ tập dữ liệu để đảm bảo chất lượng đầu vào cho mô hình học máy. Các tiêu chí kiểm tra như sau:

2.3.1 Không còn giá trị thiếu

Toàn bộ các giá trị thiếu (NaN) trong các cột quan trọng đã được xử lý thông qua:

- Điền trung vị cho giá trị thiếu trong `financial_stress`.
- Loại bỏ các dòng có giá trị không hợp lệ trong `cgpa`.
- Điền mode hoặc gom nhóm `city` hiếm vào `Others`.

Kết quả: Dữ liệu không còn thiếu giá trị và đảm bảo hoàn chỉnh.

2.3.2 Các biến phân loại được mã hóa đồng nhất

Tất cả các biến phân loại đã được xử lý và mã hóa nhất quán:

- Biến Yes/No \rightarrow nhị phân 0/1
- Biến có thứ tự \rightarrow Label Encoding

- Biến danh mục → One-Hot Encoding

Việc này giúp mô hình học máy có thể hiểu đúng cấu trúc dữ liệu và xử lý chính xác từng loại biến.

2.3.3 Biến định lượng được chuẩn hóa

Các biến định lượng như `age`, `cgpa`, `academic_pressure` đã được chuẩn hóa về phân phối chuẩn bằng `StandardScaler`, đảm bảo:

- Không có biến nào thống trị vì đơn vị đo quá lớn.
- Các thuật toán như KNN, Logistic Regression hoạt động ổn định.

2.3.4 Các đặc trưng mới giúp tăng giá trị học máy

Các đặc trưng được tạo thêm như `sleep_adequate`, `high_academic_pressure`, `stress_interaction` đều mang ý nghĩa ngữ nghĩa rõ ràng và phản ánh những giả thuyết thực tiễn. Chúng hỗ trợ mô hình học tốt hơn và góp phần giải thích kết quả về sau.

Kết luận: Bộ dữ liệu sau tiền xử lý đã đạt tiêu chuẩn để đưa vào các thuật toán học máy. Đây là nền tảng vững chắc cho các bước mô hình hóa, phân tích và khai phá tri thức trong các chương tiếp theo.

📄 Kích thước cuối cùng: (27860, 73)
Số Null còn lại: 0

	id	gender	age	academic_pressure	cgpa	study_satisfaction	sleep_duration	dietary_habits	suicidal_thoughts	work/study_hours	financial_str
count	27861.000000	27861.000000	2.786100e+04	2.786100e+04	2.786100e+04	2.786100e+04	27861.000000	27861.000000	27861.000000	2.786100e+04	2.786100e+04
mean	70434.708912	0.442841	5.279148e-17	-1.887433e-16	-7.681141e-16	-1.438377e-16	2.387294	1.903942	0.632605	7.268302e-18	1.147841e
std	40633.834254	0.498731	1.000019e+00	1.000019e+00	1.000019e+00	1.000019e+00	1.127916	0.797780	0.482104	1.000018e+00	1.000018e
min	2.000000	0.000000	-1.594302e+00	-2.274721e+00	-1.794696e+00	-2.163693e+00	0.000000	0.000000	0.000000	-1.930955e+00	-1.488479e
25%	35046.000000	0.000000	-8.627941e-01	-8.264577e-01	-9.343891e-01	-6.938826e-01	1.000000	1.000000	0.000000	-8.518804e-01	-7.827117e
50%	70680.000000	0.000000	-1.673803e-01	-1.023261e-01	7.612978e-02	4.102269e-02	2.000000	2.000000	1.000000	2.271839e-01	-9.694442e
75%	105810.000000	1.000000	8.518484e-01	6.218056e-01	8.613303e-01	7.780280e-01	3.000000	3.000000	1.000000	7.667311e-01	5.988228e
max	140688.000000	1.000000	6.763382e+00	1.345937e+00	1.598736e+00	1.511033e+00	4.000000	3.000000	1.000000	1.306268e+00	1.294590e

Hình 2.12: Dữ liệu sau khi xử lý

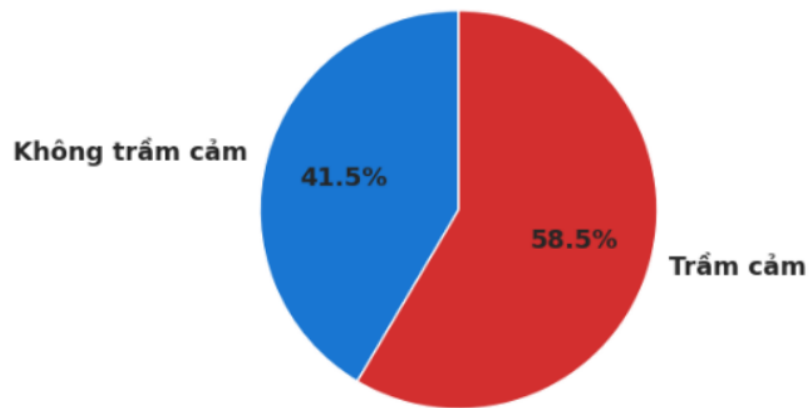
Chương 3

KHÁM PHÁ VÀ TRỰC QUAN HÓA DỮ LIỆU

Trước khi tiến hành xây dựng mô hình dự đoán hay khai thác các yếu tố tác động đến trạng thái trầm cảm, việc phân tích và hiểu rõ phân bố tổng thể của các biến đóng vai trò vô cùng quan trọng. Trong chương này, các biểu đồ và thống kê mô tả sẽ được sử dụng để khám phá dữ liệu, kiểm tra sự phân phối của biến mục tiêu và các đặc trưng đầu vào, từ đó phát hiện các xu hướng tiềm ẩn.

3.1 Phân phối biến mục tiêu

Biến mục tiêu `depression_status` là nhãn phân loại nhị phân cho biết sinh viên có đang trải qua trầm cảm hay không.



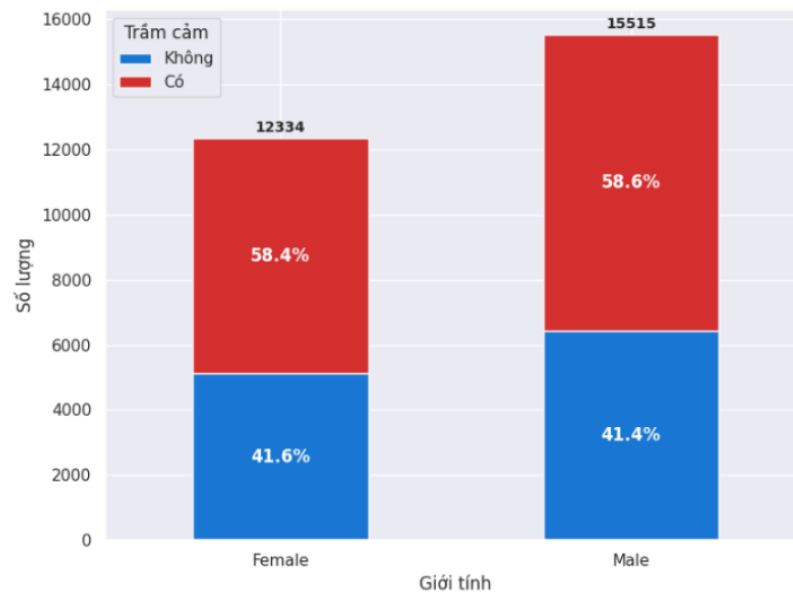
Hình 3.1: Phân phối trạng thái trầm cảm trong dữ liệu

Tỉ lệ trầm cảm chiếm đa số, tỉ lệ trầm cảm ở sinh viên khoảng (58,5%) lớn hơn tỉ lệ không trầm cảm (41,5%). Cho thấy đây trầm cảm đang là một vấn đề “nhức nhối” đối với sinh viên trong môi trường đại học. Đây là một vấn đề ta nên quan tâm và phân tích.

Phân bố này khá cân bằng dù tỉ lệ trầm cảm cao hơn nhưng vẫn chưa tới mức bị mất cân đối trầm trọng. Do đó, các phân tích thống kê và xây dựng mô hình dự đoán sẽ ít bị ảnh hưởng bởi vấn đề dữ liệu mất cân bằng nghiêm trọng.

3.2 Phân tích một số biến phân loại

3.2.1 Giới tính

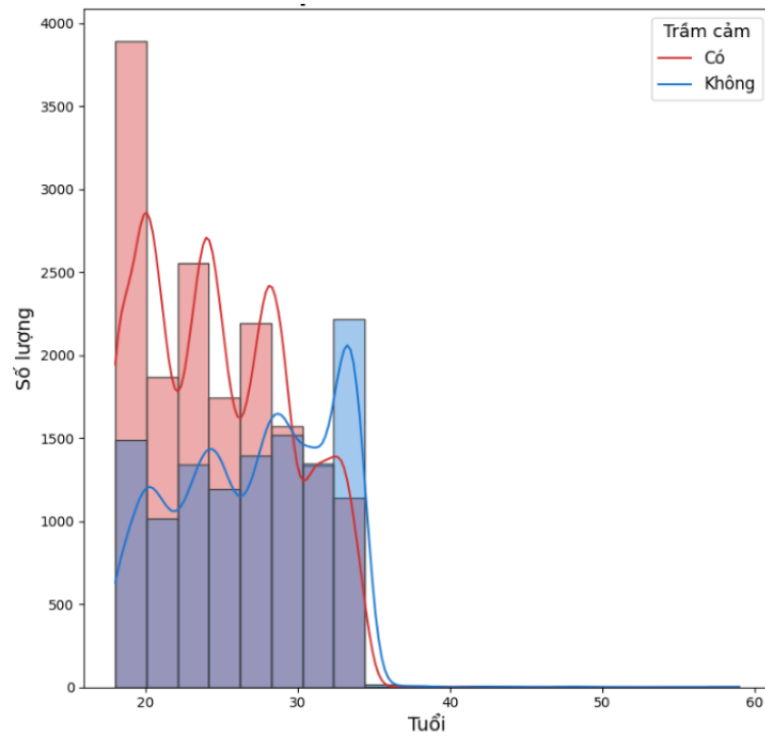


Hình 3.2: Phân phối trầm cảm theo giới tính

Nhìn chung, số lượng Male và Female có chênh lệch nhưng tỉ lệ trầm cảm ở cả 2 giới tính khá giống nhau. Điều này cho thấy mức độ trầm cảm không ảnh hưởng đến giới tính. Nghĩa là nam vẫn có thể bị trầm cảm và nữ cũng không ngoại lệ.

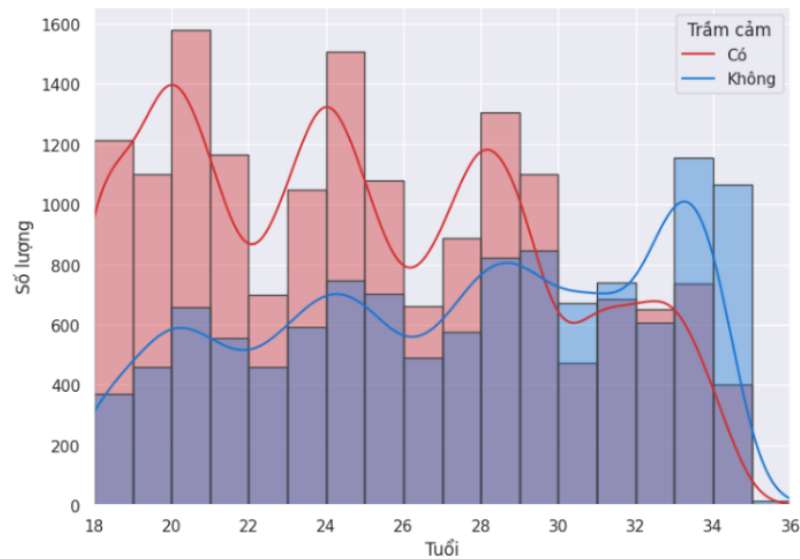
=> Giới tính dường như không phải là yếu tố quyết định chính đến khả năng bị trầm cảm trong tập mẫu này.

3.2.2 Độ tuổi



Hình 3.3: Phân phối trầm cảm theo nhóm tuổi

- Chủ yếu tuổi phân bố tập trung từ dưới 20 đến 36 nên ta sẽ vẽ lại biểu đồ tập trung ở độ tuổi đó để có thể đưa ra thêm các phân tích sâu hơn cho bộ dữ liệu.

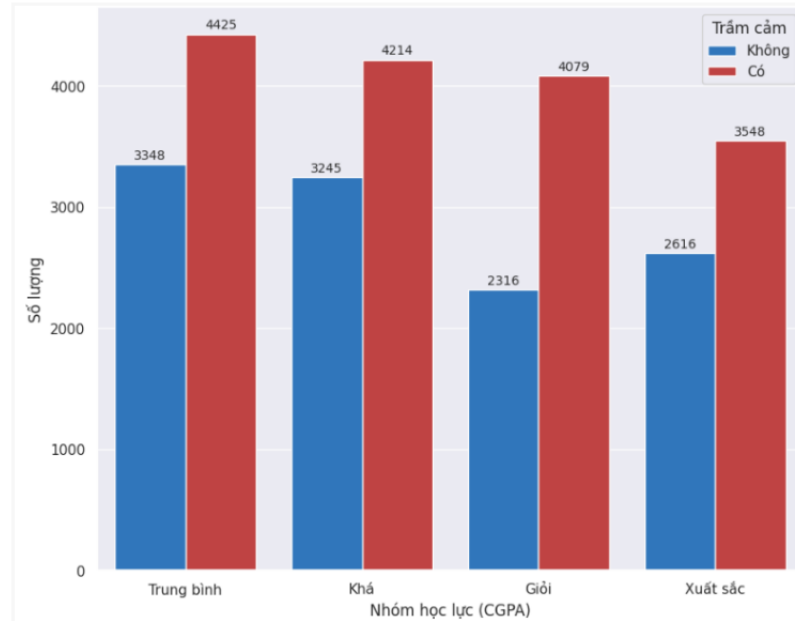


Hình 3.4: Phân phối trầm cảm theo nhóm tuổi từ 18 -36

- Ở độ tuổi trẻ (18 – 24), số bạn có trầm cảm luôn nhỉnh hơn rõ rệt so với số bạn không trầm cảm.
- Trong khoảng 25 – 30 tuổi, hai nhóm có mức độ gần nhau hơn, nhưng nhóm “có trầm cảm” vẫn hơi chiếm ưu thế cho đến khoảng 28 tuổi.
- Từ 31 – 36 tuổi ngược lại, số bạn không trầm cảm vượt lên, đặc biệt đỉnh ở 34 tuổi.

=> **Tóm lại:** Sinh viên mới vào đại học (18 – 22) dễ gặp stress, áp lực học tập, dẫn đến tỷ lệ trầm cảm cao. Sự “ổn định” khi lớn hơn, những bạn 30 – 36 tuổi có thể đã thích nghi tốt hơn với việc học/tìm việc và ít bị trầm cảm. Vùng “giao thoa” (25 – 30 tuổi), đây là đối tượng tiềm ẩn – cần kết hợp thêm biến `academic_pressure`, `social_support`,... đều giải thích tại sao cùng lứa tuổi vẫn có bạn trầm cảm khác nhau.

3.2.3 Học lực (CGPA)



Hình 3.5: Tỷ lệ trầm cảm theo nhóm học lực

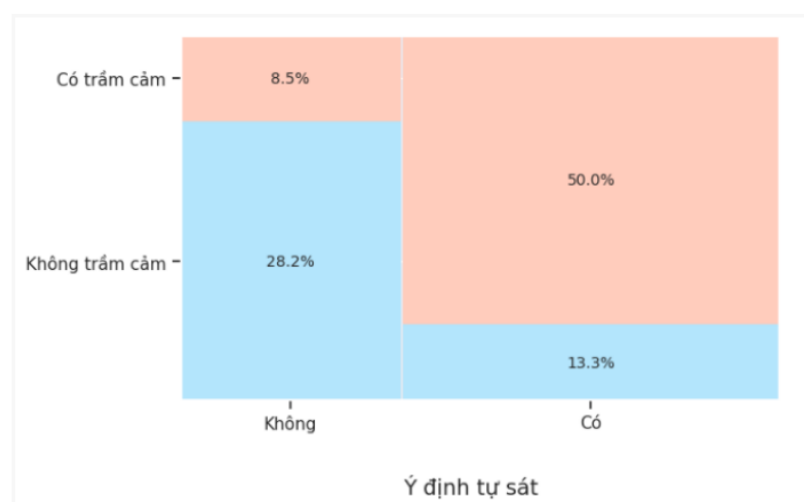
- Dễ nhận thấy ở mọi nhóm học lực, số sinh viên bị mắc chứng trầm cảm luôn cao. Điều này cho biết trầm cảm là một vấn đề khá phổ biến ở mọi đối tượng học lực, không chỉ riêng học lực mạnh hay yếu.
- Số lượng sinh viên trầm cảm giảm dần từ Trung bình đến Xuất sắc, điều này có thể là do số lượng sinh viên của các học lực này cũng giảm dần.
- Tỷ lệ trầm cảm của nhóm Giỏi cao bất thường:
 - “Trung bình”: $4427/7777 \approx 56.9\%$
 - “Khá”: $4215/7462 \approx 56.5\%$
 - “Giỏi”: $4081/6397 \approx 63.8\%$
 - “Xuất sắc”: $3551/6167 \approx 57.6\%$

- Điều này có thể là do áp lực duy trì thành tích học tập, kỳ vọng cao.

- Nhóm “Xuất sắc” có số lượng trầm cảm thấp nhất về số tuyệt đối và tương đối, có thể do nhóm này đã đạt được sự tự tin, cân bằng tâm lý, hoặc số lượng mẫu cũng ít hơn các nhóm còn lại.

⇒ Đây cũng là một khía cạnh nên đào sâu để đưa ra các suy luận đúng đắn.

3.2.4 Suy nghĩ tiêu cực



Hình 3.6: Tỷ lệ trầm cảm theo ý định tự tử

- Phần lớn sinh viên có ý định tự sát cũng bị trầm cảm chiếm 50% số mẫu – đây là tỷ lệ cực kỳ cao. Điều này khẳng định mối liên hệ rất chặt giữa trầm cảm và ý định tự sát trong tập dữ liệu.
- Sinh viên không có ý định tự sát nhưng vẫn bị trầm cảm mặc dù chỉ với tỷ lệ thấp (8,5%), nhưng cũng chỉ ra ý định tự sát chỉ là nguy cơ, không phải là điều kiện đủ cho việc bị trầm cảm.
- Một bộ phận nhỏ sinh viên không bị trầm cảm nhưng vẫn có ý định tự sát (13,3%).

- Số sinh viên “bình thường” (không trầm cảm, không ý định tự sát) chỉ chiếm khoảng 28.2% tổng số mẫu, cho thấy sức khỏe tâm thần là vấn đề lớn trong tập dữ liệu này.

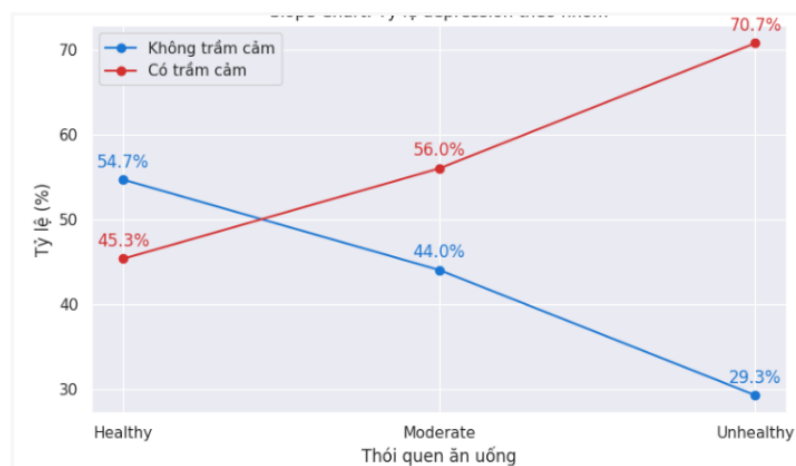
3.3 Thói quen ăn uống

Phân phối 'dietary_habits':

	dietary_habits	Count	Percentage
0	Unhealthy	10306	36.99%
1	Moderate	9906	35.56%
2	Healthy	7637	27.41%
3	Others	12	0.04%

Hình 3.7: Kiểm tra phân phối thói quen ăn uống

Kiểm tra các giá trị của cột `dietary_habits` cho thấy nhóm “Others” có tỉ lệ rất thấp (≈ 0.004), ta sẽ điền mode vào các giá trị đó để không làm mất giá trị các cột khác.

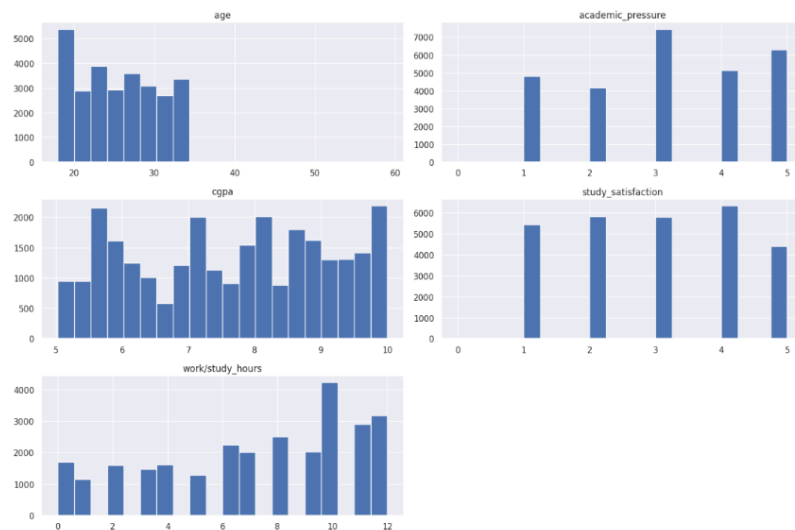


Hình 3.8: Tỷ lệ trầm cảm theo thói quen ăn uống

- Tỷ lệ trầm cảm ở nhóm có thói quen ăn uống “Healthy” là thấp nhất.
- Nhóm “Unhealthy” có tỷ lệ trầm cảm cao nhất.

⇒ **Chế độ ăn uống không lành mạnh có liên quan mạnh đến nguy cơ trầm cảm ở sinh viên.** Nhóm “Moderate” cũng có tỷ lệ trầm cảm khá cao, cần được quan tâm, không nên chỉ chú ý đến nhóm ăn uống cực kỳ kém. Đáng chú ý, ngay cả nhóm có chế độ ăn “Healthy” cũng có tỷ lệ trầm cảm gần 50%, cho thấy thói quen ăn uống chỉ là một trong các yếu tố ảnh hưởng đến trầm cảm.

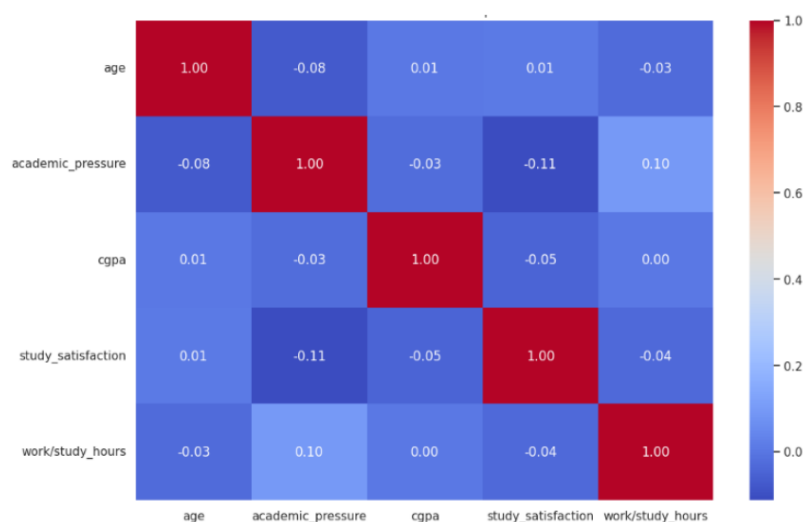
3.4 Sự phân tán các biến số dạng số



Hình 3.9: Biểu đồ phân tán các biến số

Sinh viên chủ yếu trong độ tuổi 18 - 25, có CGPA khá cao, thường chịu áp lực học tập trung bình đến cao và dành nhiều thời gian cho học tập/làm việc. Đa số sinh viên cảm thấy hài lòng với việc học của mình.

3.5 Phân tích ma trận tương quan



Hình 3.10: Biểu đồ nhiệt thể hiện sự tương quan giữa các biến số

- `academic_pressure` và `study_satisfaction` có hệ số tương quan âm đáng chú ý là -0.11 , cho thấy áp lực học tập cao hơn có xu hướng đi kèm với sự hài lòng về học tập thấp hơn, tuy nhiên tương quan này rất yếu.
- `academic_pressure` và `work/study_hours` có tương quan dương nhẹ (0.10), thể hiện rằng áp lực học tập cao hơn một chút khi sinh viên dành nhiều thời gian hơn cho việc học/làm, tuy nhiên cũng khá yếu và không rõ ràng.
- Biến `age`, `cgpa`, và `work/study_hours` gần như không có mối liên hệ đáng kể với các biến còn lại.

Tóm lại, các biến số này khá độc lập với nhau. Khi xây dựng mô hình khai phá dữ liệu, các biến này có thể được giữ lại mà không cần lo ngại nhiều về hiện tượng đa cộng tuyến. Tuy nhiên, cần lưu ý rằng các biến này có thể có

mối quan hệ phi tuyến hoặc phức tạp hơn mà heatmap này chưa thể hiện được.

Chương 4

ÁP DỤNG CÁC MÔ HÌNH KHAI PHÁ DỮ LIỆU

Trong chương này, các mô hình học máy được triển khai nhằm dự đoán trạng thái trầm cảm của sinh viên dựa trên các đặc trưng hành vi, học tập và tâm lý. Các mô hình bao gồm: Logistic Regression, Decision Tree, Random Forest, XGBoost, Naive Bayes, và các phương pháp khai phá luật kết hợp và phân cụm như Apriori, K-Means, và Agglomerative Clustering.

4.1 Logistic Regression

4.1.1 Cơ sở lý thuyết

Logistic Regression là mô hình phân loại tuyến tính, sử dụng hàm sigmoid để ánh xạ đầu ra về khoảng $[0,1]$, đại diện cho xác suất xảy ra trầm cảm. Hàm mất mát được tối ưu là Binary Cross-Entropy:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad ; \quad \mathcal{L} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

4.1.2 Lý do lựa chọn

Logistic Regression là mô hình nền tảng, dễ diễn giải, cho phép kiểm tra trọng số của từng biến đầu vào. Điều này giúp xác định yếu tố nào có ảnh hưởng lớn đến xác suất trầm cảm.

4.1.3 Quá trình thực thi

Mô hình được huấn luyện trên dữ liệu đã chuẩn hóa và mã hóa bằng One-Hot Encoding. Không sử dụng regularization để bảo toàn khả năng giải thích của mô hình. Dự đoán được đánh giá thông qua các chỉ số:

- Độ chính xác (Accuracy)
- Precision, Recall theo macro trung bình
- F1-score tổng hợp

4.2 Decision Tree

4.2.1 Cơ sở lý thuyết

Cây quyết định là một thuật toán học máy thuộc nhóm mô hình có giám sát, được sử dụng cho cả phân loại và hồi quy. Mô hình này biểu diễn các quyết định dưới dạng một cấu trúc cây gồm các nút, nhánh và lá.

Nguyên lý hoạt động:

1. Chọn thuộc tính (feature) tốt nhất để phân tách dữ liệu tại mỗi nút, dựa trên các chỉ số như:
2. Tạo nhánh mới cho mỗi giá trị của thuộc tính đã chọn dựa trên:

- Entropy & Information Gain (ID3, C4.5)
- Gini Index (CART)

3. Lặp lại quá trình cho từng nhánh con đến khi:

- Dữ liệu tại nút là đồng nhất (cùng 1 lớp)
- Không còn thuộc tính nào để phân tách
- Đạt tới độ sâu tối đa hoặc số lượng mẫu tối thiểu

Tiêu chí chọn thuộc tính phân tách:

- **Entropy:** Đo mức độ hỗn loạn/tạp trong dữ liệu.
- **Information Gain:** Đo lượng mức độ giảm Entropy khi phân chia dữ liệu theo thuộc tính nào đó.
- **Gini Index:** Đo xác suất chọn sai nhãn nếu chọn ngẫu nhiên theo phân phối nhãn tại nút.

4.2.2 Lý do lựa chọn

Cây quyết định là một trong những mô hình dễ hiểu và dễ trực quan hóa nhất, giúp người đọc, kể cả những người không chuyên về dữ liệu, có thể nhanh chóng nắm bắt được cách thức mô hình đưa ra quyết định thông qua các nhánh và nút trong cây. Tính minh bạch và khả năng giải thích cao là yếu tố rất quan trọng đối với các bài toán trong lĩnh vực giáo dục, khoa học xã hội hay những nghiên cứu cần sự rõ ràng, dễ trình bày kết quả cho các bên liên quan. Bên cạnh đó, mô hình không đòi hỏi các giả định phức tạp về phân phối dữ liệu hoặc mối quan hệ tuyến tính giữa các biến, cho phép áp dụng hiệu quả trên nhiều dạng dữ liệu thực tế vốn đa dạng và thường xuyên tồn tại các mối quan hệ phi tuyến.

4.3 Random Forest

4.3.1 Cơ sở lý thuyết

Random Forest là mô hình học tổ hợp thuộc nhóm *ensemble*, xây dựng nhiều cây quyết định (*decision trees*) và tổng hợp kết quả bằng cơ chế voting (phân loại) hoặc trung bình (hồi quy). Các cây được huấn luyện trên tập con dữ liệu bootstrap và chọn ngẫu nhiên một tập con thuộc tính tại mỗi nút chia:

- Giảm phương sai của mô hình gốc (Decision Tree)
- Tăng khả năng tổng quát hóa

Công thức voting phân loại:

$$\hat{y} = \text{majority_vote}(y_1, y_2, \dots, y_K)$$

4.3.2 Lý do lựa chọn

Random Forest có khả năng mô hình hóa dữ liệu phức tạp, không yêu cầu chuẩn hóa dữ liệu và xử lý tốt các biến phân loại. Đồng thời, mô hình cho phép đánh giá độ quan trọng của từng đặc trưng đầu vào, hỗ trợ phân tích nguyên nhân dẫn đến trầm cảm.

4.3.3 Quá trình thực thi

Mô hình được xây dựng với thư viện `sklearn.ensemble.RandomForestClassifier` cùng các tham số tối ưu hóa:

- `n_estimators = 100`: số lượng cây
- `max_depth = 20`: giới hạn độ sâu để tránh overfitting
- `random_state = 42`: đảm bảo tái lập kết quả

4.4 XGBoost

4.4.1 Cơ sở lý thuyết

XGBoost, viết tắt của *Extreme Gradient Boosting*, là một thuật toán học có giám sát được xây dựng dựa trên phương pháp boosting theo gradient. Trong quá trình huấn luyện, các cây quyết định được xây dựng một cách tuần tự nhằm khắc phục sai số của những cây trước đó. Mỗi cây mới đóng vai trò điều chỉnh mô hình thông qua việc tối thiểu hóa hàm mất mát bằng kỹ thuật *gradient descent*, nhưng thay vì tối ưu trên không gian tham số, quá trình này diễn ra trên không gian hàm mục tiêu.

So với các phương pháp boosting truyền thống, XGBoost thể hiện nhiều điểm cải tiến đáng kể. Một trong những đặc trưng nổi bật là khả năng điều chuẩn (*regularization*) linh hoạt, cho phép mô hình kiểm soát tốt hơn hiện tượng quá khớp. Ngoài ra, thuật toán được thiết kế tối ưu về mặt hiệu năng, không chỉ hỗ trợ xử lý dữ liệu thiếu một cách hiệu quả mà còn tận dụng được khả năng tính toán song song và bộ nhớ đệm, từ đó rút ngắn đáng kể thời gian huấn luyện khi làm việc với tập dữ liệu lớn. Nhờ những ưu điểm này, XGBoost hiện là một trong những lựa chọn phổ biến và mạnh mẽ trong nhiều bài toán học máy hiện đại.

4.4.2 Lý do lựa chọn

XGBoost được lựa chọn làm thuật toán chính cho bài toán phân loại chất lượng đánh giá trò chơi (tốt/xấu) nhờ khả năng thích ứng cao với đặc điểm của tập dữ liệu khảo sát. Với ưu thế xử lý hiệu quả dữ liệu dạng bảng đã qua mã hóa và làm sạch, XGBoost đặc biệt phù hợp với các đặc trưng phân loại được biểu diễn dưới dạng *one-hot encoding*.

Bên cạnh đó, thuật toán này tích hợp cơ chế điều chỉnh trọng số cho các lớp mất cân bằng thông qua tham số `scale_pos_weight`, giúp cải thiện độ chính xác trong bối cảnh nhãn phân loại không đồng đều. Khả năng kiểm soát hiện tượng quá khớp cũng được đảm bảo nhờ vào các kỹ thuật điều chuẩn và cắt tỉa cây (*pruning*) được tích hợp sẵn.

Trong thực tiễn, XGBoost thường mang lại hiệu suất vượt trội so với các mô hình truyền thống như *Logistic Regression* hay *Decision Tree*, đặc biệt khi số lượng đặc trưng lớn và có sự tương tác phức tạp giữa các biến.

4.4.3 Quá trình thực thi

Mô hình XGBoost được xây dựng thông qua lớp `XGBClassifier` từ thư viện `xgboost`, sử dụng tập dữ liệu huấn luyện đã được xử lý và chuẩn hóa trong giai đoạn tiền xử lý. Quá trình huấn luyện được thực hiện với tập `X_train` và `y_train`, trong đó các tham số mô hình được thiết lập một cách có chủ đích nhằm tối ưu hóa hiệu suất đồng thời duy trì khả năng tổng quát.

Các siêu tham số được lựa chọn dựa trên quá trình thử nghiệm và điều chỉnh thủ công. Cụ thể:

- `max_depth = 10`: kiểm soát độ sâu của cây và hạn chế mô hình học quá mức.
- `n_estimators = 200`: xác định số lượng cây được xây dựng trong quá trình boosting.
- `learning_rate = 0.05`: giúp làm chậm quá trình học để mô hình cập nhật ổn định hơn.
- `random_state = 42`: đảm bảo tính tái lập trong các lần chạy khác nhau.

4.5 Naive Bayes

4.5.1 Cơ sở lý thuyết

Thuật toán **Naive Bayes** là một phương pháp phân loại dựa trên định lý Bayes, với giả định đơn giản (“naive”) rằng các đặc trưng đầu vào là độc lập có điều kiện với nhau. Mặc dù giả định này hiếm khi đúng trong thực tế, nhưng nó giúp đơn giản hóa tính toán và thường vẫn đem lại kết quả phân loại hiệu quả.

Công thức tính xác suất hậu nghiệm được sử dụng trong mô hình là:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Thuật toán hoạt động bằng cách tính xác suất xảy ra của từng lớp dựa trên dữ liệu quan sát và gán nhãn lớp với xác suất hậu nghiệm lớn nhất. Tùy theo loại dữ liệu đặc trưng, các biến thể phổ biến của thuật toán Naive Bayes có thể được sử dụng như:

- **GaussianNB**: dữ liệu liên tục
- **MultinomialNB**: cho dữ liệu rời rạc dạng đếm
- **BernoulliNB**: cho dữ liệu nhị phân

4.5.2 Lý do lựa chọn

Trong các biến thể của Naive Bayes, Gaussian Naive Bayes được lựa chọn để triển khai, bởi mô hình này đặc biệt phù hợp với các đặc trưng dạng số liên tục đã được chuẩn hóa thông qua các phương pháp như **StandardScaler** hoặc **MinMaxScaler**. Việc chuẩn hóa giúp dữ liệu tiệm cận giả định phân phối chuẩn mà GaussianNB yêu cầu.

4.6 Apriori – Luật kết hợp

4.6.1 Cơ sở lý thuyết

Thuật toán Apriori là một kỹ thuật nền tảng trong lĩnh vực khai phá luật kết hợp (Association Rule Mining), được sử dụng để khám phá các mối quan hệ dạng “nếu – thì” giữa các mục thường xuất hiện cùng nhau trong tập dữ liệu. Mục tiêu của thuật toán là tìm ra các tập mục phổ biến (frequent itemsets) có độ hỗ trợ (support) đủ cao, từ đó sinh ra các luật kết hợp có ý nghĩa thống kê.

Một luật kết hợp được biểu diễn dưới dạng $A \rightarrow B$, với các chỉ số đánh giá chính gồm:

- **Support:** Tỷ lệ xuất hiện của tập hợp $A \cup B$ trong toàn bộ dữ liệu, phản ánh mức độ phổ biến của luật.
- **Confidence:** Xác suất để mục B xảy ra khi biết rằng A đã xảy ra, thể hiện độ tin cậy của luật.
- **Lift:** Tỷ lệ so sánh giữa xác suất đồng thời xảy ra của A và B so với xác suất chúng xảy ra độc lập; nếu $\text{Lift} > 1$, điều đó cho thấy giữa A và B tồn tại mối liên hệ dương.

Thuật toán hoạt động theo nguyên tắc mở rộng dần các tập mục từ nhỏ đến lớn, khai thác tính chất bắc cầu của tập phổ biến: nếu một tập con không thỏa ngưỡng support, thì các siêu tập của nó cũng không cần xét đến. Nhờ đó, Apriori giúp rút gọn đáng kể không gian tìm kiếm và tập trung vào những mối quan hệ có giá trị phân tích cao trong dữ liệu.

4.6.2 Lý do lựa chọn

Thuật toán Apriori được lựa chọn trong nghiên cứu này vì phù hợp với mục tiêu khám phá mối liên hệ giữa các đặc trưng hành vi, lối sống và tình trạng tâm lý của học sinh – sinh viên. Những mối quan hệ dạng “nếu – thì” có thể hỗ trợ trong việc hiểu rõ hơn các yếu tố liên quan đến trầm cảm, suy nghĩ tiêu cực,... Ngoài ra, Apriori là phương pháp phổ biến, dễ triển khai và có thể diễn giải trực quan bằng các chỉ số support, confidence và lift – phù hợp với phân tích mô tả và hỗ trợ ra quyết định chính sách giáo dục – tâm lý.

4.6.3 Tiền xử lý dữ liệu

Trong nghiên cứu này, dữ liệu đầu vào là kết quả khảo sát hành vi và tâm lý của học sinh – sinh viên. Để có thể áp dụng thuật toán Apriori, vốn yêu cầu dữ liệu ở dạng nhị phân (0/1), quá trình xử lý dữ liệu được thực hiện tuần tự nhằm chuẩn hóa định dạng:

- Trước tiên, các biến định tính như giới tính được nhị phân hóa. Ví dụ: `gender_male = 1` nếu là nam.
- Các biến phân loại như `city`, `degree`, `dietary_habits`, `age_group`, `cgpa_group` được mã hóa bằng **one-hot encoding**.
- Các biến liên tục như `cgpa`, `academic_pressure`, `work/study_hours` được phân loại thành các nhóm định tính (“Low”, “Medium”, “High”). Riêng `sleep_duration` được chia thành hai nhóm: “Adequate” và “Inadequate”.

Toàn bộ dữ liệu sau đó được chuyển thành định dạng nhị phân bằng `astype(int)`. Chỉ những biến có ý nghĩa phân tích – bao gồm các đặc trưng rời rạc đã mã hóa và các biến mục tiêu như `suicidal_thoughts`, `depression`,

family_mental_history – được giữ lại để phục vụ khai phá luật kết hợp.

4.7 Phân cụm – KMeans & Agglomerative Clustering

4.7.1 KMeans

Cơ sở lý thuyết

KMeans là một thuật toán phân cụm thuộc nhóm học không giám sát, thường được sử dụng để chia tập dữ liệu thành K cụm sao cho các điểm dữ liệu trong cùng một cụm có tính chất tương đồng cao nhất.

Nguyên lý hoạt động: tối thiểu hóa tổng bình phương khoảng cách (SSE) giữa các điểm dữ liệu với tâm cụm gần nhất. Quá trình thuật toán:

1. Khởi tạo: Chọn ngẫu nhiên K tâm cụm ban đầu (centroid).
2. Gán nhãn: Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất (thường dùng khoảng cách Euclidean).
3. Cập nhật tâm cụm: Tính lại vị trí tâm cụm bằng trung bình cộng các điểm trong cụm.
4. Lặp lại: Thực hiện hai bước trên cho đến khi không còn sự thay đổi trong gán cụm hoặc đạt tới số lần lặp tối đa.

Công thức toán học:

$$SSE = \sum_{i=1}^n \min_{k \in \{1, \dots, K\}} \|x_i - \mu_k\|^2$$

Với x_i là điểm dữ liệu, μ_k là tâm cụm thứ k , và K là số cụm.

Lý do lựa chọn

Đây là mô hình phân cụm đơn giản và dễ hiểu nhất trong bài toán học không giám sát. Tuy vậy, đơn giản nhưng thuật toán rất mạnh trong việc xác định các nhóm đối tượng tương đồng trong dữ liệu mà không cần phải có sẵn nhãn phân loại. Điều này đặc biệt hữu ích đối với các bộ dữ liệu lớn, đa dạng hoặc chưa rõ cấu trúc, giúp nhà phân tích khám phá những phân khúc ẩn hoặc mẫu hành vi đặc biệt mà các phương pháp truyền thống khó nhận biết. Ngoài ra, KMeans có tốc độ xử lý nhanh, hiệu quả về mặt tính toán và kết quả phân cụm hỗ trợ trực quan hóa dữ liệu rõ ràng, giúp kiểm chứng, đánh giá và truyền đạt kết quả dễ dàng hơn. Việc áp dụng KMeans còn góp phần nâng cao giá trị khai phá dữ liệu, hỗ trợ ra quyết định như phân khúc khách hàng, phát hiện bất thường hoặc làm đầu vào cho các mô hình phân tích tiếp theo.

4.7.2 Agglomerative Clustering

Cơ sở lý thuyết

Agglomerative Clustering là một thuật toán phân cụm phân cấp (Hierarchical Clustering) theo phương pháp gộp dần (bottom-up). Đây là một trong hai cách tiếp cận chính của phân cụm phân cấp, ngược lại với phương pháp chia tách dần (divisive).

Ở Agglomerative Clustering, ban đầu mỗi điểm dữ liệu được coi là một cụm riêng biệt. Sau đó, thuật toán lần lượt hợp nhất các cụm gần nhau nhất thành các cụm lớn hơn, quá trình này lặp lại cho đến khi tất cả điểm dữ liệu gộp thành một cụm duy nhất hoặc đạt tới số cụm mong muốn.

Nguyên lý hoạt động:

1. **Khởi tạo:** Mỗi điểm dữ liệu là một cụm riêng biệt.
2. **Tính toán khoảng cách:** Xác định khoảng cách giữa tất cả các cặp cụm.
3. **Gộp cụm:** Hợp nhất hai cụm gần nhau nhất thành một cụm mới.
4. **Lặp lại:** Tiếp tục gộp các cụm gần nhau nhất cho đến khi đạt số cụm mong muốn hoặc không còn cụm nào để gộp nữa.

Các phương pháp đo khoảng cách giữa cụm

- **Single linkage:** Khoảng cách giữa hai cụm là khoảng cách nhỏ nhất giữa một cặp điểm bất kỳ, mỗi điểm thuộc một cụm.
- **Complete linkage:** Khoảng cách lớn nhất giữa một cặp điểm bất kỳ.
- **Average linkage:** Trung bình các khoảng cách giữa mọi cặp điểm.
- **Ward linkage:** Dựa trên phương sai tổng trong cụm.

Lý do lựa chọn

Agglomerative Clustering là một lựa chọn phù hợp cho các bài toán phân cụm khi cần khai thác các cấu trúc phân cấp hoặc chưa biết rõ số lượng cụm trước. Thuật toán này không yêu cầu xác định trước số cụm, giúp dễ dàng khám phá dữ liệu và nhận diện các phân cấp tự nhiên giữa các đối tượng, điều mà các phương pháp như KMeans không làm được. Bên cạnh đó, Agglomerative Clustering có thể xử lý tốt dữ liệu có hình dạng cụm phức tạp, không yêu cầu các cụm phải có dạng hình cầu như KMeans.

KẾT QUẢ VÀ PHÂN TÍCH

Trong chương này, kết quả từ các mô hình được đánh giá và phân tích chi tiết. Mục tiêu là không chỉ đo lường hiệu suất về mặt chỉ số (accuracy, precision, recall, F1-score) mà còn rút ra được các ý nghĩa thực tiễn từ các đặc trưng đầu vào ảnh hưởng đến trạng thái trầm cảm. Ngoài ra, một số kết quả từ luật kết hợp và phân cụm sẽ giúp khám phá tri thức tiềm ẩn trong dữ liệu.

5.1 Logistic Regression

5.1.1 Hiệu suất mô hình

Sau khi được huấn luyện trên tập huấn luyện và đánh giá trên tập kiểm tra, mô hình Logistic Regression cho thấy hiệu suất ổn định và đáng tin cậy. Cụ thể, độ chính xác (Accuracy) của mô hình đạt 83.9%, cho thấy tỷ lệ phân loại đúng các trường hợp trầm cảm và không trầm cảm là khá cao. Chỉ số F1-score trung bình đạt 0.84, phản ánh sự cân bằng tốt giữa độ chính xác (precision) và khả năng bao phủ (recall) của mô hình đối với cả hai lớp. Đặc biệt, chỉ số ROC-AUC đạt 0.917, cho thấy mô hình có khả năng phân biệt hai lớp rất tốt, ngay cả khi ngưỡng phân loại thay đổi.

Bảng 5.1 – Báo cáo hiệu suất Logistic Regression

Lớp	Precision	Recall	F1-score	Support
0 (Không trầm cảm)	0.79	0.83	0.81	2312
1 (Trầm cảm)	0.88	0.84	0.86	3261
Accuracy	0.839			
Macro Avg	0.83	0.84	0.84	5573
Weighted Avg	0.84	0.84	0.84	5573

Báo cáo phân loại chi tiết cho thấy: precision của lớp trầm cảm đạt 0.88, recall đạt 0.84, và F1-score là 0.86 – những con số này thể hiện rằng mô hình có khả năng phát hiện đúng phần lớn sinh viên có nguy cơ trầm cảm, đồng thời kiểm soát tốt tỷ lệ báo động giả. Ma trận nhầm lẫn tương ứng là:

$$\begin{bmatrix} 1923 & 389 \\ 506 & 2755 \end{bmatrix}$$

Trong đó, 2755 sinh viên bị trầm cảm đã được phát hiện đúng, và chỉ có 506 trường hợp bị bỏ sót (false negatives), điều này cho thấy mô hình phù hợp với mục tiêu phát hiện sớm nguy cơ tâm lý trong môi trường học đường.

5.1.2 Phân tích và diễn giải

Một trong những ưu điểm nổi bật của Logistic Regression là khả năng diễn giải các trọng số (hệ số hồi quy) ứng với từng đặc trưng đầu vào. Các hệ số hồi quy thể hiện ảnh hưởng của từng biến đến log-odds của xác suất trầm cảm. Các hệ số dương phản ánh mối liên hệ đồng biến (biến tăng \rightarrow xác suất trầm cảm tăng), trong khi hệ số âm cho thấy mối quan hệ nghịch biến.

Bảng 5.2 – Một số trọng số đặc trưng của mô hình

Biến đầu vào	Trọng số (Coefficient)
suicidal_thoughts	2.56
academic_pressure	1.18
financial_stress	0.80
age	-0.58
dietary_habits	-0.53

Trong kết quả thu được, đặc trưng **suicidal_thoughts** (suy nghĩ tự tử) có hệ số cao nhất với giá trị 2.56, cho thấy đây là yếu tố dự báo mạnh nhất cho tình trạng trầm cảm. Theo sau đó là **academic_pressure** (áp lực học tập) với hệ số 1.18 và **financial_stress** (căng thẳng tài chính) với hệ số 0.80, phản ánh rõ vai trò của áp lực học tập và tài chính trong sức khỏe tâm lý sinh viên. Đáng chú ý, biến **age** có hệ số âm (-0.58), cho thấy sinh viên lớn tuổi hơn có xu hướng ít trầm cảm hơn, có thể do họ đã thích nghi tốt hơn với môi trường học thuật hoặc có khả năng kiểm soát cảm xúc cao hơn.

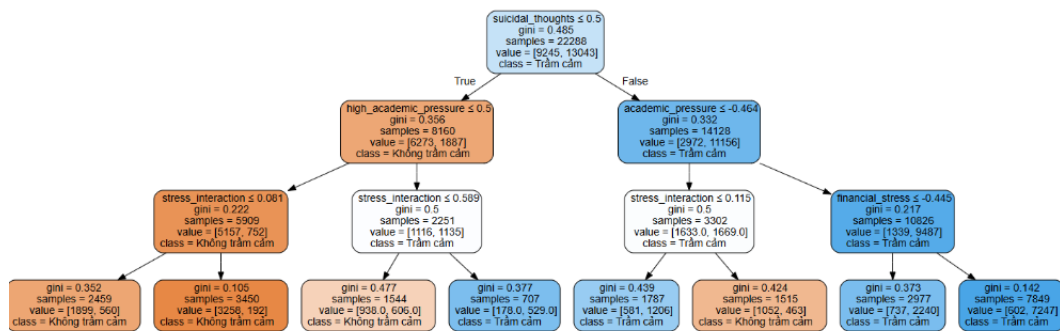
Các hệ số dương thể hiện mối quan hệ tích cực với khả năng trầm cảm, trong khi hệ số âm thể hiện mối quan hệ nghịch. Cụ thể, ý nghĩ tự tử là yếu tố ảnh hưởng mạnh nhất. Áp lực học tập và căng thẳng tài chính cũng là các yếu tố nổi bật liên quan đến trạng thái trầm cảm. Ngược lại, các đặc trưng như tuổi và thói quen ăn uống lành mạnh có xu hướng giúp giảm nguy cơ.

Nhìn chung, Logistic Regression không chỉ cho thấy hiệu suất dự đoán cao mà còn mang lại khả năng giải thích mô hình rõ ràng. Điều đặc biệt hữu ích trong các nghiên cứu giáo dục, xã hội hoặc y tế cộng đồng. Việc xác định các đặc trưng ảnh hưởng mạnh đến nguy cơ trầm cảm là cơ sở quan trọng để thiết kế các chương trình hỗ trợ và can thiệp sớm cho sinh viên.

5.2 Decision Tree

5.2.1 Trực quan hóa cây quyết định

Một trong những ưu điểm của mô hình Decision Tree là khả năng trực quan hóa quá trình ra quyết định thông qua các điều kiện phân nhánh. Hình dưới đây minh họa một phần cấu trúc cây quyết định được huấn luyện từ tập dữ liệu:



Hình 5.1: Cây quyết định phân loại trầm cảm

Trong cây trên, các điều kiện phân tách ở các nút cho thấy rõ vai trò của một số đặc trưng đầu vào trong việc phân biệt sinh viên có và không có dấu hiệu trầm cảm. Cụ thể:

- **suicidal_thoughts** (có hay không từng có ý nghĩ tự tử) là điều kiện gốc – thể hiện vai trò mạnh mẽ trong việc xác định tình trạng trầm cảm.
- Các nhánh tiếp theo liên quan đến **academic_pressure**, **financial_stress** và **stress_interaction**, cho thấy sự kết hợp giữa áp lực học tập và tài chính đóng vai trò phân tách quan trọng.
- Mỗi nút lá trong cây biểu thị số lượng mẫu, phân bố theo hai lớp, và lớp được dự đoán cuối cùng (trầm cảm hoặc không).

5.2.2 Hiệu suất mô hình

Mô hình cây quyết định (Decision Tree) được đánh giá bằng phương pháp cross-validation với 3 folds trên tập dữ liệu đã được xử lý. Kết quả cho thấy độ chính xác trung bình của mô hình đạt 82.4%, với các fold dao động rất nhỏ (82.2% đến 82.6%). Điều này chứng tỏ mô hình có khả năng tổng quát hóa tốt, ít bị ảnh hưởng bởi overfitting và hoạt động ổn định trên các phần phân mảnh dữ liệu khác nhau.

Bảng 5.3 – Độ chính xác theo cross-validation (3 folds)

- Accuracy trung bình: 0.824
- Accuracy từng fold: [0.8258, 0.8224, 0.8247]

Mô hình khi được đánh giá trên toàn bộ tập kiểm tra cho độ chính xác (Accuracy) đạt 82.1%. Cụ thể, precision của lớp trầm cảm đạt 84%, recall đạt 85% và F1-score là 0.85 – thể hiện khả năng nhận diện tốt các sinh viên đang có dấu hiệu trầm cảm. Trong khi đó, nhóm không trầm cảm có recall chỉ đạt 75% và F1-score là 0.78, tức mô hình còn nhầm lẫn một số sinh viên không trầm cảm thành trầm cảm.

Bảng 5.4 – Báo cáo phân loại Decision Tree

Lớp	Precision	Recall	F1-score	Support
0 (Không trầm cảm)	0.79	0.75	0.78	2312
1 (Trầm cảm)	0.84	0.85	0.85	3261
Accuracy	0.821			
Macro Avg	0.82	0.81	0.82	5573
Weighted Avg	0.82	0.82	0.82	5573

Ma trận nhầm lẫn tương ứng như sau:

$$\begin{bmatrix} 1796 & 516 \\ 479 & 2782 \end{bmatrix}$$

5.2.3 Phân tích và nhận xét

Mô hình cây quyết định đã cho thấy hiệu quả đáng kể trong việc phát hiện sinh viên trầm cảm, với recall nhóm này đạt đến 85% – tức là phát hiện đúng đa số sinh viên thật sự bị trầm cảm, chỉ bỏ sót 479 trên tổng số 2782 trường hợp. Đây là điểm rất quan trọng trong các hệ thống sàng lọc vì mô hình ưu tiên không bỏ sót các ca nguy cơ.

Tuy nhiên, mô hình vẫn còn tỷ lệ báo động giả tương đối cao. Có 516 sinh viên không trầm cảm bị dự đoán nhầm là trầm cảm (false positives), dẫn đến recall nhóm không trầm cảm chỉ là 75% và precision là 79%. Điều này thể hiện sự đánh đổi trong độ nhạy và độ chính xác khi áp dụng mô hình thiên về phát hiện trầm cảm hơn là bảo toàn tính đúng đắn tuyệt đối cho lớp âm tính.

Nhìn chung, hiệu suất tổng thể của Decision Tree ở mức tốt với độ chính xác

82% và các chỉ số macro trung bình đều ở mức 0.82, cho thấy mô hình khá cân bằng giữa hai lớp. Ngoài ra, một ưu điểm khác là khả năng trực quan hóa, dễ giải thích cho người dùng không chuyên.

5.3 Random Forest

5.3.1 Hiệu suất mô hình

Trong nghiên cứu này, mô hình Random Forest được triển khai nhằm dự đoán nguy cơ trầm cảm của sinh viên dựa trên tập dữ liệu đã được xử lý tiền xử lý và biến đổi. Mô hình được xây dựng với các tham số cơ bản gồm 100 cây quyết định (`n_estimators = 100`), sử dụng trọng số cân bằng (`class_weight = 'balanced'`) để đối phó với vấn đề mất cân bằng lớp, và `random_state = 42` nhằm đảm bảo tính tái lập.

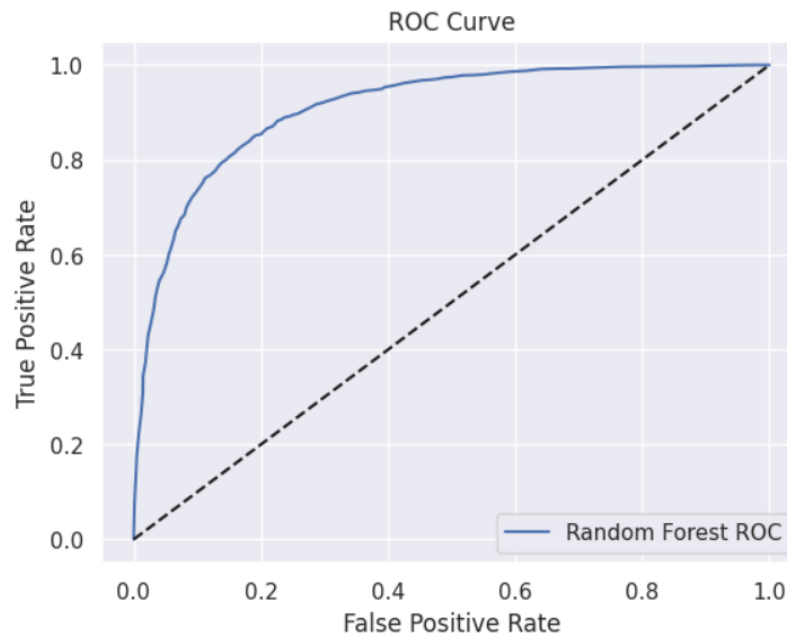
Sau quá trình huấn luyện, mô hình được đánh giá trên tập kiểm tra với các chỉ số hiệu suất như sau:

Bảng 5.5 – Hiệu suất mô hình Random Forest				
Lớp	Precision	Recall	F1-score	Support
0 (Không trầm cảm)	0.82	0.78	0.80	2312
1 (Trầm cảm)	0.85	0.88	0.86	3261
Accuracy	0.835			
Macro Avg	0.83	0.83	0.83	5573
Weighted Avg	0.83	0.84	0.83	5573

Ma trận nhầm lẫn thu được như sau:

$$\begin{bmatrix} 1797 & 515 \\ 402 & 2859 \end{bmatrix}$$

Mô hình cũng được đánh giá bằng đường cong ROC (Receiver Operating Characteristic), thể hiện sự so sánh giữa độ nhạy (True Positive Rate) và tỷ lệ báo động sai (False Positive Rate). Kết quả cho thấy diện tích dưới đường cong (ROC-AUC) đạt 0.910, chứng tỏ khả năng phân biệt rất tốt giữa hai lớp (có và không trầm cảm), bất kể ngưỡng phân loại.



Hình 5.2: Đường cong ROC của mô hình Random Forest

5.3.2 Phân tích và diễn giải

Một trong những ưu điểm nổi bật của Random Forest là khả năng đo lường độ quan trọng của các đặc trưng đầu vào đối với kết quả dự đoán. Kết quả phân tích feature importance cho thấy những biến có ảnh hưởng lớn nhất đến xác suất trầm cảm gồm:

Bảng 5.6 – Top 10 đặc trưng quan trọng nhất

Biến đầu vào	Trọng số quan trọng (importance)
suicidal_thoughts	0.1887
academic_pressure	0.0878
stress_interaction	0.072
financial_stress	0.0644
high_academic_pressure	0.0603
age	0.0600
id	0.0568
cgpa	0.0547
work/study_hours	0.0540
study_satisfaction	0.0342

Các kết quả trên phù hợp với giả thuyết ban đầu và các nghiên cứu trong lĩnh vực tâm lý học. Biến `suicidal_thoughts` tiếp tục là yếu tố ảnh hưởng lớn nhất, điều này hoàn toàn hợp lý trong bối cảnh mô hình phân loại trầm cảm. Các yếu tố như `academic_pressure`, `financial_stress`, và đặc biệt là `stress_interaction` – một biến do nhóm nghiên cứu tạo ra từ sự kết hợp giữa áp lực học tập và tài chính – cũng thể hiện rõ vai trò quyết định trong mô hình.

Điều này cho thấy mô hình không chỉ có hiệu suất cao về mặt dự đoán mà còn cung cấp cơ sở đáng tin cậy để nhận diện các yếu tố nguy cơ, hỗ trợ các chương trình tư vấn hoặc sàng lọc tâm lý trong môi trường học đường.

5.4 XGBoost

5.4.1 Hiệu suất mô hình

Sau khi hoàn thiện quá trình xử lý và chuẩn bị dữ liệu, mô hình XGBoost đã được huấn luyện trên tập dữ liệu khảo sát hành vi và trạng thái tâm lý của học sinh - sinh viên nhằm phân loại đầu ra thành hai nhóm: có hoặc không có dấu hiệu trầm cảm.

Kết quả đánh giá cho thấy mô hình đạt độ chính xác tổng thể (Accuracy) là 83%, một con số cho thấy khả năng dự đoán mạnh mẽ và ổn định trên tập kiểm tra. Báo cáo phân loại cho thấy chỉ số F1-score trung bình đạt 0.82, precision và recall đều duy trì ở mức cao đối với cả hai lớp, đặc biệt là lớp trầm cảm.

Bảng 5.7 – Báo cáo hiệu suất mô hình XGBoost

Lớp	Precision	Recall	F1-score	Support
0 (Không trầm cảm)	0.81	0.77	0.79	2312
1 (Trầm cảm)	0.84	0.87	0.86	3261
Accuracy	0.83			
Macro Avg	0.83	0.82	0.82	5573
Weighted Avg	0.83	0.83	0.83	5573

Ma trận nhầm lẫn thu được từ mô hình XGBoost như sau:

$$\begin{bmatrix} 1786 & 526 \\ 420 & 2841 \end{bmatrix}$$

5.4.2 Phân tích và diễn giải kết quả

Mô hình cho thấy hiệu suất đặc biệt tốt ở lớp trầm cảm (positive class), với recall đạt 0.87 – tức là phát hiện chính xác tới 87% các trường hợp có dấu hiệu trầm cảm. Đây là một yếu tố then chốt trong các hệ thống cảnh báo và hỗ trợ sức khỏe tinh thần, nơi việc bỏ sót các trường hợp có nguy cơ cao là điều cần hạn chế tối đa.

Precision của lớp 1 cũng đạt 0.84, nghĩa là trong số các dự đoán là “trầm cảm”, có 84% là chính xác. Độ chênh lệch vừa phải giữa precision và recall cho thấy mô hình duy trì được sự cân bằng giữa phát hiện đúng (true positive) và tránh cảnh báo sai (false positive) – rất quan trọng trong thực tiễn khi áp dụng vào các hệ thống gợi ý can thiệp.

Ngược lại, lớp không trầm cảm có recall là 0.77, tức khoảng 23% trường hợp bị trầm cảm nhẹ có thể bị đánh giá nhầm là bình thường. Điều này phần nào phản ánh bản chất của dữ liệu là hơi mất cân bằng (lớp 1 nhiều hơn lớp 0), đồng thời cũng cho thấy định hướng thiết kế mô hình thiên về đảm bảo không bỏ sót ca bệnh tiềm ẩn.

Tổng thể, mô hình XGBoost cho thấy độ chính xác cao, các chỉ số macro/weighted đều ổn định quanh mức 0.82 – 0.83. Mô hình phù hợp cho các hệ thống hỗ trợ ra quyết định liên quan đến sàng lọc trầm cảm trong trường học, hệ thống tư vấn học đường, hoặc các ứng dụng số trong lĩnh vực chăm sóc sức khỏe tâm thần.

5.5 Naive Bayes

5.5.1 Hiệu suất mô hình

Mô hình Gaussian Naive Bayes được lựa chọn để triển khai trong trường hợp này vì đặc trưng dữ liệu đã được chuẩn hóa hoặc chuẩn tắc hóa (scaling), với phần lớn các biến đầu vào là liên tục và có giá trị âm/dương. Đây là điều kiện phù hợp với giả định phân phối chuẩn (Gaussian) của mô hình.

Sau quá trình huấn luyện, mô hình được đánh giá với kết quả như sau:

Bảng 5.8 – Báo cáo hiệu suất mô hình Gaussian Naive Bayes				
Lớp	Precision	Recall	F1-score	Support
0 (Không trầm cảm)	0.88	0.57	0.70	2312
1 (Trầm cảm)	0.76	0.94	0.84	3261
Accuracy	0.79			
Macro Avg	0.82	0.76	0.77	5573
Weighted Avg	0.81	0.79	0.78	5573

Ma trận nhầm lẫn thu được:

$$\begin{bmatrix} 1321 & 991 \\ 181 & 3080 \end{bmatrix}$$

5.5.2 Phân tích và diễn giải kết quả

Mặc dù Gaussian Naive Bayes là mô hình đơn giản, hiệu suất đạt được vẫn khá ấn tượng – đặc biệt ở lớp trầm cảm (positive class), với recall đạt tới

94%. Điều này cho thấy mô hình có khả năng phát hiện tốt hầu hết các sinh viên có nguy cơ, một yếu tố rất quan trọng trong các ứng dụng hỗ trợ tâm lý, nơi mục tiêu chính là không bỏ sót người thật sự cần hỗ trợ.

Tuy nhiên, precision của lớp 1 chỉ đạt 0.76 tức là trong số các dự đoán là “có trầm cảm”, chỉ khoảng 76% là chính xác. Số lượng báo động sai (false positive) tương đối cao với 991 sinh viên không trầm cảm bị dự đoán nhầm. Điều này có thể chấp nhận được trong bối cảnh phòng ngừa sớm, vì mô hình ưu tiên “báo động để không bỏ sót” hơn là “im lặng để tránh báo động nhầm”.

Trong khi đó, precision của lớp 0 (không trầm cảm) đạt 0.88 nhưng recall lại thấp (0.57), cho thấy mô hình có xu hướng gán nhiều mẫu vào lớp 1 – tức là thiên về cảnh báo.

Tổng thể, với độ chính xác 79%, chỉ số macro F1 0.77 và khả năng phát hiện tốt các ca trầm cảm, mô hình Naive Bayes là lựa chọn đáng cân nhắc trong các hệ thống cần phản hồi nhanh, đơn giản và có định hướng nhạy cảm với lớp dương tính.

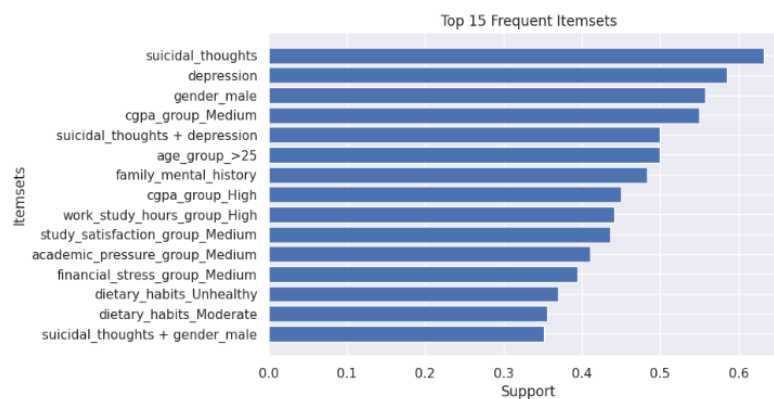
Ghi chú: Lý do lựa chọn GaussianNB

Trong bài toán này, dữ liệu đầu vào đã được chuẩn hóa nên các biến đặc trưng mang giá trị liên tục (có thể âm hoặc dương). Vì vậy:

- **MultinomialNB**: không phù hợp – yêu cầu dữ liệu rời rạc, không âm.
- **BernoulliNB**: chỉ dùng cho dữ liệu nhị phân (0 hoặc 1).
- **GaussianNB**: phù hợp nhất – hỗ trợ dữ liệu liên tục và đã chuẩn hóa.

5.6 Apriori – Khai phá luật kết hợp

5.6.1 Các tập mục phổ biến



Hình 5.3: Phân bố các tập mục phổ biến

Từ biểu đồ phân bố trên, phân tích tần suất trên tập dữ liệu nhị phân hóa cho thấy một số đặc trưng hành vi tâm lý xuất hiện với tần suất cao. Tiêu biểu, biến `suicidal_thoughts` (từng có suy nghĩ tự tử) xuất hiện trong khoảng 64% mẫu, và biến `depression` (có dấu hiệu trầm cảm) chiếm khoảng 59%. Bên cạnh đó, các đặc trưng như `gender_male`, `cgpa_group_Medium`, `family_mental_history`, và `academic_pressure_group_Medium` cũng có tần suất đáng kể, cho thấy đây là các yếu tố phổ biến và tiềm ẩn vai trò đáng kể trong việc hình thành các luật kết hợp.

5.6.2 Các luật dẫn đến trầm cảm

Từ tập các mục phổ biến, thuật toán Apriori đã phát hiện nhiều luật kết hợp có hậu đề là `depression`. Các luật được sắp xếp theo độ hỗ trợ (support), độ tin cậy (confidence), và hệ số nâng (lift) giúp đánh giá mức độ liên kết và ý nghĩa thống kê.

	antecedents	consequents	\
91	(dietary_habits_Unhealthy, academic_pressure_g...	(depression)	
61	(academic_pressure_group_Medium, suicidal_thou...	(depression)	
135	(financial_stress_group_High, academic_pressur...	(depression)	
102	(financial_stress_group_Medium, academic_press...	(depression)	
54	(academic_pressure_group_Medium, suicidal_thou...	(depression)	
84	(academic_pressure_group_Medium, suicidal_thou...	(depression)	
42	(academic_pressure_group_Medium, gender_male, ...	(depression)	
5	(academic_pressure_group_Medium, suicidal_thou...	(depression)	
38	(financial_stress_group_High, suicidal_thoughts)	(depression)	
46	(academic_pressure_group_Medium, cgpa_group_Me...	(depression)	

	support	confidence	lift
91	0.128495	0.949602	1.622722
61	0.143355	0.948018	1.620015
135	0.108646	0.935993	1.599466
102	0.121424	0.935047	1.597850
54	0.145364	0.927198	1.584437
84	0.130828	0.924423	1.579695
42	0.154697	0.919565	1.571393
5	0.284304	0.918803	1.570091
38	0.167187	0.918013	1.568742
46	0.153476	0.914066	1.561996

Hình 5.4: Các luật kết hợp có hậu quả là depression

Một điểm nổi bật là biến `academic_pressure_group_Medium` xuất hiện trong hầu hết các luật có confidence cao nhất (từ 0.91 đến 0.95). Điều này phản ánh rằng mức độ áp lực học tập trung bình không chỉ phổ biến mà còn liên hệ mạnh với tình trạng trầm cảm. Lý giải điều này có thể xuất phát từ việc nhóm này vừa không được hỗ trợ mạnh mẽ như nhóm áp lực cao, vừa không đủ “dư dả tâm lý” như nhóm áp lực thấp tạo nên trạng thái tiềm ẩn rủi ro.

Đáng chú ý, các tổ hợp như:

- (dietary_habits_Unhealthy, academic_pressure_group_Medium)
- (academic_pressure_group_Medium, suicidal_thoughts)
- (financial_stress_group_High, academic_pressure_group_Medium)

đều có chỉ số `lift` > 1.6, cho thấy khả năng xảy ra trầm cảm trong các tổ hợp này cao hơn nhiều so với kỳ vọng nếu các yếu tố độc lập với nhau.

5.6.3 Các luật dẫn đến suy nghĩ tự tử

Tương tự, Apriori cũng khai phá được các luật có hậu đề là `suicidal_thoughts`. Gần như tất cả các luật này đều có sự xuất hiện của `depression` trong điều kiện đầu vào khẳng định rằng trầm cảm là yếu tố trung tâm dẫn đến nguy cơ suy nghĩ tự tử.

	antecedents	consequents
143	(depression, study_satisfaction_group_Medium, ...	(suicidal_thoughts)
132	(depression, financial_stress_group_Medium, wo...	(suicidal_thoughts)
134	(depression, age_group_>25, work_study_hours_g...	(suicidal_thoughts)
118	(depression, cgpa_group_Medium, financial_stre...	(suicidal_thoughts)
69	(depression, cgpa_group_Medium, work_study_hou...	(suicidal_thoughts)
130	(depression, cgpa_group_Medium, study_satisfac...	(suicidal_thoughts)
12	(depression, work_study_hours_group_High)	(suicidal_thoughts)
80	(depression, family_mental_history, work_study...	(suicidal_thoughts)
148	(depression, study_satisfaction_group_Medium, ...	(suicidal_thoughts)
74	(depression, sleep_adequate)	(suicidal_thoughts)

	support	confidence	lift
143	0.106565	0.871697	1.377949
132	0.111267	0.871276	1.377283
134	0.110046	0.868063	1.372205
118	0.115287	0.865302	1.367840
69	0.137396	0.864889	1.367188
130	0.111482	0.863978	1.365747
12	0.257528	0.863729	1.365354
80	0.131510	0.863133	1.364411
148	0.103550	0.862997	1.364197
74	0.135063	0.862875	1.364004

Hình 5.5: Các luật kết hợp có hậu quả là `suicidal_thoughts`

Ngoài ra, các biến như `study_satisfaction_group_Medium`, `sleep_adequate` = 0, hoặc thời gian học/làm việc cao cũng xuất hiện đồng thời, phản ánh các yếu tố hành vi học tập, đời sống ảnh hưởng đáng kể đến trạng thái tâm lý tiêu cực.

Một luật tiêu biểu:

- (depression, study_satisfaction_group_Medium, academic_pressure_group_Medium) \Rightarrow suicidal_thoughts

Confidence = 0.8717, Lift = 1.3779

Điều này cho thấy rằng sinh viên đang trầm cảm, đồng thời không hài lòng

với việc học và chịu áp lực ở mức trung bình, có nguy cơ cao hơn rất nhiều trong việc hình thành ý định tự tử.

5.6.4 Các luật bảo vệ ($\text{Lift} < 1$)

Bên cạnh các luật có xu hướng tiêu cực, thuật toán Apriori cũng phát hiện được các luật có $\text{lift} < 1$, tức là những mối liên hệ nghịch chiều – khi xuất hiện đồng thời sẽ làm giảm xác suất xảy ra trầm cảm. Phần lớn các luật này liên quan đến nhóm tuổi trên 25 ($\text{age_group_}>25$), thể hiện vai trò "bảo vệ" của yếu tố tuổi tác.

Khi kết hợp với các yếu tố như hài lòng học tập mức trung bình, không có áp lực tài chính nghiêm trọng, hoặc không ăn uống kém, xác suất trầm cảm được ghi nhận là giảm đáng kể. Những phát hiện này có giá trị thực tiễn trong việc phân tầng nguy cơ và định hướng chính sách can thiệp.

5.6.5 Diễn giải tổng thể

Tổng kết lại, các luật kết hợp được khai phá từ tập dữ liệu cho thấy trầm cảm ở sinh viên không phải là kết quả của một yếu tố đơn lẻ mà là sự tổng hợp giữa nhiều điều kiện tâm lý – hành vi. Những yếu tố như áp lực học tập trung bình, suy nghĩ tiêu cực, căng thẳng tài chính, chất lượng giấc ngủ, hoặc sự không hài lòng với việc học đều đóng vai trò then chốt trong các luật có lift cao.

Ngược lại, các yếu tố như tuổi lớn hơn 25, kết quả học tập ổn định, hoặc sự hài lòng tương đối với môi trường học lại mang vai trò bảo vệ.

Kết quả này không chỉ xác thực hiệu quả của khai phá luật kết hợp trong phân tích dữ liệu giáo dục về sức khỏe tâm thần, mà còn làm cơ sở xây dựng hệ thống cảnh báo và hỗ trợ chính sách dựa trên dữ liệu.

5.7 Phân cụm – KMeans và Agglomerative Clustering

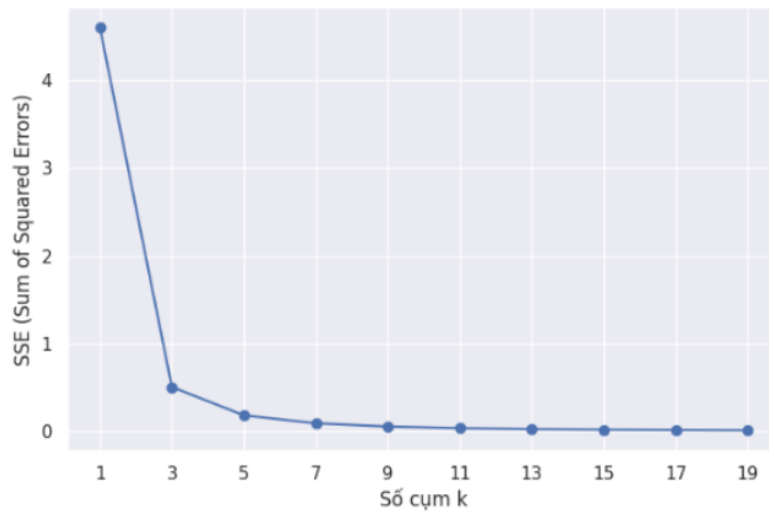
5.7.1 KMeans

Xác định số cụm K

Việc xác định số cụm K trong bài toán phân cụm là một bước quan trọng, giúp tối ưu hóa độ đồng nhất trong mỗi cụm cũng như sự khác biệt giữa các cụm. Nếu K quá nhỏ, các đối tượng không đồng nhất có thể bị “gộp” vào cùng cụm, làm mất đi tính đa dạng. Ngược lại, nếu chọn K quá lớn, dữ liệu có thể bị chia nhỏ quá mức, dẫn đến việc tạo ra các cụm không thực sự tồn tại trong thực tế (overfitting).

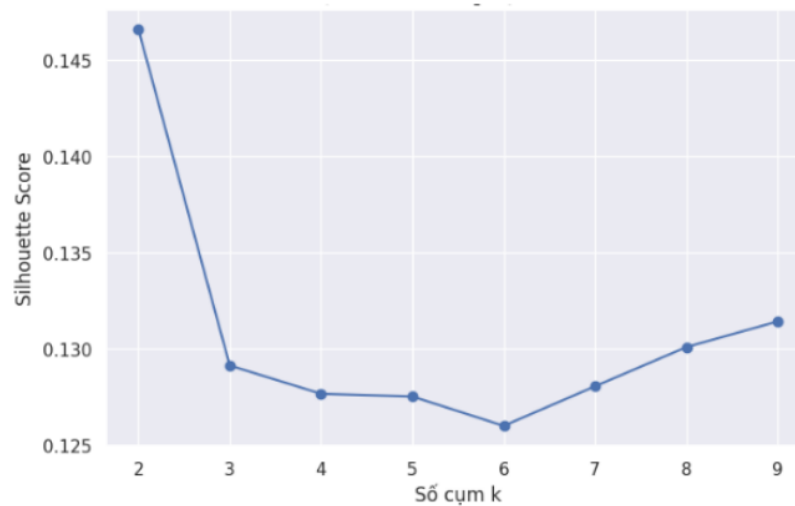
Trong nghiên cứu này, hai phương pháp phổ biến được áp dụng để xác định giá trị K tối ưu:

- **Phương pháp “Khuỷu tay” (Elbow Method)** – đánh giá sự giảm dần của tổng sai số bình phương (SSE) theo từng giá trị K.
- **Hệ số Silhouette** – đo lường mức độ tách biệt giữa các cụm, giá trị càng gần 1 càng tốt.



Hình 5.6: Xác định số cụm K bằng phương pháp Elbow

Từ biểu đồ Elbow, có thể nhận thấy sau $K = 3$, đường cong bắt đầu “phẳng”, cho thấy việc tăng số cụm không còn giúp cải thiện rõ rệt SSE, gợi ý rằng 3 là giá trị K hợp lý. Tuy nhiên, khi đánh giá bằng Silhouette Score:

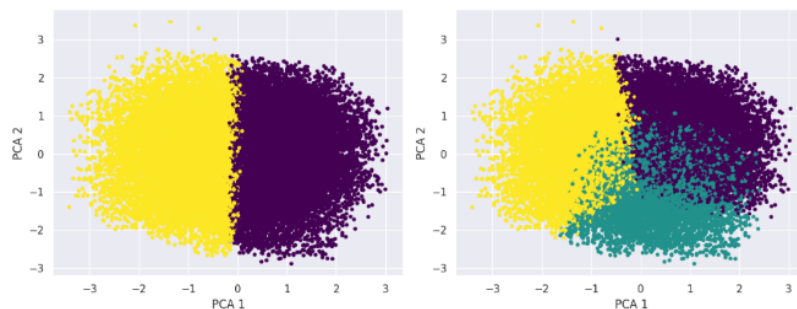


Hình 5.7: Silhouette Score theo từng giá trị K

Giá trị Silhouette Score đạt cao nhất tại $K = 2$, gợi ý rằng mô hình với hai cụm phản ánh cấu trúc dữ liệu tốt hơn so với ba cụm. Do đó, hai giá trị $K = 2$ và $K = 3$ đều được thử nghiệm để trực quan hóa và so sánh.

Trực quan hóa và đánh giá kết quả phân cụm

Dữ liệu sau khi giảm chiều bằng PCA được trực quan hóa để đánh giá hiệu quả phân cụm với $K = 2$ và $K = 3$:



Hình 5.8: Phân cụm KMeans với $K = 2$ (trái) và $K = 3$ (phải)

Quan sát biểu đồ phân cụm:

- Với $K = 2$: Dữ liệu được chia thành hai cụm tương đối cân bằng, ranh giới nằm chủ yếu trên trục PCA1. Hai cụm tách biệt rõ ràng, ít điểm chồng lấn.
- Với $K = 3$: Xuất hiện thêm một cụm mới (màu vàng), tuy nhiên hai cụm còn lại (xanh và tím) có ranh giới mờ nhạt hơn, chồng lấn khá nhiều.

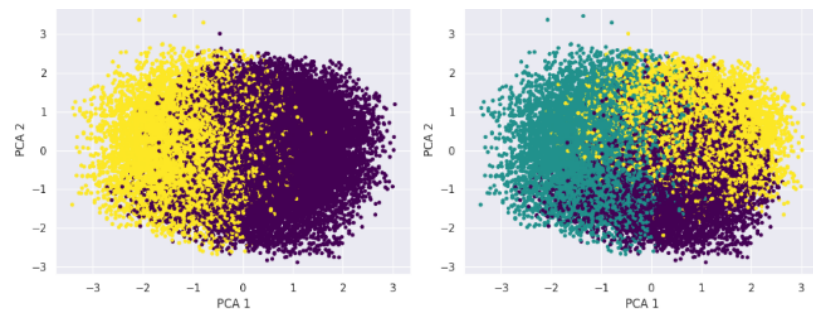
Tóm lại: Với dữ liệu hiện tại, mô hình KMeans với $K = 2$ phản ánh rõ hai nhóm đối tượng chính trong dữ liệu. Việc ép thành ba cụm có thể khiến một trong hai nhóm lớn bị chia nhỏ mà không mang lại phân tách thực sự có ý nghĩa.

Mặc dù KMeans không cung cấp nhãn rõ ràng để gán ý nghĩa cho từng cụm, nhưng kỹ thuật này hữu ích trong việc khám phá cấu trúc ẩn, từ đó giúp các nhà nghiên cứu hiểu thêm về sự phân bố của các nhóm sinh viên theo đặc trưng hành vi và tâm lý.

5.7.2 Agglomerative Clustering

Agglomerative Clustering là một kỹ thuật phân cụm phân cấp (hierarchical clustering), thực hiện theo hướng “gộp dần từ dưới lên” (bottom-up). Mỗi điểm dữ liệu ban đầu được xem là một cụm riêng lẻ, sau đó các cụm gần nhau nhất được kết hợp lặp lại cho đến khi đạt số cụm mong muốn.

Trong nghiên cứu này, thuật toán được áp dụng với hai giá trị cụm là $k = 2$ và $k = 3$ để trực quan hóa và đánh giá khả năng tách cụm của dữ liệu. Dữ liệu đã được giảm chiều bằng PCA để dễ biểu diễn trong không gian 2 chiều.



Hình 5.9: Agglomerative Clustering với $k = 2$ (trái) và $k = 3$ (phải)

Phân tích kết quả

- **Cấu trúc cụm không rõ rệt:** Các cụm thu được từ Agglomerative Clustering không có ranh giới rõ ràng, phần lớn các điểm nằm chồng lấn, thể hiện khả năng tách cụm yếu.
- **Với $k = 2$:** Dữ liệu được chia thành hai nhóm tương đối cân bằng, ranh giới phân tách chủ yếu theo trục PCA1. Tuy nhiên, vẫn có nhiều điểm nằm ở vùng giao nhau giữa hai cụm, dẫn đến phân tách thiếu sắc nét.
- **Với $k = 3$:** Xuất hiện thêm một cụm mới, nhưng ba cụm không tách biệt rõ. Một số điểm của các cụm nằm xen kẽ nhau, cho thấy dữ liệu có

độ phân tán cao và khó phân loại chính xác bằng cách phân cụm đơn thuần.

- **Tính phân cụm yếu:** Kết quả này phù hợp với chỉ số Silhouette thấp trong phân tích trước, cho thấy dữ liệu không có cấu trúc phân cụm tự nhiên rõ ràng mà phù hợp hơn với mô hình học có giám sát.

Nhận định

Mặc dù Agglomerative Clustering là một thuật toán mạnh trong nhiều trường hợp dữ liệu có cấu trúc phân cấp rõ rệt, song đối với tập dữ liệu này, khả năng phân biệt giữa các nhóm sinh viên không đạt hiệu quả cao. Nguyên nhân có thể đến từ sự tương đồng trong các đặc trưng hành vi và tâm lý giữa các đối tượng, khiến thuật toán khó xác định ranh giới cụm rõ ràng. Do đó, mô hình học có giám sát vẫn là lựa chọn phù hợp hơn để dự đoán trạng thái tâm lý trong trường hợp này.

5.8 Tổng kết hiệu suất các mô hình

Sau khi triển khai và đánh giá năm mô hình học máy khác nhau trên tập dữ liệu đã xử lý, bảng sau đây tổng hợp các chỉ số chính: Accuracy, Precision, Recall và F1-score trung bình (macro) của từng mô hình:

Bảng 5.9 – So sánh hiệu suất giữa các mô hình

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)
Logistic Regression	0.8394	0.8340	0.8383	0.8357
Decision Tree	0.8215	0.8165	0.8150	0.8157
XGBoost	0.8303	0.8267	0.8218	0.8239
GaussianNB	0.7897	0.8180	0.7579	0.7664
Random Forest	0.8355	0.8323	0.8270	0.8293

Đánh giá tổng hợp

Logistic Regression đạt kết quả cao nhất về độ chính xác (Accuracy = 0.8394) và F1-score (0.8357), đồng thời có giá trị Precision và Recall cao và ổn định (lần lượt là 0.8340 và 0.8383). Điều này cho thấy Logistic Regression là mô hình hiệu quả trong việc phân biệt hai nhóm đánh giá tốt và tệ dựa trên các đặc trưng đầu vào.

Random Forest là mô hình xếp thứ hai với độ chính xác 0.8355 và F1-score 0.8293, chứng minh khả năng tổng quát hóa tốt và cân bằng giữa Precision (0.8323) và Recall (0.8270). Với ưu thế về khả năng xử lý dữ liệu phức tạp và chống overfitting, Random Forest là lựa chọn tiềm năng bên cạnh Logistic Regression.

XGBoost đạt Accuracy 0.8303 và F1-score 0.8239. Tuy thấp hơn Logistic Regression và Random Forest nhưng vẫn nằm trong nhóm mô hình có hiệu suất cao, đồng thời cho thấy khả năng tối ưu tốt nhờ các kỹ thuật boosting.

Decision Tree tuy đơn giản nhưng cho kết quả tương đối ổn định với Accuracy 0.8215 và F1-score 0.8157. Tuy nhiên, mô hình này có thể gặp vấn đề quá khớp khi áp dụng trên dữ liệu mới.

Gaussian Naive Bayes là mô hình có hiệu suất thấp nhất trong số các mô hình được đánh giá, với Accuracy chỉ đạt 0.7897 và F1-score 0.7664. Nguyên nhân chủ yếu đến từ giả định độc lập giữa các đặc trưng đầu vào mà mô hình này yêu cầu, điều không phù hợp với bản chất dữ liệu thực tế.

Đánh giá tổng quan mô hình phân cụm

Ngoài các mô hình phân loại, nghiên cứu còn áp dụng hai kỹ thuật phân cụm không giám sát là KMeans và Agglomerative Clustering để khám phá cấu

trúc tiềm ẩn trong dữ liệu.

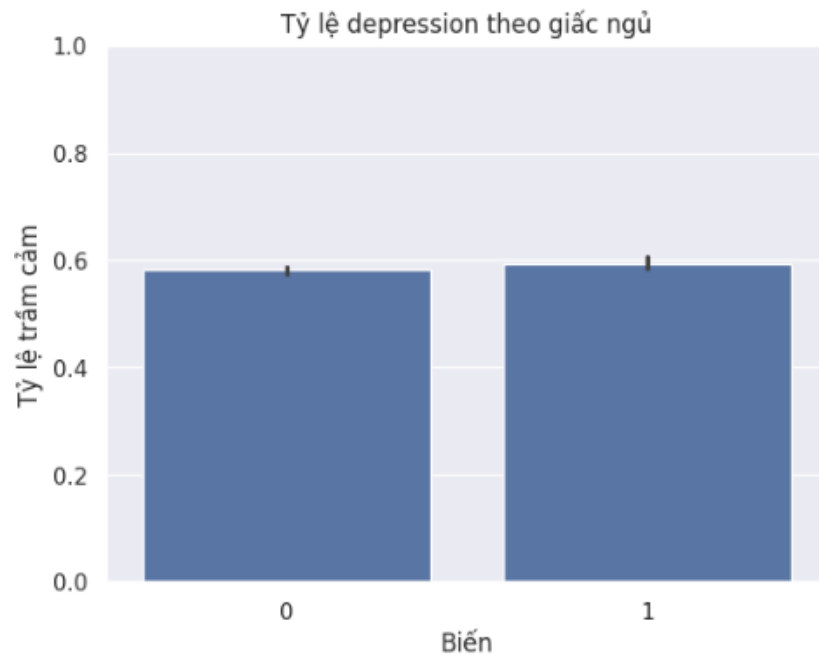
Kết quả cho thấy:

- **KMeans**: với $k = 2$, dữ liệu được phân chia thành hai cụm tương đối rõ ràng theo PCA1, Silhouette Score đạt giá trị cao nhất, thể hiện khả năng phân tách khá tốt. Tuy nhiên, mô hình không cung cấp nhãn cụ thể, và việc giải thích cụm còn hạn chế.
- **Agglomerative Clustering**: phân cụm cho thấy các nhóm bị chồng lấn nhiều, không có biên ranh rõ, phù hợp với việc Silhouette Score thấp. Điều này cho thấy cấu trúc phân cấp không rõ ràng trong dữ liệu hiện tại.

Nhìn chung, kỹ thuật phân cụm hỗ trợ khám phá cấu trúc tiềm ẩn nhưng không đạt hiệu quả cao như các mô hình phân loại có giám sát trong bài toán này.

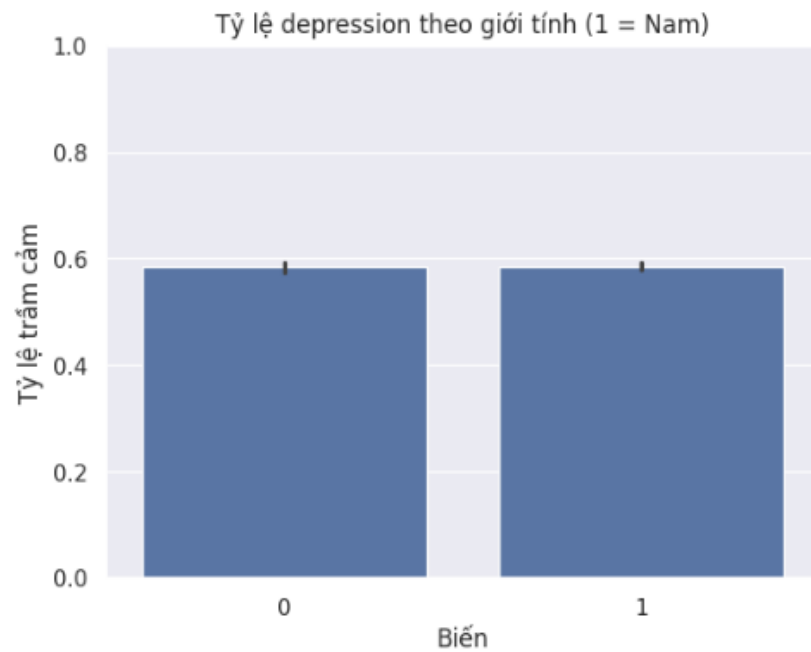
5.9 Phân tích mối quan hệ giữa các biến

Trong phần này, các biểu đồ được xây dựng nhằm trực quan hóa mối quan hệ giữa các yếu tố hành vi, xã hội và học tập với tỷ lệ trầm cảm của sinh viên.



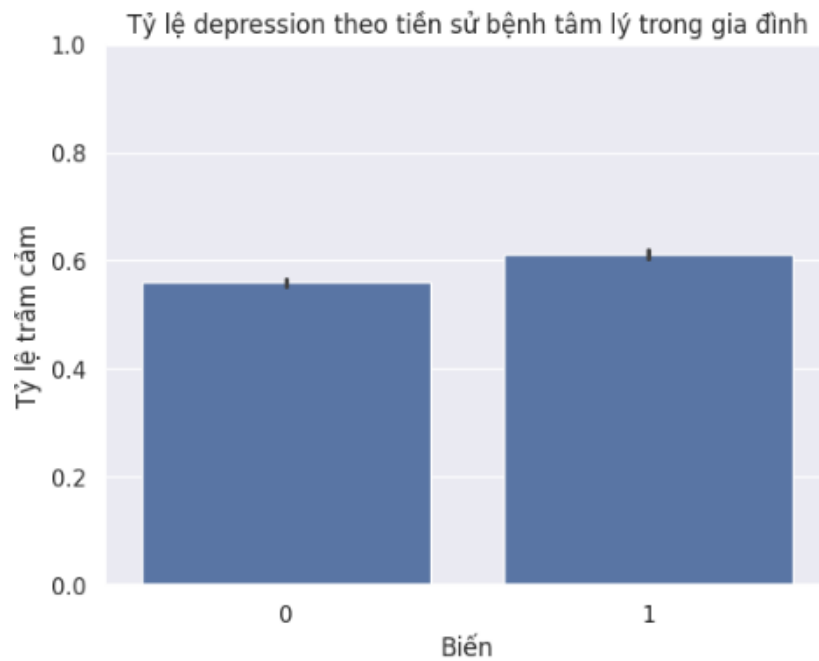
Hình 5.10: Tỷ lệ trầm cảm theo giấc ngủ

- **Giấc ngủ (sleep_adequate):** Nhóm có giấc ngủ không đầy đủ (0) và đầy đủ (1) có tỷ lệ trầm cảm gần tương đương nhau, cho thấy giấc ngủ có thể chưa phải yếu tố phân biệt rõ trong bộ dữ liệu này.



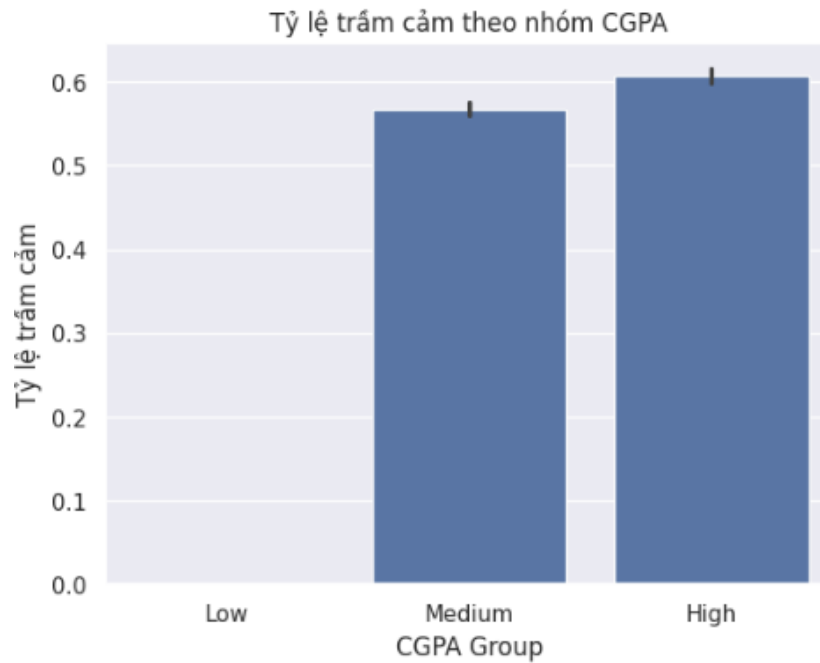
Hình 5.11: Tỷ lệ trầm cảm theo giới tính

- **Giới tính (gender):** Tỷ lệ trầm cảm giữa nam và nữ gần tương đương, không có sự chênh lệch rõ rệt.



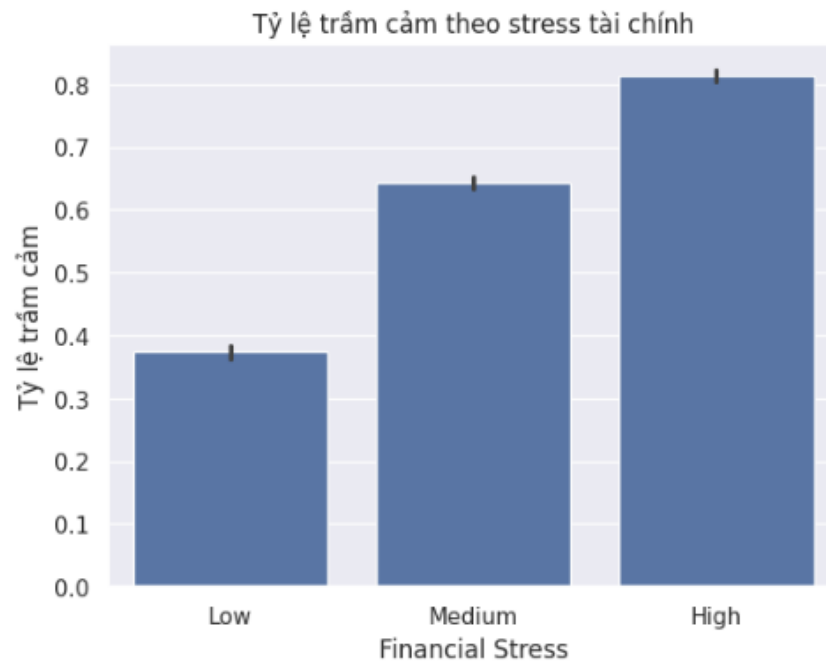
Hình 5.12: Tỷ lệ trầm cảm theo tiền sử bệnh tâm lý

- **Tiền sử bệnh tâm lý gia đình:** Nhóm có người thân từng mắc bệnh tâm lý có tỷ lệ trầm cảm cao hơn đáng kể.



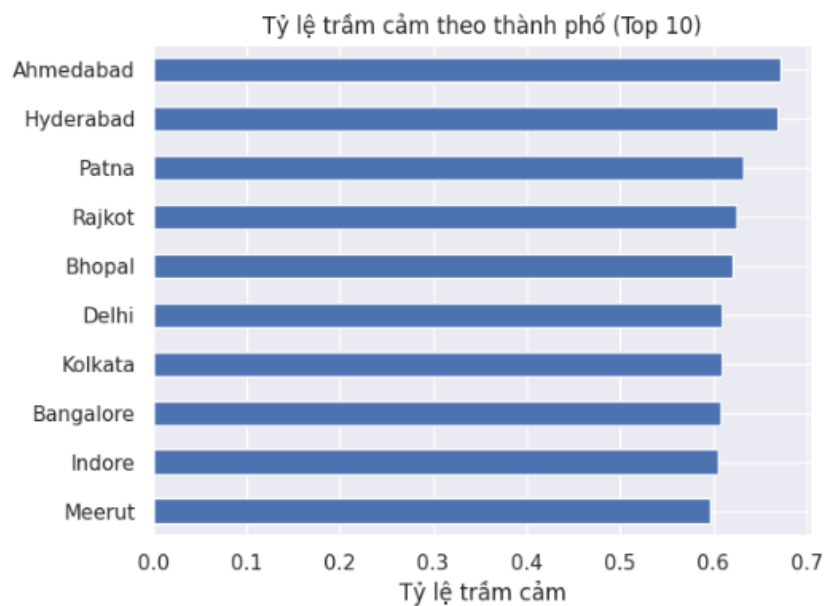
Hình 5.13: Tỷ lệ trầm cảm theo nhóm CGPA

- **Điểm CGPA:** CGPA càng cao, tỷ lệ trầm cảm có xu hướng tăng nhẹ, phản ánh áp lực học tập có thể ảnh hưởng đến tâm lý.



Hình 5.14: Tỷ lệ trầm cảm theo mức stress tài chính

- **Căng thẳng tài chính:** Nhóm có mức stress cao có tỷ lệ trầm cảm lên đến hơn 80%, cho thấy yếu tố tài chính là một trong những yếu tố nguy cơ mạnh.



Hình 5.15: Tỷ lệ trầm cảm theo thành phố

- **Thành phố:** Một số thành phố như Ahmedabad, Hyderabad, Patna có tỷ lệ trầm cảm cao hơn trung bình, gợi ý ảnh hưởng của điều kiện sống/áp lực khu vực.

5.10 Kiểm định giả thuyết thống kê

Để kiểm định mối liên hệ giữa các yếu tố và trầm cảm, nhóm thực hiện các kiểm định thống kê sử dụng kiểm định Chi-square (χ^2) với mức ý nghĩa $\alpha = 0.05$.

Bảng 5.10 – Kết quả kiểm định Chi-Square cho các giả thuyết

Giả thuyết	Phép kiểm	p-value	Kết luận
H1: sleep_duration vs depression	Chi-square	0.018	Có ảnh hưởng
H2: city vs depression	Chi-square	1.6×10^{-20}	Có khác biệt
H3: suicidal_thoughts vs depression	Chi-square	< 0.00001	Có liên quan
H4: dietary_habits vs depression	Chi-square	0.0082	Có tương quan

Giả thuyết kiểm định và kết quả

- **H1:** Thời gian ngủ có ảnh hưởng đáng kể đến tỷ lệ trầm cảm. ($p = 0.0184 < 0.05$) \Rightarrow Bác bỏ H_0 .
- **H2:** Có sự khác biệt về tỷ lệ trầm cảm giữa các thành phố. ($p = 1.6 \times 10^{-20}$) \Rightarrow Bác bỏ H_0 .
- **H3:** Suy nghĩ tiêu cực có liên quan đến trầm cảm. ($p < 0.00001$) \Rightarrow Bác bỏ H_0 .
- **H4:** Thói quen ăn uống có tương quan đến tình trạng trầm cảm. ($p = 0.0082$) \Rightarrow Bác bỏ H_0 .

Diễn giải

- Các kiểm định thống kê khẳng định những yếu tố hành vi và cảm xúc như suy nghĩ tiêu cực, căng thẳng tài chính, hay giấc ngủ đều có liên quan đến tỷ lệ trầm cảm.
- Kết quả phù hợp với phân tích luật kết hợp và mô hình học máy, củng cố độ tin cậy trong phát hiện các yếu tố rủi ro.

Chương 6

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Tổng kết nội dung nghiên cứu

Trong đề tài này, nhóm nghiên cứu đã tiến hành khai phá và phân tích bộ dữ liệu khảo sát sức khỏe tâm thần sinh viên với mục tiêu nhận diện nguy cơ trầm cảm dựa trên các yếu tố hành vi, học tập và tâm lý. Quá trình thực hiện bắt đầu từ bước tiền xử lý dữ liệu, bao gồm chuẩn hóa tên cột, xử lý giá trị thiếu, mã hóa các biến phân loại, đến việc áp dụng các mô hình học máy nhằm phân loại và nhận diện các yếu tố nguy cơ.

Cụ thể, nhóm đã triển khai và so sánh năm mô hình học máy bao gồm Logistic Regression, Decision Tree, Random Forest, XGBoost và Gaussian Naive Bayes. Đồng thời, để hiểu sâu hơn về mối quan hệ giữa các đặc trưng, thuật toán khai phá luật kết hợp (Apriori) được sử dụng nhằm phát hiện những tổ hợp biến có khả năng dẫn đến trầm cảm hoặc suy nghĩ tiêu cực. Bên cạnh đó, hai thuật toán phân cụm K-Means và Agglomerative Clustering cũng được áp dụng nhằm tìm kiếm các nhóm sinh viên có đặc điểm tương

đồng. Cuối cùng, các giả thuyết về mối quan hệ giữa các biến và tình trạng trầm cảm được kiểm định thông qua phương pháp thống kê Chi-Square.

Kết quả cho thấy:

- Mô hình Logistic Regression và Random Forest đạt hiệu suất tốt nhất với độ chính xác trên 83%, F1-score cao và ổn định giữa hai lớp.
- Các yếu tố có ảnh hưởng mạnh đến trầm cảm bao gồm: suy nghĩ tiêu cực, áp lực học tập, căng thẳng tài chính, điểm CGPA, thời gian học tập.
- Các luật kết hợp trích xuất từ Apriori cho thấy nguy cơ trầm cảm thường đến từ sự kết hợp của nhiều yếu tố rủi ro hơn là một yếu tố đơn lẻ.
- Phân cụm không mang lại hiệu quả phân nhóm cao do dữ liệu không tách rời rõ ràng.
- Các kiểm định thống kê đã xác nhận vai trò của một số yếu tố quan trọng với p-value rất nhỏ.

6.2 Hạn chế của đề tài

Bên cạnh các kết quả tích cực, đề tài vẫn tồn tại một số hạn chế. Trước hết, bộ dữ liệu khảo sát có sự mất cân bằng đáng kể giữa các nhóm thuộc tính như giới tính, khu vực cư trú hoặc ngành học, điều này có thể ảnh hưởng đến độ khái quát hóa của mô hình khi áp dụng cho các nhóm ít đại diện. Thứ hai, phần lớn đặc trưng trong bộ dữ liệu là biến phân loại đơn giản, chưa khai thác được các thang đo định lượng chi tiết như mức độ stress theo thang điểm, độ hài lòng học tập theo thang Likert, hoặc số giờ học/ngủ thực tế. Việc thiếu các đặc trưng sâu như vậy làm giảm khả năng phản ánh chính xác hành vi và trạng thái tâm lý của sinh viên.

Ngoài ra, các mô hình được sử dụng trong nghiên cứu chủ yếu thuộc nhóm tuyến tính và cây quyết định, tuy dễ diễn giải nhưng chưa thể hiện được sức mạnh học phức tạp của dữ liệu. Các mô hình hiện đại như mạng học sâu (Deep Neural Networks), SVM hoặc mô hình kết hợp phức tạp chưa được triển khai. Cuối cùng, dữ liệu khảo sát chỉ được thu thập tại một thời điểm, chưa phản ánh được tiến trình thay đổi tâm lý của sinh viên theo thời gian, điều này hạn chế khả năng áp dụng các phân tích theo chiều thời gian hoặc dự báo trong tương lai.

6.3 Hướng phát triển

Để khắc phục những hạn chế nêu trên và nâng cao chất lượng phân tích trong tương lai, nhóm đề xuất một số hướng phát triển như sau. Thứ nhất, mở rộng tập dữ liệu cả về quy mô và chiều sâu bằng cách tích hợp thêm các đặc trưng về môi trường học, mối quan hệ xã hội, lịch sử học tập, hoặc thông tin cá nhân liên quan đến sức khỏe tinh thần. Thứ hai, áp dụng các mô hình học máy nâng cao như mạng học sâu, SVM, hoặc các kỹ thuật ensemble có khả năng học phi tuyến và biểu diễn đặc trưng mạnh mẽ hơn.

Bên cạnh đó, việc xây dựng một hệ thống cảnh báo nguy cơ trầm cảm có thể mang lại giá trị thực tiễn cao, hỗ trợ nhà trường và các tổ chức giáo dục trong việc chăm sóc sức khỏe tinh thần cho sinh viên. Một hướng đi khác là tích hợp thêm phân tích cảm xúc từ các nguồn dữ liệu văn bản như phản hồi học tập, email, hoặc bài đăng mạng xã hội, từ đó giúp mô hình có cái nhìn toàn diện hơn về trạng thái tâm lý của người học. Cuối cùng, nếu có thể mở rộng sang khai thác dữ liệu theo thời gian thực hoặc chuỗi thời gian, các mô hình có thể dự đoán sớm và phát hiện các dấu hiệu chuyển biến tiêu cực từ giai đoạn đầu.

TÀI LIỆU THAM KHẢO

- [1] Ikram Harouni. The modern methods of data analysis in social research: Python programming language and its pandas library as an example—a theoretic study. *Social Empowerment Journal*, 6(1):56–70, 2024.
- [2] Pranali Dhawas et al. Big data analysis techniques: data preprocessing techniques, data mining techniques, machine learning algorithm, visualization. In *Big Data Analytics Techniques for Market Intelligence*, pages 183–208. IGI Global Scientific Publishing, 2024.
- [3] Fred Torres Cruz, Evelyn Eliana Coaquira Flores, and Sebastian Jarom Condori Quispe. Prediction of depression status in college students using a naive bayes classifier based machine learning model. *arXiv preprint arXiv:2307.14371*, 2023.
- [4] Xinqiao Liu and Jingxuan Wang. Depression, anxiety, and student satisfaction with university life among college students: a cross-lagged study. *Humanities and Social Sciences Communications*, 11, 2024.
- [5] Malik Muhammad Qirtas et al. The relationship between loneliness and depression among college students: Mining data derived from passive sensing. *arXiv preprint arXiv:2308.15509*, 2023.

PHÂN CÔNG CÔNG VIỆC NHÓM

Bảng phân công nhiệm vụ của nhóm

MSSV	Họ tên phụ trách	Nội dung công việc	Ghi chú
2251050038	Nguyễn Vĩ Khang	Xử lý đặc trưng (Data Transforming), huấn luyện mô hình, viết nội dung, trình bày báo cáo	100%
2251050008	Trương Thái Bảo	Khám phá và trực quan hóa dữ liệu (EDA & Visualization), huấn luyện mô hình, viết nội dung	100%
2251050025	Thạch Nhật Hào	Làm sạch dữ liệu (Data Cleaning), huấn luyện mô hình, viết nội dung	100%