**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN, ĐHQG-HCM**

**KHOA KHOA HỌC MÁY TÍNH**



**LẬP TRÌNH PYTHON CHO MÁY HỌC**

**ĐỀ TÀI: Phát triển và Đánh giá Mô hình Học máy cho Dự đoán Nguy cơ Trầm cảm**

**Môn học: CS116.P22 – Lập trình Python cho máy học**

**Giảng viên hướng dẫn: Nguyễn Vinh Tiệp**

**Thành viên nhóm 4:**

| 1. Dương Tấn Lộc | MSSV: 23520854 | Trưởng nhóm |
| --- | --- | --- |
| 1. Nguyễn Việt Hùng | MSSV: 23520571 | Thành viên |
| 1. Nguyễn Việt Nguyên | MSSV: 23521054 | Thành viên |

Tp. Thủ Đức 6/2025

**MỤC LỤC**

[I. GIỚI THIỆU 3](#_mvtzl6ic0kxg)

[II. BỘ DỮ LIỆU & TIỀN XỬ LÝ 3](#_f78ssky6e6gu)

[A. Mô tả bộ dữ liệu 3](#_ecfq26xuvakc)

[B. Các điểm nổi bật của Exploratory Data Analysis (EDA) 4](#_j1n64tr0l9j9)

[1. Mất cân bằng lớp 4](#_9mldl1509dh3)

[2. Giá trị khuyết có cấu trúc 5](#_wkd8bdvua5ul)

[3. Mối quan hệ giữa Gender và Depression 5](#_xf25tvfpexjq)

[4. Mối quan hệ giữa CGPA và Depression 6](#_xzje6fllgdaf)

[5. Mối quan hệ giữa Age và Depression 7](#_5u7nhabrymhl)

[6. Mối quan hệ giữa Sleep Duration và Depression 8](#_6fh7h063fvzl)

[7. Mối quan hệ giữa Dietary Habits và Depression 9](#_5r4udf3l8513)

[8. Nhận định từ các chỉ số tự đánh giá việc học tập, làm việc 10](#_2kc3poc6g6lk)

[9. Một vài thách thức với các biến phân loại đa giá trị 12](#_6yu6qq7stl6k)

[C. Các bước tiền xử lý 15](#_bwy3h2608p6k)

[D. Feature Engineering 16](#_w0a06edcaqy6)

[Loại bỏ các đặc trưng không cần thiết: các biến định danh như ‘Id’ và ‘Name’ 16](#_3pdc9g3rrnc)

[1. Ma trận tương quan 17](#)

[2. Statistical test 18](#)

[3. Model-based methods 19](#)

[III. HUẤN LUYỆN VÀ TINH CHỈNH MÔ HÌNH 24](#_ucnmlo6fq78e)

[A. Chia tập dữ liệu 24](#_yx91xfkpbfar)

[B. Chọn mô hình cho bài toán 24](#_1plwpeg43ad6)

[C. Tinh chỉnh mô hình 24](#_gixnxrfittlb)

[IV. KẾT QUẢ VÀ ĐÁNH GIÁ 25](#_33ztkh27y52w)

[A. Metrics 25](#_vug3qj1wanun)

[B. Kết quả và phân tích 26](#_few7fbecnb4n)

[1. Đánh giá trên test set 26](#_nrlq9udxk1oq)

[V. KẾT LUẬN 32](#_vt9ke2n0z3yy)

# **I. GIỚI THIỆU**

Rối loạn sức khỏe tâm thần là một vấn đề đang ngày càng gia tăng trên phạm vi toàn cầu, ảnh hưởng nghiêm trọng đến con người, cộng đồng và nền kinh tế trên toàn thế giới. Các tình trạng như trầm cảm, lo âu và rối loạn lưỡng cực có tác động lớn đến cuộc sống hàng ngày và có thể dẫn đến những khó khăn lâu dài. Việc tìm kiếm và chẩn đoán sớm những tình trạng này là rất quan trọng. Nó cho phép mọi người nhận được sự giúp đỡ và điều trị kịp thời, điều này có thể ngăn ngừa những vấn đề nghiêm trọng hơn trong tương lai.

Các phương pháp chẩn đoán truyền thống thường phụ thuộc vào cách đánh giá cá nhân của bác sĩ và báo cáo của chính bệnh nhân. Những phương pháp này có thể tốn rất nhiều thời gian và đôi khi dẫn đến sự không nhất quán. Vì các tình trạng sức khỏe tâm thần có thể đến rồi đi, và mọi người mô tả các triệu chứng của họ theo những cách khác nhau, nên cần có những cách nhanh chóng để kịp thời phát hiện tình trạng sức khoẻ cho bệnh nhân đồng thời có thể mở rộng và dựa trên dữ liệu để phát hiện sớm các vấn đề.

Mục tiêu: Bài báo cáo này hướng đến xây dựng và đánh giá các mô hình học máy nhằm dự đoán các tình trạng sức khỏe tâm thần, sử dụng bộ dữ liệu công khai từ Kaggle có tiêu đề "Exploring Mental Health Data". Mục tiêu chính là xác định các đặc điểm hành vi và nhân khẩu học có thể là dấu hiệu tiềm ẩn của các vấn đề sức khỏe tâm thần.

# **II. BỘ DỮ LIỆU & TIỀN XỬ LÝ**

## **A. Mô tả bộ dữ liệu**

Bộ dữ liệu ‘Exploring Mental Health Data’ là bộ dữ liệu phục vụ cho việc dự đoán sức khỏe tâm thần. Dữ liệu bao gồm một loạt các đặc trưng đa dạng, bao gồm thông tin nhân khẩu học, các triệu chứng do người được khảo sát báo cáo, các yếu tố lối sống và các đặc điểm khác có liên quan đến sức khỏe tinh thần. Bộ dữ liệu như vậy thường chứa một hỗn hợp dữ liệu số, phân loại và đôi khi là dữ liệu văn bản, phản ánh bản chất nhiều mặt của các đánh giá sức khỏe tâm thần.

Bộ dữ liệu được sử dụng trong đồ án này chủ yếu dựa trên các phản hồi khảo sát và thông tin tự khai báo. Loại dữ liệu này mang đến những hạn chế nhất định, vì nó bao gồm các thành kiến cá nhân và góc nhìn chủ quan. Thay vì phản ánh các chỉ số khách quan về mặt thể chất hoặc lâm sàng, dữ liệu chủ yếu thể hiện cách mà mỗi cá nhân cảm nhận về cảm xúc và hành vi xã hội của bản thân. Do đó, việc tiền xử lý cẩn thận là rất cần thiết để quản lý sự không nhất quán và cải thiện chất lượng dữ liệu phục vụ cho huấn luyện mô hình.

Bộ dữ liệu bao gồm 19 feature columns và 1 target column, được phân loại như sau:

* Thông tin định danh: id, Name
* Các biến đánh giá bản thân (được đánh giá theo thang điểm từ 1 đến 5):
  + Academic Pressure: Áp lực học tập
  + Work Pressure: Áp lực công việc
  + Study Satisfaction: Mức độ hài lòng với việc học
  + Job Satisfaction: Mức độ hài lòng với công việc
  + Financial Stress: Căng thẳng tài chính
* Biến số:
  + Age: Tuổi
  + CGPA: Điểm trung bình tích luỹ
  + Work/Study Hours: Số giờ làm việc/ học
* Biến phân loại nhị phân:
  + Gender: Giới tính
  + Working Professional or Student: Người đi làm hay học sinh/sinh viên
  + Have you ever had suicidal thoughts?: Đã từng có ý nghĩ tự tử hay chưa?
  + Family History of Mental Illness: Tiền sử gia đình có người mắc bệnh tâm lí
* Biến phân loại đa giá trị:
  + City: Thành phố cư trú
  + Profession: Nghề nghiệp
  + Sleep Duration: Thời gian ngủ
  + Dietary Habits: Thói quen ăn uống
  + Degree: Bằng cấp chuyên môn
* Biến mục tiêu: Depression: Trầm cảm (binary: 0 = Không trầm cảm, 1 = Trầm cảm)

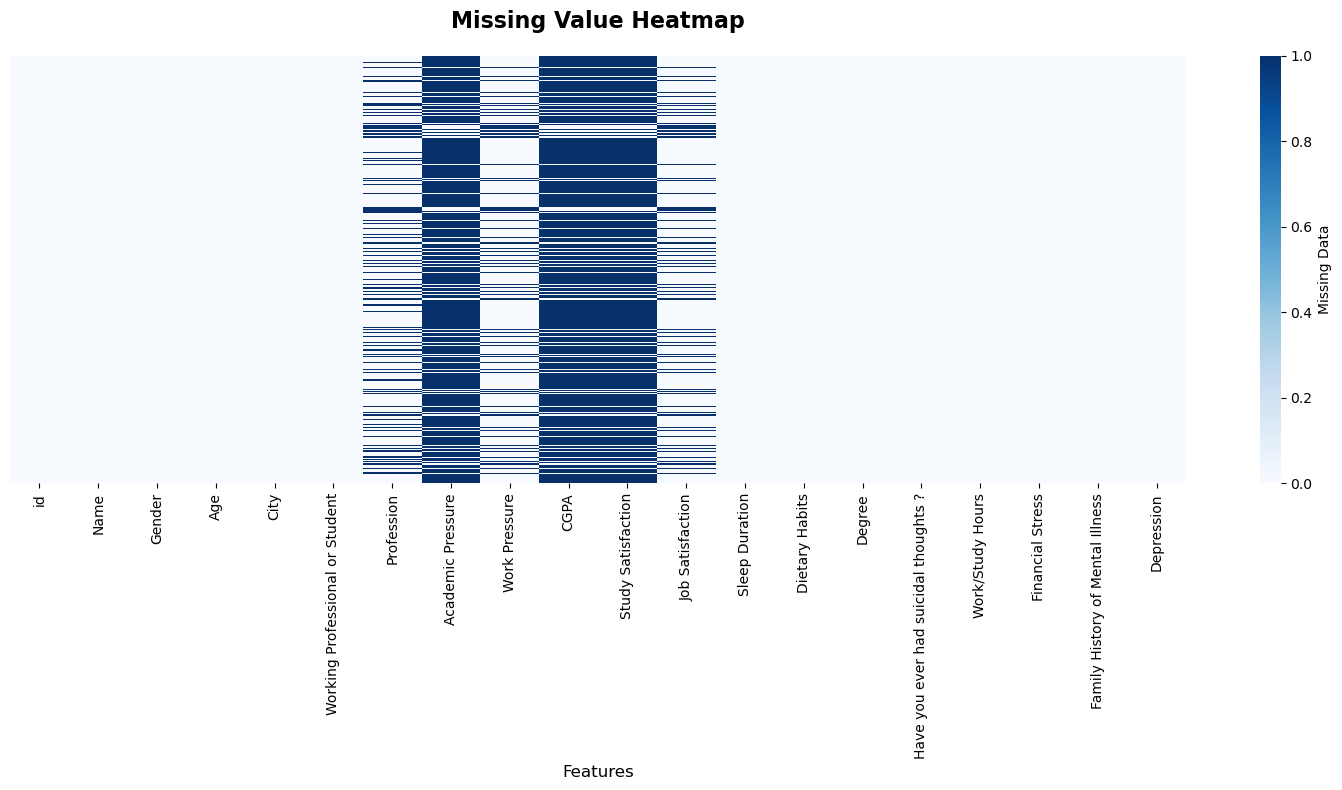
## **B. Các điểm nổi bật của Exploratory Data Analysis (EDA)**

### **Mất cân bằng lớp**

Một quan sát đáng chú ý là sự mất cân bằng rõ rệt giữa hai lớp trong biến mục tiêu. Sự mất cân bằng này sẽ cần được giải quyết trong quá trình huấn luyện mô hình để ngăn mô hình bị thiên lệch (biased) về phía lớp đa số và có khả năng hoạt động kém trên lớp thiểu số. Việc mất cân bằng này phản ánh đúng thực tế – tỷ lệ người mắc trầm cảm thường thấp hơn số không mắc, đặc biệt trong các khảo sát dân số.

### **Giá trị khuyết có cấu trúc**

Khi trực quan hóa các giá trị null bằng heatmap, một pattern rõ ràng và có cấu trúc đã xuất hiện: nếu một cá nhân được xác định là "student", họ luôn thiếu thông tin liên quan đến "professional", "job satisfaction" và "working pressure". Ngược lại, những cá nhân không phải là sinh viên (hàm ý rằng họ đang làm việc) thì không có thông tin gì về tình trạng sinh viên.



*Hình 2: Heatmap các giá trị Null cho thấy sự thiếu dữ liệu có cấu trúc*

Điều này cho thấy một mối quan hệ loại trừ lẫn nhau (mutually exclusive) giữa hai nhóm biến này, trong đó sự hiện diện của thông tin trong một tập hợp đồng nghĩa với sự vắng mặt ở tập hợp kia. Cấu trúc này là một phát hiện có giá trị và sẽ ảnh hưởng đến cách các giá trị bị thiếu này được xử lý trong giai đoạn tạo đặc trưng (feature engineering).

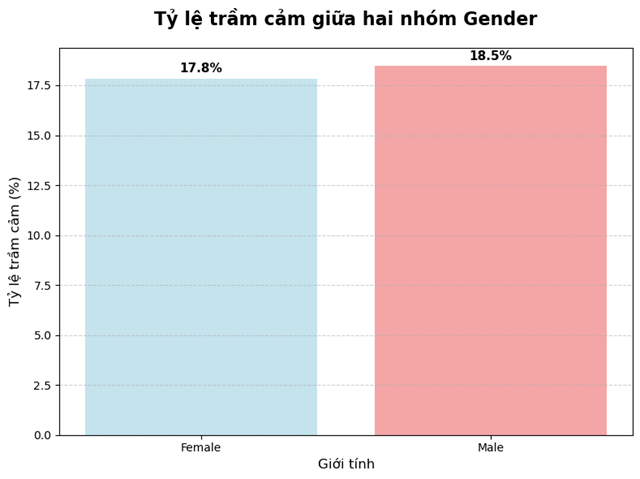
### **Mối quan hệ giữa Gender và Depression**

Dù thực tế đã có một vài nghiên cứu cho thấy nữ giới có khả năng mắc trầm cảm cao gấp đôi so với nam giới do các yếu tố:

* Sinh học: sự thay đổi hormon trong chu kỳ kinh nguyệt
* Xã hội: Phụ nữ có thể phải đối mặt với áp lực xã hội, công việc và gia đình cao hơn, dẫn đến căng thẳng và ảnh hưởng đến sức khỏe tâm thần.

Tuy nhiên bộ dữ liệu này không phản ánh được xu hướng ấy. Có thể là do các vấn đề về thu thập, khai báo, phân bố mẫu hoặc cách định nghĩa nhãn.

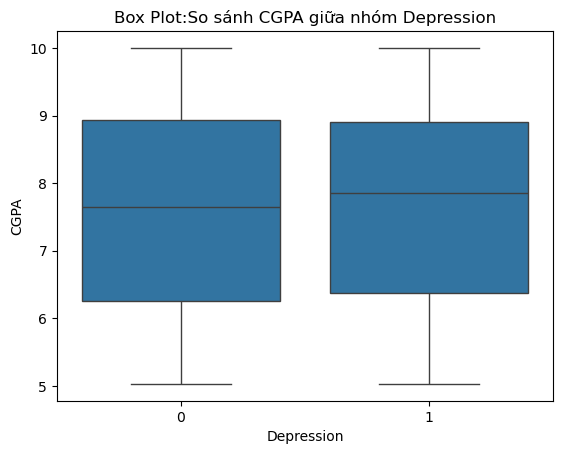
Theo phân tích, tỉ lệ trầm cảm của nam giới và nữ giới chệnh lệch không đáng kể. Điều này cho thấy Nam và Nữ có thể không phải là yếu tố quyết định việc mắc trầm cảm trong bộ dữ liệu này.



*Hình 3: Tỉ lệ trầm cảm dựa trên giới tính*

### **Mối quan hệ giữa CGPA và Depression**

Thông thường, những sinh viên có điểm trung bình học tập cao hơn sẽ có áp lực về học tập nhiều hơn so với sinh viên điểm thấp. Nhưng đối đối với dữ liệu của tập dữ liệu này xu hướng đó cũng không được thể hiện.



*Hình 4: Boxplot so sánh điểm trung bình tích luỹ giữa hai nhóm 'Trầm cảm' và 'Không trầm cảm'*

Phân bố của điểm trung bình tích luỹ trải từ khoảng 6 tới 9 điểm ở cả 2 class 0 và 1, không cho thấy sự phân tán hay trung vị khác biệt đáng kể giữa nhóm có trầm cảm và không trầm cảm.

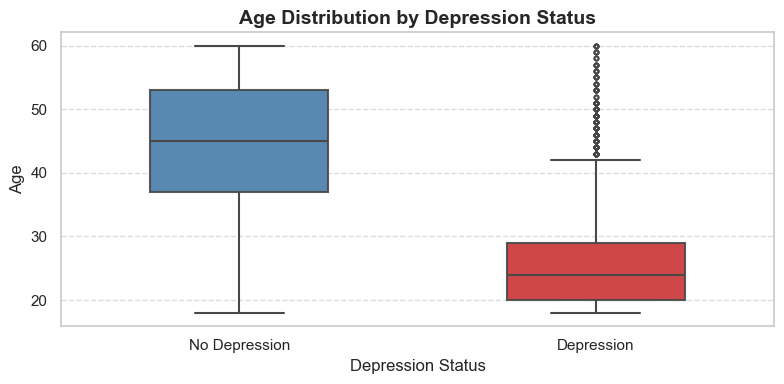
Từ điều đó có thể suy luận rằng CGPA không phải là yếu tố chính hoặc không có mối liên hệ mạnh đối với tình trạng trầm cảm trong bộ dữ liệu này.

### **Mối quan hệ giữa Age và Depression**

Trong số các biến số, Age cho thấy một tác động đáng kể đến trầm cảm, tiết lộ một pattern đáng chú ý. Các cá nhân được phân loại là "Depression" dường như trẻ hơn trung bình so với những người trong nhóm "No Depression". Biểu đồ hộp cho thấy một sự phân tán rộng hơn về độ tuổi trong nhóm "No Depression", trong khi nhóm "Depression" có độ tuổi tập trung chủ yếu ở độ tuổi trong khoảng từ 20 tới 30, với một số outliers ở độ tuổi lớn hơn.

Thực tế, dễ dàng nhận thấy được ở độ tuổi trung niên 40 tới trên 50 tuổi lúc mà con người đã ổn định được việc làm, con cái của họ đã có thể tự lập, thì họ đã không cần phải lo nghĩ quá nhiều điều nữa tức tỉ lệ trầm cảm không cao.

Trong khi đó, những người ở độ tuổi khoảng từ 20 tới 30, do có nhiều áp lực nên thường mắc trầm cảm hơn. Thực tế trong độ tuổi này có khá nhiều áp lực đến từ việc học tập do môi trường đại học vào các năm cuối thường sẽ có nhiều đồ án cùng với việc thực tập, ngoài ra đối với những người không còn học thì có các áp lực từ tìm kiếm việc làm, áp lực tài chính do không còn phụ thuộc tài chính vào bố mẹ. Tất cả các nguyên nhân trên có thể dẫn đến lý do cho việc phân bố trầm cảm tập trung chủ yếu từ 20 tới 30 tuổi.



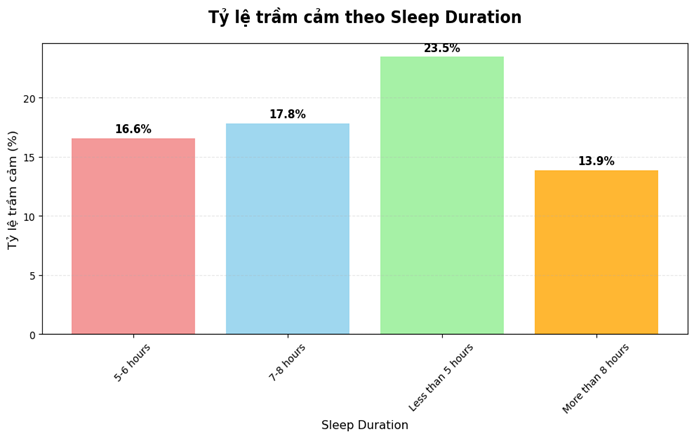
*Hình 5: Phân bố độ tuổi theo trạng thái trầm cảm cho thấy xu hướng những cá nhân trầm cảm thường trẻ hơn đáng kể*

Điều này cho thấy rằng tuổi tác có thể là một yếu tố quan trọng trong việc dự đoán tình trạng sức khỏe tâm thần.

### **Mối quan hệ giữa Sleep Duration và Depression**

Thực tế cho thấy việc thiếu ngủ dù do khách quan hay chủ quan thì đều dẫn đến những hệ luỵ nghiêm trọng về sức khoẻ, trong đó có sức khoẻ tinh thần. Cụ thể, rất nhiều nghiên cứu khoa học đã chứng minh việc thời lượng ngủ quá ít dẫn đến thiếu ngủ, thiếu ngủ có thể gây ra nhiều ảnh hưởng tiêu cực đến sức khỏe tinh thần, bao gồm thay đổi tâm trạng, tăng nguy cơ mắc các bệnh tâm lý như trầm cảm và lo âu.

Đối chiếu với phân tích dữ liệu thời lượng ngủ của bộ dữ liệu, rõ ràng tỷ lệ trầm cảm của thời lượng ngủ ít hơn 5 tiếng là cao nhất, còn ngủ hơn 8 tiếng tỷ lệ trầm cảm là thấp nhất.



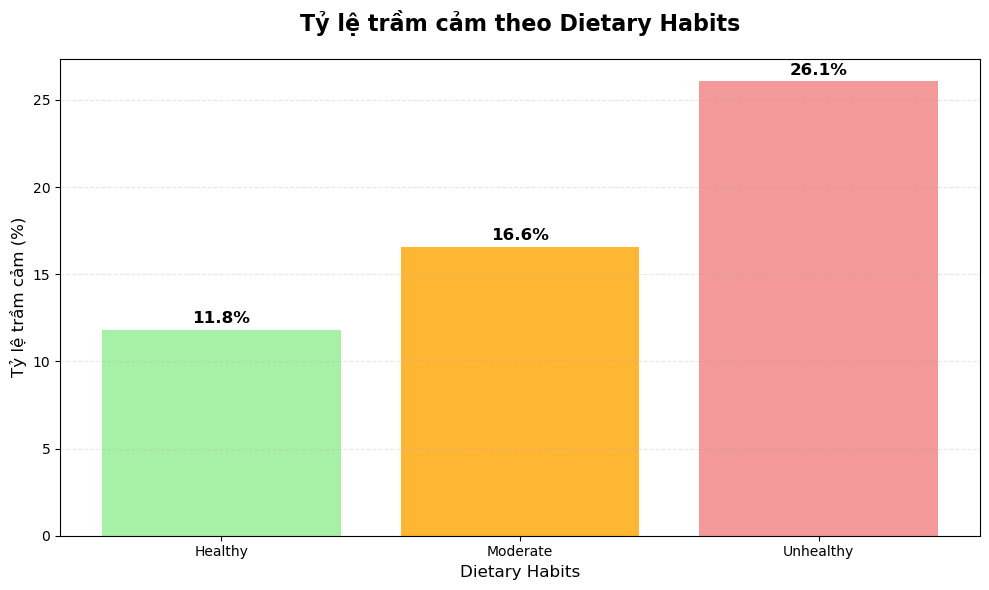
*Hình 6: Tỉ lệ trầm cảm dựa trên thời gian ngủ*

Với phân tích trên, thời lượng ngủ có thể là một yếu tố có ảnh hưởng tới việc dự đoán tình trạng trầm cảm.

### **Mối quan hệ giữa Dietary Habits và Depression**

Giống với Sleep Duration, Dietary Habits cũng là một trong những thói quen sống của con người. Việc thói quen sống không lành mạnh sẽ trực tiếp ảnh hưởng đến cả sức khoẻ thể chất và tinh thần. Cụ thể việc ăn uống không lành mạnh (bao gồm ăn uống với giờ giấc thất thường, ăn uống thiếu chất,..) sẽ làm tăng nguy cơ trầm cảm, lo âu. Việc thiếu hụt các chất dinh dưỡng thiết yếu có thể làm suy yếu hệ thần kinh và ảnh hưởng đến quá trình sản xuất các chất dẫn truyền thần kinh, từ đó tác động đến tâm trạng và cảm xúc.

Với dữ liệu từ bộ dữ liệu, rõ ràng việc ăn uống không lành mạnh (Unhealthy) dẫn đầu trong tỉ lệ trầm cảm, tiếp sau là ăn uống bình thường (Moderate) và tỉ lệ trầm cảm thấp nhất là ăn uống lành mạnh (Healthy).

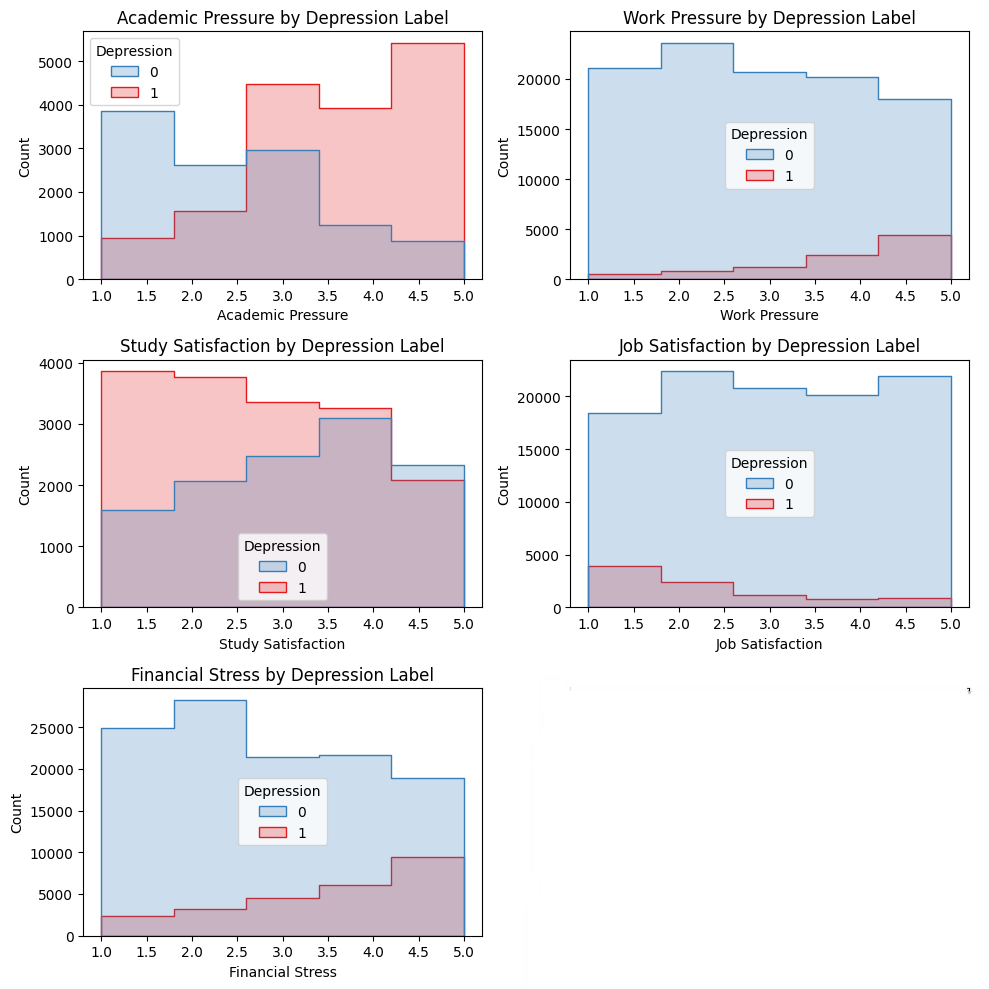


*Hình 7: Tỉ lệ trầm cảm dựa trên thói quen ăn uống*

Từ đó thấy được ăn uống càng thiếu lành mạnh thì khả năng trầm cảm càng cao. Thói quen ăn uống có thể là một yếu tố quan trọng cho việc xác định trầm cảm.

### **Nhận định từ các chỉ số tự đánh giá việc học tập, làm việc**

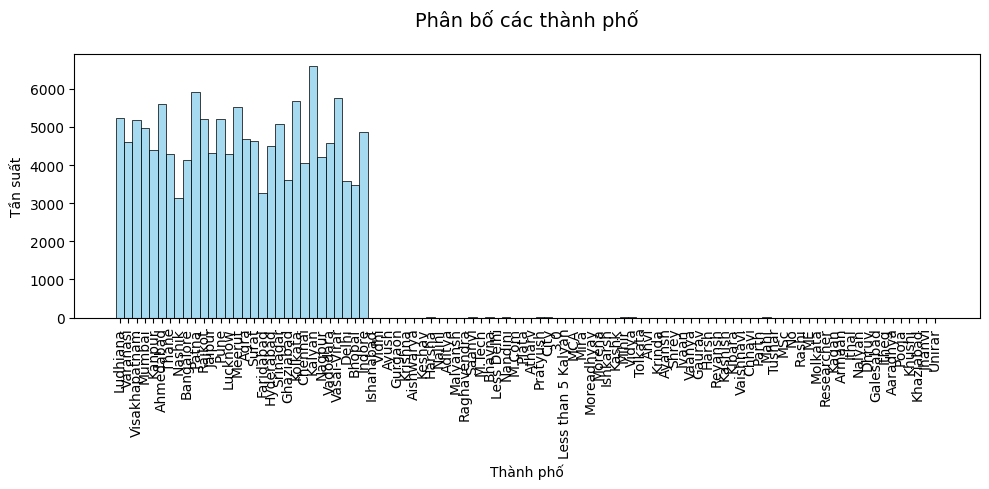
Việc phân tích các chỉ số tự đánh giá bao gồm: ‘Academic Pressure’, ‘Work Pressure’, ‘Study Satisfaction’, ‘Job Satisfaction’ và ‘Financial Stress’ (mỗi yếu tố được tính điểm từ 1 đến 5), cho thấy một xu hư rõ rệt. Thông qua các biểu đồ histogram được phân loại theo trạng thái trầm cảm, có thể quan sát thấy rằng các cá nhân thể hiện "điểm số tích cực" (có nghĩa là điểm áp lực thấp hơn và điểm hài lòng cao hơn) có xu hướng có khả năng bị trầm cảm thấp hơn.



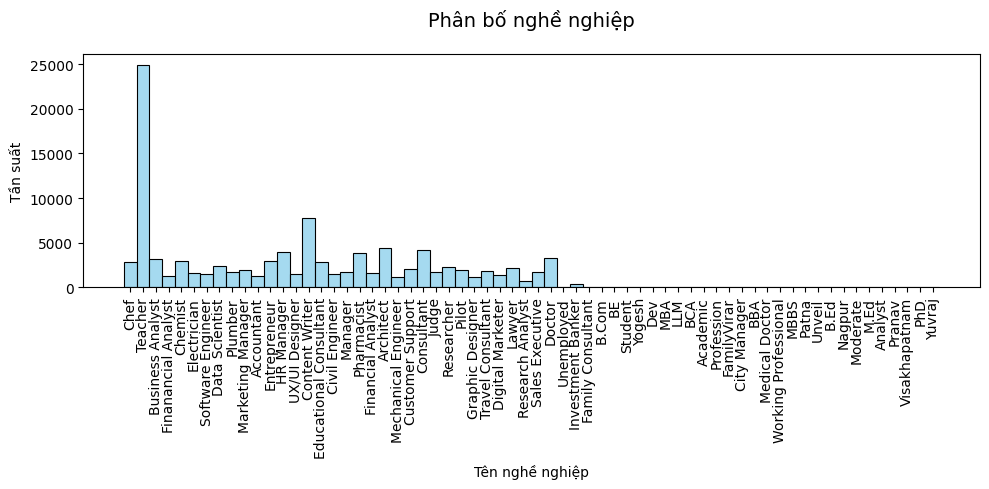
*Hình 8: Histogram của các chỉ số tự đánh giá theo trạng thái trầm cảm cung cấp bằng chứng trực quan cho mối quan hệ này*

Điều này phản ánh mối liên hệ hợp lý giữa cảm nhận chủ quan về căng thẳng/hài lòng và nguy cơ trầm cảm, và là yếu tố đáng lưu ý trong việc phân tích và xây dựng mô hình.

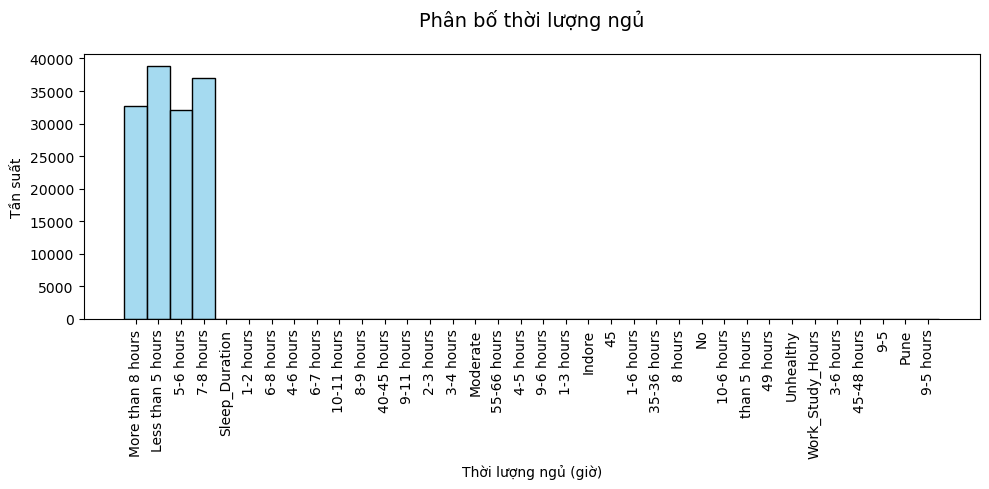
### **Một vài thách thức với các biến phân loại đa giá trị**



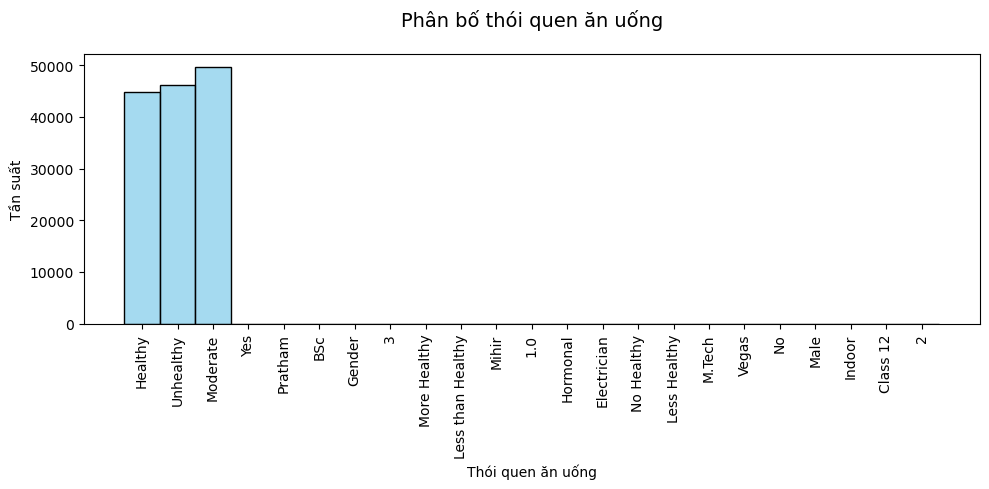
*Hình 9: Tần suất xuất hiện của các thành phố trong tập dữ liệu huấn luyện*



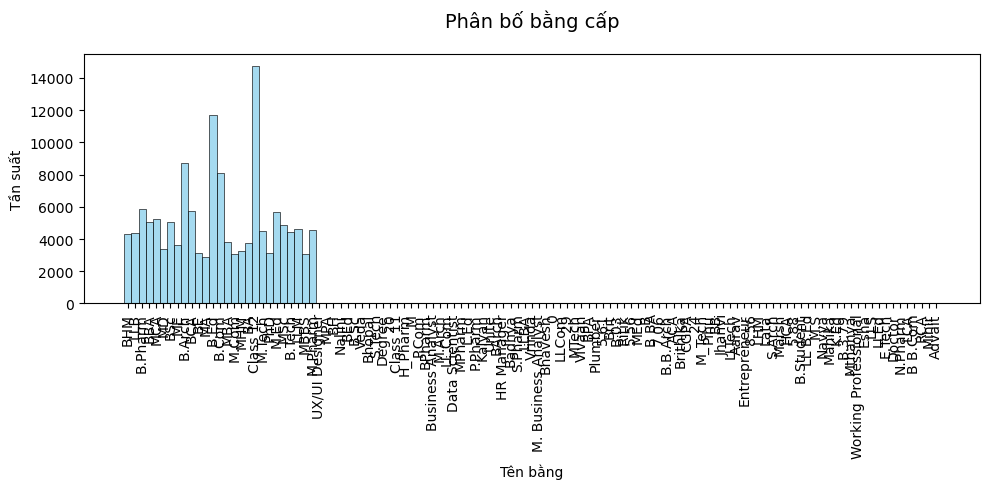
*Hình 10: Tần suất xuất hiện của các nghề nghiệp trong tập dữ liệu huấn luyện*



*Hình 11: Tần suất xuất hiện của các thời lượng ngủ (giờ) trong tập dữ liệu huấn luyện*



*Hình 12: Tần suất xuất hiện của các thói quen ăn uống trong tập dữ liệu huấn luyện*



*Hình 13: Tần suất xuất hiện của các bằng cấp trong tập dữ liệu huấn luyện*

Một số biến multi-category, đặc biệt là ‘City’ (Hình 9), ‘Profession’ (Hình 10), ‘Sleep Duration’ (Hình 11), ‘Dietary Habits’ (Hình 12) và ‘Degree’ (Hình 13), có các vấn đề đáng kể về chất lượng dữ liệu. Các vấn đề này bao gồm:

* Tên gọi không đồng nhất cho cùng một ý nghĩa: Các categories tồn tại với các tên khác nhau nhưng truyền tải cùng một ý nghĩa (ví dụ: ‘Less than Healthy’ và ‘Less Healthy’ trong ‘Dietary Habits’, hoặc các cách viết khác nhau của "B.Tech" như ‘B.Tech’, ‘B.T’ trong ‘Degree’).
* Danh mục không phù hợp: Một số categories rõ ràng là xuất hiện sai ngữ cảnh hoặc không liên quan tới biến (ví dụ: ‘Unhealthy’, ‘Work\_Study\_Hours’, ‘Pune’ xuất hiện trong ‘Sleep Duration’).
* Danh mục hiếm: Một số categories chỉ xuất hiện một lần, không có ý nghĩa phân tích rõ ràng và có thể gây nhiễu cho mô hình.

Trên phương diện thực tế các vấn đề trên xảy ra trong bộ dữ liệu có thể do các lí do:

* Nhập liệu thủ công không kiểm soát: người tham gia khảo sát nhập thông tin tự do, không giới hạn lựa chọn có sẵn dẫn đó sự không nhất quán trong cách viết.
* Thiếu chuẩn hoá và kiểm tra ràng buộc dữ liệu: hệ thống đầu vào không áp dụng quy tắc kiểm tra tính hợp lệ cho từng trường thông tin, do đó các giá trị không đúng ngữ cảnh có thể được ghi nhận.
* Thiếu tập danh mục chuẩn: không có danh sách các nhãn định nghĩa trước cho từng biến phân loại, dẫn đến nhiều cách biểu diễn khác nhau cho cùng một ý tưởng.
* Tập dữ liệu được tổng hợp từ nhiều nguồn khác nhau: Dữ liệu có thể được thu thập từ nhiều lần khảo sát, nhiều biểu mẫu, hoặc nhà cung cấp khác nhau, mỗi bên sử dụng chuẩn đầu vào khác nhau, gây ra sự phân tán và trùng lặp.
* Danh mục hiếm do tính cá nhân hóa quá mức hoặc lỗi đánh máy: Một số người tham gia có thể nhập giá trị duy nhất, hoặc sai chính tả, dẫn đến các nhãn chỉ xuất hiện một lần duy nhất gây nhiễu.

Nhìn chung những vấn đề này phần lớn là hậu quả của quy trình thu thập dữ liệu không được kiểm soát chặt chẽ, cần phải có các biện pháp tiền xử lý phù hợp để giải quyết.

## **C. Các bước tiền xử lý**

1. **Xử lý Null**

Xử lý Null của ‘CGPA’: Do tỷ lệ giá trị khuyết lên tới 80,2%, cột CGPA được giữ nguyên thay vì nội suy hay loại bỏ, vì cả hai lựa chọn đều không phù hợp trong bối cảnh dữ liệu này.

Xử lý các giá trị Null có cấu trúc: Các đặc trưng mới đã được tạo ra nhằm phản ánh toàn diện hơn trạng thái sức khoẻ tinh thần bao gồm:

* Aca/Work Satisfaction: kết hợp giữa Academic Satisfaction và Job Satisfaction.
* Aca/Work Pressure: kết hợp giữa Study Pressure và Work Pressure.

1. Xử lý Outlier

Như đã phân tích trên phần EDA, các feature có tính chất đặc biệt như ‘Degree’, ‘City’, ‘Profession’, ‘Dietary Habits’ và ‘Sleep Duration’ được xử lý bằng cách: các categories có ý nghĩa tương tự được mapping với một biểu diễn thống nhất duy nhất. Ngoài ra, các categories hiếm khi xảy ra hoặc các giá trị null trên các features khác nhau đã được thay thế bằng Null để đảm bảo tính đồng nhất và cải thiện chất lượng.

1. **Encoding**

Chuyển đổi biến ‘Sleep Duration’ thành dạng biến số: Biến ‘Sleep Duration’ đã được chuyển đổi từ categorical thành numerical để cho phép mô hình diễn giải theo cách liên tục và tận dụng tốt hơn các mối quan hệ tuyết tính tiềm ẩn.

Mã hoá cho ‘Dietary Habits’: biến ‘Dietary Habits’ gồm ba mức Unhealthy, Moderate, Healthy đã được ordinally encoded để bảo toàn bản chất xếp hạng vốn có của nó.

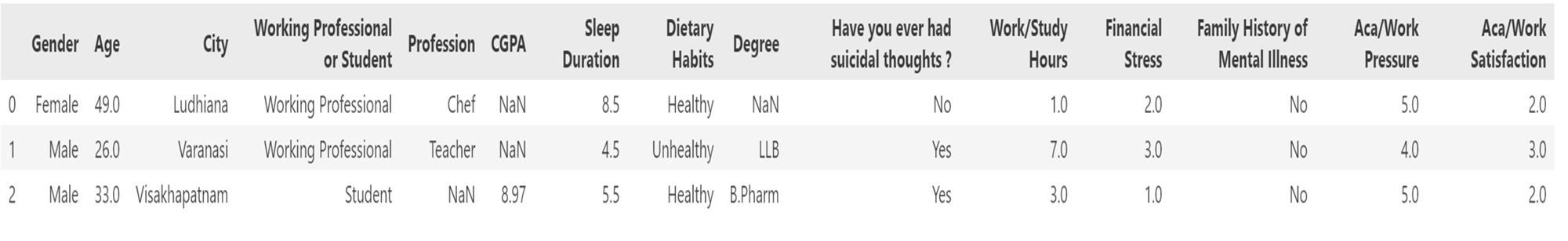
1. **Các phương pháp phục vụ cho các chiến lược feature engineering**

Nhóm tạo ra 2 bộ dữ liệu để phân tích:

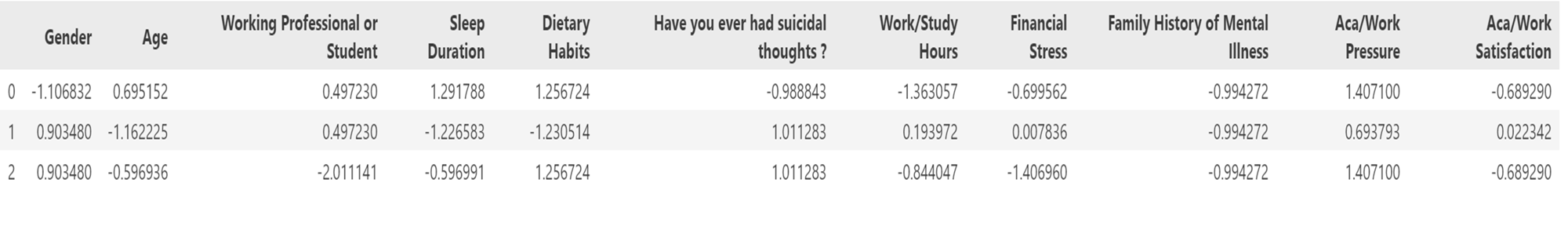
* Bộ dữ liệu hỗn hợp dữ liệu gốc X1: Cho phương pháp model-based có thể xử lí Categorical data không cần encoding.
* Bộ dữ liệu số X2 (bỏ những đặc trưng không thể encoding): được encoding và normalize để sử dụng cho 2 mô hình Logistic và Decision Tree của feature engineering.

Bộ dữ liệu gốc vẫn là bộ để huấn luyện mô hình.

X1 có kích thước (140656, 15)



X2 có kích thước (140656, 11)

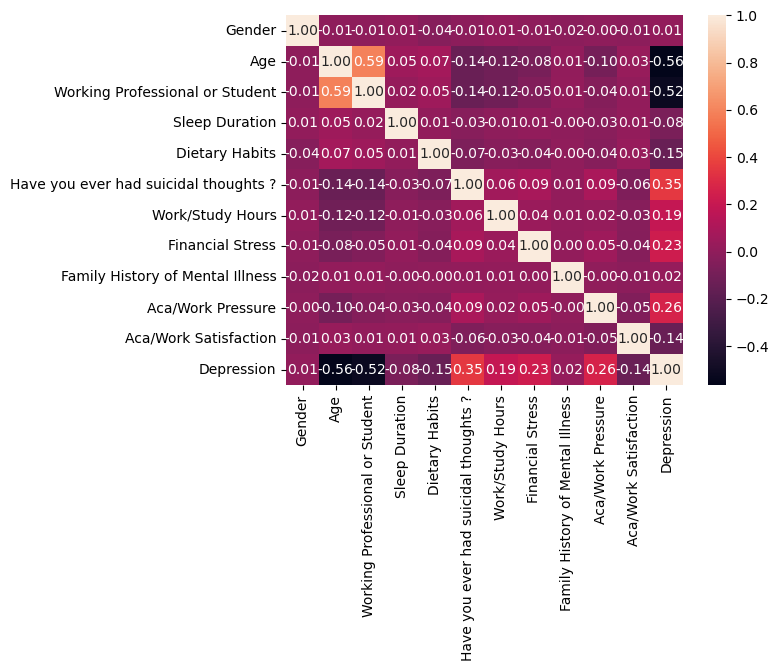


## **D. Feature Engineering**

### Loại bỏ các đặc trưng không cần thiết: các biến định danh như ‘Id’ và ‘Name’

Chọn lựa đặc trưng, nhóm kết hợp sử dụng 3 chiến lược sau:

### **Ma trận tương quan**



Từ ma trận tương quan nhận thấy được:

‘Gender’ và ‘Family History of Mental Illness’ có tương quan với cột mục tiêu Depression rất thấp lần lượt là 0.01 và 0.02.

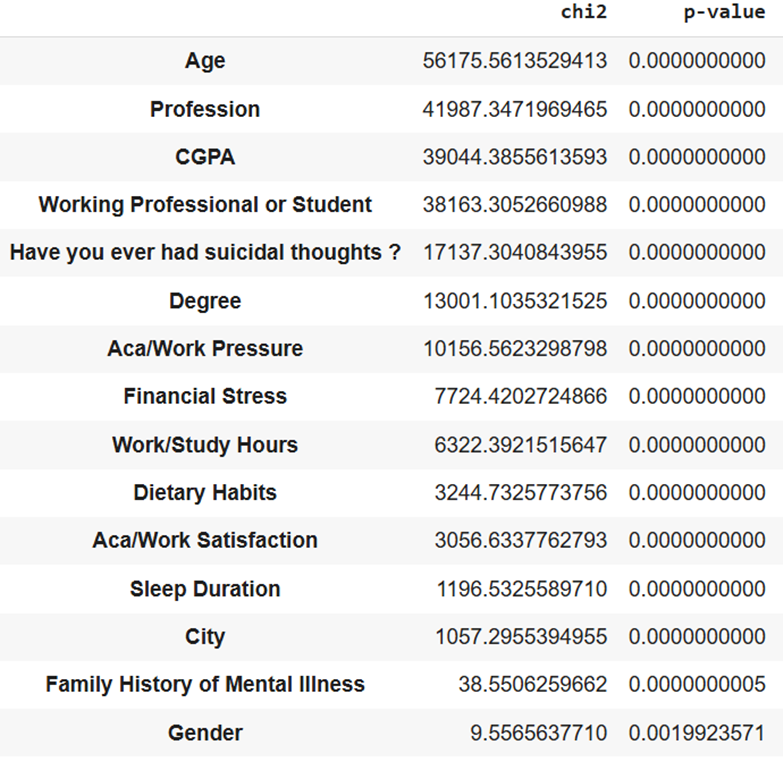
Ngoài ra còn có hai đặc trưng là ‘Age’ và ‘Working Professional or Student’ có mối tương quan cao với nhau. Dù vậy cũng không nên loại bỏ hai đặc trưng này. Vì:

* Tương quan của 2 cột này đối với cột target là rất cao sao với các cột khác.
* Mặc dù Age và Working Professional or Student có tương quan 0.59, nhưng không phải là (0.8 - 0.9) tức là vẫn còn thông tin riêng biệt ở mỗi cột. Ví dụ: Hai người cùng là “Working Professional” nhưng tuổi có thể rất khác nhau.
* ‘Age’ là một biến định lượng (số tuổi cụ thể). ‘Working Professional or Student’ là một biến phân loại
* Ngưỡng tương quan chưa đủ cao 0.59 chưa tới mức quá lớn như 0.8 trở lên

Tóm lại với phương pháp chọn bằng ma trận tương quan chỉ nên cân nhắc loại bỏ hai đặc trưng là ‘Gender’ và ‘Family History of Mental Illness’.

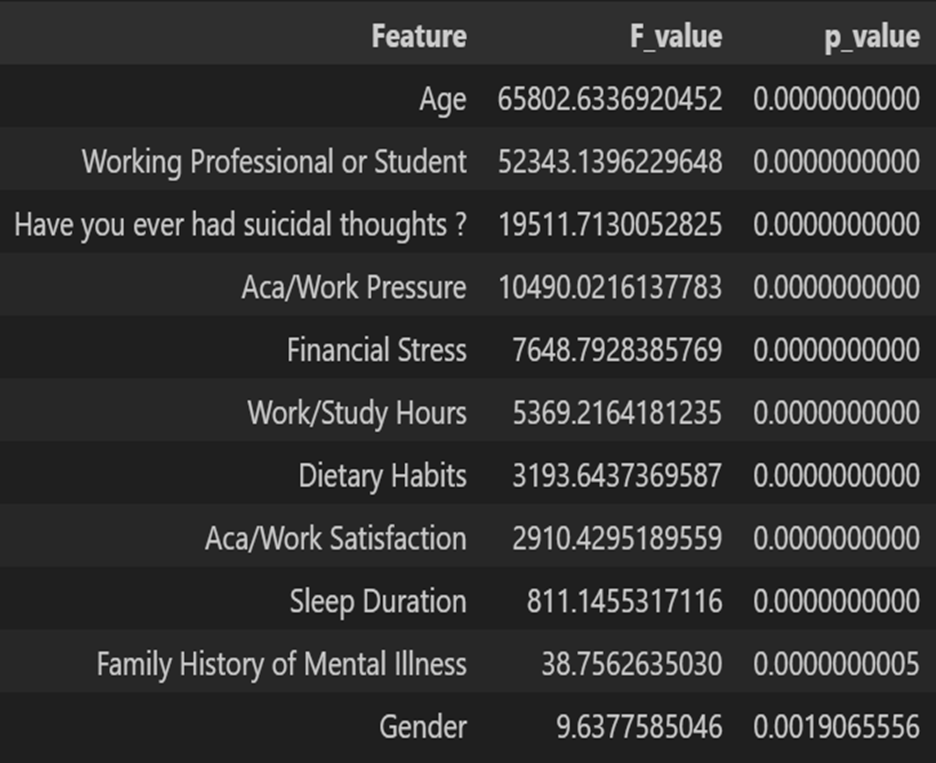
### **Statistical test**

1. **Chi-square test**



* Tất cả các biến đều có mối liên hệ thống kê với Depression (p-value < 0.05)
* Nhóm có Chi² càng lớn (và p-value ≈ 0) như Age, Profession, CGPA, Working Professional or Student, Have you ever had suicidal thoughts?,… cho thấy mối liên hệ từ mạnh tới rất mạnh với nguy cơ trầm cảm.
* Ngược lại, những biến như Gender và Family History of Mental Illness có Chi² nhỏ hơn đáng kể, nghĩa là mức độ liên hệ ít hơn so với các biến khác, cân nhắc loại bỏ.

1. **ANOVA**



Các feature có giá trị F\_value cao (F\_value > 100) và p\_value thấp (p\_value < 0.01) như:

* Age
* Working Professional or Student
* Have you ever had suicidal thoughts?
* Aca/Work Pressure
* Financial Stress
* Work/Study Hours
* Aca/Work Satisfaction
* Dietary Habits
* Sleep Duration

Có sự khác biệt giữa các nhóm lớn hơn nhiều so với độ biến thiên trong cùng một nhóm. Tức Feature đó phân biệt tốt giữa các nhãn.

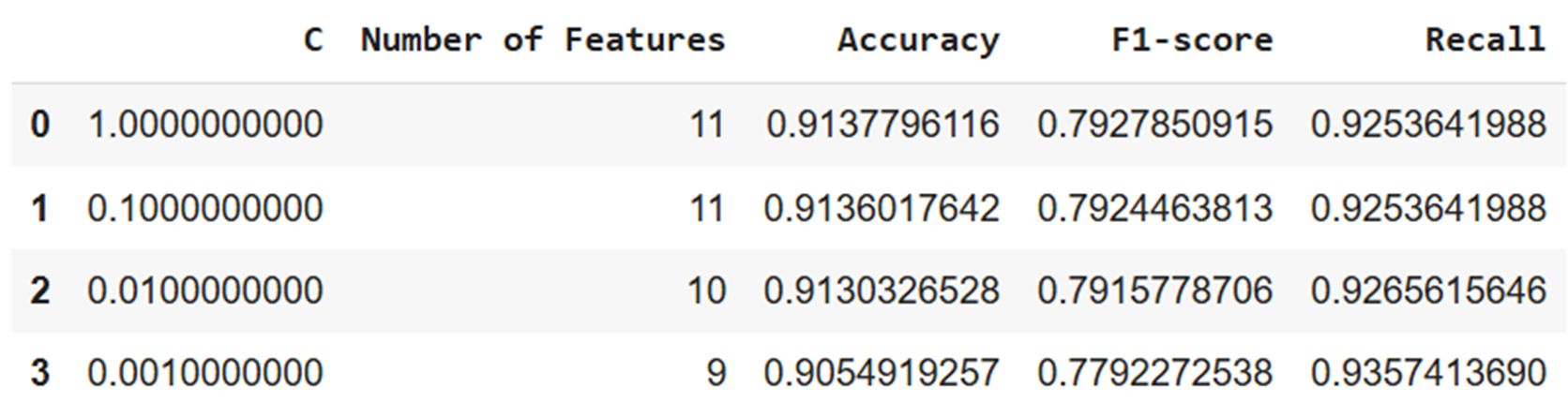
Có khả năng sự khác biệt do ngẫu nhiên rất thấp. Feature thật sự liên quan đến nhãn và nên được giữ lại

Tuy nhiên, cột: Family History of Mental Illness và Gender có F\_value không quá cao (< 100) nhưng p\_value thấp (< 0.01) nên cần được kết hợp thêm với các chiến lược khác để quyết định có nên loại bỏ hay không

### **Model-based methods**

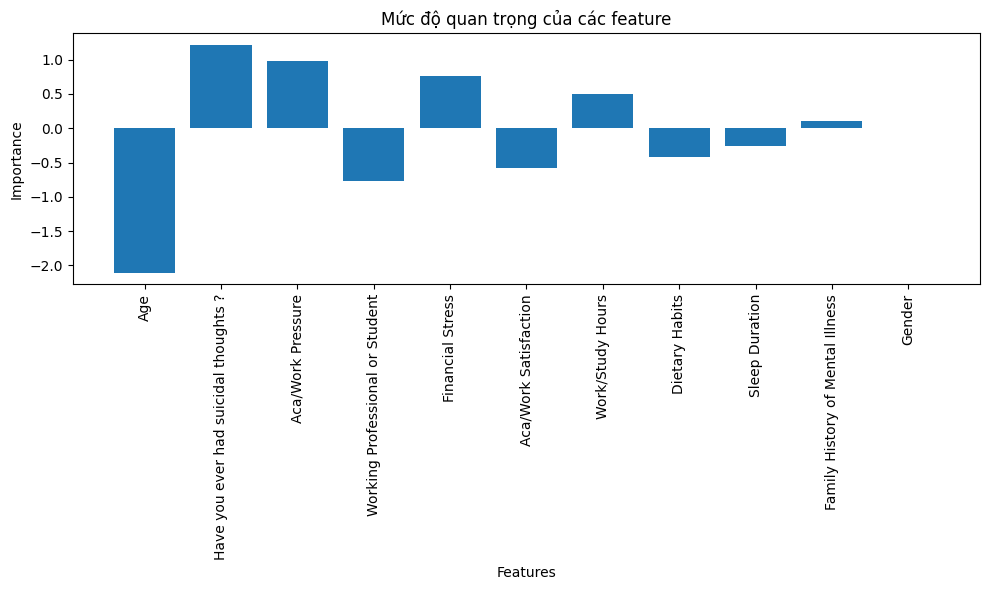
1. Mô hình Logistic Regression với siêu tham số L1

* Chọn siêu tham số C để quyết định mức độ phạt của L1:



Nếu muốn giảm số feature mà mất rất ít performance nhóm đã quyết định chọn C = 0.01:

* Giảm 1 biến (11→10)
* F1 giảm
* Recall tăng nhẹ
* Accuracy giảm không đáng kể



Top 5 biến ảnh hưởng mạnh nhất:

Age (~-2.11)

Have you ever had suicidal thoughts ? (~1.22)

Aca/Work Pressure (~0.98)

Working Professional or Student (~-0.77)

Financial Stress (~0.76)

→ Đây là những biến có giá trị tuyệt đối của hệ số lớn nhất, nghĩa là mỗi biến này sẽ có ảnh hưởng mạnh nhất đến việc bị trầm cảm.

Các biến ở mức trung bình:

Aca/Work Satisfaction (~-0.58)

Work/Study Hours (~0.51)

Dietary Habits (~-0.42)

Sleep Duration (~-0.25)

→ Vẫn có đóng góp nhưng không nổi bật bằng nhóm trên. Có thể cân nhắc giữ lại nếu muốn phân tích toàn diện, hoặc bỏ bớt nếu cần đơn giản hóa mô hình.

Các biến ít quan trọng:

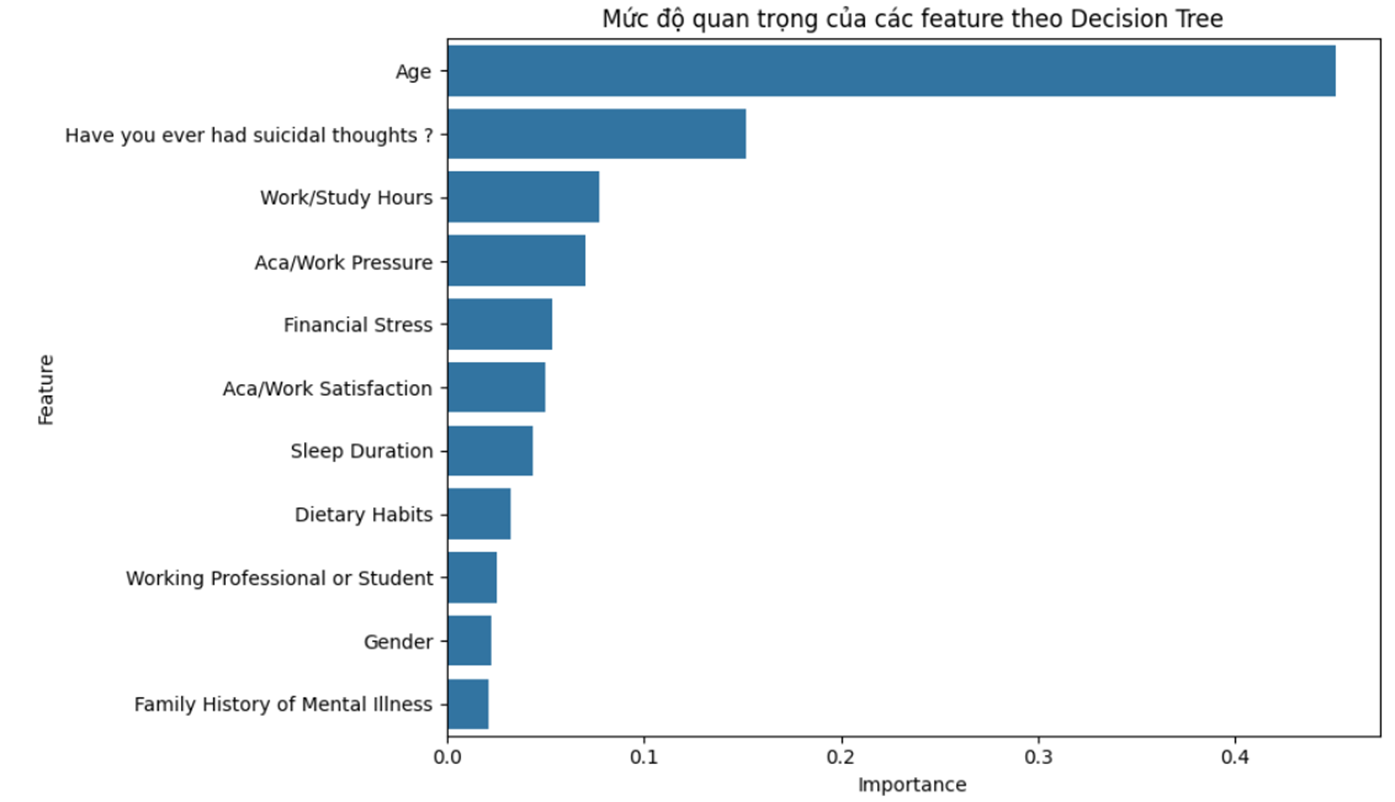
Family History of Mental Illness (~0.11)

Gender (0.00)

→ Hệ số nhỏ, cho thấy đóng góp rất yếu. . Đặc biệt Gender đã bị ép về 0%.

Tóm lại, ở mô hình Logistic Regression nhóm lựa chọn bỏ feature Gender.

1. Mô hình Decision Tree



* Age một mình chiếm gần 45%. ➔ Đây là feature cực mạnh trong việc phân loại.
* Suicidal thoughts cũng chiếm 15%, và đây là biến quan trọng lâm sàng.
* Work/Study Hours, Aca/Work Pressure (~7% mỗi cái)
* Financial Stress, Aca/Work Satisfaction (~5% mỗi cái)
* Working Professional or Student?

→ Góp phần khá tốt, chứng tỏ môi trường học tập/làm việc có ảnh hưởng tâm lý.

* Các features yếu: Gender, Family History of Mental Illness .

→ Gần như không ảnh hưởng nhiều trong dự đoán (~2%).

Tóm lại, với mô hình Decision Tree, nhóm lựa chọn loại bỏ Gender, Family History of Mental Illness

1. Các mô hình Boosting Tree thông qua kĩ thuật Shap values

Kĩ thuật SHAP (SHapley Additive exPlanations) là một công cụ mạnh mẽ để diễn giải các mô hình học máy.

Với kĩ thuật SHAP và model xử lí được categorical data với mô hình xử lí (CatBoost,...), ta có thể đánh giá các cột này.

* Huấn luyện mô hình Boosting Tree

Fit mô hình Gradient Boosting (ví dụ XGBoost, LightGBM) trên tập huấn luyện để làm nền tảng cho tính SHAP.

* Khởi tạo TreeExplainer

Dùng `shap.TreeExplainer(model)` để chuẩn bị công cụ tính SHAP values cho mô hình tree-based.

* Tính SHAP values

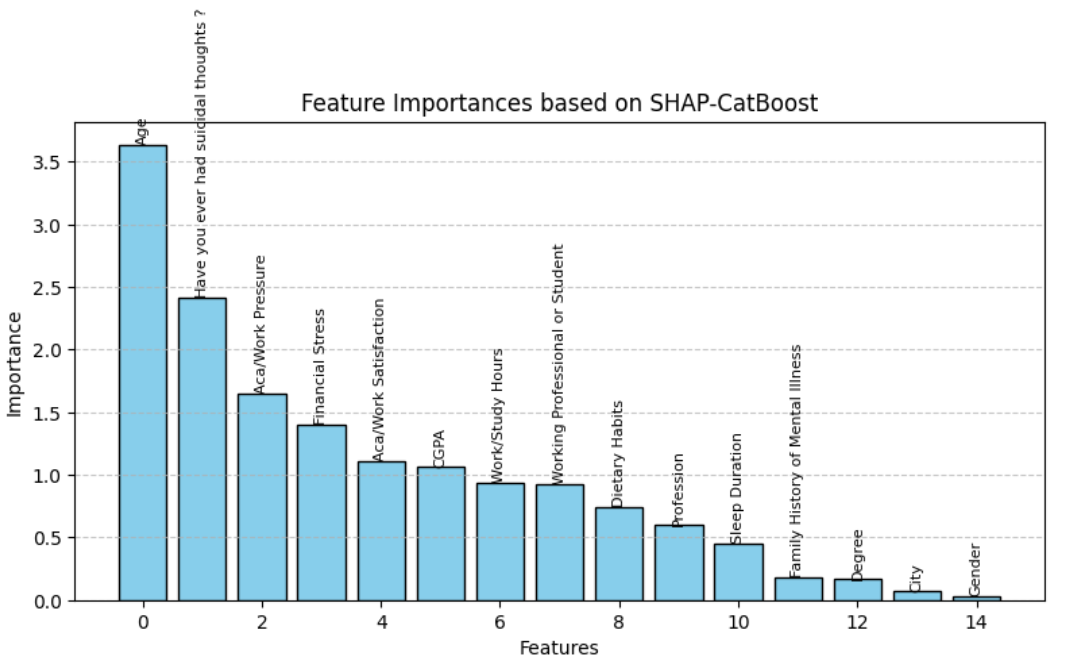
Gọi `explainer.shap\_values(X)` để thu được ma trận SHAP values (số mẫu × số đặc trưng).

* Tổng hợp tầm quan trọng

Tính trung bình tuyệt đối SHAP values mỗi đặc trưng để có thứ tự ưu tiên toàn cục.

* Chọn đặc trưng

Áp dụng ngưỡng, đánh giá qua performance của model hoặc phân phối elbow.



Tổng hợp qua 3 phương pháp đã thực hiện, nhóm đã chọn các đặc trưng:

* Age
* Have you ever had suicidal thoughts ?
* Aca/Work Pressure
* Financial Stress
* Aca/Work Satisfaction
* CGPA
* Work/Study Hours
* Working Professional or Student
* Dietary Habits
* Profession

Đã loại bỏ 4 đặc trưng:

* Family History of Mental Illness
* Degree
* City
* Gender

Với 2 phương pháp ban đầu chỉ xác định loại bỏ 2 đặc trưng là Family History of Mental Illness và Gender. Tuy nhiên ở phương pháp cuối, nhóm nhận thấy Degree và City có đóng góp vào kết quả dự đoán thấp hơn cả Family History of Mental Illness nên nhóm đã quyết định loại bỏ luôn cả Degree và City.

# **III. HUẤN LUYỆN VÀ TINH CHỈNH MÔ HÌNH**

Phần này trình bày chi tiết quá trình lựa chọn các mô hình học máy cũng như phương pháp huấn luyện và tinh chỉnh mô hình. Mục tiêu là xác định các mô hình hiệu quả nhất cho bài toán dự đoán sức khỏe tâm thần, đồng thời đảm bảo kết quả đạt được có độ tin cậy cao.

## **A. Chia tập dữ liệu**

Bộ dữ liệu đã qua xử lý trước được phân chia thành các tập train và test.

Tỉ lệ split được sử dụng là:

* 80% cho training
* 20% cho testing

## **B. Chọn mô hình cho bài toán**

Hai mô hình dựa trên cây và một mô hình Ensemble kết hợp hai mô hình đã được chọn cho bài toán phân loại này: LightGBM, CatBoost và Ensemble. Các mô hình này được sử dụng rộng rãi vì hiệu suất mạnh mẽ và khả năng nắm bắt các quan hệ phi tuyến tính. Chúng thể mạnh mẽ trước dữ liệu ngoại lai và mất cân bằng nhãn.

* LightGBM: Được tối ưu hóa cho tốc độ và hiệu quả bộ nhớ, LightGBM sử dụng một chiến lược tăng trưởng leaf-wise và xử lý các bộ dữ liệu lớn với native support cho các categorical features.
* CatBoost: Được thiết kế để xử lý hiệu quả dữ liệu categorical bằng các techniques như ordered boosting, giúp giảm overfitting và ngăn ngừa rò rỉ dữ liệu.
* Ensemble (stacking): là mô hình kết hợp của 2 mô hình trên tận dụng các ưu điểm của 2 mô hình thành phần để cải thiện hiệu suất tổng thể vời meta-learner là Logistic Regresion.

Class weights đã được sử dụng để đảm bảo học cân bằng trong quá trình huấn luyện. Trọng số 1 đã được gán cho lớp "no depression" (lớp 0) và trọng số 4 đã được gán cho class "depression" (lớp 1).

## **C. Tinh chỉnh mô hình**

Việc lựa chọn siêu tham số phù hợp đóng vai trò quan trọng trong việc nâng cao hiệu suất mô hình. Để đạt được điều này, chiến lược tinh chỉnh kết hợp giữa **Random Search** và **Bayesian Optimization** đã được áp dụng nhằm cân bằng giữa phạm vi tìm kiếm và hiệu quả thời gian.

Cụ thể, **Random Search** được sử dụng đầu tiên để khảo sát nhanh một không gian siêu tham số rộng với chi phí tính toán thấp. Sau đó, điểm có hiệu suất tốt nhất từ quá trình này được chọn làm điểm khởi đầu cho **Bayesian Optimization** sử dụng thuật toán **TPE (Tree-structured Parzen Estimator)**. Cách tiếp cận kết hợp này giúp tăng khả năng tìm ra cấu hình tối ưu trong thời gian ngắn hơn.

| **Model** | **Hyperparameters** |
| --- | --- |
| **LightGBM** | 'max\_bin': 205,  'learning\_rate': 0.04508184739922241,  'num\_leaves': 67,  'max\_depth': 4,  'min\_data\_in\_leaf': 96,  'lambda\_l1': 1.987419215841184,  'lambda\_l2': 3.4964600099442964,  'feature\_fraction': 0.987878776112855,  'bagging\_fraction': 0.7118762375349093,  'bagging\_freq': 9, 'min\_child\_samples': 38 |
| **CatBoost** | 'iterations': 260,  'learning\_rate': 0.01684410510828215,  'depth': 5,  'l2\_leaf\_reg': 3.443242463040269,  'border\_count': 244,  'bagging\_temperature': 0.7683018048075387  'random\_strength': 2.8472974936188695 |

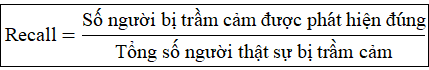
# **IV. KẾT QUẢ VÀ ĐÁNH GIÁ**

Phần này trình bày những phát hiện thực nghiệm từ các mô hình, mô tả chi tiết hiệu suất so sánh của chúng.

## **A. Metrics**

Do dữ liệu có sự chênh lệch số lượng giữa hai nhãn, việc chọn Accuracy dường như không phù hợp với bài toán và bộ dữ liệu này.

Bên cạnh đó, vì đây là mô hình sàn lọc ban đầu cho người sử dụng nhận ra vấn đề tâm lí của bản thân, việc để xót lại người bị trầm cảm có thể dẫn đến người đó gánh chịu trầm cảm trong thời gian sau. Bởi vậy, nhóm chọn Recall làm metric chính:



Nếu mô hình có Recall cao, nghĩa là nó có khả năng phát hiện phần lớn người thật sự bị trầm cảm, tránh để họ bị bỏ qua - điều rất quan trọng trong y tế và sức khỏe tâm thần.

Bên cạnh đó, F1-score cũng được sử dụng làm metric phụ, giúp kiểm soát sự cân bằng giữa Precision và Recall.

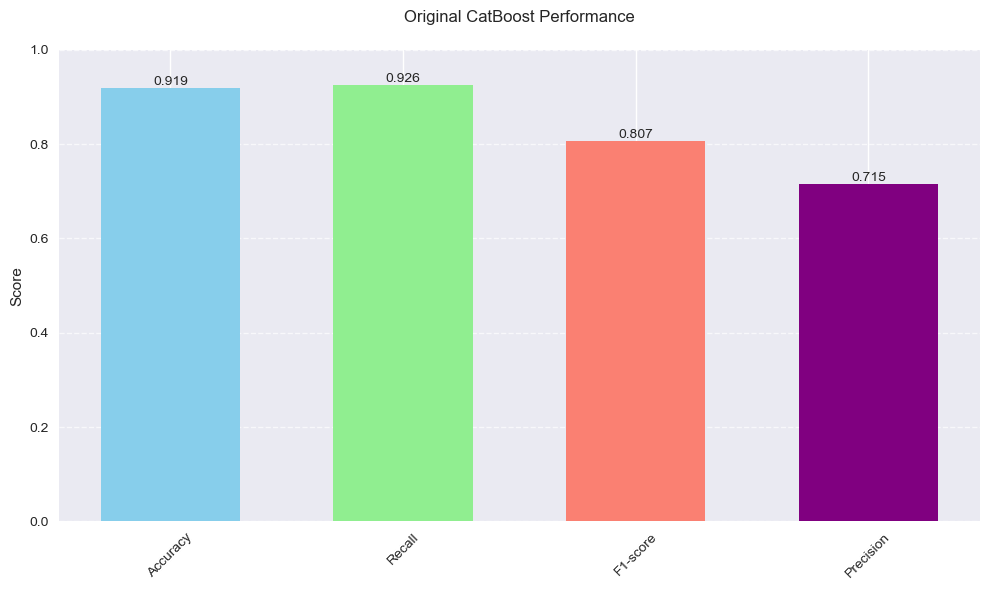
Mô hình cũng được đánh giá qua metrics khác để so sánh hiệu quả trên nhiều phương diện:

* Accuracy: Tính chính xác chung, được giải thích một cách thận trọng do sự mất cân bằng.
* Precision: Phản ánh trong số các trường hợp mô hình dự đoán là trầm cảm, thì bao nhiêu là đúng.
* ROC-AUC: ROC curve là đường cong thể hiện mối quan hệ giữa True Positive Rate (Recall) và False Positive Rate ở nhiều ngưỡng phân loại khác nhau. đo lường khả năng mô hình phân biệt giữa hai lớp.

## **B. Kết quả và phân tích**

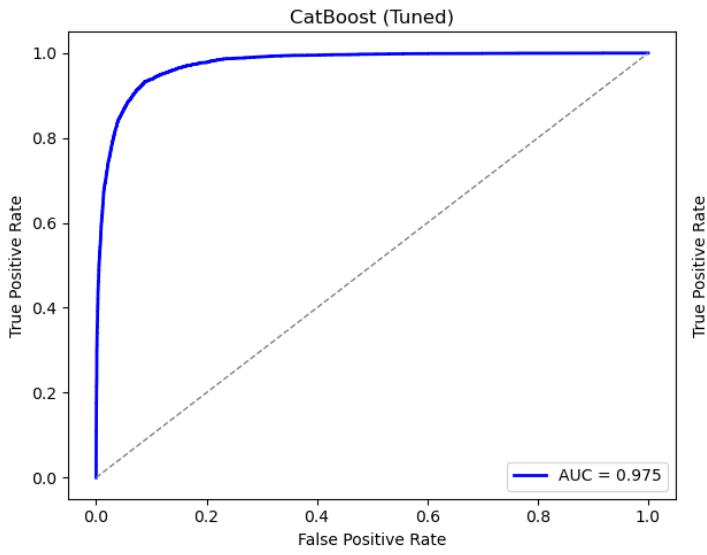
### **Đánh giá trên test set**

1. Mô hình CatBoost



*Hình 14: Điểm đánh giá của từng metric cho mô hình CatBoost*

* Recall đạt 0.926 chứng tỏ mô hình có độ nhạy rất tốt có thể chọn được tới 92.6% người thực sự trầm cảm
* F1-score ở mức 0.8 cho thấy mô hình vẫn có thể cân bằng tốt giữa Precision và Recall.



Đường ROC cong mạnh về bên trái chứng tỏ mô hình rất tốt trong việc phân biệt 2 lớp.

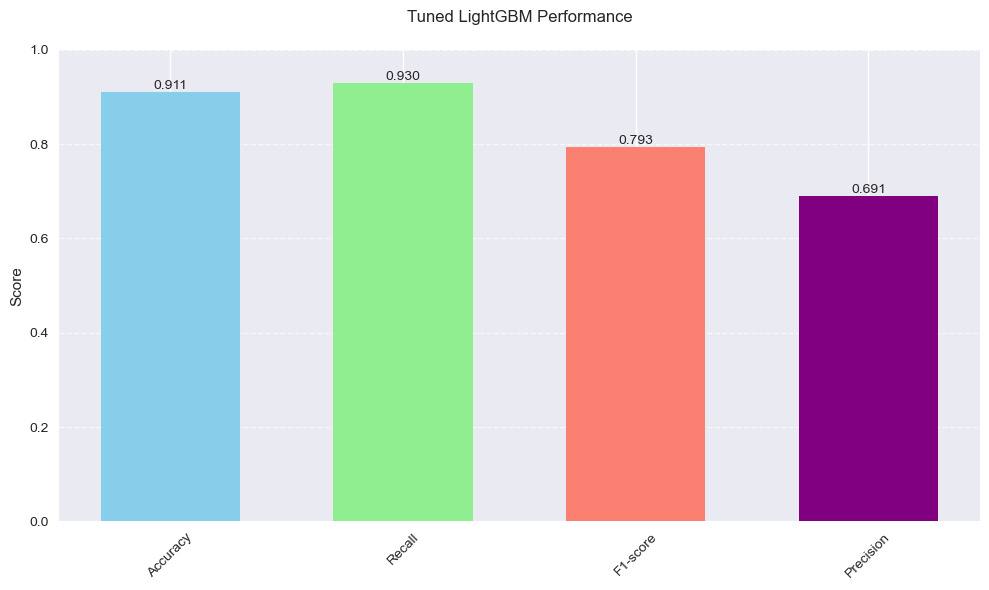
AUC = 0.975 cho thấy hiệu suất mô hình cao.

Kiểm tra overfitting:

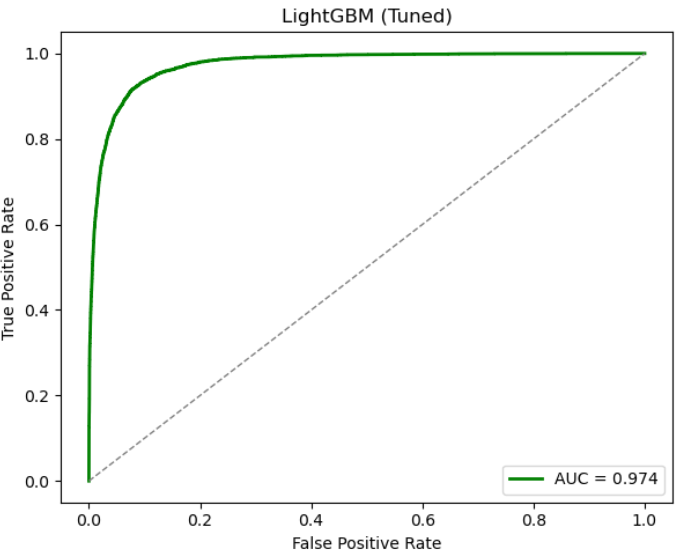
| **CatBoost (Tuned)** | **Train** | **Test** |
| --- | --- | --- |
| **Accuracy** | 0.9139 | 0.9159 |
| **Recall** | 0.9324 | 0.9325 |
| **F1-score** | 0.7971 | 0.8021 |
| **Precision** | 0.6961 | 0.7037 |

Mô hình hoàn toàn không bị overfitting vì các chỉ số trên train và test gần như tương đương.

1. Mô hình LightGBM



Precision và F1-score thấp hơn một chút so với CatBoost. Điều này cho thấy LightGBM có xu hướng dự đoán dương tính giả nhiều hơn, gây giảm độ tin cậy.



Mô hình LightGBM đạt hiệu suất gần như tương đương với CatBoost, chỉ thấp hơn rất nhẹ. Mô hình vẫn có hiệu suất cao với khả năng phân biệt tốt giữa 2 lớp ‘Trầm cảm’ và ‘Không trầm cảm’.

Kiểm tra overfitting:

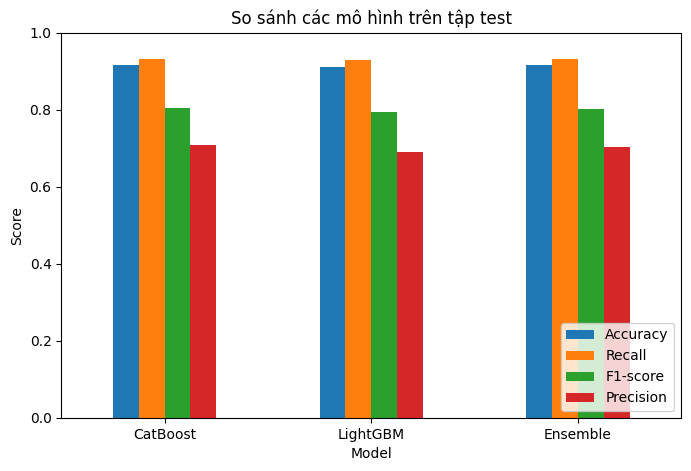
| **LightGBM (Tuned)** | **Train** | **Test** |
| --- | --- | --- |
| **Accuracy** | 0.9115 | 0.9112 |
| **Recall** | 0.9358 | 0.9304 |
| **F1-score** | 0.7932 | 0.7929 |
| **Precision** | 0.6883 | 0.6909 |

Mô hình LightGBM cũng không bị overfitting.

**2. Ensemble model**

Ngoài ra, Stacking Ensemble cũng được thử nghiệm với:

* Base Learner: CatBoost và LightGBM đã được tuning.
* Meta-learner: Logistic Regression, kết quả dự đoán từ CatBoost và LighGBM được chuyển tiếp vào mô hình meta-learner với mong muốn có được mô hình tốt hơn.



Qua hình vẽ so sánh giữa 3 mô hình: CatBoost, LightGBM và Stacking. Ta thấy Stacking Ensemble có hiệu quả không tốt hơn so với mô hình con.

Số lượng mẫu giống nhau bị dự đoán sai ở cả ba mô hình là **2156** mẫutrong khi riêng Ensemble Stacking dự đoán sai **2375 mẫu**. Tỷ lệ dự đoán sai giống nhau chiếm đến 90% tổng dự đoán sai cho thấy 2 mô hình con học từ bộ dữ liệu giống nhau, sai những lỗi giống nhau.

Vậy mô hình ensemble trên chưa phù hợp với bài toán và bộ dữ liệu hiện tại.

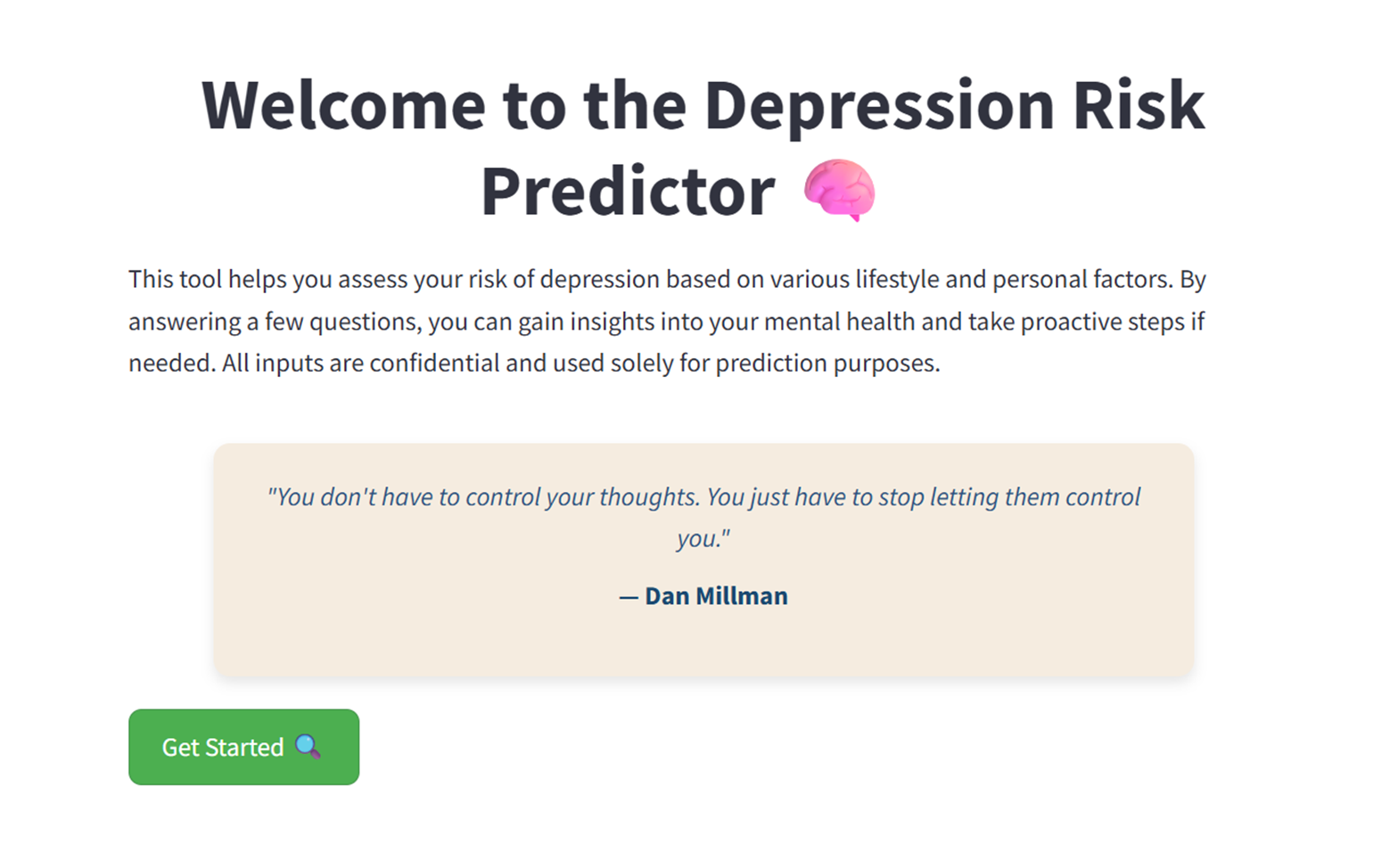
**V. Triển khai mô hình**

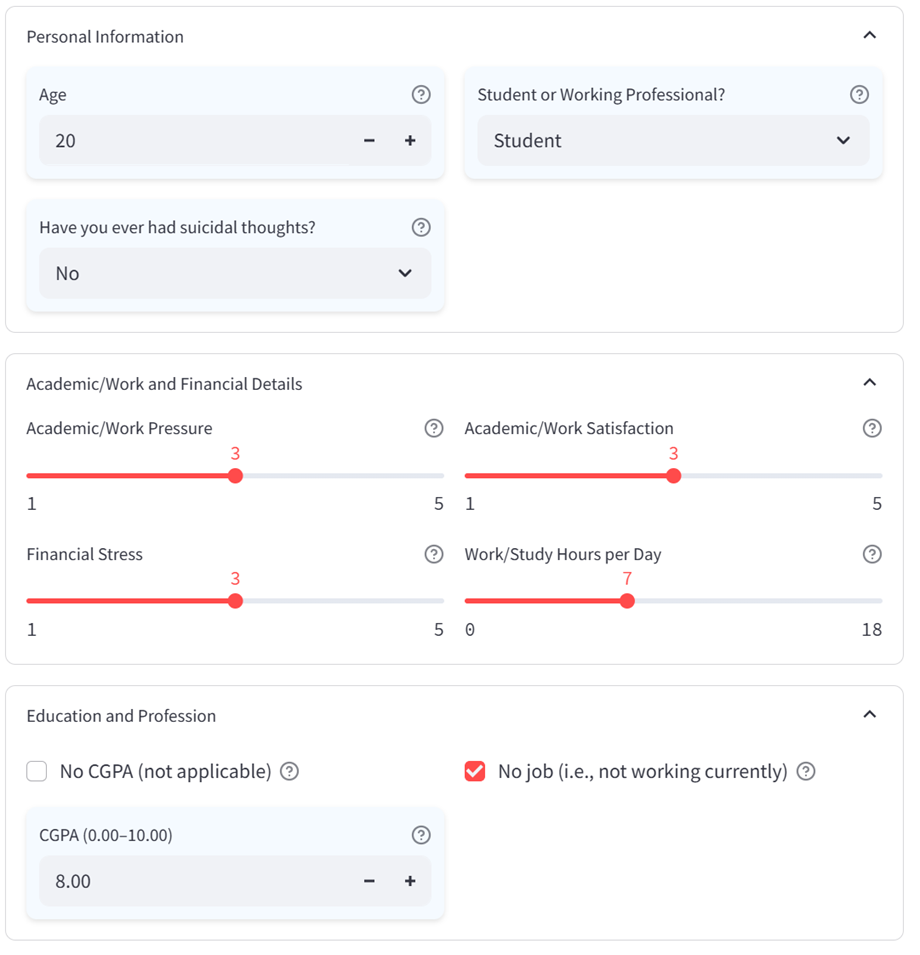
Mô hình **CatBoost** đã được lựa chọn để triển khai đến người dùng cuối nhờ hiệu suất vượt trội, đặc biệt về các chỉ số Recall và F1-score, cùng với thời gian suy luận (inference time) ngắn, phù hợp cho các ứng dụng thời gian thực.

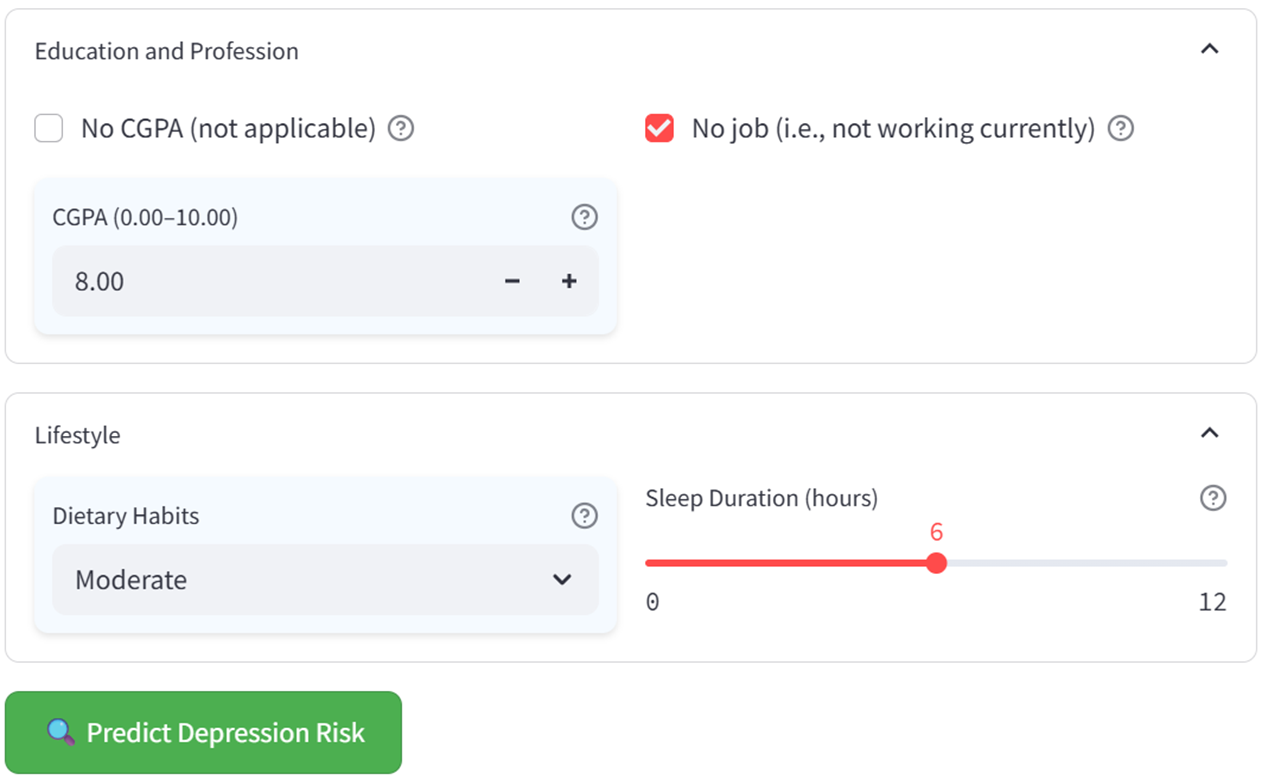
Người dùng sẽ cung cấp các thông tin cần thiết về bản thân thông qua giao diện nhập liệu, sau đó nhấn nút “Predict Depression Risk**”** để nhận kết quả dự đoán. Thay vì chỉ trả về kết quả nhị phân (có/không có nguy cơ trầm cảm), hệ thống hiển thị **xác suất bị trầm cảm**, kèm theo **mức đánh giá định tính** (thấp, trung bình hoặc cao), cùng với lời khuyên phù hợp.

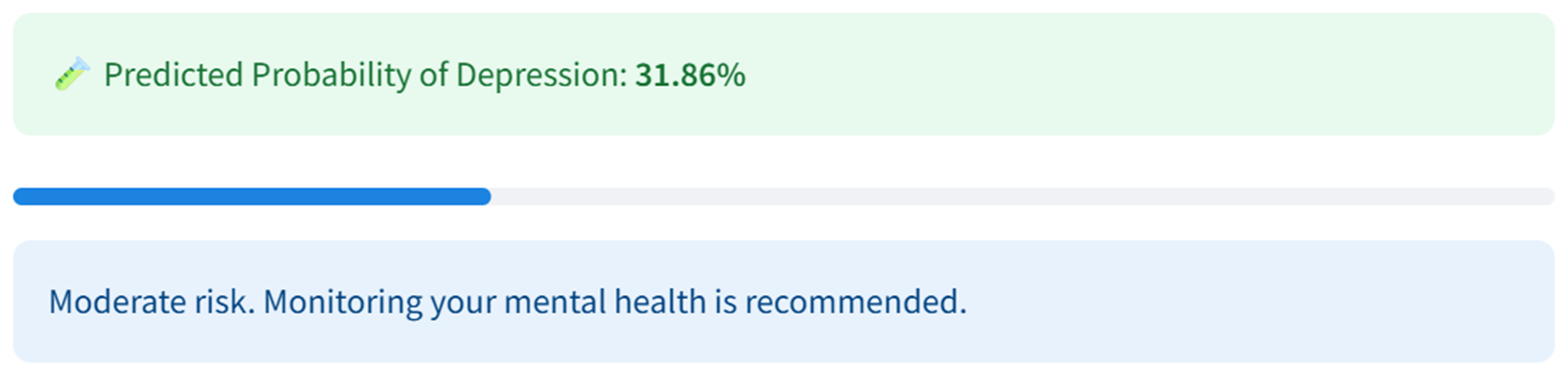
Việc hiển thị xác suất giúp người dùng hiểu rõ hơn về mức độ rủi ro cũng như độ tin cậy của kết quả, từ đó hỗ trợ ra quyết định cá nhân một cách chủ động và có cơ sở hơn.

Một vài hình ảnh của web:









# **V. KẾT LUẬN**

Bài báo cáo trình bày xây dựng và đánh giá các mô hình học máy để dự đoán nguy cơ trầm cảm, sử dụng bộ dữ liệu "Exploring Mental Health Data".

Sau quá trình phân tích dữ liệu chuyên sâu (EDA) và tiền xử lý, các yếu tố quan trọng ảnh hưởng đến trầm cảm được phát hiện ra như tuổi tác, thời lượng ngủ, thói quen ăn uống và các chỉ số tự đánh giá về áp lực/hài lòng trong học tập/công việc và tài chính.

Trong số các mô hình được thử nghiệm (LightGBM, CatBoost và Stacking Ensemble) , CatBoost đã cho thấy hiệu suất tốt hơn, đặc biệt ở chỉ số Recall (khả năng phát hiện đúng người trầm cảm) và F1-score. CatBoost được chọn để triển khai do điểm tốt nhất và thời gian suy luận nhanh. Hệ thống được triển khai cung cấp cho người dùng xác suất và mức độ rủi ro trầm cảm cùng với lời khuyên phù hợp, nhằm hỗ trợ việc phát hiện sớm và đưa ra quyết định chủ động về sức khỏe tâm thần.

Hướng phát triển tiếp:

* Bổ sung thêm các dữ liệu mới và đặc trưng mới vì các thông tin của bộ dữ liệu gốc chủ yếu do người dùng tự đánh giá, có thể có thiên vị (bias).
* Sử dụng các mô hình phức tạp hơn như học sâu để có hiệu suất cao hơn.
* Tích hợp các phương pháp giải thích mô hình khác để tăng cường tính minh bạch và đáng tin cậy của hệ thống.

TÀI LIỆU THAM KHẢO

[1] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24.

[2] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[3] Walter Reade and Elizabeth Park. Exploring Mental Health Data. Kaggle. <https://kaggle.com/competitions/playground-series-s4e11>, 2024.

[4] Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. Ieee Access, 10, 99129-99149.

[5] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. Ensemble machine learning: Methods and applications, 157- 175.

[6] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.