# **FPT SOFTWARE**

**2025.07 Python for Machine Learning HN25\_CPL.V\_PML\_01 Nhóm 09**

# **FPT SOFTWARE ACADEMY**

# --------🙞🙞🕮🙜🙜--------



**BÁO CÁO NHÓM**

**Chủ đề:**

**ME/CFS vs Depression Classification Dataset**

**NHÓM THỰC HIỆN: NHÓM 09**

**Hà Nội, 07/2025**

**FPT SOFTWARE**

**FPT SOFTWARE ACADEMY**

**--------🙞🙞🕮🙜🙜--------**



**BÁO CÁO NHÓM**

***Chủ đề:***

**ME/CFS vs Depression Classification Dataset**

|  |  |
| --- | --- |
| **NHÓM 09** | |
| Nguyễn Ngọc Hải | Haui\_2022605947 |
| Nguyễn Thành Công | Haui\_2022606702 |
| Vũ Minh Đức | NEU\_11221425 |
| Nguyễn Hoàng Nguyên | NEU\_11224818 |
| Nguyễn Trọng Vỹ | NEU\_11227025 |

**Hà Nội, 07/2025**

**MỤC LỤC**

[DANH MỤC BIỂU ĐỒ 3](#_Toc203534468)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 5](#_Toc203534469)

[1.1. Lý do chọn đề tài 5](#_Toc203534470)

[1.2. Mục tiêu dự án 5](#_Toc203534471)

[1.3. Mô tả nguồn và đặc điểm dữ liệu 5](#_Toc203534472)

[1.4. Giải thích ý nghĩa các cột 6](#_Toc203534473)

[1.5. Các đặc trưng quan trọng 7](#_Toc203534474)

[1.6. Vấn đề có thể giải quyết bằng học máy 7](#_Toc203534475)

[CHƯƠNG 2: PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU (EDA) 9](#_Toc203534476)

[2.1. Thống kê mô tả dữ liệu numerical 9](#_Toc203534477)

[2.2. Phân tích dữ liệu categorical 11](#_Toc203534480)

[2.3. Trực quan hóa dữ liệu 13](#_Toc203534519)

[2.3.1. Histogram các biến định lượng 13](#_Toc203534520)

[2.3.2. Boxplot các biến định lượng 14](#_Toc203534521)

[2.4. Phân tích mối quan hệ giữa các thuộc tính 18](#_Toc203534528)

[CHƯƠNG 3: NHẬN XÉT & ĐỊNH HƯỚNG 22](#_Toc203534530)

[3.1. Tóm tắt các phát hiện quan trọng từ dữ liệu 22](#_Toc203534531)

[3.2. Những đặc trưng có thể ảnh hưởng lớn đến mô hình 22](#_Toc203534532)

[3.3. Nhận xét chung khi phân tích mối quan hệ 22](#_Toc203534533)

[3.4. Giải pháp tiền xử lý dữ liệu 23](#_Toc203534534)

[3.5. Định hướng mô hình học máy 23](#_Toc203534535)

[3.6. Bài học rút ra 23](#_Toc203534536)

[3.7. Hướng phát triển trong tương lai 24](#_Toc203534537)

[KẾT LUẬN 25](#_Toc203534538)

[TÀI LIỆU THAM KHẢO 27](#_Toc203534539)

# **DANH MỤC BIỂU ĐỒ**

[Biểu đồ 1: Biểu đồ tổng quan Histogram 14](#_Toc203534665)

[Biểu đồ 2: Biểu đồ Boxplot 15](#_Toc203534666)

[Biểu đồ 3: Biểu đồ Barchart 16](#_Toc203534667)

[Biểu đồ 4: Biểu đồ Scatter plot 18](#_Toc203534668)

[Biểu đồ 5: Biểu đồ Heatmap 19](#_Toc203534669)

**MỞ ĐẦU**

Với sự phát triển mạnh mẽ của công nghệ thông tin và các công cụ phân tích dữ liệu, việc ứng dụng khoa học dữ liệu vào nghiên cứu và hỗ trợ ra quyết định trong nhiều lĩnh vực ngày càng trở nên quan trọng. Khối lượng dữ liệu lớn cùng khả năng xử lý nhanh chóng đã giúp các nhà nghiên cứu và tổ chức y tế khám phá sâu hơn về đặc điểm bệnh học, từ đó đưa ra những hướng điều trị phù hợp và kịp thời hơn bao giờ hết. Trong bối cảnh đó, việc áp dụng các kỹ thuật phân tích dữ liệu và học máy vào lĩnh vực y tế đã trở thành xu thế tất yếu, giúp nâng cao chất lượng nghiên cứu và hiệu quả chẩn đoán trong môi trường y khoa hiện đại.

Nhận thức được tầm quan trọng của việc khai thác dữ liệu y tế, đặc biệt là trong nghiên cứu về hội chứng mệt mỏi mạn tính (ME/CFS) và trầm cảm – hai tình trạng bệnh lý có nhiều triệu chứng chồng lấn, nhóm chúng tôi đã thực hiện đề tài phân tích dữ liệu này. Mục tiêu của đề tài là tiến hành phân tích, trực quan hóa và nhận diện các đặc trưng quan trọng từ tập dữ liệu được cung cấp, qua đó hỗ trợ quá trình nhận định và phân loại bệnh một cách chính xác, nhanh chóng. Hệ thống phân tích cần đảm bảo cung cấp thông tin rõ ràng, dễ hiểu, phục vụ cho cả nhu cầu nghiên cứu lẫn ứng dụng thực tiễn.

Mục tiêu của dự án không chỉ dừng lại ở việc tìm ra các mối quan hệ quan trọng giữa các biến số mà còn hướng tới việc đề xuất mô hình học máy phù hợp, góp phần hỗ trợ chẩn đoán bệnh hiệu quả hơn. Thông qua đề tài này, nhóm chúng tôi hy vọng sẽ góp phần nâng cao giá trị ứng dụng của phân tích dữ liệu trong lĩnh vực y tế, đồng thời mở ra những định hướng nghiên cứu mới trong việc ứng dụng công nghệ dữ liệu vào chăm sóc sức khỏe.

**CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI**

**1.1. Lý do chọn đề tài**

Trong bối cảnh hiện nay, các rối loạn sức khỏe tâm thần và thể chất đang ngày càng được quan tâm nghiên cứu. Đặc biệt, hội chứng mệt mỏi mạn tính (ME/CFS) và trầm cảm là hai tình trạng có nhiều triệu chứng chồng lấn như mệt mỏi kéo dài, chất lượng giấc ngủ kém, khó tập trung, hoặc suy giảm hoạt động xã hội.  
Tuy nhiên, việc phân biệt và đánh giá hai tình trạng này gặp nhiều khó khăn do sự tương đồng về biểu hiện lâm sàng. Vì vậy, nhóm chúng tôi lựa chọn đề tài “Phân tích dữ liệu về ME/CFS và Trầm cảm” nhằm khám phá đặc điểm dữ liệu, tìm hiểu mối quan hệ giữa các chỉ số sức khỏe, từ đó hỗ trợ định hướng xây dựng mô hình dự đoán chẩn đoán trong tương lai.

**1.2. Mục tiêu dự án**

Mục tiêu chính của dự án bao gồm:

1. Thực hiện phân tích khám phá dữ liệu (EDA) để hiểu rõ đặc điểm phân bố và xu hướng của các biến số trong bộ dữ liệu.
2. Nhận diện các thuộc tính quan trọng có khả năng phân biệt giữa các nhóm chẩn đoán (ME/CFS, Trầm cảm hoặc cả hai).
3. Đề xuất định hướng áp dụng mô hình học máy cho các bước phân tích nâng cao hoặc dự đoán trong tương lai.

**1.3. Mô tả nguồn và đặc điểm dữ liệu**

Bộ dữ liệu được cung cấp cho nhóm bao gồm 1.000 bản ghi, mỗi bản ghi tương ứng với một người tham gia khảo sát. Bộ dữ liệu được thu thập từ một dự án nghiên cứu sức khỏe cộng đồng, trong đó mỗi người tham gia được đánh giá qua nhiều khía cạnh: nhân khẩu học, lâm sàng, thói quen sinh hoạt và tình trạng chẩn đoán.  
 Dữ liệu được lưu trữ dưới dạng bảng, gồm cả biến định lượng (numerical) và biến định tính (categorical). Một số cột có giá trị khuyết (missing values) cần được xem xét xử lý trước khi huấn luyện mô hình.

**1.4. Giải thích ý nghĩa các cột**

Bảng dưới đây liệt kê chi tiết các cột dữ liệu và ý nghĩa của chúng:

|  |  |
| --- | --- |
| **Tên cột** | **Ý nghĩa** |
| age | Tuổi của người tham gia khảo sát. |
| gender | Giới tính (Male/Female). |
| sleep\_quality\_index | Chỉ số chất lượng giấc ngủ (thang điểm 0–10, cao hơn nghĩa là chất lượng tốt hơn). |
| brain\_fog\_level | Mức độ sương mù não (0–10, cao hơn nghĩa là nặng hơn). |
| physical\_pain\_score | Điểm đau thể chất (0–10, cao hơn nghĩa là đau nhiều hơn). |
| stress\_level | Mức độ căng thẳng (0–10). |
| depression\_phq9\_score | Điểm trắc nghiệm PHQ-9 đánh giá trầm cảm (thang điểm chuẩn). |
| fatigue\_severity\_scale\_score | Điểm thang đo mức độ mệt mỏi. |
| pem\_duration\_hours | Số giờ kéo dài triệu chứng PEM (post-exertional malaise). |
| hours\_of\_sleep\_per\_night | Số giờ ngủ trung bình mỗi đêm. |
| pem\_present | Cờ nhị phân (0/1) cho biết có xuất hiện PEM hay không. |
| work\_status | Tình trạng công việc hiện tại (Working/Not working/Partially working…). |
| social\_activity\_level | Mức độ hoạt động xã hội (Low/Medium/High). |
| exercise\_frequency | Tần suất tập thể dục (Daily/Often/Rarely/Never). |
| meditation\_or\_mindfulness | Có thực hành thiền/chánh niệm hay không (Yes/No). |
| diagnosis | Nhãn chẩn đoán: ME/CFS, Depression, hoặc Both. |

**1.5. Các đặc trưng quan trọng**

Bộ dữ liệu mang những đặc điểm mô phỏng sát thực tế, giúp việc phân tích và xây dựng mô hình học máy trở nên thực tiễn hơn:

* Giá trị thiếu (Missing values):Hầu hết các thuộc tính đều tồn tại giá trị thiếu ở mức 1–5%, phản ánh đúng những khó khăn thường gặp trong quá trình thu thập dữ liệu sức khỏe ngoài đời thực. Đây là yếu tố đòi hỏi nhóm cần cân nhắc các kỹ thuật tiền xử lý như loại bỏ, thay thế hoặc ước lượng giá trị.
* Nhiễu có kiểm soát (Controlled noise):Tất cả các đặc trưng số (numerical features) đều được chèn nhiễu có kiểm soát, nhằm đảm bảo dữ liệu không hoàn toàn tách lớp một cách hoàn hảo. Điều này giúp mô phỏng dữ liệu thực tế, nơi luôn tồn tại biến thiên và sai số đo lường.
* Logic chẩn đoán mô phỏng thực tiễn:Nhãn chẩn đoán trong bộ dữ liệu được thiết lập dựa trên các heuristics lâm sàng (clinical-like heuristics). Nhờ đó, các mô hình học máy được huấn luyện trên bộ dữ liệu này có thể hỗ trợ tốt hơn trong các tình huống ra quyết định thực tế trong lĩnh vực y tế.

**1.6. Vấn đề có thể giải quyết bằng học máy**

Dựa trên đặc điểm dữ liệu, có thể đặt ra các bài toán học máy như sau:

* Phân loại nhị phân:

Xây dựng mô hình phân loại để phân biệt giữa hai nhóm chính: ME/CFS và Depression.  
➝ Ứng dụng: hỗ trợ bác sĩ xác định tình trạng bệnh nhân dựa trên các đặc trưng lâm sàng.

* Phân loại đa lớp:

Xây dựng mô hình phân loại nhiều lớp để phân biệt ME/CFS, Depression và Both.  
➝ Ứng dụng: cung cấp chẩn đoán chi tiết hơn, hỗ trợ cá nhân hóa phương pháp điều trị.

**CHƯƠNG 2: PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU (EDA)**

**2.1. Thống kê mô tả dữ liệu numerical**

Đối với các biến định lượng, nhóm tiến hành tính toán các chỉ số cơ bản như giá trị trung bình (mean), trung vị (median), giá trị nhỏ nhất – lớn nhất (min – max), và độ lệch chuẩn (std).

Các biến số được phân tích bao gồm: age, sleep\_quality\_index, brain\_fog\_level, physical\_pain\_score, stress\_level, depression\_phq9\_score, fatigue\_severity\_scale\_score, pem\_duration\_hours, hours\_of\_sleep\_per\_night.

*Bảng thống kê mô tả cho các cột dạng số*

|  | **age** | **sleep\_**  **quality\_**  **index** | **Brain**  **\_fog**  **\_level** | **Physical**  **\_pain**  **\_score** | **Stress**  **\_level** | **Depression**  **\_phq9**  **\_score** | **Fatigue**  **\_severity**  **\_scale**  **\_score** | **pem\_**  **duration**  **\_hours** | **hours\_of**  **\_sleep**  **\_per\_**  **night** | **Pem**  **\_present** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1000.00 | 953.00 | 952.00 | 966.00 | 952.00 | 978.00 | 979.00 | 976.00 | 979.00 | 1000.00 |
| **mean** | 44.38 | 5.47 | 5.61 | 5.52 | 5.46 | 12.27 | 6.41 | 23.11 | 6.57 | 0.60 |
| **std** | 15.32 | 2.57 | 2.65 | 2.57 | 2.57 | 6.31 | 2.21 | 13.82 | 2.04 | 0.49 |
| **min** | 18.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 3.00 | 0.00 |
| **25%** | 31.75 | 3.10 | 3.30 | 3.32 | 3.30 | 9.00 | 6.30 | 11.00 | 4.80 | 0.00 |
| **50%** | 45.00 | 5.60 | 5.80 | 5.60 | 5.40 | 10.00 | 7.00 | 23.00 | 6.60 | 1.00 |
| **75%** | 57.00 | 7.70 | 7.90 | 7.80 | 7.70 | 16.00 | 7.50 | 35.00 | 8.35 | 1.00 |
| **max** | 70.00 | 10.00 | 10.00 | 10.00 | 10.00 | 27.00 | 10.00 | 47.00 | 10.00 | 1.00 |

***Nhận xét chi tiết về một số biến quan trọng***

* Gender  
  Cột gender thể hiện sự phân bố giới tính trong tập dữ liệu. Tỷ lệ phần trăm của từng giới tính cho phép đánh giá sự cân bằng giới tính trong mẫu nghiên cứu. Nếu một giới tính chiếm tỷ lệ áp đảo, điều này có thể ảnh hưởng đến khả năng tổng quát hóa của mô hình đối với các nhóm giới tính còn lại.
* Age  
  Tuổi trung bình của bệnh nhân là 44.38 tuổi, trải dài từ 18 đến 70 tuổi. Tuổi có khả năng ảnh hưởng đến sự xuất hiện và biểu hiện của cả ME/CFS và trầm cảm. Một số bệnh lý có xu hướng phổ biến hơn ở một độ tuổi nhất định hoặc các triệu chứng có thể biểu hiện khác nhau tùy theo nhóm tuổi.
* Fatigue\_Severity\_Scale\_Score  
  Mức độ mệt mỏi trung bình là dao động từ 0 đến 10. Mệt mỏi là triệu chứng cốt lõi của ME/CFS và cũng phổ biến ở người bị trầm cảm. Mức độ nghiêm trọng của mệt mỏi có thể là một yếu tố phân biệt quan trọng, đặc biệt khi kết hợp với các triệu chứng khác như PEM.
* Sleep\_Quality\_Index  
  Chất lượng giấc ngủ trung bình là 5.47, với khoảng giá trị từ 1 đến 10. Rối loạn giấc ngủ là triệu chứng thường gặp ở cả ME/CFS và trầm cảm, nhưng bản chất rối loạn giấc ngủ có thể khác nhau giữa hai nhóm, do đó biến này có thể đóng vai trò là một chỉ báo quan trọng.

***Thống kê biến mục tiêu diagnosis***

* Kết quả thống kê cho thấy phân bố các nhãn chẩn đoán như sau:

|  |  |  |
| --- | --- | --- |
| **Diagnosis** | **Số lượng (Count)** | **Tỷ lệ (%)** |
| Depression | 401 | 40.1% |
| ME/CFS | 397 | 39.7% |
| Both | 202 | 20.2% |

* **Nhận xét về biến diagnosis:**
* Cột diagnosis là biến mục tiêu, cho biết bệnh nhân được chẩn đoán mắc ME/CFS, Trầm cảm hoặc cả hai.
* Sự phân bố giữa các nhóm bệnh tương đối cân bằng, trong đó nhóm Depression chiếm 40.1%, nhóm ME/CFS chiếm 39.7% và nhóm Both chiếm 20.2%.
* Việc phân bố gần như đồng đều giữa hai nhóm chính (Depression và ME/CFS) giúp mô hình học máy tiềm năng giảm nguy cơ thiên lệch. Tuy nhiên, nhóm Both có tỷ lệ thấp hơn, điều này có thể khiến mô hình khó phân loại chính xác hơn đối với nhóm này.
* Trong quá trình huấn luyện mô hình, cần lưu ý đến nhóm thiểu số để tránh hiện tượng mô hình thiên về dự đoán nhóm đa số.

**2.2. Phân tích dữ liệu categorical**

Nhóm tiến hành thống kê mô tả trên các biến dạng phân loại (categorical) gồm: gender, work\_status, social\_activity\_level, exercise\_frequency, meditation\_or\_mindfulness, diagnosis.

*Bảng tổng quan các cột categorical*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cột** | **Số bản ghi** | **Số giá trị thiếu** | **Số giá trị hợp lệ** | **Số loại giá trị** |
| gender | 1000 | 0 | 1000 | 2 |
| work\_status | 1000 | 47 | 953 | 3 |
| social\_activity\_level | 1000 | 40 | 960 | 5 |
| exercise\_frequency | 1000 | 39 | 961 | 5 |
| meditation\_or\_mindfulness | 1000 | 11 | 989 | 2 |
| diagnosis | 1000 | 0 | 1000 | 3 |

*Nhận xét:* Một số cột có giá trị thiếu, tỷ lệ dao động từ 1% đến 5% (ví dụ: work\_status có 4.7% giá trị thiếu). Điều này cần được cân nhắc xử lý trước khi huấn luyện mô hình.

***Thống kê chi tiết từng cột***

1. Gender

* Male: 524 bản ghi (52.4%)
* Female: 476 bản ghi (47.6%)  
  Phân bố giới tính khá cân bằng, giúp mô hình tránh thiên lệch theo giới.

2. Work Status

* Partially working: 335 (33.5%)
* Working: 320 (32.0%)
* Not working: 298 (29.8%)
* Thiếu dữ liệu: 47 (4.7%)  
  Các nhóm phân bố tương đối đồng đều, nhưng cần xử lý dữ liệu thiếu.

3. Social Activity Level

* Very low: 203 (20.3%)
* High: 200 (20.0%)
* Low: 190 (19.0%)
* Very high: 186 (18.6%)
* Medium: 181 (18.1%)
* Thiếu dữ liệu: 40 (4.0%)  
  Phân bố mức độ hoạt động xã hội khá đều, không có nhóm nào chiếm ưu thế vượt trội.

4. Exercise Frequency

* Rarely: 210 (21.0%)
* Often: 202 (20.2%)
* Daily: 194 (19.4%)
* Sometimes: 184 (18.4%)
* Never: 171 (17.1%)
* Thiếu dữ liệu: 39 (3.9%)  
  Phân bố khá đồng đều, có thể sử dụng làm đặc trưng tiềm năng.

5. Meditation or Mindfulness

* Yes: 505 (50.5%)
* No: 484 (48.4%)
* Thiếu dữ liệu: 11 (1.1%)  
  Phân bố gần như cân bằng giữa hai nhóm, ít khả năng gây thiên lệch.

6. Diagnosis (biến mục tiêu)

* Depression: 401 (40.1%)
* ME/CFS: 397 (39.7%)
* Both: 202 (20.2%)  
  Phân bố tương đối cân bằng, trong đó nhóm Both chiếm tỷ lệ thấp hơn.

**Nhận xét tổng quan:**

Các biến dạng phân loại trong dữ liệu có phân bố tương đối đồng đều, không có sự mất cân đối nghiêm trọng giữa các nhóm. Một số cột có giá trị thiếu cần được xử lý để đảm bảo chất lượng dữ liệu đầu vào cho mô hình. Phân bố cân bằng này là điều kiện thuận lợi để huấn luyện mô hình học máy mà không gặp nhiều vấn đề về thiên lệch lớp.

**2.3. Trực quan hóa dữ liệu**

Để hiểu rõ hơn về sự phân bố và mối quan hệ giữa các thuộc tính, nhóm tiến hành trực quan hóa dữ liệu với nhiều loại biểu đồ khác nhau. Việc trực quan hóa giúp làm nổi bật các xu hướng, kiểm tra sự cân bằng dữ liệu và phát hiện các giá trị ngoại lai.

**2.3.1. Histogram các biến định lượng**

Các biểu đồ histogram thể hiện phân bố của các biến số định lượng như age, sleep\_quality\_index, brain\_fog\_level, physical\_pain\_score, stress\_level, depression\_phq9\_score, fatigue\_severity\_scale\_score, pem\_duration\_hours, và hours\_of\_sleep\_per\_night.

***Nhận xét:***

* Phần lớn các biến có phân bố khá đồng đều.
* Riêng biến depression\_phq9\_score có xu hướng tập trung nhiều ở mức điểm 7–10.
* Biến fatigue\_severity\_scale\_score có đỉnh tập trung rõ rệt ở mức 7–8.

A group of blue bars

AI-generated content may be incorrect.

Biểu đồ 1: Biểu đồ tổng quan Histogram

**2.3.2. Boxplot các biến định lượng**

Các boxplot thể hiện phân bố trung vị, tứ phân vị và giá trị ngoại lai của các biến định lượng.

***Nhận xét:***

* depression\_phq9\_score xuất hiện một số giá trị ngoại lai vượt trội (khoảng trên 27 điểm).
* fatigue\_severity\_scale\_score xuất hiện nhiều giá trị ngoại lai.
* pem\_duration\_hours cũng có độ phân tán lớn và tồn tại ngoại lai.

*A group of blue squares

AI-generated content may be incorrect.*

**A group of blue squares

AI-generated content may be incorrect.**

**A group of blue squares

AI-generated content may be incorrect.**

Biểu đồ 2: Biểu đồ Boxplot

**2.3.3. Bar chart các biến phân loại theo nhóm chẩn đoán**

Các biểu đồ cột dưới đây thể hiện tần suất từng giá trị của các biến phân loại (gender, work\_status, meditation\_or\_mindfulness, exercise\_frequency, social\_activity\_level) phân tách theo nhóm chẩn đoán.

***Nhận xét:***

* Phân bố giới tính giữa các nhóm chẩn đoán khá cân bằng.
* work\_status và social\_activity\_level không có nhóm nào chiếm ưu thế vượt trội, nhưng có sự khác biệt nhẹ giữa các nhóm chẩn đoán.
* Tần suất tập luyện (exercise\_frequency) và thói quen thiền/chánh niệm (meditation\_or\_mindfulness) cũng phân bố tương đối đồng đều giữa các nhóm.

*A group of colorful bars

AI-generated content may be incorrect.*

Biểu đồ 3: Biểu đồ Barchart

**2.3.4. Scatter plot giữa các biến số định lượng và biến mục tiêu**

Nhằm quan sát rõ hơn mối quan hệ giữa các biến số định lượng với biến mục tiêu diagnosis, nhóm thực hiện các biểu đồ scatter plot. Trục hoành thể hiện biến mục tiêu diagnosis, trục tung thể hiện giá trị của từng biến số định lượng.

***Nhận xét từ các biểu đồ:***

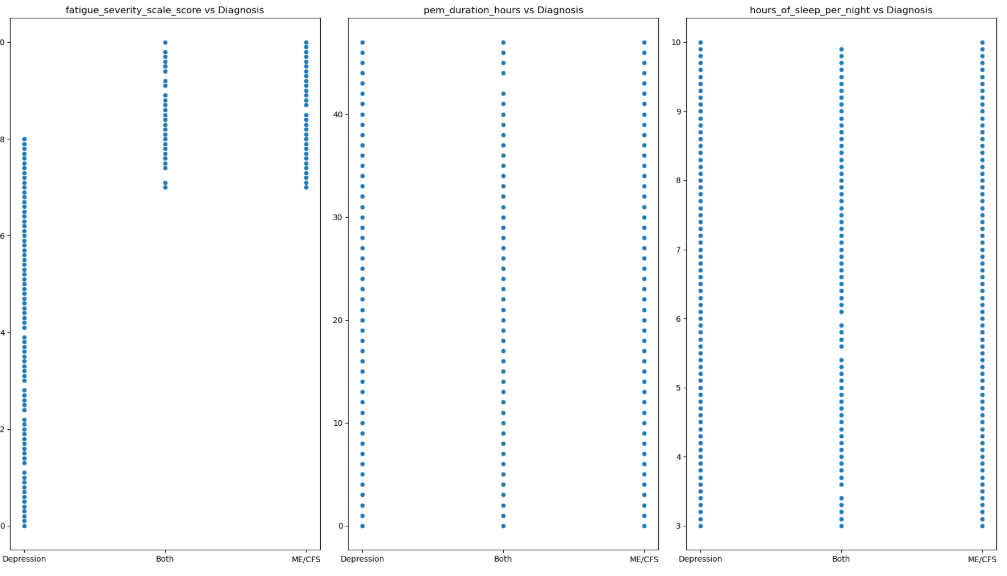
* Đa số các biến (age, sleep\_quality\_index, brain\_fog\_level, physical\_pain\_score, stress\_level,hours\_of\_sleep\_per\_night, pem\_duration\_hours) phân bố tương đối đồng đều giữa các nhóm chẩn đoán, không thể hiện sự chênh lệch rõ rệt.
* Depression\_phq9\_score:
  + Nhóm Depression và Both tập trung chủ yếu ở mức điểm trên 10.
  + Nhóm ME/CFS tập trung ở khoảng 0–10 điểm.  
    → Đây là biến có khả năng phân biệt tương đối rõ giữa các nhóm.
* Fatigue\_severity\_scale\_score:
  + Nhóm Depression tập trung chủ yếu ở khoảng 0–8 điểm.
  + Nhóm ME/CFS và Both có xu hướng tập trung ở mức điểm cao hơn, khoảng7–10.  
    → Biến này cũng cho thấy sự khác biệt giữa các nhóm, có thể đóng vai trò quan trọng trong mô hình phân loại.

A group of blue dots

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.



Biểu đồ 4: Biểu đồ Scatter plot

**2.4. Phân tích mối quan hệ giữa các thuộc tính**

Để kiểm tra mối quan hệ giữa các biến số định lượng, nhóm đã loại bỏ toàn bộ các dòng chứa giá trị khuyết trong các cột numerical trước khi tính toán ma trận tương quan.

***Số dòng dữ liệu còn lại sau khi loại bỏ giá trị null:***  
 761 dòng dữ liệu hợp lệ được sử dụng để tính toán ma trận tương quan.

***Biểu đồ heatmap tương quan***

Biểu đồ dưới đây thể hiện hệ số tương quan Pearson giữa các cặp biến số định lượng trong dữ liệu:

A screenshot of a graph

AI-generated content may be incorrect.

Biểu đồ 5: Biểu đồ Heatmap

***Nhận xét:***

* fatigue\_severity\_scale\_score và pem\_present: hệ số tương quan **0.59**, thể hiện mối tương quan thuận khá mạnh. Người có điểm mệt mỏi càng cao thì khả năng xuất hiện PEM (Post-Exertional Malaise) càng lớn.
* depression\_phq9\_score và fatigue\_severity\_scale\_score: hệ số tương quan **-0.20**, cho thấy mối tương quan nghịch nhẹ. Mức độ trầm cảm cao có xu hướng đi kèm với mức độ mệt mỏi thấp hơn.
* depression\_phq9\_score và pem\_present: hệ số tương quan **-0.35**, cho thấy mối tương quan nghịch trung bình. Người có điểm trầm cảm cao thì khả năng xuất hiện PEM giảm.
* Các mối tương quan còn lại đều dao động từ **-0.08** đến **0.02**, cho thấy gần như không có mối quan hệ đáng kể giữa các biến đó.

***Kết luận:***  
 Phân tích này giúp nhóm xác định được những cặp biến có mối quan hệ đáng chú ý, từ đó gợi ý lựa chọn đặc trưng cho các bước xây dựng mô hình sau này.

Dưới đây là phần trình bày chuyên nghiệp cho mục Phân tích mối quan hệ giữa các biến số numerical và categorical kèm nhận xét của bạn được diễn đạt lại mạch lạc, phù hợp đưa vào báo cáo:

**2.5. Phân tích mối quan hệ giữa các biến số numerical và categorical**

Nhóm tiến hành phân tích phân bố các biến số định lượng theo từng nhóm của biến phân loại (diagnosis) thông qua các biểu đồ boxplot và scatter plot. Mục đích nhằm quan sát sự khác biệt về trung vị, khoảng giá trị và phát hiện các giá trị ngoại lai giữa các nhóm.

***Nhận xét:***

* Đối với hai biến fatigue\_severity\_scale\_score và depression\_phq9\_score: Mặc dù phân bố trung vị của các nhóm chẩn đoán trong biến categorical diagnosis khá tương đồng và không thể hiện sự khác biệt quá rõ rệt, nhưng xuất hiện một số giá trị ngoại lai (outliers) rất cao hoặc rất thấp. Các giá trị ngoại lai này có thể ảnh hưởng trực tiếp đến kết quả phân tích và cần được cân nhắc xử lý trong giai đoạn tiền xử lý dữ liệu.
* Đối với các biến numerical khác: Phân bố giữa các nhóm chẩn đoán hầu như tương đương nhau, không xuất hiện sự khác biệt đáng kể và cũng không quan sát thấy nhiều giá trị ngoại lai bất thường.

***Kết luận:***  
 Kết quả này cho thấy phần lớn các biến numerical có sự phân bố đồng đều giữa các nhóm categorical. Tuy nhiên, cần đặc biệt chú ý xử lý hoặc đánh giá ảnh hưởng của các outliers ở fatigue\_severity\_scale\_score và depression\_phq9\_score để tránh gây nhiễu cho mô hình học máy trong các bước tiếp theo.

**CHƯƠNG 3: NHẬN XÉT & ĐỊNH HƯỚNG**

**3.1. Tóm tắt các phát hiện quan trọng từ dữ liệu**

* Dữ liệu gồm cả biến số định lượng và định tính, trong đó số lượng bản ghi giữa các nhóm chẩn đoán khá cân bằng. Tuy nhiên, nhóm Depression và ME/CFS chiếm tỷ lệ lớn hơn, trong khi nhóm Both có tỷ lệ thấp hơn.
* Các giá trị trong các biến dạng phân loại (categorical) nhìn chung cân bằng, do đó bản thân các biến này không phải là yếu tố phân loại quá mạnh cho biến mục tiêu diagnosis.
* Một số biến định lượng có nhiều giá trị ngoại lai (ví dụ: fatigue\_severity\_scale\_score, depression\_phq9\_score) và một lượng nhỏ giá trị thiếu, điều này cần được xử lý trước khi huấn luyện mô hình.
* Ma trận tương quan cho thấy tồn tại một số mối liên hệ đáng chú ý, điển hình như:
  + fatigue\_severity\_scale\_score và pem\_present có tương quan thuận khá mạnh.
  + depression\_phq9\_score và pem\_present có tương quan nghịch trung bình.

**3.2. Những đặc trưng có thể ảnh hưởng lớn đến mô hình**

* fatigue\_severity\_scale\_score: Thể hiện rõ sự khác biệt giữa các nhóm chẩn đoán, có nhiều thông tin quan trọng.
* depression\_phq9\_score: Phản ánh mức độ trầm cảm, có sự phân bố khác nhau giữa các nhóm.
* pem\_present: Có mối liên hệ khá mạnh với mức độ mệt mỏi.
* Các biến categorical như gender, work\_status, social\_activity\_level, exercise\_frequency, meditation\_or\_mindfulness có phân bố khá đồng đều nên ít khả năng phân loại mạnh, nhưng vẫn có thể giữ lại để kiểm chứng.

**3.3. Nhận xét chung khi phân tích mối quan hệ**

* Dữ liệu có một lượng giá trị null nhưng ảnh hưởng không đáng kể đến tổng thể phân tích.
* Xuất hiện nhiều giá trị ngoại lai, ảnh hưởng đến một số biến quan trọng.
* Tồn tại các mối quan hệ cả mạnh và yếu giữa các biến, cần cân nhắc khi lựa chọn đặc trưng.

**3.4. Giải pháp tiền xử lý dữ liệu**

Để nâng cao chất lượng dữ liệu đầu vào cho mô hình học máy, nhóm đề xuất:

* Loại bỏ các bản ghi có giá trị null hoặc thay thế giá trị thiếu bằng median hoặc mean.
* Chuyển đổi các giá trị dạng chữ (categorical) sang số thông qua mã hóa (encoding).
* Xử lý các giá trị ngoại lai bằng cách giới hạn (capping) hoặc loại bỏ.
* Cân nhắc loại bỏ một số đặc trưng ít liên quan đến biến mục tiêu.

**3.5. Định hướng mô hình học máy**

* Loại bài toán: Phân loại đa lớp (diagnosis: Depression, ME/CFS, Both).
* Mô hình đề xuất: RandomForestClassifier, vì:
  + Phù hợp với dữ liệu phức tạp, không đòi hỏi tính tuyến tính.
  + Có khả năng xử lý tốt dữ liệu hỗn hợp (biến số và biến phân loại).
  + Chống overfitting tốt hơn nhiều mô hình khác.
  + Có thể tự động đánh giá và lựa chọn các đặc trưng quan trọng.

**3.6. Bài học rút ra**

Qua quá trình thực hiện, nhóm đã học được:

* Hiểu rõ hơn về quy trình phân tích dữ liệu từ bước làm sạch, khám phá đến trực quan hóa.
* Cải thiện kỹ năng trực quan hóa dữ liệu để nhận diện mối liên hệ giữa các biến.
* Biết cách xử lý các vấn đề thường gặp trong dữ liệu, điển hình như:
  + Thiếu dữ liệu (missing values) và cách thay thế bằng giá trị trung bình, trung vị hoặc loại bỏ.
  + Phát hiện và xử lý các giá trị ngoại lai để tránh ảnh hưởng tiêu cực đến mô hình.

**3.7. Hướng phát triển trong tương lai**

* Tiếp tục mở rộng tập dữ liệu từ nhiều nguồn khác nhau để tăng tính tổng quát và độ tin cậy của mô hình.
* Thử nghiệm nhiều mô hình khác, bao gồm cả machine learning và deep learning, để tối ưu kết quả dự đoán.
* Nghiên cứu triển khai một hệ thống dự đoán theo thời gian thực, ví dụ: ứng dụng hỗ trợ tự đánh giá sức khỏe tinh thần cho người dùng.
* Tích hợp thêm dữ liệu từ các thiết bị đeo tay (wearables) như nhịp tim, giấc ngủ, mức vận động… nhằm cải thiện độ chính xác của mô hình.

# **KẾT LUẬN**

Qua quá trình thực hiện đề tài, nhóm đã tiến hành đầy đủ các bước từ làm sạch dữ liệu, phân tích khám phá (EDA) cho đến trực quan hóa nhằm hiểu rõ đặc điểm của tập dữ liệu liên quan đến hội chứng mệt mỏi mạn tính (ME/CFS) và trầm cảm. Các phân tích đã chỉ ra rằng phần lớn các biến số có phân bố tương đối đồng đều, trong đó một số đặc trưng như fatigue\_severity\_scale\_score, depression\_phq9\_score và pem\_present thể hiện mối quan hệ đáng kể với biến mục tiêu, qua đó hứa hẹn trở thành những yếu tố quan trọng trong việc xây dựng mô hình phân loại.

Bên cạnh đó, quá trình phân tích cũng giúp nhóm nhận diện được các vấn đề cần xử lý trong dữ liệu như sự tồn tại của giá trị thiếu và các ngoại lai, đồng thời đề xuất những giải pháp tiền xử lý phù hợp. Dựa trên đặc điểm dữ liệu và yêu cầu bài toán, nhóm đề xuất sử dụng mô hình Random Forest để tiến hành phân loại, bởi mô hình này phù hợp với dữ liệu phức tạp, xử lý tốt cả biến định tính và định lượng, đồng thời hạn chế được tình trạng quá khớp.

Đề tài không chỉ giúp nhóm củng cố kiến thức về quy trình phân tích dữ liệu và ứng dụng học máy mà còn mở ra những hướng phát triển tiếp theo, bao gồm mở rộng nguồn dữ liệu, thử nghiệm thêm nhiều mô hình khác nhau và hướng tới xây dựng hệ thống dự đoán hỗ trợ chẩn đoán bệnh trong thực tế. Nhóm hy vọng kết quả nghiên cứu này sẽ là nền tảng hữu ích cho các nghiên cứu sâu hơn, góp phần thúc đẩy việc ứng dụng công nghệ dữ liệu trong lĩnh vực y tế ngày càng hiệu quả và thiết thực.

**LỜI CẢM ƠN**

Nhóm 09 xin chân thành cảm ơn Anh Trịnh Bá Tú đã tận tình hỗ trợ, định hướng và góp ý để nhóm hoàn thiện đề tài. Chúng mình cũng xin cảm ơn các thành viên trong nhóm đã phối hợp chặt chẽ, nỗ lực và chia sẻ kinh nghiệm trong suốt quá trình thực hiện. Sự giúp đỡ và hợp tác quý báu này là động lực để nhóm hoàn thành báo cáo một cách tốt nhất.

# **TÀI LIỆU THAM KHẢO**

* + 1. Mohamedasak. (2025, July 12). *ME/CFS vs Depression RF and XGB Classification*. Kaggle. https://www.kaggle.com/code/mohamedasak/me-cfs-vs-depression-rf-and-xgb-classification
    2. ME/CFS vs Depression Classification Dataset. (2025, June 8). Kaggle. <https://www.kaggle.com/datasets/storytellerman/mecfs-vs-depression-classification-dataset>
    3. Programming with Mosh. (2020, September 17). Python Machine Learning Tutorial (Data Science) [Video]. YouTube. <https://www.youtube.com/watch?v=7eh4d6sabA0>