**ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**



**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH DỰ ĐOÁN ẢNH HƯỞNG CỦA THỜI GIAN SỬ DỤNG MÀN HÌNH THIẾT BỊ ĐIỆN TỬ TRÊN**

**MÔ HÌNH HỒI QUY LOGISTIC**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn**  **Lớp**  **Nhóm**  **Sinh viên thực hiện** | **: TS. Nguyễn Mạnh Cường**  : 20251IT6077002  : 03  **:** 1. Nguyễn Thị Ngọc Châm - 2022601173  2. Nguyễn Ngọc Huyền - 2022600391  3. Phạm Thị Phương - 2022606751 |

**Hà Nội, tháng 12 năm 2025**

**ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**



**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH DỰ ĐOÁN ẢNH HƯỞNG CỦA THỜI GIAN SỬ DỤNG MÀN HÌNH THIẾT BỊ ĐIỆN TỬ TRÊN**

**MÔ HÌNH HỒI QUY LOGISTIC**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn**  **Lớp**  **Nhóm**  **Sinh viên thực hiện** | **: TS. Nguyễn Mạnh Cường**  : 20251IT6077002  : 03  **:** 1. Nguyễn Thị Ngọc Châm - 2022601173  2. Nguyễn Ngọc Huyền - 2022600391  3. Phạm Thị Phương - 2022606751 |

**Hà Nội, tháng 12 năm 2025**

**PHIẾU HỌC TẬP CÁ NHÂN/NHÓM**

**I. Thông tin chung**

1. Tên lớp : HTTT01. Khóa : 17.

2. Họ và tên sinh viên: **Nguyễn Thị Ngọc Châm.** Mã sinh viên: 2022601173.

3. Tên nhóm : Nhóm 3.

**II. Nội dung học tập**

**1. Tên chủ đề : Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy logistic.**

**2. Hoạt động của sinh viên:**

-Hoạt động/Nội dung 1: *Thành lập nhóm học tập, lập kế hoạch làm bài tập lớn, thực hiện nghiên cứu phát biểu bài toán.*

Mục tiêu/chuẩn đầu ra: L3.

-Hoạt động/Nội dung 2: *Tìm hiểu các kỹ thuật phổ biến giải quyết bài toán. Tìm hiểu kỹ thuật chính sẽ sử dụng để thực nghiệm trong bài tập lớn.*

Mục tiêu/chuẩn đầu ra: L3.

-Hoạt động/Nội dung 3: *Tiến hành các bước thu thập, tiền xử lý dữ liệu; sử dụng các công cụ phù hợp để thực nghiệm; tổng hợp, so sánh, đánh giá kết quả. Xây dựng chương trình demo (nếu có), viết báo cáo thí nghiệm/ thực nghiệm.*

Mục tiêu/chuẩn đầu ra: L3.

**3. Sản phẩm nghiên cứu:** *Báo cáo thí nghiệm/ Thực nghiệm + Chương trình demo (nếu có).*

**III. Nhiệm vụ học tập**

1. Tổng hợp kiến thức đã học trong học phần, ứng dụng kiến thức, kỹ năng học được để giải quyết một bài toán thực tế liên quan tới phân tích dữ liệu.
2. Hoàn thành bài tập lớn theo đúng thời gian quy định (từ ngày 21/10/2025, đến ngày 21/12/2025).
3. Nộp bài và báo cáo sản phẩm theo chủ đề được giao trước giảng viên và những sinh viên khác.

**IV. Học liệu thực hiện báo cáo**

1. Tài liệu học tập : Các tài liệu hướng dẫn thực hiện bài tập lớn do giảng viên cung cấp, các tài liệu, code mẫu tham khảo trong bài giảng/ bài thực hành và trên mạng internet.

2. Phương tiện, nguyên liệu thực hiện bài tập lớn: sử dụng các công cụ phù hợp: Excel, Weka, PyCharm, Anaconda, Jupiter Notebook, Google Collab, R, Apache Hadoop, Apache Spark……

**PHIẾU HỌC TẬP CÁ NHÂN/NHÓM**

**I. Thông tin chung**

1. Tên lớp : HTTT01. Khóa : 17.

2. Họ và tên sinh viên: **Nguyễn Ngọc Huyền.** Mã sinh viên: 2022600391.

3. Tên nhóm : Nhóm 3.

**II. Nội dung học tập**

**1. Tên chủ đề : Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy logistic.**

**2. Hoạt động của sinh viên:**

-Hoạt động/Nội dung 1: *Thành lập nhóm học tập, lập kế hoạch làm bài tập lớn, thực hiện nghiên cứu phát biểu bài toán.*

Mục tiêu/chuẩn đầu ra: L3.

-Hoạt động/Nội dung 2: *Tìm hiểu các kỹ thuật phổ biến giải quyết bài toán. Tìm hiểu kỹ thuật chính sẽ sử dụng để thực nghiệm trong bài tập lớn.*

Mục tiêu/chuẩn đầu ra: L3.

-Hoạt động/Nội dung 3: *Tiến hành các bước thu thập, tiền xử lý dữ liệu; sử dụng các công cụ phù hợp để thực nghiệm; tổng hợp, so sánh, đánh giá kết quả. Xây dựng chương trình demo (nếu có), viết báo cáo thí nghiệm/ thực nghiệm.*

Mục tiêu/chuẩn đầu ra: L3.

**3. Sản phẩm nghiên cứu:** *Báo cáo thí nghiệm/ Thực nghiệm + Chương trình demo (nếu có).*

**III. Nhiệm vụ học tập**

1. Tổng hợp kiến thức đã học trong học phần, ứng dụng kiến thức, kỹ năng học được để giải quyết một bài toán thực tế liên quan tới phân tích dữ liệu.
2. Hoàn thành bài tập lớn theo đúng thời gian quy định (từ ngày 21/10/2025, đến ngày 21/12/2025).
3. Nộp bài và báo cáo sản phẩm theo chủ đề được giao trước giảng viên và những sinh viên khác.

**IV. Học liệu thực hiện báo cáo**

1. Tài liệu học tập : Các tài liệu hướng dẫn thực hiện bài tập lớn do giảng viên cung cấp, các tài liệu, code mẫu tham khảo trong bài giảng/ bài thực hành và trên mạng internet.

2. Phương tiện, nguyên liệu thực hiện bài tập lớn: sử dụng các công cụ phù hợp: Excel, Weka, PyCharm, Anaconda, Jupiter Notebook, Google Collab, R, Apache Hadoop, Apache Spark……

**PHIẾU HỌC TẬP CÁ NHÂN/NHÓM**

**I. Thông tin chung**

1. Tên lớp : HTTT02. Khóa : 17.

2. Họ và tên sinh viên: **Phạm Thị Phương.** Mã sinh viên: 2022606751.

3. Tên nhóm : Nhóm 3.

**II. Nội dung học tập**

**1. Tên chủ đề : Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy logistic.**

**2. Hoạt động của sinh viên:**

-Hoạt động/Nội dung 1: *Thành lập nhóm học tập, lập kế hoạch làm bài tập lớn, thực hiện nghiên cứu phát biểu bài toán.*

Mục tiêu/chuẩn đầu ra: L3.

-Hoạt động/Nội dung 2: *Tìm hiểu các kỹ thuật phổ biến giải quyết bài toán. Tìm hiểu kỹ thuật chính sẽ sử dụng để thực nghiệm trong bài tập lớn.*

Mục tiêu/chuẩn đầu ra: L3.

-Hoạt động/Nội dung 3: *Tiến hành các bước thu thập, tiền xử lý dữ liệu; sử dụng các công cụ phù hợp để thực nghiệm; tổng hợp, so sánh, đánh giá kết quả. Xây dựng chương trình demo (nếu có), viết báo cáo thí nghiệm/ thực nghiệm.*

Mục tiêu/chuẩn đầu ra: L3.

**3. Sản phẩm nghiên cứu:** *Báo cáo thí nghiệm/ Thực nghiệm + Chương trình demo (nếu có).*

**III. Nhiệm vụ học tập**

1. Tổng hợp kiến thức đã học trong học phần, ứng dụng kiến thức, kỹ năng học được để giải quyết một bài toán thực tế liên quan tới phân tích dữ liệu.
2. Hoàn thành bài tập lớn theo đúng thời gian quy định (từ ngày 21/10/2025, đến ngày 21/12/2025).
3. Nộp bài và báo cáo sản phẩm theo chủ đề được giao trước giảng viên và những sinh viên khác.

**IV. Học liệu thực hiện báo cáo**

1. Tài liệu học tập : Các tài liệu hướng dẫn thực hiện bài tập lớn do giảng viên cung cấp, các tài liệu, code mẫu tham khảo trong bài giảng/ bài thực hành và trên mạng internet.

2. Phương tiện, nguyên liệu thực hiện bài tập lớn: sử dụng các công cụ phù hợp: Excel, Weka, PyCharm, Anaconda, Jupiter Notebook, Google Collab, R, Apache Hadoop, Apache Spark……

**KẾ HOẠCH THỰC HIỆN BÀI TẬP LỚN**

Tên lớp: 20251IT6077002 Khóa: K17.

Họ và tên sinh viên:

1. Nguyễn Thị Ngọc Châm
2. Nguyễn Ngọc Huyền
3. Phạm Thị Phương

Tên nhóm: Nhóm 03

Tên chủ đề: Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy logistic

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tuần** | **Người thực hiện** | **Nội dung công việc** | **Kết quả đạt được** | **Phương pháp thực hiện** |
| 1 | Cả nhóm | Tìm và thu thập tài liệu cho bài toán. Tìm hiểu yêu cầu về bài toán. | Tìm được bộ dữ liệu thích hợp với yêu cầu bài toán | Sưu tầm tài liệu  Sưu tầm dữ liệu  Nghiên cứu tài liệu |
| 2 | Phạm Thị Phương | Viết báo cáo chương 1 | Hoàn thành viết báo cáo về tổng quan, giới thiệu bài toán | Nghiên cứu bài toán  Viết tài liệu |
| 3 | Phạm Thị Phương | Tiền xử lý dữ liệu | Hoàn thành việc tiền xử lý dữ liệu (loại bỏ được các giá trị trùng lặp và điền khuyết) | Nghiên cứu tài liệu  Thực nghiệm |
| 3 | Nguyễn Ngọc Huyền | Nghiên cứu về mô hình dự báo Logistic | Hoàn thành | Nghiên cứu tài liệu  Thực nghiệm |
| 3 | Nguyễn Thị Ngọc Châm | Nghiên cứu các biểu đồ phân tích mô tả | Hoàn thành | Nghiên cứu tài liệu  Viết báo cáo |
| 4 | Nguyễn Thị Ngọc Châm | Phân tích mô tả | Hoàn thành việc phân tích mô tả trên tập dữ liệu đã tiền xử lý | Nghiên cứu tài liệu |
| 4 | Nguyễn Ngọc Huyền | Viết báo cáo phần lý thuyết phân tích dự báo | Hoàn thành | Viết báo cáo |
| 5 | Nguyễn Thị Ngọc Châm | Vẽ các mô hình trong phần phân tích mô tả | Hoàn thành vẽ các mô hình | Thực nghiệm |
| 5 | Nguyễn Ngọc Huyền | Thực hiện phân tích mô tả | Hoàn thành việc xây dựng mô hình, huấn luyện dữ liệu và đánh giá mô hình | Thực nghiệm |
| 6 | Phạm Thị Phương | Code đề mô sản phẩm | Hoàn thành việc xây dựng website | Nghiên cứu tài liệu  Thực nghiệm |
| 6 | Nguyễn Thị Ngọc Châm | Demo dữ liệu với Power BI | Hoàn thành xử lý dữ liệu bằng Power Query, vẽ biểu đồ bằng Power BI | Nghiên cứu tài liệu  Thực nghiệm |
| 6 | Nguyễn Ngọc Huyền | Chỉnh sửa báo cáo | Hoàn thành | Viết báo cáo |
| 7 | Cả nhóm | Họp hoàn thiện | Hoàn thiện báo cáo | Họp kết thúc |

Ngày     tháng    năm 2025

**XÁC NHẬN CỦA GIẢNG VIÊN**

(Kí, ghi rõ họ tên)

Nguyễn Mạnh Cường

**BÁO CÁO HOẠT ĐỘNG NHÓM**

Tên lớp: 20251IT6077002 Khóa: K17.

Họ và tên sinh viên:

1. Nguyễn Thị Ngọc Châm

2. Nguyễn Ngọc Huyền

3. Phạm Thị Phương

Tên nhóm: Nhóm 03

Tên chủ đề: Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy logistic

| **Tuần** | **Người thực hiện** | **Nội dung công việc** | **Kết quả đạt được** | **Kiến nghị với giảng viên hướng dẫn** |
| --- | --- | --- | --- | --- |
| 1 | Cả nhóm | Tìm và thu thập tài liệu cho bài toán. Tìm hiểu yêu cầu về bài toán. | Hoàn thành |  |
| 2 | Phạm Thị Phương | Viết báo cáo chương 1 | Hoàn thành |  |
| 3 | Phạm Thị Phương | Tiền xử lý dữ liệu | Hoàn thành |  |
| 3 | Nguyễn Ngọc Huyền | Nghiên cứu về mô hình dự báo Logistic | Hoàn thành |  |
| 3 | Nguyễn Thị Ngọc Châm | Nghiên cứu các biểu đồ phân tích mô tả | Hoàn thành |  |
| 4 | Nguyễn Thị Ngọc Châm | Phân tích mô tả | Hoàn thành |  |
| 4 | Nguyễn Ngọc Huyền | Viết báo cáo phần lý thuyết phân tích dự báo | Hoàn thành |  |
| 5 | Nguyễn Thị Ngọc Châm | Vẽ các mô hình trong phần phân tích mô tả | Hoàn thành |  |
| 5 | Nguyễn Ngọc Huyền | Thực hiện phân tích mô tả | Hoàn thành |  |
| 6 | Phạm Thị Phương | Code đề mô sản phẩm | Hoàn thành |  |
| 6 | Nguyễn Thị Ngọc Châm | Demo dữ liệu với Power BI | Hoàn thành |  |
| 6 | Nguyễn Ngọc Huyền | Chỉnh sửa báo cáo | Hoàn thành |  |
| 7 | Cả nhóm | Họp hoàn thiện | Hoàn thành |  |

Ngày tháng năm 2025

**XÁC NHẬN CỦA GIẢNG VIÊN**

(Kí, ghi rõ họ tên)

Nguyễn Mạnh Cường

# MỤC LỤC

[MỤC LỤC i](#_Toc217250383)

[DANH MỤC HÌNH ẢNH iv](#_Toc217250384)

[DANH MỤC BẢNG BIỂU vii](#_Toc217250385)

[LỜI CẢM ƠN 1](#_Toc217250386)

[LỜI NÓI ĐẦU 2](#_Toc217250387)

[CHƯƠNG 1: TỔNG QUAN VỀ VẤN ĐỀ VẤN ĐỀ NGHIÊN CỨU 4](#_Toc217250388)

[1.1. Kiến thức tổng quan 4](#_Toc217250389)

[1.1.1. Trí tuệ nhân tạo ( Artificial Intelligence) 4](#_Toc217250390)

[1.1.2. Học máy (Machine Learning) 6](#_Toc217250391)

[1.1.3. Phân tích dữ liệu là gì? 8](#_Toc217250392)

[1.2. Tổng quan về bài toán phân tích dự đoán 9](#_Toc217250393)

[1.3. Bài toán dự đoán ảnh hưởng của thời gian sử dụng màn hình 11](#_Toc217250394)

[1.3.1. Giới thiệu chung 11](#_Toc217250395)

[1.3.2. Mục đích, ý nghĩa của đề tài 11](#_Toc217250396)

[1.3.3. Đầu vào và đầu ra của bài toán 13](#_Toc217250397)

[1.3.4. Thuận lợi và khó khăn khi giải quyết bài toán 14](#_Toc217250398)

[1.3.5. Ứng dụng của việc giải quyết bài toán 14](#_Toc217250399)

[CHƯƠNG 2: CÁC PHƯƠNG PHÁP VÀ KĨ THUẬT 16](#_Toc217250400)

[2.1. Tiền xử lý dữ liệu 16](#_Toc217250401)

[2.2. Các phương pháp phân tích 17](#_Toc217250402)

[2.2.1. Phương pháp phân tích mô tả 17](#_Toc217250403)

[2.2.1.1. Phương pháp phân tích trên từng biến 18](#_Toc217250404)

[2.2.1.2. Phương pháp phân trên nhiều biến 19](#_Toc217250405)

[2.2.2. Phương pháp phân tích hồi quy 21](#_Toc217250406)

[2.3. Các công cụ 22](#_Toc217250407)

[2.3.1. Ngôn ngữ Python 22](#_Toc217250408)

[2.3.2. Weka 23](#_Toc217250409)

[2.3.3. Excel 24](#_Toc217250410)

[2.4. Kết luận chương 2 25](#_Toc217250411)

[CHƯƠNG 3: THỰC NGHIỆM 26](#_Toc217250412)

[3.1. Dữ liệu thực nghiệm 26](#_Toc217250413)

[3.2. Môi trường thực nghiệm 29](#_Toc217250414)

[3.2.1. Google Colab 29](#_Toc217250415)

[3.2.2. PyCharm 30](#_Toc217250416)

[3.2.3. Các thư viện sử dụng 31](#_Toc217250417)

[3.3. Quy trình thực nghiệm 32](#_Toc217250418)

[3.3.1. Đặt mục tiêu 32](#_Toc217250419)

[3.3.2. Tiền xử lý dữ liệu 32](#_Toc217250420)

[3.3.2.1. Kiểm tra dữ liệu 33](#_Toc217250421)

[3.3.2.2. Làm sạch dữ liệu 33](#_Toc217250422)

[3.3.3. Phân tích mô tả 38](#_Toc217250423)

[3.3.3.1. Phân tích đơn biến 38](#_Toc217250424)

[3.3.3.2. Phân tích đa biến 42](#_Toc217250425)

[3.3.3.3. Phân tích với Power BI 55](#_Toc217250426)

[3.3.4. Phân tích hồi quy logistic 59](#_Toc217250427)

[3.3.4.1. Hồi quy logistic 59](#_Toc217250428)

[3.3.4.2. Huấn luyện mô hình 62](#_Toc217250429)

[3.3.5. Kết luận chương 3 89](#_Toc217250430)

[CHƯƠNG 4: XÂY DỰNG SẢN PHẨM 91](#_Toc217250431)

[4.1. Công cụ và công nghệ sử dụng 91](#_Toc217250432)

[4.1.1. Giới thiệu về Flask 91](#_Toc217250433)

[4.1.2. Giới thiệu về Bootstrap 91](#_Toc217250434)

[4.1.3. Giới thiệu về thư viện Scikit-learn 92](#_Toc217250435)

[4.1.4. Giới thiệu về thư viện Joblib 92](#_Toc217250436)

[4.2. Chuẩn bị tài nguyên xây dựng chương trình 92](#_Toc217250437)

[4.3. Mô tả chương trình 93](#_Toc217250438)

[4.4. Demo sản phẩm 95](#_Toc217250439)

[KẾT LUẬN 100](#_Toc217250440)

[TÀI LIỆU THAM KHẢO 102](#_Toc217250441)

# DANH MỤC HÌNH ẢNH

[Hình 1.1: Quy trình phân tích dữ liệu 8](#_Toc217250444)

[Hình 2.1: Phương pháp phân tích mô tả 17](#_Toc217250445)

[Hình 3.1: Dữ liệu 20 dòng đầu của dataset 27](#_Toc217250446)

[Hình 3.2: Phần mềm Google Colab 30](#_Toc217250447)

[Hình 3.3: Phần mềm Pycharm 31](#_Toc217250448)

[Hình 3.4: Các thư viện cần thiết 32](#_Toc217250449)

[Hình 3.5: Tóm tắt bộ dataset 33](#_Toc217250450)

[Hình 3.6: Kiểm tra dữ liệu thiếu 34](#_Toc217250451)

[Hình 3.7: Bộ dữ liệu sau khi điền khuyết 35](#_Toc217250452)

[Hình 3.8: Đoạn mã kiểm tra dữ liệu trùng lặp 36](#_Toc217250453)

[Hình 3.9: Đếm các dòng trùng lặp 37](#_Toc217250454)

[Hình 3.10: Dữ liệu sau khi loại bỏ các dòng trùng lặp 38](#_Toc217250455)

[Hình 3.11: Biểu đồ Boxplot của nhóm hành vi sử dụng thiết bị 39](#_Toc217250456)

[Hình 3.12: Biểu đồ Boxplot của nhóm chỉ số sức 39](#_Toc217250457)

[Hình 3.13: Biểu đồ Histogram 41](#_Toc217250458)

[Hình 3.14: Biểu đồ heatmap 43](#_Toc217250459)

[Hình 3.15: Biểu đồ Boxplot của App\_work\_time và health\_status 45](#_Toc217250460)

[Hình 3.16: Biểu đồ Boxplot của App\_entertainment\_time và health\_status 46](#_Toc217250461)

[Hình 3.17: Biểu đồ Boxplot của Sleep\_duration và health\_status 48](#_Toc217250462)

[Hình 3.18: Biểu đồ đường của App\_Social\_Media\_Time với stress\_level 50](#_Toc217250463)

[Hình 3.19: Biểu đồ scatter của Stress\_Level vs Daily\_Screen\_Time 52](#_Toc217250464)

[Hình 3.20: Biểu đồ Scatter của Well\_Being\_Score vs Sleep\_Duration 54](#_Toc217250465)

[Hình 3.21: Nhập dữ liệu cần xử lý 55](#_Toc217250466)

[Hình 3.22: Cửa sổ dữ liệu chưa xử lý với Power Query 56](#_Toc217250467)

[Hình 3.23: Xóa, thay thế dữ liệu trống hoặc lỗi bằng giá trị trung bình 56](#_Toc217250468)

[Hình 3.24: Dữ liệu sau làm sạch bằng Power Query 57](#_Toc217250469)

[Hình 3.25: Kết quả thu được sau khi vẽ Power BI 59](#_Toc217250470)

[Hình 3.26: Mã nguồn đọc dữ liệu 62](#_Toc217250471)

[Hình 3.27: Mã nguồn đổi giờ sang phút 62](#_Toc217250472)

[Hình 3.28: Mã nguồn xóa các giá trị không hợp lệ 63](#_Toc217250473)

[Hình 3.29: Mã nguồn xử lý ngoại lai 64](#_Toc217250474)

[Hình 3.30: Biểu đồ boxplot sau khi xử lý ngoại lai 65](#_Toc217250475)

[Hình 3.31: Mã nguồn log-transform 66](#_Toc217250476)

[Hình 3.32: Mã nguồn tạo biến tỷ lệ 66](#_Toc217250477)

[Hình 3.33: Mã nguồn xử lý vô cực và giá trị thiếu 66](#_Toc217250478)

[Hình 3.34: Mã nguồn tạo các biến tương tác 67](#_Toc217250479)

[Hình 3.35: Mã nguồn tạo biến mục tiêu 67](#_Toc217250480)

[Hình 3.36: Mã nguồn chọn các đặc trưng 68](#_Toc217250481)

[Hình 3.37: Mã nguồn chuẩn hóa và chia tập dữ liệu 68](#_Toc217250482)

[Hình 3.38: Mã nguồn xây dựng mô hình 69](#_Toc217250483)

[Hình 3.39: Mã nguồn huấn luyện theo Hold-out 70](#_Toc217250484)

[Hình 3.40: Mã nguồn huấn luyện theo K-Fold 72](#_Toc217250485)

[Hình 3.41: Mã nguồn đánh giá mô hình (Hold-Out) 75](#_Toc217250486)

[Hình 3.42: Kết quả đánh giá trên tập kiểm tra (Hold-Out) 77](#_Toc217250487)

[Hình 3.43: Mã nguồn và kết quả đánh giá theo Stratified K-FOLD 79](#_Toc217250488)

[Hình 3.44: Mã nguồn xây dựng biến mục tiêu 81](#_Toc217250489)

[Hình 3.45: Mã nguồn feature Engineering 82](#_Toc217250490)

[Hình 3.46: Mã nguồn làm sạch dữ liệu 82](#_Toc217250491)

[Hình 3.47: Mã nguồn lựa chọn đặc trưng 83](#_Toc217250492)

[Hình 3.48: Mã nguồn chia tệp train/ test và chuẩn hóa 83](#_Toc217250493)

[Hình 3.49: Mã nguồn huấn luyện mô hình 84](#_Toc217250494)

[Hình 3.50: Mã nguồn đánh giá mô hình 84](#_Toc217250495)

[Hình 3.51: Kết quả đánh giá 85](#_Toc217250496)

[Hình 3.52: Ma trận nhầm lẫn 86](#_Toc217250497)

[Hình 3.53: ROC Curve 87](#_Toc217250498)

[Hình 3.54: Precision, Recall, F1-score theo từng lớp 88](#_Toc217250499)

[Hình 3.55: ROC Curve của Extreme Stress 89](#_Toc217250500)

[Hình 4.1: Sơ đồ usecase 93](#_Toc217250501)

[Hình 4.2: Giao diện người dùng 96](#_Toc217250502)

[Hình 4.3: Mã nguồn giao diện nhập dữ liệu 96](#_Toc217250503)

[Hình 4.4: Mã nguồn kiểm tra dữ liệu vào 97](#_Toc217250504)

[Hình 4.5: Kết quả dự đoán 97](#_Toc217250505)

[Hình 4.6: Mã nguồn dự đoán 97](#_Toc217250506)

[Hình 4.7: Biểu đồ dữ liệu đầu vào và dự đoán 99](#_Toc217250507)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1: Bảng mô tả usecase 93](#_Toc217249795)

# LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến Trường Công Nghệ Thông Tin và Truyền Thông - Đại học Công Nghiệp Hà Nội đã đưa môn học “Phân tích dữ liệu lớn” vào chương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – Thầy Nguyễn Mạnh Cường đã dạy, truyền đạt những kiến thức quý báu cho chúng em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học Phân tích dữ liệu lớn của thầy, chúng em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này.

Bộ môn “Phân tích dữ liệu lớn” là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài báo cáo khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong thầy xem xét và góp ý để bài báo cáo của chúng em được hoàn thiện hơn.

Lời cuối cùng em xin chúc thầy luôn dồi dào sức khỏe, luôn vui vẻ và thành công trong cuộc sống!

Chúng em xin chân thành cảm ơn!

***Sinh viên thực hiện***

Nguyễn Thị Ngọc Châm

Nguyễn Ngọc Huyền

Phạm Thị Phương

# LỜI NÓI ĐẦU

Trong thời đại công nghệ số phát triển mạnh mẽ, các thiết bị điện tử như điện thoại thông minh, máy tính bảng và máy tính cá nhân đã trở thành một phần không thể thiếu trong cuộc sống con người. Tuy nhiên, cùng với sự tiện ích mà chúng mang lại, việc sử dụng các thiết bị này quá mức đang làm dấy lên nhiều lo ngại về sức khỏe tinh thần và thể chất, đặc biệt là ở giới trẻ. Trong bối cảnh đó, việc phân tích và dự đoán ảnh hưởng của thời gian sử dụng màn hình đến sức khỏe tinh thần trở thành một hướng nghiên cứu có ý nghĩa thực tiễn sâu sắc.

Với sự phát triển vượt bậc của trí tuệ nhân tạo (AI), học máy (Machine Learning) và phân tích dữ liệu (Data Analysis), các mô hình thống kê tiên tiến như hồi quy logistic đang được ứng dụng rộng rãi trong việc phân tích và dự đoán mối quan hệ giữa các yếu tố tác động đến hành vi con người. Mô hình hồi quy logistic, với khả năng xử lý các biến phụ thuộc dạng nhị phân (chẳng hạn như “có nguy cơ ảnh hưởng sức khỏe tinh thần” hoặc “không ảnh hưởng”), được xem là công cụ hữu hiệu giúp khám phá các yếu tố then chốt và mức độ ảnh hưởng của việc sử dụng màn hình tới trạng thái tâm lý người dùng.

Đề tài “Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy logistic” nhằm khai thác sức mạnh của học máy và phân tích dữ liệu để xác định, dự đoán và cảnh báo sớm nguy cơ ảnh hưởng đến sức khỏe tinh thần do việc sử dụng thiết bị điện tử quá nhiều. Bằng cách kết hợp giữa phân tích thống kê, mô hình hóa dữ liệu và công nghệ AI, nghiên cứu hướng tới việc xây dựng một mô hình dự báo có độ chính xác cao và có thể ứng dụng thực tiễn trong lĩnh vực giáo dục, y tế và quản lý sức khỏe cộng đồng.

Nội dung của báo cáo sẽ bao gồm các chương như sau:

**Chương 1. Tổng quan về vấn đề nghiên cứu**: Trình bày bối cảnh, lý do lựa chọn đề tài, các khái niệm liên quan đến sức khỏe tinh thần và việc sử dụng thiết bị điện tử, đồng thời giới thiệu các phương pháp phân tích dữ liệu và mô hình hồi quy logistic.

**Chương 2. Phương pháp và công cụ nghiên cứu**: Trình bày cơ sở lý thuyết về hồi quy logistic, các kỹ thuật xử lý dữ liệu, lựa chọn biến, và mô tả chi tiết các công cụ lập trình, thư viện được sử dụng trong nghiên cứu.

**Chương 3. Thực nghiệm và kết quả**: Trình bày quá trình xây dựng, huấn luyện và đánh giá mô hình dự đoán; phân tích ý nghĩa thống kê của các biến đầu vào; và đánh giá độ chính xác, hiệu quả của mô hình qua các chỉ số đo lường.

**Chương 4. Ứng dụng và triển khai hệ thống**: Mô tả việc tích hợp mô hình dự báo vào một hệ thống phân tích hoặc giao diện ứng dụng, giúp người dùng có thể nhập dữ liệu và nhận được dự đoán, cảnh báo về nguy cơ ảnh hưởng sức khỏe tinh thần.

**Phần kết luận:** Tổng hợp lại kết quả đạt được, nêu bật đóng góp của đề tài, đồng thời đề xuất các hướng phát triển và mở rộng trong tương lai.

Thông qua việc thực hiện đề tài, người nghiên cứu không chỉ củng cố kiến thức về trí tuệ nhân tạo, học máy và thống kê, mà còn góp phần tạo ra một công cụ hỗ trợ hữu ích cho việc theo dõi, quản lý và nâng cao nhận thức cộng đồng về sức khỏe tinh thần trong thời đại số. Đề tài là minh chứng cho sự kết hợp giữa khoa học dữ liệu và ý nghĩa xã hội, mở ra những triển vọng mới trong việc ứng dụng công nghệ vào việc bảo vệ sức khỏe con người.

# TỔNG QUAN VỀ VẤN ĐỀ VẤN ĐỀ NGHIÊN CỨU

## Kiến thức tổng quan

### Trí tuệ nhân tạo ( Artificial Intelligence)

Vào năm 1943, Warren McCulloch và Walter Pitts bắt đầu thực hiện nghiên cứu ba cơ sở lý thuyết cơ bản: Triết học cơ bản và chức năng của các nơ ron thần kinh; phân tích các mệnh đề logic; lý thuyết dự đoán Turing. Các tác giả đã nghiên cứu đề xuất mô hình neural nhân tạo, mỗi neural đặc trưng bởi hai trạng thái “bật”, “tắt” và phát hiện mạng neural có khả năng học.

Trí tuệ nhân tạo (AI) được thiết lập bởi John McCarthy tại hội thảo đầu tiên về chủ đề này vào mùa hè năm 1956. Đồng thời, ông cũng đề xuất ngôn ngữ lập trình Lisp, một trong những ngôn ngữ lập trình hàm tiêu biểu, được sử dụng trong lĩnh vực AI. Sau đó, Alan Turing đưa ra “Turing test” như một phương pháp kiểm chứng hành vi thông minh.

Marvin Minsky và Seymour Papert đưa ra các chứng minh đầu tiên về giới

hạn của các mạng neural đơn giản. Ngôn ngữ lập trình logic Prolog ra đời và được phát triển bởi Alan Colmerauer. Ted Shortliffe xây dựng thành công một số hệ chuyên gia đầu tiên trợ giúp chẩn đoán y học, các hệ thống này sử dụng ngôn ngữ luật để biểu diễn tri thức và suy diễn.

Vào đầu năm 1980, những nghiên cứu thành công liên quan đến AI như các hệ chuyên gia (expert systems), một dạng của chương trình AI mô phỏng tri thức và các kỹ năng phân tích của một hoặc nhiều chuyên gia con người. AI được áp dụng trong logic, khai phá dữ liệu, chẩn đoán y học và nhiều lĩnh vực ứng dụng khác trong công nghiệp. Sự thành công dựa vào nhiều yếu tố: Tăng khả năng tính toán của máy tính, tập trung giải quyết các bài toán con cụ thể, xây dựng các mối quan hệ giữa AI và các lĩnh vực khác giải quyết các bài toán tương tự và một sự chuyển giao mới của các nhà nghiên cứu cho các phương pháp trong toán học vững chắc và chẩn đoán khoa học chính xác.

**Các lĩnh vực của trí tuệ nhân tạo:**

* Lập luận, suy diễn tự động: Khái niệm lập luận và suy diễn được sử dụng rất phổ biến trong lĩnh vực AI. Lập luận là suy diễn logic, dùng để chỉ một tiến trình rút ra kết luận (tri thức mới) từ những giả thiết đã cho (được biểu diễn dưới dạng cơ sở tri thức). Như vậy, để thực hiện lập luận người ta cần có các phương pháp lưu trữ cơ sở tri thức và các thủ tục lập luận trên cơ sở tri thức đó.
* Biểu diễn tri thức: Muốn máy tính có thể lưu trữ và xử lý tri thức thì cần có các phương pháp biểu diễn tri thức. Các phương pháp biểu diễn tri thức ở đây bao gồm các ngôn ngữ biểu diễn và các kỹ thuật xử lý tri thức. Một ngôn ngữ biểu diễn tri thức được đánh giá là “tốt” nếu nó có tính biểu đạt cao và tính hiệu quả của thuật toán lập luận trong ngôn ngữ đó. Tính biểu đạt của ngôn ngữ thể hiện khả năng biểu diễn một phạm vi rộng lớn các thông tin trong một miền ứng dụng. Tính hiệu quả của các thuật toán lập luận thể hiện chi phí về thời gian và không gian dành cho việc lập luận.
* Lập kế hoạch: Khả năng suy ra các mục đích cần đạt được đối với các nhiệm vụ đưa ra và xác định dãy các hành động cần thực hiện để đạt được mục đích đó.
* Xử lý ngôn ngữ tự nhiên: Là một nhánh của AI, tập trung vào các ứng dụng trên ngôn ngữ của con người. Các ứng dụng trong nhận dạng tiếng nói, nhận dạng chữ viết, dịch tự động, tìm kiếm thông tin...
* Hệ chuyên gia: Cung cấp các hệ thống có khả năng suy luận để đưa ra những kết luận. Các hệ chuyên gia có khả năng xử lý lượng thông tin lớn và cung cấp các kết luận dựa trên những thông tin đó. Có rất nhiều hệ chuyên gia nổi tiếng như các hệ chuyên gia y học MYCIN, đoán nhận cấu trúc phân tử từ công thức hóa học DENDRAL...

Một số ứng dụng của trí tuệ nhân tạo:

* Lĩnh vực giáo dục: Mô hình trường học thông minh hay học trực tuyến ngày một phát triển nhờ ứng dụng AI vào quá trình tương tác học tập giữa nhà trường và học sinh, sinh viên tạo ra một hệ sinh thái về giáo dục mà ở đó nhà trường và học sinh tương tác với nhau một cách thuậnlợi và nhanh chóng thông qua hệ sinh thái này.
* Lĩnh vực y tế: Chẩn đoán, điều trị và theo dõi bệnh giúp nâng cao chăm sóc sức khỏe con người, giảm chi phí chữa bệnh cho người dân, trong đó kể cả các dự án như điều trị ung thư bằng AI, robot chăm sóc y tế...
* Công nghiệp: Nhận dạng hình ảnh, giọng nói, robot thông minh, xe tự hành, hệ thống tương tác thực ảo... đã phát triển mạnh mẽ nhờ sự phát triển của mạng Neural học sâu, hệ thống xử lý phân tán song song (Parallel Distributed processing).

### Học máy (Machine Learning)

Học máy (Machine Learning) là một ngành khoa học nghiên cứu các thuật

toán cho phép máy tính có thể học được các khái niệm (concept). Có hai loại phương pháp học máy chính:

* Phương pháp quy nạp: máy học hoặc phân biệt các khái niệm dựa trên dữ liệu đã thu thập được trước đó. Phương pháp này cho phép tận dụngđược nguồn dữ liệu rất nhiều và sẵn có.
* Phương pháp suy diễn: máy học hoặc suy diễn dựa vào các luật. Phương pháp này cho phép tận dụng được các kiến thức chuyên ngành để hỗ trợ máy tính.

Machine Learning được chia thành 3 nhánh chính: supervised learning (học có giám sát), unsupervised learning (học không có giám sát), và reinforcement learning (học tăng cường).

* Học có giám sát được dùng trong trường hợp một thuộc tính (nhãn) có sẵn cho một tập dữ liệu nhất định (tập huấn luyện), nhưng thiếu và cần được dự đoán cho các trường hợp khác.
* Học không có giám sát thì ngược lại, nó được sử dụng trong trường hợp khám phá các mối quan hệ tiềm ẩn trong một tập dữ liệu không được gán nhãn (các mục không được chỉ định trước).
* Học tăng cường thì nằm giữa 2 loại trên, có một số hình thức phản hồi sẵn cho mỗi bước tiên đoán hoặc hành động, nhưng không có nhãn chính xác hoặc thông báo lỗi.

Trong lĩnh vực xử lý ảnh, bài toán phân loại mẫu và nhận dạng đối tượng

được hiểu là việc phân loại, xác định các mẫu thuộc lớp nào đó một cách chính

xác dựa vào việc đo lường về đối tượng dựa trên các đặc trưng của đối tượng. Nói một cách khác, nhận dạng đối tượng nhằm mục đích phân loại các mẫu dựa trên kiến thức có trước về đối tượng hoặc dựa vào thông tin thống kê, đo lường được trích rút ra từ các mẫu trong tập dữ liệu huấn luyện. Ngày nay, nhận dạng được ứng dụng rộng rãi trong nhiều hệ thống như nhận dạng người trong các hệ thống tương tác người máy; nhận dạng khuôn mặt và định danh cá thể, nhận dạng phương tiện giao thông trong các hệ thống giao thông thông minh, hỗ trợ lái xe; nhận dạng hành động trong các hệ thống giám sát thông minh; nhận dạng chữ viết... Lĩnh vực nhận dạng đối tượng liên quan đến các phương pháp, kỹ thuật xử lý ảnh và học máy. Để huấn luyện mô hình và nhận dạng đối tượng, các mẫu dữ liệu thường được số hóa và biểu diễn thành các vector đặc trưng trong một không gian đa chiều tương ứng, được gọi là không gian đặc trưng. Nhờ vào các vector đặc trưng này để có thể phân tích, đánh giá nhằm xây dựng được mô hình đối tượng phục vụ phân loại, nhận dạng chúng. Trong nhận dạng, các phương pháp học máy có giám sát và không giám sát đều được nghiên cứu và ứng dụng. Các kỹ thuật học máy có giám sát thường được sử dụng như cây quyết định, mạng neural, SVM, Boosting, rừng ngẫu nhiên (random forest). Các kỹ thuật học máy không giám sát như gom cụm (Clustering Algorithms), phân tích thành phần chính (Principal Component Analysis - PCA), phân tích thành phần độc lập (Independent Component Analysis).

### Phân tích dữ liệu là gì?

Phân tích dữ liệu là quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, đưa ra kết luận và hỗ trợ việc ra quyết định.

Quy trình phân tích dữ liệu:



Hình 1.1: Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu thường bao gồm các bước chính:

* Xác định mục tiêu và thu thập dữ liệu:
* *Xác định mục tiêu*: là những kết quả cụ thể mà ta muốn đạt được thông qua việc xử lý và phân tích dữ liệu. Mục tiêu này xác định hướng đi và phạm vi của quá trình phân tích, giúp ta tập trung vào việc thu thập thông tin quan trọng và thực hiện các phân để đáp ứng các yêu cầu hoặc nhu cầu cụ thể.
* *Thu thập dữ liệu*: là thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web, thiết bị cảm biến, và nhiều nguồn khác. Dữ liệu có thể là số liệu, văn bản, hình ảnh, hoặc âm thanh.
* Tiền xử lý dữ liệu: Dữ liệu thường không hoàn hảo và có thể chứa nhiễu, dữ liệu bị thiếu, hoặc không chính xác. Tiền xử lý dữ liệu bao gồm việc tóm lược dữ liệu, làm sạch dữ liệu, tích hợp dữ liệu, chuyển đổi dữ liệu, rút gọn dữ liệu và rời rạc hóa dữ liệu để chuẩn bị cho bước phân tích.
* Phân tích dữ liệu: Bước quan trọng này dựa vào kiến thức và kỹ thuật phân tích để tìm ra mối liên hệ và thông tin ẩn sau dữ liệu. Phân tích dữ liệu có thể sử dụng các phương pháp phân tích mô tả, phân tích hồi quy, phân tích sự khác biệt, thống kê, machine learning, data mining, và nhiều kỹ thuật khác.
* Kết luận và dự đoán: Dựa trên phân tích và thông tin từ dữ liệu, chúng ta có thể rút ra kết luận, hiểu rõ hơn về tình hình, và thậm chí đưa ra dự đoán cho tương lai.

## Tổng quan về bài toán phân tích dự đoán

Phân tích dự đoán (Predictive Analysis) là một lĩnh vực quan trọng trong khoa học dữ liệu và phân tích dữ liệu, tập trung vào việc sử dụng dữ liệu lịch sử kết hợp với các mô hình thống kê, thuật toán học máy và các kỹ thuật khai phá dữ liệu nhằm dự báo những kết quả có thể xảy ra trong tương lai. Thông qua việc phát hiện các mẫu, xu hướng và mối quan hệ ẩn trong dữ liệu, phân tích dự đoán hỗ trợ các tổ chức đưa ra những quyết định chính xác và chủ động hơn.

Ngày nay, phân tích dự đoán được ứng dụng rộng rãi trong nhiều lĩnh vực như tài chính, marketing, y tế, sản xuất hay thương mại điện tử. Với lượng dữ liệu ngày càng lớn và đa dạng, các tổ chức có thể tận dụng phân tích dự đoán để tối ưu hóa chiến lược, giảm thiểu rủi ro và nâng cao hiệu quả hoạt động. Một số vai trò quan trọng của phân tích dự đoán có thể kể đến:

* Dự báo xu hướng và kết quả tương lai:

Bằng cách xử lý dữ liệu lịch sử và dữ liệu thời gian thực, các thuật toán tiên tiến như hồi quy, cây quyết định, mô hình học sâu hay mô hình chuỗi thời gian có thể đưa ra dự đoán về giá trị, hành vi hoặc sự kiện trong tương lai. Điều này giúp các tổ chức chuẩn bị kế hoạch phù hợp, tận dụng cơ hội và hạn chế các rủi ro tiềm ẩn.

* Hỗ trợ ra quyết định chiến lược:

Phân tích dự đoán cung cấp những thông tin định lượng có giá trị, giúp nhà quản lý đưa ra quyết định dựa trên dữ liệu thay vì cảm tính. Việc mô phỏng các tình huống và dự đoán kết quả của từng lựa chọn giúp tổ chức xây dựng chiến lược kinh doanh đúng đắn, cải thiện hiệu suất và nâng cao khả năng cạnh tranh.

* Hiểu rõ đặc điểm và phân phối dữ liệu:

Việc phân tích phân phối dữ liệu thông qua biểu đồ tần suất hoặc biểu đồ mật độ kernel giúp hiểu sâu hơn về bản chất dữ liệu đầu vào. Đây là bước quan trọng trong tiền xử lý, ảnh hưởng trực tiếp đến chất lượng mô hình dự đoán và độ chính xác của kết quả.

Phân tích dự đoán đóng vai trò then chốt trong quá trình chuyển đổi số và ứng dụng dữ liệu vào hoạt động thực tiễn. Nhờ khả năng dự đoán tương lai và khám phá hiểu biết sâu sắc từ dữ liệu, các tổ chức có thể nâng cao hiệu quả, tối ưu hóa chiến lược, đem lại trải nghiệm cá nhân hóa cho khách hàng và thúc đẩy tăng trưởng bền vững. Trong bối cảnh cạnh tranh ngày càng mạnh mẽ, phân tích dự đoán không chỉ là một công cụ hỗ trợ ra quyết định mà còn là lợi thế chiến lược giúp tổ chức thành công trong dài hạn.

## Bài toán dự đoán ảnh hưởng của thời gian sử dụng màn hình

### Giới thiệu chung

Trong bối cảnh công nghệ phát triển mạnh mẽ, thời gian sử dụng màn hình của các thiết bị điện tử như điện thoại, máy tính bảng, máy tính xách tay và tivi ngày càng tăng. Việc sử dụng màn hình quá mức có thể ảnh hưởng tiêu cực đến sức khỏe thể chất và tinh thần, đặc biệt là giấc ngủ, thị lực, mức độ căng thẳng và khả năng tập trung.

Bài toán dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử là bài toán phân tích dữ liệu nhằm xác định mối quan hệ giữa thời gian tiếp xúc với màn hình và các chỉ số sức khỏe. Mục tiêu là xây dựng mô hình dự đoán ảnh hưởng dựa trên dữ liệu thực tế như số giờ sử dụng màn hình, độ tuổi, thói quen sinh hoạt, mức độ vận động,… Đây là bài toán quan trọng vì giúp đề xuất giải pháp điều chỉnh hành vi, bảo vệ sức khỏe và nâng cao chất lượng sống.

### Mục đích, ý nghĩa của đề tài

1. Mục đích

Mục tiêu chính của đề tài là xây dựng mô hình dự đoán mối quan hệ giữa thời gian sử dụng thiết bị điện tử và các chỉ số sức khỏe tinh thần như:

* Mức độ căng thẳng (Stress\_Level)
* Điểm hạnh phúc, sức khỏe tinh thần (Well\_Being\_Score)
* Chỉ số cảm xúc (Mood\_Rating)
* Thời lượng ngủ (Sleep\_Duration)

Thông qua việc phân tích dữ liệu lịch sử, mô hình sẽ xác định:

* Yếu tố nào trong hành vi sử dụng thiết bị ảnh hưởng mạnh nhất (Daily\_Screen\_Time, App\_Social\_Media\_Time, Phone\_Unlocks…)
* Mức độ ảnh hưởng (tăng, giảm, mạnh, nhẹ)
* Quan hệ tuyến tính giữa biến giải thích và biến mục tiêu

Từ đó, mô hình giúp dự đoán tình trạng sức khỏe tinh thần của người dùng dựa trên thói quen sử dụng thiết bị hằng ngày.

1. Ý nghĩa khoa học

Đề tài mang ý nghĩa khoa học quan trọng khi vận dụng các kiến thức về thống kê, học máy và phân tích dữ liệu để đánh giá tác động của công nghệ lên sức khỏe con người. Việc sử dụng mô hình dự đoán giúp:

* Xác định mối quan hệ định lượng giữa việc sử dụng màn hình và sức khỏe tinh thần
* Kiểm chứng các giả thuyết khoa học về tác hại của sử dụng thiết bị quá mức
* Ứng dụng quy trình khoa học dữ liệu gồm: thu thập – làm sạch – phân tích – mô hình hóa – đánh giá

Dataset sử dụng trong đề tài chứa nhiều đặc trưng quan trọng như thời gian sử dụng mạng xã hội, thời gian giải trí, công việc và số lần mở khóa điện thoại, giúp mô hình có khả năng phân tích sâu và đa chiều.

1. Ý nghĩa thực tiễn

Trong thực tế, ảnh hưởng của việc sử dụng thiết bị điện tử lên sức khỏe tinh thần thường được đánh giá chung chung và thiếu số liệu cụ thể. Đề tài mang lại giá trị thực tiễn rõ rệt:

* Giúp người dùng cá nhân nhận biết thói quen sử dụng của bản thân có ảnh hưởng như thế nào đến giấc ngủ và tinh thần.
* Phụ huynh và giáo viên có thể theo dõi, cảnh báo cho trẻ nhỏ hoặc học sinh về nguy cơ sử dụng thiết bị quá mức.
* Các nhà phát triển ứng dụng có thể xây dựng hệ thống cảnh báo hoặc đề xuất thời gian nghỉ ngơi hợp lý.
* Tổ chức y tế và tâm lý học có thể tham khảo mô hình để hỗ trợ chẩn đoán hoặc tư vấn.

Trong tương lai, mô hình dự đoán này còn có thể tích hợp vào các ứng dụng theo dõi thiết bị (Digital Well-Being) nhằm phân tích thời gian sử dụng và tự động đưa ra khuyến nghị cải thiện.

### Đầu vào và đầu ra của bài toán

1. Đầu vào

Đầu vào của bài toán là dữ liệu về hành vi sử dụng thiết bị điện tử và các chỉ số sức khỏe tinh thần, bao gồm:

* Daily\_Screen\_Time: tổng thời gian sử dụng trong ngày
* App\_Social\_Media\_Time: thời gian dùng mạng xã hội
* App\_Work\_Time: thời gian dùng để làm việc
* App\_Entertainment\_Time: thời gian dùng để giải trí
* Phone\_Unlocks: số lần mở khóa điện thoại
* Sleep\_Duration: số giờ ngủ
* Well\_Being\_Score, Stress\_Level, Mood\_Rating: các chỉ số sức khỏe tinh thần

Dataset có thể chứa giá trị ngoại lai, chênh lệch giữa các ngày hoặc các hành vi đặc thù của từng người tham gia, đòi hỏi phải tiền xử lý dữ liệu trước khi đưa vào mô hình.

1. Đầu ra

Đầu ra của bài toán là:

* Giá trị dự đoán của một chỉ số sức khỏe tinh thần Stress\_Level
* Các hệ số ảnh hưởng (trong mô hình hồi quy) giúp chỉ ra yếu tố nào tác động mạnh nhất.
* Các chỉ số đánh giá mô hình như Precision, Recall, F1…

### Thuận lợi và khó khăn khi giải quyết bài toán

1. Thuận lợi

* Dữ liệu rõ ràng và có cấu trúc, bao gồm đầy đủ các biến cần thiết cho bài toán dự đoán.
* Nhiều thư viện hỗ trợ như pandas, numpy, scikit-learn, matplotlib giúp xử lý dễ dàng.
* Mối quan hệ giữa thời gian sử dụng thiết bị và sức khỏe tinh thần đã được nghiên cứu nhiều, dễ xây dựng giả thuyết.
* Mô hình dự đoán dễ triển khai, đặc biệt với hồi quy tuyến tính do dữ liệu mang nhiều yếu tố số hóa.

1. Khó khăn

* Dữ liệu có thể chứa nhiễu, đặc biệt do hành vi người dùng không ổn định theo ngày.
* Một số biến có thể bị đa cộng tuyến (ví dụ: Daily\_Screen\_Time liên quan chặt với Social\_Media\_Time hoặc Entertainment\_Time).
* Khó đảm bảo yếu tố nhân quả, chỉ xác định được mối quan hệ chứ không chứng minh được nguyên nhân.
* Cần tiền xử lý dữ liệu kỹ để tránh mô hình bị sai lệch.

### Ứng dụng của việc giải quyết bài toán

* Xây dựng hệ thống cảnh báo sức khỏe tinh thần dựa trên thói quen sử dụng thiết bị.
* Ứng dụng trong giáo dục, giúp phụ huynh kiểm soát thời gian sử dụng thiết bị của trẻ.
* Hỗ trợ bác sĩ tâm lý theo dõi bệnh nhân và đưa ra khuyến nghị phù hợp.
* Cải thiện trải nghiệm người dùng trong các ứng dụng Digital Well-Being.
* Phát triển các mô hình AI cá nhân hóa, đưa ra đề xuất giảm stress dựa trên hành vi thiết bị thực tế.

# CÁC PHƯƠNG PHÁP VÀ KĨ THUẬT

## Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một công đoạn cực kỳ quan trọng đối với việc lưu trữ cũng như khai thác dữ liệu bởi vì dữ liệu trong thực tế thường không hoàn chỉnh, nhiễu và đôi khi không tương thích. Tiền xử lý dữ liệu bao gồm các công đoạn chính sau:

* Làm sạch dữ liệu: Quá trình làm sạch dữ liệu sẽ cố gắng thêm vào những giá trị bị thiếu, làm trơn dữ liệu bằng cách khử nhiễu, xác định các mẫu dữ liệu ngoại lai (outlier) và hiệu chỉnh, loại bỏ những thành phần không tương thích.
* Tích hợp dữ liệu: Quá trình tích hợp dữ liệu kết hợp dữ liệu từ nhiều nguồn khác nhau, tổng hợp thành một nguồn dữ liệu kết dính chặt chẽ. Dữ liệu quản lý, phân tích quan hệ kiểm định không tương thích là các kỹ thuật thường dùng để làm trơn (smooth) dữ liệu tích hợp.
* Chuyển đổi dữ liệu: Quá trình chuyển đổi dữ liệu sẽ biến đổi dữ liệu về những dạng chuẩn thuận lợi cho việc khai phá tri thức. Lấy ví dụ, là các giá trị thuộc tính có thể sẽ được chuẩn hoá sao cho nằm trong khoảng [0,1].
* Rút gọn dữ liệu: Quá trình rút gọn dữ liệu có thể được tiến hành thông qua nhiều kỹ thuật: lựa chọn tập thuộc tính, giảm chiều dữ liệu, giảm kích thước dữ liệu nhằm biểu diễn dữ liệu dưới dạng rút gọn, tối thiểu hoá sự mất mát thông tin khi rút gọn.

Tiền xử lý dữ liệu là một bước rất quan trọng trong việc giải quyết bất kỳ vấn đề nào trong quá trình nghiên cứu nhằm loại bỏ những “lỗi” ở bộ số liệu trước khi đưa vào phân tích. Hầu hết các bộ dữ liệu được sử dụng trong các vấn đề liên quan đến phân tích dữ liệu cần được xử lý, làm sạch và biến đổi trước khi sử dụng một kỹ thuật phân tích trên những bộ dữ liệu này. Các kỹ thuật tiền xử lý dữ liệu phổ biến hiện nay bao gồm: xử lý dữ liệu bị khuyết (missing data), mã hóa các biến nhóm (encoding categorical variables), chuẩn hóa dữ liệu (standardizing data), co giãn dữ liệu (scaling data), ... Những kỹ thuật này tương đối dễ hiểu nhưng sẽ có nhiều vấn đề phát sinh khi chúng ta áp dụng vào các dữ liệu thực tế. Trong phần này, chúng ta sẽ cùng nhau tìm hiểu về các kỹ thuật tiền xử lý dữ liệu và cách áp dụng chúng trong các bài toán thực tế.

## Các phương pháp phân tích

### Phương pháp phân tích mô tả

Phân tích mô tả là quá trình tóm tắt, mô tả và trình bày các đặc điểm cơ bản của dữ liệu bằng các thước đo thống kê và biểu đồ nhằm giúp dữ liệu trở nên dễ hiểu và dễ diễn giải. Thông thường, người ta hay sử dụng các kỹ thuật trong thống kê, và do vậy chúng thường được hiểu là thống kê mô tả.

Mục tiêu của phân tích mô tả:

* Chuyển đổi dữ liệu để dễ hiểu, dễ diễn giải
* Sắp đặt, sắp xếp lại và thao tác trên đó để tạo ra thông tin mô tả

Ta có sơ đồ sau:

A diagram of a flowchart

AI-generated content may be incorrect.

Hình 2.1: Phương pháp phân tích mô tả

Tùy theo dữ liệu, yêu cầu mà lựa chọn công việc

* Phân tích trên 1 biến (1 thuộc tính): giúp xác định các giá trị ngoại lai hoặc các giá trị bất thường để có thể kiểm tra lỗi nhập dữ liệu, xác định “hình d̀ạng” của phân bố dữ liệu.
* Dữ liệu số: Vẽ biểu đồ Histogram, Box & Whisker, tính các đại lượng thống kê: mean, stdev, (nếu phân bố này là đối xứng), median, quartile (nếu phân bố không đối xứng). Ngoài ra có thể tính min, max cho dữ liệu.
* Dữ liệu phi số (biến thể loại): tạo bảng tần suất (frequency table), vẽ biểu đồ thanh, biểu đồ hình tròn hoặc biểu đồ doughnut.
* Phân tích trên nhiều biến: biến có thể là định lượng hoặc thể loại nên tìm ra thông tin sơ bộ về mối liên hệ giữa các biến
* Dữ liệu số: tạo biểu đồ scatter
* Dữ liệu phi số: tạo các bảng dữ liệu thống kê
* Dữ liệu hỗn hợp: tạo bảng thống kê tổng hợp, vẽ đồ thị box-and-whisker của dữ liệu định lượng trên từng thể loại.

Các công việc chính

* Tính các đại lượng thống kê: mean, median, mode, min, max, variance, stdev, quartiles (Q1, Q2, Q3).
* Tạo bảng: bảng tần suất, bảng thống kê.

#### Phương pháp phân tích trên từng biến

Khi thực hiện phân tích trên một biến (hoặc một thuộc tính), mục tiêu chính là hiểu rõ các đặc điểm cơ bản của biến đó. Điều này thường bao gồm xác định và xử lý các giá trị ngoại lai hoặc bất thường (Outliers). Đây là các giá trị dữ liệu mà rất khác biệt so với phần lớn các giá trị khác trong tập dữ liệu. Các giá trị ngoại lai có thể xuất hiện do lỗi nhập liệu, lỗi đo lường, hoặc đơn giản là do các sự kiện hiếm gặp.

Việc xác định các giá trị ngoại lai có vai trò quan trọng và là mắt xích liên kết giữa phân tích mô tả và phân tích hồi quy, bởi vì ta có thể tiến hành làm sạch những giá trị này tại công đoạn tiền xử lý dữ liệu của phân tích hồi quy. Cụ thể với từng loại dữ liệu khác nhau, ta sẽ phân tích như sau:

* Dữ liệu số:
* Biểu đồ Histogram: Biểu đồ hiển thị tần suất xuất hiện của các khoảng giá trị dữ liệu.
* Các đại lượng thống kê: Bao gồm mean (trung bình), stdev (độ lệch chuẩn), median (trung vị), quartile (phân vị) ... Các giá trị này giúp mô tả trung bình, phương sai và phân phối của dữ liệu.
* Biểu đồ Box & Whisker (Boxplot): Biểu đồ hiển thị tổng quan giá trị đó bao gồm các giá trị đại lượng thống kê đã tính được.
* Dữ liệu phi số:
* Bảng tần suất (Frequency table): Biểu đồ liệt kê các giá trị khác nhau của biến và số lần xuất hiện của mỗi giá trị.
* Biểu đồ cột (Bar chart): Biểu đồ thể hiện tần suất của từng giá trị dữ liệu dưới dạng các cột đứng.
* Biểu đồ hình tròn hoặc donut (Pie chart, Donut chart): Biểu đồ thể hiện phần trăm tần suất của từng giá trị trong tổng số.

#### Phương pháp phân trên nhiều biến

Phân tích trên nhiều biến hướng tới việc hiểu mối quan hệ và tương tác giữa các biến trong tập dữ liệu. Điều này có thể giúp bạn phát hiện ra các mẫu, xu hướng hoặc tương quan có thể tồn tại giữa chúng.

Các mối liên hệ giữa các biến (Interrelationships) có thể là nhiều dạng khác nhau: Mối tương quan tuyến tính, tương quan không tuyến tính, tương quan ngược, … Với mỗi mối liên hệ, ta có thể phân tích và tìm ra được cách các biến tương tác và ảnh hưởng lẫn nhau.

Việc phân tích trên nhiều biến cũng có mối liên hệ mật thiết đến phân tích hồi quy khi giúp ta xác định được các giá trị ngoại lai của dữ liệu. Do là phân tích nhiều biến, vậy nên sẽ có 3 kiểu dữ liệu phân tích khác nhau: Số, phi số và hỗn hợp (cả số và phi số):

* Dữ liệu số:
* Scatter Plot (Biểu đồ Scatter): Biểu đồ thể hiện mối quan hệ giữa hai biến số. Mỗi điểm trên biểu đồ thể hiện một cặp giá trị của hai biến trên trục ngang và dọc. Biểu đồ này dùng để tìm kiếm sự tương quan giữa 2 biến số như tương quan tuyến tính hoặc không tuyến tính.
* Bảng dữ liệu thống kê (Statistical Summary Table): Tạo bảng để liệt kê các đại lượng thống kê (mean, median, stdev, …) giữa các biến số của dữ liệu.
* Dữ liệu phi số:
  + Bảng dữ liệu thống kê (Statistical Summary Table): Cũng là bảng dữ liệu thống kê nhưng với giá trị phi số, đó sẽ chỉ có giá trị tần suất xuất hiện (mode) của dữ liệu.
* Dữ liệu hỗn hợp:
  + Bảng thống kê tổng hợp: Đây là sự kết hợp giữa bảng dữ liệu thống kê của dữ liệu số và phi số. Sự kết hợp tổng quan này sẽ cho ta bao quát được phân bổ của dữ liệu.
  + Biểu đồ Box & Whisker (Boxplot): Được sử dụng để so sánh phân phối của một dữ liệu số với tần suất của một dữ liệu phi số. Biểu đồ này sẽ cho ta mối quan hệ mật thiết về sự ảnh hưởng của các giá trị phi số lên giá trị số được phân tích.

### Phương pháp phân tích hồi quy

Phân tích hồi quy là một tập hợp các phương pháp thống kê được sử dụng để ước tính các mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Nó có thể được sử dụng để đánh giá sức mạnh của mối quan hệ giữa các biến và để mô hình hóa mối quan hệ trong tương lai giữa chúng.

Phân tích hồi quy là một cách phân loại toán học để xác định biến nào trong số những biến đó thực sự có tác động. Nó trả lời các câu hỏi: Yếu tố nào quan trọng nhất? Cái nào có thể bỏ qua? Các yếu tố đó tương tác với nhau như thế nào? Và quan trọng nhất, chúng ta chắc chắn như thế nào về tất cả những yếu tố này?

Trong phân tích hồi quy, ta cần xác định một biến phụ thuộc - yếu tố chính mà ta đang cố gắng hiểu hoặc dự đoán. Phân tích hồi quy bao gồm một số biến thể, chẳng hạn như tuyến tính, nhiều tuyến tính và phi tuyến tính. Trong đó mô hình phổ biến là tuyến tính và nhiều tuyến tính. Đối với phân tích hồi quy phi tuyến, chúng thường được sử dụng cho các tập dữ liệu phức tạp hơn trong đó các biến phụ thuộc và độc lập thể hiện mối quan hệ phi tuyến.

Để phân tích hồi quy có rất nhiều phương pháp để phân tích. Dưới đây sẽ là một số phương pháp quan trọng dùng để phân tích hồi quy:

* Hồi quy tuyến tính (Linear Regression): Hồi quy tuyến tính dự đoán giá trị mục tiêu dựa trên biến độc lập bằng cách tìm đường thẳng "tốt nhất" vượt qua dữ liệu. Phương pháp này đơn giản và phù hợp với dữ liệu có mối quan hệ tuyến tính. Tuy nhiên, nó có thể không xử lý được dữ liệu phi tuyến và ảnh hưởng bởi nhiễu dữ liệu.
* Hồi quy Ridge (Ridge Regression): Hồi quy Ridge là phiên bản cải tiến của hồi quy tuyến tính bằng cách thêm hệ số điều chuẩn 12 vào hàm mất mát. Điều này giúp kiểm soát độ phức tạp của mô hình và tránh tình trạng quá khớp (overfitting). Tuy ưu điểm là giảm overfitting và xử lý đa cộng tuyến, nhưng cần lựa chọn tham số điều chuẩn chính xác.
* Hồi quy Lasso (Lasso Regression): Hồi quy Lasso cũng cải tiến từ hồi quy tuyến tính, nhưng thay vì 12, nó sử dụng hệ số điều chuẩn 11 để thúc đẩy một số hệ số về 0. Điều này dẫn đến lựa chọn biến tự động và giảm biến quan trọng. Lasso giải quyết vấn đề "chọn biến" nhưng cần phải có tham số điều chuẩn chính xác.
* Hồi quy Logistic (Logistic Regression): Hồi quy Logistic được sử dụng chủ yếu cho các bài toán phân loại, trong đó biến mục tiêu là rời rạc (thường là nhị phân). Thay vì dự đoán giá trị liên tục như hồi quy tuyến tính, hồi quy Logistic ước lượng xác suất xảy ra của một lớp bằng cách sử dụng hàm sigmoid. Phương pháp này đơn giản, dễ diễn giải và hiệu quả với dữ liệu có mối quan hệ tuyến tính giữa các biến độc lập và log-odds của biến mục tiêu, tuy nhiên hiệu suất có thể giảm khi dữ liệu có mối quan hệ phi tuyến phức tạp

## Các công cụ

### Ngôn ngữ Python

Python là một ngôn ngữ lập trình căn bản mạnh mẽ nhưng dễ dàng sử dụng. Đây là một ngôn ngữ có thể xử lý các cấu trúc dữ liệu cấp cao một cách hiệu quả và cách tiếp cận tuy đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp tinh tế và cách trình bày linh động, kết hợp với nhau cùng với tính chất thông dịch của nó, tất cả các điều đó khiến cho Python trở thành một ngôn ngữ lý tưởng trong rất nhiều lĩnh vực khác nhau, trong đó phải kể đến việc thực hiện phân tích các dữ liệu lớn. Với bộ thư viện cực kỳ lớn, Python có hầu hết các công cụ cho việc phân tích dữ liệu: pandas (cho phép thực hiện các thao tác đơn giản trong việc thu thập hay chỉnh sửa dữ liệu), scikit-learn (giúp chúng ta thực hiện hay xây dựng các mô hình liên quan trong việc phân tích dữ liệu), matplotlib (giúp chúng ta trực quan hóa bộ dữ liệu), ... Có thể kết luận rằng Python là một ngôn ngữ lập trình mạnh mẽ và linh hoạt, được sử dụng rộng rãi trong bài toán phân tích dữ liệu. Python có nhiều ưu điểm như dễ học, dễ sử dụng, mở rộng và linh hoạt ngoài ra còn có sẵn nhiều tài liệu hỗ trợ.

### Weka

Weka là một phần mềm mã nguồn mở được phát triển bởi Đại học Waikato, New Zealand, chuyên dùng cho phân tích dữ liệu và học máy. Weka cung cấp một bộ công cụ mạnh mẽ cho việc khám phá, phân tích và mô hình hóa dữ liệu thông qua các thuật toán học máy truyền thống. Phần mềm này đặc biệt phổ biến trong môi trường học thuật và nghiên cứu nhờ vào giao diện đồ họa dễ sử dụng, không yêu cầu người dùng phải lập trình để thực hiện các phân tích phức tạp. Weka hỗ trợ nhiều thuật toán học máy, bao gồm phân loại, hồi quy, phân cụm, và các thuật toán nâng cao như cây quyết định, k-nearest neighbors (KNN), Naive Bayes, SVM, và nhiều thuật toán khác. Các thuật toán này được tích hợp sẵn trong phần mềm, giúp người dùng dễ dàng áp dụng mà không cần phải cài đặt thêm thư viện hay mã nguồn.

Một điểm mạnh của Weka là giao diện đồ họa người dùng (GUI), cho phép người dùng thao tác trực quan để tải dữ liệu, chọn thuật toán, huấn luyện mô hình và đánh giá kết quả mà không cần viết mã. Điều này làm cho Weka trở thành công cụ lý tưởng cho người mới bắt đầu hoặc những người không có kinh nghiệm lập trình nhưng vẫn muốn tìm hiểu và áp dụng các kỹ thuật học máy. Weka hỗ trợ nhiều định dạng dữ liệu phổ biến như CSV, ARFF (định dạng đặc biệt của Weka), và các định dạng khác. Nó cũng cho phép người dùng thực hiện tiền xử lý dữ liệu như xử lý giá trị thiếu, chuẩn hóa, chuyển đổi dữ liệu và các kỹ thuật tiền xử lý khác.

Tuy nhiên, Weka có một số hạn chế như không hỗ trợ tốt khi làm việc với dữ liệu lớn, và khả năng mở rộng hạn chế so với các ngôn ngữ lập trình như Python. Dù vậy, Weka vẫn là một công cụ học máy tuyệt vời cho những người mới bắt đầu, các nghiên cứu học thuật và các bài toán học máy cơ bản đến trung cấp.

### Excel

Excel là một công cụ xử lý dữ liệu phổ biến, mạnh mẽ nhưng rất dễ dàng sử dụng đối với hầu hết người dùng. Đây là phần mềm hỗ trợ thao tác với các bảng dữ liệu có cấu trúc một cách hiệu quả, đồng thời cung cấp nhiều tính năng trực quan nhưng không kém phần mạnh mẽ trong việc quản lý và phân tích thông tin.

Giao diện thân thiện cùng cách trình bày linh hoạt, kết hợp với khả năng xử lý dữ liệu theo hướng thủ công lẫn tự động bằng công thức, hàm và các công cụ nâng cao, tất cả những điều đó khiến cho Excel trở thành một công cụ lý tưởng trong rất nhiều lĩnh vực khác nhau, đặc biệt là trong phân tích dữ liệu.

Với bộ hàm, tiện ích và công cụ cực kỳ phong phú, Excel có hầu hết các tính năng phục vụ cho việc phân tích dữ liệu:

* Hàm và công thức (giúp thao tác, tính toán, kiểm tra và xử lý dữ liệu một cách nhanh chóng)
* PivotTable và PivotChart (hỗ trợ tổng hợp dữ liệu và trực quan hóa một cách linh hoạt)
* Power Query (cho phép thu thập, làm sạch và biến đổi dữ liệu từ nhiều nguồn)
* Power Pivot (giúp xây dựng mô hình dữ liệu phức tạp và phân tích dữ liệu lớn hiệu quả hơn)

Có thể kết luận rằng Excel là một công cụ mạnh mẽ, linh hoạt và được sử dụng rộng rãi trong các bài toán phân tích dữ liệu. Excel sở hữu nhiều ưu điểm như dễ dùng, dễ tiếp cận, hỗ trợ mở rộng qua Add-ins và macro, đồng thời có sẵn nhiều tài liệu hướng dẫn ở mọi cấp độ

## Kết luận chương 2

Như vậy, chương 2 đã trình bày toàn diện các cơ sở lý thuyết và kỹ thuật phục vụ nghiên cứu, tập trung vào phương pháp hồi quy tuyến tính, quy trình tiền xử lý dữ liệu, và các công cụ phân tích như Python, pandas, và scikit-learn. Lý thuyết về hồi quy tuyến tính được làm rõ, nhấn mạnh khả năng mô hình hóa mối quan hệ tuyến tính giữa các đặc trưng kim cương và giá trị. Các bước tiền xử lý dữ liệu được mô tả chi tiết, đảm bảo cơ sở cho việc xây dựng mô hình chất lượng. Nội dung chương cung cấp nền tảng lý thuyết vững chắc, định hướng rõ ràng cho các bước thực nghiệm và phân tích trong chương tiếp theo.

# THỰC NGHIỆM

## Dữ liệu thực nghiệm

Để xây dựng mô hình phân tích ảnh hưởng của thời gian sử dụng thiết bị điện tử lên sức khỏe tinh thần, bộ dữ liệu được sử dụng được lấy từ nguồn “**Mental Health and Screen Time Correlation Dataset”** một bộ dữ liệu uy tín và được sử dụng rộng rãi trong các nghiên cứu về hành vi và sức khỏe. Bộ dữ liệu này bao gồm thông tin của nhiều người dùng với các đặc trưng liên quan đến thời gian sử dụng thiết bị điện tử mỗi ngày, thói quen công nghệ, mức độ hoạt động xã hội, chất lượng giấc ngủ, cũng như các chỉ số đánh giá sức khỏe tinh thần.

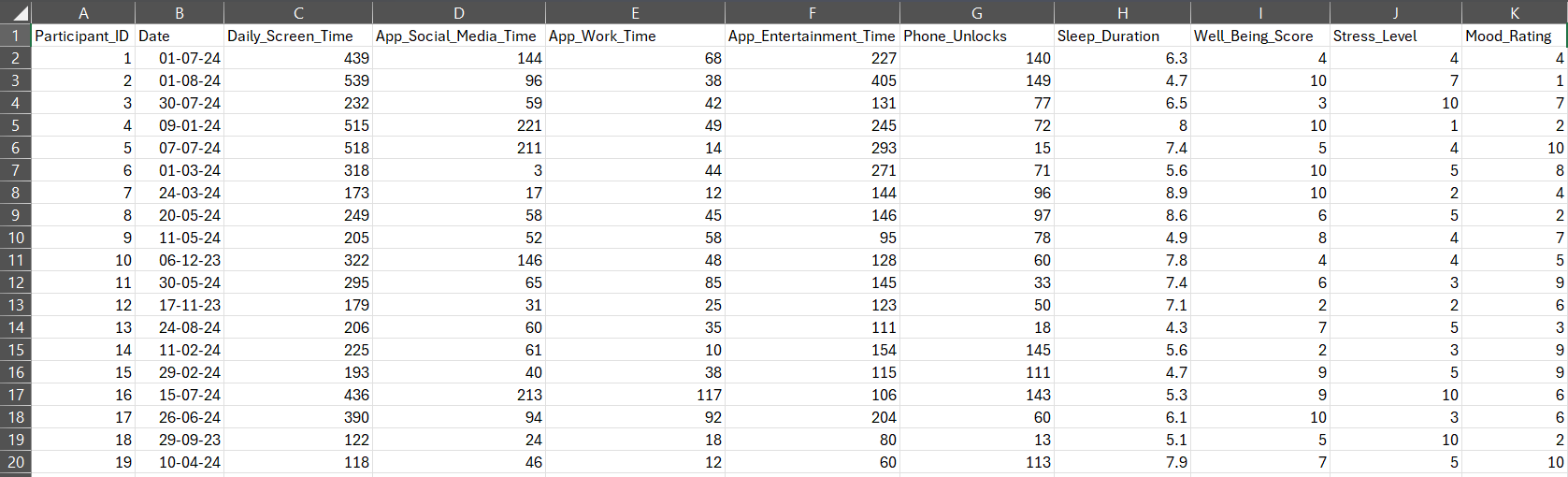
Bộ dữ liệu cung cấp một số biến mục tiêu quan trọng như mental\_health\_score (điểm sức khỏe tinh thần), hoặc mức độ căng thẳng/trầm cảm, cho phép mô hình dự đoán hoặc phân tích mối quan hệ giữa thời lượng sử dụng thiết bị và sự thay đổi về trạng thái tâm lý. Bên cạnh đó, các thuộc tính đầu vào (features) bao gồm daily\_screen\_time, chất lượng giấc ngủ, mức độ vận động, tần suất tham gia mạng xã hội và các yếu tố nhân khẩu học khác giúp mô hình hiểu rõ hơn các yếu tố ảnh hưởng.

Những dữ liệu dạng phân loại như loại thiết bị sử dụng hoặc nhóm độ tuổi sẽ được mã hóa (encoding) sang dạng số để phù hợp với các thuật toán học máy. Trong khi đó, các thuộc tính số như số giờ sử dụng thiết bị, thời gian ngủ hoặc mức độ vận động sẽ được chuẩn hóa nhằm đảm bảo quá trình huấn luyện diễn ra ổn định và hiệu quả.

Việc sử dụng bộ dữ liệu Mental Health and Screen Time đảm bảo rằng tập dữ liệu có độ bao phủ tốt, đa dạng và đủ chi tiết để phân tích các yếu tố ảnh hưởng đến sức khỏe tinh thần. Sau khi tải về, dữ liệu sẽ được kiểm tra, loại bỏ giá trị thiếu, xử lý nhiễu và làm sạch để sẵn sàng cho bước tiền xử lý tiếp theo, từ đó xây dựng được mô hình phân tích – dự đoán ảnh hưởng của thời gian sử dụng thiết bị lên sức khỏe tinh thần một cách chính xác và đáng tin cậy.

Thông tin bộ dữ liệu:

* Tên bộ dữ liệu: Mental Health and Screen Time Correlation Dataset
* Nguồn: [Mental Health and Screen Time Correlation Dataset](https://www.kaggle.com/datasets/yajkotak/mental-health-and-screen-time-correlation-dataset)

****

Hình 3.1: Dữ liệu 20 dòng đầu của dataset

Thông tin cụ thể của các cột trong dataset:

* Participant\_ID (Mã người tham gia): là mã định danh duy nhất được gán cho mỗi người tham gia nghiên cứu nhằm phân biệt các cá nhân trong bộ dữ liệu.
* Date (Ngày thu thập dữ liệu): thể hiện thời điểm ghi nhận dữ liệu của người tham gia, được đo theo lịch dương với định dạng năm–tháng–ngày. Đây là mốc thời gian, cho phép sắp xếp dữ liệu theo trình tự thời gian và phân tích sự thay đổi hành vi, thói quen sử dụng thiết bị và trạng thái tâm lý theo từng ngày.
* Daily Screen Time (Thời gian sử dụng màn hình mỗi ngày): đo tổng thời gian người tham gia sử dụng màn hình trong một ngày, bao gồm tất cả các hoạt động hiển thị nội dung trên màn hình như làm việc, học tập, giải trí và tương tác xã hội. Chỉ số này được tính bằng phút/ngày, thông qua việc cộng dồn toàn bộ thời gian bật màn hình trong ngày. Giá trị càng cao cho thấy mức độ tiếp xúc với màn hình càng lớn, có thể làm gia tăng mệt mỏi, căng thẳng tinh thần và ảnh hưởng tiêu cực đến giấc ngủ.
* App Social Media Time (Thời gian sử dụng ứng dụng mạng xã hội): đo tổng thời gian sử dụng các ứng dụng mạng xã hội trong ngày, được tính bằng phút/ngày. Thời gian được cộng dồn từ các phiên sử dụng để giao tiếp, tương tác nội dung và lướt thông tin. Chỉ số này phản ánh mức độ tương tác xã hội trực tuyến và tiếp xúc thông tin thông qua màn hình.
* App Work Time (Thời gian sử dụng ứng dụng phục vụ công việc/học tập): đo tổng thời gian sử dụng các ứng dụng phục vụ công việc hoặc học tập trong ngày, được tính bằng phút/ngày. Chỉ số này phản ánh cường độ làm việc và học tập thông qua màn hình. Giá trị cao cho thấy mức độ tập trung lớn, nhưng nếu kéo dài có thể dẫn đến quá tải tinh thần.
* Phone Unlocks (Số lần mở khóa thiết bị): đo số lần kích hoạt hoặc mở khóa màn hình trong một ngày, được tính bằng lần/ngày. Đây là biến đo tần suất, phản ánh hành vi kiểm tra thiết bị và mức độ sử dụng màn hình phân mảnh. Giá trị cao cho thấy người dùng thường xuyên bị gián đoạn và khó duy trì sự tập trung liên tục.
* Sleep Duration (Thời lượng giấc ngủ): đo tổng thời gian ngủ trong một ngày, được tính bằng giờ/ngày. Chỉ số này phản ánh chất lượng sinh hoạt cá nhân và có mối liên hệ chặt chẽ với sức khỏe tinh thần. Thời lượng ngủ hợp lý (khoảng 7–9 giờ) góp phần duy trì trạng thái tinh thần ổn định.
* Well-Being Score (Chỉ số hạnh phúc): đo mức độ hạnh phúc và sự hài lòng tổng thể trong cuộc sống, được đánh giá bằng thang điểm định lượng. Đây là chỉ số đại diện cho sức khỏe tinh thần dài hạn. Điểm càng cao cho thấy trạng thái tâm lý càng tích cực. Thang đo: 1- 10.
* Stress Level (Mức độ căng thẳng): đo mức độ căng thẳng tinh thần trong ngày, được biểu diễn bằng thang điểm định lượng. Chỉ số này phản ánh áp lực tâm lý phát sinh từ thói quen sử dụng màn hình và sinh hoạt hằng ngày. Điểm càng cao thể hiện mức độ stress càng lớn. Thang đo: 1- 10.
* Mood Rating (Đánh giá tâm trạng) đo trạng thái cảm xúc tổng thể trong ngày, được đánh giá bằng thang điểm định lượng. Đây là chỉ số phản ánh cảm xúc ngắn hạn và có thể thay đổi theo mức độ stress, giấc ngủ và thời gian sử dụng màn hình. Điểm cao thể hiện tâm trạng tích cực, trong khi điểm thấp phản ánh cảm xúc tiêu cực. Thang đo: 1- 10.

## Môi trường thực nghiệm

### Google Colab

Google Colab là một môi trường phát triển dựa trên nền tảng điện toán đám mây, được sử dụng chủ yếu để lập trình bằng Python. Nó cung cấp môi trường làm việc thông qua trình duyệt web, cho phép người dùng viết và thực thi mã Python trực tiếp mà không cần cài đặt phần mềm. Google Colab tích hợp sẵn các thư viện phổ biến cho khoa học dữ liệu, học máy và trí tuệ nhân tạo, đồng thời hỗ trợ GPU và TPU giúp tăng tốc quá trình huấn luyện mô hình. Ngoài ra, Colab còn cho phép chia sẻ notebook dễ dàng, hỗ trợ làm việc nhóm và tích hợp chặt chẽ với Google Drive.

Google Colab được phát triển bởi Google và dựa trên nền tảng Jupyter Notebook. Đây là một công cụ hướng tới sự tiện lợi và khả năng truy cập cao, đặc biệt phù hợp cho học tập, nghiên cứu và thử nghiệm nhanh các mô hình dữ liệu.

Google Colab là dịch vụ đa nền tảng, có thể sử dụng trên mọi hệ điều hành như Microsoft Windows, macOS và Linux thông qua trình duyệt web. Google Colab có hai phiên bản chính: phiên bản miễn phí, cung cấp tài nguyên tính toán cơ bản với giới hạn thời gian và tài nguyên; và phiên bản trả phí (Colab Pro/Colab Pro+), cung cấp hiệu năng cao hơn, thời gian chạy dài hơn và quyền truy cập ưu tiên vào GPU/TPU. So với phiên bản trả phí, phiên bản miễn phí có nhiều hạn chế hơn về tài nguyên và thời gian sử dụng.

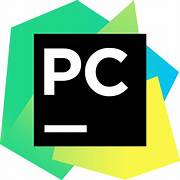


Hình 3.2: Phần mềm Google Colab

### PyCharm

PyCharm là một môi trường phát triển tích hợp (IDE) được sử dụng để lập trình bằng Python. Nó cung cấp các tính năng phân tích mã, trình gỡ lỗi đồ họa, công cụ kiểm tra đơn vị tích hợp, tích hợp với các hệ thống kiểm soát phiên bản, và hỗ trợ phát triển web với Django. PyCharm được phát triển bởi công ty JetBrains của Cộng hòa Séc và được xây dựng trên nền tảng IntelliJ của họ.

PyCharm là phần mềm đa nền tảng, hoạt động trên Microsoft Windows, macOS, và Linux. PyCharm có hai phiên bản: Professional Edition được phát hành theo giấy phép độc quyền và Community Edition được phát hành theo giấy phép Apache. PyCharm Community Edition có ít tính năng hơn so với Professional Edition.



Hình 3.3: Phần mềm Pycharm

### Các thư viện sử dụng

Một số thư viện cần thiết cho việc xử lý và phân tích dữ liệu, những thư viện này cung cấp các công cụ mạnh mẽ cho việc khám phá và xử lý dữ liệu, cũng như trực quan hóa thông tin một cách hiệu quả, cụ thể:

* Seaborn (sns) và Matplotlib (plt): Seaborn là một thư viện trực quan hóa dữ liệu dựa trên Matplotlib, giúp tạo ra các biểu đồ thống kê và trực quan hóa dữ liệu một cách dễ dàng và đẹp mắt.
* Missingno (msno): Missingno là một thư viện hỗ trợ kiểm tra và trực quan hóa dữ liệu thiếu sót (missing data) trong bộ dữ liệu, giúp phân tích dữ liệu và quyết định cách xử lý giá trị thiếu.
* Plotly Express (px), Plotly Graph Objects (go) và Make Subplots: Plotly là một thư viện mạnh mẽ cho việc tạo các biểu đồ tương tác. Plotly Express (px) cung cấp cách dễ dàng để tạo các biểu đồ phức tạp, trong khi Plotly Graph Objects (go) cho phép tùy chỉnh chi tiết hơn. Make Subplots giúp tạo ra các subplot để so sánh và hiển thị nhiều biểu đồ trên cùng một hình ảnh.
* Plotly Figure Factory (ff): Figure Factory trong Plotly là một bộ công cụ cho việc tạo các biểu đồ phức tạp và đa dạng, giúp bạn mở rộng khả năng trực quan hóa của mình.
* StandardScaler từ sklearn.preprocessing: StandardScaler là một phương pháp chuẩn hóa dữ liệu trong khoảng (0,1) hoặc theo phân phối chuẩn, giúp cân bằng và chuẩn hóa dữ liệu trước khi đưa vào mô hình học máy.

## Quy trình thực nghiệm

### Đặt mục tiêu

* Tiền xử lý dữ liệu: làm sạch bộ dữ liệu không còn bị khuyết và trùng lặp dữ liệu.
* Phân tích mô tả: thể hiện được mối quan hệ giữa các giá trị của dữ liệu, từ đó đánh giá được mối tương quan giữa chúng.
* Phân tích dự báo: xây dựng được mô hình hồi quy tuyến tính để dự báo giá kim cương.

### Tiền xử lý dữ liệu

Để bắt đầu bước tiền xử lý dữ liệu ta cần import các thư viện cần thiết cho

cả bài toán:

A computer code with text

AI-generated content may be incorrect.

Hình 3.4: Các thư viện cần thiết

#### Kiểm tra dữ liệu

Kiểm tra dữ liệu là bước đầu tiên trong quá trình tiền xử lý dữ liệu. Sau quá trình kiểm tra dữ liệu sẽ cung cấp cho chúng ta các thông tin đầy đủ về bộ dữ liệu từ đó ta có thể đưa ra các bước tiếp theo trong xử lý dữ liệu.

A screenshot of a computer program

AI-generated content may be incorrect.

Hình 3.5: Tóm tắt bộ dataset

#### Làm sạch dữ liệu

Làm sạch dữ liệu là quá trình loại bỏ các sai sót, lỗi, nhiễu và thông tin không chính xác hoặc không cần thiết khỏi tập dữ liệu ban đầu để đảm bảo dữ liệu đáng tin cậy và phù hợp cho việc phân tích và xử lý tiếp theo. Một số tác vụ chính trong quá trình làm sạch dữ liệu bao gồm:

* Kiểm tra dữ liệu bị thiếu: Dữ liệu thiếu là dữ liệu không hợp lệ và chúng ta phải xử lý bằng một số phương pháp như điền khuyết. Trước hết ta cần phải kiểm tra dữ liệu thiếu:

A screenshot of a computer code

AI-generated content may be incorrect.

Hình 3.6: Kiểm tra dữ liệu thiếu

Như vậy, ta thấy bộ dữ liệu xuất hiện giá trị bị khuyết. Sau khi điền khuyết ta có bộ dữ liệu gồm 8691 bản ghi

A screenshot of a computer program

AI-generated content may be incorrect.

Hình 3.7: Bộ dữ liệu sau khi điền khuyết

* Loại bỏ trùng lặp dữ liệu: Khi loại bỏ các dữ liệu trùng lặp giúp giảm bớt khối lượng bộ nhớ. Loại bỏ dữ liệu trùng lặp không làm thay đổi dữ liệu ban đầu mà còn giảm bớt việc phải tính toán lại do trùng dữ liệu:
  + Nếu dữ liệu không có lý do hợp lý để trùng lặp: Trong bài toán phântích giá kim cương, ta có thể loại bỏ chúng vì chúng không mang thêm giá trị thông tin.
  + Để giảm nhiễu trong mô hình học máy: Các dòng trùng lặp có thể làm lệch kết quả phân tích thống kê hoặc huấn luyện mô hình vì nó tạo ra trọng số không đồng đều cho các quan sát.
  + Để tối ưu kích thước bộ dữ liệu: Trong trường hợp bộ dữ liệu lớn, việc loại bỏ các dòng trùng lặp giúp giảm bớt khối lượng tính toán.

A computer code with black text

AI-generated content may be incorrect.

Hình 3.8: Đoạn mã kiểm tra dữ liệu trùng lặp

A screenshot of a screen

AI-generated content may be incorrect.

Hình 3.9: Đếm các dòng trùng lặp

Như vậy, ta thấy bộ dữ liệu xuất hiện 149 dòng trùng lặp. Sau khi loại bỏ, bộ dữ liệu giảm từ 8691 dòng xuống còn 8542 dòng.

**A screenshot of a computer code

AI-generated content may be incorrect.**

Hình 3.10: Dữ liệu sau khi loại bỏ các dòng trùng lặp

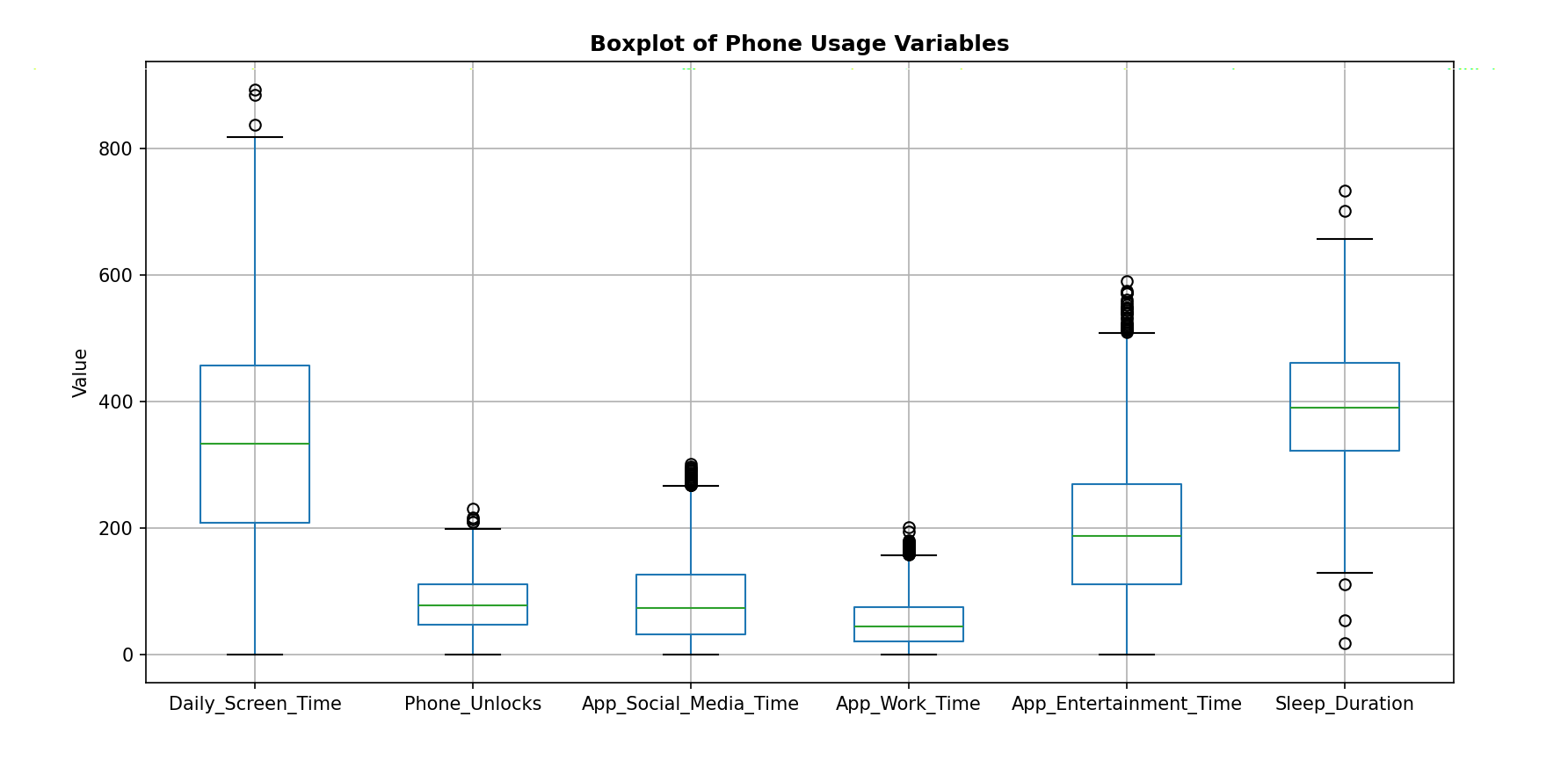
### Phân tích mô tả

Phân tích mô tả trong phân tích dữ liệu là quá trình tóm tắt, mô tả và hiểu sâu về các đặc điểm, mẫu thái và thông tin quan trọng của tập dữ liệu. Với mục tiêu đó, ta sẽ tiến hành phân tích mô tả cho bộ dữ liệu của project theo cả 2 hướng phân tích đơn biến (trên từng biến) và phân tích đa biến (trên nhiều biến) bằng cách biểu diễn dưới các biểu đồ khác nhau.

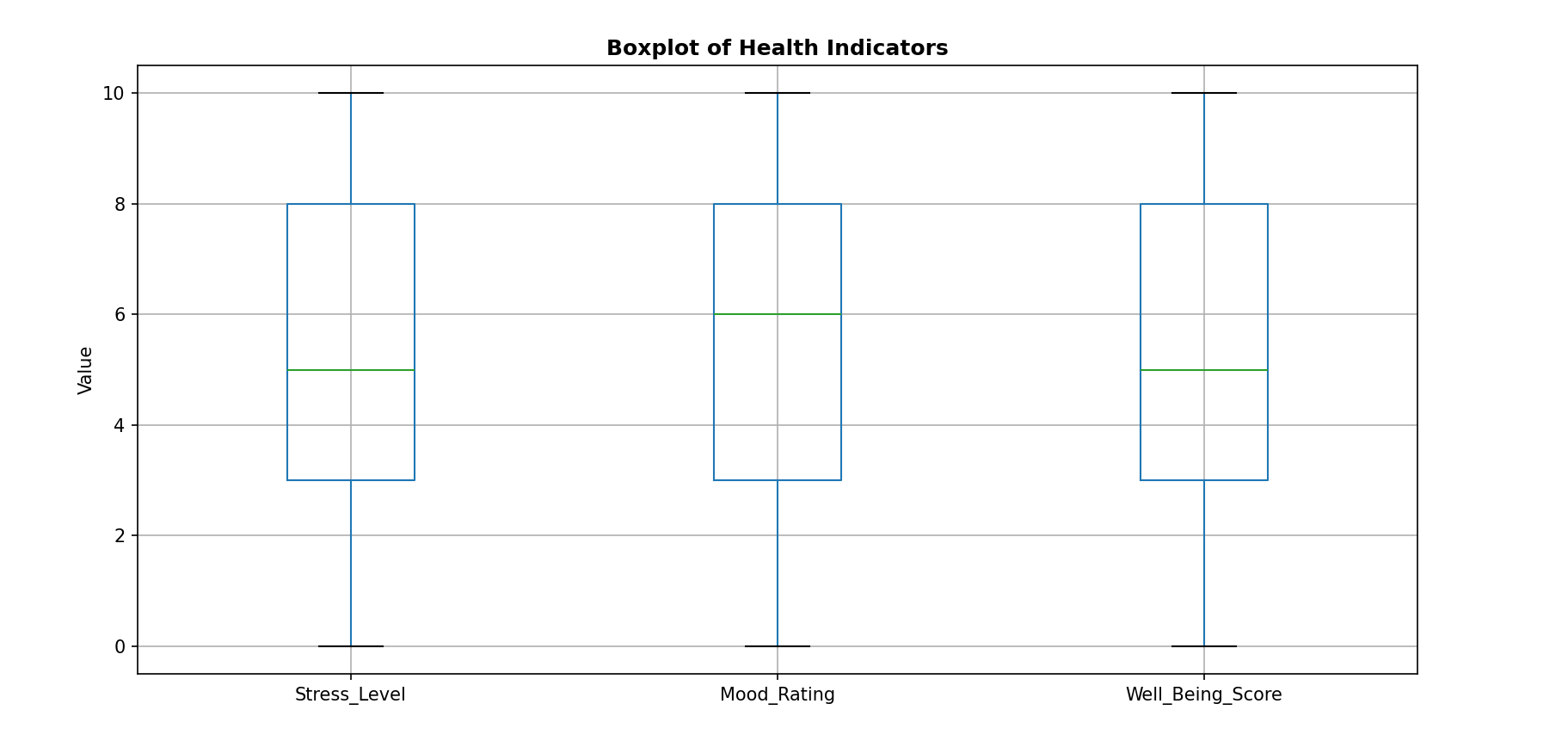
#### Phân tích đơn biến

*Biểu đồ 1: Biểu đồ hộp (Box & Whisker Plot)*

* Dạng biểu đồ: Boxplot
* Nhóm biến phân tích:
* Hành vi sử dụng điện thoại: Daily\_Screen\_Time, Phone\_Unlocks, App\_Social\_Media\_Time, App\_Work\_Time, App\_Entertainment\_Time, Sleep\_Duration
* Chỉ số sức khỏe tinh thần: Stress\_Level, Mood\_Rating, Well\_Being\_Score
* Mục đích: Biểu đồ hộp được sử dụng để đánh giá phân bố dữ liệu, độ phân tán, trung vị và các giá trị ngoại lai của cả nhóm biến hành vi sử dụng điện thoại và nhóm biến sức khỏe tinh thần, từ đó cung cấp cái nhìn tổng quan về sự khác biệt giữa các cá nhân trong bộ dữ liệu.



Hình 3.11: Biểu đồ Boxplot của nhóm hành vi sử dụng thiết bị

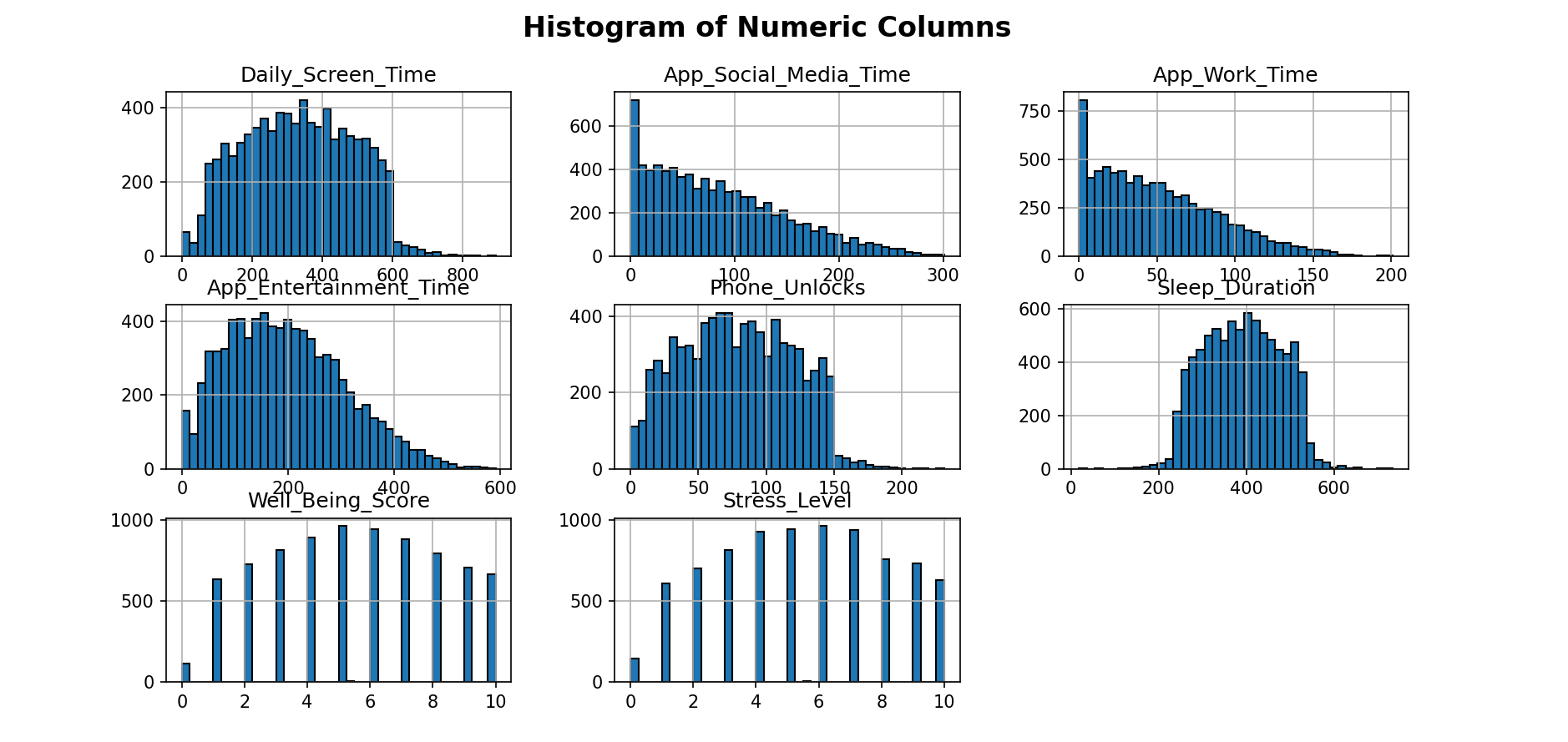


Hình 3.12: Biểu đồ Boxplot của nhóm chỉ số sức

* Nhận xét
* Nhóm biến sử dụng điện thoại cho thấy độ biến thiên lớn, đặc biệt là Daily\_Screen\_Time và App\_Entertainment\_Time. Các biến này có IQR rộng và nhiều giá trị ngoại lai cao, phản ánh sự khác biệt rõ rệt trong thói quen sử dụng điện thoại giữa người dùng.
* Phone\_Unlocks và App\_Social\_Media\_Time có trung vị ở mức trung bình nhưng vẫn xuất hiện một số ngoại lai đáng kể, cho thấy một bộ phận người dùng kiểm tra điện thoại và sử dụng mạng xã hội với tần suất cao hơn hẳn so với phần lớn mẫu dữ liệu.
* App\_Work\_Time có phân bố tương đối tập trung hơn, cho thấy thời gian sử dụng ứng dụng phục vụ công việc ít biến động hơn so với các loại ứng dụng khác.
* Sleep\_Duration có phân bố tương đối ổn định quanh trung vị, tuy nhiên vẫn tồn tại ngoại lai thấp và cao, phản ánh sự không đồng đều về thời lượng giấc ngủ giữa các cá nhân.
* Nhóm chỉ số sức khỏe tinh thần (Stress\_Level, Mood\_Rating, Well\_Being\_Score) có phân bố trải rộng trên toàn thang đo. Trong đó, Stress\_Level có mức phân tán lớn, trong khi Mood\_Rating và Well\_Being\_Score có trung vị cao hơn, cho thấy phần lớn người dùng có trạng thái tâm lý tương đối tích cực nhưng vẫn tồn tại một nhóm có sức khỏe tinh thần kém.
* Kết luận: Biểu đồ hộp cho thấy sự khác biệt rõ rệt giữa các cá nhân cả về hành vi sử dụng điện thoại lẫn sức khỏe tinh thần. Đặc biệt, sự xuất hiện của nhiều giá trị ngoại lai ở nhóm biến sử dụng điện thoại gợi ý rằng việc sử dụng điện thoại quá mức có thể là yếu tố nguy cơ ảnh hưởng đến sức khỏe tinh thần, cần được phân tích sâu hơn trong các phần phân tích mối quan hệ và mô hình hóa tiếp theo.

*Biểu đồ 2: Biểu đồ Histogram*

* Dạng biểu đồ: Histogram
* Loại phân tích: Đơn biến
* Các biến phân tích: Daily\_Screen\_Time, App\_Social\_Media\_Time, App\_Work\_Time, App\_Entertainment\_Time, Phone\_Unlocks, Sleep\_Duration, Well\_Being\_Score, Stress\_Level
* Kiểu dữ liệu: Số
* Mục đích: Biểu đồ histogram được sử dụng để quan sát dạng phân phối của từng biến, xác định xu hướng tập trung, độ lệch và sự xuất hiện của các giá trị bất thường.



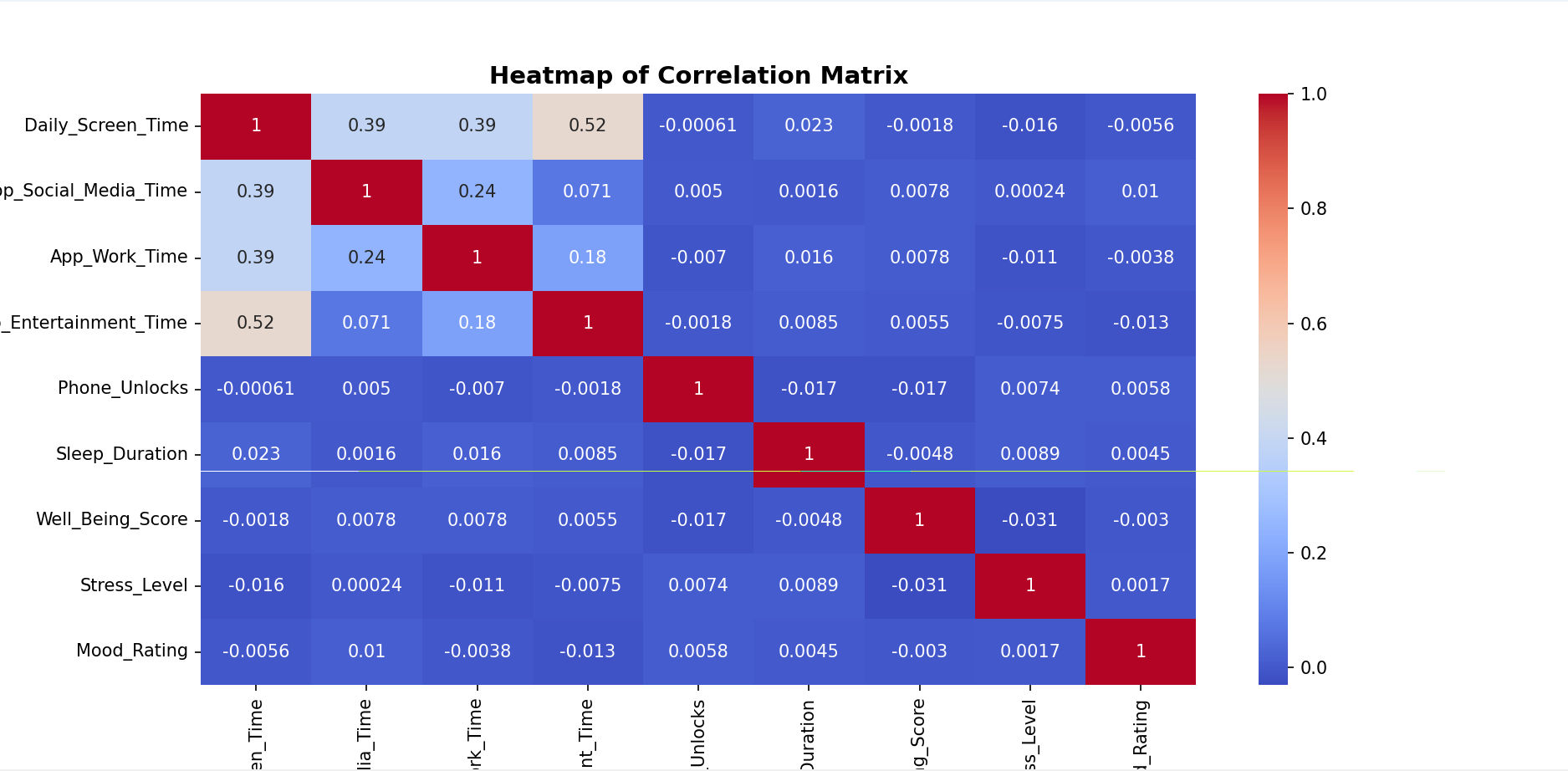
Hình 3.13: Biểu đồ Histogram

* Nhận xét:
  + Daily\_Screen\_Time và App\_Entertainment\_Time có phân phối lệch phải, cho thấy phần lớn người dùng có mức sử dụng trung bình, trong khi một số ít người có thời gian sử dụng rất cao.
  + App\_Social\_Media\_Time và App\_Work\_Time tập trung nhiều ở các giá trị thấp, phản ánh rằng đa số người dùng không dành quá nhiều thời gian cho từng loại ứng dụng riêng lẻ.
  + Phone\_Unlocks có phân phối trải rộng, cho thấy thói quen kiểm tra điện thoại diễn ra với tần suất rất khác nhau giữa các cá nhân.
  + Sleep\_Duration có phân phối gần chuẩn, tập trung quanh khoảng thời gian ngủ trung bình, cho thấy thói quen ngủ tương đối ổn định.
  + Well\_Being\_Score và Stress\_Level chủ yếu tập trung ở các mức trung bình, rất ít giá trị cực đoan.
* Kết luận: Các biểu đồ histogram cho thấy dữ liệu không hoàn toàn tuân theo phân phối chuẩn, đặc biệt đối với các biến liên quan đến thời gian sử dụng màn hình. Điều này phản ánh sự đa dạng trong hành vi sử dụng thiết bị điện tử và là cơ sở cho việc phân tích mối quan hệ giữa các biến này với sức khỏe tinh thần.

#### Phân tích đa biến

*Biểu đồ 3: Biểu đồ Heatmap*

* Dạng biểu đồ: Heatmap (Ma trận tương quan)
* Loại phân tích: Đa biến
* Các biến phân tích: *Daily\_Screen\_Time, App\_Social\_Media\_Time, App\_Work\_Time, App\_Entertainment\_Time, Phone\_Unlocks, Sleep\_Duration, Well\_Being\_Score, Stress\_Level, Mood\_Rating*
* Kiểu dữ liệu: Số (int / float)
* Mục đích: Biểu đồ heatmap được sử dụng để đánh giá mức độ và chiều hướng tương quan tuyến tính giữa các biến liên quan đến hành vi sử dụng điện thoại và các chỉ số sức khỏe tinh thần, từ đó xác định các mối quan hệ nổi bật cần được phân tích sâu hơn.

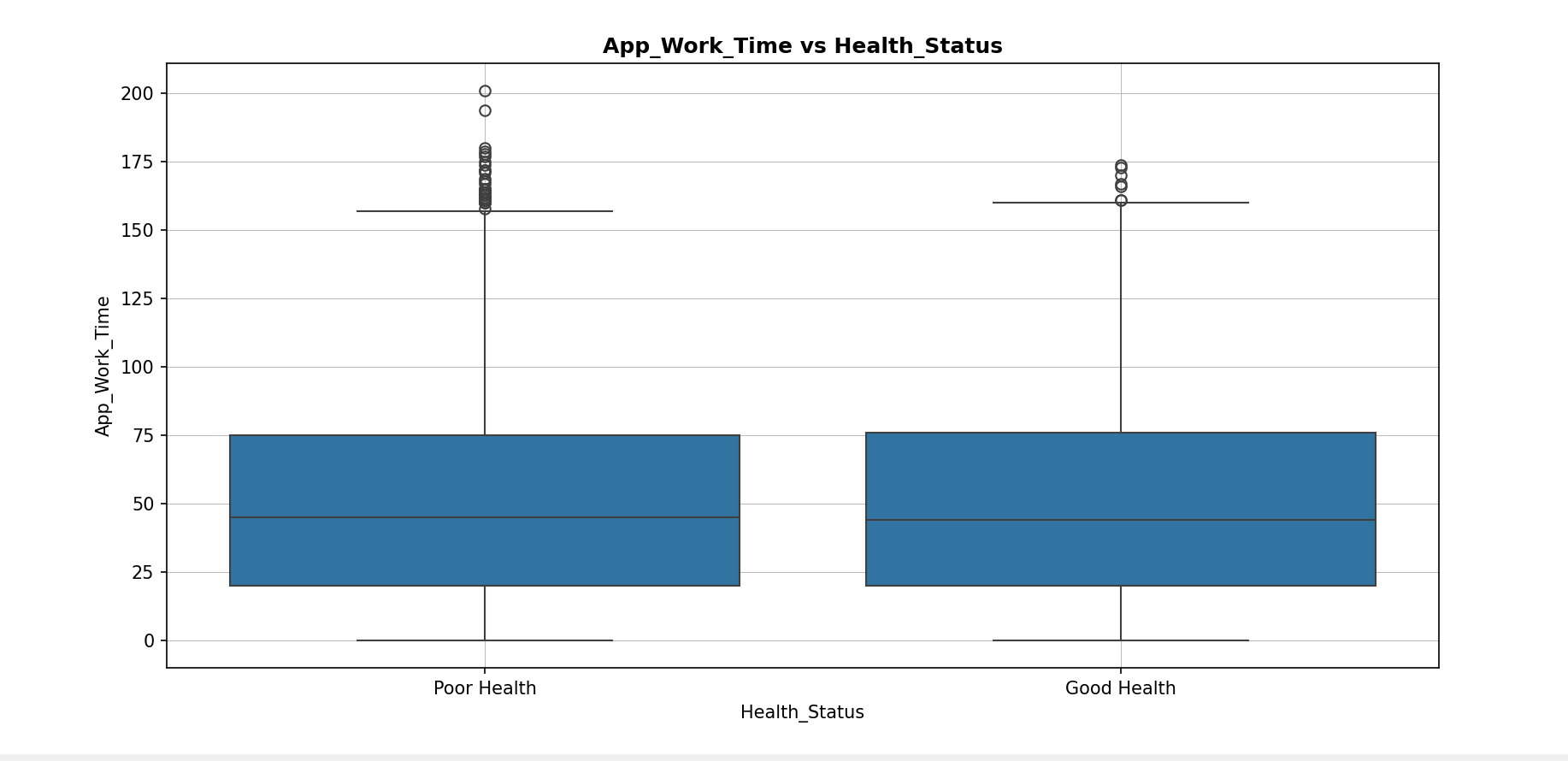


Hình 3.14: Biểu đồ heatmap

* Nhận xét
* Daily\_Screen\_Time có tương quan dương mức trung bình với *App\_Entertainment\_Time* (r ≈ 0.52), cho thấy thời gian sử dụng màn hình chủ yếu đến từ các hoạt động giải trí.
* Daily\_Screen\_Time cũng có tương quan dương nhẹ với *App\_Social\_Media\_Time* và *App\_Work\_Time* (r ≈ 0.39), phản ánh cả mạng xã hội và công việc đều đóng góp đáng kể vào tổng thời gian sử dụng điện thoại.
* Các biến hành vi sử dụng điện thoại có tương quan mạnh với nhau, nhưng tương quan rất yếu hoặc gần như bằng 0 với các chỉ số sức khỏe như *Sleep\_Duration, Well\_Being\_Score, Stress\_Level* và *Mood\_Rating*.
* Well\_Being\_Score có tương quan âm nhẹ với *Stress\_Level* (r ≈ −0.03), phù hợp với kỳ vọng rằng mức độ căng thẳng tăng lên sẽ làm giảm cảm nhận hạnh phúc, tuy nhiên mức độ ảnh hưởng là không lớn.
* Stress\_Level và Mood\_Rating gần như không có tương quan tuyến tính rõ ràng với các biến sử dụng điện thoại trong bộ dữ liệu này.
* Kết luận: Ma trận tương quan cho thấy mối liên hệ mạnh chủ yếu tồn tại giữa các biến hành vi sử dụng điện thoại, trong khi mối quan hệ tuyến tính giữa việc sử dụng điện thoại và sức khỏe tinh thần là khá yếu. Điều này gợi ý rằng tác động của việc sử dụng điện thoại đến sức khỏe không đơn giản và có thể mang tính phi tuyến, hoặc chịu ảnh hưởng đồng thời của nhiều yếu tố trung gian khác như giấc ngủ và mức độ căng thẳng.

*Biểu đồ 4: Thời gian sử dụng ứng dụng công việc theo tình trạng sức khỏe*

* Dạng biểu đồ: Box & Whisker Plot
* Loại phân tích: Song biến
* Các biến phân tích:
  + App\_Work\_Time (thời gian sử dụng ứng dụng công việc)
  + Health\_Status (Tình trạng sức khỏe: Tốt / Kém)
* Kiểu dữ liệu:
  + App\_Work\_Time: Số (liên tục)
  + Health\_Status: Phân loại (Categorical)
* Mục đích: Biểu đồ nhằm so sánh phân bố thời gian sử dụng ứng dụng công việc giữa hai nhóm tình trạng sức khỏe khác nhau, từ đó đánh giá liệu mức độ sử dụng ứng dụng phục vụ công việc có liên quan đến tình trạng sức khỏe tổng thể hay không.

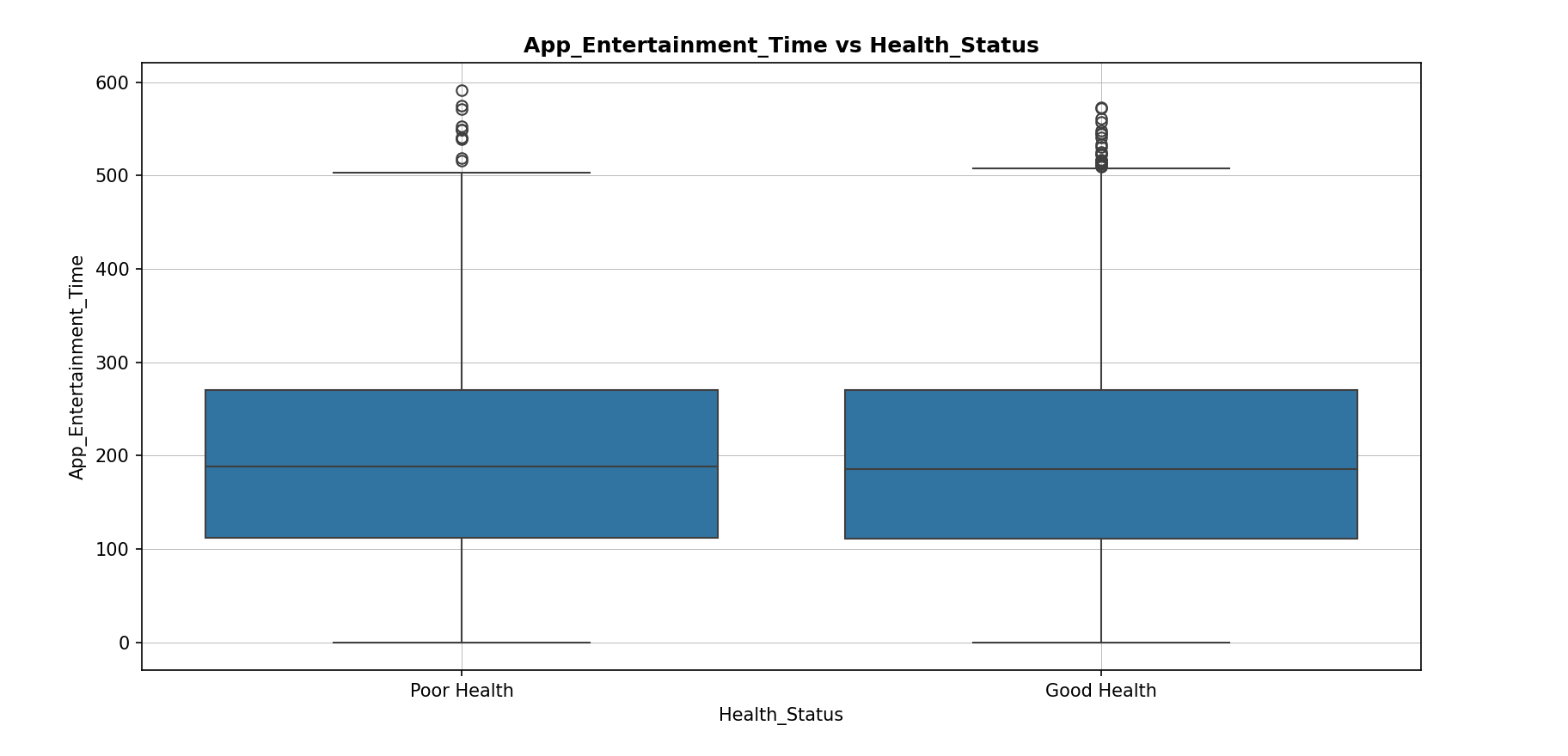


Hình 3.15: Biểu đồ Boxplot của App\_work\_time và health\_status

* Nhận xét
  + Hai nhóm Good Health và Poor Health có giá trị trung vị của App\_Work\_Time khá tương đồng, cho thấy thời gian sử dụng ứng dụng công việc trung bình không khác biệt rõ rệt giữa hai nhóm.
  + Độ phân tán (IQR) của hai nhóm tương đối giống nhau, phản ánh sự đa dạng tương đương trong hành vi sử dụng ứng dụng công việc.
  + Cả hai nhóm đều xuất hiện nhiều giá trị ngoại lai ở mức cao, cho thấy tồn tại một bộ phận người dùng dành rất nhiều thời gian cho các ứng dụng công việc, bất kể tình trạng sức khỏe.
  + Nhóm Poor Health có xu hướng xuất hiện outliers ở mức cao hơn một chút, gợi ý khả năng rằng việc sử dụng ứng dụng công việc quá nhiều có thể liên quan đến tình trạng sức khỏe kém, dù sự khác biệt chưa thực sự rõ ràng.
* Kết luận: Biểu đồ cho thấy chưa có sự khác biệt đáng kể về thời gian sử dụng ứng dụng công việc giữa nhóm sức khỏe tốt và kém. Điều này cho thấy mức độ sử dụng ứng dụng công việc không phải là yếu tố quyết định trực tiếp đến tình trạng sức khỏe tổng thể, mà có thể chịu ảnh hưởng đồng thời bởi các yếu tố khác như mức độ căng thẳng, giấc ngủ hoặc cân bằng công việc – cuộc sống.

*Biểu đồ 5: Thời gian sử dụng ứng dụng giải trí theo tình trạng sức khỏe*

* Dạng biểu đồ: Box & Whisker Plot
* Loại phân tích: Song biến
* Các biến phân tích: App\_Entertainment\_Time (thời gian sử dụng ứng dụng giải trí), Health\_Status (Tình trạng sức khỏe: Tốt / Kém)
* Kiểu dữ liệu:
  + App\_Entertainment\_Time: Số (liên tục)
  + Health\_Status: Phân loại
* Mục đích: Biểu đồ được sử dụng để so sánh mức độ sử dụng ứng dụng giải trí giữa hai nhóm tình trạng sức khỏe, nhằm đánh giá tác động tiềm ẩn của hành vi giải trí trên điện thoại đến sức khỏe tổng thể.

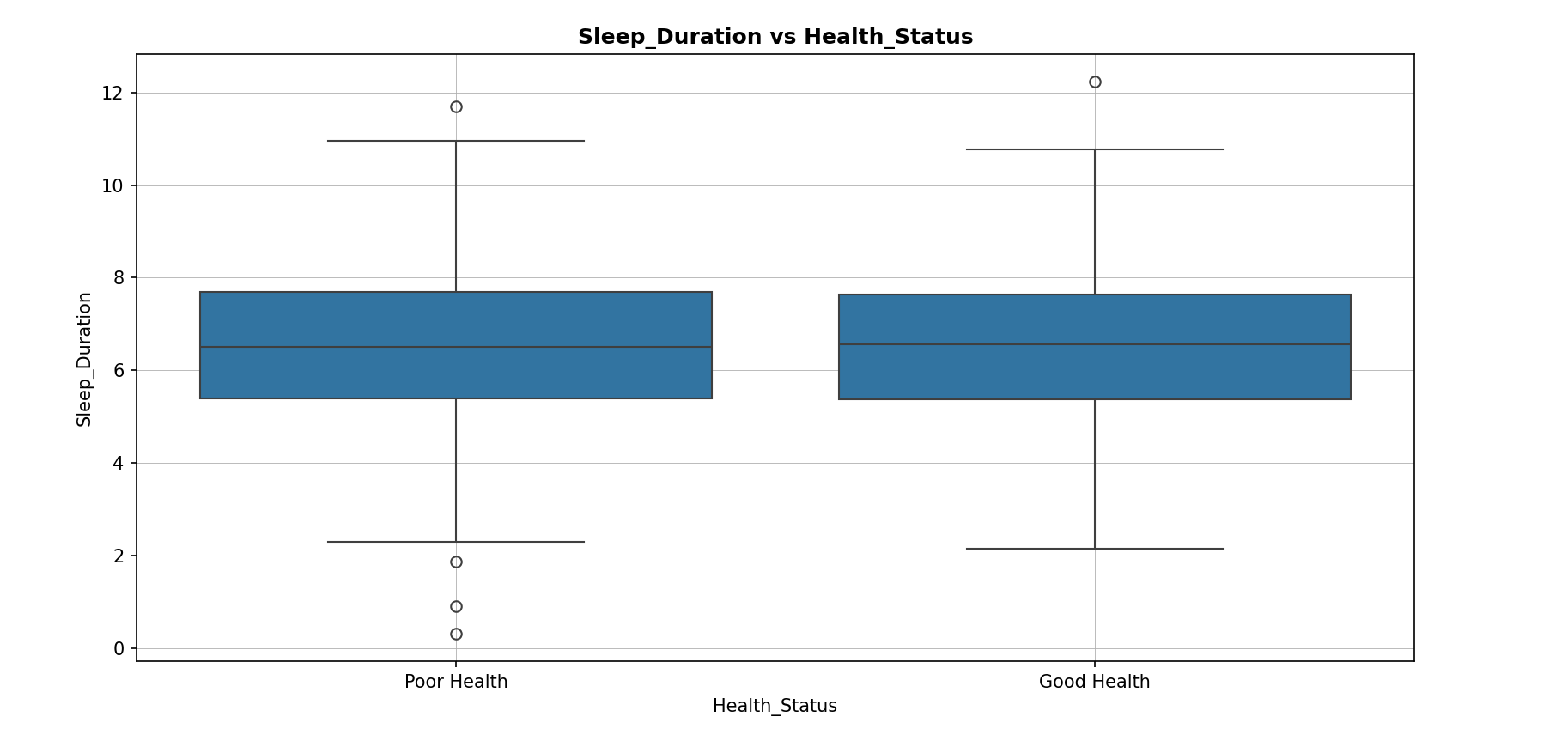


Hình 3.16: Biểu đồ Boxplot của App\_entertainment\_time và health\_status

* Nhận xét
  + Cả hai nhóm Good Health và Poor Health có giá trị trung vị gần tương đương, cho thấy mức sử dụng ứng dụng giải trí trung bình không có sự chênh lệch rõ rệt.
  + IQR của nhóm Poor Health rộng hơn, phản ánh sự khác biệt lớn hơn trong hành vi giải trí của nhóm này, bao gồm cả những người sử dụng rất ít và rất nhiều.
  + Nhiều giá trị ngoại lai ở mức cao xuất hiện ở cả hai nhóm, cho thấy một số người dùng dành thời gian rất lớn cho các ứng dụng giải trí, bất kể tình trạng sức khỏe.
  + Nhóm Poor Health có xu hướng xuất hiện nhiều outliers cực trị hơn, gợi ý rằng việc sử dụng ứng dụng giải trí quá mức có thể liên quan đến sức khỏe kém ở một bộ phận người dùng.
* Kết luận: Biểu đồ cho thấy không tồn tại sự khác biệt rõ ràng về trung vị thời gian sử dụng ứng dụng giải trí giữa hai nhóm sức khỏe. Tuy nhiên, độ phân tán lớn và sự xuất hiện của nhiều giá trị ngoại lai, đặc biệt ở nhóm sức khỏe kém, cho thấy việc sử dụng ứng dụng giải trí quá mức có thể là yếu tố rủi ro tiềm ẩn, cần được phân tích sâu hơn kết hợp với các chỉ số khác như giấc ngủ và mức độ căng thẳng.

*Biểu đồ 6: Sleep\_Duration vs Health\_Status*

* Dạng biểu đồ: Box Plot (Biểu đồ hộp)
* Loại phân tích: Song biến
* Các biến phân tích
  + Sleep\_Duration (thời lượng ngủ, tính theo giờ)
  + Health\_Status (tình trạng sức khỏe, gồm hai nhóm: Poor Health và Good Health)
* Kiểu dữ liệu
  + Biến độc lập: Health\_Status (dữ liệu phân loại – định tính)
  + Biến phụ thuộc: Sleep\_Duration (dữ liệu số liên tục)
* Mục đích: Biểu đồ được sử dụng để so sánh sự phân bố thời lượng ngủ giữa hai nhóm tình trạng sức khỏe khác nhau (sức khỏe kém và sức khỏe tốt). Thông qua biểu đồ hộp, có thể quan sát sự khác biệt về giá trị trung vị, mức độ phân tán cũng như các giá trị ngoại lai của thời gian ngủ, từ đó đánh giá mối liên hệ giữa thời lượng ngủ và tình trạng sức khỏe tổng thể.

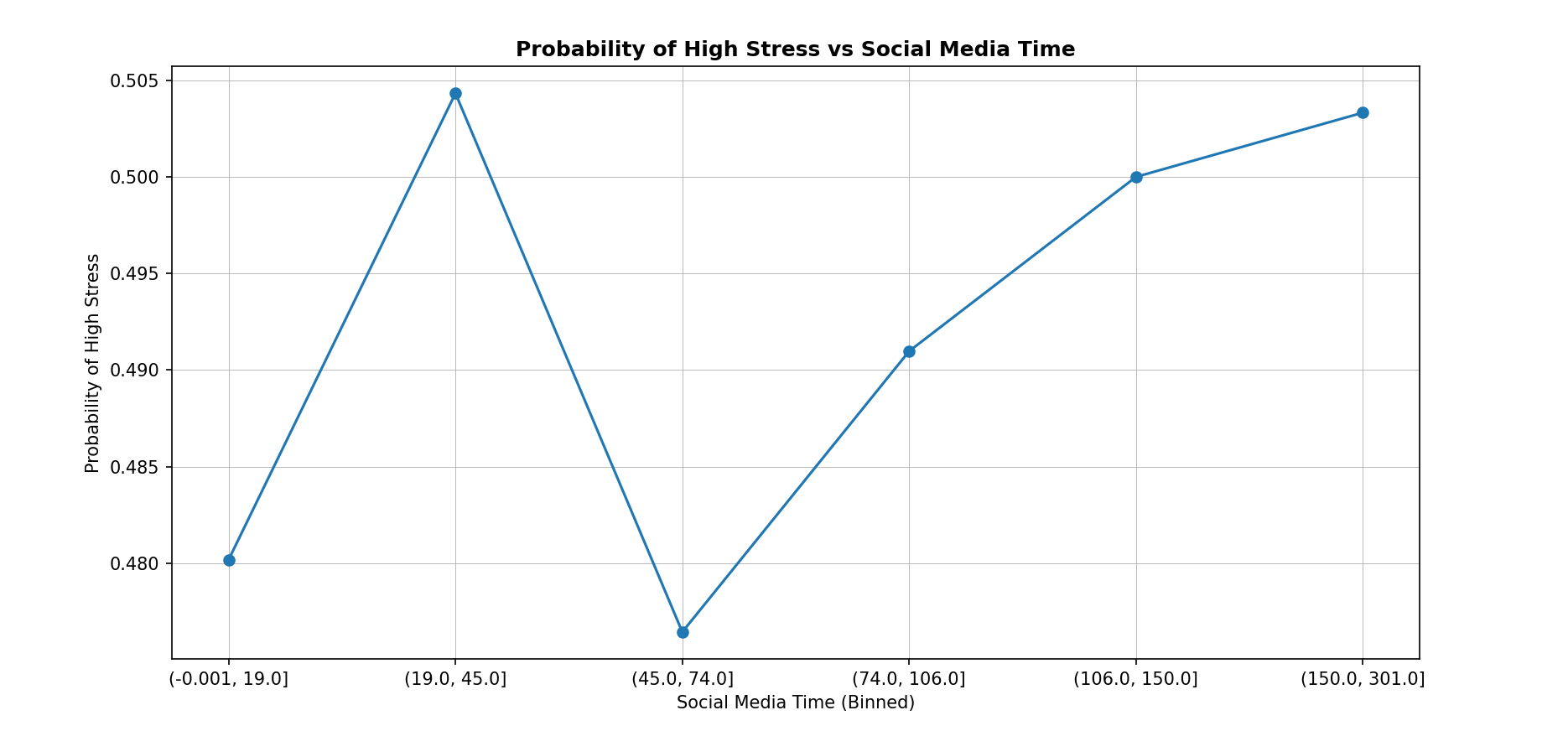


Hình 3.17: Biểu đồ Boxplot của Sleep\_duration và health\_status

* Nhận xét
  + Ở nhóm Poor Health, thời lượng ngủ có trung vị ở mức khoảng 6–6,5 giờ, với độ phân tán tương đối lớn. Nhóm này xuất hiện một số giá trị ngoại lai ở mức rất thấp (ngủ rất ít) và một vài giá trị cao, cho thấy thói quen ngủ không ổn định.
  + Ở nhóm Good Health, trung vị thời lượng ngủ cũng nằm quanh mức 6–7 giờ, tuy nhiên phân bố có xu hướng cân đối hơn và ít giá trị cực đoan thấp so với nhóm sức khỏe kém.
  + So sánh hai nhóm cho thấy thời lượng ngủ của nhóm có sức khỏe tốt nhìn chung nhỉnh hơn và ổn định hơn, trong khi nhóm sức khỏe kém có sự dao động lớn hơn, đặc biệt là hiện tượng ngủ quá ít.
  + Mặc dù khoảng giá trị giữa hai nhóm có sự chồng lấn đáng kể, nhưng sự khác biệt về mức độ phân tán và các ngoại lai cho thấy chất lượng và tính đều đặn của giấc ngủ có thể liên quan đến tình trạng sức khỏe.
* Kết luận: Biểu đồ cho thấy mối liên hệ nhất định giữa thời lượng ngủ và tình trạng sức khỏe. Những người có sức khỏe tốt thường duy trì thời gian ngủ ổn định hơn, trong khi nhóm sức khỏe kém có xu hướng ngủ thất thường và xuất hiện nhiều trường hợp ngủ quá ít. Điều này gợi ý rằng giấc ngủ đóng vai trò quan trọng trong việc duy trì sức khỏe tổng thể, và việc thiếu ngủ hoặc ngủ không đều có thể là một yếu tố liên quan đến tình trạng sức khỏe kém. Kết quả này phù hợp với các nghiên cứu về tác động của thói quen ngủ đối với sức khỏe thể chất và tinh thần.

*Biểu đồ 3: Xác suất căng thẳng cao theo thời gian sử dụng mạng xã hội*

* Dạng biểu đồ: Line Chart
* Loại phân tích: Song biến
* Các biến phân tích:
  + *App\_Social\_Media\_Time* (đã phân nhóm)
  + *Stress\_Level* (quy đổi thành xác suất căng thẳng cao)
* Kiểu dữ liệu:
  + Biến độc lập: Số (liên tục, phân nhóm)
  + Biến phụ thuộc: Xác suất
* Mục đích: Biểu đồ nhằm phân tích xu hướng thay đổi xác suất căng thẳng cao theo mức độ sử dụng mạng xã hội, dựa trên dữ liệu đã được điều chỉnh.

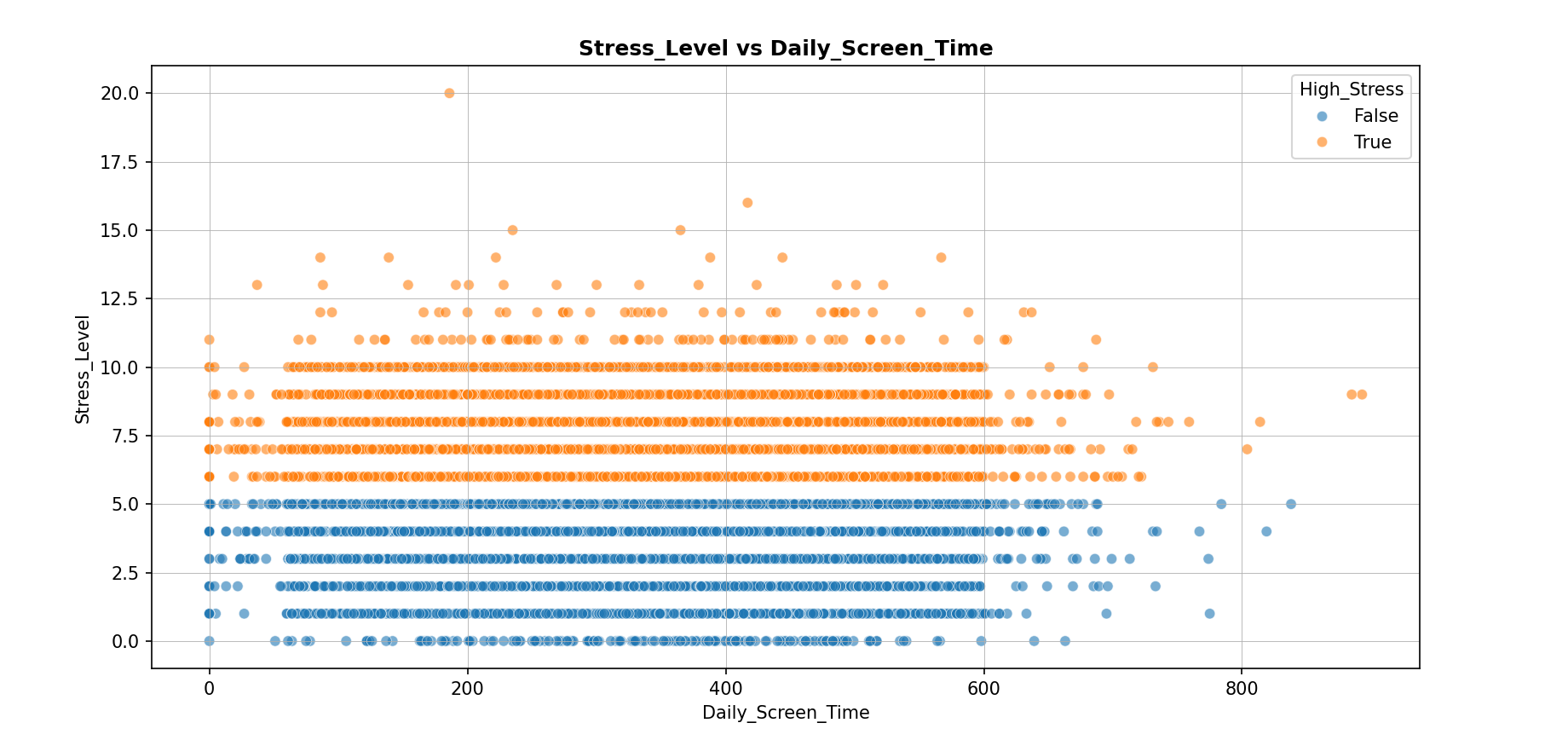


Hình 3.18: Biểu đồ đường của App\_Social\_Media\_Time với stress\_level

* Nhận xét:
* Ở nhóm sử dụng mạng xã hội thấp nhất, xác suất căng thẳng cao ở mức thấp nhất trong các nhóm, cho thấy xu hướng ít căng thẳng hơn ở nhóm người dùng ít tiếp xúc với mạng xã hội.
* Khi thời gian sử dụng mạng xã hội tăng lên mức trung bình thấp, xác suất căng thẳng cao tăng rõ rệt, phản ánh tác động ban đầu của việc gia tăng sử dụng mạng xã hội.
* Tại một khoảng trung gian, xác suất căng thẳng có xu hướng giảm nhẹ, cho thấy mối quan hệ không hoàn toàn tuyến tính.
* Ở các nhóm sử dụng mạng xã hội cao, xác suất căng thẳng cao tăng trở lại và đạt mức cao nhất, thể hiện xu hướng gia tăng nguy cơ căng thẳng khi thời gian sử dụng mạng xã hội kéo dài.
* Kết luận: Biểu đồ cho thấy mối quan hệ dương không tuyến tính giữa thời gian sử dụng mạng xã hội và xác suất căng thẳng cao. Mặc dù ma trận tương quan cho thấy mối liên hệ tuyến tính yếu, phân tích theo nhóm cho thấy việc sử dụng mạng xã hội nhiều vẫn có xu hướng làm tăng nguy cơ căng thẳng, đặc biệt ở nhóm người dùng có thời gian sử dụng cao.

*Biểu đồ 4: Stress\_Level vs Daily\_Screen\_Time*

* Dạng biểu đồ: Scatter Plot (Biểu đồ phân tán)
* Loại phân tích: Song biến
* Các biến phân tích: Daily\_Screen\_Time (thời gian sử dụng màn hình mỗi ngày), Stress\_Level (mức độ căng thẳng), Biến phân nhóm hiển thị màu: High\_Stress (căng thẳng cao: True / False)
* Kiểu dữ liệu
  + Biến độc lập: Daily\_Screen\_Time (số liên tục)
  + Biến phụ thuộc: Stress\_Level (số rời rạc)
  + Biến phân loại bổ trợ: High\_Stress (nhị phân)
* Mục đích: Biểu đồ được sử dụng để phân tích mối quan hệ giữa thời gian sử dụng màn hình hằng ngày và mức độ căng thẳng, đồng thời quan sát sự khác biệt giữa nhóm có căng thẳng cao và không cao. Việc sử dụng biểu đồ phân tán giúp nhận diện xu hướng tổng quát và mức độ phân bố của dữ liệu.

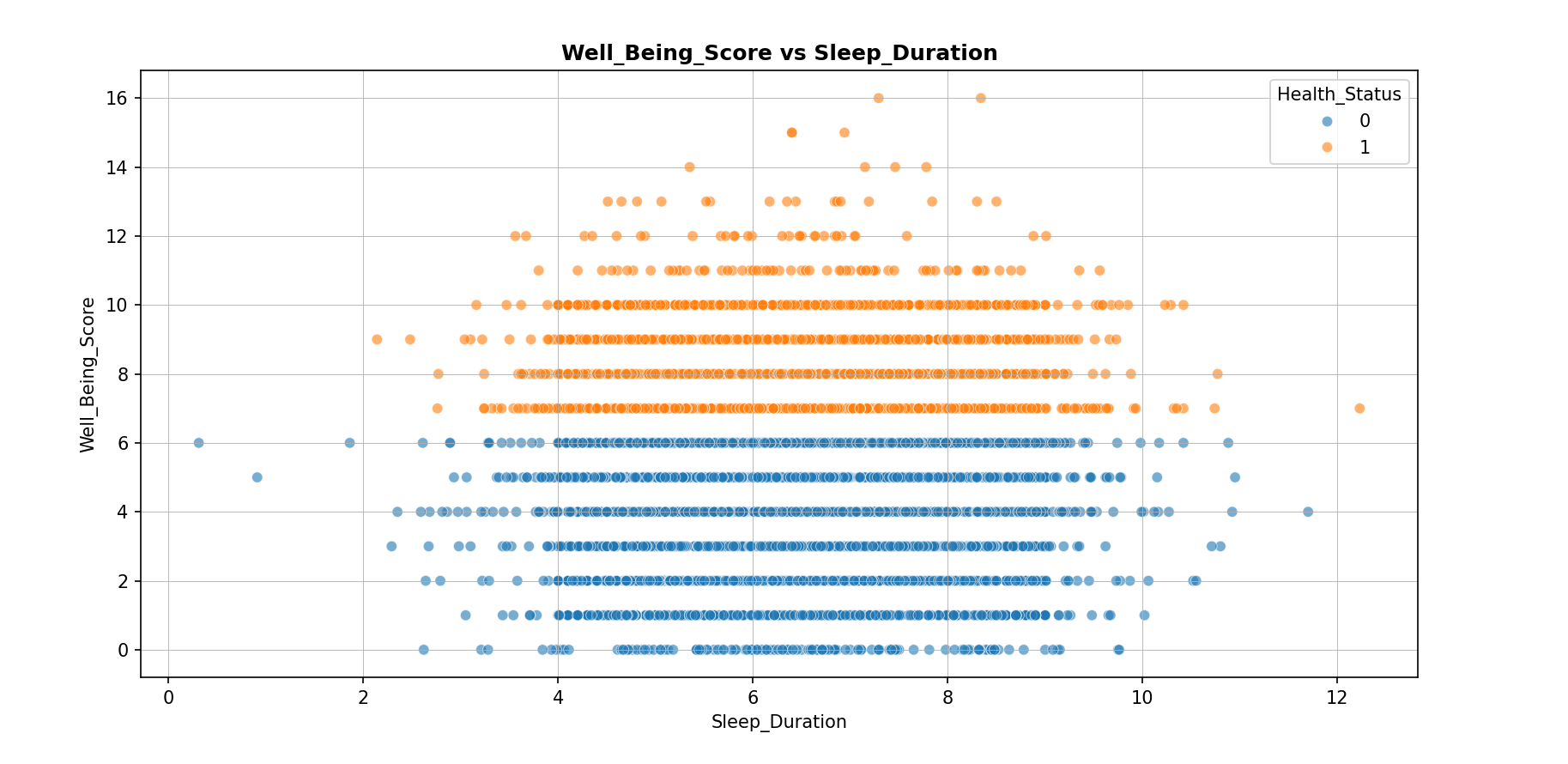


Hình 3.19: Biểu đồ scatter của Stress\_Level vs Daily\_Screen\_Time

* Nhận xét:
  + Các điểm dữ liệu cho thấy xu hướng đồng biến nhẹ giữa thời gian sử dụng màn hình và mức độ căng thẳng: khi thời gian sử dụng màn hình tăng, mức căng thẳng có xu hướng cao hơn.
  + Nhóm High\_Stress = True (màu cam) xuất hiện nhiều hơn ở vùng có thời gian sử dụng màn hình trung bình đến cao, cho thấy những người dùng thiết bị điện tử nhiều có khả năng rơi vào trạng thái căng thẳng cao hơn.
  + Nhóm High\_Stress = False (màu xanh) tập trung nhiều hơn ở mức căng thẳng thấp và trung bình, dù vẫn có sự chồng lấn đáng kể với nhóm căng thẳng cao.
  + Dữ liệu phân tán rộng, không tạo thành một đường tuyến tính rõ ràng, cho thấy căng thẳng không chỉ phụ thuộc vào thời gian sử dụng màn hình mà còn chịu ảnh hưởng của các yếu tố khác như giấc ngủ, công việc và sức khỏe cá nhân.
* Kết luận: Biểu đồ cho thấy thời gian sử dụng màn hình có mối liên hệ nhất định với mức độ căng thẳng, đặc biệt là làm tăng xác suất rơi vào trạng thái căng thẳng cao khi thời gian sử dụng kéo dài. Tuy nhiên, mối quan hệ này không hoàn toàn tuyến tính, gợi ý rằng cần xem xét thêm các biến trung gian khác khi đánh giá tác động của hành vi sử dụng thiết bị điện tử đến sức khỏe tinh thần.

*Biểu đồ 5: Well\_Being\_Score vs Sleep\_Duration*

* Dạng biểu đồ: Scatter Plot (Biểu đồ phân tán)
* Loại phân tích: Song biến
* Các biến phân tích
  + Sleep\_Duration (thời lượng ngủ)
  + Well\_Being\_Score (điểm đánh giá mức độ hạnh phúc / sức khỏe tinh thần)
  + Biến phân nhóm hiển thị màu: Health\_Status (0: sức khỏe kém, 1: sức khỏe tốt)
* Kiểu dữ liệu
  + Biến độc lập: Sleep\_Duration (số liên tục)
  + Biến phụ thuộc: Well\_Being\_Score (số rời rạc)
  + Biến phân loại bổ trợ: Health\_Status (nhị phân)
* Mục đích: Biểu đồ được sử dụng để phân tích mối quan hệ giữa thời lượng ngủ và mức độ hạnh phúc, đồng thời so sánh sự khác biệt giữa nhóm có sức khỏe tốt và sức khỏe kém. Qua đó đánh giá vai trò của giấc ngủ đối với sức khỏe tinh thần tổng thể.



Hình 3.20: Biểu đồ Scatter của Well\_Being\_Score vs Sleep\_Duration

* Nhận xét:
  + Biểu đồ cho thấy xu hướng đồng biến khá rõ: khi thời lượng ngủ tăng, điểm Well\_Being\_Score có xu hướng cao hơn.
  + Nhóm Health\_Status = 1 (sức khỏe tốt) tập trung nhiều ở vùng điểm hạnh phúc cao, đặc biệt khi thời lượng ngủ nằm trong khoảng trung bình đến cao.
  + Nhóm Health\_Status = 0 (sức khỏe kém) chủ yếu xuất hiện ở vùng điểm hạnh phúc thấp hơn, ngay cả khi thời lượng ngủ tương đương, cho thấy chất lượng sức khỏe tổng thể cũng ảnh hưởng đến cảm nhận hạnh phúc.
  + Có sự chồng lấn giữa hai nhóm, tuy nhiên sự phân tách tương đối rõ ràng về mức điểm cho thấy giấc ngủ là yếu tố quan trọng nhưng không phải duy nhất.
* Kết luận: Biểu đồ chỉ ra mối quan hệ tích cực giữa thời lượng ngủ và mức độ hạnh phúc, đặc biệt rõ rệt ở nhóm có sức khỏe tốt. Những người ngủ đủ và ổn định thường có điểm hạnh phúc cao hơn, trong khi thiếu ngủ có liên quan đến mức độ hạnh phúc thấp. Kết quả này củng cố nhận định rằng giấc ngủ đóng vai trò then chốt trong việc duy trì sức khỏe tinh thần và chất lượng cuộc sống.

#### Phân tích với Power BI

1. Làm sạch dữ liệu với Power Query

Tiền xử lý dữ liệu với công cụ Power Query Editor của Microsoft Excel:

Đầu tiên ta nhập dữ liệu cần xử lý : Data >> Get Data.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.21: Nhập dữ liệu cần xử lý

Sau khi load dữ liệu xong, cửa sổ Power Query sẽ khởi chạy và ta sẽ tiến hành làm sạch dữ liệu:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.22: Cửa sổ dữ liệu chưa xử lý với Power Query

Xóa, thay thế dữ liệu bị khuyết hoặc lỗi:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.23: Xóa, thay thế dữ liệu trống hoặc lỗi bằng giá trị trung bình

Dữ liệu sau làm sạch:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.24: Dữ liệu sau làm sạch bằng Power Query

1. Xử lý dữ liệu với Power Query

Để phục vụ cho việc trực quan hóa và phân tích mối quan hệ giữa hành vi sử dụng thiết bị và các chỉ số sức khỏe tinh thần, dữ liệu ban đầu đã được xử lý và biến đổi qua các bước chính sau:

* Phân loại mức độ sử dụng ứng dụng:

Ba biến thời gian sử dụng ứng dụng gồm *Entertainment Time*, *Work Time* và *Social Time* được chuyển đổi từ dạng số liên tục sang dạng phân loại (*Low – Medium – High*). Việc phân loại này giúp so sánh hành vi sử dụng giữa các nhóm người dùng một cách trực quan và dễ diễn giải hơn.

* Chuẩn hóa cấu trúc dữ liệu bằng thao tác Unpivot:

Ba cột phân loại (*Entertainment Level*, *Work Level*, *Social Level*) được chuyển đổi từ dạng bảng rộng sang dạng bảng dài thông qua thao tác Unpivot Columns. Kết quả tạo ra hai cột chính:

* *Type*: đại diện cho loại hoạt động (Entertainment, Work, Social)
* *Level*: đại diện cho mức độ sử dụng (Low, Medium, High)

Cách tổ chức này cho phép hiển thị các mức độ sử dụng của từng loại hoạt động trên cùng một biểu đồ cột.

* Kết hợp với các biến số định lượng:

Dữ liệu sau khi chuyển đổi được liên kết với các biến định lượng như Stress Level, Sleep Duration và Phone Unlocks nhằm phân tích sự thay đổi của các chỉ số này theo từng mức độ sử dụng và loại hoạt động.

1. Thiết kế biểu đồ trên Power BI

Để phân tích ảnh hưởng của thời gian sử dụng màn hình đến sức khỏe, các biểu đồ trực quan hóa được xây dựng trên Power BI với các biến khác nhau. Cụ thể, biểu đồ hình tròn được sử dụng để thể hiện tổng thời gian sử dụng màn hình (Daily\_Screen\_Time) theo các mức Stress Level (thấp, trung bình, cao).

Biểu đồ đường được sử dụng để mô tả mối quan hệ giữa Mood Rating và tần suất sử dụng thiết bị (Type\_use). Biểu đồ đường khác được xây dựng nhằm so sánh thời lượng ngủ trung bình (Sleep Duration) theo từng mức độ sử dụng (Level) và từng loại hình sử dụng (giải trí, công việc, mạng xã hội).

Biểu đồ cột được dùng để biểu diễn tổng mức độ căng thẳng (Stress Level) theo mức độ sử dụng thiết bị và loại hình sử dụng, giúp so sánh sự khác biệt giữa các nhóm người dùng.

A close-up of a graph

AI-generated content may be incorrect.

Hình 3.25: Kết quả thu được sau khi vẽ Power BI

Kết quả cho thấy mức độ căng thẳng có xu hướng cao hơn ở nhóm sử dụng thiết bị nhiều, trong khi thời lượng ngủ giảm dần khi mức độ sử dụng tăng. Đồng thời, tần suất mở khóa điện thoại tăng rõ rệt ở nhóm sử dụng trung bình và cao, phản ánh mức độ phụ thuộc vào thiết bị lớn hơn.

### Phân tích hồi quy logistic

#### Hồi quy logistic

1. Định nghĩa

Hồi quy logistic là một thuật toán học máy được sử dụng chủ yếu trong các bài toán phân loại, đặc biệt là phân loại nhị phân (binary classification), nơi mà đầu ra chỉ có hai giá trị có thể có (ví dụ: 0 hoặc 1). Hồi quy logistic là một phương pháp mở rộng của hồi quy tuyến tính, nhưng thay vì dự đoán một giá trị thực tế liên tục, nó dự đoán xác suất của một sự kiện thuộc vào một trong hai lớp.

Hồi quy logistic sử dụng một hàm sigmoid (còn gọi là hàm logistic) để chuyển đổi kết quả của hồi quy tuyến tính thành xác suất từ 0 đến 1.

Mô hình hồi quy logistic có dạng:

Trong đó:

* P(y=1/x) là xác suất của biến phụ thuộc y thuộc lớp 1, với các giá trị của các biến độc lập x1, x2, …, xn .
* w0, w1,..., wn là các hệ số hồi quy (weights hoặc coefficients), được học trong quá trình huấn luyện.
* x1, x2,..., xn là các biến độc lập (features).
* e là hằng số cơ bản của logarit tự nhiên.

1. Hàm logarit (sigmoid)

Hàm sigmoid là hàm kích hoạt trong hồi quy logistic, có dạng:

Trong đó: z là đầu ra của mô hình hồi quy tuyến tính.

Hàm sigmoid có đặc điểm:

z → ∞, (z) → 1, nghĩa là xác suất của sự kiện thuộc lớp 1 là gần 1.

z → -∞, (z) → 0, nghĩa là xác suất của sự kiện thuộc lớp 1 là gần 1.

z = 0, (z) = 0.5, xác suất là 50%.

Hàm sigmoid giúp đảm bảo rằng kết quả của mô hình luôn nằm trong khoảng từ 0 đến 1, phù hợp cho việc dự đoán xác suất.

1. Huấn luyện mô hình hồi quy logistic:

Quá trình huấn luyện mô hình hồi quy logistic nhằm tìm ra các hệ số hồi quy w0 , w1 ,...,wn sao cho mô hình dự đoán chính xác nhất xác suất của biến phụ thuộc y. Quá trình huấn luyện sử dụng hàm mất mát (loss function) và tối ưu hóa nó để tìm ra các hệ số tốt nhất.

Hàm mất mát trong hồi quy logistic thường được tính bằng Cross Entropy Loss (hoặc Log Loss), được sử dụng để đo lường sự khác biệt giữa xác suất dự đoán và giá trị thực tế.

Hàm mất mát cho một điểm dữ liệu có dạng:

Trong đó:

* P là xác suất dự đoán của mô hình (tính từ hàm sigmoid).
* Y là giá trị thực tế của biến phụ thuộc (0 hoặc 1).

Hồi quy Logistic được huấn luyện bằng cách tối ưu hóa hàm mất mát Cross-Entropy Loss thông qua thuật toán SAGA. Mục tiêu của SAGA là điều chỉnh các hệ số sao cho giá trị hàm mất mát được giảm thiểu, đồng thời đảm bảo sự ổn định và hội tụ nhanh hơn so với Stochastic Gradient Descent thông thường. Thuật toán SAGA tính toán gradient của một mẫu dữ liệu kết hợp với gradient trung bình của tất cả các mẫu trước đó để giảm phương sai trong quá trình cập nhật các hệ số.

Công thức cập nhật trong SAGA:

Trong đó:

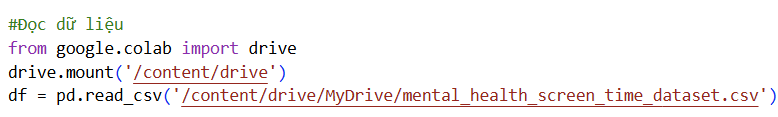
* là hệ số hồi quy cần cập nhật, là tốc độ học (learning rate).
* là gradient của hàm mất mát theo hệ số với mẫu .
* là giá trị gradient của mẫu tại bước cập nhật trước đó.
* là số lượng mẫu trong tập huấn luyện.

Nhờ cơ chế này, SAGA vừa tối ưu hóa hàm mất mát Log Loss vừa hỗ trợ regularization L1/L2 trong Elastic Net, giúp mô hình hội tụ nhanh và giảm nguy cơ overfitting.

Mô hình hồi quy logistic có thể được đánh giá thông qua các chỉ số như Accuracy, Precision, Recall, và F1-Score

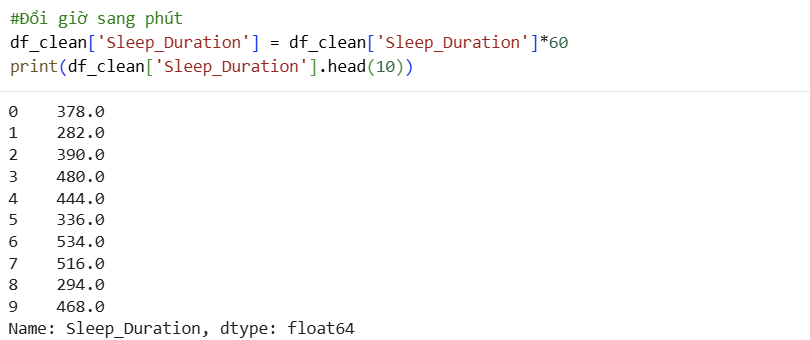
#### Huấn luyện mô hình

1. Chuẩn bị dữ liệu



Hình 3.26: Mã nguồn đọc dữ liệu

Do các cột dữ liệu như: Daily\_Screen\_Time, App\_Social\_Media\_Time, App\_Work\_Time, App\_Entertainment\_Time,... đang có đơn vị là phút nên đổi cột Sleep\_Duration đang có đơn vị là giờ sang phút.



Hình 3.27: Mã nguồn đổi giờ sang phút

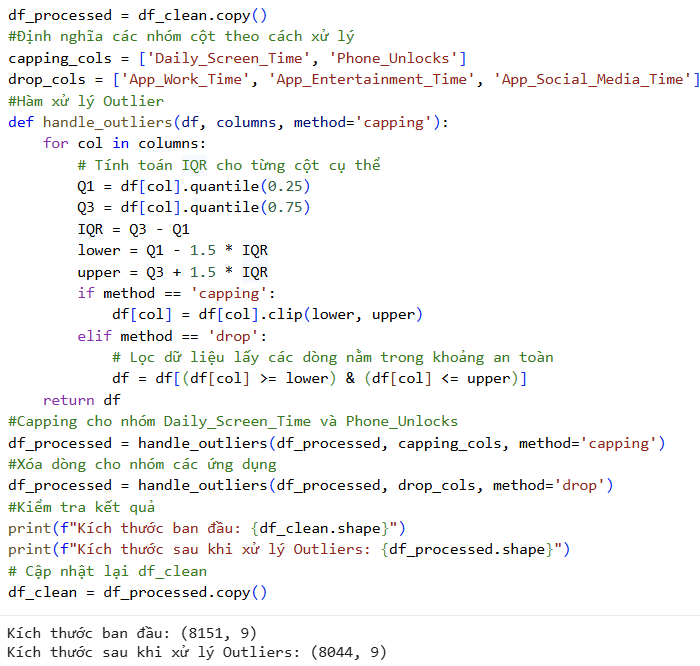
Theo mô tả, chúng ta có thể thấy ở các cột Well\_Being\_Score, Stress\_Level, Mood\_Rating có Min đều là 0 và Max lần lượt là 16, 20, 16 mà các cột trên đo trên thang 10 nên ta cần xóa các giá trị ngoài khoảng từ 0 đến 10.



Hình 3.28: Mã nguồn xóa các giá trị không hợp lệ

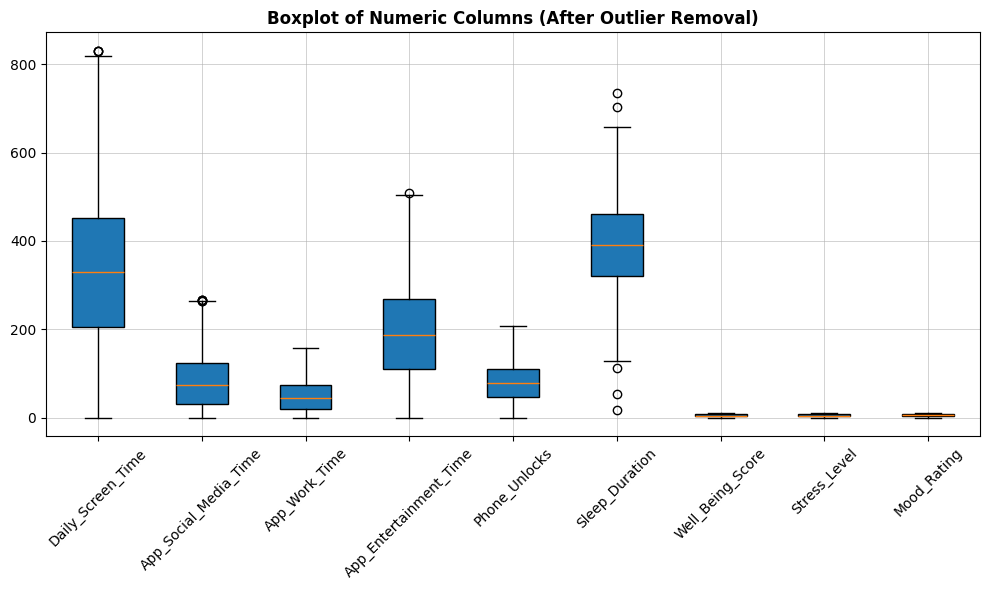
Dựa trên kết quả từ biểu đồ Box plot, tập dữ liệu ghi nhận sự hiện diện của các giá trị ngoại lai ở nhiều biến định lượng. Nhằm tối ưu hóa chất lượng dữ liệu nhưng vẫn đảm bảo duy trì quy mô mẫu (sample size) đủ lớn cho các phân tích tin cậy, chúng tôi áp dụng chiến lược tiền xử lý chọn lọc cho từng nhóm biến:

* Nhóm hành vi sử dụng thiết bị cụ thể: Với các biến ứng dụng (App\_Entertainment\_Time, App\_Work\_Time, App\_Social\_Media\_Time), do số lượng điểm dị biệt không quá lớn, chúng tôi tiến hành loại bỏ (removal) để triệt tiêu nhiễu hệ thống. Ngược lại, với biến Phone\_Unlocks, để tránh việc mất quá nhiều dữ liệu quan trọng, phương pháp Capping (Winsorization) được ưu tiên sử dụng; các giá trị cực đoan sẽ được kéo về ngưỡng $1.5 \times IQR$ thay vì xóa bỏ.
* Biến Daily\_Screen\_Time: Đây là biến số có sự biến động cao. Chúng tôi sẽ thực hiện thử nghiệm song song cả hai phương pháp Capping và Loại bỏ để đánh giá phương án nào vừa giữ được đặc trưng của dữ liệu, vừa mang lại hiệu suất mô hình tốt hơn.
* Nhóm chỉ số sức khỏe và tâm lý (Sleep\_Duration, Stress\_Level, Mood\_Rating, Well\_Being\_Score): Toàn bộ dữ liệu được giữ nguyên bản. Quyết định này dựa trên đặc thù tâm lý học hành vi: các giá trị nằm ngoài khoảng IQR không hẳn là sai số, mà thường phản ánh các trạng thái thực tế quan trọng như mất ngủ cấp tính hoặc căng thẳng cực độ—những thông tin giá trị không nên bị lược bỏ.



Hình 3.29: Mã nguồn xử lý ngoại lai

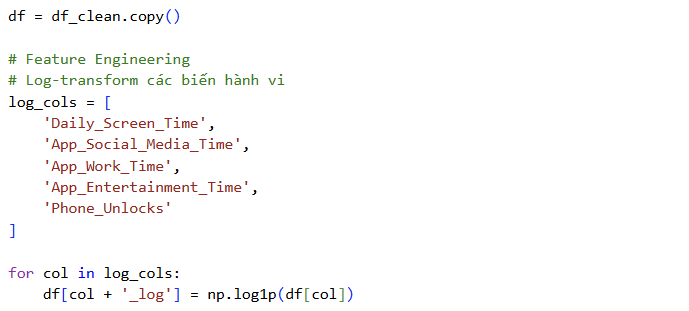
Sau khi xử lý, tập dữ liệu đã loại bỏ 107 bản ghi chứa giá trị ngoại lai gây nhiễu. Biểu đồ box plot được tái lập bên dưới cho thấy các biến số đã đạt độ tập trung tốt hơn và phân phối ổn định hơn, tạo điều kiện thuận lợi cho các bước phân tích mô hình tiếp theo.



Hình 3.30: Biểu đồ boxplot sau khi xử lý ngoại lai

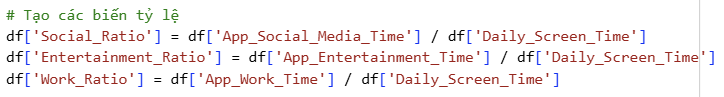
Sau khi xử lý ngoại lai, dữ liệu đã được làm sạch và giảm thiểu ảnh hưởng của các giá trị bất thường. Tuy nhiên, để mô hình có thể khai thác hiệu quả hơn các thông tin tiềm ẩn trong dữ liệu, bước Feature Engineering được thực hiện nhằm biến đổi và tạo mới các đặc trưng đầu vào có ý nghĩa hơn, phục vụ cho quá trình huấn luyện mô hình hồi quy logistic.

Đầu tiên, Log-transform các biến hành vi. Việc này giúp ổn định ảnh hưởng của các giá trị lớn, đảm bảo tính đồng nhất giữa các biến, và làm mối quan hệ giữa hành vi người dùng và khả năng căng thẳng gần tuyến tính hơn:



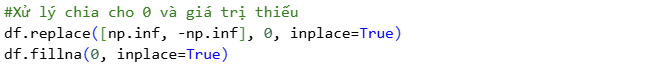
Hình 3.31: Mã nguồn log-transform

Do các biến thô có hệ số tương quan tuyến tính rất thấp, ta khởi tạo các biến tỷ lệ (Ratio features) nhằm tập trung vào cấu trúc hành vi thay vì quy mô thời gian tuyệt đối. Phương pháp này giúp bình thường hóa dữ liệu giữa các nhóm người dùng khác nhau, đồng thời làm nổi bật các tín hiệu gây căng thẳng tiềm ẩn — vốn thường bị che lấp khi chỉ phân tích các con số thời gian đơn lẻ.



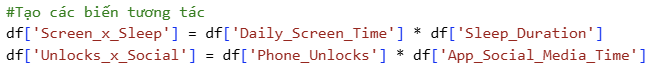
Hình 3.32: Mã nguồn tạo biến tỷ lệ

Sau khi thực hiện chia tỷ lệ, nhằm đảm bảo tính ổn định về mặt toán học cho mô hình, các giá trị không xác định (NaN) hoặc vô cực (Infinity) phát sinh do mẫu số bằng 0 được quy đổi về 0. Bước xử lý dưới đây giúp chuẩn hóa dữ liệu, loại bỏ các lỗi tính toán tiềm ẩn và đảm bảo thuật toán Logistic Regression vận hành chính xác trên toàn bộ tập mẫu.



Hình 3.33: Mã nguồn xử lý vô cực và giá trị thiếu

Song song đó, chúng tôi thiết lập các biến tương tác để ghi lại tác động cộng hưởng giữa các yếu tố hành vi mà các biến đơn lẻ chưa thể hiện rõ. Việc kết hợp các cặp đặc trưng như thời gian sử dụng máy với thời lượng ngủ (Screen\_x\_Sleep) hay tần suất mở máy với thời gian mạng xã hội (Unlocks\_x\_Social) giúp mô hình nắm bắt được các kịch bản thực tế phức tạp — nơi tác động của một yếu tố lên mức độ căng thẳng có thể bị khuếch đại hoặc giảm thiểu tùy thuộc vào sự hiện diện của yếu tố kia.



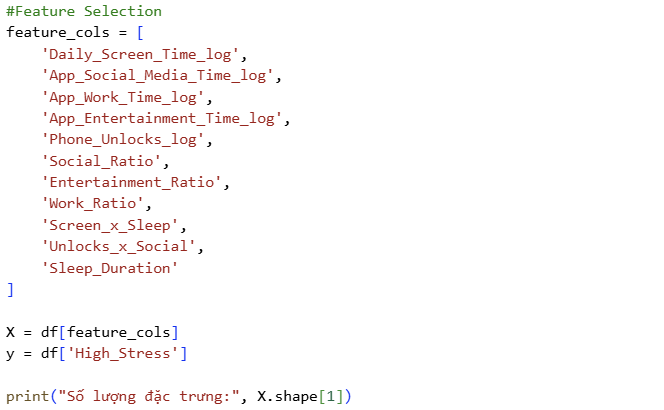
Hình 3.34: Mã nguồn tạo các biến tương tác

Tiếp theo, để chuyển đổi bài toán từ hồi quy sang phân loại nhị phân, chúng tôi khởi tạo biến mục tiêu High\_Stress. Bằng cách thiết lập ngưỡng cắt (threshold) >= 6, tập dữ liệu được phân tách rõ rệt thành hai nhóm: Căng thẳng bình thường (0) và căng thẳng cao (1).



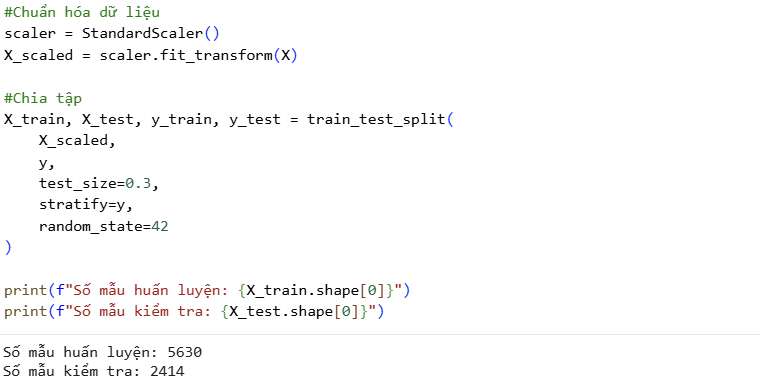
Hình 3.35: Mã nguồn tạo biến mục tiêu

Sau khi đã xác định biến mục tiêu, chúng tôi tiến hành tổng hợp danh sách các đặc trưng đầu vào (X) tối ưu để đưa vào huấn luyện mô hình. Tập dữ liệu cuối cùng bao gồm 11 biến được chọn lọc kỹ lưỡng từ các nhóm: biến hành vi đã qua xử lý Logarit, các biến tỷ lệ phản ánh thói quen sử dụng, các biến tương tác bắt giữ tác động cộng hưởng và biến thời lượng ngủ.



Hình 3.36: Mã nguồn chọn các đặc trưng

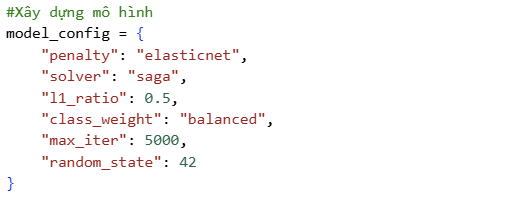
Sau khi tập đặc trưng đã được xác định, chúng tôi tiến hành hai bước kỹ thuật cuối cùng để đảm bảo tính ổn định và khách quan cho mô hình.



Hình 3.37: Mã nguồn chuẩn hóa và chia tập dữ liệu

Chúng tôi áp dụng StandardScaler để đưa các biến về cùng thang đo, tránh việc các biến có giá trị lớn gây nhiễu trọng số của mô hình. Sau đó, dữ liệu được chia theo tỷ lệ 70/30 bằng phương pháp Stratified Sampling nhằm duy trì sự cân bằng của biến mục tiêu giữa các tập huấn luyện và kiểm tra, đảm bảo tính khách quan tối đa cho kết quả dự báo.

1. Xây dựng mô hình



Hình 3.38: Mã nguồn xây dựng mô hình

Trong nghiên cứu này, mô hình hồi quy Logistic được lựa chọn để giải quyết bài toán phân loại nhị phân mức độ căng thẳng của người dùng. Nhằm nâng cao khả năng tổng quát hóa và giảm thiểu hiện tượng overfitting, mô hình được kết hợp với regularization Elastic Net, cho phép kết hợp ưu điểm của cả hai phương pháp L1 (Lasso) và L2 (Ridge).

Cụ thể, Elastic Net giúp mô hình vừa có khả năng lựa chọn đặc trưng thông qua việc đưa một số hệ số về 0, vừa kiểm soát độ lớn của các hệ số nhằm hạn chế ảnh hưởng của đa cộng tuyến giữa các biến độc lập. Tỷ lệ kết hợp giữa L1 và L2 được thiết lập thông qua tham số l1\_ratio với giá trị bằng 0.5, nhằm đạt được sự cân bằng giữa hai cơ chế regularization.

Thuật toán tối ưu SAGA được sử dụng để huấn luyện mô hình đo khả năng hỗ trợ Elastic Net và hiệu quả cao khi làm việc với tập dữ liệu có kích thước lớn. Bên cạnh đó, tham số class\_weight được thiết lập ở chế độ balanced nhằm xử lý tình trạng mất cân bằng giữa các lớp trong dữ liệu, đảm bảo mô hình không thiên lệch về lớp chiếm đa số.

Số vòng lặp tối đa (max\_iter) được đặt ở mức 5000 để đảm bảo thuật toán hội tụ đầy đủ trong quá trình huấn luyện. Cuối cùng, tham số random\_state được cố định nhằm đảm bảo tính tái lập của kết quả thực nghiệm, phục vụ cho việc so sánh và đánh giá mô hình một cách nhất quán.

1. Huấn luyện mô hình

Sau khi hoàn tất bước chuẩn bị dữ liệu và xây dựng mô hình, quá trình huấn luyện được thực hiện theo hai chiến lược chính: phương pháp Hold-out và phương pháp Stratified K-Fold Cross Validation.

A screenshot of a computer code

AI-generated content may be incorrect.

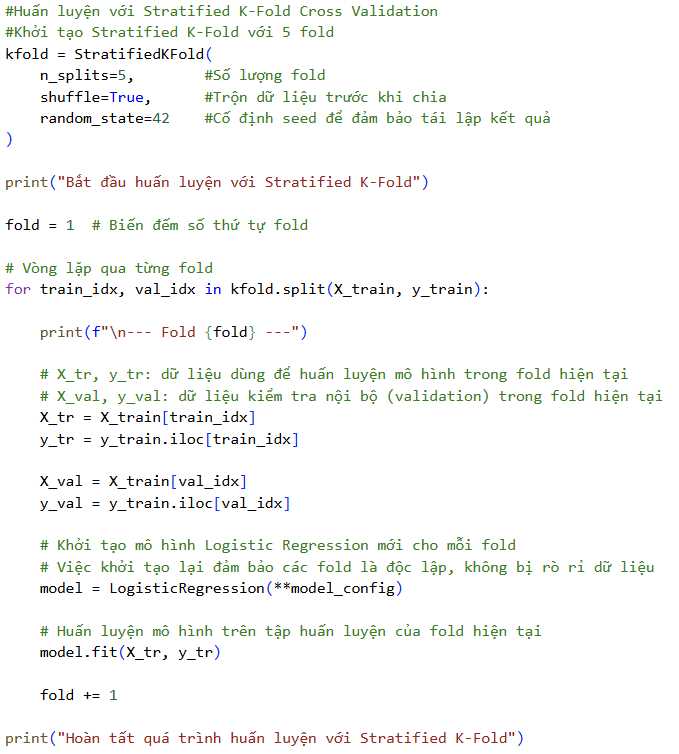
Hình 3.39: Mã nguồn huấn luyện theo Hold-out

Trong phương pháp Hold-out, mô hình hồi quy Logistic với cấu hình Elastic Net được huấn luyện trên toàn bộ tập dữ liệu huấn luyện (Xtrain, ytrain​). Ở bước này, mô hình sử dụng các đặc trưng đầu vào đã được xử lý và chuẩn hóa để học mối quan hệ giữa các biến độc lập và biến mục tiêu.

Quá trình huấn luyện được thực hiện thông qua việc tối ưu hàm mất mát Log Loss (Binary Cross-Entropy), kết hợp với cơ chế regularization Elastic Net nhằm kiểm soát độ lớn của các hệ số hồi quy và giảm thiểu hiện tượng overfitting. Kết quả của bước huấn luyện là việc xác định tập các tham số mô hình phù hợp, phản ánh mức độ ảnh hưởng của từng đặc trưng đến xác suất xảy ra trạng thái căng thẳng cao.

Mặc dù phương pháp Hold-out cho phép huấn luyện mô hình một cách đơn giản và hiệu quả trên tập dữ liệu huấn luyện ban đầu, tuy nhiên kết quả thu được vẫn có thể chịu ảnh hưởng đáng kể từ cách chia dữ liệu duy nhất. Trong trường hợp dữ liệu có sự phân bố không đồng đều giữa các lớp hoặc kích thước mẫu không quá lớn, mô hình có nguy cơ học chưa đầy đủ đặc trưng tổng thể của tập dữ liệu.

Do đó, để khắc phục những hạn chế của phương pháp Hold-out và tăng cường độ tin cậy trong quá trình huấn luyện, nghiên cứu này tiếp tục áp dụng phương pháp Stratified K-Fold Cross Validation. Phương pháp này cho phép mô hình được huấn luyện lặp lại trên nhiều tập con dữ liệu khác nhau, đồng thời đảm bảo tỷ lệ phân bố giữa các lớp được giữ nguyên trong từng fold. Nhờ vậy, quá trình huấn luyện trở nên ổn định hơn và các tham số học được phản ánh toàn diện hơn cấu trúc của dữ liệu.



Hình 3.40: Mã nguồn huấn luyện theo K-Fold

Trước hết, phương pháp Stratified K-Fold được khởi tạo với số lượng fold K = 5. Việc lựa chọn 5 fold giúp đảm bảo sự cân bằng giữa chi phí tính toán và độ tin cậy của quá trình huấn luyện. Đồng thời, tùy chọn shuffle =True được sử dụng để trộn ngẫu nhiên dữ liệu trước khi chia, giúp hạn chế ảnh hưởng của thứ tự sắp xếp dữ liệu ban đầu. Tham số random\_state = 42 được cố định nhằm đảm bảo khả năng tái lập kết quả trong các lần thực nghiệm khác nhau.

Tiếp theo, tập dữ liệu huấn luyện ban đầu (Xtrain, ytrain) được chia thành 5 tập con (fold) sao cho tỷ lệ phân bố giữa các lớp trong biến mục tiêu được giữ nguyên ở mỗi fold. Điều này đặc biệt quan trọng trong bối cảnh dữ liệu có sự mất cân bằng lớp, giúp mô hình không bị thiên lệch trong quá trình học.

Quá trình huấn luyện được thực hiện thông qua một vòng lặp, trong đó tại mỗi lần lặp, một fold được sử dụng làm tập kiểm tra nội bộ, trong khi 4 fold còn lại được kết hợp làm tập huấn luyện. Cụ thể, các chỉ số train\_idx và val\_idx xác định vị trí các mẫu dữ liệu tương ứng cho tập huấn luyện (X\_tr, y\_tr) và tập kiểm tra nội bộ (X\_val, y\_val​) trong fold hiện tại.

Ở mỗi fold, một mô hình hồi quy Logistic mới được khởi tạo với cùng cấu hình siêu tham số. Việc khởi tạo lại mô hình ở từng fold đảm bảo tính độc lập giữa các lần huấn luyện, tránh hiện tượng rò rỉ thông tin hoặc kế thừa trọng số từ các fold trước đó. Sau đó, mô hình được huấn luyện trên tập huấn luyện của fold tương ứng, cho phép mô hình học được các tham số phù hợp dựa trên các mẫu dữ liệu khác nhau.

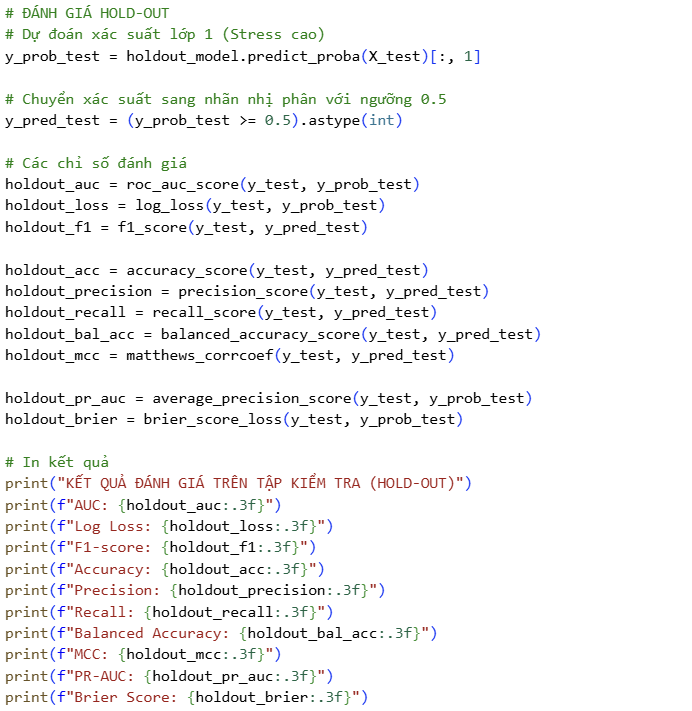
Quy trình này được lặp lại cho đến khi mỗi fold lần lượt đóng vai trò là tập kiểm tra nội bộ một lần. Kết quả của quá trình huấn luyện bằng Stratified K-Fold giúp mô hình được tiếp xúc với toàn bộ tập dữ liệu huấn luyện trong các ngữ cảnh khác nhau, từ đó nâng cao độ ổn định của các tham số học được và giảm sự phụ thuộc vào một cách chia dữ liệu cố định.

1. Đánh giá mô hình

Sau khi hoàn tất quá trình huấn luyện mô hình hồi quy Logistic với cấu hình Elastic Net, bước tiếp theo là tiến hành đánh giá hiệu năng của mô hình nhằm xác định mức độ chính xác, khả năng phân biệt và tính ổn định trong dự đoán trạng thái căng thẳng cao của người dùng. Việc đánh giá mô hình đóng vai trò quan trọng trong việc kiểm chứng mức độ phù hợp của mô hình đối với bài toán đặt ra, đồng thời làm cơ sở cho việc so sánh, phân tích và thảo luận kết quả nghiên cứu.

Trong nghiên cứu này, mô hình được đánh giá theo hai hướng chính: đánh giá trên tập kiểm tra độc lập theo phương pháp Hold-out và đánh giá thông qua Stratified K-Fold Cross Validation. Cách tiếp cận này cho phép xem xét hiệu năng mô hình không chỉ trong một lần chia dữ liệu cố định mà còn trên nhiều kịch bản phân chia khác nhau, từ đó phản ánh đầy đủ khả năng tổng quát hóa và độ ổn định của mô hình.

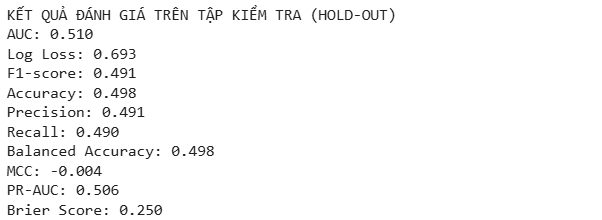
Để đảm bảo đánh giá toàn diện, nhiều chỉ số đo lường khác nhau được sử dụng, bao gồm các chỉ số dựa trên xác suất dự đoán như ROC-AUC và Log Loss, cũng như các chỉ số phân loại phổ biến như Accuracy, Precision, Recall và F1-score. Việc kết hợp đa dạng các chỉ số đánh giá giúp phân tích mô hình dưới nhiều góc độ, hạn chế sự thiên lệch khi chỉ dựa vào một thước đo duy nhất, đặc biệt trong bối cảnh dữ liệu có hiện tượng mất cân bằng lớp.



Hình 3.41: Mã nguồn đánh giá mô hình (Hold-Out)

Các chỉ số:

* ROC-AUC (Area Under the ROC Curve): đo lường khả năng của mô hình trong việc phân biệt giữa hai lớp trên toàn bộ các ngưỡng phân loại. Giá trị nằm trong khoảng [0, 1].
* Log Loss (Binary Cross-Entropy Loss): đo lường mức độ sai lệch giữa xác suất dự đoán và nhãn thực tế. Giá trị càng nhỏ, mô hình càng dự đoán tốt.
* F1-score là trung bình điều hòa giữa Precision và Recall, dùng để đánh giá sự cân bằng giữa hai chỉ số này.
* Accuracy là tỷ lệ dự đoán đúng trên tổng số mẫu.
* Precision (Độ chính xác dương tính): đo lường tỷ lệ dự đoán đúng trong các trường hợp mô hình dự đoán là lớp dương tính.
* Recall (Độ bao phủ / Sensitivity): đo lường khả năng phát hiện đầy đủ các mẫu dương tính thực sự.
* Balanced Accuracy là trung bình của Recall của lớp dương tính và lớp âm tính, giúp đánh giá công bằng khi dữ liệu mất cân bằng.
* MCC (Matthews Correlation Coefficient): đo lường mức độ tương quan giữa nhãn dự đoán và nhãn thực tế, có xét đến cả TP, TN, FP, FN. Giá trị trong khoảng [-1, 1].
* PR-AUC (Area Under Precision–Recall Curve): đánh giá sự đánh đổi giữa Precision và Recall, đặc biệt quan trọng khi dữ liệu mất cân bằng.
* Brier Score: đo lường độ chính xác của xác suất dự đoán, bằng trung bình bình phương sai lệch giữa xác suất dự đoán và nhãn thực. Giá trị càng nhỏ, mô hình càng tốt.



Hình 3.42: Kết quả đánh giá trên tập kiểm tra (Hold-Out)

Kết quả đánh giá mô hình hồi quy Logistic trên tập kiểm tra độc lập cho thấy hiệu năng của mô hình ở mức thấp, gần tương đương với mô hình dự đoán ngẫu nhiên. Cụ thể, chỉ số ROC–AUC đạt 0.510, phản ánh khả năng phân biệt giữa hai lớp của mô hình còn rất hạn chế. Giá trị này chỉ nhỉnh hơn 0.5, cho thấy mô hình chưa khai thác hiệu quả mối quan hệ giữa các đặc trưng đầu vào và biến mục tiêu.

Giá trị Log Loss bằng 0.693, xấp xỉ mức Log Loss của mô hình ngẫu nhiên trong bài toán phân loại nhị phân, cho thấy mô hình chưa ước lượng tốt xác suất xảy ra trạng thái căng thẳng cao. Các dự đoán xác suất của mô hình còn thiếu độ tin cậy và chưa được hiệu chỉnh tốt.

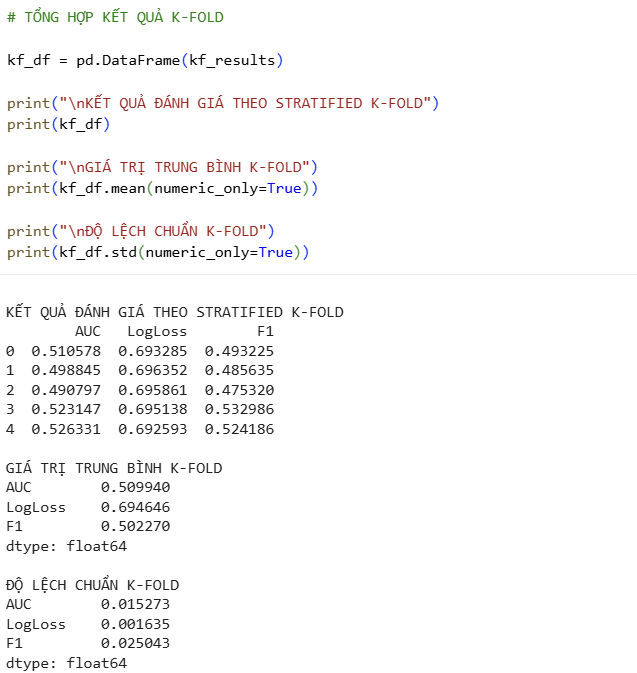
Chỉ số F1-score đạt 0.491, cho thấy sự cân bằng giữa Precision và Recall ở mức thấp. Điều này được phản ánh rõ qua Precision = 0.491 và Recall = 0.490, cho thấy mô hình vừa dự đoán sai khá nhiều trường hợp dương tính, vừa bỏ sót một tỷ lệ lớn các trường hợp căng thẳng cao thực sự. Do đó, mô hình chưa đáp ứng tốt yêu cầu phát hiện chính xác trạng thái căng thẳng cao.

Độ chính xác tổng thể Accuracy đạt 0.498, thấp hơn 0.5, cho thấy mô hình dự đoán không tốt hơn ngẫu nhiên. Đồng thời, Balanced Accuracy cũng đạt 0.498, phản ánh rằng hiệu năng phân loại trên cả hai lớp là tương đối cân bằng nhưng đều ở mức thấp.

Chỉ số Matthews Correlation Coefficient (MCC) đạt -0.004, gần bằng 0 và mang giá trị âm, cho thấy mối tương quan giữa nhãn dự đoán và nhãn thực tế gần như không tồn tại. Đây là dấu hiệu rõ ràng cho thấy mô hình hiện tại chưa có khả năng phân loại hiệu quả.

Ngoài ra, PR–AUC đạt 0.506, cho thấy khả năng cân bằng giữa Precision và Recall của mô hình trên các ngưỡng phân loại khác nhau chỉ ở mức rất thấp, đặc biệt trong bối cảnh bài toán có thể chịu ảnh hưởng bởi sự mất cân bằng lớp. Cuối cùng, Brier Score bằng 0.250 cho thấy chất lượng dự đoán xác suất của mô hình ở mức kém, gần tương đương với mô hình dự đoán ngẫu nhiên.

Tuy nhiên, kết quả đánh giá theo phương pháp Hold-out vẫn phụ thuộc vào cách chia dữ liệu cụ thể, do đó có thể chưa phản ánh đầy đủ khả năng tổng quát hóa của mô hình. Vì vậy, phương pháp Stratified K-Fold Cross Validation được áp dụng nhằm đánh giá mô hình một cách ổn định và khách quan hơn thông qua nhiều lần huấn luyện và kiểm tra trên các tập dữ liệu con khác nhau.



Hình 3.43: Mã nguồn và kết quả đánh giá theo Stratified K-FOLD

Kết quả đánh giá theo phương pháp Stratified K-Fold Cross Validation cho thấy hiệu năng của mô hình hồi quy Logistic tương đối ổn định giữa các fold. Giá trị AUC dao động trong khoảng từ 0.49 đến 0.53, với giá trị trung bình đạt 0.510, cho thấy khả năng phân biệt giữa hai lớp của mô hình chỉ nhỉnh hơn mức ngẫu nhiên (0.5). Điều này phản ánh mô hình chưa khai thác hiệu quả các đặc trưng hiện có để phân tách rõ ràng hai trạng thái mục tiêu.

Chỉ số Log Loss có giá trị trung bình khoảng 0.695 và độ lệch chuẩn rất nhỏ (0.0016), cho thấy quá trình huấn luyện trên các fold là ổn định và mô hình không bị dao động mạnh về mặt tối ưu hóa hàm mất mát. Tuy nhiên, giá trị Log Loss xấp xỉ 0.693 cũng cho thấy chất lượng dự đoán xác suất của mô hình chưa cao.

Đối với F1-score, giá trị trung bình đạt 0.502, với độ lệch chuẩn khoảng 0.025, phản ánh sự cân bằng còn hạn chế giữa Precision và Recall. Mặc dù một số fold (đặc biệt là fold 4 và fold 5) đạt F1-score cao hơn, song nhìn chung mô hình vẫn chưa thể hiện hiệu năng ổn định ở mức tốt.

Tổng hợp các chỉ số trung bình và độ lệch chuẩn cho thấy mô hình có độ ổn định tương đối tốt khi thay đổi tập dữ liệu huấn luyện, tuy nhiên hiệu năng tổng thể còn thấp, chỉ tương đương hoặc nhỉnh hơn một mô hình dự đoán ngẫu nhiên. Điều này gợi ý rằng cần xem xét cải thiện mô hình thông qua việc lựa chọn đặc trưng phù hợp hơn, thử nghiệm các tương tác phi tuyến, hoặc áp dụng các mô hình phức tạp hơn trong các bước nghiên cứu tiếp theo.

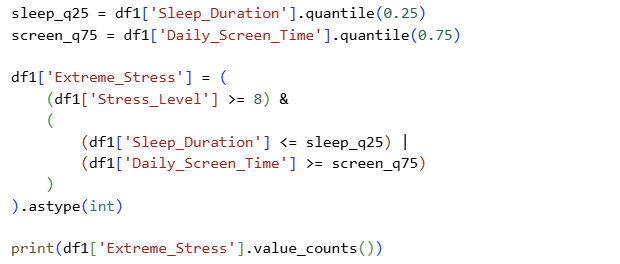
Kết quả so sánh cho thấy giá trị AUC trên tập kiểm tra theo phương pháp Hold-out đạt 0.510, gần như tương đương với giá trị AUC trung bình 0.510 thu được từ phương pháp Stratified K-Fold Cross Validation. Sự tương đồng này cho thấy mô hình không bị phụ thuộc đáng kể vào cách chia dữ liệu cụ thể, đồng thời phản ánh tính ổn định tương đối của mô hình trong quá trình huấn luyện.

Tuy nhiên, mặc dù kết quả giữa hai phương pháp là nhất quán, giá trị AUC chỉ xấp xỉ mức 0.5 cho thấy khả năng phân biệt của mô hình còn hạn chế. Do đó, cả hai phương pháp đều thống nhất chỉ ra rằng mô hình hiện tại chưa đạt hiệu năng cao và cần được cải thiện trong các bước nghiên cứu tiếp theo.

1. Phân tích bổ sung mô hình phát hiện căng thẳng cực đoan (Extreme Stress)

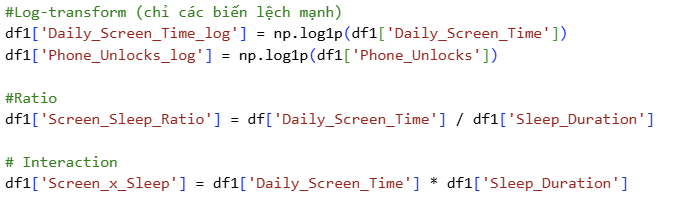
Do hiệu năng của mô hình hồi quy Logistic trong bài toán dự đoán mức độ căng thẳng tổng quát còn hạn chế, nghiên cứu tiếp tục thực hiện một phân tích bổ sung thông qua bài toán phát hiện căng thẳng cực đoan (Extreme Stress Detection). Mục tiêu của phần này là kiểm tra liệu việc tái định nghĩa biến mục tiêu và lựa chọn đặc trưng phù hợp hơn có giúp cải thiện khả năng phân biệt của mô hình hay không.

* Bước 1: Xây dựng biến mục tiêu



Hình 3.44: Mã nguồn xây dựng biến mục tiêu

* Biến Extreme\_Stress được gán giá trị 1 nếu người tham gia có mức căng thẳng rất cao (Stress\_Level ≥ 8) và kết hợp với hành vi sinh hoạt xấu: thời gian ngủ thấp (dưới 25% phân vị) hoặc thời gian sử dụng thiết bị điện tử cao (trên 75% phân vị).
* Bước 2: Feature Engineering



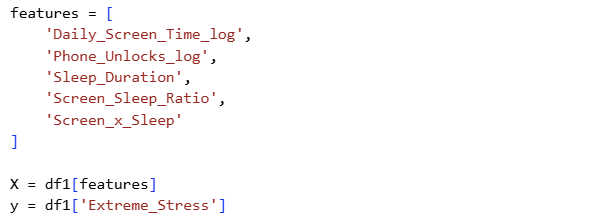
Hình 3.45: Mã nguồn feature Engineering

* Log-transform áp dụng cho các biến có phân bố lệch (Daily\_Screen\_Time, Phone\_Unlocks).
* Ratio: Screen\_Sleep\_Ratio phản ánh tỷ lệ giữa thời gian dùng màn hình và thời gian ngủ.
* Interaction: Screen\_x\_Sleep cho phép nắm bắt mối tương tác trực tiếp giữa hai yếu tố này.
* Bước 3: Làm sạch dữ liệu



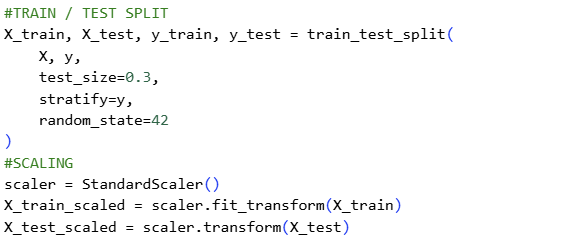
Hình 3.46: Mã nguồn làm sạch dữ liệu

* Loại bỏ NaN và giá trị vô hạn do tính toán ratio và interaction.
* Bước 4: Lựa chọn đặc trưng



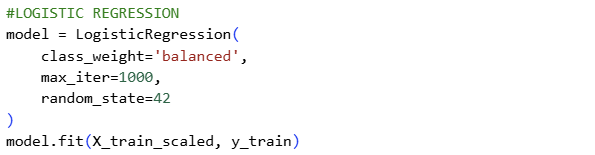
Hình 3.47: Mã nguồn lựa chọn đặc trưng

* Giữ lại 5 đặc trưng chính, trực tiếp ảnh hưởng tới trạng thái căng thẳng cực đoan.
* Bước 5: Train/Test Split và Scaling



Hình 3.48: Mã nguồn chia tệp train/ test và chuẩn hóa

* Chia dữ liệu 70/30 với stratify để giữ tỷ lệ lớp.
* Bước 6: Huấn luyện mô hình Logistic Regression



Hình 3.49: Mã nguồn huấn luyện mô hình

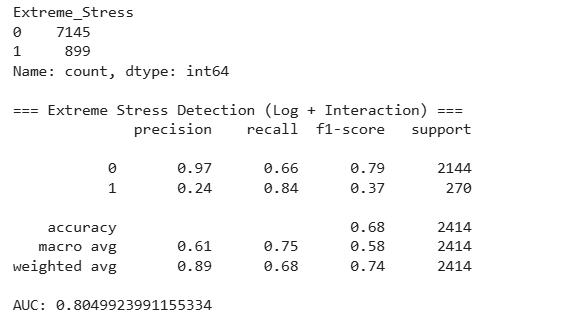
* Sử dụng class\_weight='balanced' để xử lý mất cân bằng lớp (Extreme Stress chiếm tỷ lệ thấp).
* Không sử dụng regularization phức tạp như Elastic Net để tập trung học mối quan hệ chính giữa Sleep và Screen.
* Bước 7: Đánh giá mô hình



Hình 3.50: Mã nguồn đánh giá mô hình

* Sử dụng classification report và ROC-AUC để đánh giá hiệu năng.

Kết quả:



Hình 3.51: Kết quả đánh giá

So với mô hình dự đoán mức độ căng thẳng tổng quát trước đó, có hiệu năng gần như ngẫu nhiên với AUC ≈ 0.51 và F1-score ≈ 0.49, mô hình phát hiện căng thẳng cực đoan (Extreme Stress Detection) cho thấy cải thiện đáng kể. Cụ thể, AUC của mô hình này đạt ~0.805, Accuracy = 0.68, và Recall lớp căng thẳng cao lên tới 0.84, cho thấy khả năng nhận diện các trường hợp căng thẳng nghiêm trọng tốt hơn. Mặc dù Precision của lớp căng thẳng cực đoan còn thấp (0.24), nhưng việc tập trung vào các đặc trưng log-transform, tỷ lệ và tương tác đã giúp mô hình phân biệt hiệu quả hơn giữa các trạng thái căng thẳng, vượt trội so với mô hình dự đoán tổng quát về khả năng phân loại và dự đoán các trường hợp cực đoan.

1. Trực quan kết quả

A blue squares with white text

AI-generated content may be incorrect.

Hình 3.52: Ma trận nhầm lẫn

A graph of a positive rate

AI-generated content may be incorrect.

Hình 3.53: ROC Curve

A graph with different colored bars

AI-generated content may be incorrect.

Hình 3.54: Precision, Recall, F1-score theo từng lớp

A graph of a curve

AI-generated content may be incorrect.

Hình 3.55: ROC Curve của Extreme Stress

### Kết luận chương 3

Trong chương 3, nhóm đã tiến hành đầy đủ quy trình thực nghiệm cho bài toán phân tích và dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử dựa trên mô hình hồi quy Logistic. Quy trình bao gồm các bước chính như tiền xử lý dữ liệu, phân tích mô tả và xây dựng mô hình dự đoán. Dữ liệu đầu vào được làm sạch nhằm loại bỏ các giá trị thiếu và ngoại lai, sau đó được chuẩn hóa và mã hóa phù hợp (đặc biệt với các biến phân loại) để đáp ứng yêu cầu của mô hình hồi quy Logistic.

Phân tích mô tả được thực hiện nhằm làm rõ mối quan hệ giữa hành vi sử dụng điện thoại và các chỉ số sức khỏe tinh thần, bao gồm thời gian sử dụng màn hình, thời gian sử dụng các loại ứng dụng, chất lượng giấc ngủ và mức độ căng thẳng. Kết quả phân tích cho thấy việc sử dụng điện thoại kéo dài, đặc biệt là mạng xã hội và ứng dụng giải trí, có xu hướng liên quan đến sự gia tăng nguy cơ căng thẳng và suy giảm trạng thái sức khỏe tinh thần. Những phát hiện này cung cấp cái nhìn tổng quan về đặc điểm dữ liệu và là cơ sở thực nghiệm cho việc xây dựng mô hình dự đoán ở các bước tiếp theo.

Mô hình hồi quy Logistic đã được triển khai hiệu quả để dự đoán khả năng xảy ra ảnh hưởng tiêu cực dựa trên thời gian sử dụng màn hình và các biến đầu vào liên quan. Hiệu quả của mô hình được đánh giá thông qua các chỉ số như Accuracy, Precision, Recall, F1-score và ma trận nhầm lẫn. Kết quả cho thấy mô hình có khả năng phân loại tương đối chính xác giữa các nhóm có và không có ảnh hưởng, đồng thời xác định được mức độ tác động của từng yếu tố thông qua các hệ số hồi quy. Điều này chứng minh rằng hồi quy Logistic là công cụ phù hợp và hiệu quả trong việc phân tích và dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử, góp phần hỗ trợ cho việc đưa ra các khuyến nghị về thói quen sử dụng thiết bị hợp lý.

# XÂY DỰNG SẢN PHẨM

## Công cụ và công nghệ sử dụng

### Giới thiệu về Flask

Flask là một framework web nhẹ (micro-framework) được phát triển bằng ngôn ngữ Python, cho phép xây dựng các ứng dụng web một cách nhanh chóng, linh hoạt và dễ mở rộng. Không giống như các framework web phức tạp khác, Flask cung cấp một kiến trúc tối giản, giúp người phát triển dễ dàng kiểm soát luồng xử lý của ứng dụng.

Với Flask, hệ thống có thể tiếp nhận dữ liệu đầu vào từ người dùng thông qua các biểu mẫu (form), xử lý dữ liệu phía server và trả kết quả dự đoán trực tiếp lên giao diện web. Flask hỗ trợ tốt việc tích hợp các mô hình học máy đã được huấn luyện sẵn, cho phép triển khai mô hình dự đoán vào môi trường thực tế một cách hiệu quả.

Nhờ khả năng hoạt động đa nền tảng và cộng đồng phát triển mạnh mẽ, Flask được sử dụng rộng rãi trong nghiên cứu khoa học dữ liệu, triển khai mô hình machine learning và xây dựng các hệ thống web phục vụ phân tích và dự báo.

### Giới thiệu về Bootstrap

Bootstrap là một framework CSS phổ biến giúp thiết kế giao diện web hiện đại, trực quan và thân thiện với người dùng. Bootstrap cung cấp sẵn nhiều thành phần giao diện như form nhập liệu, nút bấm, thẻ hiển thị thông tin và hệ thống lưới (grid system), giúp rút ngắn thời gian phát triển giao diện.

Trong hệ thống này, Bootstrap được sử dụng để xây dựng giao diện nhập dữ liệu và hiển thị kết quả dự đoán một cách rõ ràng, dễ quan sát. Nhờ khả năng hỗ trợ responsive, giao diện web có thể hiển thị tốt trên nhiều thiết bị khác nhau như máy tính, máy tính bảng và điện thoại di động.

### Giới thiệu về thư viện Scikit-learn

Scikit-learn là một thư viện Python mạnh mẽ, được sử dụng rộng rãi trong lĩnh vực machine learning. Thư viện này cung cấp nhiều thuật toán học máy phổ biến như hồi quy Logistic, hồi quy tuyến tính, cây quyết định và các phương pháp đánh giá mô hình.

Trong đề tài, Scikit-learn được sử dụng để xây dựng mô hình hồi quy Logistic, nhằm dự đoán xác suất mức độ stress của người dùng dựa trên các yếu tố như thời gian sử dụng điện thoại, thời gian sử dụng mạng xã hội, thời gian ngủ và số lần mở điện thoại.

### Giới thiệu về thư viện Joblib

Joblib là thư viện Python chuyên dùng cho việc lưu trữ và tải lại các đối tượng Python một cách hiệu quả, đặc biệt hữu ích trong các ứng dụng machine learning.

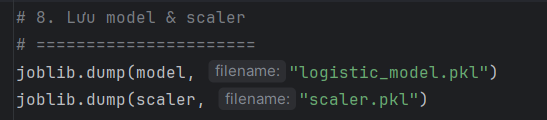
Với Joblib, hệ thống có thể:

* Lưu trữ mô hình hồi quy Logistic đã được huấn luyện.
* Lưu và tải lại bộ chuẩn hóa dữ liệu (scaler).
* Giảm thời gian huấn luyện lại mô hình khi triển khai ứng dụng.

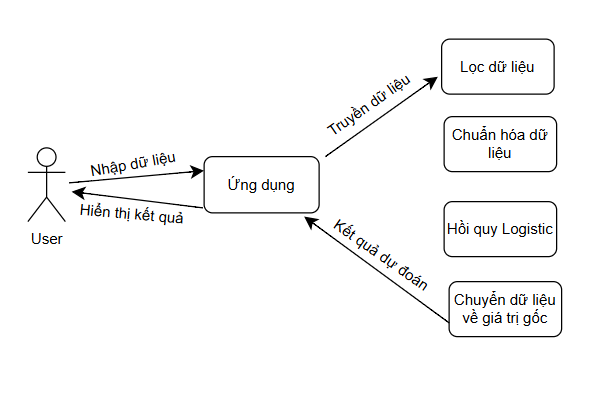
Nhờ API đơn giản và khả năng tương thích cao với các thư viện như numpy và scikit-learn, Joblib giúp tối ưu hóa quá trình triển khai mô hình học máy vào hệ thống web.

## Chuẩn bị tài nguyên xây dựng chương trình

Khi mô hình đã được huấn luyện và dữ liệu đã được chuẩn hóa ta sẽ lưu lại chúng để phục vụ cho việc dự đoán dựa trên dữ liệu người dùng đưa vào:



## Mô tả chương trình

****

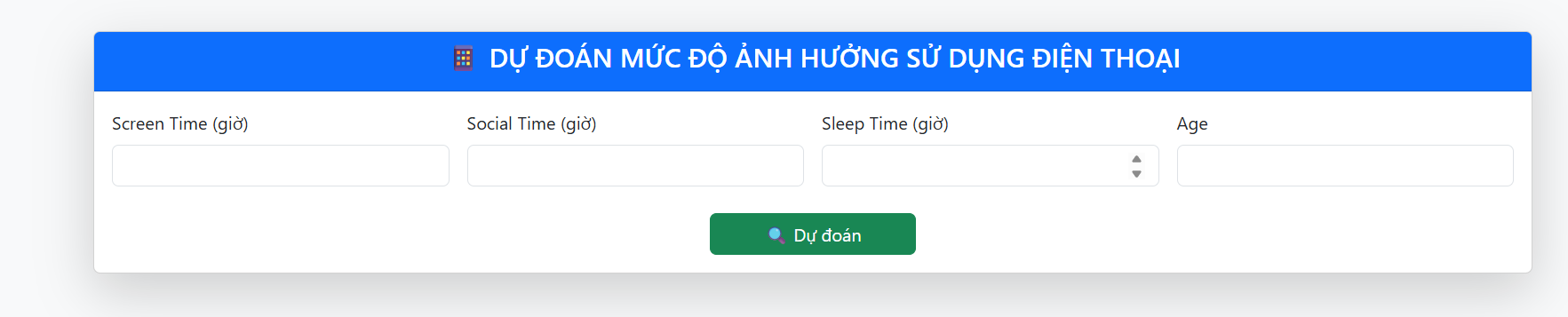
Hình 4.1: Sơ đồ usecase

Bảng 4.1: Bảng mô tả usecase

|  |  |
| --- | --- |
| Tên use case | Dự đoán mức độ ảnh hưởng dựa trên thời gian sử dụng điện thoại |
| Tóm tắt | Người dùng nhập các thông tin liên quan đến thời gian sử dụng điện thoại hằng ngày. Hệ thống kiểm tra tính hợp lệ của dữ liệu, sau đó sử dụng mô hình Logistic Regression đã được huấn luyện sẵn để dự đoán xác suất stress. Kết quả được hiển thị cho người dùng dưới dạng thông báo và biểu đồ trực quan (cột dữ liệu đầu vào và đường dự đoán màu đỏ). |
| Actor | Người dùng |
| Tiền điều kiện | Hệ thống đã được khởi động thành công.  Mô hình dự đoán (logistic\_model.pkl) và bộ chuẩn hóa dữ liệu (scaler.pkl) đã được tải vào bộ nhớ.  Người dùng truy cập được giao diện web của ứng dụng. |
| Đảm bảo tối thiểu | Người dùng được thông báo nếu dữ liệu nhập vào không hợp lệ hoặc thiếu thông tin.  Hệ thống không bị lỗi trong quá trình xử lý yêu cầu. |
| Đảm bảo thành công | Người dùng nhận được kết quả dự đoán mức độ stress dưới dạng phần trăm.  Kết quả được hiển thị trực quan bằng biểu đồ gồm:   * Các cột biểu diễn dữ liệu đầu vào. * Đường màu đỏ biểu diễn xác suất stress dự đoán. |
| Kích hoạt | Người dùng mở ứng dụng web và nhấn nút “Dự đoán” sau khi nhập đầy đủ dữ liệu. |
| Luồng sự kiện | Luồng chính   1. Người dùng mở trang web của ứng dụng. 2. Hệ thống tải mô hình Logistic Regression và scaler vào bộ nhớ. 3. Hệ thống hiển thị giao diện nhập liệu cho người dùng. 4. Người dùng nhập các dữ liệu  * Thời gian sử dụng điện thoại mỗi ngày (giờ). * Thời gian sử dụng mạng xã hội (giờ). * Thời gian ngủ (giờ). * Số lần mở khóa điện thoại mỗi ngày.  1. Người dùng nhấn nút “Dự đoán”. 2. Hệ thống kiểm tra tính hợp lệ của dữ liệu đầu vào. 3. Hệ thống chuẩn hóa dữ liệu bằng StandardScaler. 4. Hệ thống sử dụng mô hình Logistic Regression để dự đoán xác suất stress. 5. Hệ thống xử lý kết quả dự đoán và phân loại mức độ stress. 6. Hệ thống hiển thị kết quả dự đoán:  * Thông báo mức độ stress và xác suất (%). * Biểu đồ cột thể hiện dữ liệu đầu vào. * Đường dự đoán màu đỏ thể hiện xác suất stress.   Luồng thay thế  Luồng thay thế – Dữ liệu không hợp lệ (Bước 6)   * Hệ thống hiển thị thông báo lỗi cho người dùng. * Người dùng chỉnh sửa lại dữ liệu và thực hiện dự đoán lại.   Hậu điều kiện   * Kết quả dự đoán được hiển thị thành công trên giao diện web. * Người dùng có thể tiếp tục nhập dữ liệu mới để dự đoán lần tiếp theo. |

## Demo sản phẩm

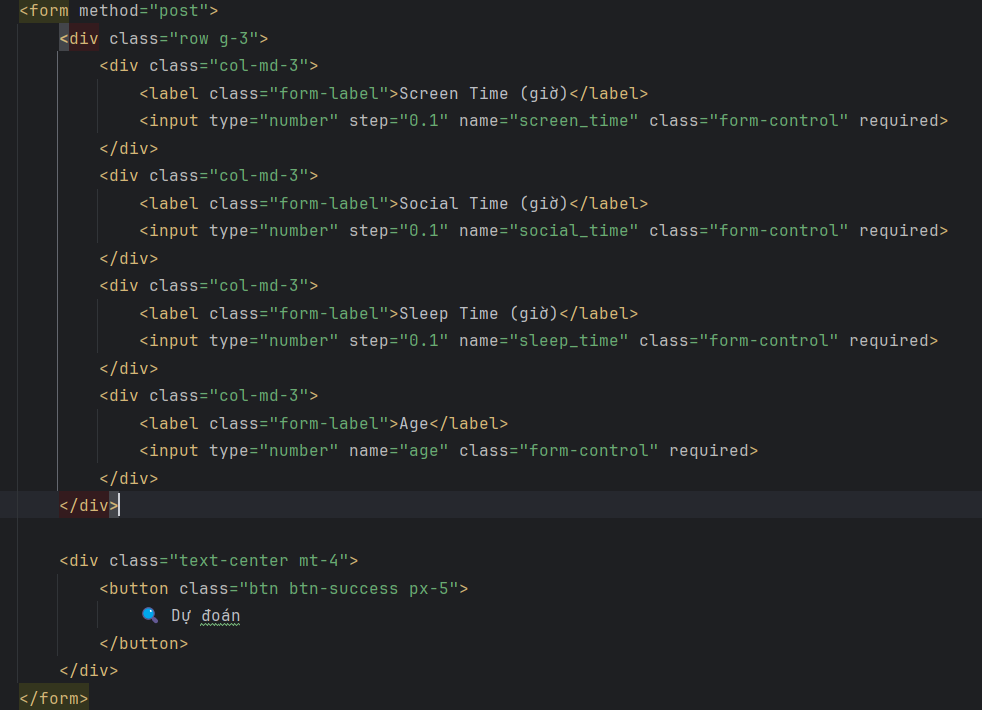
Bước 2: Giao diện nhập dữ liệu



Hình 4.2: Giao diện người dùng

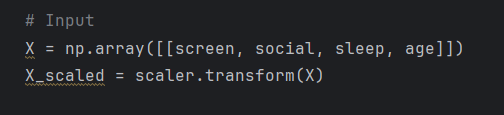
Khi truy cập vào hệ thống, người dùng sẽ thấy giao diện gồm 4 trường nhập liệu:

* Screen Time – Tổng thời gian sử dụng màn hình (giờ/ngày)
* Social Time – Thời gian sử dụng mạng xã hội (giờ/ngày)
* Sleep Time – Thời gian ngủ (giờ/ngày)
* Age – Tuổi người dùng



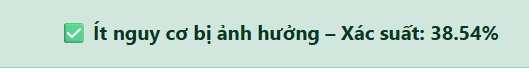
Hình 4.3: Mã nguồn giao diện nhập dữ liệu

Bước 3: Kiểm tra và xử lý dữ liệu đầu vào

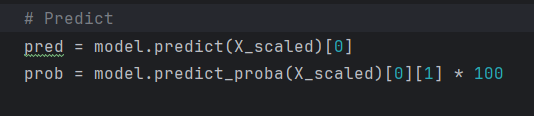


Hình 4.4: Mã nguồn kiểm tra dữ liệu vào

Bước 4: Thực hiện dự đoán



Hình 4.5: Kết quả dự đoán

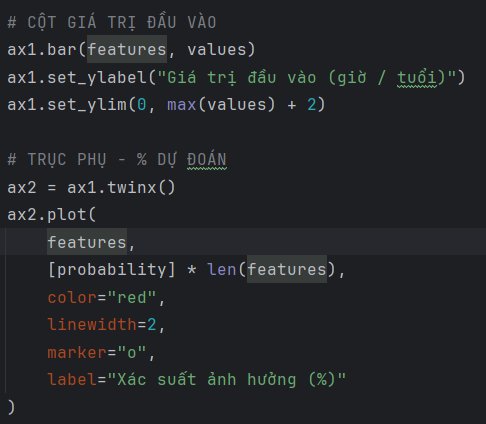


Hình 4.6: Mã nguồn dự đoán

Bước 5: Sinh biểu đồ trực quan (điểm nổi bật của demo)

Hệ thống tự động tạo biểu đồ 2 trục:

* Trục trái – Biểu đồ cột
* Thể hiện giá trị dữ liệu đầu vào:
  + Screen
  + Social
  + Sleep
  + Age
* Trục phải – Đường dự đoán
* Đường màu đỏ biểu diễn:
  + Xác suất ảnh hưởng (%)
  + Được vẽ đồng đều trên tất cả các cột để nhấn mạnh kết quả dự đoán

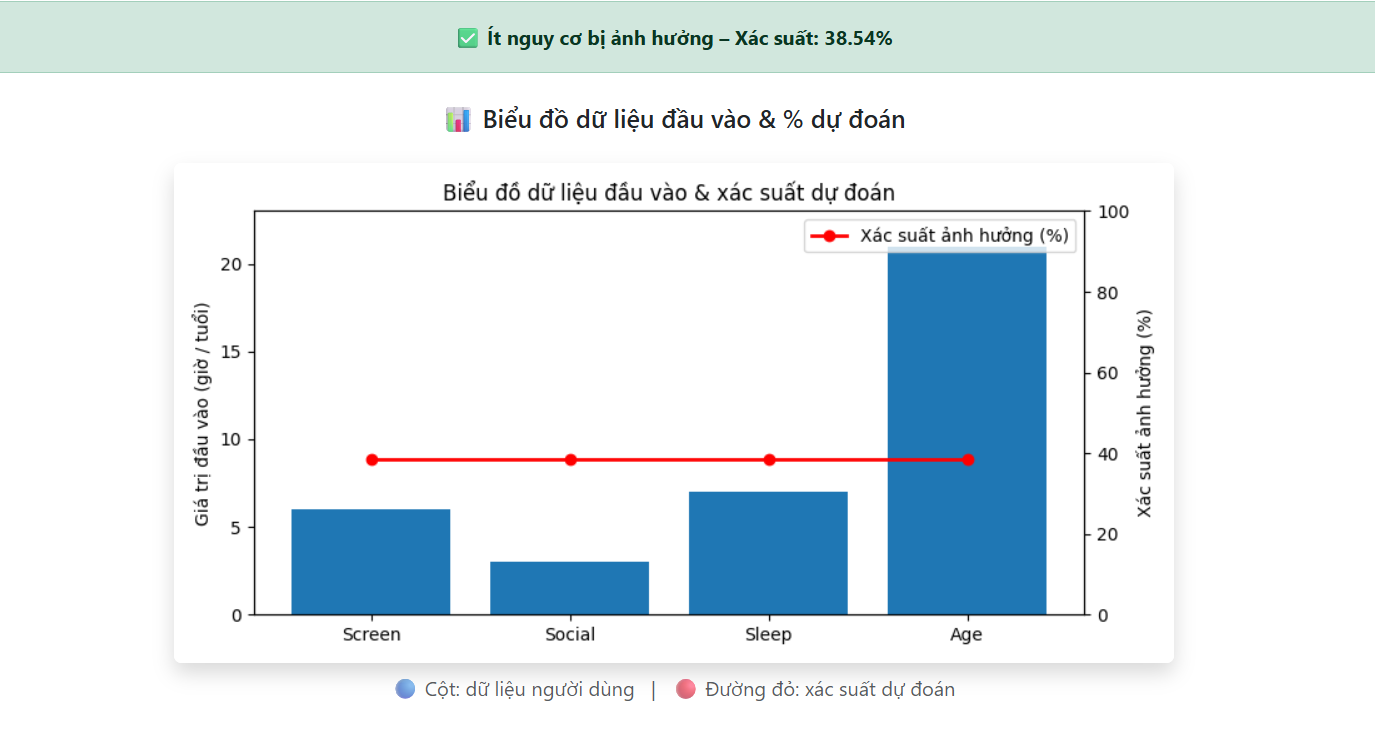


Bước 6: Hiển thị kết quả cho người dùng

Sau khi dự đoán xong, giao diện hiển thị đồng thời:

1. Thông báo kết quả
   * Màu xanh: Nguy cơ thấp
   * Màu vàng: Nguy cơ trung bình
   * Màu đỏ: Nguy cơ cao
2. Biểu đồ dự đoán
   * Cột: dữ liệu đầu vào
   * Đường đỏ: xác suất ảnh hưởng

Người dùng có thể quan sát trực tiếp mối liên hệ giữa dữ liệu nhập và mức độ ảnh hưởng.



Hình 4.7: Biểu đồ dữ liệu đầu vào và dự đoán

# KẾT LUẬN

Trong khuôn khổ đề tài *“Phân tích dự đoán ảnh hưởng của thời gian sử dụng màn hình thiết bị điện tử trên mô hình hồi quy Logistic”*, nhóm chúng em đã tập trung nghiên cứu một vấn đề mang tính thời sự và có ảnh hưởng trực tiếp đến sức khỏe tinh thần trong xã hội hiện đại, đó là tác động của việc sử dụng các thiết bị điện tử có màn hình như điện thoại thông minh, máy tính và máy tính bảng. Mục tiêu chính của đề tài là phân tích mối quan hệ giữa thời gian sử dụng màn hình và nguy cơ ảnh hưởng đến sức khỏe tinh thần, từ đó xây dựng một mô hình dự đoán nhằm hỗ trợ người dùng nhận thức rõ hơn về thói quen sử dụng thiết bị của mình.

Nhóm đã thực hiện đầy đủ các giai đoạn nghiên cứu, bao gồm xác định mục tiêu, thu thập và tiền xử lý dữ liệu, phân tích mô tả và xây dựng mô hình hồi quy. Bộ dữ liệu được sử dụng phản ánh các yếu tố liên quan trực tiếp đến hành vi sinh hoạt của người dùng như thời gian sử dụng màn hình, thời gian sử dụng mạng xã hội, thời gian ngủ và độ tuổi. Quá trình phân tích mô tả đã giúp nhóm nhận diện được các đặc điểm cơ bản của dữ liệu, xu hướng phân bố của các biến, cũng như mối quan hệ ban đầu giữa thời gian sử dụng thiết bị điện tử và nguy cơ ảnh hưởng đến sức khỏe tinh thần.

Trong giai đoạn xây dựng mô hình, nhóm đã lựa chọn mô hình hồi quy Logistic – một phương pháp phổ biến và hiệu quả trong các bài toán phân loại nhị phân. Mô hình được huấn luyện nhằm dự đoán khả năng một cá nhân có nguy cơ bị ảnh hưởng hay không dựa trên các đặc trưng đầu vào. Kết quả dự đoán cho thấy mô hình có khả năng phân loại tương đối tốt, đồng thời cung cấp xác suất xảy ra ảnh hưởng, giúp người dùng dễ dàng đánh giá mức độ rủi ro. Việc trực quan hóa kết quả bằng biểu đồ cũng góp phần nâng cao khả năng tiếp cận và hiểu biết của người dùng đối với kết quả dự đoán.

Bên cạnh những kết quả đạt được, nghiên cứu vẫn tồn tại một số hạn chế nhất định. Dữ liệu sử dụng trong đề tài chủ yếu dựa trên các yếu tố hành vi cơ bản, chưa bao gồm các yếu tố tâm lý chuyên sâu, môi trường sống, hay áp lực học tập và công việc – những yếu tố có thể tác động mạnh mẽ đến sức khỏe tinh thần. Ngoài ra, mô hình hồi quy Logistic có cấu trúc tương đối đơn giản, do đó khả năng nắm bắt các mối quan hệ phi tuyến phức tạp còn hạn chế.

Trong tương lai, nhóm dự kiến mở rộng nghiên cứu bằng cách thu thập thêm các nguồn dữ liệu đa dạng hơn, bao gồm các yếu tố tâm lý và xã hội. Đồng thời, các mô hình học máy nâng cao như cây quyết định, rừng ngẫu nhiên hoặc mạng nơ-ron sâu có thể được thử nghiệm để cải thiện độ chính xác dự đoán. Bên cạnh đó, việc triển khai mô hình dưới dạng một ứng dụng web hoặc hệ thống trực tuyến sẽ giúp đánh giá khả năng hoạt động của mô hình trong môi trường thực tế, đồng thời hỗ trợ người dùng theo dõi và điều chỉnh thói quen sử dụng thiết bị điện tử một cách khoa học hơn. Thông qua những hướng phát triển này, nhóm kỳ vọng đề tài sẽ đóng góp giá trị thiết thực cho các nghiên cứu tiếp theo và nâng cao nhận thức cộng đồng về sức khỏe tinh thần trong kỷ nguyên số.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | N. M. Cường, Slide bài giảng môn phân tích dữ liệu lớn, Đại học Công Nghiệp Hà Nội, 2025. |
| [2] | N. V. Tuấn, Phân tích dữ liệu với R, 2018. |
| [3] | R. Layton, Learning Data Mining with Python, 2016. |
| [4] | P. H. Thông, Slide bài giảng môn học máy, Đại học Công Nghiệp Hà Nội, 2025. |