BỘ GIÁO DỤC VÀ ĐÀO TẠO



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH DỮ LIỆU THÔNG MINH

**BÁO CÁO BÀI TẬP MÔN HỌC**

**NỘI DUNG: ỨNG DỤNG CÔNG CỤ AI TRONG QUY TRÌNH PHÂN TÍCH DỮ LIỆU THỰC TẾ**

**Giảng viên hướng dẫn** :Nguyễn Tiến Huy

**Lớp** :CQ2022/24

**Sinh viên thực hiện** :Đinh Viết Lợi- 22120188

Nguyễn Trần Lợi-22120190

Lê Võ Nhật Minh-22120210

Nguyễn Hữu Nghĩa-22120227

Nguyễn Lê Tấn Phát-22120262

*Hồ Chí Minh, ngày 28 tháng 4 năm 2025*

# MỤC LỤC

[MỤC LỤC 1](#_Toc196814773)

[PHẦN 1: TỔNG QUAN ĐỒ ÁN 3](#_Toc196814774)

[I. Tổng quan đồ án 3](#_Toc196814775)

[1. Giới thiệu đồ án 3](#_Toc196814776)

[2. Yêu cầu đồ án 3](#_Toc196814777)

[II. Báo cáo kết quả 3](#_Toc196814778)

[PHẦN 2: QUY TRÌNH VÀ CÔNG CỤ 4](#_Toc196814779)

[I. Vấn đề quan tâm 4](#_Toc196814780)

[II. Nguồn dữ liệu 4](#_Toc196814781)

[III. Pipeline 4](#_Toc196814782)

[1. Khám phá dữ liệu 4](#_Toc196814783)

[2. Tiền xử lý dữ liệu 5](#_Toc196814784)

[3. Trực quan hoá và phân tích 5](#_Toc196814785)

[4. Mô hình hoá dữ liệu 5](#_Toc196814786)

[IV. Công cụ AI hỗ trợ và phương thức 5](#_Toc196814787)

[1. Công cụ AI 5](#_Toc196814788)

[2. Phương thức ứng dụng 7](#_Toc196814789)

[PHẦN 3: PHÂN TÍCH DỮ LIỆU 8](#_Toc196814790)

[I. Phụ nữ và vấn đề trầm cảm ? 8](#_Toc196814791)

[II. Sinh viên chịu ảnh hưởng từ trầm cảm ra sao ? 8](#_Toc196814792)

[III. Lối sống ảnh hưởng đến sức khoẻ tinh thần như thế nào? 9](#_Toc196814793)

[IV. Áp lực tài chính, học tập ảnh hưởng tới sinh viên thế nào ? 10](#_Toc196814794)

[V. Khả năng quản lý thời gian và bệnh trầm cảm 10](#_Toc196814795)

[PHẦN 4: ĐÁNH GIÁ CÔNG CỤ AI 11](#_Toc196814796)

[I. ChatGPT 11](#_Toc196814797)

[II. Gemini 11](#_Toc196814798)

[III. Claude 11](#_Toc196814799)

[IV. Deepseek 11](#_Toc196814800)

[V. So sánh hiệu quả 11](#_Toc196814801)

[PHẦN 5: MÔ HÌNH HOÁ DỮ LIỆU 13](#_Toc196814802)

[I. Tiền xử lý dữ liệu 13](#_Toc196814803)

[II. Lựa chọn mô hình phù hợp 13](#_Toc196814804)

[III. Huấn luyện mô hình 14](#_Toc196814805)

[IV. Triển khai mô hình 14](#_Toc196814806)

[PHẦN 6: BÁO CÁO NHÓM 16](#_Toc196814807)

[I. Phương thức hoạt động 16](#_Toc196814808)

[II. Phân công công việc 16](#_Toc196814809)

[PHẦN 7: TÀI LIỆU THAM KHẢO 18](#_Toc196814810)

[I. Tài liệu nhóm 18](#_Toc196814811)

[II. Tài liệu tham khảo 18](#_Toc196814812)

# PHẦN 1: TỔNG QUAN ĐỒ ÁN

## Tổng quan đồ án

### **Giới thiệu đồ án**

* + Đồ án “Ứng dụng Công cụ AI trong Quy trình Phân tích Dữ liệu Thực tế” là đồ án phục vụ cho môn học “Phân tích dữ liệu thông minh” thuộc lớp CQ 2022/24 trong học kỳ II năm học 2024-2025.
  + Đồ án là xây dựng một giải pháp cho các bài toán dữ liệu thông thường với đầy đủ quy trình tìm hiểu, xử lý, trực quan, phân tích và xây dựng các mô hình hỗ trợ tác vụ liên quan. Tập trung hơn cả ở đồ án này là việc ứng dụng mạnh mẽ các công cụ AI để hỗ trợ các bước phân tích dữ liệu.

### **Yêu cầu đồ án**

* + Nhóm cần lựa chọn một chủ đề cụ thể để thực hiện phân tích, chủ đề được lựa chọn phải phù hợp với thực tế, có ý nghĩa khi thực hiện các phân tích, tìm hiểu và phải giải quyết được một số vấn đề.
  + Làm việc theo nhóm, tổ chức quy trình làm việc hiệu quả thông qua việc sử dụng nhiều công cụ, môi trường làm việc phù hợp.
  + Quá trình làm việc cần có sự trao đổi, đóng góp, phản biện giữa các thành viên, có các ghi nhận và phê bình đúng lúc.
  + Quy trình làm việc và các kết quả đạt được phải được xây dựng hoàn chỉnh, đảm bảo thuận lợi cho việc báo cáo kết quả.

## Báo cáo kết quả

* + Slide trình bày các nội dung về chủ đề nhóm quan tâm và tập trung vào các phân tích chuyên sâu cũng như cách thức nhóm ứng dụng các công cụ AI hỗ trợ.
  + Các file jupyter notebook là kết quả cụ thể suốt quá trình làm việc của nhóm, bao gồm các quy trình làm việc với dữ liệu và phân tích chuyên sâu.
  + Workspace ghi nhận lại quá trình làm việc nhóm, nội dung công việc và cá nhân đảm nhận.

# PHẦN 2: QUY TRÌNH VÀ CÔNG CỤ

## Vấn đề quan tâm

* + Trong bối cảnh xã hội hiện đại, tỷ lệ mắc bệnh trầm cảm đang gia tăng nhanh chóng ở nhiều độ tuổi và ngành nghề khác nhau, trở thành một vấn đề sức khỏe tinh thần đáng báo động. Báo động hơn khi biểu hiện của căn bệnh này rất khó nhận biết và ít được cộng đồng quan tâm đến.
  + Nhận thấy tầm quan trọng và mức độ ảnh hưởng sâu rộng của trầm cảm đối với đời sống cá nhân cũng như xã hội, nhóm quyết định chọn đề tài phân tích dữ liệu liên quan đến bệnh trầm cảm. Thông qua việc khai thác và phân tích dữ liệu, nhóm mong muốn hiểu rõ hơn về các yếu tố liên quan đến bệnh trầm cảm, xu hướng phát triển của nó và từ đó đề xuất những hướng tiếp cận hữu ích nhằm hỗ trợ công tác phòng ngừa và can thiệp hiệu quả.

## Nguồn dữ liệu

* + Nhằm đảm bảo việc phân tích diễn ra đúng với thực tế và có ý nghĩa về sau cần đảm bảo các yếu tố: tính xác thực, tính đa dạng và số lượng. Việc lựa chọn tập dữ liệu cần được đánh giá sơ bộ trước khi thực hiện làm việc kĩ hơn. Nguồn dữ liệu: [Student Depression Dataset](https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset)
  + ​Tập dữ liệu Student Depression Dataset do người dùng AdilShamim8 chia sẻ trên Kaggle nhằm mục đích phân tích xu hướng và các yếu tố dự đoán bệnh trầm cảm ở sinh viên. Tập dữ liệu này bao gồm thông tin về nhân khẩu học, áp lực học tập và công việc, cùng các yếu tố khác có thể liên quan đến sức khỏe tâm thần của sinh viên.
  + Tập dữ liệu này hữu ích cho việc khám phá mối liên hệ giữa các yếu tố như giới tính, áp lực học tập và tỷ lệ trầm cảm, từ đó hỗ trợ xây dựng các mô hình dự đoán và đề xuất các biện pháp can thiệp phù hợp. Đây là nguồn dữ liệu phù hợp cho các dự án phân tích dữ liệu, học máy hoặc nghiên cứu trong lĩnh vực sức khỏe tâm thần học đường.

## Pipeline

### **Khám phá dữ liệu**

Giai đoạn này tầm trung vào việc tìm hiểu các đặc tính cơ bản của bộ dữ liệu nhằm chuẩn bị cho các bước phân tích sâu hơn. Cụ thể nhóm thực hiện các công việc

* + - Xác định kích thước của tập dữ liệu.
    - Khảo sát kiểu dữ liệu và ý nghĩa của các thuộc tính, vai trò của các thuộc tính này đối với bài toán, kiểu dữ liệu có phù hợp chưa.
    - Kiểm tra và xử lý các trường hợp lỗi dữ liệu như missing value, giá trị không phù hợp.
    - Thống kê mô tả, kiểm tra dữ liệu ngoại lai.

### **Tiền xử lý dữ liệu**

* + Từ kết quả của bước khám phá dữ liệu ta sẽ tiến hành tiền xử lý dữ liệu để loại bỏ các vấn đề ở bước khám phá dữ liệu hoặc có các công đoạn tạo lớp mới, chuẩn hoá giá trị và loại bỏ các thuộc tính, giá trị không phù hợp.

### **Trực quan hoá và phân tích**

* + Đây là giai đoạn cốt lõi nhất của đồ án, tập trung vào việc thực hiện các nghiên cứu sâu vào từng chủ đề con của vấn đề trầm cảm.
  + Sử dụng các công cụ AI, công cụ tìm kiếm cần khai thác tổng quát về các vấn đề liên quan, thông tin bên ngoài, các nghiên cứu xoay quanh chủ đề trầm cảm. Thực hiện các bước trực quan hoá biểu đồ để kiểm chứng giả thuyết, phân tích xu hướng.
  + Việc phân tích không chỉ nằm ở việc thấy trên biểu đồ mà đồng thời còn phải liên hệ với thực tế, kiểm chứng tình trạng và rút ra được các nhận xét, giải thích thoã mãn với kết quả thu được.

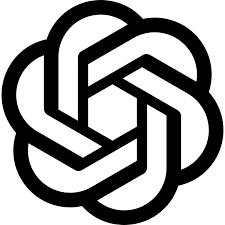
### **Mô hình hoá dữ liệu**

* + Mô hình hoá dữ liệu là công đoạn giúp nhóm tóm lại được nội dung của quá trình phân tích dữ liệu.
  + Cân nhắc các mô hình phân loại phù hợp, cải thiện quá trình huấn luyện để cho ra được kết quả tốt nhất.
  + Để triển khai kết quả khi cho mô hình học máy học từ tập dữ liệu để dự đoán tình trạng trầm cảm của một cá nhân thì nhóm sẽ thực hiện triển khai mô hình thành ứng dụng đơn giản, thân thiện.

## Công cụ AI hỗ trợ và phương thức

### **Công cụ AI**

* + *ChatGPT:*
    - ChatGPT là một mô hình ngôn ngữ lớn do OpenAI phát triển, dựa trên kiến trúc GPT (Generative Pre-trained Transformer). Được huấn luyện bằng cách học từ rất nhiều dữ liệu văn bản để hiểu và sinh ra ngôn ngữ tự nhiên.
    - Tác vụ: trả lời câu hỏi, trợ lý ảo hỗ trợ công việc (viết, tóm tắt, dịch thuật, hỗ trợ lập trình, v.v.), sáng tạo nội dung (viết bài, soạn email, kịch bản, thơ, v.v.), giải thích kiến thức phức tạp, trò chuyện tương tác tự nhiên với con người. Tương tác hội thoại tự nhiên hơn (giữ ngữ cảnh tốt qua nhiều lượt chat). Có khả năng tùy chỉnh "personality" (thái độ, phong cách trả lời) qua Custom Instructions hoặc GPTs riêng

 *Hình 1: ChatGPT*

* + Gemini:
    - Gemini là một họ các mô hình AI đa phương thức lớn (Large Multimodal Models - LMMs) được phát triển bởi Google AI. Điểm đặc biệt của Gemini là khả năng xử lý và hiểu nhiều loại thông tin khác nhau, không chỉ văn bản mà còn cả hình ảnh, âm thanh, video và mã nguồn. Điều này khác biệt so với nhiều mô hình ngôn ngữ lớn (LLMs) truyền thống thường tập trung chủ yếu vào văn bản.
    - Tác vụ: hiểu và trả lời câu hỏi phức tạp, tạo nội dung đa dạng, hỗ trợ lập trình, phân tích dữ liệu trực quan, tích hợp với Google ecosystem để hỗ trợ những tác vụ hằng ngày như: soạn email với Gmail, soạn thảo văn bản với Doc, tìm kiếm thông tin với Google.

 *Hình 2: Gemini*

* + Claude:
    - Claude là trợ lý AI thế hệ tiếp theo, được xây dựng dựa trên nghiên cứu của Anthropic về việc huấn luyện các hệ thống AI hữu ích, trung thực và vô hại. Có thể truy cập qua giao diện trò chuyện và API trong bảng điều khiển dành cho nhà phát triển, Claude có khả năng thực hiện đa dạng các tác vụ hội thoại và xử lý văn bản đồng thời duy trì độ tin cậy và tính dự đoán cao.
    - Claude có thể hỗ trợ các trường hợp sử dụng bao gồm tóm tắt, tìm kiếm, viết sáng tạo và hợp tác, hỏi đáp, lập trình và còn nhiều hơn thế nữa. Các khách hàng đầu tiên cho biết Claude ít có khả năng tạo ra kết quả gây hại, dễ trao đổi hơn và dễ điều chỉnh — nhờ đó có thể đạt được đầu ra mong muốn với ít nỗ lực hơn. Claude cũng có thể nhận hướng dẫn về tính cách, giọng điệu và cách ứng xử.

 *Hình 3: Claude*

* + Deepseek:
    - DeepSeek là một công ty phát triển AI có trụ sở tại Hàng Châu, Trung Quốc, được thành lập vào tháng 5 năm 2023 bởi Liang Wenfeng, cựu sinh viên Đại học Chiết Giang. Hiện DeepSeek hoạt động như một phòng thí nghiệm nghiên cứu AI độc lập dưới sự bảo trợ của High-Flyer – một quỹ đầu cơ định lượng tại Trung Quốc do Wenfeng đồng sáng lập. Mặc dù chưa công bố tổng số vốn đầu tư và định giá, phòng thí nghiệm này tập trung vào việc phát triển các công nghệ AI tiên tiến và thực hiện những nghiên cứu, phát triển đột phá.
    - Claude AI là trợ lý ngôn ngữ đa năng, có thể: Tóm tắt và trích xuất thông tin từ văn bản dài; Hỗ trợ viết sáng tạo, soạn thảo nội dung và hợp tác chỉnh sửa;  Trả lời câu hỏi chuyên sâu, giải thích khái niệm và tư vấn; Sinh mã, gỡ lỗi và tối ưu hóa code;  Dịch thuật giữ phong cách nguyên bản; Phân tích dữ liệu, tạo báo cáo và đưa ra insight; Triển khai chatbot chăm sóc khách hàng 24/7; Điều chỉnh giọng điệu, phong cách và cá nhân hóa theo nhu cầu người dùng.

 *Hình 4: Deepseek*

### **Phương thức ứng dụng**

* + Nhóm chủ trương thực hiện đồ án với phương hướng sử dụng đa dạng, hiệu quả các công cụ AI như một trợ lý đồng hành suốt quá trình xây dựng sản phẩm.
  + AI sẽ luôn được sử dụng như cố vấn khi xây dựng pipeline và nội dung chi tiết ở tất cả các bước để đánh giá đúng sai, hiệu quả và cải thiện tốt hơn. Đặc biệt là tại giai đoạn trực quan hoá và phân tích dữ liệu.
  + Các prompt AI hỗ trợ cần có sự mô tả rõ ràng, nội dung và ngữ cảnh cụ thể để cho ra kết quả tốt nhất. Tránh đặt các câu hỏi đóng kín làm giới hạn khả năng sáng tạo của cả người dùng và AI.
  + Sử dụng đa dạng các công cụ AI lẫn việc kiểm tra tính xác thực lại thủ công để đảm bảo kết quả được hỗ trợ bởi AI đúng khoa học.
  + AI sẽ được sử dụng như trợ lý thay vì chỉ biết vâng lệnh, các chủ đề sẽ được khai thác mở trước bằng việc cung cấp thông tin đa dạng, thuộc nhiều lĩnh vực khía cạnh khác nhau để có thể chạm tới các vấn đề sâu xa hơn.

# PHẦN 3: PHÂN TÍCH DỮ LIỆU

* Nhóm chủ trương sử dụng AI như một trợ lý hỗ trợ và nhà tư vấn xuyên suốt quá trình phân tích. Cụ thể, AI được khai thác để hỗ trợ việc trao đổi, đánh giá tính khả thi và mức độ chính xác của các bước thực hiện phân tích dữ liệu. Việc sử dụng AI không đơn thuần chỉ dừng lại ở việc cung cấp bộ dữ liệu và yêu cầu AI tự động tạo ra các phân tích, vì cách tiếp cận này có thể làm hạn chế khả năng mở rộng tư duy của cả công nghệ lẫn người sử dụng.
* Thay vào đó, nhóm lựa chọn cách tiếp cận chủ đề một cách nền tảng và toàn diện hơn: bắt đầu từ việc tìm hiểu các khái niệm cơ bản về bệnh trầm cảm, nghiên cứu những hiện tượng cụ thể cũng như các công trình khoa học liên quan. Sau khi có được cái nhìn tổng quan và lựa chọn được một chủ đề tiềm năng, có ý nghĩa thực tiễn, nhóm tiến hành kiểm chứng tính xác thực của chủ đề thông qua dữ liệu. Từ đó, nhóm tận dụng bộ dữ liệu hiện có để lựa chọn các phương pháp trực quan hóa phù hợp, đảm bảo việc phân tích vừa mang tính khoa học, vừa sâu sắc và thực tiễn.

## Phụ nữ và vấn đề trầm cảm ?

* + *Tại sao chủ đề này quan trọng:* Trầm cảm là kết quả của sự tương tác phức tạp giữa nhiều yếu tố bao gồm yếu tố sinh học, tâm lý-xã hội, yếu tố môi trường... Theo đó phụ nữ là nhóm người chịu ảnh hưởng nặng nhất từ các yếu tố trên do sự thay đổi mạnh mẽ của nội tiết tố như hormon estrogen và progesterone trong các giai đoạn quan trọng của cuộc đời như dậy thì. Thêm vào đó phụ nữ cũng là nhóm đối tượng bị ảnh hưởng nặng nề bởi xã hội như kỳ vọng, lạm dụng, bất công.
  + *Lợi ích khi phân tích chủ đề này:* Xác thực vấn đề này giúp mọi người trong xã hội có sự đối xử công bằng và hiểu rõ hơn về tình trạng sức khỏe tinh thần của phụ nữ. Đồng thời giúp chỉnh bản thân những người phụ nữ có thể nắm rõ các yếu tố ảnh hưởng đến bản thân để có sự điều chỉnh hợp lý đối với cuộc sống bản thân.
  + *AI đã hỗ trợ thế nào:*
    - ***Cung cấp nền tảng kiến thức chuyên sâu:*** AI nhanh chóng tổng hợp các báo cáo khoa học, bài nghiên cứu, và thống kê mới nhất về tình trạng trầm cảm ở phụ nữ, giúp nhóm tiết kiệm đáng kể thời gian tìm kiếm và tiếp cận chủ đề này.
    - ***Phân tích nguyên nhân chuyên sâu.***
    - ***Dự đoán chủ đề tiềm ẩn:*** Bằng cách sử dụng các kỹ thuật Topic Modeling, AI đã giúp phát hiện ra các nguyên nhân tiềm ẩn mà trước đây ít được chú ý.

## Sinh viên chịu ảnh hưởng từ trầm cảm ra sao ?

* + *Tại sao chủ đề này quan trọng:* Trầm cảm trong học đường, đặc biệt ở sinh viên và học sinh, đang là một vấn đề sức khoẻ tâm thần đáng báo động. Theo WHO, trầm cảm là một trong những nguyên nhân hàng đầu gây tàn tật và tử vong (tự tử) ở người trẻ tuổi. Sinh viên là nhóm đang trong giai đoạn chuyển giao mạnh mẽ giữa tuổi thiếu niên và trưởng thành, phải đối mặt với:
    - Áp lực học tập, thi cử, cạnh tranh. -
    - Lo âu về tương lai, nghề nghiệp, tài chính. -
    - Cô đơn, thiếu kết nối xã hội khi xa nhà hoặc trong môi trường mới.
  + *Lợi ích khi phân tích chủ đề này:* Nhận diện nhóm có nguy cơ cao để can thiệp sớm. Giúp các trường học, tổ chức xã hội hiểu rõ hơn về tác động tâm lý của việc học. Khuyến khích cải thiện môi trường học tập lành mạnh, đồng thời nâng cao nhận thức về chăm sóc sức khỏe tinh thần.
  + *AI đã hỗ trợ thế nào:*
    - **Thu thập và tổng hợp tài liệu tự động:** Sử dụng các mô hình NLP để quét, phân loại và tóm tắt hàng loạt bài báo, báo cáo nghiên cứu về trầm cảm học đường, giúp tiết kiệm thời gian và đảm bảo không bỏ sót các phát hiện quan trọng.
    - **Phân tích dữ liệu lớn & trực quan hóa**: Áp dụng các thuật toán học máy (machine learning) để xử lý dữ liệu khảo sát từ hàng nghìn học sinh, sinh viên; từ đó xây dựng dashboard

## Lối sống ảnh hưởng đến sức khoẻ tinh thần như thế nào?

* + *Tại sao chủ đề này lại quan trọng:* Chế độ ăn uống lành mạnh cung cấp các dưỡng chất cần thiết cho não bộ và cơ thể, giúp duy trì sự ổn định về hóa học và giảm viêm. Giấc ngủ đủ và chất lượng cho phép não bộ phục hồi, xử lý cảm xúc và hoạt động tối ưu. Khi cả hai yếu tố này bị xáo trộn, nó có thể làm tăng nguy cơ và mức độ nghiêm trọng của trầm cảm. Việc chú trọng đến cả chế độ ăn uống và giấc ngủ là một phần quan trọng trong việc phòng ngừa và hỗ trợ điều trị trầm cảm.
  + *Lợi ích khi phân tích chủ đề này:* Phân tích mối liên hệ giữa ăn uống, giấc ngủ và trầm cảm giúp ta hiểu rõ hơn nguyên nhân sâu xa của bệnh, từ đó phát triển các biện pháp can thiệp hiệu quả hơn dựa trên lối sống, cá nhân hóa phương pháp điều trị. Đồng thời đề xuất những khuyến nghị cụ thể để cải thiện sức khỏe tinh thần. Và cũng góp phần định hướng cho các chiến lược hỗ trợ sinh viên tại trường học hoặc trong cộng đồng.
  + *AI đã hỗ trợ thế nào:*
    - **Hỗ trợ coding.**
    - **Quan sát biểu đồ dữ liệu và phân tích mối quan hệ giữa các biến số.**
    - **Áp dụng kiến thức nền tảng để diễn giải:** về tâm lý học, sức khỏe cộng đồng và dinh dưỡng để diễn giải những xu hướng xuất hiện trong dữ liệu, thay vì chỉ mô tả bề mặt.
    - **Đào sâu vào các nguyên nhân tiềm ẩn trong bối cảnh xã hội:** chẳng hạn như áp lực học tập, kinh tế, và sự thay đổi trong lối sống hiện đại, từ đó kết nối được hành vi cá nhân với các yếu tố môi trường rộng lớn hơn.
    - **Phân tích mối quan hệ hai chiều giữa các yếu tố:** làm rõ cách thiếu ngủ, ăn uống kém có thể gây ra trầm cảm, và ngược lại, cách trầm cảm có thể tiếp tục phá vỡ thói quen ngủ nghỉ, ăn uống, tạo thành vòng luẩn quẩn tiêu cực.
    - **Đảm bảo cách diễn đạt:** vừa mang tính khoa học, vừa gần gũi với người đọc, phù hợp để sử dụng trong báo cáo, thuyết trình hoặc làm tài liệu nghiên cứu.
    - **Đưa ra thêm các đề xuất về hướng giải pháp:** như can thiệp môi trường học đường, giáo dục dinh dưỡng, hỗ trợ tâm lý để người học không chỉ dừng lại ở việc nắm bắt vấn đề mà còn định hướng hành động.

## Áp lực tài chính, học tập ảnh hưởng tới sinh viên thế nào ?

* + *Tại sao chủ đề này quan trọng:* Ảnh hưởng trực tiếp đến sức khỏe tâm thần và thể chất của sinh viên. Tác động đến hiệu suất học tập và khả năng hoàn thành chương trình học. Có thể dẫn đến các vấn đề nghiêm trọng như trầm cảm và ý tưởng tự tử. Hiểu biết về mối quan hệ này giúp các tổ chức giáo dục thiết kế chương trình hỗ trợ phù hợp.
  + *Lợi ích khi phân tích chủ đề này:* Xác định nhóm sinh viên có nguy cơ cao gặp khó khăn để can thiệp kịp thời. Phát triển các chính sách và chương trình hỗ trợ tài chính hiệu quả. Cải thiện môi trường học tập và giảm áp lực không cần thiết. Hiểu rõ mối tương quan giữa áp lực tài chính, áp lực học tập và các biến số khác như trầm cảm. Có cơ sở dữ liệu để đề xuất giải pháp cụ thể cho các vấn đề sinh viên đang gặp phải.

## Khả năng quản lý thời gian và bệnh trầm cảm

* + *Tại sao chủ đề này lại quan trọng:* Trầm cảm đang trở thành đại dịch thầm lặng ở sinh viên. Quản lý thời gian kém không chỉ ảnh hưởng đến điểm số mà còn là nguy cơ tâm lý lâu dài. Thành công học thuật chỉ có ý nghĩa khi đi kèm với sức khỏe tinh thần ổn định.
  + *Lợi ích khi phân tích chủ đề này:*Hiểu rõ mối quan hệ giữa thời gian học → stress → trầm cảm, để có thể: Thiết kế các chương trình giáo dục kỹ năng quản lý thời gian cho sinh viên ngay từ năm nhất. Tư vấn tâm lý nhắm trúng vào nhóm nguy cơ cao.
  + *AI đã hỗ trợ như thế nào:*
    - **Tiếp nhận và làm rõ bài toán.**
    - **Phân nhóm hành vi (Clustering):** Chuẩn hóa các biến số cần thiết. Áp dụng K-Means Clustering để phân loại dữ liệu thành 4 nhóm hành vi điển hình dựa trên sleep duration, study hours và CGPA.
    - **Kết nối với lý thuyết tâm lý học:** Vận dụng lý luận về stress học đường và burnout syndrome để giải thích các pattern, trích dẫn nguồn các bài báo liên quan.

# PHẦN 4: ĐÁNH GIÁ CÔNG CỤ AI

## ChatGPT

* + Điểm mạnh: khả năng giải thích và phân tích tương đối mạnh. Đưa ra nhiều phương pháp trong quá trình thực hiện, có nói rõ ưu nhược điểm và độ phù hợp của từng phương pháp. Có chức năng Research hỗ trợ phân tích chuyên sâu tương đối mạnh mẽ.
  + Điểm yếu: Đôi khi lan man dài dòng quá mức cần thiết. Có tính phí và giới hạn số lượng prompt. Khả năng bị hallucination khá cao, phải dùng song song với một AI hoặc search engine khác để verify lại câu trả lời. Hỗ trợ code thỉnh thoảng vẫn bị lạc đề với ngữ cảnh có sẵn.

## Gemini

* + Điểm mạnh: Trả lời ngắn gọn, đúng trọng tâm, có hỗ trợ trích dẫn nguồn. Không giới hạn prompt để xử lý ảnh tuần tự. Hỗ trợ code khá đúng ngữ cảnh. Code và các thư viện liên quan được up-to-date.
  + Điểm yếu: Thỉnh thoảng bị nhảy ra những ngôn ngữ kì lạ. Không hỗ trợ xử lý nhiều file ảnh cùng lúc. Đưa ra phương pháp nhưng không nêu rõ độ phù hợp của phương pháp (vd: t-test chỉ dùng cho dữ liệu có phân phối chuẩn nhưng vẫn được đề xuất dù thông tin về phân phối của dữ liệu chưa được cung cấp).

## Claude

* + Điểm mạnh: Trả lời tự nhiên mượt mà. Code logic khá tốt và sạch. Phân tích chuyên sâu tốt.
  + Điểm yếu: Ít hỗ trợ tìm kiếm thông tin thời gian thực, không thể đọc file upload quá nặng. Xử lý các câu hỏi ngắn từ người dùng.

## Deepseek

* + Điểm mạnh: Hỗ trợ tiếng Việt tốt, miễn phí, khả năng tìm kiếm & xử lý dữ liệu mạnh (hỗ trợ upload file). Có thể xử lý câu hỏi dài từ người dùng. Code sạch và logic.
  + Điểm yếu: Đôi khi chi tiết hóa quá mức cần thiết (overfit). Khi đối mặt các vấn đề phức tạp, xuất hiện lỗi logic khá nhiều. Phiên bản miễn phí bị giới hạn khá nhiều.

## So sánh hiệu quả

|  |  |  |
| --- | --- | --- |
|  | Phù hợp | Hạn chế |
| ChatGPT | Phù hợp cho phân tích chuyên sâu. Brainstorm để chọn phương pháp phù hợp. Ngôn ngữ tự nhiên, mượt mà, dễ hiểu. | Dễ bị hallucination. Hỗ trợ hình ảnh và các file đa phương tiện có giới hạn sau đó cần tính phí do đó phải lập kế hoạch sử dụng cho phù hợp. |
| Gemini | Sinh code, kiểm tra logic. Phân tích sơ bộ. Hỗ trợ tìm dẫn chứng cho luận điểm hoặc bài báo khoa học. Có thể dùng song song như công cụ để kiểm tra lại kết quả. | Không phân tích chuyên sâu nhiều biểu đồ cùng lúc và quên ngữ cảnh tương đối nhanh. |
| Claude | Coding logic khá chi tiết. Phân tích các vấn đề chuyên sâu khá tốt. | Không thể đọc file quá nặng, cần mô tả chi tiết. |
| Deepseek | Tổng hợp thông tin thời gian thực. Đưa ra các gợi ý hướng dẫn khá tốt. | Tốc độ phản hồi chậm. Giới hạn cho 1 câu trả lời khá ngắn. |

# PHẦN 5: MÔ HÌNH HOÁ DỮ LIỆU

## Tiền xử lý dữ liệu

* + Tiền xử lí dữ liệu gốc thành dữ liệu phù hợp cho quá trình huấn luyện mô hình.
  + Nhóm sử dụng hai công cụ hỗ trợ tại bước này là ChatGPT và Copilot.
  + Quá trình thực hiện sẽ bao gồm: tải file dữ liệu lên và yêu cầu tiền xử lý dữ liệu, thực hiện khám phá dữ liệu sau tiền xử lý.
  + Kết quả thu được là quy trình xử lý của ChatGPT với hiệu quả thể hiện qua: Đã loại bỏ cột gốc sau khi one-hot encoding, tránh trùng lặp thông tin, tên cột gọn gàng hơn, loại bỏ các cột không cần thiết giúp giảm thời gian huấn luyện.
  + Đánh giá cho thấy

|  |  |
| --- | --- |
| ChatGPT | Copilot |
| * Đưa ra các phân tích đầy đủ và chi tiết hơn, nhiều thông tin hơn. * Tốc độ phản hồi nhanh. | * Đưa các phân tích ngắn gọn hơn,  tập trung vào vấn đề chính. * Tốc độ phản hồi chậm. |

## Lựa chọn mô hình phù hợp

* + Mục tiêu: Xây dựng mô hình AI dự đoán xem một cá nhân có nguy cơ cao mắc trầm cảm hay không, dựa trên thông tin cá nhân và lối sống.
  + Loại bài toán: Phân loại nhị phân (Binary Classification), đầu ra là nhãn 0 (không nguy cơ) hoặc 1 (nguy cơ cao).
  + Dữ liệu đã được xử lý sẵn (one-hot encoding). Các đặc trưng bao gồm:Giới tính, Thành phố sống, Nghề nghiệp, Bằng cấp, Thói quen ăn uống, Thời lượng giấc ngủ… Đặc trưng hỗn hợp giữa biến phân loại và biến số lượng nhỏ.
  + Đặc điểm quan trọng:
    - Các biến đầu vào chủ yếu là phân loại → cần mô hình hỗ trợ one-hot tốt.
    - Không yêu cầu tính liên tục của đầu ra → bài toán classification thuần.
  + XGBoost Classifier là mô hình được lựa chọn bởi:
    - Tối ưu cho dữ liệu dạng bảng (tabular data).
    - Khả năng khai thác quan hệ phi tuyến tính rất tốt.
    - Chống overfitting nhờ Regularization (lambda, alpha).
    - Cho phép kiểm soát chi tiết quá trình train qua tham số (max\_depth, learning\_rate...).
  + Công cụ AI có sự đóng góp quan trọng ở giai đoạn này:

|  |  |
| --- | --- |
| **Công cụ AI** | **Vai trò** |
| ChatGPT 4.0 | Gợi ý danh sách mô hình phân loại phù hợp (Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM). |
| Bing Copilot | So sánh chi tiết ưu nhược từng mô hình, tư vấn chọn XGBoost do hiệu quả cao trên dữ liệu dạng bảng nhỏ/vừa. |
| Perplexity AI | Tra cứu nghiên cứu gần đây, xác nhận XGBoost vẫn được sử dụng nhiều trong các bài toán dự đoán sức khỏe tâm thần. |

## Huấn luyện mô hình

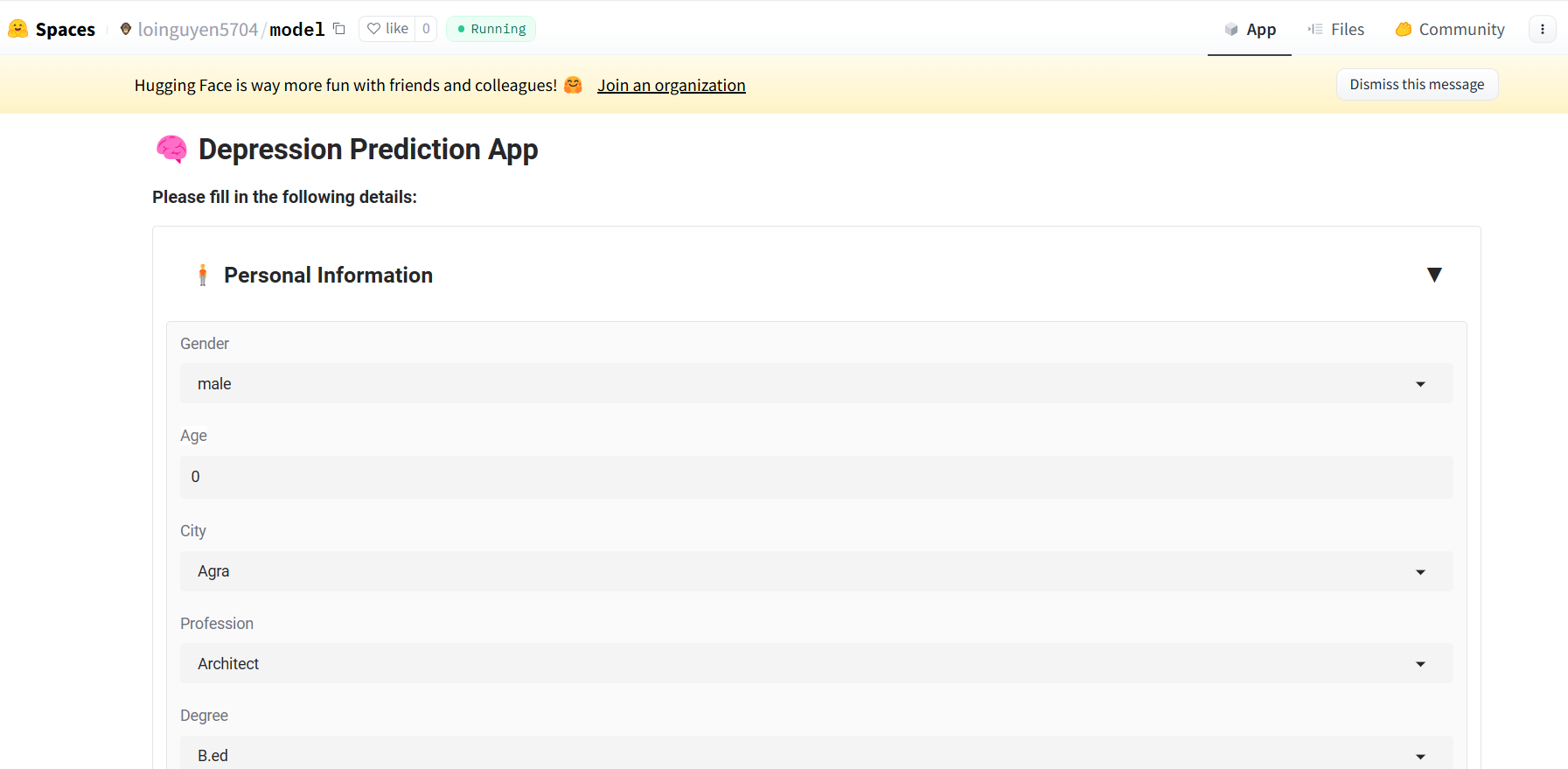
* + Dữ liệu đã được chuẩn hóa sẵn: one-hot encoding cho các biến phân loại.
  + Không cần scaling thêm. Chia dữ liệu theo tỷ lệ 80% Train / 20% Test.
  + XGBoost Classifier (xgboost.XGBClassifier) — một mô hình cây quyết định tăng cường mạnh mẽ, phù hợp cho bài toán phân loại nhị phân như dự đoán nguy cơ trầm cảm.
  + Các siêu tham số ban đầu được thiết lập và tối ưu bằng phương pháp GridSearchCV.
  + Thiết kế hàm mất mát có trọng số: Nhằm nâng cao khả năng phát hiện các trường hợp người có nguy cơ cao mắc trầm cảm, mô hình đã sử dụng một hàm mất mát tùy chỉnh (weighted\_logloss) thay cho hàm mất mát mặc định của XGBoost.
  + Đánh giá hiệu quả mô hình:
    - Accuracy: Tỷ lệ dự đoán đúng tổng thể.
    - Precision, Recall, F1-score: Đặc biệt chú trọng Recall, vì bài toán y tế ưu tiên phát hiện đúng những ca nguy cơ cao hơn là chỉ đạt độ chính xác chung.
    - Confusion Matrix: Phân tích chi tiết các trường hợp dự đoán đúng/sai.
  + Vai trò của công cụ AI:

|  |  |
| --- | --- |
| **Công cụ AI** | **Vai trò** |
| ChatGPT 4.0 | Hướng dẫn thiết kế quy trình huấn luyện (train/test split → train → evaluate). |
| Google Gemini Advanced | Gợi ý giá trị siêu tham số khởi tạo tối ưu cho XGBoost phù hợp với dữ liệu nhỏ/vừa. |
| Bing Copilot | Đề xuất thêm kỹ thuật như Cross-validation, GridSearchCV để tăng độ chính xác mô hình. |
| Perplexity AI | Tra cứu tầm quan trọng của Recall trong các bài toán y tế tâm thần để xác định ưu tiên khi tối ưu hóa mô hình. |

## Triển khai mô hình

* + **Mục tiêu bài toán:** Xây dựng một trang web đã triển khai mô hình, có thể dự đoán được người đó có khả năng đang bị trầm cảm hay không.
  + **Các công cụ sử dụng:** ChatGPT, Gradio và Hugging Face Spaces
  + **Đánh giá công cụ Gradio và Hugging Face Spaces:**

|  |  |
| --- | --- |
| Ưu điểm | Nhược điểm |
| * Miễn phí. * Dễ triển khai. * Thân thiện người dùng. | * Giao diện không thể điều chỉnh nhiều. * Không phù hợp với model nặng (vài GB trở lên). |

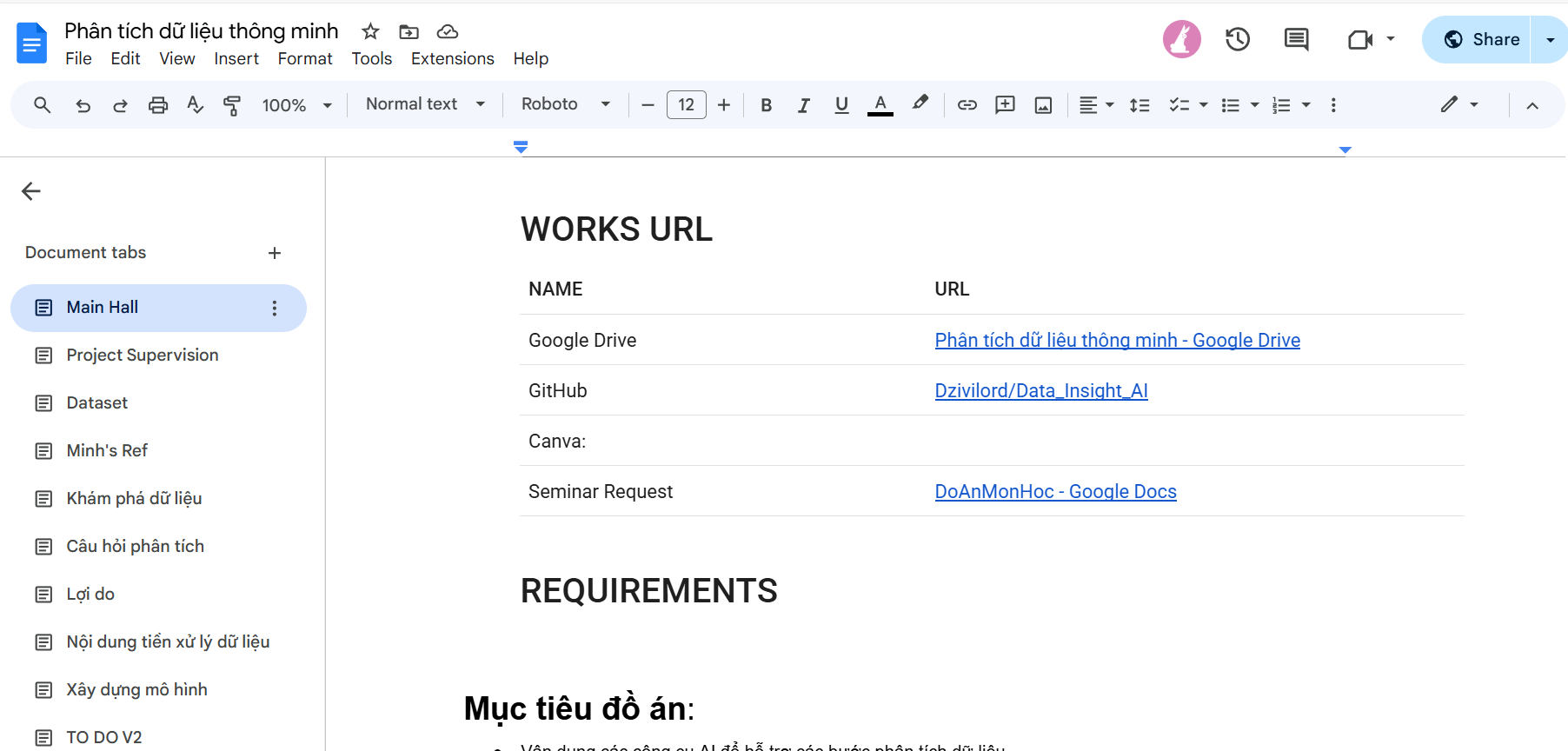
**

*Hình 5: Mô hình dự đoán*

# PHẦN 6: BÁO CÁO NHÓM

## Phương thức hoạt động

* + Nhóm hoạt động dựa trên việc khuyến khích sự trao đổi, thảo luận thường xuyên giữa các thành viên. Ngoài ra, nhóm còn áp dụng các nguyên tắc sau:
    - Chia sẻ kiến thức và kỹ năng: Các thành viên hỗ trợ thực hiện phân tích dữ liệu, lập trình và xử lý mô hình thông qua việc chia sẻ kinh nghiệm cá nhân và tài liệu tham khảo.
    - Phân chia công việc hợp lý: Dựa trên thế mạnh và sở trường của từng thành viên, nhóm phân công nhiệm vụ rõ ràng nhằm tối ưu hóa hiệu quả làm việc.
    - Trao đổi phản hồi liên tục: tất cả thắc mắc đều được các thành viên chia sẻ tại các công cụ trao đổi để nhanh chóng nhận được ý kiến từ các thành viên.
    - Chủ động và sáng tạo : Nhóm khuyến khích mỗi thành viên đề xuất ý tưởng mới, phản biện và đóng góp vào thành quả của nhóm.
    - Thống nhất quyết định dựa trên thảo luận: Các quyết định quan trọng (lựa chọn tập dữ liệu, hướng xây dựng pipeline, thiết kế triển khai...) đều được đưa ra dựa trên ý kiến đóng góp của tập thể, đảm bảo sự đồng thuận của tất cả thành viên.

**

*Hình 6: Workspace*

## Phân công công việc

|  |  |
| --- | --- |
| Thành viên | Vai trò chính |
| Đinh Viết Lợi | Điều hành nhóm, phân chia công việc, tạo lập môi trường làm việc, trao đổi và mẫu yêu cầu từng phần. Khám phá dữ liệu  Viết báo cáo tổng hợp |
| Nguyễn Trần Lợi | Tiền xử lý dữ liệu Triển khai mô hình sau huấn luyện  Quay video demo sản phẩm |
| Lê Võ Nhật Minh | Phân tích chuyên sâu các chủ đề Nghiên cứu và đánh giá công cụ ChatGPT và Gemini  Chỉnh sửa video demo |
| Nguyễn Hữu Nghĩa | Huấn luyện mô hình dự đoán  Làm slide trình bày |
| Nguyễn Lê Tấn Phát | Phân tích chuyên sâu các chủ đề Nghiên cứu và đánh giá công cụ Claude và DeepSeek  Soạn nội dung trình bày slide |

# PHẦN 7: TÀI LIỆU THAM KHẢO

## Tài liệu nhóm

* + Kho lưu trữ chung: [Folder - Google Drive](https://drive.google.com/drive/folders/1R3ho4ebXHVRL_S_EGQIxq9v95AQLt_V_)
  + Version control system: [Dzivilord/Data\_Insight\_AI](https://github.com/Dzivilord/Data_Insight_AI)
  + Slide trình bày:
  + Workspace: [Phân tích dữ liệu thông minh - Google Docs](https://docs.google.com/document/d/1Ke7d6Smshj51UX6EGZO8jHb_GZdDhh3YRSamReRSuv0/edit?tab=t.qaho97so9pg4)

## Tài liệu tham khảo

1. Community’s Code, ‘Student Depression Dataset’. Đường dẫn: [Student Depression Dataset](https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset).
2. ChatGPT- OpenAI: [OpenAI](https://openai.com/)
3. Gemini- Google: [Gemini](https://gemini.google.com/app)
4. DeepSeek- Shēndù Qiúsuǒ : [DeepSeek | 深度求索](https://www.deepseek.com/)
5. Claude- Anthropic: [Claude](https://claude.ai/login?returnTo=%2F%3F)