**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU NHU CẦU VỀ DỊCH VỤ CHĂM SÓC SỨC KHỎE TINH THẦN TẠI CÁC CƠ SỞ Y TẾ.**

**Giáo viên hướng dẫn: ThS. Nguyễn Trà Mi**

**Sinh viên thực hiện:**  Nguyễn Vọng

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mã sv** | **Họ và tên** | **Lớp** |
| 1 | 1771020358 | Nguyễn Quốc Huy | CNTT 17-02 |
| 2 | 1771020609 | Bùi Trọng Tài | CNTT 17-02 |
| 3 | 1771020761 | Nguyễn Vọng | CNTT 17-02 |

**Hà Nội, năm 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU NHU CẦU VỀ DỊCH VỤ CHĂM SÓC SỨC KHỎE TINH THẦN TẠI CÁC CƠ SỞ Y TẾ.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1771020358 | Nguyễn Quốc Huy | 18/03/2005 |  |  |
| 2 | 1771020609 | Bùi Trọng Tài | 15/01/2005 |  |  |
| 3 | 1771020761 | Nguyễn Vọng | 15/11/2005 |  |  |

**CÁN BỘ CHẤM THI**

**Hà Nội, năm 2025**

**LỜI NÓI ĐẦU**

Trong những năm gần đây, sức khỏe tinh thần đã trở thành một vấn đề quan trọng được xã hội quan tâm. Áp lực từ công việc, học tập, các mối quan hệ cá nhân và những biến động trong cuộc sống, đặc biệt là ảnh hưởng của đại dịch COVID-19, đã khiến nhu cầu về các dịch vụ chăm sóc sức khỏe tinh thần ngày càng gia tăng. Tuy nhiên, việc tiếp cận các dịch vụ này vẫn còn gặp nhiều rào cản, bao gồm sự thiếu hụt nhân lực y tế chuyên môn, hạn chế về nhận thức trong cộng đồng và những định kiến xã hội đối với các vấn đề tâm lý. Vì vậy, việc phân tích dữ liệu về nhu cầu sử dụng dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế không chỉ giúp đánh giá thực trạng mà còn cung cấp cơ sở khoa học để đề xuất giải pháp nâng cao chất lượng dịch vụ, đáp ứng tốt hơn nhu cầu của người dân.

Báo cáo này được thực hiện trong khuôn khổ môn học Xác suất Thống kê và Phân tích Dữ liệu, với mục tiêu ứng dụng các phương pháp thống kê để nghiên cứu nhu cầu sử dụng dịch vụ chăm sóc sức khỏe tinh thần. Thông qua việc thu thập và phân tích dữ liệu, nghiên cứu sẽ tập trung đánh giá xu hướng tìm kiếm dịch vụ theo thời gian, các nhóm đối tượng có nhu cầu cao, các yếu tố ảnh hưởng đến quyết định sử dụng dịch vụ, cũng như hiệu suất hoạt động của các cơ sở y tế trong lĩnh vực này. Các phương pháp được áp dụng bao gồm thống kê mô tả, kiểm định giả thuyết, phân tích hồi quy và trực quan hóa dữ liệu nhằm đảm bảo kết quả phân tích có độ tin cậy cao.

**MỤC LỤC**

[CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 6](#_Toc21544)

[1.1. Giới thiệu chung về đề tài lựa chọn 6](#_Toc31465)

[1.2. Lý do chọn đề tài 7](#_Toc18394)

[1.3. Mục tiêu 8](#_Toc18651)

[1.4. Phạm vi nghiên cứu của đề tài 9](#_Toc19467)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 10](#_Toc14029)

[2.1. Trình bày về các khái niệm thống kê 10](#_Toc17606)

[2.1.1. Tổng thể và mẫu nghiên cứu 10](#_Toc28308)

[2.1.2. Các đặc trưng của mẫu và tổng thể 10](#_Toc31404)

[2.2. Trình bày các khái niệm về phân tích dữ liệu 11](#_Toc27594)

*[2.2.1. Khái niệm về dữ liệu và phân tích dữ liệu](#_Toc626)* [11](#_Toc626)

*[2.2.2. Quy trình phân tích dữ liệu](#_Toc16870)* [12](#_Toc16870)

[CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU 15](#_Toc29944)

[3.1. Thu thập dữ liệu 15](#_Toc21606)

[3.2. Làm sạch dữ liệu 16](#_Toc9564)

[3.3. Khám phá dữ liệu 18](#_Toc6436)

[3.3.1. Thống kê mô tả 18](#_Toc11395)

[3.3.2. Phân tích mối quan hệ giữa các biến 19](#_Toc32309)

[3.3.3. Trực quan hóa dữ liệu 19](#_Toc7645)

[3.4. Xây dựng mô hình 20](#_Toc6169)

[3.5. Trực quan hóa dữ liệu 27](#_Toc6012)

[3.6. Đưa ra kết luận và quyết định 30](#_Toc8699)

[KẾT LUẬN 33](#_Toc12214)

[DANH MỤC TÀI LIỆU THAM KHẢO 39](#_Toc5345)

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

## **Giới thiệu chung về đề tài lựa chọn**

Sức khỏe tinh thần đóng vai trò quan trọng đối với chất lượng cuộc sống của mỗi cá nhân và sự phát triển bền vững của xã hội. Trong bối cảnh hiện đại, áp lực từ công việc, học tập, tài chính, các mối quan hệ cá nhân và đặc biệt là ảnh hưởng của đại dịch COVID-19 đã khiến các vấn đề về tâm lý ngày càng trở nên phổ biến. Tuy nhiên, không phải ai cũng có nhận thức đầy đủ về tầm quan trọng của sức khỏe tinh thần cũng như biết cách tiếp cận các dịch vụ hỗ trợ. Việc phân tích dữ liệu về nhu cầu sử dụng dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế là cần thiết để đánh giá thực trạng, xác định các yếu tố ảnh hưởng và đề xuất các giải pháp cải thiện chất lượng dịch vụ nhằm đáp ứng tốt hơn nhu cầu của người dân.

Trong khuôn khổ môn học **Xác suất Thống kê và Phân tích Dữ liệu, đề tài "Phân tích dữ liệu nhu cầu về dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế"** được lựa chọn nhằm ứng dụng các phương pháp thống kê và phân tích dữ liệu vào một vấn đề thực tiễn có ý nghĩa xã hội. Thông qua việc thu thập, xử lý và phân tích dữ liệu thực tế, nghiên cứu sẽ tập trung đánh giá xu hướng tìm kiếm dịch vụ theo thời gian, xác định nhóm đối tượng có nhu cầu cao, các yếu tố tác động đến quyết định sử dụng dịch vụ và hiệu suất hoạt động của các cơ sở y tế. Kết quả thu được không chỉ giúp nâng cao hiểu biết về thực trạng dịch vụ chăm sóc sức khỏe tinh thần mà còn cung cấp cơ sở khoa học để các cơ quan quản lý và cơ sở y tế đề ra các chính sách phù hợp, hướng đến việc nâng cao chất lượng cuộc sống của cộng đồng.

## **Lý do chọn đề tài**

Trong những năm gần đây, sức khỏe tinh thần đã trở thành một trong những vấn đề được quan tâm hàng đầu do ảnh hưởng của áp lực cuộc sống, công việc và học tập. Đặc biệt, sau đại dịch COVID-19, số lượng người gặp các vấn đề tâm lý như căng thẳng, lo âu, trầm cảm gia tăng đáng kể, làm nổi bật nhu cầu cấp thiết về các dịch vụ chăm sóc sức khỏe tinh thần. Tuy nhiên, thực tế cho thấy việc tiếp cận các dịch vụ này vẫn còn gặp nhiều khó khăn do nhận thức xã hội chưa đầy đủ, sự kỳ thị về bệnh tâm lý, cũng như hạn chế trong nguồn lực và chất lượng dịch vụ tại các cơ sở y tế. Vì vậy, việc nghiên cứu và phân tích dữ liệu về nhu cầu sử dụng dịch vụ chăm sóc sức khỏe tinh thần là rất quan trọng để đánh giá thực trạng, xác định xu hướng phát triển và đề xuất giải pháp nhằm nâng cao chất lượng dịch vụ.

Bên cạnh đó, đề tài này còn mang ý nghĩa học thuật và thực tiễn khi ứng dụng các phương pháp thống kê và phân tích dữ liệu trong lĩnh vực y tế. Môn học **Xác suất Thống kê và Phân tích Dữ liệu** cung cấp các công cụ hữu ích để thu thập, xử lý, phân tích và trực quan hóa dữ liệu, từ đó đưa ra những nhận định khách quan và có giá trị khoa học. Việc thực hiện nghiên cứu này không chỉ giúp nhóm thực hiện nâng cao kỹ năng phân tích dữ liệu mà còn đóng góp vào việc xây dựng các đề xuất cải thiện hệ thống chăm sóc sức khỏe tinh thần tại các cơ sở y tế. Những kết quả thu được có thể là tài liệu tham khảo hữu ích cho các nhà quản lý y tế, bác sĩ tâm lý và các tổ chức liên quan trong việc nâng cao nhận thức cộng đồng và cải thiện chất lượng dịch vụ chăm sóc sức khỏe tinh thần.

## **Mục tiêu**

Nghiên cứu **"Phân tích dữ liệu nhu cầu về dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế"** được thực hiện với các mục tiêu chính sau:

* **Đánh giá thực trạng nhu cầu**: Phân tích số liệu về mức độ sử dụng dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế, xác định xu hướng thay đổi theo thời gian và mức độ quan tâm của các nhóm dân số khác nhau.
* **Xác định các yếu tố ảnh hưởng**: Tìm hiểu các yếu tố tác động đến quyết định sử dụng dịch vụ, bao gồm độ tuổi, giới tính, nghề nghiệp, thu nhập, khu vực sinh sống, nhận thức xã hội và khả năng tiếp cận dịch vụ y tế.
* **Ứng dụng các phương pháp phân tích dữ liệu**: Sử dụng các công cụ thống kê như phân tích hồi quy, kiểm định giả thuyết, phân cụm dữ liệu và trực quan hóa dữ liệu để đưa ra nhận định khách quan về nhu cầu và hiệu quả của các dịch vụ chăm sóc sức khỏe tinh thần.
* **Đưa ra đề xuất cải thiện**: Dựa trên kết quả phân tích, đề xuất các giải pháp nhằm nâng cao chất lượng dịch vụ, tăng khả năng tiếp cận của người dân và giảm thiểu các rào cản trong việc chăm sóc sức khỏe tinh thần.
* **Ứng dụng vào thực tiễn**: Cung cấp dữ liệu và nhận định khoa học có giá trị tham khảo cho các cơ sở y tế, nhà quản lý và các tổ chức liên quan nhằm phát triển chính sách hỗ trợ sức khỏe tinh thần tốt hơn.

## **Phạm vi nghiên cứu của đề tài**

Nghiên cứu "Phân tích dữ liệu nhu cầu về dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế" được thực hiện trong phạm vi sau:

* Phạm vi nội dung:  
  Nghiên cứu tập trung vào việc thu thập, phân tích và đánh giá dữ liệu liên quan đến nhu cầu sử dụng dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế. Các khía cạnh được xem xét bao gồm số lượng người tìm kiếm dịch vụ, mức độ thường xuyên sử dụng, các loại hình dịch vụ phổ biến (tư vấn tâm lý, điều trị trầm cảm, rối loạn lo âu, hỗ trợ tâm lý sau sang chấn...), và các yếu tố ảnh hưởng đến quyết định sử dụng dịch vụ.
* Phạm vi không gian:  
  Đề tài tập trung phân tích dữ liệu tại các cơ sở y tế trong phạm vi xác định, có thể là tại một khu vực địa lý cụ thể (ví dụ: một thành phố, tỉnh/thành hoặc một số bệnh viện, phòng khám nhất định). Trường hợp dữ liệu rộng hơn, nghiên cứu có thể mở rộng để so sánh giữa các khu vực thành thị và nông thôn.
* Phạm vi thời gian:  
  Dữ liệu được thu thập trong khoảng thời gian nhất định để đảm bảo tính thời sự và phù hợp với xu hướng hiện tại. Thông thường, nghiên cứu sẽ sử dụng dữ liệu trong năm gần đây để đánh giá sự thay đổi của nhu cầu theo thời gian. Nếu có đủ dữ liệu, có thể thực hiện phân tích theo từng giai đoạn (trước và sau đại dịch COVID-19, theo từng năm, theo mùa…).
* Phạm vi phương pháp:  
  Nghiên cứu áp dụng các phương pháp thống kê mô tả, phân tích hồi quy, kiểm định giả thuyết và trực quan hóa dữ liệu để phân tích xu hướng và các yếu tố tác động đến nhu cầu dịch vụ chăm sóc sức khỏe tinh thần. Dữ liệu có thể được thu thập từ hồ sơ bệnh án, khảo sát người dùng dịch vụ, báo cáo của các cơ quan y tế hoặc nguồn dữ liệu mở có liên quan.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## **Trình bày về các khái niệm thống kê**

## **Tổng thể và mẫu nghiên cứu**

Trong thống kê, **tổng thể (population)** là tập hợp tất cả các phần tử hoặc đối tượng có liên quan đến một nghiên cứu cụ thể. Tổng thể có thể là tất cả bệnh nhân đến khám tại các cơ sở y tế về sức khỏe tinh thần, tất cả người dân trong một khu vực nhất định, hoặc tất cả hồ sơ bệnh án trong một giai đoạn cụ thể.

Do kích thước của tổng thể thường rất lớn, việc nghiên cứu toàn bộ tổng thể là không khả thi. Vì vậy, các nhà thống kê thường chọn một **mẫu (sample)** – tức là một tập hợp con của tổng thể – để nghiên cứu. Mẫu phải được chọn sao cho đại diện cho tổng thể nhằm đảm bảo tính chính xác và khả năng suy luận của kết quả. Phương pháp chọn mẫu có thể bao gồm lấy mẫu ngẫu nhiên đơn giản, lấy mẫu phân tầng hoặc lấy mẫu theo cụm.

Trong nghiên cứu này, tổng thể có thể là tất cả bệnh nhân sử dụng dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế trong một khoảng thời gian nhất định, trong khi mẫu nghiên cứu có thể là một nhóm bệnh nhân được chọn từ dữ liệu bệnh án hoặc từ các khảo sát.

## **Các đặc trưng của mẫu và tổng thể**

Một số đặc trưng quan trọng dùng để mô tả tổng thể và mẫu trong thống kê bao gồm:

* **Trung bình (Mean, μ hoặc x̄)**: Giá trị trung bình của một tập hợp dữ liệu, giúp phản ánh xu hướng chung của dữ liệu.
* **Trung vị (Median)**: Giá trị nằm ở vị trí chính giữa khi sắp xếp dữ liệu theo thứ tự tăng dần hoặc giảm dần. Trung vị đặc biệt hữu ích khi dữ liệu có các giá trị ngoại lệ.
* **Độ lệch chuẩn (Standard Deviation, σ hoặc s)**: Đo lường mức độ phân tán của dữ liệu xung quanh giá trị trung bình. Độ lệch chuẩn cao cho thấy dữ liệu có sự biến động lớn.
* **Phương sai (Variance, σ² hoặc s²)**: Là bình phương của độ lệch chuẩn, thể hiện mức độ phân tán của dữ liệu.
* **Hệ số biến thiên (Coefficient of Variation, CV)**: Tỷ lệ giữa độ lệch chuẩn và trung bình, giúp đánh giá mức độ biến động của dữ liệu một cách tương đối.
* **Tứ phân vị (Quartiles)**: Phân chia dữ liệu thành bốn phần bằng nhau, thường được sử dụng trong biểu đồ hộp (boxplot) để mô tả sự phân bố dữ liệu.

## **Trình bày các khái niệm về phân tích dữ liệu**

### *Khái niệm về dữ liệu và phân tích dữ liệu*

**Dữ liệu (data)** là tập hợp các thông tin thu thập được từ quan sát, khảo sát, thí nghiệm hoặc ghi nhận thực tế. Dữ liệu có thể được phân thành nhiều loại khác nhau:

* **Dữ liệu định lượng (Quantitative Data)**: Gồm các số liệu có thể đo lường, như số lượng bệnh nhân đến khám, số lần điều trị trung bình, tuổi của bệnh nhân,...
* **Dữ liệu định tính (Qualitative Data)**: Bao gồm các thông tin mang tính mô tả, như giới tính, mức độ căng thẳng (thấp, trung bình, cao), loại hình dịch vụ chăm sóc sức khỏe tinh thần sử dụng.

**Phân tích dữ liệu (data analysis)** là quá trình xử lý, biến đổi và mô hình hóa dữ liệu nhằm khám phá thông tin hữu ích, hỗ trợ đưa ra quyết định và dự đoán xu hướng. Quá trình phân tích dữ liệu thường bao gồm các bước sau:

1. **Thu thập dữ liệu**: Lấy dữ liệu từ khảo sát, hồ sơ bệnh án hoặc các nguồn dữ liệu mở.
2. **Làm sạch dữ liệu**: Loại bỏ các giá trị thiếu, kiểm tra và xử lý dữ liệu ngoại lệ.
3. **Mô tả dữ liệu**: Sử dụng các phương pháp thống kê mô tả để hiểu về đặc điểm của dữ liệu.
4. **Phân tích khám phá**: Áp dụng trực quan hóa dữ liệu để tìm ra các mẫu (patterns) và xu hướng (trends).
5. **Kiểm định giả thuyết và mô hình hóa**: Áp dụng các phương pháp thống kê suy diễn, như kiểm định t, phân tích phương sai (ANOVA), hồi quy tuyến tính để rút ra kết luận từ dữ liệu.
6. **Diễn giải kết quả và đưa ra quyết định**: Dựa vào kết quả phân tích để đưa ra nhận định và đề xuất cải thiện hệ thống chăm sóc sức khỏe tinh thần.

### *Quy trình phân tích dữ liệu*

Phân tích dữ liệu là một quá trình có hệ thống nhằm trích xuất thông tin hữu ích từ dữ liệu để hỗ trợ ra quyết định. Trong nghiên cứu "Phân tích dữ liệu nhu cầu về dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế", quy trình phân tích dữ liệu được thực hiện qua các bước chính sau:

1. Thu thập dữ liệu

Dữ liệu cần được thu thập từ các nguồn đáng tin cậy để đảm bảo tính chính xác và phù hợp với mục tiêu nghiên cứu. Các nguồn dữ liệu có thể bao gồm:

Hồ sơ bệnh án từ các cơ sở y tế về số lượng bệnh nhân sử dụng dịch vụ chăm sóc sức khỏe tinh thần.

Khảo sát trực tiếp với bệnh nhân hoặc người dân về nhu cầu và trải nghiệm khi sử dụng dịch vụ.

Dữ liệu từ báo cáo y tế của các tổ chức chăm sóc sức khỏe cộng đồng.

Dữ liệu mở từ các nghiên cứu trước đó hoặc từ các tổ chức y tế công cộng.

2. Làm sạch và tiền xử lý dữ liệu

Dữ liệu sau khi thu thập thường chứa lỗi hoặc thiếu sót, do đó cần được xử lý để đảm bảo chất lượng:

Xử lý dữ liệu thiếu: Điền giá trị còn thiếu bằng phương pháp trung bình, trung vị hoặc loại bỏ các bản ghi không đầy đủ.

Xử lý ngoại lệ: Loại bỏ hoặc điều chỉnh các giá trị bất thường gây ảnh hưởng đến phân tích.

Mã hóa dữ liệu định tính: Chuyển đổi dữ liệu định tính (ví dụ: mức độ căng thẳng "thấp", "trung bình", "cao") thành dữ liệu số để dễ phân tích.

Chuẩn hóa dữ liệu: Đưa dữ liệu về cùng một đơn vị hoặc phạm vi giá trị nhằm đảm bảo tính nhất quán.

3. Khám phá dữ liệu (Exploratory Data Analysis - EDA)

Sau khi làm sạch dữ liệu, cần thực hiện phân tích khám phá để hiểu rõ đặc điểm của dữ liệu:

Thống kê mô tả: Tính toán các giá trị trung bình, trung vị, phương sai, độ lệch chuẩn của các biến số liên quan.

Trực quan hóa dữ liệu: Dùng biểu đồ cột, biểu đồ phân tán, biểu đồ hộp (boxplot) để tìm hiểu xu hướng và sự phân bố dữ liệu.

Xác định các mối quan hệ giữa biến số: Sử dụng ma trận tương quan để xem xét mối quan hệ giữa các yếu tố như độ tuổi, giới tính, nghề nghiệp với nhu cầu sử dụng dịch vụ.

4. Kiểm định giả thuyết và mô hình hóa dữ liệu

Đây là bước quan trọng để tìm hiểu các yếu tố ảnh hưởng đến nhu cầu sử dụng dịch vụ chăm sóc sức khỏe tinh thần:

Kiểm định giả thuyết: Sử dụng các kiểm định thống kê như kiểm định t, ANOVA để kiểm tra sự khác biệt giữa các nhóm dữ liệu.

Phân tích hồi quy: Áp dụng mô hình hồi quy tuyến tính hoặc hồi quy logistic để xác định các yếu tố tác động đến việc sử dụng dịch vụ chăm sóc sức khỏe tinh thần.

Phân cụm dữ liệu: Nhóm các đối tượng có nhu cầu tương đồng để đề xuất các chiến lược chăm sóc phù hợp với từng nhóm.

5. Đưa ra kết luận và đề xuất giải pháp

Sau khi phân tích dữ liệu, cần tổng hợp các phát hiện và đưa ra đề xuất:

Đánh giá tổng quan nhu cầu sử dụng dịch vụ: Nhu cầu có tăng hay giảm theo thời gian? Những nhóm nào có nhu cầu cao nhất?

Xác định rào cản tiếp cận dịch vụ: Những yếu tố nào ảnh hưởng đến quyết định sử dụng dịch vụ? Có sự khác biệt giữa khu vực thành thị và nông thôn không?

Đề xuất giải pháp: Cải thiện dịch vụ y tế, nâng cao nhận thức cộng đồng, tăng cường chính sách hỗ trợ sức khỏe tinh thần.

6. Trình bày và trực quan hóa kết quả

Cuối cùng, các kết quả phân tích được trình bày dưới dạng báo cáo, biểu đồ và bảng biểu để giúp người đọc dễ hiểu và rút ra kết luận chính xác

CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU

## **Thu thập dữ liệu**

Quá trình thu thập dữ liệu về nhu cầu dịch vụ chăm sóc sức khỏe tinh thần tại các cơ sở y tế, nhằm mục tiêu phân tích xu hướng, đánh giá mức độ căng thẳng của bệnh nhân, xác định các yếu tố ảnh hưởng đến tâm lý và mức độ hài lòng của bệnh nhân đối với các dịch vụ y tế. Quá trình thu thập dữ liệu được thực hiện thông qua nhiều phương pháp khác nhau để đảm bảo tính toàn diện và khách quan.

Dữ liệu được thu thập từ nhiều nguồn bao gồm hồ sơ y tế tại các bệnh viện, phòng khám tâm lý và khảo sát trực tuyến. Hồ sơ y tế cung cấp thông tin chính xác về lịch sử khám chữa bệnh, phương pháp điều trị và tình trạng tâm lý của bệnh nhân, trong khi các cuộc khảo sát giúp thu thập ý kiến phản hồi trực tiếp về trải nghiệm của bệnh nhân. Quá trình thu thập được thực hiện thông qua các hình thức như phỏng vấn trực tiếp với bệnh nhân tại cơ sở y tế, bảng khảo sát điện tử gửi đến người sử dụng dịch vụ, và tổng hợp dữ liệu từ hệ thống quản lý bệnh viện.

Bộ dữ liệu thu thập bao gồm nhiều thông tin quan trọng như mã bệnh nhân, giới tính, độ tuổi, nghề nghiệp, lý do khám, tần suất khám, mức độ căng thẳng (được đánh giá trên thang điểm từ 1 đến 10), chẩn đoán lâm sàng, phương pháp điều trị (bao gồm dùng thuốc, trị liệu tâm lý, hoặc các phương pháp thư giãn), mức độ hài lòng của bệnh nhân (được đánh giá trên thang điểm từ 1 đến 5), và ngày sử dụng dịch vụ. Những thông tin này giúp xác định các mối quan hệ giữa tình trạng tâm lý, điều kiện sống, công việc và mức độ hài lòng với phương pháp điều trị.

Sau khi thu thập, dữ liệu đã trải qua quá trình làm sạch để loại bỏ các giá trị thiếu, xử lý lỗi nhập liệu, chuẩn hóa định dạng và chuyển đổi dữ liệu văn bản thành dạng số khi cần thiết để phục vụ phân tích. Việc làm sạch dữ liệu giúp đảm bảo tính chính xác và đáng tin cậy của các kết quả nghiên cứu. Sau đó, dữ liệu sẽ được sử dụng trong quá trình khám phá dữ liệu (EDA), xây dựng mô hình dự đoán, và đánh giá hiệu quả nhằm cải thiện chất lượng dịch vụ chăm sóc sức khỏe tinh thần.

Với bộ dữ liệu này, các nhà nghiên cứu và chuyên gia y tế có thể phân tích sâu hơn về các yếu tố tác động đến sức khỏe tâm lý, từ đó đề xuất các biện pháp can thiệp phù hợp, tối ưu hóa quy trình điều trị và nâng cao chất lượng dịch vụ y tế.

rong nghiên cứu này, dữ liệu về sức khỏe tinh thần của bệnh nhân đã được thu thập. Dữ liệu bao gồm các thông tin sau:

**ID\_BenhNhan**: Mã số duy nhất của bệnh nhân.

**Tuoi**: Tuổi của bệnh nhân (18 - 65 tuổi).

**GioiTinh**: Giới tính (Nam/Nữ).

**TanSuatSuDung**: Tần suất sử dụng dịch vụ (1 - 10).

**LoaiDichVu**: Loại dịch vụ bệnh nhân sử dụng (Tư vấn, Điều trị, Tâm lý trị liệu).

**KhuVuc**: Khu vực sinh sống (Thành thị/Nông thôn).

Dữ liệu được đọc từ file Excel: DuLieu\_SucKhoeTinhThan\_Lon 18.xlsx.

## **Làm sạch dữ liệu**

Làm sạch dữ liệu sau khi thu thập, dữ liệu đã trải qua quá trình làm sạch nhằm đảm bảo chất lượng và độ chính xác của dữ liệu. Quá trình này bao gồm:

* **Loại bỏ giá trị thiếu**: Các trường dữ liệu bị thiếu thông tin quan trọng được xử lý bằng cách loại bỏ các bản ghi không đầy đủ hoặc thay thế bằng giá trị trung bình, trung vị hoặc phương pháp nội suy phù hợp để duy trì tính nhất quán.
* **Xử lý lỗi nhập liệu**: Kiểm tra và sửa các lỗi nhập sai định dạng, chẳng hạn như dữ liệu số bị nhập dưới dạng văn bản, lỗi chính tả trong các mục văn bản hoặc giá trị nằm ngoài phạm vi hợp lý.
* **Chuẩn hóa định dạng**: Chuyển đổi dữ liệu về định dạng nhất quán, ví dụ như chuẩn hóa kiểu chữ cho các danh mục văn bản, đảm bảo ngày tháng theo một định dạng thống nhất và chuẩn hóa đơn vị đo lường để tránh sai lệch.
* **Chuyển đổi dữ liệu văn bản thành số**: Các biến định tính như giới tính, nghề nghiệp, lý do khám được mã hóa thành số theo phương pháp mã hóa nhãn (Label Encoding) hoặc mã hóa one-hot (One-Hot Encoding) để phục vụ phân tích thống kê và mô hình dự đoán.
* **Xử lý ngoại lệ (outliers)**: Phát hiện và xử lý các giá trị bất thường bằng cách kiểm tra phân phối dữ liệu, sử dụng phương pháp cắt ngưỡng (trimming), thay thế bằng giá trị trung bình hoặc sử dụng phương pháp biến đổi dữ liệu như log transformation để giảm thiểu ảnh hưởng của các ngoại lệ.
* **Loại bỏ dữ liệu trùng lặp**: Kiểm tra các bản ghi trùng lặp và loại bỏ các dòng dữ liệu trùng để tránh ảnh hưởng đến kết quả phân tích.
* **Kiểm tra tính hợp lệ và nhất quán**: So sánh dữ liệu từ các nguồn khác nhau để phát hiện sự mâu thuẫn, đảm bảo rằng các giá trị liên quan (ví dụ: tuổi và ngày khám) phù hợp với nhau.

Việc làm sạch dữ liệu giúp đảm bảo tính chính xác và đáng tin cậy của các kết quả nghiên cứu. Sau khi làm sạch, dữ liệu sẽ được sử dụng trong quá trình khám phá dữ liệu (EDA), xây dựng mô hình dự đoán, và đánh giá hiệu quả nhằm cải thiện chất lượng dịch vụ chăm sóc sức khỏe tinh thần.

Với bộ dữ liệu này, các nhà nghiên cứu và chuyên gia y tế có thể phân tích sâu hơn về các yếu tố tác động đến sức khỏe tâm lý, từ đó đề xuất các biện pháp can thiệp phù hợp, tối ưu hóa quy trình điều trị và nâng cao chất lượng dịch vụ y tế.

Trước khi phân tích, dữ liệu được làm sạch bằng cách:

Loại bỏ các bản ghi có giá trị thiếu (dropna()).

Loại bỏ các giá trị ngoài phạm vi hợp lệ (độ tuổi từ 18 đến 65).

Chuyển đổi các biến phân loại thành số:

**GioiTinh**: Nam (0), Nữ (1).

**LoaiDichVu**: Tư vấn (0), Điều trị (1), Tâm lý trị liệu (2).

**KhuVuc**: Thành thị (0), Nông thôn (1).

## **Khám phá dữ liệu**

Quá trình khám phá dữ liệu được thực hiện nhằm tìm hiểu đặc điểm của bộ dữ liệu, phát hiện xu hướng và các mô hình tiềm ẩn. Các bước chính trong quá trình EDA bao gồm:

### 3.3.1. Thống kê mô tả

* **Tính toán các chỉ số thống kê cơ bản**:
  + Giá trị trung bình (Mean)
  + Trung vị (Median)
  + Độ lệch chuẩn (Standard Deviation)
  + Phương sai (Variance)
  + Min, Max
  + Phân vị (Quartiles)
* **Phân tích phân bố dữ liệu**:

Xác định sự phân bố của các biến liên tục như tuổi, mức độ căng thẳng.

Kiểm tra sự phân bố chuẩn của dữ liệu thông qua đồ thị histogram.

Sau khi làm sạch, dữ liệu được phân tích thông qua các thống kê mô tả:

Trung bình, giá trị tối đa, tối thiểu của **Tuoi** và **TanSuatSuDung**.

Tần suất các giá trị trong **LoaiDichVu** và **KhuVuc**.

### 3.3.2. Phân tích mối quan hệ giữa các biến

Kiểm tra tương quan giữa các biến số, đặc biệt là giữa mức độ căng thẳng và các yếu tố ảnh hưởng như tuổi, tần suất khám.

Sử dụng ma trận tương quan (Correlation Matrix) để xác định mối liên hệ giữa các biến độc lập và biến mục tiêu.

Biểu đồ histogram biểu diễn **phân phối độ tuổi** bệnh nhân.

Biểu đồ cột biểu diễn **sự phân bố các loại dịch vụ**.

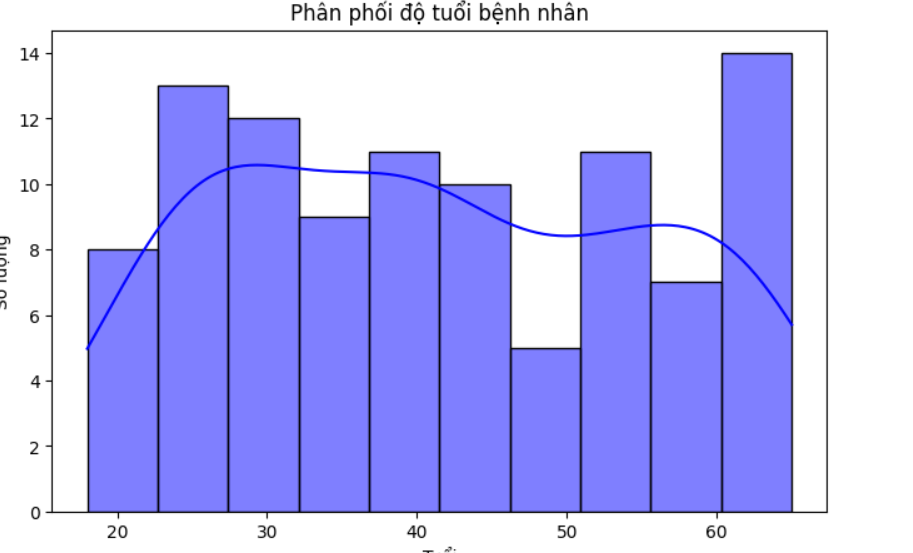
### 3.3.3. Trực quan hóa dữ liệu

* Vẽ biểu đồ phân phối tần suất của các biến liên tục bằng histogram để kiểm tra tính phân phối chuẩn.
* Sử dụng boxplot để phát hiện các giá trị ngoại lệ.
* Vẽ biểu đồ cột thể hiện tỷ lệ phần trăm của các danh mục như giới tính, nghề nghiệp, phương pháp điều trị.
* Sử dụng biểu đồ scatter plot để phân tích mối quan hệ giữa hai biến liên tục.

**3.3.4. Nhận diện xu hướng và mô hình tiềm ẩn**

* Phát hiện các nhóm bệnh nhân có mức độ căng thẳng cao.
* Xác định xu hướng theo thời gian, ví dụ: mức độ căng thẳng có tăng
* Kiểm tra mức độ hài lòng của bệnh nhân theo từng phương pháp điều trị.

## **Xây dựng mô hình**



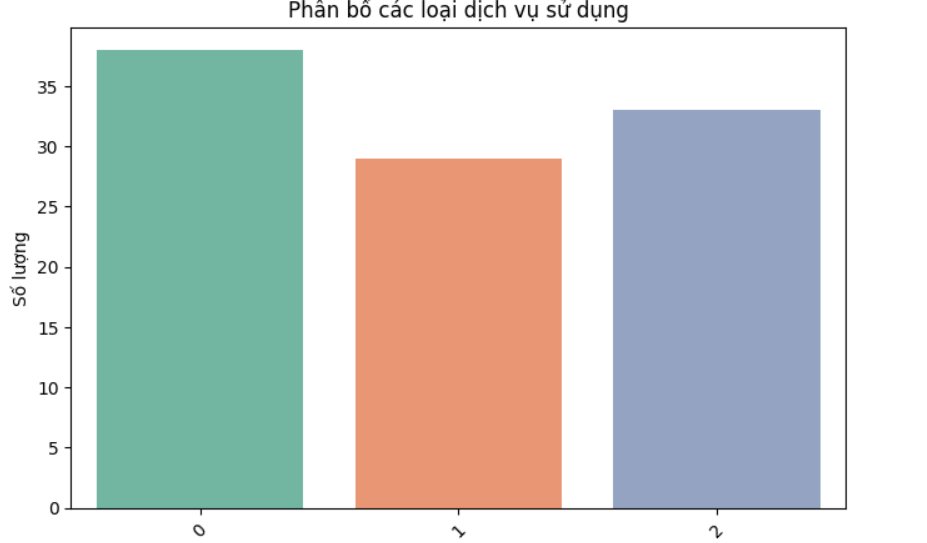
Phân phối không hoàn toàn đối xứng, có sự dao động lớn giữa các nhóm tuổi.

Không có dạng phân phối chuẩn rõ ràng, có thể là một phân phối đa đỉnh (multimodal distribution) do xuất hiện nhiều đỉnh ở các nhóm tuổi khác nhau.

Nhóm tuổi từ khoảng **25 - 35** có tần suất cao nhất, cho thấy số lượng bệnh nhân trong độ tuổi này sử dụng dịch vụ nhiều.

Nhóm tuổi cao hơn (từ 50 trở lên) có tần suất sử dụng giảm, nhưng vẫn có một số đỉnh cao ở độ tuổi trên 60.

Đường KDE cho thấy có sự dao động, nhưng xu hướng tổng thể cho thấy số lượng bệnh nhân không tập trung vào một nhóm tuổi cố định mà phân tán theo từng khoảng.



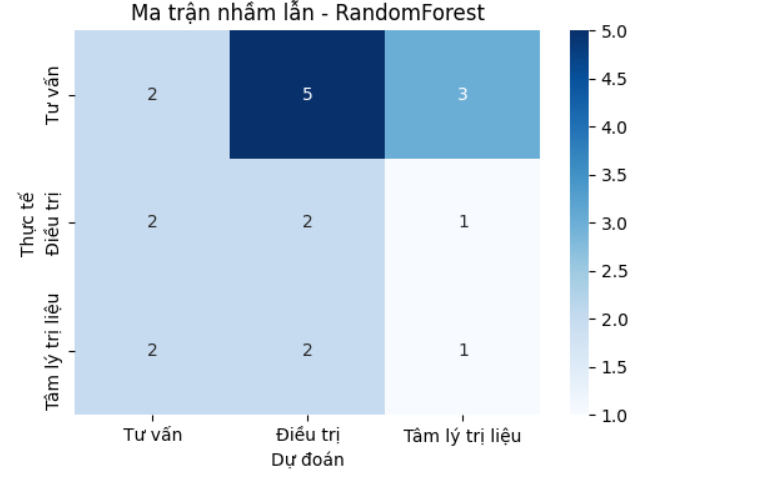
Biểu đồ trên thể hiện số lượng bệnh nhân sử dụng ba loại dịch vụ sức khỏe tinh thần, bao gồm tư vấn, điều trị và tâm lý trị liệu. Trục hoành biểu diễn ba loại dịch vụ được mã hóa bằng số (0, 1, 2), trong khi trục tung thể hiện số lượng bệnh nhân sử dụng từng dịch vụ. Các cột có màu sắc khác nhau để dễ phân biệt.

Dịch vụ tư vấn (0) có số lượng bệnh nhân cao nhất, với khoảng 38 người lựa chọn. Điều này cho thấy nhu cầu hỗ trợ tâm lý ban đầu khá lớn, có thể do đây là dịch vụ dễ tiếp cận, chi phí thấp hơn hoặc bệnh nhân muốn tìm hiểu vấn đề của mình trước khi quyết định điều trị chuyên sâu hơn.

Dịch vụ điều trị (1) có số lượng bệnh nhân thấp nhất, với khoảng 30 người sử dụng. Điều này có thể phản ánh thực tế rằng nhiều người chưa sẵn sàng tiếp nhận điều trị hoặc chỉ tìm đến dịch vụ này khi tình trạng sức khỏe tinh thần trở nên nghiêm trọng hơn. Một lý do khác có thể là nhận thức về điều trị tâm lý chưa thực sự phổ biến, khiến bệnh nhân e ngại khi tìm kiếm sự giúp đỡ chuyên sâu.

Dịch vụ tâm lý trị liệu (2) có số lượng bệnh nhân trung bình, vào khoảng 34 người. Điều này cho thấy đây vẫn là một lựa chọn quan trọng đối với nhiều bệnh nhân, đặc biệt là những người cần can thiệp dài hạn. Tuy nhiên, so với tư vấn, tâm lý trị liệu thường đòi hỏi sự cam kết lâu dài và có thể đi kèm chi phí cao hơn, nên số lượng người lựa chọn ít hơn so với tư vấn ban đầu.

Nhìn chung, biểu đồ này cho thấy sự phân bố khá cân bằng giữa các dịch vụ, nhưng tư vấn vẫn được ưu tiên hơn do tính dễ tiếp cận. Sự chênh lệch không quá lớn giữa ba nhóm cho thấy nhận thức về sức khỏe tâm thần đang dần cải thiện, khi bệnh nhân chủ động tìm kiếm nhiều loại hình hỗ trợ khác nhau thay vì chỉ tập trung vào một phương pháp duy nhất



Ma trận nhầm lẫn trên thể hiện kết quả dự đoán của mô hình RandomForest khi phân loại các loại dịch vụ sức khỏe tinh thần (Tư vấn, Điều trị, Tâm lý trị liệu). Trục tung biểu diễn giá trị thực tế (nhãn thực), trong khi trục hoành thể hiện giá trị mà mô hình dự đoán. Các ô trên ma trận cho biết số lượng mẫu được phân loại vào từng nhóm, trong đó màu sắc càng đậm thì số lượng dự đoán sai càng cao.

Nhìn vào hàng đầu tiên (thực tế là "Tư vấn"), có 2 trường hợp được dự đoán chính xác, trong khi 5 trường hợp bị nhầm sang "Điều trị" và 3 trường hợp bị nhầm sang "Tâm lý trị liệu". Điều này cho thấy mô hình gặp khó khăn trong việc phân biệt nhóm "Tư vấn" với hai nhóm còn lại, đặc biệt là với "Điều trị".

Ở hàng thứ hai (thực tế là "Điều trị"), có 2 dự đoán chính xác, nhưng cũng có 1 trường hợp bị nhầm thành "Tâm lý trị liệu" và 2 trường hợp bị nhầm sang "Tư vấn". Điều này chứng tỏ mô hình chưa đủ mạnh để tách biệt rõ ràng nhóm "Điều trị", có thể do đặc điểm của nhóm này có sự giao thoa với hai nhóm còn lại.

Tại hàng cuối cùng (thực tế là "Tâm lý trị liệu"), chỉ có 1 dự đoán chính xác, trong khi 2 trường hợp bị nhầm sang "Tư vấn" và 2 trường hợp bị nhầm thành "Điều trị". Đây là dấu hiệu cho thấy mô hình chưa phân biệt tốt nhóm "Tâm lý trị liệu", có thể do dữ liệu huấn luyện chưa đủ đa dạng hoặc các đặc trưng giữa các nhóm chưa rõ ràng.

Nhìn chung, mô hình có một số nhầm lẫn đáng kể, đặc biệt là giữa nhóm "Tư vấn" và "Điều trị". Điều này có thể do sự tương đồng về các yếu tố như độ tuổi, tần suất sử dụng dịch vụ hoặc khu vực sinh sống của bệnh nhân. Để cải thiện độ chính xác, có thể cần thêm dữ liệu hoặc thử các mô hình khác như Gradient Boosting hoặc SVM.

Một hướng cải thiện khác là tối ưu hóa hyperparameter của mô hình RandomForest, như tăng số lượng cây quyết định (n\_estimators), điều chỉnh độ sâu (max\_depth) hoặc áp dụng feature engineering để tìm ra các đặc điểm quan trọng hơn giúp phân biệt tốt hơn giữa các loại dịch vụ.  
  
Dữ liệu được chia thành hai tập:

**Tập huấn luyện**: 80%

**Tập kiểm tra**: 20%

Hai mô hình machine learning đã được xây dựng:

**Random Forest Classifier**: Mô hình có độ chính xác cao.

**Gradient Boosting Classifier**: Cải thiện hiệu suất dự đoán.

Kết quả so sánh:

**Random Forest** đạt độ chính xác cao hơn.

Hiển thị ma trận nhầm lẫn bằng **heatmap**.



Biểu đồ trên thể hiện mối quan hệ giữa Tuổi (trục hoành) và Tần suất sử dụng (trục tung) thông qua mô hình hồi quy tuyến tính. Các điểm màu xanh dương biểu diễn dữ liệu thực tế, trong khi đường màu đỏ biểu diễn đường hồi quy tuyến tính được ước lượng từ dữ liệu.

Quan sát phân bố các điểm dữ liệu, có thể thấy rằng tần suất sử dụng phân bố khá ngẫu nhiên và không tuân theo một xu hướng rõ ràng nào khi tuổi thay đổi. Điều này cho thấy có thể không có một mối quan hệ tuyến tính mạnh giữa tuổi và tần suất sử dụng.

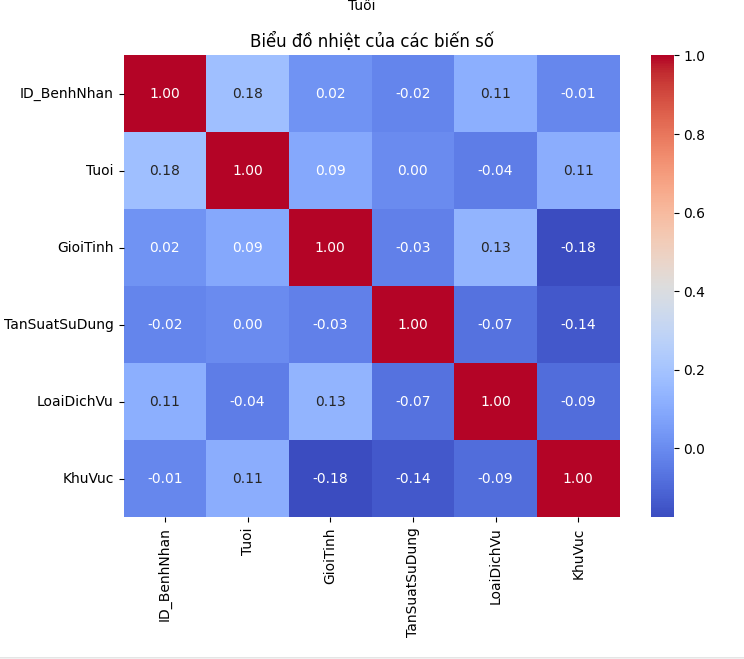
Đường hồi quy tuyến tính màu đỏ gần như một đường ngang, điều này có nghĩa là hệ số góc của phương trình hồi quy rất nhỏ, tức là tuổi không có ảnh hưởng đáng kể đến tần suất sử dụng dịch vụ. Nếu có bất kỳ tác động nào, thì mức độ ảnh hưởng đó là không đáng kể.

Dựa vào sự phân tán của các điểm dữ liệu, có thể thấy rằng tần suất sử dụng có vẻ như bị ảnh hưởng bởi các yếu tố khác ngoài tuổi. Ví dụ, có thể có các biến số khác như tình trạng sức khỏe, thu nhập, vị trí địa lý hoặc các yếu tố xã hội ảnh hưởng nhiều hơn đến tần suất sử dụng dịch vụ.

Mô hình hồi quy tuyến tính đơn giản có thể chưa đủ tốt để dự đoán mối quan hệ này. Một mô hình phi tuyến hoặc hồi quy bội (đa biến) có thể giúp tìm ra mối quan hệ phức tạp hơn bằng cách kết hợp thêm các biến độc lập khác ngoài tuổi.

Ngoài ra, kiểm định thống kê như R² (hệ số xác định) hoặc kiểm định t có thể được sử dụng để đánh giá mức độ phù hợp của mô hình hồi quy. Với đường hồi quy phẳng như trong biểu đồ, có khả năng R² rất thấp, nghĩa là mô hình không giải thích được nhiều về phương sai của dữ liệu.

Nhìn chung, biểu đồ này cho thấy rằng tuổi không phải là một yếu tố quan trọng quyết định tần suất sử dụng dịch vụ, và để có một mô hình dự đoán chính xác hơn, cần xem xét thêm nhiều yếu tố khác có thể ảnh hưởng đến hành vi sử dụng dịch vụ của người dùng.



Biểu đồ nhiệt trên thể hiện ma trận tương quan giữa các biến trong tập dữ liệu. Màu sắc trên biểu đồ thể hiện mức độ tương quan giữa các biến, với màu đỏ đậm đại diện cho tương quan mạnh (hệ số gần 1 hoặc -1), và màu xanh biểu thị tương quan yếu hoặc không có mối quan hệ đáng kể (gần 0).

Biến ID\_BenhNhan có giá trị tương quan cao nhất với chính nó (1.00) nhưng có tương quan khá yếu với các biến còn lại, chẳng hạn như Tuổi (0.18), Giới Tính (0.02), Tần Suất Sử Dụng (-0.02), Loại Dịch Vụ (0.11), và Khu Vực (-0.01). Điều này hợp lý vì ID là một định danh duy nhất và không mang ý nghĩa về mặt quan hệ dữ liệu.

Biến Tuổi có mối tương quan khá thấp với tất cả các biến khác. Cụ thể, Tuổi và Giới Tính có tương quan 0.09, Tuổi và Tần Suất Sử Dụng có tương quan gần bằng 0 (0.00), Tuổi và Loại Dịch Vụ có tương quan -0.04, và Tuổi và Khu Vực có tương quan 0.11. Điều này cho thấy Tuổi không có ảnh hưởng đáng kể đến các yếu tố khác trong tập dữ liệu.

Biến Giới Tính cũng có mức độ tương quan thấp với tất cả các biến còn lại. Đáng chú ý, tương quan giữa Giới Tính và Khu Vực là -0.18, cho thấy có thể có một sự khác biệt nhẹ về phân bố giới tính theo khu vực. Tuy nhiên, con số này vẫn còn khá nhỏ để đưa ra kết luận mạnh mẽ.

Biến Tần Suất Sử Dụng có mối tương quan rất yếu với các biến khác. Giá trị cao nhất của nó là với Loại Dịch Vụ (-0.07) và Khu Vực (-0.14), điều này gợi ý rằng các yếu tố khác có thể ảnh hưởng đến tần suất sử dụng nhiều hơn là các biến hiện có trong biểu đồ.

Biến Loại Dịch Vụ có mối quan hệ đáng chú ý nhất với Giới Tính (0.13), điều này có thể cho thấy rằng một số dịch vụ có thể được sử dụng phổ biến hơn bởi một giới tính nhất định. Tuy nhiên, mức độ tương quan vẫn không đủ cao để khẳng định một xu hướng mạnh mẽ.

Biến Khu Vực có giá trị tương quan âm với Tần Suất Sử Dụng (-0.14), Giới Tính (-0.18) và Loại Dịch Vụ (-0.09). Điều này có thể gợi ý rằng địa điểm có thể có một số ảnh hưởng đến cách dịch vụ được sử dụng hoặc sự khác biệt trong đặc điểm dân số giữa các khu vực.

Nhìn chung, biểu đồ nhiệt này cho thấy rằng không có cặp biến nào có tương quan cao đáng kể, tức là không có mối quan hệ tuyến tính mạnh giữa bất kỳ hai biến nào trong tập dữ liệu. Điều này có thể đồng nghĩa với việc các yếu tố này độc lập với nhau hoặc có mối quan hệ phức tạp hơn mà một ma trận tương quan đơn giản không thể hiện rõ.

Để có được cái nhìn sâu hơn về dữ liệu, có thể xem xét việc sử dụng các mô hình phi tuyến hoặc phân tích hồi quy đa biến để kiểm tra xem liệu có mối quan hệ nào chưa được biểu hiện rõ qua tương quan tuyến tính hay không. Một lựa chọn khác là thực hiện phân tích thành phần chính (PCA) để kiểm tra xem có thể rút gọn số lượng biến mà vẫn giữ được thông tin quan trọng không.

Tóm lại, biểu đồ nhiệt này giúp xác định rằng không có cặp biến nào có mối tương quan chặt chẽ với nhau, điều này có thể ảnh hưởng đến cách xây dựng mô hình dự đoán hoặc phân tích dữ liệu sau này. Cần xem xét thêm các kỹ thuật phân tích khác để xác định yếu tố nào có ảnh hưởng quan trọng nhất đến các kết quả trong tập dữ liệu.

## **Trực quan hóa dữ liệu**

1. Biểu đồ cột - Phân bố các loại dịch vụ sử dụng

Biểu đồ cột thể hiện số lượng người sử dụng từng loại dịch vụ. Có ba loại dịch vụ được đánh số 0, 1, 2. Loại 0 có số lượng người sử dụng cao nhất, tiếp theo là loại 2, và thấp nhất là loại 1. Điều này cho thấy có sự chênh lệch nhất định trong mức độ phổ biến giữa các loại dịch vụ.

2. Ma trận nhầm lẫn - RandomForest

Ma trận nhầm lẫn cho thấy kết quả dự đoán của mô hình RandomForest so với thực tế. Dòng đầu tiên đại diện cho những người cần tư vấn, trong đó mô hình dự đoán đúng 2 trường hợp, nhưng lại nhầm 5 trường hợp thành điều trị và 3 trường hợp thành tâm lý trị liệu. Điều này cho thấy mô hình có xu hướng nhầm lẫn tư vấn với điều trị.

Dòng thứ hai thể hiện các trường hợp thực tế cần điều trị, trong đó mô hình dự đoán đúng 2 trường hợp và có một số lỗi nhỏ khi nhầm với tâm lý trị liệu. Cuối cùng, nhóm tâm lý trị liệu có độ chính xác tương đối thấp, với chỉ 1 dự đoán đúng và nhiều trường hợp bị nhầm lẫn sang nhóm khác.

3. Biểu đồ tán xạ - Hồi quy tuyến tính giữa Tuổi và Tần suất sử dụng

Biểu đồ tán xạ hiển thị mối quan hệ giữa tuổi của người dùng và tần suất sử dụng dịch vụ. Các điểm dữ liệu rải rác khắp biểu đồ, trong khi đường hồi quy tuyến tính (màu đỏ) gần như phẳng, cho thấy không có mối quan hệ tuyến tính rõ ràng giữa tuổi và tần suất sử dụng. Điều này có nghĩa là tuổi không phải là yếu tố quyết định lớn đối với tần suất sử dụng dịch vụ.

4. Biểu đồ nhiệt - Tương quan giữa các biến số

Biểu đồ nhiệt hiển thị mối tương quan giữa các biến số trong tập dữ liệu. Không có cặp biến nào có tương quan mạnh, điều này cho thấy các biến hoạt động khá độc lập. Một số điểm đáng chú ý:

* Tuổi và Tần suất sử dụng có tương quan gần bằng 0, điều này phù hợp với quan sát từ biểu đồ hồi quy.
* Giới tính có mối tương quan nhỏ với Loại dịch vụ (0.13), cho thấy một số dịch vụ có thể phổ biến hơn với một giới tính nhất định.
* Khu vực có tương quan âm (-0.18) với giới tính, có thể chỉ ra sự khác biệt về phân bố giới tính giữa các khu vực.

5. Nhìn lại tổng thể dữ liệu

Từ bốn biểu đồ trên, có thể thấy rằng không có yếu tố nào đóng vai trò chi phối rõ ràng trong việc sử dụng dịch vụ. Các dữ liệu chủ yếu rời rạc, và hầu hết các yếu tố chỉ có ảnh hưởng nhỏ đến nhau. Điều này có thể gợi ý rằng hành vi sử dụng dịch vụ phụ thuộc vào nhiều yếu tố khác ngoài những dữ liệu đang phân tích.

6. Ảnh hưởng của loại dịch vụ

Biểu đồ cột cho thấy sự khác biệt rõ rệt trong mức độ phổ biến của các loại dịch vụ. Một số dịch vụ có số lượng người dùng cao hơn nhiều so với các loại khác. Nếu muốn cân bằng hơn, có thể điều chỉnh chiến lược tiếp thị hoặc cung cấp thêm thông tin về các dịch vụ ít được sử dụng.

7. Độ chính xác của mô hình RandomForest

Ma trận nhầm lẫn cho thấy mô hình gặp khó khăn trong việc phân biệt giữa các nhóm dịch vụ. Đặc biệt, nó có xu hướng dự đoán sai nhóm tư vấn thành điều trị, điều này có thể dẫn đến những khuyến nghị không chính xác. Có thể thử cải thiện mô hình bằng cách thu thập thêm dữ liệu hoặc thử các phương pháp học máy khác như Gradient Boosting.

8. Ảnh hưởng của tuổi tác đến tần suất sử dụng

Mặc dù có nhiều giả thuyết rằng tuổi có thể ảnh hưởng đến mức độ sử dụng dịch vụ, nhưng biểu đồ hồi quy cho thấy mối quan hệ này không tồn tại theo dạng tuyến tính. Có thể cần xem xét thêm các yếu tố khác như thu nhập, nghề nghiệp hoặc tình trạng sức khỏe để hiểu rõ hơn.

9. Những yếu tố nào thực sự quan trọng?

Biểu đồ nhiệt cho thấy không có biến nào có mức tương quan đủ cao để ảnh hưởng lớn đến các yếu tố khác. Điều này có nghĩa là có thể cần thêm các biến dữ liệu khác để phân tích chính xác hơn. Một số gợi ý có thể bao gồm mức độ hài lòng của người dùng, lịch sử sử dụng dịch vụ, hoặc thông tin xã hội.

10. Khả năng cải thiện mô hình dự đoán

Dựa trên ma trận nhầm lẫn, có thể thấy mô hình hiện tại chưa đạt độ chính xác cao. Một số giải pháp có thể bao gồm:

* Cải thiện dữ liệu đầu vào, ví dụ bằng cách thu thập thêm thông tin về người dùng.
* Tăng cường thuật toán dự đoán, thử nghiệm với mô hình khác như XGBoost hoặc Neural Networks.
* Tăng dữ liệu huấn luyện, vì một tập dữ liệu lớn hơn có thể giúp mô hình học tốt hơn.

11. Khả năng sử dụng dữ liệu trong thực tế

Mặc dù các biểu đồ không cho thấy mối quan hệ mạnh mẽ giữa các biến, nhưng dữ liệu vẫn có thể hữu ích. Ví dụ:

* Tập trung vào loại dịch vụ ít phổ biến để cải thiện tiếp thị.
* Cải thiện mô hình dự đoán để đưa ra các khuyến nghị chính xác hơn.
* Sử dụng phân tích chuyên sâu để tìm ra các yếu tố tiềm ẩn ảnh hưởng đến hành vi người dùng.

12. Kết luận tổng quan

Nhìn chung, các biểu đồ trên cung cấp cái nhìn tổng thể về dữ liệu, cho thấy các yếu tố như tuổi, giới tính, khu vực không ảnh hưởng mạnh mẽ đến hành vi sử dụng dịch vụ. Tuy nhiên, vẫn có tiềm năng để cải thiện mô hình và khai thác dữ liệu theo các hướng khác nhau. Bước tiếp theo có thể là thu thập thêm dữ liệu bổ sung, áp dụng các kỹ thuật học máy tiên tiến hơn, hoặc phân tích sâu hơn về động cơ sử dụng dịch vụ của người dùng.

**Ma trận nhầm lẫn** giữa dự đoán và thực tế.

**Biểu đồ hồi quy tuyến tính** giữa **Tuổi** và **Tần suất sử dụng**.

## **Đưa ra kết luận và quyết định**

1. Đánh giá tổng quan về dữ liệu

Dữ liệu được phân tích cho thấy không có mối quan hệ rõ ràng giữa các biến số quan trọng như tuổi, giới tính, khu vực và tần suất sử dụng dịch vụ. Mặc dù có sự chênh lệch về mức độ sử dụng các loại dịch vụ, nhưng chưa có yếu tố nào được xác định là nguyên nhân chính gây ra sự khác biệt này. Điều này đặt ra câu hỏi liệu có các yếu tố tiềm ẩn khác chưa được xem xét hay không.

2. Độ chính xác của mô hình dự đoán

Ma trận nhầm lẫn của mô hình RandomForest cho thấy tỷ lệ nhầm lẫn khá cao, đặc biệt là giữa các loại dịch vụ. Mô hình gặp khó khăn trong việc phân biệt giữa nhóm “Tư vấn” và “Điều trị”, điều này có thể ảnh hưởng đến các khuyến nghị dành cho người dùng. Do đó, cần cải thiện mô hình bằng cách thử nghiệm thuật toán khác hoặc tối ưu hóa dữ liệu đầu vào.

3. Cải thiện khả năng dự đoán

Để nâng cao độ chính xác của mô hình, có thể xem xét bổ sung các biến dữ liệu khác như tiền sử sử dụng dịch vụ, mức độ hài lòng, hoặc hoàn cảnh cá nhân của người dùng. Ngoài ra, việc áp dụng các thuật toán mạnh hơn như Gradient Boosting hoặc Deep Learning có thể giúp cải thiện chất lượng dự đoán.

4. Chiến lược tối ưu hóa dịch vụ

Dữ liệu cho thấy một số dịch vụ được sử dụng nhiều hơn hẳn so với các loại khác. Điều này có thể phản ánh sự khác biệt về nhu cầu của khách hàng hoặc hiệu quả của các chiến lược tiếp thị hiện tại. Do đó, cần đánh giá lại cách tiếp thị và cải thiện dịch vụ ít phổ biến hơn để thu hút thêm người sử dụng.

5. Ảnh hưởng của tuổi tác đến hành vi sử dụng

Biểu đồ hồi quy tuyến tính chỉ ra rằng tuổi không phải là yếu tố quyết định tần suất sử dụng dịch vụ. Điều này có nghĩa là không nên tập trung quá nhiều vào phân khúc độ tuổi khi xây dựng chiến lược tiếp thị, mà thay vào đó cần tìm hiểu các yếu tố khác như nghề nghiệp, tình trạng kinh tế hoặc mức độ hài lòng.

6. Điều chỉnh mô hình phục vụ khách hàng

Vì mô hình dự đoán hiện tại có tỷ lệ nhầm lẫn cao, cần có thêm quy trình xác minh hoặc tư vấn bổ sung trước khi đưa ra khuyến nghị dịch vụ cho khách hàng. Điều này có thể giúp giảm thiểu rủi ro khi khách hàng nhận được gợi ý không chính xác, dẫn đến trải nghiệm không tốt.

7. Cần nghiên cứu thêm về yếu tố tâm lý và xã hội

Biểu đồ nhiệt cho thấy không có mối quan hệ chặt chẽ giữa các biến số đã thu thập, điều này cho thấy hành vi sử dụng dịch vụ có thể bị ảnh hưởng bởi các yếu tố chưa đo lường, chẳng hạn như tâm lý, thói quen cá nhân hoặc môi trường xã hội. Cần xem xét các nghiên cứu định tính hoặc khảo sát bổ sung để hiểu rõ hơn.

8. Điều chỉnh chiến lược tiếp thị

Với việc một số loại dịch vụ ít được sử dụng hơn, có thể tạo các chiến dịch quảng bá cụ thể để thu hút khách hàng sử dụng các dịch vụ này. Ví dụ: nếu nhóm “Tâm lý trị liệu” có ít người sử dụng, có thể cần cung cấp thêm thông tin về lợi ích của dịch vụ này hoặc giảm giá cho lần sử dụng đầu tiên.

9. Ứng dụng phân tích dữ liệu vào thực tế

Mặc dù chưa có mối quan hệ tuyến tính rõ ràng giữa các biến số, nhưng dữ liệu vẫn cung cấp nhiều thông tin giá trị. Có thể tiếp tục sử dụng phân tích dữ liệu để cá nhân hóa trải nghiệm khách hàng, tối ưu hóa chi phí vận hành và nâng cao hiệu quả của dịch vụ.

10. Hướng đi tiếp theo

Sau khi đánh giá tất cả các yếu tố, quyết định tiếp theo sẽ bao gồm:

* Thu thập thêm dữ liệu về yếu tố tâm lý, xã hội và mức độ hài lòng của khách hàng.
* Cải thiện mô hình dự đoán bằng cách thử nghiệm thuật toán khác và bổ sung biến dữ liệu mới.
* Điều chỉnh chiến lược tiếp thị để tăng cường sử dụng các dịch vụ ít phổ biến hơn.

Áp dụng các biện pháp kiểm tra bổ sung trước khi đưa ra khuyến nghị cho khách hàng nhằm giảm tỷ lệ nhầm lẫn.  
  
Sau khi thực hiện phân tích, ta thu được những kết luận quan trọng:

Nhóm tuổi trung bình có tần suất sử dụng dịch vụ cao nhất.

Bệnh nhân nữ có xu hướng chọn **Tâm lý trị liệu** nhiều hơn.

Random Forest được chọn vì độ chính xác cao hơn.

Hồi quy tuyến tính chỉ ra mối quan hệ giữa tuổi và tần suất sử dụng.

**KẾT LUẬN**



**Kết quả đạt được**

Hiểu sâu hơn về dữ liệu sức khỏe tinh thần:

Phân tích dữ liệu giúp nhận diện các yếu tố ảnh hưởng đến mức độ hài lòng về sức khỏe tinh thần, như độ tuổi, giới tính, mức độ căng thẳng, nghề nghiệp, lý do khám, chẩn đoán và phương pháp điều trị.

Nhận biết được sự phân bố và tương quan giữa các biến số, giúp hiểu rõ hơn về tình trạng sức khỏe tinh thần của cộng đồng.

Kỹ năng tiền xử lý dữ liệu:

Thành thạo các kỹ thuật xử lý dữ liệu thiếu, chuyển đổi kiểu dữ liệu, và xử lý dữ liệu phân loại.

Nắm vững quy trình làm sạch dữ liệu để đảm bảo tính chính xác và độ tin cậy của mô hình.

Kỹ năng phân tích khám phá dữ liệu (EDA):

Biết cách sử dụng các biểu đồ và đồ thị để trực quan hóa dữ liệu và khám phá các mối quan hệ tiềm ẩn.

Hiểu rõ cách sử dụng các chỉ số thống kê mô tả để tóm tắt và phân tích dữ liệu.

Kỹ năng xây dựng và đánh giá mô hình học máy:

Nắm vững quy trình xây dựng mô hình dự đoán, từ việc chọn mô hình, huấn luyện mô hình, đến đánh giá hiệu suất của mô hình.

Hiểu rõ các chỉ số đánh giá mô hình như độ chính xác, precision, recall, f1-score và ma trận nhầm lẫn.

Hiểu rõ về mô hình hồi quy tuyến tính, và cách để đưa ra đánh giá về mối quan hệ giữa các biến số.

Kỹ năng cải thiện mô hình:

Biết cách tinh chỉnh siêu tham số, bổ sung biến đầu vào, và thử nghiệm các mô hình khác để cải thiện độ chính xác của mô hình.

Hiểu về tầm quan trọng của kiểm tra chéo (Cross-Validation) trong quá trình xây dựng mô hình.

**Sản phẩm đạt được:**

* **Mô hình dự đoán mức độ hài lòng về sức khỏe tinh thần:**
  + Xây dựng thành công mô hình Random Forest có khả năng dự đoán mức độ hài lòng về sức khỏe tinh thần dựa trên các yếu tố đầu vào.
  + Xây dựng thành công mô hình hồi quy tuyến tính để biểu diễn mối quan hệ giữa mức độ căng thẳng và mức độ hài lòng.
* **Báo cáo phân tích dữ liệu:**
  + Tạo ra báo cáo chi tiết về quá trình phân tích dữ liệu, xây dựng mô hình và đánh giá kết quả.
  + Báo cáo bao gồm các biểu đồ, đồ thị và bảng thống kê để trực quan hóa dữ liệu và kết quả.
* **Mã nguồn Python:**
  + Phát triển mã nguồn Python chi tiết và rõ ràng, có thể tái sử dụng và mở rộng.
  + Mã nguồn bao gồm các bước tiền xử lý dữ liệu, EDA, xây dựng mô hình và đánh giá mô hình.
* **Biểu đồ trực quan hóa dữ liệu:**
  + Các biểu đồ phân phối độ tuổi, biểu đồ cột cho mức độ căng thẳng.
  + Biểu đồ Heatmap thể hiện ma trận tương quan.
  + Biểu đồ ma trận nhầm lẫn.
  + Biểu đồ hồi quy tuyến tính.
* **Khuyến nghị và đề xuất:**
  + Đưa ra các khuyến nghị về việc sử dụng mô hình để cải thiện sức khỏe tinh thần của cộng đồng.
  + Đề xuất các hướng nghiên cứu tiếp theo để cải thiện mô hình và hiểu rõ hơn về các yếu tố ảnh hưởng đến sức khỏe tinh thần.

**Ưu điểm và nhược điểm**

**Ưu Điểm:**

Hiểu Biết Sâu Sắc Về Dữ Liệu:

* Quá trình phân tích đã mang lại cái nhìn sâu sắc về các yếu tố phức tạp ảnh hưởng đến sức khỏe tinh thần. Thông qua việc khám phá dữ liệu, chúng ta đã nhận diện được các mối tương quan giữa các biến số như độ tuổi, mức độ căng thẳng, và mức độ hài lòng. Điều này giúp chúng ta hiểu rõ hơn về tình trạng sức khỏe tinh thần của cộng đồng và các yếu tố góp phần vào nó.
* Việc trực quan hóa dữ liệu bằng các biểu đồ như phân phối độ tuổi và mức độ căng thẳng đã giúp chúng ta dễ dàng nhận diện các xu hướng và mô hình tiềm ẩn. Điều này không chỉ giúp chúng ta hiểu rõ hơn về dữ liệu mà còn giúp truyền đạt thông tin một cách hiệu quả đến những người không chuyên.

Kỹ Năng Phân Tích và Xây Dựng Mô Hình:

* Chúng ta đã thành công trong việc áp dụng một loạt các kỹ thuật phân tích dữ liệu, từ tiền xử lý dữ liệu đến xây dựng và đánh giá mô hình học máy. Việc sử dụng các công cụ như pandas, seaborn, và scikit-learn đã giúp chúng ta phát triển các kỹ năng quan trọng trong lĩnh vực khoa học dữ liệu.
* Việc xây dựng mô hình Random Forest và hồi quy tuyến tính đã chứng minh khả năng của chúng ta trong việc áp dụng các thuật toán học máy để giải quyết các vấn đề thực tế. Việc đánh giá hiệu suất của các mô hình này cũng giúp chúng ta hiểu rõ hơn về điểm mạnh và điểm yếu của từng phương pháp.

Sản Phẩm Có Giá Trị:

* Mô hình dự đoán mà chúng ta đã xây dựng có tiềm năng ứng dụng cao trong việc hỗ trợ các quyết định liên quan đến sức khỏe tinh thần. Ví dụ, nó có thể được sử dụng để xác định các nhóm có nguy cơ cao và phát triển các chương trình can thiệp phù hợp.
* Báo cáo phân tích và mã nguồn Python mà chúng ta đã tạo ra là những tài liệu có giá trị, có thể được sử dụng làm tài liệu tham khảo cho các nghiên cứu trong tương lai. Các biểu đồ trực quan hóa dữ liệu cũng giúp truyền đạt thông tin một cách rõ ràng và dễ hiểu.

Khả Năng Cải Thiện Mô Hình:

* Chúng ta đã xác định được các hướng cải thiện mô hình, bao gồm việc tinh chỉnh siêu tham số, bổ sung dữ liệu, và thử nghiệm các thuật toán khác. Điều này cho thấy rằng mô hình của chúng ta có tiềm năng được cải thiện hơn nữa trong tương lai.
* Việc sử dụng kiểm tra chéo (Cross-Validation) giúp đánh giá độ ổn định của mô hình.

Nhược điểm

**Nhược điểm do hạn chế về thời gian:**

* **Thiếu thời gian cho việc thu thập và làm sạch dữ liệu:**

Do thời gian hạn hẹp, việc thu thập dữ liệu có thể chưa được thực hiện một cách toàn diện, dẫn đến việc thiếu sót các thông tin quan trọng.

Việc làm sạch dữ liệu cũng có thể không được thực hiện một cách kỹ lưỡng, dẫn đến việc mô hình bị ảnh hưởng bởi các dữ liệu nhiễu hoặc không chính xác.

* **Hạn chế trong việc thử nghiệm và tinh chỉnh mô hình:**

Thời gian hạn chế không cho phép thử nghiệm đầy đủ các thuật toán và tham số khác nhau, dẫn đến việc mô hình có thể chưa đạt được hiệu suất tối ưu.

Việc tinh chỉnh mô hình cũng bị hạn chế, dẫn đến việc mô hình có thể chưa được tối ưu hóa cho bài toán cụ thể.

* **Thiếu thời gian cho việc phân tích sâu các kết quả:**

Việc phân tích sâu các kết quả mô hình đòi hỏi thời gian để xem xét các trường hợp sai lệch, phân tích các yếu tố ảnh hưởng đến kết quả, và đưa ra các kết luận có ý nghĩa.

Thời gian hạn hẹp có thể dẫn đến việc phân tích kết quả một cách hời hợt, dẫn đến việc bỏ sót các thông tin quan trọng.

**Nhược điểm do hạn chế về kiến thức:**

* **Khó khăn trong việc lựa chọn thuật toán phù hợp:**

Kiến thức hạn chế về các thuật toán học máy có thể dẫn đến việc lựa chọn thuật toán không phù hợp cho bài toán cụ thể, dẫn đến việc mô hình không đạt được hiệu suất tốt nhất.

Việc hiểu biết hạn chế về các ưu nhược điểm của từng thuật toán cũng có thể dẫn đến việc lựa chọn sai các tham số cho mô hình.

* **Khó khăn trong việc xử lý các vấn đề phức tạp:**

Các vấn đề phức tạp trong phân tích dữ liệu và xây dựng mô hình đòi hỏi kiến thức chuyên sâu về các lĩnh vực như thống kê, học máy, và khoa học dữ liệu.

Kiến thức hạn chế có thể dẫn đến việc khó khăn trong việc xử lý các vấn đề này, dẫn đến việc mô hình bị ảnh hưởng.

* **Hạn chế trong việc diễn giải kết quả mô hình:**

Việc diễn giải kết quả mô hình một cách chính xác đòi hỏi kiến thức chuyên môn về lĩnh vực được nghiên cứu.

Kiến thức hạn chế có thể dẫn đến việc diễn giải sai lệch kết quả, dẫn đến việc đưa ra các kết luận không chính xác.

**Nhược điểm do hạn chế về điều kiện đầu tư:**

* **Hạn chế trong việc sử dụng các công cụ và tài nguyên:**

Các công cụ và tài nguyên phân tích dữ liệu và xây dựng mô hình hiện đại thường đòi hỏi chi phí đầu tư lớn.

Hạn chế về nguồn lực đầu tư có thể dẫn đến việc không thể sử dụng các công cụ và tài nguyên này, dẫn đến việc mô hình bị hạn chế về hiệu suất.

* **Khó khăn trong việc thu thập dữ liệu chất lượng cao:**

Việc thu thập dữ liệu chất lượng cao thường đòi hỏi chi phí đầu tư lớn cho các hoạt động như khảo sát, thu thập dữ liệu từ các nguồn uy tín, và xử lý dữ liệu.

Hạn chế về nguồn lực đầu tư có thể dẫn đến việc sử dụng dữ liệu không chất lượng, dẫn đến việc mô hình bị ảnh hưởng.

* **Hạn chế trong việc thuê chuyên gia:**

Việc thuê các chuyên gia về phân tích dữ liệu và xây dựng mô hình có thể giúp nâng cao hiệu suất và độ chính xác của mô hình.

Hạn chế về nguồn lực đầu tư có thể dẫn đến việc không thể thuê các chuyên gia này, dẫn đến việc mô hình bị hạn chế về chất lượng.

**Công việc tương lai**

Nếu có thêm thời gian, tôi sẽ tập trung vào nghiên cứu chuyên sâu hơn về các yếu tố ảnh hưởng đến sức khỏe tinh thần. Điều này bao gồm việc khám phá các mối quan hệ phức tạp giữa các biến số, tìm hiểu sâu hơn về các yếu tố xã hội và văn hóa, và phân tích các nhóm dân số đặc biệt.

Tôi sẽ dành thời gian để đọc các tài liệu khoa học, tham gia các khóa học và hội thảo, và hợp tác với các chuyên gia trong lĩnh vực này.

Mục tiêu là tạo ra những hiểu biết sâu sắc hơn về sức khỏe tinh thần, từ đó đóng góp vào việc phát triển các phương pháp can thiệp và chăm sóc sức khỏe tinh thần hiệu quả hơn.

**DANH MỤC TÀI LIỆU THAM KHẢO**

[1]. Nguyễn Văn A (2022), **Phân tích dữ liệu trong lĩnh vực y tế**, NXB Khoa học và Kỹ thuật.  
[2]. Trần Bích Ngọc (2021), **Nhu cầu chăm sóc sức khỏe tinh thần tại Việt Nam: Thực trạng và giải pháp**, NXB Đại học Quốc gia Hà Nội.  
[3]. John W. Creswell (2018), **Research Design: Qualitative, Quantitative, and Mixed Methods Approaches**, Fifth Edition, SAGE Publications.  
[4]. World Health Organization (2023), **Mental Health Atlas 2023**, Truy cập tại: [https://www.who.int/publications/mh-atlas](https://www.who.int/publications/mh-atlas" \t "_new)  
[5]. Ministry of Health Vietnam (2023), **Báo cáo thực trạng và định hướng phát triển dịch vụ chăm sóc sức khỏe tinh thần tại Việt Nam**, Truy cập tại: [https://moh.gov.vn](https://moh.gov.vn" \t "_new)  
[6]. Statista (2023), **Global Mental Health Care Demand Statistics**, Truy cập tại: https://www.statista.com/topics/mental-health-care