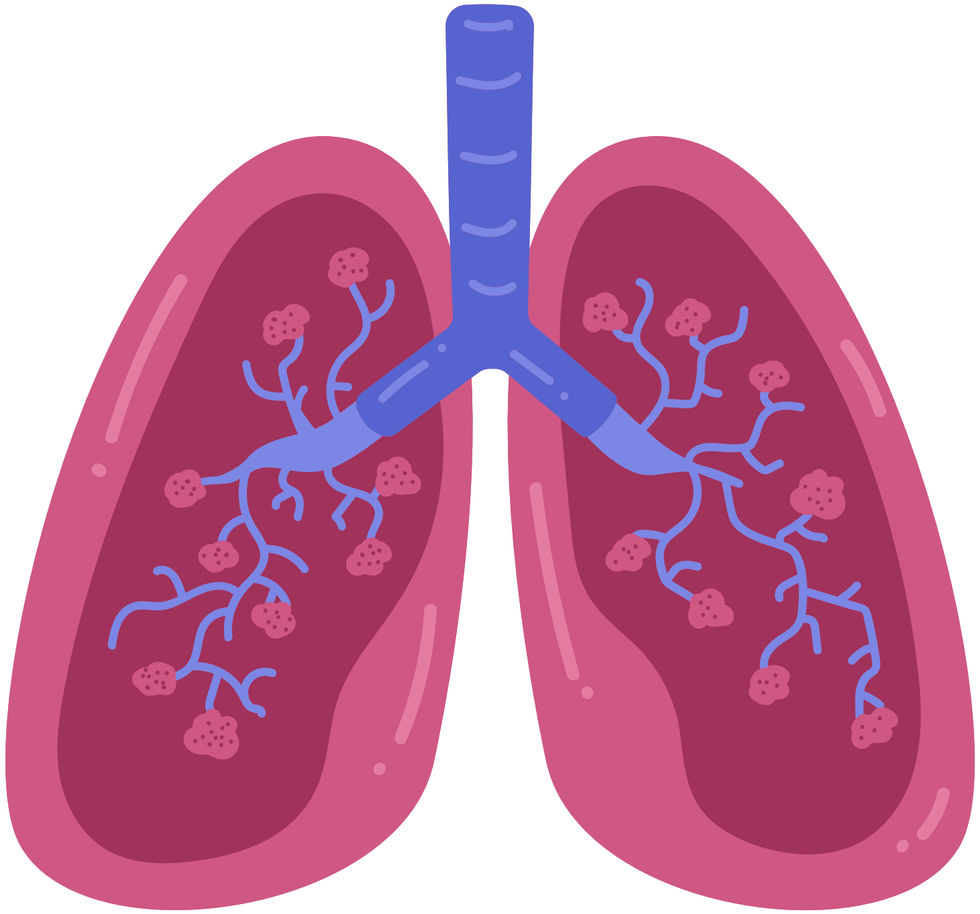
PHÂN TÍCH DỮ LIỆU BỆNH PHỔI



Nhóm 3 : 22KHDL
22127064 - Phạm Thành Đạt
22127225 - Trần Thị Thiên Kim
22127357 - Phạm Trần Yến Quyên
22127374 - Lê Thanh Tâm
22127449 - Mai Đức Vân

PHÂN TÍCH DỮ LIỆU BỆNH PHỔI

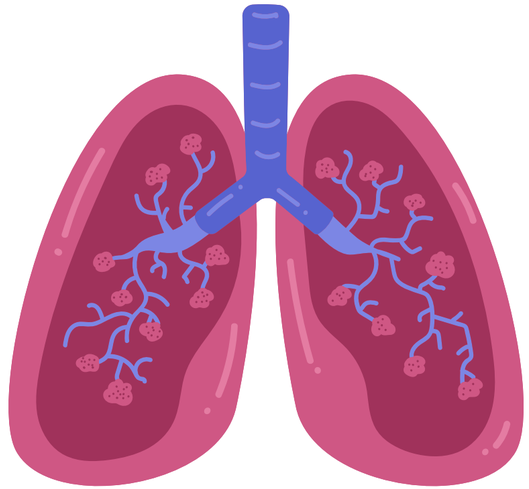
I. Giới thiệu dự án

II. Chuẩn bị dữ liệu

III. Tiền xử lý dữ liệu

IV. Cấu trúc Dashboard

V. Mở rộng và phát triển



I. GIỚI THIỆU DỰ ÁN

- Mục tiêu
- Công nghệ sử dụng

MỤC TIÊU

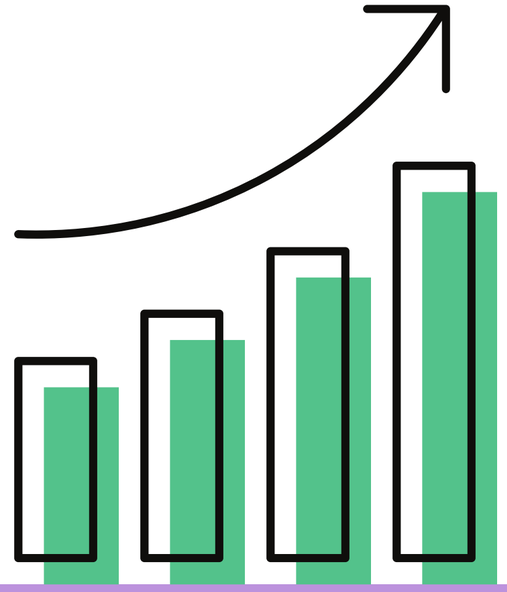
- Xây dựng một ứng dụng web dựa trên **Streamlit** để phân tích dữ liệu bệnh phổi.
- Giúp người dùng ***(bác sĩ, nhà nghiên cứu, bệnh nhân)*** dễ dàng ***khám phá dữ liệu***, hiểu xu hướng bệnh, và có thể dự đoán nguy cơ mắc bệnh phổi.
- Hỗ trợ trực quan hóa dữ liệu để dễ dàng phát hiện ***mối quan hệ giữa các chỉ số sức khỏe và bệnh phổi***.

CÔNG NGHỆ SỬ DỤNG

- **Ngôn ngữ:** Python
- **Thư viện chính**:
  - Streamlit – Xây dựng giao diện web
  - Pandas, NumPy – Xử lý dữ liệu
  - Matplotlib, Seaborn, Plotly – Trực quan hóa dữ liệu
  - sklearn.preprocessing – Chuẩn hóa dữ liệu

II. CHUẨN BỊ DỮ LIỆU

- **NGUỒN DỮ LIỆU**
  - Tổng quan dữ liệu
  - Cấu trúc dữ liệu



NGUỒN DỮ LIỆU

TỔNG QUAN VỀ DỮ LIỆU

1. Mô tả chung

- **Nguồn gốc:**
  - Tập dữ liệu được đăng tải trên nền tảng ***Kaggle*** vào ngày 25/2/2025 bởi người dùng *Samiksha Dalvi*.
  - Dữ liệu được thu thập tại ***3 trung tâm điều trị bệnh phổi*** tại khu vực sinh sống của tác giả và dữ liệu dàn trải lên tới ***3 năm.***
- **Mục đích:** Hỗ trợ trong việc phát triển các **mô hình học máy** và học sâu nhằm **phân loại** các loại bệnh phổi .

SAMIKSHA DALVI · UPDATED 13 DAYS AGO

36 Code Download

Lungs Diseases Dataset

Lung Disease Dataset: Patient Information, Smoking Status, Treatment and more

Data Card | Code (5) | Discussion (0) | Suggestions (0)

About Dataset

Here's a creative description for your lung disease dataset with emojis:

🫁🩺 **Lung Disease Dataset**

This dataset captures detailed information about patients suffering from various lung conditions. It includes:

- 🧑‍🤝‍🧑 **Age & Gender**: Patient demographics to understand the spread across age groups and gender.
- 🚬 **Smoking Status**: Whether the patient is a smoker or non-smoker.
- 🌬️ **Lung Capacity**: Measured lung function to assess disease severity.
- 🫁 **Disease Type**: The specific lung condition, like COPD or Bronchitis.
- 💊 **Treatment Type**: Different treatments patients received, including therapy, medication, or surgery.

Usability
10.00

License
CC0: Public Domain

Expected update frequency
Never

Tags

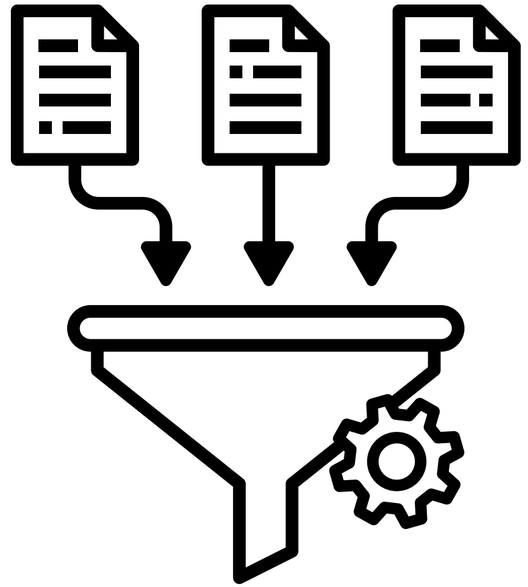
Health | Biology | Health Conditions

2. Cấu trúc dữ liệu

- **Thông tin bệnh nhân:**
  - **Tuổi & Giới tính:** Thông tin về độ tuổi và giới tính của bệnh nhân.
  - **Tình trạng hút thuốc:** Xác định bệnh nhân có hút thuốc hay không.
- **Dung tích phổi:** Đánh giá chức năng phổi để xác định mức độ nghiêm trọng của bệnh.
- **Loại bệnh phổi:** Xác định loại bệnh phổi cụ thể, như COPD hoặc Viêm phế quản, Ung thư phổi, Viêm phổi
- **Phương pháp điều trị:** Các phương pháp điều trị mà bệnh nhân đã nhận, bao gồm liệu pháp, thuốc hoặc phẫu thuật.
- **Số lần nhập viện:** Số lần bệnh nhân nhập viện để quản lý tình trạng bệnh.
- **Tình trạng hồi phục:** Cho biết liệu bệnh nhân có hồi phục sau điều trị hay không.

III. TIỀN XỬ LÝ DỮ LIỆU

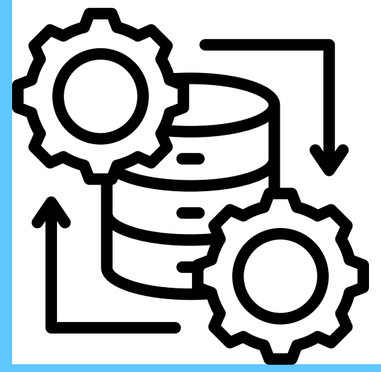
- Xử lý dữ liệu thiếu
- Chuẩn hóa tên cột và ánh xạ giá trị
- Mã hóa biến phân loại
- Mã hóa biến mục tiêu
- Chuẩn hóa dữ liệu



III. TIỀN XỬ LÝ DỮ LIỆU

I. Xử lý dữ liệu thiếu

- Điền giá trị thiếu của cột số **(Tuổi, Dung Lượng Phổi, Số Lần Khám)** bằng trung vị **(median)**.
- Điền giá trị thiếu của cột phân loại **(Giới Tính, Tình Trạng Hút Thuốc, Loại Bệnh, Loại Điều Trị, Phục Hồi)** bằng chế độ **(mode)**.



III. TIỀN XỬ LÝ DỮ LIỆU

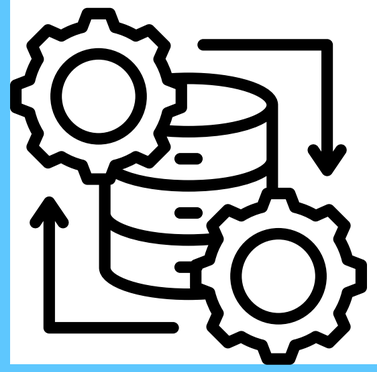
II. Chuẩn hóa tên cột và ánh xạ giá trị

- Đổi tên các cột sang tiếng Việt để dễ hiểu.
- Chuyển đổi giá trị của các cột phân loại như **Tình Trạng Hút Thuốc, Giới Tính, Phục Hồi, Loại Bệnh, Loại Điều Trị** sang tiếng Việt.

III. TIỀN XỬ LÝ DỮ LIỆU

I. Xử lý dữ liệu thiếu

- Điền giá trị thiếu của cột số **(Tuổi, Dung Lượng Phổi, Số Lần Khám)** bằng trung vị **(median)**.
- Điền giá trị thiếu của cột phân loại **(Giới Tính, Tình Trạng Hút Thuốc, Loại Bệnh, Loại Điều Trị, Phục Hồi)** bằng chế độ **(mode)**.



III. TIỀN XỬ LÝ DỮ LIỆU

III. Mã hóa biến phân loại (One-hot encoding)

- Chuyển đổi các biến phân loại thành biến giả **(dummy variables)**.
- Loại bỏ một số cột dư thừa sau khi mã hóa.

IV. Mã hóa biến mục tiêu (Recovered)

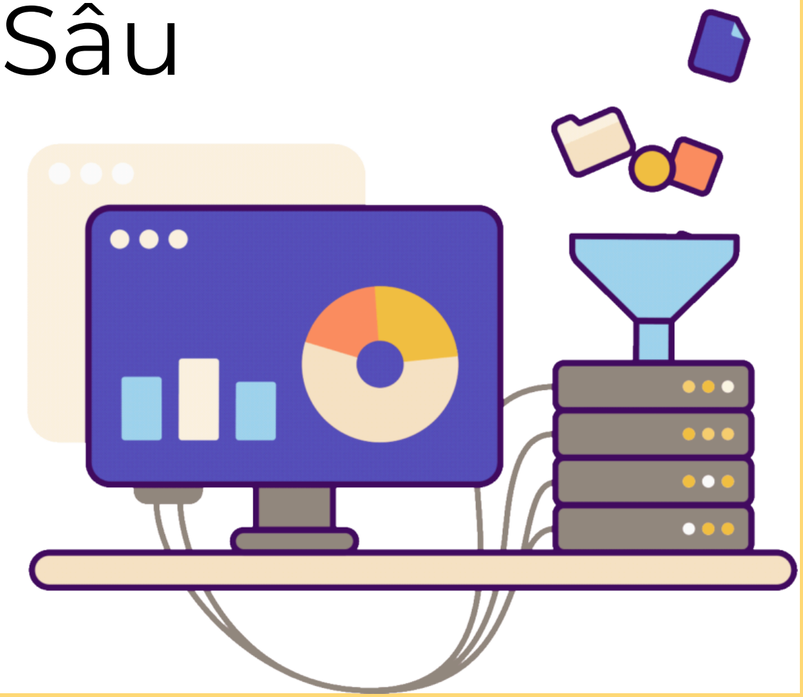
- Chuyển đổi giá trị **yes → 1** và **no → 0**.

V. Chuẩn hóa dữ liệu (Scaling/ Normalization)

- Áp dụng ***Min-Max Scaling*** để đưa dữ liệu số **(Tuổi, Dung Lượng Phổi, Số Lần Khám)** về khoảng **[0,1]**.

IV. CẤU TRÚC DASHBOARD

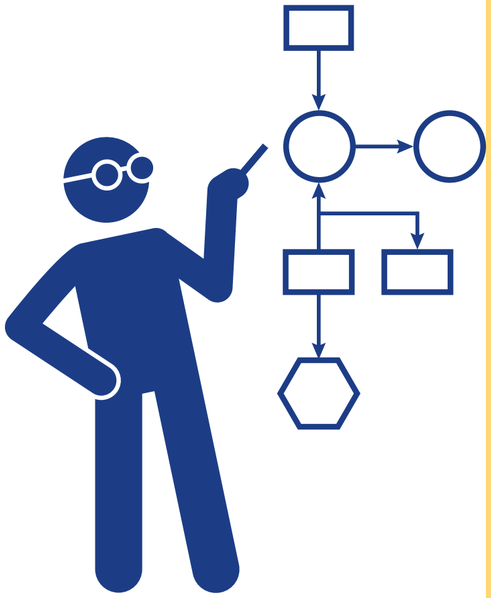
- **Trang 1:** Giới Thiệu Dữ Liệu
- **Trang 2:** Thống Kê Mô Tả
- **Trang 3:** Phân Tích Chuyên Sâu
- **Trang 4:** Nhận Xét Chung



IV. CẤU TRÚC DASHBOARD

Trang 1: Giới Thiệu Dữ Liệu

- **Nguồn dữ liệu**: Tập dữ liệu từ Kaggle về bệnh phổi (hen suyễn, viêm phế quản, COPD, ung thư phổi, viêm phổi).
  - **Các cột dữ liệu chính**: Nhân khẩu học: Tuổi, Giới Tính, Tình Trạng Hút Thuốc.
  - **Sức khỏe**: Dung Tích Phổi, Loại Bệnh, Loại Điều Trị.
  - **Lịch sử khám chữa**: Số Lượt Khám Bệnh, Tình Trạng Hồi Phục.
- **Dữ liệu thô**: Hiển thị bảng dữ liệu mẫu đã dịch sang tiếng Việt.



Deploy

Tải Dữ Liệu

Tải lên tệp CSV

Drag and drop file here
Limit 200MB per file • CSV

Browse files

lung_disease_data.csv
224.1KB

Chọn Trang

- Giới Thiệu Dữ Liệu
- Thống Kê Mô Tả
- Phân Tích Chuyên Sâu
- Nhận Xét Chung

Nguồn Gốc Dữ Liệu

- Dữ liệu được lấy từ nền tảng Kaggle: Lung Disease Prediction.
- Tập dữ liệu bao gồm thông tin về bệnh nhân mắc các bệnh phổi như hen suyễn, viêm phế quản, COPD, ung thư phổi, và viêm phổi.
- Dữ liệu bao gồm các thông tin nhân khẩu học, tình trạng hút thuốc, dung tích phổi, số lượt khám bệnh, và kết quả hồi phục.

Mô Tả Dữ Liệu

- 🧑‍🤝‍🧑 **Age:** Tuổi của bệnh nhân (số nguyên).
- ♀️ **Gender:** Giới tính (Male/Female).
- 🚬 **Smoking Status:** Tình trạng hút thuốc (Yes/No).
- 🌡️ **Lung Capacity:** Dung tích phổi của bệnh nhân (số thực, đơn vị lít).
- 🫁 **Disease Type:** Loại bệnh phổi (Asthma, Bronchitis, COPD, Lung Cancer, Pneumonia).
- 💊 **Treatment Type:** Loại điều trị (Medication, Surgery, Therapy).
- 🏥 **Hospital Visits:** Số lượt khám bệnh (số nguyên).
- ✅ **Recovered:** Bệnh nhân đã hồi phục chưa? (0: No, 1: Yes).

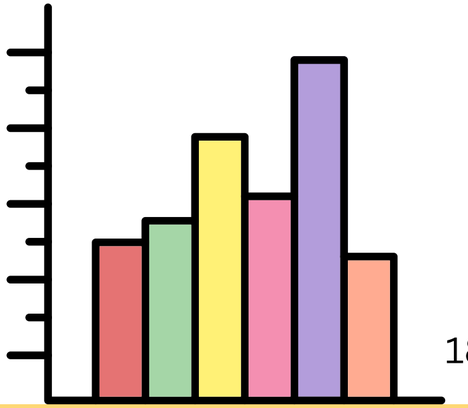
Xem Toàn Bộ Dữ Liệu Thô (đã dịch sang tiếng Việt)

	Tuổi	Giới Tính	Tình Trạng Hút Thuốc	Dung Tích Phổi	Loại Bệnh	Loại Điều Trị	Số Lượt Khám Bệnh	Hồi Phục
0	71	Female	No	4.49	COPD	Therapy	14	1
1	34	Female	Yes	None	Bronchitis	Surgery	7	0
2	80	Male	Yes	1.95	COPD	None	4	1
3	40	Female	Yes	None	Bronchitis	Medication	1	0
4	43	Male	Yes	4.6	COPD	Surgery	None	1

IV. CẤU TRÚC DASHBOARD

Trang 2: Thống Kê Mô Tả

- **Thống kê cơ bản:** Trung bình, tối thiểu, tối đa của các biến số.
- **Phân phối dữ liệu:** Biểu đồ histogram (biến số) và biểu đồ cột (biến phân loại).
- **Nhận xét:**
  - Tuổi trung bình, dung tích phổi, số lượt khám bệnh.
  - Tỷ lệ hồi phục của bệnh nhân.



Deploy

Tải Dữ Liệu

Tải lên tệp CSV

Drag and drop file here

Limit 200MB per file • CSV

Browse files

lung_disease_data.csv

224.1KB

Chọn Trang

- Giới Thiệu Dữ Liệu
- Thống Kê Mô Tả
- Phân Tích Chuyên Sâu
- Nhận Xét Chung

🫁 Bảng Điều Khiển Phân Tích Bệnh Phổi

2. Thống Kê Mô Tả

Thông Tin Dữ Liệu

	Tuổi	Dung Tích Phổi	Số Lượt Khám Bệnh	Hồi Phục
count	4,900	4,900	4,900	4,900
mean	54.4498	3.5019	7.5286	0.5086
std	20.1269	1.4612	3.9964	0.5
min	20	1	1	0
25%	37	2.22	4	0
50%	54	3.48	8	1
75%	72	4.8	11	1
max	89	6	14	1

Phân Phối Dữ Liệu

Chọn một biến số để xem phân phối biến của biến numerical:

IV. CẤU TRÚC DASHBOARD

Trang 3: Phân Tích Chuyên Sâu

- Tổng quan dữ liệu:
  - Tổng số bệnh nhân: 5200
  - Tuổi trung bình: 54.42
  - Dung tích phổi trung bình: 3.50
  - Tỷ lệ hút thuốc: 53.90
- Biểu đồ phân tích:

IV. CẤU TRÚC DASHBOARD

Trang 4: Nhận Xét Chung

- **Phát hiện chính:**
  - Hút thuốc ảnh hưởng đáng kể đến dung tích phổi.
  - Một số loại bệnh có tỷ lệ hồi phục thấp hơn.
  - Sự khác biệt giữa các nhóm tuổi về sức khỏe phổi.
- **Hạn chế:**
  - Dữ liệu thiếu, kích thước mẫu chưa đủ lớn.
  - Đơn vị đo chưa được chuẩn hóa hoàn toàn.

Deploy

Tải Dữ Liệu

Tải lên tệp CSV

Drag and drop file here
Limit 200MB per file • CSV

Browse files

lung_disease_data_clea…
306.4KB

Chọn Trang

- Giới Thiệu Dữ Liệu
- Thống Kê Mô Tả
- Phân Tích Chuyên Sâu
- Nhận Xét Chung

🫁 Bảng Điều Khiển Phân Tích Bệnh Phổi

4. Nhận Xét Chung

- **Tổng Quan về Dữ Liệu và Kết Quả Phân Tích:**
  - **Mối Tương Quan giữa Hút Thuốc và Dung Tích Phổi:** Dữ liệu cho thấy những bệnh nhân hút thuốc có xu hướng có dung tích phổi thấp hơn (xem boxplot).
  - **Phổ Biến của Bệnh:** Các loại bệnh phổi phổ biến nhất trong tập dữ liệu cần được xác định từ biểu đồ phân bố loại bệnh.
  - **Tuổi Trung Bình của Bệnh Nhân:** 54.42 tuổi.
  - **Tỷ Lệ Hút Thuốc:** 53.90%.
  - **Số Lượt Khám Bệnh:** Trung bình 7.56 lượt.
  - **Giới Tính:** Nam: 46.60%, Nữ: 53.40%.

Hạn Chế

- **Kích Thước Mẫu:** Kích thước mẫu có thể không đủ lớn để đưa ra kết luận chắc chắn.
- **Thiếu Sót Dữ Liệu:** Một số hàng có thể thiếu giá trị (nếu có).
- **Đơn Vị Đo:** Dung tích phổi được đo bằng lít, không chuẩn hóa.

V. MỞ RỘNG VÀ PHÁT TRIỂN

- Kết luận
- Mở rộng & Phát triển

V. MỞ RỘNG VÀ PHÁT TRIỂN

Kết luận:

- Dashboard **trực quan hóa** dữ liệu bệnh phổi, giúp **phân tích mối quan hệ** giữa tuổi, giới tính, hút thuốc và hồi phục.
- Các **biểu đồ và phân tích chuyên sâu** hỗ trợ hiểu rõ **xu hướng** bệnh và hỗ trợ quyết định lâm sàng, dù còn hạn chế về kích thước mẫu và dữ liệu thiếu.

Mở rộng và Phát triển

- ***Tích hợp dữ liệu mới:*** Mở rộng tập dữ liệu với nhiều nguồn hơn để tăng độ chính xác và đa dạng.
- ***Cải thiện mô hình dự đoán:*** Ứng dụng Machine Learning để dự đoán nguy cơ mắc bệnh và khả năng hồi phục.
- ***Tăng tính tương tác:*** Bổ sung công cụ lọc dữ liệu nâng cao, cho phép người dùng tùy chỉnh phân tích theo nhu cầu.
- ***Ứng dụng thực tế***: Xây dựng ứng dụng web hoặc mobile để người dùng dễ dàng tiếp cận và theo dõi dữ liệu.

Cảm ơn Thầy và các bạn đã lắng nghe