**TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC**

****

**BÁO CÁO DỰ ÁN TỐT NGHIỆP**

**Chuyên ngành Xử Lý Dữ Liệu  
Phân tích dữ đoán tình hình**

**tai nạn giao thông**

Các thành viên:

* Nguyễn Phạm Đăng Khoa – PS39283 (Nhóm trưởng)
* LÊ Hoàng Bảo – PS38302
* Phạm Minh Hải – PS33057
* Nguyễn Phúc Hậu – PS42890
* Nguyễn Đăng Khoa – PS39225
* Cao Đặng Nhật Anh – PS40602

GV HD: Ths. Văn Công Khanh

TP.HCM 12 - 2023

# LỜI CẢM ƠN

Đầu tiên, nhóm PRO229 chúng em xin chân thành gửi lời cảm ơn đến trường cao đẳng FPT Polytechnic đã cho chúng em một môi trường học tập tốt và được thực hành trong quá trình học tập. Đặc biệt, chúng em muốn gửi lời cảm ơn chân thành nhất đến thầy Văn Công Khanh – giảng viên hổ trợ cho dự án tốt nghiệp của chúng em, thầy đã rất tâm huyết, hỗ trợ và truyền đạt cho chúng em những kiến thức mới bổ ích. Trong thời gian thựuc hiện dự án dưới sự dẫn dắt của thầy, chúng em đã cải thiện được bản thân ngày càng trở nên tốt hơn, hoàn thiện dự án một cách tốt nhất và cũng học được rất nhiều kinh nghiệm từ thầy.

Là những sinh viên của chuyên ngành Xử lý dữ liệu, chúng em đã được học những kiến thức và thực hành với nhiều công cụ khác nhau. Song chúng em cũng được khám phá thêm nhiều kiến thức mới mẻ khi thực hành nhiều dự án thực tế mà nhà trường đã tạo điều kiện. Và đây là dự án quan trọng đánh dấu bước ngoặt trong cuộc đời của chúng em, dự án tốt nghiệp với đề tài “Phân tích dữ đoán tình hình tai nạn giao thông”.

Tuy chúng em đã được học và củng cố kiến thức nhưng cũng không thể tránh được những sai sót. Dưới đây là bản báo cáo dự án tốt nghiệp, cũng là thành quả và công sức của chúng em trong suốt thời gian qua. Chúng em mong nhận được sự đánh giá và góp ý của hội đồng thi, từ đó chúng em có thể rút kinh nghiệm và cải thiện hơn.

Cuối cùng, chúng em xin chúc quý Thầy/Cô sức khỏe và hạnh phúc.

Chúng em xin chân thành cảm ơn!

NHẬN XÉT CỦA GIẢNG VIÊN

**Giảng viên 1:**

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

**Giảng viên 2:**

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

......................................................................................................

MỤC LỤC

[1 Giới thiệu dự án 4](#_Toc165991907)

[1.1 Giới thiệu 4](#_Toc165991908)

[1.2 Yêu cầu của công ty 4](#_Toc165991909)

[1.3 Lập kế hoạch dự án 4](#_Toc165991910)

[2 Phân tích yêu cầu khách hàng 5](#_Toc165991911)

[2.1 Phân tích yêu cầu 5](#_Toc165991912)

[2.2 Câu chuyện dữ liệu 5](#_Toc165991913)

[2.2.1 Đặt vấn đề 5](#_Toc165991914)

[2.2.2 Xác định câu chuyện 6](#_Toc165991915)

[2.2.3 Xác định rõ đối tượng 6](#_Toc165991916)

[2.2.4 Xác định câu chuyện chi tiết 6](#_Toc165991917)

[2.2.5 Trình bày dữ liệu 6](#_Toc165991918)

[2.2.6 Những điều cần lưu ý 6](#_Toc165991919)

[2.3 Kiến trúc hệ thống 7](#_Toc165991920)

[2.3.1 Kiến trúc 7](#_Toc165991921)

[2.3.2 Giải thích 7](#_Toc165991922)

[2.4 Giải thích về bộ dữ liệu khách hàng 7](#_Toc165991923)

[2.4.1 Các khái niệm 7](#_Toc165991924)

[2.4.2 Các trường dữ liệu 7](#_Toc165991925)

[3 Làm sạch và chuyển đổi dữ liệu 8](#_Toc165991926)

[3.1 Chuẩn bị dữ liệu 8](#_Toc165991927)

[3.1.1 Giải pháp lưu trữ dữ liệu 8](#_Toc165991928)

[3.1.2 Giải pháp phân bố dữ liệu 8](#_Toc165991929)

[3.2 Làm sạch dữ liệu 8](#_Toc165991930)

[3.2.1 Các vấn đề ảnh hưởng tới dữ liệu 8](#_Toc165991931)

[3.2.2 Các tiêu chí đánh giá chất lượng dữ liệu 9](#_Toc165991932)

[3.2.3 Các bước làm sạch dữ liệu 9](#_Toc165991933)

[3.3 Chuyển đổi dữ liệu 9](#_Toc165991934)

[3.3.1 Các trường hợp cần chuyển đổi 9](#_Toc165991935)

[3.3.2 Các kỹ thuật chuyển đổi 9](#_Toc165991936)

[3.3.3 Trình bày các phép chuyển đổi trong dự án 9](#_Toc165991937)

[4 Xử lý dữ liệu 10](#_Toc165991938)

[4.1 Chuẩn hóa dữ liệu 10](#_Toc165991939)

[4.1.1 Trình bày các bước chuẩn hóa trong dự án 10](#_Toc165991940)

[4.2 Mô hình hóa dữ liệu 10](#_Toc165991941)

[4.2.1 Các loại mô hình hóa 10](#_Toc165991942)

[4.2.2 Các tiêu chí đánh giá mô hình dữ liệu 10](#_Toc165991943)

[4.2.3 Trình bày các bước mô hình hóa 10](#_Toc165991944)

[4.2.4 Trình bày các bước tạo bảng dữ liệu 11](#_Toc165991945)

[5 Trực quan hóa dữ liệu 11](#_Toc165991946)

[5.1 Các kỹ thuật trực quan hóa 11](#_Toc165991947)

[5.2 Các nguyên tắc trực quan hóa 11](#_Toc165991948)

[5.3 Trình bày tạo các report cho dự án 11](#_Toc165991949)

[5.3.1 Tạo visual thống kê chi tiết 11](#_Toc165991950)

[5.3.2 Tạo visual thống kê tổng thể 12](#_Toc165991951)

[6 Xây dựng báo cáo 12](#_Toc165991952)

[6.1 Dashboard và report 12](#_Toc165991953)

[6.2 Xây dựng báo cáo 12](#_Toc165991954)

[6.2.1 Dashboard vs Report 12](#_Toc165991955)

[6.2.2 Dashboard 13](#_Toc165991956)

[6.2.3 Report 13](#_Toc165991957)

[6.2.4 Bookmark 13](#_Toc165991958)

[7 KẾT LUẬN 14](#_Toc165991959)

[7.1 Báo cáo 14](#_Toc165991960)

[7.1.1 Các bước viết báo cáo 14](#_Toc165991961)

[7.1.2 Tổng hợp 14](#_Toc165991962)

[7.2 Khó khăn 14](#_Toc165991963)

[7.3 Thuận lợi 14](#_Toc165991964)

[7.4 Hướng phát triển 14](#_Toc165991965)

# Giới thiệu dự án

## Giới thiệu

* *Giới thiệu đề tài*

Tai nạn giao thông là một trong những vấn đề nhức nhối của xã hội hiện đại, gây thiệt hại nghiêm trọng về người và tài sản, ảnh hưởng đến trật tự an toàn giao thông và sự phát triển bền vững của mỗi quốc gia. Theo thống kê của Tổ chức Y tế Thế giới (WHO), mỗi năm có hàng triệu người tử vong và bị thương do tai nạn giao thông, trong đó phần lớn nguyên nhân đến từ yếu tố con người, điều kiện đường sá và thời tiết.

Trong thời đại dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI) phát triển mạnh mẽ, việc khai thác dữ liệu để **phân tích xu hướng, dự đoán nguy cơ tai nạn giao thông** là hướng tiếp cận mang lại nhiều giá trị thực tiễn. Từ dữ liệu lịch sử, ta có thể phát hiện các yếu tố nguy cơ, xác định khu vực hoặc thời điểm tiềm ẩn rủi ro cao, hỗ trợ cơ quan chức năng trong việc ra quyết định, quy hoạch hạ tầng, và tuyên truyền an toàn giao thông.

Vì những lý do đó, nhóm em lựa chọn đề tài:

**“Phân tích và dự đoán tình hình tai nạn giao thông”**

Nhằm ứng dụng kiến thức đã học **về xử lý dữ liệu, phân tích dữ liệu và trực quan hóa bằng Power BI** vào một bài toán có ý nghĩa xã hội cao.

## Mục TIÊU VÀ PHẠM VI

### Mục tiêu

* Thu thập và phân tích dữ liệu tai nạn giao thông từ bộ dữ liệu công khai **US Accidents (2016–2023)** trên Kaggle.
* Làm sạch, chuẩn hóa và chuyển đổi dữ liệu phục vụ cho việc trực quan hóa và mô hình hóa.
* Xây dựng các **mô hình dự đoán** (machine learning) nhằm xác định các yếu tố ảnh hưởng và khả năng xảy ra tai nạn.
* Thiết kế **bảng điều khiển (dashboard)** trực quan thể hiện xu hướng tai nạn theo khu vực, thời gian và điều kiện thời tiết.
* Đưa ra **nhận xét và kiến nghị** nhằm giảm thiểu nguy cơ tai nạn giao thông trong tương lai.

### Phạm Vi Thực Hiện

* **Dữ liệu:** Sử dụng bộ dữ liệu *US Accidents (2016–2023)*, với hơn 7 triệu bản ghi chứa thông tin về vị trí, thời gian, điều kiện thời tiết, mức độ nghiêm trọng và các yếu tố liên quan đến tai nạn tại Hoa Kỳ.
* **Thời gian thực hiện:** Trong học kỳ cuối năm 2025.
* **Công cụ sử dụng:** Python (pandas, numpy, scikit-learn, matplotlib), Tableau để trực quan hóa dữ liệu, và Excel để kiểm tra chéo dữ liệu đầu vào.
* **Phạm vi nghiên cứu:** Tập trung vào việc phân tích và dự đoán **mức độ nghiêm trọng (Severity)** và **tần suất tai nạn** theo các yếu tố môi trường, thời gian, và vị trí địa lý.

## SẢN PHẨM MỤC TIÊU

* *Sản phẩm cuối cùng của đề tài bao gồm:*

1. **Tập dữ liệu đã được làm sạch và chuẩn hóa** từ bộ dữ liệu gốc.
2. **Dashboard Tableau** thể hiện trực quan xu hướng và yếu tố ảnh hưởng đến tai nạn.
3. **Mô hình dự đoán mức độ nghiêm trọng tai nạn,** sử dụng các thuật toán như Logistic Regression, Decision Tree, hoặc Random Forest.
4. **Báo cáo tổng hợp kết quả phân tích và đề xuất giải pháp**, trình bày theo mẫu báo cáo dự án của trường FPT Polytechnic.
5. **Mã nguồn (script)** xử lý, huấn luyện và đánh giá mô hình, được lưu trữ trên GitHub để thuận tiện cho việc kiểm tra và tái sử dụng.

## 1.4 CẤU TRÚC BÁO CẤU

# tỔNG QUAN DỮ LIỆU VÀ CÔNG NGHỆ

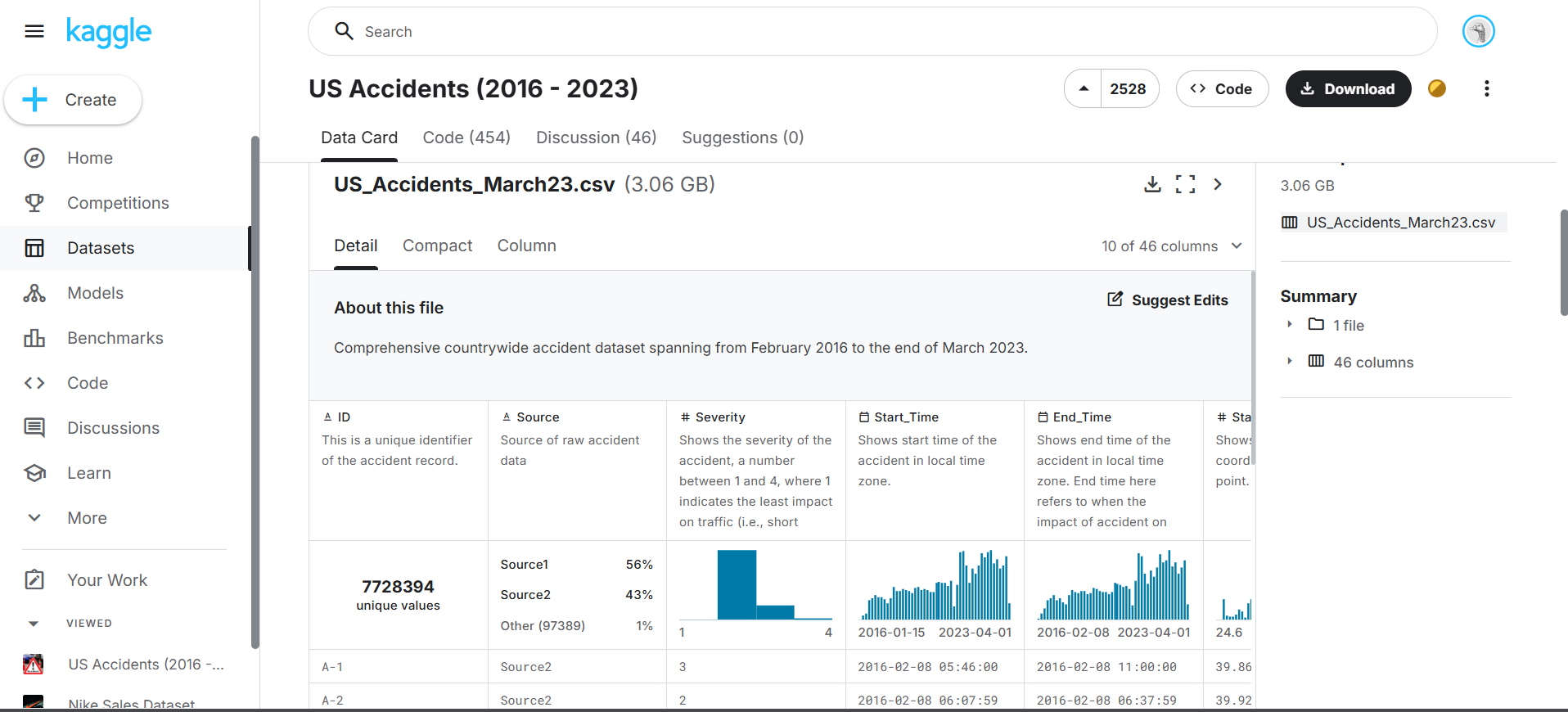
## MÔ TẢ DỮ LIỆU

### Nguồn Dữ Liệu

Dữ liệu sử dụng trong dự án được thu thập từ **Kaggle – bộ dữ liệu US Accidents (2016–2023)**, được xây dựng bởi *Sobhan Moosavi* (truy cập tại:

<https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>).

Hình ảnh



Bộ dữ liệu chứa **hơn 7,4 triệu bản ghi** về các vụ tai nạn giao thông được ghi nhận tại **49 bang của Hoa Kỳ** trong giai đoạn từ năm 2016 đến 2023. Dữ liệu được tổng hợp tự động từ các nguồn báo cáo giao thông thời gian thực như: **API của MapQuest, Bing, và các hệ thống cảm biến giao thông**.

Mục tiêu chính của bộ dữ liệu là cung cấp thông tin toàn diện để phục vụ cho **nghiên cứu, phân tích xu hướng và dự đoán tai nạn giao thông**, từ đó giúp cải thiện an toàn và hiệu quả giao thông.

### Cấu Trúc Và Định Dạng dữ liệu

Bộ dữ liệu ở dạng **CSV (Comma Separated Values)**, gồm **46 cột thông tin (attributes)**. Một số cột quan trọng gồm:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên Cột** | **Mô tả** |
| 1 | ID | Mã định danh duy nhất của mỗi bản ghi tai nạn. |
| 2 | Source | Nguồn dữ liệu gốc thu thập thông tin về tai nạn. |
| 3 | Severity | Mức độ nghiêm trọng của tai nạn (1-4). |
| 4 | Start\_Time | Thời điểm tai nạn bắt đầu (theo múi giờ địa phương). |
| 5 | End\_Time | Thời điểm kết thúc ảnh hưởng của tai nạn đến giao thông. |
| 6 | Start\_Lat | Vĩ độ GPS tại điểm bắt đầu của khu vực xảy ra tai nạn. |
| 7 | Start\_Lng | Kinh độ GPS tại điểm bắt đầu của khu vực xảy ra tai nạn. |
| 8 | End\_Lat | Vĩ độ GPS tại điểm kết thúc khu vực bị ảnh hưởng. |
| 9 | End\_Lng | Kinh độ GPS tại điểm kết thúc khu vực bị ảnh hưởng. |
| 10 | Distance(mi) | Chiều dài đoạn đường bị ảnh hưởng bởi tai nạn (dặm). |
| 11 | Description | Mô tả bằng lời về tình huống tai nạn. |
| 12 | Street | Tên đường nơi xảy ra tai nạn. |
| 13 | City | Thành phố nơi xảy ra tai nạn. |
| 14 | County | Quận/hạt nơi xảy ra tai nạn. |
| 15 | State | Bang nơi xảy ra tai nạn. |
| 16 | Zipcode | Mã bưu điện khu vực tai nạn. |
| 17 | Country | Quốc gia nơi xảy ra tai nạn. |
| 18 | Timezone | Múi giờ của vị trí tai nạn. |
| 19 | Airport\_Code | Mã sân bay của trạm khí tượng gần nhất. |
| 20 | Weather\_Timestamp | Thời gian ghi nhận dữ liệu thời tiết (địa phương). |
| 21 | Temperature(F) | Nhiệt độ tại thời điểm tai nạn (°F). |
| 22 | Wind\_Chill(F) | Cảm giác lạnh do gió gây ra (°F). |
| 23 | Humidity(%) | Độ ẩm không khí (%). |
| 24 | Pressure(in) | Áp suất không khí (inch thủy ngân). |
| 25 | Visibility(mi) | Tầm nhìn (dặm). |
| 26 | Wind\_Direction | Hướng gió tại thời điểm xảy ra tai nạn. |
| 27 | Wind\_Speed(mph) | Tốc độ gió (dặm/giờ). |
| 28 | Precipitation(in) | Lượng mưa (inch). |
| 29 | Weather\_Condition | Tình trạng thời tiết (mưa, tuyết, sương mù, v.v.). |
| 30 | Amenity | Có công trình tiện ích gần khu vực tai nạn. |
| 31 | Bump | Có gờ giảm tốc gần khu vực tai nạn. |
| 32 | Crossing | Có lối băng qua đường gần đó. |
| 33 | Give\_Way | Có biển nhường đường gần khu vực. |
| 34 | Junction | Có ngã ba, ngã tư, giao lộ gần khu vực. |
| 35 | No\_Exit | Khu vực không có lối ra gần nơi xảy ra tai nạn. |
| 36 | Railway | Có đường ray xe lửa gần vị trí tai nạn. |
| 37 | Roundabout | Có vòng xuyến gần khu vực. |
| 38 | Station | Có ga tàu hoặc trạm xe buýt gần khu vực tai nạn. |
| 39 | Stop | Có biển dừng gần nơi xảy ra tai nạn. |
| 40 | Traffic\_Calming | Có biện pháp giảm tốc giao thông gần khu vực. |
| 41 | Traffic\_Signal | Có đèn tín hiệu giao thông gần nơi xảy ra tai nạn. |
| 42 | Turning\_Loop | Có vòng quay đầu xe gần khu vực. |
| 43 | Sunrise\_Sunset | Thời điểm xảy ra (ban ngày/ban đêm) theo mặt trời mọc/lặn. |
| 44 | Civil\_Twilight | Thời điểm xảy ra (ngày/đêm) theo chạng vạng dân dụng. |
| 45 | Nautical\_Twilight | Thời điểm xảy ra (ngày/đêm) theo chạng vạng hàng hải. |
| 46 | Astronomical\_Twilight | Thời điểm xảy ra (ngày/đêm) theo chạng vạng thiên văn. |

Nhóm đã thảo luận và bỏ một số cột không cần thiết:

**'Source', 'End\_Lat', 'End\_Lng', 'Description', 'Country', 'Timezone', 'Airport\_Code', 'Weather\_Timestamp', 'Wind\_Chill(F)', 'Pressure(in)', 'Wind\_Direction', 'Amenity', 'Bump', 'Give\_Way', 'No\_Exit', 'Railway'**

Bộ dữ liệu hiện tại còn 30 cột quan trọng để phục vụ cho việc trực quan hoá

### đặc điểm dữ liệu

* Dữ liệu có **phân bố không đồng đều** giữa các bang; California, Florida và Texas chiếm tỷ trọng lớn nhất.
* Thời gian xảy ra tai nạn tập trung vào **giờ cao điểm buổi sáng (7–9h)** và **buổi chiều (16–18h)**.
* Các yếu tố thời tiết như **mưa, sương mù, tầm nhìn thấp** có mối tương quan mạnh với mức độ nghiêm trọng.
* Dữ liệu chứa một số **giá trị thiếu (missing values)** ở các cột như *Visibility*, *Wind\_Speed*, *Weather\_Condition*, cần được xử lý trong bước tiền xử lý dữ liệu (Chương 3).

## CÁC CÔNG CỤ, THƯ VIỆN VÀ CÔNG NGHỆ SỬ DỤNG

### Ngôn ngữ lập trình và môi trường làm việc

* **Python 3.11**: ngôn ngữ chính dùng để xử lý, làm sạch và phân tích dữ liệu.
* **Jupyter Notebook / Google Colab**: môi trường thực thi mã nguồn, dễ dàng trực quan hóa và trình bày kết quả.
* **Power BI**: công cụ tạo **dashboard** và biểu đồ tương tác trực quan hóa dữ liệu sau khi xử lý.
* **Excel**: được sử dụng để rà soát và kiểm tra nhanh dữ liệu thô, thống kê sơ bộ.

### thư viện python

* **pandas**: đọc, làm sạch và thao tác dữ liệu dạng bảng.
* **numpy**: hỗ trợ tính toán số học, vector, và mảng lớn.
* **matplotlib / seaborn**: trực quan hóa dữ liệu bằng biểu đồ, heatmap, scatter plot, histogram.
* **scikit-learn**: xây dựng và huấn luyện mô hình học máy (machine learning).
* **xgboost / lightgbm**: thử nghiệm các mô hình dự đoán hiệu năng cao.
* **datetime / time**: xử lý thời gian, trích xuất đặc trưng ngày, tháng, giờ, mùa.

### các công cụ hỗ trợ khác

* **Git & GitHub:** quản lý mã nguồn và phiên bản mô hình.
* **SQL (Microsoft SQL Server):** lưu trữ và truy vấn dữ liệu lớn.
* **Google Drive:** lưu trữ dữ liệu trung gian và chia sẻ trong nhóm.

## tổng quan mô hình, thuật toán liên quan

### học máy trong dự đoán tai nạn giao thông

Trong bài toán này, nhóm sử dụng các **thuật toán học có giám sát (Supervised Learning)** để dự đoán **mức độ nghiêm trọng (Severity)** của tai nạn, dựa trên các yếu tố môi trường và thời tiết.

Các mô hình được xem xét bao gồm:

* **Logistic Regression:** mô hình hồi quy nhị phân, dễ hiểu và giải thích tốt.
* **Decision Tree:** xây dựng cây phân nhánh theo điều kiện, phù hợp với dữ liệu có nhiều biến phân loại.
* **Random Forest:** tổ hợp nhiều cây quyết định, giúp cải thiện độ chính xác và giảm overfitting.
* **XGBoost:** mô hình boosting hiệu quả, tối ưu hóa cho dữ liệu lớn, hiệu suất cao.

### Tiêu chí đánh giá mô hình

Các mô hình sẽ được đánh giá bằng những chỉ số phổ biến:

* **Accuracy** (Độ chính xác)
* **Precision** (Độ chính xác theo lớp)
* **Recall** (Độ bao phủ)
* **F1-Score** (Trung bình điều hòa giữa Precision và Recall)
* **ROC-AUC** (Đường cong đặc trưng hoạt động của mô hình)

Các chỉ số này giúp đánh giá mô hình toàn diện hơn, đặc biệt trong trường hợp dữ liệu bị mất cân bằng giữa các mức độ tai nạn.

### Những điều cần lưu ý

Khi áp dụng các mô hình học máy trong việc **phân tích và dự đoán tình hình tai nạn giao thông**, cần đặc biệt lưu ý đến các yếu tố sau:

* Chất lượng dữ liệu đầu vào:Dữ liệu cần được làm sạch, chuẩn hóa và loại bỏ các giá trị sai lệch. Nếu dữ liệu đầu vào có nhiều nhiễu hoặc thiếu thông tin, kết quả mô hình sẽ không phản ánh đúng thực tế.
* Cân bằng dữ liệu (Data Imbalance): Trong bộ dữ liệu tai nạn giao thông, số lượng vụ tai nạn nhẹ thường nhiều hơn vụ nghiêm trọng. Việc mất cân bằng dữ liệu có thể khiến mô hình dự đoán thiên lệch.

→ Cần áp dụng kỹ thuật như **Resampling** hoặc **SMOTE** để cân bằng các lớp dữ liệu.

* Cần áp dụng kỹ thuật như **Resampling** hoặc **SMOTE** để cân bằng các lớp dữ liệu: Không phải tất cả các biến đều có tác động mạnh đến kết quả.  
  → Nên chọn lọc những đặc trưng quan trọng (ví dụ: thời tiết, thời gian, tầm nhìn, tốc độ gió, khu vực) để giảm nhiễu và tăng độ chính xác.
* Tránh overfitting (quá khớp): Khi mô hình học quá kỹ dữ liệu huấn luyện, nó có thể mất khả năng dự đoán chính xác cho dữ liệu mới.  
  → Cần chia dữ liệu thành tập **train/test**, dùng **cross-validation** và điều chỉnh **tham số mô hình (hyperparameter tuning)**.
* Giải thích mô hình (Model Interpretability): Ngoài độ chính xác, việc giải thích được **tại sao mô hình đưa ra kết quả đó** cũng rất quan trọng — đặc biệt khi ứng dụng trong lĩnh vực an toàn giao thông.

→ Ưu tiên các mô hình có khả năng diễn giải tốt như **Decision Tree**, hoặc dùng công cụ hỗ trợ giải thích như **SHAP** hoặc **LIME**.

* Đánh giá mô hình định kỳ: Dữ liệu thực tế luôn thay đổi theo thời gian (ví dụ do chính sách, cơ sở hạ tầng, thời tiết).

→ Cần **tái huấn luyện mô hình định kỳ** để duy trì hiệu quả dự đoán.

## Kiến trúc hệ thống

### Mục tiêu thiết kế hệ thống

Hệ thống được thiết kế nhằm **tự động hóa quá trình thu thập, xử lý, phân tích và trực quan hóa dữ liệu tai nạn giao thông**.

Mục tiêu chính:

* Tạo luồng xử lý dữ liệu khép kín, từ đầu vào đến đầu ra.
* Dễ dàng mở rộng, cập nhật và tái huấn luyện mô hình.
* Cho phép người dùng cuối (giảng viên, nhà phân tích) **trực quan hóa dữ liệu và kết quả dự đoán trực tiếp**.

### mô hình kiến trúc tổng thể

Hệ thống được thiết kế gồm **5 tầng chính**, được mô tả trong **Hình 2.1** dưới đây.

Hình 2.1. Kiến trúc tổng thể của hệ thống phân tích và dự đoán tai nạn giao thông

|  |  |  |
| --- | --- | --- |
| Tầng | Thành phần chính | Chức năng |
| Tầng thu thập dữ liệu (Data Collection Layer) | Kaggle Dataset (US Accidents 2016–2023) | Nguồn dữ liệu đầu vào được tải xuống dưới định dạng |
| Tầng tiền xử lý dữ liệu (Data Preprocessing Layer) | Python (Pandas, NumPy) | Làm sạch, xử lý giá trị thiếu, loại bỏ nhiễu, chuẩn hóa định dạng thời gian. |
| Tầng mô hình hóa (Modeling Layer) | Scikit-learn, XGBoost | Xây dựng và huấn luyện mô hình dự đoán mức độ nghiêm trọng của tai nạn (Severity). |
| Tầng lưu trữ và truy vấn (Storage Layer) | SQL Server, Google Drive | Lưu trữ dữ liệu trung gian và kết quả mô hình để phục vụ truy vấn và báo cáo. |
| Tầng trình bày và trực quan (Visualization Layer) | Power BI, Excel | Trực quan hóa dữ liệu, hiển thị dashboard, biểu đồ thống kê, và báo cáo kết quả dự đoán. |

### quy trình luồng dữ liệu(data flow)

Quy trình hoạt động của hệ thống như sau:

 **Thu thập dữ liệu** từ Kaggle.

 **Làm sạch và xử lý** dữ liệu bằng Python.

 **Lưu trữ dữ liệu** sạch vào SQL Server hoặc file CSV trung gian.

 **Huấn luyện mô hình dự đoán** bằng Scikit-learn hoặc XGBoost.

 **Đưa kết quả mô hình vào Power BI** để trực quan hóa và đánh giá.

Quy trình này giúp nhóm dễ dàng:

* Cập nhật dữ liệu mới định kỳ.
* Tối ưu mô hình khi có thay đổi trong dữ liệu thực tế.
* Tạo báo cáo động, trực quan và dễ hiểu.

### ưu điểm của kiến trúc hệ thống

 **Tự động hóa cao:** giảm thao tác thủ công khi phân tích dữ liệu.

 **Dễ bảo trì:** từng tầng hoạt động độc lập, có thể thay thế hoặc mở rộng.

 **Tính trực quan:** kết quả cuối cùng được trình bày rõ ràng trên Power BI.

 **Khả năng mở rộng:** dễ tích hợp thêm các nguồn dữ liệu khác như thời tiết, giao thông thời gian thực, bản đồ địa lý.

### kết luận tiểu mục

Kiến trúc hệ thống được xây dựng theo hướng **mở, linh hoạt và dễ mở rộng**, cho phép nhóm sinh viên triển khai toàn bộ quy trình **phân tích – dự đoán – trực quan hóa dữ liệu tai nạn giao thông** một cách hiệu quả và có tính ứng dụng cao.

# Quy trình xử lý dữ liệu

## Tổng quan quy trình xử lý dữ liệu

Quá trình xử lý dữ liệu là bước trọng tâm trong dự án phân tích và dự đoán tai nạn giao thông.  
Nhóm đã thực hiện quy trình **6 bước truyền thống**, đảm bảo dữ liệu được làm sạch, chuẩn hóa và sẵn sàng cho việc xây dựng mô hình dự đoán.

Sơ đồ 3.1: Quy trình xử lý dữ liệu truyền thống (6 bước)



|  |  |  |
| --- | --- | --- |
| Bước | Tên Bước | Mô Tả Công Việc Chính |
| 1 | |  | | --- | |  |  |  | | --- | | Xác định mục tiêu phân tích | | Xác định mục tiêu cần đạt được: phân tích xu hướng, phát hiện yếu tố rủi ro, dự đoán mức độ nghiêm trọng (Severity). |
| 2 | Thu thập dữ liệu | Thu thập dữ liệu từ nguồn Kaggle – US Accidents Dataset (2016–2023). |
| 3 | Làm sạch dữ liệu (Data Cleaning) | Loại bỏ dữ liệu trùng lặp, giá trị thiếu, định dạng sai; chuẩn hóa đơn vị và kiểu dữ liệu. |
| 4 | Thực hiện phân tích (Data Analysis) | Phân tích thống kê, khám phá đặc trưng (EDA) và nhận diện xu hướng. |
| 5 | Trình bày kết quả (Visualization) | Trực quan hóa dữ liệu bằng biểu đồ, dashboard trên Power BI. |
| 6 | Đánh giá và cải thiện | Kiểm tra tính chính xác, hiệu suất và cập nhật mô hình dựa trên phản hồi thực tế. |

Quy trình trên đảm bảo **luồng dữ liệu liền mạch từ đầu vào đến đầu ra**, hỗ trợ quá trình **phân tích, mô hình hóa và trực quan hóa** hiệu quả.

## làm sạch và xử lý dữ liệu

### các bước xử lý dữ liệu

**Xử lý dữ liệu trùng lặp (Duplicate Records):**

* Kiểm tra các dòng trùng ID hoặc cùng thời gian, tọa độ.
* Giữ lại bản ghi hợp lệ nhất, loại bỏ các bản ghi trùng.

**Xử lý giá trị thiếu (Missing Values):**

* Với cột thời tiết (Temperature, Humidity), thay thế giá trị thiếu bằng **trung bình hoặc trung vị**.
* Với cột danh mục (Weather\_Condition, City), thay thế bằng **giá trị xuất hiện nhiều nhất (mode)**.

**Xử lý định dạng thời gian:**

* Chuẩn hóa cột Start\_Time và End\_Time về định dạng ISO (YYYY-MM-DD HH:MM:SS).
* Tách các trường **ngày, tháng, giờ, thứ** để phục vụ phân tích theo thời gian.

**Chuẩn hóa dữ liệu văn bản:**

* Loại bỏ ký tự đặc biệt, viết hoa đầu từ hoặc thống nhất chữ thường.
* Chuẩn hóa tên thành phố và bang (City, State) theo định dạng chuẩn ISO US.

**Phát hiện và xử lý ngoại lệ (Outliers):**

* **Loại bỏ các giá trị bất thường**
* **Sử dụng biểu đồ Boxplot để phát hiện điểm ngoại lệ.**

## phân tích và chuyển đổi dữ liệu

### khám phá dữ liệu (exploratory data analysis – eda)

**Giai đoạn EDA giúp hiểu rõ cấu trúc và mối quan hệ giữa các biến trong bộ dữ liệu.**Nhóm đã thực hiện một số phân tích sau:

**Phân bố tai nạn theo bang:**

→ California, Florida và Texas chiếm tỷ lệ cao nhất.

**Phân tích theo thời gian:**

→ Tai nạn xảy ra nhiều nhất vào giờ cao điểm (7h–9h sáng, 16h–18h chiều).

Phân tích theo điều kiện thời tiết:

→ Tỷ lệ tai nạn nghiêm trọng cao hơn khi trời mưa, sương mù, tuyết hoặc tầm nhìn < 5 km.

**Phân bố mức độ nghiêm trọng (Severity):**

→ Dữ liệu thiên lệch về các vụ tai nạn nhẹ (mức 2), cần cân bằng lớp dữ liệu.

Các biểu đồ được thể hiện trong Power BI giúp người đọc trực quan hóa được xu hướng theo thời gian, khu vực, và điều kiện khí hậu.

### TRÍCH XUẤT VÀ CHỌN LỌC ĐẶC TRƯNG(FEATURE ENGINEERING)

**Nhằm phục vụ cho mô hình dự đoán, nhóm tiến hành:**

* Tạo thêm các đặc trưng mới **(Derived Features):**
* **Hour**, **DayOfWeek**,  **Month** từ cột **Start\_Time**.
* **IsWeekend** (1 nếu là Thứ 7 – Chủ nhật, 0 nếu là ngày thường).
* **IsNight** (1 nếu xảy ra ban đêm, dựa vào cột Sunrise\_Sunset).
* Mã hóa dữ liệu phân loại **(Categorical Encoding):**
* Biến đổi cột **Weather\_Condition**, **State,** **City** thành dạng số bằng kỹ thuật **Label Encoding** hoặc **One-Hot Encoding**.
* **Chuẩn hóa dữ liệu (Normalization):**
* Chuẩn hóa các giá trị số như **Temperature**, **Visibility**, **Wind\_Speed** để giảm sai lệch giữa các biến.
* **Chia dữ liệu thành các tập Train/Test:**
* Dữ liệu được chia theo tỷ lệ **80% huấn luyện / 20% kiểm thử**.

### LƯU TRỮ DỮ LIỆU SAU XỬ LÝ

**Sau khi làm sạch và trích xuất đặc trưng:**

* **Dữ liệu được lưu lại dưới định dạng .csv hoặc đưa vào SQL Server để phục vụ mô hình hóa.**
* **Các file được đặt tên theo quy chuẩn:**

**cleaned\_accidents.csv,** **encoded\_accidents.csv,** **features\_ready.csv.**

* **Dữ liệu này sẽ là đầu vào cho Chương 4 – Xây dựng mô hình dự đoán tai nạn giao thông.**

#### Trình bày các bước làm sạch

…

#### Trình bày các bước làm sạch trong phạm vi dự án

* *Trình bày cụ thể các bước tác động cụ thể đến bộ dữ liệu dự án (từng cột, từng dòng).*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

…

## Chuyển đổi dữ liệu

* *Các trường hợp và kỹ thuật chuyển đổi dữ liệu.*

### Các trường hợp cần chuyển đổi

* *Nêu các trường hợp cần chuyển đổi dữ liệu*

…

### Các kỹ thuật chuyển đổi

* *Trình bày các kỹ thuật chuyển đổi*
* *Kỹ thuật nào được chọn áp dụng cho dự án.*

…

### Trình bày các phép chuyển đổi trong dự án

* *Trình bày cụ thể các bước tác động cụ thể đến bộ dữ liệu dự án (từng cột, từng dòng).*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

…

# Xử lý dữ liệu

## Chuẩn hóa dữ liệu

* *Chuẩn hóa dữ liệu:*
  + *Power Query*
  + *Chỉnh sửa table, column, data type*

### Trình bày các bước chuẩn hóa trong dự án

* *Trình bày cụ thể các bước chuẩn hóa dữ liệu bằng Power Query tác động cụ thể đến bộ dữ liệu dự án (từng cột, từng dòng).*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

**…**

## Mô hình hóa dữ liệu

* *Mô hình hóa dữ liệu:*
  + *Data model*
  + *Data modeling*

### Các loại mô hình hóa

* *Trình bày các loại mô hình*
* *Mô hình nào đang được áp dụng cho bộ dữ liệu dự án ?*

…

### Các tiêu chí đánh giá mô hình dữ liệu

* *Trình bày các tiêu chí đánh giá mô hình dữ liệu tốt ?*
* *Mô hình đang dùng trong dự án đáp ứng các tiêu chí nào ?*

…

### Trình bày các bước mô hình hóa

* *Trình bày cụ thể các bước mô hình hóa dữ liệu bằng Power BI trên bộ dữ liệu của dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

…

### Trình bày các bước tạo bảng dữ liệu

* *Trình bày cụ thể các bước tạo bảng dữ liệu bằng Power BI trên bộ dữ liệu của dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

…

# Trực quan hóa dữ liệu

## Các kỹ thuật trực quan hóa

* *Trình bày các kỹ thuật trực quan hóa*
* *Kỹ thuật nào đang được áp dụng cho dự án ? Vì sao ?*

…

## Các nguyên tắc trực quan hóa

* *Trình bày các nguyên tắc trong trực quan hóa*
* *Nguyên tắc nào bạn cho rằng quan trọng nhất trong phạm vi dự án này ? Vì sao ?*

…

## Trình bày tạo các report cho dự án

* *Tạo các bộ lọc, biểu đồ theo nhu cầu dự án.*
* *Có thể phân chia theo đối tượng trong mô hình.*
* *Có thể phân chia theo chức năng hoặc page.*
* *Trong dự án cho môn học này thì phân chia theo page.*

### Tạo visual thống kê chi tiết

* *Tạo các visual dùng trong trang thống kê chi tiết.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo visual filter theo

…

#### Tạo visual filter theo

…

#### Tạo visual filter theo

…

#### Tạo visual tiếp theo ….

…

### Tạo visual thống kê tổng thể

* *Tạo các visual dùng trong trang thống kê tổng thể.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo visual filter theo ngày giao dịch

…

#### Tạo visual thống kê Tổng GTGD

…

#### Tạo visual thống kê Tổng KLGD

…

#### Tạo visual tiếp theo ….

…

# Xây dựng báo cáo

## Dashboard và report

* *Cách tối ưu hóa:*
  + *Dashboard*
  + *Report*

…

## Xây dựng báo cáo

* *Xây dựng báo cáo:*
  + *Dashboard*
  + *Report*

### Dashboard vs Report

* *Trình bày sự khác nhau giữa Dashboard và Report.*
* *Giải thích vì sao làm dashboard, report trong dự án này ?*

…

*Giải thích:* …

### Dashboard

* *Trình bày các loại dashboard có trong dự án.*
* *Giải thích vì sao làm các dashboard này ?*

…

*Giải thích:* …

### Report

* *Trình bày các loại report có trong dự án.*
* *Giải thích vì sao làm các report này ?*

#### Tạo report Detail

*Giải thích:* …

…

#### Tạo report Overview

*Giải thích:* …

…

### Bookmark

* *Trình bày các loại bookmark có trong dự án.*
* *Giải thích vì sao làm các bookmark này ?*

#### Tạo bookmark slicer

*Giải thích:* ….

#### Tạo bookmark slicer chọn

…

*Giải thích:* …

#### Tạo bookmark slicer chọn

…

*Giải thích:* …

#### Tạo visual tiếp theo ….

…

*Giải thích:* …

# KẾT LUẬN

## Báo cáo

### Các bước viết báo cáo

* *Trình bày các bước viết báo cáo phân tích dữ liệu*

…

### Tổng hợp

* *Tổng hợp các câu trả lời trong Câu Chuyện Dữ Liệu ở Phân Tích Khách Hàng, kết hợp với các bước viết báo cáo phân tích, để tạo thành bản báo cáo hoàn chỉnh.*

*Gợi ý:*

…

## Khó khăn

## Thuận lợi

## Hướng phát triển