GROUP 8 - ASSIGMENT PROJECT

(Accident Severity Prediction)

Giới thiệu về project dự đoán mức độ nghiêm trọng của tai nạn ở nước Anh:

Trong bối cảnh gia tăng số lượng tai nạn giao thông trên toàn cầu, việc hiểu và dự đoán mức độ nghiêm trọng của tai nạn trở nên ngày càng quan trọng. Theo số liệu của các cơ quan thống kê, tai nạn giao thông là một trong những nguyên nhân hàng đầu gây tử vong và thương tật, ảnh hưởng không chỉ đến sức khỏe của người dân mà còn đặt ra nhiều thách thức cho hệ thống giao thông và y tế. Ở Anh, số liệu cho thấy mức độ nghiêm trọng của tai nạn giao thông có xu hướng tăng lên, đặc biệt là ở các khu vực đô thị đông đúc.

Tai nạn giao thông có thể xảy ra vì nhiều lý do khác nhau, bao gồm điều kiện thời tiết, trạng thái của đường, hành vi của người lái xe và các yếu tố khác như tốc độ và loại phương tiện. Việc dự đoán chính xác mức độ nghiêm trọng của tai nạn không chỉ giúp giảm thiểu thiệt hại về người và tài sản mà còn hỗ trợ các cơ quan chức năng trong việc triển khai các biện pháp phòng ngừa hiệu quả hơn.

Ngày nay, với sự phát triển của công nghệ và sự gia tăng dữ liệu lớn trong lĩnh vực giao thông, việc áp dụng các phương pháp phân tích dữ liệu, như khai thác dữ liệu (data mining) và học máy (machine learning), để dự đoán mức độ nghiêm trọng của tai nạn đã trở nên khả thi. Các nhà nghiên cứu và chuyên gia giao thông đang tích cực khai thác dữ liệu từ các nguồn khác nhau để xây dựng các mô hình dự đoán hiệu quả.

Dự án này nhằm mục đích phát triển một mô hình dự đoán mức độ nghiêm trọng của tai nạn giao thông tại Anh dựa trên các yếu tố như điều kiện mặt đường, điều kiện thời tiết và đặc điểm của phương tiện trong quá khứ. Kết quả của dự án sẽ không chỉ góp phần nâng cao an toàn giao thông mà còn cung cấp thông tin quý giá cho các chính sách và chiến lược cải thiện an toàn đường bộ trong tương lai.

Trong project này, chúng ta sẽ áp dụng các phương pháp máy học cơ bản để dự đoán xem mức độ nghiêm trọng của tai nạn khi sảy ra dựa trên tập dữ liệu về Tai Nạn **UK Road Safety: Traffic Accidents and Vehicles**. Tập dữ liệu bao gồm những thông tin như sau:

Accident Index	Mã được gán cho từng vụ tai nạn
1st Road Class	Phân loại của con đường chính nơi xảy ra tai nạn.
1st Road Number	Số đường chính nơi xảy ra tai nạn.
2nd Road Class	Phân loại của con đường phụ nơi xảy ra tai nạn
2nd Road Number	Số đường phụ nơi xảy ra tai nạn.
Accident Severity	Mức độ nghiêm trọng của tai nạn.
Carriageway Hazards	Các mối nguy hiểm hiện diện trên mặt đường tại thời điểm xảy ra tai nạn.
Date	Ngày xảy ra tai nạn.
Day of Week	Ngày trong tuần xảy ra tai nạn.
Did Police Officer Attend Scene of Accident	Có cảnh sát đến hiện trường tai nạn không.
Junction Control	Loại kiểm soát giao lộ tại nơi xảy ra tai nạn.
Junction Detail	Chi tiết loại giao lộ.
Latitude	Vĩ độ của địa điểm xảy ra tai nạn.
Light Conditions	Điều kiện ánh sáng tại thời điểm xảy ra tai.
Local Authority (District)	Quận nơi xảy ra tai nạn.
Local Authority (Highway)	Cơ quan quản lý tại nơi xảy ra tai nạn.
Location Easting OSGR	Tọa độ Đông địa điểm xảy ra tai nạn.

Location Northing OSGR	Tọa độ Bắc của địa điểm xảy ra tai nạn.
Longitude	Kinh độ của địa điểm xảy ra tai nạn.
LSOA of Accident Location	Khu vực thống kê địa phương cấp thấp (LSOA) nơi xảy ra tai nạn.
Number of Casualties	Số lượng thương vong trong vụ tai nạn.
Number of Vehicles	Số lượng phương tiện liên quan trong vụ tai nạn.
Pedestrian Crossing-Human Control	Kiểm soát viên.
Pedestrian Crossing-Physical Facilities	Cơ sở vật chất cho người đi bộ.
Police Force	Lực lượng cảnh sát quản lý khu vực xảy ra tai nạn.
Road Surface Conditions	Điều kiện bề mặt đường.
Road Type	Loại đường.
Special Conditions at Site	Các điều kiện đặc biệt tại nơi xảy ra tai nạn.
Speed Limit	Giới hạn tốc độ tại nơi xảy ra tai nạn (km/h).
Time	Thời gian xảy ra tai nạn.
Urban or Rural Area	Khu vực đô thị hoặc nông thôn nơi xảy ra tai nạn.
Weather Conditions	Điều kiện thời tiết tại thời điểm xảy ra tai nạn.
Year	Năm xảy ra tai nạn.
InScotland	Tai nạn xảy ra ở Scotland hay không.

Phương pháp Phân tích

Trong project này chúng ta sẽ sử dụng các giải thuật máy học khác nhau để dự đoán, ba giải thuật dùng trong project này là Random Forest, Logistic Regression và XGBoost. Với Target là Accident Severity và các Feature được chọn là Light Conditions, Number of Casualties, Number of Vehicles, Road Surface Conditions, Road Type, Speed limit, Urban or Rural Area, Weather Conditions.

- **Light Conditions**: Ảnh hưởng đến khả năng nhìn thấy và sự an toàn khi lái xe, với tai nạn thường xảy ra nhiều hơn trong điều kiện ánh sáng kém.
- **Number of Casualties**: Chỉ số quan trọng phản ánh mức độ nghiêm trọng của tai nạn, cho thấy số lượng người bị thương hoặc thiệt mạng.
- **Number of Vehicles**: Số lượng phương tiện tham gia có thể cho thấy mức độ phức tạp của tình huống, với các vụ tai nạn liên quan đến nhiều phương tiện thường nghiêm trọng hơn.
- **Road Surface Conditions**: Tình trạng mặt đường ảnh hưởng đến khả năng điều khiển phương tiện, với mặt đường trơn trượt làm tăng nguy cơ xảy ra tai nạn.
- **Road Type**: Loại đường ảnh hưởng đến tốc độ và cách thức điều khiển, với tai nạn trên đường cao tốc có thể nghiêm trọng hơn so với trên đường phố.
- **Speed Limit**: Giới hạn tốc độ quy định mức độ nhanh chậm cho phép, với tốc độ cao làm tăng mức độ nghiêm trọng của tai nạn.
- **Urban or Rural Area**: Đặc điểm khu vực (đô thị hay nông thôn) ảnh hưởng đến mật độ giao thông và các yếu tố khác có liên quan đến tai nạn.
- **Weather Conditions**: Điều kiện thời tiết ảnh hưởng đến khả năng lái xe và điều kiện đường, làm tăng khả năng xảy ra tai nạn và mức độ nghiêm trọng của chúng.

1. Random Forest

- **Ưu điểm**: Random Forest là một thuật toán mạnh mẽ và ổn định cho các bài toán phân loại. Nó sử dụng nhiều cây quyết định (decision trees) và kết hợp kết quả của chúng để cải thiện độ chính xác và giảm thiểu nguy cơ overfitting.
- **Khả năng xử lý dữ liệu không đồng nhất**: Random Forest có khả năng làm việc tốt với dữ liệu có tính không đồng nhất và tính chất phức tạp, ví dụ như các yếu tố trong tai nạn giao thông có thể là dạng định lượng (số lượng phương tiện) hoặc định tính (tình trạng mặt đường).
- **Tính dễ diễn giải**: Thuật toán này cũng dễ dàng giúp giải thích tầm quan trọng của các feature, giúp hiểu rõ yếu tố nào ảnh hưởng nhiều đến dự đoán.

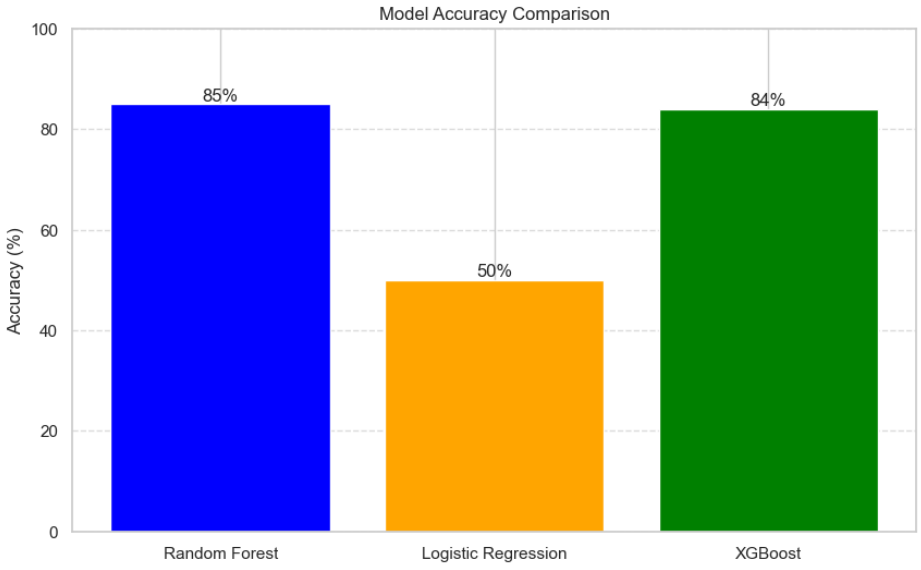
2. Logistic Regression

- **Tính đơn giản và dễ giải thích**: Logistic Regression là một thuật toán đơn giản nhưng hiệu quả trong các bài toán phân loại nhị phân (hoặc đa lớp khi mở rộng). Mô hình này giúp cung cấp dự đoán xác suất của các lớp, từ đó dễ hiểu hơn.
- **Khả năng tránh overfitting**: Với các thuật toán tuyến tính đơn giản như Logistic Regression, nguy cơ overfitting thấp hơn so với các mô hình phức tạp. Điều này giúp kiểm soát kết quả dự đoán với các bộ dữ liệu có thể bị nhiễu.
- **Hiệu quả trong dự đoán**: Logistic Regression có thể đưa ra kết quả nhanh chóng và chính xác, phù hợp với bài toán phân loại mức độ nghiêm trọng của tai nạn.

3. XGBoost

- **Hiệu năng cao**: XGBoost là một trong những thuật toán boosting tiên tiến nhất và được tối ưu để đạt hiệu suất cao trong thời gian ngắn, phù hợp cho các bộ dữ liệu lớn hoặc phức tạp.
- **Khả năng khai thác tính phức tạp của dữ liệu**: Thuật toán này có khả năng khai thác tối đa mối quan hệ phi tuyến giữa các feature, giúp cải thiện độ chính xác khi dự đoán mức độ nghiêm trọng của tai nạn.
- **Kiểm soát overfitting tốt**: Với khả năng điều chỉnh các tham số, XGBoost giúp giảm thiểu overfitting và cải thiện độ tổng quát của mô hình.

Hiệu suất của các mô hình



Phân chia công việc trong project

Họ và Tên	MSSV	Công việc
Cao Trần Tiến	SE184692	- Báo cáo project - Tìm tập dữ liệu liên quan - Giải thuật XGBoost
Nguyễn Hoàng Gia Huy	SE182631	- Giải thuật Logistic Regression
Đoàn Minh Đức	SE183915	- Làm việc với SQL - Data Visualization bằng Power BI
Lê Quang Tiến	SE182880	- Giải thuật Random Forest