US Accident Severity Prediction (2016–2023)

# Objective

This project aims to predict the severity level of traffic accidents across the United States using historical data from 2016 to 2023. The goal is to uncover key contributing factors, develop a reliable predictive model, and generate insights that can help reduce severe accidents and guide policy and safety interventions.

# Key Findings

* Weather Conditions:

Clear weather had the highest total accidents, but severe cases (severity 3 and 4) were more common in fog, thunderstorms, and heavy rain.

* Time of Day & Week:

Most accidents occurred on weekdays, particularly Wednesday–Thursday. However, evening hours (8–11 PM) showed a higher proportion of severe accidents.

* Hourly Risk Pattern:

Severity rates spiked after 10AM and 6 PM, especially between 10-12 AM and 6–10 PM, despite most volume occurring during morning rush hours.

* Monthly Trends:

Accident severity increased from spring to winter, peaking in November and December, potentially due to weather and holiday travel.

* Location Type:

Rural areas showed a 30% higher rate of severe accidents than urban zones, possibly due to higher speeds and limited infrastructure.

# Data & Modeling Approach

The dataset consists of millions of traffic accident records with attributes like weather, location type, time, road surface, and visibility. Severity is rated on a 1–4 scale, with 4 being the most severe. Due to hardware limitations, only 100,000 accident records were utilized for insights and modeling, as the local machine couldn’t efficiently process the full dataset.

To train the model:

- Data Preprocessing: Categorical columns are transformed to numeric values by target encoding method and Standard Scaler to standardize features.  
- Feature Selection: Top 30 features were selected using SHAP values, preserving nearly 100% of the model's predictive power.  
- Model Testing: Both Gradient Boosting Regressor (GBR) and Random Forest were trained. GBR performed better.  
- Hyperparameter Tuning: The tuned GBR model yielded:  
 - MAE: 0.048  
 - RMSE: 0.049  
 - R² Score: 0.805  
These results demonstrate strong model performance in predicting accident severity based on environmental and contextual factors.

# Next Steps

* Enhance data by adding vehicle type, demographics, and live traffic feeds.
* Improve facility and test advanced models for further accuracy improvement.
* Deploy the model into an interactive dashboard or API for public agencies.
* Collaborate with local governments to implement targeted safety campaigns and infrastructure improvements based on high-risk times and conditions.