UK Road Safety – Accident Severity Analysis

Project F7

KATERYNA KANTSYR
MELESSE PERSCHEWSKI

Why analyze UK road accidents?

- Thousands of accidents occur yearly in the UK, causing human and economic loss
- Data-driven insights can help prevent the most severe cases
- Our project aims **to analyze conditions that increase accident severity and build a predictive model.**

Goals

Analyse where, when and **why** accidents happen

Identify key factors influencing **the severity of an accident.**

Predict whether **an accident will be** ***Slight/ Serious/Fatal*** based on recorded conditions.

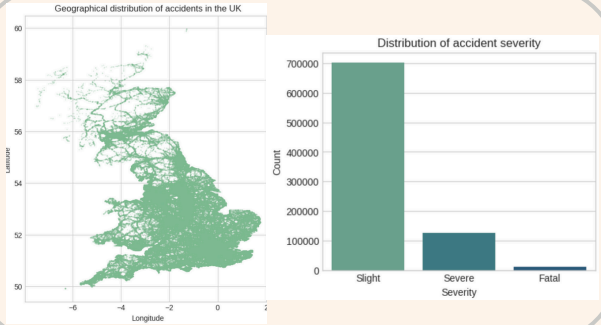
Data Overview

- **Dataset:** UK Department for Transport (Kaggle)
- **Period**: 2012–2017
- **Records**: ~830,000
- **Features**: 34 (Date, Time, Location, Weather, Lighting, Speed Limit, etc.)

Approach and Methodology

- The project combines **exploratory data analysis** and **predictive modeling** to understand and forecast accident severity in the UK.
- After initial data inspection, we performed **extensive data cleaning**, **removing invalid records**, **duplicates**, and **unrealistic numeric ranges**.
- **Categorical features** such as weather, lighting, and road conditions were examined, grouped where appropriate, and **one-hot encoded for modeling**.
- **Numeric features** were **standardized**, and **new engineered variables** were **introduced** (e.g., hour bins, weekday/weekend indicator) to capture temporal patterns.
- The dataset was then **split** into train, eval, and test sets (70/15/15), and **class imbalance** was addressed through **SMOTE oversampling** and **undersampling**

We tested several models, including **Random Forest**, **HistGradientBoosting**, and **Artificial Neural Network** (ANN), optimizing thresholds to maximize the F1-score.



Key Results

- **Model Performance**
  - **ANN achieved the highest recall** for severe accidents (0.82)
  - **Random Forest** provided **the best trade-off between accuracy and interpretability** (accuracy = 0.77, F1 = 0.63)
  - **Gradient Boosting** delivered **consistent** performance across all metrics (F1 = 0.61)
- **Feature Importance**
  - **Top predictors:** Number of Vehicles, Speed Limit, Urban/Rural Area, Time of Day, and Weather Conditions.
- **Exploratory Findings**
  - **Accidents peak on weekdays, especially Fridays and during daylight hours.**
  - **Most accidents occur under clear weather and dry roads,** highlighting behavioral rather than environmental causes.
  - **Urban areas** show more frequent but less severe accidents than **rural zones.**
- **Practical Implications**
  - The developed models help identify high-risk conditions and locations, enabling targeted interventions.

Data Science Methods

- **Data Preprocessing:** cleaned and validated data, removed duplicates, invalid coordinates, and unrealistic values; standardized formats and ensured consistency across variables.
- **Feature Engineering:** created temporal and contextual features (hour bins, weekday/weekend, daylight indicator) to enhance model interpretability.
- **Encoding and Scaling:** applied One-Hot Encoding for categorical variables and normalization for continuous fields.
- **Modeling:** trained and compared different models using stratified data and class balancing techniques.
- **Evaluation:** assessed models with Precision, Recall, F1-score, ROC-AUC, and Confusion Matrices to ensure balanced performance.
- **Feature Importance:** identified the strongest predictors of accident severity.