Predicting Injury Severity in Road Accidents: A Real-Time Classification Approach

Column Description and Feature Engineering

David Cebulla (1922129)
Gabriel Himmelein (1649181)
Lukas Ott (1842341)
Artur Loreit (2268917)
Aaron Niemesch (1836924)

October 25, 2025

Submitted to
Data and Web Science Group
Dr. Sven Hertling
University of Mannheim

1 What is the problem you are solving?

Modern assistive technologies in vehicles, such as the mandatory eCall system, automatically transmit accident data to emergency services. While this data includes location and passenger numbers, it lacks information on injury severity—a critical gap for first responders. The French National Interministerial Observatory for Road Safety (ONISR) database contains incident severity information concluded up to 30 days post-accident. Using this data as a baseline, our goal is to develop a classifier that can predict injury severity in real-time, helping first responders to take appropriate and timely precautions.

2 What data will you use?

Our analysis will use the "Annual Road Traffic Injury Databases" from 2019-2023, provided by ONISR. The data is supplied in four files per year (characteristics, locations, users, and vehicles) and includes features such as road conditions, weather, and user information. Recognizing the one-to-many relationship where each accident involves multiple participants, we will define our unit of analysis at the **individual level**. The `users` table will serve as the base for our dataset. For each user, we will merge the corresponding accident characteristics, location, and vehicle information using the `Num_Acc` key. For users without an associated vehicle, such as pedestrians, vehicle-specific features will be handled as a distinct 'Not Applicable' category, creating a comprehensive record for every individual.

3 How will you solve the problem?

3.1 What preprocessing steps will be required?

Our primary task is to predict the injury severity for each individual involved in an accident. The **target variable** is the ordinal `grav` feature, categorized into four classes of increasing severity: **Not Injured, Lightly Injured, Severely Injured, and Killed**. This individual-level approach directly addresses the relational data structure and allows the use of user-specific features.

We will use a chronological train-test split, with 2022 data for training and 2023 for testing. For hyperparameter optimization, we will use a **GroupKFold** cross-validation strategy with the accident ID (`Num_Acc`) as the group identifier. This ensures that participants from a single accident remain in the same fold, which handles their non-independence and leads to a more robust model evaluation.

Key preprocessing steps will include:

- **Real-Time Feature Selection**: A rigorous selection to prevent data leakage by exclusively using features available immediately at the accident scene.

- **Inter-Vehicle Feature Engineering**: To account for interactions between vehicles, we will engineer features summarizing the context of other parties involved, such as the number and types of other vehicles (e.g., 'truck_involved', 'motorcycle_involved').
- **Handling Class Imbalance**: Our primary strategy is using the built-in `class_weight` model parameter. If needed, we will explore advanced oversampling techniques like SMOTE as a secondary step.
- **Data Cleaning and Formatting**: This includes consistent handling of null values, correcting input errors, and formatting time fields.
- **Dimensionality Reduction**: We will evaluate **Principal Component Analysis (PCA)** to manage potential multicollinearity.

3.2 Which algorithms do you plan to use?

Our methodology integrates unsupervised and supervised learning, with clustering serving as a key step in feature engineering. After applying PCA, we will use **K-Means Clustering** and **DBSCAN** on the principal components to identify "accident personas." We will ensure these clusters are interpretable through "cluster profiling" with the original features. The resulting cluster assignments may then be engineered into a new categorical feature to potentially enhance our classification models.

Subsequently, we will train and evaluate several **supervised classification models**, prioritizing those that can leverage the ordinal nature of our target variable:

- **Ordinal Logistic Regression**: An interpretable statistical baseline that respects the ordinal data structure.
- A **Random Forest Classifier** and **Gradient Boosting** (e.g., XGBoost): Powerful ensemble models for high performance and feature importance rankings.
- **Ordinal Forest**: A specialized version of the Random Forest designed for ordinal outcomes, which we will compare against the standard implementation.

This selection allows us to test if ordinal-aware models provide an advantage. Finally, a **Feature Importance Analysis** using the tree-based models will provide a data-driven ranking of significant risk factors.

4 How will you measure success?

For clustering, success will be measured by the **Silhouette Score** and a qualitative review of the interpretability of the identified "accident scenarios."

For classification, all models must significantly outperform a **"Most Frequent Class"** baseline. As simple accuracy is an insufficient metric for our skewed data, our evaluation will be based on a suite of metrics. We will use **confusion matrices** for error analysis

and the **weighted F1-Score** to balance precision and recall. Given the ordinal nature of the target variable, we will also use metrics that penalize distant errors more heavily, such as **Weighted Cohen's Kappa**. For visualization, we will primarily generate **Precision-Recall Curves**, which are well-suited for imbalanced data.

5 What do you expect your results to look like?

Our primary goal is to create a model capable of classifying injury severity to a satisfactory degree. We expect the clustering analysis to uncover specific, evidence-based "accident personas," such as low-speed urban collisions, high-velocity rural incidents, and collisions involving vulnerable road users. From the feature importance analysis, we expect to confirm that vehicle type (motorcycle vs. car) is a dominant predictor of severe injury. We also anticipate that factors like road category, lighting conditions, driver's age, and the use of safety equipment will be highly significant in predicting an individual's outcome.