

Project Report

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

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

November 23, 2025

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

1 Application Area and Goals (0.5 pages)

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 Structure and Size of the Data Set

Our analysis uses 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 defined our unit of analysis at the **individual level**. The **users** table serves as the base for our dataset.

Directly after merging the accident characteristics, locations, users, and vehicle tables, and prior to any further preprocessing steps, the consolidated dataset comprised a total of 619,817 records and 71 columns. Processing the yearly data reveals an annual volume ranging between approximately 105,000 and 133,000 records per year. The final consolidated dataset was then partitioned into a training and validation set of **493,214 samples** (covering 2019–2022) and a test set of **125,505 samples** (covering 2023).

The features we have selected are listed below. They are categorized into Original Features (retained from raw data) and Engineered Features (calculated to model complex relationships). The final dataset consists of **57 features**, comprising **23 original columns** retained from the raw data and **34 engineered features** calculated to model complex relationships.

Time context

{time_of_day, hour_sin/cos, day_of_week_sin/cos, month_sin/cos, day_of_year_sin/cos}

Captures seasonality and daily patterns using cyclical sine/cosine transformations for time units. Also includes **time_of_day**, which groups hours into broader categories (e.g., Night, Rush Hour).

Geospatial location

{latitude, longitude}

GPS coordinates (WGS84) utilized for spatial analysis and mapping.

Environment & Roadway

{lighting_ordinal, weather_ordinal, location, infrastructure, accident_situation, horizontal_alignment, reserved_lane_present, speed_limit, road_complexity_index, surface_quality_indicator}

Combines physical site attributes (urban/rural status, alignment, infrastructure type) with risk-based ordinals for lighting and weather. Engineered indices quantify road complexity (0–10 scale) and surface quality.

Crash dynamics & Maneuvers

{type_of_collision, initial_point_of_impact, fixed_obstacle_struck, mobile_obstacle_struck, main_maneuver_before_accident, impact_score, impact_delta}

Describes the collision mechanics, including maneuvers, impact points, and obstacles. Derived metrics like `impact_score` and `impact_delta` quantify the relative risk based on vehicle size differences.

Vehicle attributes & Involvement

{motor_type, vehicle_category_simplified, vehicle_category_involved_[type]}

Specifies the primary vehicle's type and motorization. Includes binary flags indicating if specific other vehicle types (e.g., heavy trucks, bicycles, buses) were involved in the accident.

Personal attributes

{role, sex, age, age_group, position, pedestrian_location, pedestrian_action}

Covers demographic data (age, sex), the user's role (driver, passenger, pedestrian), seating position, and specific actions or locations for pedestrians at the time of the accident.

Safety equipment usage

{used_belt, used_helmet, used_child_restraint, used_airbag}

Binary variables indicating the use of protective gear (seatbelts, helmets) or the deployment of airbags.

Accident Persona Clustering

{cluster}

A categorical feature derived from unsupervised K-Prototypes clustering. It groups accidents into distinct "personas" (e.g., 0, 1, 2) based on a combination of numerical and categorical attributes to capture complex, non-linear patterns in the data.

Target / Outcome

{injury_target}

The engineered ordinal target variable classifying injury severity into three levels: 0 (Uninjured), 1 (Lightly Injured), and 2 (Hospitalized or Dead).

3 Preprocessing

Given the nature of the "Annual Road Traffic Injury Database" as a raw database output provided by ONISR in multiple separate tables with complex many-to-one relationships, extensive preprocessing was required. The pipeline is designed to transform the disjointed raw data into a singular, model-ready tabular representation.

3.1 Data Standardization and Key Generation

The initial step involved normalizing column names to English equivalents. A critical challenge was the inconsistency in user identification across years. Data from 2022 onwards included a distinct `id_user`, whereas data from 2019 and 2020 did not. To ensure a consistent unit of analysis across all years, we implemented a synthetic key generation strategy. For older data, we generated a unique identifier by combining the accident ID with a cumulative count of users within that accident (e.g., 2019000001_U1). This ensured that every individual involved in an accident could be uniquely tracked and merged with their respective vehicle and accident characteristics.

3.2 Advanced Data Merging strategies

A user-centric view was adopted for merging, treating each participant as an independent instance. While accident characteristics (weather, time) could be joined directly, the relationship between users, their vehicles, and opposing vehicles required complex logic.

Vehicle Antagonist Resolution

A significant predictor of injury severity is the disparity between the user’s vehicle and the ”opposing” entity (the antagonist). Since a simple join cannot determine which of the multiple vehicles in an accident caused the injury, we engineered a selection algorithm. We assigned an `impact_score` to vehicle categories based on mass and risk (e.g., HGV Truck = 6, Bicycle = 2). For multi-vehicle accidents, the pipeline identifies the ”antagonist” vehicle as the one involved in the same accident with the highest impact score, excluding the user’s own vehicle. For pedestrians, the striking vehicle is explicitly identified. This allows us to calculate an `impact_delta`, representing the structural disadvantage of the user (e.g., a cyclist hit by a truck results in a high negative delta).

Location Deduplication

Contrary to the dataset description, multiple location entries were found for single accident IDs, likely due to first responders logging entries for every intersecting street. To resolve this, we implemented a `completeness_score`. We assigned weights to critical columns (Road Category: 2.0, Speed Limit: 2.0, others: 1.0). For each accident, the location entry with the highest weighted score, indicating the most data-rich description of the scene, was selected, ensuring the model trains on the highest quality data available.

3.3 Handling Missing Values (Imputation Strategy)

We distinguished between ”structural missingness” (values that should not exist) and ”data quality missingness” (values that are unknown).

- **Structural Missingness:** Pedestrians, by definition, do not have a vehicle category or motor type. For these cases, we explicitly imputed a value of `-1` or `'none'`

to indicate "Not Applicable," preventing the model from treating these as missing data.

- **Data Quality Missingness:** For users who *should* have data (e.g., drivers) but lack it, we imputed a value of 0 or 'Unknown'.
- **Other Vehicle Imputation:** If no opposing vehicle was involved, columns related to the "other" vehicle were set to -1. However, if a second vehicle ID existed but its characteristics were missing, we imputed **Unknown** to differentiate this state from single-vehicle accidents.

Finally, rows missing the target variable `injury_severity` were dropped, as they cannot be used for supervised learning.

3.4 Feature Engineering

To capture complex non-linear relationships, we generated 34 new features across four domains.

Temporal Cyclical Features

Raw timestamps are ill-suited for many models due to the discontinuity between 23:59 and 00:00. We decomposed time into cyclical components using Sine and Cosine transformations for hours, days of the week, and months. Additionally, we bucketed hours into a `time_of_day` feature (Night, Morning Rush, Midday, Evening Rush) to assist tree-based models in identifying high-level patterns.

Road Complexity Index

We hypothesized that complex road environments increase accident probability but might decrease severity due to lower speeds. We engineered a `road_complexity_index`, a composite score normalized between 0 and 10. This index aggregates weighted scores from:

- **Intersection Type:** High weights for roundabouts and multi-branch intersections.
- **Road Category:** Higher weights for urban communal ways vs. motorways.
- **Traffic Regime:** Penalties for variable assignment lanes.
- **Lane Count:** Higher complexity for multi-lane roads.

Complementing this, a binary `surface_quality_indicator` was created, set to 1 only if both the pavement condition was normal and the longitudinal profile was flat.

Vehicle and User Attributes

Vehicle categories were simplified from over 30 specific codes into 6 broad classes (Bicycle, Powered 2-3 Wheeler, Light Motor Vehicle, HGV/Truck, Bus/Coach, Other) to reduce dimensionality. For users, we transformed the `year_of_birth` into an `age` feature and further binned it into sociologically relevant `age_groups` (e.g., Child/Teen, Senior). Safety equipment flags (seatbelts, helmets, airbags) were consolidated from three separate columns into binary "Used/Not Used" indicators to resolve data sparsity.

3.5 Feature Selection & Cleaning

Following engineering, the dataset underwent a rigorous cleaning process. Invalid data, such as speed limits exceeding 130 km/h or negative values for age, were filtered out. We removed high-cardinality identifiers (IDs, address strings) and columns with excessive missingness (e.g., `width_central_reservation`) that offered little predictive value. Crucially, the raw target variable `injury_severity` (4 classes) was re-mapped to an ordinal `injury_target` (0: Uninjured, 1: Light Injury, 2: Hospitalized/Dead) to address class imbalance and better reflect the triage needs of first responders. The final dataset consists of 57 refined features ready for the modeling pipeline.

4 Data Mining

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.

5 Evaluation

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.

6 Results

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.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
Gemini 2.5/3 Pro	Rephrasing	Throughout	+++
Gemini 2.5/3 Pro	Code Generation and Code Debugging	Throughout	+++

Unterschrift

Mannheim, den 30. Oktober 2025