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 (1 page min)

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.

TODO: Size of one year and size of training/val and test set.

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 **56 features**, comprising **23 original columns** retained from the raw data and **33 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.

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

3.1 Data Merging

Given the nature of the "Annual Road Traffic Injury Database" as a raw database output provided by OSNIR in multiple separate tables with existing many-to-one relationships between them, extensive preprocessing procedures, besides the renaming of columns to English equivalents, needed to be implemented to arrive at a singular tabular representation, compatible with our chosen data mining models.

While accidents in the dataset are uniquely identified via an ID column in each table easing the process, strategies to merge the information contained in the multiple rows stilled needed to applied. A naive encoding using additional columns for every entry was deemed infeasible due to the variance in relation counts observed. Instead, a user-centric view representing the accident from the perspective of each participant was adapted.

Adopting this approach, using the user table as basis for data frame merging, meant that the many-to-one relation between the general accident circumstances and the users could be resolved via a simple join operation, as the perspective of each participant were

to be treated independently of each other. While initially sparking concerns regarding overfitting and data leakage, both were addressed through starkly differing injury scores between participants of the same accident and a temporal split for the test data.

Since every participant can at most be related to one vehicle, a left join of the data suffices per individual, but looses valuable information regarding the vehicles involved in the accident as a whole. This data however is crucial for data mining models as the risk of injury for a specific participant is directly correlated with their position in it. As an example for this correlation, a pedestrian being hit by a truck will have a higher risk of injury compared to one being hit by a bike. To capture this information per single row, new features were engineered. While the first captures the number of vehicles types involved in a particular accident, the second contains the vehicle information pertaining to the "antagonist" of the participant, defined as either the vehicle a pedestrian was hit by, which is explicitly reported in the dataset, or as the most impactful vehicle of a different participant. As an example, if choosing between a bike and a truck, the truck will be considered the "antagonist".

Contrary to the dataset description provided by OSNIR, multiple location entries were found to correspond to the same accident entries. Analysis revealed that this artifact correlated with the presence of intersections, forming the theory that first responders entered all streets intersecting at this location instead of only the main road. A spot-check of randomly selected accident GPS location data also seems to support this. Since all location entries are thus valid, one of the available location entries is selected during merging based on the completeness of the entry.

3.2 Feature Selection & Data Cleaning

Besides the many-to-one anomalies observed in the raw dataset during the previous step, attribute values themselves also exhibited irregularities that needed to be removed. The anomalies observed can be categorized into three types: missing, invalid and irrelevant values.

Missing & Irrelevant Values

Missing Values in the raw dataset are denoted by either empty cells or the value "-1". Since their presence, according to the official documentation, is sometimes used to encode special meaning depending on the attribute in question (e.g. NA values for sex and birth year for hit-and-run drivers), special care needed to be taken to evaluate each attribute individually.

Table 1 offers an overview over the handling of specific attribute missing values. While the category "Little information / Many missing values" drops columns with significant gaps in knowledge (e.g. carriageway_width with >50%), columns pertaining to the use case are eliminated as they are not realistically available in the case of real time prediction. In addition, IDs and highly unique attributes are removed as patterns applicable for classification are unlikely to be inferred from them.

Category	Columns
Special meaning	year_of_birth, sex
IDs and high-cardinality identifiers	id_accident, id_vehicle, id_user, number_vehicle, department_code, commune_code, postal_address, road_number, road_number_index, road_number_letter, latitude, longitude
Little information / Many missing values	width_central_reservation, nearest_reference_marker_distance, nearest_reference_marker, carriageway_width, number_occupants_in_public_transport
Use Case	injured_pedestrian_alone, trip_purpose, pedestrian_location, pedestrian_action

Table 1: Attributes with Special Meaning and Dropped Columns

Invalid Values

Analysis of the dataset revealed attributes that contained anomalous values. While the attribute number_of_traffic_lanes contained cells with the value #VALEURMULTI, an artifact from the database export, speed limits above the legal maximum (>140) were also observed, necessitating removal. Besides the dropping of rows, whitespace in attributes such as IDs were also removed for more efficient merging.

3.3 Feature Engineering

4 Preprocessing (1 page min)

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.

For each user, we merged 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 are handled as a distinct 'Not Applicable' category, creating a comprehensive record for every individual.

We use a chronological train-test split, with 2019-2022 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.

5 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.

6 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.

7 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.