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

# Feature Description

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 Introduction

This document describes all features available in the final dataset (after the `F_feature_selection` step) used for model training. The features are divided into two main groups:

1. **Original Features:** Columns originating from the raw datasets (renamed in `a_rename.py`) that were retained.
2. **Engineered Features:** New columns calculated in `c_feature_engineering.py` or `b_merge_tables.py` to model complex relationships.

Both groups are further subdivided by data type: **Nominal** (Categorical, no order), **Ordinal** (Categorical, with order), and **Numerical** (Continuous).

# 2 Target Variable (The Feature to Predict)

This is the single, engineered column that the model is trained to predict.

## 2.1 Ordinal

- `injury_target`: (Engineered) Our new, ordinal target variable representing the severity of the injury. It is derived from the original `injury_severity` column.
  - `0`: Uninjured (from Original: 1)
  - `1`: Lightly Injured (from Original: 4)
  - `2`: Severe (Hospitalized or Killed) (from Original: 2, 3)

# 3 Original Features (Retained from Source Data)

These columns are loaded from the raw data, renamed, and are retained after all processing and feature selection steps.

## 3.1 Nominal Features (Categorical)

- `location`: Location of the accident. *(Original: `agg`)*
  - `1`: Outside built-up area
  - `2`: In built-up area
- `type_of_collision`: Type of collision. *(Original: `col`)*
  - `-1`: Not specified (imputed)
  - `1`: Two vehicles - frontal
  - `2`: Two vehicles - from behind
  - `3`: Two vehicles - from the side

  - `4`: Three vehicles and more - in chain
  - `5`: Three vehicles and more - multiple collisions
  - `6`: Other collision
  - `7`: Without collision

- **`reserved_lane_present`**: Indicates the existence of a reserved lane. *(Original: `vosp`)*
  - `-1`: Not specified
  - `0`: Not applicable
  - `1`: Bicycle path
  - `2`: Bicycle lane
  - `3`: Reserved lane

- **`horizontal_alignment`**: Horizontal alignment (plan) of the road. *(Original: `plan`)*
  - `-1`: Not specified
  - `1`: Straight section
  - `2`: Left curve
  - `3`: Right curve
  - `4`: In "S" shape

- **`infrastructure`**: Special infrastructure at the accident site. *(Original: `infra`)*
  - `-1`: Not specified
  - `0`: None
  - `1`: Underground - tunnel
  - `2`: Bridge - overpass
  - `3`: Interchange ramp or connection
  - `4`: Railway
  - `5`: Equipped junction
  - `6`: Pedestrian zone
  - `7`: Toll zone
  - `8`: Construction site
  - `9`: Others

- **`accident_situation`**: Situation of the accident. *(Original: `situ`)*
  - `-1`: Not specified
  - `0`: None

  - `1`: On roadway
  - `2`: On emergency lane
  - `3`: On shoulder
  - `4`: On sidewalk
  - `5`: On bicycle path
  - `6`: On other special lane
  - `8`: Others

- **`sex`**: Sex of the user. *(Original: `sexe`)*
  - `1`: Male
  - `2`: Female

- **`pedestrian_location`**: Location of the pedestrian at the time of the accident. *(Original: `locp`)*
  - `-1`: Not specified
  - `0`: Not applicable
  - `1`: On roadway, >50m from pedestrian crossing
  - `2`: On roadway, <50m from pedestrian crossing
  - `3`: On pedestrian crossing, without traffic light
  - `44`: On pedestrian crossing, with traffic light
  - `5`: On sidewalk
  - `6`: On shoulder
  - `7`: On refuge or emergency lane
  - `8`: On parallel lane
  - `9`: Unknown

- **`pedestrian_action`**: Action of the pedestrian. *(Original: `actp`)*
  - `-1`: Not specified
  - `0`: Not specified or not applicable
  - `1`: Moving in the same direction as the striking vehicle
  - `2`: Moving in the opposite direction of the vehicle
  - `3`: Crossing
  - `4`: Masked / Hidden
  - `5`: Playing - running
  - `6`: With animal
  - `9`: Other

  - `A`: Getting on/off vehicle
  - `B`: Unknown

- **`fixed_obstacle_struck`**: Fixed obstacle struck by the primary vehicle. *(Original: `obs`)*. Imputed with -1 (N/A) for pedestrians.
  - `-1`: Not specified / N/A
  - `0`: Not applicable
  - `1`: Parked vehicle
  - `2`: Tree
  - `3`: Metal guard rail
  - `4`: Concrete guard rail
  - `5`: Other guard rail
  - `6`: Building, wall, bridge pier
  - `7`: Road sign support or emergency call box
  - `8`: Post
  - `9`: Street furniture
  - `10`: Parapet
  - `11`: Island, refuge, high bollard
  - `12`: Curb
  - `13`: Ditch, embankment, rock wall
  - `14`: Other fixed obstacle on roadway
  - `15`: Other fixed obstacle on sidewalk or shoulder
  - `16`: Road exit without obstacle
  - `17`: Culvert

- **`mobile_obstacle_struck`**: Mobile obstacle struck by the primary vehicle. *(Original: `obsm`)*. Imputed with -1 (N/A) for pedestrians.
  - `-1`: Not specified / N/A
  - `0`: None
  - `1`: Pedestrian
  - `2`: Vehicle
  - `4`: Vehicle on rail
  - `5`: Domestic animal
  - `6`: Wild animal
  - `9`: Other

- **initial_point_of_impact**: Initial point of impact on the primary vehicle. *(Original: choc)*. Imputed with -1 (N/A) for pedestrians.
  - -1: Not specified / N/A
  - 0: None
  - 1: Front
  - 2: Front right
  - 3: Front left
  - 4: Rear
  - 5: Rear right
  - 6: Rear left
  - 7: Right side
  - 8: Left side
  - 9: Multiple impacts (rollover)
- **main_maneuver_before_accident**: Main maneuver of the primary vehicle. *(Original: manv)*. Imputed with -1 (N/A) for pedestrians/passengers, 0 (Unknown) for drivers.
  - -1: Not specified / N/A
  - 0: Unknown
  - 1: Without change of direction
  - ... (Full list in original file)
  - 26: Other maneuvers
- **motor_type**: Motorization type of the primary vehicle. *(Original: motor)*. Imputed with -1 (N/A) for pedestrians/passengers, 0 (Unknown) for drivers.
  - -1: Not specified / N/A
  - 0: Unknown
  - 1: Hydrocarbon (Gasoline/Diesel)
  - 2: Hybrid electric
  - 3: Electric
  - 4: Hydrogen
  - 5: Human (e.g., bicycle)
  - 6: Other

  **fixed_obstacle_struck_other**, **mobile_obstacle_struck_other**, **initial_point_of_impact_oth**
  **main_maneuver_before_accident_other**, **motor_type_other**: *Note on _other columns:* These features describe the **highest-impact "other" vehicle** involved

in the accident (determined in `b_merge_tables.py`). They follow the same code definitions as their primary counterparts. In `d_handle_missing_values.py`, `NaN` values are imputed with `-1` (N/A) if no second vehicle was involved, or `0` (Unknown) if a second vehicle was present but the data was missing.

### 3.2 Ordinal Features (Ordered Categories)

- **`position`**: Position occupied by the user in the vehicle. *(Original: `place`)*. (e.g., `1`: Driver, `2-9`: Passenger seats).

### 3.3 Numerical Features (Continuous)

- **`latitude`**: Latitude (WGS84). *(Original: `lat`)*
- **`longitude`**: Longitude (WGS84). *(Original: `long`)*
- **`speed_limit`**: Authorized speed limit at the accident site. *(Original: `vma`)*

## 4 Engineered Features (Created by the Pipeline)

These columns are newly calculated in `c_feature_engineering.py` or `b_merge_tables.py`.

### 4.1 Nominal Features (Categorical)

- **`time_of_day`**: Categorical time bucket derived from `hour`.
  - `'Night'`: 00:00 - 05:59, 20:00 - 23:59
  - `'Morning_Rush'`: 06:00 - 09:59
  - `'Midday'`: 10:00 - 15:59
  - `'Evening_Rush'`: 16:00 - 19:59
- **`age_group`**: Age bracket derived from `age`.
  - `'child_teen'`: 0-17
  - `'young_adult'`: 18-24
  - `'adult'`: 25-39
  - `'middle_aged'`: 40-64
  - `'senior'`: 65+
  - `'Unknown'`: Imputed value
- **`role`**: Simplified user role, derived from `user_category`.
  - `'driver'`: (Original: 1)
  - `'passenger'`: (Original: 2)
  - `'pedestrian'`: (Original: 3)

  - `'other'`: (Imputed value)

- **`vehicle_category_simplified`**: Simplified vehicle category for the primary vehicle. (e.g., `'light_motor_vehicle'`, `'hgv_truck'`, `'bicycle'`, `'unknown'`, etc.)
- **`vehicle_category_simplified_other`**: Simplified vehicle category for the *other* vehicle. (Includes `'none'` for no other vehicle).
- **`used_belt`**: (Binary) 1 if user used a seatbelt, 0 otherwise.
- **`used_helmet`**: (Binary) 1 if user used a helmet, 0 otherwise.
- **`used_child_restraint`**: (Binary) 1 if a child restraint was used, 0 otherwise.
- **`used_airbag`**: (Binary) 1 if airbag was deployed/used, 0 otherwise.
- **`surface_quality_indicator`**: (Binary) 1 if `pavement_condition` = "Normal" (1) AND `longitudinal_profile` = "Flat" (1), 0 otherwise.
- **`vehicle_category_involved_bicycle`**: (Binary) 1 if at least one bicycle was involved in the accident, 0 otherwise.
- **`vehicle_category_involved_bus_coach`**: (Binary) 1 if a bus/coach was involved.
- **`vehicle_category_involved_hgv_truck`**: (Binary) 1 if a heavy truck was involved.
- **`vehicle_category_involved_light_motor_vehicle`**: (Binary) 1 if a light motor vehicle/car was involved.
- **`vehicle_category_involved_other`**: (Binary) 1 if an "other" vehicle type was involved.
- **`vehicle_category_involved_powered_2_3_wheeler`**: (Binary) 1 if a moped/motorcycle was involved.
- **`cluster`**: Assignment to one of 3 "accident personas" identified via unsupervised K-Prototypes clustering. Used to capture complex non-linear relationships between attributes.
  - `0`: Persona 0
  - `1`: Persona 1
  - `2`: Persona 2

### 4.2 Ordinal Features (Ordered Categories)

- **`day_of_week`**: Day of the week, where Monday=0 and Sunday=6.
- **`lighting_ordinal`**: A new ordinal scale for lighting conditions (risk-based).
  - `0`: Good (Original: 1 - Full day)
  - `1`: Medium (Original: 5 - Night, light on)
  - `2`: Poor (Original: 2 - Twilight)
  - `3`: Very Poor (Original: 3, 4 - Night, light off/none)
- **`weather_ordinal`**: A new ordinal scale for weather conditions (risk-based).
  - `0`: Good (Original: 1 - Normal)
  - `1`: Okay (Original: 8 - Overcast)
  - `2`: Slight Risk (Original: 2 - Light rain, 7 - Dazzling)
  - `3`: Medium Risk (Original: 6 - Wind, 3 - Heavy rain)
  - `4`: High Risk (Original: 5 - Fog, 4 - Snow)
- **`road_complexity_index`**: Index (scaled 0-10) assessing road complexity. Based on `intersection`, `road_category`, `traffic_regime`, and `number_of_traffic_lanes`. Higher value = more complex.
- **`impact_score`**: Weighted score (0-6) based on `vehicle_category_simplified`. (e.g., Truck=6, Car=4, Bicycle=2, unknown=1).
- **`impact_score_other`**: Weighted score (0-6) for the *other* vehicle. (e.g., Truck=6, Car=4, n/a=0).
- **`impact_delta`**: The *directional* difference: `impact_score` - `impact_score_other`. A negative value implies higher risk (e.g., Car vs. Truck = 4 - 6 = -2).

### 4.3 Numerical Features (Continuous)

- **`age`**: User's age at the time of the accident. Imputed with 0 if unknown.
- **`hour_sin`** / **`hour_cos`**: Cyclical features (Sine/Cosine) for the hour of the day.
- **`day_of_week_sin`** / **`day_of_week_cos`**: Cyclical features for the day of the week.
- **`month_sin`** / **`month_cos`**: Cyclical features for the month.
- **`day_of_year_sin`** / **`day_of_year_cos`**: Cyclical features for the day of the year.