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 intelligent and connected vehicle systems, such as the mandatory European eCall service, are designed to automatically transmit crucial accident data to emergency centers when a crash occurs. These transmissions typically include location, timestamp, and the number of passengers but lack information on the severity of injuries: a critical shortcoming for emergency services, as this information is vital for prioritizing rescue efforts and optimizing medical response times.

To address this gap, we focus on leveraging historical accident data provided by the French National Interministerial Observatory for Road Safety (ONISR). This organization maintains the official "Bulletins d'Analyse des Accidents Corporels de la Circulation" (BAAC), a national database that records all injury accidents on public roads in France. The dataset captures detailed multi-table information covering the circumstances of each accident (Caractéristiques), the location and infrastructure (Lieux), the vehicles involved (Véhicules), and the individual users (Usagers). These data entries include rich contextual variables such as road category, lighting and weather conditions, vehicle type, and the user's role and behavior at the time of the accident.

Each accident record specifies the injury outcome (grav) for every participant, coded as uninjured, lightly injured, hospitalized, or killed. Although the ONISR database provides this information retrospectively (sometimes up to 30 days after the event) it represents a uniquely valuable source for supervised learning aimed at approximating these outcomes in real time.

The goal of this project is to develop a machine learning classifier capable of predicting injury severity immediately after an accident, based solely on the information that would realistically be available to emergency responders or vehicle telematics systems at the moment of impact. Such a model can significantly enhance first-responder coordination by providing an automated injury risk assessment, enabling faster triage and more efficient resource allocation.

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.

Aspect	Features	Description
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. Includes time of day categories.
Geospatial	latitude, longitude	GPS coordinates (WGS84) for spatial analysis.
Environment	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 with risk-based ordinals. Engineered indices quantify road complexity and surface quality.
Crash dynamics	type_of_collision, initial_point_of_impact, fixed_obstacle_struck, mobile_obstacle_struck, main_maneuver_before_accident, impact_score, impact_delta	Describes collision mechanics. Derived metrics quantify relative risk based on vehicle size differences.
Vehicle attributes	motor_type, vehicle_category_simplified, vehicle_category_involved[type]	Specifies primary vehicle type. Includes binary flags for involvement of specific other vehicle types.
Personal attributes	role, sex, age, age_group, position, pedestrian_location, pedestrian_action	Covers demographic data, user's role, seating position, and pedestrian actions.
Safety equipment	used_belt, used_helmet, used_child_restraint, used_airbag	Binary variables indicating use of protective gear or airbag deployment.
Clustering	cluster	Categorical feature from K-Prototypes clustering, grouping accidents into distinct "personas".
Target	injury_target	Engineered ordinal target: 0 (Uninjured), 1 (Lightly Injured), 2 (Hospitalized/Dead).

Table 1: Feature overview and descriptions from the ONISR accident dataset.

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

4.1 Experimental Setup

We partition the dataset into a training set (2019–2022) and a test set (2023). Our primary evaluation metric is the macro F1-score, chosen to prioritize recall for the critical "severely injured" class without introducing the sensitivity of arbitrary cost matrices. Instead of using a simple majority-vote baseline, we implemented a domain-specific baseline that relies exclusively on the speed-limit feature to predict accident severity:

$$\hat{y} = \begin{cases} \text{uninjured}, & \text{if speed\_limit} \leq 50, \\ \text{injured}, & \text{if speed\_limit} < 100, \\ \text{severely injured}, & \text{otherwise.} \end{cases}$$

Given the ordinal nature of the target variable, we also considered weighted Cohen's kappa. However, considering the high cost of missing a severe injury in an emergency response context, optimizing macro-F1 ensures that the minority class (severe injuries) is not overwhelmed by the majority class (uninjured).

Model	Hyperparameter	Search Space	Optimal Value
Ridge	alpha	$\{0.05, 0.10, \ldots, 10\}$	1.5
BRF	n_estimators	$[50, 400] \cap \mathbb{Z}$	400
	max_depth	$\{3, \ldots, 20\}$	18
	min_samples_leaf	$\{1, \ldots, 20\}$	1
	replacement	{`True`, `False`}	`False`
	sampling_strategy	{`all`, `not minority`}	`all`
CatBoost	iterations	$\{1000, \ldots, 5000\}$	4139
	learning_rate	$[0.01, 0.2]$	0.01
	depth	$\{4, \ldots, 10\}$	10
	l2_leaf_reg	$[10^{-2}, 10]$	0.1887
	border_count	$\{32, \ldots, 255\}$	255

Table 2: Search space and optimal hyperparameters for Ridge, BRF and CatBoost.

4.2 Classification: Model Selection

To gain an initial understanding of the difficulty of the prediction task, we experimented with a broad range of machine learning models while applying only minimal hyperparameter tuning including models based on logistic regression (ordinal, lasso, ridge), ensembles (Random Forests, HistGradientBoosting, CatBoost) and simple neural networks. Across these experiments, we observed that all machine learning models substantially outperformed the baseline method. While originally considered, methods explicitly designed to leverage the ordinal structure of the target variable did not achieve better performance compared to other approaches in terms of both F1 and misclassifications error types, eliminating them from consideration. Applying Occam's Razor, our focus was narrowed to three models for fine-tuning and selection: Ridge Classification, Balanced Random Forest and CatBoost. While the first is Logistic Regression with applied L2-regularization and the second a random forest with rebalancing during subsampling, CatBoost is a boosting algorithm similar to XGBoost chosen for its improved support for categorical data, fitting to our data composition.

Considering the size of our dataset and computational constraints, model selection in terms of hyperparameter tuning was conducted for each model using 3-fold cross validation with shuffling to avoid biases from the original ordering, utilizing Bayesian Optimization and F1-Macro for the ensembles. While Ridge Regression could only be tuned regarding its regularization strength α, both ensemble methods offered more options regarding tree pruning relevant for avoiding overfitting. Table 2 shows the hyperparameters considered. Approaches selection was similarly achieved using both F1-Macro and confusion matrices, used to evaluate recall/misclassifications on (high) injury cases, with the same 3-fold cross validation.

4.3 Clustering Strategy

To identify distinct "Accident Personas," we explored four clustering algorithms, each chosen for a specific capability:

- **K-Prototypes:** A hybrid extension of K-Means that handles mixed data types (numerical and categorical) natively. This is critical for our dataset, where features like 'road category' or 'vehicle type' carry significant semantic weight that cannot be captured by Euclidean distance alone.
- **HDBSCAN:** A density-based algorithm chosen for its ability to detect noise and clusters of varying shapes. Unlike partition-based methods, it does not force every point into a cluster, which is useful for identifying "outlier" accident types.
- **Agglomerative Clustering:** Used with Gower distance to explore hierarchical structures in the mixed dataset, offering a visual way to assess natural groupings via dendrograms.

Cluster Profiling Methodology: To derive meaningful profiles from the clustered dataset, we conducted a structured categorical overrepresentation analysis. For each cluster and each categorical feature, we computed the conditional distribution $P(X = x \mid C = c)$ and compared it to the global distribution $P(X = x)$. A category was considered *cluster-characteristic* if it satisfied the following criteria:

- **Lift filter:** A minimum overrepresentation of $\text{lift}(x, c) = \frac{P(X=x|C=c)}{P(X=x)} > 1.5$.
- **Support filter:** The category must represent more than 3% of observations within the cluster.
- **Dominance filter:** Features were excluded if the same dominant category appeared in all clusters (e.g., "Daylight").

5 Evaluation

5.1 Hyperparameter Tuning

We optimized hyperparameters using macro-F1 as the objective function. For the **Ridge classifier**, a grid search yielded an optimal regularization parameter $\alpha = 2.0$. For the ensemble models, we employed Bayesian Optimization (30 iterations) with stratified 3-fold cross-validation.

The tuning process for the **Balanced Random Forest** revealed that a high model complexity was necessary to capture the nuances of severe accidents. The optimal configuration required a large number of estimators ($n = 400$) and a significant maximum depth ($depth = 18$). Furthermore, we found that disabling replacement in the bootstrap sampling improved diversity among the trees.

In contrast, the optimization for **CatBoost** converged towards parameters that mitigated overfitting. The optimal learning rate was relatively low (0.0194), and the depth

was moderate (10). The regularization parameter ($l2\_leaf\_reg = 0.01$) suggested that the model relied heavily on the natural structure of the categorical splits rather than aggressive weight penalties.

5.2 Classifier Evaluation

We evaluated the models on the held-out test set (2023). This dataset preserves the natural distribution of the problem (47% non-injured, 36% injured, 16% heavily injured). We report class-wise precision, recall, and F1-scores. This granular evaluation is critical because a high global accuracy could hide a model's failure to detect severe injuries: the most costly error in our domain.

5.3 Clustering Evaluation

Clustering performance was assessed via Silhouette scores (Table 3). While HDBSCAN and Agglomerative Clustering yielded higher raw scores (> 0.08), deeper inspection revealed they produced degenerate solutions where over 85% (HDBSCAN) or 100% (Agglomerative) of the data collapsed into a single giant cluster. Such partitions offer no analytical value for differentiating accident types. In contrast, K-Prototypes with $k = 5$ produced a more balanced distribution (largest cluster $\approx 39\%$) with clear semantic distinctions, offering the best trade-off between mathematical cohesion and practical interpretability.

Algorithm	**Param**	**Silhouette**	**Largest Cluster [%]**
K-Prototypes	5	0.041	38.8
Agglomerative	3	0.107	100.0
HDBSCAN	10	0.089	85.4

Table 3: Comparison of clustering algorithms (abbreviated).

6 Results

6.1 Identified Accident Personas

Clustering analysis revealed five accident personas (features with lift > 1.5):

6.2 Classification Results

This section presents the performance of the three trained classification models. A summary of the core evaluation metrics is provided in table 5.

Performance Metrics

CatBoost achieves the highest macro and micro F1 and Cohen's κ, followed by BRF and RC, while the baseline performs substantially worse.

Cluster	Persona Description
0	Midday pedestrian-related accidents with older road users in semi-urban areas.
1	Night-time accidents involving younger adults (18–30 years) under low visibility.
2	Morning rush-hour collisions linked to commuter traffic and congestion.
3	High-complexity urban intersection accidents with notable pedestrian involvement.
4	Low-speed urban maneuver collisions (parking, turning) in narrow streets, usually minor but risky for vulnerable users.

Table 4: Identified accident personas (features with lift > 1.5).

	RC			**BRF**			**CB**			**BL**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Uninjured	0.83	0.74	0.78	0.87	0.73	0.80	0.85	0.77	0.81	0.50	0.64	0.56
Injured	0.62	0.50	0.55	0.63	0.61	0.62	0.64	0.63	0.64	0.34	0.30	0.32
Severe	0.42	0.74	0.54	0.47	0.72	0.56	0.51	0.65	0.57	0.16	0.08	0.10
Macro F1		0.62			0.66			0.67			0.33	
Micro F1		0.66			0.69			0.71			0.40	
Cohen's Kappa		0.6021			0.6343			0.6488			0.0910	
Macro AUC		-			0.72296			0.72929			-	
Micro AUC		-			0.78559			0.80404			-	

Table 5: Test Performance: Ridge Classifier, Balanced Random Forest, CatBoost, Baseline.

All models predict **Uninjured** cases reliably, with CatBoost performing best. The **Injured** class is more ambiguous, with many misclassifications toward Uninjured or Severe, while **Severely Injured** cases are generally identified correctly, though models tend to overpredict this class. Figure **??** shows the normalized confusion matrices, highlighting these consistent patterns across models.

Micro and macro AUC scores confirm that CatBoost produces the most separable precision-recall curves (micro_AUC=0.80404; macro_AUC=0.72929), slightly outperforming the Balanced Random Forest (micro_AUC = 0.78559, macro_AUC = 0.72296). Ridge and the baseline cannot be evaluated due to lack of calibrated probabilities.

Feature importance

Table 6 lists the top three features per model. Ridge emphasizes vehicle counts, BRF highlights collision type and impact intensity, and CatBoost leverages a broader mix of demographic, situational, and crash-mechanics variables, which likely contributes to its superior performance.

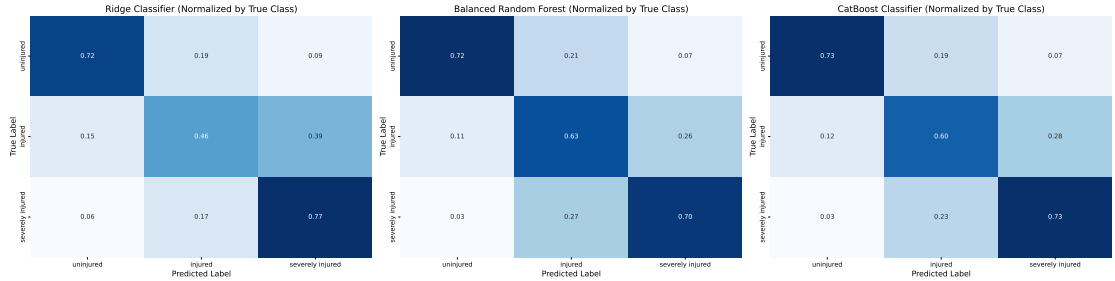

Figure 1: Generalization Performance: Confusion Matrices

Model	Feature 1	Feature 2	Feature 3
RC	mobile_obstacle_struck_1	vehicle_category_other_none	num_light_motor_vehicle
BRF	mobile_obstacle_struck	impact_delta	fixed_obstacle_struck
CB	type_of_collision	mobile_obstacle_struck	age_group

Table 6: Three most important features identified by each classifier. Ridge importance is derived from coefficient magnitudes, BRF importance from mean decrease in impurity, and CatBoost importance from built-in feature attribution.

Summary

The results highlight the benefits of flexible, non-linear models, especially CatBoost, in capturing the complex interactions underlying road traffic injury severity. The CatBoost confusion matrix shows that about 95% of severely injured cases are at least classified as injured, so a real-time system could flag most severe accidents early and support faster emergency response. However, this comes at the cost of inefficient resource allocation, because also around 5% of non-injury cases would be predicted as severely injured and trigger unnecessary deployments, illustrating a fundamental trade-off between saving lives and resource efficiency. Moreover, the remaining 5% of severely injured cases that are predicted as uninjured are a critical limitation: such a system cannot replace emergency calls or human judgment but should instead be viewed as a decision-support tool that helps responders prioritize and detect severe accidents that might otherwise be recognized too late.

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