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 30, 2025

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

1 Application Area and Goals

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

To address this gap, we leverage historical accident data provided by the French National Interministerial Observatory for Road Safety (ONISR). This organization maintains the "Bulletins d'Analyse des Accidents Corporels de la Circulation" (BAAC), a national database recording all injury accidents on public roads in France. The dataset captures detailed information covering the circumstances of each accident (Caractéristiques), the location (Lieux), the vehicles involved (Véhicules), and the individual users (Usagers). Contextual variables include road category, lighting, weather conditions, vehicle type, and user 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, 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 information realistically available to vehicle telematics systems at the moment of impact. Such a model enhances 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 23 original features (retained from raw data) and 34 engineered features (calculated to model complex relationships), comprising a total of 57 features.

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 engineered 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 Feature Selection & Cleaning

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. Hereof, 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". Missing values for the driver, indicating a hit and run, were imputed with `0` or `Unknown` to keep the information.
- **Data Quality Missingness:** Columns with $\approx 5\%$ missing values were dropped, alongside rows missing values in columns without special NaN meaning.
- **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.

Rows missing the target variable `injury_severity` were dropped as well, as they cannot be used for supervised learning. In addition, an isolation forest was used for simple outlier elimination considering the dataset's size.

3.4 Feature Engineering

To capture complex non-linear relationships, we generated 34 new features across four domains. The final dataset consists of 57 refined features ready for the modeling pipeline.

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 Resampling

Regarding the target value of injury severity, the dataset exhibits a clear class imbalance towards less severe cases (47% non-injured, 36% injured, and 16% heavily injured) biasing models trained towards the majority class. To address this, the raw target variable `injury_severity` (4 classes) was re-mapped to an ordinal `injury_target` (0: Uninjured, 1: Light Injury, 2: Hospitalized/Dead), both adjusting for updates received up to 30 days afterwards and better reflecting the triage needs of first responders.

Given the importance of correctly assessing cases of heavy injury, resampling was applied to the dataset to remove class imbalance. Both over- (SMOTE) and undersampling were considered and tested. Given the large amount of data at disposal (>400k) and considering compute constraints, random undersampling (130k) was ultimately chosen.

4 Data Mining

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. This metric ensures that the performance on the minority class (severe injuries) is given equal weight to the majority class (uninjured), preventing the model from being biased towards the more frequent, less severe cases. 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:

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

4.1 Classification: Model Selection

To establish a baseline understanding of the prediction task's complexity, we initially experimented with a diverse range of machine learning algorithms with minimal hyperparameter tuning. This exploratory phase included logistic regression variants (Ordinal, Lasso, Ridge), tree-based ensembles (Random Forests, Histogram Gradient Boosting, CatBoost), and simple neural networks.

Consistent with our hypothesis, all machine learning models substantially outperformed the rule-based baseline. Interestingly, models explicitly designed for ordinal regression did not yield significant improvements in F1-scores or specific error reduction compared to standard classifiers, leading to their exclusion. Applying the principle of parsimony (Occam's Razor), we narrowed our focus to three distinct architectures for further optimization, each selected for its specific theoretical strengths in addressing our data characteristics:

- **Ridge Classifier:** A linear model incorporating L2 regularization. It was selected for its computational efficiency and ability to mitigate multicollinearity, providing a robust linear baseline. To address class imbalance, it was coupled with random undersampling.

- **Balanced Random Forest (BRF):** An ensemble method that directly addresses the class imbalance problem by iteratively undersampling the majority class during the bootstrap aggregation (bagging) process. This ensures that each decision tree in the forest is trained on a balanced subset of data.

- **CatBoost:** A gradient boosting algorithm chosen for its superior native handling of categorical features, which constitute a significant portion of our dataset. Its ordered boosting technique is particularly effective at reducing overfitting on small datasets or when using high-cardinality features.

Given the dataset size and computational constraints, hyperparameter tuning was performed using 3-fold cross-validation with shuffling to mitigate ordering bias. We employed Bayesian Optimization to maximize the F1-Macro score. While Ridge Regression tuning was limited to the regularization strength α, the ensemble methods provided a broader range of hyperparameters, particularly regarding tree pruning to prevent overfitting. The considered hyperparameter spaces are listed in Table 2.

Final model selection relied on both the F1-Macro score and an analysis of confusion matrices, with a specific focus on recall and misclassification rates for severe injury cases. Based on these criteria and the results from the cross-validation, CatBoost was selected as the optimal model.

4.2 Clustering: Strategy

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

Model	**Hyperparameter**	**Search Space**	**Best Value**	**F1-Macro**
Ridge	alpha	$\{0.05, 0.10, \ldots, 10\}$	1.5	0.654
BRF	n_estimators	$[50,\ 400] \cap \mathbb{Z}$	400	0.674
	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	0.696
	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.

- **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").

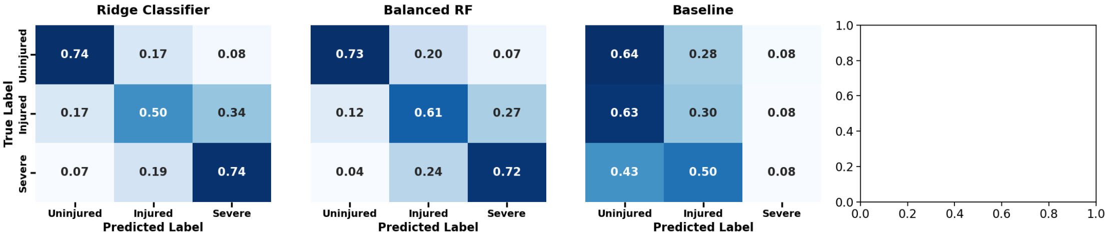

Figure 1: Generalization Performance: Confusion Matrices

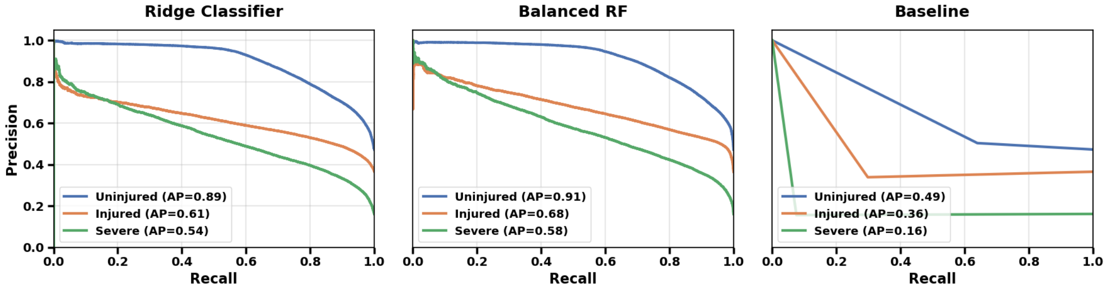

Figure 2: Generalization Performance: PR-Curves

5 Evaluation

5.1 Classification

The models were evaluated on the held-out test set (2023), which exhibits a similar target value distribution to previous years, thus naturally ensuring stratification. For the hyperparameters, the optimal values listed in Table 2 were used. Each model was systematically evaluated using an automated pipeline to report the metrics listed in subsubsection 4. Cohen's kappa was additionally included to evaluate ordinal performance.

CatBoost achieves the highest F1 and Cohen's κ, followed by BRF and RC, while the baseline performs substantially worse, following the results obtained in model selection. While models reliably predict low severity cases (e.g. F1 $[0.78, 0.80]$), differentiation between injured and heavily injured cases remains challenging, contributing the most to performance lost.

Figure 1 provides deeper insight into these classification errors. While the models rarely commit extreme misclassifications (confusing uninjured with heavily injured), the decision boundary between injured and heavily injured is not well defined, resulting in mix-ups between the two.

The Multi-class Precision-Recall curves of Figure 2 corroborate these findings. The curves for the "Uninjured" class (blue) consistently demonstrate high precision across a wide range of recall values, reflecting the model's robustness in correctly identifying non-injury cases. In contrast, the "Severe" class (green) exhibits a sharper decline, visually confirming the trade-off inherent in balancing high recall for critical cases against the precision of those alerts.

Rank	RC	BRF	CB
1	mobile_obstacle_struck_1	mobile_obstacle_struck	type_of_collision
2	vehicle_category_other_none	impact_delta	mobile_obstacle_struck
3	sex_2 (Female)	fixed_obstacle_struck	age_group
4	mobile_obstacle_struck_4	type_of_collision	initial_point_of_impact
5	vehicle_2_3_wheeler	speed_limit	speed_limit

Table 3: Feature Importances per Classifier

Later iterations of the data mining process attempted to alleviate this issue through various means, including OneVsOne, hierarchical (first 0 vs {1,2}, then 1 vs 2) and stacking classification, as well as regression with custom thresholds, but were unable to achieve meaningful improvements. Considering this result and based on the misclassification costs involved, model training and selection could be altered to prefer one over the other, depending on the exact requirements.

	RC			**BRF**			**CB**			**BL**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Uninjured	0.84	0.72	0.78	0.88	0.72	0.79	0.88	0.73	0.80	0.50	0.64	0.56
Injured	0.59	0.46	0.52	0.62	0.63	0.63	0.63	0.60	0.62	0.34	0.30	0.32
Severe	0.40	0.77	0.53	0.47	0.70	0.56	0.46	0.73	0.57	0.16	0.08	0.10
Macro F1		0.61			0.66			0.66			0.33	
Cohen's κ		0.59			0.63			0.6342			0.0910	

Table 4: Test Performance: Ridge Classifier, Balanced Random Forest, CatBoost, Baseline.

Feature Importance Analysis

Table 3 presents the top three features for each model. Ridge importance is determined by coefficient magnitudes, BRF importance by mean decrease in impurity, and CatBoost importance by built-in feature attribution. The feature importance analysis demonstrates that the nature of the obstacle struck is critical for injury prediction, as evidenced by the consistent appearance of `mobile_obstacle_struck` (`mos`) across all classifiers and the inclusion of `impact_delta` in BRF. Collisions involving a pedestrian or a train (`mos_1` and `4`) serve as strong indicators for injury outcomes. Furthermore, variables such as `type_of_collision` and `initial_point_of_impact` highlight that the precise point of impact, in conjunction with vehicle speed, is highly relevant for prediction.

5.2 Clustering Evaluation

Clustering performance was evaluated via Silhouette scores (Table 6). Although HDBSCAN and Agglomerative Clustering yielded higher raw scores (> 0.08), further inspec-

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

tion revealed degenerate solutions where over 85% (HDBSCAN) or 100% (Agglomerative) of the data collapsed into a single predominant cluster. Such partitions offer no analytical value for differentiating between accident types. In contrast, K-Prototypes with $k = 5$ produced a more balanced distribution (largest cluster $\approx$ 39%) with clear semantic distinctions, offering the optimal 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 6: Comparison of clustering algorithms (abbreviated).

6 Results

Micro and macro AUC scores identify CatBoost as the optimal candidate for the targeted real-time telematics application, yielding the most separable precision-recall curves (micro_AUC = 0.80; macro_AUC = 0.73). It slightly outperforms the Balanced Random Forest, validating the necessity of flexible, non-linear models to capture complex injury risks within the ONISR data.

In the context of eCall goals, the results illustrate a critical balance between safety and efficiency. CatBoost classifies $\approx$ 95% of severely injured cases as at least injured, directly meeting the primary objective of accelerating medical response for high-risk accidents. However, this sensitivity incurs a cost in resource allocation, as $\approx$ 5% of non-injury cases are predicted as severe, potentially triggering unnecessary deployments. Furthermore, the misclassification of the remaining 5% of severe cases confirms that the system cannot replace human judgment. Instead, it serves as a decision-support tool, enhancing emergency coordination by flagging severe accidents that standard protocols might recognize 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. November 2025