DataScientest • com

# Road Accidents in France based on Annual Road Traffic Accident Injury Database (2005 - 2023)

# Contents

# Introduction to the project

## Overview

This project focuses on predicting the severity of road accidents in France by leveraging historical data collected from 2005 to 2023. The dataset, sourced from the BAAC (Bulletin d'Analyse des Accidents Corporels de la Circulation), contains comprehensive records of injury-related accidents, including detailed attributes about environmental, temporal, behavioural, and situational factors.

Accidents vary widely in their consequences making severity prediction a critical area of study. Understanding what conditions lead to more severe accidents can inform targeted measures for prevention and mitigation.

The societal and economic impact of severe road accidents is profound, including increased healthcare costs, long-term disabilities, infrastructure damage, insurance claims, and loss of productivity. Predictive modelling of severity enables better planning and resource allocation for governments, urban planners, and emergency services.

From a technical perspective, this project involves cleaning and analysis of large-scale datasets related to road accidents. These include geographic, meteorological, temporal, and behavioral variables. The project leverages machine learning models to identify patterns and predict accident likelihood.

## Objectives

The main objectives include:

- Studying and cleaning the dataset.
- Extracting relevant characteristics for severity prediction.
- Developing a model to evaluate accident severity.
- Once trained, the model will be validated against historical data.

# Understanding and manipulation of data

## Framework

Each accident involving injuries, occurring on public roads, involving at least one vehicle, and resulting in at least one medically, treated victim is recorded by a law enforcement unit (e.g., police, gendarmerie). This information is compiled into the Injury Accident Analysis Bulletin. These bulletins form the national BAAC File, administered by the National Interministerial Road Safety Observatory (ONISR). The data is available under https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2023/

The datasets cover all injury-related road accidents across:

- Mainland France
- Overseas Departments: Guadeloupe, French Guiana, Martinique, Réunion, and Mayotte (since 2012)
- Overseas Territories: Saint-Pierre and Miquelon, Saint-Barthélemy, Saint-Martin, Wallis and Futuna, French Polynesia, New Caledonia (since 2019)

The annual data files include:

- Caractéristiques – Accident characteristics
- Lieux – Location data
- Véhicules – Vehicle details
- Usagers – Information on people involved

An additional dataset, vehicules-immatricules-baac (registered vehicles, 2009–2022), is available but not used currently due to incomplete yearly coverage, which could hinder machine learning performance.

## Relevance

The dataset contains dozens of categorical and numeric features allowing for rich feature engineering. Latitude (lat) and longitude (long) data enable spatial analysis and mapping of severity hotspots.

Several variables have direct or indirect impact on predicting injury severity (grav), which is the target variable. This variable categorizes accident outcomes as: 1 = Unharmed, 2 = Killed, 3 = Hospitalized, 4 = Light injury.

The following variables are especially relevant for modelling severity:

1. Environmental conditions: atm (atmospheric conditions), lum (lighting), surf (surface state), prof (road profile), plan (road curvature)
2. Location and infrastructure: agg (urban/rural), catr (road type), infra (infrastructure), nbv (number of lanes)
3. Temporal factors: an, mois, jour, hrmn (date and time of accident)
4. Collision and maneuver dynamics: col (type of collision), choc (point of impact), manv (driver maneuver), obs/obsm (obstacles)
5. Vehicle and user data: catv (vehicle type), catu (user role), secu / secu1-3 (safety equipment), an_nais (user birth year), sexe (user gender)

Data Limitations:

- Incomplete or missing data: Many fields include missing entries, encoded as empty cells, zeros, periods, or -1 for "not specified."
- Hospitalization data: Since 2018, the classification of hospitalized injured persons has changed due to updates in the law enforcement data entry process. As a result, this indicator is not comparable with earlier years and has not been certified by the Public Statistics Authority since 2019.
- Runaway users: From 2021 onward, users fleeing the accident scene are included, but their demographic and injury data are often missing.
- Class imbalance: Fatal cases are underrepresented compared to minor injuries, necessitating techniques such as oversampling or class reweighting.

## Pre-processing and feature engineering

Given the scale and complexity of the dataset of nearly two decades and compiled from multiple relational tables including accident, location, vehicle, and user-level information, significant data cleaning and preprocessing were essential to ensure analytical quality and model robustness.

### 1) Data Integration

The original data was fragmented across several tables provided annually by the French BAAC system. These files were merged using the unique accident identifier (num_acc) to produce a unified dataset capturing all relevant dimensions.

### 2) Outlier Treatment and Data Validation

Outlier detection was performed on numerical fields using descriptive statistics and interquartile range (IQR) methods. The values in the nbv (number of traffic lanes) column that were below 1 or above 10 were deemed implausible and replaced with missing values. Similarly, road width (larrout) values exceeding 200 meters or reported as zero were considered invalid. Age values were derived from year of birth and accident year and capped at 120 years.

Categorical variables were cleaned by referencing official codebooks. Values not conforming to the expected coding scheme, including placeholders such as -1, were marked as missing.

### 3) Datetime Feature

A unified datetime column was created by combining the original year (an), month (mois), day (jour), and time (hrmn) fields. The hrmn values were cleaned to handle inconsistent formats (e.g., 12:30, 45) and defaulted to midnight when missing.

From this timestamp, key features were extracted:

- hour – Hour of the accident
- dayofweek – Numeric day of the week (0 = Monday)

### 4) Geographic Coordinates Cleaning

To ensure reliable spatial analysis, a combined strategy was applied to clean and complete GPS data.

Raw latitude and longitude values were inconsistently formatted (e.g., comma decimals, long integers). These were standardized to float values and invalid entries (e.g., zeros) were set to missing.

For records lacking valid GPS data, geographic centroids were used based on INSEE commune codes (dep + com). A reference file containing official commune centroids (https://www.data.gouv.fr/en/datasets/communes-de-france-base-des-codes-postaux/) allowed us to impute missing locations, raising GPS coverage.

Then GPS–INSEE consistency check was performed. Coordinates were validated against the corresponding INSEE code:

- Mainland codes had to fall within France's geographic bounds.
- Overseas codes were expected outside these bounds.

About 11% of records showed mismatches and had their GPS values were marked as missing values.

### 5) Safety Equipment Processing

Safety equipment use was recorded differently before and after 2019 and required standardization.

- Before 2019: The secu column used a two-digit code where the first digit indicates the type of equipment (e.g., 1 = seat belt, 2 = helmet) and the second digit shows usage: 1 = used, 2 = not used, 3 = unknown
- From 2019: Up to three separate fields (secu1, secu2, secu3) captured equipment types using single-digit codes. Values -1 (not specified) and 8 (unknown) were treated as missing.

To unify this information, four summary columns were created: belt_status, helmet_status, child_device_status, reflective_vest_status.
Each is coded as: 1 = used, 0 = not used, -1 = not specified and NaN.

### 6) Data Cleaning Summary

To optimize the dataset for modeling, the following columns were dropped. The rationale falls into four main categories: high missingness, redundancy, low relevance, and formatting issues.

- High Missingness: vma, motor, lartpc, env1, occutc, pr, pr1, helmet_status, child_device_status, reflective_vest_status, manv, trajet
- Low Relevance: locp, actp, etatp – pedestrian-specific, mostly missing. id_usager, id_vehicule, num_veh, num_acc. – not used in prediction.
- Redundant or Derived: an_nais, com, dep, hrmn_clean, insee_code, lat_clean, long_clean
- Formatting Issues: adr, voie, v1, v2 – unstructured or inconsistent

### 7) Normalization and Standardization

To ensure consistency across features and improve model performance, the following preprocessing steps were applied within the pipeline:
*Numeric Features:*

- Missing values imputed using the median
- Standardized with StandardScaler, transforming each variable to have zero mean and unit variance (z-score normalization)

*Categorical Features:*

- Missing values imputed using the most frequent category
- Encoded with OneHotEncoder to convert categorical values into binary indicators

### 8) Dimensionality Reduction Strategy

To control feature space complexity and reduce redundancy, several dimensionality reduction techniques were implemented:

*Variance Threshold*

Applied a threshold of 0.01 to one-hot encoded categorical features. Near-constant variables were removed to eliminate noise without losing meaningful information.

*Categorical Feature Redundancy via Cramér's V*

Pairwise Cramér's V values were computed to detect associations between categorical features. place vs. catu: Cramér's V = 1.0 (perfect association). Since seat position is fully determined by user category (driver, passenger, pedestrian), the variable place was dropped. agg vs. catr: Cramér's V = 0.70 (strong association). To avoid redundancy, these features were merged into a composite variable agg_catr, capturing both urban context and road type.

*High-Cardinality Reduction: catv (Vehicle Category)*

The original variable contained over 40 categories, many rarely observed. Categories were regrouped into 9 semantically meaningful clusters (e.g., Micromobility, Light Vehicles, Trucks, Public Transport).

**9) Feature Engineering**

To enhance predictive performance, additional derived features were introduced:

- rush_hour: Binary flag for peak traffic hours (7–9 AM, 4–7 PM).
- season: Season of the year derived from the accident month (Winter, Spring, Summer, Autumn).
- age_bins: Discretized age groups (<18, 18–25, 26–40, 41–60, 60+).
- user_belt_status: Composite variable combining user category (catu) with seatbelt usage (belt_status).

# Visualizations and Statistics

### 1) Temporal Patterns in Road Accident Occurrence

To understand the temporal dynamics of road accidents, we analysed the frequency of accidents across different hours of the day and days of the week. Figure 1 highlights pronounced daily and weekly trends. Accident frequency shows a clear temporal structure, with significantly more accidents occurring between 7 AM and 7 PM, peaking around 5–6 PM, which corresponds to evening rush hour. A secondary peak is observed during 6–9 AM, reflecting morning commutes. In contrast, the lowest accident rates occur overnight (1–5 AM), when road traffic is minimal.

Fridays consistently record the highest accident counts, followed by other weekdays, while Sundays show the lowest overall incident levels.

To assess the statistical significance of these distributions, we applied Chi-squared goodness-of-fit tests. The distribution of accidents across hours of the day was significantly non-uniform ($\chi^2$ = 2,824,590.70, $p$ < 0.0001), as was the distribution across days of the week ($\chi^2$ = 82,564.14, $p$ < 0.0001). These findings confirm that both hour and day are influential factors in the temporal occurrence of traffic accidents.

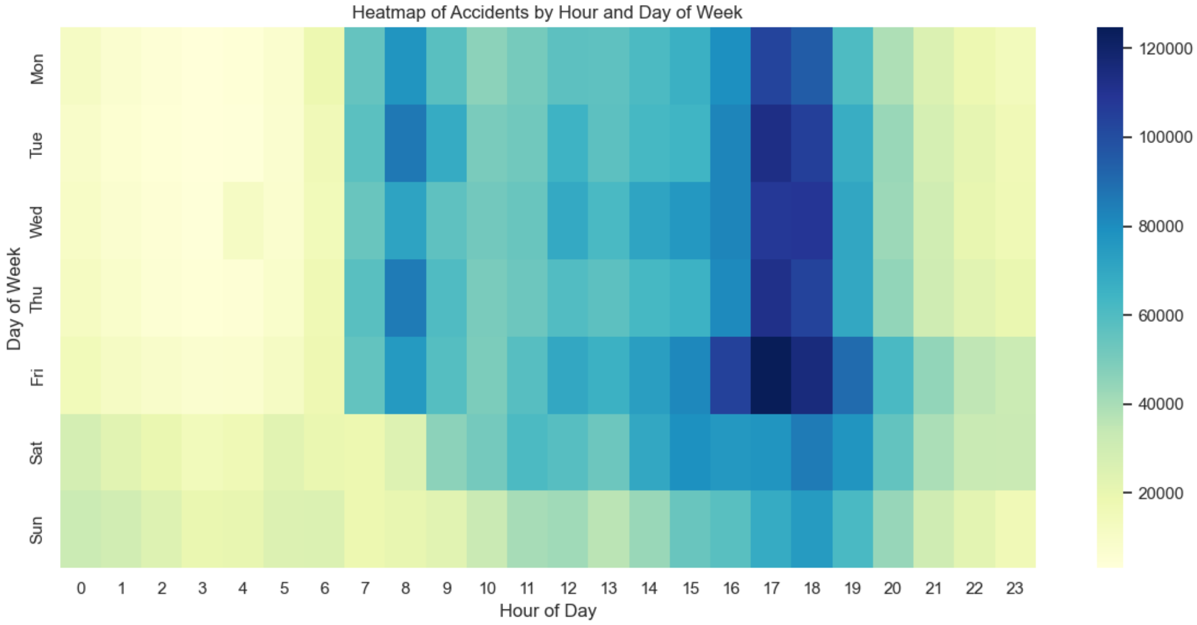


**Figure 1. Heatmap of road accidents by hour of day and day of the week.** Accident counts are highest during weekday rush hours, especially in the late afternoon.

Figure 2 illustrates the monthly evolution of road accident counts over the 18-year period. A clear long-term decline is observed, reflecting potential improvements in road infrastructure, vehicle safety, and enforcement measures. Seasonal fluctuations are also visible, with accident counts tending to peak during certain months, likely corresponding to holidays or changes in driving behaviour.

A notable disruption occurs in 2020, with a steep drop in accident frequency aligned with the onset of the COVID-19 pandemic and national lockdowns.

Given these recurring temporal patterns and disruptions, this trend provides a strong foundation for time series and machine learning modelling.

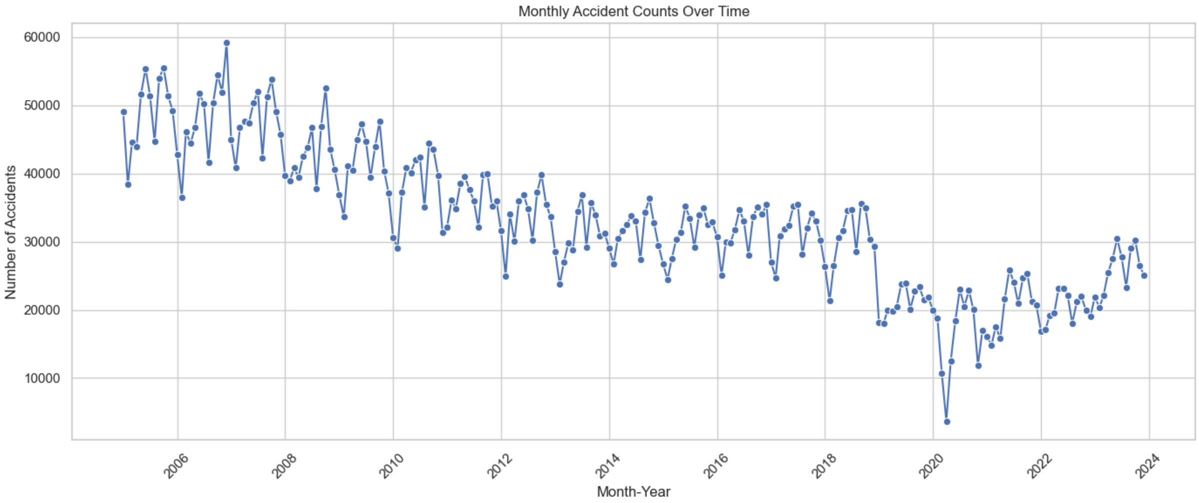


**Figure 2. Monthly Road Accident Trends**

### 2) Severity Distribution

To assess the outcomes of road accidents, we analyzed the distribution of injury severity levels reported for individuals involved. The results reveal strong class imbalance, with most cases falling into non-severe categories (**Figure 3**): nearly 45% of individuals were unharmed, and 37% sustained only light injuries. More serious cases like hospitalizations and fatalities are markedly less frequent, comprising around 16% and 2% of the data, respectively.

This significant class imbalance highlights a modelling challenge. Standard classification models may be biased toward predicting the majority classes (unharmed/light injuries), underestimating the rarer yet critical outcomes like fatalities. Addressing this will require the application of balancing techniques such as class weighting, synthetic oversampling (e.g., SMOTE), or others.

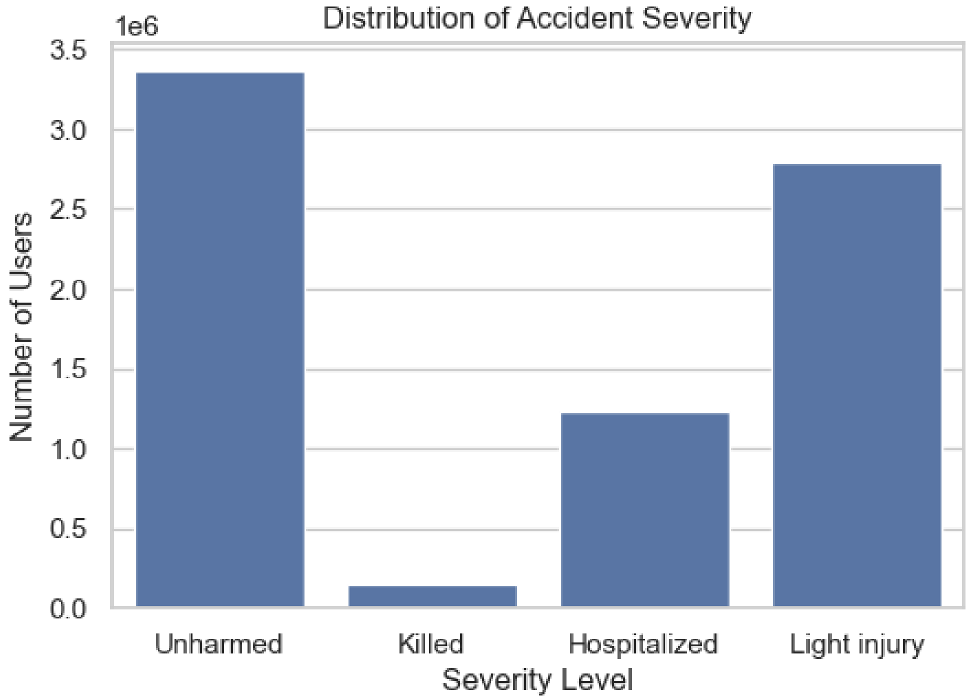


**Figure 3. Distribution of Road Accident Severity**

To understand how accident outcomes differ by participant type, we examined severity distribution across user roles: drivers, passengers, and pedestrians (**Figure 4**). Drivers show the highest share of unharmed outcomes, while pedestrians are more likely to be hospitalized or killed, confirming their greater exposure and risk in traffic environments.

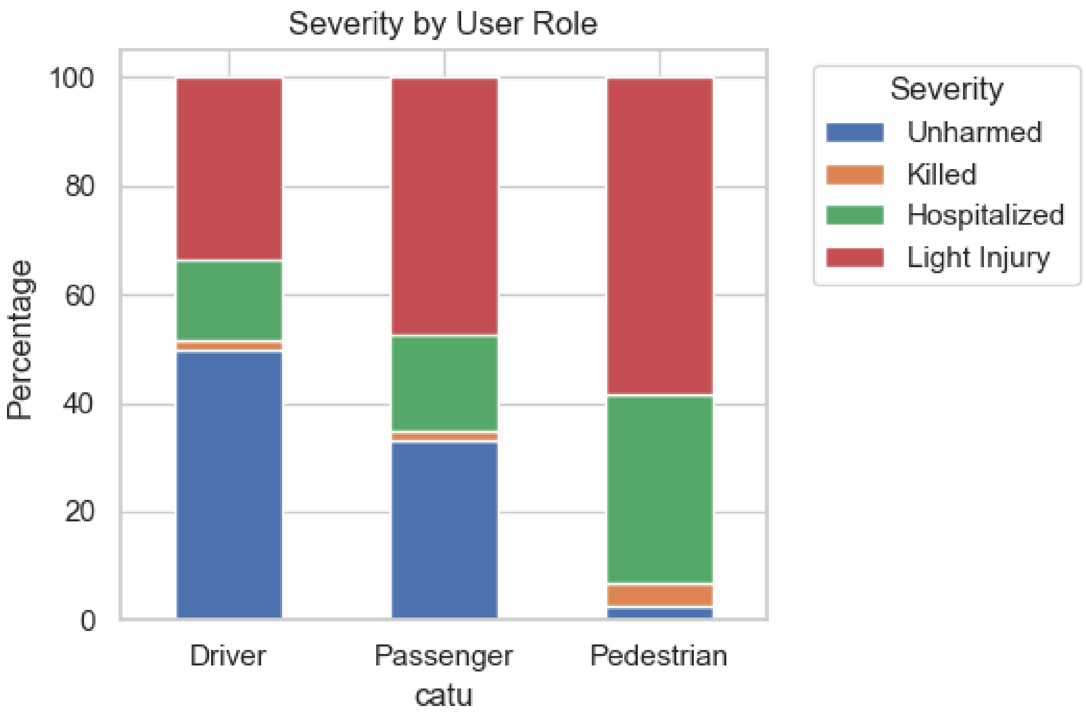


**Figure 4. Distribution of injury severity by user role.**

To assess the effectiveness of seat belts, we analysed accident severity among users of light vehicles (category 07), comparing outcomes between individuals who were wearing seat belts and those who were not (**Figure 5**). The chart clearly shows that users wearing seat belts had a much higher likelihood of remaining unharmed and significantly lower probabilities of hospitalization or death. This underlines the critical role of seat belts in reducing injury severity. Although the seat belt status variable has a relatively high proportion of missing data (approximately 30%), it remains a potentially valuable predictor for modeling accident severity. Given the strong correlation observed between seat belt usage and reduced injury severity (as shown in Figure 5), retaining this feature could enhance the predictive performance of machine learning models.

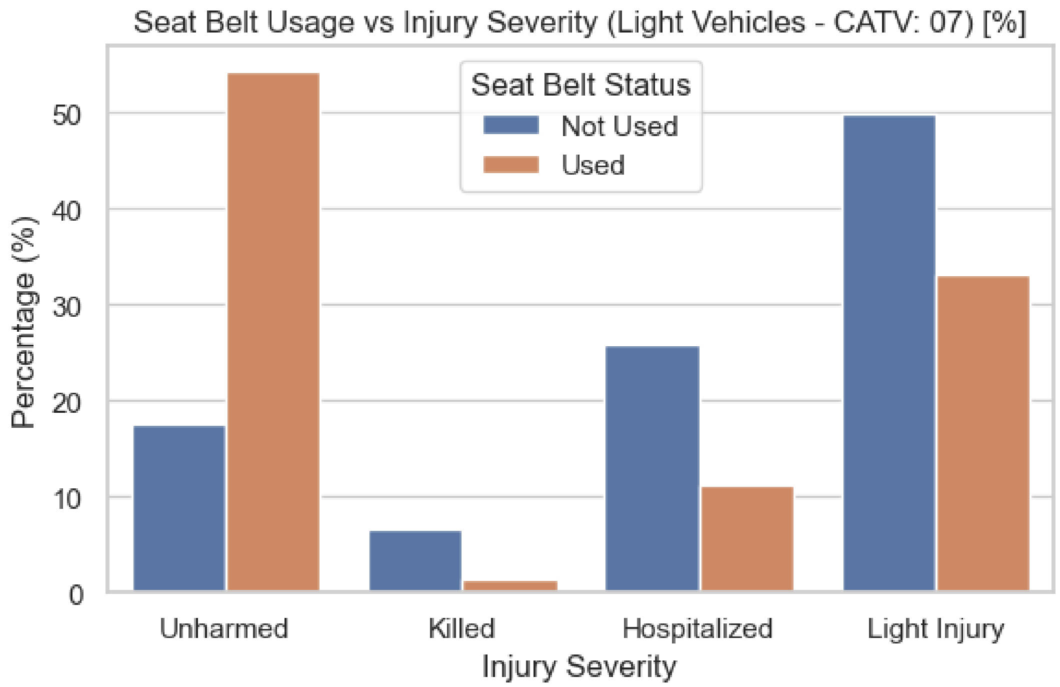


**Figure 5. Relationship between seat belt usage and injury severity for light vehicle users.**

Urban and rural environments present differing risk profiles in traffic accidents due to variations in road type, speed, density, and pedestrian presence. The chart below (**Figure 6**) examines how accident severity varies by location type. The distribution indicates that rural areas report a higher absolute number of "Unharmed" and "Light injury" outcomes, likely reflecting higher vehicle speeds and longer distances travelled. Interestingly, the number of fatalities is also slightly higher in rural areas, despite lower traffic density, which supports the idea that higher speed and delayed emergency response contribute to greater severity. In contrast, urban areas show a relatively higher share of hospitalizations and fatalities compared to their accident volume, likely influenced by vulnerable road users such as pedestrians (**Figure 5**) and cyclists. A chi-square test confirms this association ($\chi^2$ = 200,759.36, $p$ < 0.0001), highlighting location as an important variable for modeling accident severity.

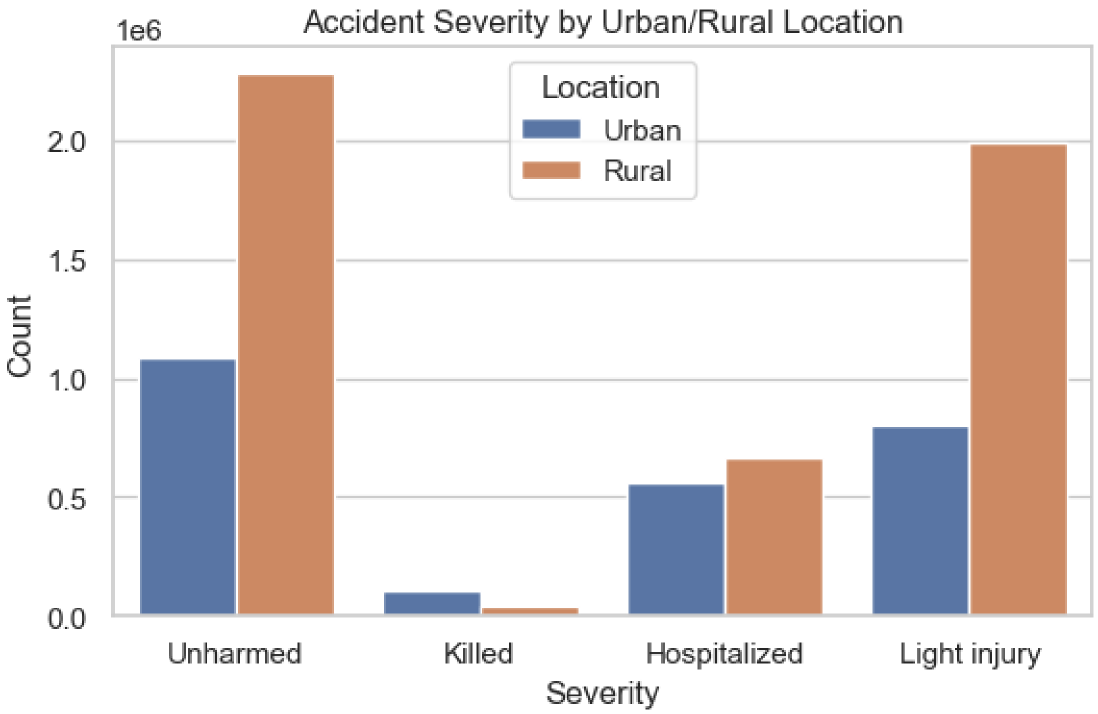


**Figure 6. Accident Severity by Urban/Rural Location**.

To better understand spatial trends in accident severity, **Figure 7** displays the distribution of road accidents across France using GPS coordinates. The colour coding highlights the severity of each case, allowing us to visually assess geographic risk patterns. This figure reveals dense accident clusters around major urban centres, as well as along national highway corridors. These hotspots coincide with areas of high traffic flow and population density, underscoring the role of infrastructure in accident concentration. While rural areas show broader dispersion, urban areas exhibit more intense clusters, which may correspond to complex intersections, pedestrian zones, or congestion.

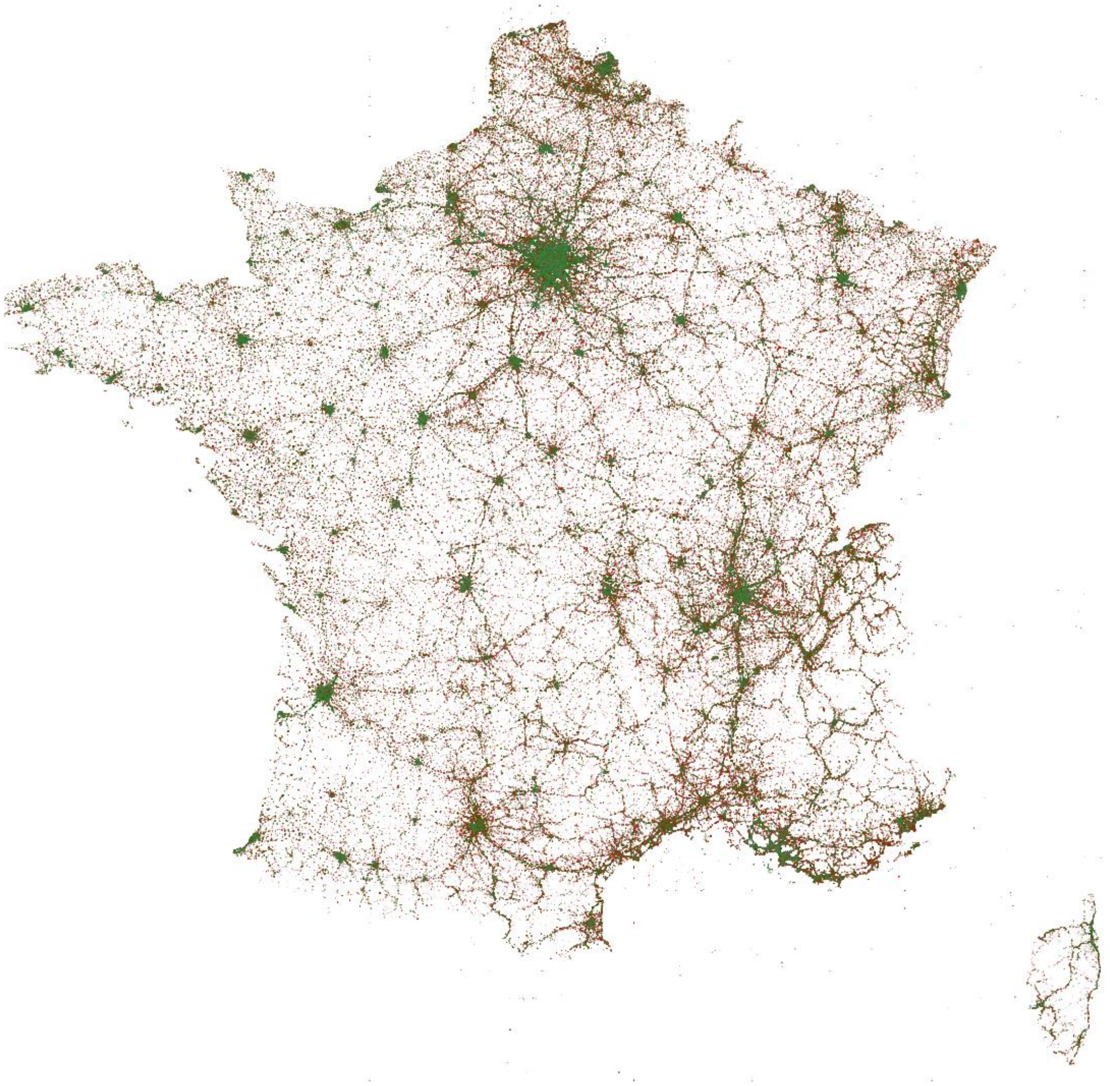


**Figure 7. Geospatial Distribution of Road Accidents in France**.
Legend: Green – Unharmed, Orange – Light injury, Red – Hospitalized, Purple – Killed

### 3) Demographic Patterns

Demographic factors, particularly age and gender, reveal notable disparities in accident involvement and outcomes. **Figure 8** shows that men are involved in significantly more accidents than women across all severity categories. The difference is particularly large in the "Unharmed" and "Light injury" categories, suggesting broader exposure or riskier behaviour among male drivers. These findings highlight the need to consider both behavioural and demographic features in predictive modelling. The statistically significant difference by sex (Chi$^2$ = 91,552.68, $p < 0.0001$) suggests that gender should be retained as an explanatory feature.

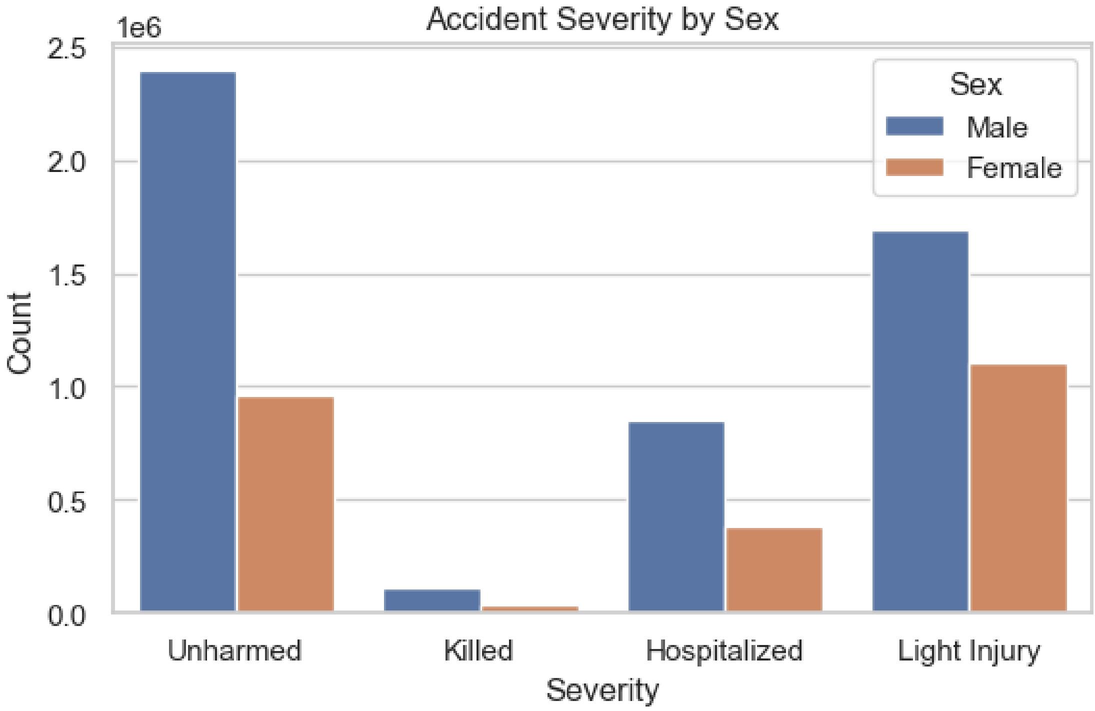


**Figure 8. Accident Severity by Sex.** The higher counts for males may reflect increased exposure or risk-taking behaviours, such as higher speeds or more frequent driving.

Similarly, age and derived age groups offer valuable granularity for model training. **Figure 9** further dissects this pattern by age. The majority of accidents involve individuals aged 15–34, with a sharp peak in the 20–24 range. This age group shows the highest accident frequency for both men and women, though the male count is consistently higher across all age brackets. Older age groups (especially 65+) are underrepresented in accident counts but may still face higher vulnerability when involved.

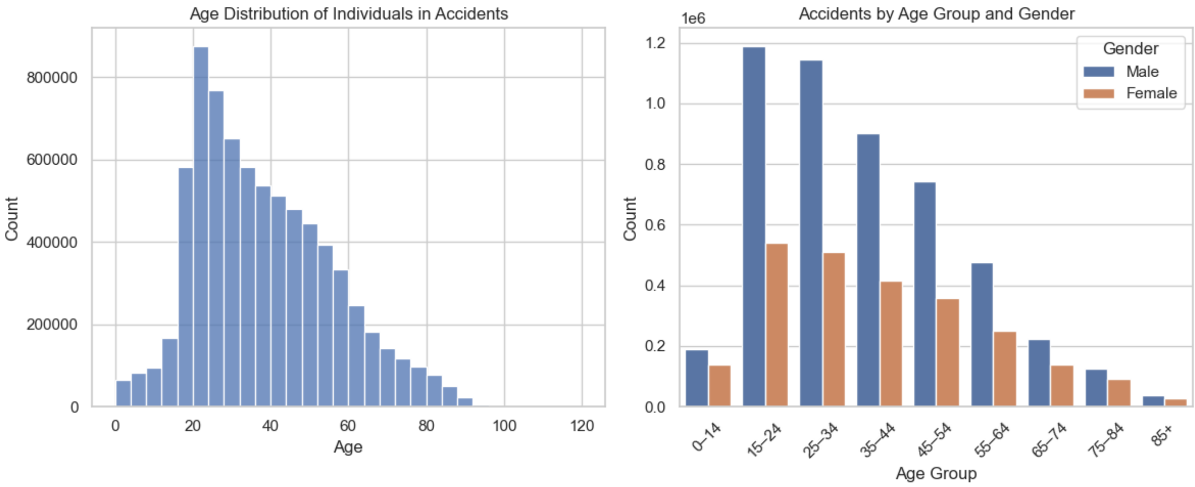


**Figure 9. Age Distribution and Accident Involvement by Gender.** The left panel shows the overall age distribution of individuals involved in accidents, with a clear

peak in the 20–24 age group. The right panel breaks down accident counts by age group and gender. Young males (15–34 years) are the most represented demographic, with male counts consistently exceeding those of females across all age groups.

### 4) Environmental Impact on Severity

Environmental conditions, such as lighting and weather, play a critical role in the dynamics of road accidents and their outcomes. Poor visibility, slippery roads, and reduced driver perception in adverse conditions can significantly elevate the risk and severity of collisions. To better understand these effects, we analyzed the relationship between accident severity and two key environmental variables: lighting conditions at the time of the crash (**Figure 10**), and prevailing atmospheric conditions (**Figure 11**).

**Figure 10** displays the percentage distribution of accident outcomes across different lighting conditions. Accidents during daylight hours show the highest proportion of "Unharmed" outcomes. In contrast, nighttime without public lighting is associated with a noticeable increase in severe outcomes, particularly fatalities and hospitalizations. The presence of functional public lighting at night appears to mitigate risk, with severity distributions closer to daylight conditions.

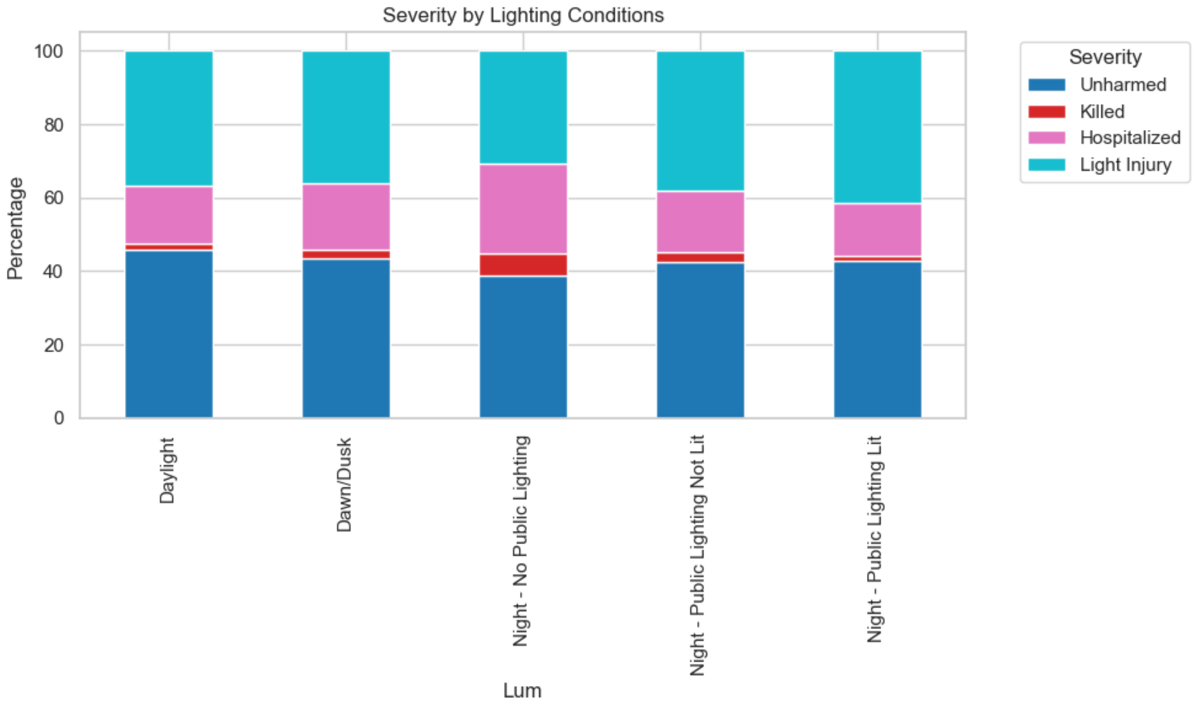


**Figure 10. Severity by Lighting Conditions**

**Figure 11** illustrates how weather impacts accident severity. Under normal and cloudy conditions, most individuals are "Unharmed" or experience "Light Injury." However, in adverse weather—such as fog, snow, or storms—the proportion of "Hospitalized" and "Killed" increases. Snow and fog, in particular, show elevated fatality rates, highlighting the need for heightened caution and potential predictive weighting for these environmental factors.

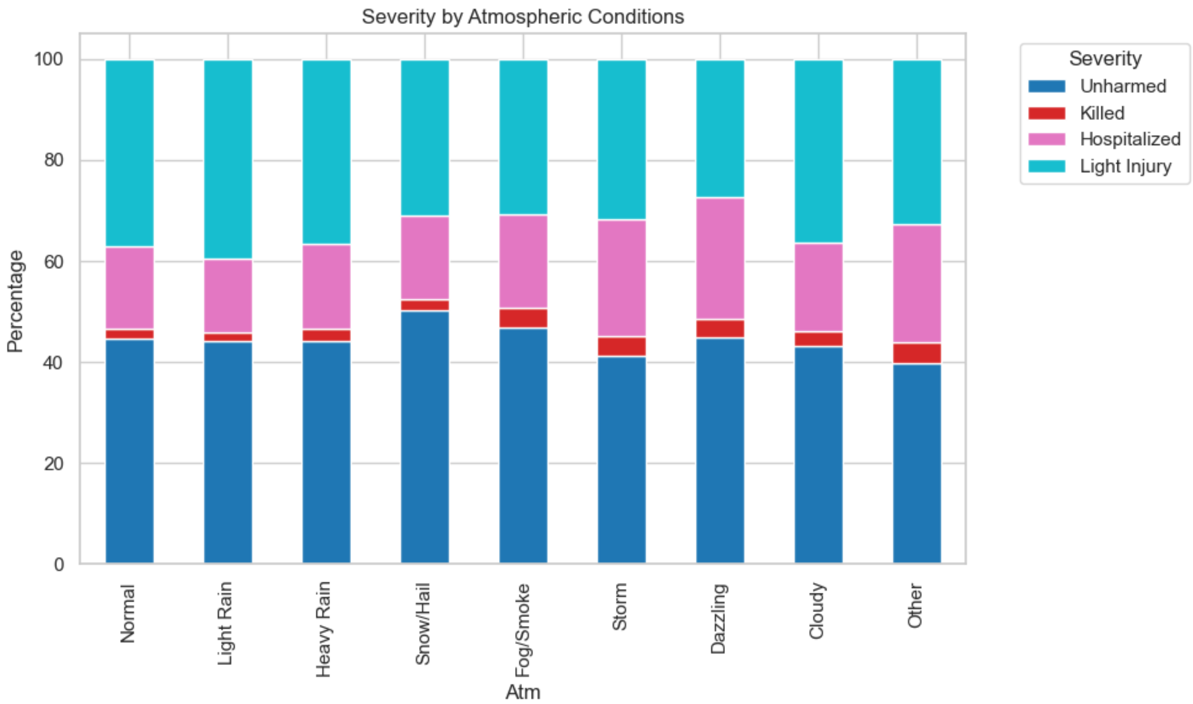


**Figure 11. Severity by Atmospheric Conditions**

# Stages of the project

## Classification of the problem

### Type of Machine Learning Problem & Task Context

This project is formulated as a supervised classification problem. Each accident record is labeled with an outcome variable (grav), representing the severity of injury (unharmed, light injury, hospitalized, killed). The task relates to injury severity prediction in road traffic accidents. Similar to applications such as fraud detection or medical risk prediction, the objective is to identify high-risk cases (e.g., hospitalized or killed) given accident, vehicle, and user characteristics.

### Performance Metrics

In the context of road safety, correctly identifying severe or fatal accidents is far more critical than minimizing false positives. Thus, the main performance metric used to compare models is Recall for the "Killed" class.

The dataset is highly imbalanced, with the majority of cases falling under Minor/None injuries (>80%), while Killed cases represent fewer than 2%. This imbalance means that a naive model predicting “Minor/None” for all cases could achieve high accuracy, but would fail completely at identifying fatalities. For this reason, Recall on the "Killed" class is prioritized over accuracy, ensuring the model does not overlook the most critical cases.

To further mitigate imbalance and improve recall on severe cases, we experimented with target variable reformulations:

- 4-class: Unharmed, Light injury, Hospitalized, Killed (original labels).
- 3-class: Unharmed/Light injury combined, keeping Hospitalized and Killed separate.
- 2-class: Minor/None (unharmed + light injury) vs. Severe (hospitalized + killed).

These aggregations reduce fragmentation of minority classes and shift the predictive task toward distinguishing severe outcomes, thereby improving the trade-off between accuracy and recall for high-risk cases.

To provide a balanced evaluation, we also report:

- Precision, Recall, and F1-score for all severity classes.

- Accuracy as an overall measure, though less informative due to class imbalance.
- Macro and Weighted Averages to account for imbalance and highlight performance across minority classes.
- Confusion Matrices to visualize misclassification patterns.

## Model choice and optimization

To address class imbalance and examine how label granularity influences performance, we trained models under three target definitions (2-class, 3-class, and 4-class). The algorithms compared were:

- Logistic Regression (simple, interpretable baseline)
- Decision Tree (captures non-linear rules, easy to interpret)
- Random Forest (robust ensembles; imbalance handled via undersampling, cost-sensitive learning, or balanced bootstraps.)
- Bagging: Balanced Bagging with Random Forest (reduces variance and improves recall on minority class through balanced resampling)
- Boosting: XGBoost (strong gradient boosting; tested with class weights and undersampling), LightGBM (efficient on large data), CatBoost (effective with categorical data)

These methods are well-suited to tabular, imbalanced data, and results showed that tree-based ensembles consistently outperformed linear and single-tree baselines. For parameter tuning, we applied Grid Search (without cross-validation) to reduce computational overhead. In addition, all benchmarking tests were run on a 20% stratified sample of the dataset (before train/test split) to further accelerate experimentation. These methodological choices allowed us to evaluate a wide range of models efficiently on a very large dataset. A complete catalog of configurations and metrics is provided in **Appendix Table 1**.

To visualize and compare the performance of the tested models, we produced a set of plots summarizing key evaluation metrics across the different class setups. **Figure 12** highlights the trade-off between overall accuracy and balanced performance (Macro-F1). **Figure 13** focuses on the models' ability to correctly identify severe accidents through recall on the severe class. **Figure 14** combines both perspectives, comparing Macro-F1 with Severe-class F1 to show which models achieve the best balance between overall fairness and critical case detection.

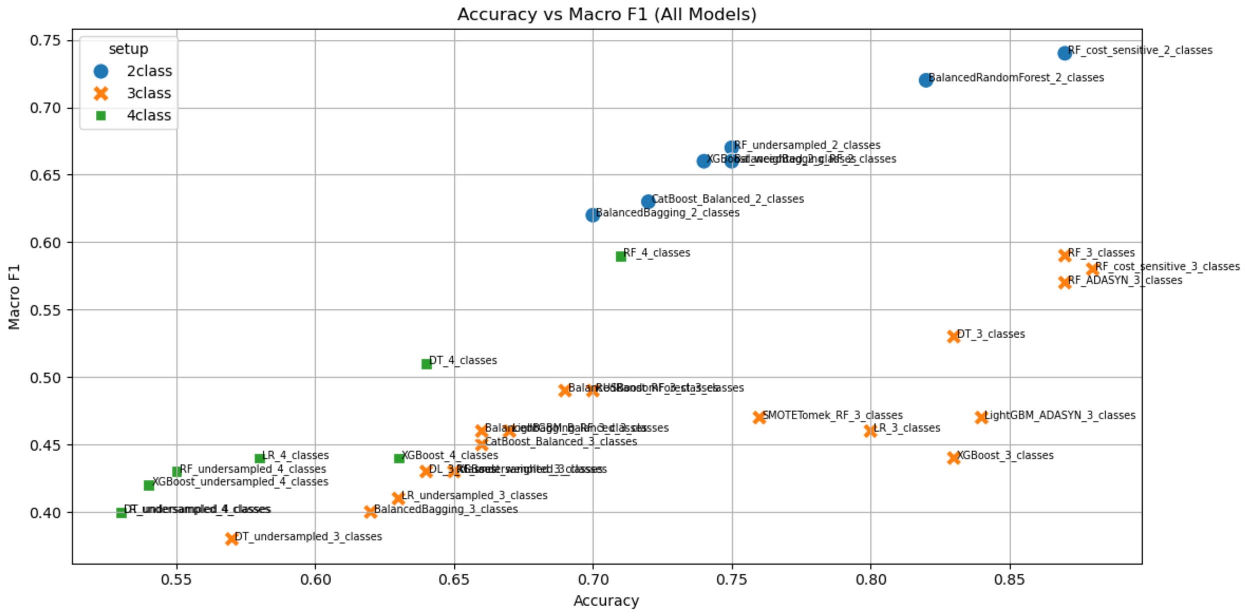


**Figure 12. Accuracy vs. Macro-F1 across models and class setups.** 2-class models cluster at the top-right (Macro-F1 ≈ 0.60–0.75), showing the best balanced performance. 3-class models often reach higher accuracy (≈ 0.80–0.87) but lower Macro-F1 (≈ 0.40–0.58), indicating they favor the majority Unharmed/Light injury group. 4-class models are lowest on both axes; splitting minority labels further makes the task harder.

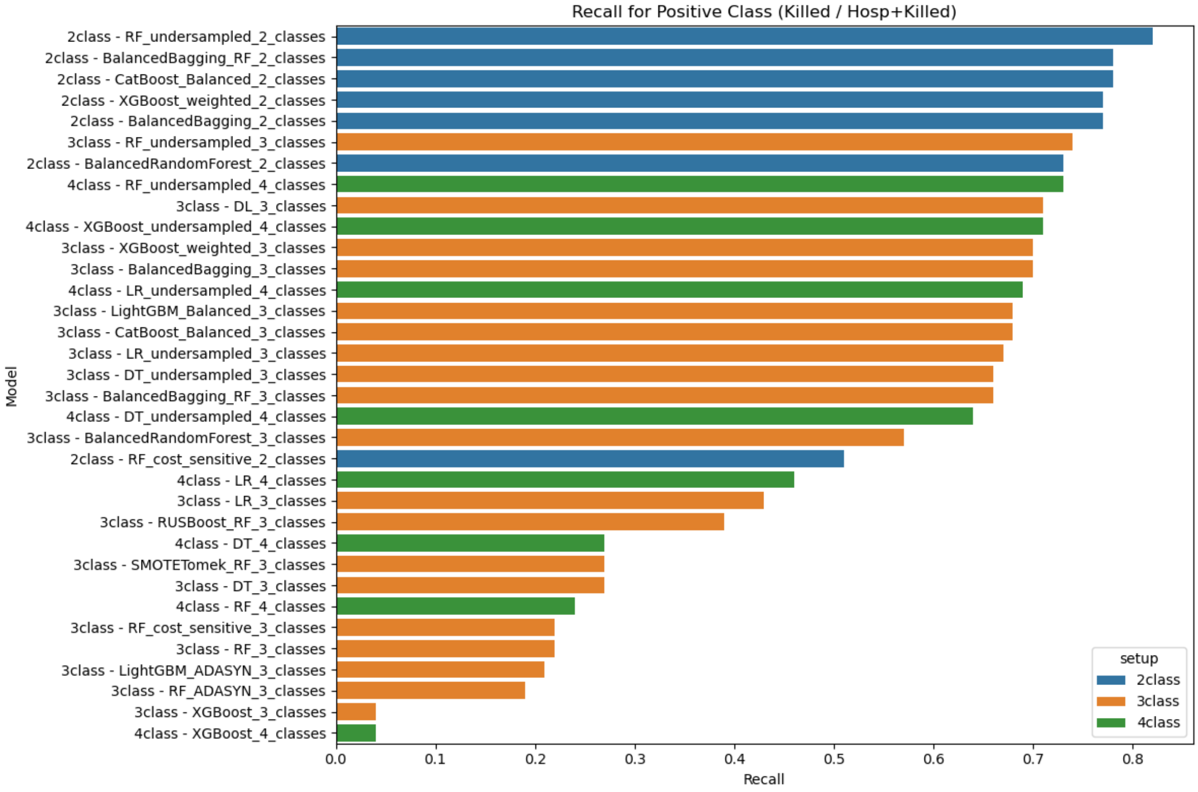


**Figure 13. Recall for the positive (Killed / Hospitalized + Killed) class across models.** The 2-class setup enables the highest severe-class recall (up to ~0.80),

while 3-class models achieve moderate recall and 4-class models trail behind, reflecting the difficulty of detecting rare categories when the label space is more fragmented.

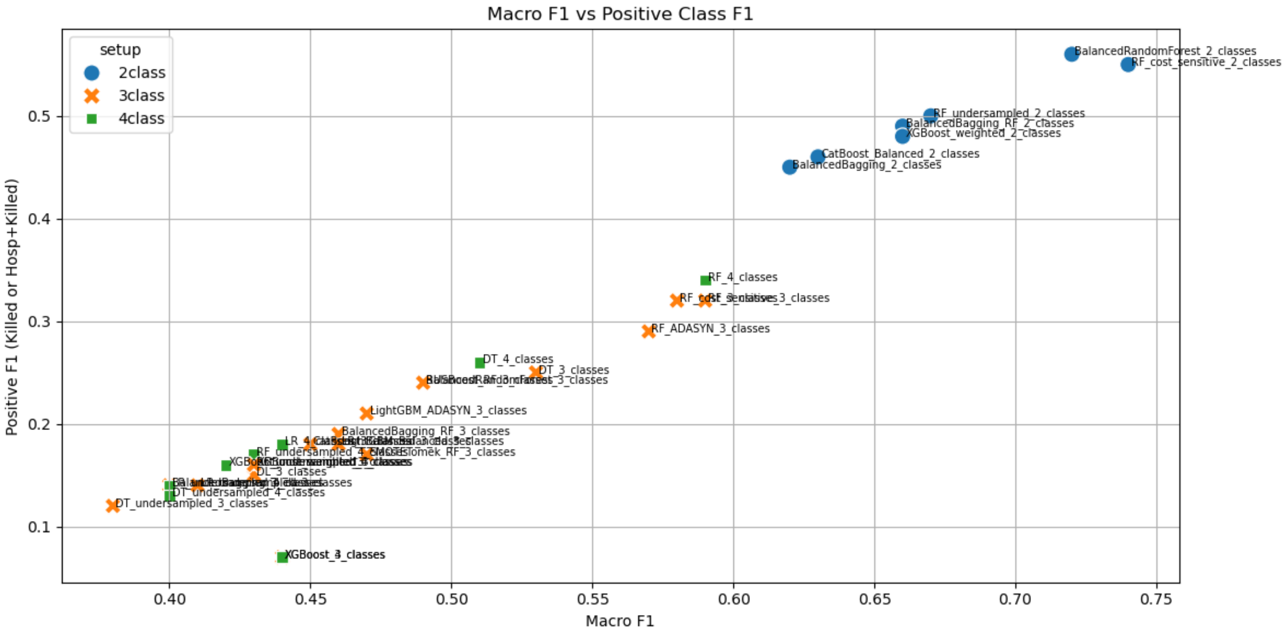


**Figure 14. Macro-F1 vs. Positive-class F1 across models and setups.** 2-class points sit in the top-right, balancing overall macro performance with strong severe-class F1. 3-/4-class points are lower, reflecting the difficulty of keeping both balance and severe-class precision/recall when labels are fragmented.

When the objective is identifying severe outcomes, the 2-class framing with balanced ensembles provides the best performance. The 3-class framing offers a reasonable compromise when a distinction between *Hospitalized* and *Killed* is needed, though at the cost of lower balanced metrics. The 4-class framing is most suitable for descriptive analytics and monitoring rather than predictive deployment, as it further fragments already rare classes.

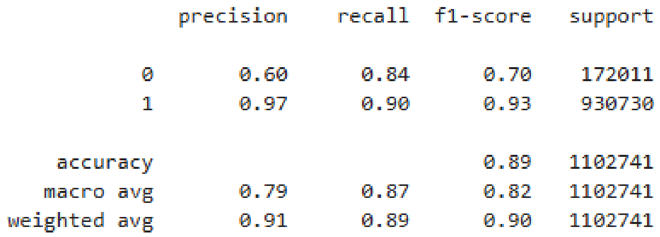
After benchmarking, we selected the 2-class formulation and the BalancedRandomForestClassifier (BRF) as the final model. This approach was retrained on the full dataset and chosen for its strong severe-class recall and robustness to imbalance.

## Interpretation of results

The final selected model is a Balanced Random Forest Classifier (2-class setup: Severe vs. Minor/None), an ensemble method based on decision trees, specifically adapted for imbalanced datasets. Unlike standard Random Forests, it builds each tree on a balanced bootstrap sample (equal number of severe and non-severe cases), which prevents the model from being dominated by the majority class. This makes it particularly well-suited for predicting rare but critical outcomes, such as severe accidents in road safety data.

To reduce computation during testing, earlier experiments were run on a 20% sample, but the final model was retrained on the complete data.

The model achieves high overall accuracy while prioritizing recall of severe accidents, aligning with the project's safety objective (**Figure 15**). Errors were examined using confusion matrix. The main issue observed was false negatives on the severe class (severe accidents classified as non-severe). To mitigate this, we applied probability threshold tuning. After adjusting the decision threshold to 0.30, severe-class recall improved to 0.97 with an F1-score of 0.95, prioritizing sensitivity to high-risk cases. In road safety, false positives (flagging non-severe as severe) are far less costly than false negatives (missing a fatal/hospitalized accident) — so lowering the threshold makes practical sense.



```
              precision    recall  f1-score   support

           0       0.60      0.84      0.70    172011
           1       0.97      0.90      0.93    930730

    accuracy                           0.89   1102741
   macro avg       0.79      0.87      0.82   1102741
weighted avg       0.91      0.89      0.90   1102741
```

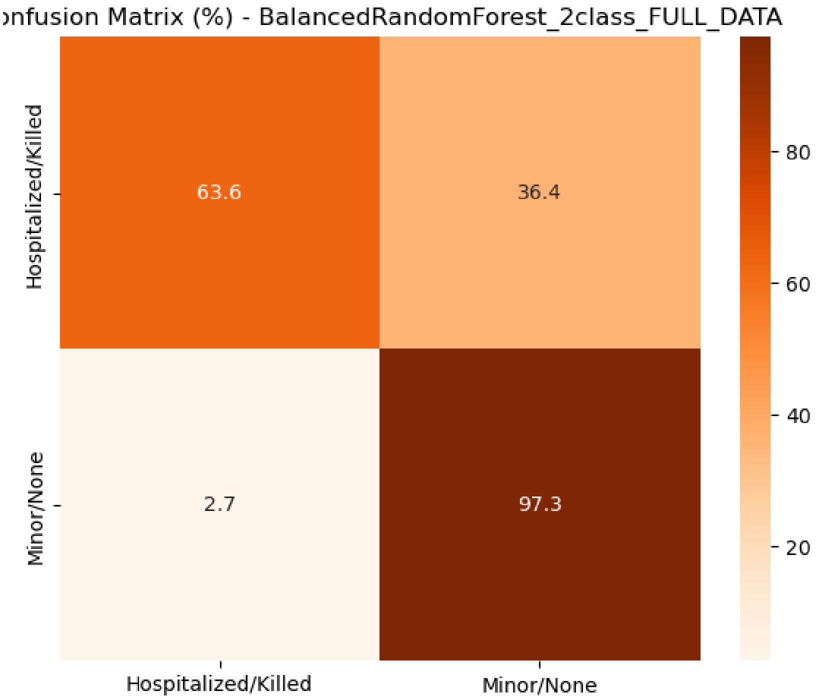


**Figure 15. Classification report and normalized confusion matrix for the final Balanced Random Forest (2-class, full data) model.**

To better understand the drivers of model predictions, we applied SHAP (SHapley Additive exPlanations) to the final Balanced Random Forest model. **Figure 16** shows the distribution of SHAP values for the most influential features, with color indicating the feature value (red = high, blue = low) and position showing the impact on the prediction.

In the top 5 categories (3 out of 5) **Seatbelt use (belt_status, also belt_user_type)** is strongest driver. Cases with belt not used (0 or -1) increase the probability of severe outcomes, while belt used (1) reduces risk. Consistent with domain knowledge, wearing a seatbelt is the most effective protective factor against severe outcomes.

2 out of top 5 are linked to **Road category and area (agg_catr combinations)**. Accidents on departmental roads outside urban areas (**cat__agg_catr_1_3)** tend to increase the risk of severe outcomes. Higher speeds and less controlled environments in non-urban settings likely explain this effect. Accidents on urban communal roads (**cat__agg_catr_2_4**) push predictions toward *Non-severe*. These roads usually have lower speed limits, traffic calming, and quicker emergency access, all reducing severity.

These results align with domain knowledge: lack of protective equipment and risky road categories are the main factors associated with higher injury severity. SHAP analysis provided trust and transparency in the model's decision-making but was not used for performance tuning.

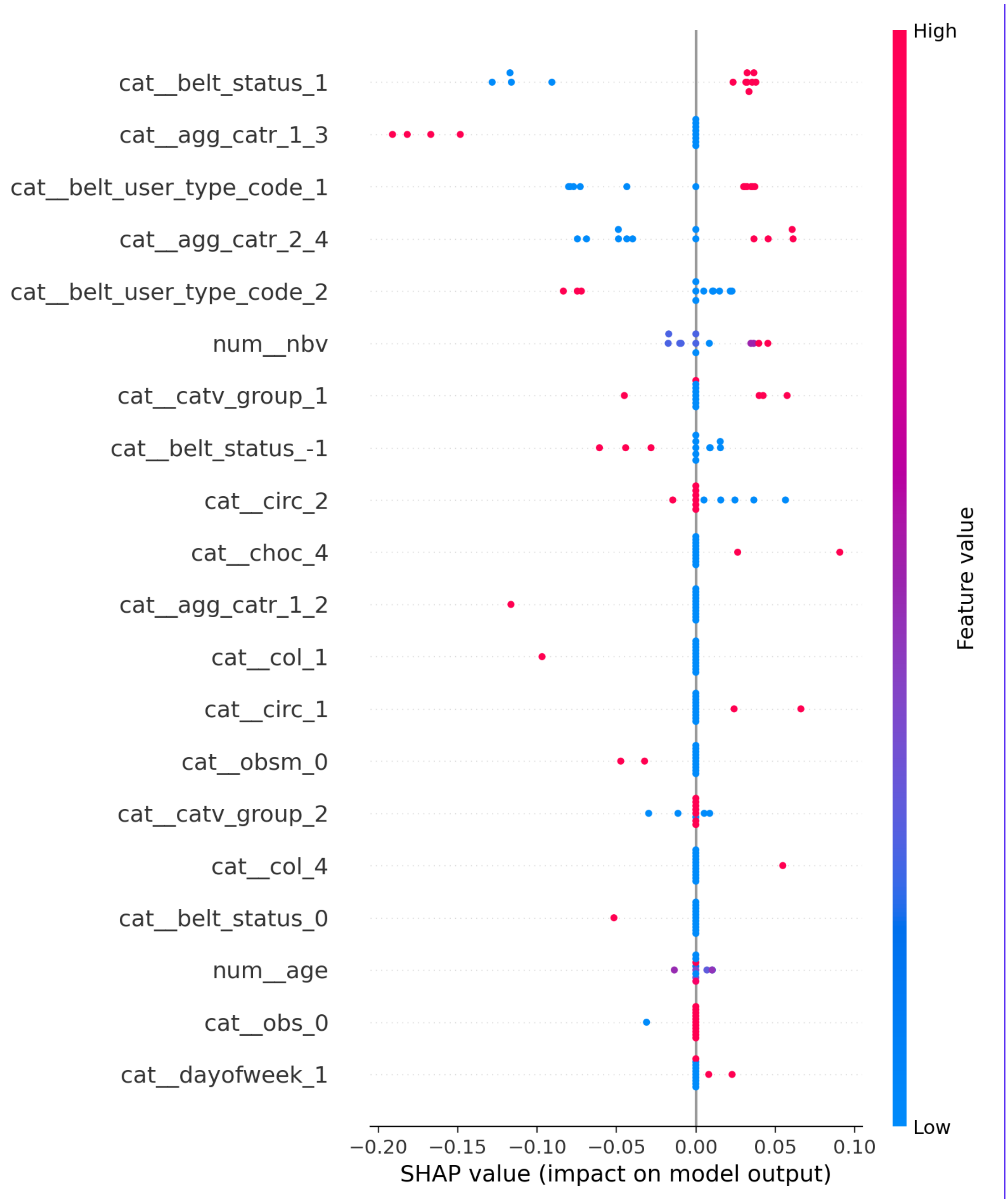


**Figure 16. SHAP summary plot for the Balanced Random Forest (2-class) model.** Left (negative SHAP values): increases the likelihood of *Severe (Hospitalized/Killed)*. Right (positive SHAP values): increases the likelihood of *Non-severe (Minor/None)*. **cat__belt_status_1** – Seatbelt used (=1); **cat__agg_catr_1_3** – Road category: outside

urban, departmental road; **cat__belt_user_type_code_1** – Driver with seatbelt; **cat__agg_catr_2_4** – Road category: urban, communal road; **cat__belt_user_type_code_2** – Driver without seatbelt; **num__nbv** – Number of lanes; **cat__catv_group_1** – Vehicle group: micromobility & cycles; **cat__belt_status_-1** – Seatbelt not specified (-1); **cat__circ_2** – Traffic regime: two-way; **cat__choc_4** – Initial impact: rear; **cat__agg_catr_1_2** – Road category: outside urban, national road; **cat__col_1** – Collision type: head-on; **cat__circ_1** – Traffic regime: one-way; **cat__obsm_0** – Mobile obstacle = 0 (none); **cat__catv_group_2** – Vehicle group: light vehicles; **cat__col_4** – Collision type: chain (3+ vehicles); **cat__belt_status_0** – Seatbelt not used (=0); **num__age** – Age; **cat__obs_0** – Fixed obstacle = 0 (not applicable); **cat__dayofweek_1** – Day of week = Monday.

To sum up, the main performance gains came from:

- Reformulating the task into 2-class (Severe vs. Non-Severe).
- Using Balanced Random Forest to handle imbalance.
- Threshold tuning based on error analysis.
- Feature engineering (belt status, type of read and area (agg_catr), type of vehicle (catv_group) and others.

# Conclusion drawn

## Difficulties encountered during the project

The main scientific obstacle encountered during this project was the strong class imbalance present in the dataset. Fatal accidents accounted for only about 2% of all cases, which created significant challenges in building predictive models. Standard classifiers tended to be biased toward the majority classes, leading to poor recall for the most critical category. As a result, much of the effort had to be directed toward testing and selecting models that could adequately cope with this imbalance.

Another difficulty was that finding the right model architecture and tuning it to handle the imbalance took considerably more time than expected. In particular, exploring several models and evaluating their performance on imbalanced data required many iterations. This extended the forecast for model development and validation compared to the initial plan.

Working with the datasets themselves also presented obstacles. Since the data spanned multiple years and sources, there was no single consistent format. Different files used varying delimiters, encodings, and header conventions, which slowed down the integration phase. In addition, there were large numbers of missing values (NaNs) in important variables, requiring additional cleaning and preprocessing steps. The datasets were also very large, which caused operations such as merging, aggregation, and training to run much longer than anticipated.

The relevance of the approaches did evolve during the project. Multi-class prediction setups (three or four classes) were tested, but their predictive performance was consistently lower than expected. This led to a strategic shift toward a two-class formulation (severe vs. non-severe accidents), which proved to be more accurate and more meaningful for the objectives of the study.

Finally, there were substantial computational and IT-related difficulties. The combination of large datasets and complex models put pressure on available resources. Memory issues were frequent, especially when merging data, running ensemble classifiers, and computing SHAP values for model explainability. Computational time was also a limiting factor, since certain tasks such as training large models or running explainability analyses took hours to complete. While storage capacity was sufficient, the main limitations came from RAM and processing power, which at times slowed the progress of the project. For this reason, most tests and experiments were carried out on only 20% of the dataset in order to save time and avoid repeated crashes. Once the models and methods were finalized, the full dataset was used for training and evaluating the final model.

## Report

My main contribution to the project was conducting a comprehensive evaluation of multiple machine learning models across different formulations of the problem. First, I carried out the preprocessing of the datasets, which involved harmonizing files coming from different years and formats, handling missing values, and ensuring that key identifiers and variables were aligned. Second, I performed an extensive evaluation of different machine learning models for accident severity prediction. By testing a wide range of algorithms under different formulations, I was able to identify the most effective setup and highlight the limits of others.

Since the last iteration, the modeling strategy has not changed. After exploring multiple families of models, the final choice remained the Balanced Random Forest in a 2-class setup (severe vs. non-severe accidents). This model was selected because it offered the best trade-off between recall on severe accidents and overall balanced performance. No further modifications were made in the last phase, as the results already showed clear improvements over other approaches.

The benchmark included a 4-class Logistic Regression model, which performed poorly on the minority (severe) class. In particular, the 4 classes setup only achieved an accuracy of 0.58, with macro recall at 0.49 and a very low positive precision (0.11) and positive F1-score (0.18). This confirmed that the multi-class formulation was not suitable for capturing rare severe cases.

In contrast, the Balanced Random Forest (2-class) setup achieved an accuracy of 0.82, macro recall of 0.78, and macro F1-score of 0.72. Most importantly, for the severe class, it reached a positive recall of 0.73, a significant improvement compared to the benchmark, though at the cost of lower precision (0.45). The positive F1-score for the severe class rose to 0.56, more than triple that of the 4-class logistic regression. These results demonstrated that the 2-class formulation with an imbalance-aware ensemble method was far more effective in meeting the project’s goals.

The project set out several main objectives:

- Studying and cleaning the dataset: This was fully achieved through a preprocessing pipeline that harmonized formats, handled missing values, and produced a merged dataset ready for modeling.
- Extracting relevant characteristics for severity prediction: This was achieved using SHAP explainability methods, which highlighted key predictors such as seatbelt usage, road category, and vehicle type.
- Developing a model to evaluate accident severity: This objective was achieved, although the problem was reformulated as a binary classification (severe vs. non-severe) rather than multi-class.
- Validating the model against historical data: The model was tested on held-out historical records. While not a full-scale deployment validation, the

evaluation confirmed its improved performance compared to the benchmark.

The final model, by distinguishing between severe and non-severe accidents, can be integrated into road safety monitoring and decision-support processes. Traffic authorities could use it to identify high-risk situations and prioritize preventive measures. Urban planners could incorporate its outputs into infrastructure design by understanding which contexts are more likely to lead to severe accidents. Furthermore, public safety campaigns could leverage the model's explainability results, for example, emphasizing the critical role of seatbelt usage, to design targeted interventions.

## Continuation of the project

There are several avenues for improvement that could be pursued to increase the performance and robustness of the model. One promising direction is feature engineering, by deriving new variables from the existing data, creating interaction terms, or incorporating contextual information such as temporal and geographic indicators. Such features could provide richer patterns for the model to capture accident severity. Another important step would be hyperparameter tuning with larger computational resources, since the experiments were constrained by memory and processing limits. With more powerful infrastructure, a more exhaustive search of the parameter space could yield further gains in performance. Finally, an interesting extension would be to refine the target variable itself. Instead of treating severity as a binary problem (severe vs. non-severe), future work could attempt a more nuanced categorization, for example distinguishing between accidents that led to hospitalization and those that resulted in fatalities. This would provide a finer-grained understanding of outcomes while still addressing the limitations observed in the 3-class and 4-class setups tested during the project.

Beyond performance improvements, the project has contributed to an increase in scientific knowledge in the field of road safety analytics. The results confirmed that binary classification is a more reliable formulation than multi-class models for predicting accident severity when dealing with strongly imbalanced data. Furthermore, the explainability analysis (via SHAP) highlighted key features such as seatbelt usage, road category, and vehicle type, reinforcing and in some cases challenging existing assumptions about the factors most associated with severe outcomes.

# Bibliography

Ministère de l’Intérieur, Observatoire national interministériel de la sécurité routière. (2023). Bases de données annuelles des accidents corporels de la circulation routière (2005–2023) [Data set]. data.gouv.fr. https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2023/

Institut National de l’Information Géographique et Forestière. (2023). Communes de France – Base des codes postaux [Data set]. data.gouv.fr. https://www.data.gouv.fr/en/datasets/communes-de-france-base-des-codes-postaux/

# Appendices

**Appendix Table 1. Complete list of models, class setups, key hyperparameters, and metrics.**

| model | accuracy | macro_precision | macro_recall | macro_f1 | positive_precision | **positive_recall** | positive_f1 |
|---|---|---|---|---|---|---|---|
| RF_undersampled_2_classes | 0.75 | 0.66 | 0.78 | 0.67 | 0.36 | **0.82** | 0.5 |
| BalancedBagging_RF_2_classes | 0.75 | 0.65 | 0.76 | 0.66 | 0.36 | **0.78** | 0.49 |
| CatBoost_Balanced_2_classes | 0.72 | 0.64 | 0.74 | 0.63 | 0.33 | **0.78** | 0.46 |
| XGBoost_weighted_2_classes | 0.74 | 0.65 | 0.75 | 0.66 | 0.35 | **0.77** | 0.48 |
| BalancedBagging_2_classes | 0.7 | 0.63 | 0.73 | 0.62 | 0.31 | **0.77** | 0.45 |
| RF_undersampled_3_classes | 0.65 | 0.43 | 0.63 | 0.43 | 0.09 | **0.74** | 0.16 |
| BalancedRandomForest_2_classes | 0.82 | 0.7 | 0.78 | 0.72 | 0.45 | **0.73** | 0.56 |
| RF_undersampled_4_classes | 0.55 | 0.43 | 0.55 | 0.43 | 0.1 | **0.73** | 0.17 |
| DL_3_classes | 0.64 | 0.44 | 0.6 | 0.43 | 0.08 | **0.71** | 0.15 |
| XGBoost_undersampled_4_classes | 0.54 | 0.43 | 0.54 | 0.42 | 0.09 | **0.71** | 0.16 |
| XGBoost_weighted_3_classes | 0.65 | 0.43 | 0.62 | 0.43 | 0.09 | **0.7** | 0.16 |
| BalancedBagging_3_classes | 0.62 | 0.42 | 0.6 | 0.4 | 0.08 | **0.7** | 0.14 |
| LR_undersampled_4_classes | 0.53 | 0.42 | 0.52 | 0.4 | 0.08 | **0.69** | 0.14 |
| LightGBM_Balanced_3_classes | 0.67 | 0.45 | 0.63 | 0.46 | 0.1 | **0.68** | 0.18 |
| CatBoost_Balanced_3_classes | 0.66 | 0.45 | 0.62 | 0.45 | 0.1 | **0.68** | 0.18 |
| LR_undersampled_3_classes | 0.63 | 0.42 | 0.59 | 0.41 | 0.08 | **0.67** | 0.14 |
| BalancedBagging_RF_3_classes | 0.66 | 0.45 | 0.62 | 0.46 | 0.11 | **0.66** | 0.19 |
| DT_undersampled_3_classes | 0.57 | 0.41 | 0.57 | 0.38 | 0.07 | **0.66** | 0.12 |
| DT_undersampled_4_classes | 0.53 | 0.41 | 0.5 | 0.4 | 0.07 | **0.64** | 0.13 |
| BalancedRandomForest_3_classes | 0.69 | 0.47 | 0.63 | 0.49 | 0.15 | **0.57** | 0.24 |
| RF_cost_sensitive_2_classes | 0.87 | 0.76 | 0.72 | 0.74 | 0.61 | **0.51** | 0.55 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| LR_4_classes | 0.58 | 0.44 | 0.49 | 0.44 | 0.11 | **0.46** | 0.18 |
| LR_3_classes | 0.8 | 0.46 | 0.52 | 0.46 | 0.12 | **0.43** | 0.18 |
| RUSBoost_RF_3_classes | 0.7 | 0.47 | 0.57 | 0.49 | 0.17 | **0.39** | 0.24 |
| DT_3_classes | 0.83 | 0.53 | 0.54 | 0.53 | 0.24 | **0.27** | 0.25 |
| SMOTETomek_RF_3_classes | 0.76 | 0.45 | 0.5 | 0.47 | 0.12 | **0.27** | 0.17 |
| DT_4_classes | 0.64 | 0.51 | 0.52 | 0.51 | 0.24 | **0.27** | 0.26 |
| RF_4_classes | 0.71 | 0.64 | 0.56 | 0.59 | 0.56 | **0.24** | 0.34 |
| RF_cost_sensitive_3_classes | 0.88 | 0.71 | 0.53 | 0.58 | 0.6 | **0.22** | 0.32 |
| RF_3_classes | 0.87 | 0.71 | 0.54 | 0.59 | 0.61 | **0.22** | 0.32 |
| LightGBM_ADASYN_3_classes | 0.84 | 0.51 | 0.46 | 0.47 | 0.22 | **0.21** | 0.21 |
| RF_ADASYN_3_classes | 0.87 | 0.7 | 0.52 | 0.57 | 0.59 | **0.19** | 0.29 |
| XGBoost_3_classes | 0.83 | 0.63 | 0.41 | 0.44 | 0.47 | **0.04** | 0.07 |
| XGBoost_4_classes | 0.63 | 0.56 | 0.43 | 0.44 | 0.47 | **0.04** | 0.07 |

**Appendix Table 2. Description of Code Files**

| Folder / File | Description | Outputs |
|---|---|---|
| src/features/<br>build_features.py | Loads yearly accident datasets (caracteristiques, lieux, vehicules, usagers) from 2005–2023, merges them into one dataset, standardizes variable types, and engineers new features | accidents_processed.csv (cleaned raw)<br>accidents_cleaned.csv (invalid/missing handled)<br>df_for_ml.csv (ML-ready) |
| src/models/<br>train_model.py | Trains a Balanced Random Forest for 2-class classification (Minor/None vs. Hospitalized/ Killed). Uses ColumnTransformer for preprocessing (numeric: imputation + scaling; categorical: imputation + OHE + variance filter). Performs GridSearchCV for hyperparameter tuning and adjusts threshold for recall/F1 balance. | Classification report<br>Confusion matrices (raw + %)<br>Best hyperparameters<br>Trained model (Joblib)<br>JSON results |
| src/models/<br>predict_model.py | Loads a trained pipeline + tuned threshold, applies identical preprocessing, generates predictions with probabilities for new data (CSV or DataFrame). | Prediction CSVs for unseen data |
| src/visualization/<br>visualize.py | Runs SHAP explainability suite on the trained model. Generates summary plots, bar/violin plots, dependence and interaction plots, SHAP heatmaps, and force plots for specific cases (false negatives, high-risk predictions). | PNG visualizations<br>shap_streamlit_summary.json (for dashboards) |