Project Outline

Multi-Factor Analysis of French Road Accident Severity

David Cebulla (1922129)
Gabriel Himmelein (1649181)
Lukas Ott (1842341)
Artur Loreit (2268917)
Aaron Niemesch (1836924)

October 14, 2025

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

1 What is the problem you are solving?

Emerging assistive technologies in modern cars enable the collection of previously undisclosed data. This offers the opportunity not only to improve accident prevention and injury reduction but also to enhance the actions taken in the accident scenario itself. An example of this is eCall, an assistive feature that has been mandatory for new cars sold in the European Union since 2018. It is capable of automatically calling emergency services while transmitting crucial data, such as the incident location and the number of passengers involved. While already helpful, the absolute severity of the accident in question remains unknown, which is crucial for evaluating the actions to be taken by first responders.

The National Interministerial Observatory for Road Safety (ONISR) in France administers the Injury Analysis Bulletin, a database containing all accident-related data from 2005 to 2023. This data includes the severity of incidents, which is concluded up to 30 days afterwards, depending on changes to a person's health status in the aftermath of the accident. Using this as a baseline, our goal is to evaluate the feasibility of developing an injury severity classifier for real-time predictions of accident data, helping first responders to take appropriate precautions.

2 What data will you use?

The analysis will be based on a subset of the "Annual Road Traffic Injury Databases" dataset from 2019 to 2023, following changes to reporting procedures and features. The data is provided in multiple files broken down by general characteristics, locations, users, and vehicles involved in accidents recorded in those years. The recorded features include types of vehicles used, impact location, road type, surface, speed limits, weather and lighting conditions, as well as user information such as age, sex, and safety equipment used. The injury severity is subject to change for up to 30 days following new information. The data is publicly available from the French government's open data portal, data.gouv.fr, and the official portal for European data, data.europe.eu, provided by the ONISR (Observatoire national interministériel de la sécurité routière). The data will be downloaded as four separate CSV files for each year. The first major task will be to merge these files using the common `Num_Acc` field.

3 3 How will you solve the problem?

3.1 What preprocessing steps will be required?

We will follow a structured approach involving focused data selection, merging, cleaning, and applying a variety of machine learning models. The preprocessing will begin with a chronological train-test split to ensure manageable training times and data relevancy. The complete dataset for the year 2022 will be used as the training set, while the complete dataset for 2023 will serve as our final, unseen test set. On the training set, we will perform hyperparameter optimization with K-Fold Cross-Validation. This strategy simulates a real-world scenario of predicting future outcomes based on the most recent available historical data. If initial results are promising and local compute allows, we can incrementally add more recent years (e.g., 2021, 2020) to the training set to see if it improves model performance. For both the training and testing years, the four separate data files will be combined into a single master table using `Num_Acc` as the key. Feature names will be replaced by their English translations to improve readability.

Initial exploratory data analysis revealed multiple preprocessing steps that need to be taken in advance for model training. While most of the features are already provided in an ordinal encoding, sparing the need to conduct the encodings ourselves, special care must be put into handling null values. This is because different features employ different methods to encode them (e.g., empty field, -1, 0). Another point of consideration is the quantity and likely meaning of null values. Even though most features comply with the 5% null-value requirement, some exceed it by a wide margin (e.g., $> 80\%$). These are either limited to features deemed irrelevant for our purposes after manual inspection or ones where the null value encodes a special meaning (e.g., not a pedestrian). Other identified issues, such as input errors (e.g., road names in road numbers field) and unsuited time formats (HH:MM), were resolved by either dropping the respective column (if irrelevant), applying suitable time formatters, or explicitly encoding null values with their context specific meaning. One of the bigger challenges associated with the dataset is that multiple user, location, or vehicle entries can belong to a single accident. To enable classification by a model, care must be taken to ensure a consistent feature encoding that includes all relevant data in a single example. Simple feature engineering approaches, such as mirroring features for users, locations, and vehicles (e.g., $x_{feature_a}$, etc.), are possible but introduce problems due to the varying number of entries each accident can have. Therefore, more complex feature engineering approaches like Principal Component Analysis (PCA) will need to be tested and evaluated.

3.2 Which algorithms do you plan to use?

To better understand the underlying risk factors impacting injury severity, we propose an initial combined usage of unsupervised and supervised machine learning models. This will be relevant for improving feature selection and engineering for the classification models. The employment of K-Means Clustering and DBSCAN as an initial unsupervised step will help with data exploration and the identification of natural groupings (accident scenarios), deepening our understanding in the earlier stages. Subsequently, several supervised classification models will be iteratively trained, evaluated, and compared with the goal of predicting injury severity. These will include Naive Bayes as a fast, simple baseline model, a Random Forest Classifier, Gradient Boosting (e.g., XGBoost), and Support Vector Machines (SVC). Finally, we will conduct a Feature Importance Analysis. Both Random Forest and Gradient Boosting models can rank features by their predictive power, which will give us a clear, data-driven list of the most significant risk factors. Using this information, the set of features will be iteratively reduced, which will lower the risk of models overfitting to the data.

4 How will you measure success?

To measure the success of our models, we will use different evaluation methods for clustering and classification. For the K-Means and DBSCAN clustering, success will be measured by the Silhouette Score combined with a qualitative review to ensure the identified "accident scenarios" are logical and interpretable. For all classification models, we recognize that simple accuracy is an insufficient metric. This is because the class distribution of the predicted variable is skewed (e.g., "killed" is a rare outcome). After using confusion matrices to determine which kinds of mistakes the model made, we will employ the weighted F1-Score. This metric provides a balanced measure of precision and recall, which is crucial for accounting for the imbalanced nature of the dataset. Additionally, given the ordinal nature of the target variable (not injured, slightly injured, heavily injured, killed), cost based evaluation procedures also influencing model training will be considered. For better visualization, we will also ensure the use of ROC Curves.

5 What do you expect your results to look like?

Our primary goal is to create models capable of classifying the injury severity of people involved in accidents to a satisfactory degree. In the process, we also expect to uncover specific, evidence-based "accident personas" and risk hierarchies from the analysis. The clustering models are anticipated to identify distinct profiles of accidents, such as low-speed urban collisions, characterized by high frequency but low severity, high-velocity rural incidents, with lower frequency but high severity, and collisions involving vulnerable road users like motorcyclists, which result in disproportionately high severity for the rider. From the feature importance analysis, we expect to confirm that vehicle type (motorcycle vs. car) is a dominant predictor of severe injury. We also predict that factors like road category and lighting conditions will be highly important. Finally, we anticipate that the driver's age and the use of safety equipment will be highly significant factors.