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

Contents

[**Relevance** 4](#_Toc199233065)

[**Pre-processing and feature engineering** 4](#_Toc199233066)

[**Visualizations and Statistics** 4](#_Toc199233067)

**Report 1: exploration, data visualization and data pre-processing report**

**Introduction to the project**

**Overview**

This project aims to predict the severity of road accidents in France using historical data collected between 2005 and 2023. The data is sourced from the BAAC (Bulletin d’Analyse des Accidents Corporels de la Circulation) and includes detailed records of injury accidents.

Accidents have a significant economic impact in terms of healthcare costs, property damage, insurance payouts, and productivity loss. This project aims to contribute to accident prevention and resource optimization, which can translate into substantial savings for governments and corporations. Predictive insights can also enable better allocation of emergency services and inform urban planning to minimize high-risk zones.

From a technical perspective, this project involves the collection, 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.
* Scoring risk zones using meteorological information and geographic data
* 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.

All datasets can be joined using the num\_acc identifier.

## **Relevance**

* Which variables seem most relevant to you with regard to your objectives?
* What is the target variable?
* What features of your dataset can you highlight?
* Are you limited by some of your data?

## **Pre-processing and feature engineering**

* Did you have to clean and process the data? If yes, describe your treatment process.
* Did you have to carry out normalization/standardization type transformations of your data? If yes, why?
* Are you considering dimension reduction techniques in the modeling part? If yes, why?

Clean and normalize

Handle missing or inconsistent data across fields. This includes:

Replacing placeholder values (-1, ., 0) with proper missing value indicators

Standardizing data types (e.g., dates, coordinates)

Removing or flagging incomplete records

Feature engineering

Extract or derive new variables that may improve model accuracy, such as:

Aggregated weather and lighting conditions

Geographical zones or clusters based on GPS

Temporal variables like time of day, weekday/weekend, season

User demographics (age groups, sex, role)

Further steps will include model training, evaluation, and visualization of severity predictions and high-risk zones.

Runaway Users:

Since 2021, data on users who fled the scene has been included.

This results in missing information such as sex, age, and injury severity (unharmed, slightly injured, hospitalized).

Missing Data:

Most variables across the four main files may contain:

Empty cells

Zeros

Periods (.)

These indicate either that law enforcement did not provide the data or that the information was not considered relevant.

Some categories also use -1 to denote 'not specified'.

Hospitalized Injured Persons:

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.

## **Visualizations and Statistics**

* Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?
* Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)
* Present the statistical analyzes used to confirm the information present on the graphs.
* Draw conclusions from the elements noted above allowing them to project themselves into the modeling part