

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Analyse des données</b>	<b>2</b>
2.1	La base de données . . . . .	2
2.2	Répartition des données . . . . .	2
2.2.1	Catégories de véhicule . . . . .	3
2.2.2	Gravité de l'accident . . . . .	3
<b>3</b>	<b>Préparation des données</b>	<b>3</b>
<b>4</b>	<b>Apprentissage</b>	<b>3</b>
<b>5</b>	<b>Audit du modèle</b>	<b>3</b>

# 1 Introduction

Ce projet a pour objectif d’analyser les accidents de la circulation routière afin de pouvoir dire à partir des données d’un véhicule accidenté si l’accident est mortel ou non. Les données sont des données libres mises à disposition par le *Ministère de l’intérieur et des Outre-Mer*. Le jeu de donnée correspond aux accidents de 2005 à 2022 en France. Nous allons dans une première partie analyser ces données afin d’extraire les informations utiles à l’apprentissage et de pouvoir repérer d’éventuelles sources de biais pour notre modèle.

## 2 Analyse des données

### 2.1 La base de données

La base de données est composée de plusieurs parties : *usagers*, *vehicules*, *lieux* et *caracteristiques*. Nous avons joint ces quatres parties pour obtenir un dataframe contenant une cinquantaine de colonnes. Ci-dessous une rapide présentation des différentes données disponibles.

Attribut	Description
<i>Num_Acc</i>	Numéro d’identifiant de l’accident
<i>jour mois</i>	Jour de l’accident, mois de l’accident
<i>an</i>	Numéro d’identifiant de l’accident.
<i>hrmn</i>	Numéro d’identifiant de l’accident.
<i>lum</i>	Numéro d’identifiant de l’accident.
<i>dep</i>	Numéro d’identifiant de l’accident.
<i>com</i>	Numéro d’identifiant de l’accident.
<i>agg</i>	Numéro d’identifiant de l’accident.
<i>int</i>	Numéro d’identifiant de l’accident.
<i>atm</i>	Numéro d’identifiant de l’accident.
<i>col</i>	Numéro d’identifiant de l’accident.
<i>adr</i>	Numéro d’identifiant de l’accident.
<i>lat</i>	Numéro d’identifiant de l’accident.
<i>long</i>	Numéro d’identifiant de l’accident.
<i>catr</i>	Numéro d’identifiant de l’accident.

### 2.2 Répartition des données

Afin de pouvoir conserver les données utiles pour l’apprentissage nous avons analysé la répartition des différentes données dans notre dataframe. Nous avons ainsi pu faire différentes observations.

Voici quelques-unes d’entre elles qui nous sont ensuite utiles pour la préparation de données.

### **2.2.1 Catégories de véhicule**

La base de données nous donne beaucoup de catégories différentes. Nous avons cependant pu remarquer que la majorité des véhicules sont dans seulement 5 catégories.

### **2.2.2 Gravité de l'accident**

En affichant l'effectif d'accidents mortels nous avons pu remarquer qu'ils ne représentent qu'une infime partie des accidents. Le peu de données sur ces accidents ne nous permet pas l'apprentissage d'un modèle. C'est la raison pour laquelle nous avons décidé de nous intéresser non pas à la mortalité à l'échelle d'une personne mais plutôt à l'échelle d'un accident. Plus précisément à l'échelle d'un véhicule car cela nous permet de conserver plus d'informations (à l'échelle d'un accident on aurait dû enlever trop d'informations pour conserver seulement les attributs plus généraux à l'accident).

## **3 Préparation des données**

## **4 Apprentissage**

## **5 Audit du modèle**