Table des matières

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 Découverte du jeu de données

2.1 La base de données

La base de données est composée de plusieurs tables : *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**
Num_Acc	Numéro d'identifiant de l'accident
jour mois	Jour de l'accident, mois de l'accident
an	Numéro d'identifiant de l'accident.
hrmn	Numéro d'identifiant de l'accident.
lum	Numéro d'identifiant de l'accident.
dep	Numéro d'identifiant de l'accident.
com	Numéro d'identifiant de l'accident.
agg	Numéro d'identifiant de l'accident.
int	Numéro d'identifiant de l'accident.
atm	Numéro d'identifiant de l'accident.
col	Numéro d'identifiant de l'accident.
adr	Numéro d'identifiant de l'accident.
lat	Numéro d'identifiant de l'accident.
long	Numéro d'identifiant de l'accident.
catr	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 des 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.

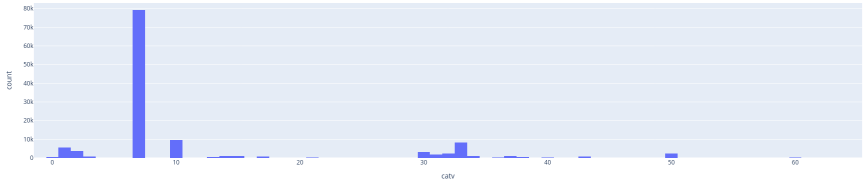

FIGURE 1 – Répartition des catégories de véhicules

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. Nous nous mettons pour cela au niveau d'un véhiule 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).

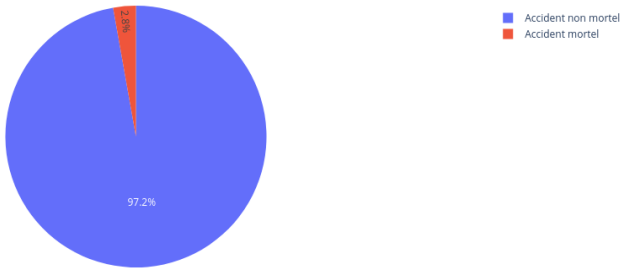

FIGURE 2 – Proportion d'accidents mortels

3 Préparation des données

3.1 Préparation des attributs utiles

À partir des observations précédentes, nous avons supprimé les attributs moins intéressant pour l'apprentissage et nous avons modifié certains attributs afin d'en extraire les informations intéressantes.
Les attributs supprimés sont : *voie*, *v1*, *v2*, *pr*, *pr1*, *lartpc*, *larrout*, *num_veh*, *occutc*, *adr*, *senc*, *etatp*, *actp*, *manv*, *jour*, *com*, *hrmn*, *motor*, *place*, *vosp*, *locp*.

Nous avons effectué les modifications suivantes :

- Création d'un attribut *mortal* qui vaut 1 si le véhicule est impliqué dans un accident mortel, 0 sinon.
- À partir de l'attribut *sexe* nous avons créé un attribut *sexe_conducteur* qui garde seulement le sexe du conducteur du véhicule.
- Nous avons utilisé l'année de naissance et l'année de l'accident pour récupérer l'âge du conducteur.
- L'attribut *vma* a été découpé en 4 catégories de vitesse.
- Pour les attributs *catv* et *vatr* nous avons gardé les valeurs le plus représentées dans la base de données.

3.2 One Hot Encoding

Une grande partie des attributs sont des attributs numériques mais sont tout de même des attributs catégoriels.

4 Analyse des données

5 Apprentissage

6 Audit du modèle