PREDICTION DE LA GRAVITE DES ACCIDENTS ROUTIERS EN FRANCE

Modélisation Supervisée, Feature Engineering et Optimisation Automatisée

Nathan WOHL
Robin KHATIB
Alexandru NITESCU

Partie 1: Introduction

La sécurité routière consiste un enjeu majeur de santé publique. Chaque année, plusieurs centaines de milliers d'accidents sont enregistrés en France, avec des conséquences humaines et économiques importantes. L'analyse des facteurs influençant la gravité d'un accident permet d'anticiper les risques et d'orienter les politiques de préventions.

Dans ce contexte, l'objectif de ce projet est de développer un modèle d'apprentissage supervisé capable de **prédire la gravité d'un accident routier** à partir des informations décrivant l'accident. Les données utilisées proviennent des bases annuelles d'accident corporels de la circulation routière publiées sur *data.gouv.fr*, comprenant pour chaque année les fichiers caractéristiques, lieux, véhicules et usagers.

Notre étude a été menée en plusieurs étapes :

- chargement, nettoyage et fusions des données ;
- exploration statistique et construction de nouvelles variables pertinentes ;
- essais de modélisation en configuration **multiclasse** (4 niveaux de gravités) ;
- analyse des performances, identification d'importantes difficultés (déséquilibre massif des classes et fortes limites de prédiction) ;
- ajustements méthodologiques, incluant la transformation de la gravité en **variable binaire** ;
- optimisation avancée des hyperparamètres via Optuna ;
- interprétation du modèle final.

Les résultats finaux montrent qu'une approche binaire « **Accident non grave** » et « **Accident grave** » permet d'obtenir un modèle performant et cohérent avec les attentes de la consigne, atteignant un score F1 supérieur à 0,73 après optimisation, tout en conservant une interprétabilité claire des facteurs de risque.

Partie 2 : Données et Préparation

2.1 Sources des données

Les données proviennent de la base publique « *Accident corporels de la circulation routière* » disponible sur *data.gouv.fr*.

Chaque année est composée de quatre fichiers CSV distincts :

- **Caractéristiques** : informations temporelles et environnementales (date, heure, météo, luminosité, type de route)
- **Lieux** : type d'infrastructure, localisation, régulation, intersection.
- **Véhicules** : catégorie du véhicule, manœuvre, obstacle, choc…
- **Usagers** : âge, sexe, place dans le véhicule…

Pour limiter le volume et accélérer le traitement, seuls **trois ans de données** ont été sélectionnés. Cela représente plusieurs centaines de milliers d'accidents, déjà largement suffisants pour entrainer des modèles. Les années **2021-2022 et 2023** ont été choisie car pour l'année 2024 tous les documents n'étaient pas disponibles sur le site.

2.2 Fusion initiale

Les fichiers ont été fusionnés autour de la clé commune **Num_Acc**, permettant de créer un dataset complet contenant, pour chaque accident répertorié, l'ensemble des informations pertinentes. Cette fusion a été réalisée via des jointures merge successives en left-join.

2.3 Nettoyage

Plusieurs opérations de nettoyage ont été effectuées afin de rendre le dataset exploitable :

- **Standardisation des variables brutes** : conversion de l'heure, harmonisation du mois, création d'un indicateur simples(weekend, nuit), et recalcul de l'âge avec imputation des valeurs manquantes.
- **Nettoyage des valeurs incohérentes ou manquantes** : gestion des formats erronés, remplacement par des valeurs neutres.
- **Encodage des variables catégorielles** : uniformisation de toutes les colonnes descriptives (type de choc, manœuvre…)
- **Création de variable agrégées et d'interactions** : nombre d'usagers/véhicules par accident, condition combinées (nuit + pluie, choc frontale…)
- **Construction de la cible binaire** : regroupement des classes BAAC en non grave et grave.

Le dataset final obtenu est propre, harmonisé et directement utilisable pour la modélisation.

2.4 Construction de la variable cible

2.4.1 Première approche : variable multiclasse

Dans une première version du projet, la gravité a été conservée selon la classification BAAC officielle :

- 1 → Indemne
- 2 → Blessé léger
- 3 → Blessé hospitalisé
- 4 → Tué

Cette variable a été ré-encodée en 4 classes (0 à 3) afin de permettre une prédiction multiclasse.

Cependant, malgré plusieurs essais de modèles et de features engineering, les performances obtenues étaient faibles et instables (fort déséquilibre, bruit dans les données), avec des scores F1 très bas sur plusieurs classes et une forte variabilité.

2.4.2 Version finale : variable binaire

La variable cible a été reconstruite en regroupant les catégories :

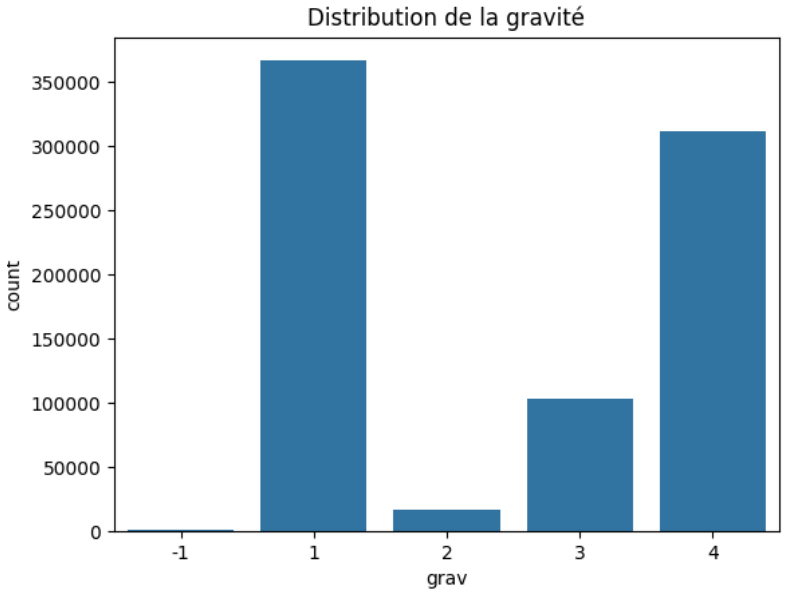
- **Non grave (0)** : indemne + blessé léger
- **Grave (1)** : blessé hospitalisé + tué

Ce regroupement permet de réduire fortement le déséquilibre, de rendre la classification plus robuste, d'amélioré la fiabilité des indicateurs (accuracy, F1-score), et de stabiliser les modèles sans overfitting. C'est cette version binaire qui constitue **la cible finale** utilisé dans les modèles optimisés.

Partie 3 : Première modélisation : approche multiclasse (échec initial)

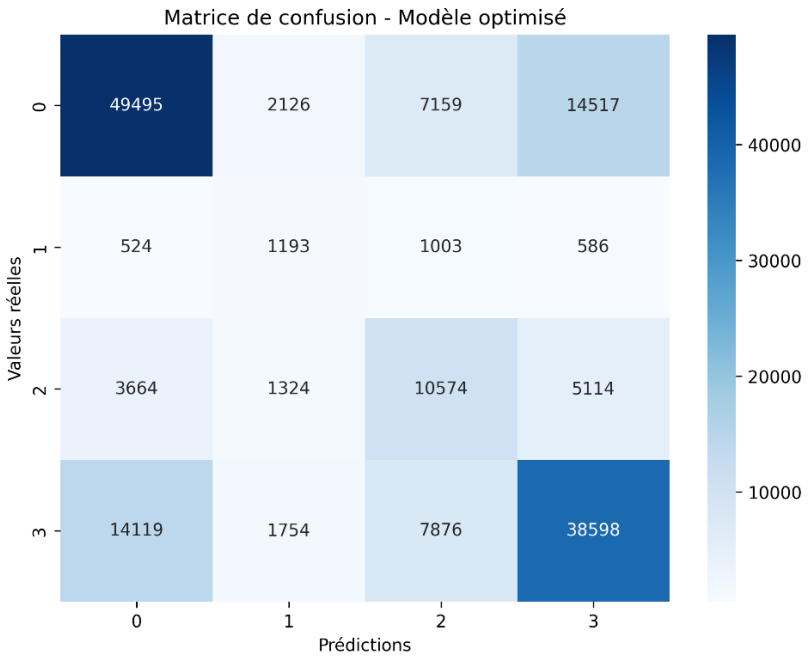
L'objectif initial était de prédire la gravité de l'accident selon la classification BAAC (4 classes de sévérité). Notre première phase a donc été menée en mode **classification multiclasse** en utilisant trois modèles standarts : **Random Forest**, **XGBoost** et **LightGBM.**

3.1 Distribution initiale de la variable cible (multiclasse)



On constate un fort déséquilibre entre les classes : la majorité des cas sont « indemnes » (1) (qu'on aura tranformé en (0)) ou « tués » (4)(pareil transformé en classe (3)), tandis que « blessé léger » et « blessé hospitalisé » sont très rare. Cette distribution rend la classification multiclasse très difficile. Les modèles risquent de prédire surtout les classes majoritaires.

3.2 Résultats des premiers modèles multiclasse (RandomForest, XGBoost, LightGBM)



Cette matrice de confusion montre que les modèles se trompent énormément sur les classes « blessé léger » et « hospitalisé » (ici 1 et 2). La classe 1 est presque impossible à reconnaître. La classe 2 est souvent confondue avec 0 ou 3.

Les performance du modèle retenue (LightGBM) étaient beaucoup trop basses : **Accuracy** = 57% et **F1-score** était inferieur à 0.50 ce qui montre que les classes étaient globalement très mal détectées.

3.3 Tentatives d'amélioration

Malgré plusieurs essais pour améliorer les performances, notamment l'ajout de nouvelles variables (nuit + pluie, route rapide + pluie, indicateurs sur le type de choc ou de véhicule, variables temporelles…), les résultats sont restés très faibles. Les classes « blessé léger » et « hospitalisé » restaient systématiquement mal identifiées, et aucun des ajustements testés n'a permis de réduire cette confusion. Cela confirme donc que la difficulté vient surtout du déséquilibre et du manque de séparation réelle entre les quatre niveaux de gravité, plutôt que d'un manque de variables ou d'un choix de modèle inadapté.

Au final, l'approche multiclasse ne se montre pas exploitable dans ce contexte. Il est donc plus pertinent de regrouper la cible en deux catégories ("accident non grave" / "accident grave") afin d'obtenir une modélisation plus stable, plus fiable et plus simple à interpréter.

Partie 4 : Modélisation binaire

Suite à l'échec de l'approche multiclasse, la variable cible a été reformulée en deux catégories :

- **0 : accident non grave** (indemne + blessé léger),
- **1 : accident grave** (blessé hospitalisé + tué)

Cette nouvelle formulation permet de stabiliser les prédictions, de simplifier le problème et d'obtenir des résultats beaucoup plus pertinents.

4.1 Construction de la cible binaire

La cible binaire grav_bin a été créée à partir des données BAAC :

- grav $\in \{\text{indemne, blesse léger}\} \rightarrow 0$
- grav $\in \{\text{hospitalisé, tué}\} \rightarrow 1$

Ce regroupement réduit le déséquilibre extrême observé précédemment.

4.2 Séparation train/ test et équilibrage

Après nettoyage, les données ont été divisées en :

- 80% pour l'entraînement
- 20% pour le test

Comme la « grave » (1) reste minoritaire, un SMOTE binaire a été appliqué sur le jeu de d'entraînement afin de générer des observations synthétiques et équilibrer les classes.

Cela évite au modèle de toujours favoriser la classe majoritaire.

4.3 Modèles testés

Trois modèles adaptés à la classification binaire ont été évalués :

- Random Forest
- XGBoost
- LightGBM

Ces modèles ont été entraînés sur les données équilibrées par SMOTE puis évalués sur le set de test réel.

Les premières performances (accuracy et F1-score) étaient déjà largement supérieures à celles obtenues en multiclasse, confirmant la pertinence de la transformation binaire.

4.4 Sélection du meilleur modèle

Parmi les trois modèles, **LightGBM** a obtenu les meilleurs scores F1 sur les données non vues, tout en conservant un temps d'entrainements faible et une bonne stabilité.

C'est donc ce modèle qui a été sélectionné pour la phase d'optimisation.

4.5 Optimisation des hyperparamètres avec Optuna

Pour améliorer encore la performance, une recherche automatique d'hyperparamètres a été menée avec **Optuna ,** en maximisant directement le F1-score sur un jeu de validation.

L'algorithme a exploré :

- profondeur maximale,
- nombres d'arbres,
- nombre de feuilles,

- learning rate ,
- taux d’échantillonnage
- colonnes utilisées

Après plusieurs essais, Optuna a proposé une configuration LightGBM nettement plus performante que les paramètres par défaut.

4.6 Performances du modèle final optimisé

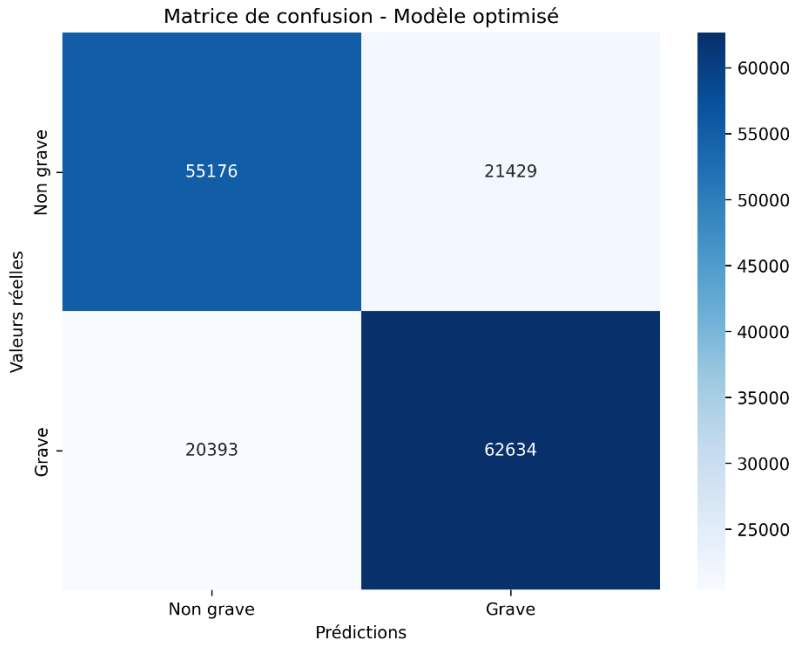
Le modèle optimisé a été ré-entraîné avec les meilleurs paramètres et évalué sur les données de test.

Il obtient :

- **Accuracy** : ~ 0.74
- **F1-score** : ~ 0.74

Ces scores montrent que le modèle arrive à détecter efficacement les accidents graves, tout e conservant un bon équilibre entre précision et rappel.

La matrice de confusion ci-dessous illustre le comportement du modèle sur les deux classes :

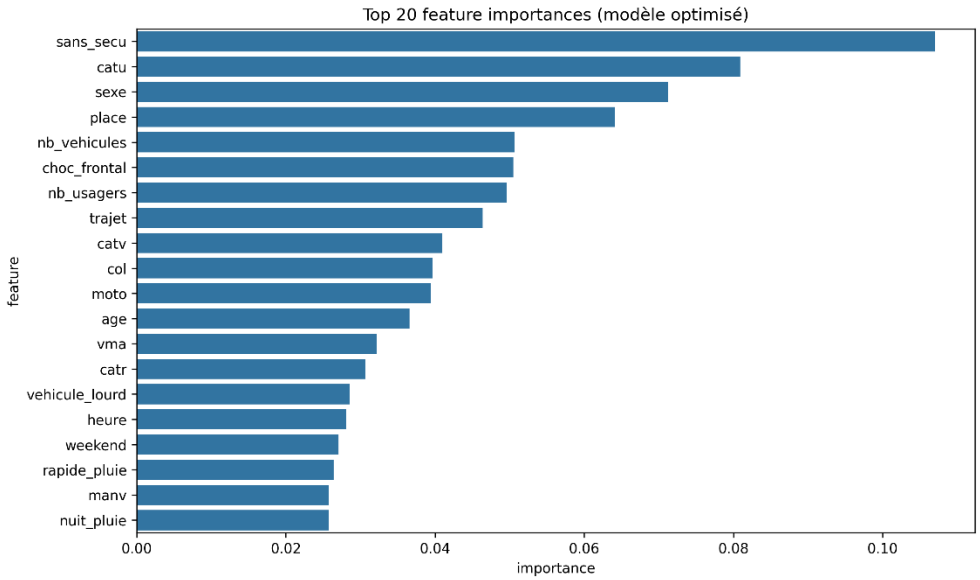


On observe que la majorité des accidents graves et non graves sont correctement identifiés.

4.7 Interprétation du modèle

Importances des variables

La figure ci-dessous présente les vingt variables les plus importantes selon le modèle LightGBM optimisé, classées par ordre décroissant d'influence.



L'analyse des importances du modèle LightGBM optimisé montre que les facteurs les plus déterminants sont :

- **sans_secu** (absence de ceinture)
- **catu** (catégorie d'usagers)
- **sexe**
- **place** dans le véhicule
- **nb_vehicules** et **nb_usagers**
- **choc_frontal** …

Ces résultats sont cohérents avec la sécurité routière : l'absence de ceinture, la catégorie d'usagers ou encore la violence du choc sont des éléments fortement liés à la gravité d'un accident.

Analyse SHAP

Les valeurs SHAP confirment les tendances déjà observées. Elles montrent :

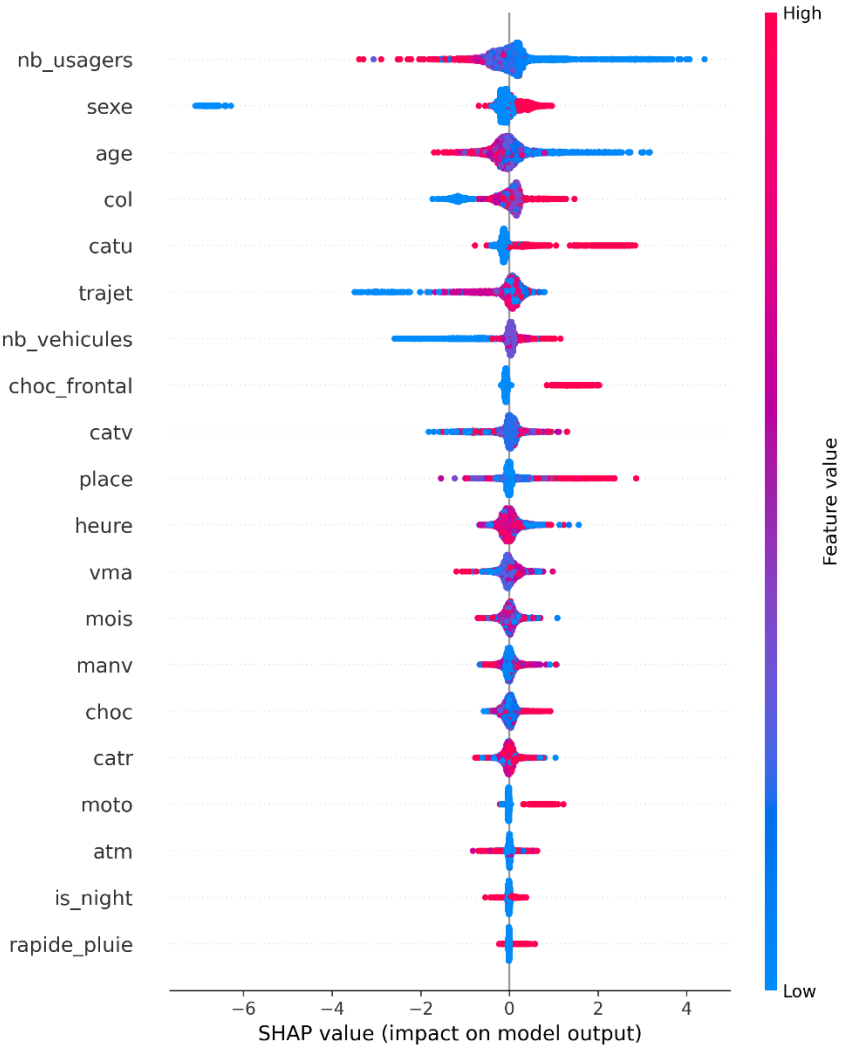
- l'impact des variables liées aux usagers (**catu**, **sexe**, **âge**),
- l'importance de la dynamique du choc (**col**, **choc_frontal**, **manv**),
- l'effet de l'exposition (**nb_usagers**, **nb_vehicules**),

- l'influence du contexte temporel et météorologique (**heure**, **mois**, **atm**).

Le graphique SHAP ci-dessous montre, pour chaque observation, comment la valeur d'une variable déplace la prédiction vers un accident *non grave* (valeurs SHAP négatives) ou *grave* (valeurs positives).

Les points rouges correspondent à des valeurs élevées de la variable, tandis que les points bleus indiquent des valeurs faibles.

Il est normal que certaines variables importantes globalement (comme **sans_secu**) n'apparaissent pas en haut du graphique : le summary plot met en avant les variables qui ont le plus contribué **localement** dans l'échantillon utilisé pour SHAP, ce qui peut différer de l'importance globale.



Partie 5 : Conclusion

Ce projet avait pour objectif de prédire la gravité d'un accident routier à partir des données détaillées issues du fichier BAAC.

L'analyse a montré que l'approche de base en classification multiclasse était difficilement exploitable, en raison du déséquilibre massif entre les classes, de la faible différence entre les niveau « blessé léger » et « hospitalisé », et des performances basses malgré les tentatives d'amélioration (nouvelles variables, équilibrage…).

Pour obtenir un modèle plus robuste, la cible a donc été reformulée en deux catégories :

Accident non grave (0) & **Accident grave (1)**

Sur cette nouvelle version binaire, le modèle optimisé (LightGBM) a comme performance :

- **Accuracy** : 0.738
- **F1-score** : 0.7374

Cela montre une bonne capacité à identifier les accidents graves tout en conservant un équilibre entre précision et rappel.

La matrice de confusion montre que les deux classes sont correctement distinguées.

L'interprétation du modèle est faite sur deux outils complémentaires.

Les importances globales montrent que des variables comme sans_secu, catu, sexe ou le choc comptent parmi les facteurs les plus déterminants.

Les valeurs SHAP, mettent en évidence l'impact local de certaines caractéristiques, du type de chic e de l'exposition (nombre d'usagers, de véhicules).

On peut conclure que les deux approches convergent vers des facteurs cohérents avec la sécurité routière.s