Application de l'analyse de classification aux accidents de la circulation routière en France

Juan Diego Alfonso, Laure Gentili , Tiavina Andriamisaina, Marwan Ait Addi

Novembre 2023

Résumé

Depuis la popularisation de l'automobile, les chercheurs étudient les accidents de la route dans le but d'améliorer la sécurité des usagers. [1]. Dans cette étude, nous cherchons à utiliser des techniques modernes de science des données pour classifier les accidents de la route en France, afin d'identifier les facteurs de risque associés. Pour ce faire, nous utilisons des données provenant de data.gouv.fr, que nous analysons à l'aide d'un algorithme de forêt aléatoire. Les résultats de cette étude nous permettrons de déterminer les facteurs qui influent la gravité d'un accident et de proposer des moyens pour les éviter.

Table des matières

1 Introduction

En 2022, l'Observatoire national interministériel a enregistré 3 267 décès liés à des accidents de la route en France métropolitaine, soit une hausse de 11% par rapport à l'année 2021. La sécurité routière est un enjeu majeur dans la société actuelle, ayant des répercussions directes sur la vie des individus.

Malgré les efforts déployés en terme d'infrastructures et de sensibilisation, le nombre de décès et de blessés sur les routes françaises reste trop significatif. Il est important de comprendre et d'identifier les facteurs déterminants afin de concevoir des stratégies toujours plus efficaces pour réduire le nombre de victimes.

Afin de saisir et de prévenir les facteurs menant à ces accidents, l'adoption d'une approche moderne et automatique s'avère nécessaire. Dans cette optique, cet article se focalise sur l'utilisation d'un outil d'apprentissage automatique, le RandomForestClassifier, dans le but de classifier les accidents routiers en France. Bien que de nombreux articles ont déjà traités la question des accidents routiers en France, la plupart n'ont utilisé que des approches limités, telles que des statistiques descriptives, qui ne sont pas suffisamment efficaces pour explorer en profondeur la complexité de ce sujet.

2 État de l'art

Les études évaluées dans divers articles ont permis d'identifier l'évolution de la collecte de données sur les accidents de la route. Cette collecte s'est grandement améliorée au fil du temps, visant à accroître sa précision. L'objectif était d'enrichir les modèles existants et de mieux identifier les inconvénients et les avantages des éléments trouvés. Ces données ont également été soumises à des comparaisons avec celles obtenues par d'autres méthodes de collecte.

L'étude présentée dans cet article se repose sur les données collectées et publiées par le gouvernement français, accessibles sur data.gouv.fr. Ces informations offrent une compilation détaillée des données recueillies de 2008 à 2022.

En tenant compte d'autres articles scientifiques, nous avons pu identifier les valeurs nécessaires à prendre en considération dans nos études précédentes, le type d'informations recueillies, et comment les traiter de manière plus efficace, en les classant selon les méthodes choisies dans cet article.

A partir des années 70, les chercheurs s'intéressent déjà aux accidents routiers en France.

- En 1970, J.L'Hoste [2] dresse une première analyse du recueil des données prises par les forces de l'ordre. Il examine la méthodologie des américains Baker, J-S et al de 1960 afin de proposer une amélioration dans la collecte d'informations et l'analyse des accidents de la route.
- En mai 1964, le ministère des transports du Québec [1] publie un dossier d'analyse et de classification des accidents de la route. Ainsi, nous pouvons accéder au formulaire de rapport d'accident utilisé par les forces de l'ordre et nous permet de mieux comprendre comment les accidents sont notés et enregistrés dans des bases de données pour être ensuite analysées.
- En 1975, J.Vallin et J-C.Chesnais [3] examinent les données depuis 1953, se concentrant sur l'évolution de la mortalité routière en France, en utilisant des statistiques descriptives. Ils concluent à une insuffisance dans les moyens mis en place pour lutter contre les accidents de la route. Bien que précurseure, cette analyse est bien trop légère et ne permet pas de comprendre en profondeur comment lutter contre ces accidents.
- En 2003, Hélène Fontaine [4] établit un premier lien entre la gravité d'un accident routier et l'âge du conducteur. En se penchant sur les caractéristiques des conducteurs âgés et la proportion des accidents impliquant ces derniers, elle conclut que la gravité d'un accident augmente avec l'âge du conducteur. Bien qu'elle ne se penche que sur une variable, son analyse très complète permet d'identifier avec précision un premier facteur déterminant dans les accidents de la route.
- En 2011, S.Shanti et Dr.R.Geetah Ramani [5] étudient la classification des collisions de véhicules en utilisant des algorithmes de Data Mining. Leur analyse très complète compare plusieurs algorithmes, démontrant l'efficacité de l'algorithme Random Tree pour classifier les accidents. Ainsi, cette étude constitue le pilier de notre analyse.

3 Dataset

Les données exploitées proviennent du ministère de l'Intérieur et répertorient tous les accidents corporels de la circulation enregistrés par les forces de l'ordre. Initialement, ces données étaient organisées en quatre tableaux distincts, chacun détaillant différentes caractéristiques des accidents :

- **Caractéristiques** : Ce tableau offre une description des circonstances générales entourant chaque accident.
- **Lieux** : Ce tableau fournit des informations sur l'emplacement précis de chaque accident.
- **Véhicules** : Ce tableau présente des détails sur le ou les véhicules impliqués dans chaque incident.
- **Usagers** : Ce tableau offre des informations sur les personnes impliquées dans chaque accident.

Par la suite, nous avons consolidé ces données dans un unique tableau afin de faciliter le processus de nettoyage. Le jeu de données utilisée regroupent ainsi 126 658 accidents répertoriés. En annexe, vous trouverez une liste des exhaustive des caractéristiques que nous avons retenues pour notre analyse.

3.1 Nettoyage, collecte et pré-traitement des données

Lors de l'évaluation initiale de nos données, nous avons constaté qu'elles bénéficiaient déjà d'une structuration et d'une normalisation efficace grâce à un pré-nettoyage préalable effectué par la source des données. Ce pré-nettoyage a grandement simplifié notre processus ultérieur. Les

données, provenant d'une source fiable, étaient organisées en tableaux facilement identifiables et ne présentaient pas d'irrégularités majeures.

Néanmoins, quelques ajustements ont été nécessaires pour adapter les données à notre méthodologie d'étude. Certains formats ne correspondaient pas avec notre approche d'analyse, nécessitant ainsi des modifications appropriées. De plus, nous avons repéré quelques valeurs supplémentaires non pertinentes pour notre étude.

La nécessité de nettoyer des caractères spéciaux ou d'éliminer les valeurs nulles ne s'est pas présentée, car ces problèmes étaient absents de notre ensemble de données. Les valeurs aberrantes identifiées ont été gérées efficacement grâce à l'encodage à chaud (one hot encoding).

Notre objectif principal lors du nettoyage des données était de parvenir à extraire les informations les plus pertinentes pour notre étude. Nous avons accordé une priorité particulière aux éléments clés, visant à rendre les variables de notre modèle aussi précises que possible. Cette approche nous a permis d'identifier des valeurs entre les ensembles de données qui méritaient d'être éliminées ou incluses dans nos résultats finaux.

En somme, bien que nos données étaient initialement bien traitées, nous avons effectué un nettoyage minutieux pour garantir la précision et la pertinence de notre analyse.

Suite à cette étape, nous avons procédé au pré-traitement des données analytiques présentées dans l'analyse descriptive. Nous avons joint les quatre tables mentionnées précédemment en utilisant la valeur commune `num_acc`, permettant ainsi l'unification de toutes les données d'une seule année en un seul ensemble de données.

3.2 Statistiques descriptives du dataset

Avant d'analyser les liens entre les différents facteurs liés aux accidents, il est pertinent d'explorer les données afin de mieux les comprendre et de sélectionner les variables pertinentes. Dans cette perspective, nous avons préalablement effectué une analyse descriptive du jeu de données.

Catégorie usager	**Fréquence**	**Pourcentage**
Conducteur	94418	74.55%
Passager	22675	17.90%
Piéton	9565	7.55%

TABLE 1 – Répartition des catégories d'usagers

Tranche d'âge	**Fréquence**	**Pourcentage**
1920-1940	2377	1.88%
1940-1960	13111	10.35%
1960-1980	31507	24.88%
1980-2000	48515	38.30%
2000 et après	28270	22.32%
Non renseigné	2878	2.27%

TABLE 2 – Répartition des tranches d'âge

Type de collision	**Fréquence**	**Pourcentage**
Deux véhicules - frontale	15053	11.88%
Deux véhicules – par l'arrière	18481	14.59%
Deux véhicules – par le côté	39844	31.46%
Trois véhicules et plus – en chaîne	8368	6.61%
Trois véhicules et plus - collisions multiples	6545	5.17%
Autre collision	30310	23.93%
Sans collision	7931	6.26%
Non renseigné	126	0.10%

TABLE 3 – Répartition des types de collisions

Genre	Fréquence	Pourcentage
Homme	84793	66.95%
Femme	39121	30.89%
Non renseigné	2744	2.17%

Table 4 – Répartition des genres

Gravité de blessure	Fréquence	Pourcentage
Indemne	53628	42.34%
Tué	3550	2.80%
Blessé hospitalisé	19260	15.21%
Blessé léger	49979	39.46%
Non renseigné	241	0.19%

Table 5 – Répartition de la gravité de blessure

4 Méthodologie

4.1 Choix de l'algorithme de classification

Pour sélectionner notre algorithme de classification, nous nous sommes appuyé sur l'étude de S.Shanti et Geetah Ramani [5], démontrant que l'algorithme de classification Random Forest est le plus performant pour notre cas d'utilisation.

Étant donné que les approches des articles antérieurs se sont révélées trop superficielles, il est important de sélectionner une méthode plus complexe, complète et adaptée à notre jeu de données. Le RandomForestClassifier, en tant qu'algorithme d'apprentissage automatique, offre l'avantage de pouvoir capturer des schémas complexes au sein des données en identifiant des interactions non linéaires entre les différentes variables. En optant pour cette approche moderne, notre objectif est de dépasser les limites des analyses conventionnelles et de fournir des conclusions plus nuancées sur les causes profondes des accidents routiers en France.

L'algorithme Random Forest a été choisi pour plusieurs raisons :

— Robustesse aux données déséquilibrées : Dans les données d'accidents de la route, il est possible que certaines classes (par exemple, les types d'accidents) soient sous-représentées, et Random Forest peut gérer ce déséquilibre.

— Gestion des variables catégorielles : Les données d'accidents de la route peuvent comporter de nombreuses variables catégorielles, comme le type de véhicule ou les conditions météorologiques. Random Forest peut traiter ce type de données sans avoir besoin de les transformer.

— Importance des caractéristiques : Random Forest fournit une mesure de l'importance des caractéristiques, aidant ainsi à comprendre quels facteurs contribuent le plus aux accidents.

— Prévention du surapprentissage : Grâce à l'utilisation de multiples arbres de décision et à l'agrégation de leurs résultats, Random Forest peut éviter le surapprentissage, un problème courant dans les modèles de machine learning.

— Flexibilité : Random Forest peut être utilisé pour des tâches de classification et de régression, le rendant flexible pour différents types d'analyses.

En implémentant le RandomForestClassifier sur un ensemble de données suffisamment grand, notre objectif est d'identifier les caractéristiques spécifiques des accidents routiers qui sont les plus prédictives de leur gravité et de leurs conséquences.

4.2 Implémentation, training dataset et Hyperparamètres

Nous avons opté pour l'implémentation des Random Forest de scikit-learn. Pour ce faire, une adaptation de notre jeu de données a été nécessaire pour utiliser des matrices numpy. Le training set est composé de 70% de nos exemples, sélectionnés aléatoirement avec la fonction sample de la librairie pandas. Afin de tester notre classifieur, nous utilisons le reste des données. Nous avons choisis de classifier sur la colonne **grav**, représentant la gravité de l'accident, bien qu'il soit possible de choisir une autre colonne.

Pour déterminer les meilleurs hyperparamètres, nous avons effectué une recherche exhaustive (en brute force) des meilleurs paramètres auprès d'une liste prédéfinie de paramètres possibles.

5 Résultats

L'algorithme RandomForestClassifier analyse les données en utilisant la gravité de l'accident comme variable cible. L'objectif est de déterminer si l'algorithme prédit correctement celle-ci afin de comprendre quels facteurs interviennent dans la prédiction de cette variable.

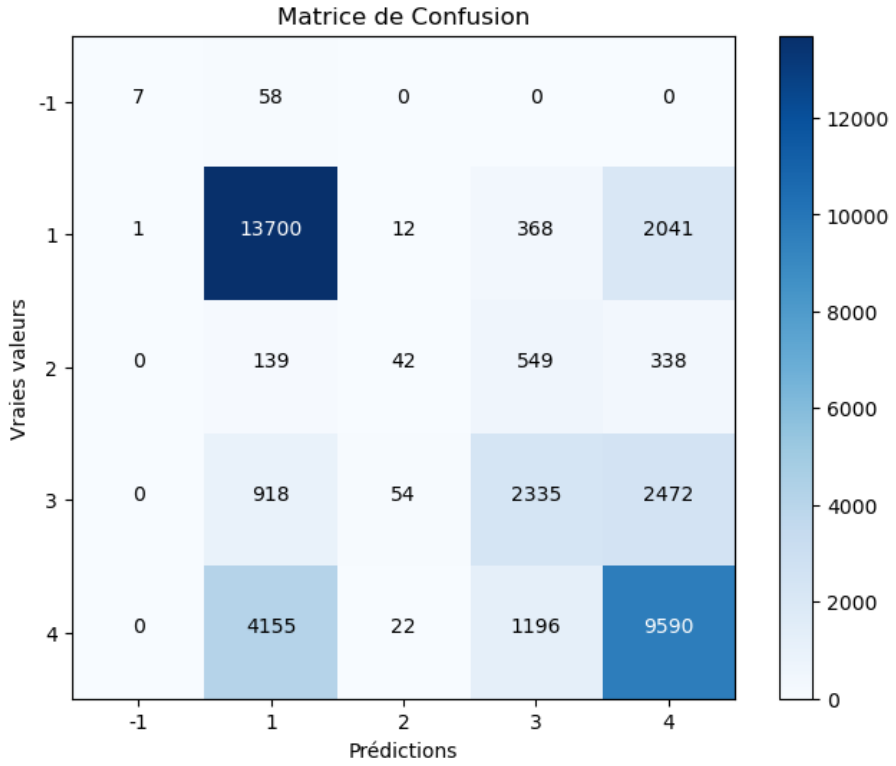

FIGURE 1 – Resultat de la classification produite par le RandomTreeClassifier : Gravité en variable cible. (-1 : Non renseigné 1 : Indemne 2 : Tué 3 : Blessé hospitalisé 4 : Blessé léger)

Gravité	Précision	Rappel	F1-Score	Support
Non renseigné	0.88	0.11	0.19	65
Indemne	0.72	0.85	0.78	16122
Tué	0.32	0.04	0.07	1068
Blessé hospitalisé	0.52	0.40	0.46	5779
Blessé léger	0.66	0.64	0.65	14963

TABLE 6 – Évaluation des performances du modèle. Variable cible : Gravité

Nous pouvons voir que le modèle choisi n'est pas exactement égal pour tout les cas.

— Non renseigné : Bien que la précision soit élevée (0,88), le rappel est très faible (0,11), indiquant des difficultés du modèle à identifier correctement cette classe. Cela pourrait être dû à un manque d'exemples d'entraînement, comme le suggère le faible support (65).
— Indemne : Le modèle semble bien performer pour cette classe, avec une précision de 0,72 et un rappel de 0,85.
— Tué : Le modèle rencontre des difficultés à classer correctement cette classe, avec une précision et un rappel faibles (0,32 et 0,04 respectivement). Cela pourrait être dû à un manque d'exemples d'entraînement, comme le suggère le support relativement faible (1068).
— Blessé hospitalisé : Le modèle montre une performance modérée pour cette classe, avec une précision de 0,52 et un rappel de 0,40.
— Blessé léger : Le modèle affiche une performance décente pour cette classe, avec une précision de 0,66 et un rappel de 0,64.

Sur le graphique illustrant l'importance des caractéristiques, nous constatons que les caractéristiques les plus pertinentes sont celles liées aux équipements de sécurité `secu1` et `secu3` utilisés dans chaque accident. Ces caractéristiques s'avèrent également pertinentes pour déterminer le type de blessure résultant de l'accident(`col`).

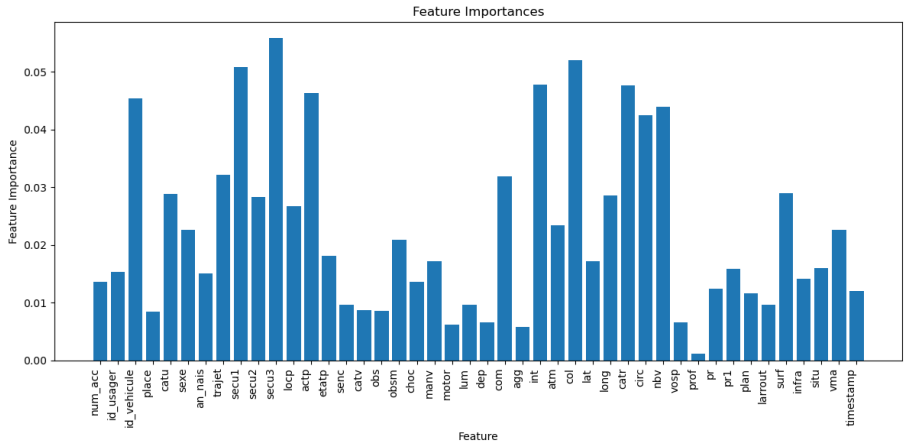

FIGURE 2 – Pertinence des variables dans la prediction de la gravité de l'accident

Ces informations peuvent être complétées par l'action du piéton(`actp`), qui fournit des indications précieuses sur les circonstances de l'accident. De plus, l'identifiant du véhicule (`id_vehicule`) offre un aperçu intéressant des types de véhicules les plus susceptibles de causer des blessures lors d'un accident.

En somme, ces caractéristiques jouent un rôle crucial dans la prédiction de la gravité des accidents de la route et peuvent nous aider à développer des stratégies plus efficaces pour prévenir ces accidents et minimiser leurs conséquences.

6 Conclusion

La classification des accidents de la route en France à l'aide de cet algorithme révèle les facteurs les plus importants à prendre en compte pour limiter les victimes.

La présence d'équipement de sécurité semble être le facteur le plus déterminant. La sensibilisation sur la nécessité de ces équipements est crucial et pourrait contribuer sans doute à réduire le nombre de blessés.

Le type de route et la localisation de l'accident semblent également jouer grandement sur la gravité des accidents. Il est donc important d'identifier ces lieux problématiques et de prendre les mesures nécessaires.

Enfin, dans les transports en commun, il est à souligner que certaines places offrent une meilleur sécurité en cas d'accident.

Références

[1] Gouvernement Canadien Ministère de la voirie. Classification et analyse des accidents, 1964.

[2] J. L'Hoste. Une Étude clinique des accidents de la circulation routiÈre. *Le Travail Humain*, 1970.

[3] J. Vallin and J.-C. Chesnais. Les accidents de la route en france. mortalité et morbidité depuis 1953. *Population (French Edition)*, 30(3) :443–478, 1975.

[4] Hélène Fontaine. Âge des conducteurs de voiture et accidents de la route : Quel risque pour les seniors ? *Recherche - Transports - Sécurité*, 79-80 :107–120, 2003.

[5] S.Shanthi and Dr.R.Geetha Ramani. Classification of vehicle collision patterns in road accidents using data mining algorithms. *nternational Journal of Computer Applications*, 35(12) :30–37, 2011.

A Choix des caractéristiques

Name	Description
num_acc	Identifiant de l'accident
id_usager	Identifiant unique de l'usager
id_vehicule	Identifiant unique du véhicule repris pour chacun de
place	Permet de situer la place occupée dans le véhicule p
catu	Catégorie d'usager
grav	Gravité de blessure de l'usager
sexe	
an_nais	Année de naissance de l'usager
trajet	Motif du déplacement au moment de l'accident
secu1 - secu2 - secu3	Présence et utilisation d'un équipement de sécurité
locp	localisation du piéton (0 si sans objet)
actp	Action du piéton
etatp	Cette variable permet de préciser si le piéton accide
senc	Sens de circulation
catv	catégorie du véhicule
obs	Obstacle fixe heurté
obsm	Obstacle mobile heurté
choc	Point de choc initial
manv	Manoeuvre principale avant l'accident
motor	Type de motorisation du véhicule
lum	Conditions d'éclairage
dep	Code du département
://www.overleaf.com/project/65607c210d395f473a6af25a agg	Localisation (Hors agglo ou en agglo)
int	intersection
atm	Conditions atmosphériques
col	Type de collision
lat	Latitude
long	Longitude
catr	Catégorie de route
circ	Régime de circulation
nbv	Nombe total de voies de circulation
vosp	Signale l'existence d'une voie réservée
prof	"Profil en long, décrit la déclivité de la route à l'end
pr	Numéro de la borne amont
pr1	Distance par rapport à la borne amont
plan	Tracé en plan
larrout	"Largeur de la chaussée hors bande d'arrêt d'urgenc
surf	Etat de la surface
infra	Aménagement
situ	Situation
vma	Vitesse maximale autorisée sur le lieu et au moment
timestamp	Heure de l'accident