Projet SARA

Projet réalisé par :

Cécile Pilon
Fadimatou Abdoulaye
Stéphane Maillard
Christophe Levra
Abdoulaye Tall

Encadré par :

Christophe - *Datascientest*

DataScientest • com

REMERCIEMENTS

Toute l'équipe projet SARA remercient DataScientest pour la qualité de la formation en Data Science et la détermination des encadreurs pendant tout le cursus. Les Masterclass, Boost Hours et les sessions de gestion de carrière ont été fortement appréciés et bénéfiques pour nos apprentissages en Machine Learning. Nous aimerions également adresser nos remerciements à notre mentor, Christophe, qui nous a accompagnés tout au long de ce projet. Sa disponibilité et son soutien sont très appréciés.

SOMMAIRE

Liste des figures

Liste des tables

1. Introduction

1.1 CONTEXTE

Les accidents de la route constituent une préoccupation majeure de santé publique pour les Pouvoirs publics, les Forces de l'ordre et l'ensemble de la société.

Chaque année, un grand nombre d'accidents se produisent, entraînant des blessures graves, voire la perte de vies humaines. Dans le but de réduire ces conséquences tragiques, il devient impératif de comprendre les facteurs qui influencent la gravité des accidents et de développer des outils prédictifs précis. Leur prévention nécessite la collaboration de nombreux acteurs (autorités gouvernementales, forces de l'ordre, experts en sécurité routière, chercheurs universitaires, compagnies d'assurances, associations de victimes d'accidents de la route, etc.) et une approche multidisciplinaire pour réduire leur fréquence et leur gravité.

Le présent projet, nommé par l'équipe projet "SARA (**Système d'Analyse et de Risque Automobile**) ", vise à utiliser des méthodes de Data Science, en se basant sur des données historiques, pour concourir à cette démarche collaborative de prédiction de la fréquence et de la gravité des accidents routiers en France.

L'exploitation de ces bases occulte néanmoins certaines données spécifiques relatives aux usagers et aux véhicules et à leur comportement dans la mesure où la divulgation de ces données porterait atteinte à la protection de la vie privée des personnes physiques aisément identifiables ou ferait apparaître le comportement de telles personnes alors que la divulgation de ce comportement pourrait leur porter préjudice.

L'importance de ce projet réside dans sa capacité à fournir des informations essentielles aux décideurs, aux autorités locales et nationales ainsi qu'aux acteurs de la sécurité routière. En anticipant la gravité potentielle d'un accident avant qu'il ne se produise, il devient possible de mettre en place des mesures préventives ciblées, d'ajuster les politiques de sécurité routière et de mobiliser les ressources adéquates pour réduire les risques et les conséquences néfastes.

Au cours de ce projet, nous mettrons en œuvre des techniques d'analyse de données avancées, telles que l'apprentissage automatique et l'exploration de données, afin de développer un modèle prédictif robuste et fiable.

Nous évaluerons également la performance de ce modèle en utilisant des métriques appropriées pour mesurer sa précision, sa sensibilité et sa spécificité.

Nous espérons que les résultats de cette étude pourront contribuer à améliorer la sécurité routière en France en fournissant des informations prédictives précieuses pour réduire le nombre d'accidents graves et les conséquences qui en découlent.

En conclusion, ce projet vise à combler une lacune importante dans la prévention des accidents routiers en France en développant un modèle de prédiction de la gravité des accidents. En exploitant les avantages de l'analyse de données, nous espérons contribuer à la réduction des risques et à l'amélioration de la sécurité routière, pour le bien-être de tous les usagers de la route.
Ce projet présente **un intérêt technique et scientifique ainsi que des enjeux multiples (économique, réglementaire, environnemental, sociétal).**

1.1.1 Point de vue technique et scientifique

Dans le domaine de la Data Science, ce projet portant sur la prédiction de la gravité des accidents routiers en France, présente un intérêt technique et scientifique.

Ce projet nécessite l'utilisation d'outils Python et de bibliothèques comme pandas, numpy, matplotlib, seaborn, missingno, scikit-learning.

Il implique l'utilisation d'algorithmes de Machine Learning et de techniques statistiques pour traiter un jeu de données volumineux, déséquilibré et diversifié.

Le projet nécessite l'application de méthodes de manipulation et de nettoyage des données, et d'extraction des variables caractéristiques pertinentes. Il nécessitera également l'utilisation de techniques de modélisation prédictive permettant un apprentissage automatique, de visualisation des résultats et enfin, d'évaluation des performances du modèle.
Ces modèles de prédiction contribueront à la compréhension scientifique des causes et des conséquences des accidents de la route. Ainsi, ces informations pourront concourir à la prévention et à la réduction de leur impact.

1.1.2 Point de vue économique

La réduction du nombre d'accidents routiers et de leur gravité est un enjeu économique important en France.

Les accidents routiers entraînent des pertes humaines, des blessures graves, des problèmes psychiques, des coûts médicaux et d'indemnisation élevés, des dommages matériels et des perturbations du trafic routier.

En développant un modèle de prédiction de la gravité des accidents, **ce projet vise à contribuer à la prévention des risques d'accident et à l'amélioration de la sécurité routière. Ce qui peut avoir un impact économique positif en réduisant les coûts associés aux accidents et en favorisant une utilisation plus efficace des ressources et investissements alloués à la sécurité routière.**

1.1.3 Point de vue règlementaire

Les accidents de la route sont régis par des réglementations et des lois visant à assurer la sécurité des usagers de la route. Les autorités publiques collectent et maintiennent des bases de données sur les accidents routiers survenus sur le territoire français.

Le projet utilisera ces données pour analyser les tendances, identifier les facteurs de risque.

Les résultats de ce projet pourraient donc avoir des implications réglementaires en matière de sécurité routière. En effet, des évolutions réglementaires pourraient, en partie, expliquer l'évolution de la fréquence et de la gravité des accidents routiers (ex : rehaussement de la vitesse autorisée, suppression de la perte de points pour les excès de vitesse inférieurs à 5 km/heure).
Par ailleurs, les autorités compétentes pourraient utiliser les informations issues de ce projet pour renforcer les réglementations visant à réduire les accidents et à améliorer la sécurité sur les routes.

1.1.4 Point de vue facteurs environnementaux

Les différents paramètres environnementaux, tels que les conditions météorologiques, les caractéristiques des routes et les divers autres facteurs environnementaux locaux, peuvent jouer un rôle important dans la survenance et la gravité des accidents de la route.

En prenant en compte ces facteurs environnementaux, le projet cherchera à évaluer leur impact sur les accidents routiers. Ces résultats pourraient aider à développer des stratégies de prévention adaptées à chaque contexte environnemental.

1.1.5 Point vue sociétal et de la santé publique

Les accidents de la route ont un impact significatif sur la société dans son ensemble et plus spécifiquement, dans le domaine de la santé publique.

Ils peuvent entraîner des pertes de vies humaines, des blessures, des traumatismes physiques et psychiques avec des conséquences émotionnelles, sociales et économiques pour les victimes, leurs proches et, voir, pour leur environnement professionnel.

Ce projet pourra contribuer à une meilleure compréhension des facteurs de risque et des conséquences.

Les résultats du projet pourront donc être utilisés par les professionnels de la santé publique pour sensibiliser les citoyens français aux comportements sécuritaires (mise en place de mesures de prévention ciblées) et améliorer la prise en charge des victimes de la route et de leurs proches.

Pour conclure sur la présentation du contexte, **en prenant en considération les aspects techniques, scientifiques, économiques, réglementaires, environnementaux, sociétaux précités,** ce projet peut fournir des informations utiles contribuant à améliorer la sécurité routière, à réduire les accidents graves et à faciliter la prise de décision en matière de prévention et de planification des mesures de sécurité routière.

Il pourra également être utile pour adapter la prise en charge des victimes et de leurs proches (ex : adaptation des ressources matérielles, des effectifs et des expertises des unités de soins en fonction de la zone géographique et de la saisonnalité).

Ce projet peut ainsi bénéficier à plusieurs acteurs parties prenantes de la sécurité routière telles que les autorités gouvernementales, les forces de l'ordre, les experts en sécurité routière, les chercheurs universitaires, les compagnies d'assurance et les associations de victimes d'accidents de la route.

Ces acteurs pourront utiliser les enseignements résultant des livrables de ce projet pour définir leurs actions en matière de sécurité routière.

1.2 OBJECTIFS ET ENJEUX

Selon les estimations de l'Observatoire national interministériel de la sécurité routière (ONISR), **3 260 personnes** ont perdu la vie sur les routes de France métropolitaine en 2022, contre 2 944 en 2021 (+10,7 %), et 3 244 en 2019 (+0,5 %), année de référence.

Ces estimations provenant des sources gouvernementales, confirment la nécessité de mettre en place des outils concrets de facilitation de la prise de décision en matière de sécurité routière. Le projet SARA (Système d'Analyse et de Risque Automobile) vise à

prédire la gravité des accidents routiers en France en appliquant des techniques de Data Science et d'apprentissage automatique sur un ensemble de données historiques.

Plus spécifiquement, l'objectif est d'identifier les facteurs et les caractéristiques qui ont une influence significative sur la gravité des accidents, tels que les conditions météorologiques, l'emplacement géographique, les caractéristiques des véhicules, les caractéristiques des conducteurs, etc.

En utilisant des techniques de modélisation prédictive basées sur les variables pertinentes, l'équipe projet testera plusieurs modèles capables d'estimer la gravité des accidents avec une certaine précision. Enfin, elle évaluera les performances de chaque modèle testé pour comparer ses prédictions avec les données historiques.

L'usage de ces modèles contribuera à la compréhension scientifique des causes et des conséquences des accidents de la route, ainsi qu'à la prévention et à la réduction de leur impact.

En effet, comme développé dans la partie « Contexte », ce projet permettra notamment de fournir des informations utiles pour la prise de décision en matière de prévention des risques d'une part, et de planification et de déploiement des mesures de sécurité routière d'autre part. Il pourra également être utile pour adapter la prise en charge des victimes et de leurs proches.

2. Compréhension et manipulation des données

2.1 PRÉSENTATION DU DATASET

2.1.1 Source des données

Nous avons utilisé quatre jeux de données concernant **les lieux, les véhicules, les usagers et les caractéristiques** des accidents survenus entre 2005 et 2021 en France. Ces données brutes proviennent des "Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2021" mises à disposition, en accès public libre, sur le site gouvernemental https://www.data.gouv.fr/.
Nous avons regroupé ces quatre jeux de données à partir des observations communes (numéro d'accident, identifiant du véhicule, numéro du véhicule et l'année de l'accident) afin de produire à l'issue d'une première analyse exploratoire, un jeu de données unique et centralisé pour bien identifier et cerner les variables présentes dans ces quatre jeux de données.

2.1.2 Exploration distincte des quatre jeux de tests

L'exploration des fichiers CSV de données sur les usagers, les véhicules, les lieux et les caractéristiques routières a été menée en suivant une série d'actions pour mieux comprendre et analyser les données disponibles. Les principales étapes entreprises pour ce prétraitement sont résumées ci-dessous :

- **Chargement des fichiers CSV :**

Les fichiers CSV contenant les données sur les usagers, les véhicules, les lieux et les caractéristiques routières ont été chargés individuellement dans un notebook spécifique. Cela a permis d'accéder aux données et de les manipuler à l'aide d'outils et de bibliothèques adaptés.

Au cours de notre exploration des quatre jeux de données sur la période 2005 à 2021 issus du site gouvernemental, nous avons découvert l'absence de données sur les années 2020 et 2021 pour le jeu de données concernant les lieux d'accidents, contrairement aux trois autres jeux de données. De façon similaire, nous avons constaté dans le cas des usagers que les numéros d'accidents étaient mal reportés sur l'année 2020. Dans les deux cas, nous avons récupéré les fichiers des années spécifiées sur le site gouvernemental, puis nous l'avons intégré pour compléter nos données sur les lieux et les usagers.

❑ Affichage des informations générales :

Pour chaque fichier CSV, des informations essentielles ont été affichées, notamment le type de données, le nombre total de variables (colonnes) et le nombre de valeurs non nulles pour chacune d'entre elles. Cette étape a permis d'avoir une vue d'ensemble des données disponibles et de détecter d'éventuelles valeurs manquantes.

❑ Affichage des premières lignes :

Pour se familiariser avec le contenu des fichiers CSV, les premières lignes ont été affichées. Cela a permis d'obtenir un aperçu des données et d'identifier les noms des variables.

❑ Description des données pour chaque type de variable :

Pour chaque variable, une description statistique appropriée a été réalisée en fonction du type de données. Pour les variables numériques, cela inclut la moyenne, l'écart-type, les valeurs minimales et maximales, et les quartiles. Pour les variables catégorielles, des comptages des différentes catégories ont été effectués.

❑ Transformation des variables en date :

Si des variables contenaient des informations temporelles, elles ont été converties au format de date pour mieux faciliter leur utilisation dans des analyses ultérieures. Il s'agit de *l'année de naissance de l'usager* ainsi que *l'année, le mois, le jour, l'heure de l'accident*.

Un aperçu des actions qui ont été effectuées :

- La variable 'hrmn' a été transformée en variable 'heure', considérant que nous n'avions pas intérêt à conserver l'information des minutes.
- La variable 'an_nais' correspondant à l'année de naissance des personnes impliquées a été transformée en 'age' pour connaître l'âge exact des usagers accidentés au moment de l'accident. Les âges ainsi obtenus ont été découpés en tranches d'âge (variable 'tranches_ages') en fonction de la distribution des accidents par âge. L'objectif est de permettre l'étude des tendances et des différences de comportement en fonction de l'âge des personnes impliquées dans les accidents lors des accidents et d'identifier les groupes d'âge qui peuvent être plus vulnérables sur les routes.

❑ Transformation des variables catégorielles :

Nous constatons que les données dont nous disposons renferment un nombre important de variables catégorielles. Nous avons choisi de convertir le type de ces variables en catégories pour permettre une meilleure compréhension des données.

En effet, la conversion d'une variable catégorielle comportant des valeurs numériques en catégories présente plusieurs avantages :

- Réduction de la taille mémoire : Lorsque nous convertissons une variable catégorielle en type catégorie, cela permet de stocker les données de manière plus compacte en mémoire. Les catégories sont représentées par des entiers, et les valeurs sont stockées dans une table de correspondance. Cela est bénéfique puisque nous travaillons avec un grand ensemble de données.
- Optimisation des opérations : En utilisant des catégories, certaines opérations peuvent être optimisées, ce qui peut accélérer le traitement des données. Par exemple, lors d'un tri ou d'un regroupement (groupby), les catégories sont plus rapides à manipuler que des valeurs numériques.
- Facilitation de l'analyse : En convertissant des valeurs numériques en catégories, nous attribuons des étiquettes significatives aux catégories, ce qui facilite l'interprétation des résultats et améliore la lisibilité des analyses. Par exemple, la variable “catégorie de véhicule (‘catv’)” possédait initialement 40 modalités différentes. Nous l’avons réduit à 21 en tenant compte du type de permis nécessaire pour chaque catégorie.
- Gestion des valeurs manquantes : Lors de la conversion en catégories, nous pouvons spécifier une catégorie spéciale pour les valeurs manquantes, facilitant ainsi leur gestion lors des analyses ou du traitement des données.
- Limitation des opérations mathématiques inappropriées : Si les valeurs numériques ont une signification catégorielle plutôt que continue, les opérations mathématiques (comme les moyennes ou les sommes) sur ces valeurs peuvent ne pas avoir de sens. La conversion en catégories permet d'éviter de telles opérations inappropriées.
- Préparation pour l’apprentissage automatique : la plupart des algorithmes de machine learning requièrent des données numériques pour fonctionner. Cela signifie que nous devons transformer nos variables catégorielles en une forme numérique adaptée avant de les utiliser dans certains algorithmes.

❑ Réalisation d'une statistique descriptive :

Une analyse statistique descriptive a été réalisée pour explorer les tendances et les caractéristiques clés des données. Cela inclut des visualisations telles que des histogrammes, des diagrammes à barres et des diagrammes en boîte pour identifier les tendances et les schémas dans les données.

❑ Nouvelles modalités de variables et variables ajoutées :

En 2019, la base de données des accidents a évolué, de nouvelles modalités de certaines variables et des nouvelles variables ont été ajoutées.

Pour donner un exemple de nouvelles modalités de variables ajoutées : l'indicateur « blessé hospitalisé » n'est plus labellisé par l'autorité de la statistique publique depuis 2019.

Les nouvelles variables ajoutées sont : "motor" (Type de motorisation du véhicule), id_vehicule (identifiant du véhicule) et "secu3" (troisième équipement de sécurité utilisé).

❑ Étude des valeurs manquantes :

Les valeurs manquantes ont été identifiées et étudiées pour évaluer leur impact sur les analyses ultérieures. Différentes approches ont été envisagées pour gérer les valeurs manquantes, telles que l'imputation ou la suppression des colonnes concernées.

❑ Suppression de certaines variables :

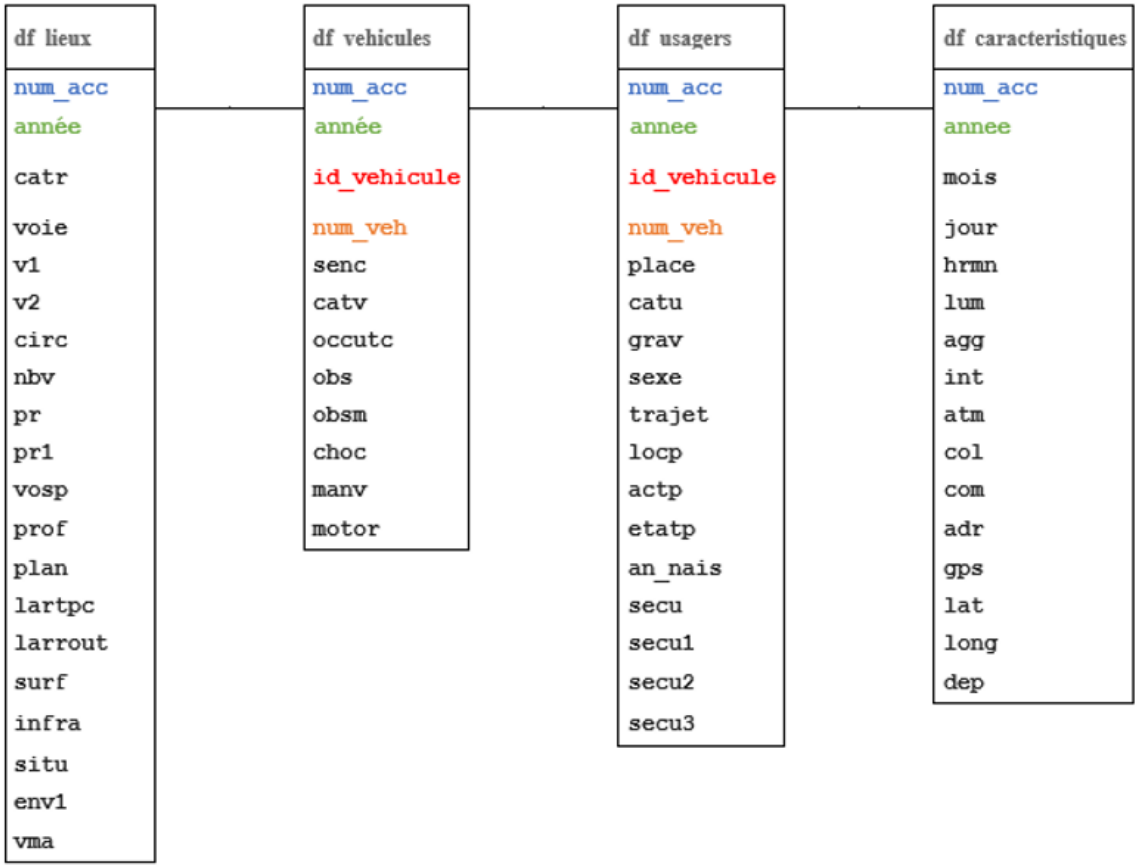
Après une évaluation approfondie des variables et de leur pertinence pour l'objectif de l'analyse, certaines variables jugées non pertinentes ou redondantes ont été supprimées pour simplifier les données et se concentrer sur les caractéristiques les plus importantes.

En somme, l'exploration des fichiers CSV de données sur les usagers, les véhicules, les lieux et les caractéristiques routières a été une étape cruciale pour comprendre la nature des données disponibles, identifier les tendances et les relations, et préparer les données pour des analyses ultérieures plus approfondies. Les actions entreprises ont permis de mieux cerner les caractéristiques des accidents routiers et d'orienter les prochaines étapes de traitement et de modélisation des données.

	Usagers	Véhicules	Caractéristiques	Lieux	Total
Catégorie	12	7	6	6	**31**
Chaîne	3	2	6	2	**13**
Entier	1	1	2	1	**5**
Décimal	2	1	2	1	**6**
Date	3	1	2	1	**7**
Total	**21**	**12**	**18**	**11**	**62**

Table 1. Récapitulatif des variables par type et par dataframe après exploration

2.1.3 Mapping Conceptuel de Données



2.1.3 Fusion des dataframes

Pour construire le dataframe global et complet, nous avons suivi une stratégie de fusion étape par étape en utilisant les variables communes entre les dataframes.

Voici l'ordre dans lequel nous avons réalisé ces fusions :

- Nous avons commencé par fusionner les dataframes "Usagers" et "Véhicules", car ces deux dataframes contiennent les variables clés 'num_acc', 'num_veh', 'annee' et 'id_vehicule' qui nous permettent d'associer les informations spécifiques sur les usagers et les véhicules impliqués dans les accidents.
- Parallèlement, nous avons fusionné les dataframes ‘caractéristiques’ avec le dataframe "Lieux". Nous avons utilisé les variables communes 'num_acc' et “annee” pour effectuer cette fusion.
- Enfin, le dataframe issu de la deuxième fusion a été fusionné avec le premier. Encore une fois, nous avons utilisé les variables communes 'num_acc' et “annee” pour réaliser cette dernière fusion.

En suivant cette approche de fusion progressive en utilisant des variables communes, nous avons réussi à créer un dataframe complet (voir Table 2.) qui rassemble toutes les informations importantes concernant les usagers, les véhicules, les lieux et les

caractéristiques des accidents, nous permettant ainsi d'effectuer des analyses approfondies et de tirer des conclusions pertinentes.

2.2 NETTOYAGE DES DONNÉES

2.2.1 Gestion des Doublons

2709 observations en doublons ont été supprimées.

2.2.2 Traitement des valeurs manquantes après fusion des jeux de données.

- **Suppression des variables avec un nombre important de valeurs manquantes**

Un certain nombre de variables contiennent un trop grand nombre de valeurs manquantes. Il s'agit des variables :

- 'id_vehicule', c'est une variable qui a été rajoutée dans la base de données en 2019. On ne la retiendra pas dans l'étude.
- 'motor', c'est également une variable qui a été rajoutée en 2019. On ne la retiendra pas non plus.
- 'adr', il s'agit de l'adresse postale du lieu de l'accident. Elle est renseignée uniquement pour les accidents survenus en agglomération et le format de saisie des données est très hétérogène et par conséquent, inexploitable (y compris, après nettoyage des données).
- 'secuTrois', c'est une variable qui a également été rajoutée dans la base de données en 2019. Elle comprend 99% de valeurs manquantes et la seule valeur renseignée dans les 1% de données restantes est "Autre". On ne retiendra donc pas cette variable dans l'étude.
- 'gps', 'lat' et 'long' : ces variables étaient initialement envisagées pour permettre l'identification des emplacements géographiques des accidents routiers en utilisant les relevés GPS. L'objectif était de dresser une cartographie des risques routiers en identifiant les endroits dangereux ou propices à ces événements.
  Cependant, lors de l'exploration de ces variables, nous avons rencontré plusieurs difficultés. La variable 'gps' contenait plus de 50% de valeurs manquantes et ne présentait aucun doublon, ce qui la rendait peu exploitable pour notre objectif spécifique.

Quant aux variables 'lat' et 'long', elles présentaient plus de 80% de valeurs manquantes. De plus, elles avaient plusieurs valeurs mal renseignées et étaient exprimées dans deux formats différents, à savoir UTM et degré décimal. Nous avons entrepris des tentatives de conversion des valeurs au format degré décimal afin d'identifier les adresses correspondantes, mais nous avons constaté que certaines valeurs se trouvaient hors de la zone UTM entre 2005 et 2021, ce qui rendait cette conversion problématique.
Dans une deuxième approche, nous avons effectué une exploration plus poussée en prenant en compte la variable 'annee' pour tenter de réaliser les conversions à partir d'une date précise. Malheureusement, nous avons remarqué qu'un très faible nombre de points GPS étaient disponibles pour effectuer des correspondances aux adresses.
En somme, nous avons constaté que ces trois variables ('gps', 'lat' et 'long') étaient très peu renseignées entre 2005 et 2021, ce qui les rend inutilisables pour notre projet. Malgré nos efforts pour explorer ces variables et réaliser la cartographie des risques routiers, leur manque de données fiables et significatives ne nous permet pas d'atteindre cet objectif spécifique.

❑ Suppression de variables intermédiaires créés lors de l'exploration et la visualisation de chaque datasets

Pour l'âge des usagers accidentés, compte tenu de la création de nouvelles variables lors de l'exploration et la visualisation de chaque datasets, après fusion des datasets, nous disposions de quatre variables en plus de la variable "tranches_ages".
Afin de simplifier notre analyse, nous avons supprimé deux des trois variables liées à l'âge, à savoir "age_acc" (exprimée en nombre de jours) et "age_acc_seconds".
Nous conserverons uniquement la variable "age_acc_an" qui donne l'âge de l'accident en années.
Par ailleurs, nous disposons également de la variable "an_naiss" qui pourrait être utile pour d'éventuelles visualisations, nous la conserverons donc.
Ces ajustements nous permettront de travailler de manière plus concise et pertinente sur les données relatives à l'âge des personnes impliquées dans les accidents.

Comme nous disposons déjà de la date de l'accident, nous avons supprimé les variables relatives à l'année, au mois et au jour de l'accident pour éviter toute redondance.

❑ Gestion des valeurs manquantes pour la variable “tranches_ages”

Nous avons choisi de supprimer les observations pour lesquelles la tranche d'âge n'était pas renseignée, car cela pourrait induire des biais dans l'analyse, en supposant que l'année de naissance n'était pas non plus renseignée.

Ces ajustements contribuent à améliorer la qualité et la fiabilité des données relatives à la sécurité des usagers, ce qui est essentiel pour des analyses pertinentes et des conclusions éclairées.

❑ Gestion des valeurs manquantes dans les variables relatives à la sécurité des usagers

Dans le cadre du traitement des variables relatives à la sécurité des usagers, nous avons effectué une analyse visuelle des données manquantes en utilisant la méthode matrix de la librairie missingno.

En examinant la matrice de visualisation des données manquantes, nous avons identifié trois variables spécifiques, à savoir 'place' (place occupée dans le véhicule par l'usager), 'secuUn' (1er équipement de sécurité utilisé), et 'secuDeux' (2nd équipement de sécurité utilisé), qui semblaient former un groupe cohérent concernant la sécurité des personnes impliquées dans l'accident.

En prenant en compte cette observation, nous avons procédé au remplacement des valeurs manquantes de ces variables par la méthode de remplacement par la valeur la plus fréquente (mode()). Cette approche nous a permis de préserver la cohérence et la validité des données relatives à la sécurité des usagers, en évitant toute distorsion dans l'analyse ultérieure.

Ces ajustements ont été effectués dans le but d'obtenir des données plus complètes et fiables, ce qui est essentiel pour réaliser des analyses précises et pertinentes sur les aspects de sécurité liés aux accidents.

❑ Gestion des valeurs manquantes dans les autres variables :

Dans la continuité de notre traitement des données manquantes, nous avons décidé de remplacer les valeurs manquantes (NAN) par la mention 'Non renseigné' pour certaines variables.

Nous avons appliqué cette approche aux variables pour lesquelles le pourcentage de données manquantes est inférieur ou égal à 5%. Cela nous permet de conserver un niveau de précision acceptable tout en évitant de perdre des informations essentielles.

Voici les variables concernées avec le pourcentage de données manquantes associé :

- 'atm' (Conditions atmosphériques) : 0.01%
- 'int' (Intersection) : 0.01% (incluant les valeurs -1 qui représentent également un manque d'information)
- 'choc' (Type de choc) : 0.01%
- 'senc' (Sens du véhicule) : 0.01%
- 'manv' (Manœuvre du véhicule) : 0.02%
- 'trajet' (Motif du trajet) : 0.02%
- 'obsm' (Obstacle fixe heurté par le véhicule) : 0.04%
- 'obs' (Obstacle mobile heurté par le véhicule) : 0.05%
- 'locp' (Localisation du piéton lors de l'accident) : 2.57%
- 'etatp' (État du piéton lors de l'accident) : 2.57%
- 'actp' (Action du piéton lors de l'accident) : 2.57%

En utilisant la méthode `fillna` avec les valeurs spécifiées, nous avons complété les données manquantes avec 'Non renseigné' pour ces variables. Cela nous permet de garantir que ces données sont prises en compte lors de nos analyses futures, tout en étant transparents sur la nature des valeurs manquantes.

Après toutes ces manipulations, nous avons obtenu un jeu de données fusionné qui compte 41 variables et 2 291797 observations.

	Dataset Fusionné
Catégorie	29
Chaîne	5
Entier	2
Décimal	3
Date	2
Total	**41**

Table 2. Nombre de variables dans le dataset fusionné

2.3 VISUALISATIONS

2.3.1 Description des variables disponibles, en particulier la variable d'intérêt ('grav')

Notre dataset résultant de la fusion des 4 rubriques possède 2192779 lignes et 64 colonnes. La majorité de ces variables peuvent être dans le contexte de l'étude considérer comme catégorielles. En effet, pour une très large majorité des variables, les modalités de celle- ci correspondent à des codes correspondant eux-mêmes aux observations des forces de l'ordre intervenues sur les lieux de l'accident.

Par exemple, notre variable cible 'grav' est encodé de la manière suivante :

- 1 = les personnes indemnes : impliquées non décédées et dont l'état ne nécessite aucun soin médical du fait de l'accident,
- 2 = les personnes tuées : personnes qui décèdent du fait de l'accident, sur le coup ou dans les trente jours qui suivent l'accident,
- 3 = les blessés dits « hospitalisés » : victimes hospitalisées plus de 24 heures,
- 4 = les blessés légers : victimes ayant fait l'objet de soins médicaux mais n'ayant pas été admises comme patients à l'hôpital plus de 24 heures.

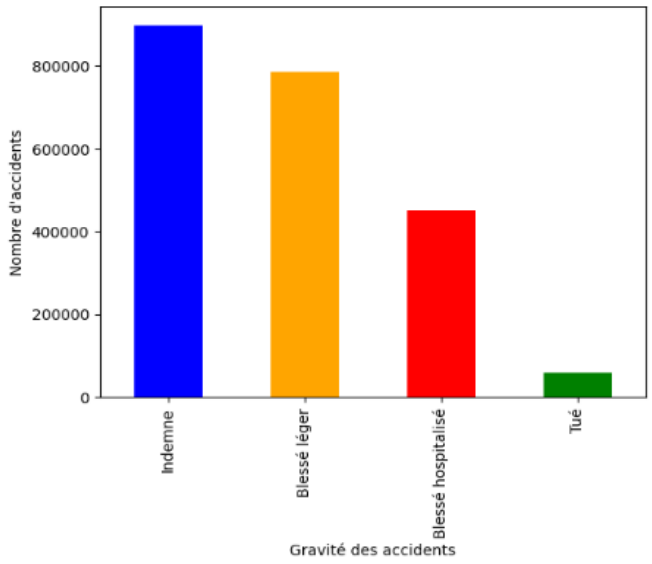


Figure 1. Distribution de la gravité des accidents

Afin de faciliter l'analyse des données, nous avons effectué une permutation dans l'ordre des modalités de la variable "grav" pour refléter un ordre de gravité croissant. Nous avons permuté les valeurs "Tué" (2) par "Blessé léger" (4) et les valeurs "Blessé léger" (4) par "Tué" (2).

De cette manière, les catégories sont désormais classées par ordre de gravité croissante, allant des accidents les moins graves (Indemne et Blessé léger) aux accidents les plus graves (Blessé hospitalisé et Tué).

Cette analyse est essentielle pour comprendre les tendances de sécurité routière et identifier les situations où des mesures préventives doivent être prises pour réduire les accidents les plus graves. Par exemple, en se concentrant sur les catégories de gravité les plus élevées, les autorités peuvent cibler les interventions et les politiques visant à améliorer la sécurité sur nos routes.

Au vue du grand nombre de variables il ne nous semble pas judicieux de retranscrire l'analyse de toutes celles-ci dans ce rapport. Nous allons plutôt résumer les observations principales.

2.3.2 Analyse temporelle

❑ Annuelle

L'analyse montre que le nombre d'accidents a légèrement diminué au fil des ans.

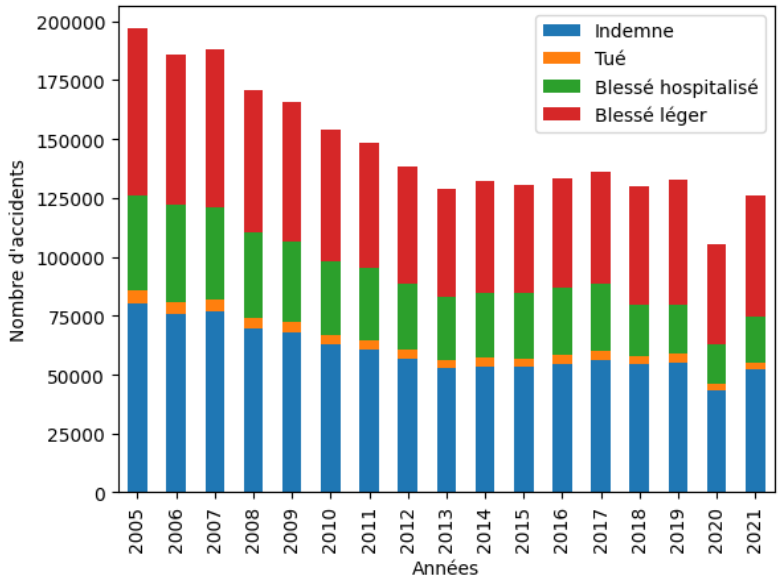


Figure 2. Distribution de nombre d'accidents par an

Le nombre d'accidents a diminué depuis 2005. On observe les effets de la politique de prévention routière, qui s'est fortement accentuée avec notamment l'apparition des radars fixes en 2003. Cependant, la distribution de la gravité des accidents n'a pas varié.

L'année 2020 est particulière et correspond aux confinements liés à l'épidémie de COVID-19.

❑ Mensuelle

Sur la période de l'étude, les mois de juin, juillet et octobre ont enregistré le plus grand nombre d'accidents. En revanche, le mois de février a enregistré le moins d'accidents.

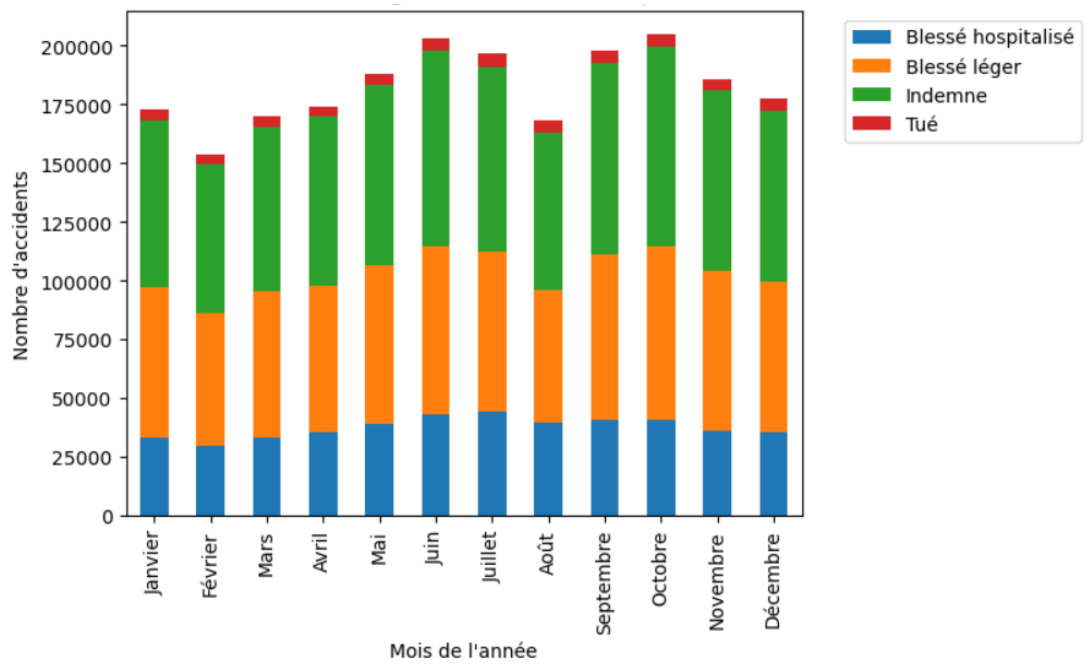


Figure 3. Distribution du nombre d'accidents par mois

❑ Hebdomadaire

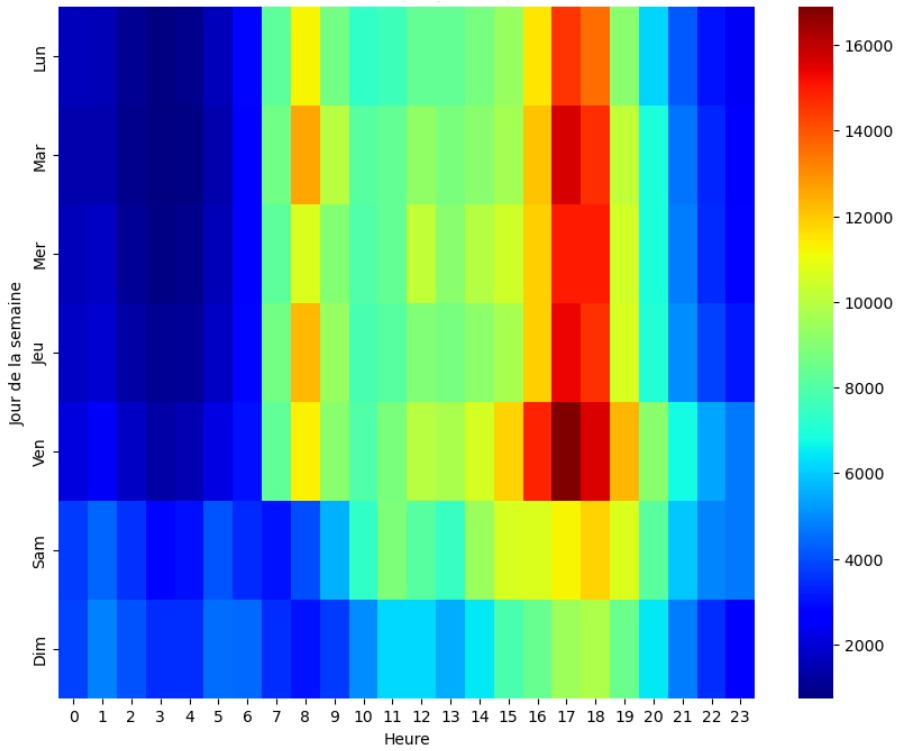


Figure 4. Distribution du nombre d'accidents par jour de la semaine et heure de la journée

On note la différence de la répartition des accidents entre la semaine et les week-ends :

- En semaine, les accidents se produisent pendant les heures de pointe correspondants aux horaires de travail.
- le week-end la répartition est plus uniforme même pendant la nuit.

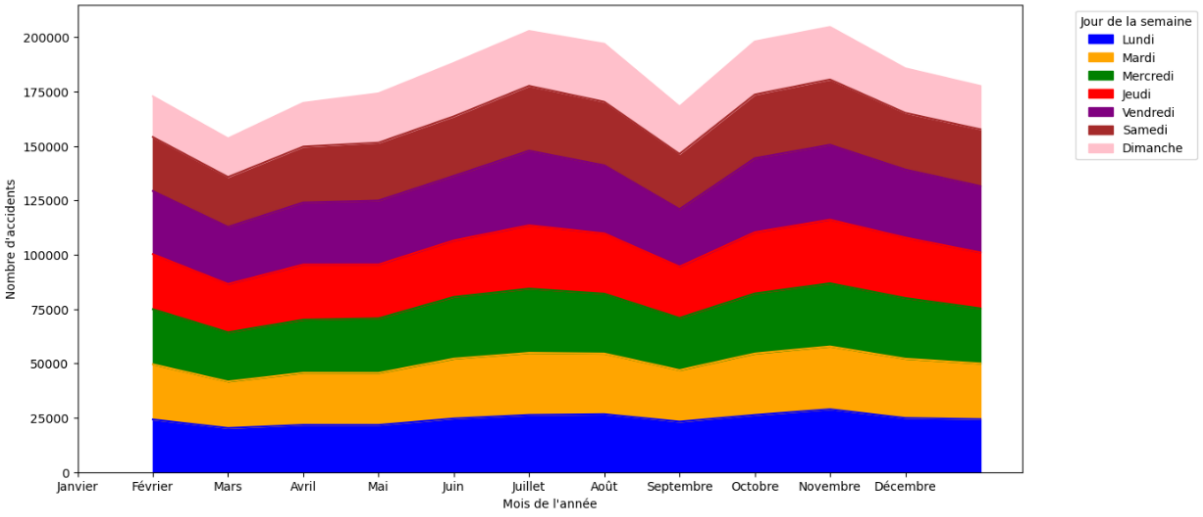


Figure 5. Evolution mensuelle du nombre d'accidents répartis sur les jours de la semaine

- **Journalier**

Selon les résultats affichés sur le graphique, nous pouvons observer que les accidents surviennent principalement le vendredi, la veille du week-end, et le samedi, peu importe le mois de l'année.
En outre, il est intéressant de noter que le nombre d'accidents est significativement élevé les dimanches du mois de juillet, tandis que les dimanches enregistrent le moins d'accidents pour les autres mois de l'année.

Ces constatations mettent en évidence des tendances importantes en termes de jours de la semaine et de mois où la prudence sur les routes est particulièrement essentielle pour assurer la sécurité de tous les usagers.

Nous pouvons croiser ces informations avec les résultats du graphique concernant les motifs de déplacement des personnes accidentées (cf. partie suivante).

2.3.3 Analyse sociologique

Si les causes des accidents de la circulation peuvent être multifactorielles, une dimension sociale reste encore à explorer : l'impact du genre sur la fréquence et la gravité des accidents routiers.
La problématique sociologique que nous abordons dans ce projet se concentre sur la question de savoir si le genre des conducteurs influe sur le nombre d'accidents de la circulation en France. Les stéréotypes de genre peuvent jouer un rôle dans les comportements au volant, les choix de véhicules et les réactions face aux situations de conduite. Ainsi, il est essentiel d'analyser comment les différences de genre peuvent influencer les taux d'accidents routiers et la gravité des conséquences pour les conducteurs et les passagers.
Il est essentiel de se pencher sur la manière dont les variables de genre jouent un rôle dans cette dynamique. Les normes sociales, les comportements au volant, les choix de véhicules et les conditions de conduite peuvent varier selon le genre, et cela peut potentiellement se refléter dans les taux de blessures graves et de décès parmi les conducteurs.

❑ Gravité des accidents selon le motif de déplacement

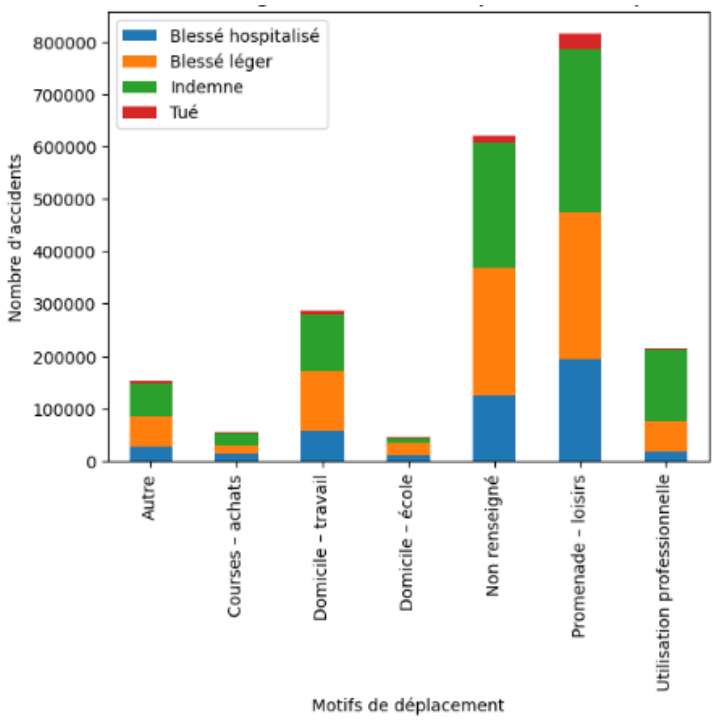


Figure 6. Distribution de la gravité des accidents selon le motif de déplacement

Il s'avère que plus de 37% des accidents sont survenus au cours d'un déplacement pour le motif 'promenade-loisirs'. Cette constatation pourrait expliquer les jours de la semaine et les mois présentant le plus grand nombre d'accidents.
En effet, nous avons indiqué ci-avant avoir observé que les accidents surviennent principalement le vendredi, la veille du week-end, et le samedi, peu importe le mois de l'année. Ces jours de la semaine sont généralement associés à des activités de loisirs et de détente, ce qui pourrait entraîner une augmentation du trafic routier pendant ces périodes.
En outre, il est intéressant de noter que le nombre d'accidents est significativement élevé les dimanches du mois de juillet, tandis que les dimanches enregistrent le moins d'accidents pour les autres mois de l'année. Le dimanche est souvent considéré comme un jour de repos et de loisirs, ce qui pourrait expliquer pourquoi les déplacements pour le motif 'promenade-loisirs' sont plus fréquents à cette période.
Ces résultats mettent en évidence des tendances importantes en termes de jours de la semaine et de mois où la prudence sur les routes est particulièrement essentielle pour assurer la sécurité de tous les usagers, en particulier lors des déplacements pour le motif “promenade-loisirs”.

❑ Gravité des accidents selon le genre des usagers

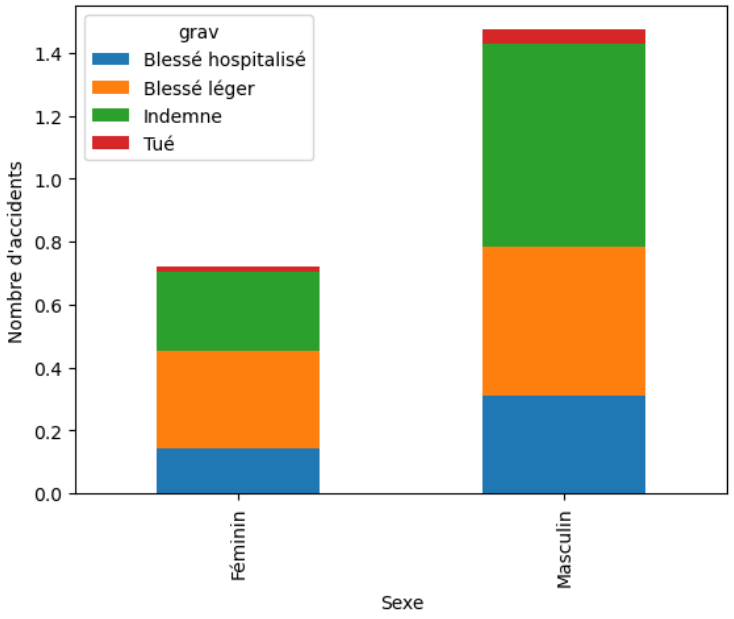


Figure 7. Répartition de la gravité des accidents selon le genre des usagers

Si les femmes semblent avoir moins d'accidents que les hommes (Figure 7.), la proportion de chaque modalité de la gravité semble être la même.

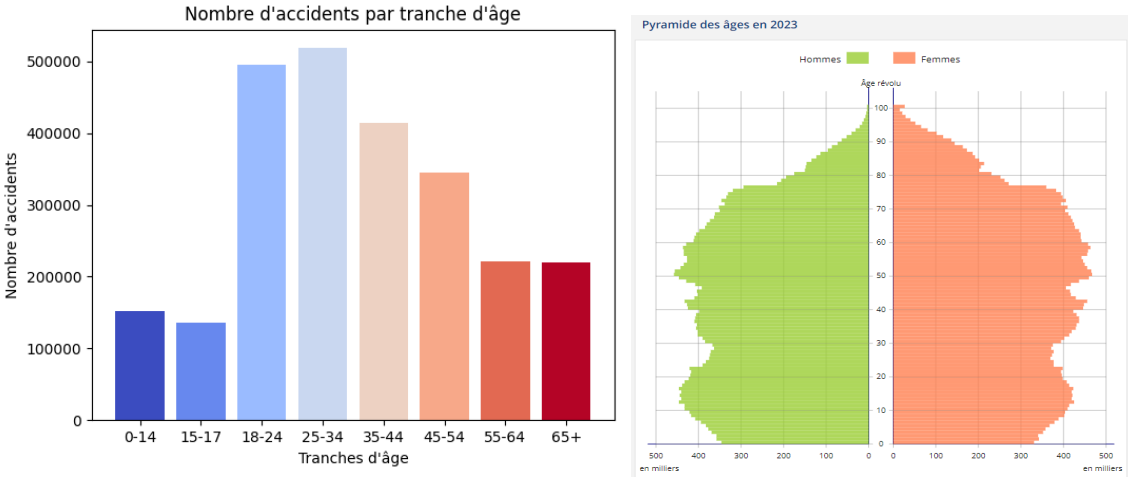


Figure 8. Répartition du nombre d'accidents par tranches d'âge

La répartition par tranche d'âge semble indiquer que les jeunes conducteurs ont plus d'accidents.

2.3.4 Analyse géographique

Ci-dessous est représentée une heatmap de la fréquence des accidents sur le territoire métropolitain. Sans surprise, les zones avec la plus forte densité d'accidents sont les zones à forte densité de population (les villes principalement) et donc de circulation.

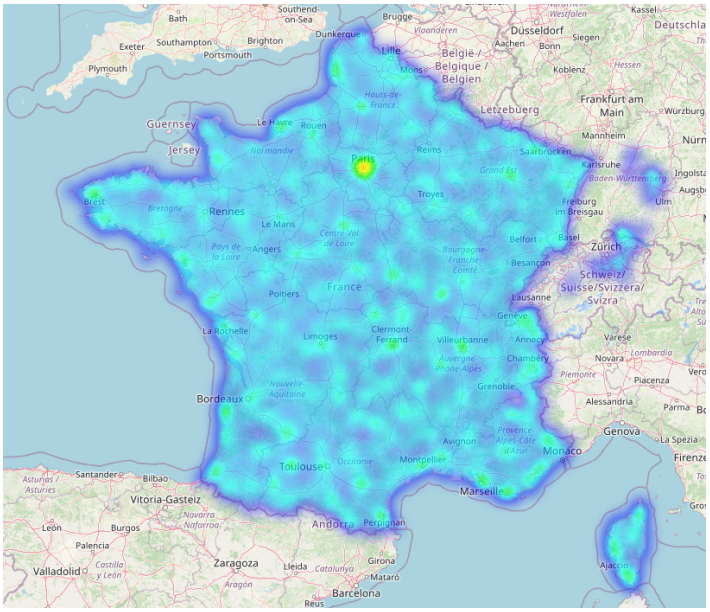
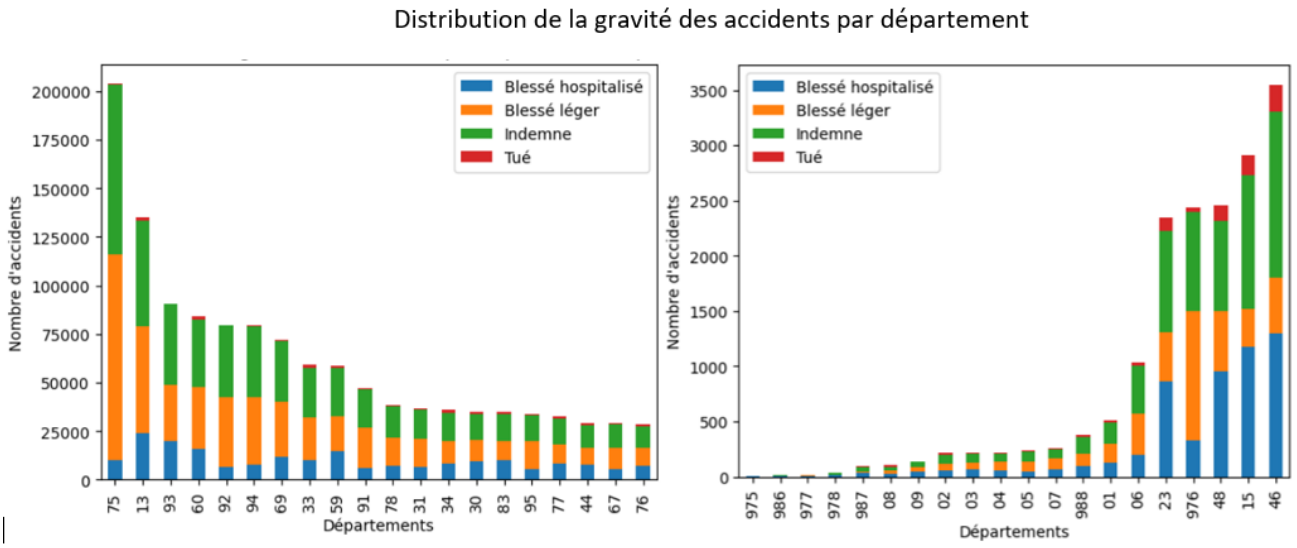


Figure 9. Fréquence des accidents en métropole



Top 20 des départements présentant le plus d'accidents

Top 20 des départements présentant le moins d'accidents

2.3.5 Autres observations

Pour un grand nombre de variables, une modalité prédomine très largement les autres, par exemple pour les conditions atmosphériques :

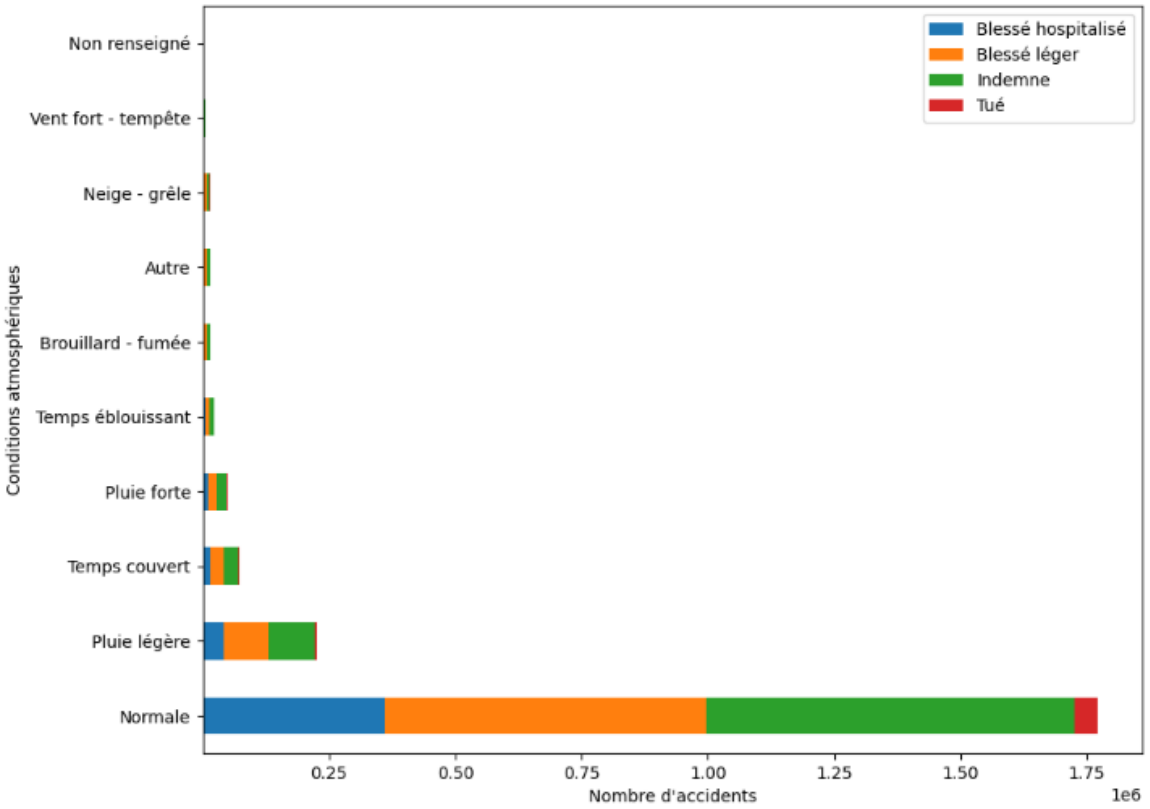


Figure 10. Répartition de la gravité des accidents selon les conditions atmosphériques

Pour d'autres variables la modalité prédominante n'apporte aucune information par exemple concernant la localisation des piétons impliqués dans l'accident :

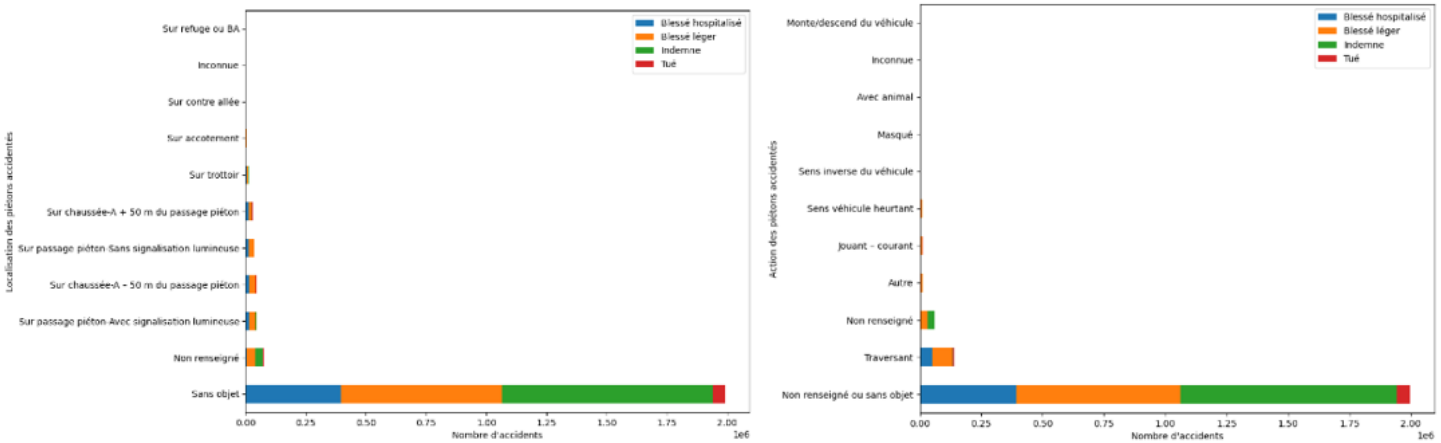


Figure 11. Répartition de la gravité des accidents selon la localisation et l'action du piéton

2.4 ANALYSES STATISTIQUES

Lors de cette étude, nous envisageons d'utiliser des analyses statistiques approfondies pour confirmer et étayer les informations présentées sur les graphiques. Ces analyses nous permettront d'approfondir notre compréhension des tendances et des relations entre les variables, en fournissant des mesures quantitatives pour étayer les observations visuelles. Nous allons utiliser des méthodes statistiques telles que la corrélation, l'analyse de variance (ANOVA), les tests de significativité et les régressions linéaires pour explorer les associations entre les différentes variables et déterminer leur impact sur les résultats des accidents routiers.

Grâce à ces analyses rigoureuses, nous allons pouvoir valider les informations clés présentées sur les graphiques, renforçant ainsi la fiabilité de nos conclusions et fournissant des preuves solides pour étayer nos recommandations en matière de sécurité routière.

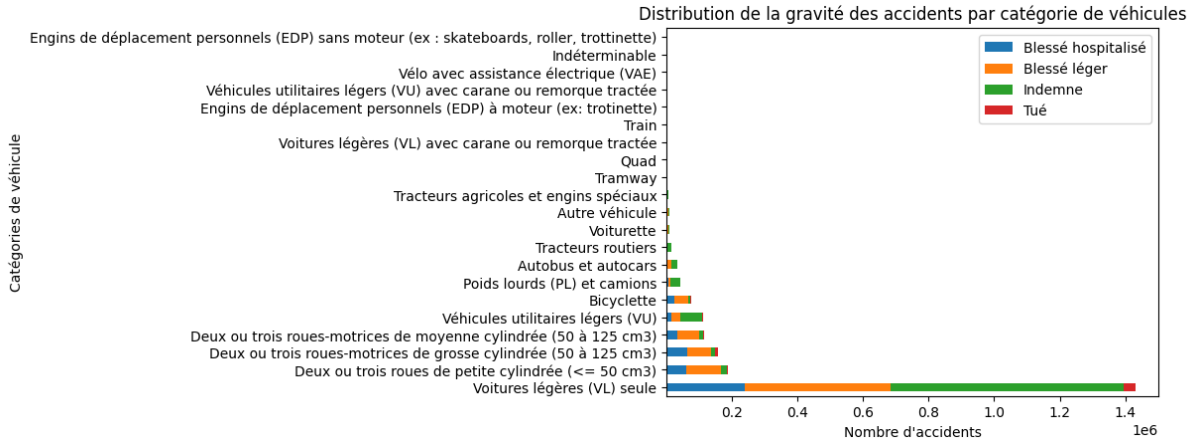
2.4.1 Analyse bivariée entre 'grav' et les autres variables catégorielles avec le test non paramétrique du Chi²

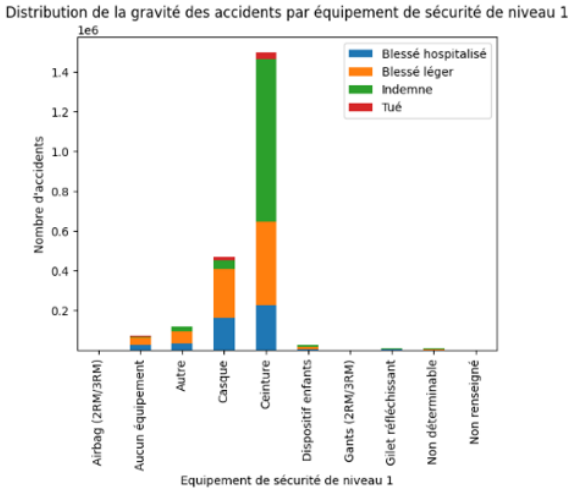
Comme indiqué précédemment, dans le cadre de cette étude, nous avons choisi d'utiliser le test du chi2 pour examiner les relations entre la variable dépendante catégorielle et chacune des variables catégorielles indépendantes ***(cf. annexe 1).***

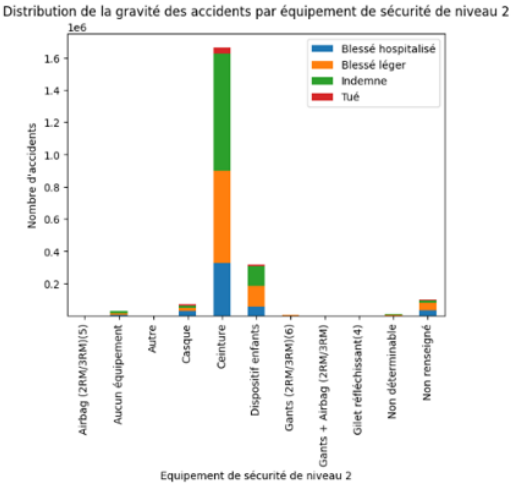
C'est un test non paramétrique utilisé pour analyser les associations entre deux variables catégorielles. Il permet de déterminer si les fréquences observées dans les différentes catégories sont significativement différentes des fréquences attendues.

Pour analyser les relations de la variable 'grav', nous avons appliqué le test de Chi2 et calculé le coefficient de Cramer.

Il est à noter qu'aucune des variables ne semble présenter une forte relation avec notre variable cible. Les deux variables ayant le coefficient de Cramer le plus élevé sont la variable "Catégorie du véhicule (‘catv’)" et 'secuDeux', qui correspond au principal équipement de sécurité utilisé par l'usager impliqué dans l'accident.







2.4.2 Analyse bivariée entre 'grav' et les autres variables catégorielles avec le test paramétrique ANOVA

L'Analyse de variance (ANOVA) est un test paramétrique utilisé pour comparer les moyennes de plus de deux groupes. L'ANOVA peut être utilisée lorsque la variable dépendante est catégorielle avec plus de deux niveaux et que la variable indépendante est catégorielle.

Après avoir utilisé l'Analyse de variance (ANOVA) pour comparer les moyennes de plus de deux groupes, nous avons présenté les résultats sous forme de statistiques F et de valeurs p ***(cf. annexe 2).***

La statistique F mesure la variance entre les groupes par rapport à la variance au sein des groupes. Une valeur de F élevée indique des différences significatives entre les moyennes des groupes.

La valeur p est la probabilité d'observer ces différences si l'hypothèse nulle est vraie. Une valeur p faible (généralement inférieure à 0,05) indique que les différences sont statistiquement significatives.

En analysant les résultats, voici quelques conclusions possibles :

- Les variables "lum", "int", "atm", "col", "catr", "vosp", "prof", "plan", "surf", "infra", "situ", "obs", "obsm", "choc", "manv", "catv_Label", "permis", "locp", "actp", "etatp", "an_nais", "secuUn", "secuDeux", "an_naiss" ont des valeurs de p très faibles (inférieures à 0,05), ce qui suggère qu'il y a des différences significatives dans la gravité des accidents entre les différentes catégories de ces variables. Ces variables jouent donc un rôle important dans l'explication des différences de gravité des accidents.
- En revanche, les variables "agg", "circ", "sexe" et "trajet" ont des valeurs de p élevées (supérieures à 0,05), ce qui indique que les différences dans la gravité des accidents entre les catégories de ces variables ne sont pas statistiquement significatives. Ces variables semblent donc moins influentes sur la gravité des accidents.

Ces conclusions concourent à la compréhension de l'impact des différentes variables sur la gravité des accidents routiers et de cibler celles qui sont les plus déterminantes.

L'ANOVA peut également être utilisée pour étudier la relation entre la gravité des accidents et les variables numériques.
L'utilisation de l'ANOVA pour les variables numériques pourrait nous permettre d'identifier les différences significatives dans la gravité des accidents en fonction des valeurs numériques spécifiques de certaines variables (par exemple : l'âge des personnes impliquées, le nombre d'occupants du véhicule, etc.). Cela vous donnerait une compréhension plus approfondie de l'impact de ces variables numériques sur la gravité des accidents routiers.

Conclusion intermédiaire

L'analyse du jeu de données révèle que peu de variables présentent une relation forte avec notre variable cible.

Cette conclusion est basée sur les résultats de deux tests statistiques, à savoir le test du Chi² et l'ANOVA, que nous avons réalisés pour évaluer les associations entre les variables explicatives et la gravité des accidents.

Ces tests nous ont permis de quantifier la relation entre chaque variable explicative et la variable cible (gravité des accidents).

Malheureusement, les résultats indiquent qu'il n'existe que peu de variables ayant une influence significative sur la gravité des accidents, ce qui rend notre objectif de construire un modèle prédictif plus complexe.

Face à cette situation, il se peut qu'il soit nécessaire d'enrichir notre dataset en ajoutant des données supplémentaires, ce qui pourrait nous permettre de capturer davantage de caractéristiques et de relations potentielles avec la gravité des accidents.

En attendant, nous avons décidé de procéder à la construction de nos premiers modèles prédictifs en utilisant les données disponibles. Nous avons ainsi entraîné deux modèles, à savoir le modèle Random Forest et l'arbre de décision, afin de voir comment ils se comportent malgré les limitations actuelles du jeu de données.

Ces modèles nous permettent de commencer à explorer les performances de prédiction et d'identifier les caractéristiques importantes dans l'estimation de la gravité des accidents. Cependant, nous sommes conscients que l'enrichissement du dataset en ajoutant des données supplémentaires pourrait améliorer considérablement les résultats de nos modèles.

En parallèle, nous continuons à peaufiner le preprocessing de nos données pour garantir que nous exploitons au mieux les informations actuellement disponibles. Une fois que nous aurons obtenu un dataset enrichi, nous pourrons alors entraîner nos modèles sur des données plus complètes et affiner leur performance de prédiction pour atteindre notre objectif de construire un modèle prédictif robuste et précis pour estimer la gravité d'un accident.

3. Modélisation

3.1 Sélection des variables et pré processing

3.1.1 Sélection des variables

Dans le processus de sélection des variables, notre objectif était de déterminer les caractéristiques les plus pertinentes pour notre analyse, avec l'intention d'améliorer la précision des modèles à tester. Nous avons suivi une approche visant à éliminer le bruit et à se concentrer sur les aspects les plus significatifs des données.

Voici les étapes clés que nous avons entreprises dans cette démarche :

- **Transformation des dates :**

Pour faciliter l'analyse temporelle, nous avons converti la colonne "date" au format datetime, tout en extrayant le mois et le jour de la semaine à partir de cette date.

- **Suppression des colonnes non pertinentes :**

Nous avons exclu certaines colonnes de l'ensemble de données, considérant qu'elles ne contribuaient pas de manière significative à notre analyse ou qu'elles étaient redondantes. Les colonnes supprimées comprennent "Unnamed: 0", "num_acc", "an_nais", "an_naiss", "age_acc_an", "num_veh", "senc", "occutc", "permis", "secuDeux", "date", et "com".

- **Gestion des types de données et des valeurs manquantes :**

Dans un souci d'efficacité, nous avons converti certaines colonnes en types de données appropriés.
Par exemple, la colonne "dep" a été transformée de type "object" à "int64", en remplaçant les valeurs '2A' par 201 et '2B' par 202. De plus, nous avons converti la colonne "place" en type "object".
De plus, nous avons éliminé les lignes contenant des valeurs manquantes pour optimiser la qualité des données.

- **Vérification finale des types de données :**

Une vérification finale a été effectuée pour garantir la cohérence des types de données après les transformations. Le jeu de données final se compose désormais de 34 colonnes, avec des types de données variés tels que "float64", "int64", et "object".

A l'issue de l'étape de pré-processing rigoureux, les variables que nous avons choisi de conserver dans notre ensemble de données final sont au nombre de 34.

Celles-ci sont réparties comme suit :

Type de variables	Désignation des variables
Explicatives	place, catu, sexe, trajet, locp, actp, etatp, secuUn, tranches_ages, catr, circ, vosp, prof, plan, surf, infra, situ, obs, obsm, choc, manv, catv_Label, lum, agg, int, atm, col, jour_de_la_semaine
Numériques	nbv, dep, heure, month, day
Cible	grav

Nous considérons que ces colonnes représentent de manière pertinente les caractéristiques essentielles pour notre analyse.

3.1.2 Pré-processing des variables

Nous avons effectué un pré-processing des données afin de les préparer pour la modélisation.
Nous avons sélectionné des méthodes de pré-processing pour chaque type de variable en fonction de leurs caractéristiques et des modèles que nous souhaitions tester.

Cette approche nous a permis d'optimiser la qualité des données pour chaque modèle testé (cf. 3.2.1 Méthodologie).

Voici un aperçu des méthodes de pré-processing que nous avons appliquées, adaptées en fonction des modèles :

- **Encodage des variables catégorielles :**

- One Hot Encoder : Appliqué de manière générale pour traiter les variables catégorielles comme des catégories distinctes sans introduire d'ordre ou de relation.

- Ordinal Encoder (Label Encoder) : Utilisé spécifiquement pour les variables catégorielles ordinales afin de conserver l'ordre inhérent des catégories.

- **Encodage de la variable cible :**

- Label Encoder : Appliqué à la variable cible pour exprimer la variable cible sous forme numérique, tout en conservant l'ordre, facilitant son utilisation dans les modèles.

- **Encodages spécifiques :**

- StandardScaler : Employé pour mettre à l'échelle les variables numériques, garantissant une comparaison équitable, particulièrement pertinent pour des modèles sensibles à l'échelle des données, tels que la régression logistique ou les méthodes basées sur la distance (k-NN).

- Binary Encoding : Utilisé pour réduire la dimensionnalité de certaines variables catégorielles.

- Frequency Encoding : Choisi pour certaines variables afin de conserver l'information sur la fréquence des catégories.

Ces actions ont été orchestrées de manière à assurer une représentation adéquate des données tout en respectant les exigences spécifiques de chaque modèle.
Par exemple, l'encodage One-Hot a été préféré pour les variables catégorielles en raison de sa simplicité et de son efficacité, tandis que l'encodage ordinal a été privilégié lorsque l'ordre des catégories avait une signification particulière.

3.2 Choix des algorithmes

3.2.1 Méthodologie

Dans la phase initiale de notre modélisation, nous avons opté pour une approche comparative, en commençant par des modèles relativement simples pour établir une base de référence. Nous avons démarré avec des modèles basiques, dont les arbres de décision et la régression logistique.

Après avoir établi une performance de référence avec les modèles de base, nous avons progressivement intégré des modèles plus sophistiqués comme Random Forest et Gradient Boosting (ou plutôt sa variante Histogram-based Gradient Boosting, plus performante sur les jeux de données volumineux). Ces modèles sont reconnus pour leur capacité à capturer des relations non linéaires complexes.

Pour chaque modèle, nous avons examiné une série de métriques pour évaluer leur performance (décrites en détail dans la section 3.2.2). L'ajustement des hyperparamètres, l'utilisation de techniques de rééchantillonnage, et l'évaluation des modèles sur un ensemble de validation ont été des composantes clés de notre méthodologie.

3.2.2 Description des métriques d'évaluation

Pour évaluer de manière complète et précise les performances des modèles dans notre étude de classification multi-classes déséquilibrée, nous avons choisi un ensemble de métriques d'évaluation qui capturent différentes facettes de la performance du modèle. Voici un aperçu des métriques utilisées :

- **F1 Score** : Le F1 score est une métrique qui combine la précision et le rappel en une seule mesure. C'est particulièrement utile dans les situations où nous avons des classes déséquilibrées, car il n'est pas biaisé vers la classe majoritaire.
- **Balanced Accuracy** : il s'agit de la moyenne des taux de vrai positifs (sensibilité) et des taux de vrai négatifs (spécificité) pour chaque classe. Cela nous permet d'obtenir une évaluation plus juste de la performance du modèle, surtout lorsque les classes sont inégalement représentées dans le jeu de données.
- **Geometric Mean Score :** c'est une mesure qui équilibre la performance sur les classes en calculant la racine carrée du produit du rappel de toutes les classes. Cela assure que la performance du modèle n'est pas dictée par la classe majoritaire et qu'il performe de manière uniforme sur toutes les classes.
- **Matrice de Confusion** : La matrice de confusion est une représentation visuelle de la performance du modèle. Elle montre explicitement quand et comment les prédictions se trompent, ce qui nous aide à comprendre le comportement du modèle.
- **Rappel de la Classe 'Tué'** : c'est une métrique spécifique que nous avons utilisée pour évaluer la performance du modèle à identifier correctement la classe 'Tué', qui est cruciale pour notre problème. Un rappel élevé indique que le modèle est capable de détecter la majorité des instances positives pour cette classe, ce qui est essentiel étant donné les conséquences graves d'une mauvaise classification sur cette classe.

Chacune de ces métriques a été sélectionnée pour assurer une évaluation approfondie et équilibrée de nos modèles, permettant d'identifier le modèle non seulement le plus performant en termes de précision globale, mais également celui qui traite le mieux les classes individuelles, y compris celles moins représentées mais critiques pour l'étude.

3.3 Expérimentation et Résultats

3.3.1 Description des expériences effectuées

Durant la phase d'expérimentation, nous avons cherché à améliorer les performances des modèles retenus lors de la première phase de comparaison.

Nous avons testé diverses techniques de rééchantillonnage afin de pallier le déséquilibre des classes. Nous avons testé plusieurs algorithmes de sélection de variables ou de réduction de dimension. Puis finalement, nous avons cherché à optimiser les hyperparamètres.

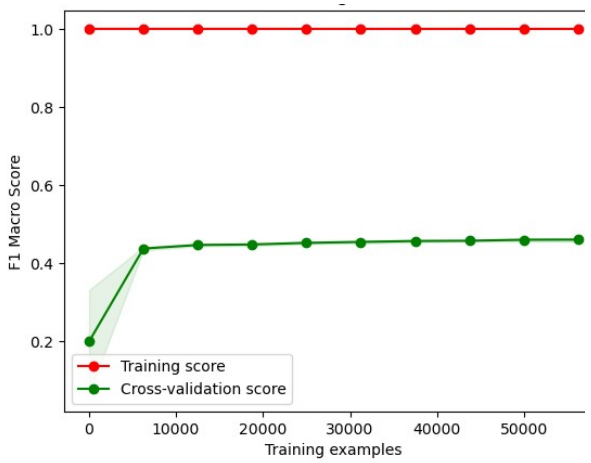
Rééchantillonnage	Feature Selection (réduction de dimension)	Hyperparamètres
RandomOverSampler	SelectKbest	RandomizedSearchCV
RandomUnderSampler	RFE	GridsearchCV
SMOTE	PCA	

Il est à noter concernant RandomizedSearchCV que nous l'avons intégré dans une boucle qui redéfinit les plages de recherche autour des meilleurs paramètres trouvés à chaque itération. L'objectif était d'identifier les plages de paramètres les plus prometteuses en évitant la recherche systématique qui aurait été trop longue avec GridSearchCV.

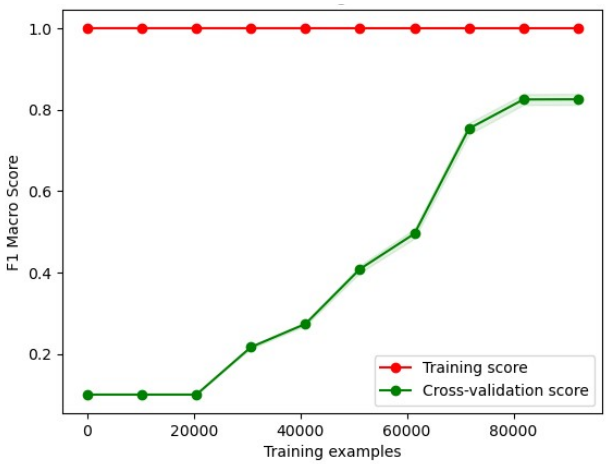
3.3.2 Présentation des résultats

a. RandomForest

L'utilisation combinée de RandomOverSampler et RandomUnderSampler a conduit à une amélioration significative des performances du modèle RandomForest.



Courbes d'apprentissage du modèle sans rééchantillonnage des classes



Courbe d'apprentissage du modèle avec l'utilisation combiné de RandomOverSampler et RandomUnderSampler

Par contre l'utilisation des algorithmes de Feature Selection n'ont pas permis d'amélioration des performances.

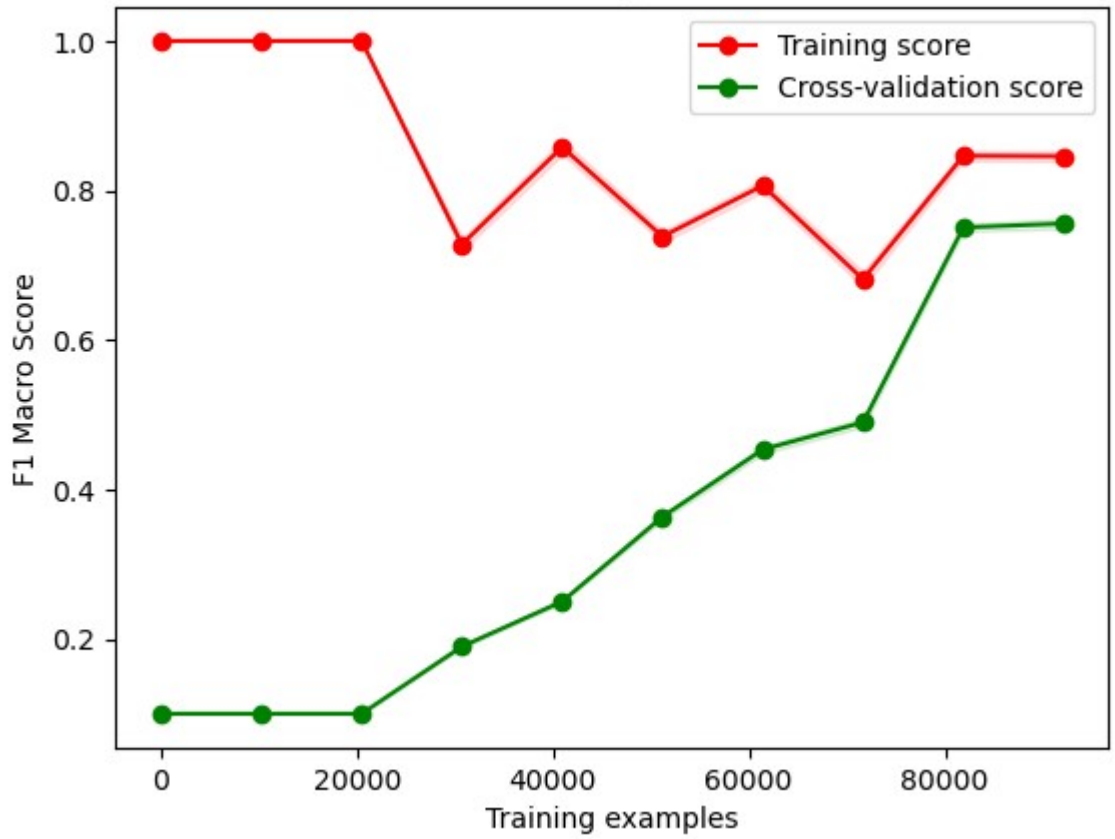
Dans une seconde étape, le principal enjeu a été de réduire le sur-apprentissage observé avec les hyper paramètres par défaut. Afin de résoudre ce problème, quatre hyper paramètres ont été modifiés afin de réduire la complexité du modèle :

- **'n_estimators' :** c'est le nombre d'arbres créé par le modèle.
- **'max_depth'** : Cet hyperparamètre contrôle jusqu'à quel niveau les arbres de la forêt peuvent croître. Si max_depth est élevé, les arbres peuvent devenir très complexes et détaillés et cela peut conduire à du surapprentissage.
- **'min_sample_split'** : c'est le nombre minimum d'échantillons qu'un nœud doit avoir avant qu'il puisse être divisé en nœuds enfants. En augmentant cette valeur, on limite la création de nouveaux nœuds. Cela conduit à des arbres moins complexes qui sont moins susceptibles de s'adapter trop précisément aux données d'entraînement, réduisant ainsi le risque de surapprentissage.
- **'min_sample_leaf'** : c'est le nombre minimum d'échantillons qu'un nœud doit avoir pour devenir le résultat final d'une division dans l'arbre. Des valeurs plus élevées pour min_samples_leaf garantissent que chaque nœud feuille a un nombre suffisant d'observations, ce qui peut empêcher l'arbre de devenir trop complexe et de s'adapter excessivement aux données d'entraînement.

Ainsi nous avons adapté les plages de recherche des meilleurs hyper paramètres pour ces trois valeurs : des valeurs plus faibles de 'n_estimators' et de 'max_depth' et des valeurs plus élevées pour 'min_sample_split' et 'min_sample_leaf'.

Hyper paramètres	valeurs par défauts	valeurs retenues
n_estimator	100	38
max_depth	pas de limite	42
min_sample_split	2	10
min_sample_leaf	1	4

Cela a permis de réduire très significativement le sur-apprentissage comme le montre la figure ci-dessous :



Courbes d'apprentissage du modèle avec des hyperparamètres adaptés pour réduire le sur-apprentissage.

Ci dessous le rapport de classification et la matrice de confusion sur l'ensemble de test :

Classe prédite	Blessé hospitalisé	Blessé léger	Indemne	Tué
Classe réelle				
Blessé hospitalisé	2258	1198	593	217
Blessé léger	1640	4019	1905	84
Indemne	642	935	6959	62
Tué	357	56	52	107

	precision	recall	f1-score	support
Blessé hospitalisé	0.46	0.53	0.49	4266
Blessé léger	0.65	0.53	0.58	7648
Indemne	0.73	0.81	0.77	8598
Tué	0.23	0.19	0.21	572
accuracy			0.63	21084
macro avg	0.52	0.51	0.51	21084
weighted avg	0.63	0.63	0.63	21084

La matrice de confusion montre des confusions significatives entre certaines classes.

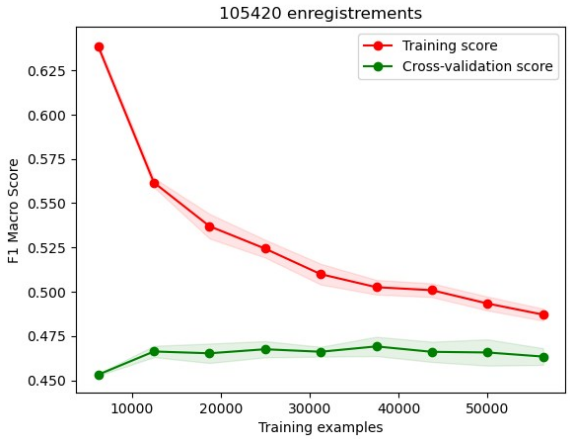
Deux exemples :

- 'Blessé léger' et 'Blessé hospitalisé' sont souvent confondues,
- La classe 'Tué' est souvent classée comme 'Blessé hospitalisé'.

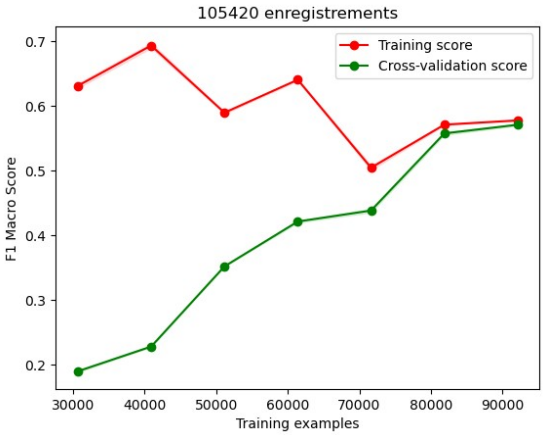
Le rapport de classification montre que le f1-score Macro n'est que de 51% sur l'ensemble de test, montrant que le modèle a du mal à généraliser sur de nouvelles données, la classe ‘’Tué’’ n'étant pas prédite correctement.

b. GradientBoostingClassifier

L'utilisation combinée de RandomOverSampler et RandomUnderSampler a la aussi conduit à une amélioration significative des performances de ce modèle comme le montre les deux figures ci-dessous:



Courbes d'apprentissage du modèle sans rééchantillonnage des classes

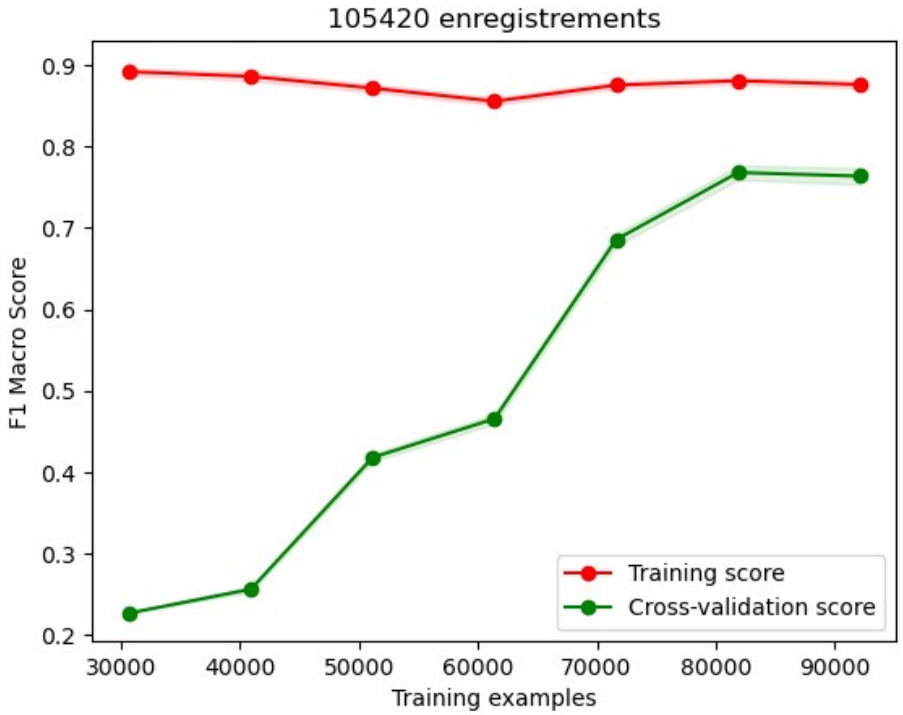


Courbe d'apprentissage du modèle avec l'utilisation combiné de RandomOverSampler et RandomUnderSampler

Une recherche des meilleurs paramètres a été effectuée :

Paramètre	**Plage de recherche**	**Valeur retenue**
n_estimators	[100, 200, 300, 400, 500]	400
min_samples_split	[2, 5, 10]	10
min_samples_leaf	[1, 2, 4]	2
max_depth	[3, 4, 5, 6, 7]	7
learning_rate	[0.01, 0.02, 0.05, 0.1]	0.1
subsample	[0.6, 0.7, 0.8, 0.9, 1.0]	0.8

Ci-dessous la courbe d'apprentissage du modèle avec les valeurs d'hyper-paramètres optimisés :



Courbes d'apprentissage du modèle avec des hyper-paramètres optimisés

Le modèle semble atteindre de bien meilleures performances sur l'ensemble de validation avec les paramètres optimisés.

Ci dessous le rapport de classification et la matrice de confusion sur l'ensemble de test :

Classe prédite	Blessé hospitalisé	Blessé léger	Indemne	Tué
Classe réelle				
Blessé hospitalisé	2178	1095	503	490
Blessé léger	1606	4188	1627	227
Indemne	639	1142	6659	158
Tué	307	57	39	169

	precision	recall	f1-score	support
Blessé hospitalisé	0.46	0.51	0.48	4266
Blessé léger	0.65	0.55	0.59	7648
Indemne	0.75	0.77	0.76	8598
Tué	0.16	0.30	0.21	572
accuracy			0.63	21084
macro avg	0.51	0.53	0.51	21084
weighted avg	0.64	0.63	0.63	21084

Les conclusions tirées de la matrice de confusion et du rapport de classification sont les mêmes que pour le modèle RandomForest.

c. Réseau de Neurones Artificiels (RNA)

Nous avons exploré l'application des techniques de Deep Learning en employant un modèle de Réseau de Neurones Artificiels (RNA) avec Keras.

Il s'agit d'un modèle DNN (Dense Neural Network) à 2 couches.

- **Architecture du modèle :**

Nous avons défini un modèle séquentiel avec deux couches Dense, utilisant des activations relu, suivies de couches Dropout pour la régularisation.

La couche de sortie utilise une activation softmax pour classer en quatre catégories.

- **Affichage de la structure :**

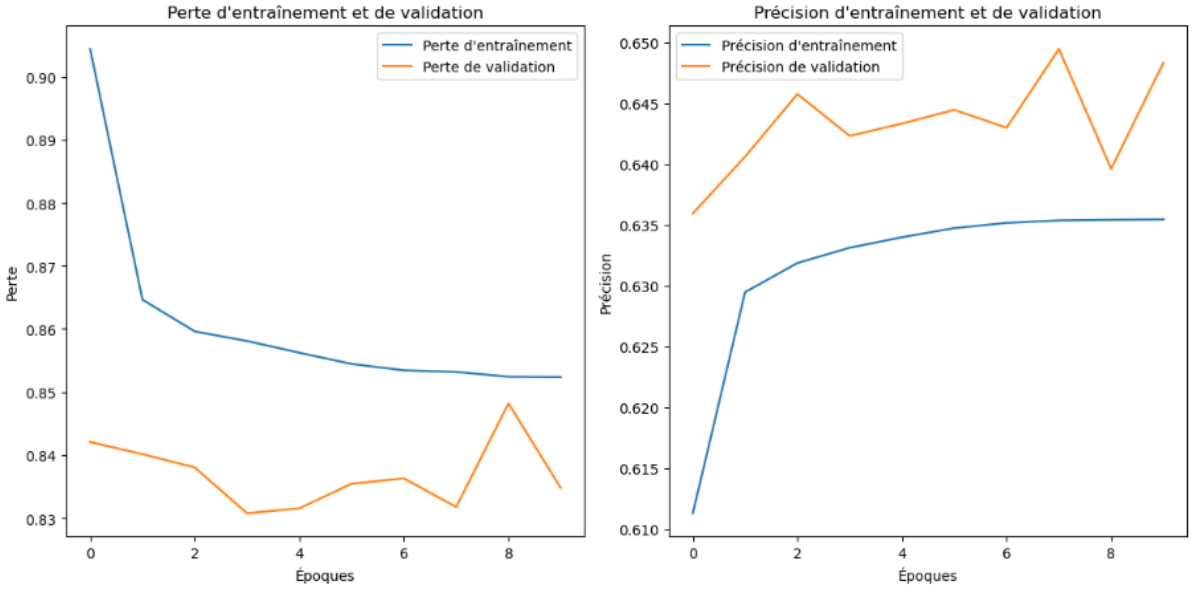
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               33024

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 dense_2 (Dense)             (None, 4)                 260

=================================================================
Total params: 41540 (162.27 KB)
Trainable params: 41540 (162.27 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

- **Entraînement et évaluation :**

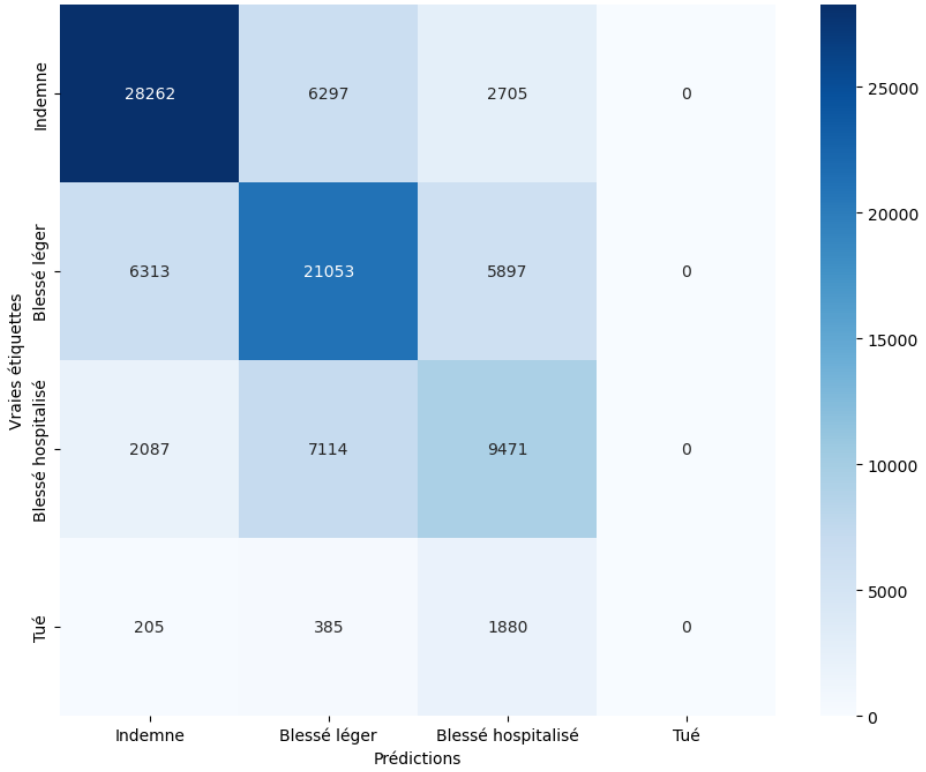
Le modèle a été entraîné sur 10 époques avec une taille de lot de 64, atteignant une précision de 64,13% sur le jeu de test.

- **Interprétation des résultats :**

L'analyse des courbes d'apprentissage suggère un début prometteur suivi d'une phase de surapprentissage.



Cependant, les performances variables, surtout pour la classe "Tué" avec un f1-score de 0%, soulignent des défis significatifs dans la généralisation du modèle.



```
Rapport de classification :
                    precision    recall  f1-score   support

           Indemne       0.77      0.76      0.76     37264
      Blessé léger       0.60      0.63      0.62     33263
Blessé hospitalisé       0.47      0.51      0.49     18672
               Tué       0.00      0.00      0.00      2470

          accuracy                           0.64     91669
         macro avg       0.46      0.47      0.47     91669
      weighted avg       0.63      0.64      0.63     91669
```

- **Recherche d'hyper paramètres :**

Malgré une tentative d'optimisation des hyperparamètres via une recherche aléatoire (RandomSearch), aucune amélioration significative des performances du modèle n'a été observée en termes de précision sur le jeu de test.

- **Pistes d'amélioration étudiées :**

Dans le but d'améliorer les performances du modèle, différentes approches ont été examinées. **Une tentative de rééquilibrage des classes** été entreprise pour corriger les disparités entre les catégories, incluant l'utilisation de poids de classe.
De plus, **une couche d'embedding a été intégrée** pour traiter les variables catégorielles, offrant ainsi une représentation plus riche et adaptable de ces données.
Des techniques telles que **l'augmentation de données**, **l'application d'un arrêt anticipé** (early stopping) pour prévenir le surapprentissage, et **l'ajout de couches de dropout** pour la régularisation du modèle ont également été examinées.

3.3.3 Récapitulatif des résultats

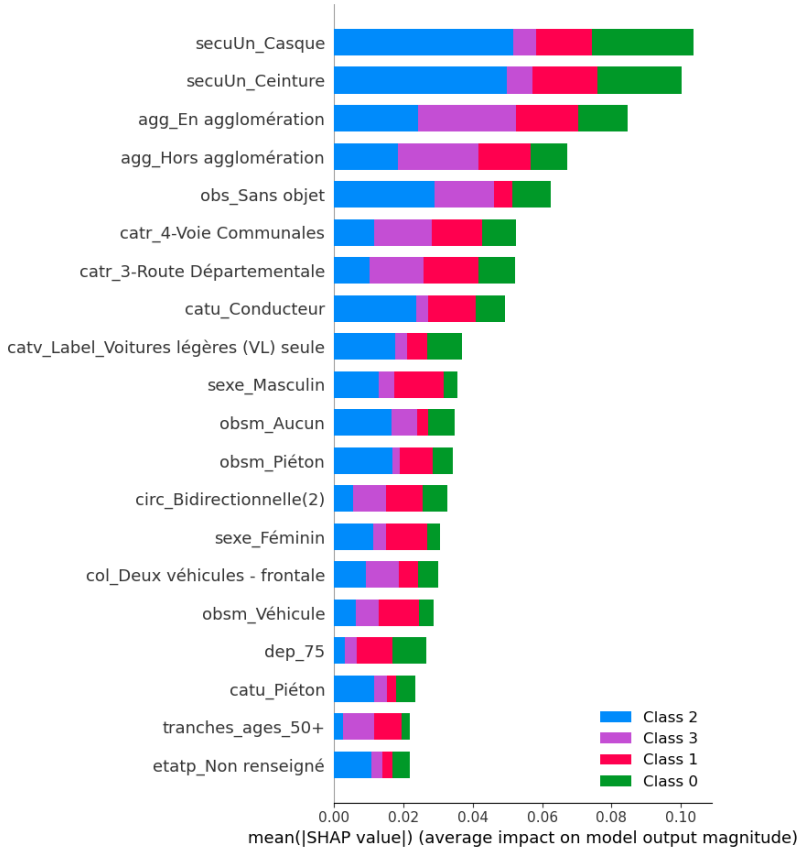
Les différents modèles développés affichent des résultats très semblables. Il est difficile de les classer selon les métriques de performances et les matrices de confusion montrent que les erreurs de classification sont elles aussi semblables. Par exemple, la classe ‘Tué’ est majoritairement prédite en classe ‘Blessé grave’ et la classe ‘Indemne’ est souvent la mieux classée.

Malgré les efforts déployés sur les différents modèles testés, la classe minoritaire est toujours mal prédite et nous ne sommes pas parvenu à déplafonner les performances des modèles sur les jeux de données test.

Nous avons décidé de tester une approche différente en regroupant des classes afin de transformer le problème en classification binaire. Cette approche sera détaillée dans la prochaine partie.

3.3.4 Interprétabilité des résultats

Le calcul de l'importance des caractéristiques a été effectué sur le modèle RandomForest à l'aide de la bibliothèque SHAP. Voici le résultat :



Classe 0 = blessé hospitalisé, **Classe 1** = blessé léger, **Classe 2** = indemne, **Classe 3** = Tué

Les équipements de sécurité et les types de route semblent avoir une grande importance dans la sévérité des blessures. Par ailleurs, il ne semble pas y avoir beaucoup de caractéristiques ayant une importance forte. Ce qui n'est pas étonnant, la phase d'analyse exploratoire nous ayant montré que les caractéristiques du jeu de données expliquent très peu la gravité.

3.4. Simplification du problème

Modèle DNN à 4 couches [CL1]

- **Architecture du modèle :**

Afin de s'affranchir de cette problématique de déséquilibre des classes, un ré-échantillonnage sur les données a été réalisé au moyen d'une fonction ‘’oversampling’’ en dupliquant aléatoirement des exemples de la classe minoritaire.

Puis, nous avons transformé notre problématique en une classification à 2 classes en modifiant la variable cible pour qu'elle prenne 2 modalités.

Une modalité non grave incluant la catégorie "Indemne" et "Blessé léger" ainsi qu'une modalité grave incluant la catégorie "Blessé hospitalisé" et "Tué".

Puis nous avons défini un modèle séquentiel avec quatre couches Dense, les trois premières implémentent une fonction d'activation ‘’tanh’’ et la quatrième, à savoir la couche de sortie, utilise une fonction d'activation ‘’softmax’’ pour classer en deux catégories.

- **Affichage de la structure :**

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Input (InputLayer)          [(None, 194)]             0
 Dense_Layer1 (Dense)        (None, 800)               156000
 Dense_Layer2 (Dense)        (None, 200)               160200
 Dense_Layer3 (Dense)        (None, 40)                8040
 Dense_Layer4 (Dense)        (None, 2)                 82
=================================================================
Total params: 324,322
Trainable params: 324,322
Non-trainable params: 0
```

- **Entraînement du modèle et évaluation :**

Pour notre problème spécifique de modélisation prédictive, nous avons opté pour l'expérimentation systématique afin de découvrir ce qui fonctionne le mieux.

Nous avons par conséquent joué sur les hyperparamètres du modèle de manière empirique (nombre de couches, nombre de neurones par couche, nombre d'epoch...) afin d'optimiser l'accuracy. Au final, une précision de 78,74 % (valeur max) a été obtenue sur les données d'entraînement, sur la base de 40 epochs et en utilisant une taille de lot de 32. La précision sur les données de validation est de 75,47 % (valeur max).

- **Interprétation des résultats :**

La matrice de confusion ci-dessous illustre que le nombre de bonnes classifications (sur la diagonale) est supérieur au nombre de mauvaises classifications.

Le taux de bonnes prédictions ('accuracy') correspond au nombre d'éléments bien classés sur le nombre total d'individus.

Classe prédite	**0**	**1**
Classe réelle		
0	208197	50489
1	49040	150622

A partir de la matrice de confusion, l'accuracy peut être calculé 'à la main', le résultat arrondi est d'ailleurs rigoureusement identique à la précision obtenue dans le rapport de classification ci-dessous.

Calcul de l'accuracy 'à la main' : (208197+150622)/(208197+150622+50489+49040) = 358819/458348 = 78.28 %.

Affichage du compte-rendu évaluatif détaillé de la performance du modèle grâce à la fonction classification_report du sous-module metrics de scikit-learn :

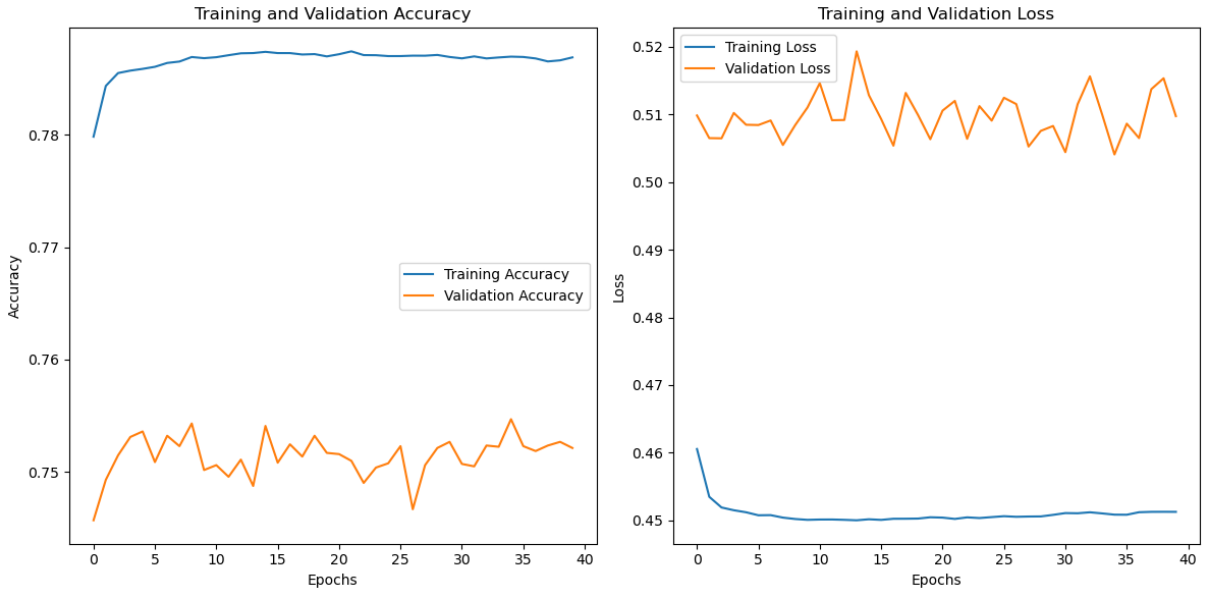
	precision	**recall**	**f1-score**	**support**
0	0.81	0.80	0.81	258686
1	0.75	0.75	0.75	199662
accuracy			**0.78**	458348
macro avg	0.78	0.78	0.78	458348
weighted avg	0.78	0.78	0.78	458348

La précision et le rappel sur la classe 0 (indemne et blessé léger) sont plutôt satisfaisants, ils sont respectivement de 81 % et 80 % : cette classe a été bien gérée par le modèle. Le F1-score, qui permet de mesurer la précision et le rappel à la fois, est de 81 %.

La précision et le rappel sur la classe 1 (blessé hospitalisé et tué) qui nous intéresse tout particulièrement sont également satisfaisants, ils sont tous les deux de 75 % : cette classe a également été bien gérée par le modèle. Le F1-score est logiquement de 75 %.

La performance de ce modèle est relativement satisfaisante et la classification binaire a permis de solutionner notre problématique de déséquilibre de classes.

De surcroît, en analysant les courbes d'apprentissage, celles-ci montrent que nous n'avons plus d'overfitting (surapprentissage) comme c'était le cas avec les modèles précédents.



La performance de ce modèle est relativement satisfaisante grâce à la classification binaire qui nous a permis de solutionner notre problématique de déséquilibre de classes.

3.5 Limitations de l'étude

Malgré notre approche méthodologique (exploration des données, sélection des variables, pré-processing des données, encodage des variables, modélisation et évaluation des modèles), notre étude présente quelques limites.

Tout d'abord, la qualité des prédictions des modèles dépend étroitement de la qualité des données d'entrée. Malgré une étape de pré-processing, la présence potentielle de biais ou de lacunes dans les données pourrait affecter les résultats.
Par exemple, des inégalités dans la déclaration des accidents par les différentes autorités compétentes, des variations dans la qualité des rapports d'accidents (dont des omissions dans les détails des circonstances de l'accident), les changements dans les pratiques de collecte de données au fil du temps peuvent créer des distorsions dans la représentation des caractéristiques des accidents, influençant ainsi les performances des modèles.

De plus, les modèles adoptés pourraient ne pas saisir de manière exhaustive toutes les situations spécifiques liées à la survenance des accidents routiers, notamment la complexité des interactions entre différentes caractéristiques de la route et la dynamique du trafic, qui ne suivent pas nécessairement des schémas prévisibles.
Par exemple, la présence d'événements imprévus tels que des travaux routiers, des changements soudains et localisés de conditions météorologiques, des pannes de véhicules ou encore, des anomalies de comportement des conducteurs, pourrait introduire des subtilités non capturées par les modèles. Cette limitation pourrait conduire à une perte de précision dans les prédictions de la gravité des accidents.

Une autre limite de notre étude découle de la nature dynamique des accidents routiers, soumise à des facteurs changeants tels que les conditions sociologiques liées à l'évolution du comportement des usagers de la route, tels que les conducteurs qui ont du mal à se séparer de leur smartphone (phénomène dit de "nomophobie") ou les piétons concentrés sur l'écran de leur smartphone (phénomène dit de "smombie").

De plus, les variations dans les politiques de sécurité routière peuvent avoir un impact significatif sur la fréquence et la gravité des accidents. Le déploiement plus large des radars automatiques et la variation du nombre de points retirés du permis de conduire sont des exemples de variation dans la répression routière, influençant ainsi le comportement des conducteurs et potentiellement impactant la fréquence et la gravité des accidents.

Les fluctuations économiques, influençant les taux de déplacement (notamment en fonction du prix du carburant pour les usagers), le nombre de véhicules sur la route, et d'autres aspects liés à la sécurité routière tels que l'entretien des véhicules et des routes, constituent également des facteurs à considérer.

La politique environnementale croissante, notamment le développement des pistes cyclables dans les grandes et moyennes agglomérations, peut également jouer un rôle important dans la dynamique des accidents.

Par ailleurs, le règlement général sur la protection des données (RGPD) limite la disponibilité de données descriptives sur l'état des usagers de la route, tels que le taux d'alcoolémie ou l'influence de stupéfiants. Ces éléments importants échappent partiellement à notre modélisation.

3.6 Suggestions pour les recherches futures

Pour surmonter les limitations évoquées et améliorer la pertinence de la prédiction de la gravité des accidents routiers, plusieurs pistes de recherche peuvent être explorées dans le futur.

- **Amélioration des données d'entrée :**

Une attention particulière devrait être accordée à l'enrichissement et à l'amélioration de la qualité de la saisie des ensembles de données. Cela implique une collecte plus exhaustive, précise et harmonisée des variables pertinentes, en mettant l'accent sur la minimisation des biais potentiels **et de lacunes au sein des données**.

L'intégration de données telles que des informations détaillées sur les conditions météorologiques en temps réel, les états de santé des conducteurs, et d'autres facteurs contextuels, pourrait renforcer la robustesse des modèles.

- **Exploration de modèles plus sophistiqués :**

L'utilisation de modèles de machine learning plus avancés, tels que les réseaux neuronaux avec des approches d'apprentissage profond. Ces modèles peuvent certainement offrir une meilleure capacité à capturer des relations complexes entre les variables, permettant ainsi une analyse plus approfondie et précise de la gravité des accidents.

- **Suivi temporel et analyse des tendances :**

La réalisation d'études longitudinales permettrait de tenir compte des variations temporelles dans les politiques de sécurité routière, les comportements des usagers de la route, et d'autres facteurs dynamiques. Cela contribuerait à une compréhension plus approfondie des variations de la gravité des accidents au fil du temps et permettrait d'ajuster les modèles en conséquence.

- **Intégration de facteurs contextuels :**

Une exploration plus poussée des facteurs contextuels, tels que les aspects sociologiques, politiques, économiques, environnementaux, et les évolutions technologiques (notamment des véhicules), pourrait enrichir la modélisation.
Cela implique de considérer de manière plus explicite l'influence des politiques de sécurité routière, des changements économiques, des développements urbains, et d'autres facteurs externes.

- **Évaluation continue des politiques de sécurité routière :**

L'évaluation constante des politiques de sécurité routière, y compris des variations dans la répression routière, peut être cruciale. Cela permettrait d'ajuster les modèles en fonction des changements dans les règlements, les amendes, ou d'autres mesures qui pourraient influencer le comportement des conducteurs.

- **Renforcement de la collaboration interdisciplinaire :**

La collaboration avec des experts dans des domaines connexes, tels que la sociologie, l'économie, la santé publique, et l'urbanisme, peut fournir des perspectives complémentaires. Une approche interdisciplinaire pourrait aider à mieux comprendre et intégrer les multiples facettes des accidents routiers.

Une approche holistique, combinant l'ensemble de ces pistes d'amélioration pourrait conduire à des résultats plus précis et généralisables.

4. Conclusions

4.1 Résultats finaux au regard des objectifs fixés

Quelles conclusions pouvons-nous tirer des prédictions de nos modèles de classification testés, en configuration multi-classes et binaire ?

Tout d'abord, nous avons pu constater que le contexte lié à l'accident ainsi que certaines caractéristiques liées à l'usager et au véhicule étaient déterminants dans la potentialité de gravité de l'accident. En effet, nos ''DataViz'' ont montré que le niveau de gravité était notamment déterminé à partir d'éléments contextuels qui peuvent diminuer ou, inversement, aggraver le risque, comme par exemple, une conduite en dehors des agglomérations, sur route départementale ou nationale, à double-sens, en soirée, sans éclairage, etc …

Nous avons souhaité prédire les accidents de la route graves, sur la base de modèle(s) qui soit(ent) capable(s) de les détecter avec une bonne fiabilité, car ils ont des conséquences plus graves que les accidents légers.

Ainsi, nous nous sommes focalisés sur les classes minoritaires « blessés hospitalisés » et « tués » que nous avons souhaité pouvoir identifier avec un niveau de précision satisfaisant.

Or, il s'est avéré que les valeurs relatives aux classes « indemnes » et « blessés légers » étaient sur-représentées par rapport à celles que nous avons ciblées, ce qui nous a posé des difficultés pour entraîner correctement nos modèles de Machine Learning.

Nous avons effectivement expérimenté différents modèles d'apprentissage supervisé et, pour chaque modèle, nous avons sélectionné les métriques qui nous apparaissaient les plus pertinentes afin d'évaluer leur performance respective et de pouvoir les comparer.

Ce fort déséquilibre entre les classes a été mis en évidence de manière flagrante grâce à la matrice de confusion et à une série de métriques comme la précision, le rappel, le f1-score et la moyenne géométrique (G-mean).

Concernant le choix des métriques, nous avons pu nous rendre compte à postériori que l'accuracy n'est pas toujours une métrique pertinente pour évaluer des modèles sur un jeu de données déséquilibré, car elle ne tient pas compte de la distribution des classes.

Aussi, cette problématique de classification multi-classes avec des données déséquilibrées nous a permis d'expérimenter **l'accuracy paradoxe**.

Les conséquences de ce déséquilibre ont été multiples :

Les modèles entraînés ont été plus ou moins biaisés en faveur des classes majoritaires, c'est-à-dire qu'ils ont eu tendance à prédire souvent ces classes, même quand ce n'était pas le cas. Cela nous a parfois conduit à des taux de prédictions qui ne reflétaient pas la réalité.

Parfois même, les classes minoritaires ont été traitées comme des valeurs aberrantes de la classe majoritaire, ce qui a été notamment le cas avec l'algorithme Régression Logistique. Celui-ci a généré des classifieurs triviaux et a classé chaque observation dans l'une des classes majoritaires.

En synthèse, notre jeu de données d'entraînement contenait trop peu d'exemples d'accidents graves ou mortels, et par conséquent, les modèles entraînés sur les jeux de test ont eu du mal à les reconnaître quand ils les ont rencontrés.

Par ailleurs, nous avons également pu mettre en évidence le fait que bon nombre de modèles avaient un pouvoir de généralisation faible, c'est-à-dire qu'ils n'étaient pas capables de s'adapter à de nouvelles données qui ne suivaient pas la même distribution que le jeu de données d'entraînement.

Dans le cadre de nos démarches d'optimisation, l'ajustement des hyperparamètres, l'utilisation de techniques de rééchantillonnage et l'évaluation des modèles sur la base de la ‘’cross validation’’ ont été des composantes clés de notre méthodologie.

Au final, deux modèles se sont révélés plus performants que les autres, à savoir « Random Forest » et « Histogram-based Gradient Boosting ».
Ces modèles étant reconnus pour leur capacité à capturer des relations non linéaires complexes sur des jeux de données déséquilibrés et relativement volumineux.

Nous avons pu nous apercevoir que les méthodes de machine learning classiques n'étaient pas bien adaptées pour la classification multi-classes sur des données déséquilibrées.

Fort de ce constat, nous avons transformé notre problématique en une classification à 2 classes en modifiant la variable cible pour qu'elle prenne deux modalités :

- une modalité non grave incluant la catégorie "Indemne" et "Blessé léger",
- une modalité grave incluant la catégorie "Blessé hospitalisé" et "Tué".

Puis, nous avons expérimenté un modèle DNN à 4 couches et il s'est avéré que la performance de ce modèle a été relativement satisfaisante. La classification binaire a permis de solutionner notre problématique de déséquilibre de classes.

4.2 Conclusion sur les utilités de notre projet.

Quels pourraient être les usages et applications avec notre modèle de prédiction optimisé ?

Avant d'évoquer les limites que nous avons rencontrées avec les données à notre disposition, nous nous sommes interrogés sur les usages et les applications pour lesquels notre modèle de prédiction « Random Forest » pourrait être utile.

En d'autres termes, nous avons humblement essayé de voir comment notre modèle pouvait être mis à profit dans le cadre d'une utilisation par les services publics ou des compagnies d'assurances.

Nous avons tout d'abord pensé à des applications potentielles dans les domaines de l'assistance et de la prévention.

***Dans le domaine de la prévention, on pourrait imaginer lancer des campagnes de prévention ou de sensibilisation** : - On constate que les populations les plus âgées (+ de 66 ans) représentent une population plus risquée avec un taux d'accidents mortels plus élevé.

A ces conducteurs séniors identifiés par le score comme plus risqués, et qui vont rouler le plus souvent sur des routes départementales, en dehors des agglomérations et dans certaines régions où la météo est généralement défavorable, etc…, il pourrait sans doute être pertinent de leur proposer un suivi à travers des bilans de santé gratuits afin

de vérifier leur bonne vision et leurs capacités de réflexe pour faire face, en cas de situation de danger.

****Par ailleurs, dans le domaine de l'assistance, on pourrait imaginer un générateur d'itinéraires** qui va prendre en compte les localisations des accidents corporels de la route que nous avons exploité dans le cadre de notre modélisation prédictive des accidents mortels.

De cette façon, en exploitant les données à disposition sur le site data.gouv.fr, notre algorithme va repérer les zones accidentogènes et proposer un trajet qui sera moins risqué a priori. Dans ce sens, un partenariat pourrait être mis en place entre un assureur et un éditeur d'applications GPS.

Ainsi, des notifications d'alerte pourraient, par exemple, être proposées au moment de la recherche d'itinéraires avec le GPS dans le cas où les trajets proposés présentent des situations risquées, à savoir des routes départementales et/ou nationales, en dehors d'agglomérations, avec des conditions d'éclairage faible, …, qui plus est, dans des zones où plusieurs accidents corporels ont d'ores et déjà été détectés par le passé.

Dans ce cas, des itinéraires alternatifs moins accidentogènes pourraient être proposés tout en respectant une limite de durée supplémentaire de trajet.

*****Toujours dans le domaine de l'assistance,** on pourrait envisager d'appliquer notre modèle de prédiction de la gravité des accidents de la route pour **estimer le degré d'intervention à apporter et prioriser ainsi certaines interventions comme celles des pompiers.**

En effet, à l'occasion d'un accident corporel, en fonction du contexte dans lequel s'est déroulé l'accident, les équipes d'assistance pourraient déployer des moyens d'intervention adaptés selon le degré de gravité identifié à travers le score mis en place (quitte à neutraliser certaines données du score dans le cas où on n'est pas en mesure de les déterminer) associé à d'autres règles additionnelles, et ainsi agir plus efficacement en cas d'urgence vitale détectée avec un renforcement des équipes de secours.

De la même façon, en cas d'accidents simultanés sur une même période à proximité, ce score pourrait permettre de déterminer un degré de priorité en complément d'autres règles additionnelles dans le cas où les équipes de secours doivent intervenir au plus vite.

******Enfin, nous pourrions également entrevoir des perspectives dans le cadre des voitures autonomes.**

Au-delà de considérations au niveau de la prévention et de l'assistance des assurés, à partir des taux d'accidents mortels évoqués précédemment pour les différentes

modalités de certaines variables, dans le cadre de la mise en circulation des voitures autonomes, il pourrait être utile de tenir compte des statistiques descriptives décrites dans les algorithmes de ces voitures.

Par exemple, en cas d'accident inévitable, **on pourrait s'appuyer sur les statistiques dans la prédiction de la gravité des accidents corporels de la route où un choix d'obstacle est à effectuer** (par exemple : percuter un arbre ou un panneau de signalisation, a priori le panneau de signalisation est moins dangereux).

Les véhicules devront tout d'abord être en mesure de classifier les différents obstacles fixes ou mobiles à travers notre algorithme de machine learning puis, pourraient se servir des probabilités mentionnées pour réaliser le choix après avoir identifié chaque type de classe d'obstacles.

Les limites par rapport aux données disponibles liées aux accidents corporels.

Il est important nous semble-t-il de souligner que notre modèle a été entraîné exclusivement sur l'environnement fermé des accidents corporels. Il s'agit de fait d'un modèle de prédiction de la gravité des accidents corporels de la route.

Cependant, les performances de notre modèle auraient sans doute pu encore être améliorées car ce jeu de données ne tenait pas compte d'éléments qui peuvent nous sembler essentiels dans la prédiction des accidents mortels.

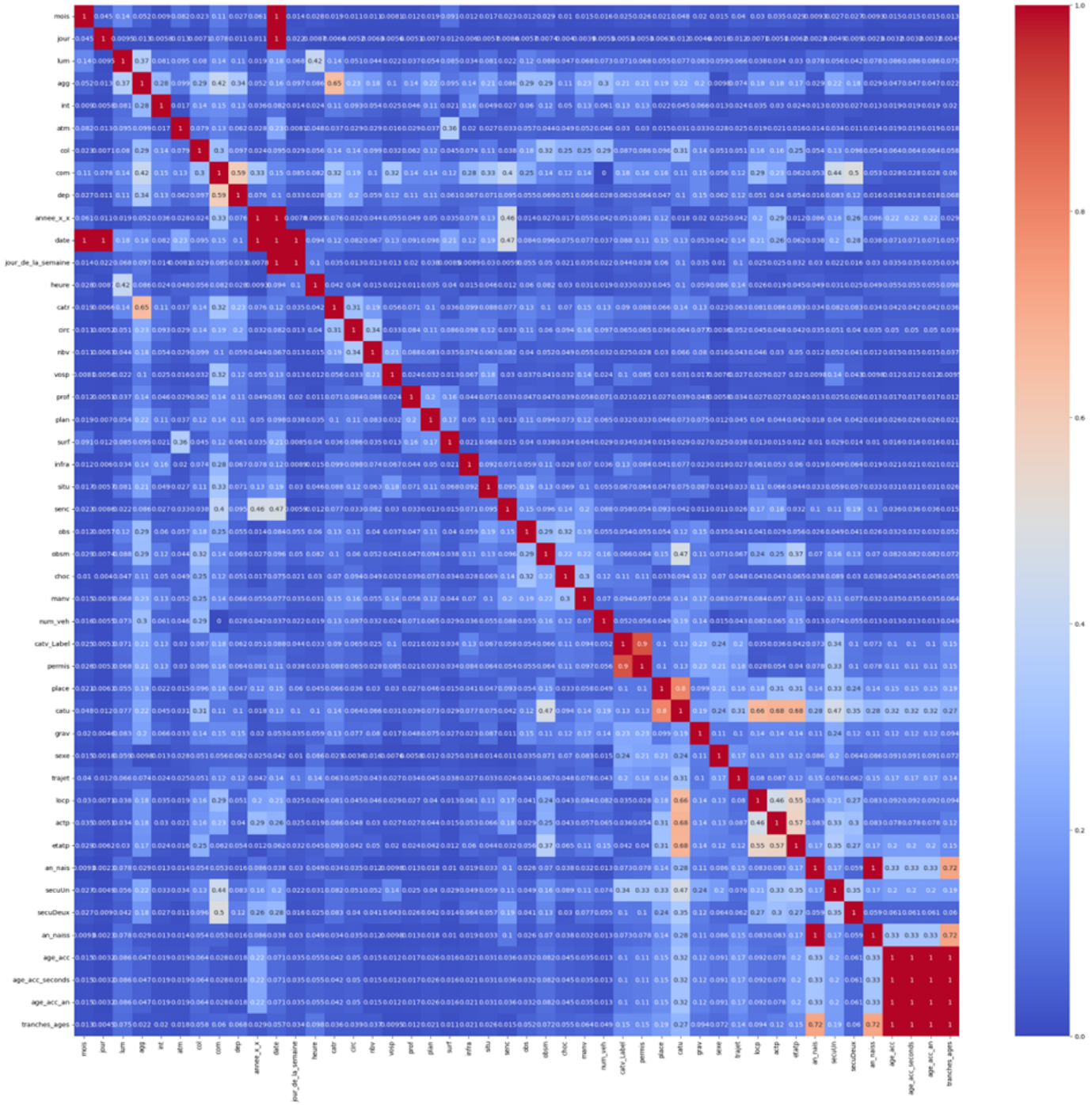
En effet, nous n'avons pas tenu compte tout d'abord de la prise en compte de facteurs de risque humain qui ont été occultés dans les bases de données mises à notre disposition par la sécurité routière (ONISR). Parmi les principaux facteurs figure la vitesse.

Dans le rapport du bilan de la sécurité routière 2017, **la vitesse excessive** ou inadaptée est la première cause de mortalité sur les routes de France puisqu'elle apparaît dans un accident mortel sur trois. Viennent ensuite **l'abus d'alcool au volant et la conduite sous l'emprise de stupéfiants** en deuxième cause des accidents mortels avec respectivement 20 et 23%.

Le **refus de priorité et l'inattention** interviennent également respectivement dans 12% et 9% des accidents mortels.

Les bases de données à disposition du grand public occultent ce type de données spécifiques relatives aux usagers et aux véhicules et à leur comportement dans la mesure où la divulgation de ces données porterait atteinte à la protection de la vie privée des personnes physiques aisément identifiables ou ferait apparaître le comportement de telles personnes alors que la divulgation de ce comportement pourrait leur porter préjudice (avis de la CADA – 2 janvier 2012).

Annexe 1 : Test Chi²



Annexe 2 : Test ANOVA

ANOVA result for the categorical variable 'lum':
F-statistic = 5.915649939400401
p-value = 0.004607122463436838

ANOVA result for the categorical variable 'agg':
F-statistic = 1.2890383353519088
p-value = 0.2995343615910299

ANOVA result for the categorical variable 'int':
F-statistic = 8.258408179813243
p-value = 1.3376089918331384e-05

ANOVA result for the categorical variable 'atm':
F-statistic = 7.83520118460214
p-value = 7.694476591819332e-06

ANOVA result for the categorical variable 'col':
F-statistic = 3.4738537871413717
p-value = 0.010321773795376358

ANOVA result for the categorical variable 'catr':
F-statistic = 5.228051323939453
p-value = 0.0010084100383082441

ANOVA result for the categorical variable 'circ':
F-statistic = 6.397805370627458
p-value = 0.003277040335518708

ANOVA result for the categorical variable 'vosp':
F-statistic = 8.272959178115414
p-value = 0.0009880549157575429

ANOVA result for the categorical variable 'prof':
F-statistic = 6.560039027236189
p-value = 0.0029316643813264015

ANOVA result for the categorical variable 'plan':
F-statistic = 6.399622930307336
p-value = 0.0032729251309046655

ANOVA result for the categorical variable 'surf':
F-statistic = 7.833291160533194
p-value = 7.712457730765182e-06

ANOVA result for the categorical variable 'infra':
F-statistic = 8.24016940699886
p-value = 1.6449694214668471e-06

ANOVA result for the categorical variable 'situ':

F-statistic = 7.3795911247344845
p-value = 3.524204829563489e-05

ANOVA result for the categorical variable 'obs':
F-statistic = 6.852162737114381
p-value = 1.035158799298044e-08

ANOVA result for the categorical variable 'obsm':
F-statistic = 6.009504221329476
p-value = 0.0004012775440742489

ANOVA result for the categorical variable 'choc':
F-statistic = 5.125531531485514
p-value = 0.0001649193669987099

ANOVA result for the categorical variable 'manv':
F-statistic = 7.626589384387409
p-value = 2.83576585360372e-13

ANOVA result for the categorical variable 'catv_Label':
F-statistic = 5.620969841011969
p-value = 6.517130846716092e-08

ANOVA result for the categorical variable 'permis':
F-statistic = 5.547610244823238
p-value = 6.384057997751008e-07

ANOVA result for the categorical variable 'catu':
F-statistic = 4.479521582167622
p-value = 0.0446495327313231

ANOVA result for the categorical variable 'sexe':
F-statistic = 1.7338882098705672
p-value = 0.23597204745697542

ANOVA result for the categorical variable 'trajet':
F-statistic = 4.347322552572357
p-value = 0.005272908635831232

ANOVA result for the categorical variable 'locp':
F-statistic = 7.5130272042557555
p-value = 4.356110657740322e-06

ANOVA result for the categorical variable 'actp':
F-statistic = 7.070246259693944
p-value = 1.0130205295853895e-09

ANOVA result for the categorical variable 'etatp':
F-statistic = 7.098193831092538
p-value = 0.00534126850862662

ANOVA result for the categorical variable 'an_nais':

F-statistic = 3.230673931086638
p-value = 5.871791515094458e-16

ANOVA result for the categorical variable 'secuUn':
F-statistic = 4.457491631945983
p-value = 0.0009001217075645784

ANOVA result for the categorical variable 'secuDeux':
F-statistic = 7.118338549924037
p-value = 7.566966551075305e-06

ANOVA result for the categorical variable 'an_naiss':
F-statistic = 3.230673931086638
p-value = 5.871791515094458e-16

ANOVA result for the categorical variable 'tranches_ages':
F-statistic = 1.3764810688337263
p-value = 0.2793523133323504