RAPPORT PROJET ACCIDENTOLOGIE

DATASCIENTEST FEVRIER 2024

THINEY Rémi

Avant-propos

Titulaire d'un Master en Méthodes Informatiques Appliquées à la Gestion des Entreprises, j'exerce actuellement en tant qu'enseignant de la conduite et de la sécurité routière. Fort de cette double expertise, j'ai souhaité approfondir mes compétences en science des données en réalisant un projet d'application du machine learning à la sécurité routière.

Ce projet s'inscrit dans le cadre de la formation Data Scientist de la promotion Février 2024 de DataScientest. Initialement engagé dans un travail de groupe, j'ai choisi de poursuivre cette étude de manière autonome afin d'explorer pleinement les enjeux liés à la prédiction de la gravité des accidents de la route à partir de données historiques.

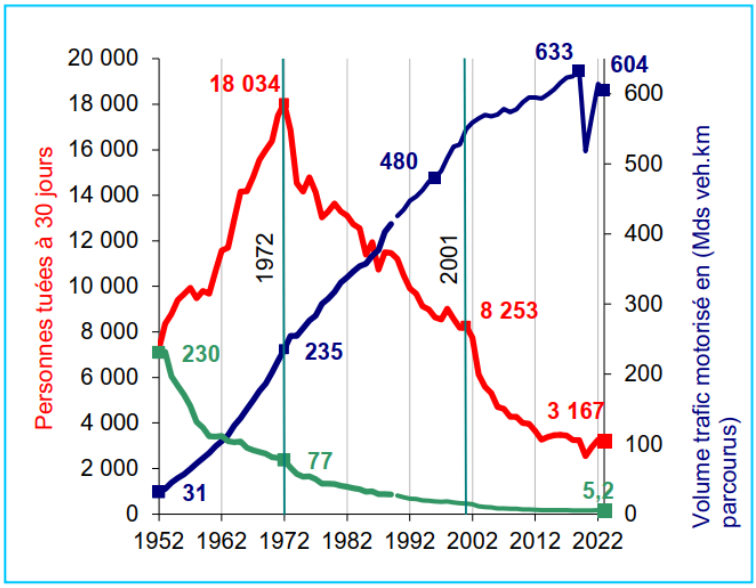
En complément du présent rendu écrit, ce projet fera l'objet d'une soutenance orale visant à présenter les méthodologies employées, les résultats obtenus et les perspectives d'amélioration du modèle développé.

Importance Métier

Historique de l'accidentologie routière en France

L'accidentologie routière en France constitue un champ d'étude crucial, mobilisant des disciplines variées telles que l'épidémiologie, la sociologie des transports et l'ingénierie de la sécurité. Dès le début du XXe siècle, la quantification des accidents révèle une croissance exponentielle du nombre de victimes. En 1924, 2 246 décès sont recensés, un chiffre qui s'élève à 5 608 en 1930, avant d'atteindre un sommet provisoire de 6 221 en 1933, pour redescendre à 5 650 en 1937. Cette tendance, intrinsèquement liée à l'essor de l'automobile et à l'absence de régulation efficace, reflète les risques accrus du développement de la mobilité motorisée.

Évolution comparée de la mortalité et de la circulation routière entre 1952 et 2023*



Les données de trafic fournies par le SDES ont été rebasées en 2024 pour les années allant de 1990 à 2023
*2023 : données du trafic provisoires

L'institutionnalisation des premières bases de données accidentologiques s'amorce en 1938 avec la mise en place du Bulletin d'Analyse des Accidents Corporels (BAAC), dispositif qui introduit une approche systématique de la collecte de données. En 1951, cette structuration s'affine avec la publication du premier bilan détaillé de l'accidentalité routière en France, amorçant une rationalisation des politiques publiques en matière de prévention.

L'année 1972 constitue un point d'inflexion critique : avec 16 545 décès recensés à six jours et environ 18 000 à trente jours, la mortalité routière atteint son paroxysme. Ce phénomène alarmant suscite une mobilisation sociale et politique inédite, notamment incarnée par l'opération "Mazamet ville morte", un signal fort de la société civile dénonçant l'inaction des pouvoirs publics face à ce fléau.

En réaction, une série de réformes structurelles est mise en œuvre pour enrayer cette tendance. Parmi les mesures adoptées figurent la limitation de la vitesse maximale autorisée (VMA) hors agglomération, l'obligation du port de la ceinture de sécurité et du

casque pour les motocyclistes, ainsi qu'une intensification des contrôles routiers. Parallèlement, l'innovation technologique devient un levier fondamental de la sécurisation des déplacements, avec l'intégration progressive d'équipements tels que les ceintures de sécurité à trois points, les freins à disque et les systèmes d'absorption des chocs.

L'impact de ces politiques et avancées technologiques sur la mortalité routière est manifeste. En 2001, le nombre de décès sur les routes françaises s'établissait à 8 253, contre seulement 3 167 en 2023, soit une baisse de plus de 80 % depuis le pic de 1972. Cette réduction spectaculaire illustre l'efficacité des stratégies de prévention mises en œuvre au fil des décennies. Cependant, la dynamique actuelle souligne la nécessité d'une approche systémique, tenant compte des nouveaux enjeux liés aux comportements des usagers, à la mutation des infrastructures et aux innovations de la mobilité (véhicules autonomes, transition énergétique, etc.). L'étude de ces transformations demeure une priorité pour anticiper les risques émergents et adapter les stratégies de prévention.

Impacts potentiels

L'insécurité routière demeure un enjeu majeur dans nos sociétés modernes. Malgré les avancées en matière de réglementation et de sécurité automobile, les accidents de la route restent une cause importante de mortalité et de blessures graves. La mobilité est un élément fondamental du quotidien, englobant les déplacements individuels, professionnels, logistiques et de transports publics. Pour améliorer la sécurité routière, l'exploitation des données historiques à l'aide de technologies avancées telles que le machine learning et le deep learning s'avère être une solution prometteuse.

La mobilité routière regroupe plusieurs catégories qui doivent être prises en compte dans l'analyse des risques d'accidents. La mobilité individuelle englobe les trajets effectués en voiture, en moto, à vélo ou en trottinette. La mobilité professionnelle concerne les déplacements liés aux activités économiques, notamment ceux des chauffeurs routiers, des taxis et des VTC. La mobilité collective comprend les transports en commun tels que les bus, les tramways et les autocars. Enfin, la mobilité logistique englobe la livraison de marchandises et le fret routier.

L'intégration du machine learning dans l'analyse des accidents permet de traiter et d'interpréter d'énormes volumes de données afin d'identifier les tendances et d'anticiper les situations à risque. Grâce aux modèles prédictifs, il devient possible de mieux comprendre les facteurs de gravité d'un accident et d'agir en amont.

L’application du machine learning à l’analyse de l’accidentologie routière engendre des avancées notables en matière de modélisation des risques et d’optimisation des politiques publiques. L’exploitation des données historiques permet d’affiner la compréhension des facteurs contributifs aux accidents et de hiérarchiser les zones les plus dangereuses. En conséquence, les pouvoirs publics peuvent élaborer des stratégies fondées sur une approche quantitative rigoureuse pour l’aménagement du territoire, la régulation du trafic et l’amélioration des infrastructures routières.

L’impact sur la sensibilisation des usagers est également significatif. Grâce aux modèles prédictifs, il devient possible d’anticiper les comportements à risque et d’adapter les campagnes de prévention en fonction des profils de conducteurs et des conditions de circulation. Par ailleurs, l’implémentation d’alertes intelligentes en temps réel, exploitant des flux de données dynamiques, constitue une opportunité pour minimiser les comportements dangereux et améliorer la prise de décision des automobilistes.

L’optimisation des services d’urgence représente un autre axe fondamental. L’analyse des données spatio-temporelles des accidents permet de rationaliser la répartition des ressources et d’améliorer la réactivité des secours. La prédiction des situations critiques, fondée sur des modèles d’apprentissage profond, contribue à la mise en place de protocoles d’intervention plus efficaces et à la réduction des délais d’intervention.

Enfin, le secteur privé tire également parti de ces avancées. Les compagnies d’assurances affinent leurs modèles actuariels en intégrant des variables prédictives issues de l’analyse des données d’accidents, ce qui permet d’ajuster les politiques tarifaires et d’inciter à des comportements plus prudents. Les auto-écoles peuvent personnaliser les formations en fonction des zones à risque identifiées, et les entreprises possédant des flottes de véhicules peuvent développer des stratégies de prévention ciblées pour réduire leur sinistralité. Ainsi, l’intégration du machine learning à l’accidentologie routière s’impose comme un levier décisif pour renforcer la sécurité et la résilience des systèmes de transport modernes.

L’application du machine learning à l’accidentologie routière représente une avancée majeure pour la prévention des accidents et l’amélioration de la sécurité routière. En exploitant les données historiques de manière intelligente, les pouvoirs publics, les entreprises et les usagers peuvent bénéficier d’une meilleure anticipation des risques et d’une réduction significative du nombre d’accidents. Cette approche constitue

une nouvelle étape vers une mobilité plus sûre et plus efficiente.

Travaux Existants

Différentes études ont exploré l'application des réseaux de neurones profonds à l'analyse des risques routiers. Le travail de Muhammad Monjurul Karim, Yu Li, Ruwen Qin et Zhaozheng Yin, intitulé A system of vision sensor based deep neural networks for complex driving scene analysis in support of crash risk assessment and prevention, met en avant l'utilisation de capteurs visuels et de réseaux de neurones pour l'analyse des scènes de conduite et la prévention des risques d'accidents.

Dans Deep Learning Serves Traffic Safety Analysis: A Forward-looking Review, Abolfazl Razi et ses collègues passent en revue les différentes approches d'apprentissage profond dédiées à l'analyse de la sécurité routière. Leur étude met en lumière l'intérêt des modèles prédictifs pour améliorer la gestion des infrastructures et la prévention des incidents.

Une autre contribution notable est celle de Noushin Behboudi, Sobhan Moosavi et Rajiv Ramnath, avec leur étude Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques, qui propose une synthèse des techniques de machine learning appliquées à l'analyse et à la prédiction des accidents de la route. Ils mettent en avant des méthodes supervisées et non supervisées pour traiter efficacement les grandes bases de données accidentologiques.

Le travail de Victor Adewopo et Nelly Elsayed, Smart City Transportation: Deep Learning Ensemble Approach for Traffic Accident Detection, s'inscrit dans une perspective de mobilité intelligente en appliquant des modèles d'apprentissage profond pour la détection en temps réel des accidents et l'optimisation de la gestion du trafic.

La plateforme AVATAR du Cerema (Analyse et Visualisation Automatique de données de Trafic Routier) illustre une initiative française visant à exploiter les données de circulation pour analyser automatiquement les risques et identifier des tendances accidentogènes.

Les technologies de computer vision jouent un rôle fondamental dans le développement des véhicules autonomes et dans l'amélioration des systèmes d'assistance à la conduite qui équipent de plus en plus les véhicules modernes. En intégrant des

capteurs avancés tels que les caméras haute résolution, les lidars et les radars, ces systèmes permettent une analyse en temps réel de l'environnement routier. Grâce aux algorithmes de deep learning, les véhicules sont capables de détecter les obstacles, d'identifier les panneaux de signalisation et d'anticiper les comportements des autres usagers de la route. Cette capacité d'interprétation visuelle ne se limite pas aux véhicules totalement autonomes, mais bénéficie également aux technologies d'aide à la conduite, comme le freinage d'urgence assisté, l'alerte de franchissement de ligne et l'adaptation automatique de la vitesse. Ces innovations contribuent significativement à la réduction des risques d'accident en optimisant la prise de décision en fonction des conditions dynamiques du trafic et en assistant le conducteur dans les situations critiques.

L'application du machine learning à la prédiction de la gravité des accidents routiers représente une avancée significative dans l'amélioration de la sécurité routière. En exploitant les données historiques et en affinant les modèles prédictifs, il devient possible d'identifier les scénarios les plus à risque et de mettre en place des stratégies adaptées pour réduire le nombre et la gravité des accidents.

Défis

La mise en place d'un modèle de machine learning pour prédire la gravité d'un accident de la route repose sur des défis majeurs liés à la complexité du phénomène accidentel. Un accident est un événement multifactoriel, influencé par l'humain, le véhicule et l'environnement (HVE). La capacité du modèle à capturer cette interaction complexe déterminera en grande partie sa pertinence et sa fiabilité.

Un autre enjeu clé concerne l'interprétabilité des résultats. Les modèles de machine learning, notamment les approches profondes, offrent des prédictions précises mais sont souvent perçus comme des « boîtes noires ». Il est crucial de garantir une certaine transparence dans l'analyse des facteurs influençant la gravité des accidents afin de permettre aux décideurs et experts de comprendre et d'exploiter efficacement ces prédictions.

Enfin, la qualité des données joue un rôle déterminant. Les bases de données publiques contiennent souvent des informations incomplètes, biaisées ou hétérogènes, ce qui peut affecter la robustesse du modèle. Une attention particulière doit être portée à la collecte, au nettoyage et à la structuration des données afin d'assurer des résultats exploitables et pertinents pour la prévention et la gestion des risques routiers.

Problématique

L'objectif de notre étude est de développer un modèle de machine learning permettant de prédire la gravité des blessures subies par un usager impliqué dans un accident de la route. Cette problématique revêt une importance capitale dans l'optimisation des interventions des secours et l'amélioration des politiques de prévention routière.

Notre variable cible est la gravité des blessures de l'usager, catégorisée en quatre classes :

- Indemne
- Blessé léger
- Blessé grave
- Tué

Cependant, cette distribution est fortement déséquilibrée, les cas d'usagers indemnes et légèrement blessés étant majoritaires. Une analyse approfondie de cette distribution sera effectuée lors de l'exploration des données afin d'ajuster les choix méthodologiques en conséquence.

Le modèle doit être capable de classifier les accidents selon la gravité des blessures des usagers impliqués. Toutefois, la structure en quatre classes représente un défi en raison du déséquilibre des données et des différences marquées entre les catégories. Pour améliorer la performance du modèle et réduire les erreurs de classification, nous envisageons une binarisation de la variable cible en regroupant certaines classes :

- Groupe 1 : Indemnes et blessés légers
- Groupe 2 : Blessés graves et tués

Cette approche vise à faciliter l'apprentissage du modèle en augmentant la séparation entre les catégories et en garantissant une meilleure prise en compte des cas critiques.

Le choix de la métrique de performance est déterminé par les objectifs finaux du modèle. Dans notre cas, il est essentiel de maximiser le recall, afin de privilégier la détection des accidents graves. Cette approche implique que nous acceptons un taux plus

élevé de faux positifs, mais permet d'identifier un maximum de situations critiques nécessitant une intervention prioritaire. Un compromis sera à trouver entre recall et précision pour garantir un modèle à la fois efficace et utilisable en contexte réel.

Ressources

Il existe de précieuses ressources qui pourront m'aider lors de ce projet, la plus importante: l'Observatoire national interministériel de la sécurité routière, qui produit, notamment, des bilans et analyses statistiques annuelles sur le sujet.

Afin d'approfondir ma compréhension des enjeux liés à mon projet, j'ai engagé des interactions approfondies avec les acteurs impliqués dans la collecte et la saisie des données relatives à la mortalité routière en France. Ces échanges, notamment avec les forces de l'ordre responsables du bulletin d'analyse des accidents de la circulation (BAAC), ont considérablement enrichi ma démarche en me fournissant des informations précieuses sur les pratiques de terrain.

La documentation sur le BAAC, bien qu'essentielle, offre une vision théorique et parfois datée de la structure des données et des méthodologies employées. Pour pallier cette limite, j'ai mené des entretiens avec des agents responsables de la saisie des données. Ces échanges ont mis en lumière les disparités existantes entre les méthodologies, les outils logiciels, et les consignes appliquées par les différentes forces de l'ordre. Ces variations influencent directement la comparabilité et la cohérence des données. Par exemple, l'utilisation de matériels et de logiciels hétérogènes engendre des erreurs de saisie et des incohérences structurelles dans les données, une problématique que je devrai aborder lors de l'étape de prétraitement.

Au cours de mes échanges, j'ai également pu explorer les processus opérationnels associés à la collecte des données sur les lieux des accidents. L'importance du pré-rapport, saisi dans un délai réglementaire après l'accident, a été soulignée. Bien que ce document améliore l'immédiateté des données, la pression opérationnelle peut affecter leur qualité, nécessitant une vérification rigoureuse par les services spécialisés. Cette étape de validation, bien qu'indispensable, n'élimine pas totalement les biais humains. Ces biais, qu'ils soient liés à l'interprétation ou aux choix contraints par les options disponibles, nécessitent une modélisation attentive pour garantir la robustesse et la précision des prédictions.

En complément de ces entretiens, mes échanges avec d'autres acteurs de la sécurité routière, tels que les enseignants de la conduite, les formateurs certifiés (BAFM), et les inspecteurs du permis de conduire, ont élargi ma compréhension des dynamiques

systémiques liées à l'accidentologie. Ces perspectives pratiques enrichissent ma capacité à formuler des hypothèses pertinentes sur les déterminants des accidents.

Exploration des données

Sources des données

Dans le cadre de ce projet de prédiction de la gravité des accidents de la route, nous exploitons des données issues d'une source publique, le Bulletin d'Analyse des Accidents Corporels (BAAC). Ces données sont considérées comme fiables, car elles sont collectées par des agents humains directement sur les lieux des accidents ainsi que lors d'une phase ultérieure d'enquête. Toutefois, ce mode de collecte peut introduire des biais, notamment liés à l'interprétation des circonstances ou à des erreurs de saisie.

Les données se structurent en quatre catégories principales :

- Les lieux des accidents : informations descriptives du lieu de l'accident.
- Les caractéristiques des accidents : variables temporelles et contextuelles comme la date, les conditions météorologiques et autres circonstances générales.
- Les véhicules impliqués : type et nombre de véhicules concernés par chaque accident.
- Les usagers impliqués : profil des personnes concernées, incluant les conducteurs, passagers et piétons.

L'ensemble des données couvre la période 2005 à 2022, ce qui permet d'analyser l'évolution des tendances accidentologiques sur près de deux décennies. Chaque année, quatre fichiers correspondant aux quatre catégories précédentes sont générés, représentant un total de 72 fichiers à traiter. Cette volumétrie constitue un défi en termes de stockage et de traitement des données, nécessitant des méthodes adaptées pour l'analyse et la modélisation.

Il est important de noter que la méthodologie de saisie des données a évolué au fil des années, ce qui peut impacter la comparabilité des données historiques. Jusqu'en 1966, les tués étaient comptabilisés immédiatement ou dans un délai de trois jours après l'accident. Entre 1967 et 2004, cette période d'observation a été étendue à six jours. Depuis 2005, la définition officielle considère les décès survenus dans les trente jours suivant l'accident.

Par ailleurs, à partir de 2018, les processus de saisie par les forces de l'ordre ont évolué, rendant difficile la comparaison directe avec les années antérieures. Ces ajustements doivent être pris en compte lors de l'interprétation des tendances et de la modélisation des prédictions.

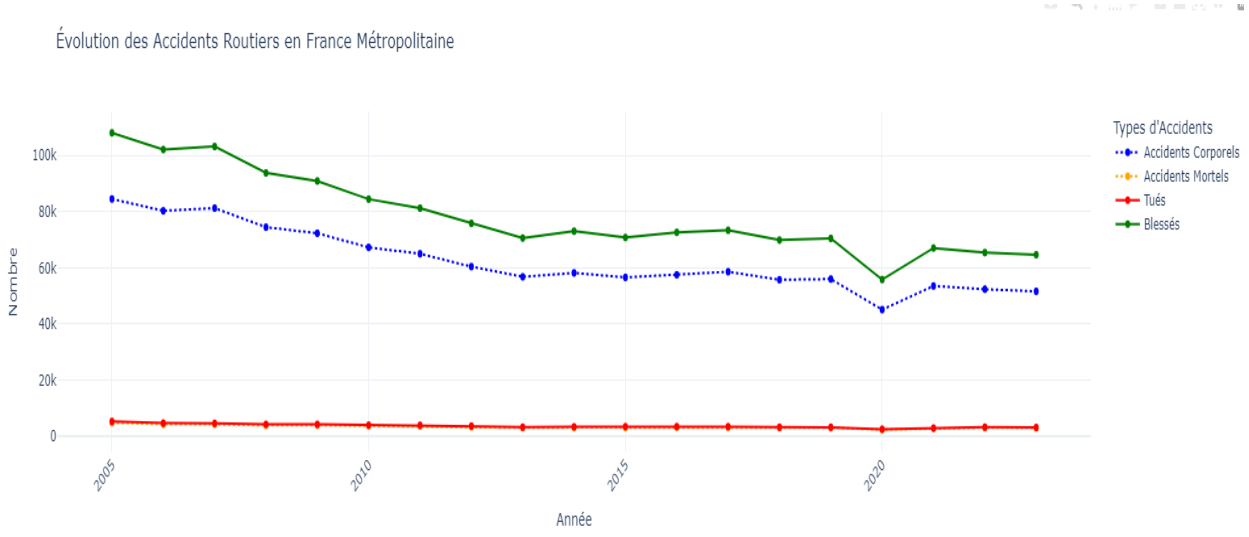
Enfin, ces données ne portent pas atteinte à la protection de la vie privée des usagers, aucun élément permettant d'identifier personnellement les individus impliqués n'étant inclus dans les fichiers disponibles.

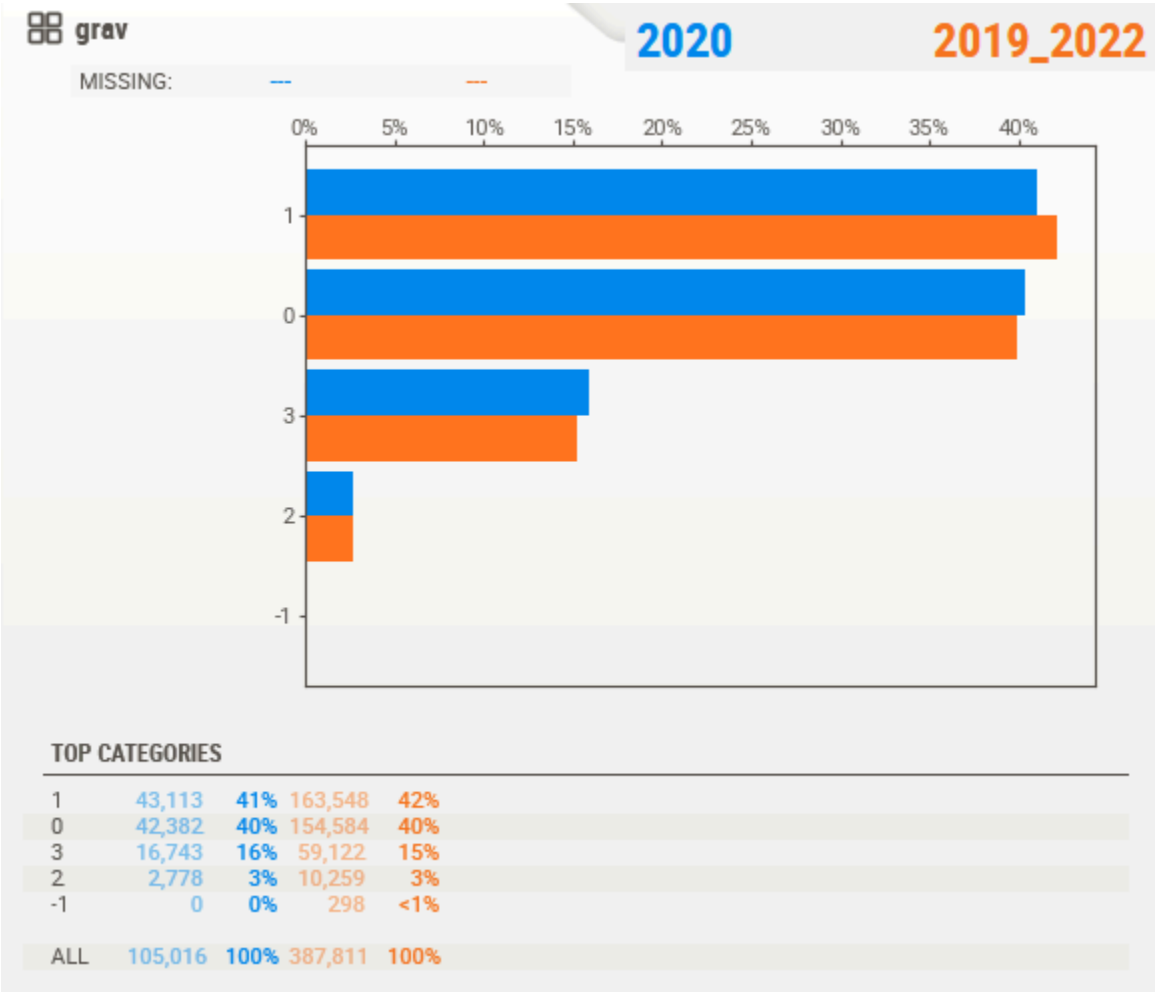
L'exploration de ces données est une étape clé pour comprendre leur structure, identifier d'éventuelles valeurs manquantes ou incohérentes, et adapter les choix méthodologiques pour le développement du modèle de machine learning.

Le cas COVID-19

L'évolution des processus de saisie des données en 2018 a conduit à limiter le champ d'étude aux années 2019 à 2022. Dans ce contexte, l'année 2020 représente un cas particulier en raison des restrictions liées à la pandémie de COVID-19, qui ont fortement impacté la mobilité des Français, notamment lors des périodes de confinement.

Une étude spécifique a donc été menée afin d'évaluer l'effet de cette année exceptionnelle sur les données d'accidentologie. En comparant les données de 2020 avec celles de 2019, 2021 et 2022, on observe une baisse du nombre d'accidents, ce qui correspond à la réduction du trafic en France métropolitaine durant les périodes de restrictions.





Répartition des classes de la variable cible 'griv' pour les années 2020 et 2019 à 2022 (2020 exclue).

Cependant, l'analyse de la répartition des classes de gravité des accidents ne montre pas de différences significatives entre ces années. Les variables disponibles dans notre jeu de données ne permettent pas de déceler un changement marqué du comportement des usagers de la route durant cette période.

Les tendances dans la répartition des autres variables restent tout à fait semblable, à tout niveau. L'étude des corrélations entre les variables

Ainsi, bien que le volume total des accidents ait diminué en 2020, l'impact sur la distribution des classes de gravité semble limité. Par conséquent, il paraît raisonnable d'inclure l'année 2020 dans le champ d'étude du modèle de machine learning sans ajustement spécifique. Cette approche permettra de tirer pleinement parti des données

disponibles tout en conservant une cohérence dans l'analyse et la modélisation des prédictions de gravité des accidents.

Séléction des variables

L'élaboration d'un modèle de machine learning nécessite une identification rigoureuse des variables influentes. Une sélection adéquate des caractéristiques permet d'améliorer la robustesse des prédictions, d'atténuer les biais et de garantir une meilleure généralisation du modèle. Cette étape repose sur une exploration approfondie des données disponibles, en vue d'éliminer les variables redondantes et d'optimiser la représentation des phénomènes accidentogènes.

L'approche adoptée repose sur un processus itératif où chaque phase d'exploration des données est suivie d'un traitement, avant d'être réévaluée à la lumière des résultats obtenus. Cette méthode permet d'affiner progressivement la sélection des variables, d'identifier d'éventuels biais et d'optimiser la pertinence des données intégrées dans le modèle. L'ajustement continu du pipeline d'analyse garantit ainsi une meilleure adéquation entre les variables retenues et la capacité prédictive du modèle.

Un notebook Jupyter intitulé exploration.ipynb a été conçu afin d'automatiser l'analyse des variables. Cette approche permet une évaluation systématique de chaque attribut, incluant la description des modalités, la typologie des variables, la proportion de valeurs manquantes, la distribution des catégories et leur évolution temporelle, la présence de doublons.

Cette analyse initiale a abouti à la production d'un rapport détaillé de 86 pages, documentant chaque variable avec des informations quantitatives et qualitatives. Ce document a ensuite été enrichi par Erika, qui y a intégré des analyses numériques complémentaires. Le fichier résultant, Exploration des variables - rapport EM V240624.pdf, demeure disponible en tant que référence analytique.

Par ailleurs, une synthèse des variables et de leurs descriptions a été compilée dans Description des variables.pdf. Une version structurée et exploitable sous format JSON a été produite sous le nom descvar.json, facilitant l'intégration des informations dans les pipelines de traitement des données.

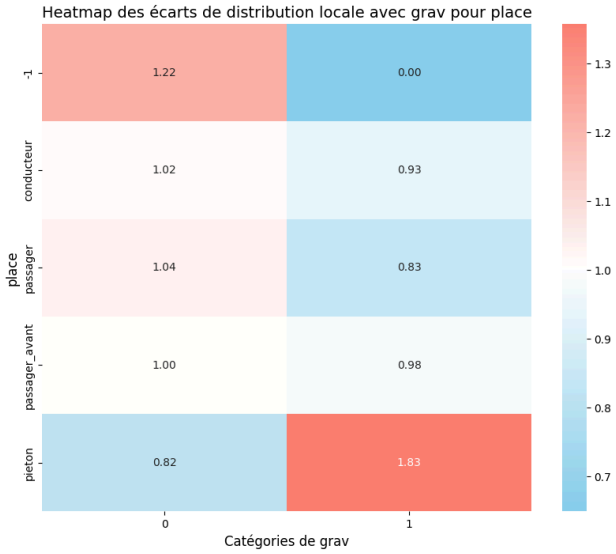
L'analyse des variables a permis d'identifier celles ayant une influence significative sur la gravité des accidents, en se basant sur des critères statistiques et sur la pertinence contextuelle. Plutôt que de traiter exhaustivement l'ensemble des variables, l'accent est mis sur les facteurs les plus déterminants :

- place : position de l'usager dans le véhicule, facteur déterminant en cas de choc frontal ou latéral.
- catu : catégorie d'usager (conducteur, passager, piéton, etc.), influençant directement l'exposition au risque.
- grav : variable cible indiquant la gravité des blessures subies.
- sexe et an_nais : variables démographiques pouvant moduler les risques en fonction des profils d'usagers.
- trajet : contexte du déplacement (professionnel, personnel), influant sur le comportement de conduite.
- locp et actp : localisation et action du piéton, cruciales pour l'analyse des accidents impliquant des usagers vulnérables.
- vosp : présence d'une voie réservée.
- catv : type de véhicule impliqué, impactant la cinétique de l'accident.
- obs / obsm : identification des obstacles heurtés, précisant les circonstances de l'accident.
- manv : manœuvre du véhicule, permettant de caractériser la dynamique de l'incident.
- jour, mois, an : dimensions temporelles influençant la variabilité saisonnière des accidents.
- lum : conditions de luminosité, facteur de visibilité critique.
- agg et int : contexte urbain et type d'intersection, impactant la fréquence des collisions.
- atm : conditions atmosphériques, affectant l'adhérence et la perception.
- col : type de collision, structurant l'intensité du choc.
- catr, circ et vma : attributs liés au réseau routier et aux conditions de circulation.
- surf, prof et plan : caractéristiques physiques de la chaussée, modulant les risques d'accident.

- infra et situ : aménagements et situations spécifiques pouvant influencer la sévérité des blessures.
- choc: point d'impact initial lors de l'accident.
- reg: régions, initialement départements mais entrainait une dimensionnalité trop élevée.
- secu: les équipements de sécurité utilisés.
- imply_x: implication d'un type de véhicule dans l'accident hors véhicule courant.

Variables catégorielles

A. Place



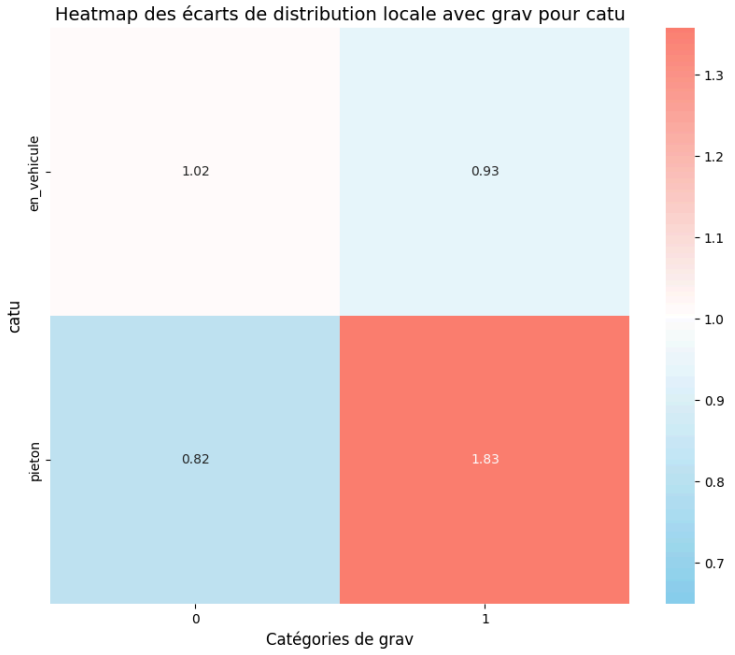
L'une des dimensions cruciales est la position de l'usager dans le véhicule (place), qui peut influencer de manière significative l'issue d'un accident. La heatmap ci-dessus met en évidence les écarts entre la distribution globale de la variable grav et sa distribution locale en fonction de la position des usagers impliqués dans un accident, permettant ainsi de quantifier les différences et d'identifier les biais éventuels.

L'analyse des écarts de distribution met en évidence des variations significatives selon la position des usagers. Les piétons présentent une forte disparité entre les catégories de gravité, avec un coefficient de 1.83 pour la classe grav=1 (accidents graves) et 0.82 pour grav=0 (accidents non graves). Cette sur-représentation des cas graves est cohérente avec le fait que les piétons sont plus vulnérables lors d'un impact. Les conducteurs affichent un coefficient de 1.02 en grav=0 et 0.93 en grav=1, ce qui indique une répartition plus équilibrée des niveaux de gravité dans cette catégorie. Les passagers avant et arrière présentent également des valeurs proches de 1, illustrant une distribution relativement homogène des blessures. On note une légère sous-représentation des passagers arrières dans les accidents graves avec 0.83.

L'exploration des écarts de distribution entre la position des usagers et la gravité des accidents confirme la nécessité d'intégrer cette variable dans le modèle.

B. Catu - Catégorie d'usager

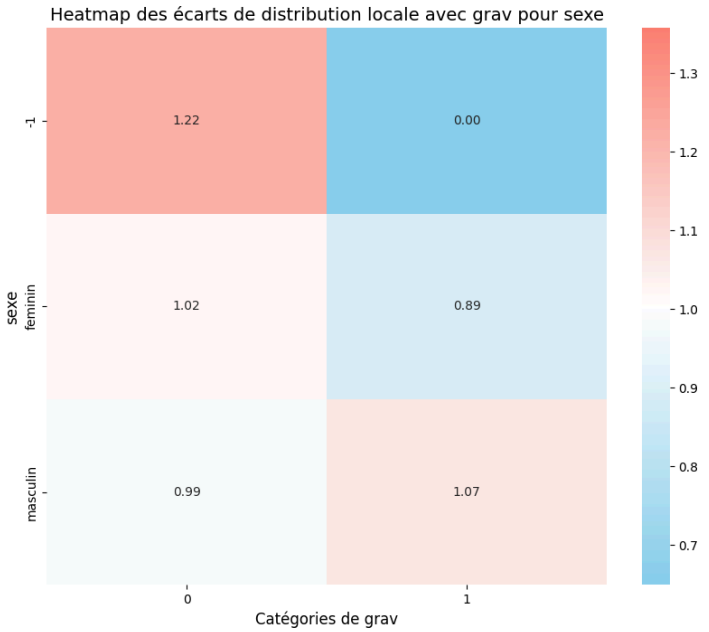
Une autre des dimensions fondamentales à explorer est la catégorie d'usager (catu), qui regroupe notamment les conducteurs, passagers et piétons.



L'analyse des écarts de distribution met en évidence des différences notables selon la catégorie d'usager. Les piétons montrent une forte disparité entre les classes de gravité, avec une sur-représentation des cas graves (grav=1) mesurée à 1.83, tandis que leur proportion dans les cas moins graves (grav=0) est sous-représentée à 0.82. Cette tendance est conforme à la vulnérabilité intrinsèque des piétons, qui subissent généralement des blessures plus sévères en cas de collision. Et corobore les conclusions établies par la variable place.

Les usagers en véhicule, comprenant conducteurs et passagers, présentent une répartition plus équilibrée avec un coefficient de 1.02 pour grav=0 et 0.93 pour grav=1. Cette homogénéité indique que la gravité des blessures est moins influencée par la catégorie d'usager dans ce groupe, en comparaison avec les piétons.

A la lumière de cette analyse, malgré l'importance de la catégorie d'usager, il apparait qu'il s'agit d'une variable redondante avec la colonne place qui contient déjà l'information.



C. Sexe

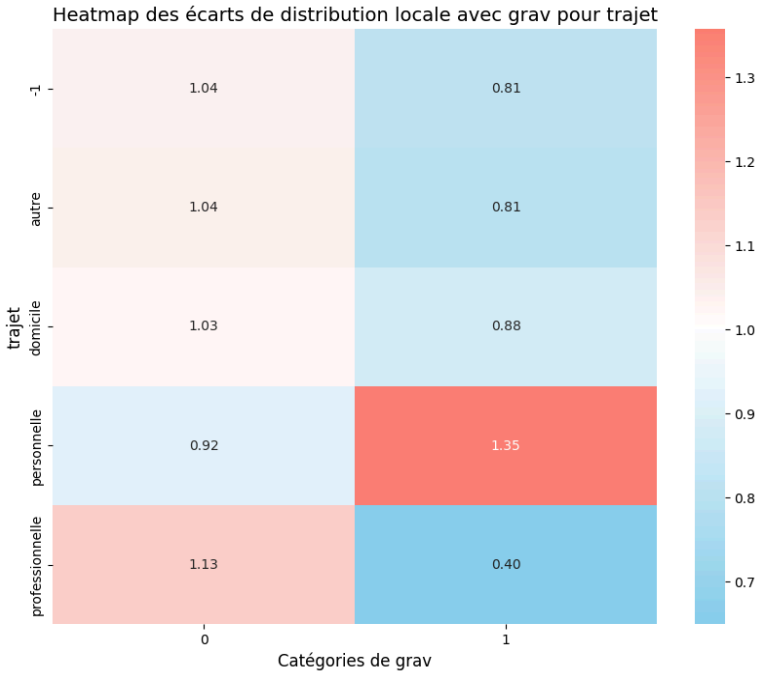
Après avoir étudié la catégorie d'usager, nous nous intéressons désormais au sexe des usagers impliqués.

L'analyse des écarts de distribution révèle des différences notables entre les sexes. Les femmes présentent une très légère sur-représentation dans les cas de faible gravité (grav=0), avec un coefficient de 1.02, et une sous-représentation dans les cas graves (grav=1), mesurée à 0.89. À l'inverse, les hommes affichent une répartition inversée avec 0.99 pour grav=0 et une légère sur-représentation des cas graves à 1.07. Ces résultats suggèrent une exposition plus importante des hommes aux blessures graves lors des accidents, ce qui pourrait être lié à des différences de comportement au volant, d'exposition au risque ou de type d'accident rencontré.

L'écart observé justifie l'inclusion de la variable sexe dans le modèle prédictif.

D. Trajet

Dans la continuité de notre étude sur les variables explicatives influençant la gravité des accidents, nous nous intéressons ici à la nature du trajet au moment de l'accident. La distinction entre trajets personnels, professionnels ou autres peut avoir un impact significatif sur la distribution des niveaux de gravité.



L'analyse des écarts de distribution met en évidence des tendances distinctes en fonction du type de trajet. Les accidents survenant lors de trajets personnels montrent une sur-représentation des cas graves (grav=1) avec un coefficient de 1.35, ce qui indique une exposition plus importante aux accidents graves dans ce contexte. À l'inverse, les trajets professionnels présentent une sous-représentation marquée des cas graves (grav=1) avec un coefficient de 0.40, suggérant un niveau de gravité moindre pour ces accidents. Ce phénomène pourrait être lié à des comportements plus prudents des conducteurs dans un cadre professionnel ou à des différences structurelles comme le type de véhicules utilisés.

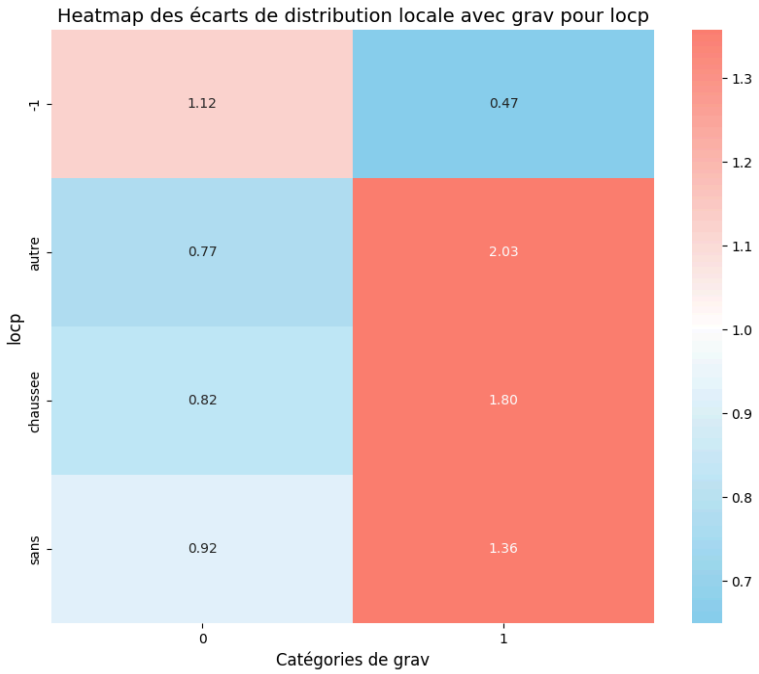
Pour une analyse plus approfondie des accidents impliquant des véhicules utilitaires légers (VULs), il est recommandé de se référer à l'étude Études détaillées d'accidents impliquant des véhicules utilitaires légers de Thierry Serre, Christophe Perrin, Maxime Dubois-Lounis et Claire Naude. Cette étude apporte des éléments de

compréhension supplémentaires sur les spécificités de ces véhicules en matière de risques et de gravité des accidents. Et peut expliquer l'analyse précédente.

Les trajets domicile-travail et les autres types de trajets affichent des écarts moins prononcés, avec des coefficients proches de l'équilibre, respectivement 1.03 pour grav=0 et 0.88 pour grav=1. Ces valeurs indiquent une distribution relativement homogène, ne présentant pas de décalage majeur entre les classes de gravité.

L'influence significative du type de trajet sur la gravité des accidents justifie l'inclusion de cette variable dans le modèle prédictif. Elle apporte des informations importantes pour le modèle.

E. Locp - localisation piéton

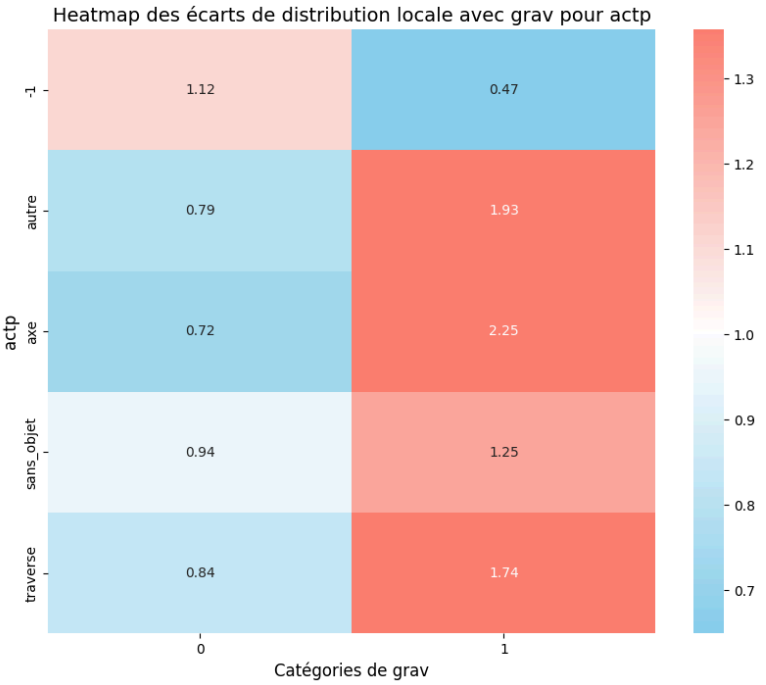


Nous analysons ici l'impact de la localisation du piéton au moment de l'accident. Cette variable est essentielle pour comprendre les conditions dans lesquelles les piétons sont les plus vulnérables.

L'analyse des écarts de distribution révèle une sur-représentation marquée des accidents graves (grav=1) pour les piétons localisés sur la chaussée, avec un coefficient de 1.80. Il est important de noter que cette modalité ne précise pas la présence ou l'absence d'un passage piéton, ce qui limite l'interprétation en termes d'infrastructure sécurisée ou non. La catégorie autre, qui regroupe les piétons situés sur trottoirs, accotements, refuges ou bandes d'arrêt d'urgence, affiche une sur-représentation encore plus marquée des cas graves (grav=1) avec un coefficient de 2.03. Cette observation suggère que les piétons impliqués dans des accidents en dehors des voies de circulation classiques subissent des blessures plus sévères, possiblement en raison de l'effet de surprise ou d'un impact à haute vitesse.

Les écarts de distribution observés confirment que la localisation du piéton est une variable essentielle pour la prédiction de la gravité des accidents.

F. Actp - action piéton



Nous nous intéressons ici à l'action du piéton au moment de l'accident.

L'analyse des écarts de distribution révèle une sur-représentation marquée des accidents graves (grav=1) lorsque le piéton est situé sur l'axe de la chaussée, avec un coefficient de 2.25. Cette valeur indique une forte exposition au risque lorsque le piéton se trouve dans des zones où la circulation est dense et rapide. Les piétons impliqués dans des accidents tout en étant en train de traverser affichent également une surexposition aux accidents graves (grav=1) avec un coefficient de 1.74, ce qui reflète un danger significatif lors des traversées.

La catégorie autre présente un coefficient de 1.93 pour les cas graves, ce qui suggère que certaines actions spécifiques, telles que le fait d'être masqué, de courir ou jouer, d'interagir avec un animal, de monter ou descendre d'un véhicule, ou encore d'autres comportements atypiques, exposent les piétons à des risques accrus.

Encore une fois, la modalité sans_objet représente un accident n'impliquant pas de piéton.

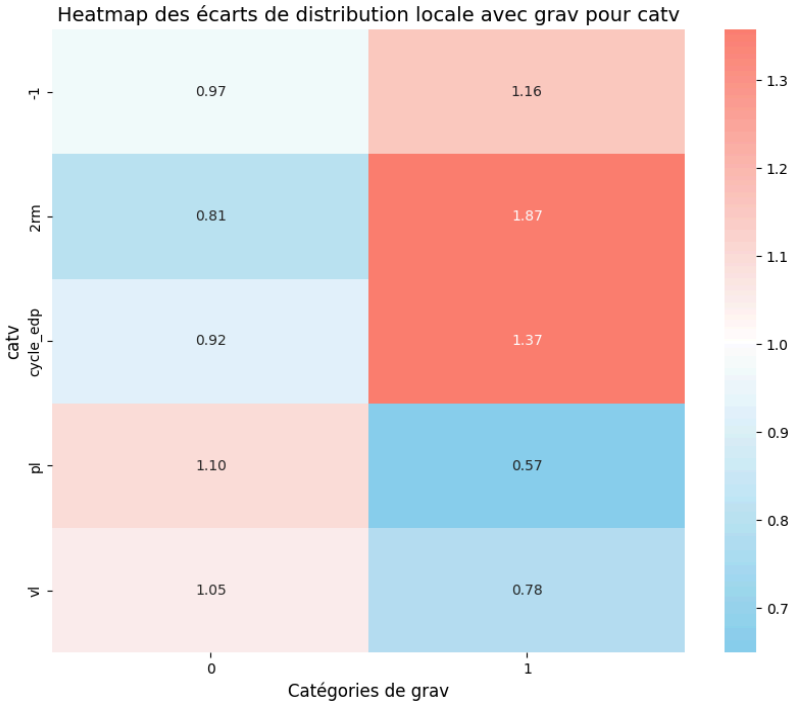
Ces observations confirment l'importance d'inclure la variable actp dans le modèle prédictif. Cependant, une attention particulière devra être apportée aux variables de la galaxie piéton afin d'éviter une sur pondération par le modèle.

G. catv - catégorie de véhicule

Nous nous penchons ici sur la catégorie de véhicule impliqué. Ce facteur joue un rôle déterminant dans l'intensité d'un impact et les blessures subies par les usagers de la route.

L'analyse des écarts de distribution met en évidence des différences notables entre les catégories de véhicules. Les accidents impliquant des deux-roues motorisés

(2rm) sont fortement sur-représentés dans les cas graves (grav=1), avec un coefficient de 1.87. Cette observation est cohérente avec la vulnérabilité inhérente des usagers de deux-roues, qui sont moins protégés en cas de choc. De même, les cycles et engins de déplacement personnel (cycle_edp) présentent une sur-représentation des cas graves avec un coefficient de 1.37, confirmant un risque accru pour ces usagers en raison de l'absence de structure protectrice.



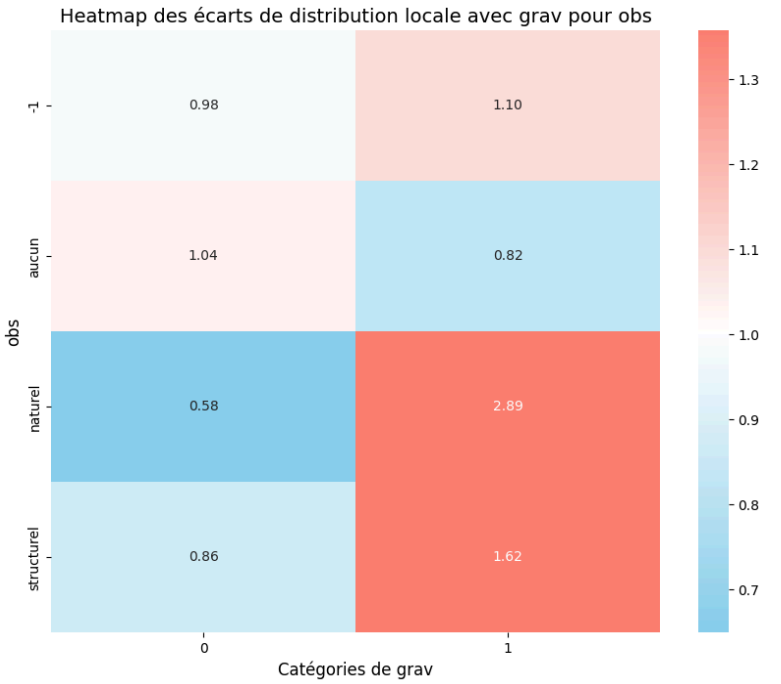
En revanche, les poids lourds (pl) et véhicules légers (vl) présentent une sous-représentation des cas graves (grav=1), avec des coefficients respectifs de 0.57 et 0.78. Cela peut s'expliquer par le fait que les occupants de ces véhicules bénéficient d'une meilleure protection structurelle, notamment grâce aux dispositifs de sécurité avancés. Cependant, il est important de noter que si la gravité des blessures pour les occupants de ces véhicules est moindre, leur implication dans des accidents peut augmenter la gravité des blessures pour les autres usagers de la route.

L'importance de la catégorie de véhicule dans la gravité des accidents justifie pleinement son intégration dans le modèle prédictif.

H. Obs - Obstacle fixe heurté

Cette variable, nommée obs, peut jouer un rôle significatif dans l'intensité des blessures subies par les occupants des véhicules.

L'analyse des écarts de distribution révèle des tendances marquées selon la nature de l'obstacle. Les accidents impliquant un obstacle de type naturel (comme un arbre ou un rocher) sont



fortement sur-représentés dans les cas graves (grav=1), avec un coefficient de 2.89. Cette observation peut être attribuée à la rigidité et à l'inertie de ces obstacles, qui absorbent peu l'énergie cinétique, augmentant ainsi la gravité des impacts.

De leur côté, les obstacles structurels (tels que des barrières ou des murs) présentent également une sur-représentation des cas graves avec un coefficient de 1.62. Bien que ces obstacles puissent être conçus pour amortir les chocs dans certains cas, leur rigidité reste un facteur aggravant dans de nombreux scénarios d'accidents.

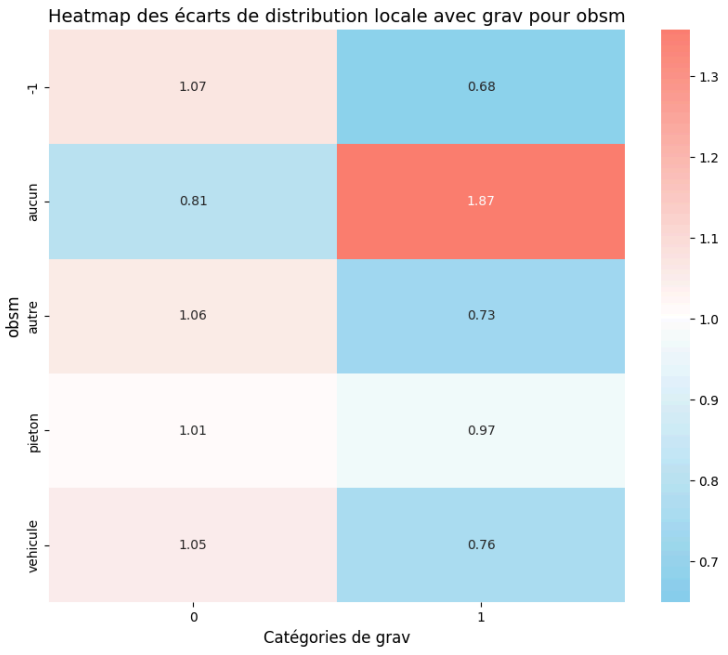
Pour des recommandations et analyses supplémentaires concernant la sécurité des infrastructures et leur rôle dans la gravité des accidents, consulter l'étude intitulée La sécurité des infrastructures, une analyse et des recommandations par le comité des experts au Conseil National de Sécurité Routière par Marie Line Gallenne et Vincent Ledoux, disponible dans l'archive ouverte HAL (HAL Id: hal-01317212). Cette ressource offre une perspective détaillée sur l'influence des infrastructures routières dans la sécurité des usagers.

En revanche, la modalité aucun, indiquant l'absence d'un obstacle fixe percuté, est associée à une sous-représentation des cas graves (grav=1), avec un coefficient de 0.82. Cela reflète le fait que les accidents sans impact contre un obstacle fixe tendent à être moins sévères, souvent limités aux collisions entre véhicules ou à des pertes de contrôle sans choc majeur.

Ces résultats soulignent l'importance de prendre en compte la variable obs dans le modèle prédictif. Les écarts significatifs observés pour les obstacles naturels et structurels justifient une attention particulière à ces catégories.

L. Obsm - Obstacle mobile heurté

Dans la continuité de notre analyse des variables influençant la gravité des accidents, nous nous intéressons ici à la nature de l'obstacle mobile percuté lors d'un accident. Cette variable, nommée obsm, peut avoir un impact significatif sur l'intensité des blessures subies, selon qu'il s'agisse d'un véhicule, d'un piéton ou d'un animal (catégorie autre).



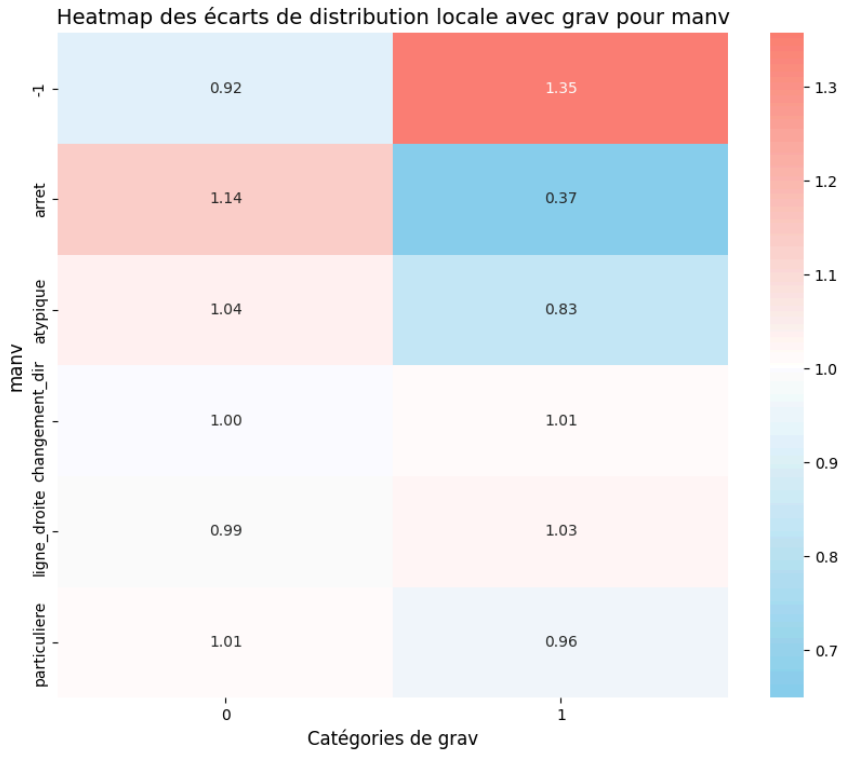
L'analyse des écarts de distribution révèle que les accidents où aucun obstacle mobile n'a été

percuté (aucun) sont sur-représentés dans les cas graves (grav=1), avec un coefficient de 1.87. Cela peut s'expliquer par le fait que ces accidents sont souvent liés à des pertes de contrôle ou des collisions contre des obstacles fixes, augmentant la gravité.

En revanche, les collisions avec un véhicule présentent une sous-représentation dans les cas graves (grav=1), avec un coefficient de 0.76.

Les accidents impliquant un piéton montrent une répartition relativement équilibrée entre les gravités, avec un coefficient de 0.97 pour les cas graves, reflétant une variabilité selon les circonstances spécifiques de l'accident. Ce qui est étonnant après analyse des variables précédentes.

Enfin, la catégorie autre, qui fait principalement référence à des collisions avec des animaux, affiche une sous-représentation des cas graves (grav=1), avec un coefficient de 0.73, ce qui peut s'expliquer par la nature imprévisible mais souvent moins létale de ces impacts.



M. Manv - Manoeuvre principale avant l'accident

Cette variable, nommée manv, reflète les actions des conducteurs ou usagers immédiatement avant l'impact et peut jouer un rôle significatif dans l'intensité des blessures subies.

L'analyse des écarts de distribution révèle que les accidents où un véhicule était à l'arrêt présentent une sous-représentation marquée des cas graves (grav=1), avec un coefficient de 0.37, ce qui s'explique par l'énergie cinétique moindre lors de l'impact. En revanche, les manoeuvres classées comme atypiques affichent une sur-représentation des cas graves (grav=1), avec un coefficient de 1.35, suggérant des comportements imprévisibles ou des situations complexes avant l'accident.

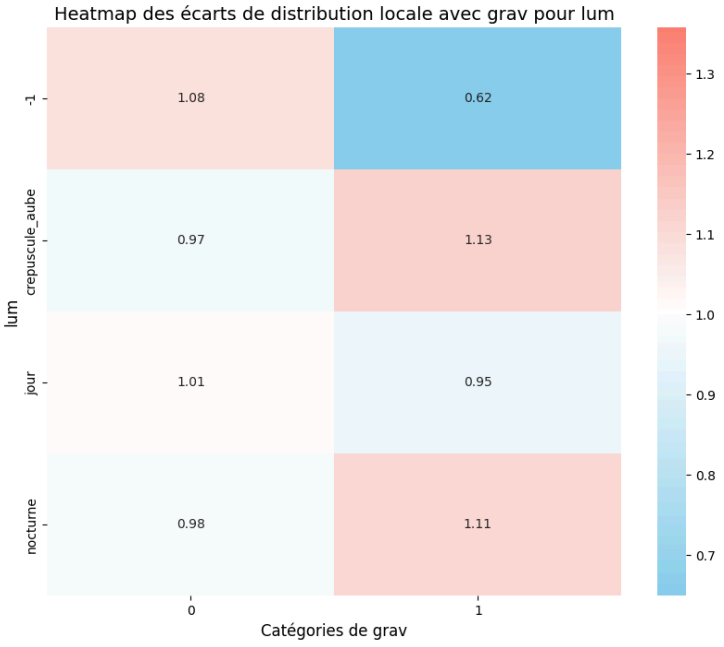
Les manoeuvres impliquant un changement de direction ou un déplacement en ligne

droite présentent des coefficients proches de l'équilibre, respectivement 1.03 et 1.01, ce qui indique une répartition homogène des gravités. Cela reflète des scénarios d'accidents plus classiques avec des impacts modérés. Enfin, les actions classées comme particulières affichent une distribution légèrement déséquilibrée en faveur des accidents moins graves (grav=0), avec un coefficient de 1.01.

La variable manv est pertinente pour la prédiction de la gravité des accidents, notamment en raison de l'écart significatif observé pour les manoeuvres atypiques.

N. Lum - Conditions lumineuses

Cette variable, nommée lum, est cruciale pour comprendre l'influence des périodes de la journée ou des conditions d'éclairage sur l'intensité des impacts et les blessures subies.



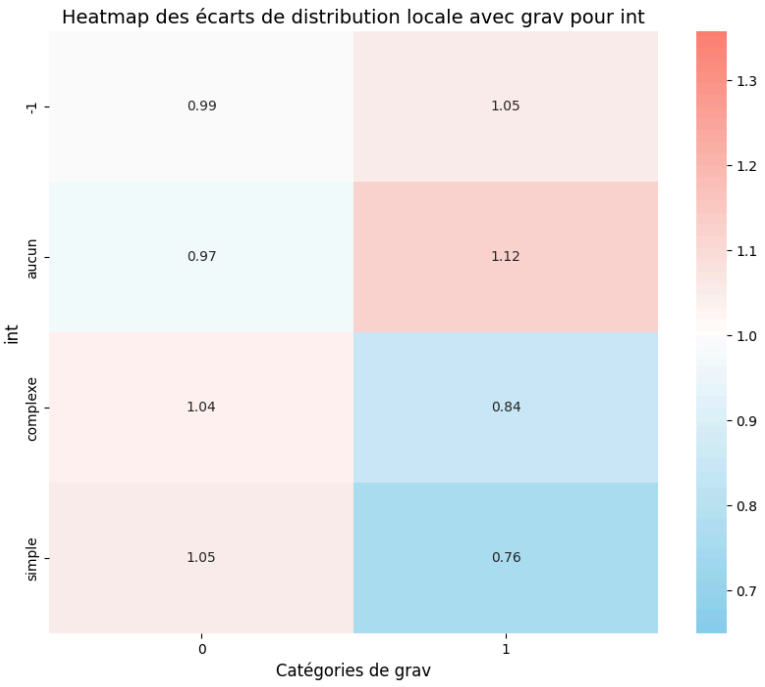
L'analyse des écarts de distribution révèle que les accidents survenant au crépuscule ou à l'aube présentent une légère sur-représentation des cas graves (grav=1), avec un coefficient de 1.13. Cette observation peut s'expliquer par une visibilité réduite pendant ces périodes, combinée à des conditions d'éclairage souvent sous-optimales.

Les accidents se produisant de nuit montrent également une sur-représentation des cas graves (grav=1), avec un coefficient de 1.11, reflétant les défis liés à la conduite dans des conditions d'obscurité, malgré l'utilisation d'éclairage artificiel.

En revanche, les accidents survenant en plein jour affichent une sous-représentation des cas graves (grav=1), avec un coefficient de 0.95, indiquant que la visibilité optimale en journée contribue à atténuer la gravité des blessures.

Ces résultats confirment que la variable lum est un facteur pertinent pour la prédiction de la gravité des accidents.

O. Int - type d'intersection



Nous nous intéressons ici au type d'intersection où s'est produit l'accident.

L'analyse des écarts de distribution révèle que les accidents survenant à des intersections simples présentent une sous-représentation des cas graves (grav=1), avec un coefficient de 0.76. Cela peut s'expliquer par une configuration moins complexe, réduisant les risques de collisions graves.

En revanche, les intersections complexes affichent une sous-représentation plus marquée des cas graves (grav=1), avec un coefficient de 0.84. Cette configuration pourrait permettre une meilleure régulation des flux de trafic, mais elle nécessite également une analyse plus fine pour comprendre les causes sous-jacentes.

Les accidents ne se produisant à aucune intersection montrent une légère sur-représentation des cas graves (grav=1), avec un coefficient de 1.12, ce qui reflète souvent des collisions à grande vitesse sur des tronçons routiers rectilignes ou moins régulés.

Ces résultats confirment que la variable int est un facteur pertinent pour la prédiction de la gravité des accidents.

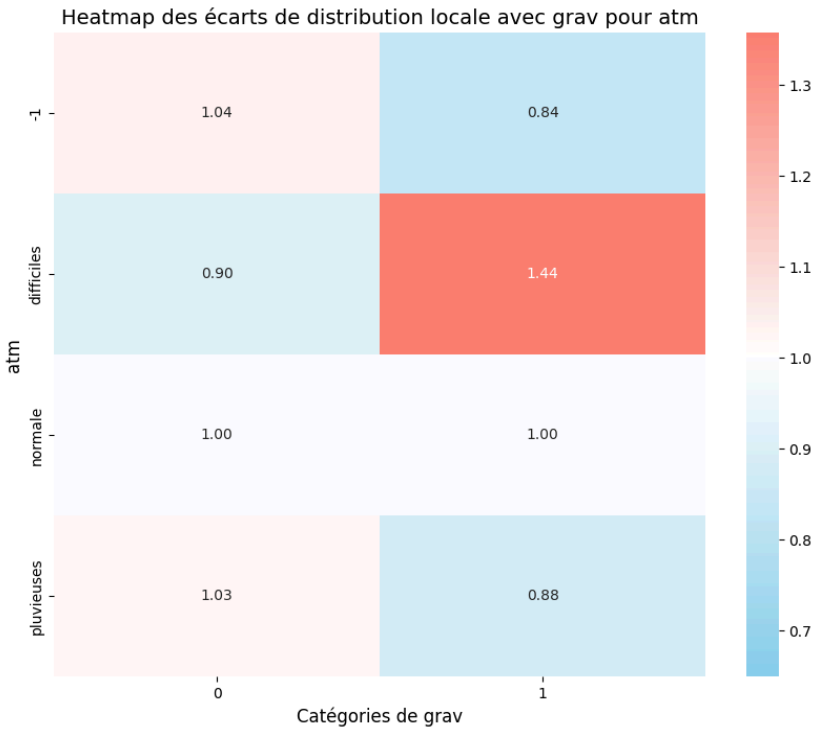
P - Atm - Condition atmosphérique

Nous nous concentrons ici sur les conditions atmosphériques au moment de l'accident.

L'analyse des écarts de distribution révèle que les accidents survenant dans des conditions difficiles (brouillard, neige, verglas, etc.) présentent une sur-représentation marquée des cas graves (grav=1), avec un coefficient de 1.44. Ces conditions augmentent les risques en raison d'une visibilité réduite et d'une adhérence moindre, rendant les collisions plus sévères.

Les accidents dans des conditions normales montrent une distribution équilibrée entre les cas graves et non graves, avec des coefficients de 1.00 pour les deux catégories. Cela reflète une situation standard où les facteurs météorologiques n'amplifient ni ne réduisent la gravité des impacts.

En revanche, les accidents survenant dans des conditions pluvieuses présentent une sous-représentation des cas graves (grav=1), avec un coefficient de 0.88. Cela peut s'expliquer par une conduite plus prudente des usagers dans de telles conditions, bien que les risques de collisions restent présents.

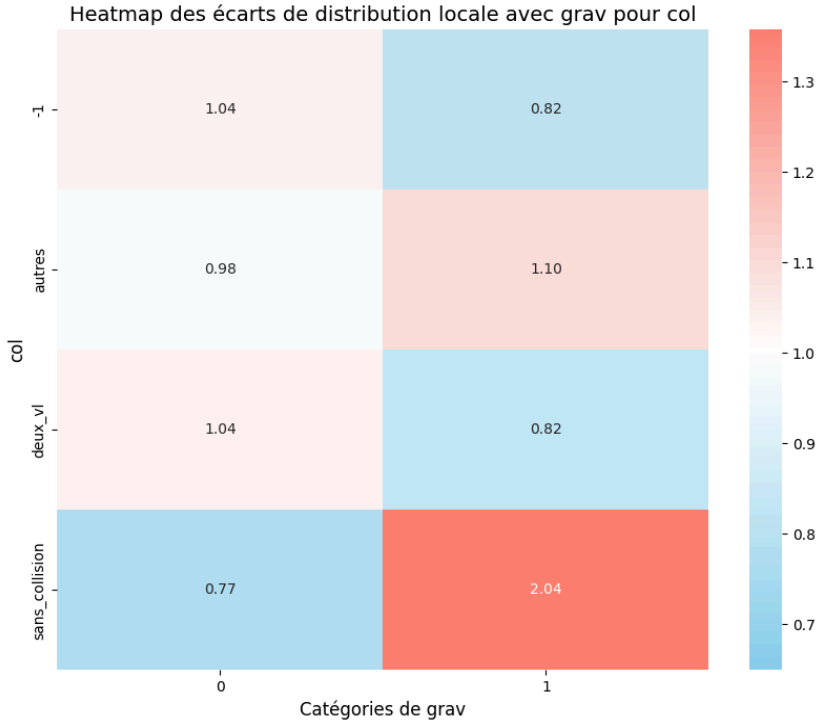


Les écarts significatifs observés pour les conditions difficiles justifient une prise en compte particulière dans le modèle.

Q. Col - Type de collision

Nous nous intéressons ici au type de collision impliqué dans l'accident.

L'analyse des écarts de distribution révèle que les accidents classés comme sans collision (exemple : sortie de route, tonneaux, etc.) présentent une sur-représentation marquée des cas graves (grav=1), avec un coefficient de 2.04. Ces situations impliquent souvent des pertes de contrôle à grande vitesse ou des impacts isolés contre des



obstacles fixes, augmentant la gravité des blessures.

En revanche, les collisions entre deux véhicules présentent une sous-représentation des cas graves (grav=1), avec un coefficient de 0.82, suggérant que les dispositifs de sécurité des véhicules et la répartition des forces lors de l'impact réduisent la gravité des blessures.

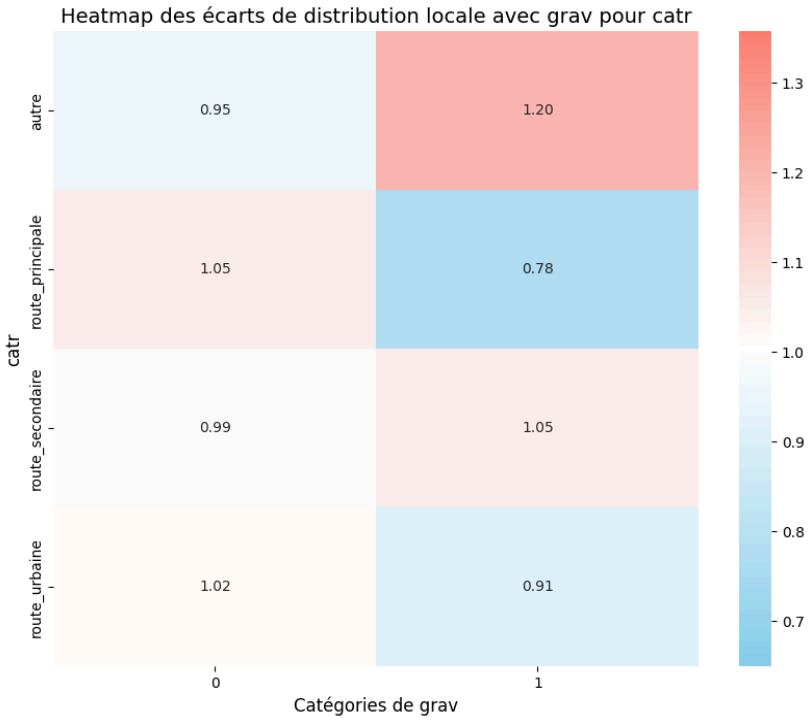
Les collisions classées dans la catégorie autres (impliquant des accidents entre plus de deux véhicules) montrent une légère sur-représentation des cas graves (grav=1), avec un coefficient de 1.10, ce qui peut refléter la complexité accrue de ces scénarios et le risque plus élevé d'impacts multiples.

Ces résultats confirment que la variable col est pertinente pour la prédiction de la gravité des accidents.

R. Catr - Catégorie de route

Nous nous intéressons ici aux catégories de routes sur lesquelles surviennent les accidents.

L'analyse des écarts de distribution révèle que les accidents survenant sur des routes classées dans la catégorie autres (regroupant des routes spécifiques non incluses dans les principales classifications) présentent une sur-représentation des cas graves (grav=1), avec un coefficient de 1.20. Cette observation pourrait s'expliquer par des conditions de circulation plus variables, des infrastructures moins standardisées ou des limitations de vitesse différentes.

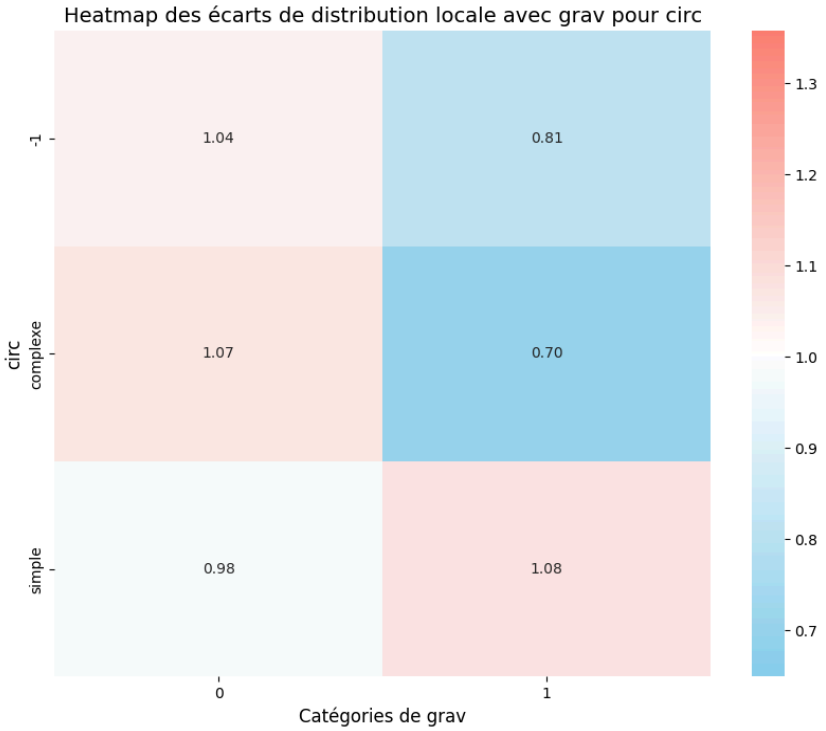


En revanche, les accidents sur routes principales présentent une sous-représentation marquée des cas graves (grav=1), avec un coefficient de 0.78. Ces axes étant généralement bien aménagés et mieux sécurisés, cela pourrait expliquer cette moindre gravité des accidents.

Les routes secondaires et les routes urbaines affichent des distributions plus équilibrées, avec des coefficients respectifs de 1.05 et 0.91 pour les cas graves (grav=1), indiquant une gravité relativement proche de la moyenne.

Ces résultats confirment que la variable catr est un facteur pertinent pour la prédiction de la gravité des accidents. Une attention particulière pourrait être portée aux routes de la catégorie autres.

S. Circ - régime de circulation



Nous nous concentrons ici sur les régimes de circulation en vigueur au moment de l'accident.

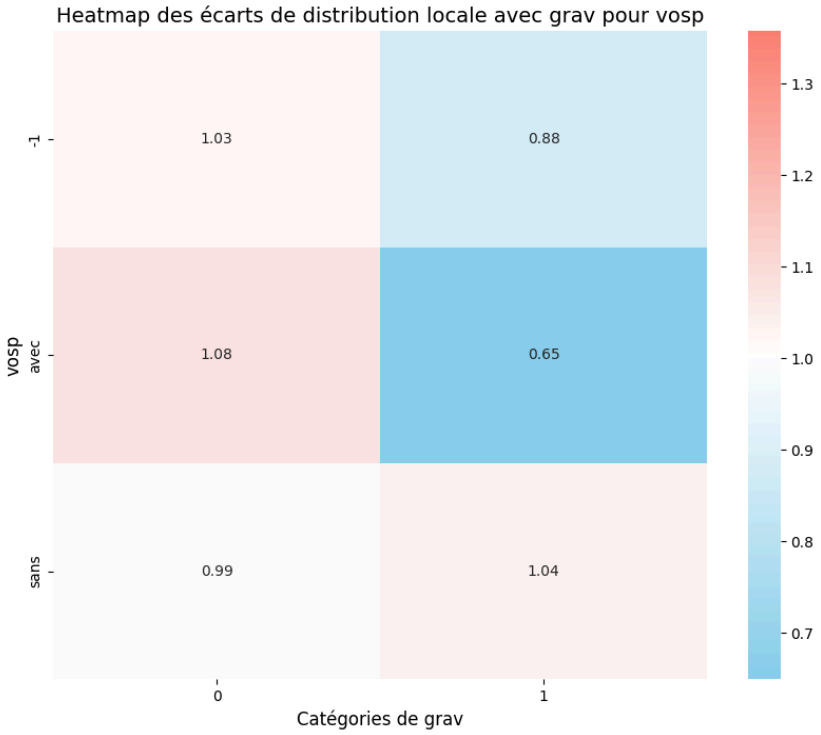
L'analyse des écarts de distribution révèle que les accidents survenant dans un régime de circulation simple présentent une sur-représentation des cas graves (grav=1), avec un coefficient de 1.08. Cette observation pourrait s'expliquer par des vitesses moyennes plus élevées sur ces axes, réduisant les possibilités d'évitement en cas de situation dangereuse.

En revanche, les accidents survenant dans un régime de circulation complexe montrent une sous-représentation des cas graves (grav=1), avec un coefficient de 0.70. Cela suggère que la présence d'intersections, de croisements multiples et d'éléments de régulation de la circulation pourrait limiter la gravité des collisions en réduisant la vitesse moyenne des véhicules.

Ces résultats ne sont pas forcément surprenants si l'on considère le principe de l'homéostasie du risque routier. Cette théorie suggère que les usagers de la route ajustent leur comportement en fonction des perceptions du danger, ce qui pourrait expliquer pourquoi les régimes de circulation plus complexes, bien que plus réglementés, ne présentent pas nécessairement une gravité d'accident plus élevée. Pour une analyse approfondie de cette approche, voir Théorie de la décision et risques routiers par Claudine Pérez-Diaz.

Ces résultats confirment que la variable circ est pertinente pour la prédiction de la gravité des accidents.

T. Vosp - Présence de voie réservée



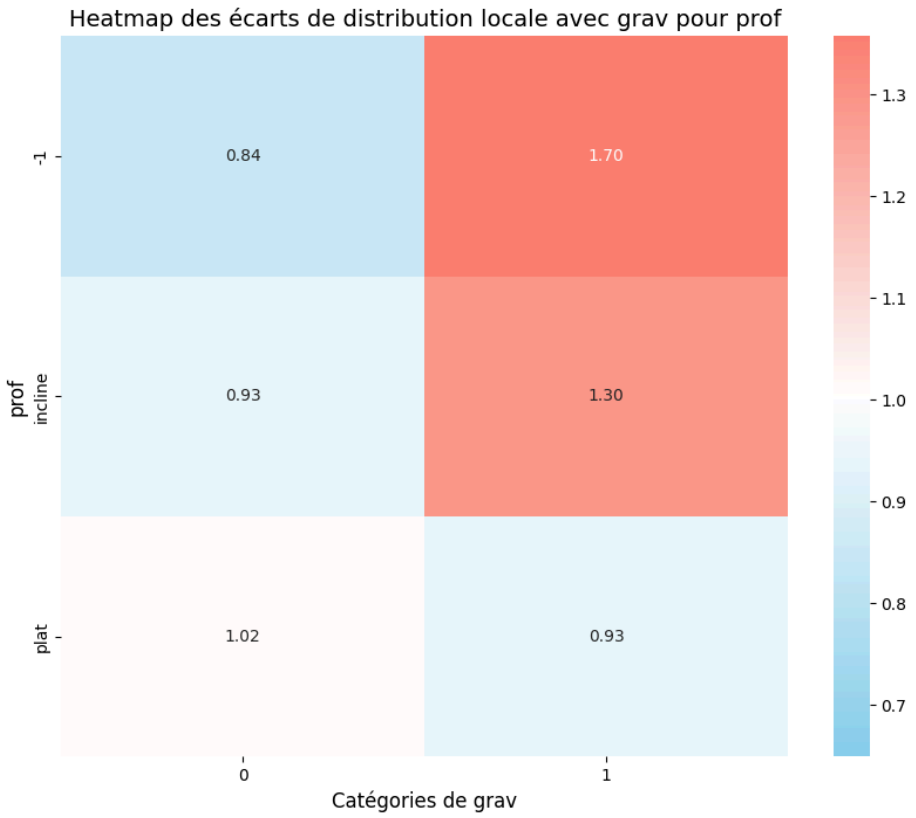
Nous nous concentrons ici sur la présence ou non d'une voie réservée sur la chaussée où s'est produit l'accident.

L'analyse des écarts de distribution révèle que les accidents survenant sur des routes avec une voie réservée présentent une sous-représentation marquée des cas graves (grav=1), avec un coefficient de 0.65. Cela pourrait être dû à une meilleure organisation du trafic et à la séparation des flux de véhicules, réduisant ainsi la gravité des impacts en cas d'accident.

En revanche, les accidents survenant sans voie réservée affichent une sur-représentation des cas graves (grav=1), avec un coefficient de 1.04. L'absence de séparation des flux de circulation pourrait favoriser des collisions plus violentes, notamment lors de manœuvres de dépassement ou en cas de trafic dense.

Ces résultats confirment que la variable vosp est pertinente pour la prédiction de la gravité des accidents.

U. Prof - profil de la route



Dans la continuité de notre analyse des variables influençant la gravité des accidents, nous nous concentrons ici sur le profil de la route, autrement dit la déclivité de la chaussée.

L'analyse des écarts de distribution révèle que les accidents survenant sur des routes en pente (montée ou descente) présentent une sur-représentation des cas graves (grav=1), avec un coefficient de 1.30. Cette observation peut s'expliquer par la difficulté accrue de freinage en descente et la perte de contrôle plus fréquente en montée, augmentant ainsi la gravité des impacts.

Les routes plates, en revanche, montrent une répartition équilibrée entre les accidents graves et non graves, avec un coefficient de 0.93 pour les cas graves. Cela reflète une stabilité accrue et une meilleure maîtrise du véhicule sur ce type de chaussée.

V. Plan - Tracé de la route

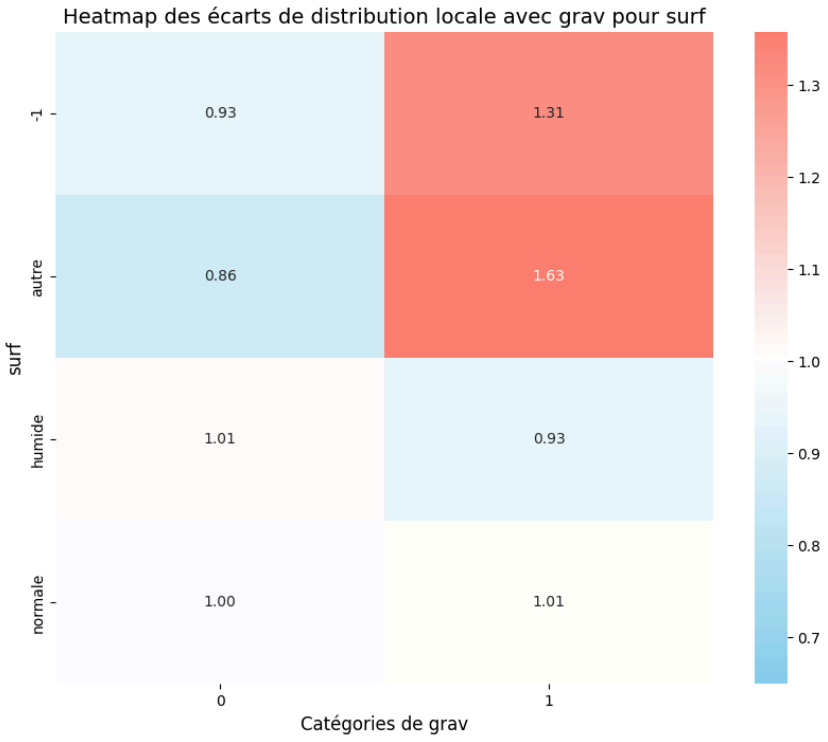
Cette variable, nommée plan, joue un rôle crucial dans la dynamique des accidents en influençant la visibilité, les trajectoires des véhicules et les marges de manœuvre disponibles pour les conducteurs.

L'analyse des écarts de distribution révèle que les accidents survenant sur des routes en courbe présentent une sur-représentation marquée des cas graves (grav=1), avec un coefficient de 1.55. Cette observation peut être attribuée aux pertes de contrôle plus fréquentes, notamment à vitesse élevée, ainsi qu'à la réduction des distances de visibilité qui limite les possibilités d'anticipation.

Les accidents sur des routes rectilignes, en revanche, montrent une distribution plus équilibrée, avec un coefficient de 0.88 pour les cas graves. Ce résultat peut s'expliquer par le fait que les lignes droites permettent une meilleure anticipation des obstacles et des autres usagers, réduisant ainsi la sévérité des impacts.

Ces résultats confirment que la variable plan est pertinente pour la prédiction de la gravité des accidents. Les routes en courbe, associées à un risque accru de perte de contrôle et de collisions sévères, devraient être prises en compte de manière spécifique dans le modèle.

W. Surf - Etat de surface



Cette variable, nommée surf, joue un rôle fondamental dans l'adhérence des véhicules et leur capacité à réagir efficacement face aux imprévus.

L'analyse des écarts de distribution révèle que les accidents survenant sur des routes présentant un état dégradé ou atypique (autre) présentent une sur-représentation significative des cas graves (grav=1), avec un coefficient de 1.63. Cette observation peut être

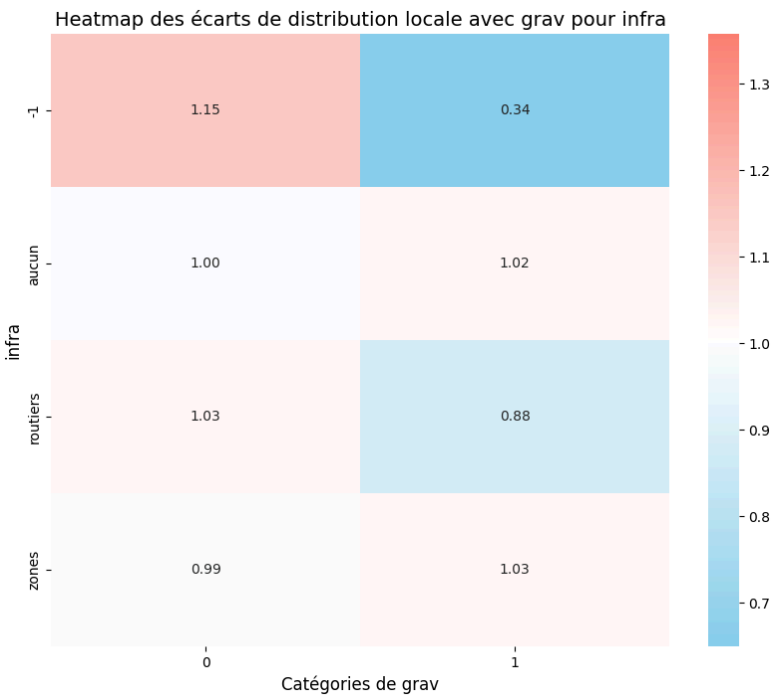
attribuée à des conditions d'adhérence imprévisibles ou à des infrastructures défectueuses pouvant causer des pertes de contrôle.

Les accidents survenant sur des routes humides montrent une distribution relativement équilibrée, avec un coefficient de 0.93 pour les cas graves. Bien que l'adhérence soit réduite, il est possible que les conducteurs adoptent une conduite plus prudente, compensant ainsi le risque accru.

Les routes normales présentent une distribution neutre, avec un coefficient de 1.01 pour les cas graves, ce qui reflète la situation standard où l'état de la surface n'influence ni positivement ni négativement la gravité des impacts.

X. Infra - Aménagement

Nous nous concentrons ici sur la présence d'aménagements routiers.

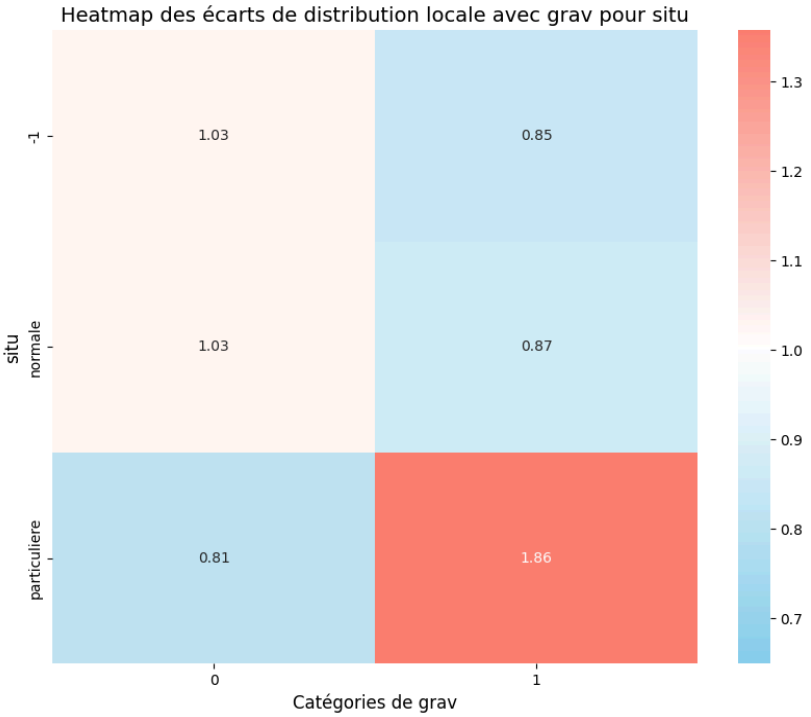


L'analyse des écarts de distribution révèle que les accidents survenant sur des infrastructures classées routiers, comprenant des éléments tels que les tunnels, ponts, bretelles d'échangeurs et carrefours aménagés, présentent une sous-représentation des cas graves (grav=1), avec un coefficient de 0.88. Cela peut être attribué à une meilleure régulation du trafic et à des infrastructures conçues pour réduire les impacts graves.

En revanche, les accidents survenant dans des zones spécifiques, comprenant des passages à niveau, zones piétonnes, péages, chantiers et autres configurations particulières, montrent une répartition légèrement plus élevée des cas graves (grav=1), avec un coefficient de 1.03. Ces environnements présentent des risques accrus liés à l'interaction avec d'autres usagers vulnérables ou des infrastructures temporaires.

Les accidents sur des routes sans aménagements spécifiques présentent une répartition équilibrée, avec un coefficient de 1.02 pour les cas graves, ce qui reflète une situation standard où l'aménagement ne joue ni un rôle protecteur ni aggravant.

Z. Situ - Situation de l'accident



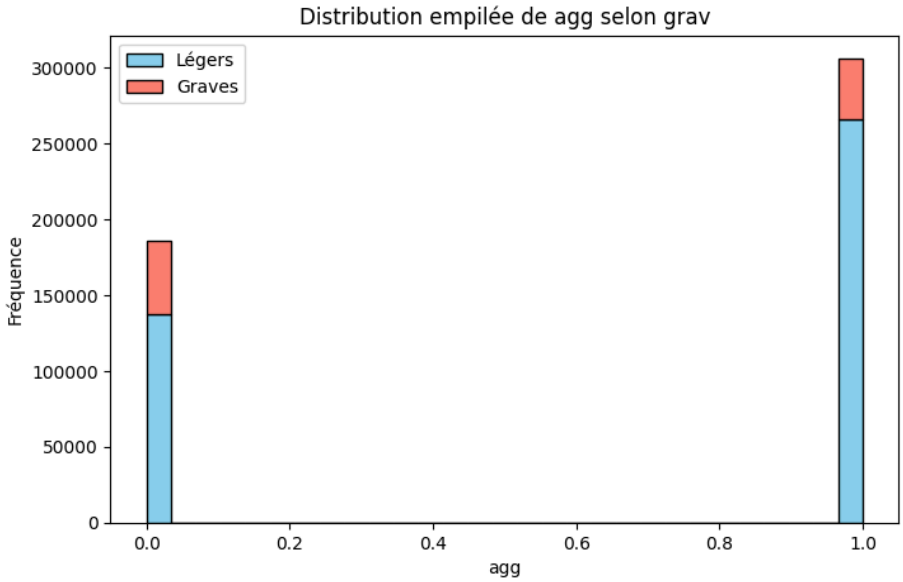
Cette variable, nommée situ, décrit l'emplacement exact où l'accident s'est produit, ce qui peut jouer un rôle important dans l'intensité et les conséquences des collisions.

L'analyse des écarts de distribution révèle que les accidents survenant dans des situations particulières, telles que les bandes d'arrêt d'urgence, accotements, trottoirs, pistes cyclables ou autres voies spéciales, présentent une sur-représentation marquée des cas graves (grav=1), avec un coefficient de 1.86. Cette observation peut s'expliquer par la vulnérabilité accrue des usagers impliqués dans ces situations et par l'éventuelle présence d'obstacles fixes.

En revanche, les accidents survenant dans une situation normale présentent une distribution plus équilibrée, avec un coefficient de 0.87 pour les cas graves. Cela suggère que les infrastructures standards offrent un niveau de sécurité plus homogène, réduisant les impacts les plus sévères.

ZA. Agg - Agglomération (BINAIRE)

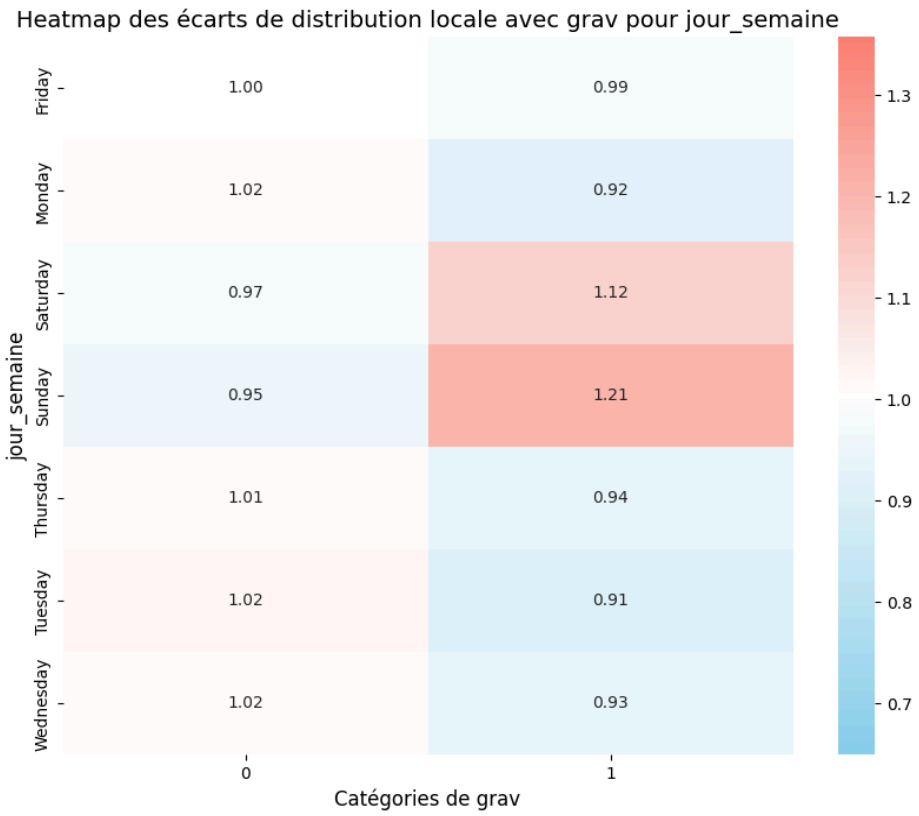
La localisation d'un accident en agglomération ou hors agglomération (agg) constitue un paramètre fondamental pour comprendre la gravité des accidents de la route.



L'analyse des distributions révèle une disparité significative entre les accidents survenant en agglomération et ceux survenant hors agglomération. Les accidents hors agglomération, bien que moins fréquents en volume, présentent une proportion plus élevée de blessures graves. Ce phénomène peut être attribué à plusieurs facteurs, notamment des vitesses autorisées plus élevées, une infrastructure routière différente, et une prise en charge médicale potentiellement plus tardive en raison d'un éloignement des centres de secours.

À l'inverse, les accidents en agglomération sont plus nombreux en raison d'une densité de trafic plus importante. Cependant, la proportion d'accidents graves y est relativement plus faible, ce qui peut s'expliquer par des vitesses généralement plus réduites, la présence de régulations spécifiques comme les zones à vitesse limitée, ainsi que l'influence d'infrastructures conçues pour minimiser la gravité des impacts.

ZB. Jour_semaine



Il est essentiel d'examiner l'impact du jour de la semaine sur la distribution des accidents graves et légers. Cette analyse permet d'évaluer les variations hebdomadaires et de déterminer si certains jours sont plus propices aux accidents graves.

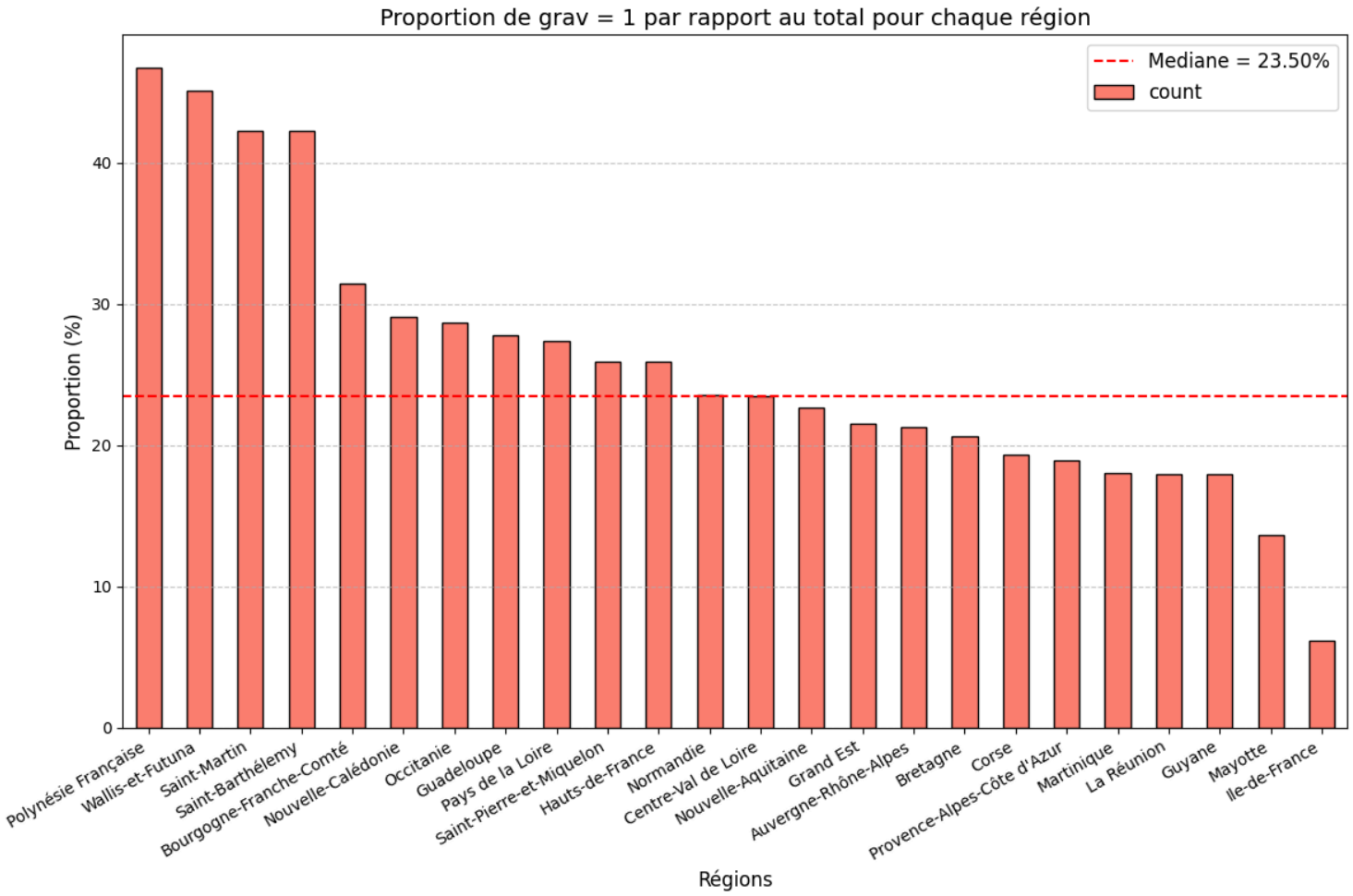
L'analyse met en évidence des variations intéressantes selon le jour de la semaine. Le week-end, comprenant le samedi et le dimanche, présente une surreprésentation des accidents graves, avec un écart particulièrement marqué le dimanche (coefficient de 1,21). Cela suggère que les accidents survenant ce jour-là sont plus fréquemment graves par rapport à la moyenne.

En semaine, du lundi au vendredi, les écarts sont plus proches de la moyenne,

indiquant une légère sous-représentation des accidents graves les mardis, mercredis et jeudis (coefficients autour de 0,91-0,94). Ces jours apparaissent ainsi comme relativement plus sûrs en termes de gravité des accidents.

Ces résultats confirment des tendances bien établies en sécurité routière. Le week-end est souvent associé à une augmentation des accidents graves, probablement en raison de comportements plus risqués comme la conduite sous influence et la vitesse excessive, ainsi que de trajets moins encadrés par des impératifs professionnels.

ZC. Reg - Régions



Il est crucial d'examiner la répartition régionale des accidents graves. Cette analyse permet d'identifier les disparités géographiques et d'évaluer si certaines régions sont plus exposées aux accidents graves que d'autres.

Nous avons comparé la proportion d'accidents graves (grav = 1) par rapport au total des accidents dans chaque région. Une médiane nationale a été calculée pour servir de point de référence et faciliter l'interprétation des écarts régionaux.

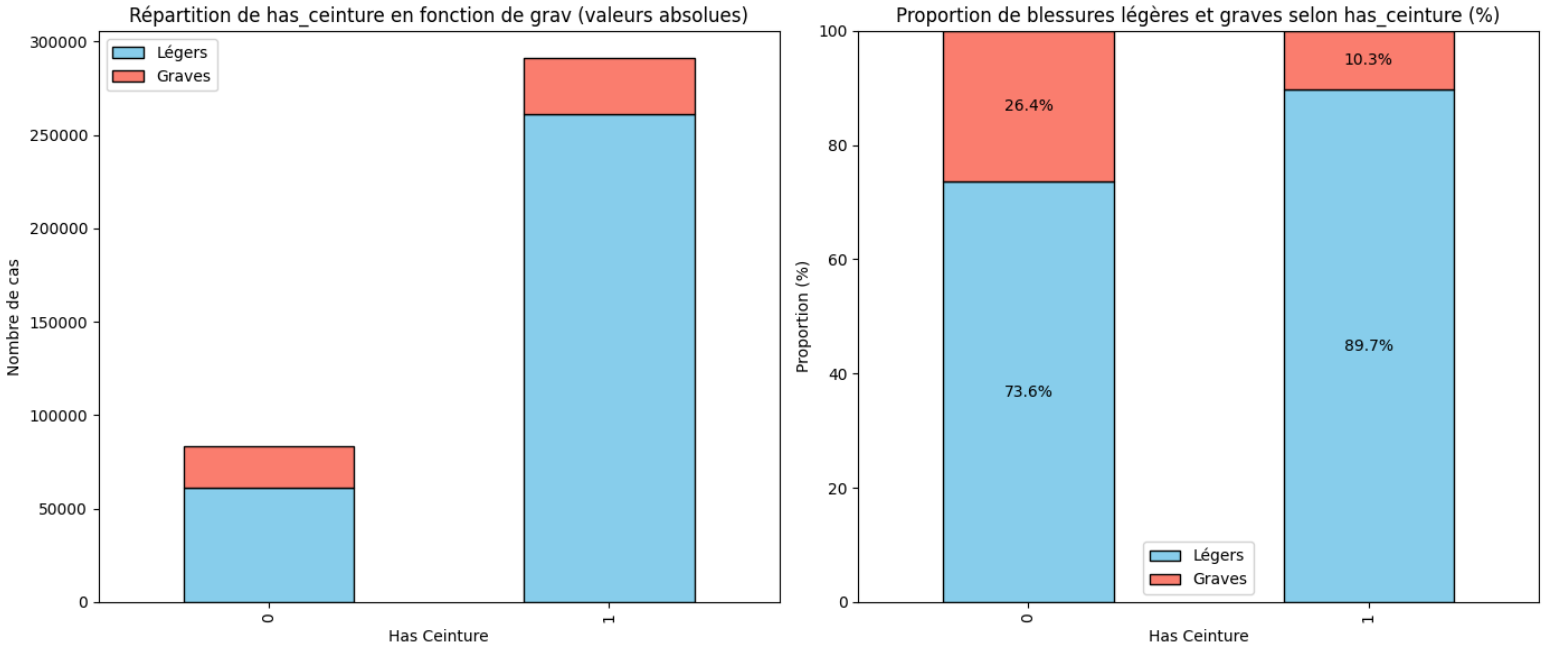
L'analyse révèle des disparités significatives entre les régions. Certaines présentent une proportion nettement plus élevée d'accidents graves, notamment la Polynésie Française, Wallis-et-Futuna, et Saint-Martin, où la proportion dépasse 40 %. Ces valeurs suggèrent un risque accru de gravité des accidents dans ces territoires.

Par ailleurs, l'analyse des chiffres en Polynésie Française peut être éclairée par le Bilan 2023 des chiffres de la sécurité routière en Polynésie française, publié par le Haut-Commissariat de la République en Polynésie française. Ce rapport, bien que postérieur à notre champ d'étude, met en avant des facteurs clés contribuant à la gravité des accidents dans cette région, notamment la consommation d'alcool et de stupéfiants ainsi que la vitesse excessive. Il souligne également que plus de 70 % des tués dans cette zone géographique sont des usagers de deux-roues, ce qui pourrait expliquer la proportion élevée d'accidents graves observée.

À l'inverse, l'Île-de-France affiche la plus faible proportion d'accidents graves, en dessous de 15 %. Ce résultat peut être influencé par une densité de trafic élevée, où les accidents, bien que fréquents, sont souvent de moindre gravité en raison de vitesses moyennes plus faibles.

Les régions métropolitaines comme la Bourgogne-Franche-Comté, l'Occitanie et la Normandie présentent des proportions légèrement supérieures à la médiane nationale, ce qui pourrait être lié à des infrastructures routières spécifiques ou des comportements de conduite distincts.

ZD. Has_ceinture

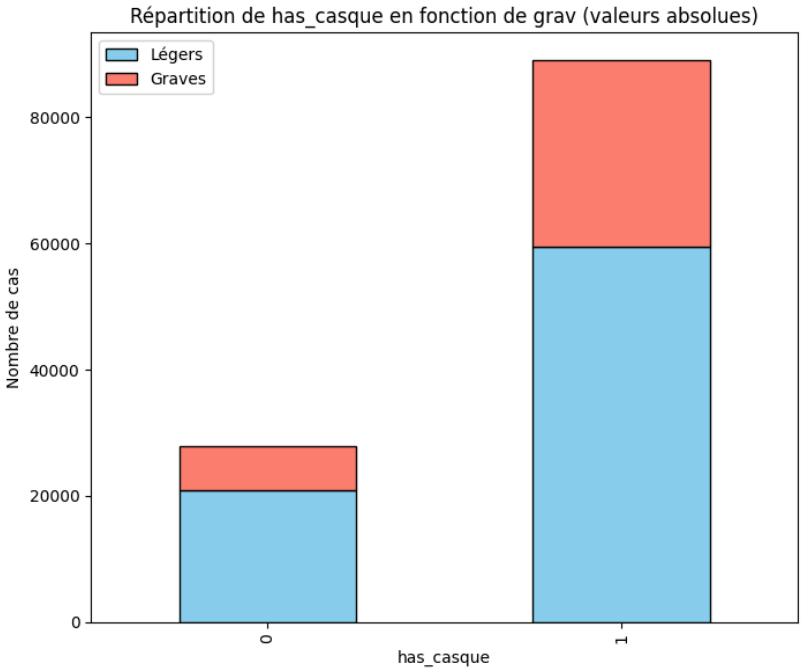


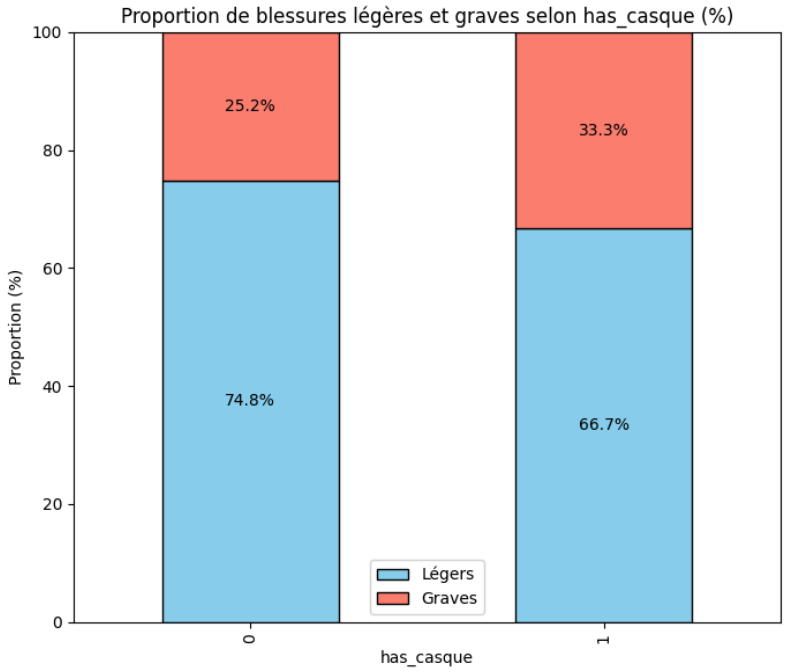
Nous nous intéressons à la relation entre le port de la ceinture de sécurité (has_ceinture) et la gravité des blessures subies, en nous concentrant uniquement sur les véhicules de tourisme et les poids lourds.

Les données révèlent une différence notable entre les usagers de ces catégories de véhicules qui portent une ceinture et ceux qui ne la portent pas. Les accidents impliquant des personnes ne portant pas de ceinture présentent une proportion plus élevée de blessures graves. Le graphique montre clairement que la part des blessures graves est significativement plus importante pour les individus n'ayant pas attaché leur ceinture.

À l'inverse, lorsque la ceinture est portée, la proportion d'accidents entraînant des blessures graves diminue nettement. Ce constat souligne le rôle crucial de la ceinture de sécurité dans la protection des occupants des véhicules de tourisme et des poids lourds.

ZE. Has_casque



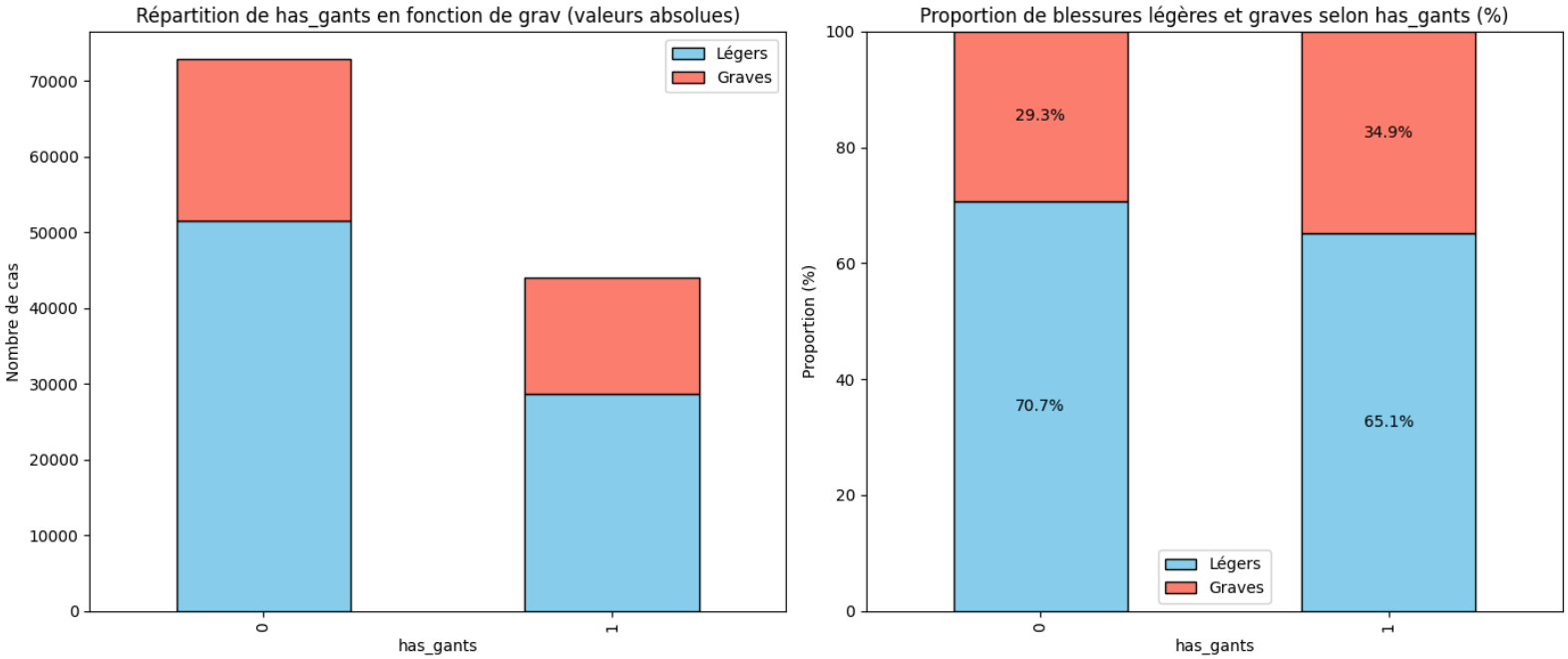


Nous nous intéressons à la relation entre le port du casque (has_casque) et la gravité des blessures subies, en nous concentrant uniquement sur les usagers des deux-roues motorisés et des engins de déplacement personnel (EDP).

Le port du casque est un comportement largement observé parmi les usagers des deux-roues motorisés et des engins de déplacement personnel (EDP). Cette adoption est notamment liée à la réglementation qui impose le port du casque pour les usagers des deux-roues motorisés. En revanche, dans le cas des EDP, les cadres réglementaires émergents à l'époque du champ d'étude pourraient expliquer des comportements d'adoption variables.

Il est intéressant de noter que les usagers portant un casque présentent une proportion plus élevée de blessures graves (33,3 %) par rapport à ceux qui ne le portent pas (25,2 %). Cependant, il est fondamental de ne pas confondre corrélation et causalité. Ces résultats ne suggèrent pas que le port du casque augmente intrinsèquement le risque de blessures graves. Au contraire, ils révèlent probablement des facteurs sous-jacents, comme une plus grande exposition à des accidents à haute énergie, plus fréquents parmi les usagers respectant cette mesure de sécurité. Ces facteurs doivent être intégrés dans une analyse plus large pour éviter des conclusions erronées.

ZF. Has_gants



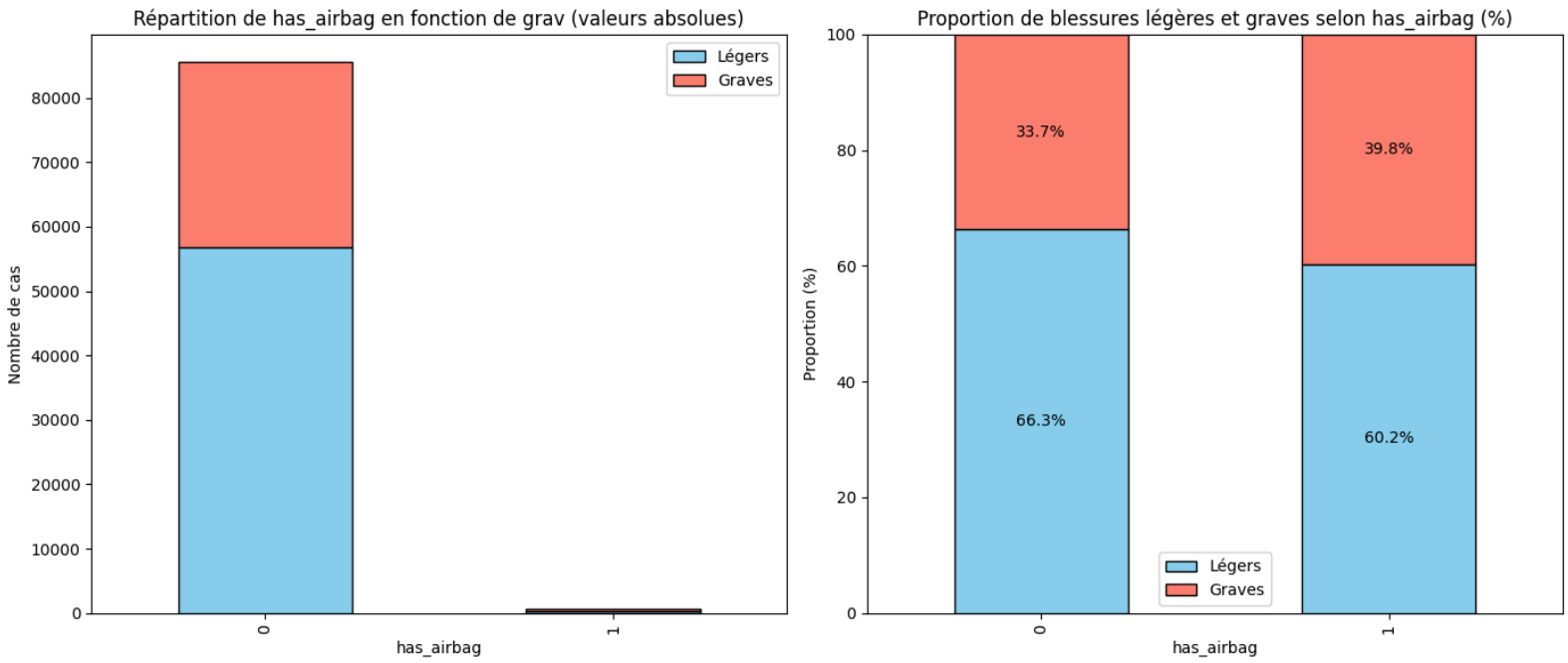
Nous nous intéressons à la relation entre le port des gants (has_gants) et la gravité des blessures subies, en nous concentrant uniquement sur les usagers des deux-roues motorisés et des engins de déplacement personnel (EDP).

Le port des gants, bien que moins systématique que celui du casque, demeure une pratique essentielle, particulièrement parmi les usagers des deux-roues motorisés. Chez les utilisateurs d'EDP, cette pratique est nettement moins répandue, en raison de l'absence d'obligation réglementaire. En revanche, les deux-roues motorisés sont soumis à une réglementation stricte, rendant le port des gants obligatoire.

Les données révèlent une proportion plus élevée de blessures graves parmi les usagers portant des gants (34,9 %) par rapport à ceux qui n'en portent pas (29,3 %). Cette observation ne doit pas être interprétée comme une indication que le port des gants accroît le risque de blessures graves. Au contraire, ces chiffres reflètent probablement des biais liés à des facteurs contextuels, tels que la vitesse ou la nature des accidents rencontrés par les usagers équipés. Les gants, en tant qu'élément de sécurité, jouent un rôle crucial dans la réduction de certaines blessures, bien qu'ils ne puissent offrir une protection complète face à l'ensemble des traumatismes potentiels. En effet, pour les usagers sans carrosserie, l'équipement de sécurité réduit les risques, mais reste insuffisant pour prévenir l'intégralité des blessures lors d'un accident.

De plus, la qualité des gants, leur ajustement et les circonstances spécifiques de l'accident influencent fortement leur efficacité. Enfin, l'adoption des gants est possiblement plus fréquente parmi les usagers particulièrement exposés à des conditions à risque, ce qui pourrait en partie expliquer la prévalence plus importante de blessures graves dans ce groupe.

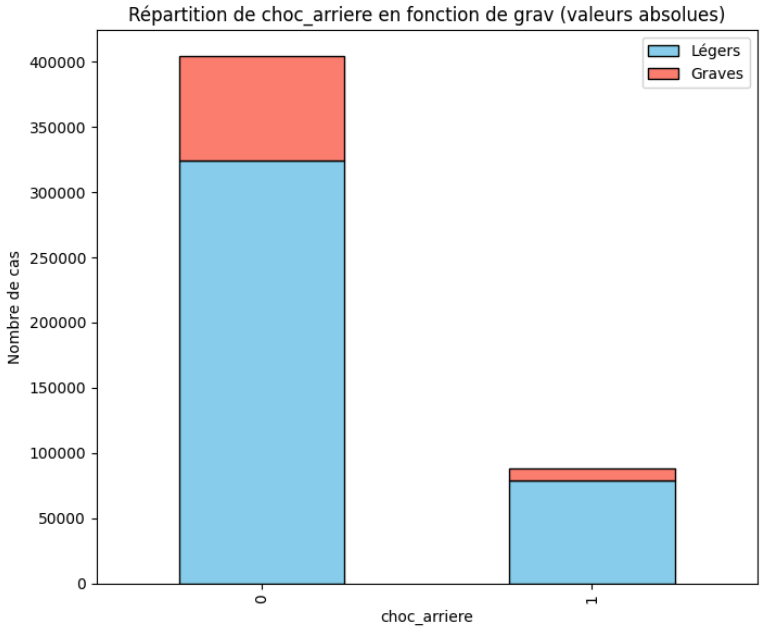
ZG. Has_airbag - airbag 2 roues

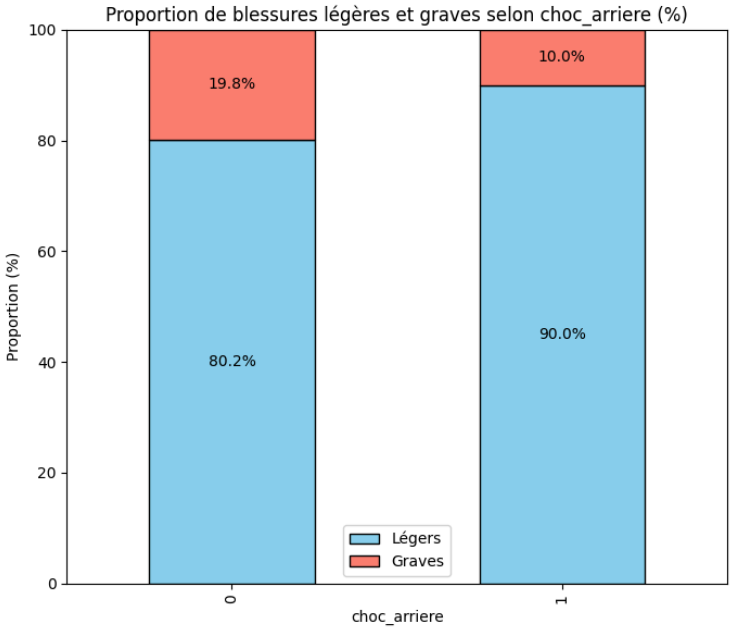


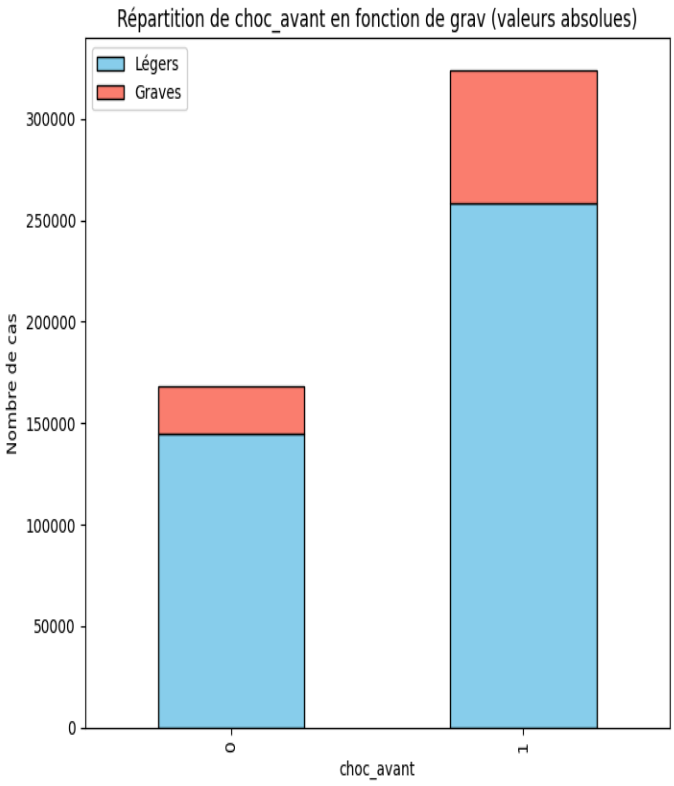
Nous nous intéressons à la relation entre le port de l'airbag (has_airbag) et la gravité des blessures subies, en nous concentrant uniquement sur les usagers des deux-roues motorisés.

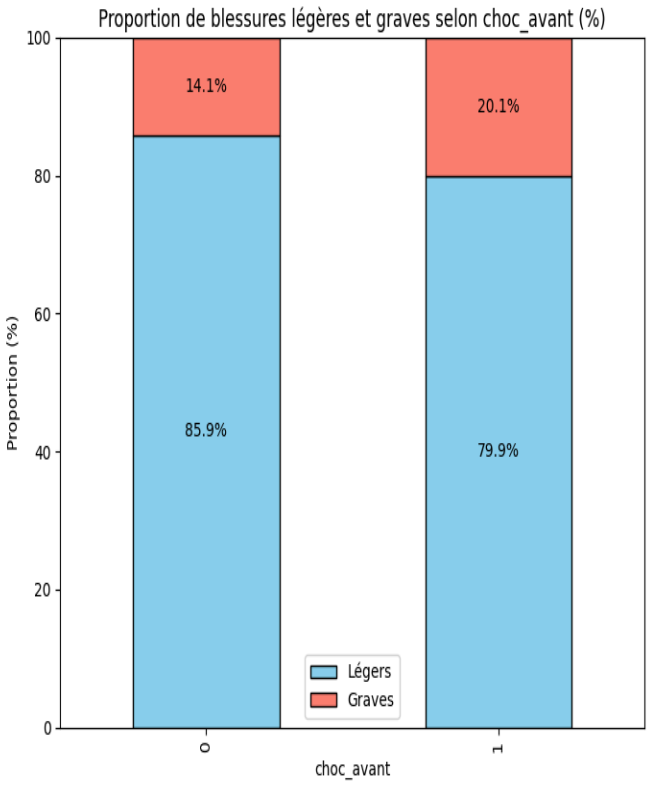
Le port de l'airbag pour les usagers des deux-roues motorisés reste une pratique marginale, comme en témoignent les données disponibles. Malgré son faible taux d'adoption, il est essentiel d'évaluer son impact sur la gravité des blessures. Les résultats montrent une proportion légèrement plus élevée de blessures graves parmi les usagers équipés d'un airbag (39,8 %) par rapport à ceux qui n'en portent pas (33,7 %). Cette différence ne doit cependant pas être interprétée comme une causalité directe entre l'utilisation de l'airbag et un risque accru de blessures graves.

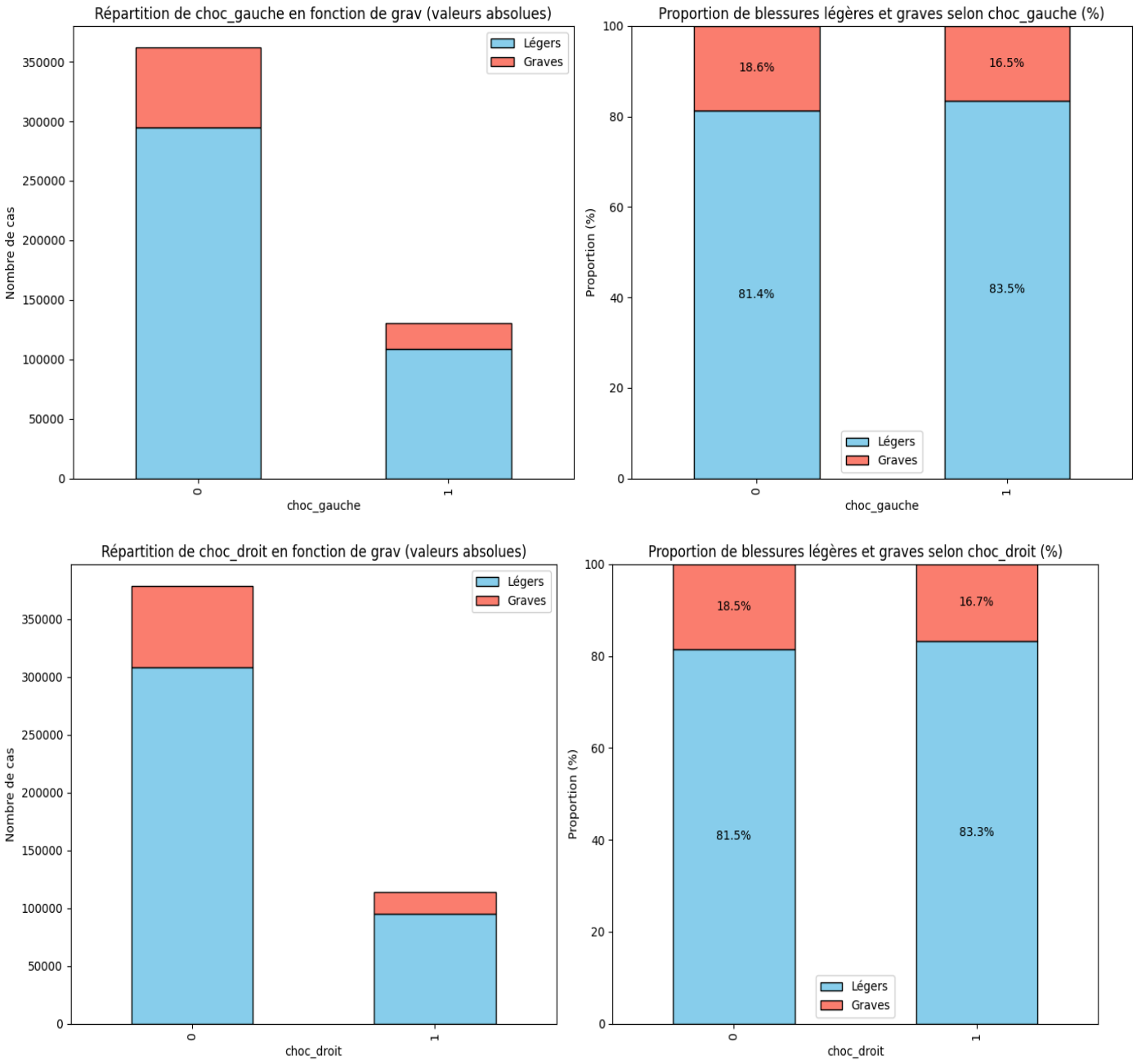
ZH. Choc_avant, choc_arriere, choc_gauche et choc_droit











Nous nous intéressons à la relation entre les points d'impact du véhicule (choc avant, arrière, gauche et droit) et la gravité des blessures subies.

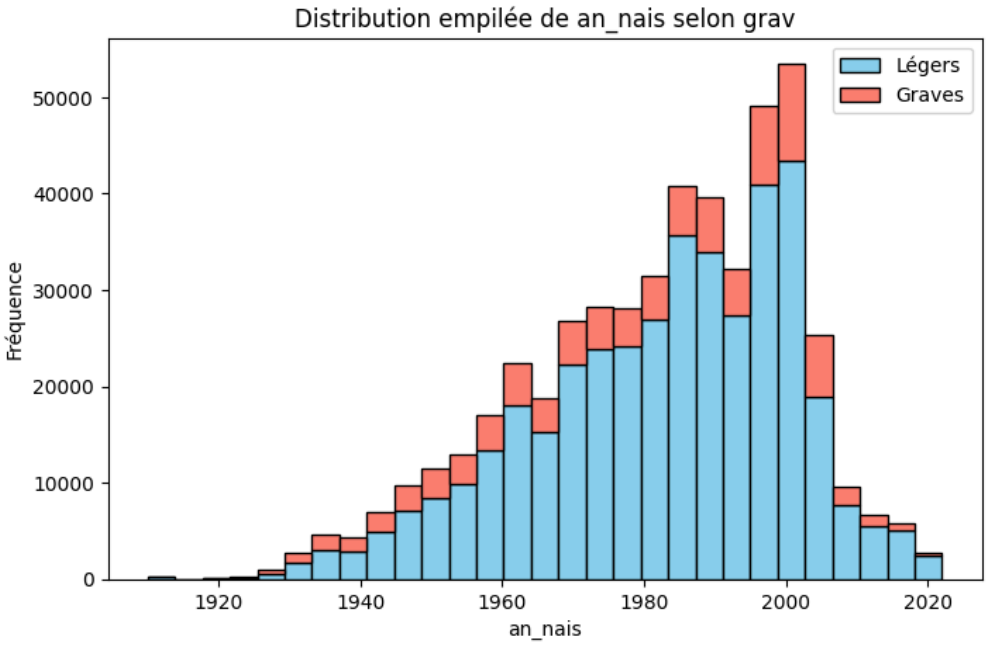
Les données révèlent des différences marquées dans la gravité des blessures en fonction des points d'impact. Lors des collisions frontales, le choc avant est associé à une proportion relativement élevée de blessures graves (20,1 %) par rapport à l'absence de choc avant (14,1 %). Cette tendance s'explique par la force importante généralement concentrée à l'avant du véhicule lors de ces impacts.

Les chocs arrière, quant à eux, montrent une proportion de blessures graves plus faible (10,0 %) en comparaison avec les accidents sans choc arrière (19,8 %). Ce phénomène peut être lié à la protection accrue offerte par la structure arrière du véhicule et aux caractéristiques moins sévères des collisions par l'arrière.

En ce qui concerne les impacts latéraux, les collisions sur le côté gauche entraînent une proportion de blessures graves de 16,5 %, légèrement inférieure à celle des accidents sans choc gauche (18,6 %). Cette différence pourrait s'expliquer par la position du conducteur et la configuration de l'accident. Du côté droit, les proportions sont similaires, avec 16,7 % de blessures graves en cas de choc, contre 18,5 % en l'absence de collision sur ce côté. Les différences entre les chocs gauche et droit pourraient être influencées par la structure asymétrique du véhicule et la répartition des passagers.

Variables quantitatives

A. An_nais - Année de naissance



L'année de naissance des usagers impliqués dans un accident (an_nais) est un facteur clé dans l'étude de la gravité des accidents de la route.

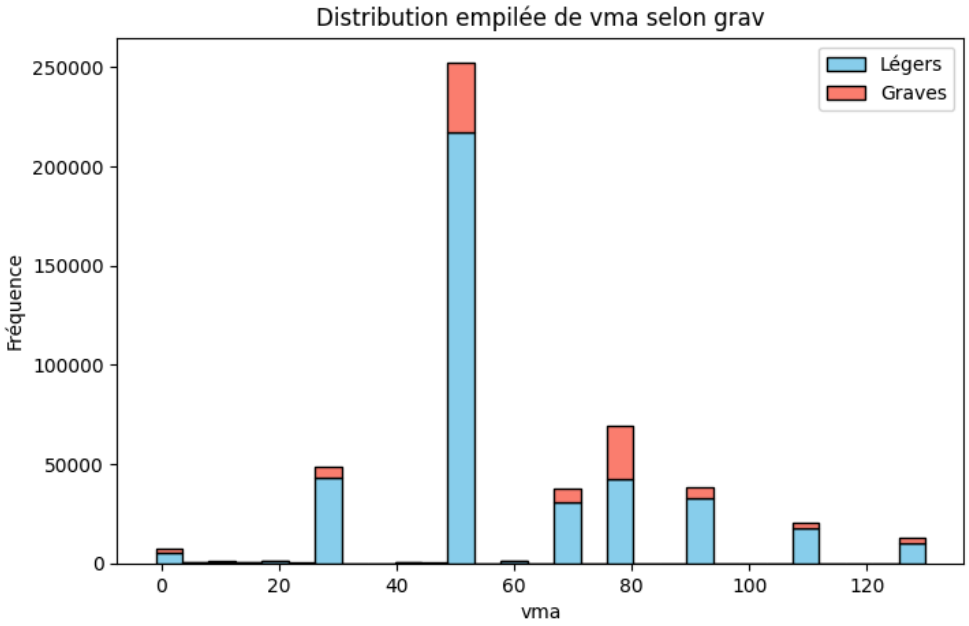
L'analyse des distributions met en évidence plusieurs tendances intéressantes. On observe une concentration accrue des accidents parmi les usagers nés après les années 1980, ce qui reflète la structure démographique des conducteurs en circulation. Parmi ces accidents, les

blessures graves apparaissent plus fréquentes dans deux tranches d'âge distinctes. D'une part, les jeunes conducteurs, nés après 2000, affichent une proportion plus importante de blessures graves, ce qui peut être attribué à un manque d'expérience, une prise de risque plus marquée ou une conduite plus dynamique. D'autre part, les usagers plus âgés, nés avant 1950, présentent également une probabilité accrue d'accidents graves, probablement en raison d'une fragilité physiologique plus importante et d'un temps de réaction plus long.

Dans l'ensemble, la répartition des accidents suit une distribution cohérente avec l'évolution démographique, avec un pic d'accidents parmi les conducteurs nés entre 1970 et 2000.

L'inclusion de la variable an_nais dans le modèle de prédiction de la gravité des accidents permettrait de mieux comprendre l'influence de l'âge des usagers, bien que cette variable encapsule également des effets générationnels. Toutefois, il est probablement plus pertinent de ne conserver que l'âge, qui constitue un facteur plus directement lié à la vulnérabilité et à l'expérience de conduite.

B. Vma - Vitesse maximale autorisée



La vitesse maximale autorisée (vma) sur le lieu d'un accident est un facteur déterminant dans la gravité des blessures subies par les usagers impliqués.

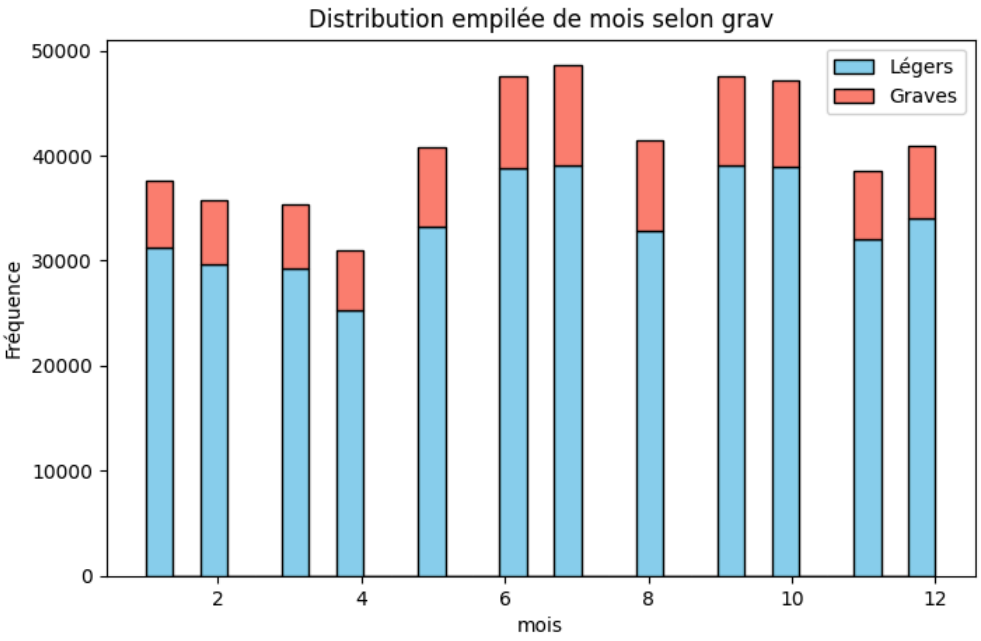
L'analyse des distributions montre que la grande majorité des accidents se produisent dans des zones où la vitesse maximale autorisée est de 50 km/h, correspondant aux environnements urbains. Cependant, les accidents graves sont proportionnellement plus fréquents sur les routes où la vitesse est plus élevée, notamment à 80 km/h et au-delà. Cela s'explique par l'augmentation de l'énergie cinétique en cas de collision, réduisant ainsi les chances de survie et augmentant la

sévérité des blessures.

On observe également une fréquence notable d'accidents sur des routes limitées à 30 km/h, bien que la proportion d'accidents graves y soit relativement plus faible. Ce phénomène peut être lié aux zones à forte densité piétonne où les collisions, bien que plus fréquentes, sont moins létales.

L'intégration de vma dans le modèle de prédiction de la gravité des accidents est essentielle pour capturer l'influence de la vitesse sur la dangerosité des collisions.

C. Mois



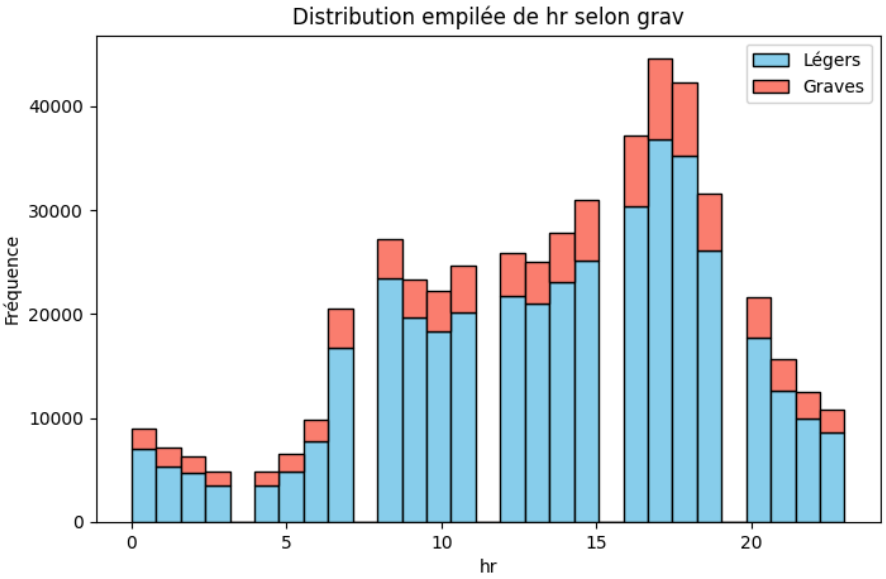
En comparant la distribution globale de la gravité des accidents (grav) avec sa distribution locale en fonction du mois de l'année (mois), nous pouvons identifier d'éventuelles variations saisonnières influençant la probabilité d'accidents graves.

L'analyse des distributions révèle une variation notable de la fréquence des accidents en fonction des mois. On observe une augmentation des accidents en période estivale, avec un pic en juin et juillet, mois durant lesquels la circulation routière est généralement plus dense en raison des départs en vacances et des conditions météorologiques favorables aux déplacements. Cette augmentation s'accompagne d'une proportion accrue d'accidents graves, ce qui pourrait être lié à une plus grande présence de véhicules sur les routes rapides et une augmentation de la vitesse moyenne pratiquée.

En revanche, les mois hivernaux présentent une fréquence d'accidents plus modérée, bien que la proportion d'accidents graves y demeure significative. Les conditions climatiques difficiles, telles que la neige et le verglas, ainsi que la luminosité réduite peuvent contribuer à la sévérité des accidents, malgré une réduction potentielle

du volume de circulation.

D. Hr - Heure

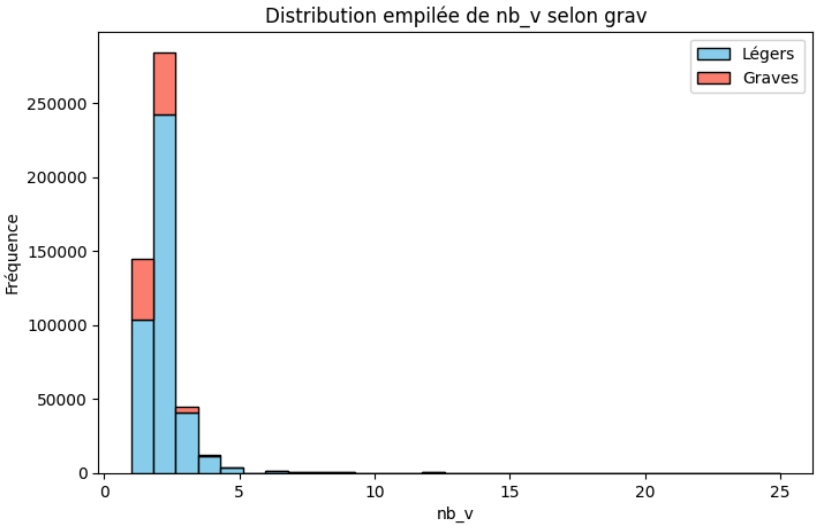


L'analyse de la distribution des accidents selon l'heure de la journée met en évidence des variations marquées en fonction des plages horaires. On observe notamment deux pics distincts, l'un en début de journée vers 8 heures et un autre plus prononcé en fin d'après-midi, autour de 17 heures. Ces tendances correspondent aux périodes de forte circulation associées aux trajets domicile-travail et domicile-école.

La proportion d'accidents graves suit généralement la tendance des accidents légers, mais elle semble particulièrement accentuée en fin de journée. Cela pourrait s'expliquer par une fatigue accrue des conducteurs, des conditions de circulation plus denses et une prise de risque potentiellement plus élevée.

De manière intéressante, la nuit et les premières heures du matin (0h - 5h) présentent une fréquence d'accidents plus faible en valeur absolue, mais une proportion relative d'accidents graves plus importante. Cela peut être lié à une vitesse plus élevée des véhicules en circulation, une visibilité réduite et une potentielle consommation d'alcool ou de stupéfiants chez certains conducteurs.

E. Nb_v - Nombre de véhicules impliqués



Nous nous intéressons à la relation entre le nombre de véhicules impliqués (nb_v) et la gravité des blessures subies par les personnes concernées.

La majorité des accidents impliquent un ou deux véhicules, ce qui se reflète dans la forte concentration des barres sur ces

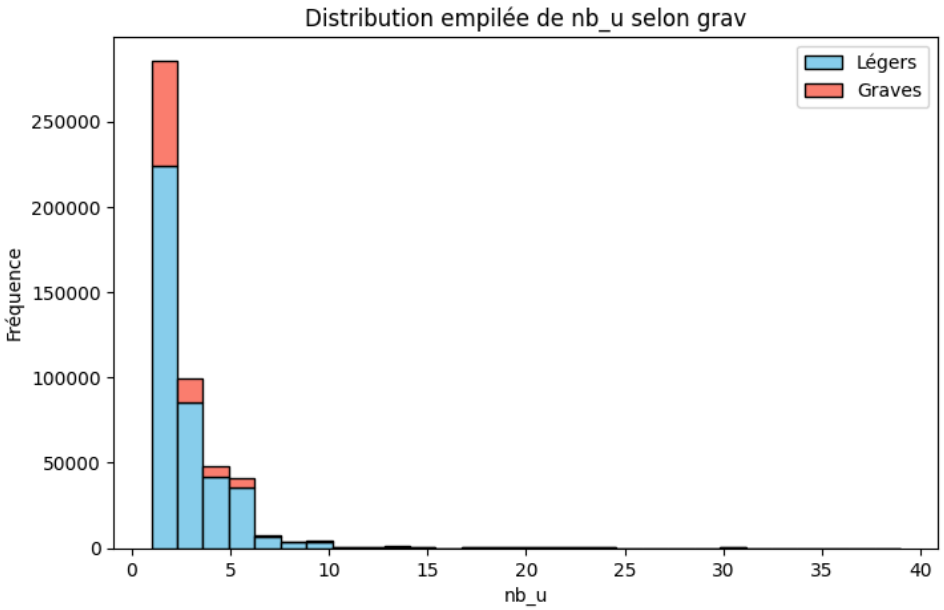
valeurs. Cependant, les données du tableau indiquent que la proportion d'accidents graves tend à diminuer avec l'augmentation du nombre de véhicules impliqués. Pour les accidents avec un seul véhicule, environ 28,6 % des cas sont graves. Cette proportion diminue progressivement, atteignant environ 6,1 % pour six véhicules impliqués.

Toutefois, une légère remontée du pourcentage d'accidents graves est observée à partir de sept véhicules impliqués. Par exemple, pour neuf véhicules impliqués, 8,8 % des cas sont graves, alors que pour 10 véhicules, aucun cas grave n'a été enregistré dans l'échantillon analysé. Ces fluctuations peuvent être dues à la faible fréquence des accidents avec un grand nombre de véhicules.

Lorsque le nombre de véhicules dépasse cinq, la fréquence des accidents diminue drastiquement. Toutefois, même dans ces configurations rares, on constate que des accidents graves peuvent encore survenir.

Les données montrent une tendance générale à la diminution du pourcentage d'accidents graves à mesure que le nombre de véhicules augmente, bien que des variations existent.

F. Nb_u - Nombre d'usagers impliqués



Nous nous intéressons à la relation entre le nombre d'usagers impliqués (nb_u) et la gravité des blessures subies.

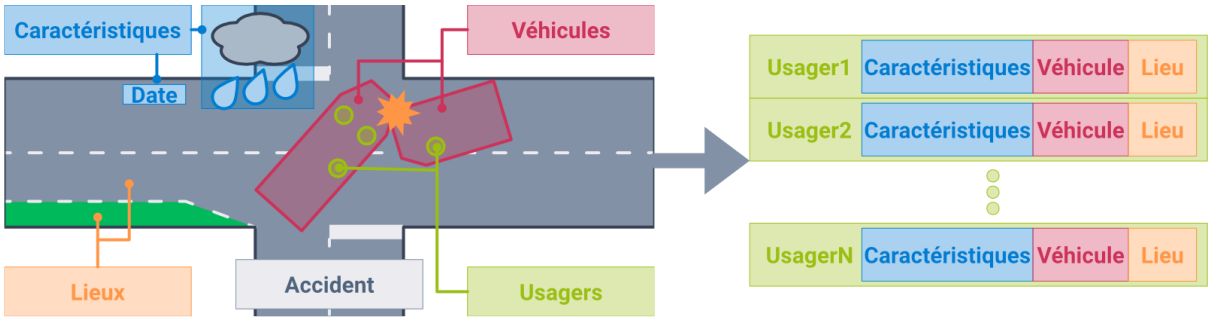
Les données révèlent que les accidents impliquant un seul usager sont les plus fréquents, mais ils présentent également une proportion élevée de blessures graves (environ 47,4 %). Dès que le nombre d'usagers augmente, cette proportion diminue significativement, atteignant environ 14,4 % pour trois usagers et restant aux alentours de 10-12 % jusqu'à six usagers impliqués.

À partir de sept usagers, la part des blessures graves continue de diminuer,

stabilisant autour de 8-10 % avec quelques fluctuations. Pour les configurations avec plus de dix usagers, la proportion d'accidents graves devient marginale, tombant sous la barre des 5 % dans la plupart des cas. Toutefois, le faible nombre de cas dans ces catégories impose une certaine prudence dans l'interprétation des résultats.

Prétraitement

Fonctionnement général



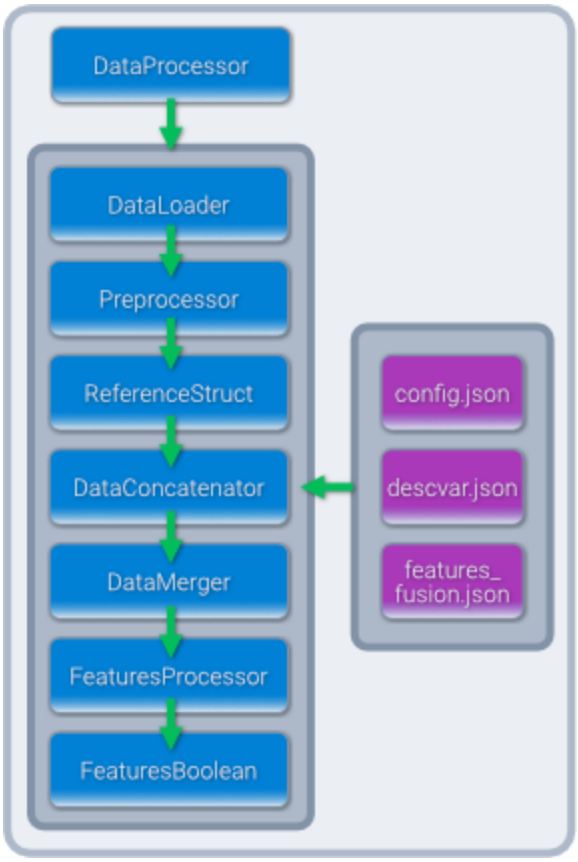
Objectif du prétraitement: fusionner les données pour avoir une liste d'usagers capturant le contexte de l'accident.

Le script data_processing.py constitue la colonne vertébrale du prétraitement initial des données. Il orchestre l'ensemble des étapes permettant de charger, nettoyer, structurer, fusionner et sauvegarder les données avant leur exploitation par un modèle de machine learning. Ce pipeline modulaire assure un traitement automatisé et reproductible des jeux de données bruts.

Le traitement des données commence par le chargement des jeux de données en parallèle via la classe DataLoader, en fonction des préfixes et années définis dans la configuration. Cette approche permet une ingestion rapide et efficace des fichiers nécessaires. Une fois les données chargées, elles subissent un prétraitement via la classe Preprocessor, qui applique diverses transformations pour harmoniser les formats, traiter les valeurs manquantes et assurer une meilleure qualité des données.

L'un des éléments clés du pipeline est l'uniformisation des structures de données. La classe ReferenceStructureExtractor est chargée d'extraire une structure de référence et de l'appliquer à tous les jeux de données afin de garantir une homogénéité dans les formats et colonnes utilisées. Ce processus est essentiel pour éviter toute incohérence lors des étapes suivantes.

Une fois cette structure appliquée, les données sont consolidées à travers un processus de concaténation et de fusion. DataConcatenator s'occupe de regrouper les jeux de données selon des critères spécifiques avant que DataMerger ne les fusionne en un ensemble cohérent et complet. Ces étapes permettent d'agréger les informations issues de différentes sources tout en éliminant les redondances éventuelles.



L'amélioration des données est poursuivie avec FeaturesProcessor, qui applique des transformations avancées sur les caractéristiques du jeu de données pour en extraire des informations pertinentes. En complément, FeaturesBoolean crée des variables booléennes dérivées afin d'enrichir la représentation des données. Ces transformations sont essentielles pour préparer le dataset final en vue de la modélisation.

Enfin, avant d'enregistrer les résultats finaux, une étape de traitement des doublons est effectuée avec Preprocessor. Une fois les données nettoyées et consolidées, elles sont sauvegardées sous forme de fichier CSV dans l'emplacement spécifié par la configuration. Tout au long du processus, un système de journalisation permet de suivre les différentes étapes et de mesurer le temps d'exécution pour optimiser les performances.

DataLoader

Le script data_loading.py est un élément fondamental du pipeline de traitement des données. Il est responsable du chargement des fichiers nécessaires aux modules de prétraitement et de modélisation. Ce composant a évolué pour répondre aux contraintes de performance, notamment lorsqu'il était exécuté sur une machine peu puissante. Afin d'optimiser le processus, une approche multi-threadée a été mise en place, permettant de réduire significativement les temps de chargement.

Le DataLoader gère l'ensemble du processus de chargement des fichiers CSV. Il commence par détecter dynamiquement le délimiteur des fichiers pour assurer une lecture correcte des données. Il prend également en charge différents encodages afin

d’éviter les erreurs de décodage lors de l’ouverture des fichiers. En cas d’échec avec un encodage donné, une tentative est effectuée avec un autre encodage défini dans la configuration.

Un aspect clé du DataLoader est la gestion de la structuration des noms de fichiers. Étant donné que leur format a changé selon les années, le script adapte dynamiquement la syntaxe à utiliser pour retrouver les bons fichiers sources. Cette flexibilité garantit une compatibilité avec différentes versions des données historiques.

Initialement, le chargement des fichiers se faisait de manière séquentielle, ce qui entraînait des délais importants, surtout lorsque de nombreux fichiers étaient à traiter. Pour pallier ce problème, une approche basée sur le multi-threading via ThreadPoolExecutor a été intégrée. Cette optimisation permet de charger plusieurs fichiers en parallèle, ce qui réduit drastiquement le temps d’exécution.

L’utilisation du multi-threading est particulièrement efficace ici, car le chargement des fichiers est une tâche principalement liée à l’entrée/sortie (I/O bound). Ainsi, en parallèle, plusieurs fichiers peuvent être ouverts et lus sans être limités par la puissance de calcul du processeur. Cette amélioration a apporté un gain de temps significatif, rendant le pipeline plus efficace et rapide.

Preprocessor

Le script preprocessor.py est conçu pour appliquer des traitements standardisés aux jeux de données avant leur exploitation dans le pipeline. Il assure une certaine flexibilité en fonction du type de fichier traité, ce qui permet d’anticiper d’éventuelles évolutions des formats de données.

Le Preprocessor est structuré pour appliquer des transformations spécifiques aux jeux de données en fonction de leur nature. Il commence par une étape de suppression de colonnes superflues, définies dynamiquement en fonction du type de fichier. Cette approche permet d’éliminer les informations non pertinentes et d’optimiser la qualité des données utilisées dans les étapes ultérieures.

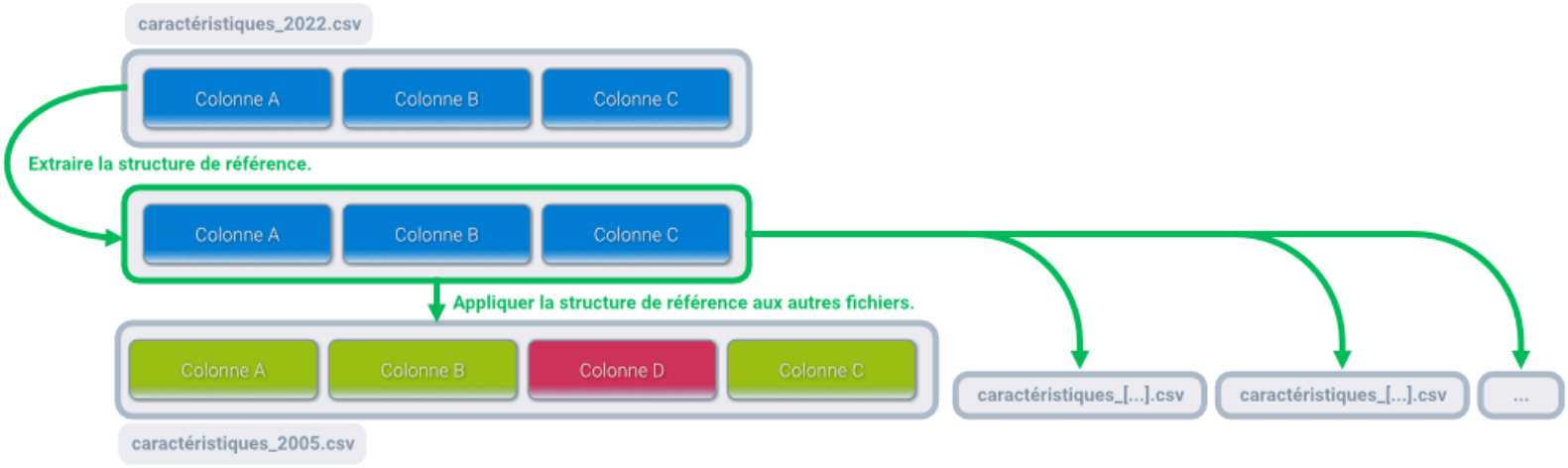
Ensuite, le script applique des modifications spécifiques selon le préfixe des fichiers. Par exemple, pour les fichiers de type "caracteristiques", il renomme certaines colonnes pour assurer une cohérence avec les autres jeux de données. Cette standardisation facilite la fusion et l’analyse des données à plus grande échelle.

Un autre aspect important du Preprocessor est la gestion des doublons. Il intègre une méthode dédiée à la détection et à la suppression des entrées en double, garantissant

ainsi l'intégrité des données finales. Cette étape est cruciale pour éviter des biais dans les analyses et améliorer la fiabilité des modèles prédictifs.

ReferenceStructureExtractor

Le script reference_structure.py est conçu pour extraire et appliquer une structure de référence aux jeux de données. Il garantit une cohérence dans la disposition des colonnes et les types de données entre les différentes sources, facilitant ainsi leur fusion et leur analyse. Cette harmonisation est d'autant plus nécessaire que les fichiers



ont largement évolué au fil des années, rendant leur structuration initiale hétérogène.

Le ReferenceStructureExtractor est chargé d'analyser la dernière version d'un jeu de données pour en extraire une structure de référence, comprenant les noms de colonnes et les types de données associés. Cette structure est ensuite appliquée aux jeux de données futurs afin d'assurer une harmonisation systématique.

L'application de la structure de référence implique la suppression des colonnes non pertinentes et l'ajout de colonnes manquantes, remplies par des valeurs vides. Le script veille également à forcer la conversion des types de données pour correspondre à la structure de référence, tout en gérant les erreurs potentielles liées à ces conversions.

Cette classe joue un rôle essentiel dans l'uniformisation des jeux de données en garantissant leur compatibilité structurelle. Il précède les étapes de concaténation et de fusion des jeux de données, et son exécution est nécessaire pour assurer la réussite de ces étapes futures. Grâce à cette approche, il facilite leur intégration dans le pipeline de traitement et réduit les risques d'erreurs lors de l'analyse et de la modélisation.

DataConcatenator

Le script data_concatenator.py est chargé de l'agrégation des jeux de données issus des différentes années de traitement. Son objectif principal est d'assurer la continuité et la cohérence des données en fusionnant les fichiers tout en appliquant des ajustements spécifiques à certaines colonnes critiques.

Le DataConcatenator prend en charge la concaténation des jeux de données pour chaque catégorie définie par un préfixe. Une fois les fichiers fusionnés, il s'assure que les données restent homogènes et utilisables pour les étapes suivantes du pipeline de traitement.

Une attention particulière est portée aux fichiers de type "usagers", notamment sur la colonne id_usager. Certains fichiers présentent des valeurs manquantes dans cette colonne, ce qui peut poser problème lors des traitements ultérieurs. Pour y remédier, le script convertit d'abord ces valeurs en format numérique, puis attribue des identifiants uniques aux valeurs manquantes afin d'éviter toute incohérence dans les analyses futures.

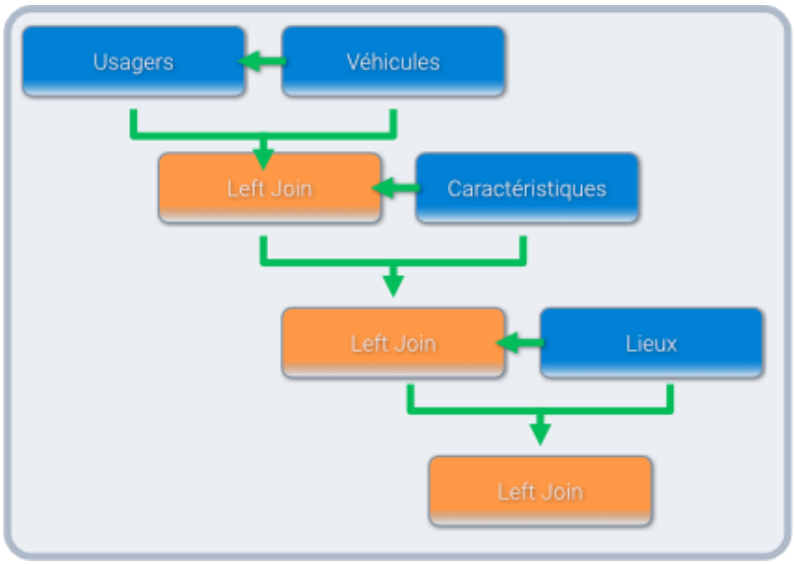
Enfin, après la concaténation et les ajustements nécessaires, les fichiers sont enregistrés dans un format standardisé. L'export des fichiers concaténés est réalisé avec une nomenclature qui reflète la période de traitement, garantissant ainsi une bonne traçabilité des données.

En consolidant les fichiers tout en traitant les incohérences, il prépare efficacement les données pour les phases ultérieures du pipeline, notamment la fusion et la modélisation. Son intégration permet une meilleure qualité des données et assure leur cohérence sur l'ensemble des années analysées.

DataMerger

Le script data_merger.py est chargé de la fusion des différentes sources de données préparées lors des étapes précédentes du pipeline. En combinant les fichiers concaténés des différentes entités (caractéristiques, lieux, usagers et véhicules), il construit une base de données unifiée et exploitable pour la modélisation.

Le DataMerger récupère les fichiers précédemment traités et les fusionne selon des clés communes. Il commence par fusionner les données des usagers et des véhicules, en utilisant les identifiants de véhicule et d'accident. Une fois cette première jointure réalisée, le fichier résultant est enrichi avec les caractéristiques de l'accident, puis avec



les informations liées aux lieux.

L'ensemble du processus est réalisé en respectant une approche par jointures successives (merge en mode left), afin de garantir que toutes les données disponibles soient correctement alignées sans perte d'information critique. À chaque étape, des logs sont générés pour assurer un suivi de l'avancement et permettre une détection rapide d'éventuelles anomalies.

FeaturesProcessor

Le script features_processor.py est dédié à l'amélioration et à la transformation des variables présentes dans le jeu de données final. Il applique un ensemble de traitements visant à harmoniser, restructurer et préparer les caractéristiques pour la phase de modélisation. Son rôle est crucial pour garantir une qualité optimale des données avant leur exploitation.

La classe FeaturesProcessor commence par charger une configuration définissant les modalités de transformation de certaines variables. Il utilise notamment le fichier features_fusion.json, qui permet de regrouper les modalités de certaines variables afin de réduire la dimensionnalité du problème et ainsi optimiser la modélisation. Il s'assure que les colonnes concernées sont converties en chaînes de caractères avant d'appliquer les remplacements de valeurs définis dans le fichier de configuration.

L'une des principales transformations effectuées concerne la gestion des valeurs manquantes et la correction des types de données. Le script remplace certaines valeurs spécifiques par NaN et applique des conversions de type pour garantir l'homogénéité des variables. De plus, il ajuste les identifiants (id_vehicule, id_usager) et calcule des indicateurs tels que le nombre de véhicules (nb_v) et d'usagers (nb_u) impliqués par accident.

D'autres traitements portent sur la création de nouvelles variables, notamment la

variable age, dérivée de l'année de naissance et de l'année de l'accident. Il effectue aussi une correction des dates en combinant les colonnes an, mois et jour, ce qui permet d'obtenir une structure temporelle exploitable.

Enfin, certaines variables sont transformées pour correspondre à des catégories normalisées. Grav, qui constitue la variable cible du modèle, est réorganisée en réduisant ses modalités afin de mieux distinguer les niveaux de gravité des accidents. De même, des ajustements sont appliqués aux vitesses maximales (vma) et aux localisations (dep transformé en reg).

FeaturesProcessor joue un rôle déterminant dans la préparation des données en appliquant des transformations spécifiques pour garantir leur qualité et leur cohérence. Il assure une harmonisation des modalités, bien que le regroupement des valeurs manquantes reste une étape intermédiaire. Ces dernières ne sont pas encore pleinement gérées à ce stade et seront traitées dans un pipeline ultérieur, plus proche de la modélisation finale. De même, l'encodage et la standardisation des variables ne sont pas encore appliqués ici afin de limiter les risques de data leakage. Son intégration dans le pipeline permet ainsi d'assurer une transition progressive et contrôlée vers l'étape de modélisation.

FeaturesBoolean

Le script features_boolean.py est conçu pour générer des variables booléennes à partir de certaines caractéristiques du jeu de données. En transformant des informations catégorielles en indicateurs binaires, il facilite l'interprétation des données et leur utilisation dans des modèles prédictifs.

Le FeaturesBoolean applique des transformations spécifiques à deux types de données principales : les équipements de sécurité et les types de chocs.

La première transformation concerne les équipements de sécurité des usagers. Le script crée des indicateurs booléens pour vérifier la présence de ceintures, casques, gants, airbags et autres équipements de protection. Ces informations sont extraites de plusieurs colonnes (secu1, secu2, secu3) et consolidées en de nouvelles variables booléennes, permettant une meilleure interprétation et exploitation par les modèles.

Ensuite, le script traite les informations relatives aux types de chocs subis par les véhicules. Il génère des indicateurs booléens pour identifier les chocs avant, arrière, gauche et droit. Cette transformation permet de simplifier la représentation des impacts et facilite l'analyse des facteurs influençant la gravité des accidents.

Après la création des variables booléennes, les colonnes sources correspondantes sont supprimées afin d'éviter toute redondance inutile dans le jeu de données.

Il permet une meilleure exploitation des informations liées à la sécurité et aux types de chocs tout en préparant efficacement le jeu de données pour les étapes de modélisation.

DepToReg - départements vers régions

Le script dep_to_reg.py est conçu pour assurer la conversion des codes de départements en codes de régions afin d'harmoniser les données géographiques. En utilisant un fichier de correspondance externe, il garantit une transformation fiable et cohérente, essentielle pour les analyses spatiales et les modèles prédictifs exploitant cette information.

Le jeu de données contient des informations GPS, ce qui permettait plusieurs approches pour conserver une quantité significative d'informations géographiques. Deux solutions ont été envisagées :

1. Entraîner un modèle de clustering sur les coordonnées GPS des accidents afin d'identifier des zones homogènes.
2. Regrouper les départements en régions, une solution plus simple à mettre en œuvre et qui réduit la dimensionnalité du problème tout en conservant une information territoriale pertinente.

Dans un souci de simplicité et d'efficacité, la seconde approche a été privilégiée, permettant une standardisation rapide des données géographiques tout en limitant l'augmentation de la complexité du modèle.

Le DepToReg charge un fichier CSV contenant les correspondances entre les départements et les régions. Lors de cette étape, il s'assure que les colonnes nécessaires (departement et region) sont bien présentes et formate les codes des départements pour éviter les erreurs de correspondance.

Le processus de transformation applique un mapping direct des codes de départements aux régions correspondantes. Une attention particulière est portée aux valeurs absentes ou mal formatées, qui sont traitées par une correction automatique. Si certaines valeurs restent non reconnues, elles sont journalisées pour permettre une correction manuelle ultérieure.

Features_fusion.json

Le fichier features_fusion.json est un élément clé du pipeline de prétraitement des données. Il définit un ensemble de règles permettant d'agréger certaines modalités de variables catégorielles afin de réduire la dimensionnalité du problème. Cette approche vise à améliorer la lisibilité des données et à optimiser la modélisation en limitant le nombre de catégories distinctes.

Les jeux de données comportent de nombreuses variables catégorielles ayant un nombre élevé de modalités, ce qui peut complexifier l'entraînement des modèles. Pour y remédier, features_fusion.json regroupe des valeurs similaires sous des catégories plus générales. Cela permet de simplifier l'interprétation des données et d'éviter des problèmes liés à la mal-distribution des modalités rares.

Une autre finalité de ce fichier est la gestion des valeurs manquantes. Certaines modalités sont spécifiées comme à convertir en valeurs nulles (NaN), ce qui permet un traitement différé de ces données dans un pipeline ultérieur, réduisant ainsi le risque de fuites de données (data leakage).

Chaque variable catégorielle est décrite avec trois types d'informations :

- Type de donnée (dtype) : Indique si la variable doit être traitée comme une catégorie ou un entier.
- Regroupement des modalités (modalities) : Définit les nouvelles catégories sous lesquelles les valeurs originales sont regroupées.
- Valeurs à remplacer par NaN (to_nan) : Spécifie les valeurs à exclure pour un traitement ultérieur.

Le fichier features_fusion.json joue un rôle central dans la réduction de la complexité du jeu de données en facilitant la gestion des modalités catégorielles. Il assure une structuration plus cohérente des données et permet d'anticiper le traitement des valeurs manquantes, contribuant ainsi à une meilleure préparation des données avant la modélisation. Toutefois, le choix de regroupement des modalités peut entraîner des biais, un compromis conscient fait pour réduire les temps de calcul, en particulier sur des machines moins performantes. Cette approche améliore la robustesse du pipeline et garantit une meilleure exploitabilité des données dans les phases analytiques et prédictives.

Modélisation

Transformation des données

Imputation des valeurs manquantes

La gestion des valeurs manquantes est une étape essentielle dans le traitement des données en machine learning. Si elles ne sont pas correctement prises en charge, ces valeurs peuvent introduire des biais ou entraîner des erreurs lors de l'entraînement ou de l'évaluation du modèle. Pour remédier à ce problème, nous utilisons la classe SimpleImputer de la bibliothèque sklearn.impute, qui offre des solutions efficaces pour remplacer ces valeurs par des estimations appropriées.

Lors des étapes préliminaires de prétraitement, les valeurs manquantes ont été identifiées et standardisées. Cela a été réalisé en remplaçant les marqueurs spécifiques, tels que -1 pour les données numériques et "-1" pour les données catégoriques, par la valeur universelle NaN. Cette uniformisation facilite leur traitement ultérieur tout en garantissant une cohérence dans l'approche. Une fois ces valeurs manquantes bien définies, les colonnes du jeu de données ont été classées en deux catégories : les colonnes numériques et les colonnes catégoriques. Cette distinction repose sur les types de données observés, permettant ainsi d'appliquer des stratégies d'imputation adaptées à chaque catégorie.

Les données ont ensuite été scindées en deux ensembles distincts : un jeu d'entraînement, contenant 80 % des observations, et un jeu de test, qui regroupe les 20 % restants. Cette séparation vise à préserver l'intégrité de l'évaluation en s'assurant que les données utilisées pour valider le modèle ne sont pas contaminées par les informations du jeu d'entraînement. Ce principe, fondamental en machine learning, garantit une évaluation réaliste des performances du modèle.

Pour les colonnes numériques, les valeurs manquantes ont été remplacées par la moyenne des observations disponibles dans le jeu d'entraînement. Cette stratégie statistiquement neutre permet de réduire les écarts tout en minimisant l'introduction de biais. En ce qui concerne les colonnes catégoriques, les valeurs manquantes ont été imputées à l'aide de la modalité la plus fréquente dans le jeu d'entraînement. Cette méthode vise à préserver les relations existantes entre les catégories et à limiter l'apparition d'effets indésirables, comme l'introduction de nouvelles modalités artificielles.

Ces stratégies d'imputation, appliquées uniquement sur le jeu d'entraînement, renforcent la robustesse du modèle tout en assurant une séparation stricte entre les données d'entraînement et celles de test. Elles permettent également de garantir la compatibilité des données avec les algorithmes de machine learning, qui nécessitent généralement des ensembles de données complets. En mettant en œuvre cette approche, nous avons ainsi renforcé la qualité du jeu de données tout en minimisant les risques de biais pouvant compromettre les performances du modèle.

Encodage et standardisation

Dans leur état brut, les données disponibles ne sont pas exploitables directement par notre modèle. Il est essentiel de les transformer dans un format adapté, permettant une interprétation correcte et une optimisation des performances de l'apprentissage automatique. Pour ce faire, nous utilisons les classes OneHotEncoder et MinMaxScaler de la bibliothèque sklearn.preprocessing, qui offrent des solutions robustes et adaptées pour la gestion des données numériques et catégorielles.

La transformation des valeurs numériques est réalisée à l'aide de MinMaxScaler. Cet outil applique une normalisation linéaire des données, en les redimensionnant dans une plage définie, comprise entre 0 et 1. Chaque valeur est ainsi recalculée proportionnellement par rapport au minimum et au maximum de la colonne correspondante, ce qui permet de préserver les relations relatives tout en réduisant l'influence des écarts d'échelle entre les différentes variables.

Pour les variables catégorielles, nous appliquons la classe OneHotEncoder. Cette méthode consiste à représenter chaque modalité d'une colonne catégorielle par une série de variables binaires (colonnes supplémentaires), où chaque nouvelle colonne indique la présence ou l'absence d'une modalité spécifique. Par exemple, une colonne ayant trois catégories distinctes, comme "rouge", "vert" et "bleu", sera transformée en trois colonnes distinctes, chacune prenant la valeur 1 si la catégorie correspondante est présente, et 0 sinon. Cette technique permet au modèle de traiter les catégories de manière équitable, sans leur attribuer un ordre implicite qui pourrait introduire des biais.

Ces transformations sont appliquées de manière cohérente à la fois sur le jeu d'entraînement et sur le jeu de test, tout en respectant une séparation stricte pour éviter les fuites de données entre ces ensembles. Dans le cas de MinMaxScaler, les paramètres de normalisation (minimum et maximum) sont calculés uniquement sur le jeu d'entraînement, puis appliqués au jeu de test. De même, pour OneHotEncoder, les

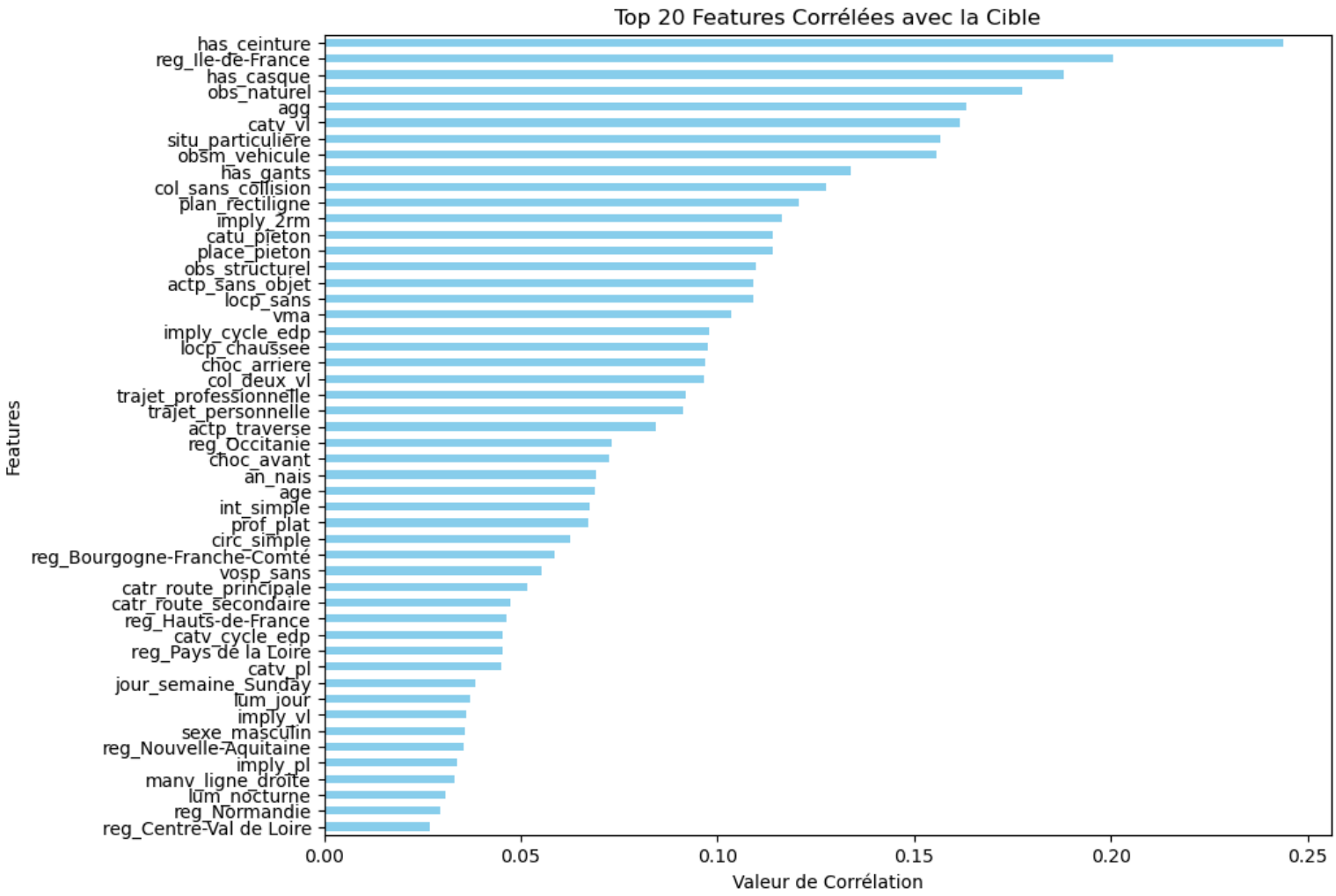
modalités identifiées sur le jeu d'entraînement sont utilisées pour encoder les données de test, garantissant ainsi une cohérence dans le traitement tout en préservant l'intégrité des évaluations. Dans une optique de diminution de la dimentionalité, on supprime la première colonne produite par OneHotEncoder.

En adaptant ainsi les données numériques et catégorielles, nous avons créé un jeu de données parfaitement compatible avec notre modèle, tout en minimisant les risques de biais et en maximisant la pertinence des informations pour l'apprentissage.

Features selection

Pour optimiser les performances de notre modèle et réduire les temps de calcul, il est important de réduire la dimensionnalité des données. La présence d'un grand nombre de variables explicatives peut non seulement ralentir les algorithmes d'apprentissage, mais également introduire du bruit qui nuit à la précision des prédictions. Une réduction efficace de la dimensionnalité permet de diminuer ces risques tout en améliorant la lisibilité des résultats et la robustesse du modèle. Elle peut également prévenir les problèmes de surapprentissage, en se concentrant sur les caractéristiques les plus pertinentes.

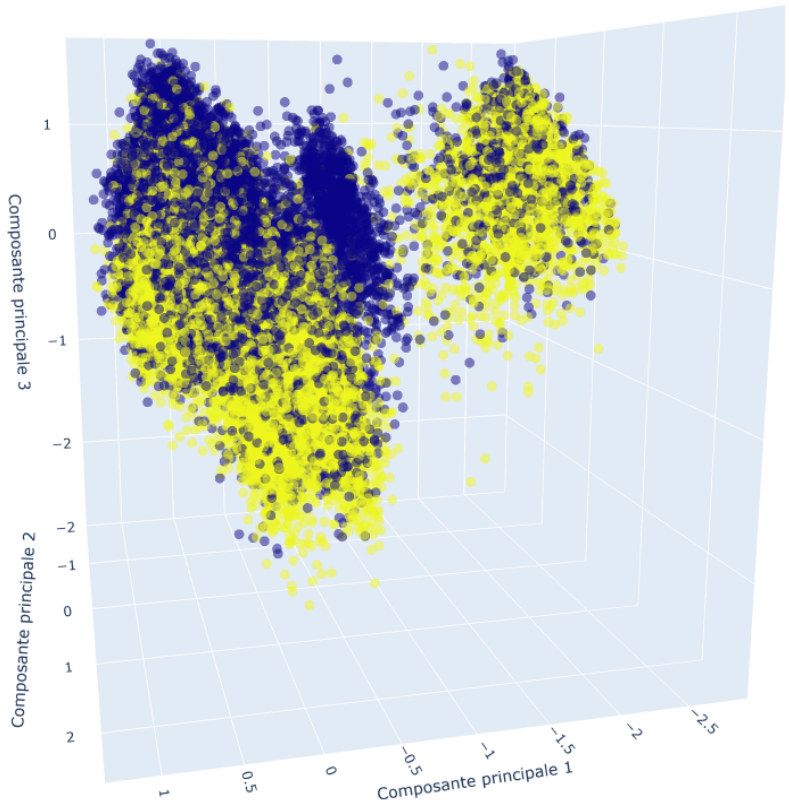
Bien que nous ayons déjà effectué une sélection des features lors d'une analyse approfondie de chaque variable explicative, cette étape se concentre maintenant sur l'application d'algorithmes de réduction de dimension. Ces algorithmes permettent de cibler les variables dont la contribution globale au modèle est limitée. Dans ce cadre, je commence par utiliser VarianceThreshold, un outil proposé par la bibliothèque sklearn.feature_selection. Cet algorithme supprime automatiquement les variables ayant une très faible variance, car elles apportent peu d'informations discriminantes. Une faible variance indique que la variable reste globalement constante à travers les observations, ce qui la rend peu utile pour le modèle, quel que soit le type de tâche d'apprentissage.



En complément, j'examine également les corrélations entre les différentes variables explicatives et la variable cible. Cet examen est réalisé à titre indicatif et n'influence pas directement la réduction des dimensions, mais il me permet d'identifier d'éventuelles problématiques ou biais dans les données. Par exemple, les variables comme has_ceinture, qui indique le port de la ceinture de sécurité, ou reg_ile-de-france, pourraient conduire à des biais dans le modèle si elles influencent de manière disproportionnée les prédictions. Ces observations seront essentielles pour anticiper les risques et envisager des ajustements au moment de l'interprétation des résultats.

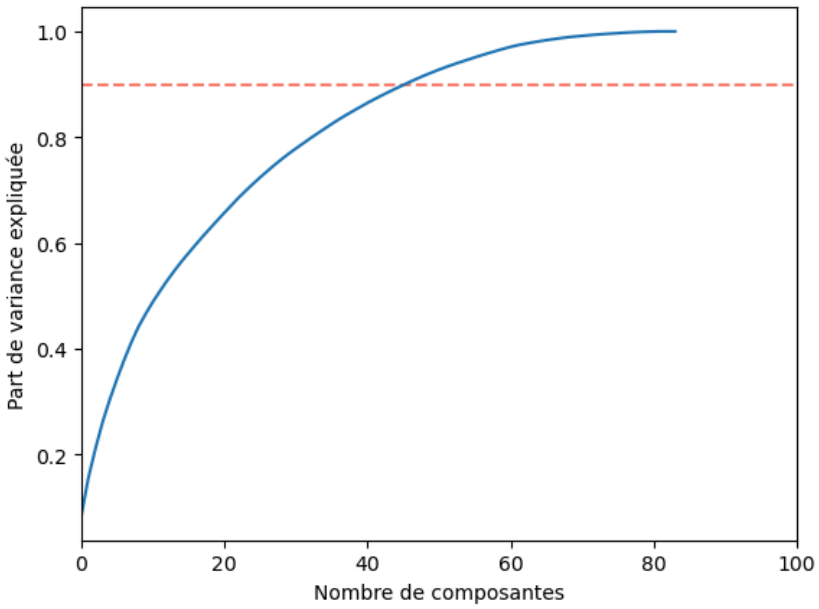
PCA

Pour poursuivre la réduction de la dimensionnalité de nos données, nous avons utilisé l'analyse en composantes principales (PCA). La PCA est une méthode de réduction qui consiste à transformer les données initiales en un ensemble de nouvelles variables, appelées composantes principales, tout en conservant autant que possible la variance totale des données. Chaque composante est une combinaison linéaire des variables d'origine et est orthogonale aux autres, ce qui garantit qu'il n'y a pas de redondance entre elles.



Dans un premier temps, nous avons effectué une PCA à trois composantes afin de représenter visuellement nos données dans un espace tridimensionnel. Cette visualisation a pour objectif de nous permettre d'observer comment les classes se distribuent dans cet espace. Comme le montre le graphique correspondant, les données restent largement entremêlées, ce qui indique que la tâche de discrimination sera difficile pour les modèles. Cependant, certaines zones présentent des regroupements en fonction des classes, ce qui suggère que les modèles pourront mieux discriminer dans ces régions, où les observations d'une même classe sont voisines.

Ensuite, nous avons cherché à déterminer combien de composantes principales étaient nécessaires pour conserver 90 % de la variance des données initiales. L'analyse montre qu'il faut 47 composantes pour atteindre ce seuil, ce qui permet de réduire de près de moitié le nombre de colonnes tout en conservant une grande partie de l'information. Cette réduction substantielle est essentielle pour améliorer les performances computationnelles et simplifier les calculs sans sacrifier significativement

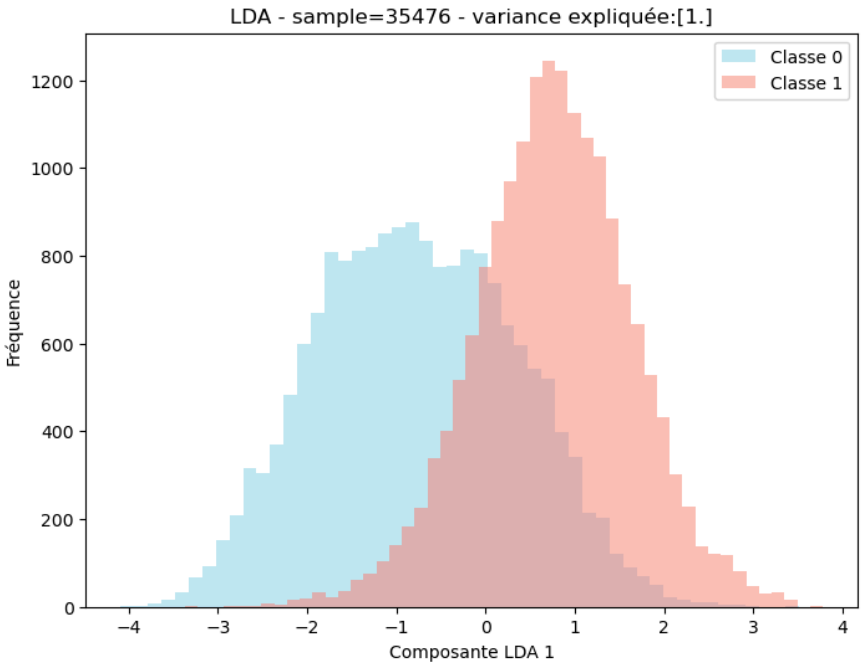


la qualité des données utilisées.

Toutefois, il est important de noter qu'un inconvénient majeur de la PCA réside dans la perte d'interprétabilité des données. À partir de cette étape, il devient plus difficile de relier directement une composante principale à une variable spécifique du jeu de données d'origine. Cela nécessite une attention particulière lors de l'interprétation des résultats et de l'analyse des modèles basés sur ces nouvelles données transformées. Néanmoins, cette approche nous permet de travailler avec des données plus compactes et mieux adaptées aux étapes suivantes du processus de modélisation.

LDA

Une autre approche testée pour réduire la dimensionnalité des données est l'analyse discriminante linéaire (LDA, Linear Discriminant Analysis). LDA est une méthode supervisée qui vise à maximiser la séparation entre les classes en projetant les données dans un espace de dimensions réduites. Contrairement à la PCA, qui est non supervisée et se concentre uniquement sur la variance des données, LDA utilise les informations sur les étiquettes de classe pour maximiser la séparation entre elles.



Dans notre cas, le problème implique deux classes distinctes. LDA réduit donc les données à n−1n - 1n−1 dimensions, soit une seule composante dans ce contexte. Cette

composante représente une projection optimisée pour maximiser la distinction entre les deux classes, tout en minimisant la variance intra-classe.

La représentation graphique obtenue montre les distributions des deux classes projetées sur cette composante unique. Bien que la classe 1 se regroupe majoritairement autour de la valeur 1 et la classe 0 autour de -1, il est clair que les deux surfaces présentent une intersection significative. Cette intersection indique qu'une partie non négligeable des observations est située dans une zone où la distinction entre les deux classes devient floue. Cela risque de compliquer les prédictions des modèles en raison d'un manque de séparation nette.

Après avoir analysé ces résultats, j'ai décidé de ne pas intégrer LDA dans notre pipeline. Bien que la méthode offre une réduction importante de la dimensionnalité, elle est moins adaptée à notre problématique actuelle, car elle ne parvient pas à séparer efficacement les classes de manière suffisamment distincte. Les données projetées dans cet espace unique présentent encore une ambiguïté trop importante pour être exploitées avec une performance optimale. Par conséquent, d'autres méthodes de réduction ou des transformations plus adaptées seront privilégiées pour maximiser la précision et la robustesse des modèles.

Choix du modèle

Dans la quête d'un modèle performant pour résoudre notre problématique de classification, plusieurs essais seront réalisés avec différents algorithmes. Cependant, certaines observations préliminaires sur la structure des données et leur forte interdépendance contextuelle orientent déjà nos choix initiaux. Nous avons constaté que les classes présentent un fort chevauchement et une distribution complexe, ce qui rend les modèles linéaires probablement moins adaptés à cette tâche. Ces derniers risquent d'avoir des difficultés à capturer les relations non linéaires qui semblent dominer dans les données.

Pour cette première phase exploratoire, nous adoptons une démarche simple et rapide, en nous concentrant sur une analyse sommaire des performances des modèles. À ce stade, aucune optimisation poussée, comme le réglage des hyperparamètres ou l'utilisation de validation croisée stratifiée, n'est envisagée. L'objectif est d'identifier les pistes les plus prometteuses sans entrer dans des ajustements complexes.

DecisionTree

Dans cet esprit, j'ai décidé de commencer par les modèles basés sur des arbres de décision, en utilisant l'implémentation DecisionTreeClassifier de la bibliothèque sklearn.tree. Les arbres de décision fonctionnent en divisant les données en sous-groupes successifs, selon des critères qui minimisent l'entropie ou maximisent le gain d'information à chaque étape. Ils permettent de modéliser des relations non linéaires complexes, ce qui les rend bien adaptés à notre problématique.

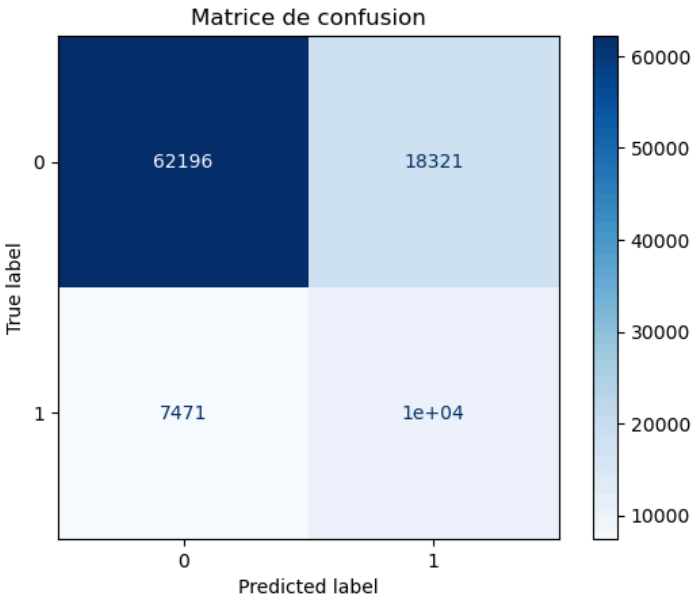
Les résultats obtenus avec DecisionTreeClassifier sont déjà encourageants. La matrice de confusion montre que le modèle parvient à capturer une partie importante de la structure des données, malgré le mélange observé entre les classes. Ce premier essai valide l'hypothèse selon laquelle les modèles d'arbres sont particulièrement adaptés à notre problème. En effet, ils semblent capables de gérer efficacement les dépendances contextuelles présentes dans nos données.

Déséquilibre des classes

Face au déséquilibre significatif des classes dans nos données, et afin de réduire les temps de calcul qui pourraient poser problème sur ma machine, j'ai opté pour l'utilisation de la méthode de sous-échantillonnage via RandomUnderSampler de la bibliothèque imblearn.under_sampling. Cette méthode équilibre les classes en réduisant le nombre d'observations dans la classe majoritaire. Bien que cela entraîne une perte d'information, j'ai choisi d'accepter ce compromis dans cette phase exploratoire. De plus, cette méthode n'est appliquée que sur les jeux d'entraînement, afin de ne pas altérer la distribution naturelle des données dans le jeu de test.

Il aurait également été possible d'utiliser des techniques de sur-échantillonnage, comme le SMOTE, mais en raison des caractéristiques de ma machine et de la nécessité de limiter les calculs, le sous-échantillonnage s'est avéré une option plus adaptée. Cette étape reste dans l'objectif d'une analyse sommaire et d'une première sélection d'un modèle viable, sans chercher pour l'instant à optimiser les performances.

Avec cette approche, le DecisionTreeClassifier manque encore de précision pour la classe 1, avec un score de seulement 0,36. De manière générale, le modèle peine à bien prédire cette classe, en partie à cause du déséquilibre qui persiste dans le jeu de test, où la classe 0 reste largement majoritaire. Cependant, un point positif réside dans le faible taux de faux négatifs, ce qui est un aspect crucial pour l'application visée. Cette observation, bien qu'encourageante, met en évidence les limites du modèle et justifie

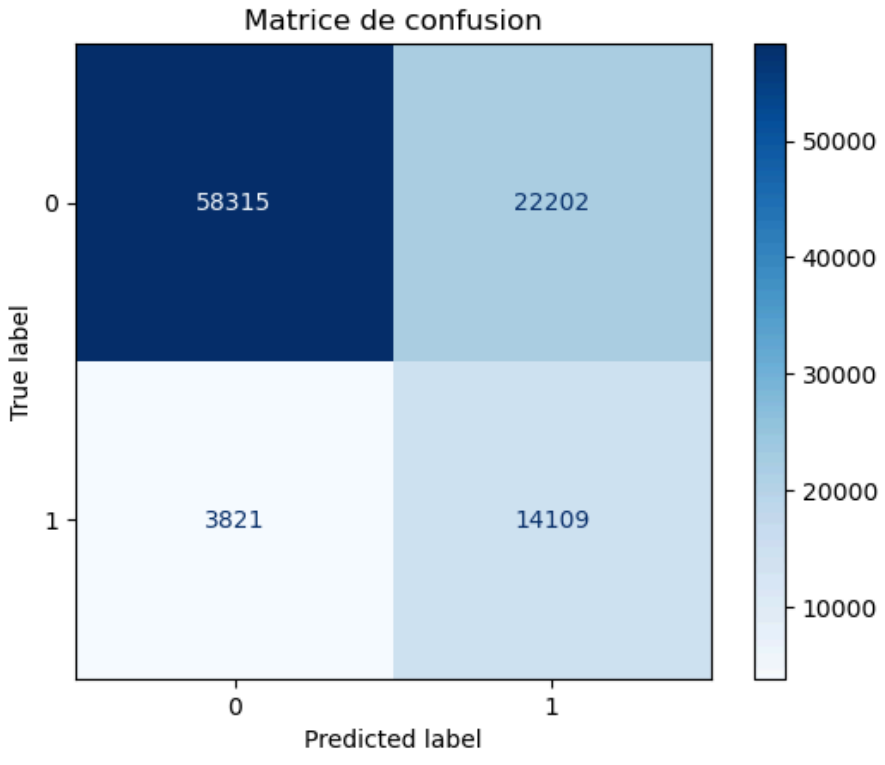


l'exploration de méthodologies complémentaires ou plus avancées pour mieux gérer ce déséquilibre.

Ces résultats prometteurs incitent à poursuivre avec des modèles d'arbres plus avancés, comme les forêts aléatoires ou les méthodes de boosting, afin d'explorer leur potentiel pour améliorer davantage les performances. Cette stratégie permettra de confirmer si les arbres constituent effectivement la famille de modèles la plus pertinente pour notre jeu de données.

Adaboost

AdaBoost, ou Adaptive Boosting, est une méthode d'ensemble qui combine plusieurs modèles faibles, généralement des arbres de décision peu profonds, pour créer un modèle plus robuste. À chaque itération, les observations mal classées par les modèles précédents reçoivent un poids plus élevé, poussant ainsi le modèle suivant à se concentrer sur ces erreurs. Cette approche permet d'améliorer progressivement les performances tout en réduisant les erreurs globales.



Les résultats obtenus avec AdaBoost sont très encourageants en termes de rappel pour la classe 1, qui atteint presque 0,8, une nette amélioration par rapport au modèle précédent. Cela indique que le modèle parvient à identifier beaucoup plus d'instances de la classe 1. Cependant, cette amélioration du rappel s'accompagne d'une baisse significative de la précision pour la classe 1, qui chute à 0,39. En

conséquence, le F1-score de la classe 1 reste faible, à 0,52, montrant un déséquilibre entre la capacité à détecter la classe 1 et la capacité à limiter les erreurs dans ces prédictions.

Un point positif est la réduction continue des faux négatifs, ce qui est essentiel pour notre application. Cependant, le prix à payer est une augmentation notable des faux positifs, au point que le modèle devient moins précis que le hasard pour la classe 1. Ces résultats mettent en lumière les limites d'AdaBoost dans ce contexte, malgré des améliorations significatives sur certains aspects. Cette étape souligne la nécessité de poursuivre les expérimentations pour trouver un équilibre entre rappel et précision, tout en maintenant une bonne robustesse générale.

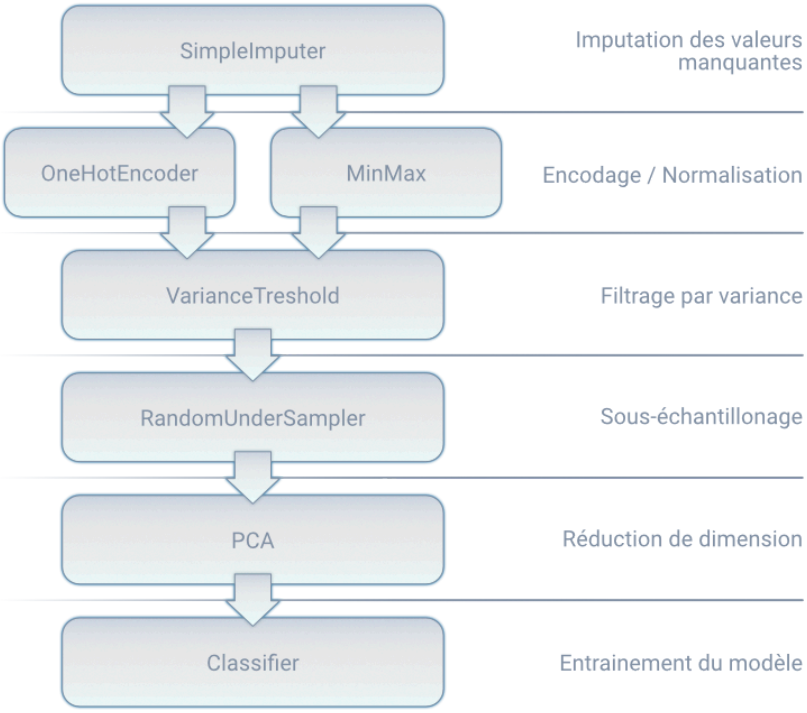
XGBoost

Poursuivant notre exploration des modèles basés sur des arbres, j'ai testé XGBoost (Extreme Gradient Boosting), une méthode puissante et flexible. XGBoost repose sur le principe du boosting par gradient, où une série de modèles faibles, généralement des arbres de décision, est construite de manière séquentielle. Chaque modèle corrige les erreurs de ses prédécesseurs en minimisant une fonction de perte. Grâce à son implémentation efficace, XGBoost offre des performances exceptionnelles et une grande rapidité, tout en intégrant des outils de régularisation pour prévenir le surapprentissage.

Les résultats obtenus avec XGBoost sont nettement supérieurs à ceux des modèles précédents. Le modèle a considérablement gagné en précision pour la classe 1, résolvant en partie les problèmes d'équilibre entre rappel et précision rencontrés avec AdaBoost. Le F1-score de la classe 1 atteint désormais 0,80, un résultat très encourageant qui indique une bonne balance entre la capacité à identifier les instances positives et à limiter les erreurs de classification. Ce niveau de performance montre que XGBoost est particulièrement bien adapté à notre problématique.

Avec ces résultats prometteurs, nous avons désormais une base solide pour passer à la phase d'optimisation des hyperparamètres. Cette étape sera cruciale pour affiner davantage le modèle et maximiser ses performances dans un cadre opérationnel.

Construction d'un Pipeline



Dans le cadre de mes expérimentations, la survenue de data leakage a révélé la nécessité d'une approche rigoureuse pour la transformation des données. Afin de garantir la réplicabilité des résultats et d'assurer une cohérence dans l'entraînement de modèles variés, j'ai mis en place un pipeline de prétraitement structuré. L'enjeu principal réside dans l'automatisation et la reproductibilité des transformations appliquées aux données, évitant ainsi toute fuite d'information entre les phases d'entraînement et de test. Un pipeline bien conçu permet d'intégrer chaque étape de préparation des données dans un flux unique, minimisant ainsi le risque d'erreurs et garantissant un traitement homogène quelles que soient les itérations du modèle.

Le pipeline mis en place s'articule autour de plusieurs étapes essentielles. Je commence par le traitement des valeurs manquantes à l'aide de SimpleImputer, permettant de remplacer les données absentes par une valeur appropriée selon la nature de la variable. Ensuite, les variables catégoriques sont encodées grâce à OneHotEncoder, suivi d'une mise à l'échelle des variables numériques via MinMaxScaler. Ces transformations sont cruciales pour garantir que les caractéristiques du jeu de données soient adaptées aux exigences des algorithmes d'apprentissage, notamment en préservant la distribution relative des valeurs.

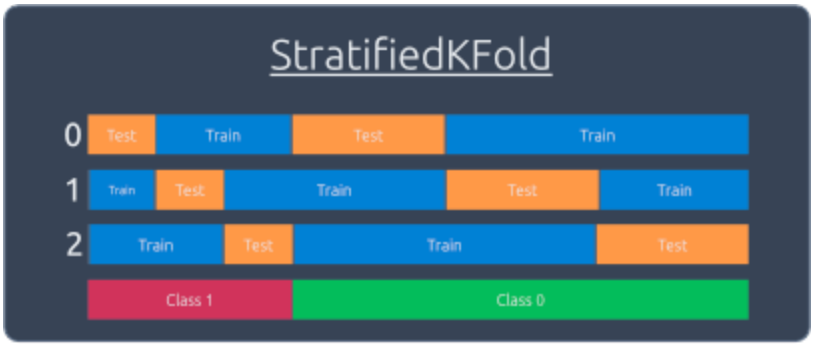
Une fois ces transformations réalisées, il est nécessaire d'effectuer un filtrage des caractéristiques afin d'éliminer celles qui ne contribuent pas significativement à la variance des données. J'utilise pour cela VarianceThreshold, qui supprime les variables ayant une variance trop faible, ce qui permet de réduire la dimensionnalité sans altérer l'information essentielle. Par la suite, afin de gérer le déséquilibre potentiel des classes, j'applique un sous-échantillonnage à l'aide de RandomUnderSampler. Cette étape est

particulièrement importante lorsque la distribution des classes est biaisée, car elle empêche le modèle de favoriser systématiquement la classe majoritaire.

Enfin, avant l'entraînement du modèle, j'applique une réduction de dimension via PCA (Analyse en Composantes Principales). Cette technique permet de projeter les données dans un espace de plus faible dimension tout en préservant un maximum de variance, facilitant ainsi l'apprentissage en réduisant la complexité du modèle. Une fois ces transformations réalisées, les données ainsi préparées sont utilisées pour l'entraînement du modèle, garantissant une évaluation plus fiable et limitant le risque de data leakage rencontré précédemment.

Optimisation des hyperparamètres

Pour optimiser les hyperparamètres de mon modèle, j'ai choisi d'utiliser RandomizedSearchCV. Contrairement à GridSearchCV, qui évalue exhaustivement toutes les combinaisons possibles de paramètres, cette approche sélectionne aléatoirement un sous-ensemble de combinaisons, permettant ainsi de réduire considérablement le temps de calcul. Bien que cette méthode soit statistiquement moins exhaustive, elle offre un compromis pertinent entre performance et rapidité, un critère essentiel dans mon cas puisque l'entraînement s'effectue sur une machine aux ressources limitées.



Afin de garantir une évaluation robuste du modèle, l'optimisation repose sur une validation croisée StratifiedKFold. Cette technique divise le jeu de données en plusieurs sous-ensembles tout en préservant la répartition des classes à chaque itération. Ainsi, chaque pli du modèle est testé sur des données qu'il n'a pas vues lors de son entraînement, réduisant le risque de surapprentissage et garantissant une estimation plus fiable de la performance.

L'ensemble du processus d'optimisation est guidé par la métrique balanced_accuracy, qui permet de mieux évaluer les performances du modèle dans un contexte de classes déséquilibrées. Une fois les expérimentations terminées, je conserve le pipeline ayant obtenu les meilleurs résultats selon cette métrique, garantissant ainsi une approche optimisée et adaptée à la nature de mes données.

MLFlow

Pour assurer un suivi rigoureux de mes expérimentations et faciliter la comparaison des modèles, j'ai mis en place un enregistrement systématique des runs à l'aide de MLflow. Ce choix me permet de centraliser toutes les informations essentielles, garantissant ainsi une traçabilité optimale des performances et une analyse fine des évolutions au fil du temps. En plus des métriques classiques telles que accuracy, precision, recall et F1-score, j'enregistre également le seuil de décision du modèle, un élément clé pour ajuster son comportement en fonction des besoins applicatifs.

Afin de conserver des éléments visuels facilitant l'interprétation des résultats, j'intègre dans chaque run plusieurs artifacts, notamment la matrice de confusion, la courbe ROC-AUC et la courbe precision-recall. Ces visualisations me permettent d'évaluer non seulement la performance globale du modèle, mais aussi la manière dont il se comporte en fonction des seuils de classification. Grâce à cette approche adoptée dès les premières itérations, j'ai pu observer précisément la progression du modèle et identifier rapidement les choix les plus impactants sur ses performances.

				Metrics			
	Run Name	Created	Duration	f1_score	recall	accuracy	precision
	Evaluation_RandomForest...	11 days ago	1.3min	0.846890448...	0.861001350...	-	-
	Evaluation_RandomForest...	11 days ago	1.3min	0.846705451...	0.860869300...	-	-
	Evaluation_RandomForest...	10 days ago	1.6min	0.845737547...	0.860402043...	-	-
	Evaluation_RandomForest...	10 days ago	1.7min	0.845342846...	0.860188730...	-	-
	Evaluation_XGBoost	16 days ago	2.4s	0.846675252...	0.860107468...	-	-
	Evaluation_XGBoost	16 days ago	2.1s	0.846675252...	0.860107468...	-	-
	Evaluation_XGBoost	16 days ago	2.4s	0.846675252...	0.860107468...	-	-
	Evaluation_XGBoost	16 days ago	2.3s	0.846675252...	0.860107468...	-	-
	Evaluation_XGBoost	17 days ago	2.2s	0.846675252...	0.860107468...	-	-
	Evaluation_XGBoost	17 days ago	2.3s	0.846675252...	0.860107468...	-	-

J'ai choisi d'utiliser joblib pour la persistance du pipeline. Cette approche me permet de sauvegarder l'intégralité du processus, incluant toutes les transformations appliquées aux données ainsi que le modèle entraîné. Ainsi, je peux charger directement le pipeline sans avoir à réexécuter l'ensemble des étapes de prétraitement, ce qui simplifie considérablement l'inférence sur de nouvelles données. En enregistrant le modèle sous cette forme, je m'assure également que les paramètres et les hyperparamètres optimisés sont conservés intacts, garantissant des performances cohérentes d'une exécution à l'autre. Cette solution facilite non seulement le partage et la

reproductibilité des résultats, mais aussi l'intégration du modèle dans un environnement de production.

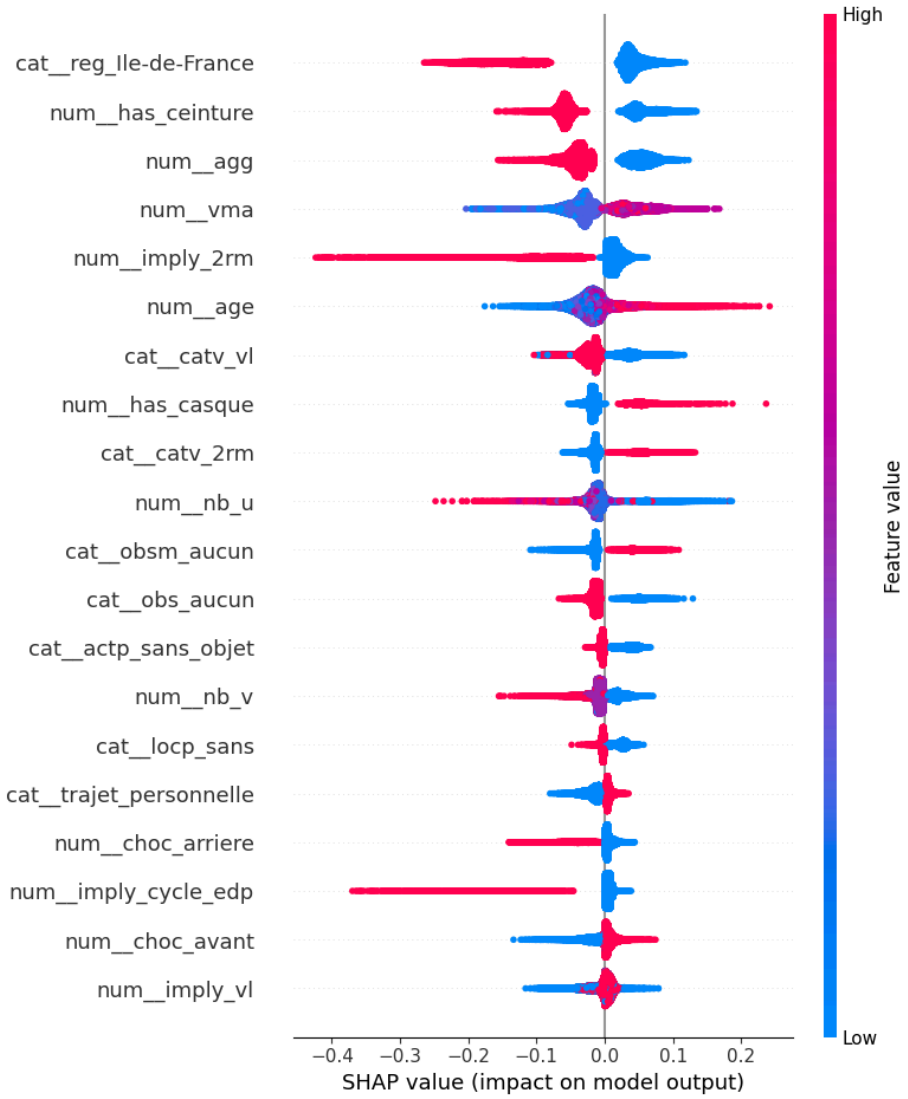
Environnement Conda

Afin de garantir la reproductibilité et la stabilité de mon travail, j'ai choisi d'utiliser un environnement Conda pour l'ensemble de mes expérimentations. L'un des principaux avantages de Conda réside dans sa capacité à gérer les dépendances de manière isolée, évitant ainsi les conflits entre différentes versions de bibliothèques, ce qui est essentiel en machine learning où les mises à jour peuvent impacter le comportement des algorithmes. Cet environnement me permet également de figer les versions des packages utilisés, assurant ainsi que les résultats obtenus restent cohérents même en cas de réexécution ultérieure du pipeline. Grâce à cette approche, je peux travailler sur différentes configurations sans risquer d'altérer mon projet principal, garantissant ainsi un cadre stable et contrôlé pour le développement et l'expérimentation de mes modèles.

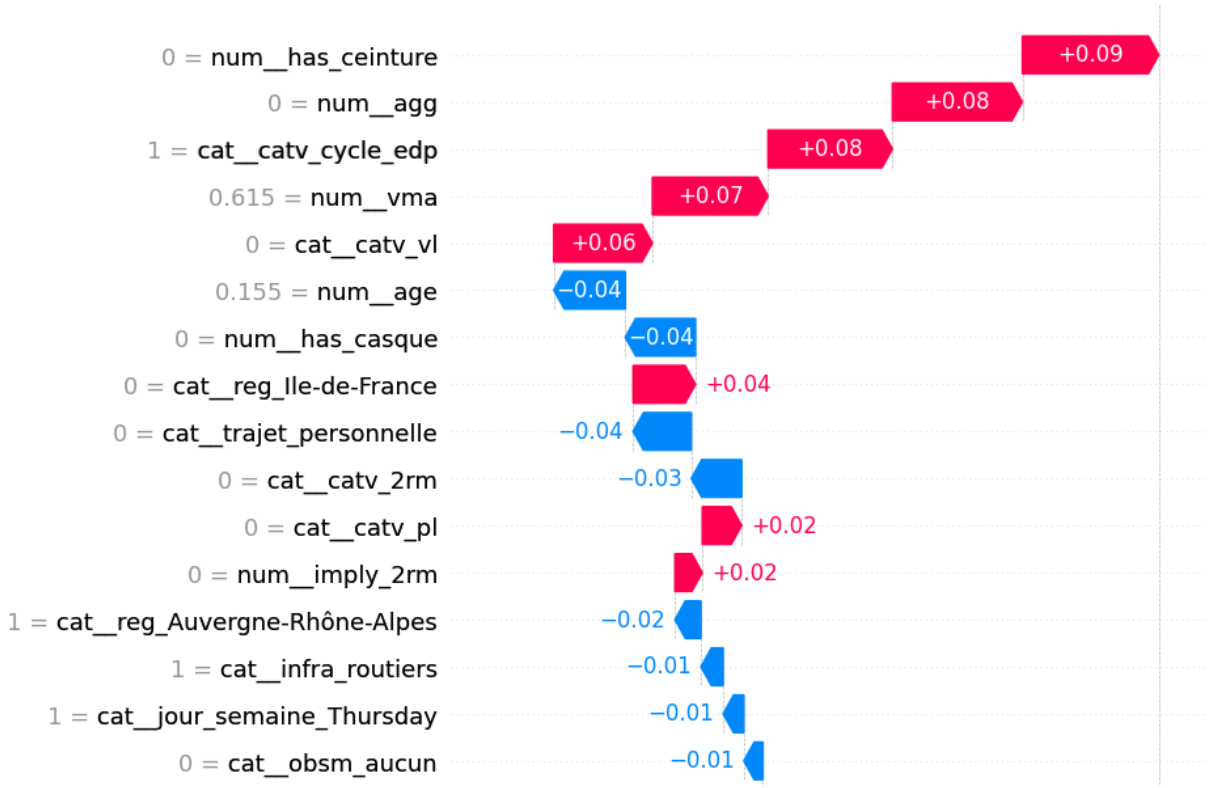
SHAP

Pour interpréter les résultats de mon modèle, j'ai utilisé SHAP, une méthode permettant d'analyser l'impact de chaque variable sur la décision prise. Cependant, l'utilisation de PCA dans le pipeline complique cette interprétation, car la transformation projette les données dans un nouvel espace où les colonnes d'origine ne sont plus directement accessibles. Afin de lever cette contrainte et d'obtenir des explications exploitables, j'ai temporairement retiré PCA du pipeline pour cette phase d'analyse.

L'un des premiers éléments étudiés est le summary plot, qui met en évidence les variables ayant le plus d'influence sur les prédictions du modèle. Il ressort notamment que certaines colonnes, comme cat_reg_Ile-de-France, semblent jouer un rôle important en réduisant la gravité des blessures lorsqu'elles sont présentes. De même, num__has_ceinture apparaît comme un facteur déterminant dans la prise de décision du modèle, ce qui souligne l'impact des mesures de sécurité sur l'évaluation du risque.



Cette analyse confirme un problème que nous soupçonnions déjà : le manque de contextualisation du modèle. En particulier, il peine à différencier correctement les types de véhicules, ce qui suggère un besoin d'amélioration dans la prise en compte de ces informations. Pour approfondir cette observation, j'ai analysé une prédiction erronée à l'aide d'un waterfall plot. Le modèle a classé un accident comme grave alors qu'il ne l'était pas, mettant en avant l'absence du port de la ceinture comme l'un des principaux facteurs expliquant cette classification. Or, en examinant les données, il apparaît que l'usager impliqué était en EDPM (Engin de Déplacement Personnel Motorisé), un type de véhicule pour lequel le port de la ceinture n'est pas pertinent. Le modèle semble ainsi pénaliser deux fois cet usager pour une caractéristique qui, dans son contexte, n'a pas la même signification, ce qui souligne un biais important dans son raisonnement.



Conclusion

L'analyse des résultats obtenus jusqu'à présent montre que mon modèle n'est pas encore prêt à être mis en production. Les performances observées restent insuffisantes pour garantir une prise de décision fiable, et des ajustements sont nécessaires avant d'envisager un déploiement opérationnel.

Une évaluation basée uniquement sur l'accuracy s'avère inadaptée, en particulier dans le cas de notre jeu de données, qui présente un déséquilibre marqué entre les classes. Cette métrique, bien qu'intuitive, ne permet pas de mesurer efficacement la capacité du modèle à identifier correctement les cas critiques. À ce titre, le recall et le F1-score sont des indicateurs plus pertinents, car ils tiennent compte du compromis entre la précision et la capacité du modèle à détecter les accidents graves.

L'optimisation du modèle ne peut se faire indépendamment des objectifs métiers. Il est crucial de déterminer si les faux positifs ou les faux négatifs représentent le plus grand risque. Une mauvaise classification peut avoir des conséquences importantes, notamment en matière de prévention et d'intervention. Si une fausse alerte entraîne un coût inutile, un accident grave mal classifié peut compromettre des actions correctives

nécessaires. Ainsi, l'ajustement du modèle doit être guidé par ces enjeux stratégiques.

L'utilisation d'un pipeline de traitement des données est essentielle pour assurer la robustesse du modèle. Une attention particulière doit être portée aux transformations appliquées afin d'éviter tout risque de data leakage. Une amélioration trop rapide et significative des performances doit immédiatement soulever des interrogations sur l'intégrité du pipeline et la pertinence des données utilisées. Chaque amélioration doit être justifiée et analysée de manière rigoureuse pour garantir la fiabilité des résultats.

Compte tenu des contraintes matérielles, plusieurs solutions ont été mises en place pour permettre l'exécution du projet sur une machine de faible puissance. Parmi celles-ci, la parallélisation du chargement des données initiales a permis de réduire le temps de traitement et d'optimiser les ressources disponibles. Ce type d'adaptation est fondamental pour garantir l'efficacité du développement et de l'expérimentation dans des conditions réelles.

L'une des principales limites du modèle actuel réside dans sa difficulté à contextualiser les accidents. L'analyse des performances suggère qu'il serait plus pertinent d'entraîner plusieurs modèles spécifiques en fonction des types de véhicules impliqués. Cette approche permettrait de mieux capturer les spécificités propres à chaque catégorie d'usagers et d'améliorer la précision des prédictions.

Par ailleurs, il est important de rappeler que les accidents routiers résultent de nombreux facteurs, dont une partie échappe à notre base de données. Nos données ne capturent pas l'intégralité des paramètres qui influencent un accident, et les biais de saisie constituent une problématique majeure. Bien que des efforts aient été déployés pour transformer et nettoyer les données, cet aspect reste perfectible. Une meilleure structuration et intégration des données pourrait contribuer à améliorer significativement la qualité des prédictions.

Enfin, ce projet a été une opportunité précieuse pour appréhender les défis concrets de la mise en œuvre d'un modèle de machine learning dans des conditions réelles. Il illustre l'écart existant entre la théorie et la pratique, soulignant la nécessité d'une expérience continue pour affiner les compétences et anticiper les contraintes liées à l'opérationnalisation des modèles. Seule une pratique régulière permettra d'acquérir une expertise suffisante pour déployer des solutions robustes et adaptées aux besoins du terrain.