ROYAUME DU MAROC
UNIVERSITE HASSAN 1
FACULTE DES SCIENCES ET TECHNIQUES SETTAT

2023/2024

RAPPORT DE PROJET DE FIN D'ETUDE

Pour l'Obtention du Diplôme :

LICENCE SCIENCES ET TECHNIQUES EN SYSTEME D'INFORMATION ET TRANSFORMATION DIGITAL

Analyse Prédictive des Risques d'Accidents Routiers au Maroc : Identification des Facteurs Clés et Développement de Solutions Prédictives à l'Aide de l'Apprentissage Automatique

En collaboration avec:

Réalisé par:	Encadré a la FST par:	Encadré a la NARSA par:
Chaymae Moudnib	**M.Touil Achraf**	**M.Ahmed bardan**

Soutenu le 25 Juin 2024 devant les membres de Jury:
Pr.Maha Rezzai : Présidente , FST Settat.
Pr. Tahiry Karim : Rapporteur, FST Settat.
Pr. Touil Achraf : encadrant, FST Settat.

I.DÉDICACE

Je dédie ce travail, fruit de nombreuses heures de dévotion et de persévérance :

À mes chers parents,
Moustapha Moudnib et **Amina Abouabdellah,**
Votre amour et votre soutien inconditionnels sont les pierres angulaires de ma vie. Que cette réalisation soit le reflet de vos sacrifices et le symbole de vos aspirations pour moi.

À mes professeurs respectés,
Pour votre guidance éclairée et votre patience, sans lesquelles ce parcours n'aurait pas été le même.

À ma sœur de cœur, **Alae moudnib**,
Ton soutien sans faille et tes encouragements ont été un phare dans les moments d'incertitude.

À mes précieuses amies : **Yassmine Bouthaich ,Salwa dardour**, **Khadija Jouah**, **Nour elhouda Othmani**.

Pour toutes ces années d'amitié, de rires et de soutien, et pour être les sœurs que le destin m'a données.

À tous ceux qui, par leur bienveillance ou leur collaboration, ont contribué à la réussite de ce projet,
Votre influence, bien que parfois silencieuse, a été une source constante d'inspiration et de motivation.

Et à tous ceux que je n'ai pas nommés mais qui occupent une place chère dans mon cœur,
Sachez que votre rôle dans ma vie ne passe pas inaperçu et que votre soutien est gravé dans ma mémoire à jamais.

II.REMERCIEMENT

Je souhaite exprimer mes sincères remerciements et ma gratitude à toutes les personnes qui ont contribué à la réalisation de ce projet, une expérience à la fois enrichissante et formatrice.

Un merci tout particulier à **M.Touil Achraf** , mon directeur de mémoire, pour son accompagnement, ses conseils précieux et son soutien constant. Sa rigueur et son expertise ont été des atouts indispensables dans l'élaboration de ce travail.

Je suis également reconnaissante envers **M.Ahmed Bardan**, Chef de Département d'Observatoire national de la sécurité routière au Maroc, encadrant externe, qui a apporté un regard neuf et des compétences complémentaires essentielles à la réussite de ce projet. Sa collaboration a été très enrichissante et a permis d'élargir les perspectives de cette étude.

Un grand merci aussi à **M.Bilal Dafa**, mon coencadrant externe, pour ses contributions précieuses et son soutien technique. Sa participation active a été un pilier dans l'aboutissement de ce projet.

Je tiens à remercier **M.Hicham DIOURI**, Chef de Division Études, Surveillance et Expertise en Sécurité Routière au sein du NARSA, pour son expertise et sa disponibilité.

Un remerciement spécial à Madame **Wafaa Dachery**, Chef de filière, pour son soutien, ses conseils avisés et son engagement envers le succès des étudiants. Sa présence et son encadrement ont été extrêmement précieux tout au long de ce parcours académique.

Ma gratitude s'étend à tous les membres du jury, pour leur disponibilité, leurs commentaires constructifs et le temps consacré à l'évaluation de ce travail.

Je suis aussi très reconnaissante envers le **NARSA** pour avoir mis à disposition les ressources nécessaires à notre recherche et pour leur coopération et leur soutien.

Un merci affectueux à ma famille et à mes amis, pour leur amour, leur soutien sans faille et leur encouragement constant, qui ont été des sources de motivation inestimables.

Enfin, je remercie tous ceux qui ont contribué indirectement à ce projet, chaque interaction ayant apporté une pierre à l'édifice de cette recherche.

III.RÉSUMÉ

Le rapport examine le problème pressant des accidents de la route au Maroc, un enjeu de taille pour la sécurité publique. Face à cette problématique, il devient impératif que les autorités développent des stratégies proactives pour anticiper et prévenir les incidents routiers, essentiels pour la sauvegarde des vies et la sécurité routières.

Nous proposons une approche novatrice : la création d'un modèle prédictif avancé qui utilise les données historiques locales pour anticiper les risques d'accidents. Ce modèle se concentrerait spécifiquement sur les zones rurales, exploiterait les technologies de pointe en matière d'analyse de données et d'apprentissage automatique. L'objectif serait de déterminer les facteurs de risque clés et de prédire les points noirs potentiels, permettant ainsi une intervention ciblée et préventive.

L'efficacité de ce modèle ne se limiterait pas à la prévention; elle permettrait également une gestion plus stratégique des ressources. En identifiant les zones à haut risque, les autorités pourraient optimiser l'allocation des ressources comme les patrouilles de police et les équipements de sécurité, assurant ainsi une utilisation judicieuse et économique des fonds publics.

Ce projet s'inscrit dans le cadre de la stratégie de sécurité routière "NARSA" pour la période 2017-2026, visant à réduire de moitié le nombre de décès dus aux accidents de la route d'ici 2026. En intégrant notre modèle prédictif à cette initiative globale, nous renforçons les efforts nationaux pour atteindre cet objectif ambitieux. Ainsi, notre solution ne se contente pas de répondre à un besoin immédiat mais s’aligne également sur les plans à long terme de sécurité routière au Maroc, contribuant de manière significative à la vision nationale de prévention des risques routiers.

Mot-clés: Apprentissage automatique , la sécurité routières , les risques d'accidents , NARSA, analyse de données,modèle prédictif.

IV.ABSTRACT

The report delves into the growing issue of road accidents in Morocco, a critical concern for public safety. In response to this pressing challenge, it's essential for authorities to not only react to accidents but to anticipate and prevent them, safeguarding lives and enhancing road safety.

We propose a forward-thinking solution: the creation of an advanced predictive model that leverages local historical data to foresee and mitigate accident risks. This model would particularly focus on rural areas, utilizing the latest in data analytics and machine learning technology. The goal is to unearth key risk factors and identify potential hotspots for accidents, enabling proactive and preventative measures.

More than just preventing accidents, this model would allow for smarter management of resources. By pinpointing high-risk areas, it would enable authorities to better allocate resources such as police patrols and safety equipment, ensuring that public funds are used efficiently and effectively.

This initiative is part of the broader "NARSA" road safety strategy, which spans from 2017 to 2026 and aims to cut the number of road traffic fatalities by half by 2026. Integrating our predictive model into this strategy bolsters ongoing national efforts to reach this ambitious goal. Our approach not only meets an immediate need but also aligns with Morocco's long-term road safety objectives, contributing significantly to the country's overarching vision of preventing road risks. This human-centered approach underscores our commitment to not just improving statistics, but saving lives and building safer communities.

Keywords: Machine learning, road safety, accident risks, NARSA, data analysis, predictive model.

ملخص.V

يسلط التقرير الضوء على مشكلة متفاقمة في المغرب، وهي حوادث الطرق، التي باتت تمثل تحديًا كبيرًا للسلامة العامة. في ظل هذه المشكلة، تبرز الحاجة الملحة للسلطات لتطوير استراتيجيات استباقية تتجاوز مجرد الاستجابة للحوادث، إلى توقعها ومنعها، لضمان الحفاظ على الأرواح وتعزيز الأمان على الطرق.

نقدم حلاً مبتكرًا: تطوير نموذج تنبؤي متقدم يستغل البيانات التاريخية المحلية لتقدير مخاطر الحوادث. هذا النموذج سيركز بشكل خاص على المناطق الريفية التي غالبًا ما تُغفل في الدراسات المتعلقة بسلامة الطرق، مستخدمًا أحدث التقنيات في التحليل البياناتي والتعلم الآلي. الهدف هو استكشاف العوامل الأساسية للخطر وتحديد النقاط الأكثر عرضة للحوادث للتدخل المستهدف والوقائي.

تتمثل إحدى الفوائد الرئيسية لهذا النموذج في تمكين السلطات من إدارة الموارد بكفاءة، من خلال توجيه الدوريات الشرطية وتخصيص المعدات الأمنية بناءً على التنبؤات، لضمان استخدام فعال واقتصادي للموارد العامة وتقليل الحوادث.

يتماشى هذا المشروع مع استراتيجية السلامة الطرقية الوطنية "نارسا" للفترة من 2017 إلى 2026، التي تستهدف خفض عدد الوفيات الناتجة عن حوادث الطرق بنسبة النصف بحلول عام 2026. إن دمج نموذجنا التنبؤي ضمن هذه الاستراتيجية يعزز الجهود الوطنية المبذولة لتحقيق هذا الهدف الطموح. بالتالي، يعد حلنا جزءًا لا يتجزأ من خطط السلامة على الطرق طويلة الأمد في المغرب، ويسهم بشكل فعال في الرؤية الوطنية لتقليل المخاطر على الطرق، مؤكدًا على التزامنا بحماية الأرواح وبناء مجتمعات أكثر أمانًا

الكلمات المفتاحية:التعلم الآلي، السلامة الطرقية، مخاطر الحوادث، نارسا، تحليل البيانات، النموذج التنبؤي

Table des matières

Table des matières

Table des matières

Liste des figures

Liste des figures

Liste des figures

Liste des tableaux

Liste des acronymes

NARSA : Agence nationale pour la sécurité routière.
RF: Random Forest.
KNN :K-Nearest Neighbors.
ANN :Artificial Neural Network.
AUC:(Area Under the Curve) .
ROC: (Receiver Operating Characteristic).
ANOVA: Analyse de variance.
TJMA :(Taux Journalier Moyen Annuel) Nombre moyen quotidien de véhicules en un point donné sur une année.
PKM: Distance parcourue par véhicule par kilomètre.
PKD/PKO: (Parcours Kilométrique Départ/Parcours Kilométrique Officiel) Le point de départ d'un trajet en termes de kilométrage. C'est le kilométrage enregistré lorsque le véhicule commence son parcours.
PKF/PKE:(Parcours Kilométrique Fin /Parcours Kilométrique Extrême) Le point de fin d'un trajet en termes de kilométrage. C'est le kilométrage enregistré lorsque le véhicule termine son parcours.
GR: Gendarmerie Royale.
DGSN : Direction Générale de la sureté Nationale.
EURORAP: European Road Assissement Programme.

Introduction générale

L'importance cruciale du Big Data dans la transformation digitale se renforce, devenant un pilier central de la révolution numérique. Chaque interaction dans l'espace digital génère d'importantes quantités de données, permettant de construire des bases solides et d'ouvrir de nouvelles perspectives pour relever les défis contemporains.

Dans ce contexte dynamique, nous sommes ravis de présenter notre projet innovant : le développement d'un modèle sophistiqué pour prédire les risques d'accidents de la route. Ce projet s'ancre profondément dans le domaine du Big Data et de la transformation digitale, avec l'ambition de révolutionner la manière dont nous abordons la sécurité routière.

Notre mission est claire : exploiter les vastes quantités de données disponibles pour anticiper et prévenir les accidents. Notre approche est multifacette, intégrant diverses technologies et méthodologies. Nous avons commencé par l'utilisation d'Excel pour la préparation et le prétraitement des données, puis nous avons employé R et Python pour réaliser des analyses plus complexes et mettre en œuvre des techniques de Machine Learning.

Pour optimiser notre processus et améliorer notre productivité, nous avons adopté Teamgantt pour la planification et la gestion de projet, assurant ainsi une coordination efficace des tâches. Parallèlement, nous utilisons Power BI pour une analyse approfondie des données, ce qui nous permet une exploration interactive et une visualisation détaillée.

Avec ces outils et méthodologies, nous développons des algorithmes avancés capables de détecter les facteurs de risque et de construire un modèle prédictif. Ce dernier vise à anticiper les accidents avant leur survenue, contribuant significativement à la sécurité routière.

Nous sommes convaincus que ce projet apportera une amélioration tangible à la sécurité sur les routes, en réduisant le nombre d'accidents et en créant un environnement plus sûr pour tous. Nous vous invitons à explorer en détail ce projet, qui, nous l'espérons, inspirera des mesures concrètes pour renforcer la sécurité routière et sauver des vies.

Chapitre 1 :
Contexte général du projet

Dans cette section, nous avons condensé les informations clés relatives à l'entité responsable du projet, à ses objectifs, aux parties prenantes, ainsi qu'au contexte et à la planification de celui-ci. Cette synthèse est conçue pour fournir une compréhension approfondie du cadre dans lequel le projet est déployé, éclairant ainsi les enjeux et les dynamiques à l'œuvre. Cette démarche vise à assurer une vision claire et structurée, permettant aux lecteurs de saisir pleinement l'orientation et les ambitions du projet, ainsi que le rôle et les contributions de chaque acteur impliqué.

I. Présentation de l'organisme d'accueil de stage

Figure 1 : Logo de NARSA

L'Agence Nationale de la Sécurité Routière (NARSA) joue un rôle essentiel dans la préservation de la vie et de la sécurité des usagers de la route au Maroc, un établissement d'utilité publique rattaché au ministère de l'Équipement et du Transport. NARSA a émergé comme une réponse proactive aux défis persistants de la sécurité routière dans le pays.

1. Mission de NARSA

- Contribuer à l'élaboration, à l'exécution, au suivi et à l'évaluation de la stratégie nationale de la sécurité routière.
- Établir un système intégré et inclusif de collecte des informations et des données relatives aux accidents et veiller à leur traitement, leur exploitation et leur publication.
- Établir des partenariats avec les organismes étrangers et internationaux concernés par la sécurité routière.
- Soutenir, encourager et promouvoir la recherche scientifique dans les différents domaines liés à la sécurité routière.
- Mettre en œuvre les projets relatifs à l'amélioration de la sécurité routière dans le cadre du partenariat.

2. Fiche technique de NARSA

Cette fiche technique permet d'élaborer et de fournir des informations détaillées sur l'organisation de NARSA.

Pour plus d'information sur National Road Safety Agency(NARSA) ,Visitez <https://www.narsa.ma/fr/nos-missions> (Dernière consultation le 30/05/2024).

Tableau 1 : Fiche technique de NARSA

Raison Sociale	NARSA
Secteur	La sécurité routière
Forme juridique	Institution publique
Date de création	2018
Siège social	Av. AL Arâar, Hay Riad – Rabat
Effectifs	500-1000
Email	contact@narsa.gov.ma
Site Web	https://www.narsa.ma/fr

3. Analyse SWOT de NARSA

En analysant NARSA, une perspective SWOT met en lumière ses forces, faiblesses, opportunités et menaces, offrant ainsi une vue globale du paysage stratégique de l'organisation.

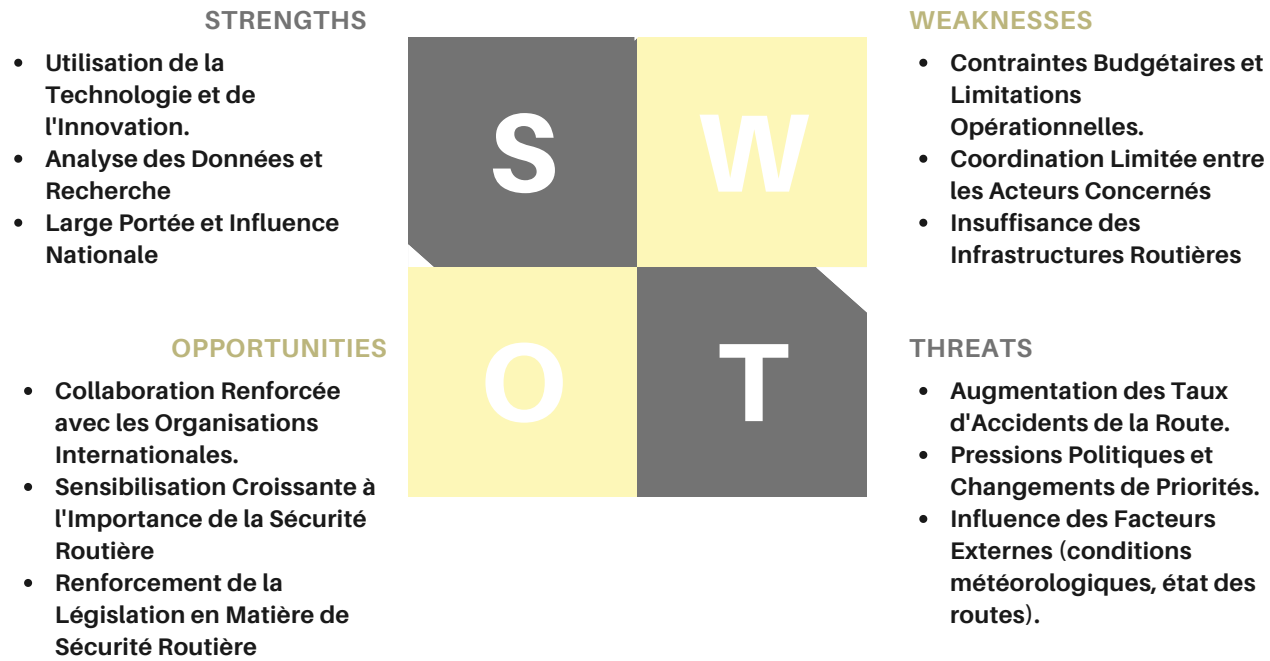


Figure 2: Analyse SWOT de NARSA

Pour plus d'information sur National Road Safety Agency(NARSA) ,Visitez <https://www.narsa.ma/fr/nos-missions> (Dernière consultation le 30/05/2024).

3. Organigramme de NARSA

La NARSA est structurée en cinq divisions principales, et notre intervention s'inscrit au sein de la division "Surveillance et Expertise en Sécurité Routière", plus précisément dans l'enceinte de l'"Observatoire National de la Sécurité Routière". Cette branche est chargée de la gestion intégrale des données relatives aux accidents de la route au Maroc. Elle coordonne la collecte, le traitement et l'analyse des données à l'échelle nationale, tout en assurant la mise à jour régulière de la base nationale des données des accidents corporels de la circulation routière.

Figure 3: Organigramme de la branche de Surveillance et Expertise en Sécurité Routière

Pour plus d'information sur National Road Safety Agency(NARSA) ,Visitez <https://www.narsa.ma/fr/nos-missions> (Dernière consultation le 30/05/2024).

Grâce à une surveillance rigoureuse de la qualité des statistiques d'accidents, l'Observatoire développe des indicateurs de risques d'accidents adaptés à différents niveaux administratifs. Il propose également des interventions ciblées pour renforcer la sécurité routière, fondées sur des analyses détaillées des données collectées. En outre, l'Observatoire collabore avec des partenaires internationaux afin d'enrichir ses méthodologies et pratiques. Ces efforts sont régulièrement partagés avec le public et les parties prenantes à travers la publication de rapports annuels qui documentent l'évolution de la sécurité routière et évaluent les performances de l'Observatoire.

II. Contexte du projet

Dans le contexte de la sécurité routière au Maroc, la transformation digitale est cruciale pour relever les défis croissants des accidents de la route. Le projet se positionne au cœur de cette transformation en développant un modèle de prédiction avancé des risques d'accidents, en se basant sur une analyse approfondie des données locales.

Avec l'adaptation d'une approche technologique, le projet vise à fournir aux autorités marocaines des outils novateurs pour prévenir les accidents. L'analyse des données et le Machine learning permettent de dégager des tendances précieuses, éclairant ainsi les défis spécifiques rencontrés sur les routes.

La transformation digitale transcende les approches traditionnelles en sécurité routière en exploitant les technologies émergentes. En améliorant la précision des prédictions, ce projet contribue significativement à réduire les accidents de la route au Maroc, préservant ainsi des vies et renforçant la sécurité routière pour les générations à venir.

.

Pour plus d'information sur National Road Safety Agency(NARSA) ,Visitez <https://www.narsa.ma/fr/nos-missions> (Dernière consultation le 30/05/2024).

1.Problématique

Figure 4 : Des chiffres sur les accidents en 2022.

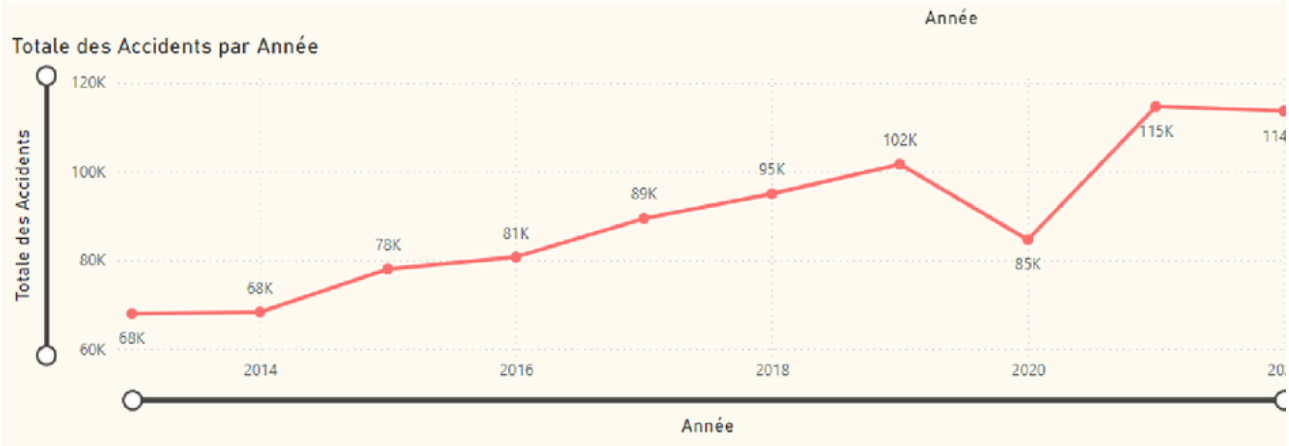


Figure 5 : L'augmentation des accidents mortels sur la Période 2013-2022.

On met en évidence une hausse alarmante des accidents de la route au Maroc, avec le nombre d'incidents ayant presque doublé, passant de près de 70 000 à environ 114 000 en quelques années seulement. L'année 2022 a été particulièrement dramatique, enregistrant 3438 morts, 10929 blessures graves et 154411 blessures légères.
Ces chiffres révèlent une augmentation préoccupante et soulignent les défis croissants en matière de sécurité routière dans le pays. Cette escalade illustre une urgence critique qui requiert une attention immédiate. La persistance de cette tendance alarmante soulève de sérieuses inquiétudes quant à la capacité à garantir la sécurité des usagers de la route, mettant en lumière la nécessité pressante d'adopter des actions concrètes pour inverser ce cours inquiétant.

2.Solution proposée :

Développer un modèle de prédiction avancé des risques d'accidents, basé sur l'analyse approfondie des données historiques locales (Hors Agglomération) et l'utilisation de technologies numériques telles que l'analyse de données et l'Apprentissage automatique.

3.Objectifs du projet

L'objectif principal du projet est de développer une solution prédictive qui permettra à NARSA de détecter et de classifier les risques d'accidents en quatre catégories distinctes : sécurité très mauvaise nécessitant une amélioration très urgente, sécurité mauvaise nécessitant une amélioration urgente, sécurité faible nécessitant une amélioration, et sécurité bonne. Cette classification s'appuie sur des techniques d'apprentissage automatique. Grâce à cette prédiction, nous visons à atteindre un ensemble d'objectifs stratégiques :

1.Prédiction des risques : Le modèle de prédiction avancé permet aux autorités d'anticiper les risques d'accidents pour une intervention préventive efficace, améliorant la sécurité routière.
2.Réduction des accidents : Identifier les facteurs de risque permet de réduire significativement le nombre d'accidents, contribuant ainsi à la sauvegarde des vies.
3.Utilisation efficace des ressources : Optimisation de l'allocation des ressources vers les zones à haut risque, maximisant l'efficacité des mesures de sécurité et minimisant les coûts.

Les contraintes:

1.Qualité des données : La précision des prédictions est conditionnée par la qualité et l'accessibilité des données historiques. Les données provenant de sources variées et hétérogènes peuvent compliquer la fiabilité des analyses.
2.Evolution des facteurs de risque : Les facteurs influençant les accidents de la route évoluent constamment, nécessitant des mises à jour régulières du modèle pour maintenir son efficacité.
3.Durée du projet : Le projet est prévu pour être réalisé dans un délai de deux mois, imposant un cadre temporel strict pour le développement et la mise en œuvre du modèle.

4.Périmètre de projet

Ce qui est inclus :

- Développement d'un modèle de prédiction avancé des risques d'accidents.
- Analyse approfondie des données historiques locales.
- Utilisation de technologies numériques telles que l'analyse de données et l'apprentissage automatique.
- Fourniture d'outils novateurs et précis aux autorités marocaines pour anticiper et prévenir les accidents de la route.
- Visualisation des données pour une meilleure compréhension des tendances et des corrélations.

Ce qui est exclu :

- Création d'un système complet de gestion de la sécurité routière.
- Développement d'autres fonctionnalités non liées à la prédiction des accidents.
- Intégration avec d'autres systèmes ou plateformes externes.
- Conception graphique complète du site web existant.

5.Parties prenantes du projet

Tableau 2 : Parties prenantes de projet

M.Ahmad Bardan	Encadrant externe Narsa : Chef de Département d'Observatoire National de Sécurité routière.
M.Achraf Touil	Encadrant interne et professeur chez FST
Chaymae Moudnib	Étudiante en Systèmes d'Information et Transformation Digitale à la Faculté des Sciences et Techniques de Settat.

III. Planification et organisation du projet

L'utilisation de la gestion de projet pour le projet de prévision des risques d'accident s'avérera avantageuse pour planifier, organiser, contrôler et coordonner les ressources afin d'atteindre avec succès les objectifs du projet.

1.Planification structurée Créez un plan complet qui décrit les ressources nécessaires, les étapes clés et les objectifs à atteindre .
2. Gestion efficace des risques: Atténuer les risques potentiels en les identifiant et en les évaluant .
3. Communication efficace : Assurer une communication claire et efficace des mises à jour du projet et des besoins en ressources à toutes les parties prenantes.

1. Planification prévisionnel: Diagramme de GANTT

Diagramme de Gantt est un outil visuel de gestion de projet qui illustre les tâches à accomplir sur une ligne de temps. Il aide à planifier et suivre l'avancement du projet en montrant les dates de début et de fin des tâches, ainsi que leur interdépendance. Cela permet de gérer efficacement les ressources et de respecter les délais.

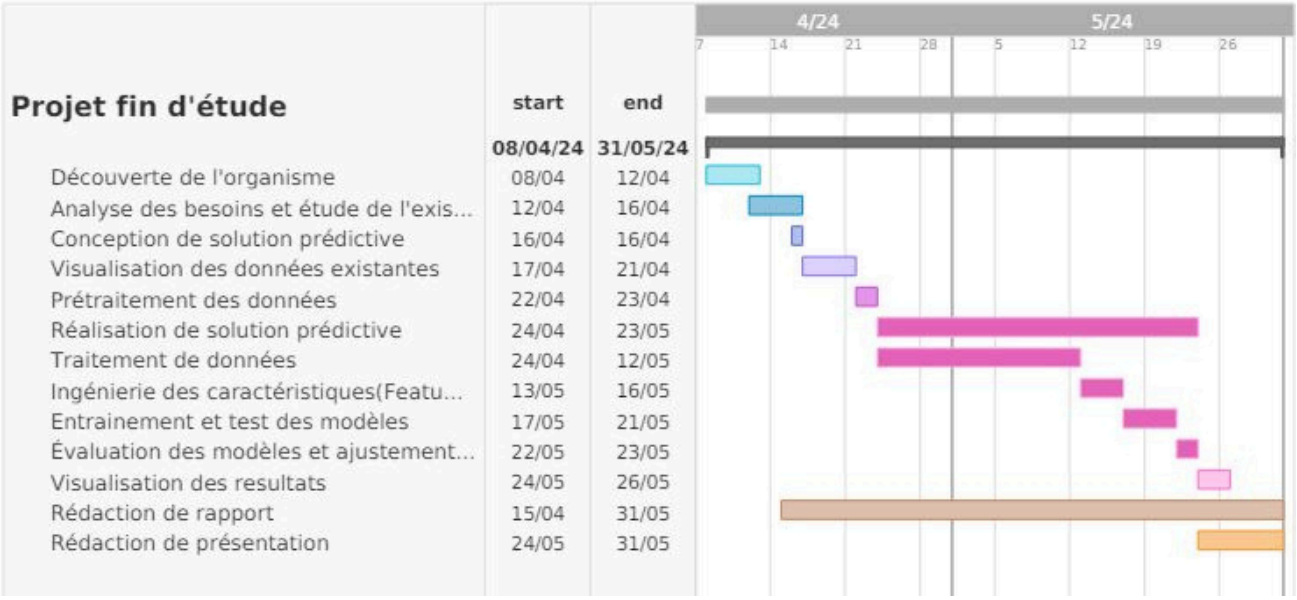


Figure 6: Diagramme de gantt

Conclusion

Le chapitre suivant, succédant à la définition des objectifs et d’élaboration de problématique ainsi que la méthodologie du projet, sera consacré à une analyse approfondie de la littérature existante. Cette revue vise à explorer et clarifier les concepts clés qui sous-tendent notre sujet d'étude.

Chapitre 2 : Revue en littérature

Pour explorer tous les aspects de la problématique abordée dans ce travail, nous examinerons les concepts et termes clés de la littérature pertinente. Cela inclut la nature des données, leur cycle de vie et leurs catégories, ainsi que des concepts tels que l'analyse des données, les types de données, l'analyse prédictive, l'apprentissage automatique, les étapes impliquées, les types d'apprentissage, les travaux connexes et les approches possibles basées sur ces travaux. Nous mettrons également l'accent sur certains aspects cruciaux de cette étude, tels que le traitement des valeurs manquantes et l'évaluation des performances des modèles de classification et de régression.

I.Introduction à l'Analyse des Données

1. Définition de l'Analyse des Données et son Importance

1.1 Qu'est-ce qu'une donnée ?

Selon Rob Kitchin professeur à l'Institut des Sciences Sociales de l'Université de Maynooth "Les données sont couramment comprises comme les matériaux bruts produits dans l'abstraction du monde en catégories, mesures et toute autre forme de représentation-nombres, caractères, symboles, images, sons, ondes électromagnétiques, bits qui constituent les fondations sur lesquelles l'information et le savoir sont créés."[1]
en conséquence , l'analyse approfondie et l'interprétation des données contribuent à une compréhension plus précise de l'environnement, renforçant ainsi la qualité des décisions prises dans divers domaines.

1.2 Cycle de Vie des Données[2]

Le cycle de vie des données désigne les différentes étapes par lesquelles les données passent depuis leur création jusqu'à leur suppression.

Figure 7 :Cycle de Vie des Données

[1] Rob Kitchin , The Data Revolution: Big Data, Open Data, Data Infrastructures and The Consequences,2014,
[2] Data life cycle , Google "Foundations: Data, Data, Everywhere"<https://www.coursera.org/learn/foundations-data?specialization=google-data-analytics> (la dernière consultation le 30/04/2024.)

1.3 Catégorie des Donnéess

Les données peuvent être classées en trois catégories principales selon leur degré de structuration:

- **Données Structurées :** Organisées clairement, souvent dans des bases de données relationnelles, et facilement accessibles via des langages de requête comme SQL, fichier Excel XSLX ou CSV.
- **Données Semi-Structurées:** Moins organisées que les données structurées, mais contenant des balises ou des marqueurs pour séparer les éléments, comme les fichiers XML et JSON.
- **Données Non Structurées:** Sans structure fixe, difficiles à traiter et à analyser, incluant des formats tels que les textes, images, vidéos et audios.

1.4 Qu'est-ce que l'Analyse des Données ?

L'analyse des données est un processus crucial qui implique l'examen, la collecte, le nettoyage, l'analyse et l'interprétation des données brutes pour en extraire des informations significatives. Ces informations sont essentielles pour prendre des décisions éclairées, identifier des tendances, comprendre les comportements et découvrir des opportunités ou des défis au sein d'une organisation.
L'importance de l'analyse des données dans le monde moderne est bien capturée par la citation de Peter Sondergaard : "L'information est le pétrole du 21e siècle, et l'analyse est le moteur à combustion." [3] Cette analogie souligne que, tout comme le pétrole a besoin d'être raffiné pour être utilisé efficacement, les données brutes doivent être analysées pour révéler leur valeur réelle.

2. Différents types d'analyses : descriptive, prédictive, prescriptive

L'analyse des données est un processus complexe qui englobe diverses méthodes pour transformer les données brutes en informations exploitables. Parmi ces méthodes:

Analyse Descriptive : Résume les données historiques pour décrire les événements passés, offrant un aperçu des performances et des tendances.
Analyse Diagnostique : Cherche à comprendre les causes sous-jacentes des événements ou des comportements observés.

[3] Peter Sondergaard <https://metropoleposition.fr/07-optimiser-la-gestion-de-ses-donnees/> (la dernière consultation le 30/04/2024.)

Analyse Prédictive : Dans son ouvrage "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die" [4], Eric Siegel, data scientist et auteur, souligne que prédire l'avenir repose sur l'art de comprendre le passé. Cette discipline utilise les données passées pour anticiper les événements futurs, permettant ainsi de formuler des prévisions précises grâce à des modèles statistiques et des algorithmes sophistiqués.
Analyse Prescriptive : Recommande des actions spécifiques pour atteindre des objectifs déterminés.

En combinant ces approches, l'analyse des données devient un outil puissant pour les organisations, leur permettant de comprendre le passé, d'analyser le présent et de planifier l'avenir avec des décisions basées sur des données solides. Ces analyses transforment les données en insights actionnables, essentiels pour la prise de décision stratégique et opérationnelle.

II.Analyse prédictive

Après avoir établi la définition de l'analyse descriptive, il devient clair qu'elle évolue vers une approche plus élaborée, offrant des modèles pouvant être exploités pour résoudre notre problème central. Ainsi, elle transcende les simples statistiques en intégrant des méthodes de prédiction telles que l'apprentissage automatique. Dans cette section, nous plongerons dans le domaine de l'apprentissage automatique et son incidence sur la sécurité routière.

2.1.Aprentissage automatique/Machine Learning

Le Machine Learning est un domaine de l'intelligence artificielle qui concerne le développement de techniques permettant aux ordinateurs d'apprendre à partir de données structurées, sans être explicitement programmés. Cette définition est bien illustrée par la citation de Tom Mitchell : "Un programme informatique apprend d'une expérience E par rapport à une tâche T et une mesure de performance P, si ses performances sur T, mesurées par P, s'améliorent avec l'expérience E." [5]

[4] Eric Siegel dans "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die",2013.
[5] Tom Mitchell ,"Machine Learning",1983.

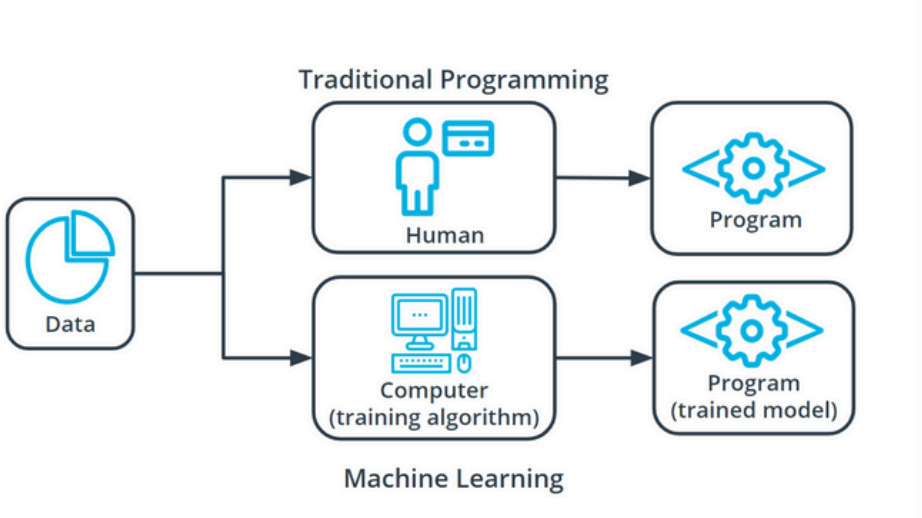


Figure 8 : L'Apprentissage automatique et la programmation traditionnel.

2.2 Les étapes de machine learning

Collection et importation de donnees	Préparation de données	Séparation de données	Entrainement de modèle	Evaluation de modèle	Testage de modèle

Figure 9 : Les étapes de machine learning

- **Préparation des données :** Cette étape essentielle implique le nettoyage, la transformation, l'analyse exploratoire (EDA pour Exploratory Data Analysis), la sélection des caractéristiques afin de les rendre optimales pour l'analyse.
- **Partitionnement des données** : Diviser les données en ensembles d'entraînement et de test pour évaluer les performances du modèle.
- **Entraînement du modèle :** Utiliser les données d'entraînement pour ajuster les paramètres du modèle et lui permettre d'apprendre des modèles à partir des données.
- **Évaluation du modèle :** Cette étape critique consiste à tester l'efficacité du modèle développé en utilisant diverses métriques de performance. Elle permet de vérifier la précision, la sensibilité, la spécificité et d'autres indicateurs pertinents pour évaluer la capacité du modèle à générer des prédictions fiables et précises. Cela inclut souvent la validation croisée et l'analyse des courbes ROC pour affiner et optimiser le modèle avant son déploiement final.
- **Test du modèle :** Tester la capacité du modèle à prédire de manière précise sur des données complètement nouvelles à partir de l'ensemble de test.

2.3 Les types de machine learning

1. Apprentissage supervisé

Le modèle apprend de données étiquetées (input) pour prédire des résultats (output) sur de nouvelles données. Il inclut :

- **Classification** : Cette technique permet de déterminer la catégorie d'un élément en fonction de l'input fourni. Par exemple, dans un contexte général, elle peut classifier un email comme spam ou non-spam. Appliquée à notre projet, la classification est utilisée pour identifier des situations potentiellement dangereuses ou des conditions de conduite à risque. Cette capacité de discrimination aide à anticiper des zones ou des moments où le risque d'accident est élevé, permettant ainsi des interventions ciblées pour améliorer la sécurité routière.

1. Classification Binaire : Distingue deux classes, comme déterminer si une route est dangereuse ou non.
2. Classification Multiclasse : Identifie plusieurs catégories distinctes, telles que les types des risques d'accident sur la route .
3. Classification Multi-label : Permet à un input d'appartenir à plusieurs classes simultanément.

- **Régression** : Prédit une valeur numérique continue pour un input (par exemple, estimer le prix d'une maison).

2. Apprentissage non supervisé

Le modèle analyse les données non étiquetées (input) pour découvrir des patterns ou des structures (output).

3. Apprentissage semi-supervisé

Utilise à la fois des données étiquetées et non étiquetées (input) pour améliorer les prédictions de résultats (output).

4. Apprentissage par renforcement

Le modèle interagit avec un environnement (input) en prenant des décisions pour maximiser une récompense (output).

2.4.Le choix d'algorithme utilisée pour l'apprentissage supervise:

Pour construire un modèle efficace, il est primordial de sélectionner soigneusement l'algorithme approprié, en tenant compte des critères de choix les plus pertinents.

1-Type de problème :

Classification : Prédiction d'une catégorie pour un input.
Modèles recommandés : Forêts aléatoires, machines à vecteurs de support (SVM), réseaux de neurones, régression logistique.
Régression : Prédiction d'une valeur continue.
Modèles recommandés : Régression linéaire, régression ridge, régression Lasso, forêts aléatoires, réseaux de neurones.
2-Taille et qualité des données :
Grand volume de données : Forêts aléatoires, réseaux de neurones, SVM.
Données bruitées ou avec valeurs manquantes : Forêts aléatoires (robustes face au bruit), techniques d'imputation pour préparer les données.
3**-Performance requise :**
Haute précision : SVM, réseaux de neurones.
Rapidité d'exécution : Régression logistique, arbres de décision.
Robustesse : Forêts aléatoires (peu sensibles aux variations dans les données d'entraînement).

2.5.Example des algorithmes d'apprentissage supervise:

2.5.1.Forêt aléatoire (Random Forest)

La forêt aléatoire, conceptualisée par Breiman en 2001, est une technique d'apprentissage supervisé adaptée tant aux tâches de classification qu'aux problèmes de régression. Ce modèle puissant repose sur le concept d'apprentissage en ensemble, qui consiste à fusionner les prédictions de plusieurs modèles pour en augmenter la précision et la fiabilité.

Dans la forêt aléatoire, l'algorithme construit un ensemble d'arbres de décision [6], chacun étant formé sur des sous-ensembles différents et aléatoires de l'ensemble de données initial. Cette stratégie est renforcée par la méthode de "Bagging" (Bootstrap Aggregating), qui vise à réduire la variance des prédictions en entraînant chaque arbre sur des échantillons aléatoires tirés avec remplacement.

[6] Pour plus de détails voir : < https://dataaspirant.com/how-decision-tree-algorithm-works/ >(la dernière consultation le 04/05/2024.)
Bagging : C'est un acronyme de **Bootstrap aggregation** est utilisé pour créer plusieurs arbres de décision, chacun formé sur un échantillon différent des données. Les résultats de tous les arbres sont ensuite combinés pour produire une prédiction finale plus précise et moins sujette au sur ajustement.

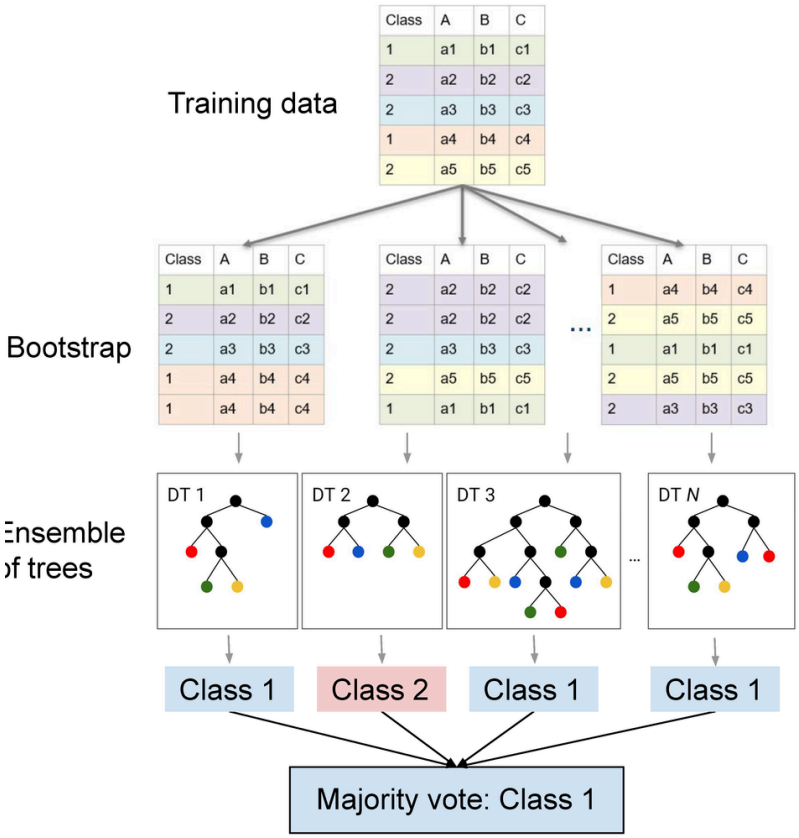


Figure 10: Algorithme de Random forest .

Pour minimiser la corrélation entre les arbres, la forêt aléatoire implémente deux approches clés :

- **Diversité des variables(le feature sampling)** : Chaque arbre est développé en utilisant un sous-ensemble aléatoire de caractéristiques, ce qui permet à différents arbres de se spécialiser dans différentes dimensions des données.
- **Diversité des échantillons(le tree bagging)** : Les arbres sont entraînés sur des échantillons distincts, assurant que les erreurs d'un arbre sont compensées par les autres.

Comment cela fonctionne?

Algorithme de Forêt aléatoire :[7]

1.Échantillonnage de données :
Nous commençons par prélever aléatoirement des échantillons de la base d'apprentissage Zi, avec remplacement. Chaque Zi est un ensemble de données distinct contenant n points, et nous créons B de ces échantillons.
2.Construction d'arbres CART :
Pour chaque échantillon Zi, nous construisons un arbre de décision CART $Gi(x)$. À chaque étape de division du nœud dans l'arbre, nous choisissons aléatoirement un sous-ensemble d'attributs parmi les p attributs disponibles. Cela permet d'introduire de la variabilité dans chaque arbre.

[7],Breiman et al. Random forests, Machine Learning, 45, 2001.

CART Gi(x):représente la prédiction pour un point de données particulier x. C'est la valeur prédite pour la cible (classe dans le cas de la classification, ou valeur continue dans le cas de la régression) en utilisant l'arbre de décision entraîné sur l'ensemble des données d'apprentissage.

3.Sélection des meilleures divisions :

Lorsque nous devons séparer un nœud dans un arbre, nous sélectionnons aléatoirement q attributs parmi les p disponibles et choisissons la meilleure division possible dans ce sous-ensemble d'attributs.

Pour un cas de :

- **Régression** : agrégation par la moyenne $G(x) = \frac{1}{B}\sum_{i=1}^{B} G_i(x).$
- **Classification**: agrégation par vote $G(x) = \text{Vote majoritaire}\,(G_1(x), \ldots, G_B(x)).$

Paramètres (valeurs par défaut) :

- **Classification** : q=√p , taille partition minimale 1 .
- **Régression** : q=p/3 , taille partition minimale 5.

2.5.2.KNN (K-Nearest Neighbors)

Le modèle K-Nearest Neighbors (KNN), développé en 1967 par Cover et Hart, est une méthode d'apprentissage supervisé fréquemment utilisée pour les tâches de classification et de régression. Ce modèle simple mais efficace repose sur le principe de proximité, où les prédictions sont basées sur la majorité des labels ou la moyenne des valeurs des k plus proches voisins.

Dans le KNN, l'algorithme identifie les 'k' points de données les plus proches d'une observation donnée dans l'espace des caractéristiques, souvent à l'aide de métriques de distance telles que la distance euclidienne. Les prédictions sont ensuite faites en fonction des labels ou des valeurs de ces voisins. Cette approche directe permet au KNN de s'adapter dynamiquement à l'évolution des données, mais requiert une sélection judicieuse de 'k' et une bonne normalisation des données pour optimiser sa performance et éviter des prédictions biaisées par des échelles de caractéristiques disproportionnées.[8]

Algorithme de KNN

.

1.Calcul de la Distance

:

La distance entre deux points xi et xj est généralement calculée en utilisant la distance Euclidienne, bien que d'autres métriques puissent également être utilisées telles que la distance de Manhattan ou la distance de Minkowski.

[8] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

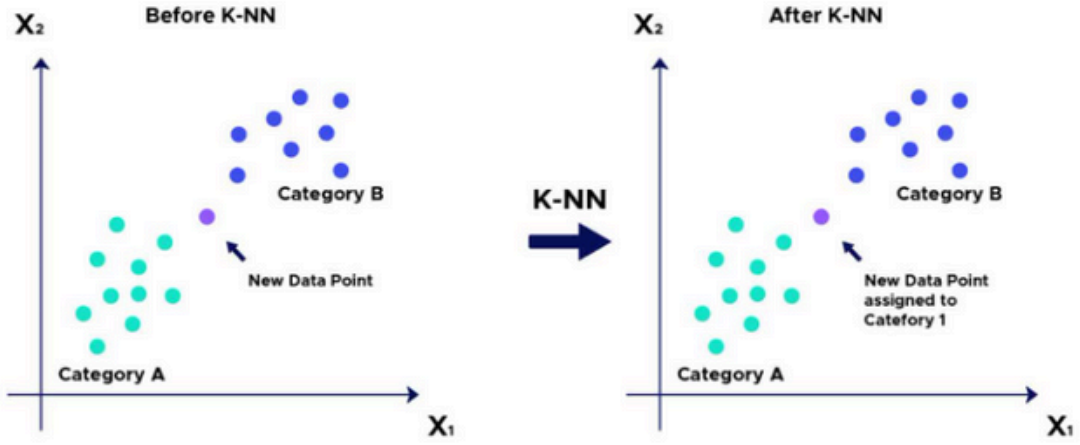


Figure 11 : Algorithme de KNN

La distance Euclidienne entre deux points dans un espace à n dimensions est définie par: :

$$Distance_E(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

La distance Manhattan: $$Distance_M(X,Y) = \sum_{i=1}^{n}|x_i - y_i|$$

La distance Minkowski: $$Distance_M(X,Y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

2.Détermination des Voisins :

Les 'k' points les plus proches de l'observation à prédire sont sélectionnés en fonction des valeurs les plus faibles de la distance calculée.

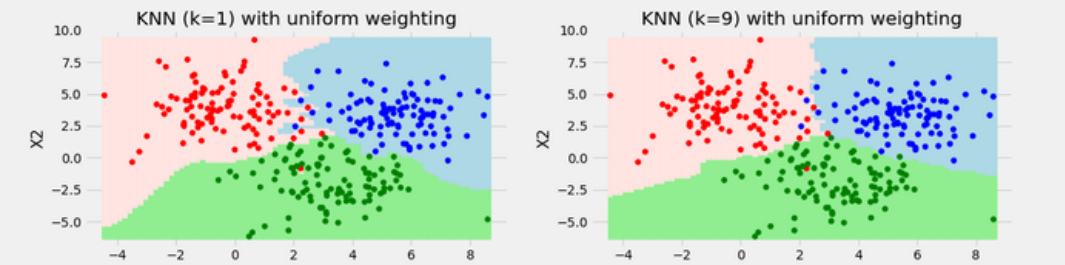


Figure 12 : Algorithme de KNN avec K=1 et K=9

3.Prédiction :

- **Classification** : La classe prédite est souvent déterminée par un vote majoritaire parmi les 'k' voisins. Chaque voisin vote pour sa classe et la classe avec le plus grand nombre de votes est choisie comme prédiction.
  Prediction= mode ($\{yi \mid xi \in$ k plus proches voisins$\}$)

- **Régression** : La valeur prédite est le résultat de la moyenne des valeurs observées chez les 'k' voisins.
  Prediction=$1/k \sum yi($ $xi \in$ k plus proches voisins)

2.5.3.ANN(Réseau de neurones artificiels)

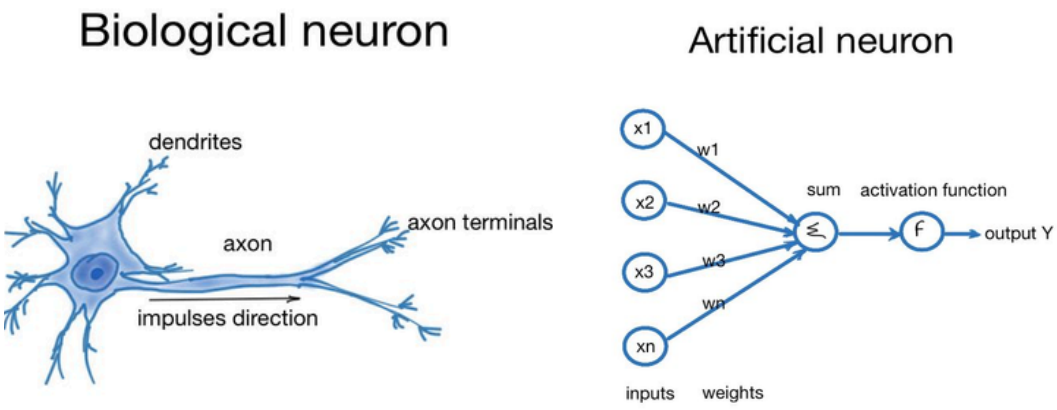


Figure 13 :Neurone biologique et le Neurone artificiel .

Les réseaux de neurones artificiels (ANN), développés à partir des travaux pionniers de McCulloch et Pitts en 1943, une méthode qui imite le cerveau humain et tente de reconnaître les principales relations dans l'ensemble de données utilisé en prétendant être celui-ci.[9]

[9] Koklu, M., I. Cinar, and Y.S. Taspinar, Classification of rice varieties with deep learning methods. Computers and electronics in agriculture, 2021. 187: p. 106285.

La structure d'un réseau de neurones artificiels:

Nœuds ou neurones : Les composants fondamentaux de l'ANN sont les neurones, similaires aux neurones biologiques. Chaque neurone est doté d'entrées, qu'il traite et génère un résultat. Couches:

- **Couche d'entrée :** couche qui reçoit la première entrée. Chaque neurone de cette couche représente un composant des données d'entrée.
- **Couche cachés :** traits cachés sous la couche d'entrée et la couche de sortie. Les neurones de ces couches traitent les données de la couche précédente et transmettent les informations traitées à la couche suivante.
- **Couche de sortie :** couche finale qui produit la sortie. Le nombre de neurones dans cette couche est égal au nombre de sorties que le modèle est censé produire.

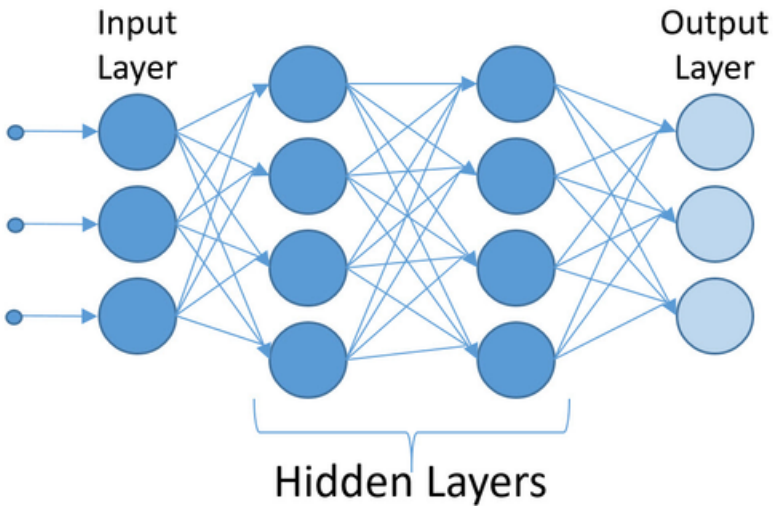


Figure 14 : Les couches de neurone.

Comment Fonctionnent les ANN ? [10]

1.Initialisation:

- Poids: Chaque connexion entre les neurones a un poids qui représente la force de la connexion. Initialement, ces poids sont généralement définis à de petites valeurs aléatoires.
- Bias: Chaque neurone a également un biais qui est ajouté à la somme pondérée des entrées avant de passer par la fonction d'activation.

[10] Pour plus d'information : <https://aws.amazon.com/fr/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain.> (la dernière consultation le 10/05/2024).

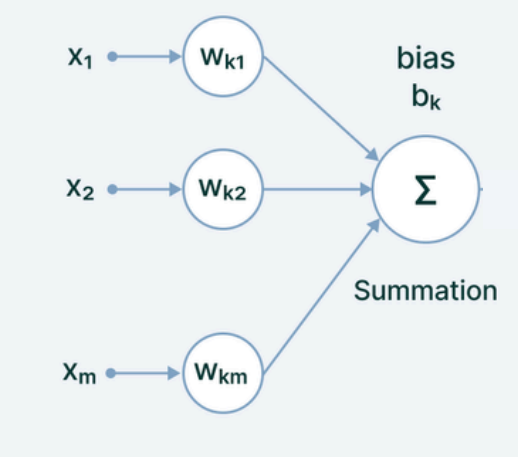


Figure 15 :Initialisation ANN

2.Propagation Avant (Forward Propagation):

- Les entrées sont alimentées dans la couche d'entrée.
- Chaque neurone dans une couche calcule la somme pondérée de ses entrées plus le biais.
- Le résultat est ensuite passé par une fonction d'activation (comme ReLU, sigmoid ou tanh) pour introduire une non-linéarité et décider si le neurone doit être activé ou non.

Ce processus continue de la couche d'entrée, à travers les couches cachées, jusqu'à la couche de sortie, produisant la sortie du réseau.

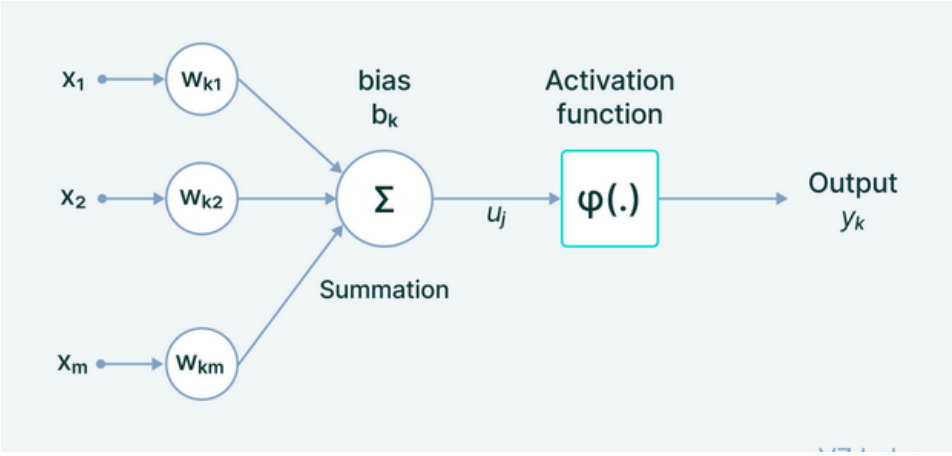


Figure 16 : Algorithme de ANN

[10] Pour plus d'information : <https://aws.amazon.com/fr/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain.> (la dernière consultation le 10/05/2024).

3.Fonctions d'Activation:

Les fonctions d'activation aident à introduire des non-linéarités dans le réseau, lui permettant de modéliser des relations complexes. Les fonctions d'activation courantes incluent:

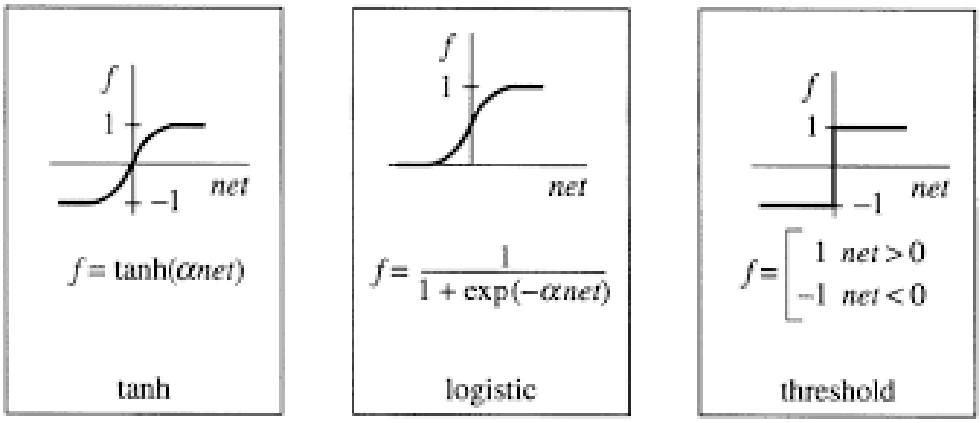


Figure 17 :Les différents fonctions d'activation

4.Fonction de perte :

- La fonction de perte mesure la différence entre la sortie prévue et la sortie réelle. Les fonctions de perte courantes incluent l'erreur quadratique moyenne (MSE) pour la régression et la perte d'entropie croisée pour la classification.

5.Rétropropagation :

- C'est le (Backpropagation) en anglais est le processus de mise à jour des poids et des biais pour minimiser la fonction de perte. Cela implique d'utiliser la règle de calcul en chaîne pour calculer le gradient de la fonction de perte par rapport à chaque poids et biais. Le gradient est ensuite utilisé pour ajuster les poids et les biais dans la direction opposée du gradient (descente du gradient) afin de réduire la perte.

2.6.Ensemble learning(Stacking)

Le stacking est une méthode d'apprentissage par ensemble où les prédictions de plusieurs modèles différents sont utilisées comme entrées pour un nouveau modèle, souvent appelé métamodèle ou modèle de niveau supérieur. Ce métamodèle est alors entraîné pour faire la prédiction finale, en exploitant les différentes perspectives apportées par les modèles initiaux.

Fonctionnement de cette type d'ensemble learning

1.Création de Modèles de Base : Plusieurs modèles prédictifs sont formés indépendamment. Ces modèles peuvent être de types très variés.
2.Prédiction des Modèles de Base : Chaque modèle de base fait une prédiction sur les données. Ces prédictions sont souvent conservées comme de nouvelles caractéristiques pour le prochain niveau de modélisation.
3.Entraînement du Métamodèle : Un nouveau modèle est formé sur les prédictions des modèles de base. Ce modèle apprend à optimiser la combinaison des prédictions des modèles de base pour améliorer la précision globale.

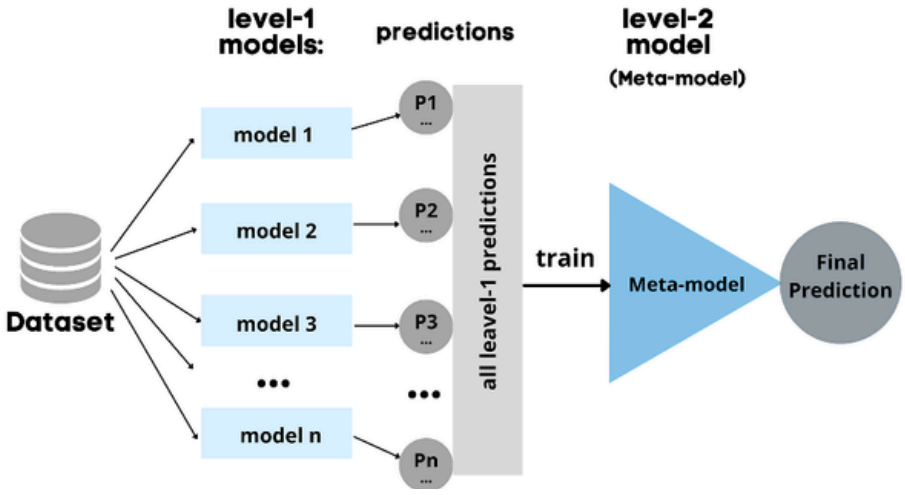


Figure 18 : Fonctionnement d'esenmble learning (stacking)

2.7.Application du machine learning dans le domaine de la sécurité routière : travaux connexes :

Au fil des dernières années, les accidents de la route se sont imposés comme une problématique majeure de santé publique à l'échelle mondiale. Ces incidents sont influencés par divers facteurs, tels que le comportement des conducteurs, les conditions environnementales et les caractéristiques des véhicules, certains ayant un impact plus prononcé sur la gravité des accidents. Poussés par la volonté de réduire ces événements tragiques, des chercheurs exploitent activement des techniques d'analyse prédictive et de fouille de données.

Ces approches visent à soutenir la prise de décision éclairée pour prévenir de nouveaux accidents, améliorer les systèmes de transport et élaborer des normes avancées pour la sécurité routière intelligente. Parmi ces recherches:

Dans leur étude, **Qasem A. Al-Radaideh (2018)** [11] a utilisé trois méthodes de classification différentes pour identifier les facteurs environnementaux qui contribuent aux accidents de la route. Les techniques utilisées étaient des arbres de décision (y compris RF, DT, J48/C4.5 et CART), des réseaux de neurones artificiels (en particulier la rétropropagation) et des machines vectorielles de support avec un noyau polynomial. Le but de cette recherche était de développer un modèle prédictif basé sur ces caractéristiques influentes. Pour valider l'efficacité de ces techniques, un véritable ensemble de données a été collecté et utilisé dans l'analyse. Les résultats de l'expérience menée par le ministère britannique des Transports ont révélé que La technique RF a atteint le plus haut niveau de précision, atteignant **80,6%**.[11]

Elle a été suivie par la méthode ANN, qui a atteint une précision de **61,4%** , et la technique SVM, qui a atteint une précision de **54,8%**. Pour améliorer le processus décisionnel et prédire la gravité des accidents, un système de décision a été établi à l'aide du modèle généré par la technique **RF**.

Un récent travail réalisé par **Gyanendra Singh(2022)** propose l'utilisation de réseaux neuronaux profonds (**DNN**) pour prédire les accidents de la route. Le modèle DNN, comprenant deux couches cachées ou plus avec de nombreux nœuds, a été appliqué à des données de 2680 accidents provenant de sections non urbaines de huit autoroutes. Les données comprenaient 16 variables explicatives liées à la géométrie de la route, au trafic et à l'environnement, collectées à partir de dossiers officiels et d'études de terrain.
Pour l'évaluation, 148 des 222 points de données sur la fréquence des accidents ont été utilisés pour l'entraînement, et les 74 restants pour tester les modèles.Le DNN a été comparé aux modèles de programmation d'expression génétique (GEP) et binomiale négative à effet aléatoire (RENB). Le DNN a obtenu un coefficient de corrélation de 0,945 (erreur quadratique moyenne = 5,908), surpassant le GEP (0,914, RMSE = 7,474) et le RENB (0,891, RMSE = 8,862), démontrant ainsi une meilleure performance pour la prédiction des accidents de la route. Bien que le GEP ait eu un coefficient de corrélation inférieur, il a quantifié les effets des différentes variables et fourni une liste classée de leur importance. [12]

[11] Q. A. Al-Radaideh et E. J. Daoud, « Data Mining Methods for Traffic Accident Severity Prediction ».

[12] Singh, G., Pal, M., Yadav, Y. et al. Deep neural network-based predictive modeling of road accidents. [5]Neural Comput & Applic 32, 12417–12426 (2020). < https://doi.org/10.1007/s00521-019-04695-8>(la dernière consultation le 11/05/2024).

[12] Pour plus de détail voir : < https://www.mdpi.com/2412-3811/5/7/61#B10-infrastructures-05-00061 > (la dernière consultation le11/05/2024).

2.8.Analyse des valeurs manquantes

Le traitement des valeurs manquantes est une étape cruciale dans la construction de modèles prédictifs. Éliminer des données peut parfois signifier la perte d'informations importantes, tandis que l'imputation peut entraîner des estimations incorrectes. Alors, comment traiter les données manquantes ?
Pour traiter efficacement l'imputation des données manquantes, il est essentiel de comprendre leurs causes, en particulier si elles ne sont pas aléatoires. Little & Rubin (1987) ont développé une typologie classant ces causes en trois catégories distinctes [13] :

1. **MCAR (Manquantes Complètement au Hasard)** : Les données manquantes de manière totalement aléatoire, sans relation avec aucune autre variable.
2. **MAR (Manquantes au Hasard)** : Les données manquantes de manière aléatoire mais sont liées à d'autres variables observées.
3. **MNAR (Manquantes Non au Hasard)** : Les données manquantes de manière non aléatoire et sont liées à la valeur de la variable manquante elle-même.

Traitement des données manquantes:

1.Suppression des Données : Cette méthode est utilisée uniquement lorsque le nombre de valeurs manquantes est faible par rapport à l'ensemble des données, afin d'éviter de biaiser le modèle, souvent utilise avec le type MCAR.
2.Suppression des variables: Cette méthode est utilisée uniquement lorsque si elles contiennent un grand nombre de valeurs manquantes et si leur suppression n'affecte pas significativement les performances du modèle.
3.Imputation des Données : Consiste à remplacer les valeurs manquantes par des estimations plausibles basées sur les autres informations disponibles et est utilisée lorsqu'il a une quantité faible à modérée de valeurs manquantes dans nos données.

- **Imputation par la moyenne/médiane/mode** : Remplacer les valeurs manquantes par la moyenne, la médiane ou la mode des données non manquantes. Cela est simple et rapide, mais peut réduire la variance.
- **Imputation par les K plus proches voisins (KNN) :** Utiliser des algorithmes comme KNN pour imputer chaque observation avec des valeurs manquantes est par la moyenne ou la médiane des k observations les plus proches.
- **Imputation par MissForest :** Stekhoven et Bühlmann (2011)[14] ont proposé une méthode de complétion basée sur les forêts aléatoires appelée missForest. Une librairie R éponyme lui est associéeUtiliser, l'algorithme MissForest qui emploie des forêts aléatoires pour imputer les valeurs manquantes de manière itérative, en tenant compte des relations complexes entre les variables.

[13] Little R.J.A. et Rubin D.B., Statistical Analysis with Missing Data, Wiley series in probability and statistics, 1987.

[14] Stekhoven D.J. et Bühlmann P., MissForest - nonparametric missing value imputation for mixed-type data, Bioinformatics Advance Access (2011).

2.9.Selection des attributs(Feature selection)

La sélection de caractéristiques est comparable à le choix des joueurs pour une équipe de football, où nous choisissons soigneusement les joueurs en fonction de leurs compétences spécifiques pour chaque position sur le terrain. De même, lorsqu'il s'agit de tâches de prédiction, nous sélectionnons les fonctionnalités les plus pertinentes pour simplifier le modèle et améliorer sa capacité à généraliser à des données inconnues. Cette approche réduit non seulement la complexité de calcul, mais facilite également l'interprétation du modèle et évite le surajustement, améliorant ainsi ses performances globales. De plus, en minimisant le nombre de fonctionnalités, nous pouvons atténuer le risque de surajustement, permettant ainsi au modèle de gérer efficacement des ensembles de données plus petits et de générer des prédictions plus fiables.

- **Corrélation :** La corrélation mesure la relation entre deux variables, indiquant si elles varient ensemble de manière positive ou négative. Elle aide à identifier les variables importantes pour les modèles prédictifs, avec un coefficient proche de 1 ou -1 signalant une forte corrélation. Cependant, il est important de noter que la corrélation ne signifie pas causalité, donc des analyses supplémentaires sont nécessaires pour établir des relations causales.

- **ANOVA(Analysis of Variance) F-score**: est utilisée en sélection de caractéristiques pour évaluer l'importance de chaque caractéristique par rapport à une variable cible. Elle compare les moyennes des valeurs de chaque caractéristique à la fois à l'intérieur des groupes et entre les groupes de la variable cible. Si la variation des valeurs est faible à l'intérieur des groupes mais élevée entre les groupes, cela indique une différence significative entre les groupes pour cette caractéristique, la rendant importante. En revanche, un faible F-score indique que la caractéristique n'est pas importante.

Les caractéristiques avec des F-scores élevés sont sélectionnées pour former un sous-ensemble plus petit et plus informatif de caractéristiques pour la construction du modèle final.

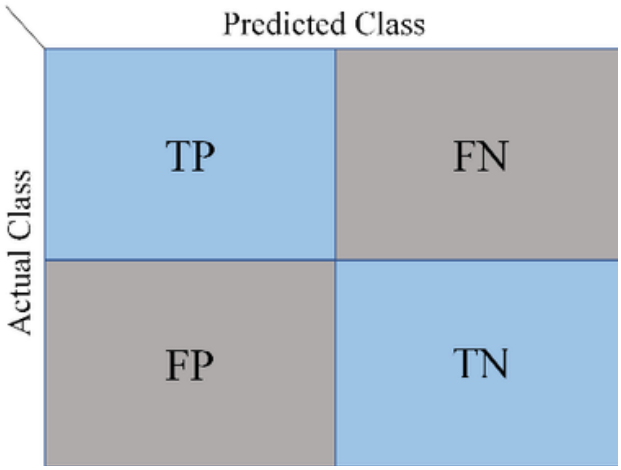
$$\text{F-score} = \frac{\text{Variance entre les groupes}}{\text{Variance dans les groupes}}$$

2.10. Evaluation de performance des modèles prédictifs(Indicateurs de performances) Cas de classification :

Dans les méthodes d'apprentissage automatique, diverses métriques de performance sont utilisées pour évaluer leur efficacité. Pour les problèmes de classification, ces métriques sont calculées à partir des valeurs obtenues via la matrice de confusion [15]. Cette étude a utilisé les métriques de performance suivantes : l'exactitude, la précision, le rappel, le score F-1, la spécificité, L'AUC (Aire Sous la Courbe) .

- **Matrice de confusion:**

La matrice de confusion est un tableau qui décrit la performance d'un modèle de classification. Ce tableau indique le nombre de prédictions correctes et incorrectes par rapport aux valeurs réelles dans un ensemble de données de test. Le tableau suivant illustre une matrice de confusion pour un modèle de classification binaire.



Échelle : 0 à 1
Une valeur plus proche de 1 indique une Précision, Recall , F1 score , Specificity, Accuracy maximale.

Figure 19 : Matrice de confusion

- **Vraie positive (True Positive, TP)** : nombre de positifs correctement classés comme positifs.
- **Fausse positive (False Positive, FP)** : nombre de négatifs incorrectement classés comme positifs.
- **Vraie négative (True Negative, TN)** : nombre de négatifs correctement classés comme négatifs.
- **Fausse négative (False Negative, FN)** : nombre de positifs incorrectement classés comme négatifs.

[15] Ropelewska, E., X. Cai, Z. Zhang, K. Sabanci, and M.F. Aslan, Benchmarking Machine Learning Approaches to Evaluate the Cultivar Differentiation of Plum (Prunus domestica L.) Kernels. Agriculture, 2022. 12(2).<https://www.mdpi.com/2077-0472/12/2/285> (la dernière consultation le 15/05/2024).

Précision (Precision) : Mesurant la proportion d'échantillons réellement positifs parmi ceux prédits comme positifs.

$$precision = \frac{TP}{TP + FP}$$

Rappel (Recall) : Mesurant la capacité du modèle à identifier tous les échantillons positifs.

$$recall = \frac{TP}{TP + FN}$$

F1 Score : La moyenne harmonique de la précision et du rappel, offrant un équilibre entre les deux mesures.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

La spécificité (Specificity) : Mesure la classification correctement des vrais négatifs parmi tous les échantillons négatifs réels.

$$specificity = \frac{TN}{TN + FP}$$

L'exactitude (Accuracy) : Mesure la proportion d'échantillons correctement classés parmi tous les échantillons.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

L'**AUC (Area Under the Curve)** pour la courbe **ROC (Receiver Operating Characteristic)** est calculée en mesurant l'aire sous la courbe qui trace le taux de vrais positifs contre le taux de faux positifs à différents seuils. Cette mesure, qui varie de 0 à 1, évalue la capacité du modèle à distinguer entre les classes, avec 1 indiquant une distinction parfaite et 0.5 une performance aléatoire. La méthode des trapèzes est souvent utilisée pour estimer cette aire, reflétant ainsi l'efficacité globale du modèle de classification.

Conclusion

Dans ce chapitre, nous avons exploré divers concepts fondamentaux liés à notre sujet, notamment l'analyse des données, l'apprentissage automatique, le traitement des valeurs manquantes, la sélection des caractéristiques, et le déséquilibrage des classes. Nous avons également abordé l'évaluation des modèles à l'aide d'indicateurs de performance, ainsi que le fonctionnement de différents modèles de classification tels que la forêt aléatoire (RF), les réseaux de neurones artificiels (ANN), et les k-plus proches voisins (KNN). En outre, nous avons examiné des travaux connexes sur la sécurité routière utilisant l'apprentissage automatique. Cette exploration a été soutenue par une documentation riche et variée, facilitant la compréhension de chaque concept utilisé dans la réalisation de notre projet de fin d'étude.

Chapitre 3 :
Analyse Intégrée des Besoins et de l'Existant avec Analyse Prédictive des Risques d'Accidents Routiers

Ce chapitre présente les concepts clés et définit les besoins précis de notre projet. Nous débutons par une analyse approfondie des données existantes pour garantir une compréhension solide, essentielle à une manipulation efficace des informations tout au long du projet. La conception de notre projet est détaillée pour faciliter le développement de la solution prédictive et l'application des techniques de visualisation des données, assurant une interprétation correcte et accessible des résultats.

Nous abordons ensuite l'analyse prédictive des risques d'accidents routiers, en mettant l'accent sur des techniques avancées pour préparer et exploiter efficacement les données. Les méthodologies classiques de traitement des données et l'ingénierie des caractéristiques sont explorées, essentielles pour améliorer les modèles prédictifs en développant des variables plus informatives.

La gestion des valeurs aberrantes et des valeurs manquantes est discutée, étant cruciales pour purifier les données et extraire des insights plus fiables. Nous explorons l'union de données provenant de différentes sources pour enrichir notre jeu de données, augmentant ainsi la robustesse et la précision de nos analyses. L'entraînement des données individuelles et l'application de la méthode de (stacking) sont examinés pour optimiser la formation du modèle, en veillant à ce que chaque observation contribue adéquatement. Nous incluons également l'utilisation de métriques d'évaluation pour évaluer notre modèle, suivi de la sauvegarde du modèle, et nous discutons les résultats élaborés par ce dernier.

En outre, nous explorons les outils et les langages de programmation qui soutiendront toutes les phases du projet, de la planification à la réalisation. Ce chapitre établit une base solide, assurant que chaque étape est alignée avec les objectifs globaux et bien préparée pour les développements futurs. Cette approche intégrée vise à enrichir notre compréhension des dynamiques des accidents routiers et à améliorer la précision des prédictions grâce à des données bien traitées, analysées et intégrées.

I. Analyse des besoins

L'analyse des besoins englobe une évaluation réaliste de nos exigences et l'examen approfondi des données existantes pour élaborer une solution prédictive optimale. Cette démarche nécessite une exploration détaillée de notre base de données pour distinguer les informations cruciales des éléments superflus, ce qui est essentiel pour prédire les zones à risque d'accidents, classifier les routes et visualiser ces risques sur une carte. Il est également impératif de maîtriser les principes de base de la sécurité routière pour assurer l'efficacité de notre projet.

1.Concept général liée au contexte

Accident: Dans le domaine de la sécurité routière peut être défini comme un événement non intentionnel impliquant au moins un véhicule en mouvement sur une voie de circulation, entraînant des conséquences dommageables telles que des blessures corporelles, des pertes de vie, des dommages matériels ou une combinaison de ceux-ci. Ces événements sont souvent le résultat de divers facteurs interagissant, tels que des erreurs humaines, des conditions routières défavorables, des défaillances mécaniques ou des comportements irresponsables.

Caractéristiques de l’accident : Cette section décrit les causes générales de l’accident, regroupant les variables générales telles que la date, les conditions lumineuses et l'état de la chaussée. Elle vise à fournir un aperçu global de l'événement.

- **Lieux :** Dans cette rubrique sont recueillies les informations relatives à l'infrastructure routière, incluant les voies empruntées par les véhicules et les usagers impliqués. Elle comprend des détails sur la route, le numéro de voie, le point kilométrique, l'accotement, la topographie, l'environnement et la signalisation.
- **Véhicule:** Cette partie présente les détails des véhicules impliqués, notamment leur type, leur utilisation et d'éventuels défauts mécaniques. Elle comprend également des informations sur les conducteurs ou les cyclistes, telles que l'âge, la profession, le port de la ceinture ou du casque, ainsi que des facteurs physiques.
- **Passagers victimes** Cette section détaille les caractéristiques des victimes , incluant leur position dans le véhicule, âge, et profession, et catégorise leur état en deux variables principales : "tué", subdivisée en "tué sur le champ" et "tué pendant le transfert à l’hôpital", et "blessé", distingué entre "blessé grave" nécessitant hospitalisation et "blessé léger" nécessitant des soins mineurs.

La rubrique des piétons spécifie des données similaires pour les piétons impliqués, ajoutant leur sexe, âge, profession, état (tué ou blessé) et facteurs physiques impactant l'accident.

- **Conducteur:** Cette section spécifique fournit des détails sur les conducteurs des véhicules impliqués, y compris leur âge, leur profession, faute liée aux conducteur ,leur état de fatigue éventuel et d'autres facteurs pertinents qui pourraient avoir influencé leur implication dans l'accident.
- **Circonstances de l'accident:** Cette rubrique englobe les détails sur les circonstances de l'accident, notamment le type de collision, les obstacles rencontrés, le point de choc initial, etc. Chaque rubrique est précédée d'un identifiant spécifiant l'unité des forces de l'ordre à l'origine du formulaire.

Voici quelques exemples des champs de ces caractéristiques.

Tableau 3 : Caractéristiques de l'accident

Rubriques	Champs
Caractéristiques de l'accident	Date d'accident ,Luminosité , Etat de chaussée,
Lieux	Numéro de route, Agglomération/ hors agglomération .
Véhicule	Type de véhicule , numéro d'ordre véhicule ,
Conducteur	Faute conducteur , sex , age .
Passagers victimes ,Piétons	Sex, age , profession .

Risque d'accident : Afin de développer une compréhension approfondie des objectifs de notre projet, nous avons consulté des experts de la NARSA pour identifier les facteurs clés qui définissent les causes de risques d'accidents. Cette consultation a mis en lumière que les principaux éléments contribuant aux accidents forment une pyramide, où chaque sommet représente un facteur critique : le comportement du conducteur, les conditions environnementales, et l'état du véhicule.

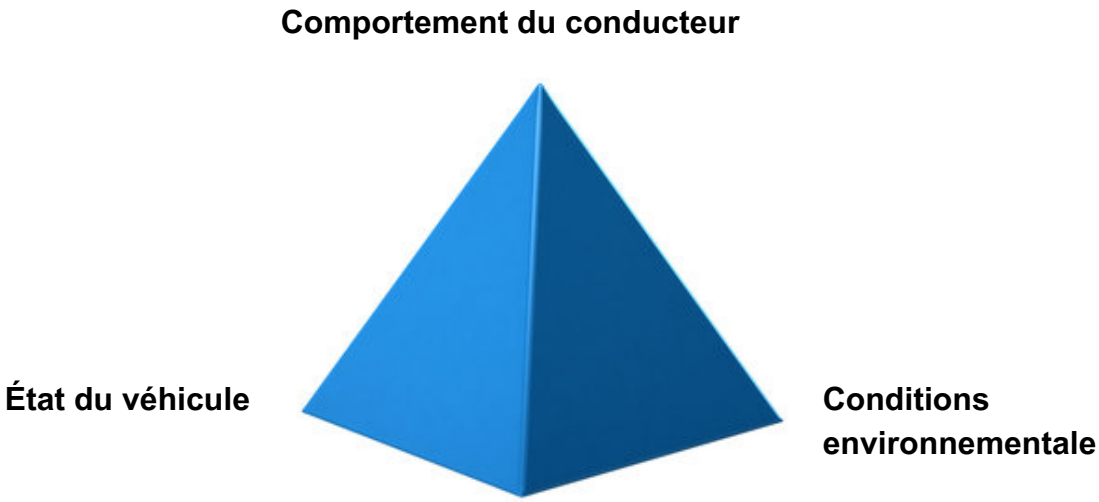


Figure 20 : Pyramide de Risque d'accident

CLASSE_ISR: Notre étude se basera sur une méthode de classification de la dangerosité des sections de routes développée par le programme EURORAP (European Road Assessment Programme). Cet indicateur prend en compte les statistiques d'accidents en se fondant sur le nombre de tués et de blessés graves sur une période suffisamment longue (5 ans). Il utilise le taux de ces victimes, c'est-à-dire le nombre de tués et de blessés graves rapporté au milliard de véhicules-kilomètres. Cet indicateur permet de diviser la dangerosité en quatre niveaux. Nous allons utiliser cette variable cible (classe_isr)) pour créer un modèle de prédiction de ces classements, en nous basant sur les caractéristiques des risques d'accidents (comportement des conducteurs, état des véhicules, conditions environnementales).

Tableau 4 : Classement de la sécurité routière

1	*sécurité très mauvaise besoin très urgent d'amélioration*
2	*sécurité mauvaise* besoin *urgent d'amélioration*
3	*Sécurité faible* besoin *d'amélioration*
4	*Sécurité bonne*

II. Etude de l’existant

L'analyse de l'existant implique une inspection minutieuse des données et des infrastructures déjà établies dans un contexte spécifique. Dans notre cas, cela englobe notamment l'examen de la manière dont ces données ont été collectées, des lieux et des procédures impliqués. De plus, cela comprend l'analyse des données historiques, la compréhension des facteurs à considérer lors de la prédiction des risques d'accident. Ces facteurs incluent les caractéristiques liées au conducteur, à l'environnement et au véhicule, ainsi que la conception de solutions prédictives pour anticiper et prévenir les routes à haut risque d'accidents.

1.Conception général de la solution

Ce schéma représente le processus que nous avons suivi pour réaliser notre solution de prédiction des risques d'accidents avec de classification des routes.

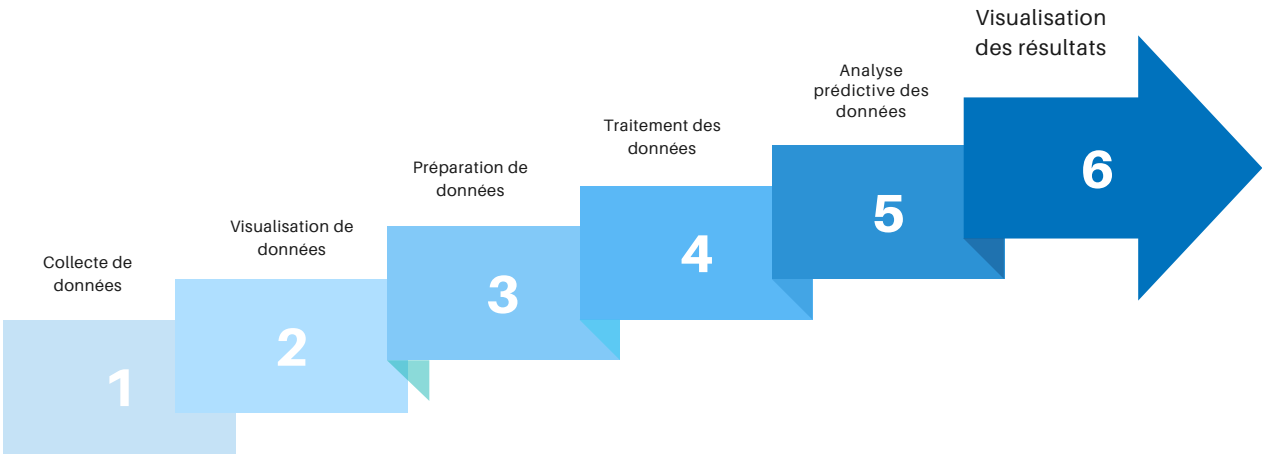


Figure 21: Conception général de la solution.

Pour une meilleure vision et compréhension de ce processus, nous allons le décomposer et expliquer chaque étape.

2.Conception détaillée de la solution

2.1.Collection de données:

La collecte des données peut être l'une des étapes les plus cruciales, où la qualité et la pertinence des données recueillies déterminent en grande partie la fiabilité et l'efficacité des analyses et des décisions qui en découlent.

Donc pour assurer cette démarche, la collecte des données pour le projet est réalisée de manière confidentielle et en étroite collaboration avec les parties pertinentes suivantes :

- La Protection Civile , la Gendarmerie Royale, la Sûreté Nationale, le Ministère de la Santé, le Ministère de la Justice , l'Agence Nationale de la Sécurité Routière , les Compagnies d'Assurance.

Les fiches dédiées à cette mission comprennent la fiche verte pour la Gendarmerie Royale (GN) et la fiche bleue pour la Sûreté Nationale (DGSN). L'Agence Nationale de la Sécurité Routière stocke ces données dans des fichiers Excel.

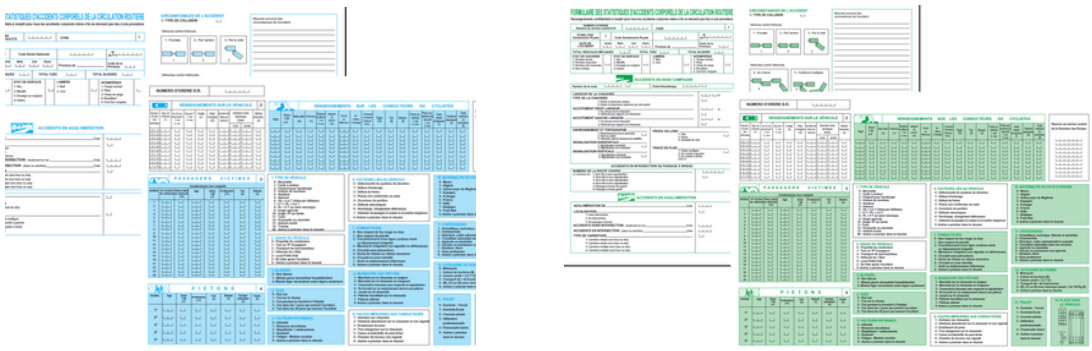


Figure 22 :les fiches de renseignement

Ainsi, il est crucial de comprendre le processus qui gère ces données pour nous permettre de les exploiter de manière plus sophistiquée et précise. Il est donc essentiel de schématiser ce processus, depuis la collecte des informations via les formulaires jusqu'à la création d'une base de données dans Excel.

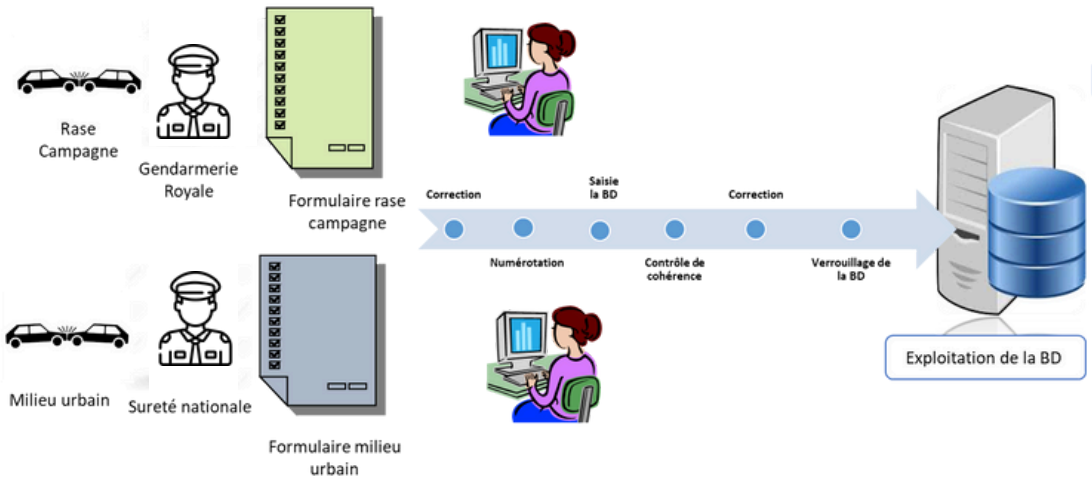


Figure 23 : Le procédure de collection et stockage de données chez NARSA

Après avoir collecté les données sous format xlsx (Excel), nous utiliserons une base de données sur les accidents de 2022, qui se compose des éléments suivants (colonnes, lignes) :

- Tableau des accidents,(46, 113626).
- Tableau des piétons,(10, 28873).
- Tableau des passagers,(10, 43910).
- Tableau des véhicules,(29, 193952).
- Tableau de conducteur_passager_pieton_vehicule,(17, 266733).

Un tableau des données des accidents couvrant la période de 2013 à 2022,(6,11).
Un tableau des données de classement ISR de l'année 2022,(11,14834).
Un tableau des données de traffic de l'année 2022,(8,992).

2.2.Visualisation de données

2.2.1.Visualisation de données sur les accidents 2022

Comprendre les données ne se limite pas à les analyser uniquement à partir des tableaux Excel. Il est crucial de tirer parti des visualisations pour appréhender l'impact des différents attributs sur l'augmentation des accidents et des décès. Cette compréhension approfondie des schémas et des relations entre les variables permet une analyse plus précise, offrant ainsi des perspectives précieuses pour améliorer la sécurité routière.

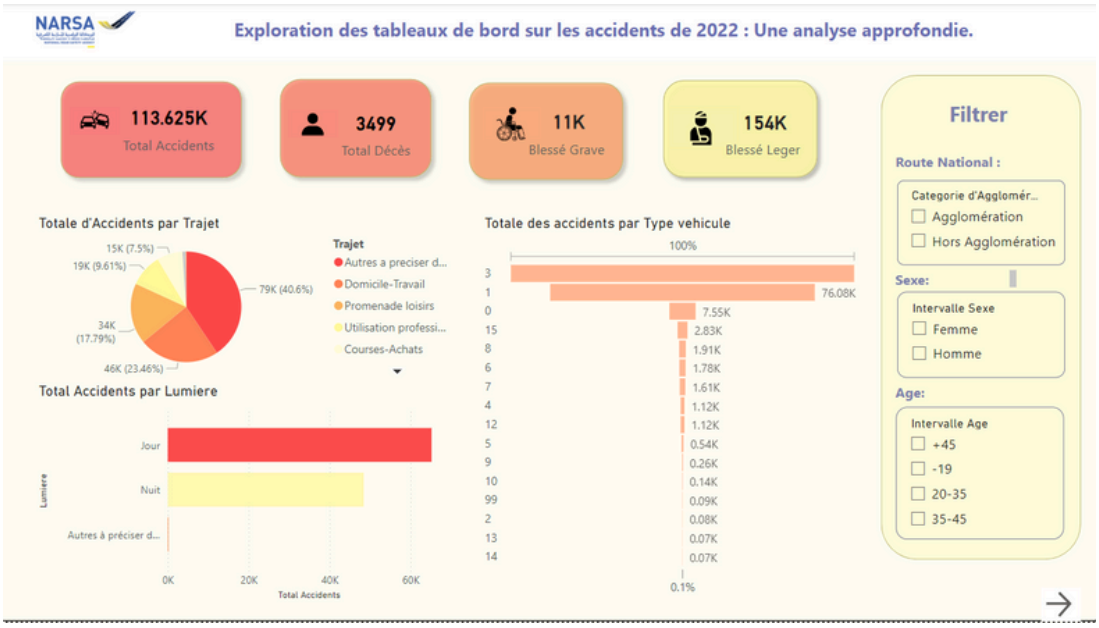


Figure 24 : Exploration des tableaux de bord sur les accidents de 2022 (Première dispositive).

Cette première diapositive, conçue avec Power BI, présente des tableaux de bord interactifs dotés de filtres, permettant d'analyser en détail les tendances des accidents survenus en 2022.

Plus de 23 % des accidents au Maroc se produisent sur le trajet travail-domicile, ce qui peut indiquer des risques accrus aux heures de pointe. Cela pourrait être dû à la fatigue, à la précipitation, à la densité de circulation, au stress, à l'inattention ou à des conditions routières défavorables

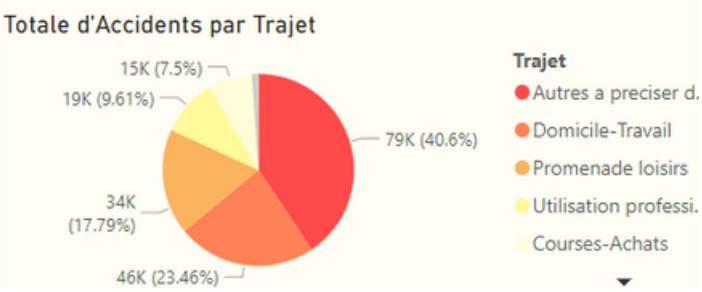


Figure 25 : Totale d'accident par rapport au trajet

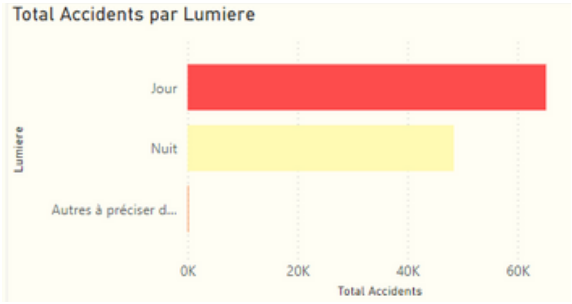


Figure 26: Totale d'accident par rapport au lumière

Le graphique indique que les accidents de jour, environ 60 000, sont 50% plus nombreux que ceux de nuit, environ 40 000. Cette différence importante peut être expliquée par des conditions de conduite plus propices aux collisions pendant la journée, avec une circulation plus dense. Le nombre d'accidents est plus faible la nuit, probablement grâce à une circulation moindre et une plus grande prudence des automobilistes.

En 2022, les voitures de tourisme, fréquemment utilisées tant pour les déplacements professionnels que personnels, sont les véhicules les plus impliqués dans les accidents, avec 98 703 incidents. Ces voitures peuvent transporter jusqu'à neuf passagers et sont souvent offertes aux salariés comme avantage en nature. Les motos, également significativement impliquées, enregistrent 76 083 accidents. Ces chiffres soulignent les risques de sécurité associés à ces modes de transport.

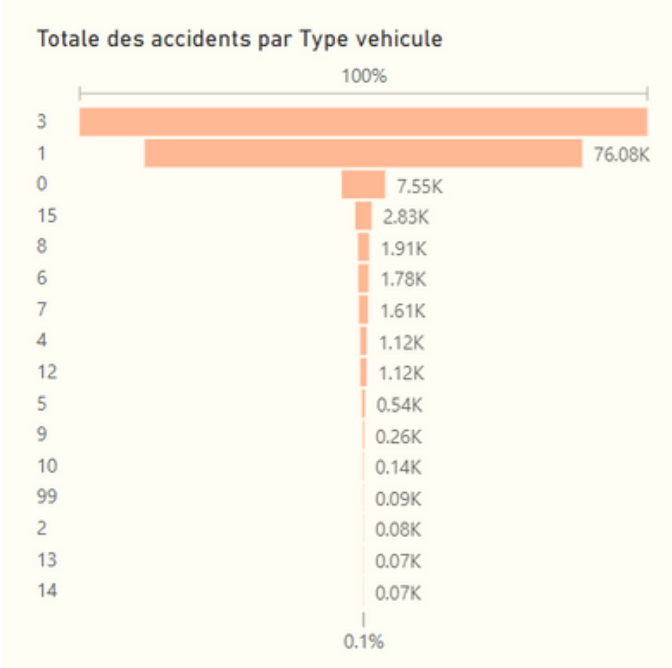


Figure 27: Totale d'accident par rapport au type véhicule impliquée

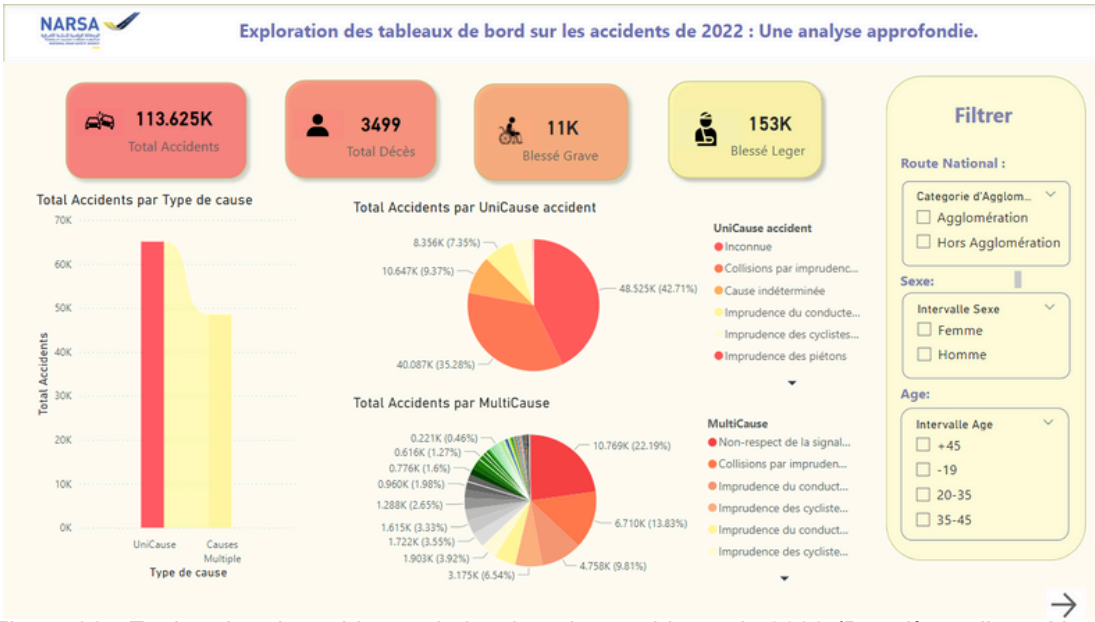
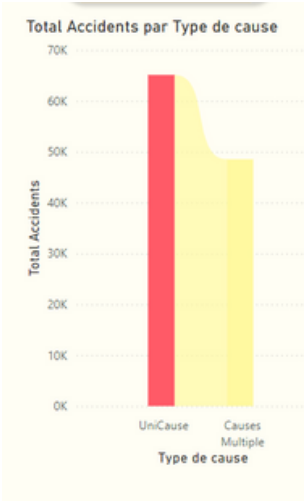


Figure 28 : Exploration des tableaux de bord sur les accidents de 2022 (Deuxième dispositive).



L'analyse du graphique illustre que les accidents de la route ne résultent pas nécessairement d'une seule cause; ils peuvent impliquer jusqu'à cinq facteurs différents. En distinguant les accidents selon qu'ils sont dus à une cause unique (unicause) ou à plusieurs causes (multicause), on observe que les accidents unicausaux sont nettement plus fréquents, représentant 70% des cas.

Figure 29 : Distribution de type de cause d'accident par le totale d'accident 2022.

L'analyse du graphique pour 2022 révèle que parmi les accidents à cause unique, 42,71% sont attribués à des causes inconnues et 35,28% à des collisions dues à l'imprudence des conducteurs impliquant plusieurs véhicules. Ces statistiques soulignent l'importance d'une meilleure collecte de données et de renforcer la sensibilisation des conducteurs pour réduire les accidents.

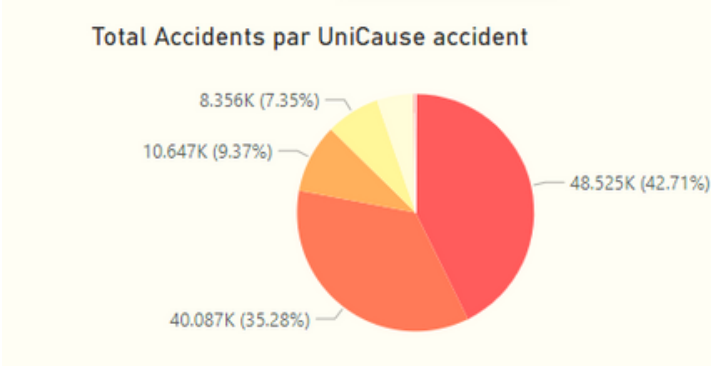


Figure 30 : Répartition des causes d'accidents à cause unique sur le total des accidents en 2022.

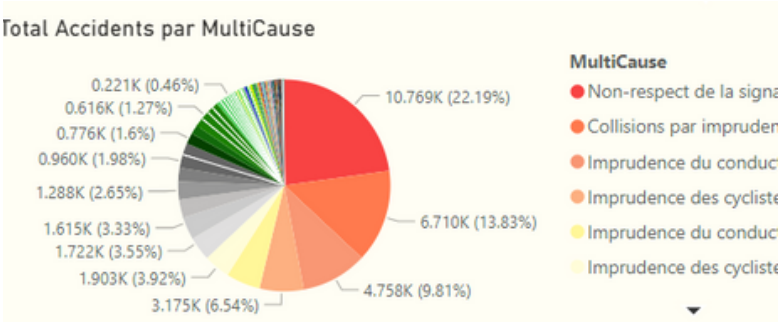


Figure 31: Répartition des causes d'accidents à cause multiple sur le total des accidents en 2022.

L'analyse du graphique dédié aux accidents impliquant des multicauses en 2022 révèle des tendances spécifiques dans la combinaison des causes. La combinaison la plus fréquente est celle entre le non-respect de la signalisation ou du régime de priorité et la collision par imprudence (avec un nombre de véhicules supérieur ou égal à 2), représentant 22,19% des cas. Elle est suivie de la combinaison inverse (collision par imprudence suivie du non-respect de la signalisation), qui constitue 13,83% des incidents. En troisième position, la combinaison de l'imprudence des conducteurs avec celle des piétons se détache, avec 9,81% des cas.

Ces résultats mettent en lumière la prévalence des erreurs de conduite et de signalisation comme facteurs principaux dans les accidents multicausaux, soulignant ainsi l'importance cruciale d'améliorer la sensibilisation à la sécurité routière et le respect des lois pour diminuer la fréquence et la gravité des accidents.

Autre chiffres :

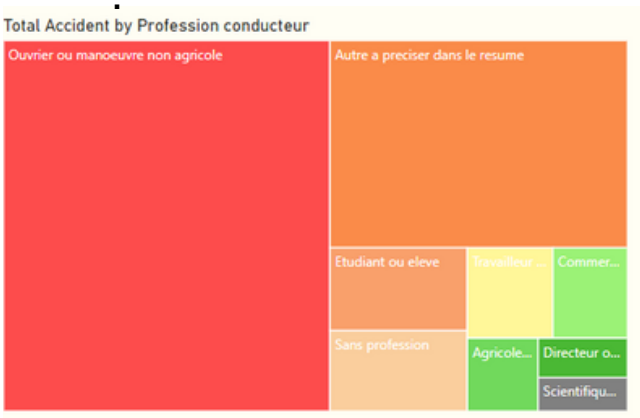


Figure 32: Répartition des accidents selon la profession du conducteur.

Il est intéressant de noter que les ouvriers et manœuvres non agricoles sont surreprésentés dans les accidents, représentant 52,31 % des cas. Cette catégorie professionnelle, souvent associée à des activités manuelles et exposée aux risques routiers, est particulièrement touchée. De plus, environ 26,66 % des accidents impliquent des conducteurs classés dans la catégorie “Autre à préciser”, ce qui suggère que certaines professions ne sont pas clairement identifiées. Enfin, les étudiants ou élèves représentent 4,92 % des accidents, une proportion non négligeable à prendre en compte dans les mesures de prévention.

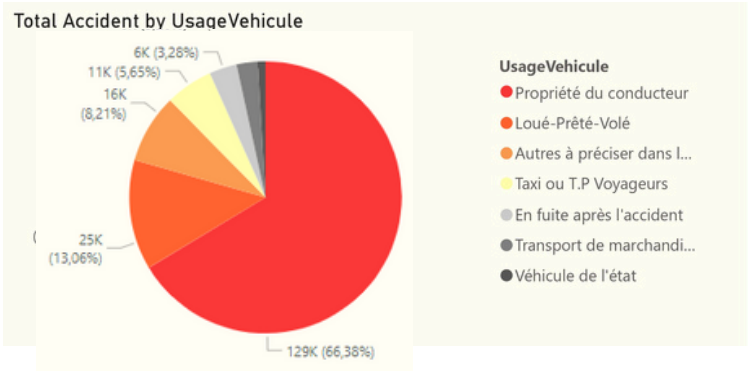


Figure 33 : Répartition des accidents selon le type d'usager du véhicule.

Il est intéressant de noter que la majorité des accidents (66,88 %) impliquent des véhicules qui sont la propriété du conducteur. Cela suggère que les conducteurs utilisant leur propre véhicule sont davantage concernés par les accidents de la route. Plusieurs facteurs peuvent expliquer cette tendance,

D'autre part, une part non négligeable (13,06 %) des accidents concerne des véhicules volés, loués ou prêtés. Cela indique que les conducteurs utilisant un véhicule qui ne leur appartient pas sont également exposés à un risque accidentogène significatif.

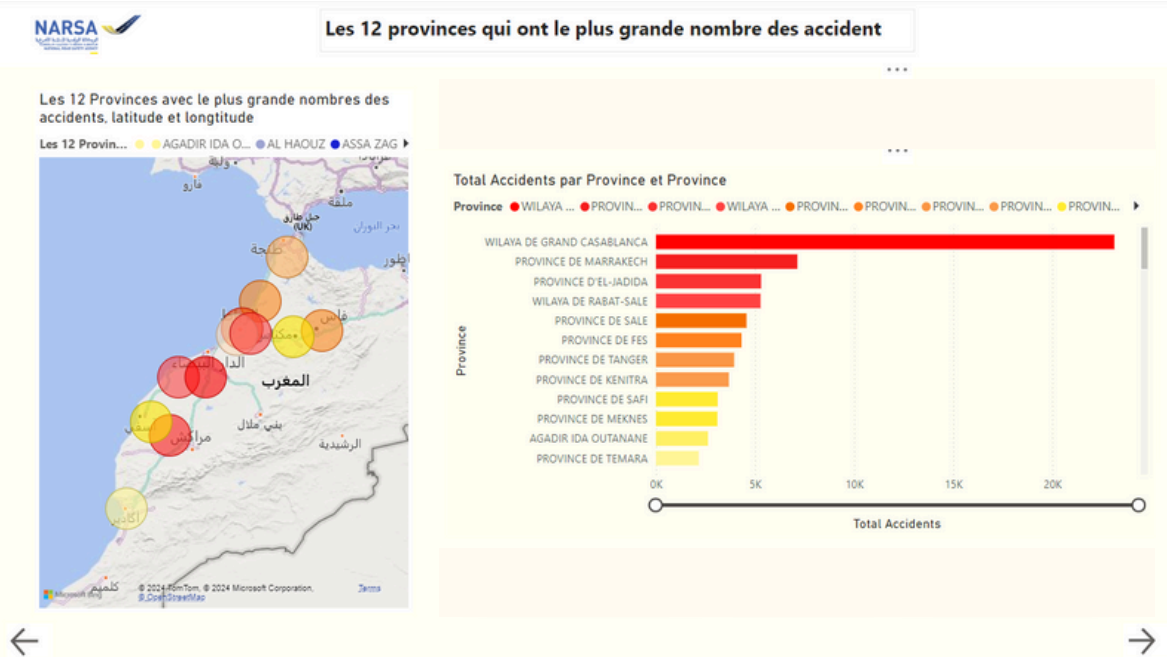


Figure 34 : Répartition des accidents selon les provinces du Maroc (Troisième dispositive)

.

L'analyse montre que les régions les plus touchées par les accidents de la route en 2022 sont :

La wilaya de Casablanca, en raison de sa forte population et de son important trafic routier, la province de Marrakech, probablement du fait de son activité économique et touristique élevée, la province d'El Jadida, malgré une démographie et une économie relativement moins importantes que les deux premières, enfin la wilaya de Rabat-Salé, en tant que capitale administrative, subissant un trafic important lié aux activités gouvernementales.

2.2.2.Visualisation de données sur les décès 2022

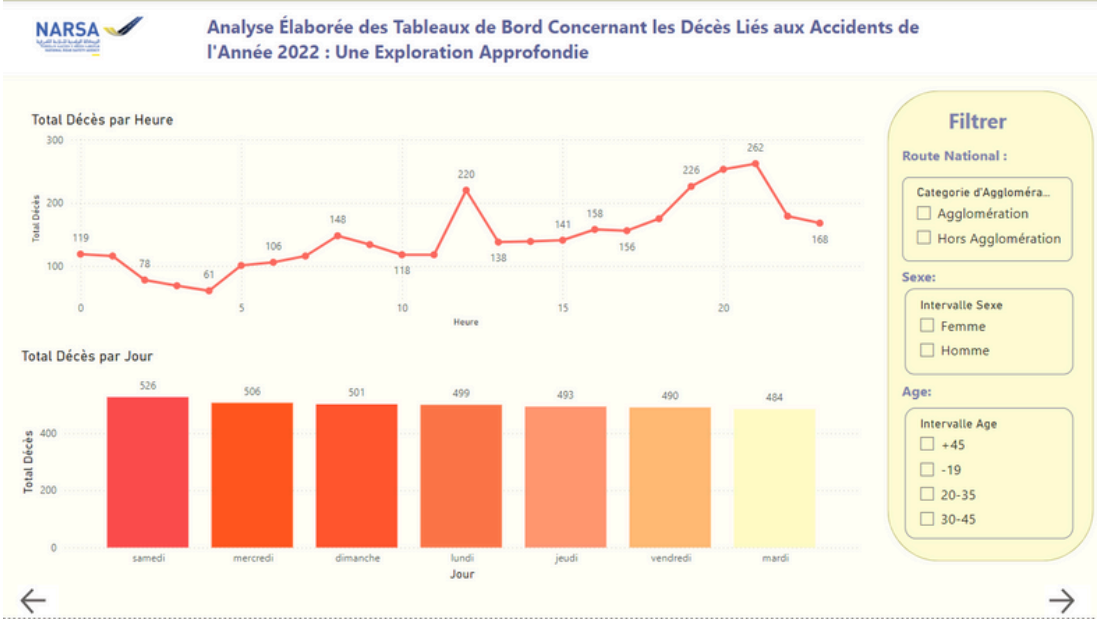


Figure 35 :Analyse Élaborée des Tableaux de Bord Concernant les Décès Liés aux Accidents de l'Année 2022(quatrième dispositive) .

Les statistiques révèlent que les accidents mortels de la route sont plus fréquents le samedi, le mercredi. Ce phénomène s'explique vraisemblablement par une circulation plus dense ces jours-là, ainsi qu'une moindre vigilance des conducteurs, le weekend étant souvent une période de détente. Par ailleurs, la tranche horaire la plus meurtrière se situe entre 21h et minuit, probablement en raison de la fatigue et de la baisse de concentration des conducteurs à ces heures avancées.

Autre chiffres :

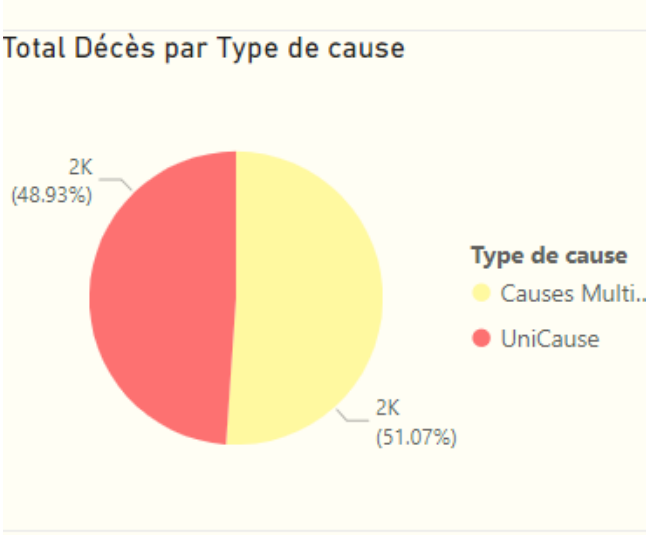


Figure 36 : Répartition des décès par rapport au type de cause .

Bien que les accidents à cause unique soient les plus fréquents, ce sont les accidents à causes multiples qui entraînent près de la moitié (48,93%) des décès.
Cela suggère que les accidents les plus graves impliquent plusieurs facteurs contributifs, sans pour autant négliger les accidents à cause unique, leur bilan humain n'étant pas très différent.
Ces accidents complexes, qu'ils soient à causes multiples ou uniques, doivent faire l'objet d'une attention particulière en matière de prévention routière.

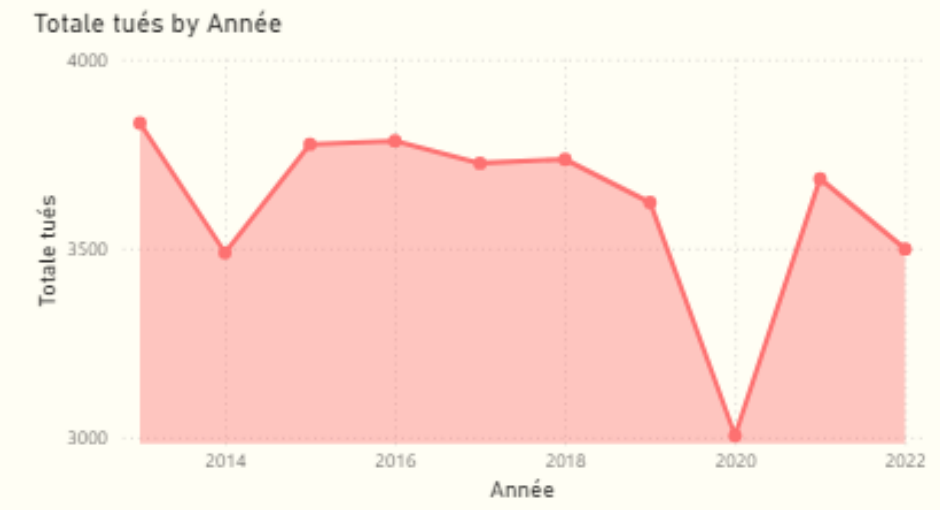


Figure 37 : Répartition des décès durant la période de (2013-2022) .

L'analyse de l'évolution du nombre de décès dans les accidents de la route de 2013 à 2022 montre des tendances contrastées. Avant 2020, le bilan était préoccupant avec 3 600 à 3 882 décès par an. En 2020, la baisse à 3 000 décès s'explique par la réduction du trafic routier liée à la pandémie. Cependant, cette amélioration n'a pas perduré, avec une hausse en 2021 et 2022, revenant à des niveaux élevés de 3 685 et 3 499 décès.

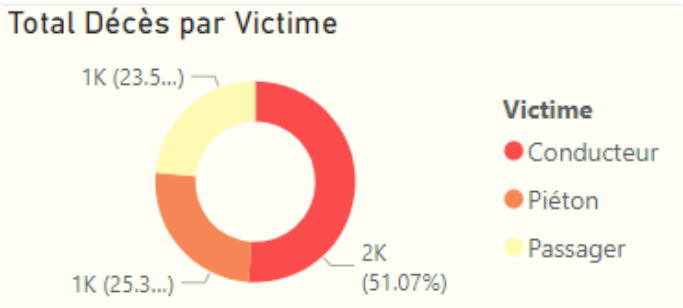


Figure 38 : Répartition des décès selon le type de victimes

L'analyse de la répartition des décès en 2022 montre que les conducteurs sont les plus touchés, représentant 51,07% des victimes. Les piétons constituent la deuxième catégorie la plus impactée, avec 25.38% des décès, suivis par les passagers à 23.55%. Ces résultats soulignent l'importance de cibler en priorité la sécurité des conducteurs, des piétons et des passagers dans les efforts de prévention routière.

2.3.Préparation de données:

Avant de progresser vers les prochaines étapes, donc dans cette étape nous traitons les zones situées en dehors des agglomérations, en répondant aux besoins de visualisation des résultats. Seules les zones hors agglomération disposent d'informations sur les points kilométriques (PKM) renseignés, qui peuvent être convertis en coordonnées GPS. De plus, nous ne traitons que les données relatives aux conducteurs pour nous concentrer sur les accidents impliquant des véhicules. Par exemple, aucun accident n'est enregistré entre deux piétons, de même pour les passagers. Cette phase implique la suppression des colonnes vides, non informatives et confidentielles, comprenant des données sensibles sur l'État remplissant le formulaire, les passagers et les piétons. En outre, elle comprend la vérification de la codification et des champs pour garantir la qualité et l'intégrité des données.

Remarque :

L'encodage des valeurs qualitatives est crucial en prétraitement des données, les transformant en format numérique pour les rendre utilisables dans l'analyse et la modélisation. C'est indispensable car de nombreux algorithmes ne peuvent pas traiter directement les données qualitatives.

2.4.Analyse de données exploratoire EDA

Le traitement des données, également connu dans la communauté de la science des données sous le nom d'EDA (Exploratory Data Analysis), représente une étape cruciale. Les data scientists y consacrent souvent plus de temps, car cette phase peut avoir un impact significatif sur le modèle final. Pour cette raison, nous allons détailler et expliquer minutieusement cette étape, afin de garantir une compréhension approfondie de ses implications et de maximiser sa valeur ajoutée pour le processus d'analyse des données. Donc cette étape implique:

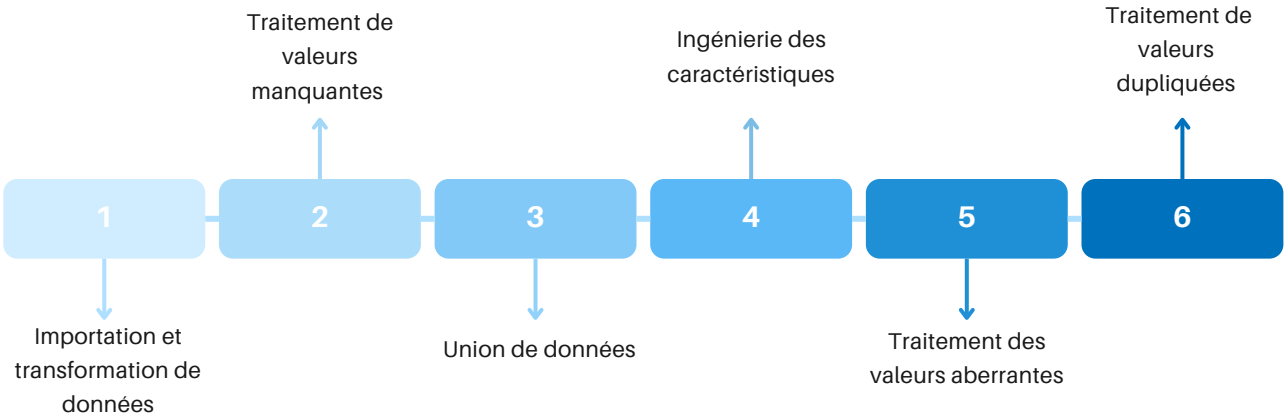


Figure 39 :Analyse de données exploratoire EDA.

- **Importation et transformation de données:** Avant d'entreprendre les diverses étapes de traitement des données, il est essentiel d'importer le jeu de données dans la base de données du logiciel pour y effectuer les traitements nécessaires. Généralement, ces données sont fournies sous forme de fichiers XLSX. Ensuite, pour faciliter la manipulation et l'analyse ultérieures, il est indispensable de transformer ces données en format CSV. Le format CSV présente plusieurs avantages, notamment sa légèreté, sa compatibilité avec de nombreux logiciels et langages de programmation, ce qui simplifie le transfert et le traitement des fichiers. Cette transformation en CSV optimise la gestion et l'analyse des données, assurant ainsi des résultats plus pertinents et exploitables.

- **Traitement des valeurs manquantes :** Cette partie implique un traitement approfondi des valeurs manquantes et l'établissement de critères pour les gérer.

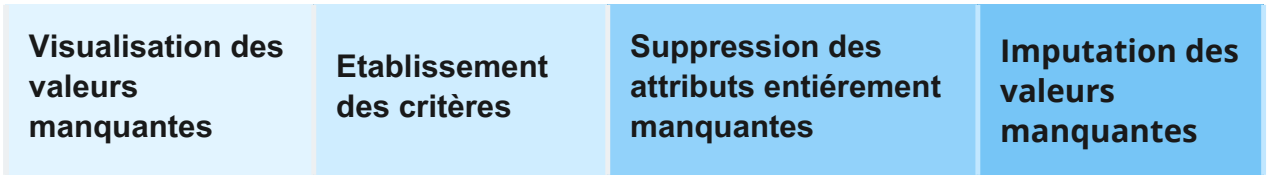


Figure 40 : Etapes de Traitement des valeurs manquantes.

- **Union de données:** Cette étape consiste à regrouper les différentes sources de données en une seule table contenant les caractéristiques nécessaires pour appliquer notre analyse prédictive.

- **Ingénierie des caractéristiques:** L'ingénierie des caractéristiques consiste à créer de nouvelles caractéristiques à partir des anciennes, et à sélectionner celles qui seront utilisées comme entrées dans les modèles prédictifs, en utilisant des techniques comme l'ANOVA pour évaluer leur pertinence. Ce processus améliore la qualité et la performance des modèles en capturant mieux les relations et les patterns dans les données.

- **Traitement des valeurs aberrantes** : Cette étape nécessite une exploration détaillée des valeurs aberrantes, suivie de l'élaboration de stratégies pour les gérer de manière appropriée.

Figure 41 : Etapes de Traitement des valeurs aberrante.

- **Traitement des valeurs dupliquées :** Identifier et éliminer les enregistrements redondants ou répétitifs dans notre ensemble de données. Ces doublons peuvent être sources de confusion et fausser les résultats de notre analyse, car ils introduisent un bruit inutile et peuvent fausser les statistiques. En supprimant ces doublons, nous nettoyons nos données et nous assurons que notre analyse repose sur des informations précises et non redondantes, ce qui améliore la qualité de nos résultats et la fiabilité de nos conclusions.

2.5.Analyse prédictive

Dans cette étape, nous construisons notre modèle en divisant le processus en sous-processus distincts. Cela permet de mieux organiser et gérer les différentes tâches impliquées dans la création du modèle. Chaque mini-processus se concentre sur une partie spécifique de la construction du modèle.

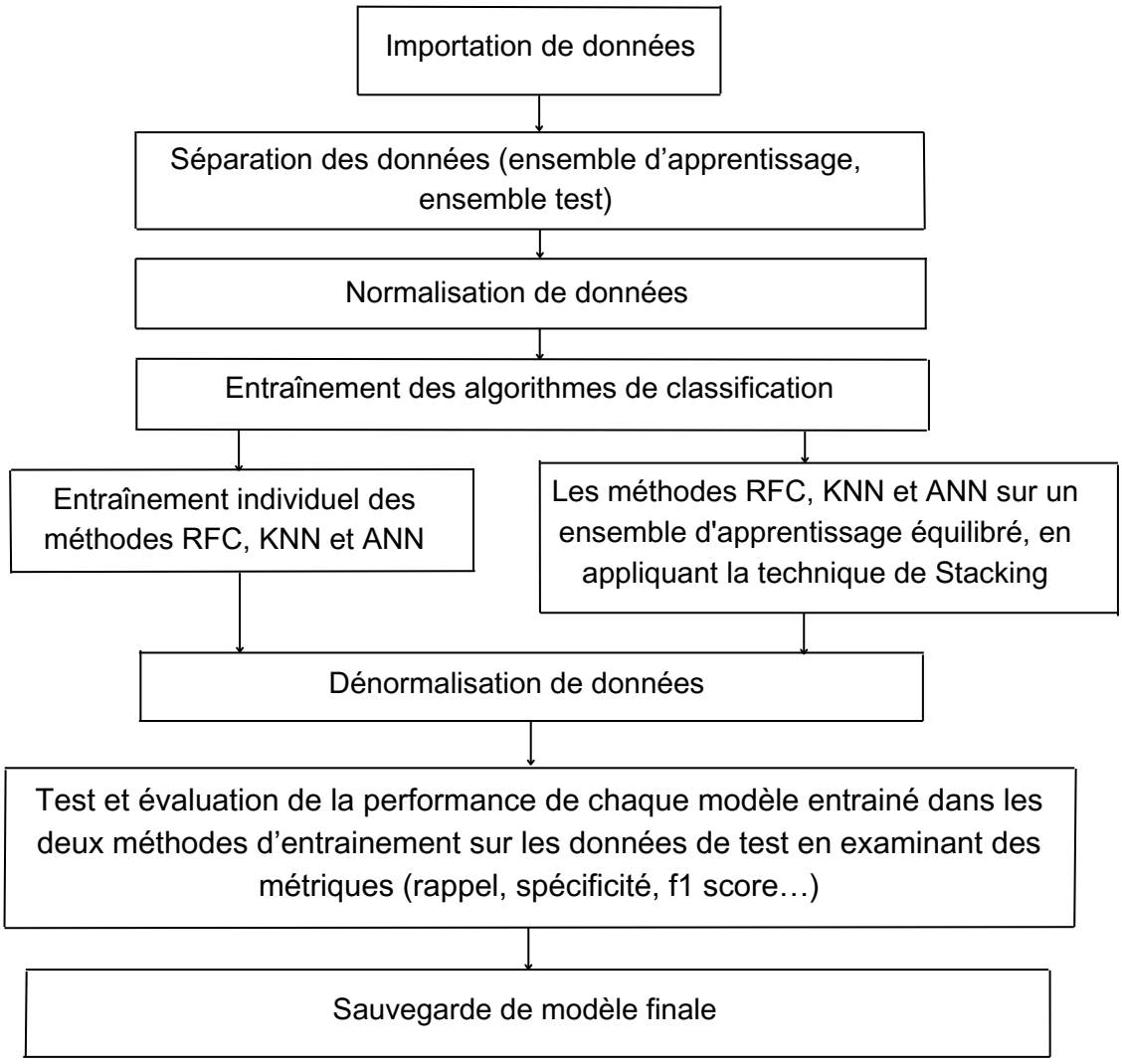


Figure 42 : Etapes d'Analyse prédictive

2.6.Visualisation des résultats

La visualisation des résultats revêt une importance capitale pour comprendre et communiquer efficacement les conclusions de travail. Dans le cadre spécifique de notre projet, l'utilisation de la cartographie nous permettra de représenter de manière visuelle les routes présentant les niveaux de risque d'accidents les plus élevés jusqu'aux plus faibles. Cette représentation géographique facilitera la compréhension des zones où une action préventive ou corrective pourrait être nécessaire, permettant ainsi une prise de décision éclairée pour améliorer la sécurité routière.

III. Langages, technologies et outils

Les outils et les langages de programmation sont essentiels pour la réalisation des processus de ce projet. Ils nous permettent de développer des méthodes et des solutions pour chaque étape, garantissant ainsi une progression structurée et efficace. Voici une description détaillée des outils et langages utilisés dans ce projet.

1.Python [16]

Python est un langage de programmation interprété, interactif, orienté objet et de haut niveau. Créé par Guido van Rossum et publié en 1991, il se distingue par sa syntaxe claire et sa lisibilité. Dans notre projet, nous utilisons Python 3.9.3 pour importer et exploiter les données, construire et améliorer notre modèle, et créer le diagramme de Gantt pour la planification des tâches. En résumé, Python est essentiel pour l'analyse des données, le développement des modèles et la gestion du projet.

Figure 43

1.1. Enivrement et outils installe pour python

- **Anaconda [17]:** Anaconda est une distribution pour Python et R, optimisée pour le calcul scientifique et l'analyse de données. Elle simplifie la gestion des paquets et des environnements.

Figure 44

- **Jupyter [18]:** Jupyter Notebook est un outil interactif populaire pour l'analyse de données et l'apprentissage automatique. Il offre une interface conviviale, supporte plusieurs langages comme Python et R, Facilite la documentation avec Markdown et LaTeX, et permet le partage facile des travaux via des fichiers .ipynb ou GitHub.

Figure 45

[16] :Pour plus d'information voir <https://www.python.org/>
[17] :Pour plus d'information voir <https://www.anaconda.com/ >
[18] :Pour plus d'information voir <https://jupyter.org/>

- **Les packages et bibliothèques utilisé du langage Python:**

Tableau 5 : Bibliothèques de Python.

Package	Logo	Utlisation	Version
Panda		Analyse et manipulation de données structurées (DataFrames).	2.2.2
Numpy		Calcul numérique et manipulation de tableaux multidimensionnels.	1.26.4
Matplotlib		Création de graphiques et visualisations statiques en 2D.	3.8.0
Seaborn		Visualisations statistiques améliorées, basées sur Matplotlib.	0.13.2
Plotly		Création de visualisations interactives et dynamiques pour l'exploration et la présentation des données.	5.19.0
skeat-learn		incontournable pour l'apprentissage automatique regorgeant d'algorithmes et de fonctionnalités inclus la classification.	1.4.2
Tensor-flow		Création, entraînement et déploiement de modèles d'apprentissage automatique, en particulier les réseaux de neurones.	2.16.1
Keras		Création de visualisations interactives et dynamiques pour l'exploration et la présentation des données.	3.3.3

2.R [19]

Dans notre projet, nous utiliserons le langage R, inventé par Ross Ihaka et Robert Gentleman dans les années 1990, principalement pour l'imputation des valeurs manquantes. Pour cette tâche, nous ferons appel à la bibliothèque MISSForest,développée par Stéphane Dray et Aurélie Siberchicot en 2012. MISSForest utilise des forêts aléatoires pour estimer les valeurs manquantes dans les ensembles de données,

Figure 46

2.1. Enivrement et outils installe pour R

Rstudio: RStudio est un IDE robuste conçu pour les utilisateurs de R. Il offre une interface conviviale, des fonctionnalités avancées d'édition, une intégration transparente des packages, une gestion efficace des projets et la prise en charge de R Markdown pour créer des documents dynamiques. C'est un outil essentiel pour l'analyse de données et la statistique.

Figure 47

Les packages et bibliothèques utilisé du langage R:

Tableau 6 : Bibliothèques de R

Package	Utlisation
readr	la lecture efficace des données, offrant des fonctionnalités améliorées par rapport aux fonctions de lecture de base
Missforest	'imputation robuste des valeurs manquantes dans les données à l'aide de forêts aléatoires

3.Outils:

3.1.Power BI [20]

Dans notre projet, nous utilisons Power BI, une plateforme de business intelligence développée par Microsoft dans les années 2010. Son objectif est de fournir un outil intuitif et puissant pour créer des visualisations de données interactives et des rapports personnalisés. En utilisant Power BI, nous visualisons nos données existantes, ce qui nous permet de mieux comprendre nos données, de découvrir des tendances et des corrélations, et de prendre des décisions stratégiques basées sur les insights obtenus.

Figure 48

[20]Pour plus d'information voir <https://www.microsoft.com/fr-fr/power-platform/products/power-bi>

[19]Pour plus d'information voir <https://www.r-project.org//>

3.2.Excel[21]

Microsoft Excel, développé par Microsoft Corporation et lancé en 1985 sous la direction de Doug Klunder et Ron Klopfer, est devenu une application de feuille de calcul incontournable. un outil essentiel dans notre projet, utilisé pour gérer notre base de données et effectuer la préparation des données. Sa convivialité et ses fonctionnalités avancées nous permettent d'entrer, d'organiser et de filtrer les données selon nos besoins.

Figure 49

3.3.Teamgantt[22]

TeamGantt est un outil de gestion de projet en ligne qui permet de construire et de suivre des diagrammes de Gantt pour la planification et la coordination des projets. Il offre des fonctionnalités pour créer des tâches, assigner des ressources et faciliter la collaboration en équipe.

Figure 50

Conclusion

Dans cette section du rapport, nous avons clarifié les concepts clés nécessaires à la compréhension de notre projet et à la définition précise de nos besoins. Une étude approfondie de nos données existantes a été réalisée pour garantir une parfaite maîtrise de celles-ci, une étape essentielle pour le développement de notre solution. Par ailleurs, la conception du projet a été soigneusement élaborée, jouant un rôle déterminant dans l'implémentation de notre solution prédictive et dans les stratégies de visualisation des données. Cette démarche nous permet de saisir pleinement les nuances de notre jeu de données.

Nous avons aussi exploré et maîtrisé les outils ainsi que les langages de programmation qui nous accompagneront tout au long du projet, de la phase de planification jusqu'à la réalisation finale et la visualisation des résultats. Cette familiarisation avec les technologies clés assure que nous disposons de toutes les ressources nécessaires pour construire une solution efficace et parfaitement adaptée à nos objectifs.

En résumé, cette partie du projet a établi des fondations solides pour la poursuite de notre travail, garantissant que chaque phase est alignée avec nos attentes et exigences. Cela positionne notre projet sur la voie du succès, en s'assurant que toutes les composantes sont harmonisées et prêtes pour la mise en œuvre de la solution.

[21]Pour plus d'information voir <https://www.microsoft.com/fr-fr/microsoft-365/excel>
[22]Pour plus d'information voir <https://www.teamgantt.com/h2?originalReferrer=https://www.google.com/>

IV. Préparation de données

Au cours de la phase de conception, nous avons souligné l'importance d'éliminer les colonnes qui ne contiennent pas d'informations pertinentes, qui sont confidentielles, ou qui sont complètement vides. De plus, nous avons décidé de retirer les données relatives aux zones d'agglomération pour nous concentrer exclusivement sur celles des zones hors agglomération, ainsi que sur les informations concernant uniquement les piétons et les passagers. Nous allons maintenant démontrer ce processus en présentant une comparaison des données avant et après cette phase de préparation. Nous avons spécialement identifié et retiré les entrées où le code d'agglomération est nul, indiquant qu'elles concernent des zones hors agglomération. Pour illustrer notre méthodologie, nous utilisons les éléments suivants :
- Tableau des accidents : 46 colonnes,113626 lignes.
- Tableau des véhicules : 29 colonnes, 193 952 lignes.
- Tableau des conducteurs, passagers, et piétons : 17 colonnes, 266 733 lignes.
- Tableau des données de classement ISR de l'année 2022 : 11 colonnes, 14 834 lignes.
- Tableau des données de trafic de l'année 2022 : 8 colonnes, 992 lignes.

1.Visualisation des valeurs manquantes dans le dataset :

La visualisation des valeurs manquantes permet de détecter et d'éliminer les colonnes entièrement vides. Nous utilisons « Matplotlib » et « Jupyter » pour créer une heatmap des données. Bien que la visualisation des données soit effectuée en Python, la préparation des données a été réalisée dans Excel.

- **Les valeurs manquantes dans la table "Accidents_2022".**

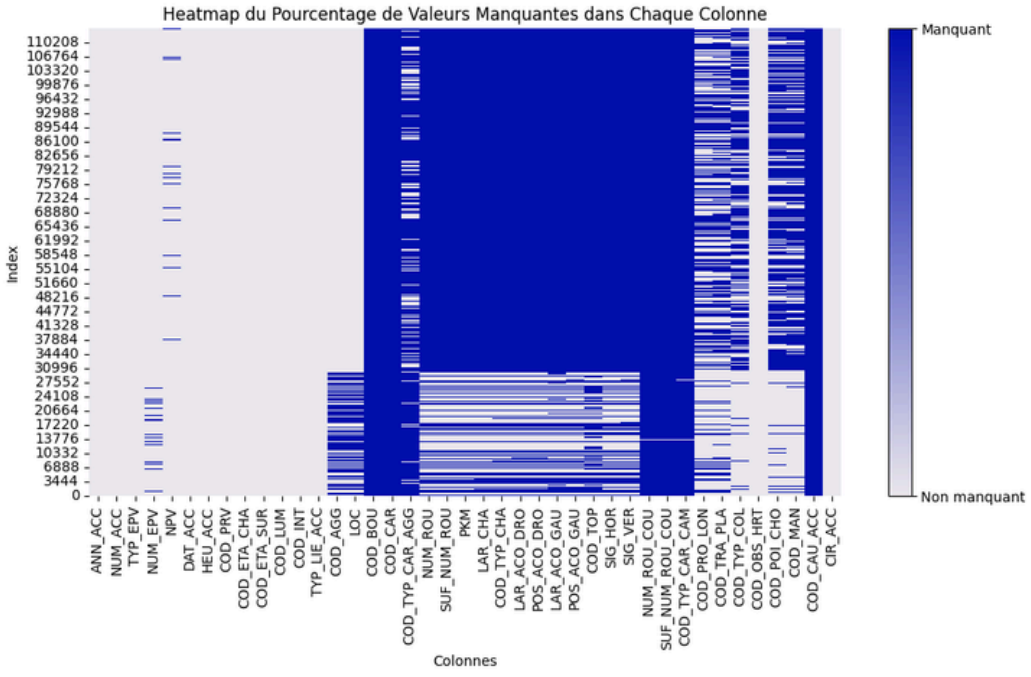


Figure 51 : Les valeurs manquantes dans la table "Accidents_2022".

D'après cette carte, nous pouvons identifier les colonnes à éliminer pour la table accident_2022: celles qui sont non informatives, confidentielles, entièrement vides, ainsi que les données relatives aux zones d'agglomération. Les colonnes à supprimer sont : : COD_AGG, COD_TOP, ANN_ACC, TYP_EPV, NUM_EPV, NPV, DAT_ACC, TYP_LIE_ACC, LOC, COD_BAU, COD_CAR, COD_TYP_CAR_AGG, SUF_NUM_ROU, SUR_NUM_ROU_COU, COD_TYP_CAR_CAM, COD_CAU_ACC, CIR_ACC, et LAR_CHA.

- **Les valeurs manquantes dans la table "Vehicule_2022".**

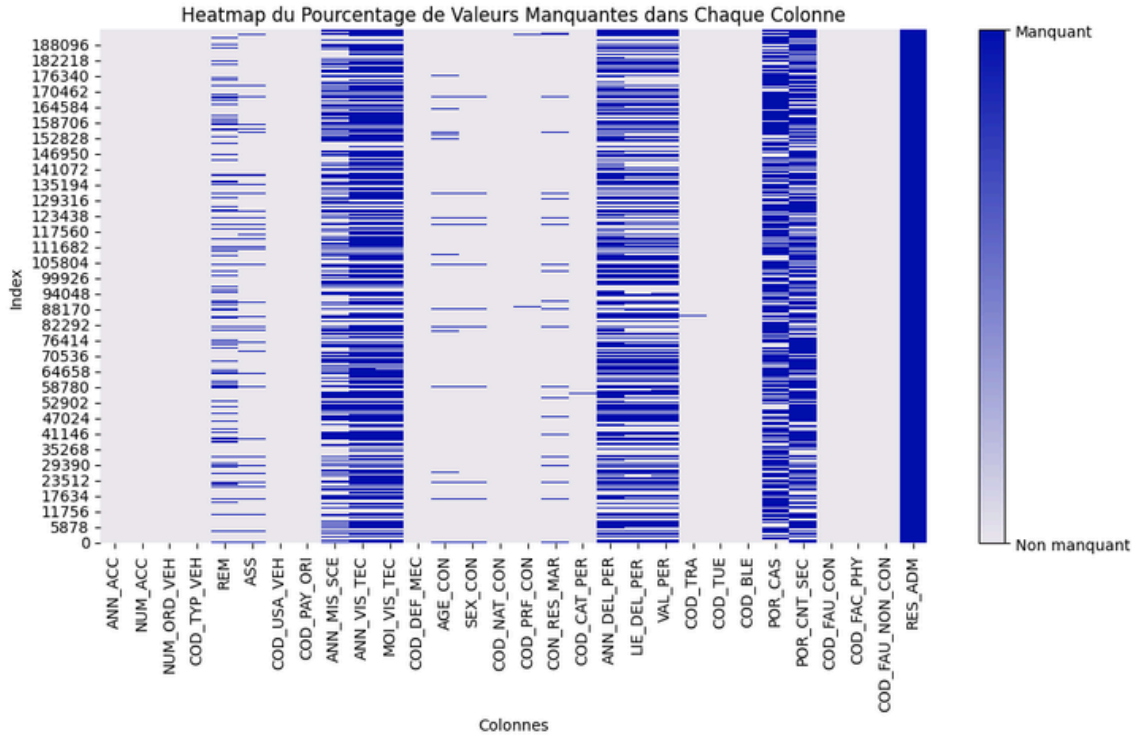


Figure 52 : Les valeurs manquantes dans la table "Vehicule_2022".

On va refait la même chose pour la table véhicule_2022 . Les colonnes à supprimer sont : RES_ADM,ANN_VIS_TEC,POR_CNT_SEC,POR_CAS,VAL_PER,LIE_DEL_PER,AMM_DEL_PER,COD_CAT_PER,CON_RES_MAR,COD_DFF_MEC,MOI_VIS_TEC,ANN_MIS_SCE,REM,ANN_ACC.

- **Les valeurs manquantes dans la table "Traffic_2022" et "Classement_isr_2022".**

Dans ces deux tables, il n'y a pas de valeurs manquantes, cependant, il existe des colonnes non informatives que nous allons éliminer, Les colonnes à supprimer sont : NB_TUE,NB_BLG,NB_BLL,NB_ACC,PROVINCE,REGION,CIRCULATION,ORIGINE, EXREMITE,LONG KM.

- **Les valeurs manquantes dans la table"conducteur_pieton_passager_2022".**

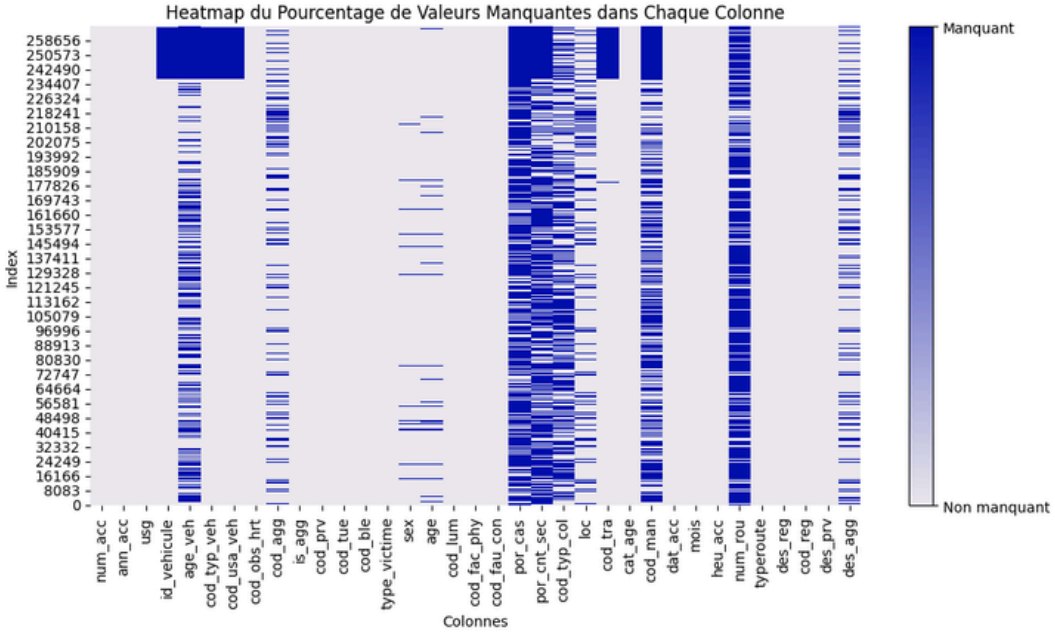


Figure 53 : Les valeurs manquantes dans la table"conducteur_pieton_passager_2022".

La table `conducteur_pieton_passager_2022` inclut des données sur les accidents impliquant piétons, passagers et conducteurs, mais elle est incomplète pour une analyse exhaustive. Pour obtenir toutes les informations nécessaires, il est impératif de la compléter avec des données provenant d'autres tables. Notre focus étant sur les conducteurs, il faut filtrer et exclure les informations concernant les piétons et les passagers. De plus, certaines colonnes redondantes des autres tables, comme NUM_ACC et NUM_ROU, sont cruciales pour correctement assembler les données. Les colonnes telles que cod_reg, des_reg, et plusieurs autres doivent être éliminées pour affiner l'analyse.

Pour résumer les résultats de la préparation, nous pouvons utiliser ce schéma table(colonne, ligne) :

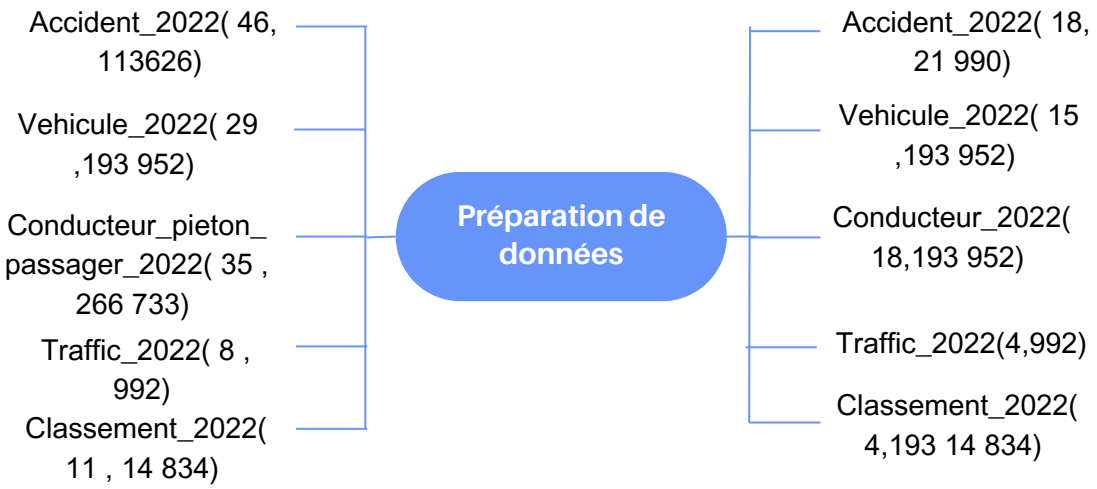


Figure 54: Résultats de préparation de données.

V.Analyse de données exploratoire EDA

Après avoir préparé notre jeu de données pour l'analyse, nous abordons maintenant l'étape cruciale qui déterminera la qualité de nos données, en conjonction avec d'autres facteurs. Cette phase est essentielle pour assurer l'intégrité et la pertinence de nos résultats analytiques.

1.Traitement de valeurs manquantes

1.1. Importation et transformation de données

Avant de procéder au traitement des données, il est essentiel d'importer les données depuis d'autres environnements (comme Excel) vers l'environnement Jupyter, et de les transformer de XLSX en CSV pour diverses raisons telles que la compatibilité, la facilité de manipulation et la performance.

1.2. Visualisation de valeurs manquantes

Pour préparer notre jeu de données pour les algorithmes de machine learning, nous devons d'abord visualiser les valeurs manquantes. Utilisant « Matplotlib » et « Jupyter », nous créons une carte des valeurs manquantes pour chaque table et calculons les pourcentages de valeurs manquantes pour chaque attribut. Cela nous aide à élaborer une stratégie pour traiter les données manquantes.

- **Les valeurs manquantes dans la table "Accidents_2022".**

Le schéma suivant représente le pourcentage des valeurs manquantes dans chaque variable de la table "Accidents_2022".

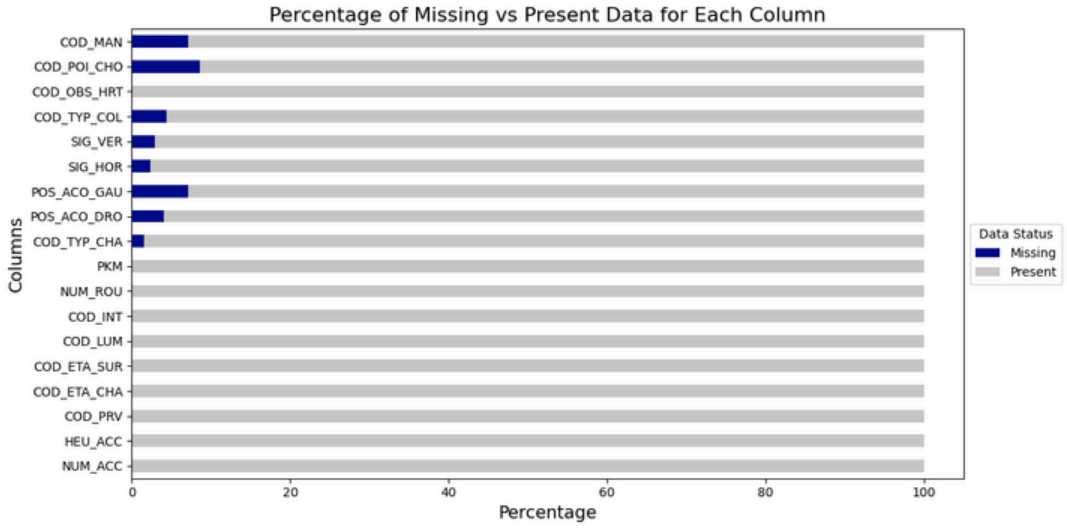


Figure 55 : Les valeurs manquantes dans la table "Accidents_2022".

Cette carte montre qu'il n'y a aucune colonne totalement vide grâce à une préparation préalable qui a éliminé les colonnes non informatives, confidentielles et celles relatives aux zones d'agglomération et aux passagers ou piétons. Les colonnes avec des valeurs manquantes sont: COD_MAN, COD_POI_CHO, COD_TYP_COL, COD_TYP_CHA, SIG_HOR, SIG_VER, POS_ACCO_DROI, POS_ACO_GAU, COD_OBS_HRT.

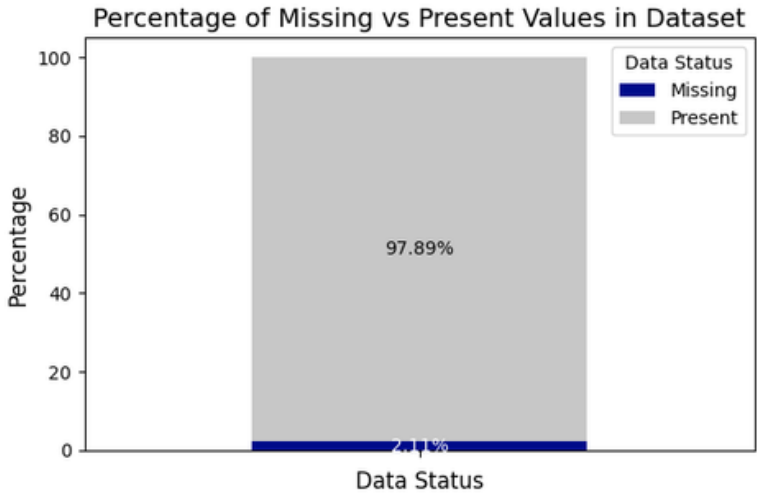


Figure 56 : Le totale es valeurs manquantes dans la table "Accidents_2022".

Cette carte indique que 2.11 % des valeurs dans la table « accidents_2022 » sont manquantes.

- **Les valeurs manquantes dans la table "Vehicule_2022".**

Le schéma suivant représente le pourcentage des valeurs manquantes dans chaque variable de la table "Vehicule_2022".

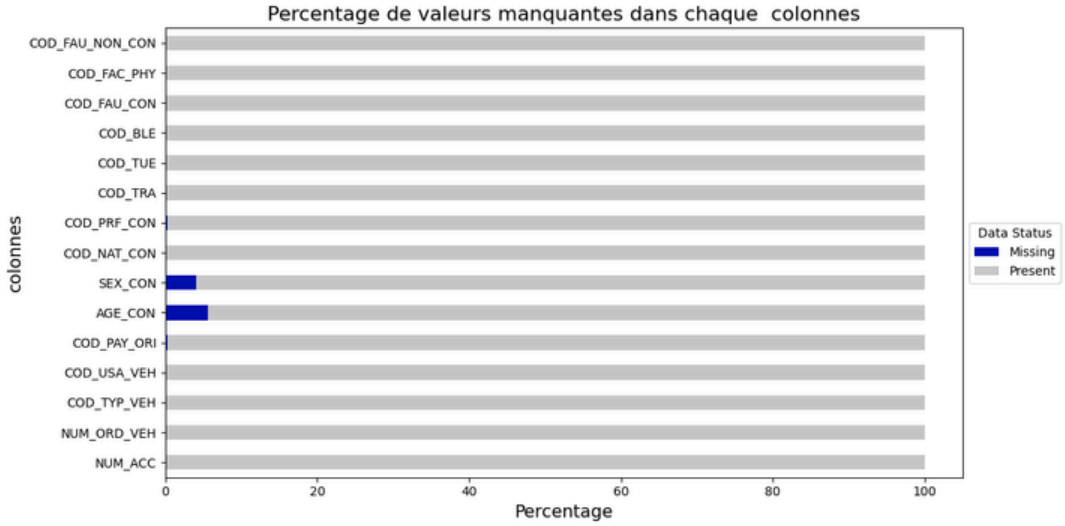


Figure 57 : Les valeurs manquantes dans la table "Vehicule_2022".

Avec des valeurs manquantes dans les colonnes suivantes: SEX_CON,AGE_CON.

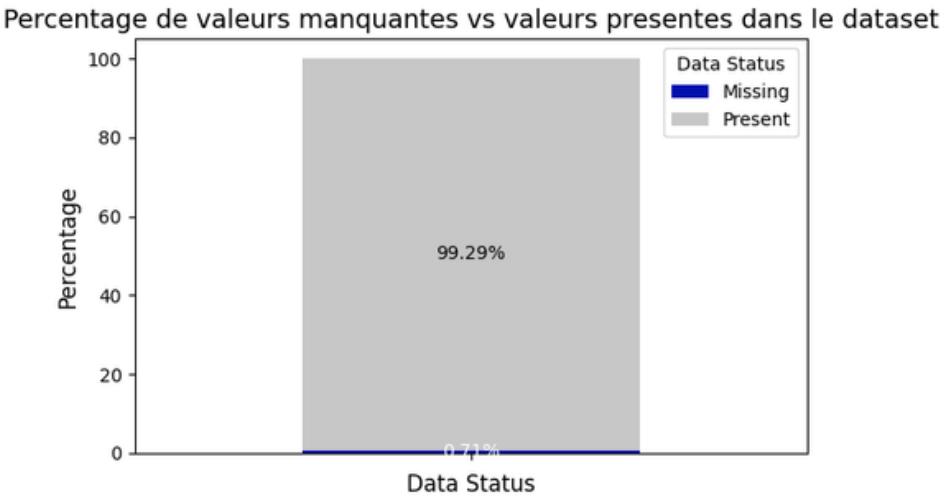


Figure 58 : Le totale Les valeurs manquantes dans la table "Vehicule_2022".

Cette carte indique que 0.71 % des valeurs dans la table « vehicule_2022 » sont manquantes.

- **Les valeurs manquantes dans la table "Conducteur_2022".**

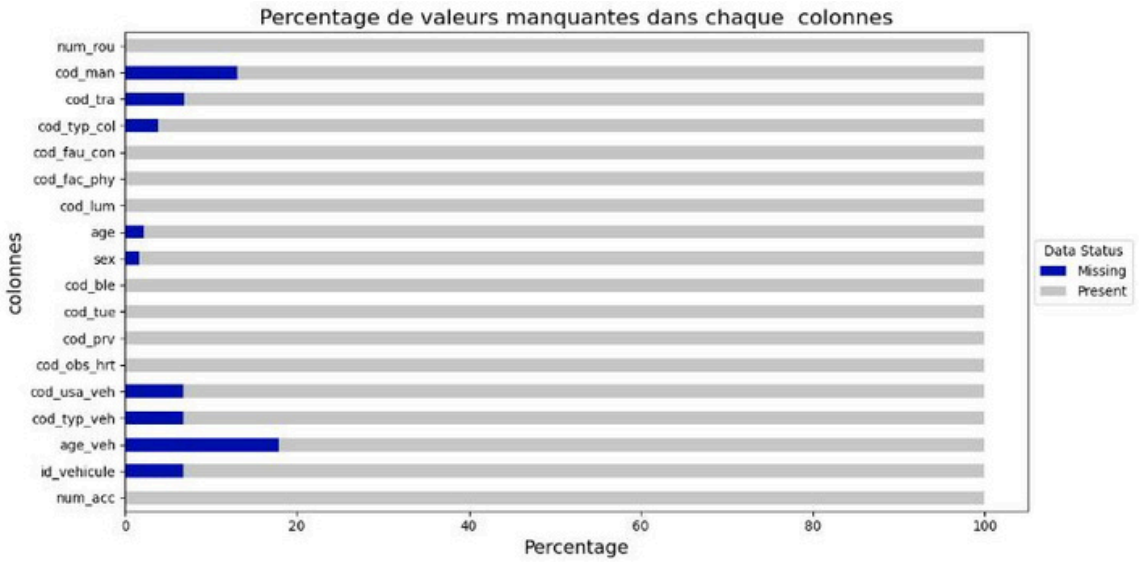


Figure 59: Les valeurs manquantes dans la table "Conducteur_2022".

Avec des valeurs manquantes dans les colonnes suivantes:
cod_man, cod_typ_col,cod_tra,age,sex,id_veh,cod_typ_veh,cod_usg_veh,cod_typ_veh

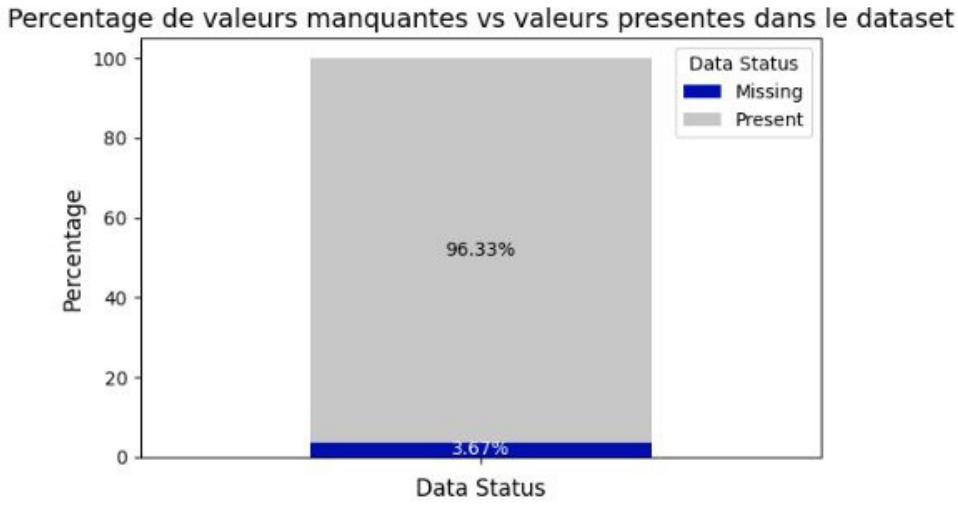


Figure 60: Le totale Les valeurs manquantes dans la table “Conducteur_2022”.

Cette carte indique que 3,76% des valeurs dans la table « conducteur_2022 » sont manquantes.

- **Les valeurs manquantes dans la table “Traffic_2022” et “Classement_isr_2022”.**

Comme nous l'avons précédemment mentionné, les deux tables intitulées « Traffic_2022 » et « Classement_isr_2022 » sont complètes et ne présentent aucune valeur manquante. Ces ensembles de données fournissent donc une base solide et fiable pour notre analyse sans nécessiter d'interventions supplémentaires pour l'imputation de données manquantes.

Bien que le nombre de données manquantes soit relativement faible dans notre jeu de données, leur importance et leur sensibilité nous empêchent de simplement les supprimer. Par conséquent, nous avons opté pour une approche d'imputation des données afin de préserver l'intégrité et la qualité de notre analyse. Cette méthode nous permet de conserver toutes les informations pertinentes tout en gérant les lacunes de manière efficace et respectueuse de la valeur intrinsèque des données. Il convient de noter une exception pour la colonne `id_vehicule`, de table ‘’Conducteur_2022’’ qui est un identifiant et ne peut pas être imputée.

1.3. Imputation de valeurs manquantes

1.3.1.Choix de méthodes d'imputation de valeurs manquantes:

Pour assurer une préparation et une réparation optimales de notre jeu de données, il est essentiel de choisir la méthode d'imputation des valeurs manquantes la plus adéquate. Pour identifier la technique d'imputation la plus appropriée pour notre base de données, nous envisageons d'évaluer plusieurs méthodes discutées dans la section « traitement des valeurs manquantes » de notre revue de littérature. Ces méthodes incluent l'imputation par l'algorithme KNN, l'imputation multiple par les équations chaînées (MissForest), ainsi que l'imputation par la médiane ou la moyenne, en fonction du type de variable (continue ou catégorique).Voici les étapes de notre approche pour évaluer ces techniques d'imputation :

1. Sélectionner une table avec un faible nombre de valeurs manquantes pour minimiser les distorsions dans l'analyse ; pour cette évaluation, nous avons choisi la table `vehicule_2022` avec 0.71% de valeurs manquantes .
2. Éliminer toutes les valeurs manquantes de cette table pour créer un jeu de données complet, que nous nommerons `vehicule_nomissing_2022`.
3. À partir de cette table complète, nous introduirons artificiellement des valeurs manquantes pour créer une version altérée du jeu, que nous appellerons `vehicule_withmissing_2022`.

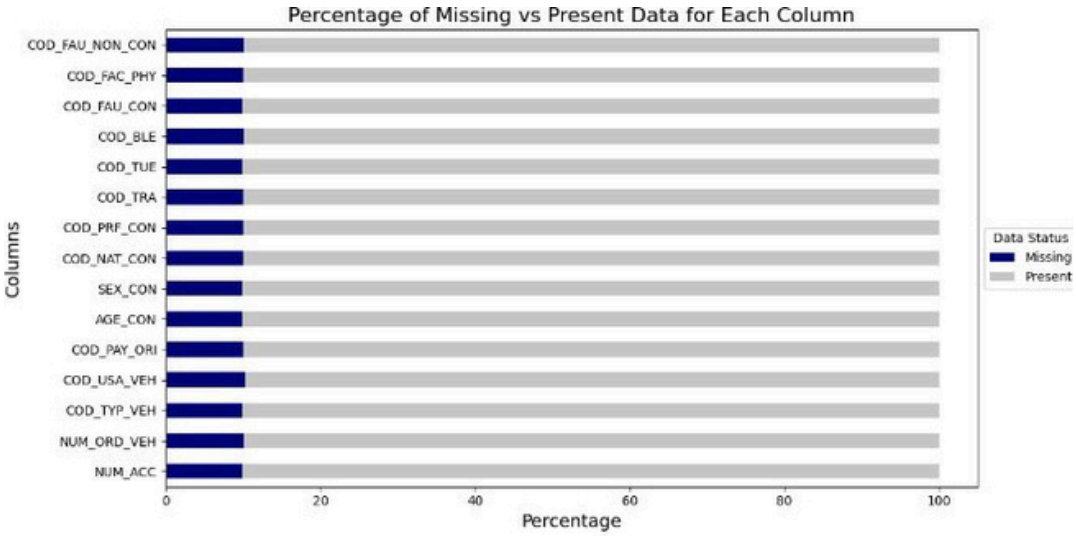


Figure 61 : Valeurs manquantes artificielles

4. Avec ces deux jeux de données un complet et un avec des valeurs manquantes nous testerons les trois méthodes d'imputation sur deux types de variables choisies : une catégorique et une continue qui sont AGE_CON et SEX_CON.
Le tableau suivant représente les résultats de l'évaluation des 3 méthodes d'imputation pour les 2 variables sexe et âge de conducteur de la table "vehicule_2022" .

Tableau 8: La résultat de choix d'imputation

Method d'Imputation	Accuracy de SEX_CON	MAE (Mean absolute error) de AGE_CON
MissForest	98.87%	1.5661
KNN	99.28%	1.1210
Mean et Mode	99.28%	1.1210

Les méthodes KNN, Mode et Médiane offrent une précision supérieure pour la colonne SEX_CON par rapport à la méthode MissForest, et une MAE plus faible pour la colonne AGE_CON. Cette constatation suggère une meilleure adéquation aux données réelles pour la classification et la régression. Pour décider entre l'utilisation de la médiane et de KNN, nous recommandons de visualiser la distribution par rapport aux données originales en utilisant les bibliothèques seaborn et matplotlib dans Jupyter. La méthode qui se rapproche le plus de ces données sera sélectionnée pour l'imputation de notre jeu de données.

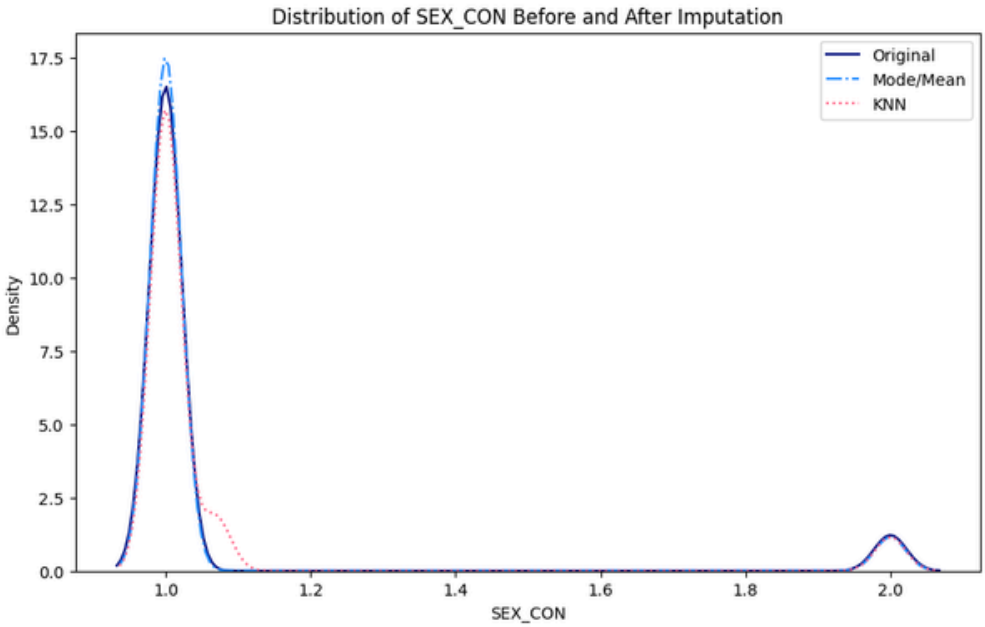


Figure 62: La distribution de sexe du conducteur avec KNN, moyenne et mode par rapport au originale.

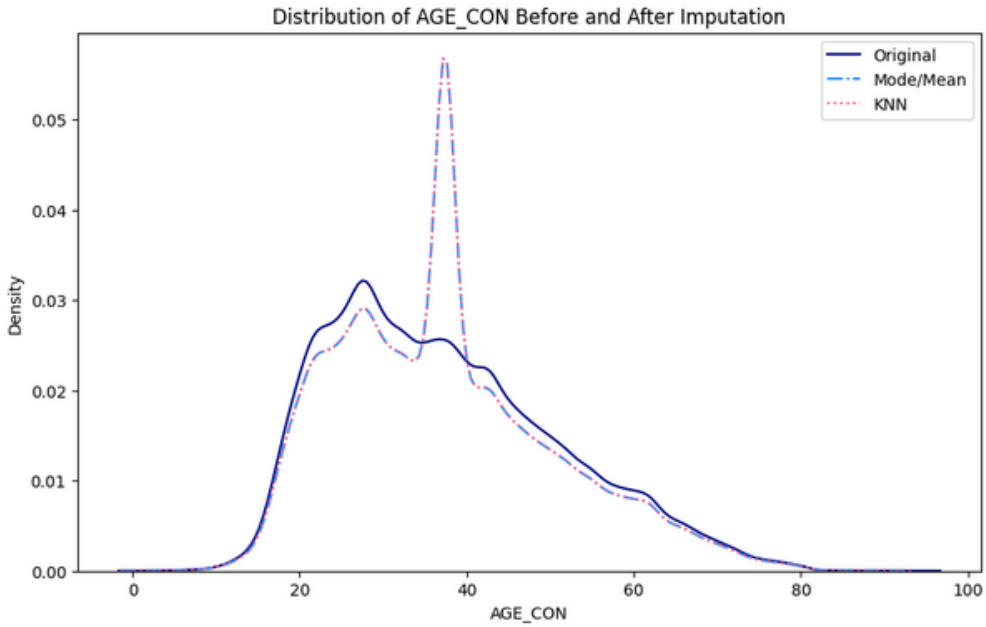


Figure 63: La distribution de l'âge du conducteur avec KNN, moyenne et mode par rapport au originale.

Dans ce cas, nous choisirons de manière préférentielle la méthode KNN pour imputer notre jeu de données, car elle présente une proximité significative avec les données originales.

2.Union de données

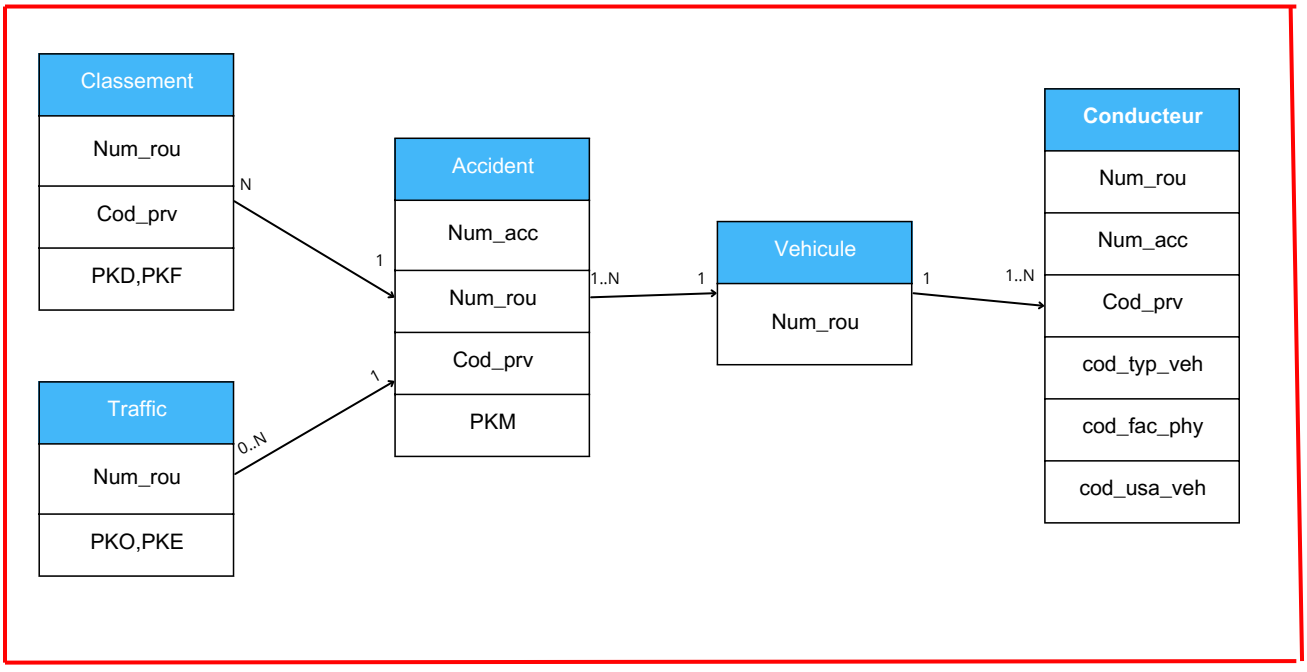


Figure 64: Les clés de jointure principales qui assurent la liaison entre les 5 tables.

Pour le PKM, il doit être sélectionné de manière à être supérieur au PK de début et inférieur ou égal au PK de fin : PKD < PKM ≤ PKF. Cette règle s'applique également pour la table de classement et pour le trafic.

3.Ingénierie des caractéristiques

Cette section se concentre principalement sur trois aspects : la création de nouvelles caractéristiques, la vérification de leur encodage, et la sélection des caractéristiques les plus pertinentes, en se basant notamment sur l'ANOVA. Nous avons initialement choisi la variable "class_isr" comme variable indépendante, également appelée variable cible. Elle est subdivisée en quatre classifications distinctes, notées de 1 à 4, dont la signification a déjà été présentée dans la phase de conception du projet.

- Tout d'abord, il est essentiel de vérifier l'encodage des caractéristiques.
- Des nouvelles caractéristiques sont créées par exemple en regroupant les 17 catégories de types de véhicules en 8, ce qui simplifie l'analyse dans un modèle prédictif.
- Le choix des caractéristiques : Pour ce faire, nous avons mis en œuvre deux méthodes clés : la corrélation et l'ANOVA.
- Initialement, notre démarche consistait à visualiser la corrélation entre les différentes caractéristiques. Cela nous permettait de saisir les motifs et les relations entre les variables. À cette fin, nous avons utilisé un heatmap, une représentation graphique qui met en évidence les niveaux de corrélation entre les différentes paires de caractéristiques. Pour créer ce heatmap, nous avons employé les bibliothèques Seaborn et Matplotlib,

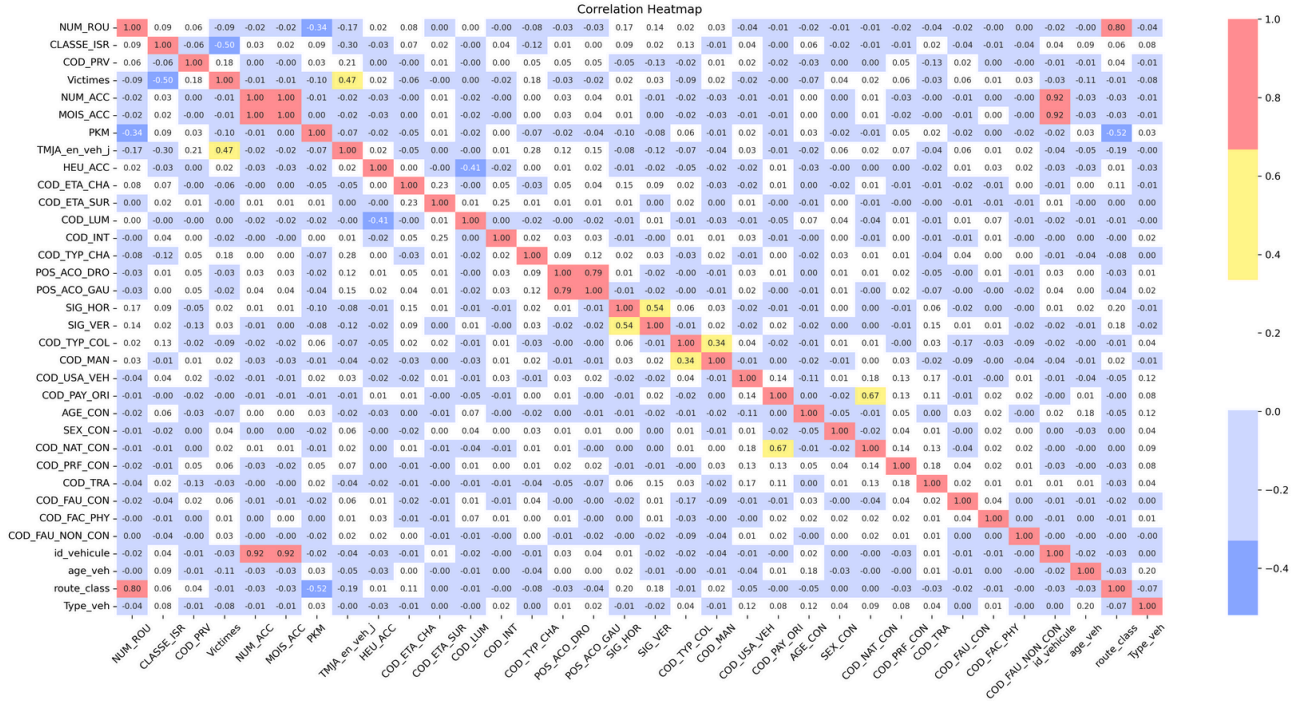


Figure 65 : La corrélation avant l'ingénierie des caractéristiques

D'après l'analyse de la corrélation effectuée à partir de ce graphique, il semble que les variables ne présentent pas de corrélation significative avec la variable cible(près a 1). Afin de mieux sélectionner les caractéristiques, nous envisageons de visualiser la distribution des variables individuellement et de les analyser manuellement. Cette approche nous permettra d'identifier les caractéristiques les plus pertinentes pour notre modèle.

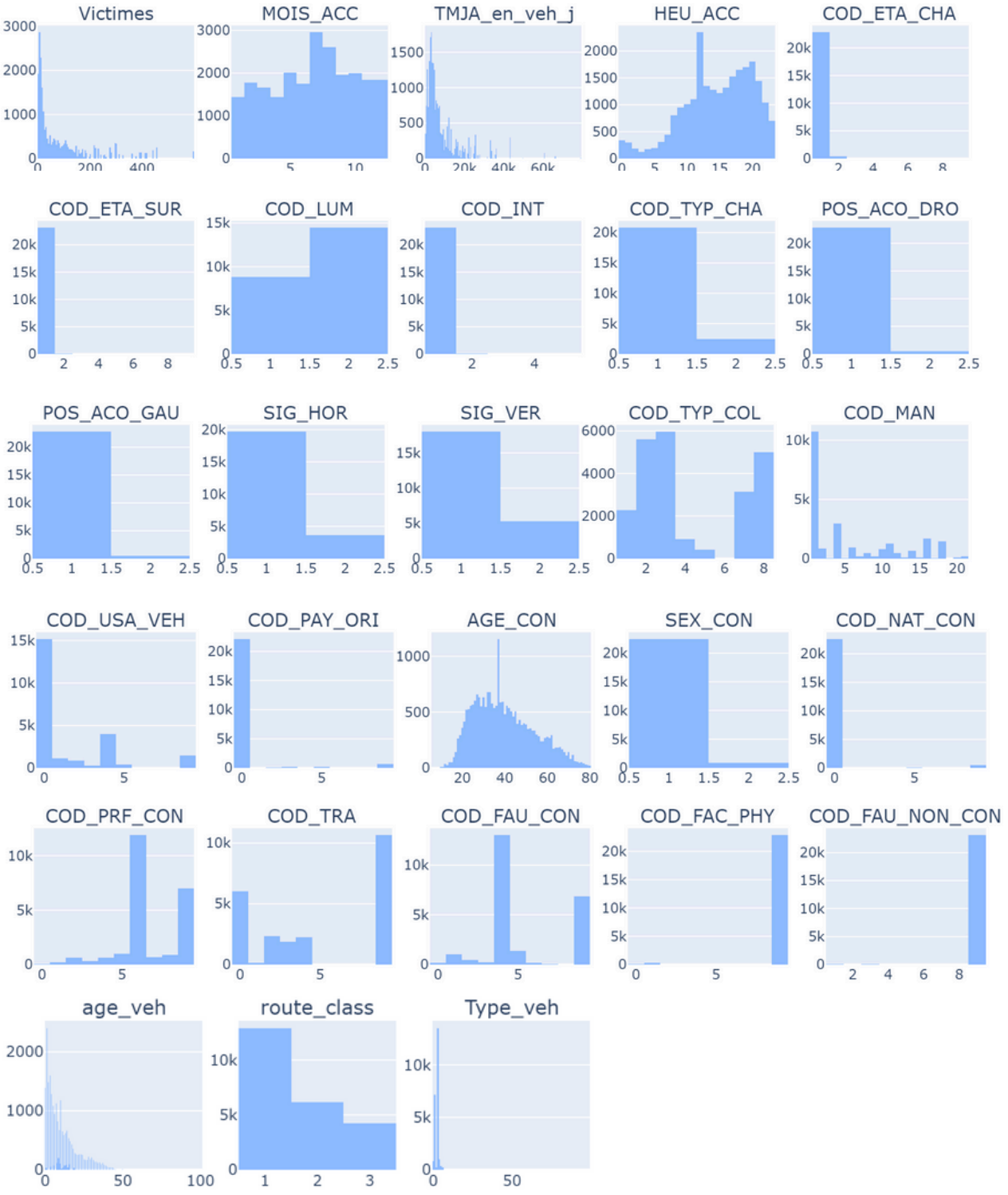


Figure 66 : Distribution des colonnes du tableau final

Après avoir examiné la distribution des variables, nous avons remarqué que certaines d'entre elles présentent plus de 95 % de leurs valeurs dans une seule catégorie, ce qui peut affecter négativement l'entraînement de notre modèle. Par conséquent, nous exclurons ces variables de notre processus de sélection lors de l'utilisation de l'ANOVA. Cependant, d'autres variables ne seront pas nécessaires pendant l'entraînement du modèle, mais nous aurons besoin de les prendre en compte lors de la visualisation des résultats. Parmi ces variables, nous trouvons : NUM_ROU, NUM_ACC, COD_PRV, PKM, et id_vehicule.

Voici la visualisation de la corrélation après avoir sélectionné les variables en utilisant l'ANOVA avec selection de k=16.

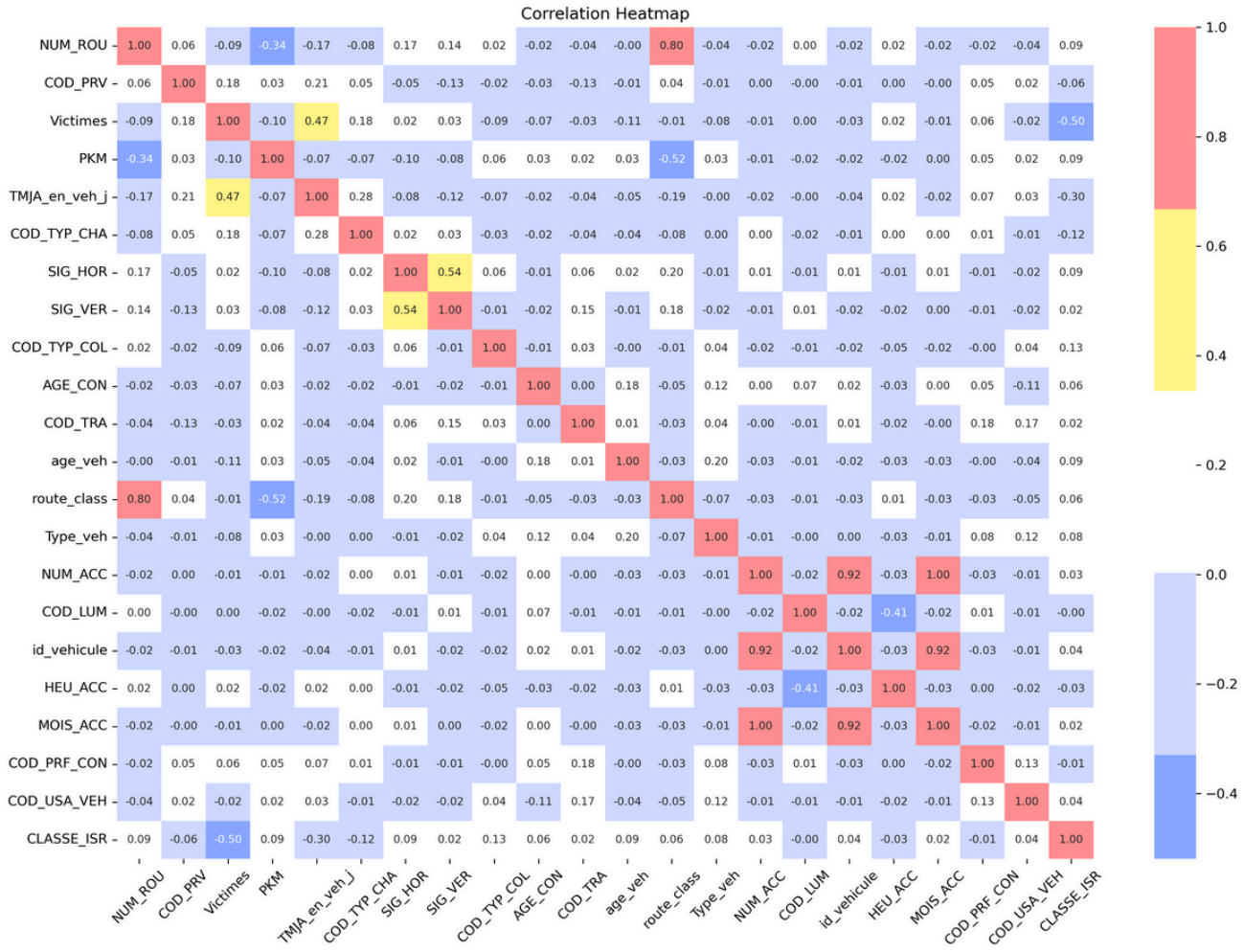


Figure 67 : La corrélation après l'ingénierie des caractéristiques

Donc, en nous basant sur les caractéristiques de risque d'accident ainsi que sur celles des accidents eux-mêmes, utiliserons la méthode ANOVA. Voici les variables que nous avons sélectionnées :

- **Comportement du conducteur :** AGE_CON, COD_PRF_CON
- **Conditions environnementales et caractéristiques sur l'accident:** COD_LUM, MOI_ACC, HEU_ACC, COD_TYP_COL, SIG_HOR, SIG_VER, COD_TYP_CHA, TYP_ROU, TJMA, Victimes.
- **État du véhicule :** TYP_VEH, COD_TRA, AGE_VEH,COD_USA_VEH.

4.Traitement des valeurs aberrantes

L'analyse des valeurs aberrantes est essentielle pour garantir la précision des interprétations des données. En employant la méthode interquartile, reconnue pour son efficacité, nous calculons ces valeurs atypiques. Grâce aux boxplots de Plotly, nous pouvons aisément les repérer, ce qui simplifie le choix des techniques appropriées de traitement, comme l'élimination ou la transformation, afin de renforcer la robustesse des analyses.

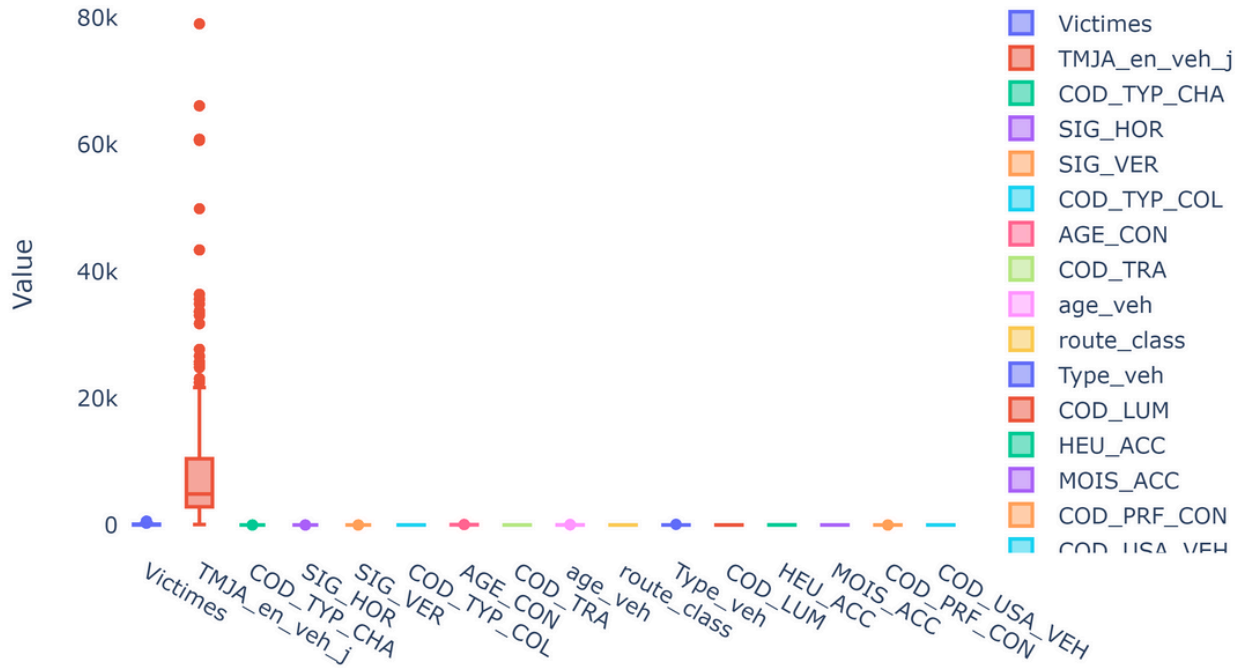


Figure 68 : Distribution des valeurs aberrantes avec le Boxplot .

Selon l'analyse du boxplot, il semble que cette table contienne un nombre significatif de valeurs aberrantes. Cependant, avant de les supprimer, il est essentiel de les examiner plus en détail. Pour ce faire, nous allons afficher les exemples de ces valeurs aberrantes pour chaque champ afin de mieux comprendre leur nature et leur impact sur nos données.

5.Traitement des valeurs dupliquées

L'absence de valeurs dupliquées dans notre tableau final est due à la qualité du processus de collecte des données. Chaque entrée unique représente un renseignement distinct qui a été soigneusement enregistré, garantissant ainsi la variété et l'unicité de notre ensemble de données. Cette précision démontre notre souci de fournir des informations complètes et exactes pour des analyses précises et fiables. En éliminant les doublons potentiels dès la collecte des données, nous avons assuré la qualité de notre tableau final, renforçant ainsi la validité de nos résultats et conclusions.

Tableau 9 : Exemple des valeurs aberrantes

Column	Number of Outliers	Examples of Outliers
Victimes	2169	303, 303, 303, 303, 303
TMJA_en_veh_j	2190	60870, 60870, 60870, 60870, 60870
COD_TYP_CHA	2494	2, 2, 2, 2, 2
SIG_HOR	3614	2, 2, 2, 2, 2
SIG_VER	5289	2, 2, 2, 2, 2
AGE_CON	21	80, 80, 80, 80, 80
age_veh	848	35, 39, 37, 36, 34
Type_veh	5	99, 99, 99, 99, 99
COD_PRF_CON	272	1, 1, 1, 1, 1

Pour une meilleure compréhension de notre tableau de données, nous avons choisi deux variables, une continue et une catégorique : "age_con" et "cod_prf_con". Bien que ces variables présentent des valeurs aberrantes telles que 2 pour "cod_prf_con" et 80 pour "age_conducteur", qui peuvent sembler inhabituelles, elles peuvent être valides dans certains contextes. Ces décisions nécessitent une évaluation par des experts pour déterminer si ces valeurs sont véritablement aberrantes. Cela souligne l'importance de consulter des spécialistes pour évaluer la pertinence des valeurs aberrantes dans nos données.

VI. Construction de model prédictive

Après une préparation et un traitement approfondis des données, la phase de construction du modèle prédictif devrait s'avérer plus aisée. Nous adopterons une démarche méthodique pour élaborer notre modèle, en nous appuyant sur deux approches distinctes d'entraînement. D'une part, nous explorerons l'entraînement de modèles individuels, en utilisant des techniques telles que K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), et Random Forest (RF). D'autre part, nous nous tournerons vers les méthodes d'ensemble learning, en particulier le stacking, pour renforcer la précision et la robustesse de notre modèle. Cette stratégie double nous permettra de comparer l'efficacité des approches et de choisir la plus adaptée à nos besoins.

1.Division de l'Ensemble de Données

Divisez vos données en ensembles d'entraînement, de validation et de test pour évaluer efficacement la performance du modèle en utilisant une méthode aléatoire. Les proportions seront les suivantes :

- 75% de données d'entraînement : Utilisées pour entraîner le modèle.
- 12.5% de données de validation : Utilisées pour ajuster les hyperparamètres et éviter le surajustement.
- 12.5% de données de test : Utilisées pour tester la performance finale du modèle afin d'évaluer son efficacité sur des données non vues.

Voici un graphique qui illustre la répartition des données d'entraînement, de validation et de test en utilisant Plotly pour le variable TJMA_EN_VEH_J.

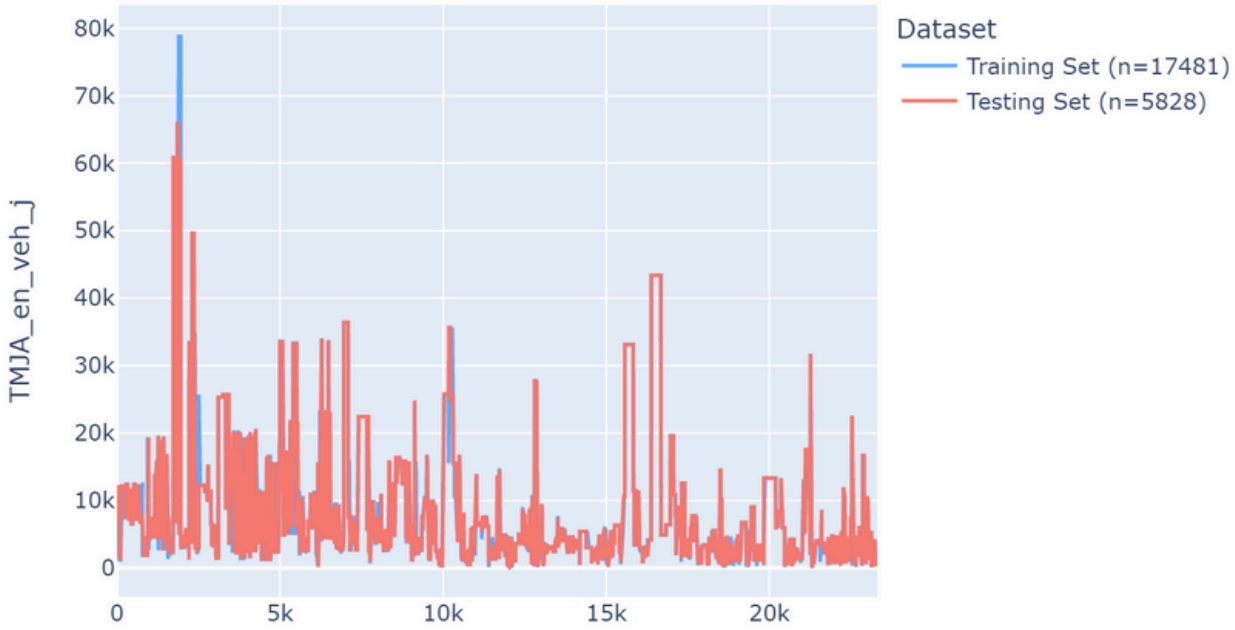


Figure 69 : Distribution de données d'entrainement et de test de colonne TJMA.

2.Entrainement individuel

Dans cette section, nous allons entraîner individuellement les modèles suivants :K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN) ,Random Forest (RF).

Chaque modèle utilise des paramètres spécifiques pour son entraînement. Voici les paramètres avec leur interprétation :

Tableau 10 : Les paramètres des trois algorithmes

Model	n_estimator	random_sta	n_neighbors	Input shape	Number of layers	Layer configuratio	Optimizer	Loss function
Random Forest	100	42	null	null	null	null	null	null
K-Nearest Neighbors	null	null	5	null	null	null	null	null
Artificial Neural Network	null	null	null	0	3	64 neurons (R	adam	sparse_catego

1.RF: Utilise 100 arbres et un `random_state` de 42, avec normalisation des données avant l'entraînement et dénormalisation après prédiction pour assurer la cohérence.
2.KNN : Opère avec 5 voisins, impliquant la normalisation des données avant l'entraînement et leur dénormalisation post-prédiction.
3.ANN: Configure 3 couches avec 64 neurones, utilisant Adam pour l'optimisation. Les données sont normalisées avant l'entraînement et dénormalisées après prédiction pour des tâches de classification multi-classes.

Chaque algorithme est attribué une durée spécifique pour l'entraînement et les tests. Il semble que l'algorithme ANN (Artificial Neural Network) nécessite plus de temps pendant la phase d'entraînement, tandis que celui de KNN (K-Nearest Neighbors) demande davantage de temps lors des tests.

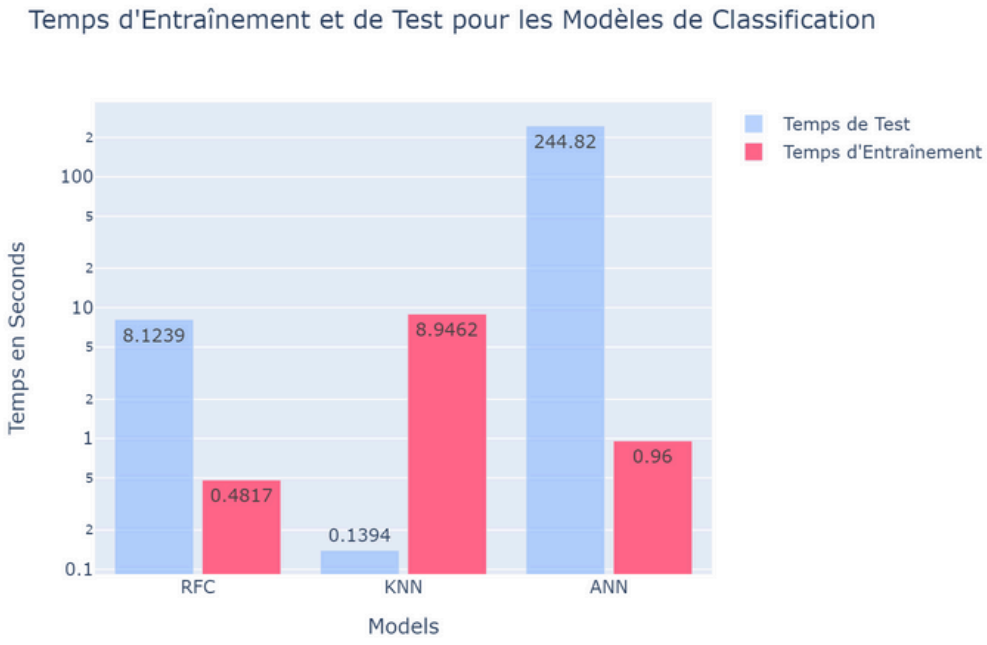


Figure 70 : Durée de test et entrainement de chaque algorithmes en second(Détaillée) .

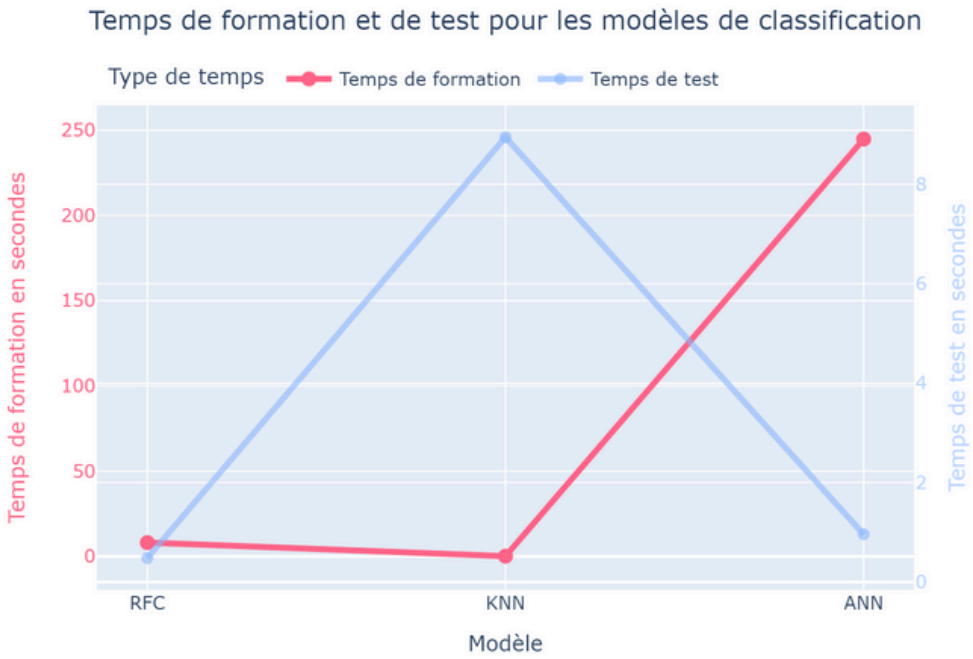


Figure 71 : Durée de test et d'entraînement de chaque algorithme (distribution linéaire)

Voici les matrices de confusion calculées pour chaque modèle (RF, KNN, ANN) afin d'évaluer leurs performances respectives. Ces matrices permettent de mesurer l'efficacité de chaque méthode d'entraînement .

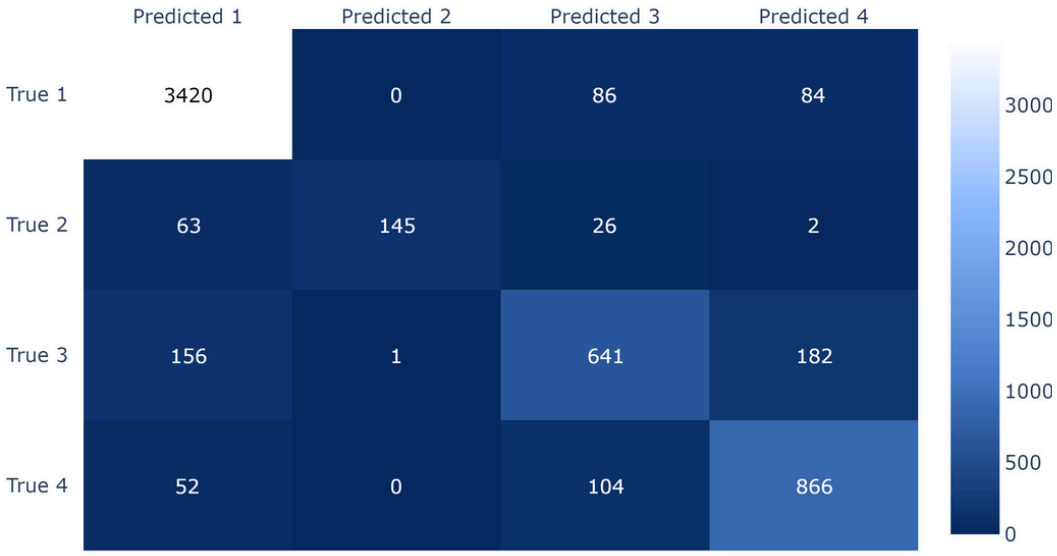


Figure 72 : Matrice de confusion de modèle RF.

Figure 73 : Matrice de confusion de modèle KNN.

Figure 74 : Matrice de confusion de modèle ANN.

Grâce aux ces matrices de confusion générées durant l'entraînement individuel de chaque algorithme, nous pouvons évaluer leurs performances en utilisant des métriques telles que l'Accuracy, le F1-Score, la Sensibilité (Recall) et la Précision. Parmi les algorithmes testés, le Random Forest se distingue par ses performances supérieures, affichant une Accuracy de 86.73%, une Précision de 93%, un F1-Score de 95%, et une Sensibilité de 94%. Ces résultats démontrent l'efficacité du Random Forest dans la classification des données étudiées.

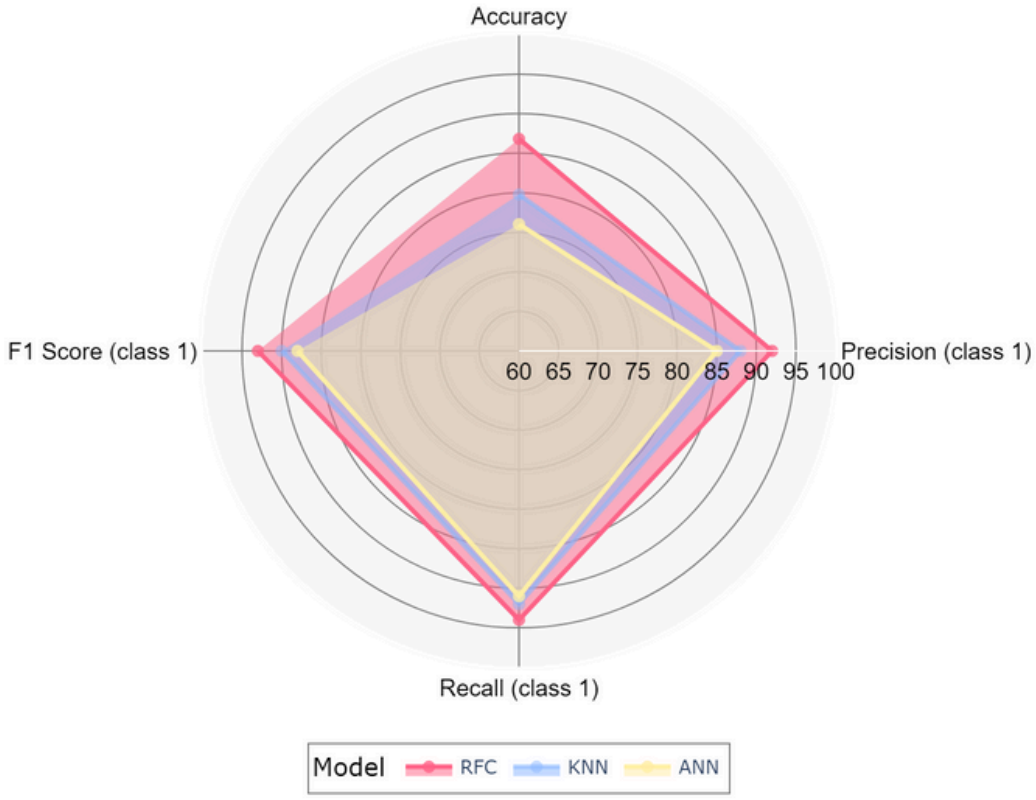


Figure 75 : Comparaison des indicateurs de performance des trois algorithmes.

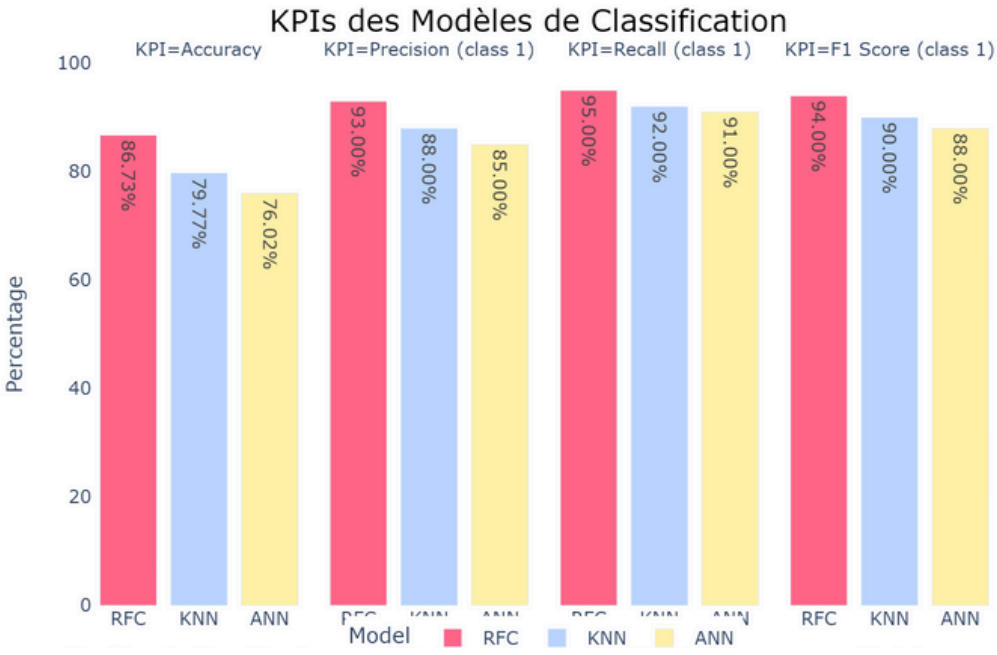


Figure 76 : Comparaison détaillée des indicateurs de performance des trois algorithmes.

Nous avons également utilisé l'AUC (Area Under the Curve) et la courbe ROC (Receiver Operating Characteristic) pour évaluer la performance des trois algorithmes. Nous avons sélectionné la classe 3 comme classe positive pour cette évaluation.

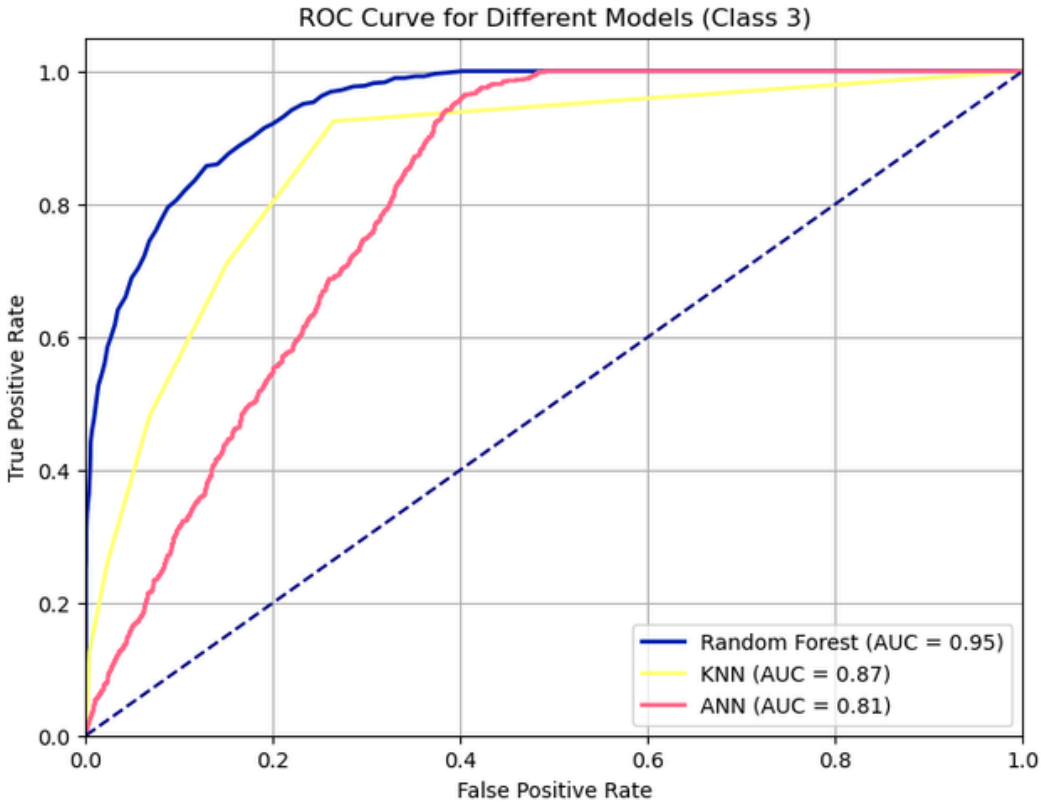


Figure 78 : La courbe ROC des trois algorithmes

Lors de l'analyse du graphique, il est clair que RF se distingue par sa performance supérieure par rapport aux autres algorithmes évalués. Son AUC de 95% indique une capacité exceptionnelle à discriminer entre les classes positives et négatives. Cette performance remarquable renforce la crédibilité et la fiabilité du modèle RF dans la résolution du problème spécifique étudié.

3.Entrainement en Ensemble learning methode de (Stacking).

L'entraînement des trois modèles, ANN, KNN et RF, est réalisé selon une méthode d'ensemble appelée stacking. Cette approche consiste à combiner les prédictions de ces modèles de base afin d'améliorer de manière significative la performance prédictive globale du système.
Voici les résultats de l'entraînement en apprentissage par ensemble, comparés à ceux du meilleur algorithme d'entraînement individuel, qui est le RF (Random Forest).

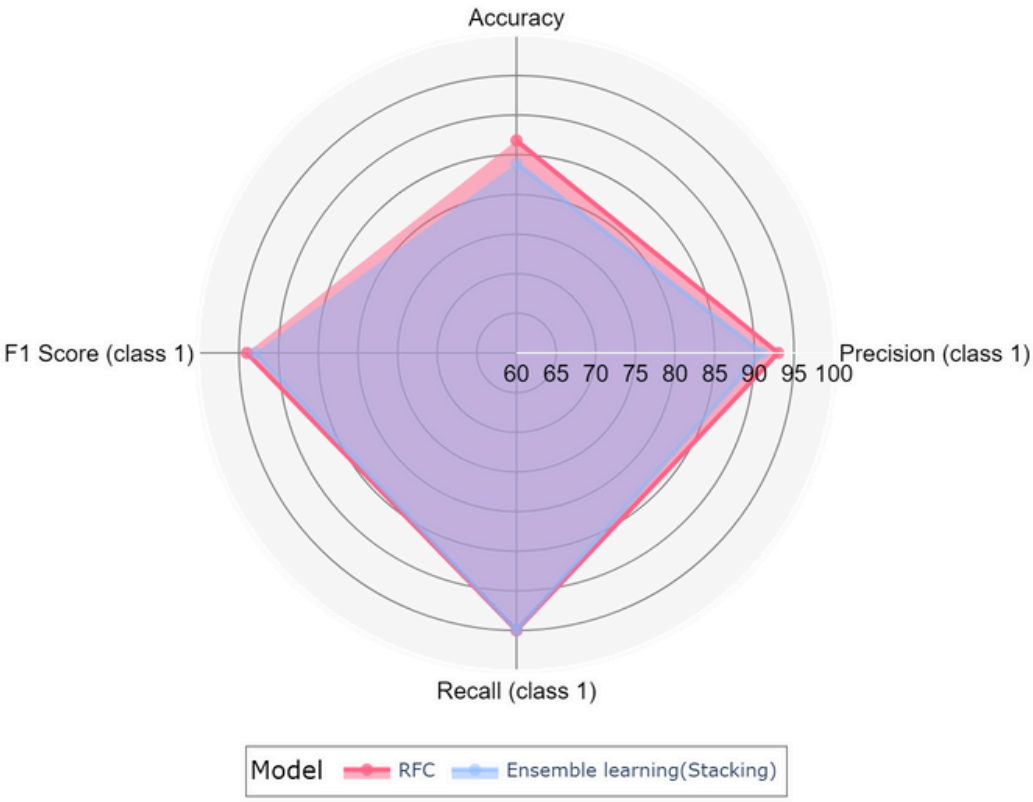


Figure 79 : Comparaison des indicateurs de performance des deux méthodes d'entrainement.

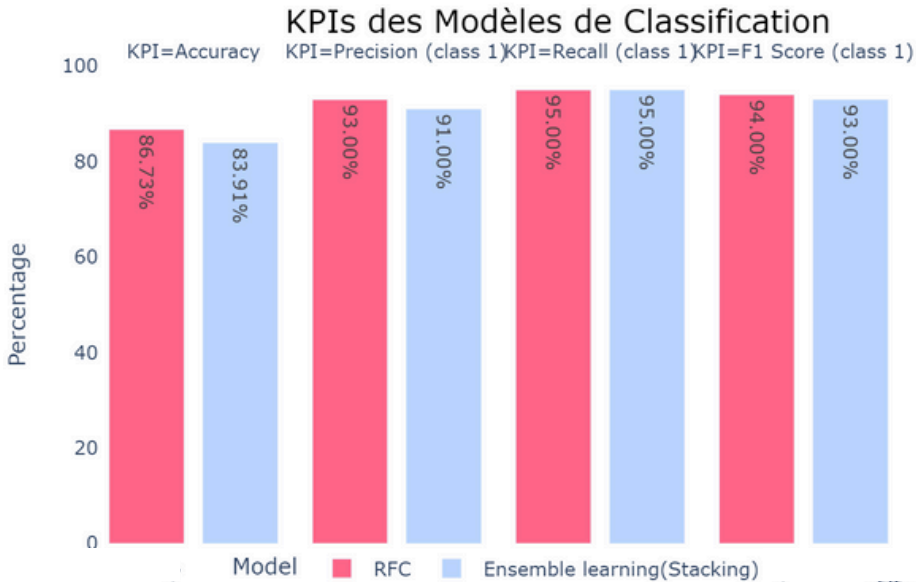


Figure 80 :Comparaison détaillée des indicateurs de performance des deux méthodes d'entrainement.

En analysant les données de la figure, nous observons que le modèle Ensemble Learning atteint une exactitude de 83,91%, tandis que le Random Forest réalise une performance de 86,73%. Cette différence de 2.82% en faveur du Random Forest démontre son efficacité supérieure. Par conséquent, nous opterons pour le modèle Random Forest pour nos résultats finaux, en raison de son avantage notable en termes des indicateurs de performances .

4.Souvgarde de Modèle finale

La sauvegarde de notre modèle finale qui le modele de RF est essentielle pour assurer sa pérennité et sa réutilisation efficace. Nous utilisons la bibliothèque `joblib` en Python pour cette tâche, car elle permet une sérialisation rapide et efficace des grands objets numpy, comme les modèles d'apprentissage automatique. En utilisant `joblib.dump`, nous sauvegardons notre modèle dans un fichier binaire, par exemple, `RFC_ACCIDENT_RISK_PREDICTION.pkl`, qui peut être facilement stocké et transféré. Pour recharger le modèle, nous utilisons `joblib.load`, ce qui restaure le modèle prêt à l'emploi. Cette méthode de sauvegarde facilite la préservation des résultats, le déploiement, la reproductibilité, et la gestion des mises à jour du modèle. `joblib` offre simplicité, efficacité, et compatibilité, rendant le processus de sauvegarde et de chargement à la fois rapide et fiable, garantissant ainsi que notre modèle peut être utilisé et amélioré en continu sans perte d'information.

VII. Visualisation de résultat

Après une analyse prédictive minutieuse, il ressort que le modèle Random Forest est supérieur aux autres alternatives. Pour présenter les résultats de manière plus explicite, nous proposons un tableau comprenant 20 échantillons. Chacun de ces échantillons est décrit par 11 caractéristiques clés, telles que la profession du conducteur, l'âge du véhicule, et divers aspects liés à l'accident, comme le trafic (TMJA) et le type de chaussée. Les échantillons incluent également des informations cruciales telles que le classement des points de risque prédit, le numéro de la route, et le point kilométrique (PKM), facilitant ainsi leur visualisation par l'Agence Nationale de Sécurité Routière (NARSA).

Pour approfondir notre compréhension des tendances régionales, une analyse détaillée par province sera également menée. Cette étude permettra de distinguer les variations de risque routier entre les provinces, en mettant en évidence celles présentant un nombre significatif de classements de niveau 1.

Un tel classement indique une sécurité routière particulièrement précaire, nécessitant une intervention urgente. L'objectif final est de permettre à la NARSA de visualiser ces données pour identifier les points qui requièrent une action ciblée et immédiate.

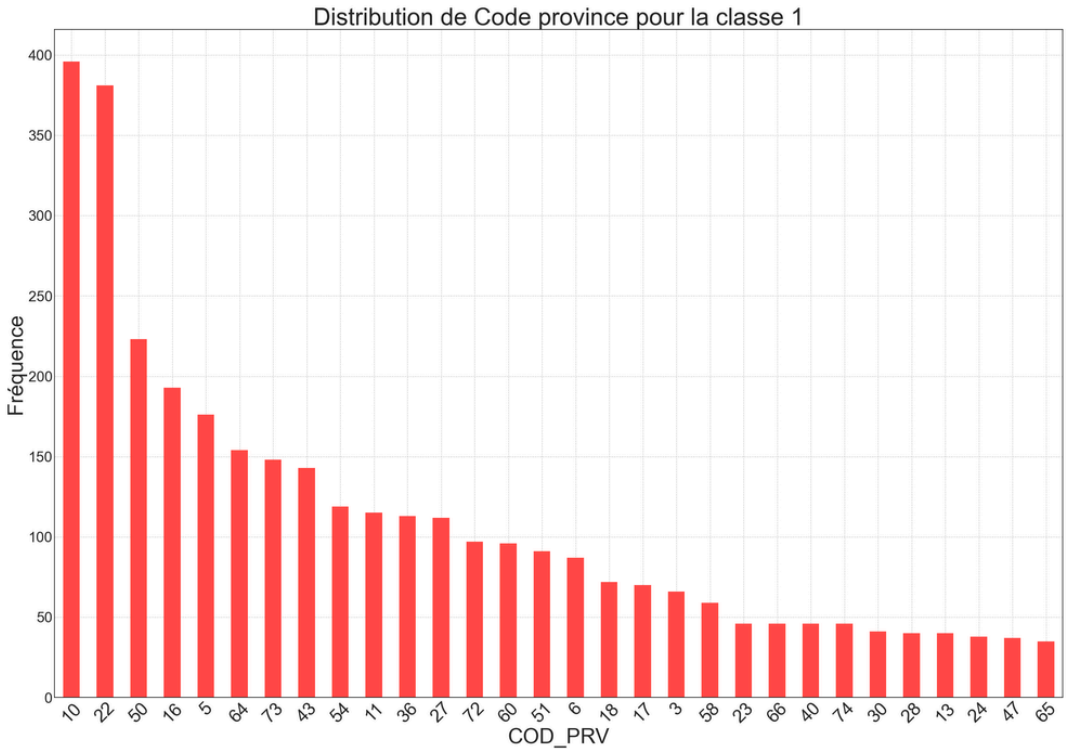


Figure 81 : Distribution des provinces par rapport au nombre de classement 1.

La province 10 détient le plus grand nombre de classements 1, représentant 399 occurrences, ce qui équivaut à 83.8% de sa valeur totale. Nous allons donc concentrer notre visualisation sur cette province afin d'examiner ses résultats en détail.

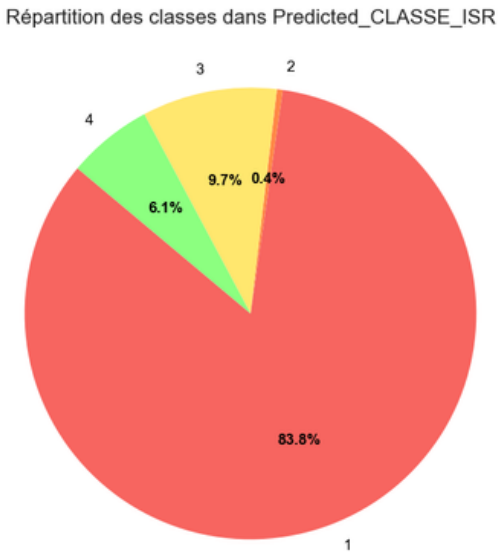


Figure 82: Distribution des classements de province El Jadida

Tableau 11 : La résultats de prédiction de classement sur la province de El Jadida

Predicted_CLASSE_ISR	PKM	NUM_ROU	COD_PRF_CON	COD_TYP_CHA	TMJA_en_veh_j	age_veh
3	461	1	6	1	12383	8
1	436	1	6	1	7195	5
1	439	1	6	1	12383	2
3	4	316	9	1	11490	14
1	18	303	9	1	4949	26
1	448	1	6	1	12383	5
4	126	202	6	1	3113	5
2	449	1	6	2	12383	1
1	54	316	5	1	3423	10
3	28	316	9	1	3423	38
1	462	1	9	1	11512	10
1	462	1	6	1	11512	7
4	5	303	9	1	4949	22
4	477	1	9	1	8225	13
2	3	316	9	1	11490	16
1	461	1	6	1	12383	6
1	28	316	8	1	3423	16
1	463	1	6	1	11512	7
1	503	1	9	2	6509	1
1	28	301	6	1	3037	4

Ainsi, ce tableau est une visualisation de 20 échantillons. Chaque échantillon est décrit par 11 caractéristiques clés, telles que la profession du conducteur, l'âge du véhicule, et divers aspects liés à l'accident, comme le trafic (TMJA) et le type de chaussée. Les échantillons incluent également des informations cruciales telles que le classement des points de risque prédit, le numéro de la route et le point kilométrique (PKM), facilitant ainsi leur visualisation par l'Agence Nationale de Sécurité Routière (NARSA). Ce tableau n'est qu'un échantillon du tableau final, qui couvre 74 provinces, 21 caractéristiques et comporte 5828 valeurs.

Maintenant, après la visualisation des données, nous allons procéder à une analyse approfondie pour identifier les schémas entre la variable cible, spécifiquement la classe 1, et les autres caractéristiques dans la province d'El Jadida. Cette étape est cruciale pour finaliser notre travail d'analyse prédictive.

- **Relations entre le classement 1 et les caractéristiques du conducteur :**

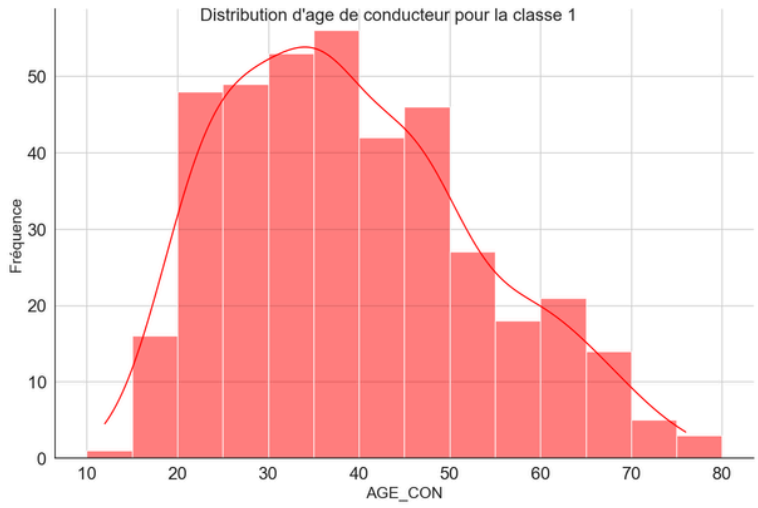


Figure 83: La distribution d'âge de conducteur par rapport au classement 1 .

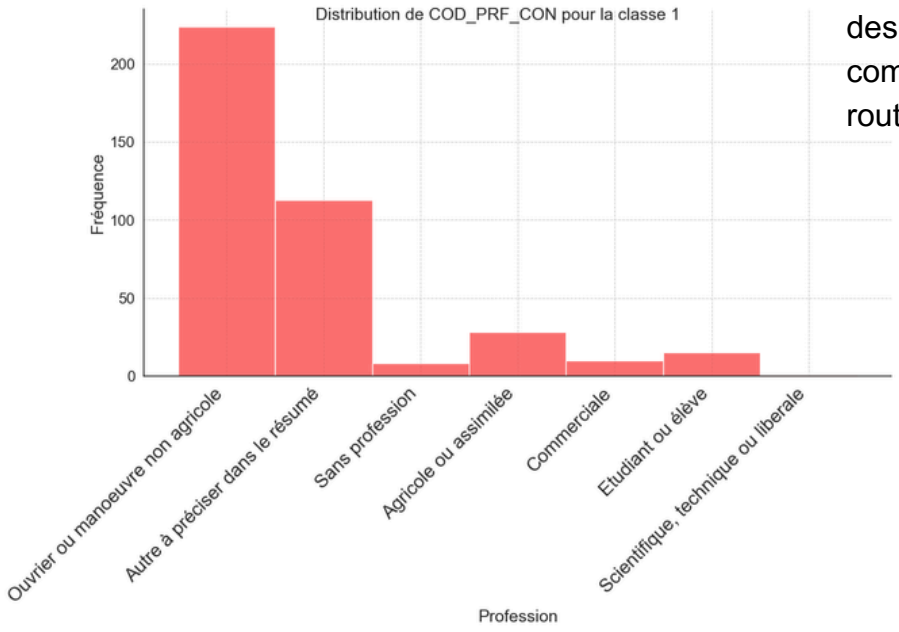


Figure 84: La distribution de profession de conducteur par rapport au classement 1

L'analyse des données révèle que le nombre de classements 1, indiquant un risque élevé ou une faible sécurité routière, est plus élevé chez les conducteurs âgés de 35 à 40 ans ainsi que chez ceux dont la profession est répertoriée comme ouvrier ou manœuvre non agricole. Ces observations suggèrent l'existence de facteurs spécifiques liés à ces tranches d'âge et à ces professions, qui peuvent contribuer à une augmentation des incidents ou des comportements à risque sur la route.

- **La relation entre le classement 1 et l'état ainsi que la condition du véhicule:**

L'analyse détaillée des facteurs affectant les classements de sécurité révèle une augmentation notable des classements de niveau 1, signe d'une sécurité routière déficiente, pour les véhicules de moins de 7 ans.
Cette tendance est également prononcée pour les véhicules appartenant directement aux conducteurs.
De plus, les trajets dont les objectifs ne sont pas clairement définis, c'est-à-dire ceux qui ne correspondent pas aux activités courantes telles que le trajet domicile-travail, domicile-école, les courses, les achats, l'utilisation professionnelle, ou les loisirs, sont également susceptibles de recevoir un classement 1.
Il est également observé que les véhicules tels que les voitures de tourisme et les cycles à moteur sont fréquemment associés à un niveau de risque plus élevé.

Ces résultats soulignent l'urgence de mettre en œuvre des stratégies de maintenance plus rigoureuses et de clarifier les intentions de trajet pour chaque déplacement afin de minimiser les risques d'accidents et d'améliorer la sécurité routière.

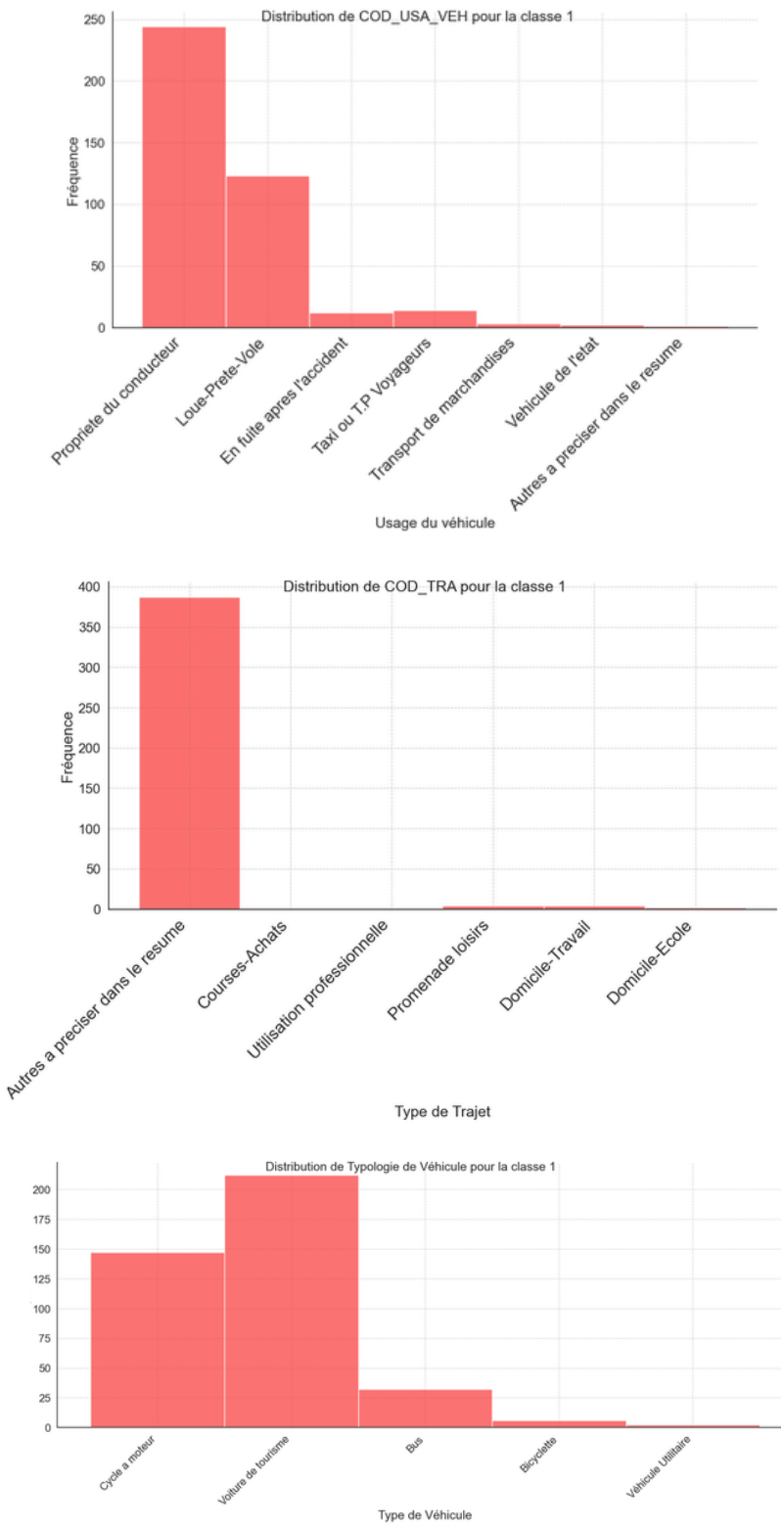


Figure 85: La distribution de trajet et type de route et le type d'usager de véhicule par rapport au classement 1

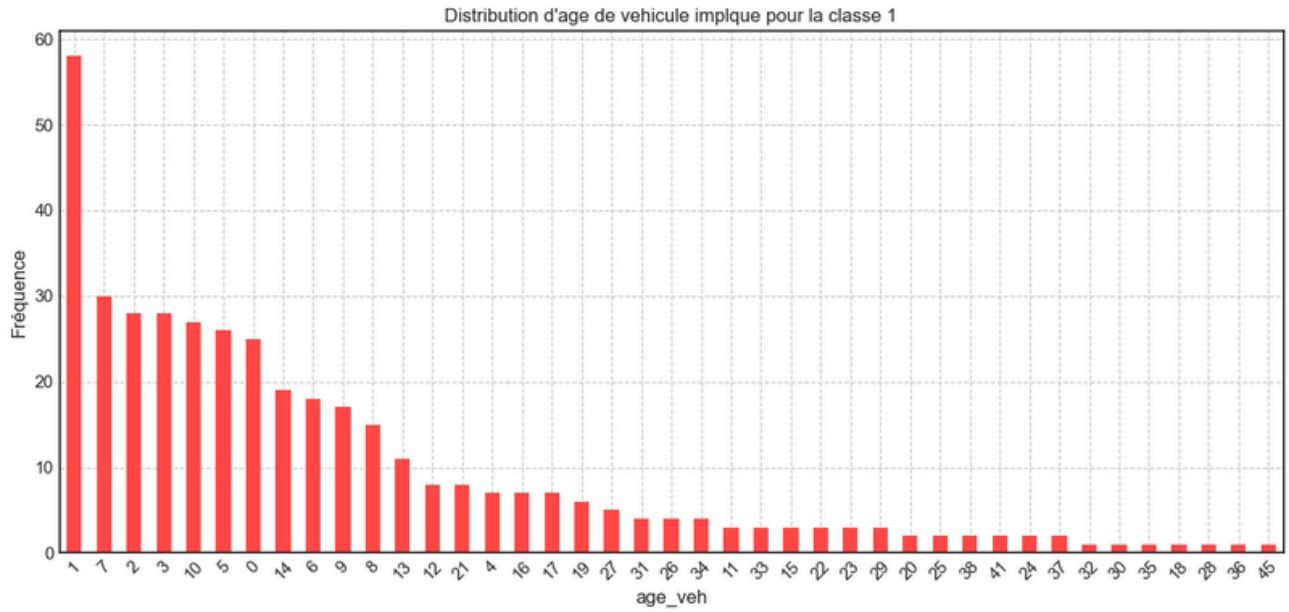
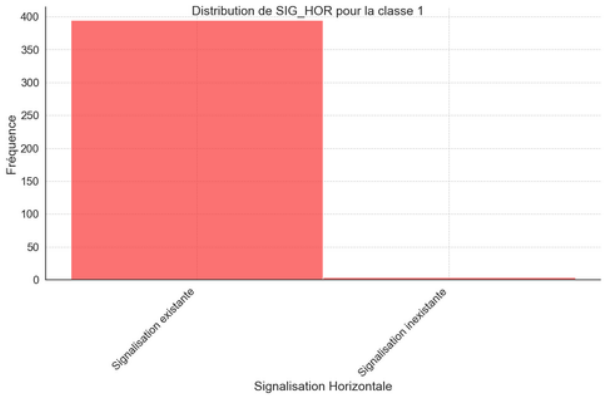
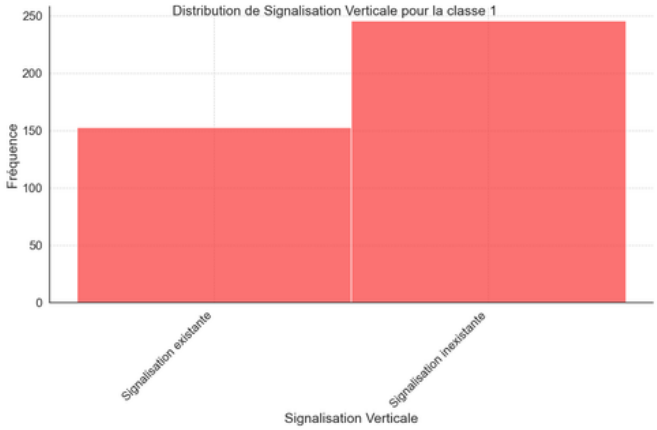
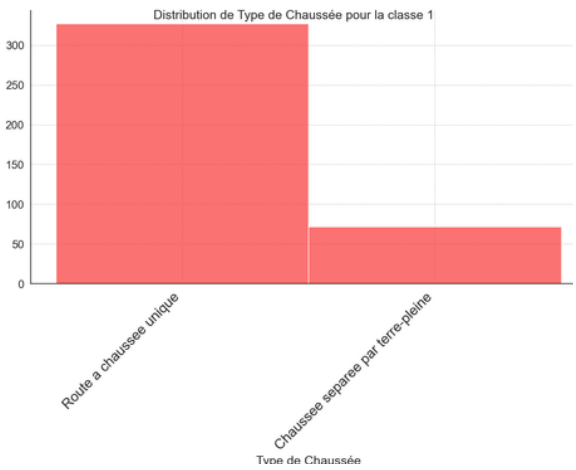


Figure 86: La distribution d'âge de véhicule par rapport au classement 1 .

- **La relation entre le classement 1 et les conditions métrologique et conditions de l'accident :**







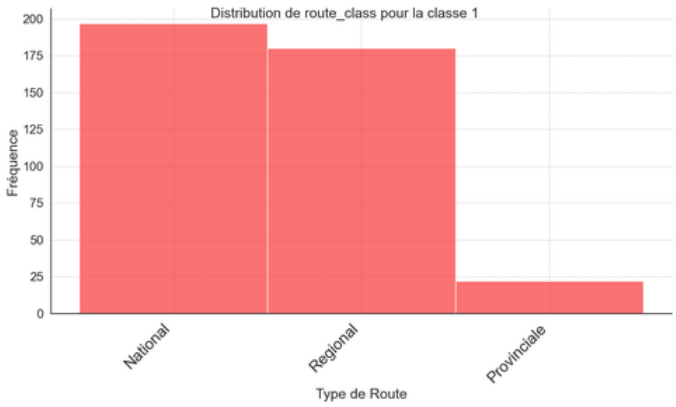


Figure 87: La distribution des conditions métrologique et conditions de l'accident par rapport au classement 1

L'observation et l'analyse rigoureuse de ces facteurs de risque routier dans ces figures met en lumière une prévalence alarmante d'incidents classés au niveau 1, révélant une crise profonde en matière de sécurité routière. Cette situation est particulièrement exacerbée par des conditions météorologiques défavorables, des configurations de trafic spécifiques, des caractéristiques de l'accident et d'autres facteurs environnementaux. Il est notable que durant les mois d'été, notamment en juillet et août, et aux heures de pointe, à midi(12) et à 20 soirée, nous observons une nette augmentation des incidents critiques. Sur les routes nationales, les points où le Trafic Moyen Journalier Annuel (TMJA) se situe à des seuils de un peu près 1900, 3500, ou 6700 véhicules correspondent à des pics d'incidents de niveau 1.
Par ailleurs, les chaussées uniques présentent des risques significativement plus élevés par rapport aux chaussées séparées par terre-pleine.

En ce qui concerne les types de collisions, les impacts latéraux ou arrière, souvent exacerbés par une forte luminosité diurne, sont directement associés à une hausse des classements de niveau 1. L'efficacité de la signalisation horizontale, combinée à l'absence de signalisation directionnelle à gauche, semble également influencer cette tendance. De plus les sections de route recensant plus de 400 victimes, sont des marqueurs clairs de zones hautement problématiques.

Face à ces constats, l'urgence de renforcer la sécurité routière se fait ressentir avec acuité, particulièrement durant les périodes et dans les zones les plus susceptibles de risques. La mise en place de mesures proactives, telles que l'amélioration de la signalisation, la gestion optimisée du trafic, et la maintenance régulière des infrastructures, est impérative. Ces actions sont essentielles non seulement pour réduire la fréquence des incidents graves mais aussi pour assurer une sécurité accrue à tous les usagers de la route, minimisant ainsi les risques et protégeant les vies.

Conclusion

Pour conclure ce chapitre, nous avons non seulement élaboré un modèle prédictif conçu pour être mis en pratique, mais également examiné les facteurs clés susceptibles de nous aider à éviter les zones à haut risque. En utilisant notre modèle de forêt aléatoire (RFC) et en analysant les facteurs ou variables de la pyramide de risque, nous avons pu établir des directives précises pour améliorer la sécurité routière. Le classement 1, signifiant une sécurité routière très mauvaise, nécessite une intervention immédiate. Nos analyses fournissent ainsi des recommandations claires, basées sur nos recherches, pour renforcer la sécurité sur les routes.

Conclusion générale

Notre étude démontre l'efficacité d'un modèle prédictif utilisant des techniques avancées de forêt aléatoire pour améliorer la sécurité routière au Maroc, en particulier dans les zones rurales. Ce modèle s'intègre parfaitement à la stratégie nationale de sécurité routière "NARSA" 2017-2026, visant à réduire les incidents et mortalités sur les routes. L'utilisation de la forêt aléatoire a permis d'identifier de manière fiable les facteurs de risque et de prédire les zones à haut risque. Cette capacité prédictive offre un potentiel immense pour des interventions ciblées et proactives, optimisant ainsi l'utilisation des ressources et renforçant les efforts de prévention. Ce projet illustre la puissance du Big Data et de l'analyse prédictive dans la sécurité publique, et marque un pas significatif vers la réduction de la mortalité routière. Les méthodologies et techniques développées serviront de modèle pour des initiatives futures, tant au niveau national qu'international.

Recommandations :

1. Formations Ciblées :
- Organiser des sessions de formation spécifiques pour les conducteurs âgés de 35 à 40 ans et pour les ouvriers, en se concentrant sur les pratiques de conduite sécuritaire afin de réduire les risques d'accidents.

2. Contrôles Techniques Renforcés :
- Imposer des contrôles techniques semestriels pour les véhicules de moins de 7 ans, ciblant en particulier les systèmes de sécurité vitaux pour prévenir les défaillances mécaniques.

3. Optimisation des Trajets :
- Promouvoir l'utilisation de technologies de navigation avancées pour aider les conducteurs à éviter les routes à haut risque et les périodes de forte affluence.

4. Amélioration de la Signalisation :
- Améliorer la signalisation routière, surtout sur les chaussées uniques où les déficits en matière de signalisation peuvent augmenter significativement les risques d'accidents.

5. Déploiement de Modèles Prédictifs :
- Appliquer la modélisation prédictive pour identifier et cibler les interventions dans les zones et aux moments où le risque d'accidents est le plus élevé, optimisant ainsi l'utilisation des ressources de patrouille.

6. Interventions Rapides:
- Établir des protocoles d'intervention rapide dans les secteurs où les statistiques montrent une concentration élevée d'accidents, pour réduire les impacts en cas d'incident.

Bibliographie

[1] Rob Kitchin , The Data Revolution: Big Data, Open Data, Data Infrastructures and The Consequences,2014.

[3] Peter Sondergaard <https://metropoleposition.fr/07-optimiser-la-gestion-de-ses-donnees/> (la dernière consultation le 30/04/2024.)

[4] Eric Siegel dans "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die",2013.

[5] Tom Mitchell ,"Machine Learning",1983.

[7],Breiman et al. Random forests, Machine Learning, 45, 2001.

[8] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[9] Koklu, M., I. Cinar, and Y.S. Taspinar, Classification of rice varieties with deep learning methods. Computers and electronics in agriculture, 2021. 187: p. 106285.

[11] Q. A. Al-Radaideh et E. J. Daoud, « Data Mining Methods for Traffic Accident Severity Prediction ».

[13] Little R.J.A. et Rubin D.B., Statistical Analysis with Missing Data, Wiley series in probability and statistics, 1987.

[14] Stekhoven D.J. et Bühlmann P., MissForest - nonparametric missing value imputation for mixed-type data, Bioinformatics Advance Access (2011).

[15] Ropelewska, E., X. Cai, Z. Zhang, K. Sabanci, and M.F. Aslan, Benchmarking Machine Learning Approaches to Evaluate the Cultivar Differentiation of Plum (Prunus domestica L.) Kernels. Agriculture, 2022. 12(2).

Webographie

Pour plus d'information sur National Road Safety Agency(NARSA) ,Visitez <https://www.narsa.ma/fr/nos-missions> (Dernière consultation le 30/05/2024).

[2]Data life cycle , Google "Foundations: Data, Data, Everywhere"<https://www.coursera.org/learn/foundations-data?specialization=google-data-analytics> (la dernière consultation le 30/04/2024)

[6] Pour plus de détails voir : < https://dataaspirant.com/how-decision-tree-algorithm-works/ >(la dernière consultation le 04/05/2024).

[10] Pour plus d'information : <https://aws.amazon.com/fr/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain.> (la dernière consultation le 10/05/2024).

[12] Singh, G., Pal, M., Yadav, Y. et al. Deep neural network-based predictive modeling of road accidents. [5]Neural Comput & Applic 32, 12417–12426 (2020). < https://doi.org/10.1007/s00521-019-04695-8> (la dernière consultation le 11/05/2024).

[12] Pour plus de détail voir : < https://www.mdpi.com/2412-3811/5/7/61#B10-infrastructures-05-00061 > (la dernière consultation le 11/05/2024).

[15] Ropelewska, E., X. Cai, Z. Zhang, K. Sabanci, and M.F. Aslan, Benchmarking Machine Learning Approaches to Evaluate the Cultivar Differentiation of Plum (Prunus domestica L.) Kernels. Agriculture, 2022. <https://www.mdpi.com/2077-0472/12/2/285> (la dernière consultation le15/05/2024).

[16] :Pour plus d'information voir <https://www.python.org/>
[17] :Pour plus d'information voir <https://www.anaconda.com/ >
[18] :Pour plus d'information voir <https://jupyter.org/>
[20]Pour plus d'information voir <https://www.microsoft.com/fr-fr/power-platform/products/power-bi>
[19]Pour plus d'information voir <https://www.r-project.org//>
[21]Pour plus d'information voir <https://www.microsoft.com/fr-fr/microsoft-365/excel>
[22]Pour plus d'information voir <https://www.teamgantt.com/h2?originalReferrer=https://www.google.com/>

Annexe

Quelque concept de sécurité routière

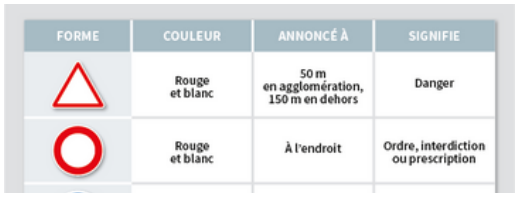
- Type de chaussée :

Séparée par terre plan

Unique

- Signalisation :

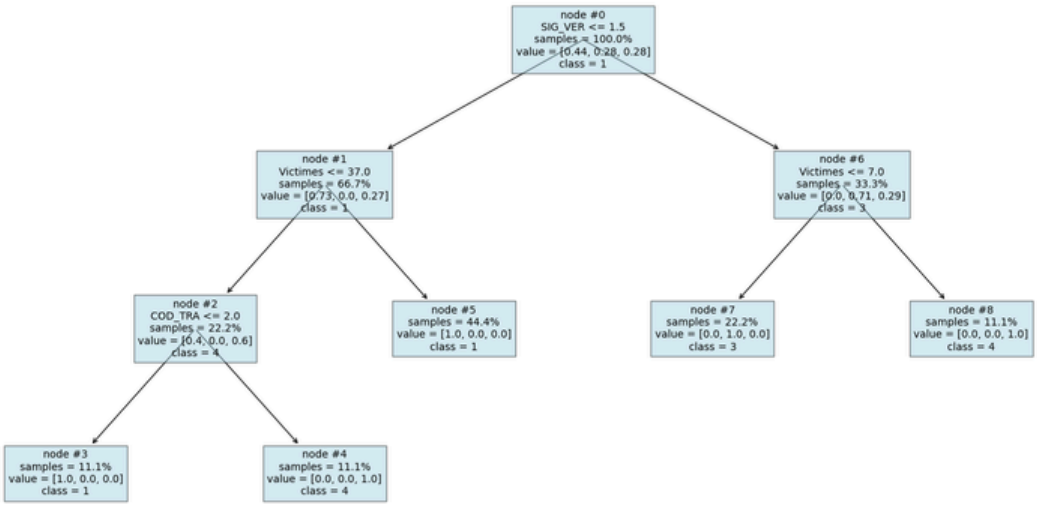
Horizontale:comprend toutes les marques peintes ou incrustées sur la surface de la route. Ces marquages guident et régulent la circulation en fournissant des informations directement sur la chaussée.



Verticale:fait référence à tous les panneaux de signalisation installés au bord des routes ou suspendus au-dessus de la chaussée.

Illustration d'un arbre de décision issu d'une forêt aléatoire de notre modèle

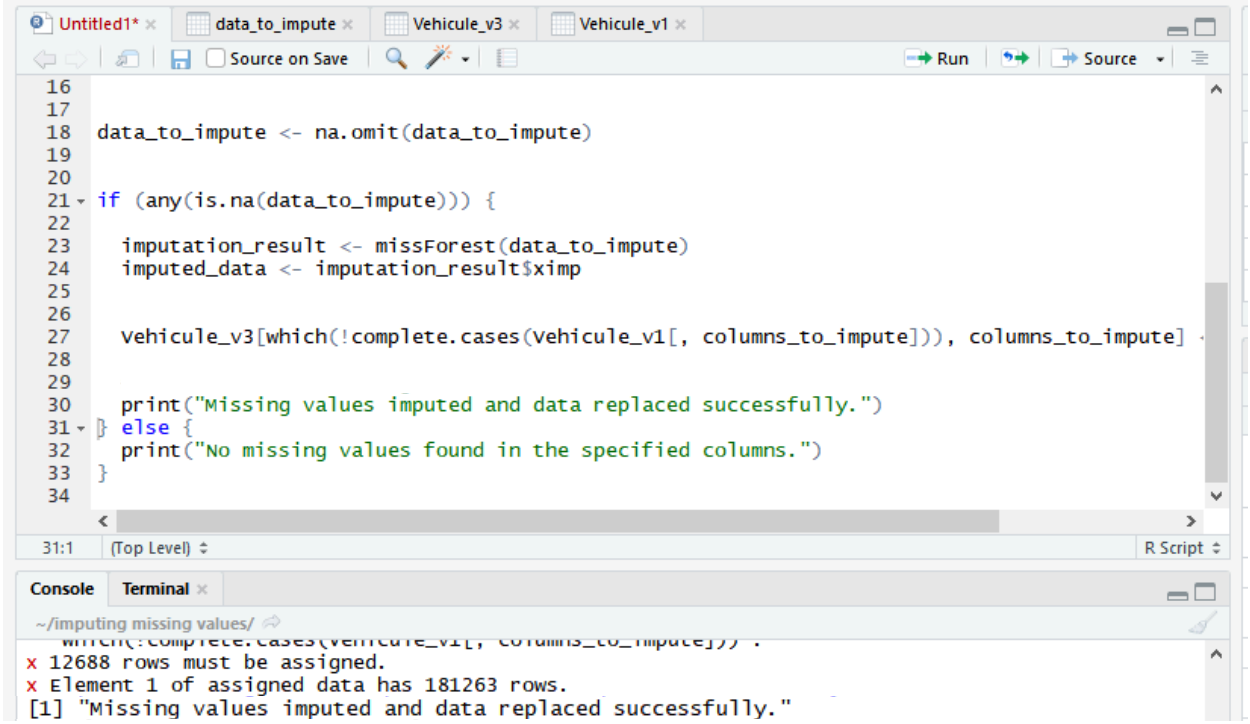
.



Annexe

Extrait de code

- **Choix d'imputation de valeurs manquantes: Algorithme MissForest dans le langage R.**



- **Exemple de création des nouvelle colonne (type de cause) utilisons code DAX (Power BI).**

```
MultiCausesAccident =
IF (
    CONTAINSSTRING ( accidents[CIR_ACC], "/" ),
    " Causes Multiple",
    " UniCause"
)

```

- **Imputation de donnees manquantes utilisons KNNimputer dans Python.**

```
import pandas as pd
from sklearn.impute import KNNImputer
import numpy as np
chemin_fichier = 'vehicule_1985.csv'
df = pd.read_csv(chemin_fichier)
categorical_cols = ['SEX_CON']
numerical_cols = ['AGE_CON']
knn_imputer_cat = KNNImputer()
df[categorical_cols] = knn_imputer_cat.fit_transform(df[categorical_cols])
knn_imputer_num = KNNImputer()
df[numerical_cols] = knn_imputer_num.fit_transform(df[numerical_cols])
print(df.head())
```