Rapport de projet Machine Learning

Arnaud Janet-Maitre, Léo Kirsch, Azou Gales

Janvier 2025

Contents

1 Présentation des données et de la problématique assurantielle

1.1 Présentation des données

Les données utilisées proviennent de l'INSEE et concernent les accidents de la route en France en 2023. Elles sont organisées de la façon suivante :

- **Caractéristiques (caract-2023.csv)** : Ce fichier contient des informations générales sur chaque accident, telles que la date, l'heure, les conditions de luminosité (`lum`), les conditions atmosphériques (`atm`), et le type de collision (`col`). Ces données permettent de comprendre le contexte global dans lequel chaque accident s'est produit.
- **Lieux (lieux-2023.csv)** : Ce fichier décrit les caractéristiques du lieu de l'accident, comme la catégorie de route (`catr`), le profil de la route (`prof`), l'état de la surface (`surf`) et la localisation précise de l'accident (`situ`). Ces informations sont essentielles pour analyser l'influence de l'environnement sur les accidents.
- **Véhicules (vehicules-2023.csv)** : Ce fichier recense les informations sur les véhicules impliqués dans chaque accident, notamment la catégorie du véhicule (`catv`), le point initial du choc (`choc`) et la manœuvre réalisée avant l'accident (`manv`).
- **Usagers (usagers-2023.csv)** : Ce fichier contient des informations sur les usagers impliqués dans les accidents, incluant leur position dans le véhicule (`place`), leur catégorie (`catu`), leur gravité (`grav`) et leur sexe (`sexe`).

Chaque fichier est relié par une clé commune, `Num_Acc`, qui identifie de manière unique chaque accident, ce qui permet de croiser les données pour une analyse approfondie.

1.2 Problématique assurantielle

Dans le cadre de ce projet, nous cherchons à prédire la gravité des accidents (`grav`), une variable issue du fichier *Usagers*, en nous basant sur l'ensemble des autres variables disponibles dans les données. Cette gravité est définie par les catégories suivantes :

- `1` : Indemne.
- `2` : Blessé léger.
- `3` : Blessé hospitalisé.
- `4` : Tué.

La gravité des accidents est une variable critique pour les assureurs automobiles, car elle impacte directement les coûts associés aux sinistres. Les prédictions sur cette gravité peuvent permettre :

- Une meilleure gestion des risques par les assureurs, en identifiant les profils ou les situations à risque.
- L'ajustement des primes d'assurance en fonction de la probabilité de gravité des sinistres.

L'objectif final est de développer un modèle prédictif capable de fournir une estimation précise de la gravité des accidents, en exploitant les caractéristiques des accidents, des lieux, des véhicules et des usagers.

2 Analyse préliminaire des données

2.1 Traitement des données

Dans cette section, nous présentons les étapes clés du traitement des données avant de les utiliser dans notre modèle prédictif.

2.1.1 Vérification de l'unicité de la clé primaire

La clé principale de la base de données est l'identifiant unique de chaque accident (`Num_Acc`). Cette variable est utilisée pour relier les différents fichiers (*Caractéristiques*, *Lieux*, *Véhicules*, *Usagers*). Une vérification préliminaire a permis d'identifier la présence de 4 doublons dans la clé `Num_Acc`, ce qui indique que ces identifiants ne sont pas totalement uniques. Les doublons ont été analysés et supprimés afin de garantir l'intégrité de la base de données fusionnée.

2.1.2 Fusion des fichiers

Les quatre fichiers (*Caractéristiques*, *Lieux*, *Véhicules*, *Usagers*) ont été fusionnés en une seule base de données en utilisant la variable commune `Num_Acc` comme clé. La fusion a permis d'obtenir un jeu de données final comportant 309337 accidents répertoriés, avec un total de 57 variables explicatives couvrant les caractéristiques générales des accidents, les informations sur les lieux, les véhicules impliqués, et les usagers.

2.1.3 Gestion des valeurs manquantes

La qualité des données a ensuite été évaluée en examinant le pourcentage de valeurs manquantes pour chaque variable. Les colonnes ayant plus de 50% de valeurs manquantes ont été considérées comme non informatives et ont été supprimées. Cela a conduit à la suppression de 3 colonnes, laissant un total de 54 variables explicatives dans la base de données finale.

Table 1: Résumé des variables explicatives importantes

Variable	**Type de donnée**	**Description**
`atm`	Catégorielle	Conditions atmosphériques (ex. : normales, pluie, brouillard, etc.)
`lum`	Catégorielle	Conditions de luminosité (ex. : plein jour, crépuscule, nuit, etc.)
`agg`	Binaire	Localisation en agglomération (1 : Oui, 2 : Non)
`col`	Catégorielle	Type de collision (ex. : frontale, par l'arrière, en chaîne, etc.)
`catr`	Catégorielle	Catégorie de la route (ex. : autoroute, route nationale, etc.)
`circ`	Catégorielle	Régime de circulation (ex. : sens unique, bidirectionnelle, etc.)
`prof`	Catégorielle	Profil de la route (ex. : plat, pente, sommet, etc.)
`surf`	Catégorielle	État de la surface (ex. : normale, mouillée, verglacée, etc.)
`catv`	Catégorielle	Catégorie du véhicule impliqué (ex. : vélo, VL, poids lourd, etc.)
`catu`	Catégorielle	Catégorie de l'usager (1 : Conducteur, 2 : Passager, 3 : Piéton)
`sexe`	Binaire	Sexe de l'usager (1 : Masculin, 2 : Féminin)
`trajet`	Catégorielle	Raison du déplacement (ex. : domicile-travail, loisirs, etc.)

2.2 Analyse bivariée

2.2.1 Etude de la sévérité et de l'heure

Pour mieux comprendre les relations entre les variables explicatives et la gravité des accidents, une analyse bivariée a été réalisée. En particulier, nous avons étudié la distribution de la gravité (`grav`) en fonction des plages horaires de la journée (matin, après-midi, soir).

Le graphique suivant (heatmap) représente la proportion des accidents pour chaque gravité et chaque plage horaire :

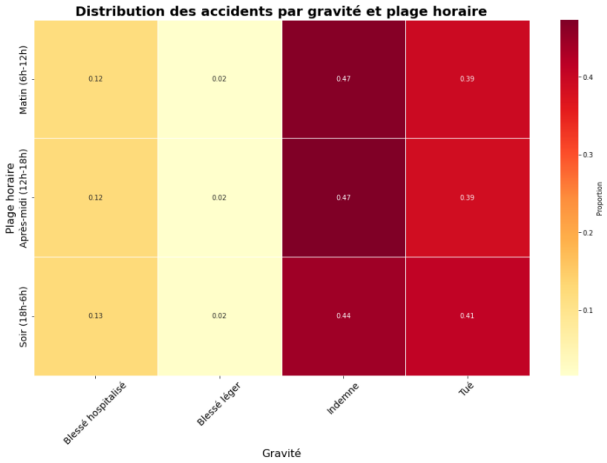

Figure 1: Distribution des accidents par gravité et plage horaire

Les proportions d'accidents où les usagers sont **indemnes** ou **tués** sont relativement similaires à travers les différentes plages horaires de la journée. Les accidents ayant conduit à un **blessé hospitalisé** ou un **blessé léger** montrent une proportion légèrement plus importante le soir (18h-6h). La distribution des accidents semble globalement homogène sur les trois plages horaires, ce qui suggère que l'heure de la journée a un impact limité sur la gravité des accidents.

2.2.2 Etude de la sévérité et des jours de la semaine

Après avoir analysé la répartition de la gravité des accidents selon les plages horaires, nous nous intéressons désormais aux jours de la semaine. Une hypothèse plausible est qu'il pourrait y avoir plus d'accidents graves les jeudis, en raison des sorties étudiantes, ou les vendredis, car c'est la fin de la semaine et c'une période où les déplacements sont plus fréquents.

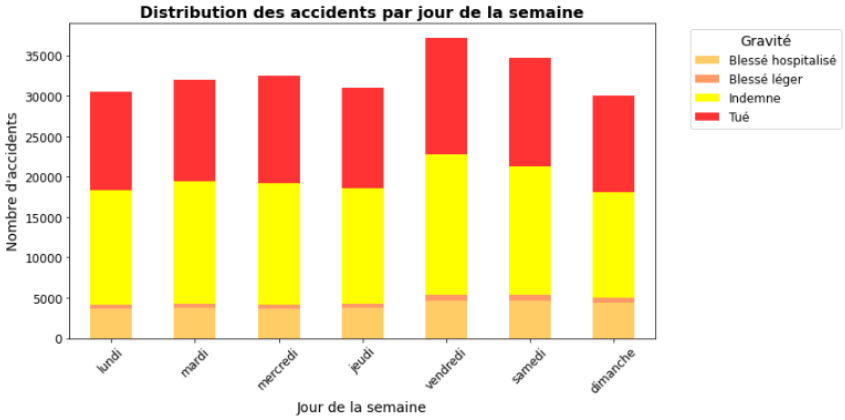

Figure 2: Répartition des accidents par gravité et jour de la semaine

En étudiant la distribution des accidents par gravité en fonction des jours de la semaine, nous observons que :

- La distribution est globalement uniforme sur l'ensemble des jours, sans différences significatives.
- Le vendredi montre une légère augmentation des accidents graves (notamment des **tués** et des **blessés hospitalisés**), ce qui pourrait s'expliquer par des comportements plus à risque en fin de semaine liés à la fatigue.

2.2.3 Répartition géographique des accidents

Pour visualiser la répartition géographique des accidents selon leur gravité, une carte a été réalisée représentant les différents niveaux de gravité des accidents (`indemne`, `blessé léger`, `blessé hospitalisé`, `tué`) à travers le territoire français.

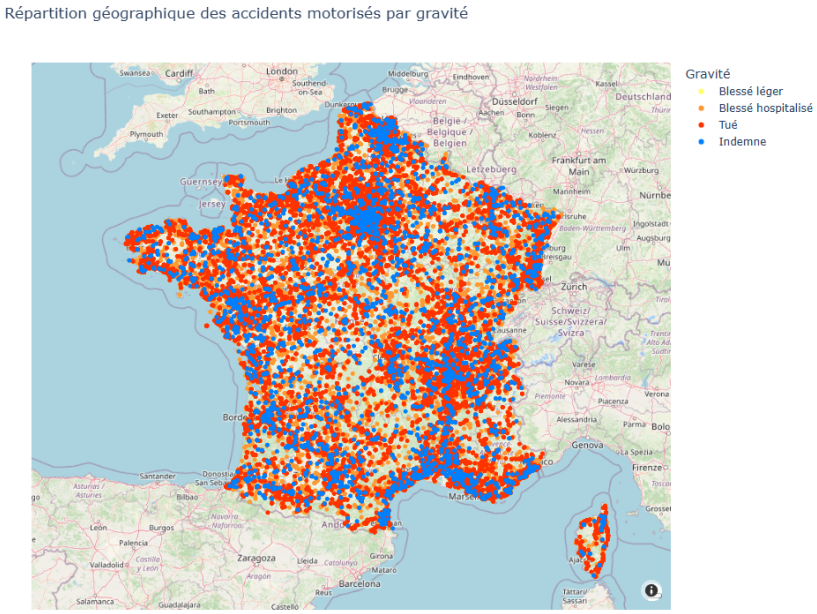

Figure 3: Répartition géographique des accidents motorisés par gravité en France

D'après la carte :

- Une forte proportion d'accidents ayant pour gravité **indemne** et **tué** est observée dans l'ensemble du territoire.
- La région parisienne (Île-de-France) montre une concentration particulièrement élevée d'accidents avec des usagers **indemnes**.
- Aucune tendance géographique claire n'apparaît pour les autres niveaux de gravité (**blessé léger** ou **blessé hospitalisé**), car ils semblent uniformément répartis.

Ces observations soulignent que la distribution géographique des accidents peut varier en fonction de la densité de population (notamment en Île-de-France), mais ne montre pas de corrélation évidente avec les niveaux de gravité en dehors de cette région.

3 Présentation des modèles utilisés

Dans le cadre de ce projet, nous avons testé trois modèles de machine learning : la régression logistique, la forêt aléatoire (*Random Forest*) et le *XGBoost*.
Il est important de noter que, pour la régression logistique uniquement, la gravité des accidents a été simplifiée en deux catégories :

- **Accidents légers** : regroupe les gravités 1 (*indemne*) et 2 (*blessé léger*).
- **Accidents très graves** : regroupe les gravités 3 (*blessé hospitalisé*) et 4 (*tué*).

Pour les modèles *Random Forest* et *XGBoost*, la classification a été réalisée directement sur les quatre classes initiales de gravité.

3.1 Régression logistique

La régression logistique est un modèle statistique souvent utilisé pour les problèmes de classification binaire. Dans ce projet, nous avons choisi de simplifier la gravité en deux classes (*léger* et *très grave*) afin de rendre ce modèle plus adapté à notre problématique.

Fonctionnement : La régression logistique modélise la probabilité $P(y = 1)$ qu'un accident appartienne à la classe *très grave* en utilisant une transformation sigmoïde appliquée à une combinaison linéaire des variables explicatives $\mathbf{x}$:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}}$$

où :

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ représente les variables explicatives.
- β_0 est l'interception ou biais du modèle.
- β_i sont les coefficients associés à chaque variable explicative x_i.

Interprétation : Les coefficients β_i permettent de quantifier l'effet des variables explicatives sur la probabilité d'un accident très grave. Par exemple, un coefficient $\beta_2 > 0$ indique qu'une augmentation de x_2 augmente la probabilité que $y = 1$ (accident très grave).

Avantages :

- Facile à interpréter, chaque coefficient ayant une signification directe.
- Adaptée pour les relations linéaires et les problèmes binaires.
- Simple à implémenter et rapide en calcul.

Limites :

- Incapacité à capturer des relations non linéaires entre les variables explicatives.
- Performances limitées dans des contextes où les classes sont déséquilibrées.

3.2 Forêt aléatoire (*Random Forest*)

Contrairement à la régression logistique, la forêt aléatoire est capable de gérer directement la classification en plusieurs classes (*indemne*, *blessé léger*, *blessé hospitalisé*, *tué*). Ce modèle repose sur une combinaison d'arbres de décision, chacun construit sur des sous-échantillons aléatoires des données.

Fonctionnement :

1. Plusieurs arbres de décision sont construits à partir de sous-échantillons aléatoires des données (tirés avec remise, selon la technique de *bagging*).
2. À chaque nœud, un sous-ensemble aléatoire des variables explicatives est sélectionné pour déterminer le critère de séparation optimal.
3. La prédiction finale est obtenue par un vote majoritaire pour la classification (ou une moyenne pour des problèmes de régression).

Interprétation : Bien que chaque arbre de décision soit interprétable individuellement, l'interprétation globale d'une forêt aléatoire est plus complexe. Cependant, des mesures comme l'importance des variables (*feature importance*) permettent de mieux comprendre les facteurs influençant les prédictions.

Avantages :

- Robuste face au bruit et au surapprentissage (*overfitting*).
- Performant pour des données avec des relations complexes et non linéaires.
- Capable de gérer des variables catégorielles et des valeurs manquantes.

Limites :

- Moins interprétable qu'un modèle linéaire comme la régression logistique.
- Gourmand en ressources computationnelles, surtout pour de grandes bases de données.

3.3 *XGBoost*

Le *XGBoost* (*eXtreme Gradient Boosting*) est une méthode de boosting avancée. Contrairement à la forêt aléatoire, où les arbres sont indépendants, le *XGBoost* construit des arbres séquentiellement, chaque nouvel arbre corrigeant les erreurs des arbres précédents.

Fonctionnement :

- À chaque étape, un arbre est ajouté pour minimiser une fonction de perte en corrigeant les erreurs des prédictions précédentes.
- La prédiction finale est donnée par une somme pondérée des arbres :

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}$$

- Une régularisation est incluse dans la fonction de perte pour éviter le surapprentissage :

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

  où ℓ représente la fonction de perte (par exemple, logarithmique pour la classification), et Ω est le terme de régularisation.

Avantages :

- Performant sur des ensembles de données complexes et déséquilibrés.
- Capable de capturer des interactions complexes entre variables.
- Rapide grâce à l'optimisation des calculs parallèles.

Limites :

- Complexe à interpréter.
- Nécessite un réglage minutieux des hyperparamètres pour atteindre son plein potentiel.

Optimisation des hyperparamètres:
Le fitting des hyperparamètres permet d'adapter le modèle XGBoost aux spécificités de nos données de gravité d'accidents. Les trois paramètres principaux jouent des rôles cruciaux :

- Le learning rate contrôle la contribution de chaque nouvel arbre au modèle final, évitant ainsi le surapprentissage tout en maintenant une bonne capacité prédictive
- Le max depth limite la complexité des arbres individuels, permettant un équilibre entre la capture des relations complexes et la généralisation
- Le nombre d'estimateurs (n estimators) détermine le nombre total d'arbres, influençant directement la capacité du modèle à capturer les motifs dans les données

Dans notre étude, cette optimisation était particulièrement cruciale car :

- La nature déséquilibrée de nos classes de gravité nécessite un paramétrage fin pour éviter les biais.
- La complexité des relations entre les variables explicatives et la gravité des accidents requiert une profondeur d'arbre adaptée.
- Le risque de surapprentissage est important étant donné la variabilité des situations d'accidents.

Sans cette optimisation, le modèle risquerait soit de sous-apprendre (ne pas capturer les relations importantes) soit de surapprendre (trop se spécialiser sur les données d'entraînement), compromettant ainsi sa capacité à généraliser sur de nouveaux cas d'accidents.

Nous avons donc procédé à une optimisation des hyperparamètres du XGBoost à travers une recherche sur grille (Grid Search) avec les paramètres suivants :

- learning rate : [0.1, 0.01, 0.05] pour contrôler la vitesse d'apprentissage
- max depth : valeurs de 2 à 10 pour définir la profondeur maximale des arbres
- n estimators : de 60 à 220 par pas de 40 pour le nombre d'arbres

Les trois modèles utilisés dans ce projet ont été sélectionnés pour leur complémentarité. La régression logistique, adaptée à la classification binaire, a permis une première analyse interprétable. La forêt aléatoire et le *XGBoost*, quant à eux, ont été exploités pour leur capacité à traiter les relations complexes et à gérer la classification multi-classes directement.

4 Résultats et performances des modèles

4.1 Résultats et analyses du modèle Random Forest

4.1.1 Analyse approfondie de la matrice de confusion

La matrice de confusion normalisée présentée ci-dessus constitue un outil essentiel pour évaluer la précision des prédictions du modèle pour chaque classe de gravité. Elle met en évidence les forces et les faiblesses du modèle dans sa capacité à classifier correctement les différents types d'accidents.

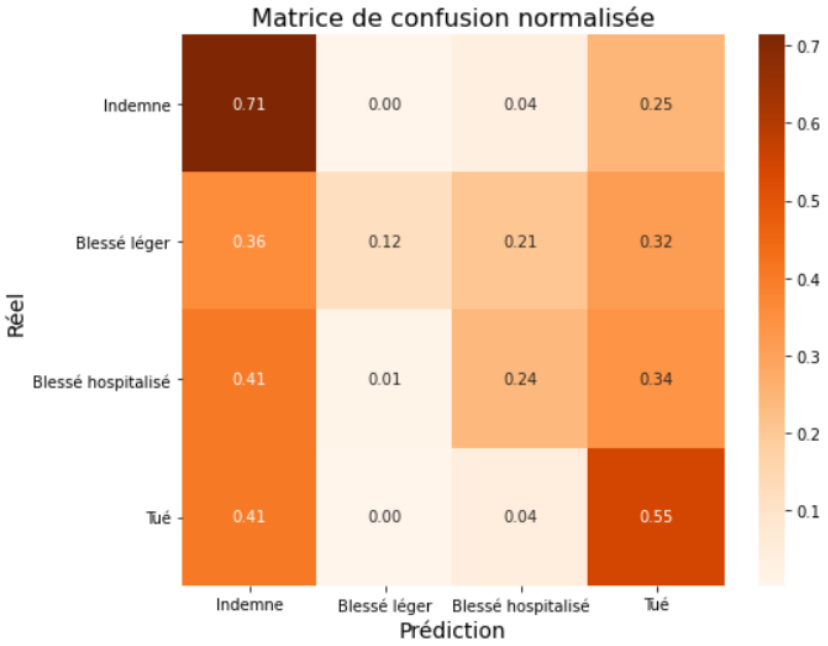

Figure 4: Matrice de confusion des prédictions d'accidents graves et non graves

La matrice de confusion met en évidence plusieurs phénomènes significatifs :

- Le taux élevé de prédiction correcte pour la classe "Indemne" (71%) suggère que les caractéristiques des accidents sans gravité sont bien captées par le modèle
- La confusion importante (41%) entre les classes "Tué" et "Indemne" indique une possible insuffisance de variables discriminantes pour les accidents mortels
- La confusion mutuelle entre blessés légers et hospitalisés (21-24%) révèle la difficulté du modèle à distinguer les niveaux de gravité intermédiaires, possiblement dû à la similarité des circonstances d'accidents

4.1.2 Interprétation des courbes ROC

Les courbes ROC (Receiver Operating Characteristic) illustrées ci-dessus offrent une visualisation de la performance du modèle pour chaque classe en termes de compromis entre le taux de vrais positifs et le taux de faux positifs. L'aire sous la courbe (AUC) fournit une mesure synthétique de la performance pour chaque classe.

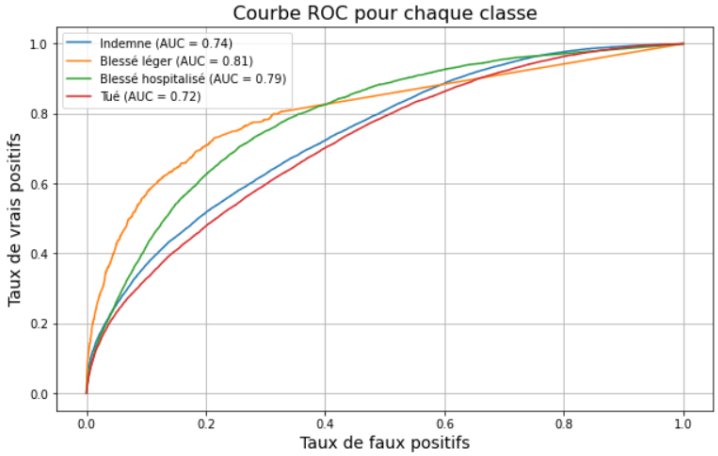

Figure 5: Courbe ROC : Évaluation de la performance du modèle

L'analyse des courbes ROC révèle des informations complémentaires :

- La pente plus raide de la courbe des blessés légers dans sa partie initiale indique une meilleure capacité à identifier les vrais positifs avec un faible taux de faux positifs
- Les courbes plus progressives pour les classes "Indemne" et "Tué" suggèrent un compromis plus délicat entre sensibilité et spécificité
- L'écart entre les courbes des différentes classes indique une hétérogénéité dans la capacité du modèle à gérer les différents types de gravité

4.1.3 Performance globale et interprétation

Les scores AUC obtenus, s'échelonnant de 0.72 à 0.8, révèlent une performance globalement satisfaisante du modèle. Cette variation des scores suggère que le modèle présente des capacités de discrimination différentes selon la gravité des accidents. La performance supérieure pour les blessés légers (AUC = 0.81) pourrait s'expliquer par des caractéristiques plus distinctives de ce type d'accidents dans les données d'entrée.

4.1.4 Implications pratiques

Ces résultats suggèrent plusieurs pistes d'amélioration :

- L'introduction de nouvelles variables explicatives spécifiques aux accidents mortels pourrait réduire la confusion avec la classe "Indemne"
- Un rééquilibrage des classes ou une stratification plus fine des données d'entraînement pourrait améliorer la discrimination entre les niveaux de blessures
- L'utilisation d'une approche d'ensemble plus sophistiquée pourrait mieux capturer les nuances entre les différents niveaux de gravité

4.2 Résultats et analyses de la régression logistique

4.2.1 Distribution des probabilités prédites

L'histogramme de distribution des probabilités prédites permet d'examiner la capacité du modèle à séparer les deux classes d'accidents (graves et non graves).

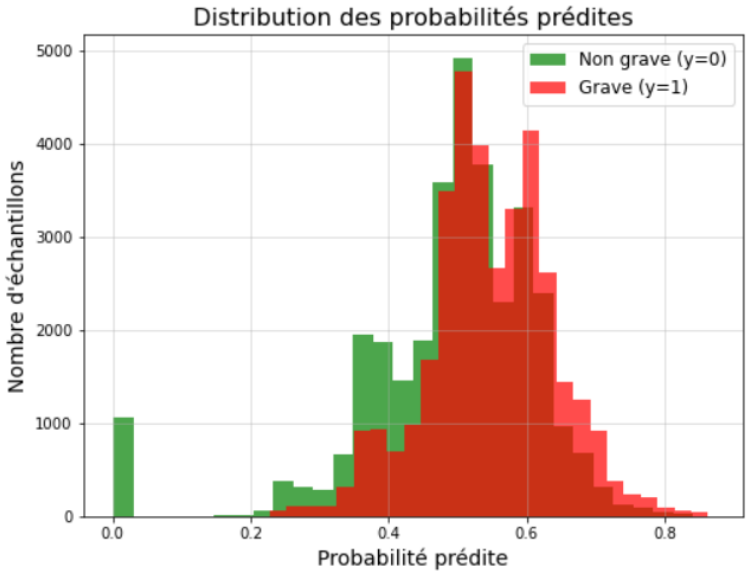

Figure 6: Distribution des probabilités prédites pour les accidents graves et non graves

L'analyse de la distribution révèle :

- Une concentration majoritaire des prédictions entre 0.4 et 0.6
- Un pic distinct pour les accidents non graves près de 0, indiquant une forte confiance du modèle pour certains cas
- Une distribution plus étalée pour les accidents graves entre 0.4 et 0.8
- Un chevauchement important des distributions, suggérant une difficulté à discriminer parfaitement les deux classes

4.2.2 Évaluation via la courbe ROC

La courbe ROC ci-dessous illustre la performance globale du modèle en termes de compromis entre sensibilité et spécificité.

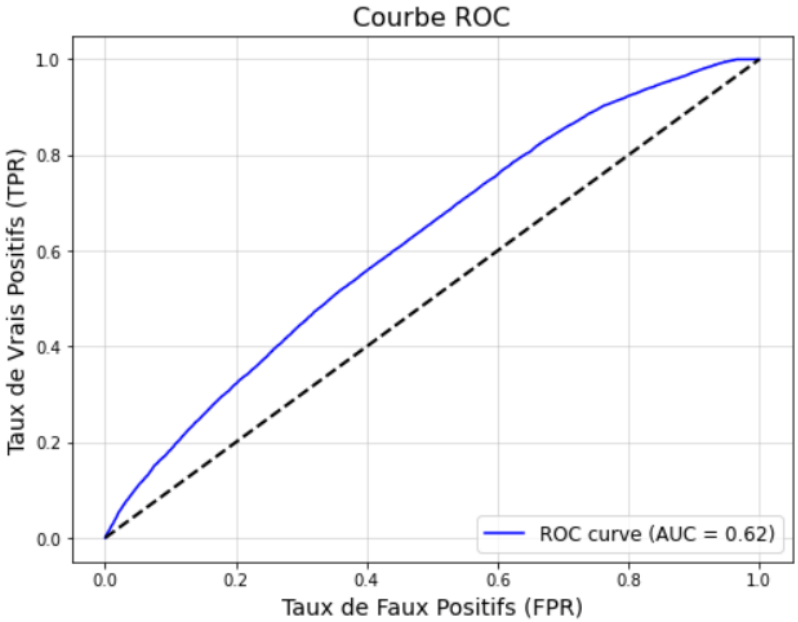

Figure 7: Courbe ROC : Évaluation de la performance du modèle

L'analyse de la courbe ROC indique :

- Un AUC de 0.62, témoignant d'une capacité de discrimination modérée
- Une courbe qui s'écarte de la diagonale de manière constante mais limitée
- Une performance supérieure à un classifieur aléatoire, mais avec une marge d'amélioration significativedeux classes

4.2.3 Analyse de la matrice de confusion

La matrice de confusion permet d'évaluer précisément la répartition des prédictions du modèle.

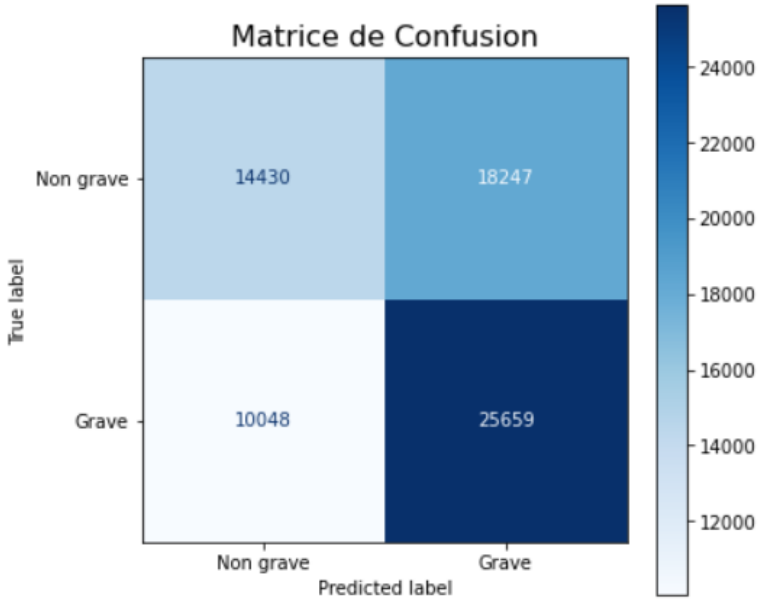

Figure 8: Matrice de confusion des prédictions d'accidents graves et non graves

L'examen détaillé révèle :

- 14,430 vrais négatifs et 25,659 vrais positifs
- Un nombre élevé de faux positifs (18,247) indiquant une tendance à la surclassification des cas graves
- 10,048 faux négatifs, représentant des cas graves non détectés
- Une précision globale limitée, suggérant la nécessité d'améliorations substantielles

4.2.4 Performance globale du modèle

L'évaluation globale du modèle de régression logistique révèle une performance modérée dans la classification des accidents. Le score AUC de 0.62 indique une capacité discriminative limitée mais supérieure à une classification aléatoire. La matrice de confusion montre que le modèle parvient à identifier correctement 25,659 cas graves et 14,430 cas non graves, mais présente également un nombre important de faux positifs (18,247) et de faux négatifs (10,048).

4.2.5 Implications et recommandations

Les résultats observés suggèrent plusieurs pistes d'amélioration :

- L'introduction de variables explicatives plus discriminantes
- L'utilisation de techniques de rééchantillonnage pour équilibrer les classes
- L'exploration d'autres algorithmes de classification plus sophistiqués
- L'ajustement du seuil de décision en fonction des coûts associés aux différents types d'erreurs

4.3 Résultats et analyses du modèle XGBoost

4.3.1 Performance globale du modèle

L'évaluation du modèle XGBoost démontre une performance variable selon les classes, avec des scores AUC s'échelonnant de 0.66 à 0.82. La classe "Blessé léger" présente la meilleure performance avec un AUC de 0.82, tandis que la classe "Tué" montre la performance la plus faible avec un AUC de 0.66.

4.3.2 Analyse de la matrice de confusion

La matrice de confusion normalisée révèle la distribution des prédictions du modèle pour chaque niveau de gravité.

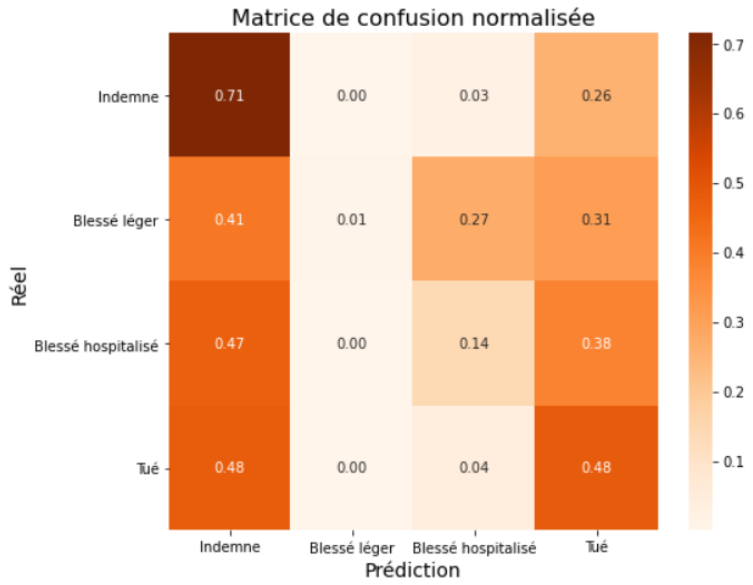

Figure 9: Matrice de confusion des prédictions d'accidents graves et non graves

L'analyse détaillée montre :

- Une prédiction correcte de 71% pour la classe "Indemne"
- Une tendance marquée à classer les accidents comme "Indemne" pour toutes les classes
- Une confusion importante pour les cas mortels, avec une répartition égale (48%) entre les prédictions "Indemne" et "Tué"
- Une difficulté à identifier correctement les blessés hospitalisés, avec 47% classés comme "Indemne"

4.3.3 Évaluation des courbes ROC

Les courbes ROC permettent d'évaluer la capacité discriminative du modèle pour chaque classe.

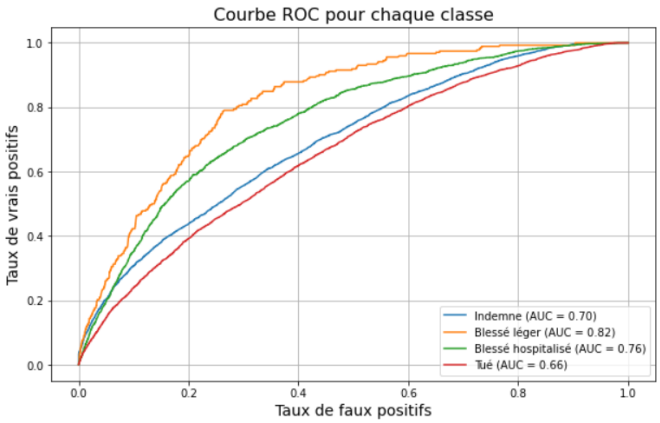

Figure 10: Courbe ROC : Évaluation de la performance du modèle

L'analyse des courbes révèle :

- Une excellente performance pour la classe "Blessé léger" (AUC = 0.82)
- Une bonne discrimination pour les blessés hospitalisés (AUC = 0.76)
- Une performance modérée pour la classe "Indemne" (AUC = 0.70)
- Une capacité limitée à identifier les cas mortels (AUC = 0.66)

4.3.4 Implications et limitations

Ces résultats suggèrent plusieurs axes d'amélioration :

- La nécessité de réduire le biais vers la classe "Indemne"
- L'importance d'améliorer la détection des cas mortels
- L'opportunité d'affiner la discrimination entre les différents niveaux de blessures
- Le besoin d'équilibrer les performances entre les classes

5 Conclusion comparative des modèles

L'analyse comparative des trois modèles (Random Forest, Régression Logistique et XGBoost) révèle des performances et caractéristiques distinctes.

5.1 Comparaison des performances

Le Random Forest présente les meilleures performances globales avec des AUC entre 0.72 et 0.81, tandis que le XGBoost montre des performances similaires mais légèrement inférieures (AUC de 0.66 à 0.82). La Régression Logistique affiche les performances les plus modestes avec un AUC global de 0.62.

5.2 Forces et faiblesses spécifiques

- Le Random Forest excelle dans l'équilibre entre les classes et la stabilité des prédictions
- Le XGBoost montre une excellente capacité pour la détection des blessés légers mais peine sur les cas mortels
- La Régression Logistique, bien que plus simple, manque de puissance discriminative

5.3 Recommandation

Le Random Forest s'impose comme le modèle le plus adapté pour cette classification de gravité d'accidents, principalement grâce à sa stabilité des performances à travers toutes les classes et sa meilleure gestion de la confusion entre classes extrêmes. Néanmoins, une approche hybride combinant les prédictions du Random Forest et du XGBoost pourrait être envisagée pour optimiser la détection des accidents mortels, particulièrement dans les cas critiques nécessitant une haute précision.