Analyse et Prédiction des Accidents de Vélo en Île-de-France

De la Classification du Risque à la Modélisation des Taux d'Accidentalité

Nicolas Huyghe David Chhoa Jérémie Masnou

Projet Science des Données Appliquée
Janvier 2026

Résumé

Cette étude analyse les accidents de vélo en Île-de-France en utilisant des techniques de machine learning. Notre approche a évolué au cours de l'étude : une première analyse basée sur le nombre brut d'accidents a révélé des limites méthodologiques importantes, nous conduisant à intégrer des données de population pour calculer des taux de risque normalisés. Nous présentons les résultats des deux approches : (1) la classification binaire du risque communal atteint un ROC-AUC de 95.6%, (2) la prédiction du nombre brut d'accidents affiche un R^2 de 89.2% mais souffre de biais liés à la taille des communes, et (3) la prédiction des taux de risque normalisés, plus robuste méthodologiquement, atteint un R^2 de 30% pour le taux par kilomètre d'aménagement. Ces résultats soulignent la complexité de modéliser le risque cycliste et l'importance du choix de la variable cible.

Mots-clés : accidents vélo, machine learning, classification, régression, Île-de-France, taux de risque

1 Introduction

1.1 Contexte

La pratique du vélo connaît un essor considérable en Île-de-France ces dernières années. Les politiques de mobilité durable, le développement des pistes cyclables et la prise de conscience environnementale ont conduit à une augmentation significative du nombre de cyclistes. Cependant, cette croissance du trafic cycliste s'accompagne d'une problématique de sécurité routière qu'il est essentiel d'analyser et de comprendre.

Les accidents de vélo représentent un enjeu majeur de santé publique. En Île-de-France, des milliers d'accidents impliquant des cyclistes sont enregistrés chaque année, allant de simples chutes à des accidents mortels. Comprendre les facteurs qui influencent ces accidents et pouvoir les prédire constitue un défi important pour les décideurs publics souhaitant améliorer la sécurité des cyclistes.

1.2 Objectifs du projet

Ce projet vise à répondre à deux questions fondamentales :

1. **Classification du risque** : Peut-on identifier les communes présentant un risque élevé d'accidents de vélo à partir des caractéristiques de leurs aménagements cyclables ?
2. **Prédiction du risque** : Peut-on prédire le niveau de risque d'accidents dans une commune en fonction de ses infrastructures cyclables et de sa population ?

1.3 Évolution de notre approche

Notre méthodologie a significativement évolué au cours de l'étude. Initialement, nous avons tenté de prédire le **nombre brut d'accidents** par commune. Cette première approche a produit des résultats apparemment excellents (R^2 de 89%), mais une analyse critique approfondie a révélé des limites méthodologiques importantes :

- Une corrélation paradoxale : les communes avec plus d'aménagements cyclables présentaient plus d'accidents
- Un effet de taille : les grandes communes (notamment Paris) dominaient complètement les résultats
- Une erreur relative élevée malgré de bons indicateurs absolus

Face à ces constats, nous avons décidé d'intégrer un nouveau jeu de données sur la **population municipale** pour calculer des **taux de risque normalisés** (accidents par habitant, accidents par km d'aménagement). Cette évolution illustre l'importance de l'analyse critique en science des données.

2 Données et Méthodologie

2.1 Sources de données

Notre étude s'appuie sur quatre jeux de données complémentaires, permettant une analyse multidimensionnelle du risque cycliste en Île-de-France.

2.1.1 Accidents de vélo

Ce dataset national contient **80 022 accidents** impliquant des cyclistes sur l'ensemble de la France, dont **22 609 en Île-de-France**. Pour chaque accident, nous disposons de :

- La localisation précise (département, commune, coordonnées GPS)
- La date et l'heure de l'accident
- La gravité (indemne, blessé léger, hospitalisé, décédé)
- Les conditions (luminosité, météo, état de la chaussée)
- Les caractéristiques de la victime (âge, sexe)

2.1.2 Aménagements cyclables

Ce dataset recense **143 060 infrastructures cyclables** en Île-de-France, incluant :

- Le type de voie (piste cyclable séparée, bande cyclable, voie partagée)
- La longueur de chaque aménagement en mètres
- Le type de revêtement (asphalte, pavés, etc.)
- La localisation par code INSEE de la commune

2.1.3 Comptages vélo

Ce dataset contient **933 757 mesures** de comptage horaire provenant de 69 compteurs automatiques répartis en Île-de-France. Ces données permettent d'estimer le trafic cycliste, bien que la couverture soit limitée.

2.1.4 Population municipale (ajouté en phase 3)

Face aux limites de notre première approche, nous avons intégré les données de **population INSEE 2021** couvrant **1 287 communes** franciliennes. Ce dataset permet de calculer des taux de risque normalisés par habitant.

2.2 Fusion et création de variables

Les quatre sources de données ont été fusionnées au niveau communal en utilisant le **code INSEE** comme clé de jointure. Ce processus d'agrégation a produit un dataset final de **1 124 communes** caractérisées par **44 variables**.

2.2.1 Variables principales créées

- `nb_accidents` : nombre total d'accidents par commune (variable cible initiale)
- `nb_accidents_graves` : accidents ayant entraîné hospitalisation ou décès
- `taux_accidents_graves` : proportion d'accidents graves sur le total
- `nb_amenagements` : nombre d'infrastructures cyclables dans la commune
- `longueur_totale_amenagements` : longueur cumulée en mètres
- `ratio_pistes_cyclables` : proportion de pistes séparées de la circulation
- `population` : nombre d'habitants (INSEE 2021)
- `taux_risque_par_km` : nombre d'accidents divisé par les km d'aménagement
- `taux_risque_par_habitant` : accidents pour 10 000 habitants
- `risque_eleve` : variable binaire (1 si le nombre d'accidents $\geq 75^{e}$ percentile)

2.3 Évolution méthodologique

Notre approche a suivi trois phases distinctes, reflétant un processus itératif d'analyse et d'amélioration.

2.3.1 Phase 1 : Approche initiale

Nous avons d'abord développé deux modèles :

- Un modèle de **classification binaire** pour identifier les communes à risque élevé
- Un modèle de **régression** pour prédire le nombre brut d'accidents

Les résultats semblaient excellents : un ROC-AUC de 95.6% pour la classification et un R^2 de 89.2% pour la régression.

2.3.2 Phase 2 : Analyse critique

Une analyse approfondie des résultats a révélé plusieurs problèmes méthodologiques :

- **Corrélation paradoxale** : nous avons observé que les communes avec plus d'aménagements cyclables avaient plus d'accidents ($r = 0.60$). Cela ne signifie pas que les aménagements sont dangereux, mais que les communes bien équipées attirent plus de cyclistes.
- **Effet de taille** : Paris, avec ses 20 arrondissements, concentre 61% des accidents de la région. Le modèle "apprenait" essentiellement à identifier Paris.
- **MAPE élevé** : malgré un R^2 de 89%, l'erreur moyenne absolue en pourcentage (MAPE) atteignait 83%, indiquant une mauvaise prédiction pour les petites communes.
- **Distribution asymétrique** : 33.6% des communes n'avaient aucun accident enregistré.

2.3.3 Phase 3 : Approche corrigée

Pour remédier à ces biais, nous avons :

- Intégré les données de **population municipale**
- Calculé des **taux de risque normalisés** (par habitant et par km)
- Appliqué des **transformations logarithmiques** pour réduire l'asymétrie

3 Analyse 1 : Classification du Risque

3.1 Définition du problème

La première question de recherche consiste à classifier les communes selon leur niveau de risque d'accidents de vélo. Nous avons défini une commune comme étant à **risque élevé** si son nombre d'accidents est supérieur ou égal au 75^{e} percentile de la distribution, soit **6 accidents ou plus**.

Cette définition binaire permet de transformer le problème en une tâche de classification supervisée, où l'objectif est de prédire si une commune appartient à la catégorie "risque élevé" ou "risque faible" en fonction de ses caractéristiques.

3.2 Modèles testés

Six algorithmes de classification ont été comparés :
— **Régression Logistique** : modèle linéaire avec pondération des classes
— **Random Forest** : ensemble d'arbres de décision
— **SVM** : machine à vecteurs de support avec noyau RBF
— **Gradient Boosting** : boosting d'arbres de décision
— **XGBoost** : implémentation optimisée du gradient boosting
— **LightGBM** : gradient boosting basé sur l'histogramme

3.3 Résultats

TABLE 1 – Performance des modèles de classification

Modèle	Accuracy	F1-Score	ROC-AUC
Régression Logistique	0.884	**0.794**	**0.956**
Random Forest	0.898	0.793	0.951
SVM (RBF)	0.880	0.791	0.953
LightGBM	0.889	0.786	0.940
XGBoost	0.884	0.780	0.941
Gradient Boosting	0.880	0.757	0.953

3.4 Analyse des résultats

La **Régression Logistique** obtient les meilleures performances globales avec :
— Un **F1-Score de 79.4%**, équilibrant bien précision et rappel
— Un **ROC-AUC de 95.6%**, indiquant une excellente capacité à distinguer les deux classes
— Une bonne stabilité en validation croisée (écart-type de 3.3%)

Ce résultat peut sembler surprenant : un modèle simple (régression logistique) surpasse des modèles plus complexes (boosting, forêts aléatoires). Cela s'explique par la nature du problème : la relation entre les features et le risque est essentiellement linéaire. Les communes à risque élevé sont celles avec beaucoup d'aménagements cyclables, car ces communes attirent plus de cyclistes.

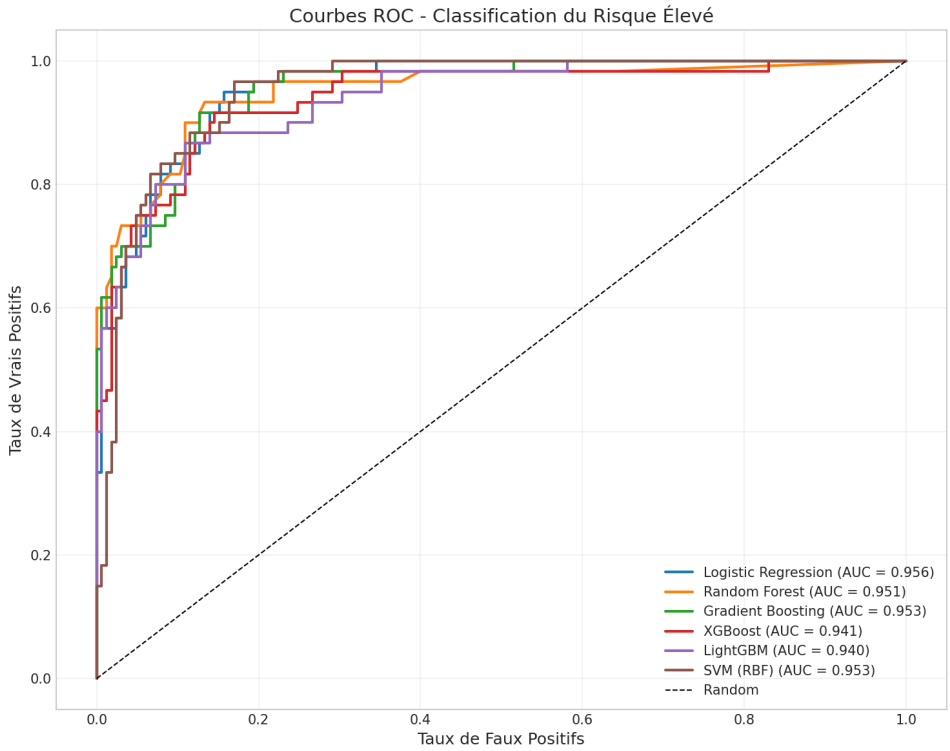

FIGURE 1 – Courbes ROC des modèles de classification. Tous les modèles atteignent un AUC supérieur à 0.94, indiquant une bonne séparation des classes.

4 Analyse 2 : Prédiction du Nombre Brut d'Accidents

4.1 Approche initiale

Notre deuxième objectif était de prédire le **nombre exact d'accidents** par commune, ce qui constitue un problème de régression. Sept algorithmes ont été comparés, allant de modèles linéaires simples à des méthodes d'ensemble avancées.

TABLE 2 – Performance des modèles de régression (nombre d'accidents)

Modèle	RMSE	R^2	CV R^2
Gradient Boosting	**37.06**	**0.892**	0.897 ± 0.025
XGBoost	37.53	0.889	0.900 ± 0.018
Random Forest	37.80	0.888	0.886 ± 0.027
Régression Linéaire	42.89	0.856	0.855 ± 0.072
Ridge	43.23	0.853	0.857 ± 0.067
Lasso	43.87	0.849	0.856 ± 0.065
LightGBM	45.41	0.838	0.825 ± 0.074

Le modèle **Gradient Boosting** obtient les meilleures performances avec un R^2 de 89.2%, ce qui signifie que le modèle explique près de 90% de la variance du nombre d'accidents. Ce résultat semblait initialement très satisfaisant.

4.2 Analyse critique : remise en question des résultats

Cependant, une analyse approfondie des résultats a révélé plusieurs problèmes méthodologiques qui remettent en question la validité de ces performances apparemment excellentes.

4.2.1 Problème 1 : Corrélation paradoxale

Nous avons observé une forte corrélation positive ($r = 0.60$) entre le nombre d'aménagements cyclables et le nombre d'accidents. Autrement dit, **plus une commune a d'aménagements, plus elle a d'accidents**.

Cette observation paradoxale ne signifie pas que les aménagements sont dangereux. Elle s'explique par un biais de confusion : les communes bien équipées en infrastructures cyclables attirent plus de cyclistes, ce qui augmente mécaniquement le nombre d'accidents en valeur absolue.

4.2.2 Problème 2 : Effet de taille (Paris)

Paris, avec ses 20 arrondissements traités comme 20 communes distinctes, concentre **61% des accidents** de toute l'Île-de-France (13 853 sur 22 609). Le modèle "apprend" donc essentiellement à identifier Paris, ce qui gonfle artificiellement le R^2.

4.2.3 Problème 3 : MAPE élevé

Malgré un R^2 de 89%, le **MAPE (Mean Absolute Percentage Error) atteint 83%**. Cette métrique, qui mesure l'erreur relative moyenne, révèle que les prédictions sont très imprécises pour les petites communes (qui sont majoritaires dans le dataset).

4.2.4 Problème 4 : Distribution asymétrique

La distribution du nombre d'accidents présente une forte asymétrie (skewness = 6.21) : 33.6% des communes n'ont aucun accident enregistré, tandis que quelques communes (Paris) en ont plus de 1 000.

4.3 Conclusion intermédiaire

Ces constats nous ont conduits à **remettre en question notre approche initiale**. Un R^2 élevé ne garantit pas la pertinence d'un modèle. Nous avons donc décidé de normaliser les données en intégrant la population et en calculant des taux de risque.

5 Analyse 3 : Prédiction des Taux de Risque

5.1 Intégration des données de population

Pour remédier aux biais identifiés dans l'approche précédente, nous avons intégré un nouveau jeu de données : la **population municipale** de chaque commune (données INSEE 2021). Cela nous permet de calculer des **taux de risque normalisés** qui éliminent l'effet de taille des communes.

5.2 Définition des taux de risque

Nous avons défini deux métriques de risque :

Taux de risque par kilomètre d'aménagement :

$$\text{Taux}_{km} = \frac{\text{Nombre d'accidents}}{\text{Longueur des aménagements (km)}} \tag{1}$$

Ce taux mesure le nombre d'accidents par kilomètre d'infrastructure cyclable. Il permet de comparer des communes indépendamment de leur niveau d'équipement.

Taux de risque pour 10 000 habitants :

$$\text{Taux}_{hab} = \frac{\text{Nombre d'accidents}}{\text{Population}} \times 10000 \tag{2}$$

Ce taux, classique en épidémiologie, normalise par la population et permet de comparer le risque entre communes de tailles différentes.

5.3 Analyse des distributions

TABLE 3 – Caractéristiques statistiques des taux de risque

Métrique	**Taux par km**	**Taux pour 10k hab**
Moyenne	1.96	9.20
Médiane	0.23	4.60
Écart-type	31.07	19.37
Skewness	25.75	6.79
Communes sans accident	33.6%	33.6%

Les deux taux présentent une forte asymétrie (skewness élevé), ce qui indique que la plupart des communes ont un faible taux de risque, tandis que quelques communes ont des taux très élevés. Pour réduire cette asymétrie, nous avons appliqué une **transformation logarithmique** : $y' = \log(1 + y)$.

5.4 Résultats de la modélisation

Sept modèles de machine learning ont été testés pour prédire chaque taux de risque.

TABLE 4 – Prédiction du taux de risque par km

Modèle	**R^2**	**MAPE**
Ridge	**0.299**	70.4%
ElasticNet	0.193	92.7%
Lasso	0.150	95.4%
LightGBM	0.148	67.5%
Random Forest	0.100	62.0%

TABLE 5 – Prédiction du taux de risque pour 10 000 habitants

Modèle	**R^2**	**MAPE**
Random Forest	**0.186**	33.6%
ElasticNet	0.124	34.6%
LightGBM	0.122	37.0%
Ridge	0.121	35.6%
Lasso	0.118	34.2%

5.5 Interprétation des résultats

Les performances obtenues avec les taux de risque sont **nettement inférieures** à celles de l'approche par nombre brut (R^2 maximum de 30% contre 89%). Loin d'être un échec, ce résultat est en réalité plus honnête et plus informatif.

5.5.1 Pourquoi les performances sont-elles plus faibles ?

Élimination du biais de taille : En normalisant par la population ou la longueur d'aménagement, le modèle ne peut plus "tricher" en identifiant simplement les grandes communes. Paris, qui dominait les données brutes, n'a plus un poids disproportionné.

Complexité réelle du risque : Le risque d'accident cycliste dépend de nombreux facteurs que nous n'avons pas dans nos données : comportement des usagers, densité du trafic automobile, conditions météorologiques, qualité de l'éclairage, etc.

Données insuffisantes : Nous disposons de seulement 17 variables explicatives, ce qui est insuffisant pour capturer toute la complexité du phénomène.

5.5.2 Importance des features

Pour le taux de risque par habitant, l'analyse de l'importance des features (Random Forest) révèle :

1. **Population** (17.1%) : les communes plus peuplées ont tendance à avoir un taux de risque différent
2. **Longueur des aménagements** (11.8%) : l'étendue du réseau cyclable influence le risque
3. **Densité population/aménagement** (10.4%) : le ratio entre population et infrastructure

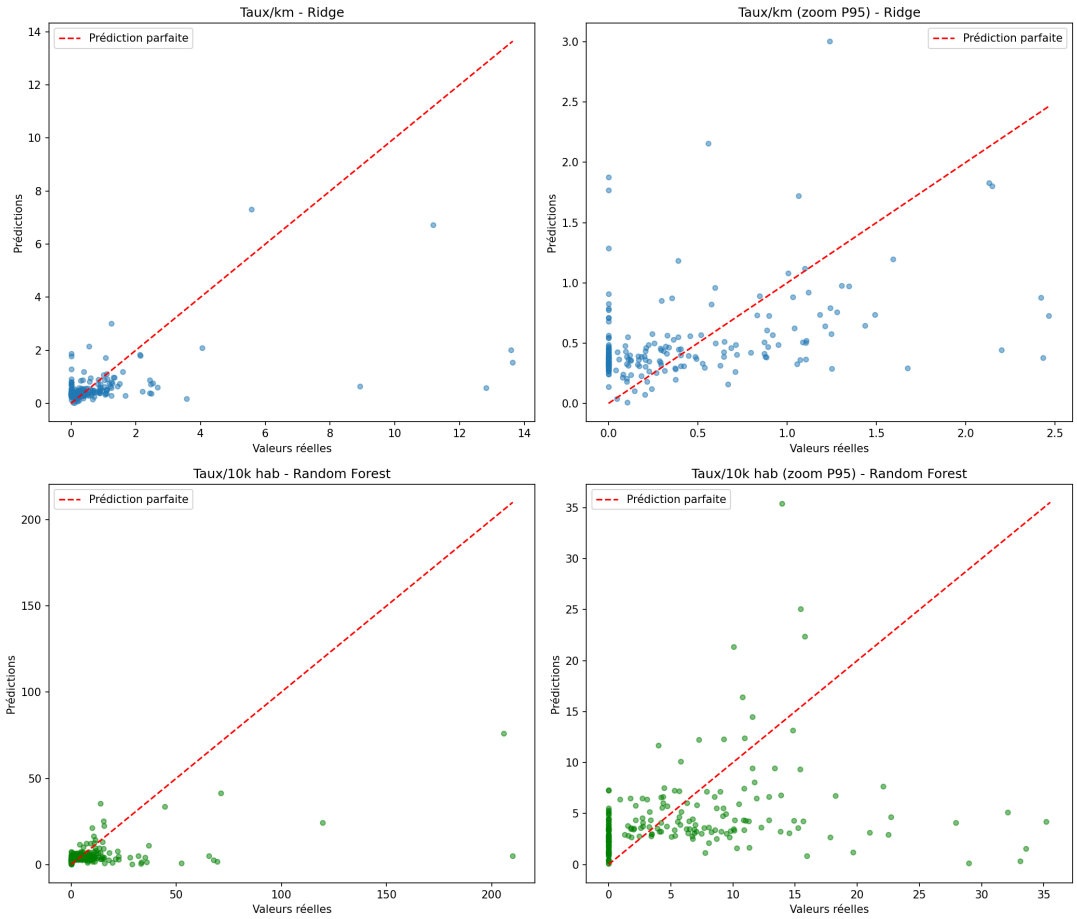

FIGURE 2 – Prédictions vs valeurs réelles pour les deux taux de risque. Les graphiques montrent une dispersion importante autour de la diagonale, confirmant que les modèles n'expliquent qu'une partie de la variance.

6 Discussion

6.1 Synthèse et comparaison des approches

Le tableau suivant résume les performances et la validité de chaque approche :

TABLE 6 – Synthèse comparative des trois analyses

Analyse	**Meilleur modèle**	**Performance**	**Validité**
Classification binaire	Rég. Logistique	AUC 96%	Élevée
Régression (nb brut)	Gradient Boosting	R^2 89%	Limitée
Taux par km	Ridge	R^2 30%	Élevée
Taux par habitant	Random Forest	R^2 19%	Élevée

Le **paradoxe apparent** (meilleure performance avec l'approche la moins valide) illustre un principe fondamental en science des données : **un R^2 élevé ne garantit pas la pertinence d'un modèle**. La régression sur le nombre brut d'accidents "triche" en exploitant l'effet de taille des communes, ce qui produit un score artificiellement élevé.

6.2 Enseignements méthodologiques

Cette étude met en évidence plusieurs bonnes pratiques :

1. **Analyser les corrélations** : Une corrélation paradoxale (aménagements → accidents) doit alerter sur un possible biais de confusion.
2. **Examiner la distribution** : Une variable cible très asymétrique nécessite une transformation ou une approche adaptée.
3. **Calculer plusieurs métriques** : Le R^2 seul peut être trompeur ; le MAPE révèle l'erreur relative.
4. **Normaliser les données** : Utiliser des taux plutôt que des valeurs absolues élimine les biais de taille.

6.3 Limites de l'étude

Plusieurs facteurs limitent la portée de nos conclusions :

— **Sous-déclaration des accidents** : Tous les accidents ne sont pas signalés aux autorités, en particulier les accidents mineurs.
— **Couverture des comptages** : Seulement 69 compteurs automatiques pour toute l'Île-de-France, ce qui limite l'estimation du trafic cycliste.
— **Variables manquantes** : Nous n'avons pas accès au trafic automobile, aux conditions météorologiques détaillées, ni aux caractéristiques urbanistiques fines.
— **Proportion de zéros** : 33.6% des communes n'ont aucun accident déclaré, ce qui pourrait nécessiter des modèles spécifiques (Zero-Inflated).

6.4 Perspectives d'amélioration

Plusieurs pistes permettraient d'améliorer les modèles :

— **Modèles Zero-Inflated** : Pour gérer explicitement les communes sans accident.
— **Données complémentaires** : Intégrer le trafic automobile, la météo, et l'urbanisme.
— **Analyse spatiale** : Prendre en compte l'autocorrélation géographique entre communes voisines.
— **Modèles séparés** : Développer un modèle spécifique pour Paris et un autre pour le reste de l'Île-de-France.

7 Conclusion

Cette étude a permis d'explorer différentes approches pour modéliser le risque d'accidents de vélo en Île-de-France, en mettant en évidence l'importance de l'analyse critique en science des données.

7.1 Principaux résultats

Classification du risque : Notre modèle de classification binaire (risque élevé vs faible) atteint un ROC-AUC de 95.6% avec une simple régression logistique. Ce modèle peut être utilisé pour identifier les communes prioritaires pour des interventions de sécurité.

Prédiction du nombre brut d'accidents : Bien que présentant un R^2 impressionnant de 89%, cette approche souffre de biais méthodologiques importants (effet de taille, corrélation paradoxale) qui limitent son interprétabilité et sa généralisabilité.

Prédiction des taux de risque : L'intégration des données de population et le calcul de taux normalisés offrent une approche méthodologiquement plus rigoureuse. Les performances plus modestes (R^2 de 20-30%) reflètent la complexité réelle du phénomène et l'insuffisance des données disponibles.

7.2 Enseignement principal

Le principal enseignement de cette étude est que **la performance brute d'un modèle ne suffit pas à évaluer sa validité**. Un R^2 élevé peut masquer des biais importants, et une analyse critique des données et des résultats est essentielle pour éviter des conclusions erronées.

7.3 Recommandations pratiques

Pour les décideurs publics et les aménageurs :

— Utiliser la **classification** pour identifier les communes prioritaires
— Privilégier les **taux de risque** pour comparer les communes entre elles
— **Ne pas interpréter** la corrélation positive entre aménagements et accidents comme un signe de dangerosité des infrastructures
— Enrichir les données avec des variables de contexte (trafic, urbanisme) pour améliorer les modèles