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

Nicolas Huyghe David Chhoa Jérémie Masnou

Projet Sciences des Données et Apprentissage 2025/2026

Résumé

Ce projet fusionne trois jeux de données ouverts (accidents vélo, aménagements cyclables, comptages du nombre de vélos) pour analyser la sécurité routière des cyclistes en Île-de-France. Nous avons abordé deux problématiques : Peut-on identifier les communes présentant un risque élevé d'accidents de vélo ? Peut-on prédire le nombre brut d'accidents par commune ? L'analyse des résultats de classification s'est révélée non concluante. Nous avons donc enrichi notre approche en intégrant un quatrième jeu de données (population INSEE) afin de prédire des taux de risque normalisés, définis de deux manières : le taux d'accidents par kilomètre d'aménagement cyclable et le taux d'accidents pour 10 000 habitants.

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

1 Introduction

1.1 Contexte

La pratique du vélo connaît un essor considérable en Île-de-France. Cependant, cette croissance s'accompagne d'une problématique de sécurité routière : des milliers d'accidents impliquant des cyclistes sont enregistrés chaque année dans la région.

1.2 Objectifs du projet

Ce projet vise à répondre à trois questions :

1. **Classification du risque** : Peut-on identifier les communes présentant un risque élevé d'accidents de vélo ?
2. **Prédiction du nombre d'accidents** : Peut-on prédire le nombre brut d'accidents par commune ?
3. **Prédiction des taux de risque** : Peut-on prédire des taux de risque normalisés (par habitant, par km d'aménagement) ?

2 Données et Méthodologie

2.1 Sources de données

Notre étude s'appuie sur quatre jeux de données, plus précisément 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

Pour normaliser nos analyses, 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**.

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)

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.

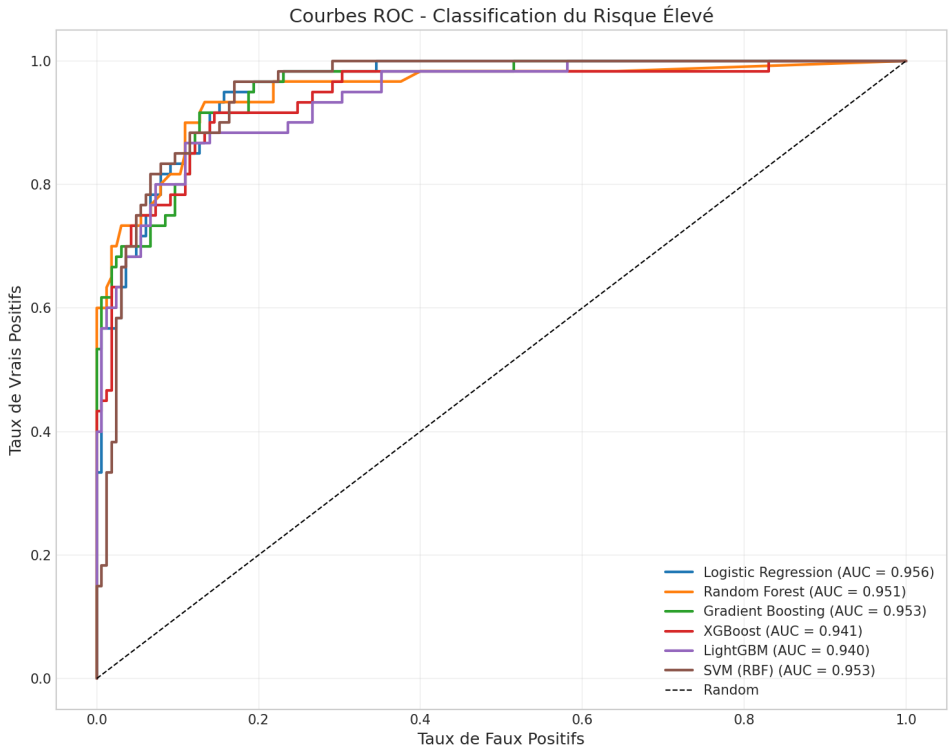

FIGURE 1 – Courbes ROC des modèles de classification. Tous les modèles atteignent un AUC supérieur à 0.94, indiquant une bonne séparation des classes.

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. On a comparé sept algorithmes.

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 des résultats

L'analyse des résultats a révélé plusieurs problèmes méthodologiques qui remettent en question la validité de ces performances apparemment excellentes.

Corrélation paradoxale : Nous observons une forte corrélation positive ($r = 0.60$) entre le nombre d'aménagements cyclables et le nombre d'accidents. Cette observation ne signifie pas que les aménagements sont dangereux, mais s'explique par un biais de confusion : les communes bien équipées attirent plus de cyclistes, ce qui augmente mécaniquement le nombre d'accidents en valeur absolue.

Effet de taille (Paris) : Paris, avec ses 20 arrondissements traités comme 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.

MAPE élevé : Malgré un R^2 de 89%, le MAPE (Mean Absolute Percentage Error) atteint 83%. Cette métrique révèle que les prédictions sont imprécises pour les petites communes, majoritaires dans le dataset.

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.

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.

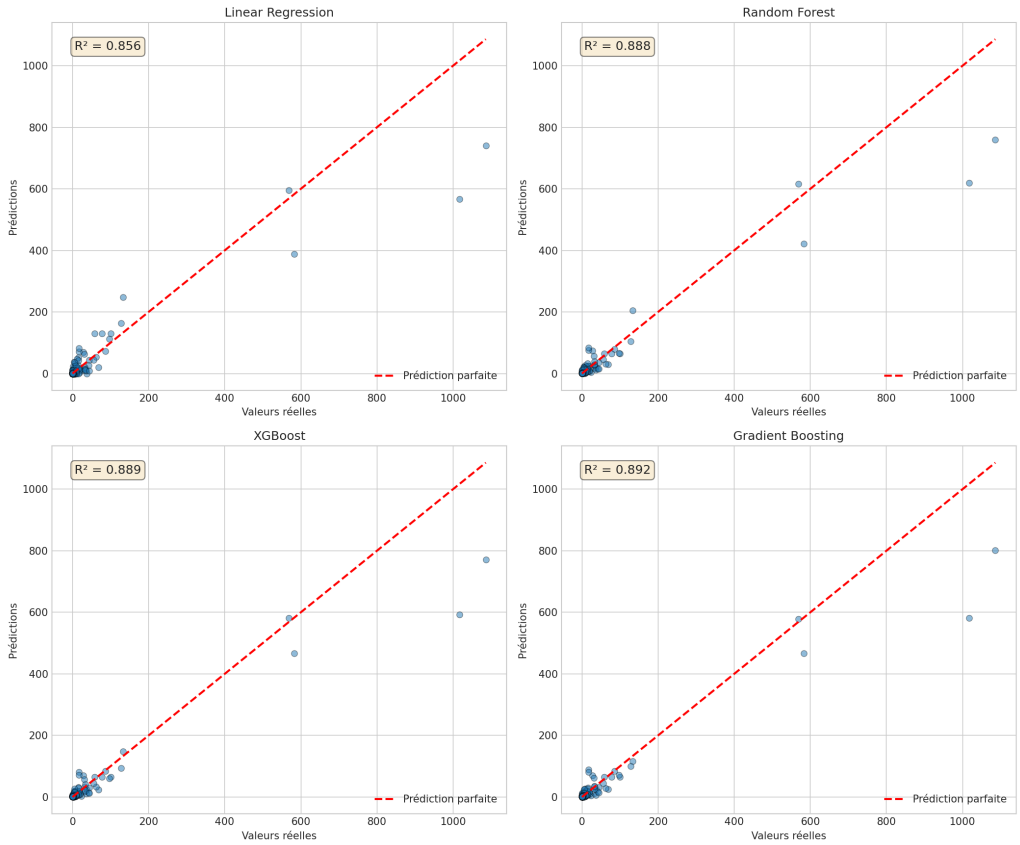

FIGURE 2 – Prédictions vs valeurs réelles du nombre d'accidents.

Le modèle prédit bien les valeurs élevées (Paris), mais sous-estime systématiquement les communes intermédiaires.

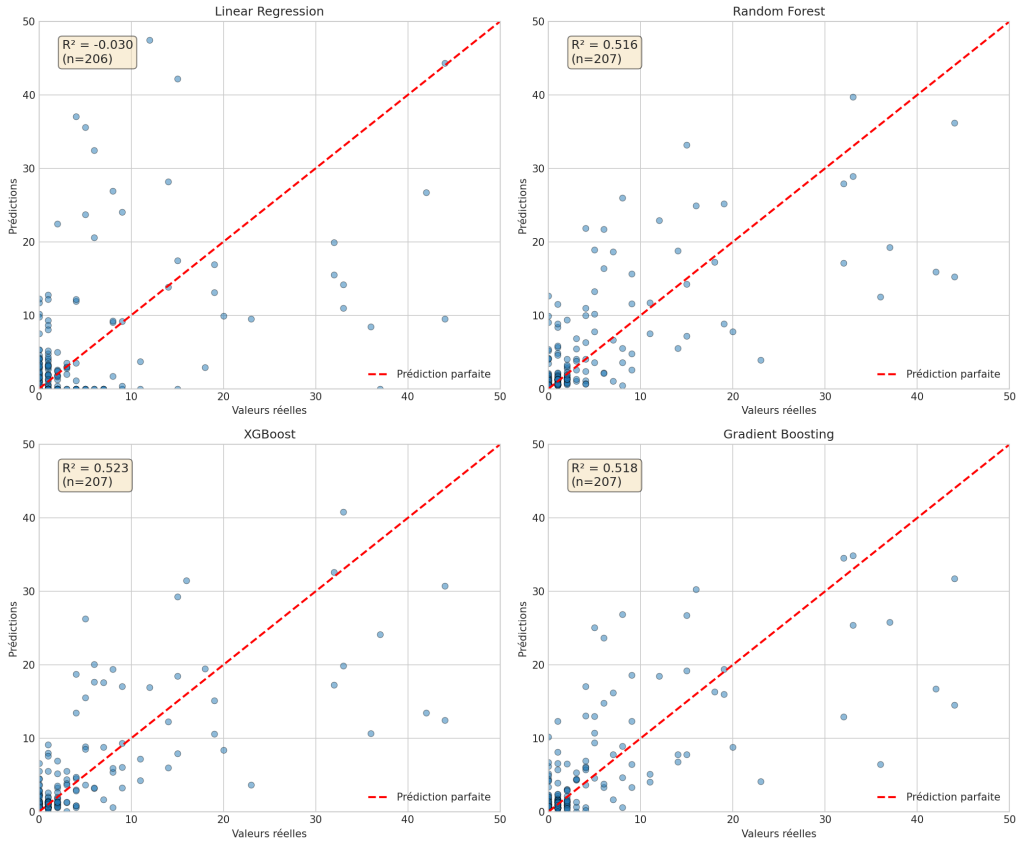

FIGURE 3 – Vue zoomée sur les communes avec moins de 50 accidents pour les prédictions vs les valeurs réelles du nombre d'accidents.

La dispersion importante autour de la diagonale confirme que le modèle est imprécis pour les petites communes.

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.
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

Cinq 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%). Ce résultat est en réalité plus honnête et plus informatif.

Élimination du biais de taille : En normalisant par la population ou la longueur d'aménagement, le modèle ne peut plus identifier simplement les grandes communes. Paris n'a plus un poids disproportionné.

Complexité réelle du risque : Le risque d'accident cycliste dépend de nombreux facteurs absents de nos données : comportement des usagers, densité du trafic automobile, conditions météorologiques, qualité de l'éclairage.

Données insuffisantes : Nous ne disposons pas assez de datasets permettant de capturer toute la complexité du problème.

Pour le taux de risque par habitant, l'analyse de l'importance des features (Random Forest) révèle que la population (17.1%), la longueur des aménagements (11.8%) et la densité population/aménagement (10.4%) sont les variables les plus influentes.

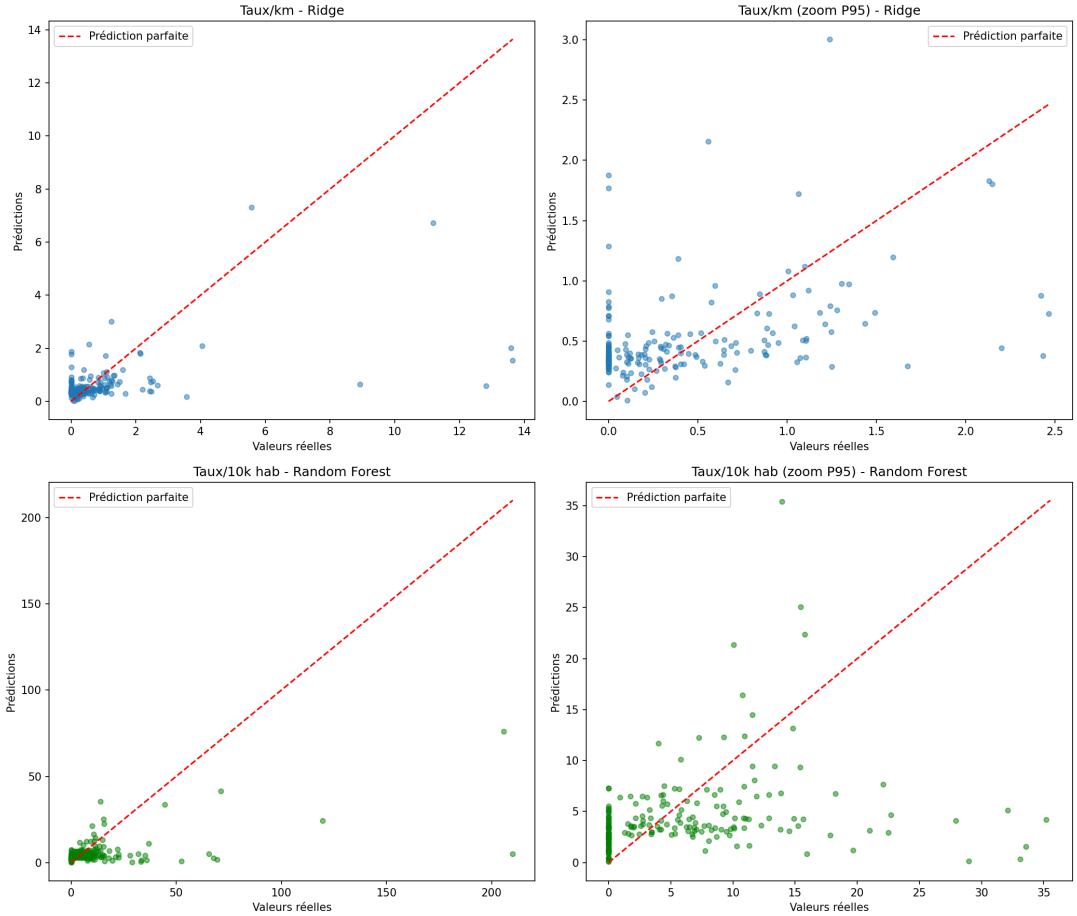

FIGURE 4 – Prédictions vs valeurs réelles pour les deux taux de risque.

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.

6.2 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.
- **Couverture des comptages** : Seulement 69 compteurs automatiques pour toute l'Île-de-France.
- **Variables manquantes** : Nous n'avons pas accès au trafic automobile ni aux caractéristiques urbanistiques fines.
- **Proportion de zéros** : 33.6% des communes n'ont aucun accident déclaré.

7 Conclusion

Ce projet a exploré différentes approches de machine learning pour modéliser le risque d'accidents de vélo en Île-de-France à partir de données ouvertes.

La **classification binaire** du risque communal (élevé vs faible) atteint un ROC-AUC de 95.6% avec une régression logistique, permettant d'identifier les communes prioritaires pour des interventions de sécurité.

La **prédiction du nombre brut d'accidents** affiche un R^2 de 89%, mais cette performance est biaisée par l'effet de taille des communes (Paris concentre 61% des accidents) et une corrélation paradoxale entre aménagements et accidents.

L'**approche par taux de risque normalisés** (par habitant ou par kilomètre d'aménagement) offre une méthodologie plus rigoureuse. Les performances plus modestes (R^2 de 20-30%) reflètent la complexité réelle du problème et l'insuffisance des données disponibles.

Cette étude illustre qu'un indicateur de performance élevé ne garantit pas la validité d'un modèle. L'analyse critique des données et des résultats reste indispensable pour éviter des conclusions erronées.