Analyse des Accidents de Vélo en Île-de-France

Projet de Science des Données Appliquée

Projet SDA
Master Data Science

Janvier 2026

Résumé

Ce rapport présente une analyse approfondie des accidents de vélo en Île-de-France en utilisant des techniques de machine learning. Nous répondons à deux questions principales : (1) identifier les communes à risque élevé d'accidents, et (2) prédire le nombre d'accidents par commune en fonction des aménagements cyclables. L'analyse s'appuie sur trois sources de données fusionnées : les accidents de vélo, les aménagements cyclables et les comptages de trafic vélo. Les résultats montrent qu'une régression logistique permet de classifier le risque avec un ROC-AUC de 95.6%, tandis qu'un modèle Gradient Boosting prédit le nombre d'accidents avec un R^2 de 89.2%.

Table des matières

1 Introduction

1.1 Contexte

La pratique du vélo connaît un essor considérable en Île-de-France ces dernières années, favorisée par les politiques de mobilité durable et le développement des infrastructures cyclables. Cependant, cette augmentation du trafic cycliste s'accompagne d'une problématique de sécurité routière qu'il convient d'analyser et de modéliser.

1.2 Objectifs du Projet

Ce projet vise à répondre à deux questions fondamentales :

1. **Classification du risque** : La commune présente-t-elle un risque élevé d'accidents de vélo ?
2. **Prédiction du nombre d'accidents** : Peut-on prédire le nombre d'accidents dans une commune en fonction de ses aménagements cyclables ?

1.3 Méthodologie

Notre approche se décompose en plusieurs étapes :
— Collecte et fusion de trois sources de données
— Préparation et nettoyage des données
— Analyse exploratoire
— Modélisation par machine learning
— Comparaison des algorithmes et conclusions

2 Description des Données

2.1 Sources de Données

Nous utilisons trois jeux de données complémentaires :

2.1.1 Accidents de Vélo

Ce dataset contient **80 022 accidents** sur l'ensemble de la France, dont **22 609 en Île-de-France**. Les principales variables sont :
— Localisation (département, commune, coordonnées GPS)
— Date et heure de l'accident
— Gravité (indemne, blessé léger, hospitalisé, tué)
— Conditions (luminosité, météo, état de la chaussée)
— Caractéristiques de la victime (âge, sexe)

2.1.2 Aménagements Cyclables

Ce dataset recense **143 060 aménagements cyclables** en Île-de-France :
— Type de voie (piste cyclable, bande cyclable, voie partagée)
— Longueur de l'aménagement
— Revêtement et état
— Localisation par commune (code INSEE)

2.1.3 Comptages Vélo

Ce dataset contient **933 757 mesures** de comptage horaire :
— Identifiant et localisation du compteur
— Comptage horaire de passages
— Date et heure de la mesure

2.2 Fusion des Données

Les données ont été fusionnées au niveau communal (code INSEE) pour créer un dataset final de **1 124 communes** avec 38 variables. La Figure 1 illustre le processus de fusion.

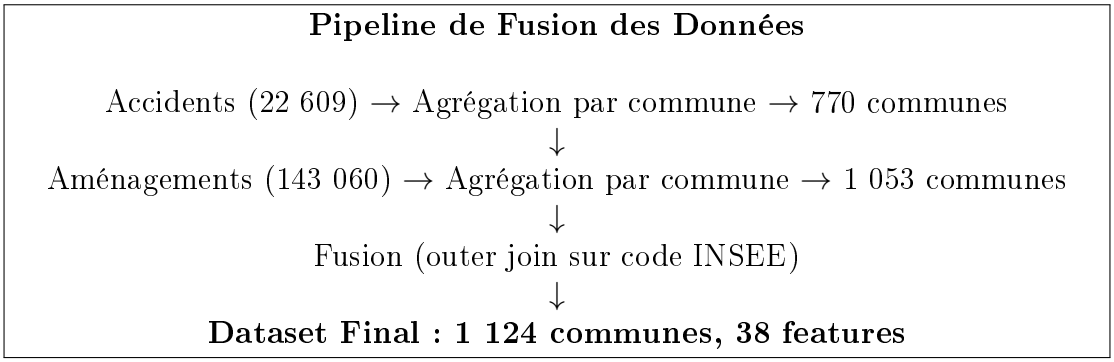

FIGURE 1 – Pipeline de préparation des données

2.3 Variables Créées

Plusieurs features ont été calculées lors de la fusion :
— `nb_accidents` : nombre total d'accidents par commune
— `nb_accidents_graves` : accidents avec hospitalisation ou décès
— `taux_accidents_graves` : ratio d'accidents graves
— `nb_amenagements` : nombre d'infrastructures cyclables
— `longueur_totale_amenagements` : longueur cumulée en mètres
— `ratio_pistes_cyclables` : proportion de pistes séparées
— `risque_eleve` : variable binaire (1 si $\geq 75^{e}$ percentile)

3 Analyse 1 : Classification du Risque

3.1 Problématique

La première question consiste à classifier les communes selon leur niveau de risque d'accidents de vélo. On définit une commune à **risque élevé** si son nombre d'accidents est supérieur ou égal au 75^{e} percentile (soit ≥ 6 accidents).

3.2 Modèles Testés

Six algorithmes de classification ont été comparés :

1. Régression Logistique (avec pondération des classes)

2. Random Forest Classifier
3. Gradient Boosting Classifier
4. XGBoost Classifier
5. LightGBM Classifier
6. SVM avec noyau RBF

3.3 Résultats

TABLE 1 – Comparaison des modèles de classification

Modèle	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.884	0.758	0.833	**0.794**	**0.956**
Random Forest	0.898	0.863	0.733	0.793	0.951
SVM (RBF)	0.880	0.739	0.850	0.791	0.953
LightGBM	0.889	0.807	0.767	0.786	0.940
XGBoost	0.884	0.793	0.767	0.780	0.941
Gradient Boosting	0.880	0.824	0.700	0.757	0.953

3.4 Analyse des Résultats

Le modèle de **Régression Logistique** obtient les meilleures performances globales avec :

— Un **F1-Score de 79.4%**, équilibrant precision et recall
— Un **ROC-AUC de 95.6%**, indiquant une excellente capacité discriminative
— Une bonne stabilité en validation croisée (CV F1 = 0.816 ± 0.033)

La Figure 2 présente les courbes ROC comparatives des modèles.

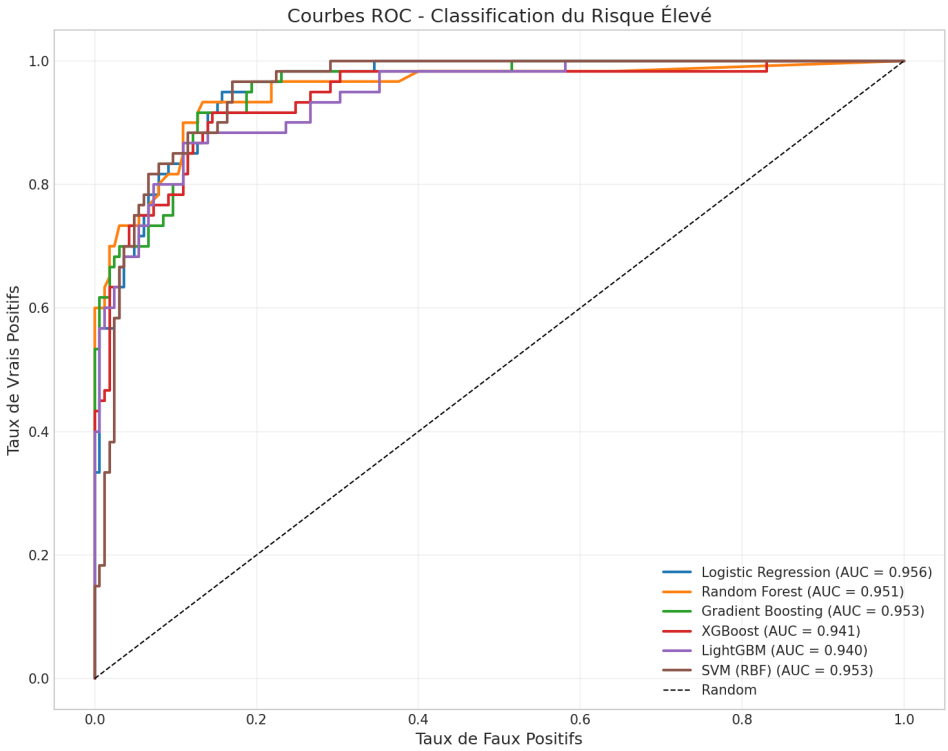

FIGURE 2 – Courbes ROC des modèles de classification

3.5 Importance des Features

L'analyse de l'importance des features révèle que les variables les plus discriminantes sont :

1. `nb_amenagements` : nombre d'infrastructures cyclables
2. `longueur_totale_amenagements` : longueur des aménagements
3. `nb_pistes_cyclables` : nombre de pistes dédiées
4. `est_paris` : indicateur d'arrondissement parisien

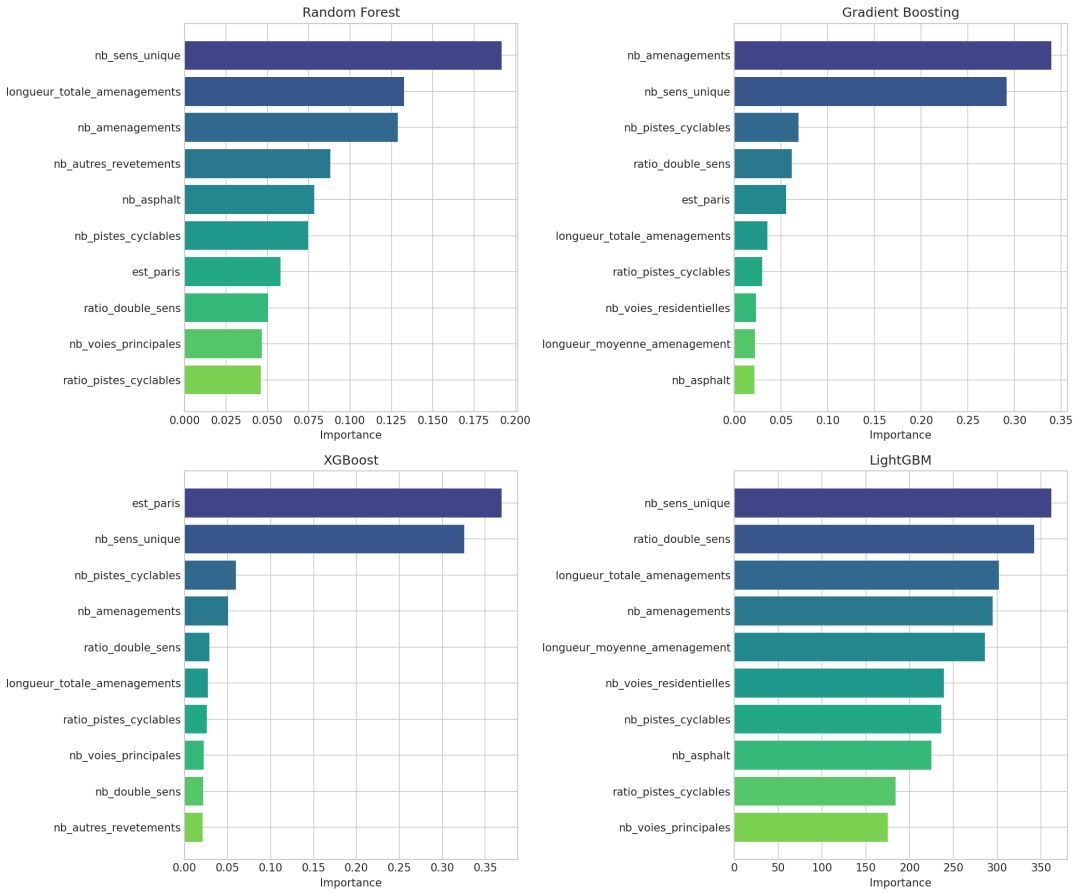

FIGURE 3 – Importance des features par modèle (Classification)

4 Analyse 2 : Prédiction du Nombre d'Accidents

4.1 Problématique

La seconde question vise à prédire le nombre exact d'accidents par commune en fonction des caractéristiques des aménagements cyclables. Il s'agit d'un problème de **régression**.

4.2 Modèles Testés

Sept algorithmes de régression ont été comparés :

1. Régression Linéaire
2. Ridge Regression (régularisation L2)
3. Lasso Regression (régularisation L1)
4. Random Forest Regressor
5. Gradient Boosting Regressor
6. XGBoost Regressor
7. LightGBM Regressor

4.3 Résultats

TABLE 2 – Comparaison des modèles de régression

Modèle	RMSE	MAE	R^2	CV R^2
Gradient Boosting	**37.06**	**8.40**	**0.892**	0.897 ± 0.025
XGBoost	37.53	8.43	0.889	$\mathbf{0.900 \pm 0.018}$
Random Forest	37.80	9.06	0.888	0.886 ± 0.027
Linear Regression	42.89	11.39	0.856	0.855 ± 0.072
Ridge Regression	43.23	11.34	0.853	0.857 ± 0.067
Lasso Regression	43.87	11.38	0.849	0.856 ± 0.065
LightGBM	45.41	10.29	0.838	0.825 ± 0.074

4.4 Analyse des Résultats

Le modèle **Gradient Boosting** obtient les meilleures performances :
— $\mathbf{R^2 = 0.892}$: le modèle explique 89.2% de la variance
— **MAE = 8.40** : erreur moyenne de 8.4 accidents par commune
— **RMSE = 37.06** : pénalise davantage les grosses erreurs
— Excellente stabilité en validation croisée

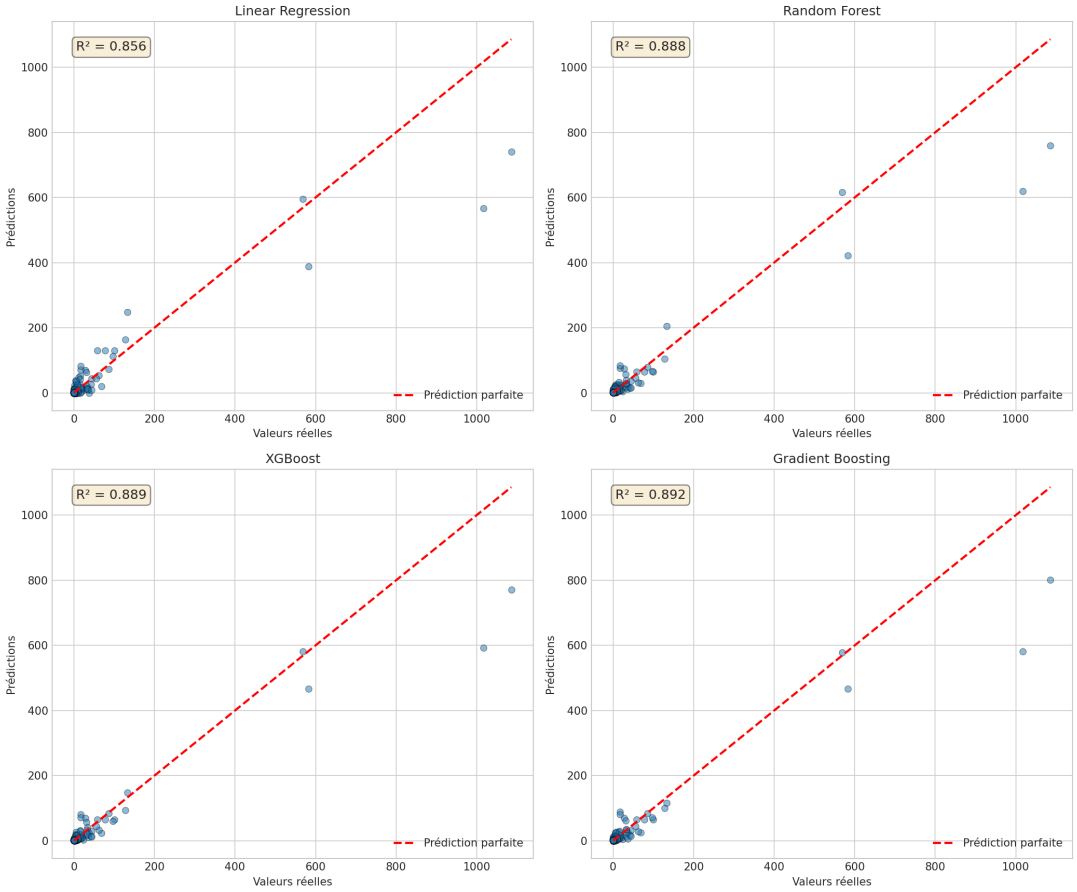

FIGURE 4 – Prédictions vs valeurs réelles pour différents modèles

4.5 Analyse des Corrélations

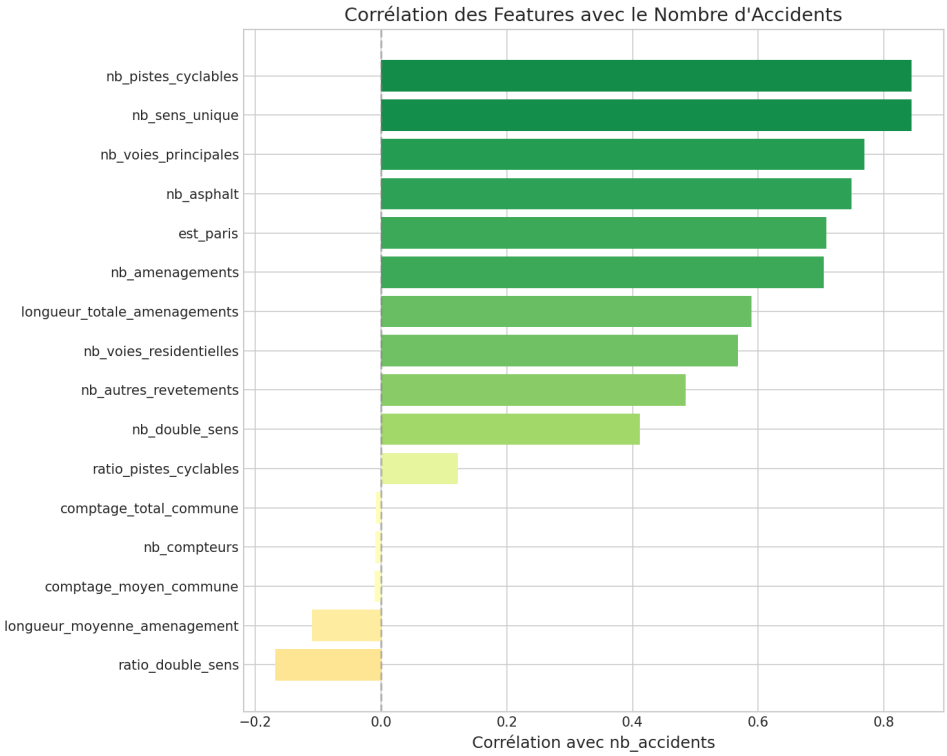

FIGURE 5 – Corrélation des features avec le nombre d'accidents

Les features les plus corrélées avec le nombre d'accidents sont :

— `nb_pistes_cyclables` : $r = 0.844$

— `nb_amenagements` : $r = 0.841$

— `longueur_totale_amenagements` : $r = 0.817$

Interprétation : La corrélation positive entre aménagements et accidents s'explique par le fait que les communes bien équipées ont aussi un trafic cycliste plus important, donc statistiquement plus d'accidents.

5 Discussion

5.1 Synthèse des Résultats

TABLE 3 – Synthèse des meilleurs modèles

Question	**Meilleur Modèle**	**Performance**
Classification du risque	Régression Logistique	ROC-AUC = 95.6%
Prédiction nb accidents	Gradient Boosting	R^2 = 89.2%

5.2 Limites de l'Étude

— **Biais de report** : tous les accidents ne sont pas déclarés

— **Données de comptage** : seuls 69 compteurs pour toute l'IDF
— **Variables manquantes** : population, densité urbaine, trafic automobile
— **Temporalité** : pas de prise en compte de l'évolution temporelle

5.3 Perspectives

Pour améliorer les modèles, on pourrait :
— Intégrer des données démographiques (population, densité)
— Ajouter des données de trafic automobile
— Effectuer une analyse temporelle (tendances, saisonnalité)
— Utiliser des modèles spatiaux (autocorrélation géographique)

6 Conclusion

Ce projet a permis de développer deux modèles prédictifs performants pour l'analyse des accidents de vélo en Île-de-France :

1. Un **modèle de classification** (Régression Logistique) capable d'identifier les communes à risque élevé avec une précision de 88% et un ROC-AUC de 96%.
2. Un **modèle de régression** (Gradient Boosting) capable de prédire le nombre d'accidents avec un R^2 de 89%, soit une erreur moyenne de 8.4 accidents par commune.

Recommandations pour les décideurs :
— Prioriser les communes à haut risque identifiées par le modèle
— Privilégier les pistes cyclables séparées de la circulation
— Améliorer l'éclairage et la signalisation sur les voies cyclables
— Développer les comptages pour mieux mesurer l'exposition au risque
Les modèles développés constituent un outil d'aide à la décision pour les collectivités souhaitant optimiser leur politique de sécurité cycliste.

Annexes

A. Structure du Projet

```
Projet_sda_top/
+-- data/
|   +-- accidentsVelo.csv
|   +-- amenagements-velo-en-ile-de-france.csv
|   +-- comptage-velo-donnees-compteurs.csv
|   +-- dataset_final_idf.csv
+-- src/
|   +-- 01_preparation_donnees.py
|   +-- 02_classification_risque.py
|   +-- 03_prediction_accidents.py
+-- outputs/
|   +-- roc_curves_classification.png
|   +-- confusion_matrix_classification.png
```

```
|   +-- feature_importance_classification.png
|   +-- predictions_vs_reel.png
|   +-- ...
+-- rapport/
|   +-- rapport.tex
+-- requirements.txt
```

B. Technologies Utilisées

- **Python 3.12**
- **pandas** : manipulation de données
- **scikit-learn** : modèles de machine learning
- **XGBoost, LightGBM** : modèles de gradient boosting
- **matplotlib, seaborn** : visualisation
- **LaTeX** : rédaction du rapport