Projet Sentinelle Écologique : Détection Intelligente de la Pollution Aquatique

1. Introduction

La pollution aquatique représente une menace majeure pour les écosystèmes et la santé humaine. La détection précoce et précise des états de pollution est cruciale pour la mise en œuvre de mesures correctives efficaces. Ce projet vise à développer un système intelligent, baptisé “Sentinelle Écologique”, capable de classifier l’état de l’eau (Sain ou Pollué) à partir de données de capteurs environnementaux. Nous explorerons et comparerons les performances de différents modèles de classification, notamment la Régression Logistique et les Machines à Vecteurs de Support (SVM) avec et sans noyau, sur un jeu de données simulant des mesures IoT complexes.

2. Méthodologie

2.1. Génération du Jeu de Données “Sentinelle Écologique”

Pour ce projet, un jeu de données synthétique a été créé pour simuler des mesures provenant de capteurs IoT déployés dans des cours d’eau. Ce dataset, nommé `eco_sentinel_dataset.csv`, comprend les caractéristiques suivantes :

- **Turbidité (NTU)** : Mesure de la clarté de l’eau, indiquant la présence de matières en suspension.
- **Oxygène Dissous (mg/L)** : Quantité d’oxygène disponible dans l’eau, essentielle à la vie aquatique.
- **État de l’Eau** : Variable cible binaire (0 pour ‘Sain’, 1 pour ‘Pollué’).

Le jeu de données a été conçu pour présenter une séparation non-linéaire complexe, où les points représentant l’eau saine sont regroupés au centre, tandis que les points d’eau polluée sont dispersés en périphérie. Cette structure permet de tester la

robustesse des modèles face à des problèmes de classification plus réalistes que des données linéairement séparables.

2.2. Modèles de Classification

Trois modèles de classification ont été entraînés et évalués :

- **Régression Logistique** : Un modèle linéaire simple, souvent utilisé comme référence.
- **Machine à Vecteurs de Support (SVM) Linéaire** : Un classifieur linéaire qui cherche à trouver l'hyperplan optimal séparant les classes.
- **Machine à Vecteurs de Support (SVM) avec Noyau RBF (Radial Basis Function)** : Un modèle non-linéaire capable de projeter les données dans un espace de dimension supérieure pour trouver une séparation optimale, même pour des données non-linéairement séparables.

2.3. Prétraitement des Données

Avant l'entraînement des modèles, les caractéristiques ont été standardisées (mise à l'échelle pour avoir une moyenne nulle et un écart-type unitaire). Cette étape est cruciale, en particulier pour les SVM, afin d'assurer que toutes les caractéristiques contribuent équitablement à la distance euclidienne et d'éviter que les caractéristiques avec de plus grandes valeurs numériques ne dominent le processus d'apprentissage.

2.4. Évaluation des Performances

Les modèles ont été évalués sur un ensemble de test distinct, en utilisant les métriques suivantes :

- **Précision (Accuracy)** : Proportion des prédictions correctes.
- **Matrices de Confusion** : Tableau résumant les prédictions correctes et incorrectes pour chaque classe.
- **Surfaces de Décision** : Visualisation graphique des frontières de décision apprises par chaque modèle.

3. Résultats et Discussion

3.1. Performances des Modèles

Le tableau ci-dessous résume les précisions obtenues par chaque modèle sur le jeu de données “Sentinelle Écologique” :

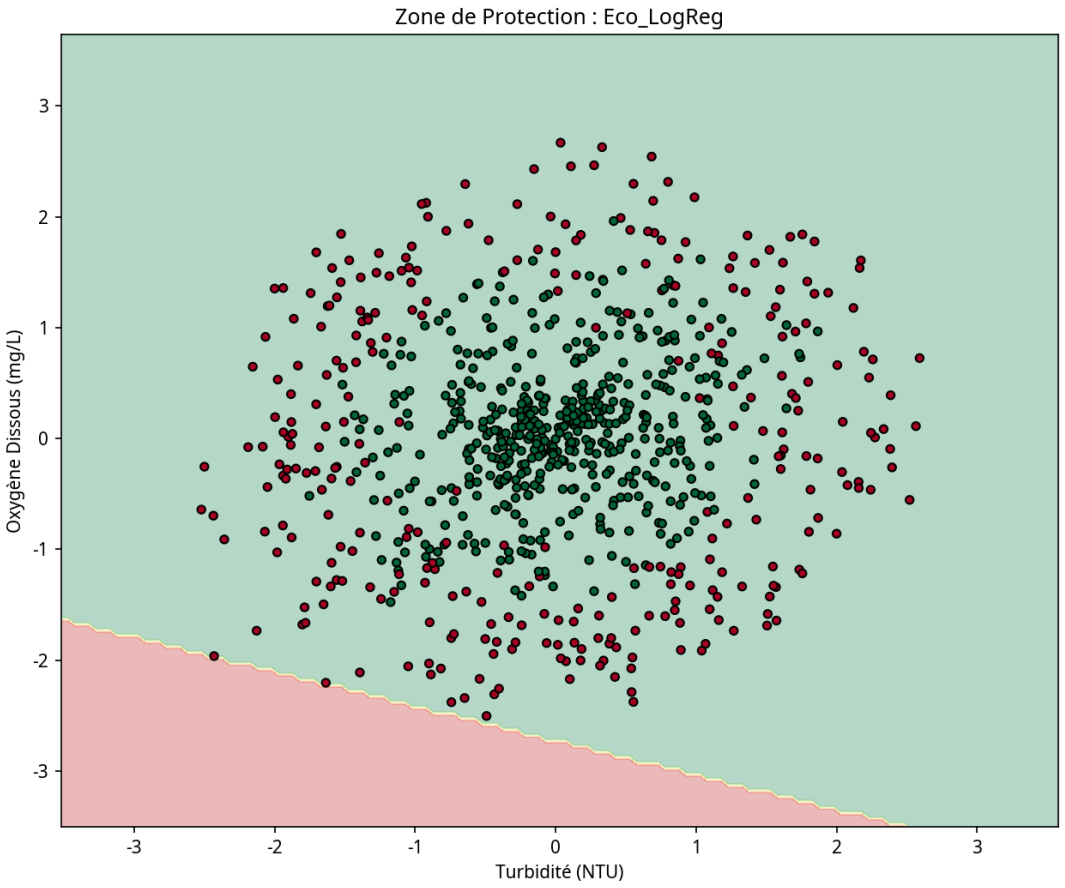
Modèle	Précision (Accuracy)
Régression Logistique	65.00%
SVM Linéaire	64.00%
SVM avec Noyau RBF	**91.50%**

Comme anticipé, la Régression Logistique et le SVM Linéaire affichent des performances relativement faibles. Cela est dû à la nature non-linéaire du problème de classification, que ces modèles linéaires peinent à résoudre efficacement. En revanche, le SVM avec noyau RBF démontre une précision nettement supérieure, atteignant 91.50%.

3.2. Surfaces de Décision

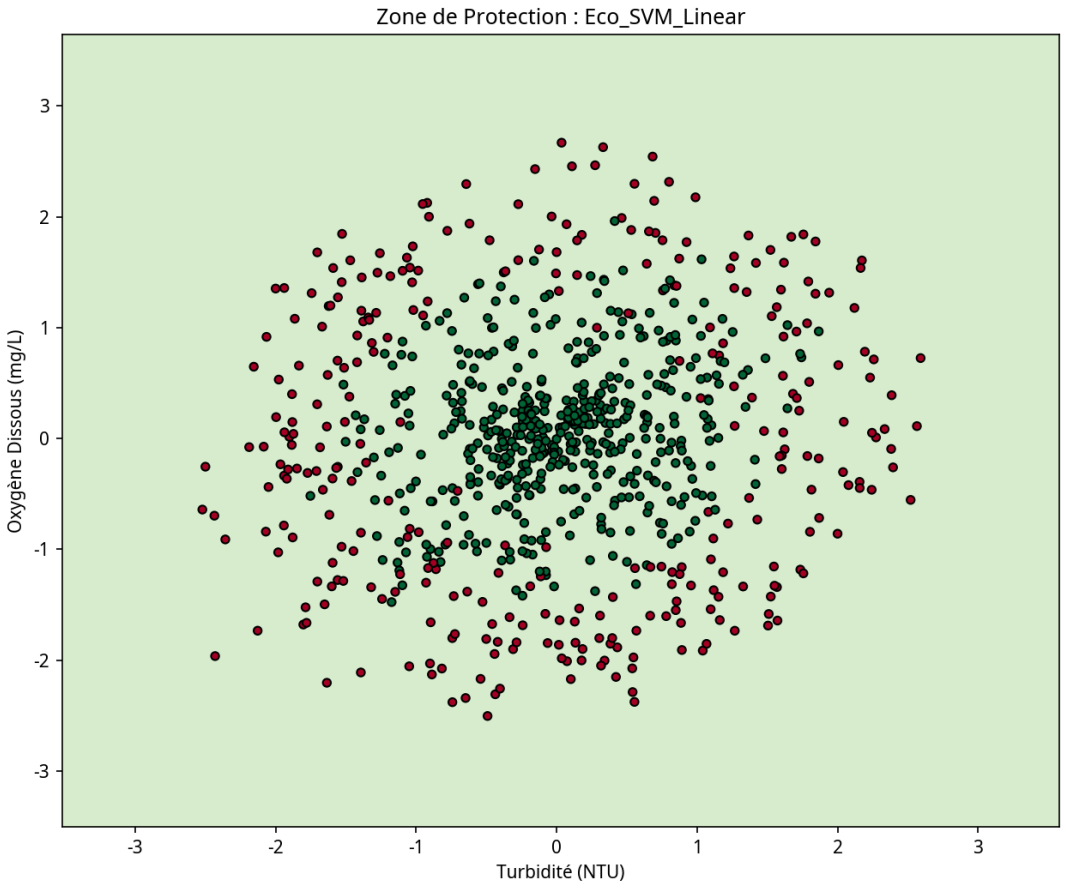
Les surfaces de décision illustrent visuellement la capacité de chaque modèle à séparer les classes ‘Sain’ et ‘Pollué’.

Régression Logistique



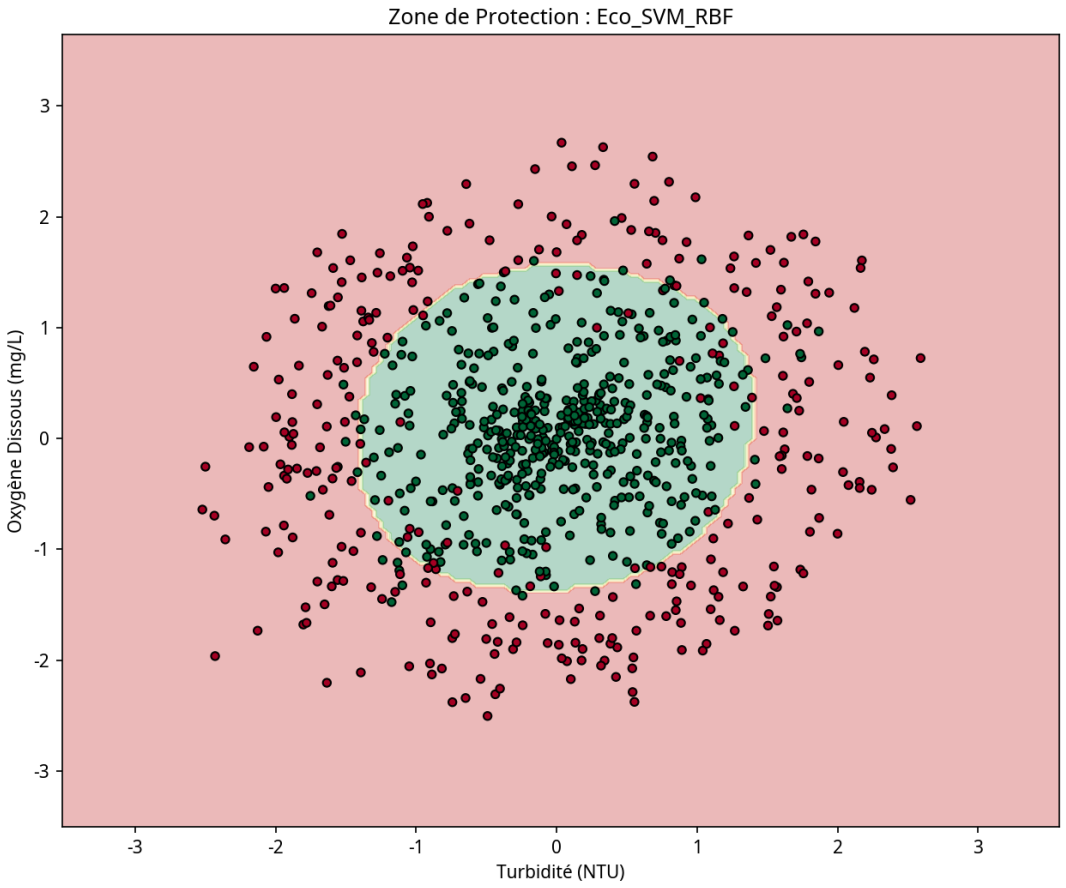
La Régression Logistique tente de trouver une frontière de décision linéaire, ce qui est clairement insuffisant pour séparer les données en forme d'anneau de notre dataset. De nombreux points sont mal classifiés.

SVM Linéaire



Similaire à la Régression Logistique, le SVM Linéaire échoue également à capturer la structure non-linéaire des données, résultant en une frontière de décision linéaire inefficace.

SVM avec Noyau RBF

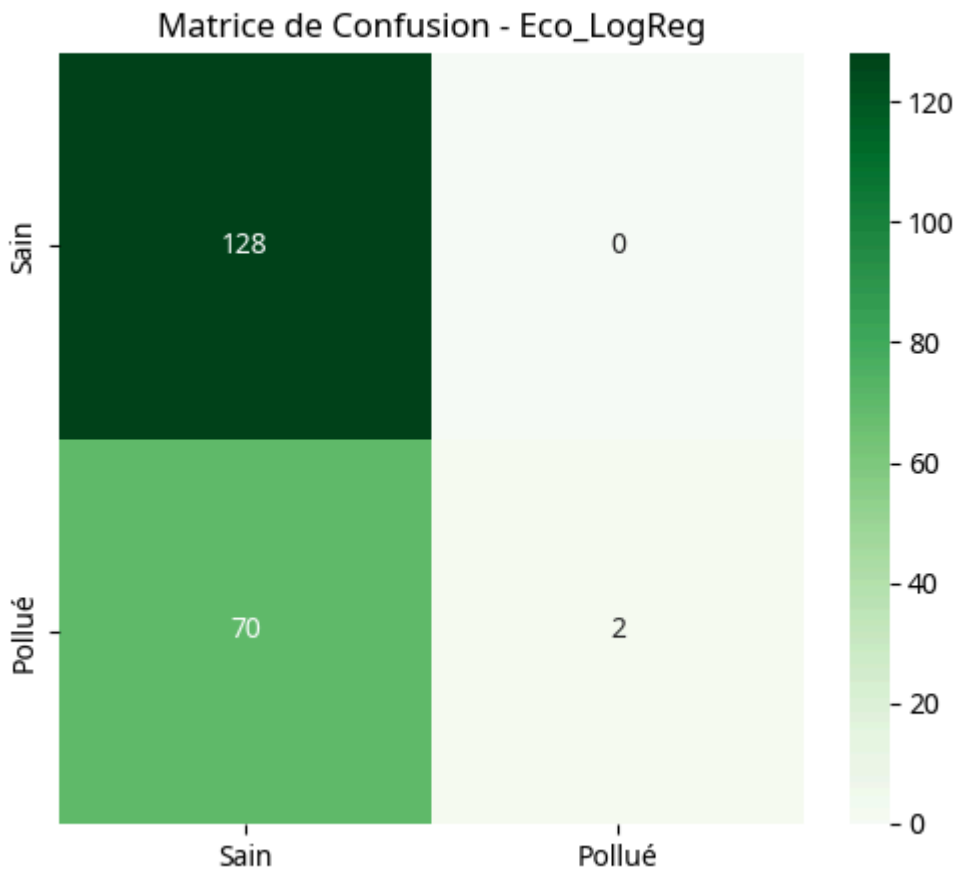


Le SVM avec noyau RBF, grâce à sa capacité à créer des frontières de décision non-linéaires, parvient à encapsuler la classe ‘Sain’ au centre et à isoler la classe ‘Pollué’ en périphérie. Cette visualisation confirme sa supériorité pour ce type de problème.

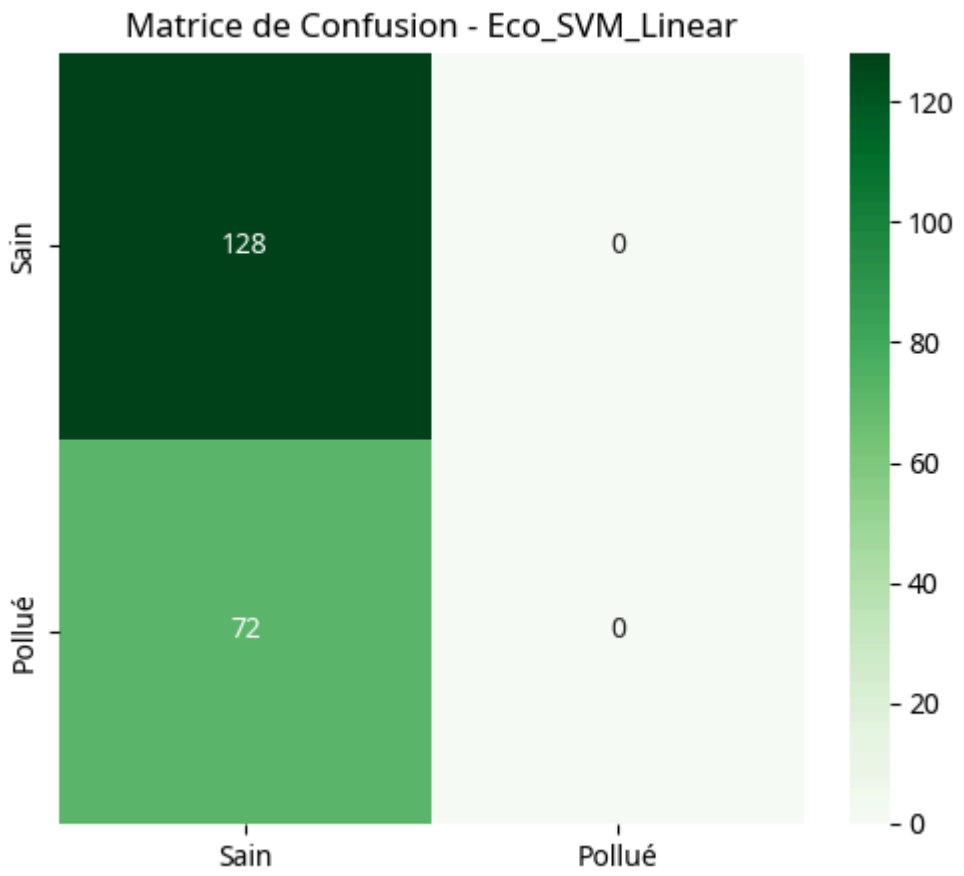
3.3. Matrices de Confusion

Les matrices de confusion fournissent une vue détaillée des erreurs de classification.

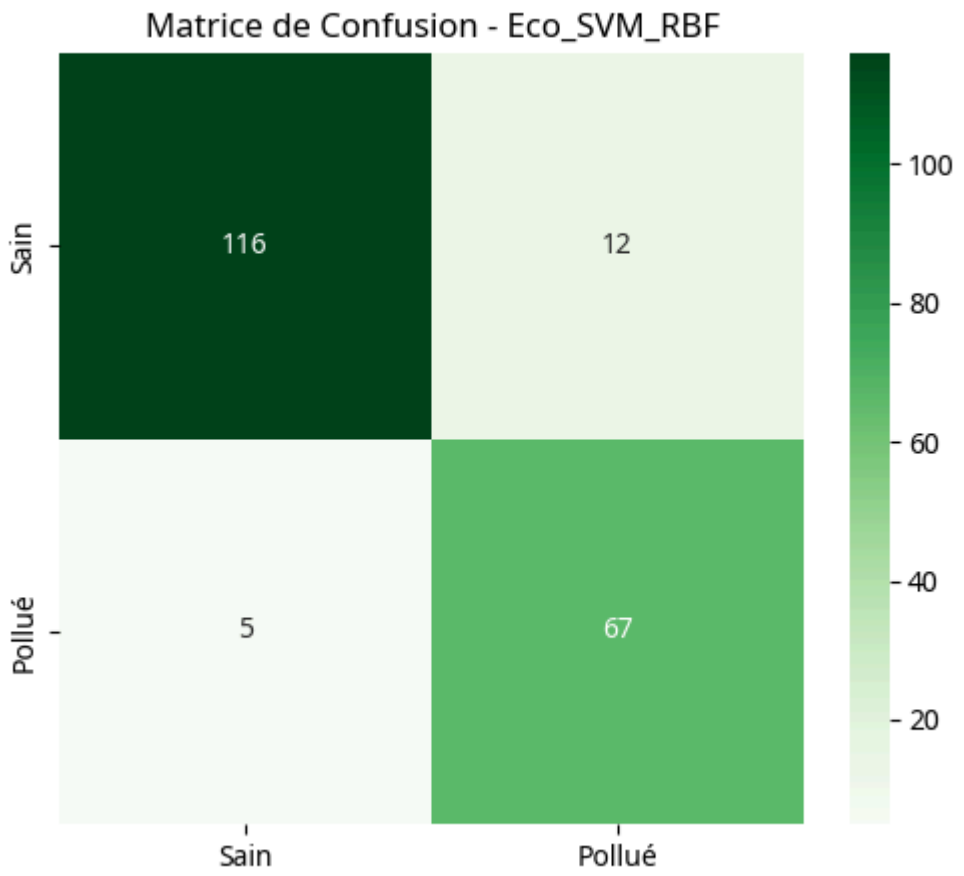
Matrice de Confusion : Régression Logistique



Matrice de Confusion : SVM Linéaire



Matrice de Confusion : SVM avec Noyau RBF



Les matrices de confusion confirment les observations précédentes. Pour la Régression Logistique et le SVM Linéaire, on observe un nombre significatif de faux positifs et de faux négatifs. En revanche, le SVM avec noyau RBF présente un nombre très faible d'erreurs, ce qui se traduit par une précision élevée et une classification quasi parfaite des états de l'eau.

4. Conclusion

Ce projet a démontré l'importance cruciale du choix du modèle de classification en fonction de la nature des données. Pour le problème de détection de pollution aquatique avec des corrélations non-linéaires, les modèles linéaires tels que la Régression Logistique et le SVM Linéaire se sont avérés inefficaces. Le **SVM avec noyau RBF** a, quant à lui, prouvé son excellence en capturant la complexité sous-jacente des données et en atteignant une précision remarquable. Le projet "Sentinelle Écologique" illustre comment l'intelligence artificielle, et en particulier les algorithmes

de machine learning avancés, peuvent jouer un rôle vital dans la protection de l'environnement en fournissant des outils de surveillance et d'alerte précis. Ce projet constitue une base solide pour des développements futurs, incluant l'intégration de données en temps réel et l'exploration d'autres caractéristiques environnementales.