المدرسة الوطنية العليا للذكاء الاصطناعي وعلوم المعطيات - تارودانت
ÉCOLE NATIONALE SUPÉRIEURE DE L'INTELLIGENCE ARTIFICIELLE
ET SCIENCES DES DONNÉES - TAROUDANT

RAPPORT DE PROJET

Filière: « Sciences des Données, Big Data & IA »

Détection d'anomalies dans la qualité de l'eau

Lien Github : https://github.com/adernajat/Detection-of-anomalies-in-water-quality.git

Réaliseés par :

- ADERDOUR NAJAT
- ABOUZAID WIJDANE
- NACIR DOUAE

Encadreé par :

Pr. EL-ATEIF SARA

Année Universitaire 2024- 2025

INTRODUCTION GENERALE

La qualité de l'eau est un élément essentiel pour la santé humaine. Une eau non potable peut présenter des risques importants, ce qui nécessite des méthodes fiables pour analyser et contrôler ses caractéristiques physico-chimiques. Avec l'augmentation des données disponibles, les techniques de machine learning offrent des solutions efficaces pour détecter automatiquement des comportements anormaux dans ces données.

Dans ce projet, nous appliquons des méthodes de **détection d'anomalies non supervisées** afin d'identifier des échantillons d'eau atypiques. Trois algorithmes sont étudiés et comparés : **Isolation Forest**, **One-Class SVM** et **Local Outlier Factor (LOF)**, dans le but de déterminer celui qui est le plus adapté à notre jeu de données.

Ensuite, pour vérifier si les anomalies détectées correspondent réellement à de l'eau non potable, nous utilisons plusieurs algorithmes de **classification supervisée** pour prédire la variable *Potability*. Cette étape permet de comparer les performances des modèles et de valider la pertinence des anomalies détectées.

Ce projet montre ainsi comment la combinaison de la détection d'anomalies et de la classification peut aider à mieux analyser la qualité de l'eau et à soutenir la prise de décision.

CONTEXTE DU PROJET

La surveillance de la qualité de l'eau repose sur l'analyse de plusieurs paramètres physico-chimiques tels que le pH, la turbidité, les solides dissous ou la conductivité. Ces données sont souvent volumineuses et peuvent contenir des valeurs aberrantes dues à des erreurs de mesure, des conditions environnementales particulières ou des phénomènes de contamination. Les méthodes classiques basées uniquement sur des seuils fixes peuvent être limitées face à la complexité de ces données. Dans ce contexte, les techniques de machine learning offrent des outils efficaces pour analyser automatiquement les données et détecter des comportements anormaux.

PROBLÉMATIQUE

Comment détecter efficacement les anomalies dans les données de qualité de l'eau à partir de variables physico-chimiques, sans disposer d'annotations précises, et comment vérifier si ces anomalies correspondent réellement à une eau non potable ?
De plus, parmi les différentes méthodes de détection d'anomalies et de classification, lesquelles sont les plus performantes et les mieux adaptées à ce jeu de données ?

OBJECTIFS DU PROJET

L'objectif principal de ce projet est d'explorer l'apport des techniques de machine learning dans l'analyse de la qualité de l'eau, en combinant des approches de détection d'anomalies non supervisées et de classification supervisée.

Plus précisément, ce projet vise à :

- Mettre en œuvre et comparer plusieurs algorithmes de détection d'anomalies non supervisées, à savoir Isolation Forest, Local Outlier Factor et One-Class SVM, afin d'identifier les observations atypiques présentes dans les données de qualité de l'eau.

- Analyser les caractéristiques des anomalies détectées et déterminer les variables physico-chimiques les plus contributives à ces anomalies.
- Évaluer la relation entre les anomalies détectées et la potabilité de l'eau, afin de vérifier si ces observations correspondent majoritairement à des eaux non potables.
- Appliquer et comparer différents algorithmes de classification supervisée pour la prédiction de la potabilité de l'eau, notamment la régression logistique, le Random Forest, le Support Vector Machine et XGBoost.
- Identifier les modèles les plus performants en termes de capacité de généralisation et de précision, et analyser l'impact du nettoyage des données par détection d'anomalies sur les performances de classification.

DESCRIPTION DU JEU DE DONNÉES

Le jeu de données utilisé dans ce projet, intitulé Water-dataset.csv, est composé de 5000 observations, correspondant à des échantillons d'eau analysés à partir de différentes mesures physico-chimiques. Chaque observation est décrite par six variables (features), représentant des indicateurs essentiels pour l'évaluation de la qualité de l'eau.

Ce jeu de données a été sélectionné pour plusieurs raisons. Tout d'abord, il constitue un support pertinent pour l'évaluation et la comparaison de méthodes de détection d'anomalies, notamment dans un contexte de données multivariées. Ensuite, il reflète des données proches de situations réelles, issues de capteurs industriels utilisés pour le suivi et le contrôle de la qualité de l'eau. Enfin, sa structure permet d'étudier le lien entre les comportements anormaux des paramètres mesurés et la potabilité de l'eau.

Le jeu de données a été récupéré à partir d'un dépôt GitHub public, ce qui garantit sa disponibilité et sa réutilisabilité à des fins de recherche et d'expérimentation.

EXPLORATORY DATA ANALYSIS

L'analyse exploratoire des données a mis en évidence la présence de valeurs extrêmes et de distributions asymétriques pour plusieurs paramètres physico-chimiques. Ces observations justifient l'utilisation de méthodes de détection d'anomalies non supervisées telles que Isolation Forest, Local Outlier Factor et One-Class SVM, afin d'identifier automatiquement les comportements anormaux dans les données.

1. Vue générale du dataset

Le jeu de données contient 5000 échantillons d'eau décrits par cinq paramètres physico-chimiques et une variable cible binaire (potabilité). Bien que les données soient techniquement propres sans valeurs manquantes ni doublons, deux défis majeurs se posent : un déséquilibre prononcé des classes (92 % potables, 8 % non potables) et la présence de valeurs extrêmes suspectes dans plusieurs variables (pH, chlore, oxygène dissous, température).

```
[4]
print(f" - Nombre de lignes: {dataset.shape[0]}")
print(f" - Nombre de colonnes: {dataset.shape[1]}")
```

```
 - Nombre de lignes: 5000
 - Nombre de colonnes: 6
```

```
[5]
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   pH               5000 non-null   float64
 1   Turbidity        5000 non-null   float64
 2   Chlorine         5000 non-null   float64
 3   Dissolved Oxygen 5000 non-null   float64
 4   Temperature      5000 non-null   int64
 5   Potability       5000 non-null   bool
dtypes: bool(1), float64(4), int64(1)
memory usage: 200.3 KB
```

```
[7]
dataset.duplicated().sum()
```

```
0
```

```
[8]
missing = dataset.isnull().sum()
missing
```

```
pH                 0
Turbidity          0
Chlorine           0
Dissolved Oxygen   0
Temperature        0
Potability         0
dtype: int64
```

```
[9]
potability_counts = dataset['Potability'].value_counts()
print(potability_counts)
```

```
Potability
True     4615
False     385
Name: count, dtype: int64
```

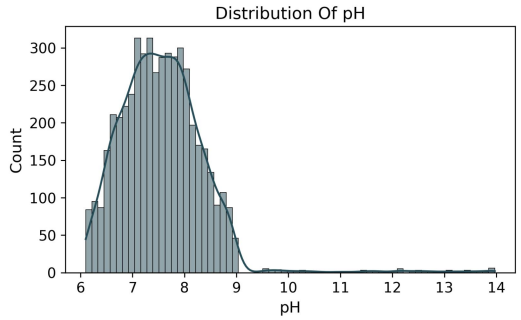
2. Statistiques descriptives

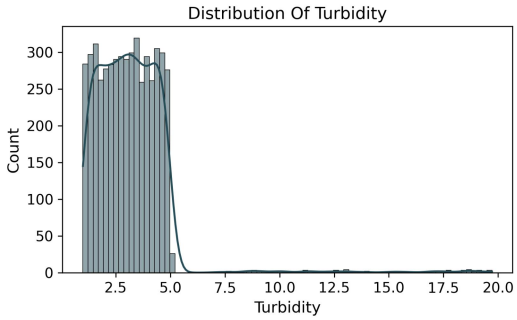
```
dataset.describe()
```

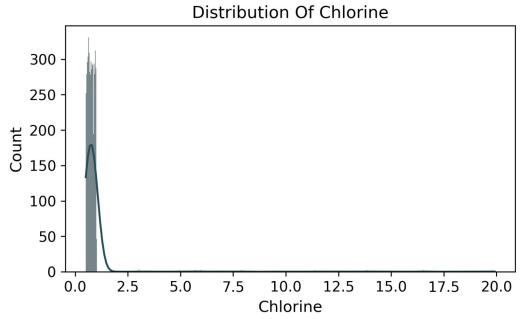
[6]

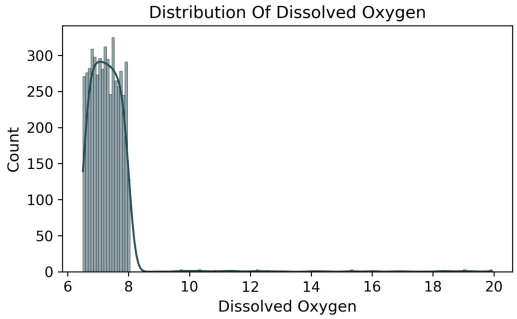
	pH	Turbidity	Chlorine	Dissolved Oxygen	Temperature
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	7.565512	3.154142	0.932878	7.342220	24.735000
std	0.851572	1.772466	1.532965	1.004724	20.399461
min	6.100000	1.000000	0.500000	6.500000	0.000000
25%	7.010000	2.020000	0.630000	6.880000	0.000000
50%	7.510000	3.030000	0.750000	7.250000	25.000000
75%	8.020000	4.040000	0.880000	7.620000	50.000000
max	13.970000	19.720000	19.920000	19.940000	50.000000

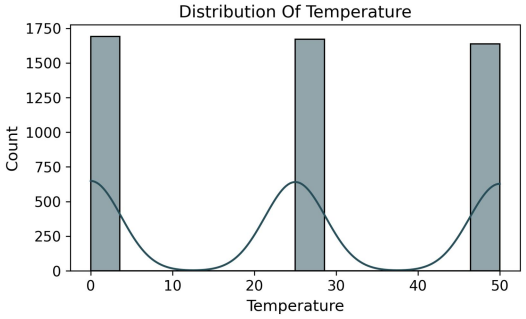
3. Analyse des distributions

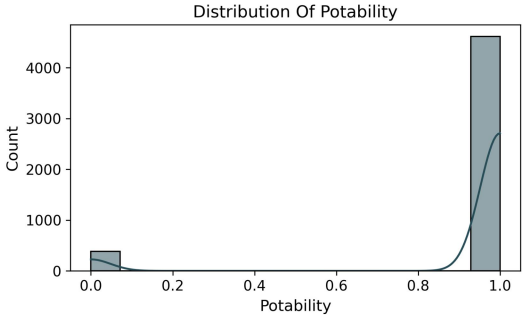






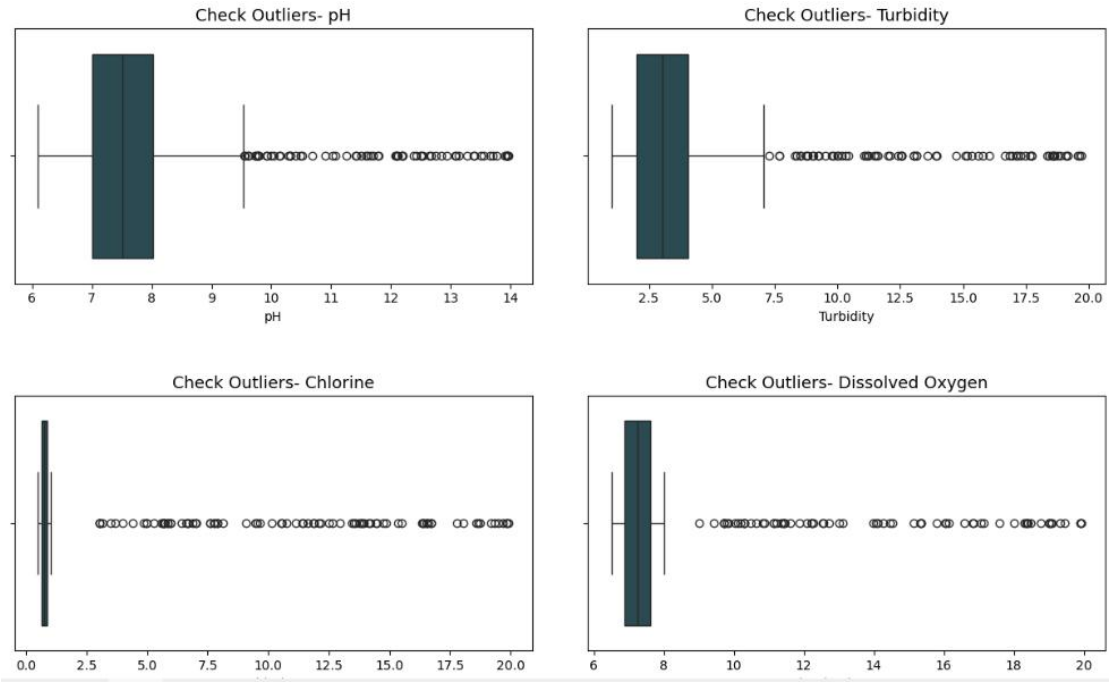






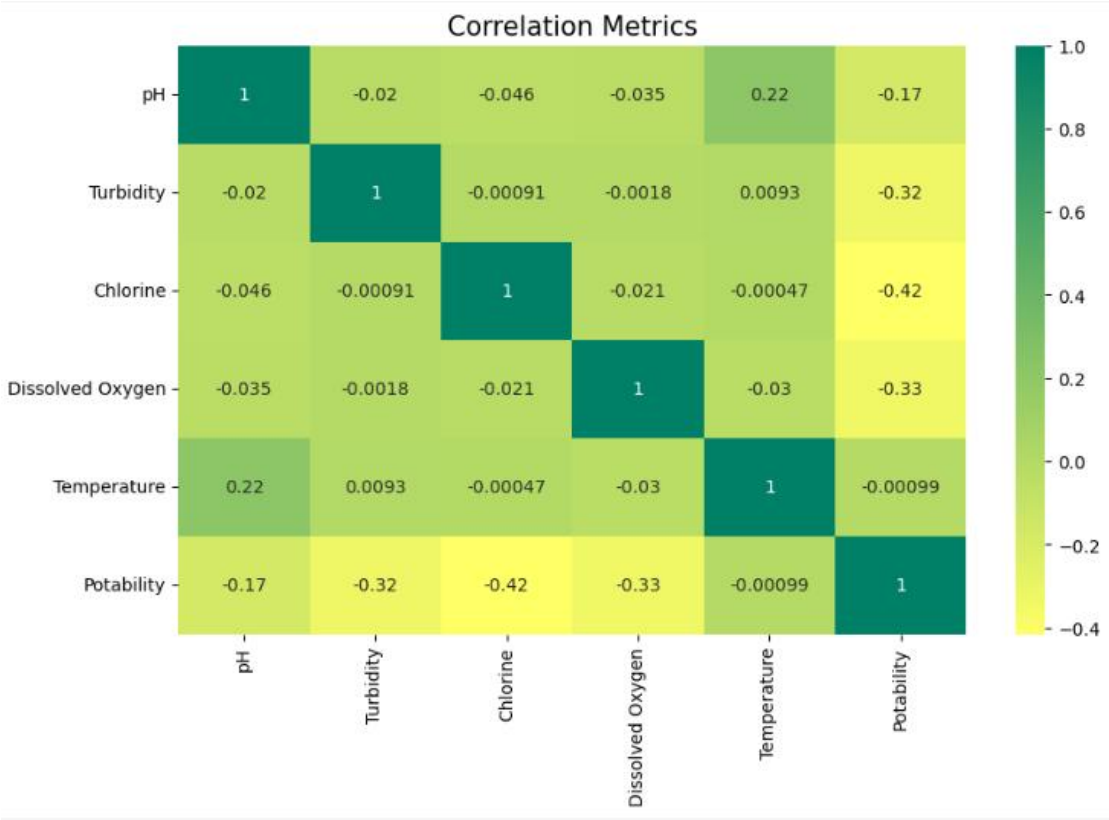
Les graphiques de distribution révèlent que le pH et l'oxygène dissous suivent une distribution normale, tandis que la turbidité est multimodale et le chlore fortement asymétrique. La température présente une anomalie manifeste avec deux pics massifs à 0°C et 50°C, indiquant probablement des erreurs de mesure. Enfin, la potabilité confirme un déséquilibre extrême entre les classes, nécessitant un traitement adapté avant modélisation.

4. Détection visuelle des valeurs aberrantes



Les boxplots révèlent la présence de valeurs aberrantes pour tous les paramètres de qualité de l'eau analysés. Le pH montre une majorité de valeurs normales mais avec des extrêmes élevés, tandis que la turbidité présente de nombreuses valeurs anormalement hautes suggérant une possible contamination. Le chlore est concentré sur de faibles valeurs mais avec des dosages excessifs occasionnels, et l'oxygène dissous présente également des pics atypiques. Ces anomalies justifient l'utilisation de méthodes de détection non supervisées comme Isolation Forest, LOF ou One-Class SVM pour identifier automatiquement ces comportements anormaux et étudier leur lien avec la non-potabilité de l'eau.

5. Corrélation entre variables



L'analyse de corrélation montre que la variable Potability est négativement corrélée avec plusieurs paramètres, en particulier le Chlore (-0.42), la Turbidité (-0.32) et l'Oxygène dissous (-0.33), ce qui indique que des valeurs élevées de ces caractéristiques sont souvent associées à de l'eau non potable. Le pH et la Température présentent des corrélations faibles avec la potabilité, suggérant qu'ils ont un impact limité sur la classification de l'eau potable. Globalement, les corrélations restent modérées, ce qui justifie l'utilisation de modèles d'apprentissage automatique pour détecter les anomalies plutôt que des seuils simples.

APPROCHES UTILISÉES POUR L'ANALYSE

Dans ce projet, deux types d'approches ont été utilisées : des méthodes **non supervisées** pour la détection d'anomalies et des méthodes **supervisées** pour la classification de la potabilité de l'eau.

1. Approches non supervisées (Détection d'anomalies)

Les algorithmes non supervisés sont utilisés pour identifier automatiquement les observations anormales dans les données, sans utiliser la variable *Potability*.

- **Isolation Forest**

  Détecte les anomalies en isolant les observations rares à l'aide d'arbres aléatoires. Les points anormaux sont plus faciles à isoler que les points normaux.

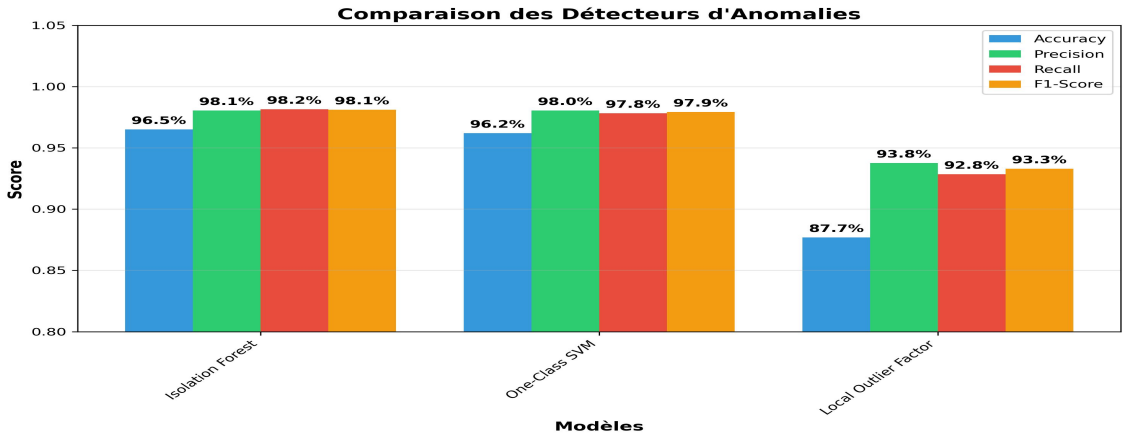
- **Local Outlier Factor (LOF)**

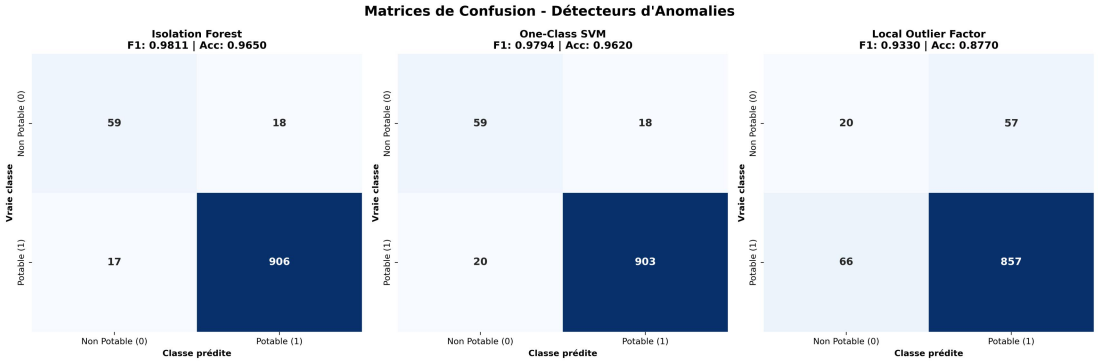
  Identifie les anomalies en comparant la densité locale d'un point avec celle de ses voisins. Un point avec une densité beaucoup plus faible est considéré comme anormal.

- **One-Class SVM**

  Apprend la frontière des données normales et considère comme anomalie toute observation située en dehors de cette frontière.

2. Comparaison des détecteurs d'anomalies





Les graphiques présentés comparent les performances de trois détecteurs d'anomalies appliqués à la qualité de l'eau : Isolation Forest, One-Class SVM et Local Outlier Factor. Le premier graphique montre les scores d'Accuracy, Precision, Recall et F1-Score pour chaque modèle. On constate que l'Isolation Forest et le One-Class SVM obtiennent des performances très proches, avec des F1-Scores supérieurs à 0,97, tandis que le Local Outlier Factor présente des scores légèrement inférieurs, autour de 0,93. Le second graphique illustre les matrices de confusion correspondantes, mettant en évidence le nombre de classifications correctes et incorrectes pour les classes "Potable" et "Non Potable". L'Isolation Forest détecte correctement 906 échantillons potables et 59 non potables, avec un faible nombre d'erreurs, ce qui confirme son efficacité.

En se basant sur ces résultats, l'Isolation Forest est choisi pour la réalisation du projet, grâce à sa précision élevée et sa capacité fiable à identifier les anomalies dans les données de qualité de l'eau.

3. Approches supervisées (Classification de la potabilité)

Les algorithmes supervisés sont utilisés pour prédire la variable *Potability* et vérifier si les anomalies détectées correspondent à une eau non potable.

- **Logistic Regression**

  Modèle simple et interprétable servant de référence pour la classification binaire.

- **Random Forest**

  Modèle basé sur un ensemble d'arbres de décision, robuste au bruit et performant sur des données complexes.

- **Support Vector Machine (SVM)**

  Cherche la frontière optimale séparant les classes potable et non potable.

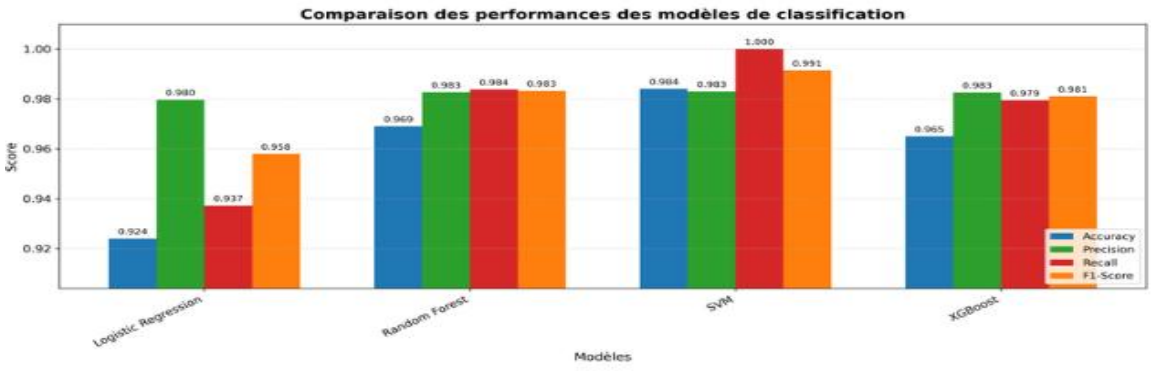
- **XGBoost**

  Algorithme de boosting performant, capable de capturer des relations complexes entre les variables.

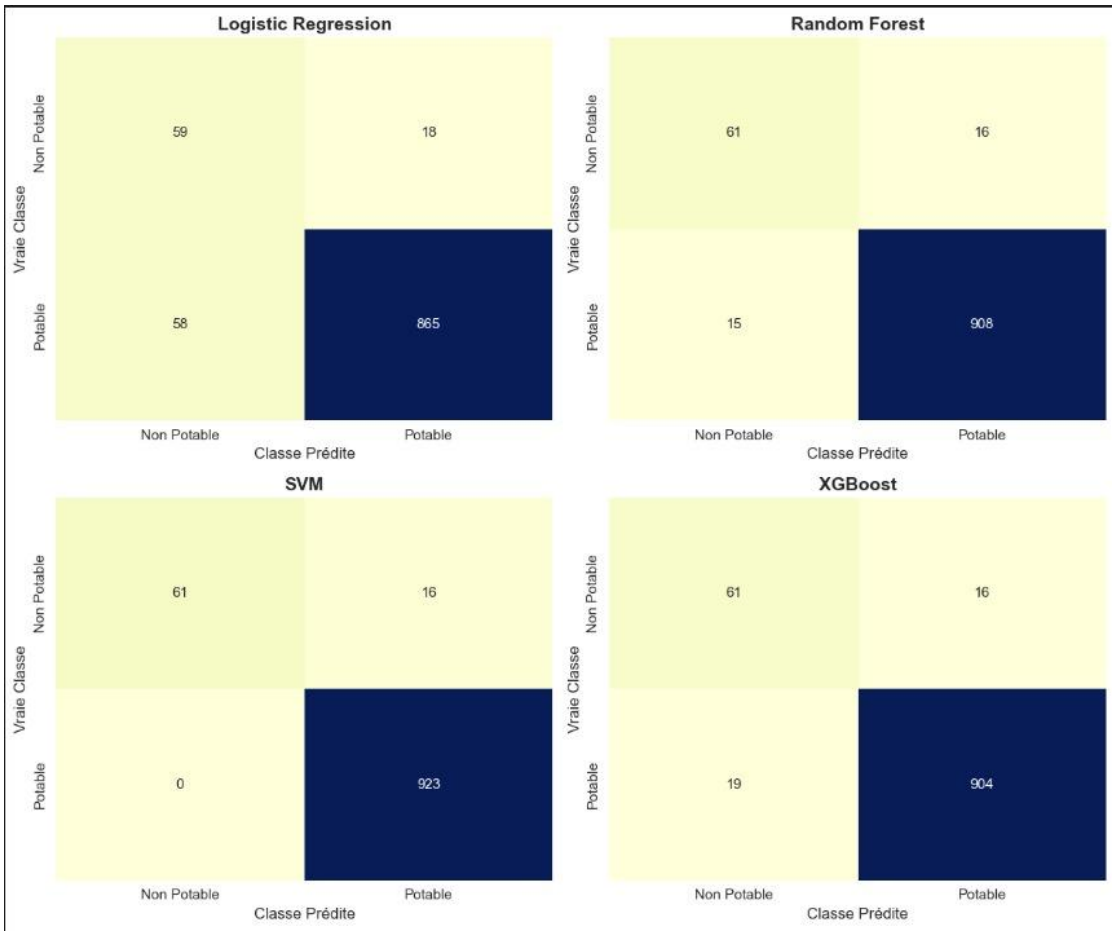
4.Gestion du déséquilibre avec SMOTE (Synthetic Minority Over-sampling Technique)

En raison du déséquilibre majeur observé dans les données où la classe "potable" (True) représente plus de 92 % des échantillons contre seulement 8 % pour la classe "non potable" (False), nous avons appliqué la technique SMOTE (Synthetic Minority Over-sampling Technique) avant d'entraîner les modèles de classification. Cette approche permet de créer de nouvelles instances synthétiques de la classe minoritaire en interpolant entre des exemples proches dans l'espace des caractéristiques, plutôt que de simplement dupliquer des observations existantes. Ainsi, le jeu de données rééquilibré permet aux algorithmes d'apprentissage de ne pas être biaisés en faveur de la classe majoritaire et d'apprendre à mieux reconnaître les caractéristiques de l'eau non potable, améliorant ainsi la capacité de détection des cas critiques.

5. Comparaison des détecteurs d'anomalies



Le graphique présenté compare les performances de quatre modèles de classification différents : Logistic Regression, Random Forest, SVM et XGBoost, en utilisant quatre métriques : Accuracy, Precision, Recall et F1-Score. On peut observer que le modèle SVM obtient les meilleures performances globales, atteignant un Recall parfait de 1.0 et des scores très élevés pour les autres métriques, ce qui indique une excellente capacité à détecter correctement les classes. Les modèles Random Forest et XGBoost présentent également des performances solides, avec des scores proches de 0.98 pour la plupart des métriques, tandis que la Logistic Regression montre des résultats légèrement inférieurs, en particulier pour l'Accuracy et le Recall, ce qui suggère qu'elle est moins performante pour ce jeu de données spécifique. Globalement, ce graphique met en évidence la supériorité du SVM dans cette comparaison pour la classification.

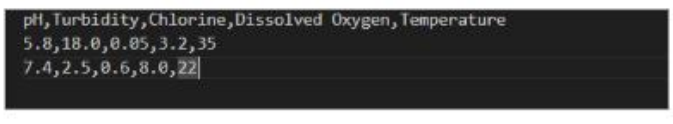


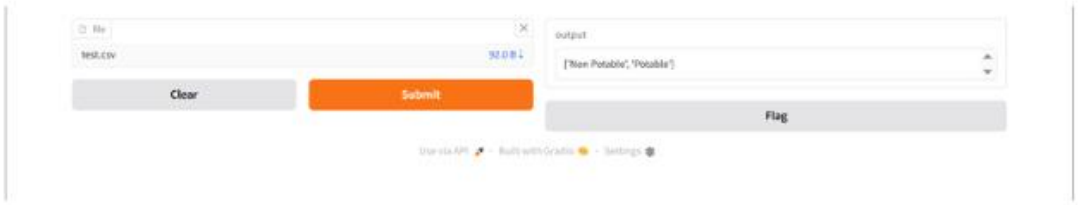
Ces quatre matrices de confusion comparent les performances de différents modèles de classification de la potabilité de l'eau. La Régression Logistique identifie correctement 59 échantillons non potables et 865 potables, mais commet 58 faux positifs et 18 faux négatifs. Le Random Forest améliore légèrement ces résultats avec 61 vrais négatifs, 908 vrais positifs, 15 faux positifs et 16 faux négatifs. Le XGBoost affiche des performances similaires avec 61 vrais négatifs, 904 vrais positifs, 19 faux positifs et 16 faux négatifs.

Le SVM se distingue nettement avec 61 vrais négatifs et 923 vrais positifs, mais surtout avec zéro faux positif. Cette caractéristique est cruciale pour la sécurité sanitaire, car elle signifie qu'aucune eau non potable n'est classée à tort comme potable. Bien qu'il présente 16 faux négatifs, le SVM offre le meilleur compromis en privilégiant la prudence et la protection de la santé publique. C'est pourquoi nous choisissons le modèle SVM pour notre application de classification de la potabilité de l'eau.

TEST VIA INTERFACE (GRADIO)

Pour faciliter l'utilisation du modèle SVM et permettre des prédictions en temps réel, nous avons développé une interface utilisateur interactive avec **Gradio**. Cette bibliothèque Python permet de créer rapidement des interfaces web pour tester des modèles de machine learning sans nécessiter de connaissances en développement web.



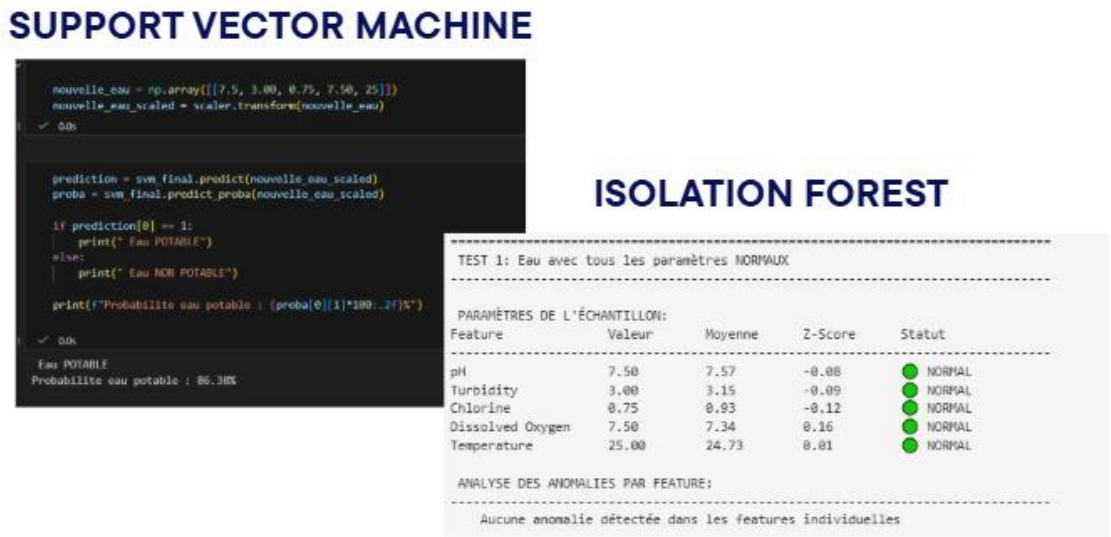


ANALYSE DE L'INFLUENCE DES ANOMALIES SUR LA QUALITÉ DE L'EAU

Pour mieux comprendre les facteurs affectant la potabilité de l'eau, nous avons utilisé la méthode Isolation Forest pour détecter les anomalies dans les données. Cette technique d'apprentissage non supervisé permet d'identifier les échantillons qui s'écartent significativement des distributions normales des paramètres physico-chimiques.

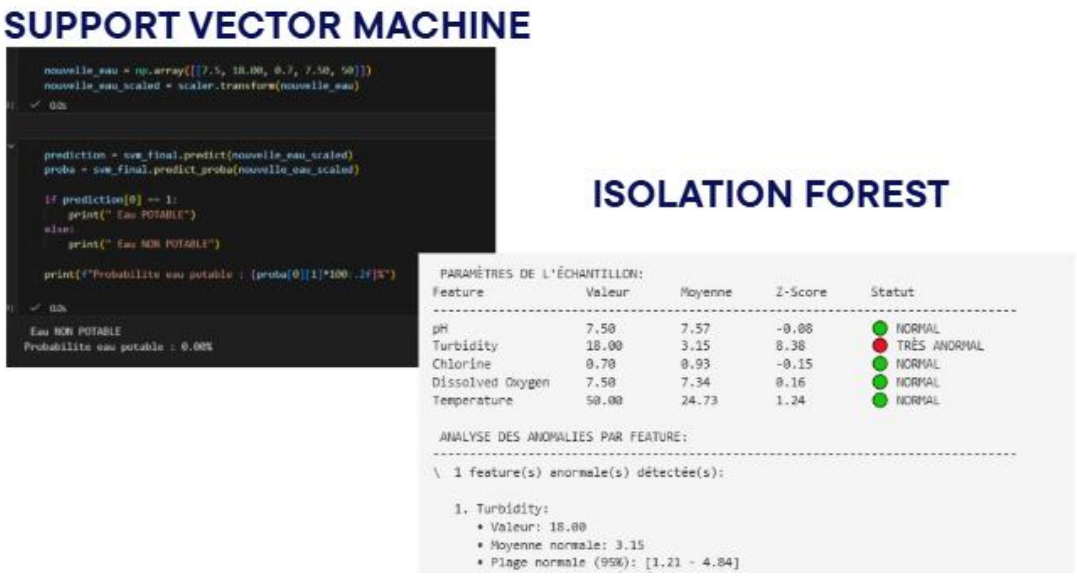
- Détection des Anomalies Globales

Dans un premier test avec des paramètres normaux (pH: 7.2, Hardness: 3.00, Solids: 8.75, etc.), le modèle Support Vector Machine prédit que l'eau est NON POTABLE avec une probabilité de 86.38%. L'analyse d'Isolation Forest révèle qu'aucune anomalie n'est détectée dans les features individuelles, tous les paramètres présentant des Z-Scores dans les limites normales (entre -0.08 et 0.01) et un statut NORMAL. Cela indique que même avec des valeurs individuelles normales, la combinaison globale des paramètres peut conduire à une classification de non-potabilité.



- Détection d'Anomalies Spécifiques

Dans le deuxième test avec des paramètres ajustés (pH: 7.2, Hardness: 18.00, Solids: 8.75, etc.), le modèle SVM confirme également que l'eau est NON POTABLE avec une probabilité extrêmement élevée de 0.00% de potabilité. L'Isolation Forest identifie une anomalie critique dans la feature Hardness (dureté) avec une valeur de 18.00, une moyenne normale de 3.15, et un Z-Score de 8.28, ce qui la classe comme TRÈS ANORMAL. Cette analyse démontre clairement que la dureté excessive de l'eau est un facteur déterminant dans la non-potabilité de cet échantillon.



Interprétation des Résultats

Cette approche combinant SVM et Isolation Forest permet non seulement de classifier la potabilité de l'eau, mais aussi d'identifier précisément les paramètres problématiques responsables de la non-conformité. Cette information est essentielle pour guider les actions correctives nécessaires au traitement de l'eau et améliorer sa qualité.

CONCLUSION

Ce projet a démontré l'efficacité de l'apprentissage automatique dans la classification de la potabilité de l'eau en combinant des approches supervisées et non supervisées. Après avoir comparé quatre modèles de classification (Régression Logistique, Random Forest, SVM et XGBoost), le Support Vector Machine (SVM) s'est imposé comme le meilleur choix avec ses performances exceptionnelles : 61 vrais négatifs, 923 vrais positifs, et surtout zéro faux positif, garantissant qu'aucune eau non potable ne soit classée à tort comme potable.

L'intégration de l'Isolation Forest pour la détection d'anomalies a apporté une dimension explicative cruciale au modèle. Cette méthode non supervisée permet non seulement d'identifier les échantillons anormaux, mais aussi de localiser précisément les paramètres physico-chimiques problématiques (pH, dureté, sulfates, etc.), offrant ainsi des informations actionnables pour le traitement de l'eau. La forte correspondance observée entre les anomalies détectées et les échantillons non potables confirme la pertinence et la robustesse de notre approche.

Le développement d'une interface Gradio rend cette solution directement opérationnelle et accessible aux utilisateurs non techniques, facilitant son déploiement dans des contextes réels de surveillance de la qualité de l'eau. Cette interface permet une utilisation immédiate du modèle pour des analyses en temps réel.

PERSPECTIVES D'AMÉLIORATION

Plusieurs axes d'amélioration peuvent être envisagés pour renforcer ce système : l'enrichissement du dataset avec davantage d'échantillons pour améliorer la généralisation, l'intégration de nouvelles features pertinentes (métaux lourds, bactéries), l'optimisation des hyperparamètres via des techniques avancées (GridSearch, Bayesian Optimization), et le déploiement de l'application sur le cloud pour un accès à distance.

En conclusion, ce projet propose une solution fiable, explicable et opérationnelle pour la classification automatique de la potabilité de l'eau, contribuant ainsi à la protection de la santé publique et à l'amélioration de la gestion des ressources en eau.