

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة الوطنية العليا للعلوم البحر وتهيئة الساحل

École Nationale Supérieure des Sciences de la Mer et de l'Aménagement du Littoral



Mémoire de fin d'études en vue de l'obtention du diplôme

D'ingénieur d'état

Domaine : Sciences et Technologie

Filière : Génie des procédés

Spécialité : Traitement et réutilisation des eaux usées

Thème :

**Vers une modélisation interprétable des stations d'épuration des eaux
usées : apport de XGBoost et SHAP
-cas de la STEP de la ville de Chlef-**

Présenter par :

BENNOUR SARRA

Soutenu le 17 juin 2026, devant le jury composé de :

Mme. GHEZALI K	MCB	ENSSMAL	Présidente
M. KELLOUCHE A	Dr	ONA	Encadrant
Mme. ATTIA N	MCA	ENSSMAL	Co-encadrante
Mme. BOUMEZBEUR M	MCB	ENSSMAL	Examinatrice

Année universitaire : 2025-2026

Remerciements

Avant tout, je remercie Dieu le Tout-Puissant de m'avoir donné de la force, la patience et la volonté pour mener ce travail à son terme.

Je tiens à exprimer ma profonde gratitude à mon encadrant Dr. KELLOUCHE Abdelhakim pour son soutien, sa disponibilité et ses précieux conseils tout au long de ce

mémoire. Merci de m'avoir accompagnée depuis le début, de m'avoir appris et guidée pas à pas, et d'avoir été présent dans chaque étape de ce travail.

Mes remerciements vont aussi à l'ingénieure de laboratoire ainsi qu'au personnel de la STEP de Chlef.

Je remercie également ma co-encadrante, Dr. ATTIA Nourhane, pour son accompagnement.

J'adresse mes remerciements aux membres du jury pour avoir accepté d'évaluer mon travail. Je tiens aussi à exprimer ma gratitude à l'ensemble des enseignants de l'ENSSMAL et à toutes les personnes, de près ou de loin, qui nous ont soutenus tout au long de cette aventure.

Dédicaces

Je dédie ce mémoire, avant tout, à moi-même. À celle qui a tenu bon quand tout devenait difficile, qui a continué à avancer même dans les moments de doute et de fatigue. À chaque effort silencieux, à chaque épreuve traversée avec patience, et à chaque fois où je suis tombée mais où j'ai choisi de me relever.

Ce travail est le reflet de tout ce chemin parcouru. Je suis fière d'avoir continué, d'avoir tenu jusqu'au bout, et d'être arrivée là où je suis aujourd'hui.

À ma mère Souad, la plus belle et la plus tendre des présences dans ma vie. Celle qui m'a toujours encouragée, qui a cru en moi même quand moi je doutais, et qui n'a jamais cessé de m'accompagner, par ses mots, son amour et ses douaa, et à tout ce qu'elle a donné sans jamais attendre en retour.

À mon père Abdelkader, le meilleur papa au monde, celui qui a toujours été présent pour moi, qui m'a encouragée à chaque étape et qui a rendu les choses plus simples quand elles devenaient difficiles. Celui qui a tout fait pour m'offrir les meilleures conditions afin que je puisse avancer et devenir la meilleure version de moi-même.

À mon frère Abderrahmane, mes sœurs Lamia et Marwa et ma belle-sœur Abir, merci du fond du cœur pour votre amour et votre présence constante à mes côtés. Vous avez toujours été là pour m'encourager, me soutenir et me donner la force d'avancer. Votre soutien m'a portée tout au long de ce parcours.

À mes étoiles Ghizlene, Ines et Wissem, mes chères nièces, ces petites âmes qui illuminent ma vie par leur présence et leur innocence.

À toute ma famille, merci pour votre soutien et votre présence à mes côtés.

À mes amies, Asma, Manel, Ibtihel, Wafa, Oumnia, Roro, Belkis, Abir, rencontrées au cours de mes études et devenues comme des sœurs pour moi. Merci pour tous les moments partagés.

Résumé

Afin d'optimiser le suivi des stations d'épuration des eaux usées, cette étude adopte une technique d'apprentissage automatique pour analyser les données physicochimiques des eaux usées de la STEP de Chlef. L'étude porte principalement sur la prédiction de paramètres importants tels que la DCO, la DBO₅ et les MES à partir des données d'exploitation de la station pour la période allant de 2011 à 2018.

L'algorithme XGBoost a été utilisé en raison de sa capacité à modéliser des relations complexes et non-linéaires entre les variables. Plusieurs configurations du modèle ont été comparées afin d'identifier celle offrant les meilleures performances prédictives.

De plus, la méthode SHAP a été utilisée pour interpréter les résultats du modèle en expliquant l'influence des paramètres sur les prédictions. Le couplage de XGBoost et SHAP permet de fournir un outil d'aide à la décision capable d'anticiper les dysfonctionnements et d'améliorer le pilotage de la STEP.

Mots-clés : station d'épuration, apprentissage automatique, analyse des données, prédiction, DCO, DBO, MES, dysfonctionnements, XGBoost, SHAP.

Abstract

To optimize the monitoring of wastewater treatment plants, this study employs a machine

learning technique to analyze the physicochemical data of wastewater from the Chlef WWTP.

The study focuses primarily on predicting key parameters such as COD, BOD, and suspended solids based on the plant's operational data for the period from 2011 to 2018.

The XGBoost algorithm was used due to its ability to model complex, non-linear relationships between variables. Several model configurations were compared to identify the one offering the best predictive performance.

Additionally, the SHAP method was used to interpret the model's results by explaining the

influence of parameters on the predictions. The combination of XGBoost and SHAP provides a decision-support tool capable of anticipating malfunctions and improving the management of the WWTP.

Keywords: wastewater treatment plant, machine learning, data analysis, prediction, COD, BOD, suspended solids, malfunctions, XGBoost, SHAP.

من اجل تحسين مراقبة محطة معالجة مياه الصرف الصحي، تعتمد هذه الدراسة تقنية التعلم الآلي لتحليل البيانات الفيزيائية والكيميائية لمياه الصرف الصحي في محطة معالجة مياه الصرف بمدينة الشلف. تركز الدراسة بشكل أساسي على التنبؤ بالخصائص الفيزيائية المهمة في المحطة استنادا الى بيانات تشغيل المحطة للفترة الممتدة من 2011 الى 2018.

تم استخدام خوارزمية XGBoost نظرا لقدرتها على نمذجة العلاقات المعقدة و غير الخطية بين المتغيرات. تمت مقارنة عدة تكوينات للنموذج من اجل تحديد التكوين الذي يقدم أفضل أداء تنبؤي.

بالإضافة الى ذلك، تم استخدام طريقة SHAP لتفسير نتائج النموذج من خلال شرح تأثيرات الخصائص الفيزيائية والكيميائية على التنبؤات. يتيح الجمع بين XGBoost SHAP توفير أداة للمساعدة في اتخاذ القرارات القادرة على توقع الأعطال وتحسين إدارة محطة معالجة مياه الصرف الصحي.

الكلمات المفتاحية : محطة معالجة مياه الصرف الصحي، التعلم الآلي، تحليل البيانات، التنبؤ، الطلب الكيميائي للأوكسجين (COD)، الطلب البيوكيميائي للأوكسجين (BOD)، المواد الصلبة العالقة، الأعطال، خوارزمية XGBoost، تفسير SHAP.

Sommaire

Introduction générale.....	1
I. Chapitre 1. Généralités sur les procédés d'épuration – Zoom sur la station d'épuration de Chlef.....	3
Introduction	4
I-1. Les types de procédé de traitement des eaux usées domestiques	4
I-1-1. Les procédés extensifs.....	4
I-1-2. Les procédés intensifs	5
I-2. Caractéristiques de la STEP de la ville de Chlef	9
I-2-1. Localisation et rôle de la STEP de la ville de Chlef.....	9
I-2-2. Filière eau	9
I-2-2-1. Prétraitement	10
I-2-2-2. Traitement biologique	10
I-2-2-3. Traitement secondaire	11
I-2-3. Filière boues	12
I-3. Les caractéristiques physico-chimiques des eaux usées.....	4
I-3-1. Paramètres physiques	7
I-3-2. Paramètres chimiques.....	7
I-3-3. Les normes de rejet des eaux usées	8
Conclusion.....	13
II. Chapitre 2. Matériels et méthodes	14
Introduction	14
II-1. Extreme Gradient Boosting (XGBoost).....	14
II-1-1. Choix de l'algorithme XGBoost.....	16
II-1-2. Domaines d'application de XGBoost	17
II-1-3. Limites de XGBoost	18
II-2. Outils d'interprétation des données analysées	18
II-2-1. Théorème SHAP (Shapley additive explanations)	19
II-2-2. Avantages et limites de SHAP (Shapley additive explanations).....	20
II-2-2-1. Les avantages	20
II-2-2-2. Les limites.....	21
II-3. L'application XGBoost & SHAP.....	21
II-3-1. Cadre d'étude de programme	21
II-3-2. Critères d'évaluation du modèle.....	24
II-4. Langage de développement et bibliothèques utilisées dans le programme	25
II-4-1. Environnement de développement	27
II-4-2. Langage Python	27
II-4-3. Bibliothèques	27
II-4-3-1. NumPy	27
II-4-3-2. Pandas	28
II-4-3-3. Matplotlib.....	28
II-4-3-4. Scikit-learn.....	29
II-4-3-5. Les bibliothèques XGBoost et SHAP	29
Conclusion.....	29
III. Chapitre 3. Résultats et discussion.	31
Introduction	31
III-1. Méthodologie appliquée.....	25
III-2. Résultats.....	31
III-2-1. Statistiques descriptives.....	31

III-2-2.	Interprétation des résultats du modèle XGBoost à l'aide de SHAP	35
III-2-3.	Analyse SHAP des eaux brutes	36
III-2-3-1.	DBO ₅	36
III-2-3-2.	DCO	39
III-2-3-3.	MES	41
III-2-4.	Analyse SHAP des eaux épurées.....	43
III-2-4-1.	DBO ₅	43
III-2-4-2.	DCO	45
III-2-4-3.	MES	47
III-2-5.	Analyse comparative des performances des configurations prédéfinies	35
III-3.	Discussion	50
III-3-1.	Impact du paramétrage du modèle XGBoost sur les performances	50
III-3-2.	Identification du modèle optimal.....	50
III-3-3.	Caractérisation statistique des eaux brutes	51
III-3-3-1.	Charge organique et variabilité	51
III-3-3-2.	Variabilité des matières en suspension	51
III-3-3-3.	Conditions physicochimiques	51
III-3-4.	Performance épuratoire des eaux traitées	51
III-3-5.	Interprétation SHAP des eaux brutes.....	51
III-3-5-1.	Prédiction de la DBO ₅	51
III-3-5-2.	Prédiction de la DCO	52
III-3-5-3.	Analyse de dépendance SHAP des MES	52
III-3-6.	Interprétation SHAP des eaux épurées	53
III-3-6-1.	Prédiction de la DBO ₅	53
III-3-6-2.	Prédiction de la DCO	53
III-3-6-3.	Analyse SHAP des MES	54
Conclusion.....		55
Conclusion générale		56

Liste des tableaux

Tableau 1 : Seuils d'acceptation des eaux usées.....	8
Tableau 2 : Normes de rejet des eaux usées	8
Tableau 3 : Tableau comparatif des algorithmes de gradient boosting.....	16
Tableau 4 : XGBoost dans différents domaines.....	17
Tableau 5 : Critères d'évaluation de XGBoost.....	24
Tableau 6 : Données – Exemple XGBoost	24
Tableau 7 : Calcul des erreurs – Exemple XGBoost.....	25
Tableau 8 : Résultats d'une étude précédente avec XGBoost	26
Tableau 9 : Statistiques descriptives des eaux brutes.....	32
Tableau 10 : Statistiques descriptives des eaux épurées	33
Tableau 11 : Tableau comparatif des résultats XGBoost avec configurations prédéfinies.....	35

Liste des figures

Figure 1 : schéma de station d'épuration à lit bactérien	5
Figure 2 : Schéma d'une station d'épuration à disques biologiques	6
Figure 3 : Schéma d'une station d'épuration à boues activées	6
Figure 4 : localisation géographique de la STEP de Chlef.....	9
Figure 5 : Dessableur et déshuileur de la STEP de Chlef	10
Figure 6 : bassin d'aération de la STEP de Chlef.....	11
Figure 7 :Bassin clarificateur de la STEP de Chlef	11
Figure 8 : l'épaississeur de la STEP de Chlef.....	12
Figure 9 :lit de séchage de la STEP de Chlef	13
Figure 10 : schéma de principe de XGBoost.....	15
Figure 11 : Exemple d'un diagramme SHAP.....	20
Figure 12 : Schéma des étapes de traitement des données	23
Figure 13 : L'interface de l'application VSCode	27
Figure 14 : Illustration de graphique avec Matplotlib	29
Figure 15 : Graphiques de distribution de la DBO ₅ et la DCO	34
Figure 16 : Graphique de synthèse SHAP de la DBO ₅ _eb (summary plot).....	36
Figure 17 : Graphique de dépendance SHAP	38
Figure 18 : Graphique de synthèse SHAP pour la DCO (summary plot)	39
Figure 19 : Graphique de dépendance SHAP de la DCO.....	40
Figure 20 : Graphique de synthèse (summary plot) SHAP des MES_ep.....	41
Figure 21 : Graphique de la dépendance SHAP	42
Figure 22 : Graphique de synthèse (summary plot) SHAP de la DBO ₅	43
Figure 23 : SHAP waterfall plot de la DBO ₅	44
Figure 24 : Graphique de synthèse (summary plot) SHAP de la DCO	45
Figure 25 : Graphique force plot SHAP de la DCO	46
Figure 26 : Graphique de synthèse (summary plot) SHAP des MES	47
Figure 27 : SHAP waterfall plot des MES_ep	49

Liste des abréviations

DBO ₅	Demande Biologique en Oxygène sur 5 jours
DCO	Demande Chimique en Oxygène
MES	Matières en Suspension
STEP	Station d'Épuration des Eaux Usées
R ²	Coefficient de détermination
RMSE	Root Mean Square Error (Erreur quadratique moyenne)
NSE	Nash-Sutcliffe Efficiency
XGBoost	eXtreme Gradient Boosting (algorithme d'apprentissage automatique)
SHAP	SHapley Additive exPlanations (méthode d'interprétation des modèles)

Introduction générale

La préservation des ressources en eau et la réduction de la pollution des milieux récepteurs constituent des enjeux environnementaux majeurs. Les eaux usées domestiques représentent une source importante de pollution organique et microbiologique, rendant leur traitement indispensable avant tout rejet dans le milieu naturel. (Bachi et al., 2022). En Algérie, cette problématique est encadrée par la réglementation nationale relative à la protection de l'environnement et à la qualité des rejets liquides, et s'inscrit dans la dynamique régionale de la Convention de Barcelone, qui vise à prévenir et à réduire la pollution marine d'origine terrestre. (*La Convention de Barcelone et ses Protocoles* | UNEPMAP, s. d.)

Dans ce cadre, la station d'épuration des eaux usées de la ville de Chlef, fonctionnant selon le procédé à boues activées à faible charge, représente une infrastructure essentielle au traitement des eaux usées urbaines. Les eaux épurées sont déversées dans l'oued Chéliff, exigeant un suivi strict de la qualité des effluents pour respecter les normes en vigueur et limiter l'impact sur le milieu récepteur.

Malgré son importance, la station d'épuration demeure un système complexe dont la modélisation est difficile. Les relations entre les paramètres physicochimiques sont dynamiques, non linéaires et influencées par de multiples facteurs. Cette complexité rend les modèles analytiques classiques insuffisants, d'où l'intérêt croissant pour les méthodes d'apprentissage automatique, capables de capter des relations complexes directement à partir des données d'exploitation.

L'algorithme XGBoost est retenu pour ses performances sur les données tabulaires complexes, sa haute précision prédictive et sa gestion efficace des non-linéarités. Afin de pallier le caractère « boîte noire » des modèles d'apprentissage automatique, la méthode SHAP, basée sur la théorie des jeux coopératifs de Shapley, permet de quantifier et de visualiser la contribution de chaque variable à chaque prédiction, aussi bien au niveau global que local. Le couplage XGBoost–SHAP constitue ainsi une approche pertinente pour l'analyse des données et l'aide à la décision opérationnelle.

Dans ce contexte, la question centrale de ce travail est : dans quelle mesure l'optimisation des configurations de l'algorithme XGBoost associée à l'interprétabilité SHAP permet-elle d'améliorer la modélisation du traitement des eaux usées de la STEP de Chlef et de fournir

aux exploitants des outils d'aide à la décision pour mieux piloter le procédé et garantir un bon fonctionnement de la STEP ?

Ce mémoire s'inscrit dans la continuité de travaux antérieurs appliquant XGBoost avec une configuration par défaut, conduisant à des performances insuffisantes. Six configurations prédéfinies de l'algorithme sont testées et comparées sur les données physicochimiques de la station (2011–2018) afin d'identifier la plus performante. La méthode SHAP complète cette démarche en révélant les variables les plus influentes, permettant aux exploitants d'anticiper les dérives de qualité et de garantir un effluent conforme aux normes. L'étude vise ainsi à transformer les résultats du modèle en décisions concrètes : ajustement des paramètres opérationnels, détection précoce des dysfonctionnements.

Ce travail est structuré en trois parties :

- Le premier chapitre présente les généralités sur le traitement des eaux usées, les caractéristiques de la STEP de Chlef, les paramètres physicochimiques étudiés et les normes de rejets applicables.
- Le deuxième chapitre présente le cadre théorique et méthodologique sur l'algorithme XGBoost, la méthode SHAP ;
- Le troisième chapitre présente et interprète les résultats obtenus : comparaison entre les configurations, les statistiques descriptives, l'analyse SHAP et discussion des résultats

*Généralités sur les procédés
d'épuration – Zoom sur la station
d'épuration de Chlef*

Introduction

Les eaux usées domestiques constituent une source de matières organiques (dissoutes ou particulières) qui contiennent du carbone, de l'azote et du phosphore. Lorsque ces nutriments atteignent des concentrations excessives, ils peuvent polluer le milieu dans lequel ils sont rejetés – les sols et les eaux de surface – et parfois, causer des dommages écologiques.(Morin-Crini et al., 2017)

Donc, il est essentiel d'intégrer un traitement approprié à la gestion des eaux usées. Cela implique que les eaux usées traitées peuvent être recyclées pour l'agriculture ou d'autres applications, tout en s'assurant qu'elles ne renferment pas de substances nuisibles. Le traitement des eaux résiduaires est crucial pour maintenir la pureté de l'eau et réduire son impact sur l'environnement et la santé humaine.(Sbaihia, 2023)

I-1. Les types de procédés de traitement des eaux usées domestiques

Le traitement des eaux usées repose sur de nombreux procédés, le choix du procédé de traitement dépend des caractéristiques des eaux résiduaires ainsi que de la superficie disponible.(Charfi, s. d.)

On distingue deux grandes catégories qu'on détaillera dans la suite.

I-1-1. Les procédés extensifs

Le lagunage naturel ou aéré, les filtres plantés et l'infiltration-percolation sur sable représentent des procédés extensifs de traitement des eaux usées.

Le lagunage naturel est un procédé d'épuration qui repose sur l'écoulement gravitaire de l'effluent à travers des bassins peu profonds, en présence des cultures libres de micro-organismes, notamment des algues qui produisent de l'oxygène par photosynthèse, en milieu ouvert.(Ruto et al., s. d.)

Les filtres plantés et l'infiltration-percolation reposent sur une biomasse fixée sur un support granulaire (sable, gravier), assurant le traitement par filtration et activité microbienne. Ces techniques ont été élaborées pour de petites et moyennes collectivités.(European Commission, 2001)

On désigne ces méthodes comme « extensives » car elles requièrent davantage d'espace et de temps de séjour, tout en consommant peu ou pas d'énergie.

I-1-2. Les procédés intensifs

Les procédés biologiques intensifs représentent les méthodes les plus avancées au niveau des stations d'épuration urbaines. On distingue trois procédés majeurs :

Lit bactérien : dans le lit bactérien, les micro-organismes se fixent sur un support, appelé garnissage, sous forme d'un biofilm. C'est une couche dense de bactéries qui peut produire des polymères, lui permettant de créer un film et de s'accrocher à un support. Le garnissage est alimenté avec l'eau usée à traiter suite à une décantation primaire ou un simple tamisage fin. Le passage de l'eau dans le système est très court, de l'ordre de quelques minutes. (Racault et Seguret, 2004)

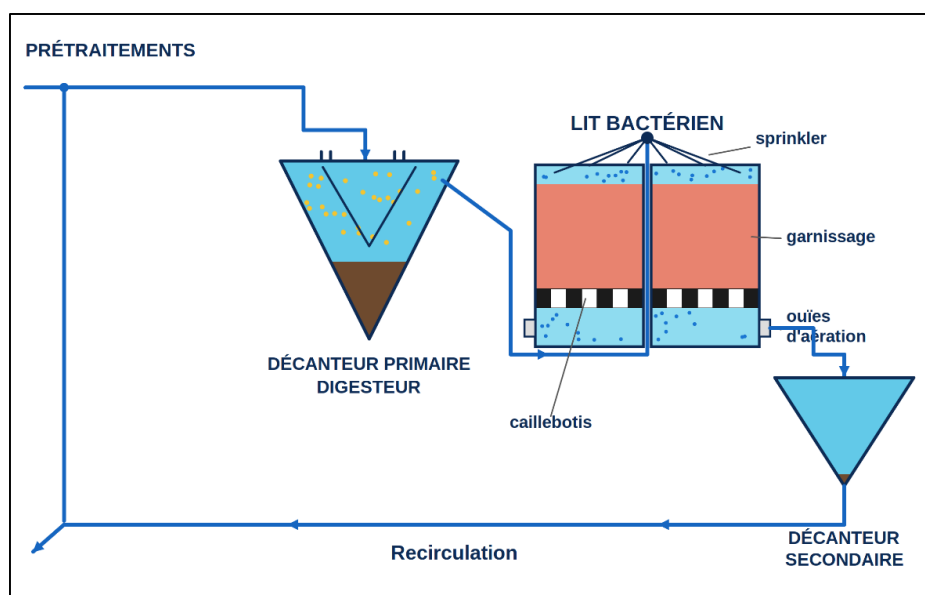


Figure 1 : schéma de station d'épuration à lit bactérien (*Traitement par le procédé des lits bactériens, 2021*)

Disques biologiques : le principe est basé sur le traitement par cultures microbiennes fixées sur des supports grossiers. La méthode classique comprend principalement un décanteur primaire, suivi de disques en rotation partiellement immergés, et enfin d'un clarificateur. (Durot et al., s. d.)

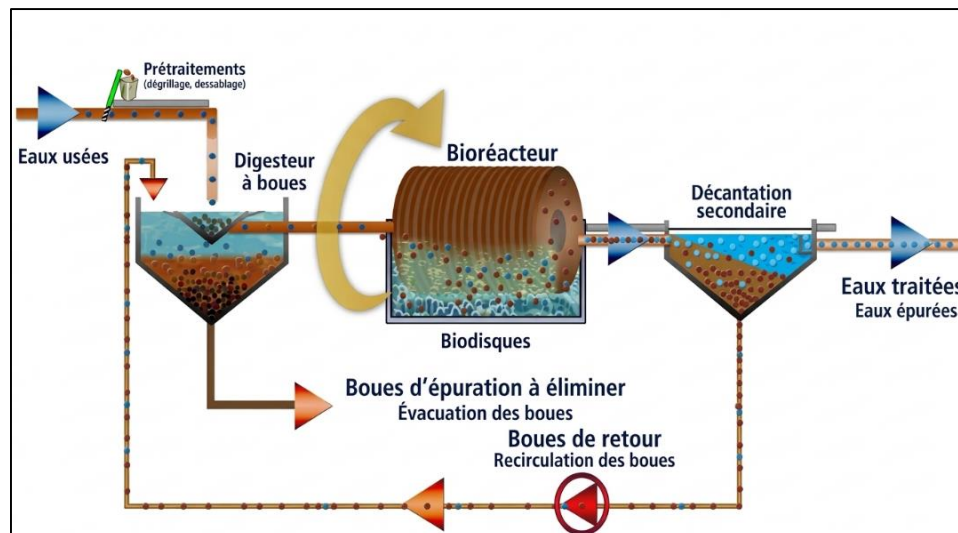


Figure 2 : Schéma d'une station d'épuration à disques biologiques (*Siden.lu - Biodisques, s. d.*)

Boues activées : les procédés à boues activées (PBA) représentent l'une des techniques les plus adoptées dans les stations d'épuration urbaines. Il s'agit de procédés biologiques fonctionnant en continu afin de produire un effluent conforme aux normes de rejet en vigueur. Le traitement repose sur deux étapes fondamentales :

- La biodégradation de la matière polluante par des microorganismes cultivés sous forme de flocs au sein d'un bassin d'aération.
- La séparation de la biomasse du liquide traité par décantation. (Oliveira et al., 2017)

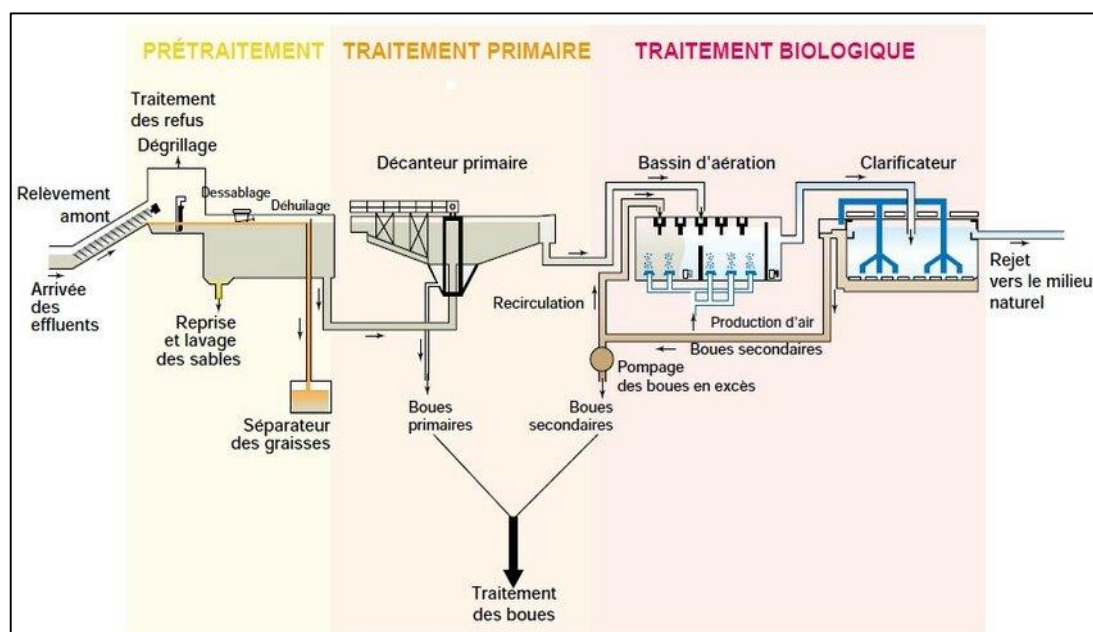


Figure 3: Schéma d'une station d'épuration à boues activées (*Zehraoui, 2011*)

I-2. Les caractéristiques physico-chimiques des eaux usées

Les paramètres physico-chimiques constituent des facteurs d'information essentiels pour l'évaluation de la qualité des eaux usées.(Morin-Crini et al., 2017)

En effet, la demande chimique en oxygène (DCO), la demande biochimique en oxygène sur cinq jours (DBO₅) et les matières en suspension totales (MES) sont des paramètres clés des polluants organiques notamment les macro-polluants présents dans l'environnement. Par ailleurs, des facteurs tels que le pH, la température, l'oxygène dissous et la conductivité influencent les micro-organismes responsables de la biodégradation de la matière organique. Par conséquent, afin de prévenir les impacts environnementaux néfastes des eaux usées, il est essentiel de connaître leur composition en polluants organiques à l'entrée de la station d'épuration. (M Sawadogo, B Gérard, P Windinpsidi, A Bere, 2023)

I-2-1. Paramètres physiques

- **La Température** est un paramètre important pour évaluer la qualité de l'eau. Elle influe sur de nombreux phénomènes, notamment la vitesse des réactions chimiques en milieu aquatique, la diminution de la solubilité des gaz et l'amélioration du goût et de la couleur de l'eau. (M Yasin, T Ketema, K Bacha, 2015)
- **Les matières en suspension (MES)** correspondent aux particules solides non dissoutes présentes dans l'eau. La connaissance de la concentration des éléments colloïdaux dans les eaux usées est essentielle pour évaluer l'impact de la pollution sur le milieu aquatique.(Ayyach A.1 et al., 2016)

I-2-2. Paramètres chimiques

- **Demande chimique en oxygène (DCO)** fournit une estimation des quantités de substances susceptibles d'être oxydées chimiquement présentes dans l'eau. (Morin-Crini et al., 2017)
- **Demande biochimique en oxygène (DBO)** représente la quantité d'oxygène consommée par les micro-organismes pour dégrader la matière organique biodégradable pendant 5 jours.
- **Potentiel d'hydrogène (pH)** indique le degré d'acidité ou d'alcalinité de l'eau.

- **Ammonium (NH_4^+)** c'est une forme réduite de l'azote, il peut être toxique à forte concentration.(Morin-Crini et al., 2017)
- **Nitrates (NO_3^-)** le nitrate est un composé chimique naturel qui s'intègre au cycle de l'azote.(Sbailhia, 2023)
- **Nitrites (NO_2^-)** sont des composés intermédiaires issus du processus de nitrification.

I-2-3. Les normes de rejet des eaux usées

Les seuils d'acceptation des eaux brutes, sont fixés dans le décret exclusif n° 09-209 du 11 juin 2009 (*Décret exclusif n° 09-209*, 2009).

Tableau 1 : Seuils d'acceptation des eaux usées

Paramètre	Seuil
Température (°C)	<30
pH	5,5-8,5
DCO (mg/l)	1000
DBO (mg/l)	500
MES (mg/l)	600
Azote total (mg/l)	150
Nitrites (mg/l)	0,1

Les normes de rejet des eaux épurées, dans le milieu naturel, en Algérie sont fixées par le Décret exclusif n° 06-141 du 19 avril 2006. (Aouel, 2006)

Tableau 2 : Normes de rejet des eaux usées (Normes Algériennes)

Paramètre	Norme
T (°C)	30
pH	Entre 6,5 et 8,5
DBO (mg /l)	35
DCO (mg /l)	120
MES (mg /l)	35
Azote	30
Couleur	Incolore
Odeur	Inodore

I-3. Caractéristiques de la STEP de la ville de Chlef

I-3-1. Localisation et rôle de la STEP de la ville de Chlef

La station d'épuration est implantée à l'ouest de la wilaya de Chlef, au niveau de la commune de Chlef. Elle a été mise en service en 2006, le procédé de traitement est à boues activées à faible charge. La capacité de cette station est de 227 528 EH et un débit journalier de 36 504 m³/j. (Kellouche, 2017)



Figure 4 : localisation géographique de la STEP de Chlef

Cette installation, dont l'objectif est de préserver l'environnement, garantit le traitement des eaux usées domestiques de la ville de Chlef. Les eaux résiduaires traitées sont par la suite déversées dans l'Oued Chéliff, qui est le milieu récepteur. (Kellouche, 2017)

Actuellement, la station d'épuration est en cours de réhabilitation pour une durée de 16 mois, sous la supervision de l'organisme de contrôle technique CTH et de l'office national d'assainissement, afin d'optimiser ses performances d'épuration et de garantir un meilleur respect des normes de rejet. Un autre projet est en train de se concrétiser, d'une durée de 18 mois, comptant la réalisation d'un traitement tertiaire en UV pour une réutilisation des eaux épurées dans l'arboriculture.

La STEP de Chlef comprend deux filières de traitement : filière eau et filière boues.

I-3-2. Filière eau

Cette filière est constituée des étapes suivantes :

I-3-2-1. Prétraitement

les prétraitements ont pour objectif d'éliminer de l'eau usée les substances qui pourraient dégrader la qualité du traitement aval ou la durabilité des installations.(RENOU, 2006)

Le dégrillage : la station est équipée de deux dégrilleurs grossiers et deux dégrilleurs fins, pour éliminer les déchets volumineux à l'aide des grilles.

Le dessablage-déshuilage : le dessableur et le déshuileur permettent respectivement de décanter les particules de sable, qui seront ensuite retirées du fond, et de faire remonter les huiles et les graisses à la surface, lesquelles sont ensuite raclées à l'aide d'un racleur de surface.



Figure 5 : Dessableur et déshuileur de la STEP de Chlef

I-3-2-2. Traitement biologique

Après le prétraitement, les eaux usées sont dirigées vers le bassin d'aération, où elles subissent un traitement biologique par boues activées. Au sein de cet ouvrage, un brassage et une aération continus ou intermittents sont essentiels à l'activité des micro-organismes épuratoires (Benyahya, s. d.) afin d'éliminer la pollution carbonée, phosphorée et azotée.

Le bassin biologique de la station est constitué de 04 bassins d'aération équipés d'aérateurs qui assurent l'apport en oxygène, les bassins présentent une structure rectangulaire en forme de serpent permettant d'augmenter le temps de séjour, la stabilisation et d'améliorer le contact entre la biomasse et la matière organique à dégrader.



Figure 6: bassin d'aération de la STEP de Chlef

I-3-2-3. Traitement secondaire

À la suite du traitement biologique, le mélange eau-boues est acheminé vers le bassin clarificateur pour assurer la séparation des boues par décantation gravitaire et l'obtention d'un effluent clarifié en surface.

Le clarificateur permet d'améliorer l'efficacité de traitement en réduisant la concentration des matières en suspension (MES).(Combe, 2016)



Figure 7 :Bassin clarificateur de la STEP de Chlef

La station fait actuellement l'objet de travaux de réhabilitation visant l'intégration d'un traitement tertiaire, en vue d'assurer la collecte des eaux usées traitées et leur transfert vers le périmètre d'irrigation moyen Cheliff, et ce pour garantir une meilleure valorisation des ressources en eau.

I-3-3. Filière boues

Les boues extraites depuis le traitement biologique, et séparées dans le clarificateur, sont dirigées vers la filière de traitement des boues afin de subir un traitement approprié. Cette dernière se compose de l'épaississement qui se fait par une décantation pour concentrer les solides.



Figure 8 : l'épaississeur de la STEP de Chlef

Ensuite la déshydratation qui se fait sur les lits de séchage, pour réduire le volume et faciliter la valorisation des boues. (*Epaississement et déshydratation des boues* | SUEZ, s. d.)



Figure 9 :lit de séchage de la STEP de Chlef

Conclusion

Ce chapitre a présenté les procédés de traitement des eaux usées extensifs et intensifs, en mettant en évidence leurs principes de fonctionnement.

Les caractéristiques de la station d'épuration de Chlef ont été présentées, en décrivant les différentes étapes de son traitement, allant du prétraitement à la gestion des boues, tout en soulignant l'importance des paramètres physico-chimiques pour le contrôle de la qualité de l'effluent, ainsi que le respect des seuils et des normes de rejets applicables.

Chapitre 2. Matériels et méthodes

Introduction

Les stations d'épuration, notamment celles utilisant le procédé à boues activées, rencontrent divers dysfonctionnements susceptibles de réduire la fiabilité de leurs performances.(Ghellai & Mohammedi, 2018)

Face à ces problèmes, il devient nécessaire de mettre en place des approches permettant de détecter précocement les anomalies afin d'améliorer l'efficacité du fonctionnement de ces installations. Dans ce contexte, les techniques d'apprentissage automatique mettent en place des modèles de prédiction simples comme la régression linéaire et logistique qui sont faciles à comprendre, mais ils sont souvent incapables de modéliser les relations complexes dans les données tabulaires. C'est pourquoi les algorithmes de gradient boosting tels qu'Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGBM) et Categorical Boosting (Catboost) ont été développés avec une haute précision de prédiction pour faciliter l'analyse des données d'exploitation et anticiper les dysfonctionnements.(C. Albertyn, 2025)

Ainsi, ce travail se focalise sur l'utilisation de l'algorithme XGBoost pour la modélisation et l'analyse des données.

II-1. Extreme Gradient Boosting (XGBoost)

Lancé par Chen et Guestrin en 2016, l'algorithme XGBoost est une méthode d'apprentissage automatique (Machine Learning en anglais) supervisée, basée sur l'implantation d'arbres de décision à gradient boosté (voir figure 1). Il est utilisé pour la régression, la classification et la détection des anomalies. À un niveau élevé, XGBoost combine séquentiellement des apprenants faibles (arbres de décision), chaque nouvel arbre corrigeant les erreurs du précédent.(M Wiens , Verone-Boyle , Henscheid , T Podichetty , Burton, 2025)

L'équation mathématique suivie par l'algorithme est donnée comme suit (Pan, 2018) :

$$\text{Fonction objective totale} = \sum_n^N l(y_i, \hat{y}_i) + \sum_k^K \Omega(f_k).$$

Avec :

n : nombre d'observations (échantillons).

K : nombre d'arbres.

$l(y_i, \hat{y}_i)$: fonction de perte entre la valeur réelle y_i et la valeur prédite $\hat{y}_i$.

$\Omega(f_k)$: terme de régularisation pour limiter la complexité de l'arbre k et pour éviter le problème de surapprentissage (overfitting).

Les performances de ce modèle dépendent fortement de la configuration des hyperparamètres, qui contrôlent la complexité du modèle et le processus d'apprentissage. (Putatunda & Rama, 2018)

Tels que :

- **Learning_rate** : contrôle la contribution de chaque nouvel arbre au modèle final.
- **n_estimators** : le nombre total d'arbres de décision construits dans le modèle.
- **max_depth** : la profondeur maximale des arbres de décision.

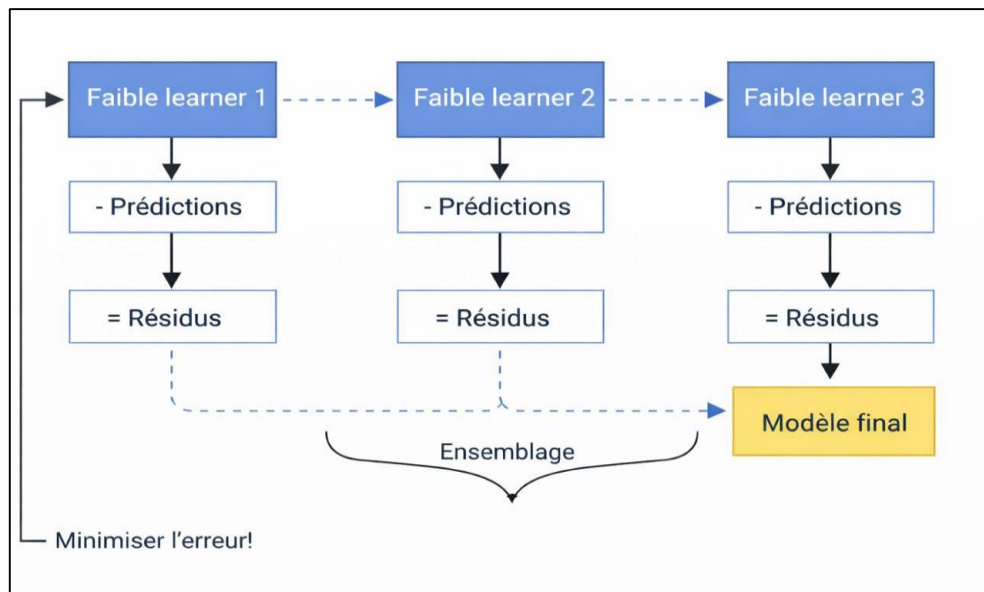


Figure 10 : schéma de principe de XGBoost

II-1-1. Choix de l'algorithme XGBoost

XGBoost offre plusieurs critères liés à sa performance et sa capacité à modéliser des relations complexes, qui rendent son choix évident pour cette étude.

- **Extensibilité** importante, cela signifie que le système a une capacité à s'adapter à une augmentation de la demande tout en conservant ses performances, l'algorithme fonctionne plus de dix fois plus vite que les algorithmes de prédiction connus.(Niazkar et al., 2024)
- **Évolutivité** (scalability), XGBoost est capable de traiter des milliards d'exemples en utilisant moins de ressources, il peut gérer des ensembles de données massifs même sur une seule machine grâce au calcul hors mémoire.(Chen & Guestrin, 2016)
- **Robustesse**, l'algorithme montre une capacité à capturer les relations complexes et non linéaires entre les variables, ce qui le rend robuste par rapport à d'autres modèles simples.(Niazkar et al., 2024)

En résumé, XGBoost est un modèle très puissant et solide en pratique, notamment dans le domaine des ressources en eau.

Le tableau comparatif suivant adapté à une étude, illustre les principales différences entre XGBoost et les autres algorithmes de gradient boosting (Rivaldo et al., 2025) :

Tableau 3 : Tableau comparatif des algorithmes de gradient boosting

Caractéristiques	XGBoost	LightGBM	CatBoost
Stratégie de croissance	Croissance par niveau (level-wise)	Par feuilles (leaf-wise)	Arbre de décision symétrique (oblivious trees)
Temps d'entraînement	Modéré	Très rapide	Lent
Prétraitement des données	Nécessite un prétraitement des données	Discrétisation en bacs	Gère les variables catégorielle sans prétraitement
Exactitude (précision)	Haute précision	Précision moyenne	Précision stable
Points forts	Excellente généralisation et stabilité	Rapide	Robustesse face aux données catégorielles

Le tableau précédent montre que l'algorithme XGBoost a marqué une haute précision prédictive et une capacité à éviter le surapprentissage (overfitting) grâce au terme de régularisation, ce qui en fait le modèle le plus performant globalement dans l'étude. LightGBM, est plus rapide à entraîner en gardant une précision correcte. CatBoost, il se distingue par sa capacité à capturer les interactions entre les caractéristiques.

Dans l'ensemble, XGBoost apparaît comme le modèle le plus fiable pour obtenir des résultats précis.

II-1-2. Domaines d'application de XGBoost

L'algorithme est largement utilisé dans différents domaines. Le tableau ci-dessous montre la diversité des applications de XGBoost dans différents domaines.

Tableau 4 : XGBoost dans différents domaines

Domaine d'application	Exemple	Références
Agriculture	Classification de type de sol.	(Ndiaye et al., 2025)
Médecine	Prédiction de la douleur.	(Cao et al., 2026)
Urbanisation	Prévision des inondations.	(Zhou et al., 2022)
Ressources en eau	• Prédire la DBO₅. • Étudier la possibilité de réutiliser les eaux usées en irrigation.	(Ching et al., 2022) (S Hasnaoui, C Hamadou, 2025)

Agriculture : dans le secteur agricole, XGBoost est utilisé pour la classification des types de sols en fonction de leurs caractéristiques physico-chimiques.(Ndiaye et al., 2025)

Domaine médical : le modèle est appliqué pour la prédiction précoce de la douleur modérée à sévère après un traitement chez les patients atteints d'un cancer, basée sur des données provenant de plusieurs centres médicaux.(Cao et al., 2026)

Urbanisation : Un modèle de prévision des inondations urbaines a été développé à l'aide de l'algorithme XGBoost ; le modèle offre une grande précision dans la prévision des pics de ruissellement.(Zhou et al., 2022)

Ressource en eau : pour une gestion efficace des ressources en eau, il est essentiel d'évaluer et de prévoir la qualité des eaux de surface, afin d'assurer que les niveaux de pollution restent conformes aux normes autorisées.(Najah Ahmed et al., 2019)

Dans ce contexte, au sein des stations d'épuration des eaux résiduaires, l'analyse et la prédiction des paramètres de fonctionnement constituent un enjeu majeur pour garantir l'efficacité du traitement et prévenir les anomalies opérationnelles. De nombreuses études ont été menées dans ce domaine :

- Des capteurs logiciels ont été développés dans une étude pour prédire la DBO5 en utilisant XGBoost, cette nouvelle méthode a été appliquée à deux cas pour évaluer sa robustesse. L'algorithme a évalué une grande diversité de valeurs de DBO5 dans les deux cas, montrant des performances cohérentes sur différents ensembles de tests. (Ching et al., 2022)
- De plus, dans une étude précédente XGBoost a été comparé à d'autres algorithmes d'intelligence artificielle pour la modélisation des données physico-chimiques de la STEP de Chlef, dans le but de déterminer la possibilité d'utiliser l'eau traitée pour l'irrigation. (S Hasnaoui, C Hamadou, 2025)

II-1-3. Limites de XGBoost

Comme tout modèle XGBoost présente certaines limites :

- La capacité de XGBoost à généraliser les résultats obtenus est influencée par la qualité et la quantité des données disponibles.
- XGBoost est incapable d'exprimer une équation mathématique en raison de sa structure en arbres de décision.
- L'algorithme nécessite de configurer divers hyperparamètres pour obtenir des résultats optimaux, ce qui rend le temps de calcul plus lent.(Niazkar et al., 2024)

II-2. Outils d'interprétation des données analysées

L'interprétation des données correspond à la capacité d'un analyste à comprendre les décisions d'un modèle, ses entrées, son architecture, ses sorties, ainsi que les relations directes ou les chemins entre ces éléments qui indiquent comment un modèle parvient à une décision ou à son raisonnement. Plusieurs méthodes d'interprétation telles que Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive Explanations

(SHAP), Partial Dependence Plots (PDP) et saliency maps ont été proposées afin de mieux comprendre les modèles de machine learning. (*Interpretable Machine Learning - an overview* | *ScienceDirect Topics*, 2023)

SHAP apparaît comme l'une des méthodes les plus efficaces pour expliquer les modèles complexes tels que XGBoost, en permettant d'analyser la contribution de chaque variable aux prédictions du modèle.

II-2-1. Théorème SHAP (Shapley additive explanations)

L'analyse SHAP est une méthode qui fournit des interprétations globales, en identifiant les variables les plus influentes et locales en expliquant une prédiction spécifique du modèle en quantifiant la contribution de chaque variable, généralement représentée sous forme de graphiques explicatifs tels que summary plot, dependence plot et force plot. Le but est de déterminer comment les modèles font la prédiction et quelles caractéristiques sont les plus significatives pour orienter la prédiction. (Li, 2022)

Grâce à cette approche, il devient plus facile de ne pas considérer le modèle XGBoost comme une boîte noire (black box), mais plutôt comme un outil compréhensible.

Mathématiquement, la valeur SHAP est donnée par (Li, 2022) :

$$\phi_j(X_j) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!} [f_{S \cup j}(X_{S \cup j}) - f_S(X_S)]$$

Avec :

- $(N = 1, \dots, p)$: ensemble des variables
- (S) : sous-ensemble sans variable (j)
- $(f_S(X_S))$: prédiction conditionnelle uniquement aux variables dans (S) .

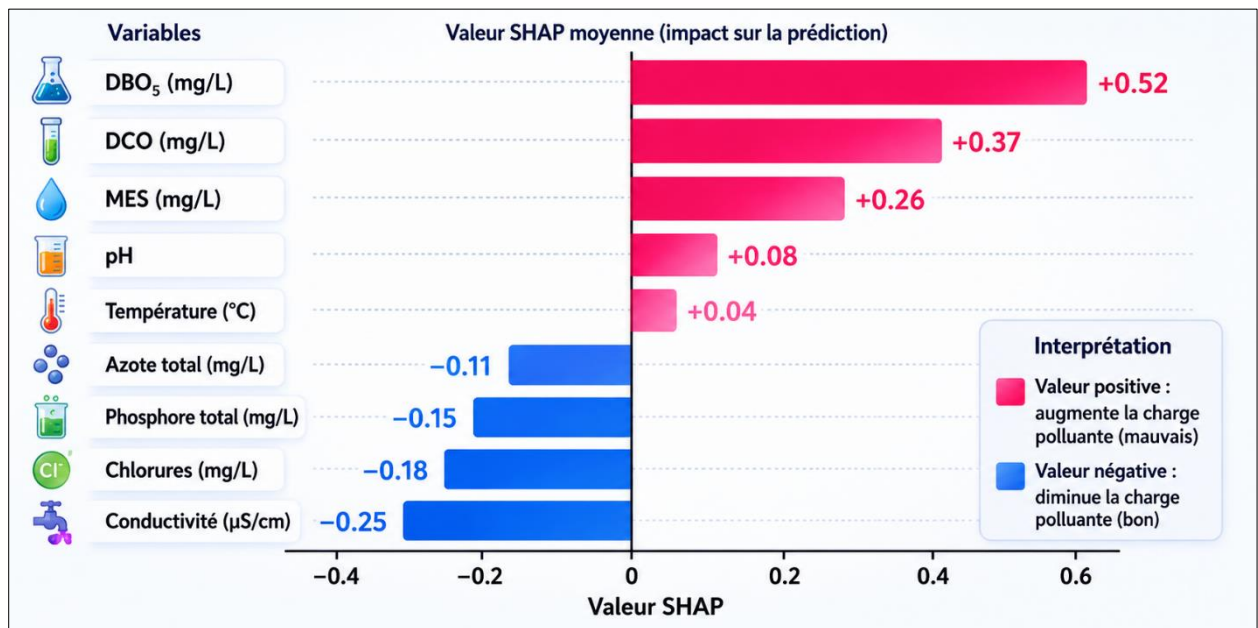


Figure 11: Exemple d'un diagramme SHAP

Ce diagramme montre la contribution des variables aux prédictions. Cette approche trouve son origine dans les valeurs de Shapley, un concept issu de la théorie des jeux coopératifs. Les valeurs de Shapley permettent une répartition équitable des gains entre les joueurs d'un jeu coopératif dans lequel ces derniers collaborent pour atteindre un objectif commun, même si leurs contributions peuvent être inégales. (Ponce-Bobadilla et al., 2024)

II-2-2. Avantages et limites de SHAP (Shapley additive explanations)

Comme tout outil SHAP n'échappe pas à la règle, il a des avantages et des limites dont il faut tenir compte.

II-2-2-1. Les avantages

- Fournir des interprétations locales et globales : permet d'obtenir des interprétations locales expliquant une prédiction spécifique en montrant la contribution de chaque variable grâce à des graphes tels que force plot et globales pour comprendre le comportement général du modèle à travers des visualisations comme dependence plot.
- Exactitude locale (local accuracy) : cela signifie que la somme des contributions de toutes les variables correspond à la prédiction du modèle, ce qui permet d'expliquer chaque prédiction individuelle. Dans la figure précédente, chaque point montre

l'effet d'une variable sur la prédiction (positif ou négatif). Ces contributions expliquent directement la valeur finale prédite par le modèle.

- Garantit une interprétation cohérente et mathématiquement justifiée des résultats basée sur la théorie des jeux coopératifs.(Lundberg & Lee, 2017)

II-2-2-2. Les limites

- L'analyse SHAP repose entièrement sur le modèle XGBoost sous-jacent. Par conséquent, elle hérite des biais éventuels du modèle, cette dépendance peut engendrer des interprétations erronées des relations entre les variables.(Takefuji, 2025)
- Le calcul exact des valeurs de SHAP est très coûteux en termes de calculs, car il nécessite d'analyser un grand nombre de combinaisons des variables, ce qui devient difficile lorsque le nombre de variables augmente.

II-3. L'application XGBoost & SHAP

II-3-1. Cadre d'étude de programme

Le programme développé dans ce travail se base sur la modélisation et l'analyse des données de la station d'épuration des eaux usées STEP de Chlef. Il a été réalisé par le Dr A. Kellouche. L'objectif est d'élaborer un pipeline automatisé permettant de traiter ces données, d'entraîner un modèle de prédiction et d'interpréter les résultats obtenus à l'aide de SHAP.(A.Kellouche, 2026)

Les données physico-chimiques des eaux usées de la STEP de Chlef sont d'abord traitées: chargement depuis Excel, nettoyage des colonnes, conversion numérique, remplacement des valeurs manquantes par la médiane, détection des outliers. Enfin, les données sont organisées et séparées.

Après le prétraitement les données sont utilisées pour entraîner un modèle d'apprentissage automatique basé sur l'algorithme XGBoost qui permet de capturer des relations complexes et non linéaires entre les variables, pour chaque paramètre cible (DCO, DBO, MES), l'algorithme construit un modèle pour prédire sa valeur à partir des variables explicatives.

Le modèle est dit performant selon des indicateurs statistiques notamment le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE), ces indicateurs permettent d'avoir une idée sur la qualité des prédictions.

Afin de mieux comprendre les résultats obtenus par le modèle XGBoost, la méthode SHAP a été utilisée, elle permet d'expliquer l'influence de chaque variable sur les prédictions du modèle et à savoir quelles variables jouent le rôle le plus important et comment elles contribuent à augmenter ou diminuer la valeur prédite.

L'utilisation de SHAP offre une meilleure lecture des résultats obtenus par XGBoost et permet de les relier logiquement aux phénomènes physico-chimiques observés au niveau de la STEP.

Le schéma ci-dessous illustre les différentes étapes du pipeline de traitement et la modélisation des données, du chargement des données jusqu'à l'interprétation des résultats pour présenter l'enchaînement des opérations réalisées au cours du processus.

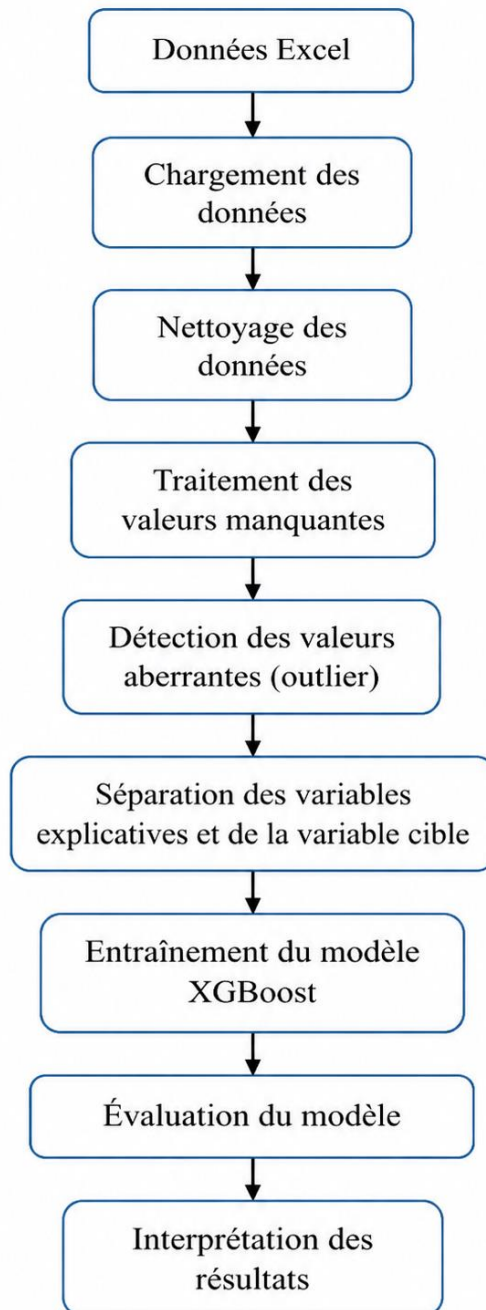


Figure 12 : Schéma des étapes de traitement des données

II-3-2. Critères d'évaluation du modèle

L'évaluation des performances du modèle s'appuie sur plusieurs critères simples permettant de juger la qualité des prédictions, parmi eux : l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE), le coefficient de détermination (R^2) et le coefficient de corrélation (R). (J.Gareth, D,Witten, 2023)

(R) et (R^2) déterminent la précision des prédictions par rapport aux données expérimentales, et la linéarité entre les prédictions et les valeurs expérimentales respectivement. On effectue les séquences d'entraînement et de test en fonction de (R^2), et on vérifie les prédictions globales en utilisant tous les indices pour confirmer la cohérence des modèles de régression. (Ekanayake et al., 2022)

Tableau 5 : Critères d'évaluation de XGBoost

Critère	Description	Exemple
R^2	Mesure la proportion de variance expliquée par le modèle, plus il est proche de 1, meilleure est la qualité de l'ajustement	0,85
R	Mesure à quel point deux variables évoluent ensemble, que ce soit dans le même sens ou en sens opposé, varie entre -1 et 1	0,92
RMSE	Indique l'écart-type des erreurs de prédiction, plus il est faible plus le modèle est précis	2,3
MAE	Mesure l'erreur moyenne absolue des prédictions, plus elle est faible plus les prédictions sont fiables	1,7

Exemple de calcul :

Pour mieux comprendre le principe de l'algorithme XGBoost, un exemple simplifié est présenté pour expliquer le processus de prédiction du modèle dans le tableau de données suivantes :

Tableau 6 : Données – Exemple XGBoost

Observation	Paramètre	Y (valeur réelle)
1	1	3
2	2	5
3	3	7

Premièrement, le modèle fait une prédiction initiale :

$$\widehat{y}_0(\text{la moyenne des valeurs } Y) = 3+5+7/3 = 5$$

Premier arbre : il va calculer les erreurs (la différence entre la valeur réelle et la prédiction) $y_i - \widehat{y}_0$ (résidu) donc :

Tableau 7 : Calcul des erreurs – Exemple XGBoost

Observation	Y	$\widehat{Y}_0$	Résidu
1	3	5	-2
2	5	5	0
3	7	5	2

On observe que les résidus traduisent l'écart entre les valeurs réelles et les prédictions initiales, qui sont ensuite corrigées par l'arbre suivant.

Donc $f_1(x)$ = résidu estimé par l'arbre 1=[-2,0,2]

XGBoost continue à ajouter d'autres arbres de manière successive afin de réduire ces erreurs, jusqu'à ce que les prédictions deviennent plus précises et que le modèle se stabilise.

Après la prédiction par XGBoost, et pour aller au-delà de la simple prédiction et comprendre le comportement du modèle, comme ce dernier reste difficile à interpréter (black box), ce projet utilise la méthode SHAP qui permet d'attribuer à chaque variable sa contribution à la prédiction et de fournir une interprétation transparente du modèle à l'aide de différentes visualisations graphiques simples et interprétables.

II-4. Méthodologie appliquée

Dans le cadre de cette étude, les données utilisées proviennent de la STEP de Chlef couvrant la période de 2011 à 2018. Ces données sont les paramètres physico-chimiques des eaux usées brutes et épurées : la demande biochimique en oxygène (DBO), la demande chimique en oxygène (DCO) et les matières en suspension (MES), considérés comme variables cibles, tandis que d'autres paramètres sont considérés comme variables explicatives.

Avant la modélisation, les données sont d'abord prétraitées afin d'assurer leur qualité, les valeurs manquantes ont été remplacées par la médiane pour limiter l'influence des valeurs extrêmes et conserver la structure globale des données.

Pour la modélisation l'algorithme XGBoost Regressor a été utilisé en raison de sa capacité à modéliser des relations complexes et non linéaires entre les variables.

L'évaluation des performances du modèle a été effectuée à l'aide des indicateurs statistiques tels que : le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE), l'erreur absolue (MAE), afin d'apprécier la qualité des prédictions et la fiabilité du modèle.

Cette étude s'inscrit dans la continuité de travaux antérieurs réalisés sur la même station, dans lesquels une configuration par défaut de modèle a été adoptée et a donné les résultats suivants (S Hasnaoui, C Hamadou, 2025) :

Tableau 8 : Résultats d'une étude précédente avec XGBoost (S Hasnaoui, C Hamadou, 2025)

Type d'eau	Paramètre	R^2	RMSE	NSE
Brute	DBO ₅	0.83	117.44	0.83
	DCO	0.63	432.41	0.63
	MES	0.26	773.77	0.26
Epurée	DBO ₅	-0.15	18.42	-0.15
	DCO	0.63	18.87	0.63
	MES	0.41	15.05	0.41

Ce tableau montre les résultats obtenus de la configuration par défaut, l'objectif de notre étude est d'aller plus loin en testant différentes configurations prédéfinies de l'algorithme XGBoost afin d'identifier celle qui donne les meilleures performances en termes de prédiction.

II-5. Langage de développement et bibliothèques utilisées dans le programme

Le choix des outils utilisés dans ce projet repose sur leur facilité d'utilisation et leur efficacité dans le domaine de l'apprentissage automatique, le projet a mis l'accent sur les

solutions permettant de manipuler, analyser et exploiter les données de façon claire et accessible.

II-5-1. Environnement de développement

Visual Studio Code (VSCode) est un éditeur de code simple et pratique, adapté à Python il permet d'écrire, tester et organiser les scripts (fichiers de code pour faire des tâches) facilement et offre des outils pour gérer les bibliothèques et visualiser les données.

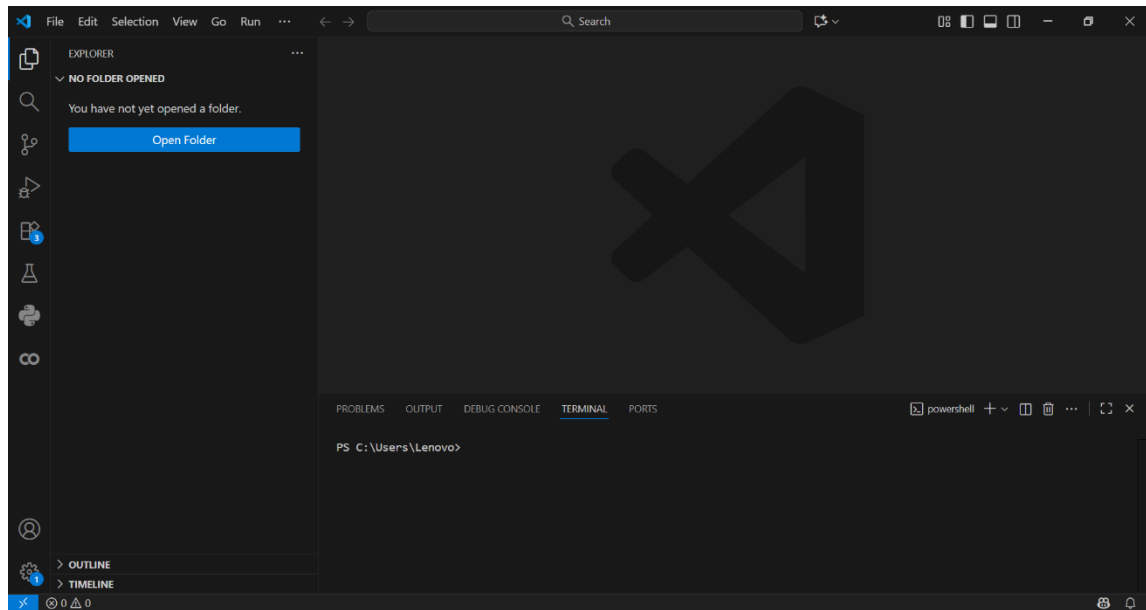


Figure 13 : L'interface de l'application VSCode

La figure ci-dessus, montre l'écran initial qui apparaît lors du lancement de VSCode

II-5-2. Langage Python

Python est un langage de programmation interprété, simple et polyvalent créé par Guido van Rossum en 1989. Il est largement utilisé dans les applications web, les datas science et le machine learning, grâce à ses bibliothèques comme NumPy, Pandas, Matplotlib et scikit-learn. Python est connu pour sa simplicité et sa lisibilité, il prend en charge plusieurs paradigmes de programmation dont la programmation structurée, orientée objet, fonctionnel et également utilisé pour l'automatisation, le scripting et le développement scientifique.(C.Dragne, 2026),(K. Chikhi, 2022)

II-5-3. Bibliothèques

II-5-3-1. NumPy

NumPy (Numerical Python) est une bibliothèque open-source disponible en Python ; elle est utile pour le calcul scientifique, elle permet de manipuler des tableaux multidimensionnels (ndarray), d'effectuer des opérations mathématiques telles que

l'algèbre linéaire, la génération de nombres aléatoires et la multiplication de matrices. NumPy fournit des fonctionnalités qui réduisent les tâches compliquées des analystes de données, elle constitue aussi une base essentielle pour plusieurs bibliothèques Python utilisées en analyse de données et en machine learning.(Gupta & Bagchi, 2024)

II-5-3-2. Pandas

Python for data analysis (Pandas) est parmi les bibliothèques de Python les plus utilisées, elle a été développée par Wes McKinney en 2008 pour améliorer l'analyse des données, elle permet d'explorer différents formats de fichiers (CSV,TXT,XLSX) et les convertir en DataFrame qui est généralement sous forme d'un tableau, Pandas peut gérer de multiples opérations sur les ensembles de données, telles que le sous-ensemble, le découpage, le regroupement et la restructuration, elle peut également traiter une grande variété d'ensembles de données dans différents formats, tels que les données matricielles et les données tabulaires hétérogènes.(Gupta & Bagchi, 2024)

II-5-3-3. Matplotlib

Matplotlib est une bibliothèque open source permettant de créer des graphiques en 2D à partir des tableaux (array) dans Python (voir figure ci-dessous), l'idée était de rendre la visualisation des données aussi simple et accessible que possible avec un minimum de commandes, elle fournit des graphiques (plots) de qualité et des figures interactives facile à manipuler, cela rend Matplotlib le meilleur choix pour visualiser les données. (*History — Matplotlib 3.10.8 documentation*, s. d.),(K. Chikhi, 2022)

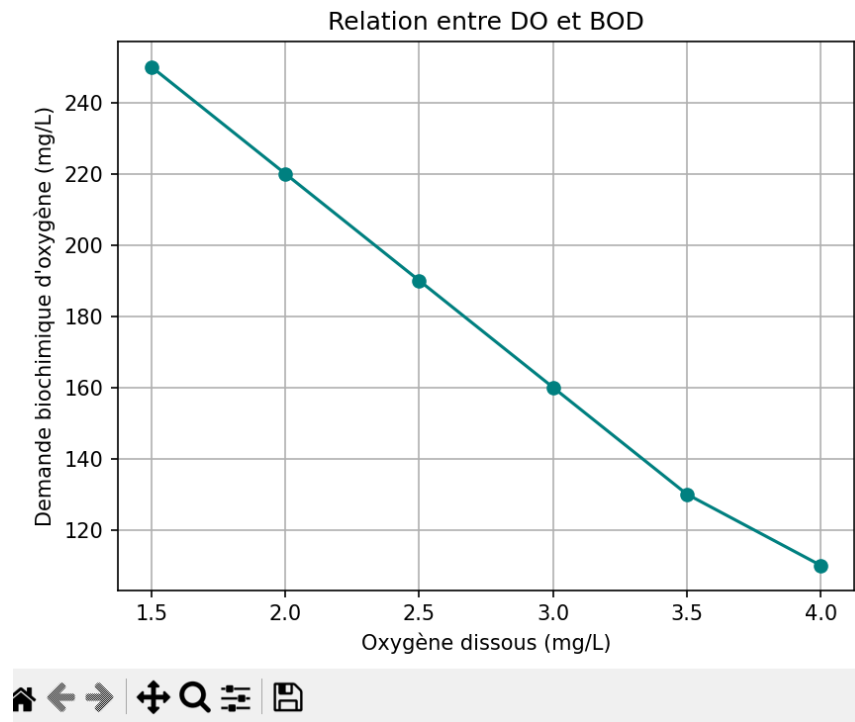


Figure 14 : Illustration de graphique avec Matplotlib

II-5-3-4. Scikit-learn

Développée en 2010, Scikit-learn est une bibliothèque Python open source contenant des algorithmes d'apprentissage automatique tels que, la classification et la régression, cette bibliothèque offre une modélisation des données et s'appuie sur NumPy, SciPy et Matplotlib.(R. Garreta & G. Moncecchi, 2013)

II-5-3-5. Les bibliothèques XGBoost et SHAP

Elles sont utilisées dans l'implémentation, XGBoost permet de réaliser les prédictions et SHAP est utilisée pour interpréter et visualiser les résultats, comme détaillé précédemment.

Conclusion

Le deuxième chapitre a permis de définir le cadre méthodologique de l'étude en mettant en évidence les atouts de l'algorithme **XGBoost** pour la modélisation des données de la STEP de Chlef. Sa capacité à représenter des relations complexes et non linéaires en fait un outil pertinent pour anticiper les dysfonctionnements et améliorer la fiabilité du traitement des eaux usées. Toutefois, la performance du modèle reste tributaire de la qualité des données et du réglage des hyperparamètres, ce qui justifie l'exploration de plusieurs configurations afin d'identifier celles offrant la meilleure précision prédictive.

L'apport de la méthode **SHAP** s'est révélé essentiel pour dépasser le caractère opaque des modèles de machine learning. En quantifiant la contribution de chaque variable aux prédictions, elle permet d'obtenir une interprétation à la fois globale et locale, renforçant ainsi la valeur opérationnelle des résultats. Cette complémentarité entre puissance de calcul et transparence analytique constitue un atout majeur pour transformer les sorties du modèle en indicateurs exploitables par les opérateurs.

Ainsi, ce chapitre prépare la discussion du chapitre suivant, où seront comparées les différentes configurations de XGBoost et interprétées à l'aide de SHAP, afin de dégager les paramètres les plus influents et proposer des pistes concrètes d'optimisation du fonctionnement de la STEP de Chlef.

Chapitre 3. Résultats et discussion.

Introduction

Ce chapitre présente les résultats obtenus par l'application du modèle XGBoost pour analyser les relations non-linéaires des paramètres physico-chimiques du procédé dit à boues activées de la STEP de Chlef pour mieux comprendre les phénomènes à l'origine des dysfonctionnements.

L'objectif est d'évaluer les performances du modèle sur des données historiques de la STEP, en comparant les configurations par défaut et personnalisées.

III-1. Résultats

Après avoir appliqué différentes configurations, nous avons obtenu, pour chaque configuration personnalisée, des résultats que nous allons présenter et interpréter dans les sections qui suivent. Nous allons commencer tout d'abord par une analyse statistique descriptive puis une interprétation via SHAP pour chacune des configurations étudiées.

III-1-1. Statistiques descriptives

Le but des statistiques descriptives est de donner des résumés des données, d'organiser et de simplifier les données pour qu'elles soient plus lisibles. Elles fournissent une description de l'ensemble des données en mettant en évidence les caractéristiques importantes telles que la centralité, le taux de dispersion et la distribution des données dans son ensemble.

Le tableau suivant présente les statistiques des eaux brutes :

Tableau 9 : Statistiques descriptives des eaux brutes

Paramètres	pH	Température	Conductivité	Turbidité	Salinité	MES	DBO ₅	DCO	Azote	nitrate	ammoniaque
Moyenne	7.345	20.914	15.359	1115	10.170	452.431	342.876	585.944	69.926	92.961	43.548
Médiane	7.340	21.400	2.440	1224	1.300	334.000	330.000	513.000	53.000	35.000	41.100
Ecart type	0.197	4.464	125.052	349.105	79.575	520.272	141.086	383.312	81.447	183.501	18.748
Min	5.950	0.400	0.440	0.600	0.100	0.800	1.000	35.000	18.000	0.300	7.000
Max	8.030	29.600	1547	1966	1529	7857	1361	4057	750.000	857.000	246.000
Coefficient de variation (%)	2.700	21.30	814.20	31.30	782.40	269.00	41.10	65.4	116.50	197.40	43.10

Ce tableau montre les statistiques des paramètres physico-chimiques des eaux usées à l'entrée de la STEP, les eaux brutes présentent une stabilité du milieu avec un pH neutre et une température modérée reflétant des conditions favorables aux processus biologiques au sein de la STEP. La pollution organique et particulaire (la DBO, MES et la DCO) affiche des valeurs moyennes élevées avec une dispersion importante, ce qui indique la présence de pics de charge plutôt qu'une pollution constante.

Le tableau suivant décrit les statistiques des eaux épurées :

Tableau 10 : Statistiques descriptives des eaux épurées

Paramètre	pH	Température	Conductivité	Turbidité	Salinité	MES	DBO ₅	DCO	Azote	Nitrate	ammoniaque
Moyenne	7.283	21.083	3.407	1059	2.016	14.262	12.143	44.775	12.124	7.097	4.015
Médiane	7.270	20.400	2.300	1162	1.200	12.000	8.000	44.000	9.000	5.200	1.700
Ecart-type	0.232	14.110	40.354	276.739	28.049	10.103	19.975	15.671	12.866	6.556	8.636
Min	1.270	2.400	0.700	1.000	0.400	0.800	0.100	0.000	0.300	0.200	0.020
Max	10.000	248.000	1413	1569	983.000	244.000	302.000	114.000	115.000	61.000	121.000
Coefficient de variation (%)	3.2	66.9	1184.3	26.1	1391.5	70.800	164.500	35.000	106.100	92.400	215.100

Ce tableau présente les statistiques des paramètres physico-chimiques des eaux épurées à la sortie de la station d'épuration, l'analyse révèle une qualité épuratoire satisfaisante et la station fonctionne correctement en conditions normales sur la majorité de la période étudiée.

Les graphiques suivants représentent la distribution des données de la DBO₅ et la DCO à l'entrée et à la sortie de la station d'épuration :

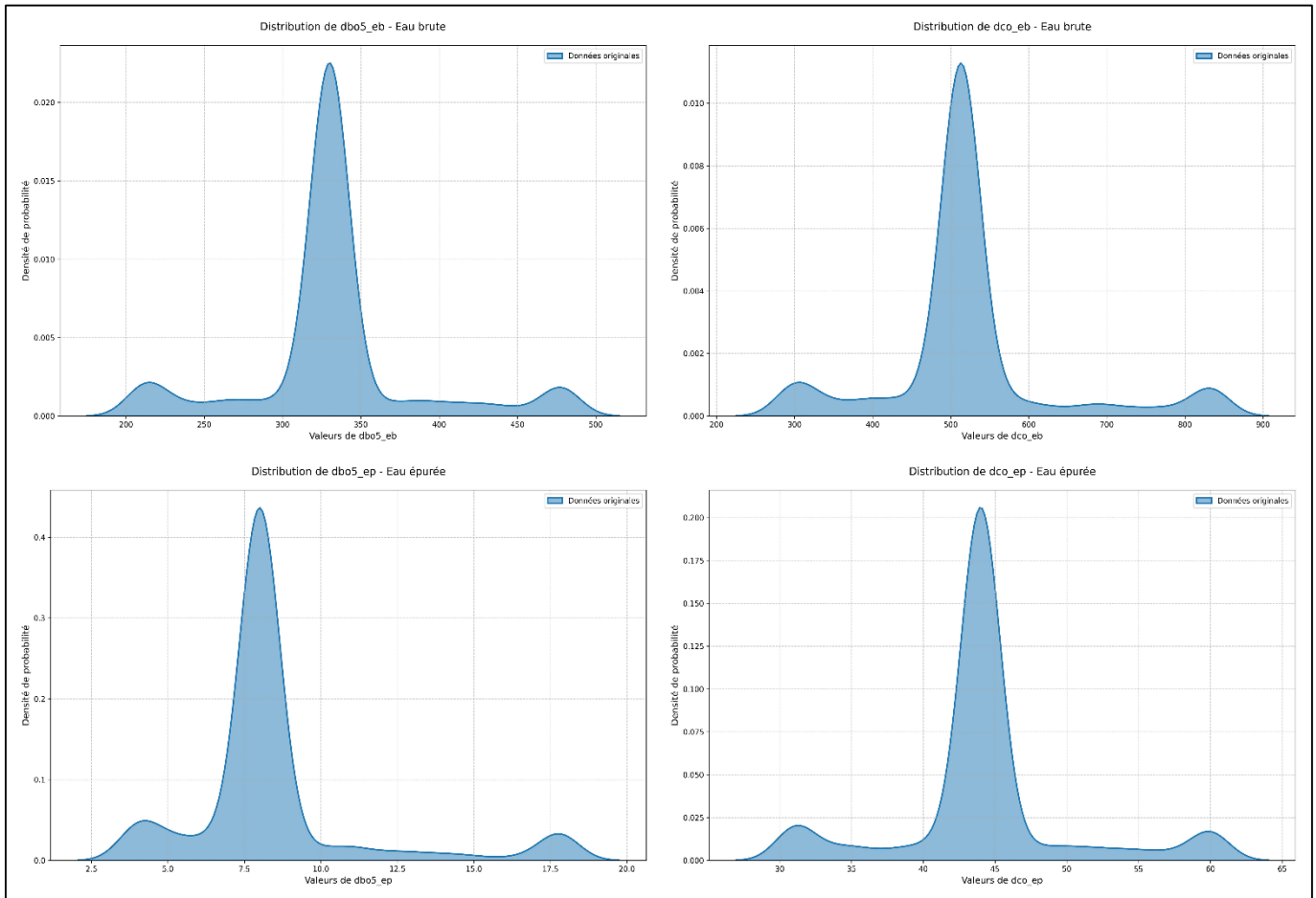


Figure 15 : Graphiques de distribution de la DBO₅ et la DCO

La demande biologique en oxygène sur 5 jours (DBO₅) et la demande chimique en oxygène (DCO) sont les deux indicateurs de référence pour évaluer la charge organique et le contrôle de la qualité des eaux usées. D'après les graphiques, la distribution de la DBO₅ de l'eau brute présente un pic principal dominant autour de 330 mg/l, ce qui représente une charge organique typique à l'entrée de la station et on observe deux pics secondaires l'un autour de 210 mg/l représente une eau usée peu chargée, et l'autre, autour de 470 mg/l (épisode de surcharge).

La distribution de la DCO des eaux brutes présente un pic principal centré autour de 510 mg/l avec des pics secondaires autour de 320 mg/l et 820 mg/l, la distribution est plus étalée que celle de la DBO₅ car la DCO capte une gamme plus large de composés.

D'après les graphiques des eaux épurées, on observe que la distribution de la DBO_5 présente un pic principal autour de 7 à 8 mg/l, ce qui indique que la station produit régulièrement une eau épurée et homogène.

La distribution de la DCO présente un pic principal autour de 44 mg/l, en dessous des normes de rejet, ce qui indique une efficacité élevée du système de traitement biologique.

III-1-2. Analyse comparative des performances des configurations prédéfinies

Dans le tableau qui suit, on a dressé les résultats des configurations appliquées sur les eaux brutes et épurées de la STEP de Chlef pour avoir une lecture comparative de ces résultats selon les indicateurs de performances (R^2 , RMSE et MAE)

Tableau 11 : Tableau comparatif des résultats XGBoost avec configurations prédéfinies

Configuration prédéfinie	Eau brute			Eau épurée		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Régularisation forte	0.6811	78.6262	50.8061	0.5633	3.3988	2.2478
Vitesse maximale	0.7133	74.9357	48.9122	0.6000	3.2566	2.1814
Equilibré	0.7733	67.4513	42.4924	0.7160	2.7486	1.8032
Haute précision	0.8065	63.0705	39.8556	0.7695	2.4770	1.6196
GPU	0.7738	67.6521	42.5336	0.7194	2.7313	1.7970
Petit jeu de données	0.7733	67.4513	42.4924	0.7160	2.7486	1.8032

La configuration haute précision présente les meilleures performances dans les deux cas avec les valeurs de R^2 les plus élevées ($R^2_{eb}=0.80$ et $R^2_{ep}=0.76$) et les plus faibles valeurs de RMSE et MAE ($MAE_{eb}=39.85$, $MAE_{ep}=1.61$ et $RMSE_{eb}=63.07$, $RMSE_{ep}=2.47$), ce qui indique une meilleure explication des données et une réduction des erreurs de prédiction.

Les autres configurations comme vitesse maximale et régularisation forte donnent des performances correctes mais inférieures, la configuration équilibrée et petit jeu de données présentent exactement les mêmes résultats, ce qui s'explique par une absence de variation réelle lors de l'exécution.

Le choix de la configuration haute précision dans cette étude apparaît comme le plus pertinent, car elle garantit une grande fiabilité et une précision des résultats pour l'eau brute et l'eau épurée.

III-1-3. Interprétation des résultats du modèle XGBoost à l'aide de SHAP

Les résultats des analyses SHAP sont présentés et interprétés pour déterminer les variables qui ont la plus grande influence sur les prédictions du modèle XGBoost. Les analyses SHAP sont faites séparément pour les eaux brutes et les eaux épurées, en considérant les variables DCO, DBO₅ et MES. Pour chaque type d'eau, une sélection des graphiques est exploitée afin de mettre en évidence les contributions des variables et de faciliter l'interprétation des résultats obtenus.

III-1-4. Analyse SHAP des eaux brutes

III-1-4-1. DBO₅

Graphique de synthèse

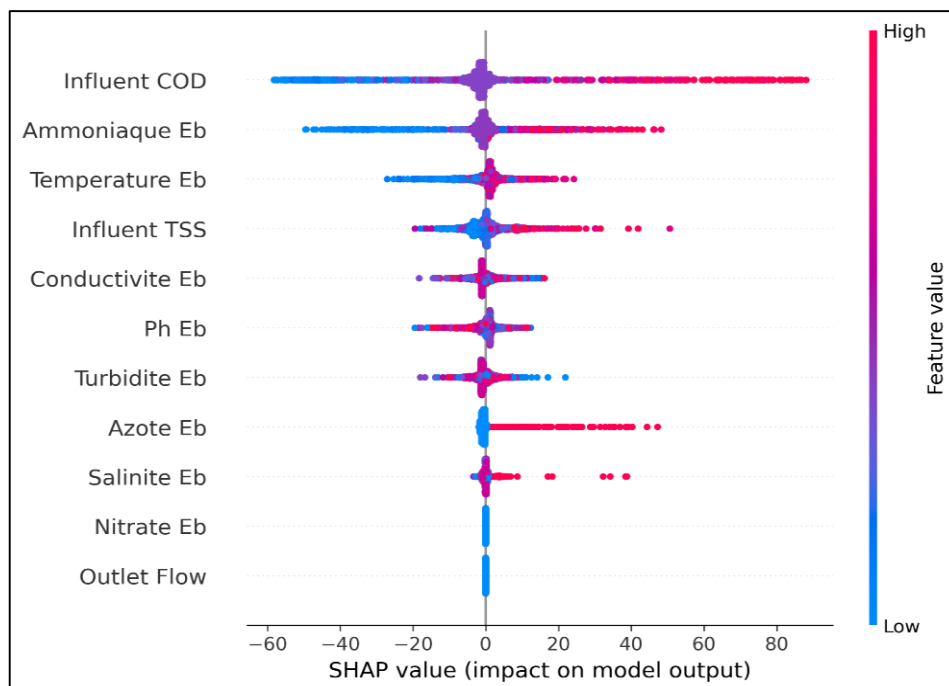


Figure 16 : Graphique de synthèse SHAP de la DBO₅_eb (summary plot)

L'analyse du graphique SHAP de la DBO₅ en entrée met en évidence des relations directement interprétables du point de vue du procédé d'épuration. La DCO apparaît comme le facteur prédominant, confirmant son rôle d'indicateur global de la charge organique. Des valeurs élevées de DCO sont associées à une augmentation de la DBO₅, traduisant une charge biodégradable plus importante. Cette relation permet, en

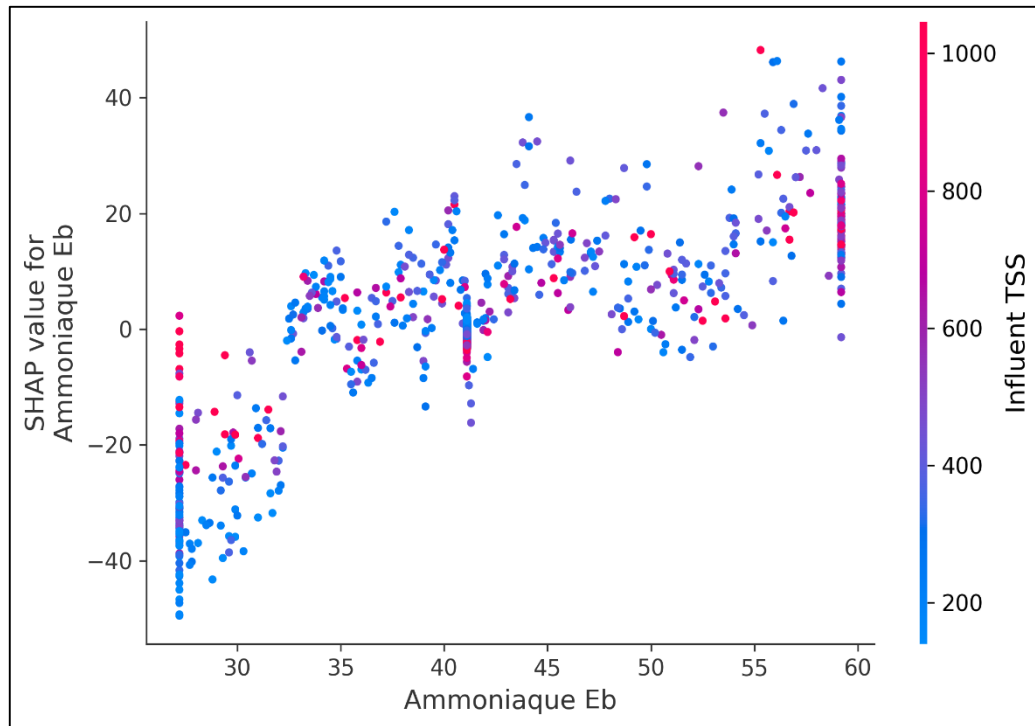
exploitation, d'anticiper les variations de la charge biologique et d'ajuster les conditions d'aération et de traitement.

L'ammoniaque constitue le second paramètre influent, avec un effet non linéaire. Des concentrations élevées traduisent généralement une charge organique azotée plus importante, issue de la dégradation de la matière organique. Elle agit ainsi comme un indicateur complémentaire de la pollution et renseigne sur les besoins potentiels en nitrification.

La température exerce une influence modérée mais significative, en lien avec les variations saisonnières. Elle agit sur la vitesse des processus biologiques et sur la concentration des effluents, ce qui peut impacter les performances du traitement.

Les autres paramètres, tels que les MES, la turbidité ou le pH, présentent une contribution plus faible. Ils interviennent principalement comme variables d'accompagnement, traduisant les conditions physicochimiques de l'effluent.

Ainsi, cette analyse met en évidence des relations cohérentes avec le fonctionnement du procédé et fournit des éléments utiles pour le pilotage opérationnel de la station.

Graphique de dépendance**Figure 17** : Graphique de dépendance SHAP

Ce graphique présente le graphe de dépendance SHAP de l'ammoniaque de l'eau brute en fonction de son impact sur la DBO₅ de l'eau brute et de l'effet de l'interaction avec les matières en suspension (MES).

D'abord, on peut observer que la relation entre l'ammoniaque et la DBO₅ est non linéaire. Pour les basses valeurs d'ammoniaque (27-35 mg/l), la majorité des valeurs SHAP est négative quelle que soit la teneur en MES représentée en couleur et cela signifie que l'ammoniaque diminue la DBO prédite. Par contre, plus l'ammoniaque augmente, les valeurs SHAP deviennent positives.

En termes de procédés, ces données correspondent au comportement d'une station d'épuration à boues activée. L'ammoniaque constitue un indicateur du taux azoté issu de la décomposition de la matière organique, et l'élévation de ce dernier est souvent liée à une pollution plus importante.

En outre, le fait que des taux élevés d'ammoniaque et de matières en suspension se combinent signifie qu'il s'agit d'une eau brute polluée, d'où une élévation de la DBO₅.

III-1-4-2. DCO

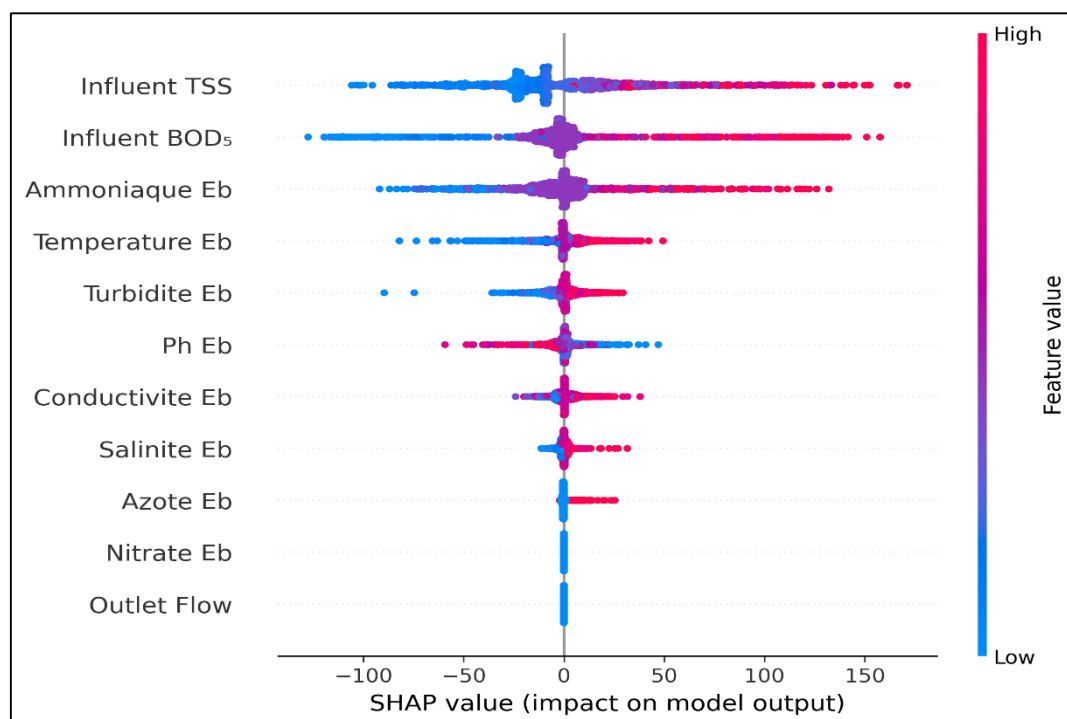


Figure 18 : Graphique de synthèse SHAP pour la DCO (summary plot)

Ce graphe représente le summary plot de la DCO des eaux brutes, on observe que les matières en suspension (MES) constituent la variable la plus explicative de la DCO de l'entrée avec des valeurs SHAP entre -120 et +160, les points rouges représentent des MES élevées qui augmentent fortement la DCO prédite, et les points en bleu sont des MES faibles qui font réduire la prédiction. La deuxième variable est la DBO₅ avec des valeurs SHAP proches à les valeurs des MES, confirmant la complémentarité entre ces deux indicateurs de la charge organique. La troisième variable est l'ammoniaque avec des valeurs SHAP très étendues de -90 à +130, un NH₃ élevé signale un effluent non dilué riche en composés azotés organiques dont la dégradation contribue à la DCO. Et la température avec des valeurs SHAP entre -60 et +50, la température influence la DCO en fonction des conditions climatiques, en été la dégradation aérobie est accélérée dans les réseaux d'assainissement ce qui peut produire des composés intermédiaires qui augmentent la DCO, et en hiver la dilution par les eaux pluviales réduit la concentration.

Les autres variables comme la turbidité, le pH et la conductivité varient entre -40 et +30.

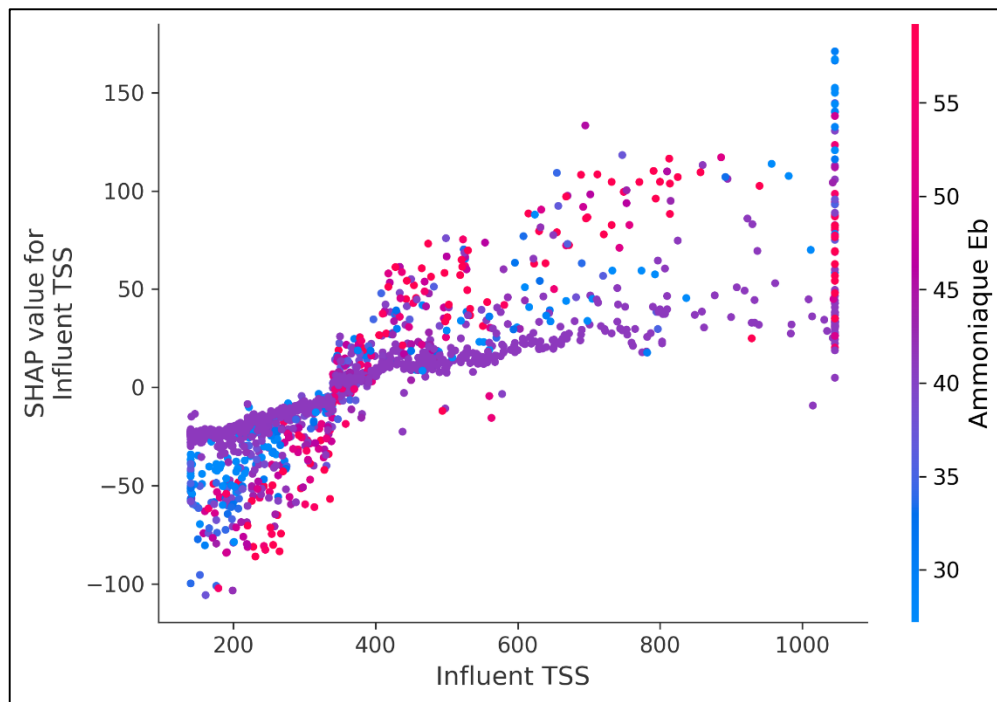


Figure 19 : Graphique de dépendance SHAP de la DCO

La figure 5 représente le graphe de la dépendance SHAP qui met en évidence l'influence des matières en suspension de l'eau brute sur la prédiction de la DCO, la couleur révèle une interaction non-linéaire entre MES et ammoniac.

On observe de faibles concentrations des MES (100 à 200 mg/l) avec des valeurs SHAP majoritairement négatives entre -100 et -20, traduisant un effet réducteur de la DCO. À partir d'un seuil de 400 mg/l l'effet devient positif, ce qui signifie que les MES participent significativement à l'augmentation de la charge organique. De 900 à 1000 mg/l, les points s'alignent verticalement avec des valeurs SHAP entre +20 et +160, la couleur révèle une relation non-linéaire entre MES et l'ammoniaque, les points en rouge représentent un (NH₃) élevé, ce qui amplifie l'impact des MES sur la DCO, les points en bleu signifient un (NH₃) faible qui fait atténuer l'impact sur la DCO même avec des MES élevés.

III-1-4-3. MES

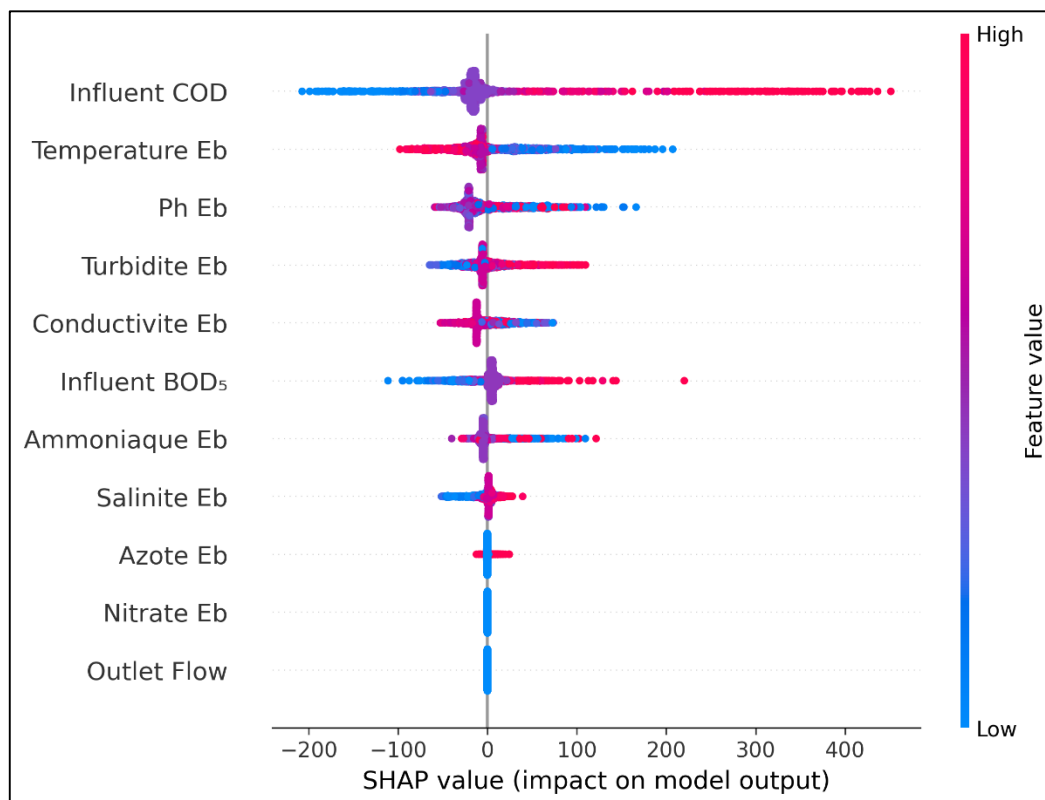


Figure 20 : Graphique de synthèse (summary plot) SHAP des MES_{eb}

La figure représente le classement des variables explicatives selon leur contribution à la prédiction des MES des eaux brutes, la DCO_{eb} (influent COD) est la variable dominante avec des valeurs SHAP de -200 à +450, ce qui témoigne d'une relation entre la charge organique totale et la charge particulaire de l'effluent. Dans un effluent urbain, les matières en suspension sont majoritairement d'origine organique et toute augmentation de la DCO s'accompagne d'une hausse des MES, la distribution des points en rouge et en bleu confirme cette relation. La température occupe la deuxième position avec une influence SHAP de -100 à +220, les températures élevées augmentent la prédiction de MES, reflétant les conditions estivales caractérisées par une concentration élevée de l'effluent en absence de dilution par les eaux pluviales, alors qu'en période hivernale, les précipitations et les températures basses entraînent une diminution de la charge particulaire. Le pH est classé en troisième rang, avec un effet majoritairement positif avec des valeurs SHAP de -50 à +180, en tenant compte du fait que les conditions alcalines favorisent la mise en suspension de certains composés organiques colloïdaux. Les autres variables contribuent modestement à la prédiction de la DCO.

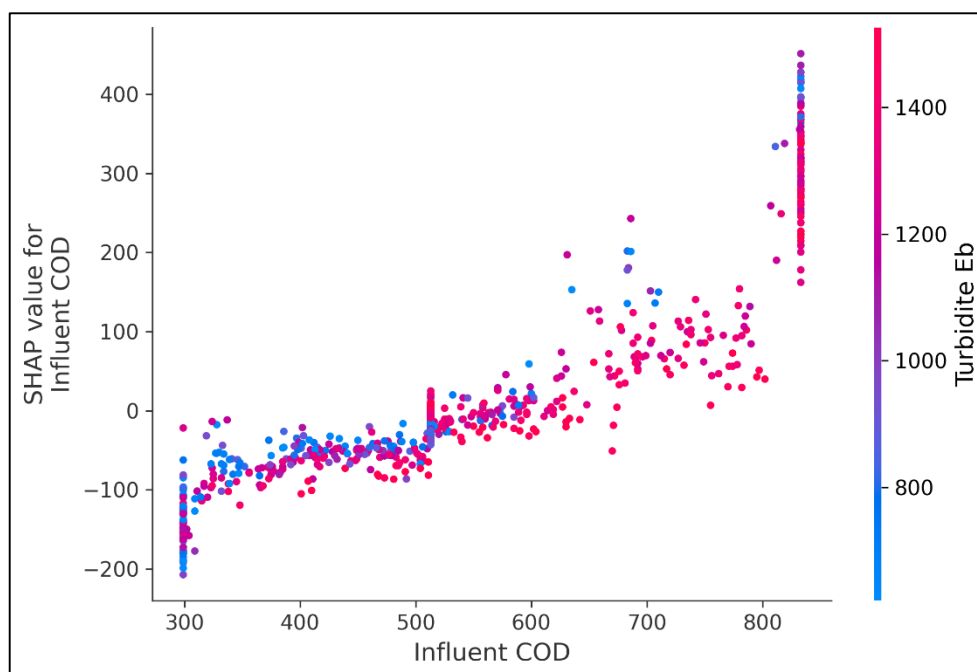


Figure 21 : Graphique de la dépendance SHAP

Le graphique de la dépendance de la DCO en entrée coloré par la turbidité de l'eau brute, illustre l'influence de la charge organique totale sur la prédiction des matières en suspension et révèle une tendance croissante, non-linéaire et fortement asymétrique

Pour les valeurs faibles de la DCO (300 à 400 mg/l) les valeurs SHAP sont négatives atteignant jusqu'à -200, ce qui signale un effluent peu concentré en matière organique particulaire, entraînant une prédiction des MES très inférieure à la moyenne.

À partir d'une concentration en DCO de 500 mg/l, on observe un point d'inflexion crucial, qui voit passer subitement les valeurs SHAP d'un signe négatif à un signe positif, marquant un seuil au-delà duquel la charge organique totale commence à contribuer positivement à la prédiction.

Entre 500 et 800 mg/l, on observe une tendance croissante et une dispersion verticale importante et l'interaction avec la turbidité devient visible, cette tendance montre un effet synergique entre DCO et turbidité dans la prédiction des MES, une forte turbidité accompagnée d'une DCO élevée augmente la prédiction des MES.

III-1-5. Analyse SHAP des eaux épurées

III-1-5-1. DBO₅

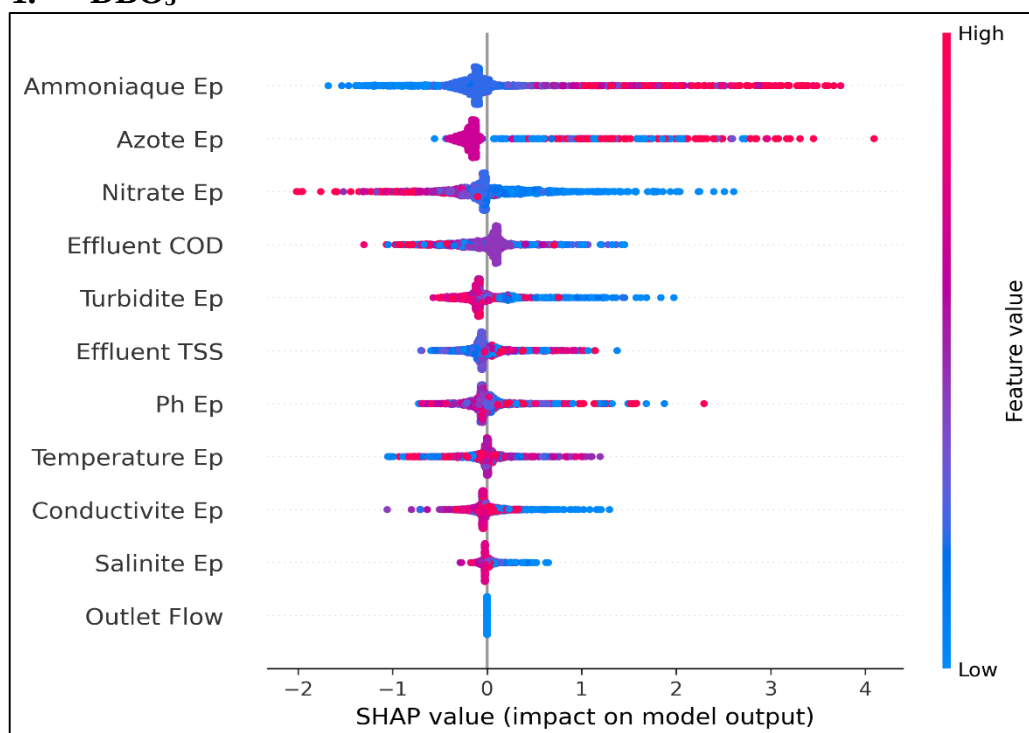


Figure 22 : Graphique de synthèse (summary plot) SHAP de la DBO₅

L'analyse des valeurs SHAP appliquée à la prédiction de la DBO₅ en eau épurée révèle une structure explicative différente de celle observée en eau brute, reflétant le changement de la nature de l'effluent après le traitement biologique par boues activées, la première différence majeure est les valeurs SHAP comprises entre -2 et +4 contrairement à l'eau brute, ce qui traduit l'efficacité du traitement biologique de la station.

L'ammoniaque en eau épurée s'impose comme le prédicteur dominant avec des valeurs SHAP de -1,5 à +4, dans une station à boues activées, la présence résiduelle d'ammoniaque en sortie indique une nitrification incomplète associée à une augmentation de la DBO₅ en sortie, ce qui peut refléter des périodes de performance biologique réduite. On observe l'azote total en deuxième position et son effet est strictement positif (les points bleus à droite avec des valeurs SHAP jusqu'au +4), ce qui indique qu'une teneur quelconque de composés azotés organiques dans l'eau sortante de la station est associée à une valeur non négligeable de la DBO₅.

Les nitrates présents dans l'eau épurée ont un comportement contraire, les valeurs en rouge (nitrates élevés) sont placées à gauche avec des valeurs SHAP négatives (jusqu'à -2), alors que les valeurs en bleu (nitrates faibles) se trouvent à droite, les nitrates élevés en sortie de

la station indiquent une nitrification parfaite, où la matière organique a été totalement décomposée, ce qui génère automatiquement une DBO₅ très faible.

La DCO, la turbidité et les autres variables présentent un impact globalement faible avec des valeurs SHAP entre -1 et +1 et centrées autour de zéro traduisant un rôle secondaire dans la variabilité de la DBO₅ en sortie.

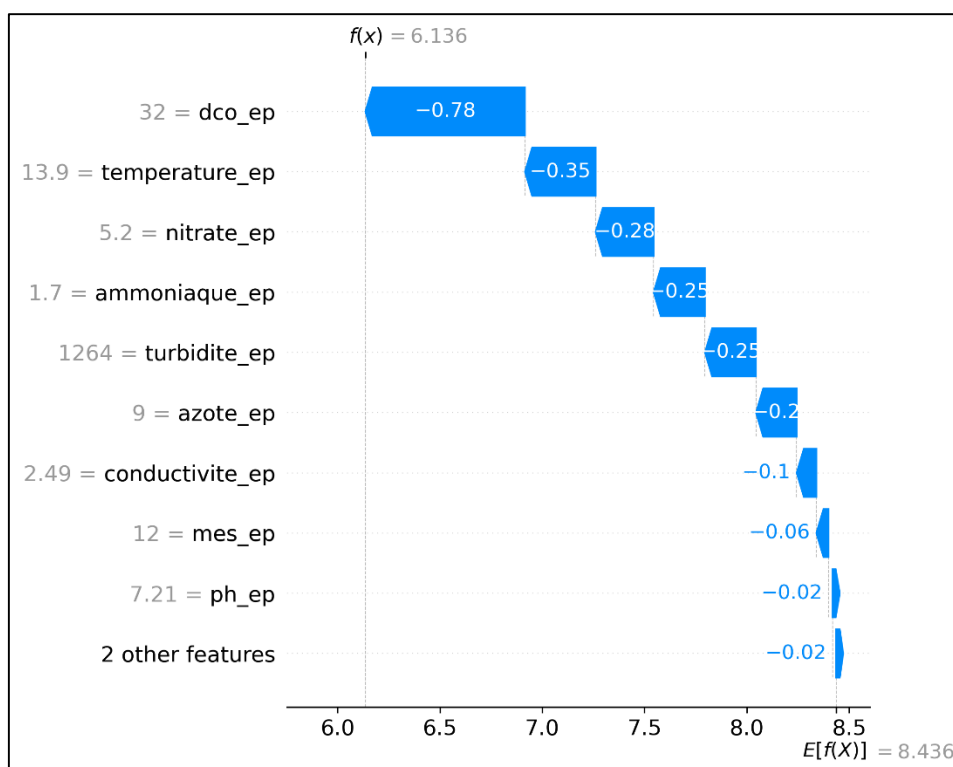


Figure 23 : SHAP waterfall plot de la DBO₅

Le graphique waterfall plot (figure 9) montre la contribution de différentes variables à la prédiction de la DBO₅ des eaux épurées. D'une observation donnée, de la valeur de base du modèle ($E[f(x)] \approx 8.44$ mg/l) on passe par des ajustements graduels dus aux effets combinés de plusieurs variables à une valeur finale prédite ($f(x)=6.14$ mg/l) qui reflète une très bonne qualité de l'effluent quant à la charge organique, aucune variable ne pousse la prédiction à la hausse, ce qui traduit une journée de fonctionnement optimal de la station où tout paramètres physicochimiques de l'effluent traité convergent vers une qualité supérieure à la moyenne du modèle. La DCO_ep constitue le contributeur dominant -0.78 (le tiers de l'écart total -2.30 mg/l), une DCO de 32 mg/l est une valeur très faible pour un effluent traité, signalant une dégradation presque complète de la matière organique

oxydable. L'effet de la température en sortie est de -0.35, une température fraîche à la sortie (13.9 °C) suggérant une condition favorable au traitement biologique. Les nitrates en eaux épurées apportent une contribution négative de -0.28 avec une valeur d'observation de 5.2 mg/l, ces résultats confirment l'interprétation établie dans summary plot (figure 8), une faible teneur en nitrates en sortie indique une bonne nitrification.

L'ammoniaque contribue négativement pour -0.25 avec une valeur d'observation de 1.7 mg/l qui est très faible pour un effluent épuré, ce qui indique que la nitrification consomme presque la totalité de l'ammoniaque, la turbidité contribue également pour -0.25, elle marque une valeur élevée à la sortie, ce qui peut indiquer la présence des particules minérales non-organiques (sable, carbonates) qui augmentent la turbidité sans contribuer à la charge organique et l'azote contribue pour -0.2.

La conductivité, les mes, le potentiel d'hydrogène et les autres variables, exercent des contributions négatives marginales entre -0.1 et -0.02.

Tous ces paramètres contribuent à une diminution de la DBO₅, indiquant un bon fonctionnement de la STEP, où la dégradation organique se fait avec succès et une maîtrise satisfaisante des autres paramètres physico-chimiques en sortie.

III-1-5-2. DCO

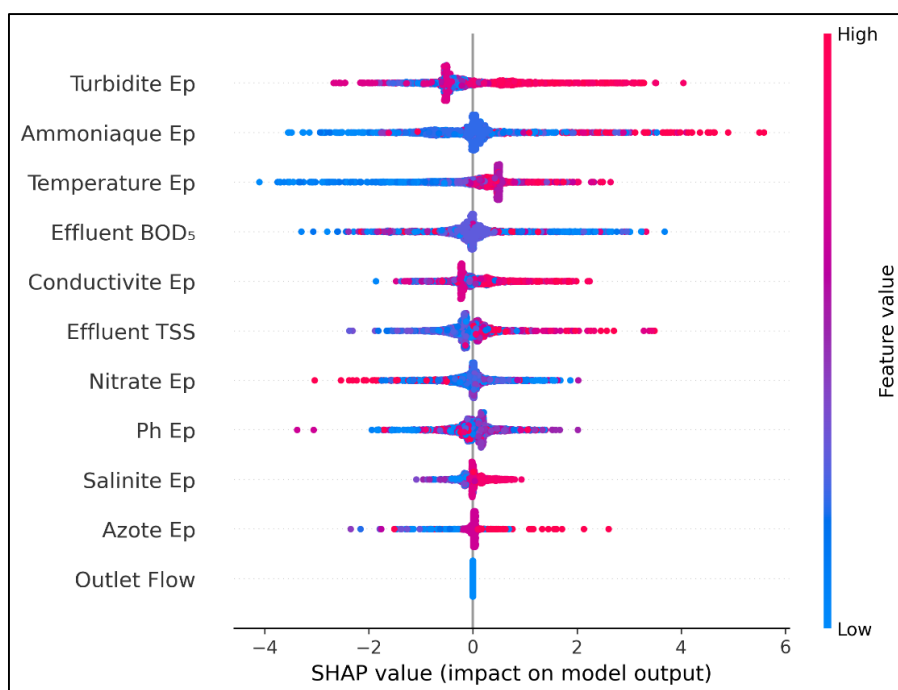


Figure 24 : Graphique de synthèse (summary plot) SHAP de la DCO

La figure 10 représente la classification des variables explicatives contribuent à la prédiction de la DCO à la sortie, l'amplitude globale des valeurs SHAP comprise entre -4 et +5, on observe que la turbidité est le facteur le plus influent avec des valeurs SHAP entre -3 et +4 et une distribution équilibrée, les points rouges représentent une turbidité élevée se répartissent près du zéro, tandis que les points bleus (turbidité faible) sont concentrés à gauche, ce qui traduit une relation complexe et non-linéaire entre turbidité et DCO résiduelle, ceci est due à la nature des particules présentes dans les eaux traitées, une turbidité élevée peut signifier qu'il y a des floccs biologiques résiduels avec une grande quantité de matières organiques qui peuvent faire monter la valeur de la DCO, ou que les particules sont minérales et inertes provenant du clarificateur sans affecter la valeur de la DCO, d'où la dispersion observée de part et d'autre du zéro.

L'ammoniaque en sortie occupe la deuxième position avec une influence majoritairement positive, avec des valeurs SHAP entre -3 et +5, cette distribution confirme l'importance de la nitrification dans la performance épuratoire globale. La température en eau épurée se classe troisième avec une influence bidirectionnelle et des valeurs SHAP comprises entre -3 et +3, les températures basses (points bleus à gauche) réduisent la prédiction de la DCO, cela peut correspondre aux périodes hivernales (la présence des eaux pluviales) ou la dilution produit un effluent moins chargé, tandis que les températures élevées (les points en rouge) font augmenter la DCO, ce qui reflète l'intensité de l'activité biologique en été.

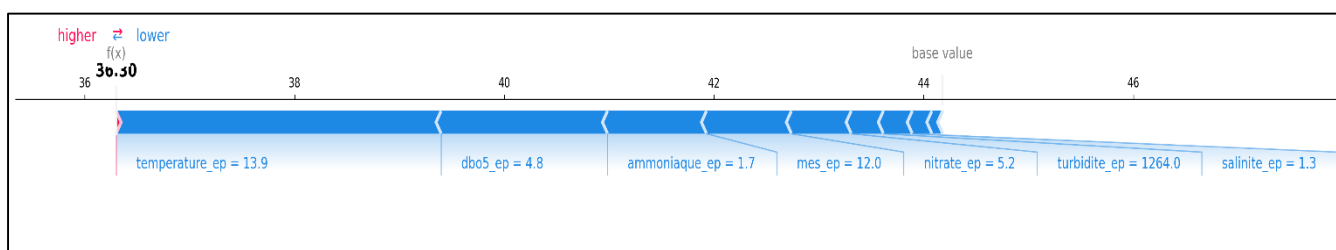


Figure 25 : Graphique force plot SHAP de la DCO

La DBO₅ résiduelle influence positivement avec des valeurs SHAP entre -2 et +3, une DBO₅ résiduelle élevée s'accompagne systématiquement d'une DCO résiduelle plus importante. La conductivité, les MES, les nitrates et le pH exercent des effets modérés influençant légèrement la DCO. La salinité et l'azote total présentent les contributions les plus faibles, tous ces résultats montrent donc que le suivi de la turbidité ainsi que de

l'ammoniac résiduel est l'indicateur d'exploitation le plus pertinent dans le cadre du contrôle des qualités de la DCO au sortir de la station d'épuration urbaine à boues activées.

Le graphique force plot (figure 11) illustre la contribution des variables explicatives à la prédiction de la DCO des eaux épurées pour une observation donnée, caractérisée par une convergence totale des forces vers la réduction de la prédiction, la valeur prédite ($f(x) \approx 36,3$ mg/l) est inférieure à la valeur de base du modèle (≈ 44 mg/l), soit un écart de -7.4 mg/l, on constate qu'aucune variable ne contribue positivement à la prédiction. La température en eau épurée ($T= 13.9^{\circ}\text{C}$) constitue le contributeur dominant, occupant la majeure partie de la barre bleue, une température de 13.9°C à la sortie est apparaît comme le signal d'un effluent produit dans des conditions de faible charge, favorisant une minéralisation complète de la matière organique, ce qui réduit la DCO résiduelle.

La DBO_5 est en deuxième position avec une contribution négative de $4,8$ mg/L, ce qui indique que la dégradation biologique de la fraction biodégradable a été quasi-complète durant cette journée. L'ammoniaque (ammoniaque= 1.7mg/l) apporte une contribution négative, ce qui indique une nitrification complète, ce qui réduit la DCO résiduelle. Les MES_ep (12mg/l), les nitrates (5.2mg/l), la turbidité (1264 NTU) et la salinité (1.3g/l) exercent des contributions négatives plus modestes, la turbidité marque un niveau élevé mais elle contribue négativement sur la DCO, ce qui confirme l'interprétation de summary plot (figure 10).

III-1-5-3. MES

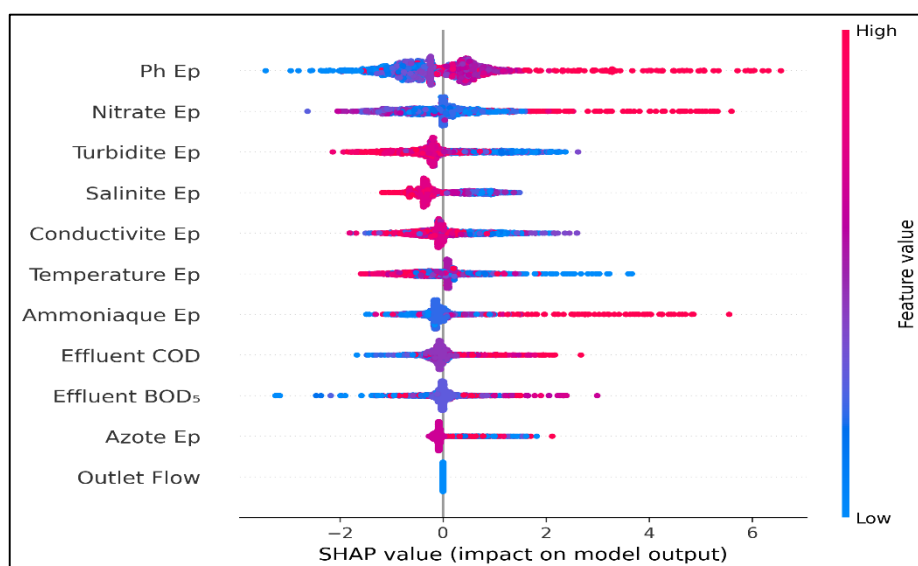


Figure 26 : Graphique de synthèse (summary plot) SHAP des MES

Le graphique summary plot représente la classification des variables explicatives qui ont contribué à la prédiction des MES des eaux épurées. Cette classification révèle une structure significative dominée par les paramètres physicochimiques de l'effluent traité plutôt que par les indicateurs de la charge organique. Les valeurs SHAP comprises entre -2 et +7 sont cohérentes avec les valeurs des MES observées à la sortie (entre 8 et 20 mg/l). le pH est le prédicteur dominant avec une distribution étalée et des valeurs SHAP de -2 à +7, le positionnement des points (rouges à droite et bleus à gauche), traduisant une relation directe entre le potentiel d'hydrogène et les matières en suspension, un pH élevé en sortie favorise l'augmentation des MES dans l'effluent final. Les nitrates occupent la deuxième position avec une influence positive pour les valeurs élevées (les points en rouge à droite) avec des valeurs SHAP jusqu'à +5.5, les nitrates élevés en sortie indiquent une nitrification intense dans le bassin biologique, le processus génère une biomasse qui peut se retrouver en suspension dans l'effluent clarifié.

La turbidité est en troisième position avec une influence importante, la turbidité élevée affecte positivement la prédiction des MES, ce qui est normal car la turbidité et les MES sont toutes les deux utilisées pour mesurer la charge particulaire. La salinité et la conductivité exercent des effets modérément positifs, traduisant que les conditions ioniques élevées dans l'effluent traité peuvent favoriser le passage de fines particules dans l'eau traitée. La température contribue positivement aux valeurs élevées avec des valeurs SHAP entre -1.5 et +3. Toutes ces observations suggèrent que les MES dans l'eau traitée dépendent surtout d'un certain nombre de paramètres relatifs à la qualité de la séparation eau-boues dans le décanteur secondaire, comme le pH, la nitrification, la température, plutôt que du contenu organique dans l'effluent, soulignant l'importance du contrôle de ces paramètres pour le respect des normes de rejet en MES.

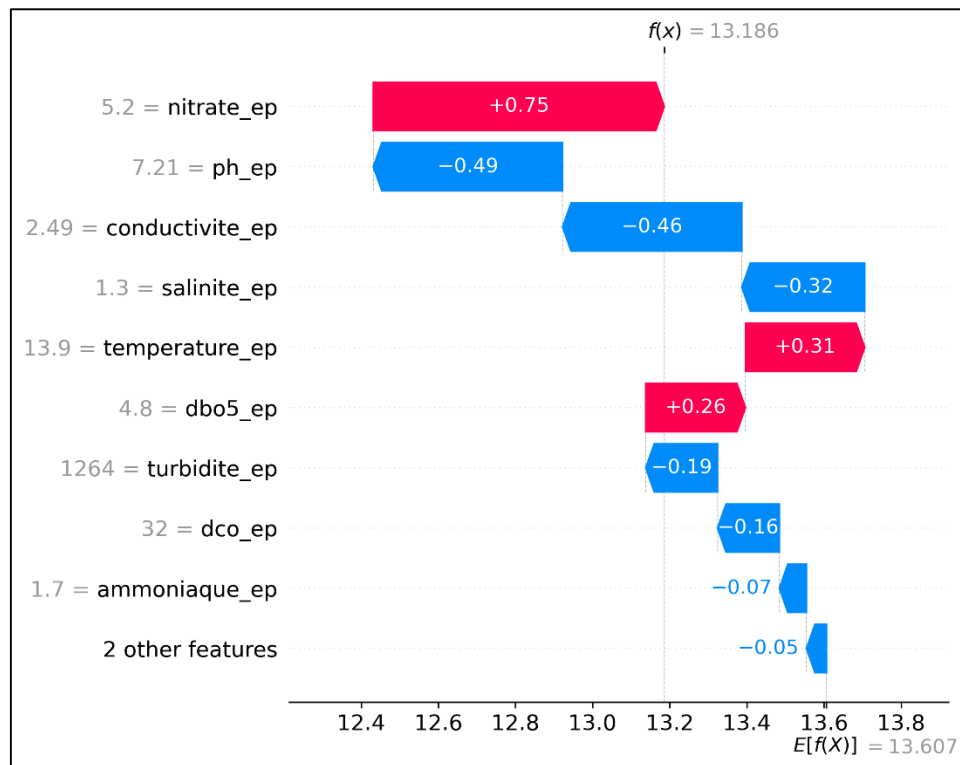


Figure 27 : SHAP waterfall plot des MES_ep

La figure 13 représente le graphique waterfall des MES des eaux épurées, la valeur de base s'établit à 13.607 mg/l, tandis que la prédiction finale est de 13.186 mg/l, soit un écart de -0.421 mg/l, ce qui traduit une journée de fonctionnement près de la moyenne de la station. Les nitrates constituent le seul contributeur positif significatif avec + 0.75, l'activité nitrifiante augmente les MES à la sortie conformément à l'interprétation de summary plot (figure 12).

L'influence de la température est positive avec +0.31 bien que les températures basses soient associées à une bonne décantabilité, le modèle capte une relation particulière de l'influence de cette température, combinée aux autres conditions de la journée et associée à une légère augmentation des MES, peut-être en raison d'une viscosité plus élevée de l'eau réduisant la vitesse de sédimentation des floccs au niveau du clarificateur. La DBO₅ apporte une contribution positive modeste de +0.26, qui révèle l'existence de composants organiques biodégradables résiduels transportés par des particules biologiques en suspension non complètement sédimentées.

Contrairement, le ph, la salinité, la conductivité, et les autres variables contribuent négativement (ph-0.49, conductivité -0.46, salinité-0.32), ces trois paramètres convergeant ainsi pour améliorer la qualité de la séparation boues-eau.

III-2. Discussion

III-2-1. Impact du paramétrage du modèle XGBoost sur les performances

Les résultats obtenus démontrent de manière explicite l'importance du paramétrage du modèle XGBoost dans la modélisation des procédés d'épuration. Comparativement à une étude précédente réalisée sur la STEP de Chlef utilisant la configuration par défaut (S Hasnaoui, C Hamadou, 2025), les nouvelles configurations ont permis des améliorations significatives des performances.

La moyenne du coefficient de détermination R^2 des eaux brutes est passée de 0,57 à 0,80, tandis que celle des eaux épurées a évolué de 0,30 à 0,77. Ces résultats confirment que le modèle par défaut présente des limitations importantes, levées grâce à l'optimisation des hyperparamètres. Cette observation est cohérente avec la littérature (Firdaus et al., 2025) et (Mamadou Ndiaye, René Boissy, Anastasie Mendy et Gallo Niang, 2025), qui souligne l'impact positif de l'optimisation sur les indicateurs R^2 , RMSE et MAE.

III-2-2. Identification du modèle optimal

Parmi les six configurations évaluées (régularisation forte, équilibrée, haute précision, vitesse maximale, GPU et petit jeu de données), la configuration « haute précision » se distingue comme la plus performante.

Elle présente des valeurs moyennes de :

- $R^2_{eb} = 0,8065$
- $R^2_{ep} = 0,7695$

avec les erreurs RMSE et MAE les plus faibles. Cette performance s'explique par :

- un nombre élevé d'estimateurs ;
- une profondeur importante des arbres ;
- un faible taux d'apprentissage.

Ces paramètres permettent une meilleure capture des relations complexes, au prix d'un temps de calcul plus élevé, conformément aux principes du boosting (Chen & Guestrin, 2016).

III-2-3. Caractérisation statistique des eaux brutes

III-2-3-1. Charge organique et variabilité

Les statistiques descriptives indiquent un profil typique d'un effluent domestique urbain caractérisé par :

- une DBO₅ moyenne de 342,876 mg/L avec un coefficient de variation de 41,1 %,
- une DCO moyenne de 585,944 mg/L avec un coefficient de variation de 65,4 %.

Ces valeurs traduisent une forte variabilité temporelle et la présence de pics de charge organique.

III-2-3-2. Variabilité des matières en suspension

Les MES présentent la variabilité la plus importante (CV = 269 %), avec des valeurs maximales dépassant 7800 mg/L. Ces extrêmes sont probablement liés à des épisodes pluvieux, des surcharges hydrauliques ou la remise en suspension de sédiments.

III-2-3-3. Conditions physicochimiques

Le pH moyen (7,37) et la température (20,9°C) indiquent des conditions favorables à l'activité biologique, notamment pour les bactéries hétérotrophes et les micro-organismes nitrifiants.

III-2-4. Performance épuratoire des eaux traitées

Les résultats montrent une performance globale satisfaisante :

- DBO₅ moyenne en sortie : 12,1 mg/L ($\approx$ 96 % d'abattement),
- DCO moyenne : 44,8 mg/L (conforme aux normes).

Ces résultats traduisent une élimination efficace de la matière organique.

III-2-5. Interprétation SHAP des eaux brutes

III-2-5-1. Prédiction de la DBO₅

La demande chimique en oxygène (DCO) apparaît comme le prédicteur dominant, avec des valeurs SHAP comprises entre -60 et +90, traduisant la forte corrélation physicochimique entre la DCO et la DBO₅. Ce rôle central confirme que la DCO constitue un indicateur majeur de la charge organique et de la qualité des effluents.

L'ammoniaque, quant à elle, présente une relation non linéaire ; son effet est négatif à faible concentration, mais devient positif à des niveaux plus élevés. Elle agit ainsi comme

un indicateur indirect de la charge organique azotée, révélant des interactions complexes avec les autres paramètres.

Enfin, la température exerce une influence notable sur les processus de biodégradation, en modulant l'activité microbienne et en introduisant un effet saisonnier sur la concentration des effluents.

Ces observations mettent en évidence l'importance de considérer simultanément les interactions entre variables pour une interprétation robuste des résultats, et soulignent le rôle complémentaire de la DCO, de l'ammoniaque et de la température dans la compréhension du fonctionnement de la station d'épuration.

III-2-5-2. Prédiction de la DCO

Les matières en suspension (MES) se distinguent comme le facteur explicatif principal, avec des valeurs SHAP pouvant atteindre +160, ce qui confirme la contribution déterminante de la fraction particulaire à la demande chimique en oxygène (DCO). Leur rôle ne s'exerce pas isolément ; la DBO₅ et l'ammoniaque participent également de manière significative, traduisant une interdépendance entre les indicateurs de pollution organique et azotée. Cette synergie met en évidence la complexité des interactions physicochimiques au sein des effluents et souligne l'importance d'une analyse conjointe des paramètres pour une compréhension robuste du fonctionnement de la station d'épuration.

Ainsi, les MES, en association avec la DBO₅ et l'ammoniaque, constituent des variables structurantes dont l'influence conditionne directement la qualité des rejets et la performance globale du procédé.

III-2-5-3. Analyse de dépendance SHAP des MES

L'analyse révèle une relation non linéaire et un seuil autour de 400 mg/L. En dessous, contributions négatives signifiant une faible charge. Au-dessus, une contribution positive d'où une dominance particulaire.

Une interaction avec l'ammoniaque est observée, traduisant la coexistence de charges organiques et azotées.

III-2-6. Interprétation SHAP des eaux épurées

III-2-6-1. Prédiction de la DBO₅

La prédiction de la DBO₅ se caractérise par une faible amplitude des valeurs SHAP, comprise entre -2 et +4, ce qui traduit une certaine stabilité du procédé et une régularité dans le comportement du modèle.

Parmi les variables explicatives, l'ammoniaque occupe une place dominante, son influence met en évidence une nitrification parfois incomplète et souligne son lien direct avec la qualité de l'effluent. Cette observation confirme le rôle de l'azote réduit comme indicateur critique du fonctionnement biologique de la station. Les nitrates, en revanche, présentent une relation négative avec la DBO₅, suggérant qu'une nitrification avancée contribue à améliorer la qualité de l'eau en réduisant la charge organique biodégradable.

L'ensemble de ces résultats illustre la complémentarité entre les paramètres azotés et la DBO₅, et met en évidence l'importance de suivre simultanément l'évolution de l'ammoniaque et des nitrates pour mieux comprendre et anticiper les performances du procédé.

Par ailleurs, le graphique montre une diminution de la DBO₅ (6,14 mg/L vs 8,44 mg/L), expliquée par :

- Une faible DCO ;
- une température favorable ;
- une nitrification efficace.

Cela confirme la cohérence locale du modèle.

III-2-6-2. Prédiction de la DCO

La prédiction de la DCO met en évidence plusieurs facteurs explicatifs dont l'influence varie selon la nature des particules et les conditions d'exploitation. La turbidité présente un effet bidirectionnel : lorsqu'elle est liée à des particules organiques, elle entraîne une augmentation de la DCO, tandis que les particules minérales exercent un effet neutre. L'ammoniaque joue également un rôle important en tant qu'indicateur de performance biologique, traduisant l'efficacité du processus de nitrification.

La température, quant à elle, agit sur l'activité microbienne et conditionne la dynamique de biodégradation, ce qui souligne son impact saisonnier sur la qualité des effluents.

L'analyse locale à travers les « force plots » confirme un bon fonctionnement global du procédé, avec une contribution positive de l'ensemble des variables à la réduction de la DCO.

Les variations observées apparaissent cohérentes avec la nature contextuelle des valeurs SHAP, qui attribuent une importance spécifique à chaque observation. Cette approche permet ainsi de saisir les effets locaux propres aux conditions d'exploitation de la STEP et d'enrichir l'interprétation des résultats par une lecture plus fine des interactions entre paramètres.

III-2-6-3. Analyse SHAP des MES

Le pH apparaît comme le facteur structurant majeur, influençant directement la stabilité des floes et l'efficacité du processus de décantation. Son rôle est déterminant pour garantir la qualité de la séparation solide-liquide et, par conséquent, la performance globale du traitement. Les nitrates traduisent quant à eux l'impact de la nitrification sur la qualité des boues et leur transfert vers l'effluent, soulignant l'importance des mécanismes biologiques dans la maîtrise des rejets. La proximité des valeurs prédites avec la valeur de base témoigne d'un fonctionnement stable du système, ce qui confirme la fiabilité du procédé dans les conditions observées.

L'approche SHAP permet ainsi d'identifier avec précision les facteurs critiques susceptibles d'entraîner des dépassements des normes en MES, offrant aux exploitants des éléments opérationnels essentiels pour anticiper et corriger d'éventuelles dérives. Cette lecture met en évidence la complémentarité entre paramètres physicochimiques et biologiques dans la compréhension du comportement de la station.

Conclusion

En conclusion, l'analyse des résultats obtenus dans ce chapitre montre la pertinence de l'approche combinant XGBoost et SHAP dans la modélisation et l'interprétation des procédés à boues activées. En termes de performances prédictives, la configuration optimisée a contribué des améliorations significatives par rapport au paramétrage par défaut, notamment concernant les variables de l'eau épurée, dont la performance était initialement plus basse. Ces résultats prouvent que l'optimisation des hyperparamètres est une étape incontournable dans le développement des modèles d'apprentissage automatique appliqués à des données environnementales réelles.

Sur le plan de l'interprétabilité, l'analyse SHAP a mis en lumière des relations physicochimiques conformes aux connaissances établies sur le fonctionnement des stations d'épuration. Cela inclut la dépendance entre la DCO et la DBO₅, le rôle de la température dans la dynamique de la charge organique, l'importance de la nitrification pour la qualité de l'effluent traité et l'influence du potentiel d'hydrogène et de la turbidité sur les processus de décantation des boues.

Ces résultats dépassent le simple cadre de la prédiction en fournissant une base d'aide à la décision opérationnelle pour les exploitants de la STEP.

Conclusion générale

La gestion durable des stations d'épuration des eaux usées représente un défi environnemental et technique de premier ordre, particulièrement dans le contexte algérien où la rareté des ressources hydrique exerce une pression croissante sur les infrastructures d'assainissement. C'est dans cette perspective que s'inscrit le présent mémoire, qui a proposé une approche innovante fondée sur le couplage de l'algorithme XGBoost et de la méthode d'interprétabilité SHAP pour la modélisation, la prédiction et le diagnostic du fonctionnement du procédé à boues activées de la station d'épuration de Chlef, à partir de données historiques d'exploitation couvrant la période 2011–2018.

Sur le plan des performances prédictives, les résultats ont clairement montré que le paramétrage du modèle XGBoost est un facteur déterminant de la qualité des prédictions. La configuration « haute précision », caractérisée par un grand nombre d'estimateurs, une profondeur d'arbres élevée et un faible taux d'apprentissage, s'est imposée comme la plus performante, avec des coefficients de détermination moyens de $R^2_{eb} = 0.81$ et $R^2_{ep} = 0.77$. Ces résultats représentent une amélioration substantielle par rapport à la configuration par défaut utilisée dans une étude précédente (S Hasnaoui, C Hamadou, 2025), donc l'optimisation des hyperparamètres est une étape incontournable dans le déploiement de modèles d'apprentissage automatique sur des données environnementales réelles.

Sur le plan de l'analyse des données physicochimiques, les statistiques descriptives ont révélé un profil d'effluent typique d'une agglomération urbaine, caractérisé par une forte charge organique à l'entrée et un rendement épuratoire globalement satisfaisant à la sortie de la station. La DBO₅ résiduelle moyenne de 12.1 mg/l et la DCO moyenne de 44.8 mg/l indiquent un fonctionnement conforme aux normes de rejets nationales.

Sur le plan de l'interprétabilité, l'analyse SHAP a constitué l'apport le plus original de ce travail, en permettant de lever le voile sur les mécanismes physicochimiques capturés par le modèle. Les résultats pour les eaux brutes confirment que le DCO est le prédicteur de la DBO₅, ce qui traduit la forte interdépendance de ces deux indicateurs de la charge organique, les résultats montrent aussi le rôle régulateur de la température sur la concentration des effluents. Pour les eaux épurées, l'analyse a mis en évidence l'importance de la nitrification, la présence résiduelle d'ammoniaque en sortie est un

indicateur d'une nitrification incomplète associée à une dégradation insuffisante de la matière organique, tandis que des teneurs élevées en nitrates signalaient au contraire une nitrification avancée et une bonne qualité de l'effluent. Le pH et la turbidité jouent un rôle sur la décantabilité des boues dans le clarificateur, soulignant l'importance du suivi de ces paramètres pour la stabilisation des MES.

Ces résultats dépassent le simple cadre de la prédiction pour offrir une base solide d'aide à la décision. La combinaison XGBoost-SHAP offre la possibilité d'identifier les paramètres qui dégradent ou améliorent la qualité de l'effluent. Dans une station d'épuration en cours de réhabilitation, ces méthodes d'analyse sont un élément complémentaire intéressant.

En perspective, l'intégration de données d'exploitation complémentaires telles que le débit et l'âge des boues permettrait d'enrichir la capacité prédictive du modèle et d'affiner les interprétations SHAP. Ainsi, l'extension de l'approche à d'autres stations d'épuration à boues activées algériennes contribuerait à évaluer la généralisation des modèles développés et à constituer un référentiel national pour la modélisation intelligente de l'épuration.

En conclusion, ce mémoire démontre le rôle majeur que joue le machine learning interprétable dans la gestion durable des stations d'épuration. Grâce à la combinaison de la puissance de prédiction de XGBoost avec la puissance analytique de SHAP, il est désormais possible de convertir des données opérationnelles brutes en outils concrets de gestion du processus de purification. Cette stratégie, mise en œuvre dans ce cas sur la STEP de Chlef, représente un pas important vers une meilleure gestion des stations d'épuration.

BIBLIOGRAPHIE

- A.Kellouche. (2026).** *Rapport d'Analyse et d'Interprétation des Modèles Prédictifs.*
- Aouel, R. E. (2006).** *Correspondant au 23 avril 2006.*
- Ayyach A.1, , FATHALLAH R.1, , HBAIZ E.M2, *, et al. (2016).** *Caractérisation physico-chimique et bactériologique des eaux usées de la station d'épuration de la ville de Dar El Gueddari (MAROC).*
- Bachi, O. E., Halilat, M. T., Bissati, S., et al. (2022).** Wastewater Treatment Performance of Aerated Lagoons, Activated Sludge and Constructed Wetlands under an Arid Algerian Climate. *Sustainability*, 14(24), 16503. <https://doi.org/10.3390/su142416503>
- Benyahya, M.** *Les virus des eaux usées et leur élimination au cours des traitements des effluents pollués.*
- C. Albertyn. (2025).** *A comparative study of gradient boosting algorithms and transformer neural networks for classification problems in the tabular data domain.* Stellenbosch university.
- Cao, J., Gong, Y., Wang, F., et al. (2026).** Development and internal-external validation of a risk prediction model for acute pain after HAIC for patients with liver cancer using Logistic Regression and XGBoost Algorithm. *Asia-Pacific Journal of Oncology Nursing*, 100923. <https://doi.org/10.1016/j.apjon.2026.100923>
- C.Dragne. (2026).** *A small revolution in Python. 10.*
- Charfi, A.** *Etude d'un procédé membranaire de traitement des eaux usées : Effet des paramètres biotiques et abiotiques sur le colmatage de la membrane.*
- Chen, T., & Guestrin, C. (2016).** XGBoost : A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 785-794. <https://doi.org/10.1145/2939672.2939785>

- Ching, P. M. L., Zou, X., et al (2022).** Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. *Environmental Research*, 210, 112953. <https://doi.org/10.1016/j.envres.2022.112953>
- Combe, M. (2016, mars 25).** Le fonctionnement d'une station d'épuration. *Natura Sciences*.
- Decret exclusif n° 09-209.** (2009).
- Durot, M.-A., Molle, P., & Prost-Boucle, S. (s. d.).** *Disques biologiques Exelio (Biorotor Pamco)*.
- Ekanayake, I. U., Meddage, D. et al (2022).** A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, 16, e01059. <https://doi.org/10.1016/j.cscm.2022.e01059>
- Epaississement et déshydratation des boues* | SUEZ. (s. d.). Consulté 3 mars 2026, à l'adresse <https://www.suez.com/fr/eau/pratiques-durables/boues-epuration/deshydratation>
- European Commission (Éd.). (2001).** *Extensive wastewater treatment processes adapted to small and medium sized communities (500 to 5,000 population equivalents)*. Publications Office.
- Firdaus, E. A., Manurung, J., et al (2025).** Optimization of XGBoost hyperparameters using grid search and random search for credit card default prediction. *Jurnal Mandiri IT*, 14(2) p.p. 269-280. <https://doi.org/10.35335/mandiri.v14i2.468>
- Ghellai, A., & Mohammedi, S. (2018).** *Contribution à l'étude du dysfonctionnement dans une station d'épuration. Cas STEP d'EL KARMA-ORAN*.
- Gupta, P., & Bagchi, A. (2024a).** Introduction to NumPy. In P. Gupta & A. Bagchi (Éds.), *Essentials of Python for Artificial Intelligence and Machine Learning* (p. 127-159). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43725-0_4

- Gupta, P., & Bagchi, A. (2024b).** Introduction to Pandas. In P. Gupta & A. Bagchi (Éds.), *Essentials of Python for Artificial Intelligence and Machine Learning* (p. 161-196). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43725-0_5
- History—Matplotlib 3.10.8 documentation.* (s. d.).[en ligne] [consulté le 4 mars 2026] disponible sur le web: <https://matplotlib.org/stable/project/history.html>
- Interpretable Machine Learning—An overview | ScienceDirect Topics.* (2023).
- J.Gareth, D,Witten,.** (2023). *An introduction to statistical learning with applications in Python.* Springer.
- K. Chikhi. (2022).** *Prédiction de la perte de clients dans une entreprise commerciale : Candia.* Béjaia : université Abdelrahmene-Mira.
- Kellouche, A. (2017).** *Cas des STEP de Chlef et Ain Defla.*
- La Convention de Barcelone et ses Protocoles | UNEPMAP.* (s. d.). [en ligne] [consulté le 8 avril 2026] disponible sur le web: <https://www.unep.org/unepmap/fr/who-we-are/barcelona-convention-and-protocols>
- Li, Z. (2022).** Extracting spatial effects from machine learning model using local interpretation method : An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, 96, 101845. <https://doi.org/10.1016/j.compenvurbsys.2022.101845>
- Lundberg, S. M., & Lee, S.-I. (2017).** A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.
- M Sawadogo, B Gérard, et al (2023).** *Diagnostic du fonctionnement de la Station d'Épuration de Kossodo : Approche par la caractérisation des eaux usées brutes entrant.* <https://www.soaphys.org/wp-content/uploads/2024/09/CJKA04-SAWADOGOMariam-et-al-2023-Journal-SOAPHYS.pdf>
- M Wiens , Verone-Boyle , et al (2025).** *A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications.* <https://doi.org/10.1111/cts.70172>

- M Yasin, T Ketema, K Bacha. (2015).** *Physico-chemical and bacteriological quality of drinking water of different sources, Jimma zone, Southwest Ethiopia.*
- Mamadou Ndiaye, René Boissy, et al (2025).** *Modélisation de la susceptibilité aux inondations par apprentissage automatique : Performance de SVM, KNN, XGBoost, RF et Perceptron multitouche (MLP) dans la moyenne vallée du fleuve Sénégal.*
- Morin-Crini, N., Winterton, P., et al (2017).** Chapitre IV. Paramètres chimiques de l'eau et rejets industriels. In *Eaux industrielles contaminées : Réglementation, paramètres chimiques et biologiques & procédés d'épuration innovants* (p. 103-144). Presses universitaires de Franche-Comté. <https://doi.org/10.4000/books.pufc.10972>
- Najah Ahmed, A., Binti Othman, F., et al. (2019).** Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578, 124084.
<https://doi.org/10.1016/j.jhydrol.2019.124084>
- Ndiaye, M., Boissy, R., Faye, et al (2025).** Évaluation comparative des algorithmes d'apprentissage automatique pour la classification des types de sols à partir de caractéristiques physico-chimiques : Application de Random Forest, XGBoost, SVM et KNN. *VertigO. La revue internationale en sciences de l'environnement.*
<https://doi.org/10.4000/13raq>
- Niazkar, M., Menapace, A., et al. (2024).** Applications of XGBoost in water resources engineering : A systematic literature review (Dec 2018–May 2023). *Environmental Modelling & Software*, 174, 105971. <https://doi.org/10.1016/j.envsoft.2024.105971>
- Oliveira, P., Coufort-Saudejaud, C., et al. (2017).** Procédé de traitement des eaux usées par boues activées : Lien entre les propriétés morphologiques des floes et l'âge de boue. *Revue des sciences de l'eau*, 30(1), 29-32. <https://doi.org/10.7202/1040060ar>
- Pan, B. (2018).** Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*. <https://doi.org/10.1088/1755-1315/113/1/012127>

- Ponce-Bobadilla, A. V., Schmitt, V., et al. (2024).** Practical guide to SHAP analysis : Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*. <https://doi.org/10.1111/cts.70056>
- Putatunda, S., & Rama, K. (2018).** A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, SPML '18*, 6-10. <https://doi.org/10.1145/3297067.3297080>
- R. Garreta & G. Moncecchi. (2013).** *Learning scikit-learn : Machine Learning in Python*.
- Racault et Seguret, Y. et F. (2004).** *Eléments de conception et de dimensionnement des lits bactériens*. <https://internet6-national-epnac.custom.hub.inrae.fr/content/download/3783/35318?version=1>
- RENOU, S. (2006).** *Analyse de cycle de vie appliquée aux systèmes de traitement des eaux usées*. France : université de lorraine disponible sur le web: https://docnum.univ-lorraine.fr/public/INPL/2006_RENOU_S.pdf
- Rivaldo, R., Taufik, R., et al. (2025).** A Comparative Study of XGBoost, LightGBM, and CatBoost Models for Customer Churn Prediction in the Banking Industry. *Jurnal Pepadun*, 6, 178-187. <https://doi.org/10.23960/pepadun.v6i2.277>
- Ruto, D. S., Jang, Z. S., et al.** Toward Sustainable Lagoon Wastewater Treatment : A Review of Nutrient Management Technologies and Their Suitability for Small Communities. *ACS Es&t Water*, 5(11), 6200-6216. <https://doi.org/10.1021/acsestwater.5c00757>
- S Hasnaoui, C Hamadou. (2025).** *Analyse statistique des paramètres physico-chimiques de la STEP de CHLEF pour optimiser la gestion de l'eau en agriculture*. Chlef: Université Hassiba Benbouali
- Sbaihia, S. (2023).** *Diagnostic de la station d'épuration de Chlef et valorisation des eaux usées épurées en agriculture [ENSH]*. <http://dspace.ensh.dz:8080/jspui/bitstream/123456789/1183/1/5-0010-23.pdf>

Siden.lu—Biodisques. (s. d.).[en ligne] [Consulté 2 mars 2026], sur le web :

<https://www.siden.lu/Biodisques-550503>

Takefuji, Y. (2025). Limitations of XGBoost-SHAP integration for interpretable machine learning in antimicrobial resistance prediction. *Journal of Infection*, 91(3), 106575.

<https://doi.org/10.1016/j.jinf.2025.106575>

Traitement par le procédé des lits bactériens. [en ligne] [consulté le 15 avril 2026] disponible sur

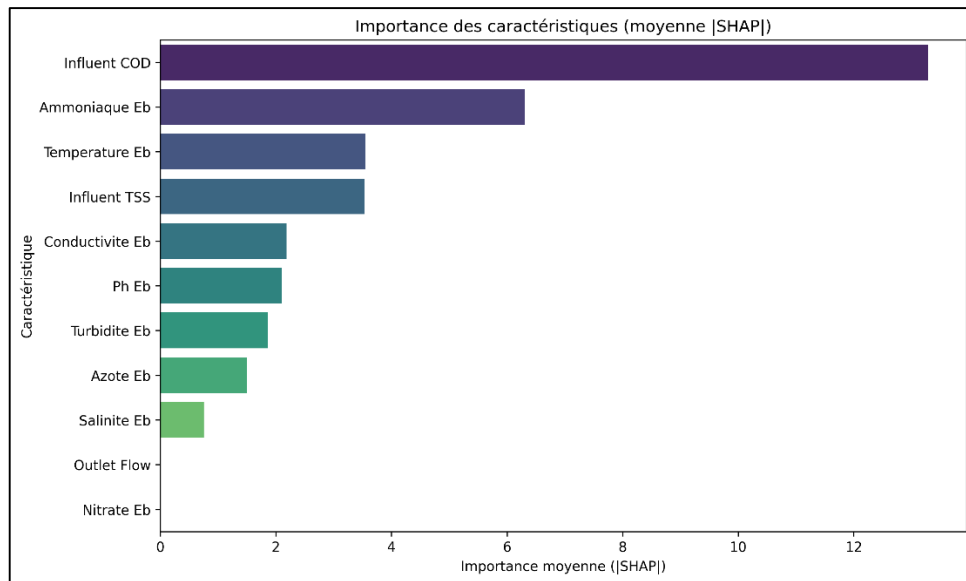
le web : [http://eau.gouvernement.lu/fr/domaines-activite/eaux-usees-](http://eau.gouvernement.lu/fr/domaines-activite/eaux-usees-pluviales/Traitements/Traitement-par-le-procede-des-lits-bacteriens.html)

[pluviales/Traitements/Traitement-par-le-procede-des-lits-bacteriens.html](http://eau.gouvernement.lu/fr/domaines-activite/eaux-usees-pluviales/Traitements/Traitement-par-le-procede-des-lits-bacteriens.html)

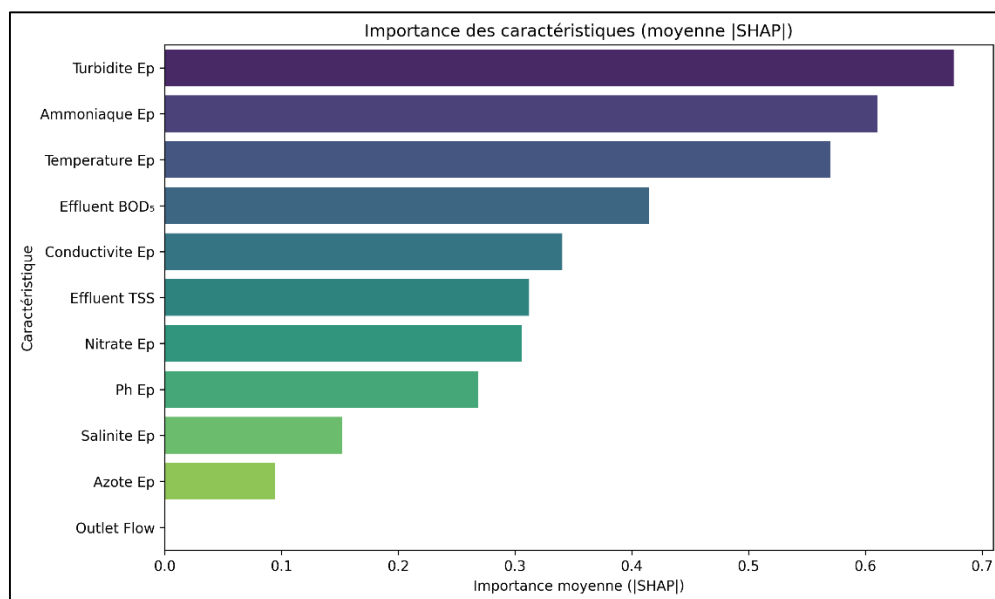
Zehraoui, R. (2011). *Évaluation d'épuration des eaux usées domestiques par boues activées, étude de la performance d'une station d'épuration de la wilaya de medea.*

Zhou, S., Liu, Z., Wang, M., et al. (2022). Impacts of building configurations on urban stormwater management at a block scale using XGBoost. *Sustainable Cities and Society*, 87, 104235. <https://doi.org/10.1016/j.scs.2022.104235>

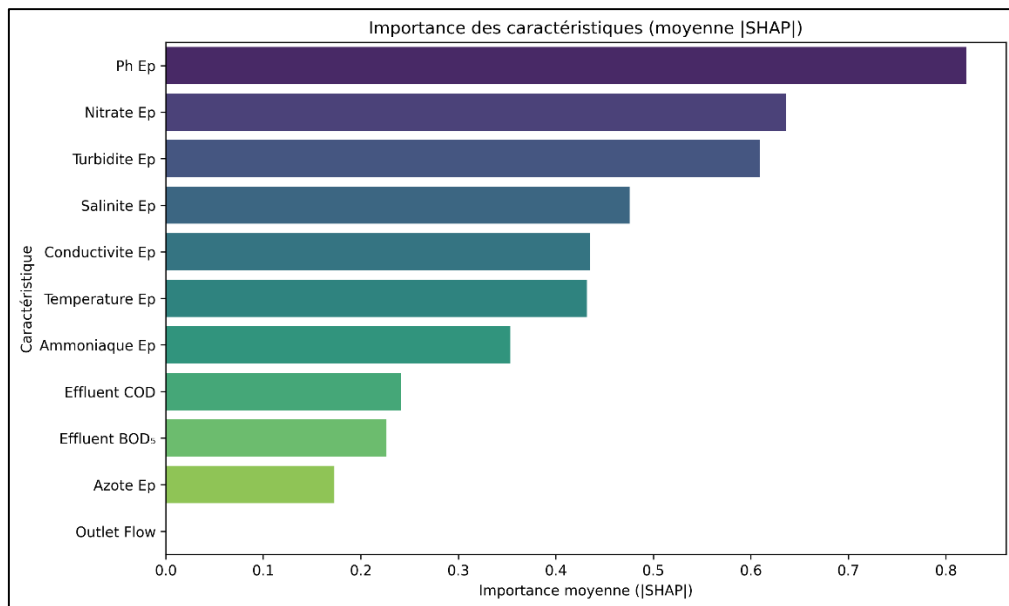
ANNEXE



A1. Importance des variables selon les valeurs SHAP moyennes de la DBO_5_{eb}



A2. Importance des variables selon les valeurs SHAP moyennes de la DCO_{ep}



A3. Importance des variables selon les valeurs SHAP moyennes des MES_ep

