Université de Strasbourg

Master d'informatique

Parcours Science de Données et Système Complexe

Entreposage et protection des données

JAIDANE Chaïma
OBERHAUSER Clément

Année universitaire 2024-2025

Table des matières

1 Introduction

L'évaluation de la qualité de l'eau est une problématique centrale dans la gestion des ressources, notamment pour garantir la préservation des écosystèmes et la santé publique. Dans ce cadre, l'indice I2M2 permet de mesurer la qualité des stations d'eau de surface. Par ailleurs, les paramètres physico-chimiques offrent des indications sur l'état des cours d'eau, bien qu'il soit difficile d'en tirer des conclusions directes sur la qualité globale.

L'objectif du projet est d'explorer la relation entre les caractéristiques physico-chimiques de l'eau et son état biologique. Il faut donc identifier les paramètres les plus influents dans ces prédictions, afin de mieux comprendre les facteurs déterminants de la qualité de l'eau. Pour ce faire nous avons suivi un axe de reflexion détaillé dans ce rapport.

Ce rapport est structuré en plusieurs sections. La première présente les données utilisées et les simplifications méthodologiques adoptées. Elle explicitera aussi la direction que nous avons décidé de prendre pour répondre à la problématique. La deuxième détaille les étapes de prétraitement des données, depuis le nettoyage jusqu'à la création des datasets. La troisième section traite la modélisation et l'analyse, notamment via l'algorithme XGBoost. Enfin, les résultats sont discutés avant de conclure sur les perspectives offertes par ce travail.

2 Présentation des données et méthodologie

2.1 Sources des données (SANDRE)

Le projet repose sur l'exploitation de données provenant de la base de données française sur les eaux : le Système d'Information sur les Ressources en Eau (SANDRE).

SANDRE : Le SANDRE (Service d'Administration Nationale des Données et Référentiels sur l'Eau) est une référence nationale pour la gestion des données relatives aux eaux superficielles et souterraines en France. Il fournit des jeux de données standardisés et fiables qui incluent des informations sur les stations de mesure (localisation, type d'eau, coordonnées géographiques) ainsi que des relevés de paramètres physicochimiques et hydrobiologiques essentiels à l'évaluation de la qualité de l'eau.

- **Dataset hydrobiologique :** Ce jeu de données contient des informations sur les prélèvements biologiques effectués sur les stations. Ces données permettent de classer les stations en fonction de leur qualité écologique (bonne ou mauvaise), sur la base des indicateurs biologiques relevés, tels que les macro-invertébrés ou les diatomées (uniquement macro-invertébrés I2M2 dans notre jeu de données).
- **Dataset physico-chimique :** Ce jeu de données regroupe les mesures des paramètres physicochimiques de l'eau (nitrates, DBO5, phosphore, oxygène dissous, etc.). Ces paramètres sont utilisés pour effectuer des prédictions sur la qualité biologique des stations, en établissant des liens entre ces mesures physicochimiques et les indicateurs biologiques.
- **Dataset des stations :** Ce dataset fournit des informations détaillées sur les stations de mesure, notamment leurs coordonnées géographiques, leur type (eau de surface ou souterraine) et leur localisation administrative. Il permet de relier les données physicochimiques et hydrobiologiques et d'attribuer chaque station à une hydroécorégion spécifique, ce qui est crucial pour les analyses géographiques et régionales.

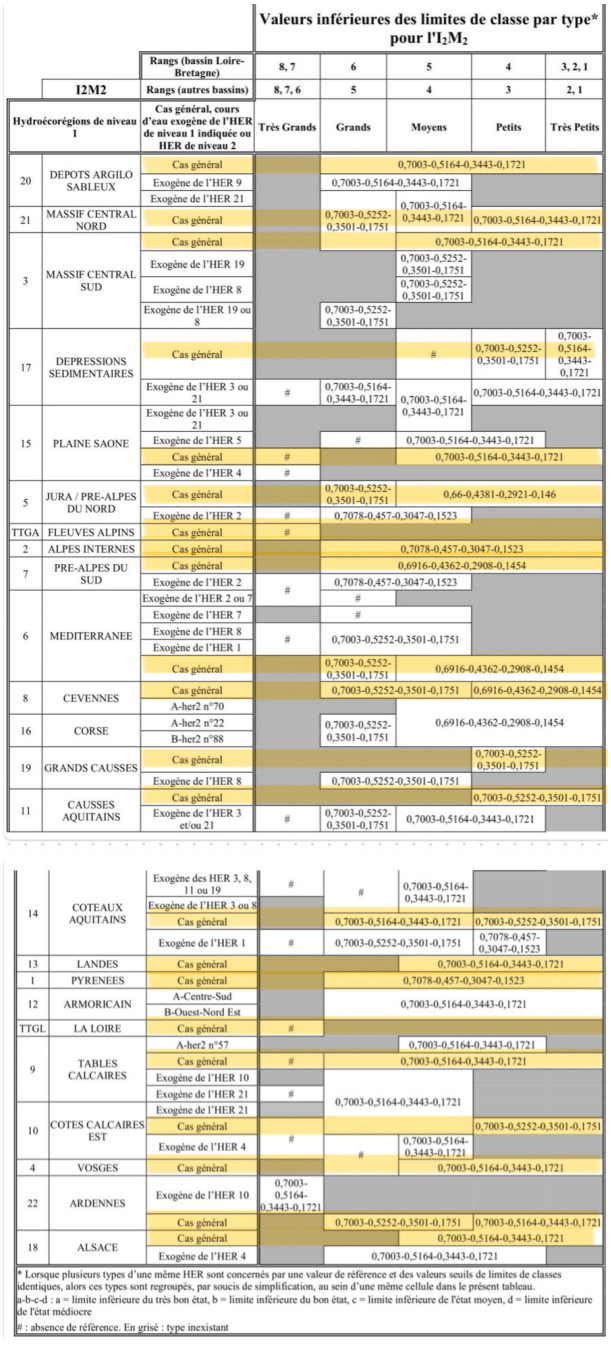
La combinaison de ces sources de données nous permet de croiser les informations sur les paramètres physicochimiques avec des indices de qualité et ainsi de répondre à notre objectif : prédire et analyser la qualité de l'eau en fonction de ces paramètres.

2.2 Description des paramètres analysés

2.2.1 Paramètres Hydrobiologiques

Comme nous l'avons mentionné précédemment, les données hydrobiologiques sont représentées par la valeur d'I2M2 d'après le tableau 1.

— **I2M2 :** L'indice I2M2 est utilisé pour évaluer la qualité de l'eau en se basant sur un ensemble de paramètres physico-chimiques et biologiques mesurés dans les stations de suivi. Cet indice, qui combine plusieurs critères environnementaux, permet de classer les stations en fonction de la qualité de l'eau qu'elles mesurent. Il est essentiel pour les gestionnaires de ressources en eau, car il fournit une vue d'ensemble de la santé des écosystèmes aquatiques.



			Valeurs inférieures des limites de classe par type* pour l'I_2M_2				
		Rangs (bassin Loire-Bretagne)	8, 7	6	5	4	3, 2, 1
	I2M2	Rangs (autres bassins)	8, 7, 6	5	4	3	2, 1
Hydroécorégions de niveau 1		Cas général, cours d'eau exogène de l'HER de niveau 1 indiquée ou HER de niveau 2	Très Grands	Grands	Moyens	Petits	Très Petits
20	DEPOTS ARGILO SABLEUX	Cas général	0,7003-0,5164-0,3443-0,1721				
		Exogène de l'HER 9		0,7003-0,5164-0,3443-0,1721			
		Exogène de l'HER 21			0,7003-0,5164-0,3443-0,1721		
21	MASSIF CENTRAL NORD	Cas général		0,7003-0,5252-0,3501-0,1751		0,7003-0,5164-0,3443-0,1721	
3	MASSIF CENTRAL SUD	Cas général			0,7003-0,5164-0,3443-0,1721		
		Exogène de l'HER 19			0,7003-0,5252-0,3501-0,1751		
		Exogène de l'HER 8			0,7003-0,5252-0,3501-0,1751		
		Exogène de l'HER 19 ou 8		0,7003-0,5252-0,3501-0,1751			
17	DEPRESSIONS SEDIMENTAIRES	Cas général			#	0,7003-0,5252-0,3501-0,1751	0,7003-0,5164-0,3443-0,1721
		Exogène de l'HER 3 ou 21	#	0,7003-0,5164-0,3443-0,1721		0,7003-0,5164-0,3443-0,1721	
15	PLAINE SAONE	Exogène de l'HER 3 ou 21			0,7003-0,5164-0,3443-0,1721		
		Exogène de l'HER 5		#	0,7003-0,5164-0,3443-0,1721		
		Cas général	#		0,7003-0,5164-0,3443-0,1721		
		Exogène de l'HER 4	#				
5	JURA / PRE-ALPES DU NORD	Cas général		0,7003-0,5252-0,3501-0,1751	0,66-0,4381-0,2921-0,146		
		Exogène de l'HER 2	#	0,7078-0,457-0,3047-0,1523			
TTGA	FLEUVES ALPINS	Cas général	#				
2	ALPES INTERNES	Cas général	0,7078-0,457-0,3047-0,1523				
7	PRE-ALPES DU SUD	Cas général	0,6916-0,4362-0,2908-0,1454				
		Exogène de l'HER 2	#	0,7078-0,457-0,3047-0,1523			
6	MEDITERRANEE	Exogène de l'HER 2 ou 7		#			
		Exogène de l'HER 7		#			
		Exogène de l'HER 8	#	0,7003-0,5252-0,3501-0,1751			
		Exogène de l'HER 1					
		Cas général		0,7003-0,5252-0,3501-0,1751	0,6916-0,4362-0,2908-0,1454		
8	CEVENNES	Cas général		0,7003-0,5252-0,3501-0,1751		0,6916-0,4362-0,2908-0,1454	
16	CORSE	A-her2 n°70		0,7003-0,5252-0,3501-0,1751	0,6916-0,4362-0,2908-0,1454		
		A-her2 n°22					
		B-her2 n°88					
19	GRANDS CAUSSES	Cas général				0,7003-0,5252-0,3501-0,1751	
		Exogène de l'HER 8		0,7003-0,5252-0,3501-0,1751			
11	CAUSSES AQUITAINS	Cas général				0,7003-0,5252-0,3501-0,1751	
		Exogène de l'HER 3 et/ou 21	#	0,7003-0,5252-0,3501-0,1751	0,7003-0,5164-0,3443-0,1721		
14	COTEAUX AQUITAINS	Exogène des HER 3, 8, 11 ou 19	#	#	0,7003-0,5164-0,3443-0,1721		
		Exogène de l'HER 3 ou 8					
		Cas général		0,7003-0,5164-0,3443-0,1721		0,7003-0,5252-0,3501-0,1751	
		Exogène de l'HER 1	#	0,7003-0,5252-0,3501-0,1751		0,7078-0,457-0,3047-0,1523	
13	LANDES	Cas général			0,7003-0,5164-0,3443-0,1721		
1	PYRENEES	Cas général		0,7078-0,457-0,3047-0,1523			
12	ARMORICAIN	A-Centre-Sud		0,7003-0,5164-0,3443-0,1721			
		B-Ouest-Nord Est					
TTGL	LA LOIRE	Cas général	#				
9	TABLES CALCAIRES	A-her2 n°57			0,7003-0,5164-0,3443-0,1721		
		Cas général	#	0,7003-0,5164-0,3443-0,1721			
		Exogène de l'HER 10					
		Exogène de l'HER 21	#	0,7003-0,5164-0,3443-0,1721			
		Exogène de l'HER 21					
10	COTES CALCAIRES EST	Cas général				0,7003-0,5252-0,3501-0,1751	
		Exogène de l'HER 4	#	#	0,7003-0,5164-0,3443-0,1721		
4	VOSGES	Cas général			0,7003-0,5164-0,3443-0,1721		
22	ARDENNES	Exogène de l'HER 10	0,7003-0,5164-0,3443-0,1721				
		Cas général		0,7003-0,5252-0,3501-0,1751		0,7003-0,5164-0,3443-0,1721	
18	ALSACE	Cas général			0,7003-0,5164-0,3443-0,1721		
		Exogène de l'HER 4		0,7003-0,5164-0,3443-0,1721			

* Lorsque plusieurs types d'une même HER sont concernés par une valeur de référence et des valeurs seuils de limites de classes identiques, alors ces types sont regroupés, par soucis de simplification, au sein d'une même cellule dans le présent tableau.
a-b-c-d : a = limite inférieure du très bon état, b = limite inférieure du bon état, c = limite inférieure de l'état moyen, d = limite inférieure de l'état médiocre
: absence de référence. En grisé : type inexistant

TABLE 1 – Valeurs limites inférieures des classes d'état en EQR par type de cours d'eau pour l'indice I2M2. Surligné en jaune sont les cas généraux de chaque HER.

2.2.2 Paramètres Physico-chimiques

Les données physicochimiques collectées dans le cadre de ce projet concernent plusieurs paramètres importants pour l'évaluation de la qualité de l'eau. Nous avons choisi de travailler avec un nombre réduit de ces paramètres afin de simplifier l'analyse tout en conservant des éléments significatifs pour la prédiction de la qualité de l'eau. Les principaux paramètres utilisés sont :

— **Nitrates :** Les nitrates sont des composés chimiques qui proviennent principalement des fertilisants agricoles. Leur présence en grande quantité dans les cours d'eau peut entraîner des phénomènes d'eutrophisation et nuire à la qualité de l'eau.

— **DBO5 (Demande Biochimique en Oxygène sur 5 jours) :** La DBO5 est un indicateur de la quantité d'oxygène consommée par les micro-organismes pour décomposer la matière organique dans l'eau. Un niveau élevé de DBO5 peut indiquer une pollution organique importante.

— **Phosphore :** Le phosphore, également un nutriment essentiel pour la croissance des plantes, peut causer de l'eutrophisation en excès, entraînant la prolifération d'algues dans les milieux aquatiques.

— **Oxygène Dissous :** L'oxygène dissous dans l'eau est vital pour la survie des organismes aquatiques. Un niveau trop faible peut entraîner la mort des espèces aquatiques et affecter l'équilibre écologique de l'écosystème.

Ces paramètres sont essentiels pour évaluer la qualité de l'eau et sont souvent utilisés dans les indices comme l'I2M2 pour établir des classifications

2.3 Choix méthodologiques

Pour simplifier l'analyse et rendre les résultats plus exploitables, plusieurs choix méthodologiques ont été adoptés. Ces simplifications visent à réduire la complexité des données tout en conservant les informations pertinentes.

1. **Sélection des données sur les Hydroécorégions (HER) :** Une hydroécorégion regroupe un ensemble de stations de mesure situées dans un même bassin versant ou une zone géographique homogène. Les stations ont été catégorisées en fonction de la qualité de l'eau selon les seuils de l'I2M2. Nous avons choisi de limiter l'analyse aux Hydroécorégions 2, 7, 11, 13, 15, 19, 20 et 9 définies dans la figure 1. Pour réduire la compléxité et le manque d'information dans le dataset nous avons choisi de nous arrêter au cas généraux des hydroécorégions définies précédents. Les valeurs de seuils associées sont représentés dans le tableau 2.

Numéro HER	Très bon état	État moyen	État médiocre	Mauvais état
2	0.7078	0.4570	0.3047	0.1523
7	0.6916	0.4362	0.2908	0.1454
9	0.7003	0.5164	0.3443	0.1721
11	0.7003	0.5252	0.3501	0.1751
13	0.7003	0.5252	0.3501	0.1751
15	0.7003	0.5164	0.3443	0.1721
19	0.7003	0.5252	0.3501	0.1751
20	0.7003	0.5164	0.3443	0.1721

TABLE 2 – Tableau des données par Hydroécorégions et seuils correspondants.

2. **Période d'analyse :** Pour limiter la quantité de données tout en assurant une diversité temporelle suffisante, nous avons concentré notre analyse sur la période de 2015 à 2018. Cette période est représentative des tendances récentes de la qualité de l'eau tout en limitant les biais liés à des données trop anciennes. Ce choix a été motivé par l'analyse des données, révélant que ces années présentent le plus grand nombre de valeurs pour chaque paramètre.
3. **Réduction du nombre de paramètres :** Afin de simplifier la modélisation, seuls quatre paramètres physico-chimiques ont été retenus : les nitrates, la DBO5, le phosphore et l'oxygène dissous. Ils ont été choisi aléatoirement pour parmi les paramètres.
4. **Agrégation des données par saison :** Étant donné que la qualité de l'eau peut varier en fonction des saisons, nous avons choisi d'agréger les données physico-chimiques par saison (été, hiver, printemps, automne). Cette approche permet de mieux capturer les effets saisonniers sur la qualité de l'eau, qui peuvent influencer la performance des modèles de prédiction.
5. **Calcul des médianes :** Plutôt que d'utiliser les valeurs brutes des mesures de chaque station et de chaque relevés, nous avons calculé les médianes des relevés de chaque station sur la période choisie (2015-2018), par saison. Cette méthode permet de réduire l'impact des valeurs aberrantes et de rendre les données plus représentatives de l'évolution générale de la qualité de l'eau.

2.4 Redéfinition de l'objectif

Afin de clarifier notre contribution dans ce projet et de mieux cibler la problématique, nous avons redéfini l'objectif comme suit :

Concevoir un modèle d'intelligence artificielle capable de prédire la qualité des stations en fonction de l'indice I2M2, à partir des données physicochimiques. L'objectif inclut également une analyse visant à identifier les paramètres physicochimiques ayant le plus d'impact sur les prédictions, afin de mettre en évidence les facteurs essentiels influençant la qualité de l'eau.

Cette redéfinition permet de structurer notre approche autour des étapes clés suivantes :

1. **Prétraitement des données** : Nettoyage et préparation des données pour l'analyse.
2. **Visualisation et sélection des paramètres** : Analyse exploratoire pour identifier les relations entre les paramètres et la qualité des stations.
3. **Entraînement des modèles d'intelligence artificielle** : Application de modèles supervisés pour effectuer les prédictions.
4. **Analyse des résultats** : Évaluation des performances des modèles et identification des paramètres les plus influents.

3 Prétraitement des Données

Le nettoyage et la préparation des données constituent une étape fondamentale pour garantir la qualité et la pertinence des analyses.

3.1 Nettoyage et transformation

3.1.1 Paramètres Hydrobiologiques

Analyse des colonnes avec des valeurs manquantes

Comme les noms de colonnes n'avaient pas à être changé, nous avons commencé par vérifier la quantité de données manquantes pour supprimer les colonnes si besoin. Les données manquantes sont présentent dans le tableau 3

Index	Valeurs Manquantes
CdStationMesureEauxSurface	False
LbStationMesureEauxSurface	False
DateDebutOperationPrelBio	False
CdSupport	False
LbSupport	False
CdProducteur	False
CdParametreResultatBiologique	False
LbLongParametre	False
RefOperationPrelBio	False
CdUniteMesure	False
SymUniteMesure	False
CdRqIndiceResultatBiologique	False
MnemoRqAna	False
NomProducteur	True
ResIndiceResultatBiologique	True
CdAccredRsIndiceResultatBiologique	True
DtProdResultatBiologique	True
CdPointEauxSurf	True
CdMethEval	True
MnAccredRsIndiceResultatBiologique	True

TABLE 3 – Données manquantes dans le dataset hydrobiologique : True pour les valeurs manquantes, False pour les valeurs présentes.

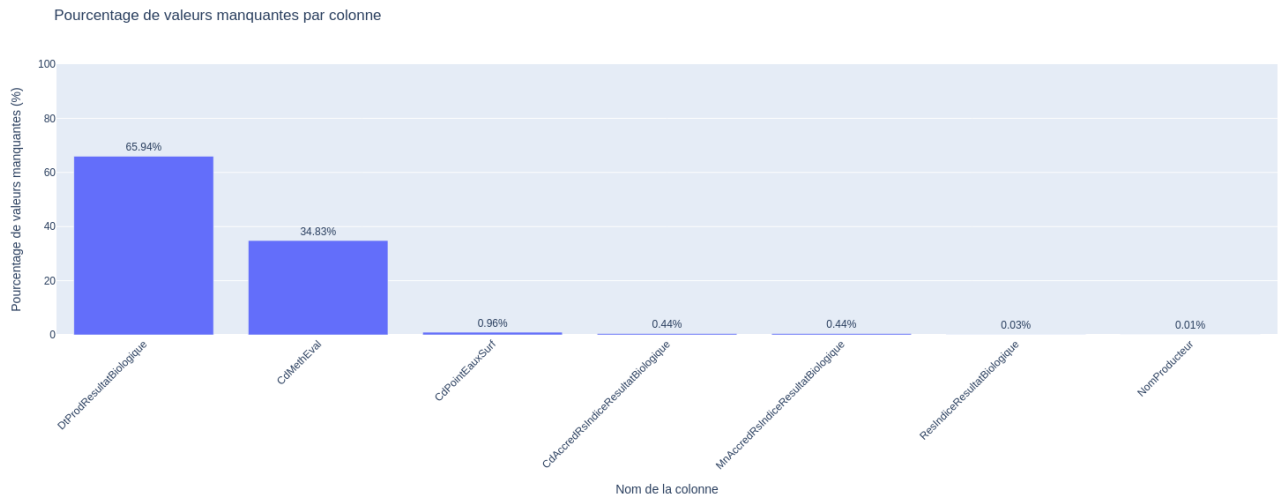


FIGURE 1 – Proportion des valeurs manquantes pour les paramètres qui ont des valeurs manquantes dans le dataset Hydrobiologiques

Certaines colonnes du dataset contiennent des valeurs manquantes. Cette section présente une analyse de certaine colonne et de leur importance et de leur impact sur les analyses effectuées.

— **DtProdResultatBiologique** : Cette colonne précise le jour, le mois et l'année où le résultat biologique a été calculé. Toutefois, ce paramètre n'est pas pertinent pour notre analyse, car nous utilisons la date de début des opérations biologiques comme référence temporelle. De plus, avec plus de 60% de valeurs manquantes, cette colonne est difficilement exploitable.

— **MnAccredRsIndiceResultatBiologique** et **CdAccredRsIndiceResultatBiologique** : Ces colonnes fournissent des informations sur l'accréditation des analyses biologiques. Les valeurs possibles pour ces paramètres sont :

1. *Analyse réalisée sous accréditation* : Analyse effectuée par un laboratoire accrédité, en conformité avec la norme ISO 17025 et les exigences du Comité Français d'Accréditation (COFRAC).
2. *Analyse réalisée hors accréditation* : Analyse réalisée sans accréditation formelle ou dans des conditions où l'accréditation n'est pas applicable.
3. *Inconnu* : Les conditions d'accréditation sont inconnues.

Afin de ne pas réduire la taille du dataset (il resterait 8428 lignes si seules les analyses accréditées étaient conservées), nous choisissons de maintenir toutes les lignes, indépendamment des conditions d'accréditation. Cependant, cette information pourra être intégrée dans des analyses futures plus approfondies.Source pour les accréditations Sandre

— **CdPointEauxSurf** : Cette colonne identifie un point de prélèvement spécifique au sein d'une station. Bien qu'elle puisse être utile pour reproduire les prélèvements par les organismes responsables, elle n'a aucun impact direct sur les analyses menées dans ce travail.

— **NomProducteur** : Ce champ indique le nom de l'intervenant ayant réalisé l'opération de prélèvement biologique. Cependant, il n'apporte aucune valeur ajoutée pour l'analyse.

— **ResIndiceResultatBiologique** : Cette colonne est essentielle car elle contient les résultats des indices biologiques qui sont au cœur de notre analyse. Les 13 lignes avec des valeurs manquantes (NaN) dans cette colonne sont supprimées afin de garantir l'intégrité des analyses.

Traitement des données manquantes

Parmi les données manquantes, il est essentiel d'évaluer le pourcentage d'absence dans l'ensemble de données afin de déterminer la meilleure stratégie de traitement. Le graphique 3 permet de prendre la décision. Les colonnes jugées non essentielles ou avec un taux élevé de valeurs manquantes sont exclues ou ignorées dans le cadre de cette étude, sauf pour **ResIndiceResultatBiologique**, où les lignes contenant des valeurs NaN sont supprimées.

Analyse des colonnes avec des valeurs complètes

Dans notre dataset, de nombreuses données ne sont pas pertinentes. Il faut supprimer les colonnes inutiles, même pour les valeurs qui n'ont pas de données manquantes. Encore une fois, pour ce faire il faut comprendre l'utilité de chaque paramètre.

Sélection des colonnes avec valeurs non manquantes

Les colonnes suivantes contiennent des valeurs non manquantes. Nous analysons leur pertinence et leur utilité dans le cadre de notre étude.

- **CdStationMesureEauxSurface** : Cette colonne identifie la station de mesure et permet de lier les données avec une station spécifique.
- **LbStationMesureEauxSurface** : Contient des descriptions des stations. Ces informations peuvent être retrouvées ailleurs et ne sont donc pas pertinentes pour cette étude.
- **DateDebutOperationPrelBio** : Colonne utilisée pour dater les prélèvements.
- **CdSupport** : Identifie le composant du milieu étudié, faisant généralement l'objet de prélèvements pour évaluer la qualité du milieu.
- **CdParametreResultatBiologique** : Ne contient qu'une seule valeur (7613) pour toutes les lignes, ce qui n'apporte aucune information.
- **LbLongParametre** : Contient le nom du paramètre testé, avec une valeur unique "Indice Invertébrés Multimétrique (I2M2)" dans toutes les lignes.
- **RefOperationPrelBio** : Contient un identifiant unique pour chaque opération de prélèvement.
- **CdUniteMesure** : Nécessite une vérification pour déterminer son rôle exact.
- **SymUniteMesure** : Semble être liée à **CdUniteMesure**. Une vérification conjointe est nécessaire pour évaluer son utilité.
- **CdRqIndiceResultatBiologique** : Cette colonne contient des codes indiquant l'état du résultat attendu. Parmi ces résultats, 13 valeurs (code `0`) ne sont pas utilisables et doivent être supprimées. Les détails des codes sont indiqués dans le tableau 4.
- **MnemoRqAna** : Colonne similaire à **CdRqIndiceResultatBiologique**, nécessitant le même traitement.

Code	Mnémonique	Libellé
0	Analyse non faite	Analyse non faite
1	Domaine de validité	Résultat supérieur au seuil de quantification et inférieur au seuil de saturation, ou résultat égal à zéro

Table 4 – Nomenclature des codes pour `CdRqIndiceResultatBiologique` et `MnemoRqAna`.

Traitement des colonnes sélectionnées

Les colonnes jugées inutiles ou avec une information redondante (LbStationMesureEauxSurface, CdParametreResultatBiologique, LbLongParametre) sont supprimées. Les colonnes essentielles (CdStationMesureEauxSurface, DateDebutOperationPrelBio, CdSupport, RefOperationPrelBio, CdRqIndiceResultatBiologique, MnemoRqAna) sont conservées après traitement des valeurs manquantes.

3.1.2 Paramètres Physicochimiques

Données au cours du temps

Pour les paramètres physico-chimiques, nous avons appliqué la même démarche. Cependant, nous avons d'abord traité les données en fonction du temps. Deux nouvelles colonnes ont ainsi été ajoutées : **Saison** et **LbSaison**, basées sur le numéro du mois. La répartition des données par saison est présentée dans la figure 4, qui a été réalisée après le nettoyage des données.

Une fois cela effectué, nous avons décidé de visualiser la quantité de données par année afin de simplifier le jeu de données en ne conservant que les années les plus représentatives. Comme le montre la figure 2, la majorité des données proviennent des années 2015 à 2018. Nous avons donc choisi de retenir ces années.

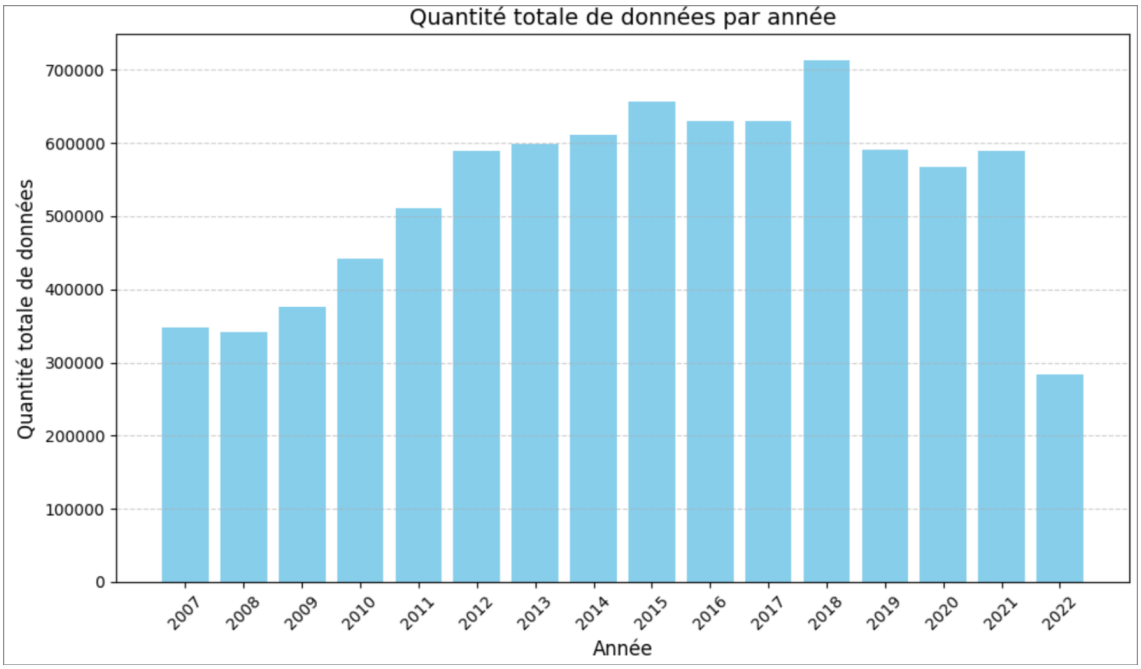


FIGURE 2 – Nombre de valeurs dans le dataset Physicochimique par années.

Choix des paramètres

Pour confirmer notre choix de sélectionner les années de 2015 à 2018, nous avons analysé le tableau 3, qui illustre le pourcentage de données disponibles pour chaque année. Cette visualisation nous a permis de constater que les années sélectionnées étaient effectivement pertinentes. Par ailleurs, nous avons observé une répartition relativement homogène des données dans l'ensemble du jeu de données.

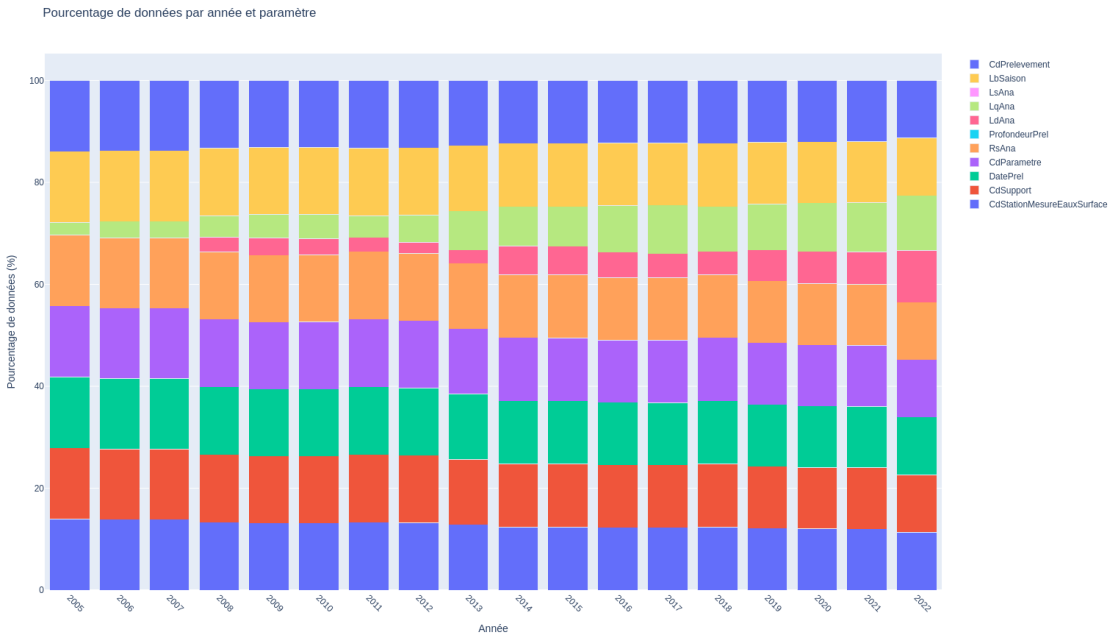


FIGURE 3 – Pourcentage de données par années et par paramètre.

Dans ce jeu de données, il y a très peu de valeurs manquantes, seulement un léger pourcentage dans RsAna, nous avons supprimé ces valeurs.

Sélection des paramètres retenus

Les paramètres suivants ont été sélectionnés pour leur pertinence et leur utilité dans le cadre de l'étude. De plus, ils permettent de courvir plusieurs dimensions essentielles. Une description détaillée est présentée ci-dessous.

Voici une reformulation de votre texte :

— **Spatiale** : `CdStationMesureEauxSurface`, identifiant unique attribué à chaque station de mesure dans le référentiel national **Sandre**. Ce paramètre est essentiel pour réaliser une jointure avec les fichiers contenant les informations sur les stations et les analyses.

— **Temporelle** : `DatePrel`, qui indique la date de début du prélèvement physico-chimique, avec une précision au jour près. Ce paramètre est fondamental pour situer les prélèvements dans le temps.

— **Qualitative et quantitative** : `RsAna`, qui contient le résultat de l'analyse physico-chimique, soit sous forme de valeur quantitative du paramètre, soit sous forme de code qualitatif. Ce paramètre est crucial pour l'analyse des données. `CdParametre`, quant à lui, définit une caractéristique du milieu, permettant d'évaluer sa qualité ou son aptitude à des usages spécifiques. Ce paramètre est indispensable pour l'analyse.

— **Gestion et traçabilité** : `CdPrelevement`, qui fournit une référence unique associée à chaque prélèvement physico-chimique et biologique, utile pour la gestion, le suivi et la traçabilité des prélèvements, notamment en ce qui concerne la facturation des prestations.

— **Contexte environnemental** : `CdSupport`, qui désigne un élément du milieu sur lequel porte l'investigation, souvent le support des prélèvements pour des analyses ultérieures. Ce paramètre est essentiel pour évaluer les caractéristiques et la qualité du milieu. `LbSaison` indique le nom de la saison au cours de laquelle le prélèvement a été effectué, fournissant ainsi une information utile pour des analyses saisonnières.

Les paramètres retenus permettent d'assurer une analyse cohérente et complète des données physico-chimiques tout en assurant une traçabilité des mesures. Les informations saisonnières (`LbSaison`) et les détails des prélèvements (`CdPrelevement`, `CdSupport`) complètent les résultats d'analyse (`RsAna`) pour fournir un cadre d'étude robuste et pertinent.

Nettoyage des données

Après avoir sélectionné les paramètres pour les années 2015 à 2018, il est nécessaire de supprimer les valeurs nulles et les doublons. Une fois cette étape réalisée, nous avons visualisé la répartition des données par saison, comme illustré dans la figure 4.

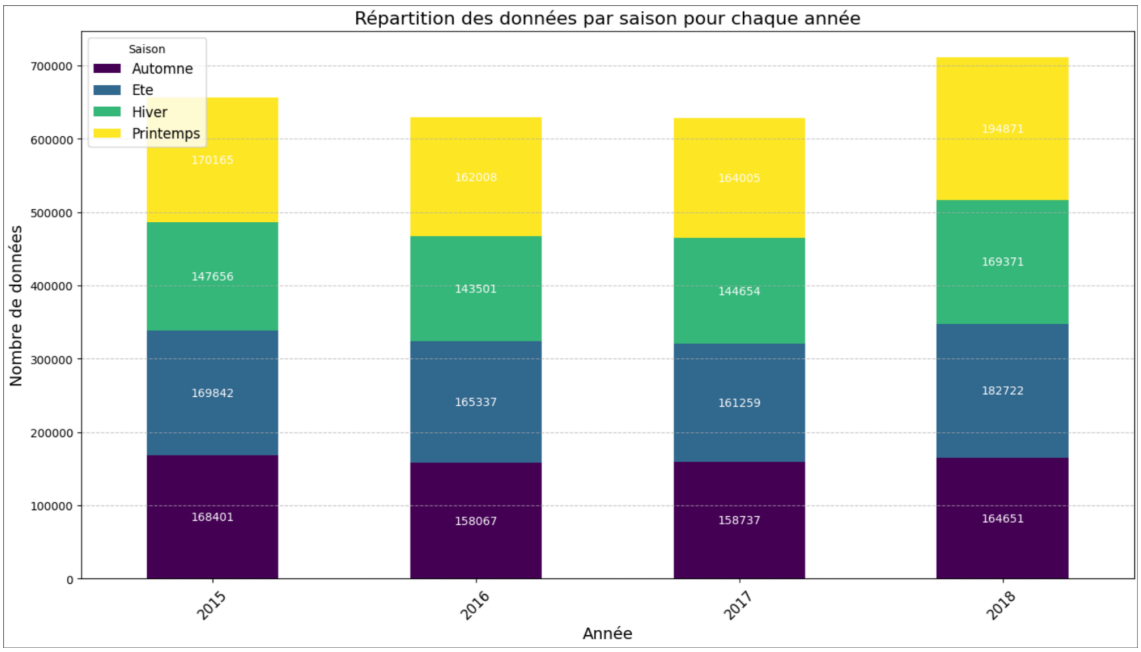


FIGURE 4 – Nombre de données par saison années et par paramètre.

Ce qui nous permet de voir que les données sont à peu près également réparties dans l'année. Ceci nous permet de travailler avec un bon jeu de données hétérogène et représentatif.

3.1.3 Traitement des Stations

Cette même démarche de traitement que nous avons réalisé sur les données hydrobiologiques et physico-chimique doit être réalisé sur les stations.

Stations et Hydroécorégions

Pour commencer notre analyse, nous avons choisi d'afficher les Hydroécorégions (HER) sur la carte de la France (Figure 5). Ce qui permet de comprendre visuellement comment les stations sont réparties sur le territoire nationale.

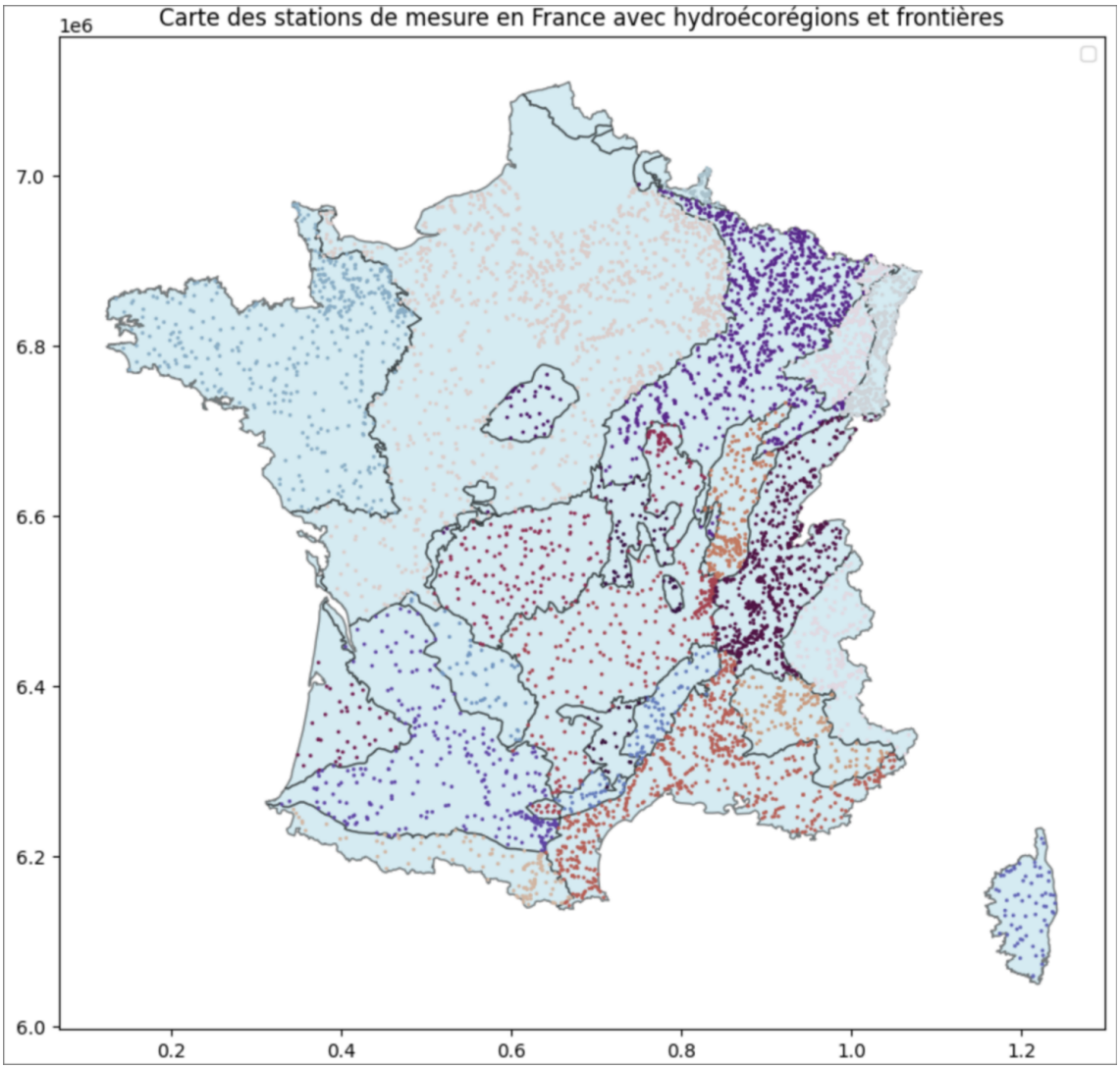


FIGURE 5 – Stations réparties sur le territoire national français (métropole uniquement) en fonction des différents HER.

Nous avons également utilisé GeoPandas, non seulement pour la visualisation, mais aussi pour classer les stations par HER dans le jeu de données.

Pertinence des stations

Pour débuter, nous avons choisi de réaliser des boxplots (figure 6) afin d'identifier d'éventuelles valeurs aberrantes dans les données. Les points situés en dehors des limites du boxplot sont considérés comme des valeurs aberrantes, que nous avons décidé de supprimer.

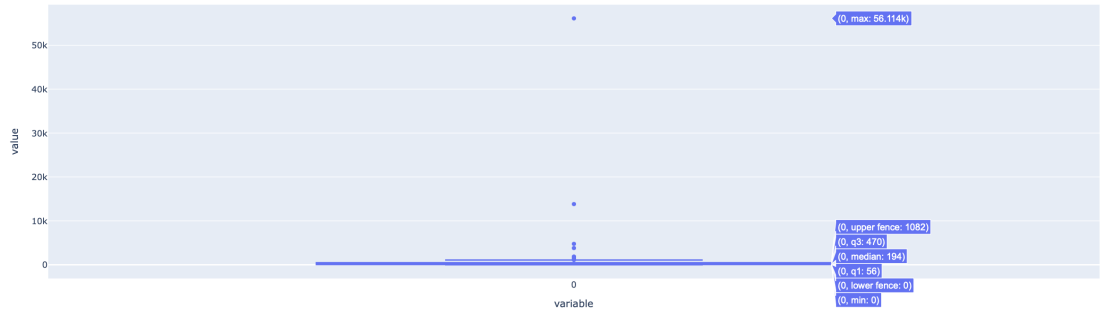


FIGURE 6 – Stations réparties sur le territoire national français (métropole uniquement) en fonction des différents HER.

Maintenant on peut se rendre compte avec la figure 7 que les données sont propres et sans valeurs abberentes dans le jeu de données.

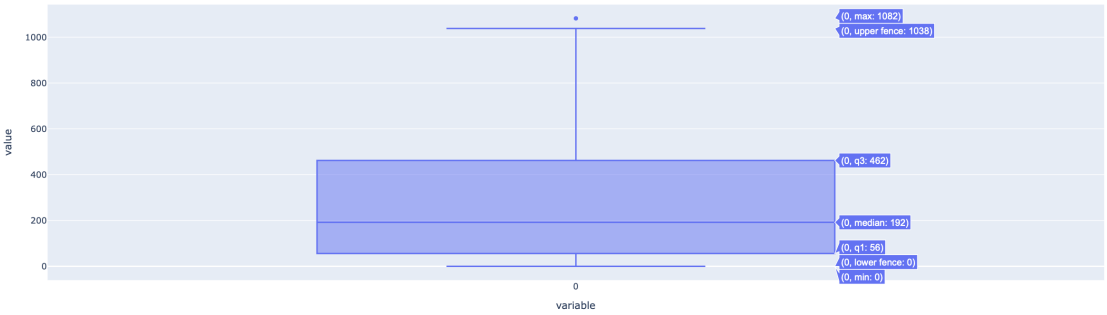


FIGURE 7 – Stations réparties sur le territoire national français (métropole uniquement) en fonction des différents HER.

Par la suite, il est nécessaire d'associer les paramètres physico-chimiques aux stations en fonction du temps. En effet, après avoir supprimé les valeurs aberrantes et exclu les années non retenues, certaines stations ne présentaient plus aucun relevé. Avant ces suppressions, le jeu de données comptait 8810 observations, contre 7466 après traitement.

Choix des paramètres

Comme pour les sections précédentes, nous avons vérifié les valeurs manquantes pour chaque colonne, plutôt que pour le nombre de relevés. Dans le tableau 5, les valeurs manquantes sont indiquées par True.

Index	Valeurs Manquantes
CdStationMesureEauxSurface	False
LbStationMesureEauxSurface	False
CodeTypEthStationMesureEauxSurface	False
CoordXStationMesureEauxSurface	False
CoordYStationMesureEauxSurface	False
CdProjStationMesureEauxSurface	False
LibelleProjection	False
LibelleTypEthStationMesureEauxSurface	False
AltitudePointCaracteritisque	False
DateCreationStationMesureEauxSurface	False
DateMAJInfosStationMesureEauxSurface	False
ComStationMesureEauxSurface	True
DateArretActiviteStationMesureEauxSurface	True
CodeNatureStationMesureEauxSurface	True
LocPreciseStationMesureEauxSurface	True
NomCoursdEau	True
LibelleNatureStationMesureEauxSurface	True
PkPointTronconEntiteHydroPrincipale	True
PremierMoisAnneeEtiage	True
FinaliteStationMesureEauxSurface	True
CdCoursdEau	True
CdEuBassinDCE	True
NomEuBassinDCE	True
DurStationMesureEauxSurface	True
CodeCommune	True
LbCommune	True
CodeDepartement	True
LbDepartement	True
CodeRegion	True
CdTronconHydrographique	True
LbRegion	True
CdEuMasseDEau	True
NomMasseDEau	True
CdEuSsBassinDCEAdmin	True
NomSsBassinDCEAdmin	True
CdBassinDCE	True
SuperficieBassinVersantReel	True
CdMasseDEau	True
SuperficieBassinVersantTopo	True

TABLE 5 – Données manquantes dans le dataset des stations de mesure : True pour les valeurs manquantes, False pour les valeurs présentes.

Données avec valeurs manquantes

Nous avons d'abord examiné la signification et l'utilité de chaque donnée dans le jeu de données. Pour débuter, nous nous sommes concentrés sur les colonnes contenant des valeurs nulles.

- **SuperficieBassinVersantReel** : Cette colonne contient uniquement des valeurs nulles.
- **PremierMoisAnneeEtiage** : Le premier mois de l'année d'étiage est le numéro dans l'année civile du premier mois de la période utilisée pour les études statistiques sur les basses eaux. Dans le cas de notre étude, ce paramètre ne sera pas utilisé pour le moment.
- **SuperficieBassinVersantTopo** : Seulement 259 valeurs sont non nulles. Cette colonne apporte trop peu d'information pour être utilisable.
- **ComStationMesureEauxSurface** : Les valeurs textuelles sont trop fluctuantes pour être analysées.

— **FinaliteStationMesureEauxSurface** : La finalité de la station constitue le but pour lequel la station de mesure a été créée. Bien que certaines valeurs comme "Impact d'un rejet domestique" ou "Impact d'un rejet d'élevage" pourraient orienter l'analyse des stations affectées, cette colonne contient de nombreuses lignes inutiles.

— **DateArrStationMesureEauxSurface** : La date d'arrêt d'activité de la station de mesure est la date à laquelle cessent les opérations de prélèvement.

— **DurStationMesureEauxSurface** : Cette donnée représente la dureté moyenne estimée d'après les analyses d'eau sur les stations du tronçon hydrographique. Bien qu'elle puisse être utile pour des analyses futures, elle n'est pas pertinente pour l'analyse actuelle.

— **PkPointTronconEntiteHydroPrincipale** : Cette mesure localise la station sur le tronçon hydrographique.

— **LibelleNatureStationMesureEauxSurface** et **CodeNatureStationMesureEauxSurface** : Ces colonnes indiquent si la mesure est manuelle ou automatique.

— **NomCoursdEau** : Cette colonne donne le nom du cours d'eau de la station. Cependant, 20 % des lignes n'ont pas de nom.

— **CdTronconHydrographique** : Ce paramètre fournit une indication précise sur la localisation.

— **CdCoursdEau** : Le code identifie l'entité hydrographique associée à la station.

— **CdBassinDCE**, **NomEuBassinDCE**, **CdEuSsBassinDCEAdmin**, et **CdEuBassinDCE** : Ces colonnes décrivent un bassin DCE, qui correspond à un district hydrographique ou une portion située sur un territoire d'État membre.

— **NomMasseDEau**, **CdMasseDEau**, et **CdEuMasseDEau** : Ces colonnes fournissent des informations sur la masse d'eau associée à la station.

— Pour toutes les autres colonnes liées à la localisation `LocPreciseStationMesureEauxSurface, LbRegion, CodeRegion, CodeCommune, LbCommune, LbDepartement, CodeDepartement`, nous n'auront pas besoin de ces informations pour réaliser les hydroécoregions .

Nous allons faire de même pour les données du tableau 5 qui n'ont pas de valeurs nulles.

Données sans valeurs manquantes

— **CdStationMesureEauxSurface** : Code pour identifier une station.

— **LbStationMesureEauxSurface** : Utile pour faciliter la visualisation des données et la localisation de la station.

— **CodeTypEthStationMesureEauxSurface** : Nature (cours d'eau ou plan d'eau) de l'entité hydrographique sur laquelle la station de mesure est localisée.

— **CoordXStationMesureEauxSurface** et **CoordYStationMesureEauxSurface** :

— **CdProjStationMesureEauxSurface** : Uniquement des valeurs 26 ("Lambert 93"). N'apporte pas d'information.

— **LibelleProjection** : Comme la colonne précédente, ne contient que les valeurs "'RGF93 / Lambert 93"'.

— **LibelleTypEthStationMesureEauxSurface** : Permet de connaître la position d'une station suivant deux valeurs : sur un cours d'eau ou sur un plan d'eau.

— **AltitudePointCaracteritisque** : Altitude de la station. Peut avoir un impact sur les résultats d'analyse.

— **DateCreationStationMesureEauxSurface** : Date à laquelle la station a été créée.

— **DateMAJInfosStationMesureEauxSurface** : Comme pour la colonne précédente.

Données retenues

Nous avons retenu plusieurs colonnes essentielles pour l'analyse. La colonne **CdStationMesureEauxSurface** permet d'identifier chaque station, tandis que **LbStationMesureEauxSurface** fournit le nom de la station, facilitant ainsi la visualisation et la localisation des données. Les coordonnées X et Y de la station (**CoordXStationMesureEauxSurface** et **CoordYStationMesureEauxSurface**) sont nécessaires pour déterminer les hydroécorégions. Le **CdEuMasseDEau** permet d'identifier la masse

d'eau européenne, et **NomSsBassinDCEAdmin** fait référence au sous-bassin administratif selon la Directive Cadre sur l'Eau (DCE), tandis que **CdEuBassinDCE** identifie le bassin DCE.

De plus, la colonne **CodeTypEthStationMesureEauxSurface** renseigne sur la nature de l'entité hydrographique où la station est située, et **DateArretActiviteStationMesureEauxSurface** indique la date d'arrêt des activités de la station. **AltitudePointCaracteritisque** est également retenue, car l'altitude de la station peut influencer les résultats des analyses. La colonne **PremierMoisAnneeEtiage** donne des informations sur le premier mois et l'année d'étiage (basses eaux), et **geometry** contient la géométrie associée à la station, utile pour les analyses géospatiales. Enfin, l'indice de jointure après une opération géographique est stocké dans la colonne **index_right**, tandis que **gid** représente un identifiant unique généré pour chaque station. Les colonnes **CdHER1** et **NomHER1** fournissent respectivement le code et le nom de l'hydroécorégion associée. Ces colonnes ont été conservées car elles sont indispensables pour une analyse complète et précise des données environnementales et géospatiales.

4 Préparation du dataset final

Pour atteindre cet objectif, il est nécessaire d'organiser les différents jeux de données afin de créer un jeu de données unique à utiliser pour le modèle.

4.1 Clustering des stations

Pour créer ce jeu de données, les relevés sont d'abord classés en deux catégories : *plutôt bonnes* et *plutôt mauvaises*, qui serviront de labels à prédire avec le modèle. Afin de présenter la méthodologie, nous nous concentrerons sur les graphiques pour l'Hydroécorégion HER20. Cependant, cette démarche a été appliquée à toutes les hydroécorégions, comme détaillé dans la section Choix méthodologiques.

Tout d'abord, nous avons classé les relevés en fonction des valeurs présentées dans le tableau 2. Nous avons sélectionné uniquement les cas généraux. Pour l'Hydroécorégion HER20, les valeurs de seuils sont présentées dans le tableau 6.

Valeur	**État écologique**
$value > 0.7003$	Très bon état
$0.5164 < value \leq 0.7003$	Bon état
$0.3443 < value \leq 0.5164$	État moyen
$0.1721 < value \leq 0.3443$	État médiocre
$value \leq 0.1721$	Mauvais état

Table 6 – Classification des états écologiques en fonction de la valeur pour le HER20.

Pour chaque relevé, nous appliquerons cette classification. Une fois cela effectué, nous procéderons à un clustering des stations en fonction des relevés, en utilisant la méthode K-means. Ce clustering permettra de définir deux classes, comme mentionné précédemment : *plutôt bon* et *plutôt mauvais*. Les stations présentant une plus grande proportion de relevés à état écologique favorable seront classées comme *plutôt bon*, et inversement, celles avec des relevés moins favorables seront classées comme *plutôt mauvais*. La figure 8 illustre la classification obtenue par le clustering, où le groupe *plutôt bon* regroupe principalement des relevés avec un état écologique positif, tout comme le groupe *plutôt mauvais* pour les relevés moins favorables.

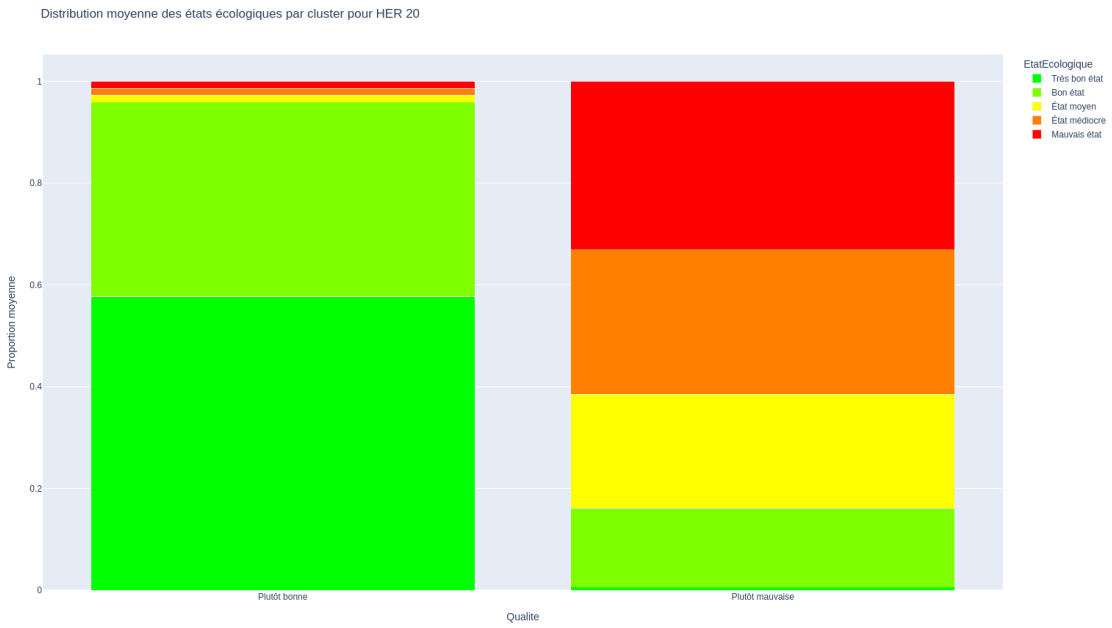


FIGURE 8 – Distribution moyenne des états écologiques par cluster (plutôt bon/plutôt mauvais) pour chaque cluster pour le HER20.

Ainsi pour chaque station, nous avons une qualité associée en fonction des proportions des états définis par les relevés. Une fois les différents dataset (un pour chaque HER) exportés, nous pouvons commencer à les exploiter avec les données physicochimiques.

4.2 Intégration aux données physicochimiques

Une fois les stations classées, il est nécessaire de traiter les données physico-chimiques. Pour cela, nous avons décidé de nous concentrer sur les paramètres de **nitrates, DBO5, phosphore et oxygène dissous**. Ces choix nous permettent de simplifier le problème tout en tenant compte de l'impact significatif de ces paramètres sur la qualité de l'eau, décision prise à la suite des recherches menées pour mieux comprendre les variables physico-chimiques.

Avant de sélectionner ces paramètres, nous avons effectué une analyse de leur quantité dans le dataset, illustrée dans la figure 9. Nous avons constaté que chaque paramètre est représenté de manière équivalente dans les données.

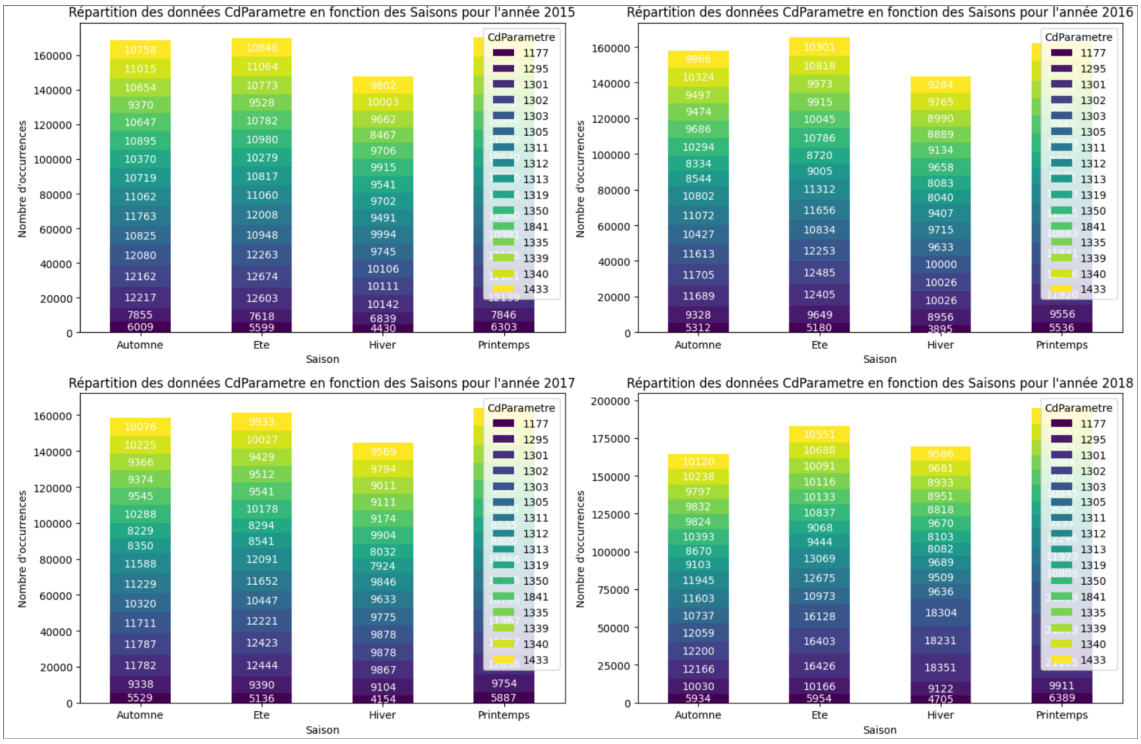


FIGURE 9 – Quantité de valeurs par parmètre physicochimique pour chaque années et chaque saisons.

LbLongParamètre	**CdParametre**
Température de l'Eau	1301
Taux de saturation en oxygène	1312
Turbidité Formazine Néphélométrique	1295
Matières en suspension	1305
Azote Kjeldahl	1319
Phosphore total	1350
Potentiel en Hydrogène (pH)	1302
Demande Biochimique en oxygène en 5 jours (D.B.O.5)	1313
Diuron	1177
Orthophosphates (PO4)	1433
Ammonium	1335
Nitrites	1339
Nitrates	1340
Carbone Organique	1841
Conductivité à 25°C	1303
Oxygène dissous	1311

TABLE 7 – Nom des paramètres et le code associé

Une fois cette sélection effectuée, l'objectif était de récupérer les relevés physico-chimiques pour chaque station. Nous avons ensuite calculé la médiane de ces relevés pour deux raisons : réduire l'impact des valeurs extrêmes et uniformiser la quantité de relevés disponibles pour chaque station. De plus, nous avons choisi de conserver les informations saisonnières dans les données, afin de déterminer si elles pourraient influencer les résultats.

Après cela, nous avons intégré au dataset la qualité de la station, déterminée à partir des données I2M2. Le jeu de données final que nous utiliserons pour entraîner et tester le modèle est présenté dans le tableau. 8.

CdStationMesureEauxSurface	medianNitrate	medianOxygeneDissous	medianDBO5	medianPhosphore	Saison_Ete	Saison_Hiver	Saison_Printemps	measureFromHB
02049500	2.30	10.200	0.75	0.0285	False	False	False	1
02049500	2.35	9.500	0.90	0.0395	True	False	False	1
02049500	2.30	12.400	0.60	0.0170	False	True	False	1
02049500	1.70	11.850	1.15	0.0230	False	False	True	1
02050000	2.75	9.200	0.75	0.0700	False	False	False	1
...	...	...	...	...	...	...	...	...
05068995	1.85	11.700	1.10	0.0300	False	False	True	0
06052650	20.70	9.430	1.00	0.1800	False	False	False	0
06052650	14.45	8.695	0.70	0.1400	True	False	False	0
06052650	17.60	10.660	1.60	0.0670	False	True	False	0
06052650	13.70	10.240	1.20	0.0960	False	False	True	0

TABLE 8 – Dataset des stations de mesure et leurs paramètres associés

5 Modèle et Analyse

Nous avons choisi d'utiliser XGBoost pour réaliser les prédictions. Ce choix est fait car c'est un algorithme qui peut apprendre des relations complexes et non linéaires comme c'est le cas dans nos données. De plus, c'est un modèles qui permet de donner l'importance des features dans chaque prédiction. En l'occurence, nous cherchons à déterminer quel est le paramètre qui a le plus d'importance dans la prédiction des données. Sur le jeu de donnée, l'algorithme XGBoost fourni les résultats suivant :

5.1 Résultats pour le HER 2

TABLE 9 – Résultats de l'algorithme XGBoost pour le HER 2

Classe	Précision	Rappel	F1-score	Support
0	0.86	0.81	0.83	52
1	0.47	0.56	0.51	16
Exactitude (accuracy)	0.75 (68)			
Moyenne macro	0.67	0.69	0.67	68
Moyenne pondérée	0.77	0.75	0.76	68

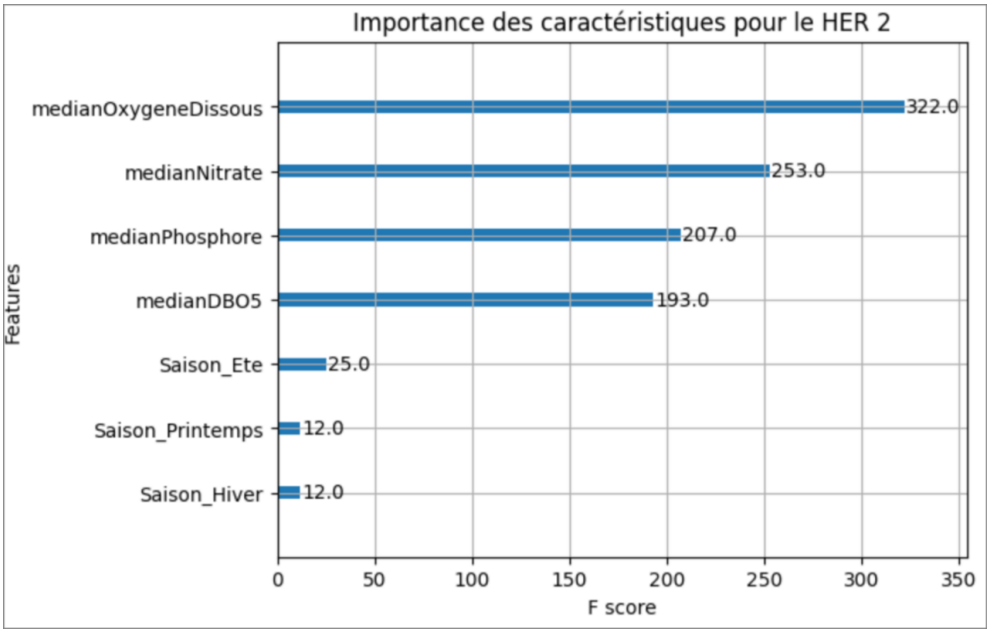


FIGURE 10 – Importance des caractéristiques pour les prédiction des données du HER 2

5.2 Résultats pour le HER 7

TABLE 10 – Résultats de l'algorithme XGBoost pour le HER 7

Classe	Précision	Rappel	F1-score	Support
0	0.50	0.42	0.45	12
1	0.89	0.92	0.90	62
Exactitude (accuracy)	0.84 (74)			
Moyenne macro	0.70	0.67	0.68	74
Moyenne pondérée	0.83	0.84	0.83	74

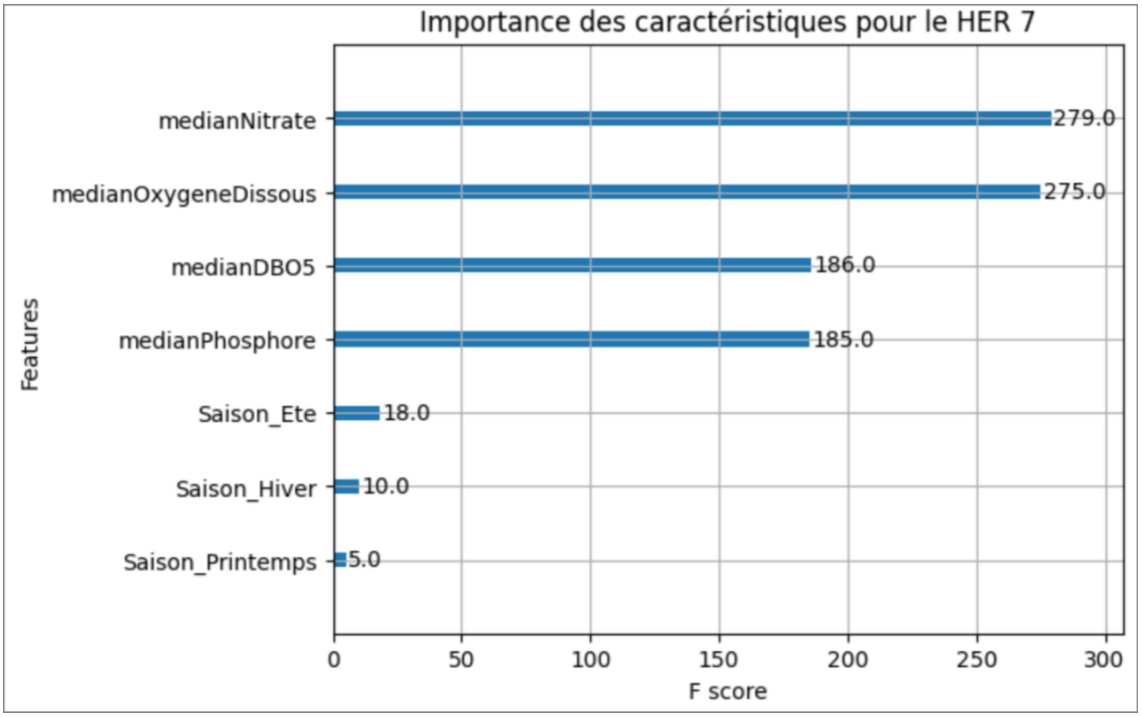


FIGURE 11 – Importance des caractéristiques pour les prédiction des données du HER 7

5.3 Résultats pour le HER 9

TABLE 11 – Résultats de l'algorithme XGBoost pour le HER 9

Classe	Précision	Rappel	F1-score	Support
0	0.48	0.48	0.48	199
1	0.62	0.62	0.62	267
Exactitude (accuracy)	0.56 (466)			
Moyenne macro	0.55	0.55	0.55	466
Moyenne pondérée	0.56	0.56	0.56	466

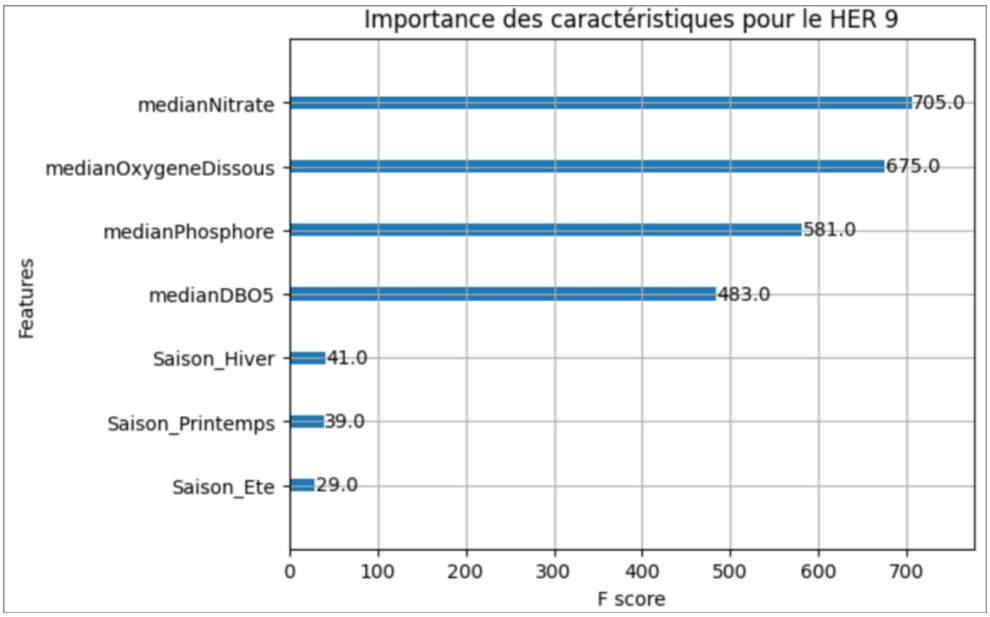


FIGURE 12 – Importance des caractéristiques pour les prédiction des données du HER 9

5.4 Résultats pour le HER 11

TABLE 12 – Résultats de l'algorithme XGBoost pour le HER 11

Classe	Précision	Rappel	F1-score	Support
0	0.59	0.67	0.63	24
1	0.58	0.50	0.54	22
Exactitude (accuracy)	0.59 (46)			
Moyenne macro	0.59	0.58	0.58	46
Moyenne pondérée	0.59	0.59	0.58	46

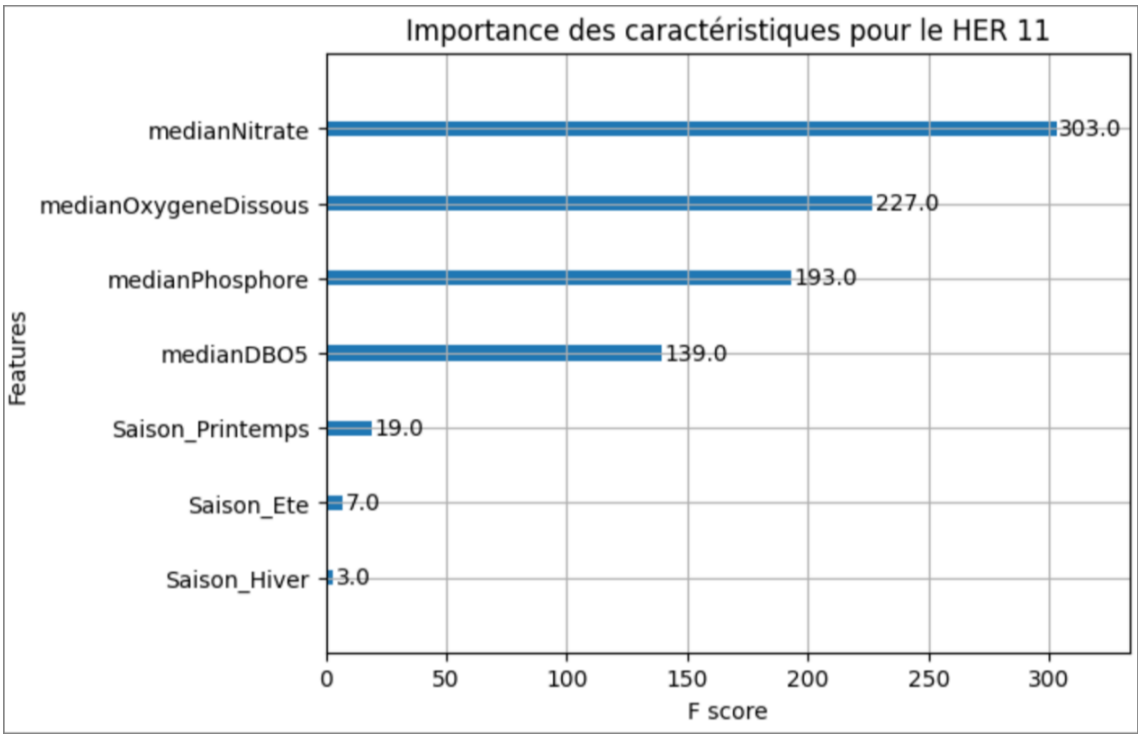


FIGURE 13 – Importance des caractéristiques pour les prédiction des données du HER 11

5.5 Résultats pour le HER 13

TABLE 13 – Résultats de l'algorithme XGBoost pour le HER 13

Classe	Précision	Rappel	F1-score	Support
0	0.67	0.82	0.74	17
1	0.70	0.50	0.58	14
Exactitude (accuracy)	0.68 (31)			
Moyenne macro	0.68	0.66	0.66	31
Moyenne pondérée	0.68	0.68	0.67	31

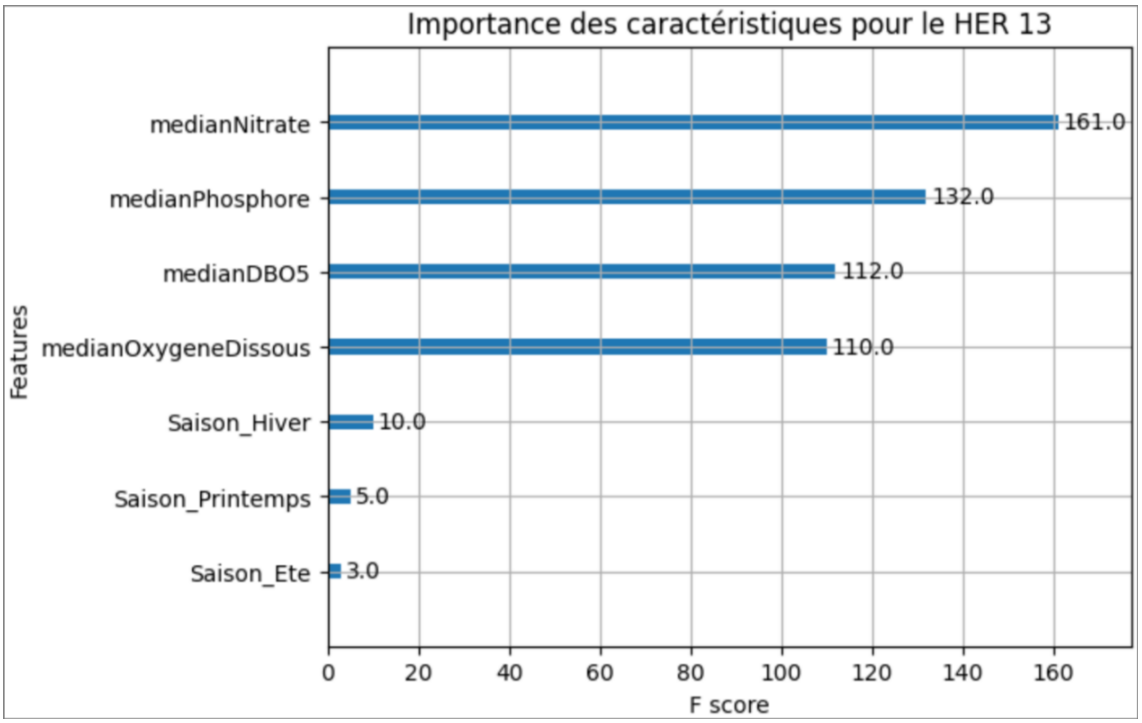


FIGURE 14 – Importance des caractéristiques pour les prédiction des données du HER 13

5.6 Résultats pour le HER 15

TABLE 14 – Résultats de l'algorithme XGBoost pour le HER 15

Classe	Précision	Rappel	F1-score	Support
0	0.67	0.84	0.74	95
1	0.70	0.47	0.56	75
Exactitude (accuracy)	0.68 (170)			
Moyenne macro	0.68	0.65	0.65	170
Moyenne pondérée	0.68	0.68	0.66	170

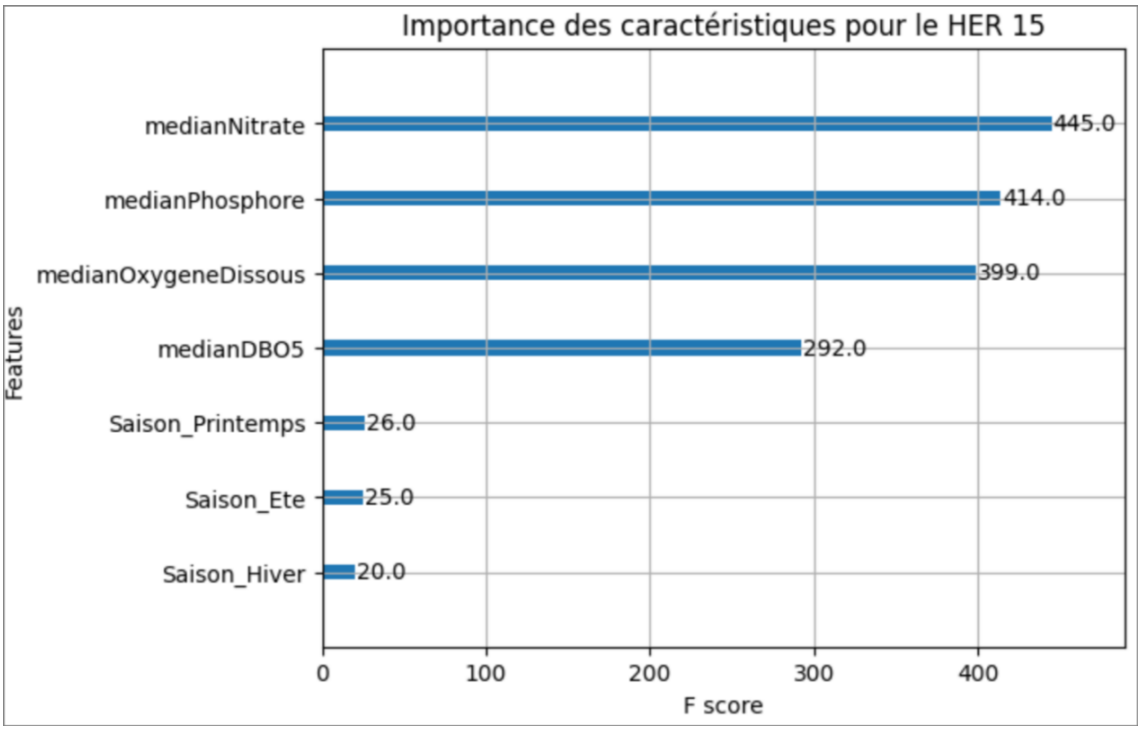


FIGURE 15 – Importance des caractéristiques pour les prédiction des données du HER 15

5.7 Résultats pour le HER 19

TABLE 15 – Résultats de l'algorithme XGBoost pour le HER 19

Classe	Précision	Rappel	F1-score	Support
0	0.78	0.88	0.82	8
1	0.90	0.82	0.86	11
Exactitude (accuracy)	0.84 (19)			
Moyenne macro	0.84	0.85	0.84	19
Moyenne pondérée	0.85	0.84	0.84	19

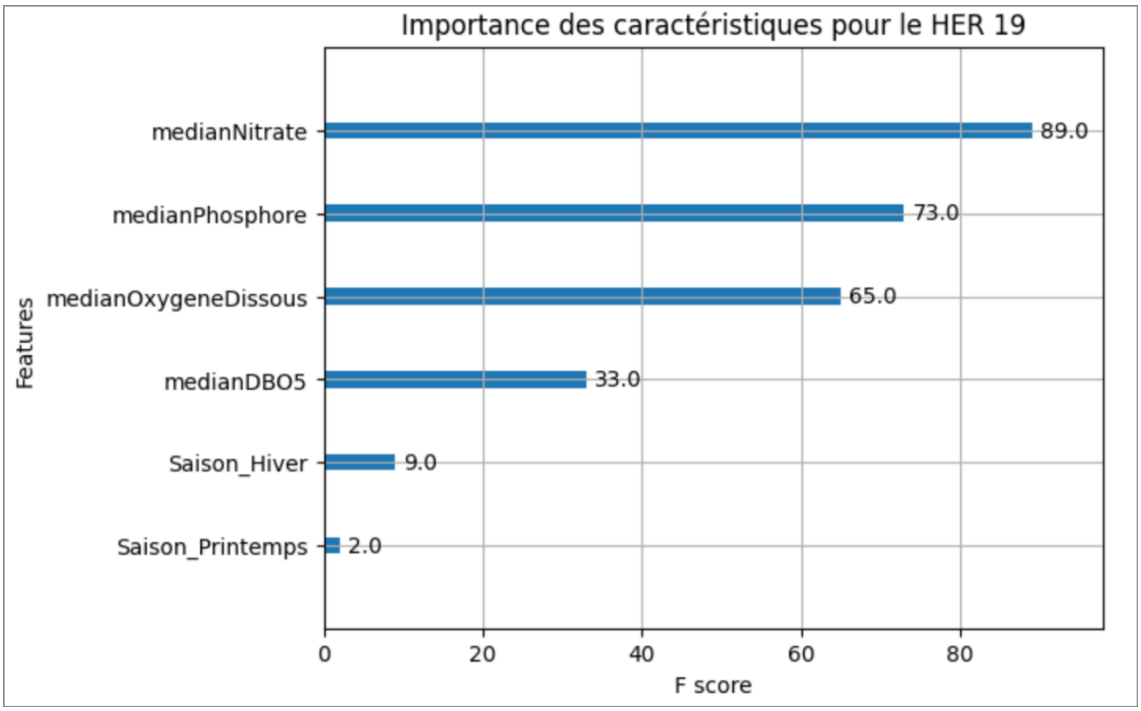


FIGURE 16 – Importance des caractéristiques pour les prédiction des données du HER 19

5.8 Résultats pour le HER 20

TABLE 16 – Résultats de l'algorithme XGBoost pour le HER 20

Classe	Précision	Rappel	F1-score	Support
0	0.73	0.83	0.78	23
1	0.43	0.30	0.35	10
Exactitude (accuracy)	0.67 (33)			
Moyenne macro	0.58	0.56	0.56	33
Moyenne pondérée	0.64	0.67	0.65	33

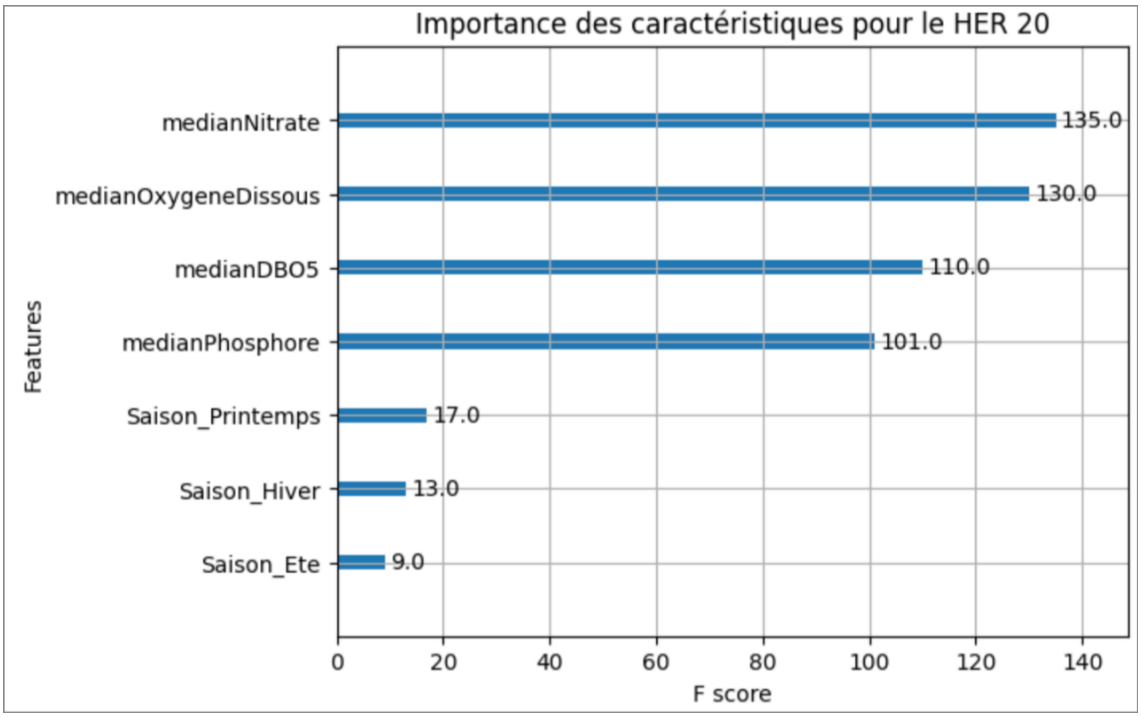


FIGURE 17 – Importance des caractéristiques pour les prédiction des données du HER 20

5.9 Explication des résultats

Pour chacun des HER, les résultats sont plus ou moins acceptables en fonction du HER. De manière générale, les résultats sont tout de même acceptables avec une grande majorité

6 Limites des simplifications effectuées

Les simplifications apportées à l'analyse ont permis de réduire la complexité du problème, mais elles comportent également plusieurs limites qu'il convient de souligner.

1. **Limites du choix de la médiane :** Le choix de la médiane pour représenter les valeurs physico-chimiques est une méthode robuste qui permet de réduire l'impact des valeurs extrêmes, mais elle présente certaines limites. En effet, la médiane ne tient pas compte de la distribution complète des données et peut masquer certaines variations importantes.
2. **Limites du choix des paramètres physico-chimiques :** Le choix des paramètres physico-chimiques (nitrates, DBO5, phosphore, oxygène dissous) a été fait en fonction de leur impact reconnu sur la qualité de l'eau, basé sur des recherches précédentes. Cependant, ce choix peut être limité, car d'autres paramètres importants comme les métaux lourds, le pH, ou la conductivité, qui influencent également la qualité de l'eau, ont été exclus. Bien que ces paramètres choisis soient représentatifs des facteurs principaux affectant la qualité de l'eau, un élargissement de cette sélection pourrait permettre d'obtenir une analyse plus complète. L'inclusion de davantage de paramètres pourrait potentiellement améliorer la précision du modèle, mais cela introduirait aussi davantage de complexité.
3. **Limites du choix des Hydroécorégions (HER) :** Le choix de ne considérer que certaines Hydroécorégions (HER) pour cette analyse est une simplification nécessaire pour limiter la portée du modèle, mais cela comporte des risques. En ne prenant pas toutes les HER, nous excluons potentiellement des régions ayant des caractéristiques distinctes qui pourraient influencer les résultats de manière significative. De plus, certaines HER non prises en compte pourraient avoir des comportements différents, et les résultats du modèle pourraient être biaisés si ces différences sont importantes. Une meilleure approche consisterait à inclure un échantillon plus large d'HER pour que le modèle puisse apprendre à partir d'une plus grande diversité de données géographiques et environnementales.

4. **Limites de l'approche par clustering :** L'utilisation du clustering pour classifier les stations en deux groupes ('plutôt bonnes' et 'plutôt mauvaises') repose sur l'idée que cette classification permet de simplifier l'analyse. Cependant, cette approche peut introduire des erreurs, notamment si certaines stations ayant des valeurs exceptionnelles sont mal classées. Par exemple, des stations avec des relevés atypiques mais représentatifs d'un état écologique spécifique pourraient être classées dans le mauvais groupe, ce qui fausserait les résultats. Ce phénomène pourrait également survenir en raison de l'application d'un algorithme de clustering comme K-means, qui suppose des clusters bien séparés et peut échouer à capturer des structures plus complexes dans les données. En conséquence, des erreurs de classification pourraient induire des résultats inexacts, avec des stations affichant des états écologiques incorrects.

En résumé, bien que ces simplifications aient été nécessaires pour rendre l'analyse plus gérable, elles comportent des limites qu'il est important de considérer. L'impact des choix effectués sur les résultats finaux doit être évalué et, si possible, des améliorations pourraient être envisagées pour rendre les conclusions plus représentatives de la complexité des données.

7 Bilan et Perspectives

Ce projet a entraîné l'exploitation d'une grande quantité de données provenant de différentes stations de mesure environnementales, ce qui a nécessité un travail approfondi de préparation et de sélection des données. L'une des étapes cruciales de cette analyse a été la définition des paramètres physico-chimiques pertinents à étudier. Après un examen minutieux, nous avons choisi de concentrer notre attention sur les nitrates, la DBO5, le phosphore et l'oxygène dissous, des variables largement reconnues pour leur impact sur la qualité de l'eau. Ce processus de sélection a permis de simplifier l'approche tout en conservant la pertinence des données analysées.

L'une des principales difficultés du projet a résidé dans la préparation des données, notamment la gestion des valeurs manquantes, l'élimination des doublons et l'harmonisation des relevés. Ce travail de prétraitement a permis de rendre les données exploitables, mais il a aussi introduit des simplifications qui limitent la précision de nos conclusions. En effet, les choix méthodologiques, comme l'utilisation de la médiane pour traiter les valeurs extrêmes ou la sélection restreinte des Hydroécorégions (HER) à étudier, ont permis de converger vers une solution, mais sans pouvoir garantir que celle-ci soit totalement représentative de la complexité du terrain.

La méthode suivie, bien qu'efficace pour réduire la portée de l'analyse et rendre le problème plus accessible, comporte des simplifications qui doivent être prises en compte dans l'interprétation des résultats. En conséquence, bien que nous ayons pu obtenir des pistes intéressantes sur l'impact des paramètres physico-chimiques sur la qualité de l'eau, nous ne pouvons pas affirmer avec certitude que la solution finale soit entièrement fiable. Toutefois, au regard des objectifs définis en début de projet, nous pouvons formuler une hypothèse solide quant à la relation entre les variables étudiées et la qualité de l'eau, ce qui représente une avancée importante dans la compréhension de ces phénomènes.

Ainsi, bien que les résultats du projet soient prometteurs, il est important de reconnaître les limites inhérentes aux simplifications apportées. Cela ouvre la voie à des recherches futures où une approche plus fine et plus complète pourrait permettre de confirmer ou d'affiner les conclusions émises ici.