République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master en Informatique

Spécialité : *Intelligence Artificielle (IA)*

Thème

Système intelligent de classification multi-niveaux de la qualité de l'eau

Encadré par

— Dr. Akli ABBAS

Réalisé par

— MECHEKAK Zahra

— MERZOUK Dyhia

2025/2026

Remerciements

Avant tout, nous rendons grâce à Allah le Tout-Puissant pour nous avoir accordé la santé, la volonté et la persévérance nécessaires à l'aboutissement de ce travail.

Nous adressons nos remerciements à notre encadreur Dr Akli Abbas, pour ses conseils et son accompagnement dans ce projet.

Nous exprimons également nos sincères remerciements aux membres du jury , ainsi que nos enseignants pour la qualité de leur formation.

Enfin, nous remercions nos familles pour leur soutien et leurs encouragements tout au long de la réalisation de ce travail.

Dédicaces

Je dédie ce travail à ma mère et à mon père, pour leurs sacrifices, leur soutien et leurs encouragements tout au long de mon parcours.

À mes deux sœurs et à mon petit frère.

À mon proche Ayoub.

À tous ceux qui m'ont soutenue et encouragée tout au long de mes études.

MECHEKAK zahra.

Dédicaces

Je dédie ce travail à Dieu, pour m'avoir donné la force et la patience nécessaires pour arriver jusqu'ici.

À moi-même, pour la persévérance et les efforts fournis tout au long de ce parcours.

À mes parents, pour leur amour inconditionnel et leurs encouragements constants.

À mes sœurs Djouher et Maria, à mon beau-frère Saïd.

À ma cousine Sabrina, et à tous ceux qui ont cru en moi dans les moments de doute.

MERZOUK dyhia.

Abstract

Water quality monitoring is essential for protecting public health and the environment. In this work, a machine learning-based approach is proposed to classify water quality using physicochemical parameters. The Water Quality Index (WQI) is used to define five quality classes : Excellent, Good, Moderate, Poor, and Dangerous. Several classification algorithms were tested after data preprocessing. To reduce dimensionality and improve performance, Principal Component Analysis (PCA) was applied. The experimental results show that ANN achieves an accuracy of 95.45%. This approach enables automatic, fast, and reliable water quality assessment and highlights the potential of machine learning for sustainable and efficient environmental monitoring.

Key words : Water quality, Machine Learning, Classification, Stacking Ensemble, Random Forest, Decision Tree, XGBoost, Intelligent System, Water Quality Index (WQI).

Résumé

La surveillance de la qualité de l'eau est essentielle pour la protection de la santé publique et de l'environnement. Dans ce travail, une approche basée sur l'apprentissage automatique est proposée pour classifier la qualité de l'eau à partir de paramètres physico-chimiques. Le Water Quality Index (WQI) est utilisé pour définir cinq classes de qualité : Excellente, Bonne, Moyenne, Mauvaise et Dangereuse. Plusieurs algorithmes de classification ont été testés après un prétraitement des données. Afin de réduire la dimensionnalité et améliorer les performances, une Analyse en Composantes Principales (ACP/PCA) a été appliquée. Les résultats expérimentaux montrent que ANN atteint une précision de 95.45% . Cette approche permet une évaluation automatique, rapide et fiable de la qualité de l'eau et met en évidence le potentiel de l'apprentissage automatique pour la surveillance environnementale durable et efficace globale.

Mots-clés : Qualité de l'eau, Apprentissage automatique, Classification, Stacking, Random Forest, Decision Tree, XGBoost, Système intelligent, Indice de qualité de l'eau (WQI).

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

AD	Arbre de Décision (*Decision Tree*)
FA	Forêt Aléatoire (*Random Forest*)
IA	Intelligence Artificielle (*AI : Artificial Intelligence*)
IoT	Internet des Objets (*Internet of Things*)
IQA	Indice de Qualité de l'Eau
KNN	K plus proches voisins (*K-Nearest Neighbors*)
ML	Apprentissage Automatique (*Machine Learning*)
NB	Naïve Bayes
pH	Potentiel Hydrogène
RL	Régression Logistique (*Logistic Regression*)
SVM	Machine à Vecteurs de Support (*Support Vector Machine*)
WQI	*Water Quality Index* (Indice de Qualité de l'Eau)
OMS	*O*rganisation Mondiale de la Santé (World Health Organization)

Introduction générale

L'eau constitue une ressource essentielle à la vie humaine, au développement économique et à l'équilibre des écosystèmes. elle est utilisée pour la consommation domestique, agriculture, l'industrie et de nombreuse autres activités essentielles. Toutefois, la croissance démographique, l'urbanisation rapide ainsi que les activités industrielles et agricoles ont entraîné une dégradation progressive de la qualité des ressources en eau dans plusieurs régions du monde [1]. Cette situation souligne l'importance de mettre en place des systèmes efficaces permettant de surveiller et d'évaluer la qualité de l'eau de maniéré continue afin de garantir la santé publique et la protection de l'environnement.

Traditionnellement,l'évaluation de la qualité de l'eau repose sur des analyses physico-chemiques et microbiologiques effectuées dans des laboratoires spécialisés [2]. Bien que ces méthodes offrent une grande précision, elles nécessitent de équipements coûteuse, des ressources humaines qualifiées et un temps d'analyse relativement important.De plus, elles ne permettent pas d'obtenir des informations en temp réel, ce qui peux retarder la détection des problèmes liés à la contamination de l'eau.

Avec l'évolution des technologie numérique, l'internet des objets (IOT) est devenu une solution prometteuse pour la surveillance intelligente des ressources hydriques. Grace a des captures connectés, il est désormais possible de mesurer en temps réel plusieurs paramètres de qualité de l'eau tels que le pH, la turbidité, la conductivité, les solides dissous totaux et la température [3].Les données collectés peuvent être transmises automatiquement vers des plateformes de traitement et d'analyse, facilitant ainsi la prise de décision et la gestion des ressources en eau.

Parallèlement, les techniques d'apprentissage automatique (Machine Learning) ont connu un développement considérable au cours des derrières années.Ces techniques permettent

d'extraire des connaissance à partir de grandes quantités de données et de construire des modèles capable de réaliser des tache de classification, de prédiction et d'aide à la décision avec une grande efficacité [4].Dans le domaine de la qualité de l'eau, plusieurs travaux ont démontré l'intérêt de l'utilisation des algorithmes de Machine Learning pour prédire l'état et détecter d'éventuelles anomalies à partir des données collectées par les capteurs[5].

Cependant, une question demeure : jusqu'à quel point est-il envisageable d'associer les technologies IoT aux algorithmes de Machine Learning afin de classifier de manière automatique et performante la qualité de l'eau en temps réel ?

Dans ce contexte, notre objectif principale est de développer un système capable de classifier l'eau en plusieurs niveaux de qualité (Excellente, Bonne, Moyenne, Mauvaise et Dangereuse) à partir de différents paramétré mesurés utilisant Machine Learning.quelques algorithmes d'apprentissage supervisé seront étudiés, comparer et optimisés afin d'identifier le modèle offrant les meilleures performances de classification.

Ce travail est structuré en trois chapitres :

• Chapitre 1 : État de l'art
Dans ce chapitre, nous présentons les concepts fondamentaux liés à la qualité de l'eau, à l'Internet des Objets et aux approches principales de Machine Learning, Deep Learning et quelque travaux Antérieurs .
• Chapitre 2 :Données, Analyse Exploratoire et Méthodologi
Dans ce chapitre, nous décrirons dataset, les étapes de prétraitement et la conception des modèles proposés pour classification multi-classe de la qualité de l'eau.
• Chapitre 3 : Implémentation des modèles et évaluation des performance
Dans ce chapitre, nous étudierons les résultats obtenus.

Enfin, une conclusion générale résume ce travail.

Chapitre 1

État de l'art

1.1 Introduction

La qualité de l'eau constitue un enjeu majeur pour la santé publique, la préservation de l'environnement et la gestion durable des ressources en eau [6]. Avec l'augmentation des activités humaines, la pollution des eaux s'intensifie, rendant les méthodes traditionnelles de surveillance souvent insuffisantes face à la grande quantité et à la complexité des données à analyser.

Dans ce contexte, les systèmes intelligents, combinant l'internet des Objets (IOT), le Machine Learning et le Deep Learning, représentent des solutions prometteuses. Ils permettent d'analyser efficacement les données, de détecter des motifs complexes et de classer automatiquement la qualité de l'eau selon plusieurs niveaux.

Ce chapitre expose les notions essentielles relatives à la qualité de l'eau, la définition de l'Internet des objets, son importance aussi son rôle dans la collecte des donnée pour l'IA et leur architecture. Il introduit ensuite les approche de l'intelligence artificielle dédiées à la classification multi-classe, le principe de fonctionnement, l'architecture générale des système intelligent, puis une revue des travaux antérieurs.

1.2 Qualité de l'eau

1.2.1 Définition de la qualité de l'eau

La qualité de l'eau se définit par les attributs chimiques, physiques, biologiques et radiologiques de l'eau, qu'il s'agisse des eaux superficielles, des eaux profondes ou des eaux souterraines. C'est une évaluation de la qualité de l'eau en fonction des nécessités d'une ou plusieurs espèces biologiques, ou de tout besoin ou objectif humain. On l'utilise souvent comme référence à un ensemble de normes qui servent à juger la conformité. Les critères les plus fréquemment employés pour juger de la qualité de l'eau se rapportent à la santé des écosystèmes, à la sûreté des interactions humaines et à l'eau destinée à la consommation [7].

1.2.2 La surveillance de la qualité de l'eau

La surveillance de la qualité de l'eau constitue un moyen essentiel pour préserver les ressources hydriques et assurer l'équilibre des écosystèmes. En s'appuyant sur la collecte et l'analysant des données, elle apporte une base scientifique solide pour la protection de l'environnement, à la maîtrise de la pollution et aux travaux de recherche. Avec l'essor continu de l'industrialisation, de l'urbanisation et des activités agricoles, la pollution de l'eau s'aggrave, ce qui renforce l'importance de cette surveillance dans la gestion durable des ressources en eau, la protection des écosystèmes et la préservation de la santé publique[8].

1.2.3 Les paramètres physico-chimiques

Ils se rapportent à des caractéristiques quantifiables telles que le pH, la température, la conductivité et la dureté et quelle composants chimiques comme les chlorures, le potassium, les ions et les sulfates.

Les paramètres physiques

La température(unité : °C)

L'intensité de la chaleur, constitue un paramètre essentiel car elle agit sur plusieurs propriétés de l'eau (densité, viscosité, teneur en oxygène dissous), avec des réflexes sur la vie aquatique. Elle peut fluctuer sous l'effet de sources naturelles (telles que l'énergie solaire) ainsi que de sources d'origine humaine (rejets industriels et eaux de refroidissement des machines) [7].

La saveur et l'odeur(unité : Taux dilution)

Résultent de causes naturelles (algues, végétation en décomposition, bactéries, champignons, composés organiques, tels que le gaz sulfurique, les sulfates et les chlorures) et artificielles (eaux usées domestiques et industrielles) [7].

La couleur (unité : mg/l platine)

Résulte de substances dissoutes, peut être causée par le fer ou le manganèse, les végétaux, les algues ou encore par l'apport d'eaux usées industrielles et domestiques [7].

La turbidité(unité : NTU)

La turbidité peut varier considérablement selon les rivières, les lacs et les périodes de l'année. C'est pourquoi la plupart des réglementations gouvernementales définissent une augmentation maximale admissible par rapport au niveau naturel propre à chaque masse d'eau. En général, une turbidité inférieure à 10 UTN est considérée comme faible, une valeur autour de 50 UTN comme modérée, tandis que des niveaux supérieurs à 100 UTN indiquent une turbidité très élevée. L'augmentation de la turbidité dans les milieux aquatiques peut résulter de phénomènes naturels, tels que de fortes pluies entraînant des particules de sol, l'érosion, la fonte des neiges, les tempêtes ou les incendies, mais aussi d'activités humaines comme l'exploitation forestière, minière, agricole ou encore le dragage[9].

Les paramètres chimiques

Potentiel hydrogène (unité :pH)

Le pH permet de mesurer les ions hydrogène présents dans l'eau. Il indique l'équilibre entre l'acidité et la basicité entre 0 à 14, avec une valeur de 7 correspondant à la neutralité.Ce paramètre reflète de divers équilibres physico-chimiques est influencé par plusieurs éléments, notamment l'origine de l'eau. On doit mesurer le pH directement sur place en utilisant un pH-mètre ou via une technique colorimétrique [9].

ph < 5	Acidité forte
pH=7	Neutre
7< pH< 8	Eaux de surface
5,5< pH< 8	Eaux de souterraines
pH=8	Alcalinité forte, évaporation intense

TABLE 1.1 – Classification des eaux d'après leur pH.

[9]

Conductivité (unité : µS/cm à 20°C)

La conductivité offre un moyen rapide de déterminer la minéralisation générale de l'eau. Elle est mesurée en déterminant la capacité de l'eau à conduire un courant électrique entre deux électrodes métalliques. Cette caractéristique est en relation inverse avec la résistivité électrique [10].

Alcalinité (unité : mg/l CaCo3)

L'alcalinité désigne la capacité de l'eau à résister aux variations de pH, c'est-à-dire son pouvoir tampon. Une eau présentant une forte alcalinité est davantage protégée contre les variations soudaines de pH lorsqu'on y ajoute des substances acides ou basiques, comme des rejets chimiques ou des eaux de ruissellement polluées [11].

Dureté TH (unité :mg/L de CaCO)

C'est la quantité de minéraux dissous qu'elle contient, principalement le calcium et le magnésium. Elle est étroitement liée à la nature des roches d'une région et aux minéraux qui les composent. Ainsi, les roches calcaires, riches en calcium, se dissolvent plus facilement au contact de l'eau.
À l'inverse, certaines roches, plus pauvres en calcium et en magnésium, sont moins sensibles à l'érosion. Dans ces régions, l'eau des rivières et des lacs possède généralement une faible dureté et est qualifiée de douce [12].

Ions majeurs

Les ions majeurs on les classe en deux groupes : les cations, incluant le calcium, le magnésium, le sodium et le potassium, et les anions, qui englobent les chlorures, les sulfates, les nitrates et les bicarbonates [9].

Nutriments

Les nutriments représentent une catégorie de paramètres chimiques que l'on peut mesurer dans l'eau. Les principaux nutriments des milieux aquatiques sont le phosphore , l'azote et le carbone. Ils sont indispensables au développement des organismes vivants, en particulier des végétaux et des micro-organismes. Néanmoins, une teneur trop élevée en nutriments peut altérer la qualité de l'eau et provoquer des déséquilibres au sein des écosystèmes [13].

métaux

Les métaux sont des éléments métalliques trouvés naturellement qui peuvent s'accumuler dans l'eau, les boues (accumulation durant les étapes de purification), les végétaux et les animaux. Les polluants majeurs de l'écosystème sont les métaux rejetés par l'homme, et leur ingestion peut entraîner des effets négatives sur la santé humaine [14].

1.3 Internet des objets

1.3.1 Définition de l'Internet des Objets

L'Internet des Objets (IoT) englobe l'ensemble des équipements physiques connectés au réseau mondial, qui s'appuient sur des capteurs et des programmes informatiques pour communiquer des données.Grâce à cela, ces appareils intelligentes peuvent collecter et échanger des données, ce qui facilite la communication et l'automatisation. L'IoT va au-delà des ordinateurs et des téléphones intelligents pour inclure des objets du quotidien comme les appareils électroménagers, les véhicules ou encore les vêtements. Cette interconnexion rend possible la surveillance, le pilotage et l'analyse à distance, ce qui améliore l'efficacité, le confort et stimule l'innovation dans de nombreux domaines, allant des maisons connectées aux services de santé, en passant par l'industrie, les transports et l'agriculture [15].

1.3.2 Architecture générale d'un système IoT

L'architecture IoT regroupe plusieurs composantes de systèmes connectés afin de garantir que les données issues des capteurs des objets soient correctement collectées, stockées et exploitées dans des entrepôts de big data, et que les actionneurs des objets exécutent les commandes transmises via une application utilisateur.
Il s'agit d'un cadre qui précise les composants physiques, l'organisation fonctionnelle et la configuration du réseau, ainsi que les procédures opérationnelles et les formats de données à utiliser. Même s'il n'existe pas de modèle d'architecture IoT unique et universellement accepté, la structure la plus simple et la plus courante s'appuie sur une architecture IoT en quatre couches [16].

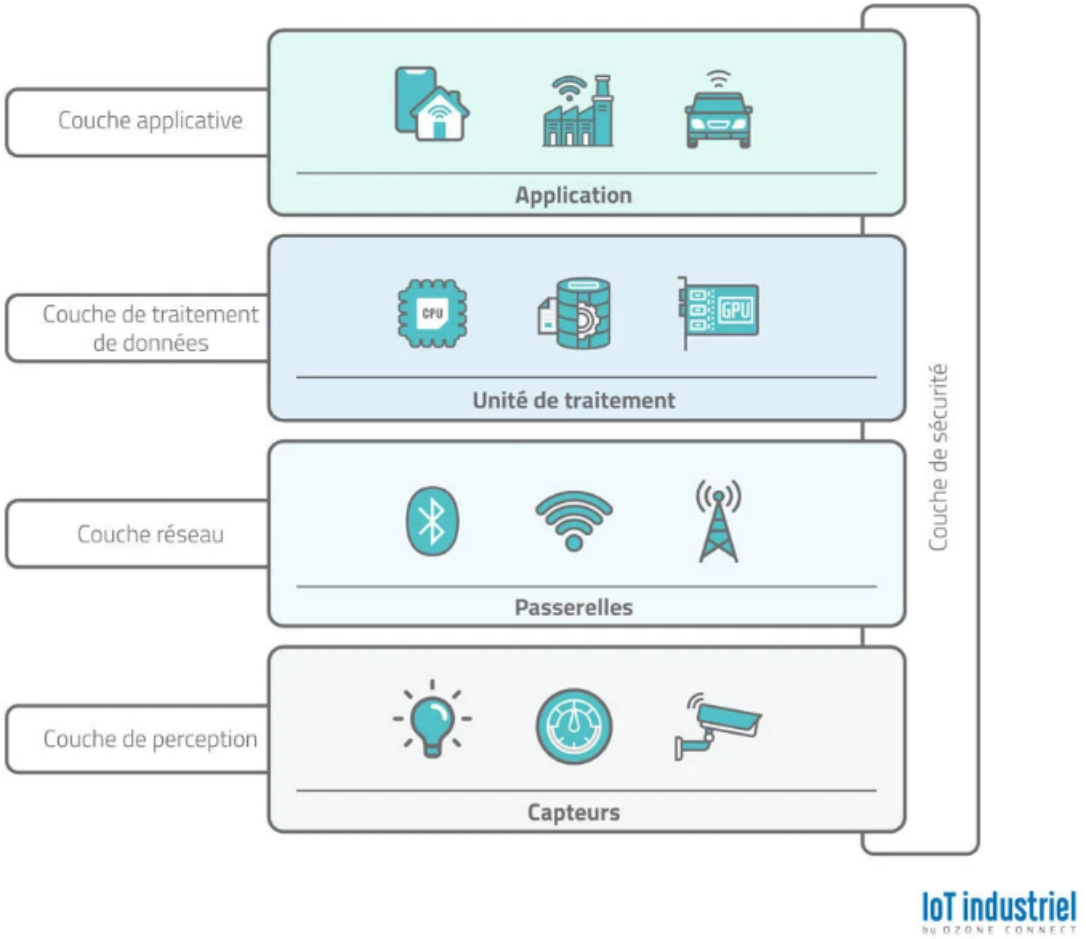

FIGURE 1.1 – Architecture IoT

[16]

1.3.3 Rôle de l'IoT dans la collecte des données pour l'IA

L'Internet des Objets constitue une source importante de données pour les applications d'intelligence artificielle. Grâce aux capteurs et équipements connectés, il devient possible d'acquérir en permanence des informations provenant de l'environnement réel. Ces informations, qu'il s'agisse de mesures physiques, environnementales ou industrielles, sont exploitées par les modèles d'apprentissage automatique afin d'identifier des tendances, détecter des comportements anormaux et produire des prédictions. L'association de l'IoT et de l'IA favorise ainsi la mise en place de systèmes autonomes capables d'améliorer la prise de décision dans de nombreux secteurs [17].

1.3.4 Capteurs IoT pour la qualité de l'eau

L'Internet des Objets (IoT) constitue une solution efficace pour la surveillance de la qualité de l'eau grâce à l'utilisation de capteurs connectés permettant la collecte et la transmission de données en temps réel. Ces capteurs, déployés sur le réseau de distribution d'eau, mesurent plusieurs paramètres physico-chimiques tels que le pH, la température, la turbidité ainsi que la présence de contaminants. Les données collectées sont transmises via des technologies de communication sans fil (telles que LoRaWAN) vers des plateformes de traitement, offrant ainsi une vision globale, continue et actualisée de l'état de l'eau. Cette transmission en temps réel permet une analyse rapide des informations et facilite la détection des anomalies. En cas de dépassement de seuils prédéfinis, des alertes automatiques sont générées afin de signaler toute dégradation de la qualité de l'eau. Cela permet aux gestionnaires d'intervenir rapidement pour limiter les risques sanitaires et environnementaux. L'utilisation des capteurs IoT présente plusieurs avantages, notamment une surveillance continue 24h/24 et 7j/7, une détection précoce des anomalies, une optimisation des traitements de l'eau, ainsi qu'une meilleure gestion des ressources et une réduction des coûts d'exploitation [18].

1.4 Intelligence artificielle

1.4.1 Définition

L'intelligence artificielle désigne un ensemble de méthodes qui imitent certains mécanismes de l'intelligence humaine, en s'appuyant sur la création et l'utilisation d'algorithmes exécutés dans un environnement informatique adaptable et extensible. Elle vise à permettre aux ordinateurs de raisonner et d'agir d'une façon proche de celle des êtres humains [19].

1.4.2 Branche de l'intelligence artificielle

— Apprentissage automatique (ML).
— apprentissage profond(DL).
— Traitement automatique du langage naturel (NLP).
— Vision par ordinateur.

— logique floue.

— Robotique intelligente [20].

Ces technologies sont aujourd'hui utilisées dans de nombreux secteurs tels que la santé, l'industrie, l'agriculture, les transports ainsi la gestion des ressources environnementales.

1.5 Machine learning

1.5.1 Définition

apprentissage automatique, est une branche de l'intelligence artificielle qui cherche à donner aux machines la capacité d'apprendre à partir de données, en utilisant des modèles mathématiques. Il s'agit plus spécifiquement du processus où des informations pertinentes sont extraites d'un ensemble de données d'apprentissage.
L'objectif de cette étape est d'obtenir les paramètres d'un modèle qui garantiront les performances optimales, en particulier lors de l'exécution de la tâche attribuée au modèle. Après avoir effectué l'entraînement, le modèle pourra par la suite être déployé en production [21].

1.5.2 Types d'apprentissage

1.5.2.1 Apprentissage supervisé :

L'apprentissage supervisé est une méthode d'apprentissage automatique qui repose sur l'utilisation de données étiquetées pour entraîner un modèle. Ce dernier apprend les relations entre les données d'entrée et les résultats attendus afin de pouvoir effectuer des prédictions précises sur de nouvelles données jamais observées auparavant [22].

L'apprentissage supervisé résout deux problèmes :

1. **Les problèmes de classification**
   Utilisent des algorithmes pour classer les données dans des catégories spécifiques. Les machines à vecteurs de support ainsi que les arbres décisionnels sont des algorithmes de classification.
   Exemple : La classification d'un courriel comme spam ou non [23].

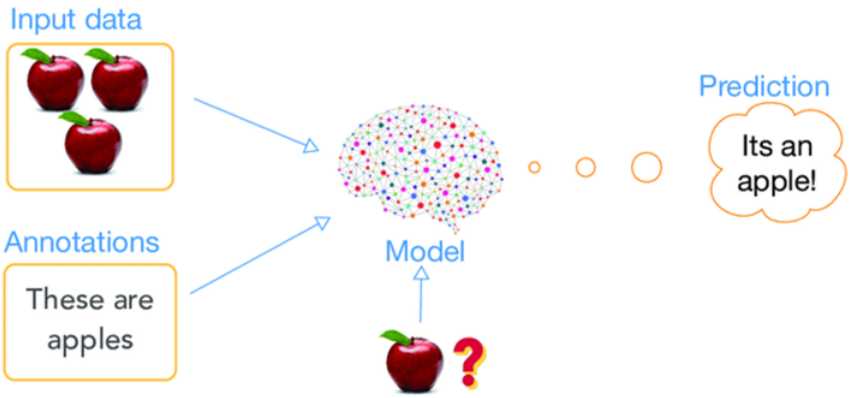

FIGURE 1.2 – Exemple d'apprentissage supervisé.

[23]

2. **Les problèmes de régression**

   Utilisent des algorithmes pour prédire une valeur numérique ou établir une relation entre des variables d'entrée et de sortie. La régression linéaire est un algorithme de régression.

   Exemple : La prévision météorologique par régression [23].

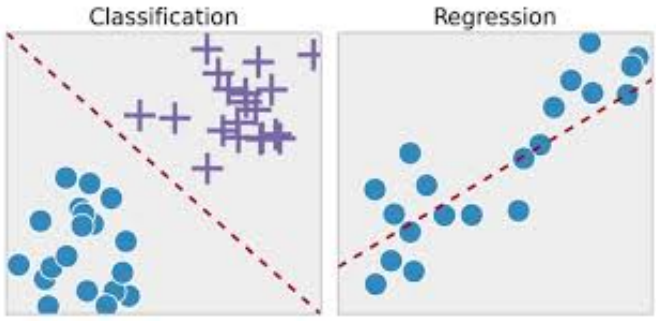

FIGURE 1.3 – Types de techniques d'apprentissage automatique supervisé.

[24]

1.5.2.2 Apprentissage non supervisé :

L'apprentissage non supervisé, repose sur l'utilisation d'algorithmes de machine learning (ML) pour examiner et organiser des ensembles de données dépourvus d'étiquettes. Ces algorithmes mettent en évidence des structures cachées ou des groupes de données, sans nécessiter d'intervention humaine [25].

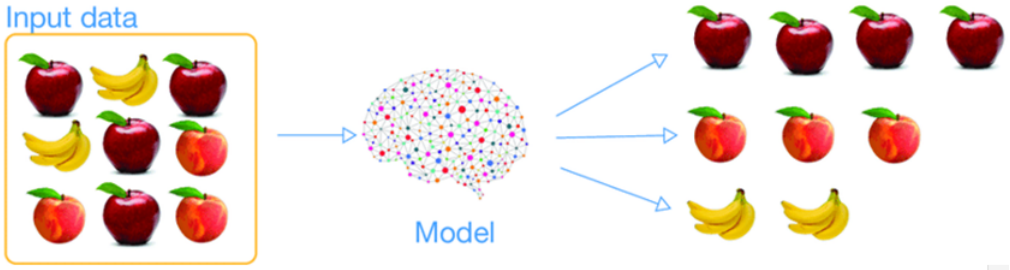

FIGURE 1.4 – Exemple d'apprentissage non supervisé.
[23]

L'apprentissage non supervisé peut être catégorisé en clustering, association et réduction de dimensionnalité.

1. **Le clustering**
   consiste à regrouper les données en clusters en fonction de leurs similitudes ou de leurs différences. Le clustering K-means est un algorithme de clustering.exemple : La segmentation du marché [23].
2. **L'association**
   détermine l'ensemble des éléments qui apparaissent ensemble dans un jeu de données. Apriori, Eclat et FP Growth sont d'algorithmes d'association. exemple : L'analyse du panier d'achat [23].
3. **La réduction de dimensionnalité**
   consiste à réduire la dimension d'un ensemble de données afin de faciliter son traitement par des algorithmes supervisés. Elle comprend deux étapes : la sélection et l'extraction des caractéristiques [23].

1.5.2.3 Apprentissage par renforcement :

Cette approche algorithmique offre aux machines et aux entités robotiques la flexibilité nécessaire pour s'adapter à des scénarios inconnus ou complexes. Sans recourir à une supervision humaine directe, le dispositif apprend de manière itérative en identifiant les décisions qui maximisent les récompenses ou mènent aux résultats les plus efficaces [26].

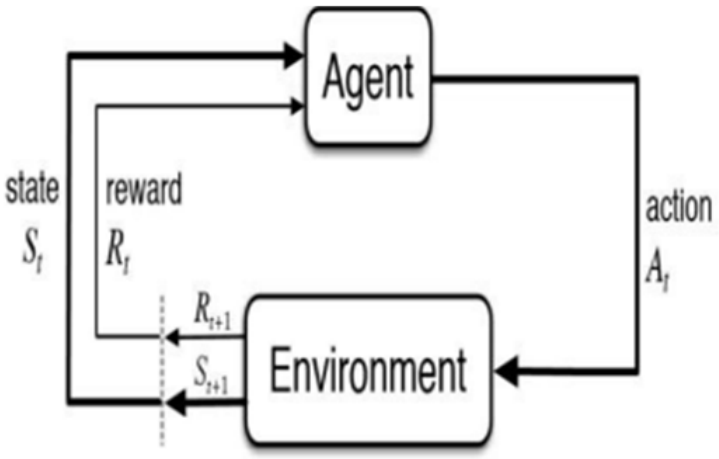

FIGURE 1.5 – Apprentissage par renforcement.

[27]

1.5.3 Algorithmes de Machine Learning

Algorithmes d'apprentissage supervisé

1. **Random Forest**

   Random Forest est une méthode d'apprentissage automatique qui repose sur un grand nombre d'arbres de décision afin d'augmenter la précision des prédictions[28].

   Fonctionnement de l'algorithme Random Forest

   Étape 1 : Nous choisissons un sous-ensemble de points de données ainsi qu'un sous-ensemble de caractéristiques pour construire chaque arbre de décision.

   Étape 2 : Nous entraînons un arbre de décision individuels sur chacun de ces échantillons.

   Étape 3 : Chaque arbre de décision produit alors sa propre prédiction.

   Étape 4 : Le résultat final est déterminé par un vote majoritaire dans le cas de la classification et par la moyenne des prédictions dans le cas de la régression [29].

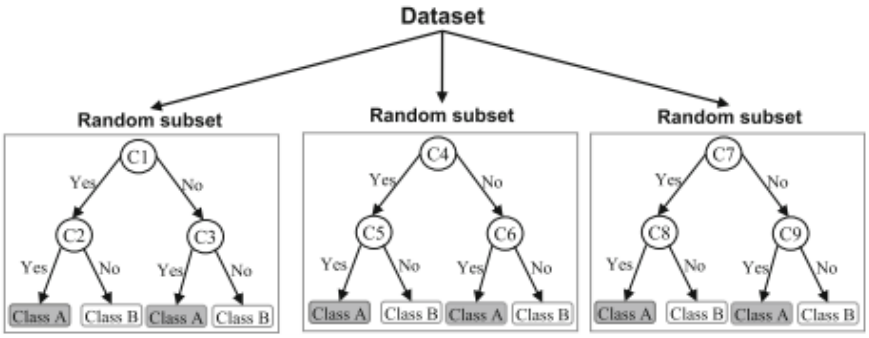

FIGURE 1.6 – Illustration de l'algorithme Random Forest

[30]

Les avantages

— L'agrégation de multiples arbres de décision diminue la probabilité que le modèle fasse du surapprentissage.

— Fournit des prédictions très précises, même avec de grands ensembles de données.

— Fonctionne bien même si les données contiennent des valeurs manquantes [28].

Les limites

— Coût de calcul élevé.

— Modèle plus complexe que d'autre [28].

2. **Support Vector Machine**

Le SVM (Machine à Vecteurs de Support) est un algorithme d'apprentissage supervisé fréquemment appliqué pour les problèmes de classification et de régression. L'algorithme SVM vise à déterminer un hyperplan qui, de manière optimale, distingue les points de données appartenant à une classe de ceux appartenant à une autre. Dans un espace en deux dimensions, cet hyperplan pourrait être une ligne, ou dans le cas d'un espace à n dimensions, il pourrait s'agir d'un plan, où n dénote le nombre de caractéristiques par observation dans l'ensemble de données. Il est possible de séparer les classes dans les données à l'aide de plusieurs hyperplans. L'algorithme SVM détermine l'hyperplan optimal, qui est celui visant à maximiser la distance entre deux classes [31].

Fonctionnement de l'algorithme SVM

Étape 1 : Mappage des données d'entrée dans un espace de représentation de dimensions supérieures

Étape 2 : Identification de l'hyperplan.

Étape 3 : Gestion du bruit et des données qui se chevauchent avec des variables Slack

Étape 4 : Optimisation des paramètres.

Étape 5 : Classification de nouvelles données [32].

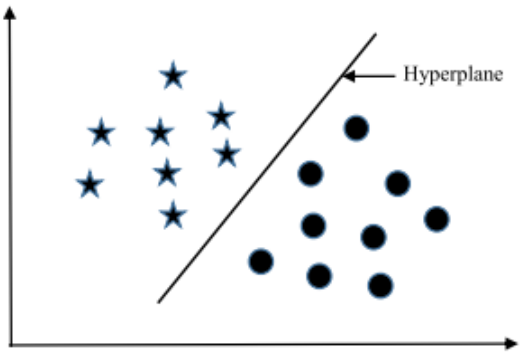

FIGURE 1.7 – Illustration de l'algorithme SVM.
[30]

Les avantages

— Efficace dans les espaces à haute dimension.

— Résistance à l'overfitting.

— Elles peuvent être utilisées pour les données linéaires et non linéaires [31].

Les limites

— Jeux de données volumineux.

— Ne s'adapte pas aux données bruitées.

— Les SVM linéaires sont faciles à comprendre et à expliquer, contrairement aux SVM non linéaires [31].

3. **Arbre de décisions**

   L'arbre de décision (Decision Tree) est un algorithme d'apprentissage automatique utilisé aussi bien pour des tâches de classification que de régression. Il permet de représenter de façon simple et visuelle les différentes étapes d'un processus décisionnel [33].

   Fonctionnement de l'algorithme arbre de décisions

   Étape 1 :Poser une question principale au niveau du nœud racine, qui est dérivée des caractéristiques de l'ensemble de données.

   Étape 2 :Diviser les données en sous-ensembles en fonction d'attributs spécifiques.

   Étape 3 :Arborescence basée sur les réponses .

   Étape 4 :Repetition de processus pour chaque nouveau sous-ensemble.

   Étape 5 :Le processus se termine lorsqu'il n'y a plus de questions utiles à poser, ce qui conduit au nœud feuille où la décision ou la prédiction finale est prise [34].

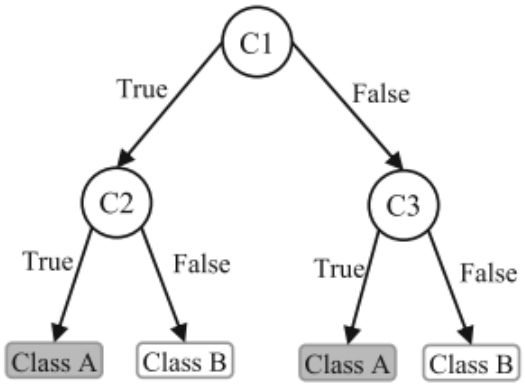

FIGURE 1.8 – Illustration d'une algorithme d'arbre de décisions.
[30]

Les avantages

— Facile à comprendre .

— Efficacité.

— Flexibilité [35].

Les limites

— Instabilité.

— Complexité.

— Prise de risques [35].

4. **k plus proches voisins (k-NN)**

L'algorithme des k plus proches voisins (KNN) compte parmi les méthodes de classification les plus anciennes et les plus élémentaires. Il assigne une étiquette à un nouvel exemple en se basant sur les « k » points de données les plus proches dans l'ensemble d'apprentissage, et constitue l'un des principaux classificateurs utilisés en apprentissage automatique. Dans l'algorithme KNN, le « K » désigne le nombre de voisins pris en compte pour le processus de vote. La sélection de différentes valeurs de « K » peut conduire à des résultats de classification distincts pour un même objet [30].

Fonctionnement de l'algorithme KNN

Étape 1 : Le choix d'un nombre de voisins (k).

Étape 2 : Calculer la distance.

Étape 3 : Déterminez les « k » voisins les plus proches .

Étape 4 : Faire une prédiction.

Étape 5 : Évaluer le modèle [36].

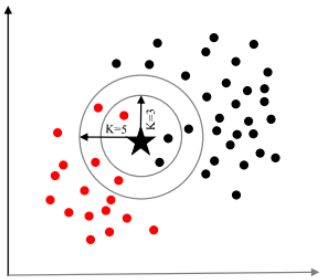

FIGURE 1.9 – Illustration d'un algorithme KNN.

[30]

Les avantages

— Mise en œuvr facile .

— Complexité réduite.

— Adaptabilité [37].

Les inconvénients

— Haute dimensionnalité .

— Sensibilité au surajustement.

— Manque de capacité de montée en charge [37].

5. **XGBoost**

XGBoost est une méthode d'apprentissage automatique basée sur les arbres de décision pour effectuer des prédictions. Plus précisément, il construit une série d'arbres de manière séquentielle, où chaque nouvel arbre vise à corriger les erreurs des précédents. Ainsi, au fil des itérations, la précision des prédictions s'améliore progressivement [38].

Fonctionnement de l'algorithme XGBoost

Étape 1 : Entraîné la premier arbre de décision sur les données.

Étape 2 : Calcul les erreurs entre les valeurs prédites et les valeurs réelles.

Étape 3 : Entraînement de l'arbre suivant sur les erreurs de l'arbre précédent pour corriger les erreurs du premier arbre.

Étape 4 : Répétition du processus

Étape 5 : Combinaison des prédictions (la somme des prédictions de tous les arbres) [39].

Les avantages

— Facile à implémenter.

— Réduction des biais éléves.

— Efficacité de calcul [40].

Les inconvénients

— Surapprentissage.

— Complexité [41].

6. **Naive Bayes**

   Le classificateur de Bayes naïf est un algorithme d'apprentissage supervisé utilisé pour des problèmes de classification, comme la catégorisation de textes. Il s'appuie sur les principes de la théorie des probabilités pour réaliser ces tâches de classification [42].

   Fonctionnement de l'algorithme Naive Bayes

   Étape 1 : Calculez la probabilité a priori associée à chaque étiquette de classe.

   Étape 2 : Évaluez la probabilité de vraisemblance de chaque attribut pour chacune des classes.

   Étape 3 : Insérez ces probabilités dans la formule de Bayes afin d'obtenir la probabilité a posteriori.

   Étape 4 : Identifiez la classe dont la probabilité a posteriori est la plus élevée et assignez l'entrée à cette classe [43].

   Les avantages

   — Simplicité.

   — Bonne évolutivité.

   — Gère des données de grande dimension[42].

   Les inconvénients

   — Sujette à la fréquence zéro .

   — Hypothèse principale irréaliste (peut conduit à des classifications incorrectes)[42].

7. **Régression logistique**

   La régression logistique est un algorithme d'apprentissage supervisé utilisé en science des données pour la classification, permettant de prédire une variable catégorielle ou discrète [44].

   Fonctionnement de l'algorithme Naive Bayes

   Étape 1 : Identifier la question.

   Étape 2 : Collecter des données historiques.

   Étape 3 : Entraîner le modèle d'analyse de régression.

   Étape 4 : Faire des prévisions pour des valeurs inconnues [45].

Les avantages

— Rapidité.

— Flexibilité.

— Simplicité [45].

Les inconvénients

— Limité aux résultats binaires.

— Surajustement et sensibilité aux valeurs aberrantes [46].

Algorithmes d'apprentissage automatique non supervisés

1. **K-Means Clustering**

   C'est l'un des algorithmes les plus utilisés en apprentissage non supervisé. Avec K-means, la machine analyse les données et les objets, puis les répartit en plusieurs clusters, de manière à ce que les objets présentant des caractéristiques similaires soient regroupés dans le même cluster, tandis que ceux ayant des caractéristiques différentes soient placés dans d'autres clusters. Les objets appartenant à un groupe donné se distinguent donc de ceux des autres clusters [47].

   Fonctionnement de l'algorithme K-Means

   Étape 1 : choisir le nombre de clusters K.

   Étape 2 : Sélectionner aléatoirement K points de données.

   Étape 3 : Associer chaque point de données au barycentre le plus proche parmi les K clusters définis.

   Étape 4 : Calculer la variation et déterminer un nouveau barycentre pour chaque cluster.

   Étape 5 : Répéter l'étape 3, c'est-à-dire réaffecter chaque point de données au barycentre le plus proche de son cluster.

   Étape 6 : Si des réaffectations continuent de se produire, revenir à l'étape 4, sinon arrêter l'algorithme.

   Étape 7 : Ainsi, le système de regroupement K-means est finalisé [47].

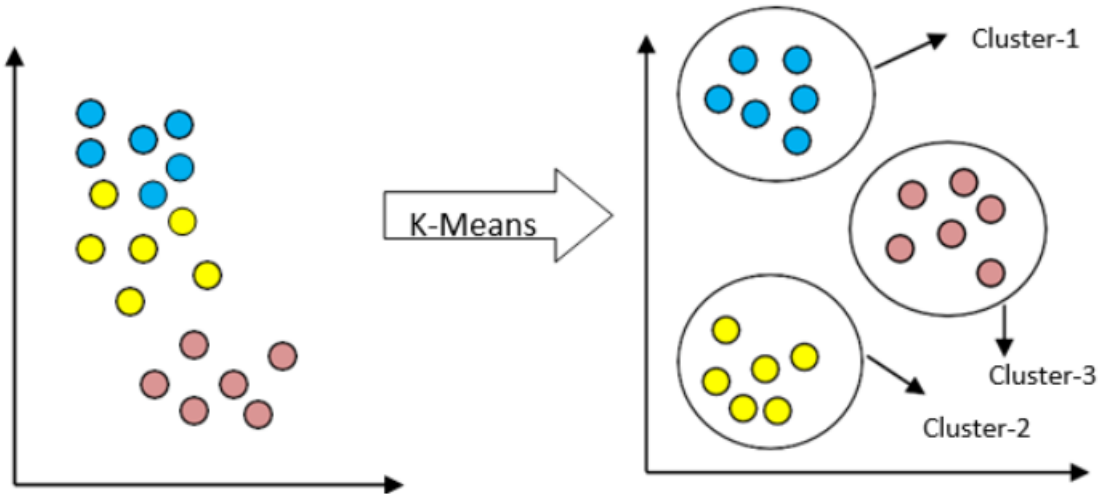

FIGURE 1.10 – Illustration d'application du clustering K-means (Avant et Après). [47]

1.6 Deep Learning

1.6.1 Définition

L'apprentissage profond est une sous-branche du Machine Learning reposant sur des réseaux de neurones artificiels à plusieurs couches. Inspiré du fonctionnement du cerveau humain, il permet aux machines d'apprendre automatiquement à partir de grandes quantités de données. Grâce aux progrès de la puissance de calcul, les modèles de Deep Learning sont devenus plus performants et capables de résoudre des problèmes complexes avec une grande précision [48].

1.6.2 Les algorithmes de Deep Learning

L'apprentissage profond utilise plusieurs types d'algorithmes, chacun adapté à des problèmes spécifiques.

1. **CNN (réseaux neuronaux convolutifs)**

   Les CNN sont spécialisés dans le traitement des images. Ils permettent de détecter automatiquement des objets, des formes ou des anomalies dans différents types d'images comme les images médicales ou satellites [49].

2. **RNN (réseaux neuronaux récurrents)**

   Les RNN sont utilisés pour les données séquentielles comme le texte, la parole et les séries temporelles. Ils prennent en compte les informations précédentes pour améliorer les prédictions [49].

3. **LSTM (Long Short-Term Memory)**

   Les LSTM sont une version améliorée des RNN, capables de mémoriser des informations sur une longue durée. Ils sont utilisés pour la traduction automatique, la reconnaissance vocale et les séries temporelles [49].

4. **RBFN (Radial Basis Function Networks)**

   Les RBFN sont des réseaux simples utilisés pour la classification, la régression et certaines tâches de prédiction [49].

1.7 Système intelligent

1.7.1 Définition

Les systèmes intelligents, également connus sous le nom de systèmes d'intelligence artificielle (IA), utilisent des algorithmes avancés pour imiter les processus cognitifs humains et effectuer des prises de décision automatisées. En associant l'apprentissage automatique, le traitement du langage naturel et l'analyse prédictive, ces systèmes sont capables d'optimiser de nombreux processus dans des domaines tels que la santé, la finance et la logistique. Leur capacité à analyser de très grands volumes de données et à s'adapter en temps réel en fait un outil particulièrement performant pour améliorer l'efficacité et favoriser l'innovation technologique [50].

1.7.2 Principe de fonctionnement général

En général, les systèmes intelligents se fondent sur l'usage de la technologie IP (Internet Protocol) et s'appuient sur des capteurs pour collecter des données provenant d'un environnement spécifique. Ces données sont par la suite transmises et partagées parmi les divers éléments du système afin d'atteindre un objectif commun. Cette liaison entre le monde réel et le monde virtuel est connue sous le nom d'Internet des Objets, qui représente une base indispensable au fonctionnement de ces systèmes.

De plus, le Big Data joue un rôle important dans le fonctionnement des systèmes intelligents, car il facilite la collecte et le traitement d'énormes volumes de données issues de diverses sources. Ces données sont par la suite analysées dans le but d'extraire des informations significatifs. En outre, grâce aux techniques d'intelligence artificielle et d'apprentissage automatique, le système a la capacité d'apprendre de ses expériences et d'optimiser ses performances progressivement [51].

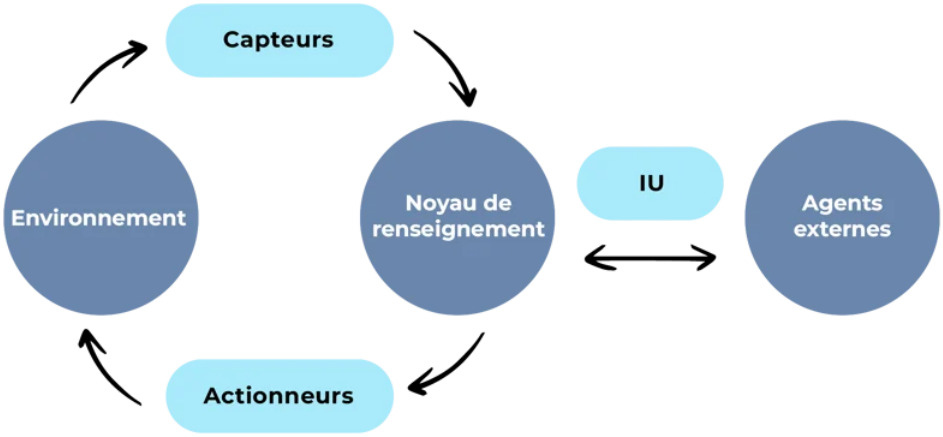

FIGURE 1.11 – Fonctionnement d'un système intelligent.
[51]

1.7.3 Architecture générale d'un système intelligent

L'architecture du système représente l'organisation générale du système. Elle décrit les différents composants, leurs rôles, ainsi que la manière dont ils interagissent entre eux. Elle repose généralement sur plusieurs modèles permettant de structurer son fonctionnement :

- Modèle fonctionnel : décrit ce que le système doit accomplir.
- Modèle physique : présente les ressources matérielles du système.
- Modèle logiciel : décrit la structure des applications et des données.
- Modèle de déploiement : montre où et comment les composants sont installés.

Elle consiste à choisir les outils, les protocoles et les normes, tout en assurant la sécurité, la fiabilité et la capacité d'évolution du système.

1.7.4 Applications des systemes intelligents

— Robots autonomes.

— Vision artificielle.

— Traitement du langage naturel.

— Systèmes experts.

— Analyse des sentiments [52].

1.8 Travaux Antérieurs

Plusieurs études ont abordé la classification de la qualité de l'eau par apprentissage automatique.

Patel et al. [53] ont proposé un modèle de prédiction de la potabilité de l'eau combinant SMOTE et intelligence artificielle explicable (XAI), appliqué au dataset Kaggle *Water Potability*. En utilisant Random Forest et Gradient Boosting, ils ont atteint une accuracy de 81%, soulignant l'importance du rééquilibrage des classes pour améliorer les performances de classification .

Uddin et al. [54] ont évalué cinq algorithmes — SVM, Naive Bayes, KNN, Random Forest et Gradient Boosting — pour la classification multi-niveaux de la qualité de l'eau côtière basée sur le WQI. Leurs résultats montrent que le Gradient Boosting obtient les meilleures performances globales pour ce type de classification multi-classes .

Ahmed et al. [55] ont développé un système de prédiction du WQI et de classification de la qualité de l'eau en utilisant plusieurs modèles d'apprentissage supervisé. Le MLP a obtenu les meilleures performances de classification avec une accuracy de 85.07%, démontrant la capacité des modèles non linéaires à capturer les relations complexes entre paramètres physico-chimiques.

1.9 Conclusion

L'Internet des objets sont des appareils intelligents communiquent entre eux permettent de collecter et d'echanger des donnees afin d'extraire des informations et de prendre des décisions en temps réel ou de manière différée.

L'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond sont fondamentalement des formes de perception machine.Il s'agit de la capacité à interpréter des données .

Dans ce chapitre, nous présenté en détail les notions fondamentales relatives à la qualité de l'eau, aussi nous avons expliqué comment l'Internet des objets peut aider les chercheurs à obtenir leurs besoin , ainsi que les concepts nécessaires à la réalisation de nos travaux.

Dans le chapitre suivant, nous décrivons en détail l'architecture de notre approche proposée.

Chapitre 2

Données, Analyse Exploratoire et Méthodologie

2.1 Introduction

Ce chapitre présente la démarche adoptée pour classifier la qualité de l'eau à l'aide de méthodes d'apprentissage automatique . Il décrit la structure du jeu de données utilisé ainsi que les principales étapes du processus, notamment l'analyse exploratoire, la construction de l'indice de qualité de l'eau (WQI) et la définition de cinq classes de qualité.

Il aborde également les étapes de prétraitement des données, incluant la gestion des valeurs manquantes, la normalisation, le rééquilibrage des classes par SMOTE et la réduction de dimension par PCA. Enfin, il présente les modèles de classification utilisés et la stratégie d'évaluation des performances

2.2 Présentation du Jeu de Données

2.2.1 Source des données

Les données utilisées dans ce travail proviennent du laboratoire de **Draa El Bordj**, situé dans la wilaya de **Bouira**. Elles sont issues d'analyses physico-chimiques réalisées sur des échantillons d'eau provenant de différents forages de la région.

Ces données ont été collectées sous forme de fichiers Excel couvrant la période allant de 2013 à 2022. Elles regroupent plusieurs paramètres environnementaux mesurant la

qualité de l'eau, notamment le pH, la température, la conductivité, la turbidité ainsi que différentes concentrations chimiques.

2.2.2 Structure du dataset

Après un processus de nettoyage , le jeu de données final contient 770 observations décrites par 15 variables physico-chimiques : pH, température (T), conductivité (Cond), turbidité (Tur), dureté totale (TH), calcium (Ca), magnésium (Mg), bicarbonates (HCO3), alcalinité totale (TAC), chlorures (Cl), sulfates (SO4), nitrates (NO3), fer (Fe), ammonium (NH4) et phosphates (PO4),ces paramètres physico-chimiques sont définis dans le Chapitre 1.

2.2.3 Variable cible (WQI) et la distribution des classes

La variable cible de ce travail est construite à partir d'un indice composite de qualité de l'eau (Water Quality Index - WQI). Cet indice est calculé à partir d'une combinaison pondérée des paramètres physico-chimiques afin de synthétiser l'information multivariée en un score unique représentatif de la qualité globale de l'eau.

Cet indice obtenu est ensuite transformé en une variable catégorielle permettant de définir cinq classes de qualité de l'eau : Dangereuse, Mauvaise, Moyenne, Bonne et Excellente.

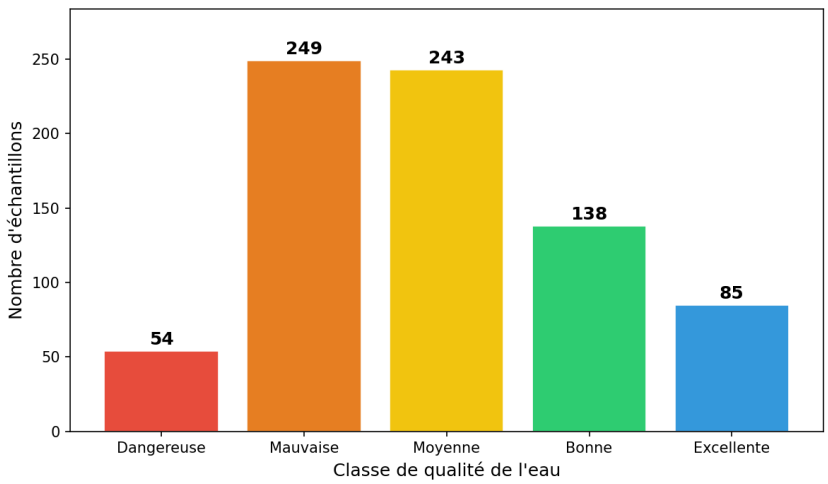

FIGURE 2.1 – Graphe-destribution des classes).

2.3 Analyse Exploratoire des Données (EDA)

2.3.1 Statistiques descriptives

Ce tableau présente les principales statistiques descriptives pour les 15 paramètres physico-chimiques, calculées sur le dataset brut.

Paramètre	Effectif	Moyenne	Écart-type	Min	Max
NH4	733	0.08	0.48	0.00	6.45
PO4	733	0.08	0.91	0.00	13.91
T	769	19.98	3.62	8.80	31.80
pH	767	7.23	0.35	6.62	9.75
Cond	769	1214.02	445.24	17.50	2990.00
Tur	767	3.84	11.11	0.05	117.00
TH	736	51.94	56.61	13.60	736.00
Ca	736	111.92	47.15	22.35	380.00
Mg	733	41.62	20.01	7.20	113.60
HCO3	736	341.59	101.92	26.50	908.90
TAC	736	35.28	45.29	4.30	439.00
Cl	733	126.41	78.30	5.68	516.17
SO4	741	151.82	192.22	2.60	1501.50
NO3	694	22.84	28.46	0.09	180.16
Fe	764	0.04	0.12	0.00	1.39

TABLE 2.1 – Statistiques descriptives des paramètres physico-chimiques

2.3.2 Analyse des corrélations

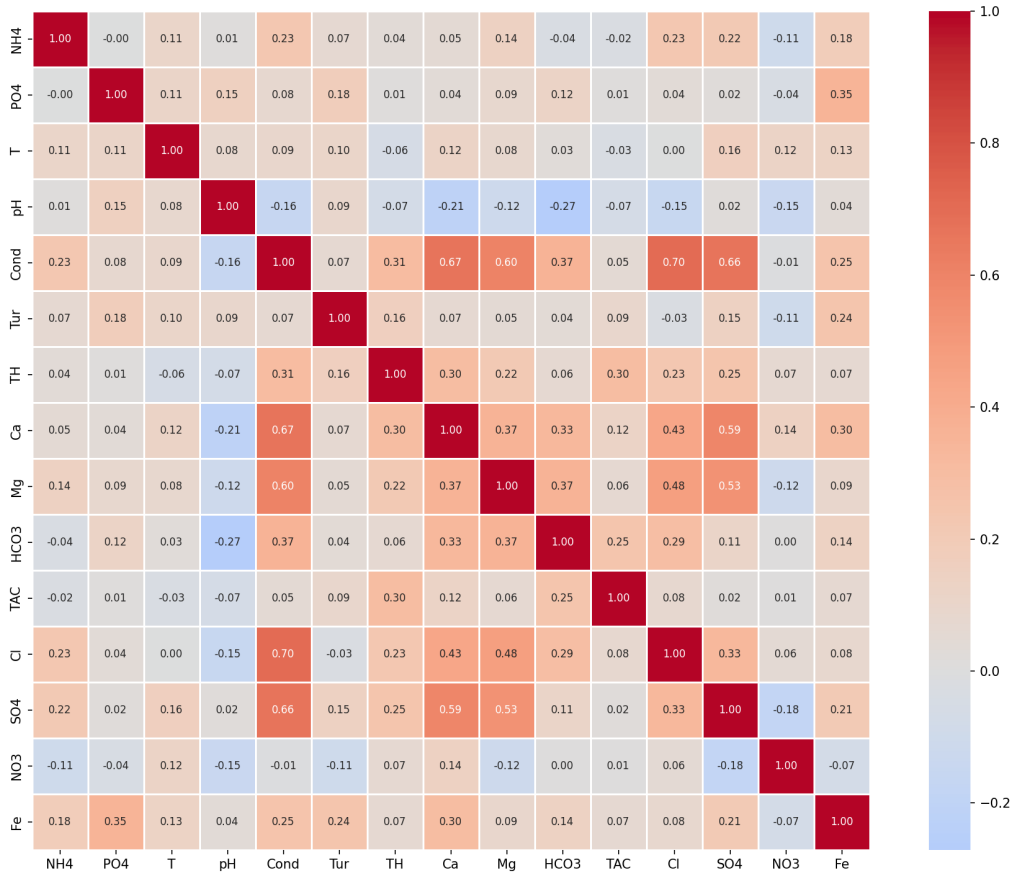

FIGURE 2.2 – La matrice de correlation des parametres .

La matrice de corrélation indique que la plupart des paramètres affichent des corrélations allant de faibles à modérées. Les relations les plus marquées concernent la conductivité et les chlorures (r = 0,70), le calcium (r = 0,67), les sulfates (r = 0,66) ainsi que le magnésium (r = 0,60). Cela implique que la conductivité a tendance à croître avec la concentration de sels minéraux dissous dans l'eau, essentiellement le calcium, le magnésium, les chlorures et les sulfates. Globalement, les corrélations restent faibles, suggérant que chaque critère fournit des données distinctes et complémentaires concernant la qualité de l'eau.

2.4 Prétraitement des Données

2.4.1 Imputation des valeurs manquantes

L'examen des valeurs manquantes révèle que tous les paramètres présentent des données incomplètes, sauf la température (T) et la conductivité (Cond). Cette figure illustre ces taux de façon comparative.

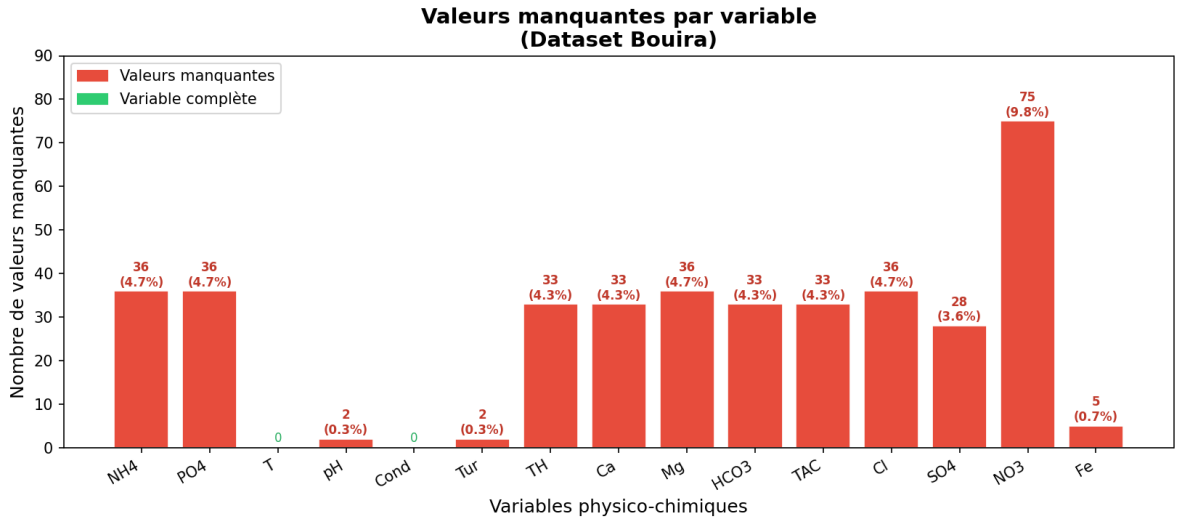

FIGURE 2.3 – Valeurs manquantes par paramètre physico-chimique. .

Ces valeurs manquantes sont traitées par **imputation par la médiane** en utilisant le code suivant :

```
imputer   = SimpleImputer(strategy='median')
X_imputed = imputer.fit_transform(X)
X_imp_df  = pd.DataFrame(X_imputed, columns=features)
```

FIGURE 2.4 – Imputation des valeurs manquantes

2.4.2 Découpage Train / Test

Notre dataset a été découpe en comme suit : ensemble d'entrainement (80%) ensemble de test (20%)

La répartition obtenue est la suivante :

Ensemble	**Nombre d'échantillons**	**Pourcentage**
Train	615	80%
Test	154	20%
Total	**769**	**100%**

TABLE 2.2 – Répartition des données en ensembles d'entraînement et de test

2.4.3 Mise à l'échelle des variables (Feature scaling)

Cette étape permet de mettre toutes les variables sur une même échelle, ce qui évite qu'une variable à grande amplitude domine les autres lors de l'apprentissage. Dans cette étude, la standardisation a été appliquée à l'aide de la méthode StandardScaler.

```
scaler  = StandardScaler()
X_train = scaler.fit_transform(X_train_raw)
X_test  = scaler.transform(X_test_raw)
```

FIGURE 2.5 – Code feature Scaling

2.4.4 Rééquilibrage des classes (SMOTE)

Le dataset présente un déséquilibre entre les quatre classes, ça peut conduire les algorithmes à privilégier les classes majoritaires, réduisant la performance sur les classes minoritaires.Pour remédier à ce problème, nous appliquons la technique**SMOTE** (*Synthetic Minority Over-sampling Technique*) [56].

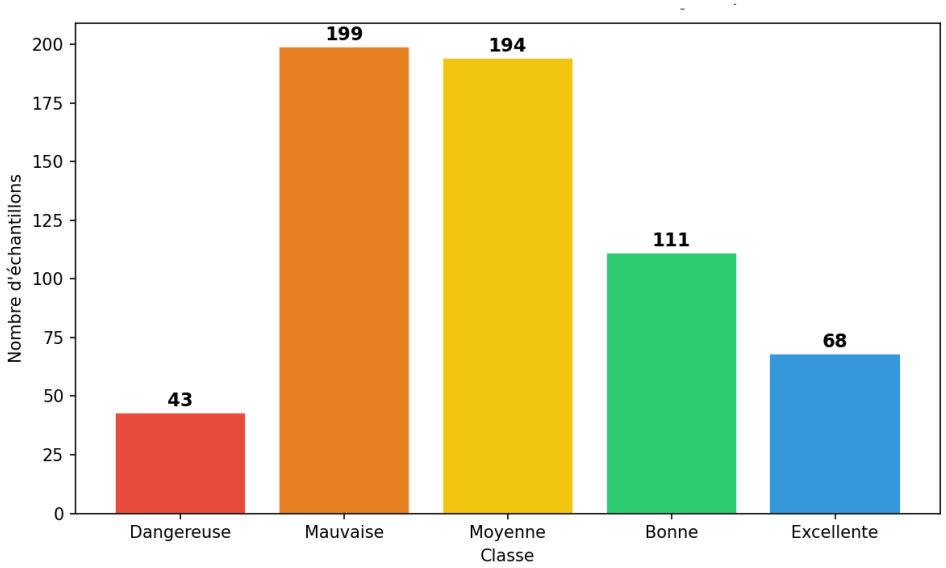

FIGURE 2.6 – Les classes avant SMOT

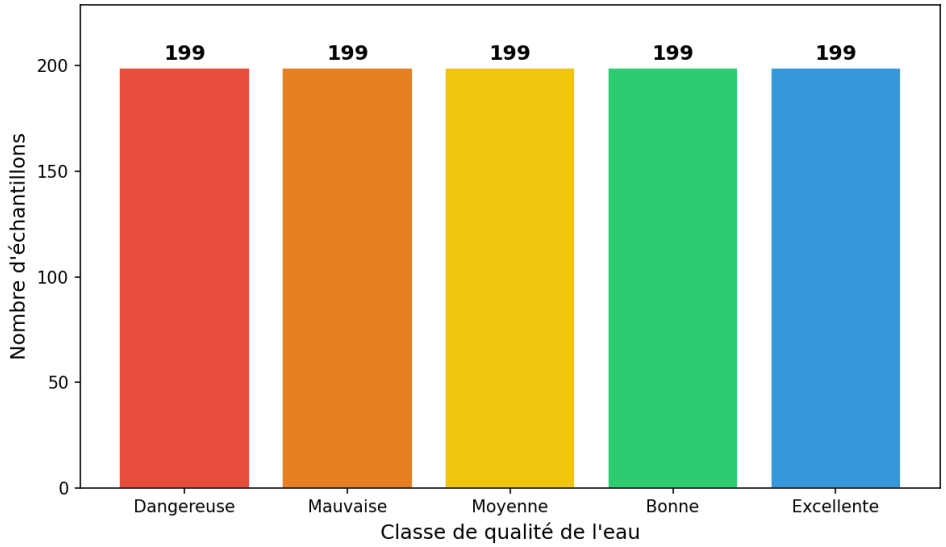

FIGURE 2.7 – Les classes après SMOT

2.4.5 Validation Croisée (Cross-Validation)

La validation croisée permet d'évaluer de manière fiable les performances d'un modèle en utilisant les données d'entraînement. Dans cette étude, une validation croisée stratifiée à 5 folds a été appliquée afin de conserver la distribution des classes dans chaque partitions.

```
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

FIGURE 2.8 – Code de validation croisée

Pour chaque fold, deux métriques sont calculées :

— **Accuracy de validation** (CV_{val}) : pourcentage de prédictions correctes.

— **F1-score pondéré** ($F1_{val}$) : indicateur global des performances du modèle.

2.4.6 Analyse en Composante (PCA)

Suite au prétraitement des données, nous avons procédé à une Analyse en Composantes Principales (PCA) dans le but de minimiser la dimensionnalité de l'ensemble des données. Les résultats montrent que **12 composantes principales** permettent d'expliquer **95,60 %** de la variance totale. Cette réduction de **15 variables à 12 composantes** conserve l'essentiel de l'information tout en simplifiant les données pour la phase de classification.

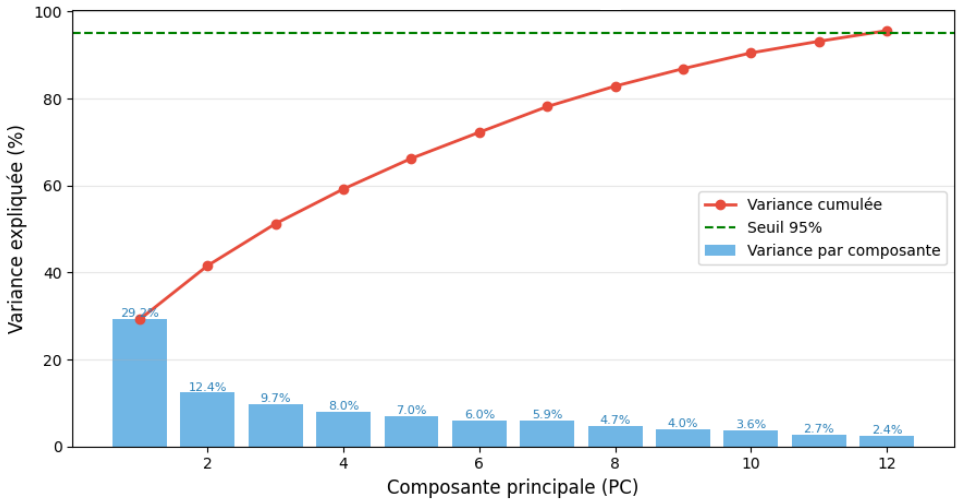

FIGURE 2.9 – PCA- variance expliquée par composante et cumulée

2.4.7 Métriques d'évaluation

L'évaluation et la comparaison des modèles reposent sur un ensemble de métriques de performance comme :

La matrice de confusion où nous comptons quatre types de résultats :

• True Positives (TP) :Nombre d'instances correctement classifiées dans leur classe réelle positive.

• True Négatives (TN) :Nombre d'instances correctement classifiées dans leur classe réelle négative.

• False positives (FP) :Nombre d'instances classifiées à tort dans la classe positive alors qu'elles appartiennent à la classe négative.

• False Négatives (FN) :Nombre d'instances classifiées à tort dans la classe négative alors qu'elles appartiennent à la classe positive. [57].

Précision Globale (Accuracy) : Proportion de prédiction correctes parmi toutes les prédictions effectuées par le système [58].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Précision :c'est La métrique qui indique combien de classes prédites qui sont correctement étiquetées. [59].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

Rappel :c'est La métrique qui indique le nombre de classes prédictives qui sont correctes[59].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

Score F1 : Le score F1 est basé sur les valeurs de précision et de rappel. Indispensable lorsque vous cherchez à trouver un équilibre entre la précision et le rappel. [59].

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

F1-score pondéré :est une métrique d'évaluation en intelligence artificielle qui calcule le F1-score global dans les problèmes de classification multi-classes[60].

$$F1_{weighted} = \sum_{i=1}^{N} w_i \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{2.5}$$

2.5 Conclusion

Dans ce chapitre, nous avons introduit le jeu de données utilisé, ainsi que les phases de prétraitement. Ces dernières visent à assurer la qualité des données recueillies pour l'évaluation à multiples niveaux de la qualité de l'eau. Le chapitre suivant sera dédié à l'évaluation et à l'analyse des résultats obtenus..

Chapitre 3

Implémentation des modèles et évaluation des performance

3.1 Introduction

Après avoir présenté notre approche proposée dans le chapitre précédent, nous abordons dans ce chapitre la phase d'implémentation et d'évaluation. Nous décrirons d'abord les différentes bibliothèques, environnements et langages de programmation utilisés pour la mise en œuvre de notre modèle. Ensuite, nous présenterons les résultats obtenus, avant de procéder à une comparaison avec ceux issus d'un projet existant.

3.2 Environnement d'implémentation et outils

3.2.1 Langage de programmation utilisé

Python

Python est un langage de programmation à la fois robuste et facile d'accès pour les débutants. Ceci offre des structures de données avancées et donne une façon aisée et efficace de travailler avec la programmation orientée objet. Grâce à sa syntaxe explicite, son typage dynamique et son mode d'exécution interprété, Python est privilégié pour la rédaction de scripts et le développement rapide d'applications dans une multitude de domaines, tout en étant compatible avec la plupart des plateformes.[61].

Les avantages de Python

— Facile à apprendre et à lire.

— Syntaxe lisible et directe.

— La popularité de Python.

— Il est extrêmement polyvalent.

— Les mises à jour de Python assurent son adaptation aux pratiques modernes [62].

3.2.2 Environnement d'implémentation

Google Colab

Google Colaboratory, est une plateforme gratuite de Google qui permet d'écrire et d'exécuter du code Python directement dans le navigateur. Elle permet d'utiliser des notebooks Jupyter sans se préoccuper de la configuration matérielle ou logicielle de l'ordinateur. Elle donne également accès à des ressources de calcul et aux principales bibliothèques d'apprentissage automatique [63].

Bénéfices de Google Colab

Google Colab est une plateforme unifiée qui propose une multitude d'avantages, tels que :

— Accessibilité

— Flexibilité dans la programmation

— Intégration avec GitHub et Google Drive

— Travail d'équipe en temps réel

— Accès facile aux bibliothèques dédiées à l'apprentissage automatique [63]

3.2.3 Bibliothèques utilisées

Pandas

Pandas est une bibliothèque Python open source utilisée pour la manipulation, l'analyse et le nettoyage des données. Elle offre des outils rapides et flexibles pour travailler avec des données tabulaires, similaires aux feuilles de calcul ou aux tables SQL [64].

NumPy

NumPy, abréviation de Numerical Python, prend en charge les grands tableaux et matrices et permet d'effectuer des opérations arithmétiques avancées sur ces tableaux. Grâce à sa puissante implémentation en C et Fortran, NumPy est la bibliothèque de choix pour le calcul en Python. Cette bibliothèque Python facilite la manipulation des tableaux et

propose des fonctions pour l'algèbre linéaire, la transformée de Fourier et le traitement des matrices [65].

Matplotlib Pyplot

Le module pyplot de Matplotlib est une interface largement utilisée qui simplifie la création de visualisations en Python [66].

Seaborn

Seaborn est une librairie Python conçue pour générer des représentations graphiques statistiques. Basée sur Matplotlib et intégrée à pandas, elle facilite la création de graphiques clairs et esthétiques grâce à des styles et palettes par défaut. Elle permet d'explorer et de mieux comprendre les relations, tendances et structures dans les données [67].

scikit-learn

scikit-learn est une bibliothèque open source de Python conçue pour simplifier le développement des modèles de machine learning. Elle propose une interface simple et cohérente adaptée aussi bien aux débutants qu'aux utilisateurs avancés. Scikit-learn prend en charge plusieurs tâches comme la classification, la régression, le clustering et le prétraitement des données. Elle fournit également des outils performants pour l'entraînement, l'évaluation et l'optimisation des modèles de manière rapide et efficace [68].

Imbalanced-learn

Imbalanced-learn est une bibliothèque Python complémentaire à scikit-learn, dédiée au traitement des jeux de données déséquilibrés. Elle implémente la technique SMOTE (*Synthetic Minority Over-sampling Technique*) permettant de générer des échantillons synthétiques pour les classes minoritaires [69].

3.3 Résultats

Dans cette étude, quatre algorithmes d'apprentissage supervisé ont été utilisés : Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) et Gradient Boosting (présentés au chapitre 2). Ces modèles ont été entraînés et évalués avant et après l'application de la PCA. Les hyperparamètres de chaque modèle ont été optimisés afin d'obtenir les meilleures performances.

3.3.1 Évaluation des modèles de classification a sans PCA

Random Forest

```
RandomForestClassifier(n_estimators=100, max_depth=8,
    max_features='sqrt', min_samples_leaf=5,random_state=42),
```

FIGURE 3.1 – Les paramètres de l'algorithme RF

	precision	recall	f1-score	support
Bonne	0.76	0.81	0.79	27
Dangereuse	0.70	0.64	0.67	11
Excellente	1.00	0.59	0.74	17
Mauvaise	0.79	0.88	0.83	50
Moyenne	0.80	0.80	0.80	49
accuracy			0.79	154
macro avg	0.81	0.74	0.76	154
weighted avg	0.80	0.79	0.79	154

FIGURE 3.2 – Performance de l'algorithme RF

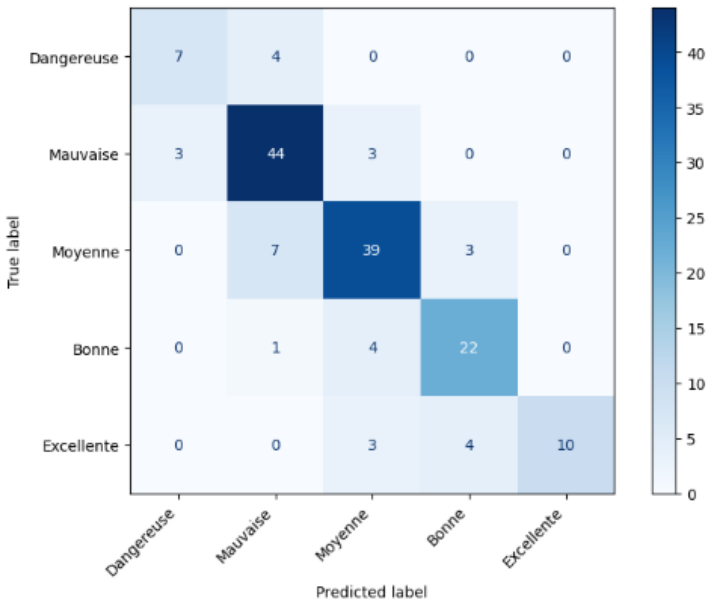

FIGURE 3.3 – Matrice de confusion de l'algorithme RF.

Les prédictions sont globalement solides sur la diagonale, surtout pour Mauvaise (44/50) et Moyenne (39/49), avec plus d'erreurs sur les classes minoritaires Dangereuse (7/11) et Excellente (10/17). Les confusions restent limitées aux classes adjacentes, sans erreurs entre extrêmes.

Gradient Boosting

```
GradientBoostingClassifier( n_estimators=200, learning_rate=0.05,  max_depth=5, random_state=42)
```

FIGURE 3.4 – Les paramètres de l'algorithme Gradient Boosting

	precision	recall	f1-score	support
Bonne	0.83	0.70	0.76	27
Dangereuse	0.62	0.45	0.53	11
Excellente	1.00	0.88	0.94	17
Mauvaise	0.80	0.90	0.85	50
Moyenne	0.81	0.86	0.83	49
accuracy			0.82	154
macro avg	0.81	0.76	0.78	154
weighted avg	0.82	0.82	0.81	154

FIGURE 3.5 – Performance de l'algorithme GBoost

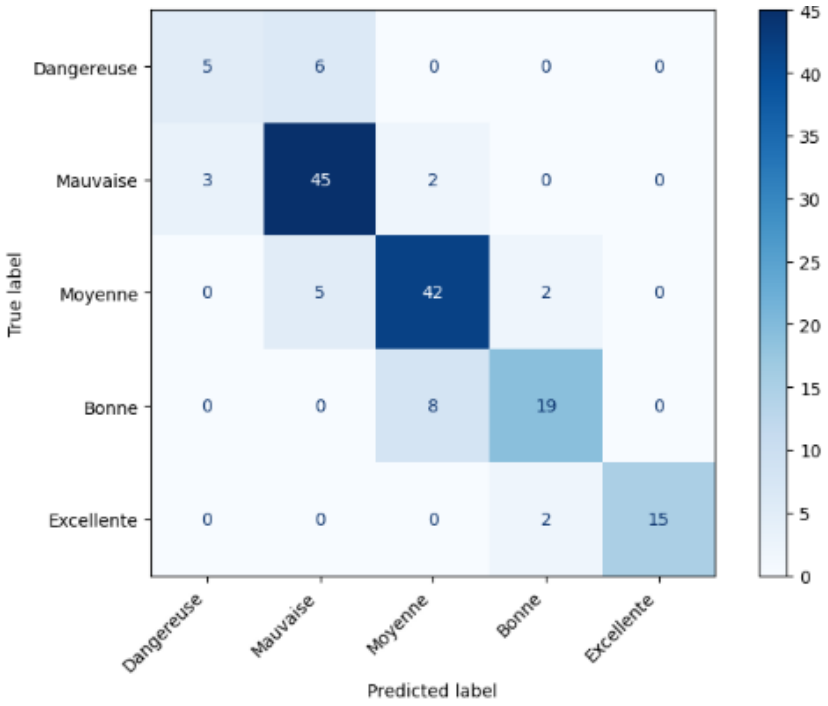

FIGURE 3.6 – Matrice de confusion de l'algorithme gboost.

Les prédictions sont solides sur la diagonale, notamment pour Mauvaise (45/50), Moyenne (42/49) et Excellente (15/17). Dangereuse reste la classe la plus faible (5/11), avec 6 cas confondus vers Mauvaise. Les erreurs restent concentrées entre classes adjacentes, sans confusion entre extrêmes.

SVM

```
SVC(kernel='rbf', C=1, gamma='scale')
```

FIGURE 3.7 – Les meilleurs paramètres de l'algorithme SVM

	precision	recall	f1-score	support
Bonne	0.81	0.93	0.86	27
Dangereuse	0.83	0.91	0.87	11
Excellente	0.94	0.88	0.91	17
Mauvaise	0.98	0.96	0.97	50
Moyenne	0.98	0.92	0.95	49
accuracy			0.93	154
macro avg	0.91	0.92	0.91	154
weighted avg	0.93	0.93	0.93	154

FIGURE 3.8 – Performance de l'algorithme SVM

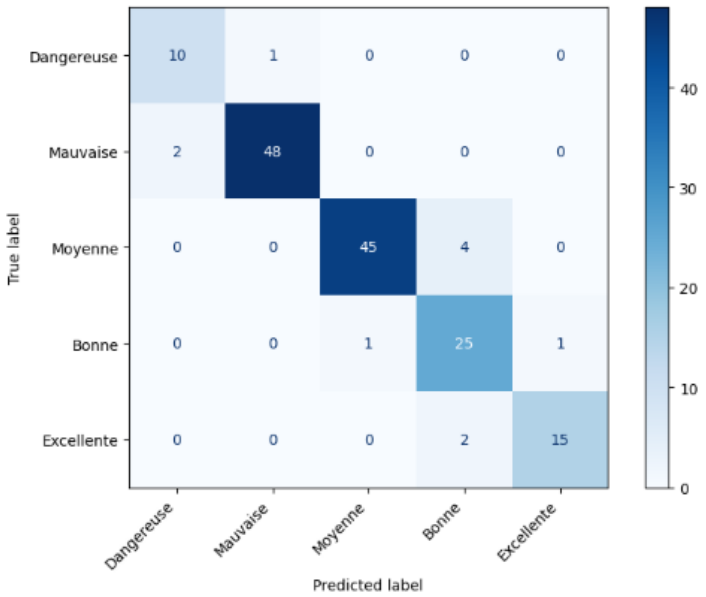

FIGURE 3.9 – Matrice de confusion de l'algorithme SVM.

Cette matrice montre d'excellentes performances sur toutes les classes, avec très peu d'erreurs : Dangereuse (10/11), Mauvaise (48/50), Moyenne (45/49), Bonne (25/26) et Excellente (15/17). Les rares confusions restent limitées aux classes adjacentes, sans erreurs entre extrêmes, confirmant un modèle très bien ajusté.

ANN

```
MLPClassifier(hidden_layer_sizes=(128, 64), activation='relu', max_iter=500,
              random_state=42),
```

FIGURE 3.10 – Les meilleurs paramètres de ANN

	precision	recall	f1-score	support
Bonne	0.96	0.89	0.92	27
Dangereuse	1.00	0.73	0.84	11
Excellente	1.00	0.94	0.97	17
Mauvaise	0.94	0.98	0.96	50
Moyenne	0.92	1.00	0.96	49
accuracy			0.95	154
macro avg	0.97	0.91	0.93	154
weighted avg	0.95	0.95	0.95	154

FIGURE 3.11 – Performance de l'algorithme ANN

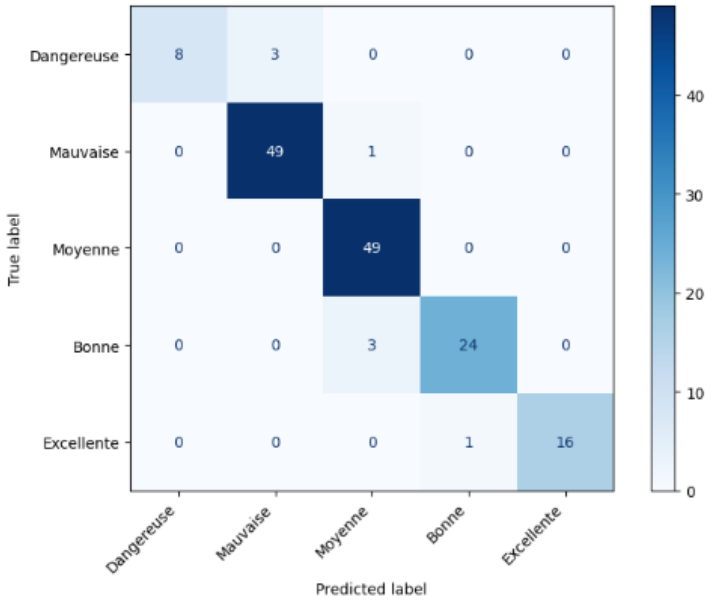

FIGURE 3.12 – Matrice de confusion de ANN.

Cette matrice montre d'excellentes performances quasi parfaites : Mauvaise (49/50), Moyenne (49/49), Bonne (24/27) et Excellente (16/17), avec Dangereuse un peu plus faible (8/11, 3 cas confondus avec Mauvaise). Les rares erreurs restent limitées aux classes adjacentes, sans confusion entre extrêmes, confirmant un modèle très bien ajusté.

3.3.2 Évaluation des modèles de classification avec PCA

Random Forest

	precision	recall	f1-score	support
Bonne	0.84	0.78	0.81	27
Dangereuse	0.75	0.82	0.78	11
Excellente	0.93	0.82	0.88	17
Mauvaise	0.94	0.94	0.94	50
Moyenne	0.90	0.96	0.93	49
accuracy			0.90	154
macro avg	0.87	0.86	0.87	154
weighted avg	0.90	0.90	0.90	154

FIGURE 3.13 – Performance de l'algorithme RF-CPA

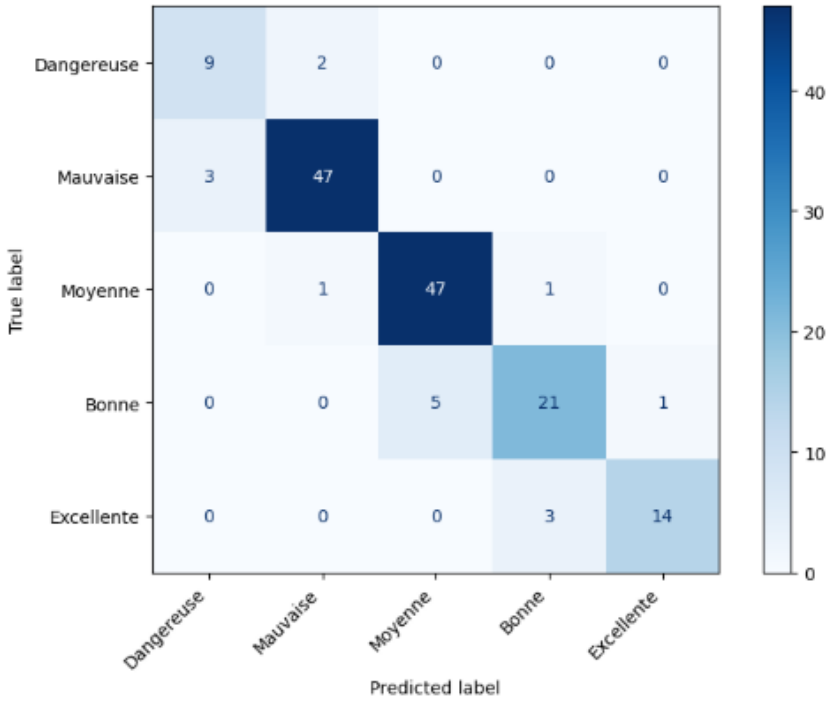

FIGURE 3.14 – Matrice de confusion de l'algorithme Random Forest-CPA.

Cette matrice montre de très bonnes performances : Dangereuse (9/11), Mauvaise (47/50), Moyenne (47/49) et Excellente (14/17), avec Bonne légèrement plus faible (21/27, 5 cas confondus avec Moyenne). Les erreurs restent limitées aux classes adjacentes, sans confusion entre extrêmes, confirmant un modèle bien ajusté.

Gradient Boosting

	precision	recall	f1-score	support
Bonne	0.88	0.81	0.85	27
Dangereuse	0.90	0.82	0.86	11
Excellente	1.00	0.88	0.94	17
Mauvaise	0.91	0.98	0.94	50
Moyenne	0.90	0.92	0.91	49
accuracy			0.91	154
macro avg	0.92	0.88	0.90	154
weighted avg	0.91	0.91	0.91	154

FIGURE 3.15 – Performance de l'algorithme GBoost-CPA

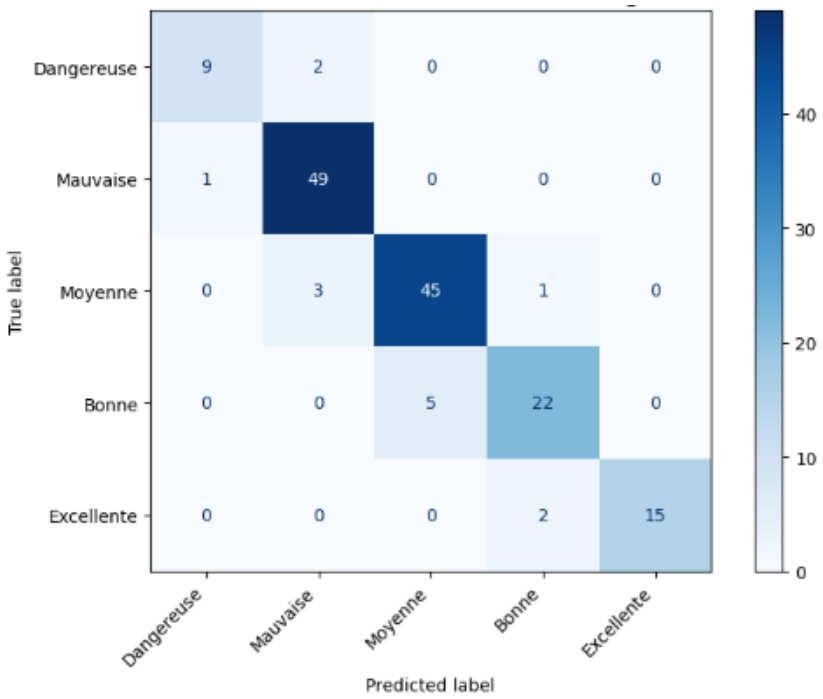

FIGURE 3.16 – Matrice de confusion de l'algorithme GBoost-CPA.

Cette matrice montre d'excellentes performances : Mauvaise (49/50), Moyenne (45/49) et Dangereuse (9/11), avec Bonne (22/27, 5 cas confondus avec Moyenne) et Excellente (15/17) également solides. Les erreurs restent limitées aux classes adjacentes, sans confusion entre extrêmes, confirmant un modèle bien ajusté.

SVM

textwidthtextwidth

FIGURE 3.17 – Performance de l'algorithme SVM-CPA

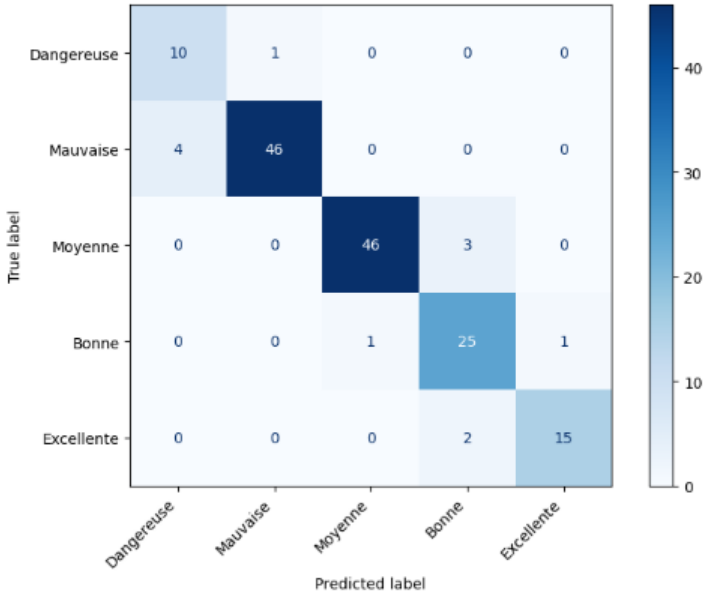

FIGURE 3.18 – Matrice de confusion de l'algorithme SVM-CPA.

Cette matrice montre d'excellentes performances : Dangereuse (10/11), Moyenne (46/49), Bonne (25/27) et Excellente (15/17), avec Mauvaise légèrement plus faible (46/50, 4 cas confondus avec Dangereuse). Les erreurs restent limitées aux classes adjacentes, sans confusion entre extrêmes, confirmant un modèle bien ajusté.

ANN

	precision	recall	f1-score	support
Bonne	0.93	0.93	0.93	27
Dangereuse	1.00	0.73	0.84	11
Excellente	1.00	1.00	1.00	17
Mauvaise	0.94	1.00	0.97	50
Moyenne	0.96	0.96	0.96	49
accuracy			0.95	154
macro avg	0.97	0.92	0.94	154
weighted avg	0.96	0.95	0.95	154

FIGURE 3.19 – Performances de ANN-CPA.

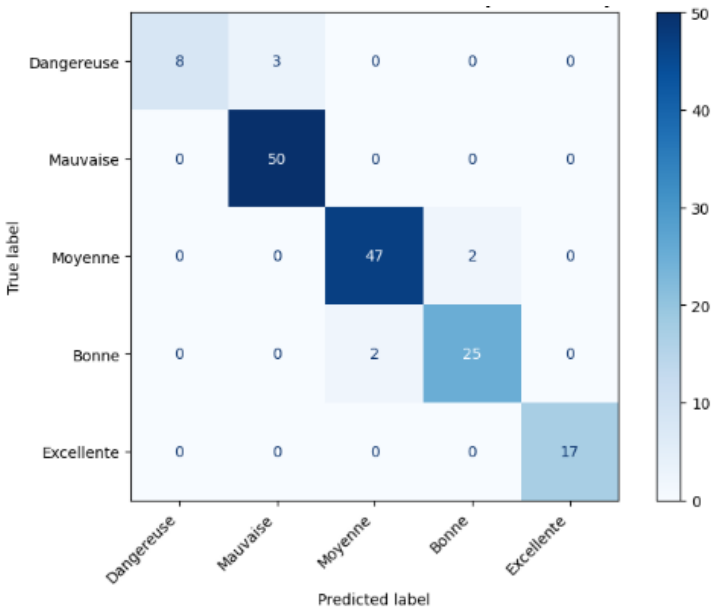

FIGURE 3.20 – Matrice de confusion de ANN-CPA.

Cette matrice montre des performances quasi parfaites : Mauvaise (50/50), Moyenne (47/49), Bonne (25/27) et Excellente (17/17), avec seulement Dangereuse un peu plus faible (8/11, 3 cas confondus avec Mauvaise). Les erreurs résiduelles restent limitées aux classes adjacentes, sans confusion entre extrêmes, confirmant un modèle très bien ajusté.

3.4 Comparaison des performances

Le tableau sauvant présente une comparaison de la confiance des algorithme utilise sans et avec PCA.

Modèle	**Test Accuracy (sans PCA)**	**Test Accuracy (avec PCA)**
Random Forest	79.22%	89.61%
Gradient Boosting	81.82%	90.91%
SVM	92.86%	92.21%
ANN	**94.81%**	**95.45%**

TABLE 3.1 – Comparaison des performances (Test Accuracy) avant et après PCA

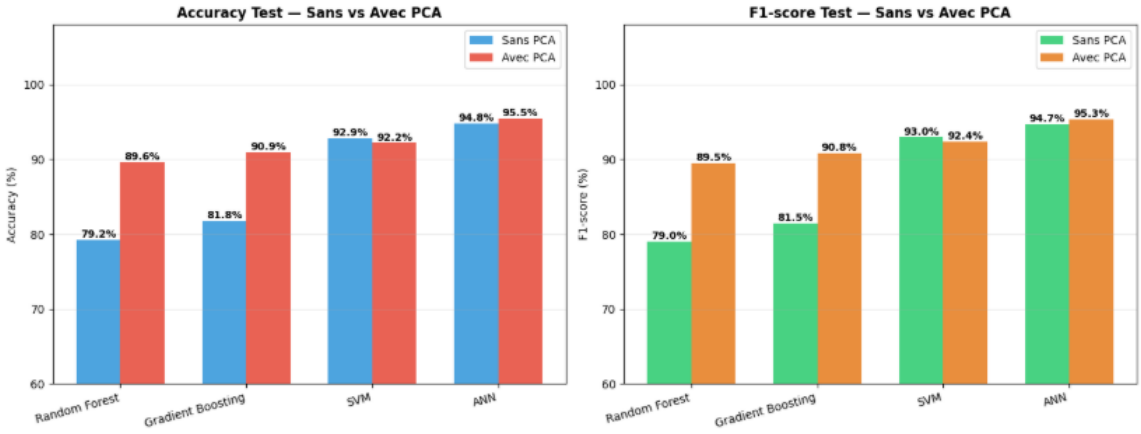

FIGURE 3.21 – Comparaison des modèles ML

On observe que l'application de la PCA améliore globalement les performances des modèles, notamment pour le Random Forest et le Gradient Boosting, tandis que le SVM et l'ANN restent globalement stables avec une légère variation. Par ailleurs, le modèle Artificial Neural Network (ANN) obtient les meilleures performances globales, atteignant la plus haute accuracy parmi les modèles testés.

3.5 Évaluation du surapprentissage des modèles

La validation croisée permet également de détecter le surapprentissage (*overfitting*) en comparant les performances sur les données d'entraînement et de validation :

$$Gap = CV_{train} - CV_{val} \tag{3.1}$$

Un gap supérieur à **5%** indique un risque de surapprentissage. Ce critère est visualisé dans la Figure

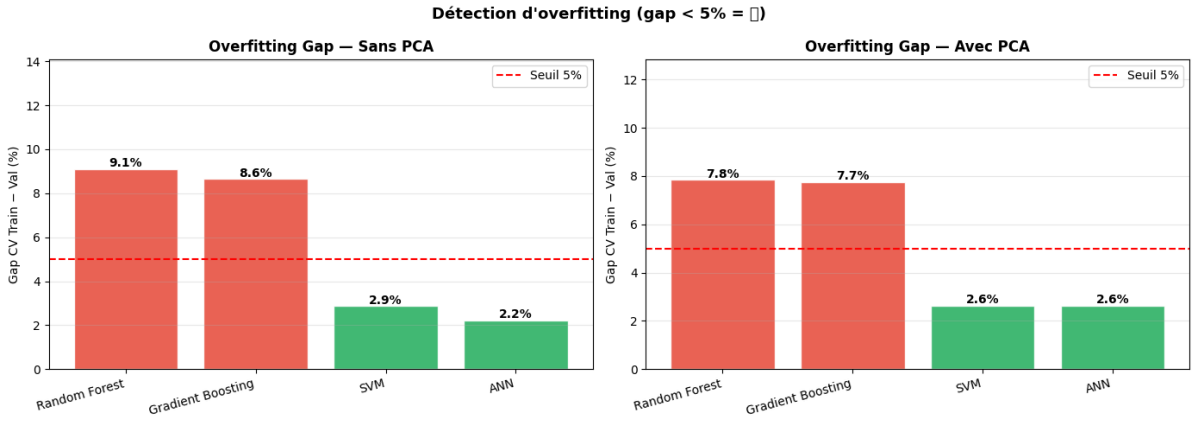

FIGURE 3.22 – Comparaison de overfeating .

Les résultats indiquent que les modèles Random Forest et Gradient Boosting présentent toujours un certain degré de surapprentissage, même si l'écart s'est atténué après l'application de l'ACP. À l'inverse, les modèles SVM et ANN présentent des écarts inférieurs à 5 %, ce qui reflète une bonne capacité de généralisation.

3.6 Conclusion

Pour terminer, ce chapitre a exposé le contexte de développement ainsi que les instruments employés pour la réalisation de l'approche suggérée. Les performances des différents modèles de classification ont été analysées et comparées avant et après l'application de l'Analyse en Composantes Principales (ACP). Les résultats montrent que l'ACP améliore globalement les performances des modèles tout en réduisant le surapprentissage pour la plupart d'entre eux. Les modèles SVM et ANN présentent les meilleurs niveaux de généralisation , indiquant une bonne stabilité face à de nouvelles données. L'ANN obtient les meilleures performances globales en termes d'accuracy et de F1-score, tandis que le SVM se distingue par le plus faible niveau de surapprentissage. Ces résultats confirment la pertinence de l'approche proposée pour la classification de la qualité de l'eau.

Conclusion Générale et Perspectives

Ce travail avait pour objectif la conception et la mise en œuvre d'un système de classification multi-niveaux de la qualité de l'eau basé sur des techniques d'apprentissage automatique. À travers les trois chapitres de ce mémoire, nous avons présenté les fondements théoriques, la méthodologie adoptée, ainsi que les résultats expérimentaux obtenus.

Nous avons montré comment les paramètres physico-chimiques de l'eau peuvent être exploités par des algorithmes de classification pour déterminer automatiquement le niveau de qualité de l'eau, sans recourir à des analyses de laboratoire coûteuses. Le système de classification développé repose sur le calcul du WQI selon la formule de Brown et al. (1972) et les normes de l'OMS, permettant une classification en quatre niveaux de qualité. Les résultats obtenus sont satisfaisants et confirment la viabilité de l'approche proposée.

Ce projet a été pour nous une expérience très enrichissante, nous permettant d'approfondir nos connaissances théoriques et pratiques dans le domaine de l'apprentissage automatique appliqué à l'environnement.

Perspectives. En guise de perspectives, nous envisageons d'améliorer davantage les performances du système en intégrant des techniques d'apprentissage profond, d'étendre l'approche à d'autres sources de données environnementales, et de déployer ce système dans un contexte réel afin de contribuer à la surveillance automatique de la qualité de l'eau au service de la santé publique.

Bibliographie

[1] World Health Organization (WHO), Guidelines for Drinking-water Quality, 4th Edition, Geneva, Switzerland, 2022.

[2] United States Environmental Protection Agency (EPA). *Water Quality Standards Handbook*. United States Environmental Protection Agency, Washington, DC, USA, 2017.

[3] Internet of Things (IoT) : A Literature Review, Journal of Computer and Communications, vol. 3, no. 5, pp. 164–173, 2015.

[4] T. M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[5] M. H. I. Dore. Climate change and changes in global precipitation patterns : What do we know ? *Environment International*, 31(8) :1167–1181, 2005.

[6] `https://sidonwater.com/fr/what-is-water-quality-and-why-is-it-important/`.

[7] `https://www.aquaportail.com/dictionnaire/definition/12555/qualite-de-l-eau`.

[8] `https://www.niuboliot.com/Product-knowledge/Significance-of-Water-Quality-Monitoring.html`.

[9] `https://www.oieau.fr/ReFEA/fiches/AnalyseEau/Physico_chimie_PresGen.pdf`.

[10] `https://lda.lozere.fr/sites/default/files/upload/analyses_physico_chimiques_des_eaux_ok.pdf`.

[11] https://datastream.org/fr-ca/guide/alcalinite.

[12] https://datastream.org/fr-ca/guide/la-durete.

[13] https://datastream.org/fr-ca/guide/les-nutriments.

[14] https://www.flandres-analyses.com/analyse-des-metaux-eaux-sols-boues-sediments-aliments/.

[15] https://www.fortinet.com/fr/resources/cyberglossary/iot.

[16] https://iotindustriel.com/iot-iiot/architecture-iot-lessentiel-a-savoir/, .

[17] https://www.ibm.com/think/topics/internet-of-things, .

[18] https://www.requea.com/surveillance-de-la-qualite-de-leau.html, .

[19] https://www.learnthings.fr/quelles-sont-les-branches-de-lintelligence-artificielle/, .

[20] Stuart Russell and Peter Norvig. *Intelligence Artificielle : Une approche moderne.* Pearson Education, Paris, France, 4ème edition, 2021.

[21] https://www.cnil.fr/fr/definition/apprentissage-automatique, .

[22] https://www.ibm.com/fr-fr/think/topics/supervised-learning, .

[23] https://devopedia.org/supervised-vs-unsupervised-learning, .

[24] https://svitla.com/blog/regression-vs-classification-in-machine-learning/, .

[25] https://www.ibm.com/fr-fr/think/topics/unsupervised-learning, .

[26] https://www.coursera.org/fr-FR/articles/reinforcement-learning, .

[27] B. Mahesh, "Machine Learning Algorithms- A Review." International Journal of Science and Research (IJSR), 2020.

[28] https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/, .

[29] https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/, .

[30] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC medical informatics and decision making, vol. 19, no. 1, pp. 1–16, 2019, .

[31] https://fr.mathworks.com/discovery/support-vector-machine.html, .

[32] http://snowflake.com/fr/fundamentals/support-vector-machine/, .

[33] https://blent.ai/blog/a/arbres-de-decision-en-machine-learning, .

[34] https://www.geeksforgeeks.org/machine-learning/decision-tree/, .

[35] https://asana.com/fr/resources/decision-tree-analysis, .

[36] https://www.servicenow.com/fr/ai/what-is-k-nearest-neighbors-algorithm.html, .

[37] https://www.ibm.com/fr-fr/think/topics/knn.

[38] https://liora.io/xgboost-tout-savoir, .

[39] https://www.geeksforgeeks.org/machine-learning/xgboost/, .

[40] https://www.jedha.co/formation-ia/algorithme-xgboost#les-avantages-de-xgboost, .

[41] https://xgboosting.com/what-is-the-xgboost-algorithm/, .

[42] https://www.ibm.com/fr-fr/think/topics/naive-bayes, .

[43] https://www.datacamp.com/fr/tutorial/naive-bayes-scikit-learn, .

[44] https://www.ibm.com/fr-fr/think/topics/logistic-regression, .

[45] https://aws.amazon.com/fr/what-is/logistic-regression/, .

[46] https://www.lemagit.fr/definition/Quest-ce-que-la-regression-logistique, .

[47] M. Suyal and S. Sharma. A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning. *Journal of Artificial Intelligence and Systems*, 6(1) :85–95, 2024.

[48] https://www.jedha.co/blog/la-formation-au-deep-learning-quapporte-t-elle.

[49] https://www.jedha.co/formation-ia/algorithmes-deep-learning.

[50] https://www.studysmarter.fr/resumes/ingenierie/genie-civil/systemes-intelligents/.

[51] https://www.algotive.ai/fr-fr/blog/syst%C3%A8mes-intelligents#:~:text=l'apprentissage.%22-,Comment%20fonctionnent%20les%20syst%C3%A8mes%20intelligents%20%3F,d'atteindre%20un%20objectif%20commun., .

[52] https://www.algotive.ai/fr-fr/blog/syst%C3%A8mes-intelligents#:~:text=Auto%2Dapprentissage%20%3A%20les%20syst%C3%A8mes%20intelligents,les%20transmettre%20par%20diff%C3%A9rents%20canaux.

[53] Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger, Komal Ladhva, Rajeev Kumar Gupta, Hashem O. Alsaab, Yusuf S. Althobaiti, and Rajnish Ratna. A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. *Computational Intelligence and Neuroscience*, 2022 :9283293, 2022. doi : 10.1155/2022/9283293. URL https://doi.org/10.1155/2022/9283293.

[54] Md. Galal Uddin, Stephen Nash, Azizur Rahman, and Agnieszka I. Olbert. Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, 169 : 808–828, 2023. doi : 10.1016/j.psep.2022.11.073. URL https://doi.org/10.1016/j.psep.2022.11.073.

[55] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan, and José García-Nieto. Efficient water quality prediction using supervised machine learning. *Water*, 11(11) :2210, 2019. doi : 10.3390/w11112210. URL https://doi.org/10.3390/w11112210.

[56] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. SMOTE : Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16 : 321–357, 2002.

[57] `https://fr.linkedin.com/pulse/evaluation-metrics-machine-learning-how-measure-model-amina-javaid-2btbf?tl=fr`.

[58] M. Benrhouma. Métriques d'Évaluation des systèmes biométriques. `https://fr.scribd.com/document/969693699/Metriques-d-Evaluation-Des-Systemes-Biometriques`, 2024. Consulté le 03 juin 2026.

[59] `https://learn.microsoft.com/fr-fr/azure/ai-services/language-service/custom-text-classification/concepts/evaluation-metrics`.

[60] A. Al-Qaness, M. Abd Elaziz, et al. Machine learning-based classification performance evaluation using weighted f1-score. *Applied Sciences*, 14(21) :9863, 2024. doi : 10.3390/app14219863. URL `https://www.mdpi.com/2076-3417/14/21/9863`.

[61] `https://docs.python.org/fr/3/tutorial/`. [consulter le : 25-04-2026].

[62] `https://liora.io/python-tout-savoir#h-quels-sont-les-avantages-de-python`. [consulter le : 25-04-2026].

[63] `https://liora.io/google-colab-tout-savoir#h-qu-est-ce-que-google-colab`. [consulter le : 26-04-2026].

[64] `https://www.geeksforgeeks.org/pandas/introduction-to-pandas-in-python/`.

[65] `https://www.geeksforgeeks.org/numpy/how-to-import-numpy-as-np/`.

[66] `https://www.geeksforgeeks.org/python/pyplot-in-matplotlib/`.

[67] `https://www.geeksforgeeks.org/python/introduction-to-seaborn-python/`.

[68] `https://www.geeksforgeeks.org/machine-learning/learning-model-building-scikit-learn-python-machine-learning-library/`.

[69] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn : A python toolbox to tackle the curse of imbalanced datasets in machine learning.

Journal of Machine Learning Research, 18(17) :1–5, 2017. URL `https://jmlr.org/papers/v18/16-365.html`.