

““
UNIVERSITÉ M’HAMMED BOUGARA BOUMERDES
Département d’Informatique / I2A

Classification de la Potabilité de l’Eau

par Méthodes de Machine Learning

Réalisé par :

SEFSAF Samia
BOUHAROUN Lyna
MZIR Massil
NECHEM Ilhem
Groupe 05

Encadré par :

Mr.BOUDAOUED

Année Universitaire 2024-2025

““

Table des matières

Introduction	1
1. Présentation détaillée du dataset	1
1.1 Origine et description générale	1
1.2 Les 9 variables mesurées	2
1.3 Analyse statistique globale	2
1.4 Problème des valeurs manquantes	3
1.5 Déséquilibre des classes	4
1.6 Relations entre variables (corrélations)	4
1.7 Conclusion de la section	5
2. Prétraitement et analyse exploratoire	5
2.1 Traitement des valeurs manquantes	5
2.2 Analyse des valeurs extrêmes (outliers)	6
2.3 Normalisation des variables	6
2.4 Division du dataset	6
2.5 Analyse en composantes principales (PCA)	6
2.5.1 variance expliquée	7
2.5.2 visualisation 2D et 3D	7
2.5.3 Bitplot	8
3. Phase non supervisée : K-Means + PCA	9
3.1 Introduction	9
3.2 Cadre théorique	9
3.2.1 l'apprentissage non supervisé	9
3.2.2 l'algorithme de clustering K_{means}	10
3.2.3 analyse en composantes principales	10
3.2.4 le dataset et ses défis	10
2.5.1 variance expliquée	7
3.3 Méthodologie	10
3.3.1 chargement et inspection des données	10
3.3.2 gestion des valeurs manquantes	11
3.3.3 standarisation et analyse en composantes	11
3.3.4 choix du nombre optimal de clusters	11
3.3.5 Décision : sélection finale	13
3.3.6 modèle $K_{means}final$	13
3.3.7 visualisation des clusters	14
3.3.8 validation externe	15
3.4 Analyse des résultats	16
3.4.1 interprétation de la visualisation 2D	16
3.4.2 limites d'apprentissage non supervisé	16
3.4.3 conclusion	17
4. Entraînement des modèles du machine learning	17
4.1 Choix et justification des algorithmes	17
4.2 K-Nearest Neighbors (KNN)	18
4.2.1 Principes théoriques	18
4.2.2 Optimisation du paramètre k	19
4.2.3 Résultats KNN Euclidien	20
4.2.4 Résultats KNN Pondéré	21

4.2.5 Visualisation (PCA 2D)	23
4.2.6 SMOTE et déséquilibre	24
4.2.7 Conclusions sur KNN	24
4.3 Réseau de Neurones Artificiels	26
4.3.1 Motivation	26
4.3.2 Configuration expérimentale	26
4.3.3 Architecture du modèle	26
4.3.4 Résultats et analyse	28
4.3.5 Conclusion et perspectives	29
4.4 Support Vector Machine(SVM)	30
4.4.1 principe théorique	30
4.4.2 choix des hyperparamètres	30
4.4.3 résultats des données brutes	30
4.4.4 Evaluation avec rééquilibrage(SMOTE)	31
4.4.5 visualisation PCA	32
4.4.6 conclusion svm	33
5. Comparaison globale des modèles	34
5.1 interprétation globale	34
Conclusion	35

Introduction

Le machine learning désigne l'ensemble des techniques permettant à un système informatique d'apprendre automatiquement à partir de données, sans être explicitement programmé pour chaque situation. Contrairement à l'approche classique, où l'on définit manuellement des règles, les modèles apprennent à identifier des motifs complexes, à généraliser l'information et à effectuer des prédictions fiables dans des contextes variés. Cette capacité d'apprentissage fait du machine learning un élément central de nombreuses applications modernes : vision artificielle, traitement du langage, sécurité informatique, systèmes de recommandation ou encore aide au diagnostic médical.

Dans ce projet, nous nous intéressons à une problématique cruciale liée à la santé publique : la prédiction automatique de la potabilité de l'eau à partir de mesures physico-chimiques. La question est de déterminer si, à partir de caractéristiques mesurées en laboratoire, un algorithme peut apprendre à distinguer l'eau potable de l'eau impropre à la consommation. Cette tâche constitue un véritable défi en classification supervisée : données complexes, classes fortement superposées, non-linéarité potentielle des relations, et déséquilibre marquant entre les étiquettes.

L'objectif de ce travail est d'appliquer et comparer différentes techniques de machine learning supervisé afin d'évaluer leur capacité à modéliser ce problème réel. Pour cela, nous avons mis en œuvre plusieurs algorithmes (K-Nearest Neighbors optimisé, Support Vector Machine et réseau de neurones artificiel), complétés par une analyse approfondie incluant l'exploration des données, le traitement du déséquilibre, la réduction de dimension par PCA et l'optimisation d'hyperparamètres. Au-delà des performances obtenues, ce projet constitue une démonstration pratique des bonnes pratiques d'apprentissage automatique et met en lumière les limitations rencontrées lorsqu'un problème présente des frontières de décision complexes.

1 Présentation détaillée du dataset

1.1 Origine et description générale

Le dataset utilisé dans ce travail est le **Water Potability Dataset**, largement employé dans les travaux académiques de classification environnementale. Il contient un ensemble d'échantillons d'eau analysés en laboratoire, chaque échantillon étant décrit par 9 caractéristiques physico-chimiques permettant d'évaluer indirectement la qualité de l'eau. L'objectif est de prédire une étiquette binaire : 0 pour eau non potable et 1 pour eau potable.

Le dataset compte 3276 échantillons, chacun représentant une mesure indépendante. Cette taille est suffisante pour entraîner des modèles supervisés mais reste modérée, ce qui rend la tâche plus délicate comparée aux grands datasets industriels.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

FIGURE 1 – Aperçu des premières lignes du dataset

1.2 Les 9 variables mesurées

Feature	Description	Rôle dans la potabilité
pH	Mesure de l'acidité ou basicité de l'eau (0-14)	L'eau potable doit se situer entre 6,5 et 8,5.
Hardness	Concentration en minéraux (calcium, magnésium)	Une dureté trop élevée peut être problématique pour la consommation.
Solids	Solides dissous totaux (TDS)	Trop de solides = mauvais goût + risques sanitaires.
Chloramines	Agents désinfectants	Trop faible → bactéries; trop élevé → toxique
Sulfate	Concentration en ions sulfate	Excès → troubles gastro-intestinaux
Conductivity	Capacité de l'eau à conduire l'électricité	Indicateur indirect de minéralisation
Organic_carbon	Teneur en carbone organique total	Mesure de la pollution organique
Trihalomethanes	Sous-produits de chlorination	Surveillés dans toutes les stations de traitement.
Turbidity	Clarté de l'eau (particules en suspension)	Élevée → risque de contamination

TABLE 1 – Les 9 variables du dataset et leur lien avec la potabilité

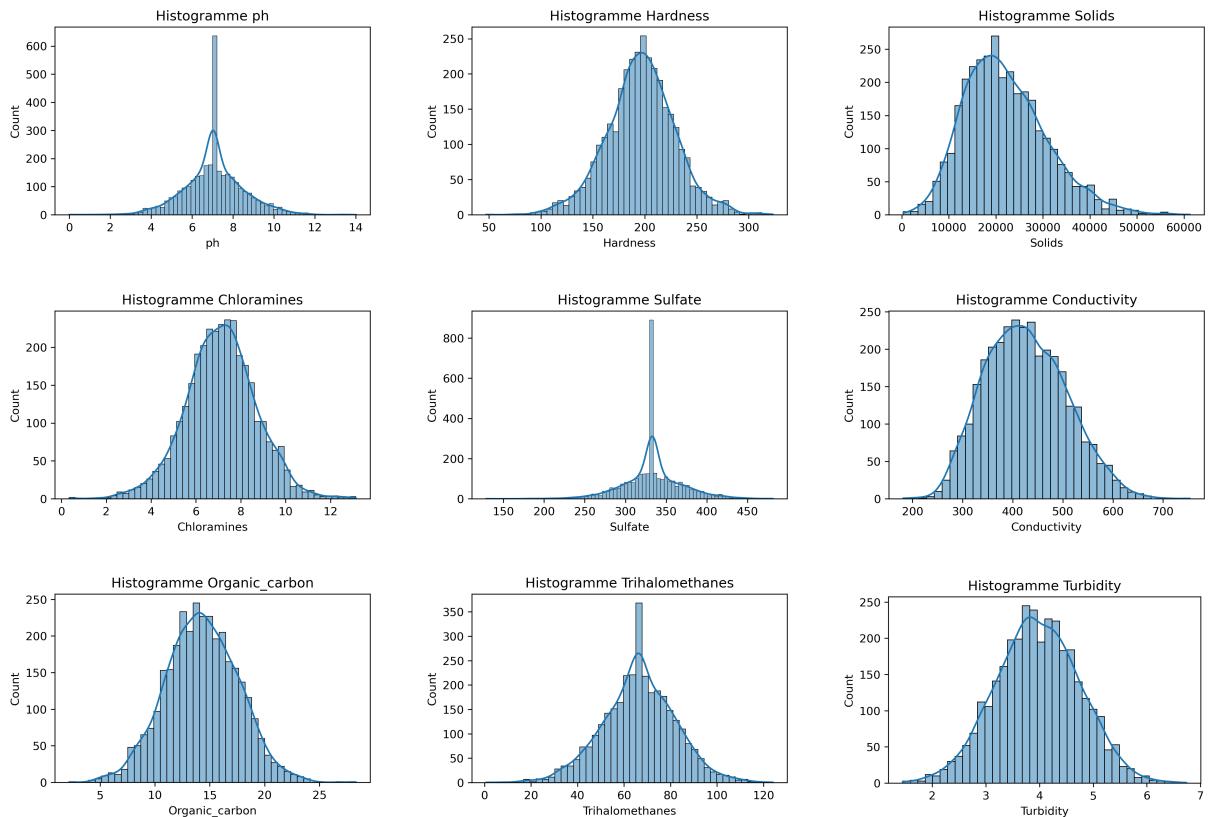


FIGURE 2 – Distributions des 9 variables physico-chimiques (histogrammes)

1.3 Analyse statistique globale

Pour chaque variable, nous avons calculé la moyenne, la médiane, la variance, l'écart-type, les valeurs minimale et maximale ainsi que la présence d'outliers. Les résultats sont présentés dans le tableau suivant :

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000

FIGURE 3 – Statistiques descriptives des variables

Ces statistiques révèlent des écarts très importants pour certaines variables (Solids, Sulfate, Conductivity), des distributions souvent non gaussiennes et fortement asymétriques, ainsi qu'une dispersion élevée qui complique la tâche de classification.

Les boxplots ci-dessous permettent de visualiser clairement les outliers et la dispersion :

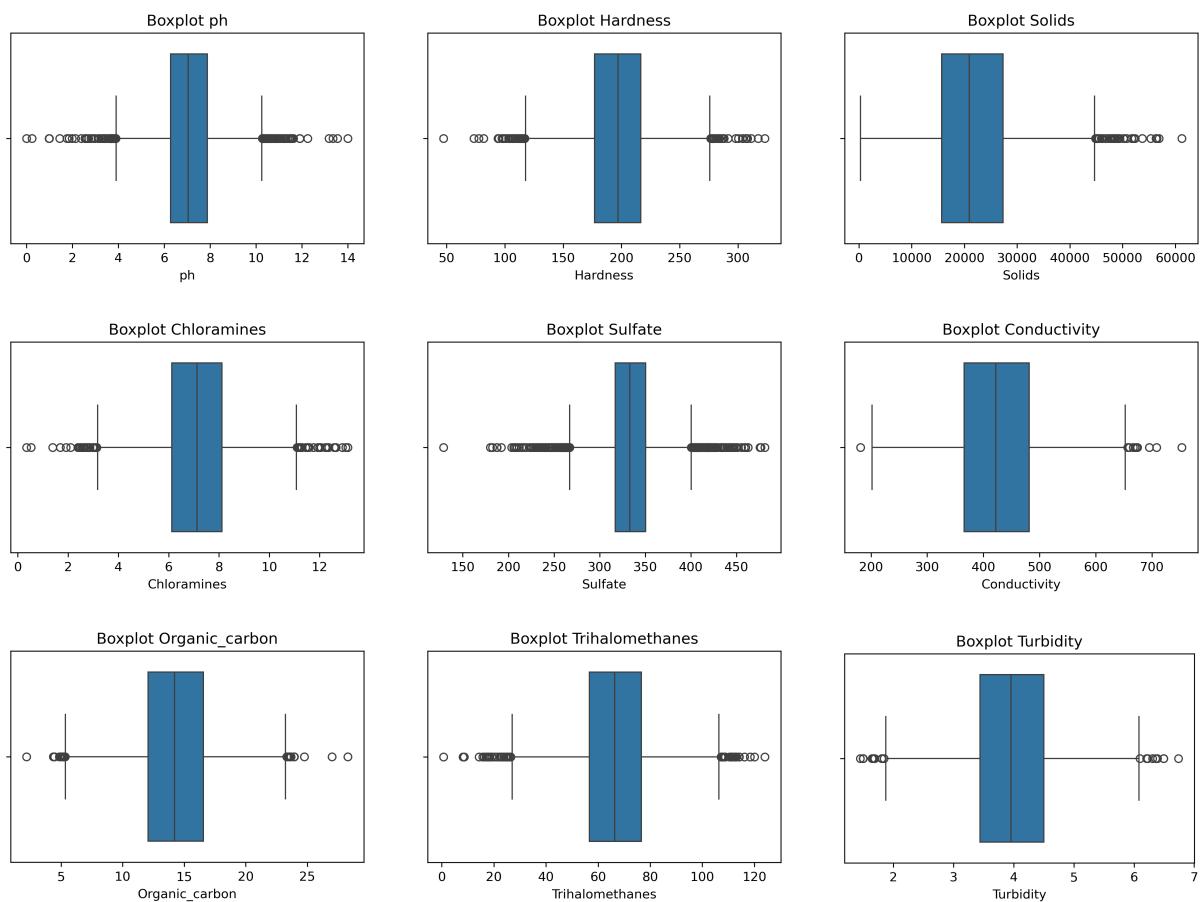


FIGURE 4 – Boxplots des 9 variables physico-chimiques

1.4 Problème des valeurs manquantes

Le dataset présente un taux significatif de valeurs manquantes, compris entre 5 % et 10 % selon les variables. Les colonnes les plus affectées sont PH , Sulfate et Trihalomethanes.

Ces absences empêchent l'entraînement direct des modèles, modifient les distributions et peuvent introduire des biais si elles ne sont pas correctement traitées. Elles ont donc été imputées ou supprimées selon la stratégie retenue lors du prétraitement.

<code>ph</code>	491
<code>Hardness</code>	0
<code>Solids</code>	0
<code>Chloramines</code>	0
<code>Sulfate</code>	781
<code>Conductivity</code>	0
<code>Organic_carbon</code>	0
<code>Trihalomethanes</code>	162
<code>Turbidity</code>	0
<code>Potability</code>	0

FIGURE 5 – somme des valeurs manquantes pour chaque variable

1.5 Déséquilibre des classes

La répartition des étiquettes est la suivante : 61 % d'échantillons non potables (classe 0) et 39 % d'échantillons potables (classe 1). Ce déséquilibre marqué constitue un obstacle majeur : un modèle naïf prédisant systématiquement « non potable » atteindrait déjà 61 % d'accuracy. Il est donc à l'origine d'un recall faible sur la classe potable et de matrices de confusion fortement déséquilibrées.

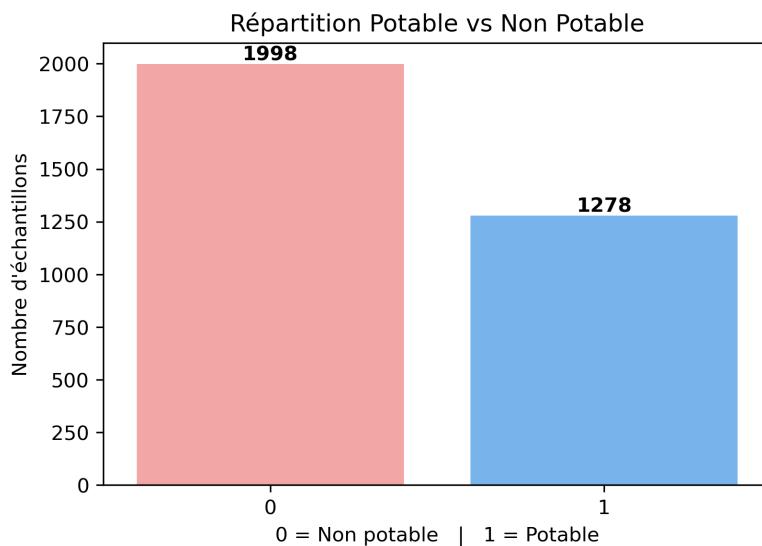


FIGURE 6 – Répartition des classes potability

1.6 Relations entre variables (corrélations)

L'analyse de corrélation montre des coefficients globalement faibles à modérés entre les variables, aucune relation dominante et des redondances très limitées. Aucun paramètre ne détermine à lui seul la potabilité ; les relations sont majoritairement non linéaires et complexes, ce qui rend la séparation entre les deux classes particulièrement difficile.

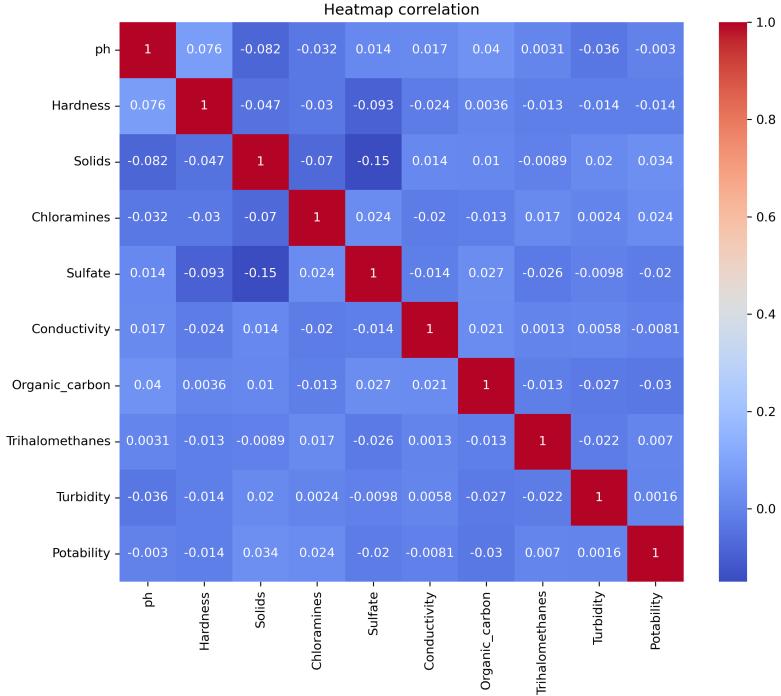


FIGURE 7 – Heatmap de corrélation

1.7 Conclusion de la section

Le dataset Water Potability présente plusieurs caractéristiques qui expliquent la difficulté de la tâche de classification : données bruitées et fortement dispersées, présence de valeurs manquantes, déséquilibre important des classes, corrélations faibles entre variables et frontières de décision très floues. Cette analyse constitue une base essentielle pour comprendre et interpréter les résultats obtenus dans la suite du projet.

2 Prétraitement et analyse exploratoire des données

Cette section présente l'ensemble des opérations réalisées avant l'entraînement des modèles : nettoyage, imputation, analyse statistique, normalisation, division des données et analyse PCA. Ce prétraitement est essentiel pour garantir la qualité des données et maximiser la capacité d'apprentissage des modèles supervisés.

2.1 Traitement des valeurs manquantes

L'examen initial du dataset a révélé des données manquantes dans trois variables : pH (15 %), Sulfate (24 %) et Trihalomethanes (5 %).

Après analyse des distributions, l'imputation suivante a été retenue :

- pH et Sulfate : remplacement par la médiane, en raison de leur caractère asymétrique et de la présence d'outliers marqués ;
- Trihalomethanes : remplacement par la moyenne, la distribution étant plus symétrique.

Cette stratégie a permis d'éliminer l'ensemble des valeurs manquantes tout en minimisant l'introduction de biais statistique.

2.2 Analyse des valeurs extrêmes (outliers)

La détection des outliers a été réalisée via la méthode de l'IQR (Interquartile Range). Elle a mis en évidence un nombre important de valeurs extrêmes dans plusieurs variables, notamment ph, Hardness, Sulfate, Chloramines et Trihalomethanes. Ces valeurs extrêmes ne sont pas des erreurs de mesure évidentes mais reflètent la variabilité naturelle des conditions physico-chimiques de l'eau. Elles peuvent, par exemple, signaler :une contamination ponctuelle,un déséquilibre temporaire du pH,une surconcentration en solides ou composés chimiques.

Afin de ne pas perdre cette information pertinente, les outliers ont été conservés dans l'ensemble du projet.

Variable	Nombre d'outliers
ph	142
Hardness	83
Solids	47
Chloramines	61
Sulfate	264
Conductivity	11
Organic_carbon	25
Trihalomethanes	54
Turbidity	19

TABLE 2 – Nombre d'outliers détectés par la méthode IQR

2.3 Normalisation des variables

Les neuf variables présentant des échelles très hétérogènes, une normalisation centrée-réduite (StandardScaler) a été appliquée a fin de présenter les features sur la même échelle . Cette étape est indispensable pour :éviter qu'une variable à grande échelle (comme Solids) domine les autres,garantir un apprentissage équilibré pour les algorithmes sensibles à l'échelle : KNN, SVM, ANN,et stabiliser les gradients lors de l'entraînement du réseau de neurones.

2.4 Division de l'ensemble de données

Le dataset a été séparé en ensemble d'entraînement (80 %) et ensemble de test (20 %), en utilisant le paramètre stratify = y afin de conserver exactement la même proportion de classes (61 % non potable, 39 % potable) dans les deux sous-ensembles. Cette précaution évite tout biais d'évaluation lié au déséquilibre initial.

2.5 Analyse en composantes principales (PCA)

L'analyse en (PCA) a été réalisée sur les données normalisées afin d'explorer la structure du dataset et la séparabilité des classes.

2.5.1 Variance expliquée

- La première composante principale (PC1) explique seulement **13,7 %** de la variance totale.
 - Les deux premières composantes cumulent **26,2 %** de la variance.
 - Pour conserver 95 % de l'information, il faut garder **8 composantes sur 9**.
- les variables sont très peu corrélées entre elles, et l'information est répartie de façon homogène sur toutes les dimensions. Il n'existe pas de directions dominantes dans l'espace des features.

Raison physique / scientifique de cette faible séparabilité

La potabilité de l'eau dépend d'un équilibre complexe entre plusieurs paramètres qui peuvent se compenser mutuellement (ex : forte concentration en solides mais bonne conductivité, pH bas mais faible turbidité, etc.). Les plages de valeurs acceptables pour l'eau potable et non potable se chevauchent largement → les deux classes occupent presque le même espace dans les 9 dimensions.

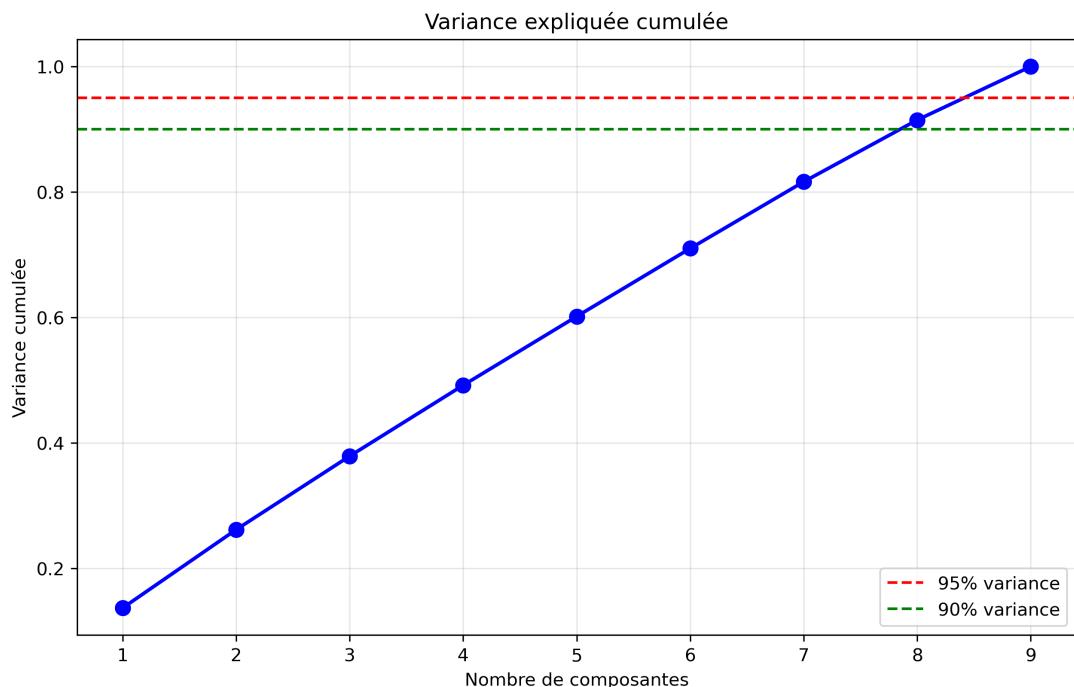


FIGURE 8 – Variance expliquée cumulée par les composantes principales

2.5.2 Visualisation 2D et 3D

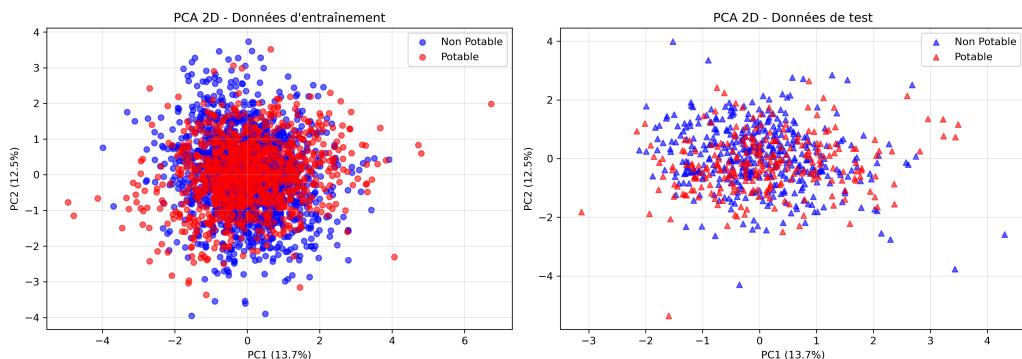


FIGURE 9 – Projection 2D des données sur PC1 et PC2 (train + test)

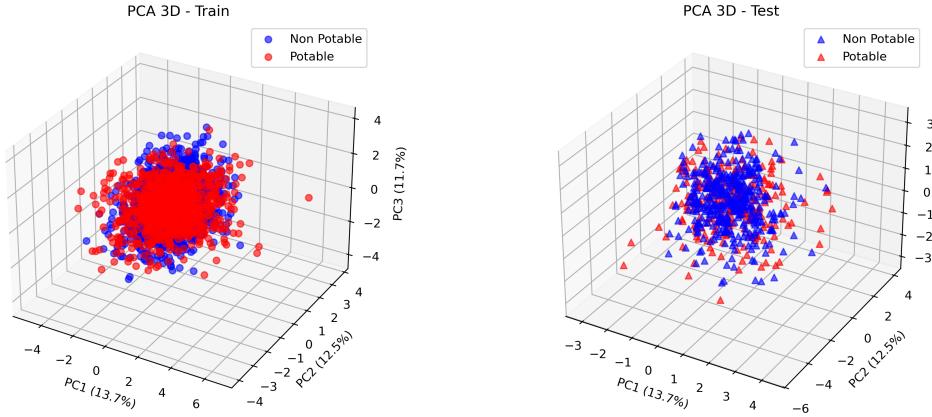


FIGURE 10 – Projection 3D des données sur PC1, PC2 et PC3

Les projections en 2D et 3D montrent un grand nuage de points fortement mélangé entre les deux classes (potable en rouge, non potable en bleu), aussi bien sur le train que sur le test.

Aucune séparation linéaire claire n'apparaît, même en 3 dimensions.

→ Cela explique les difficultés rencontrées par les modèles de classification supervisée (KNN, SVM, ANN) : les deux classes occupent presque le même espace dans les dimensions les plus informatives.

2.5.3 Biplot – Contribution des variables

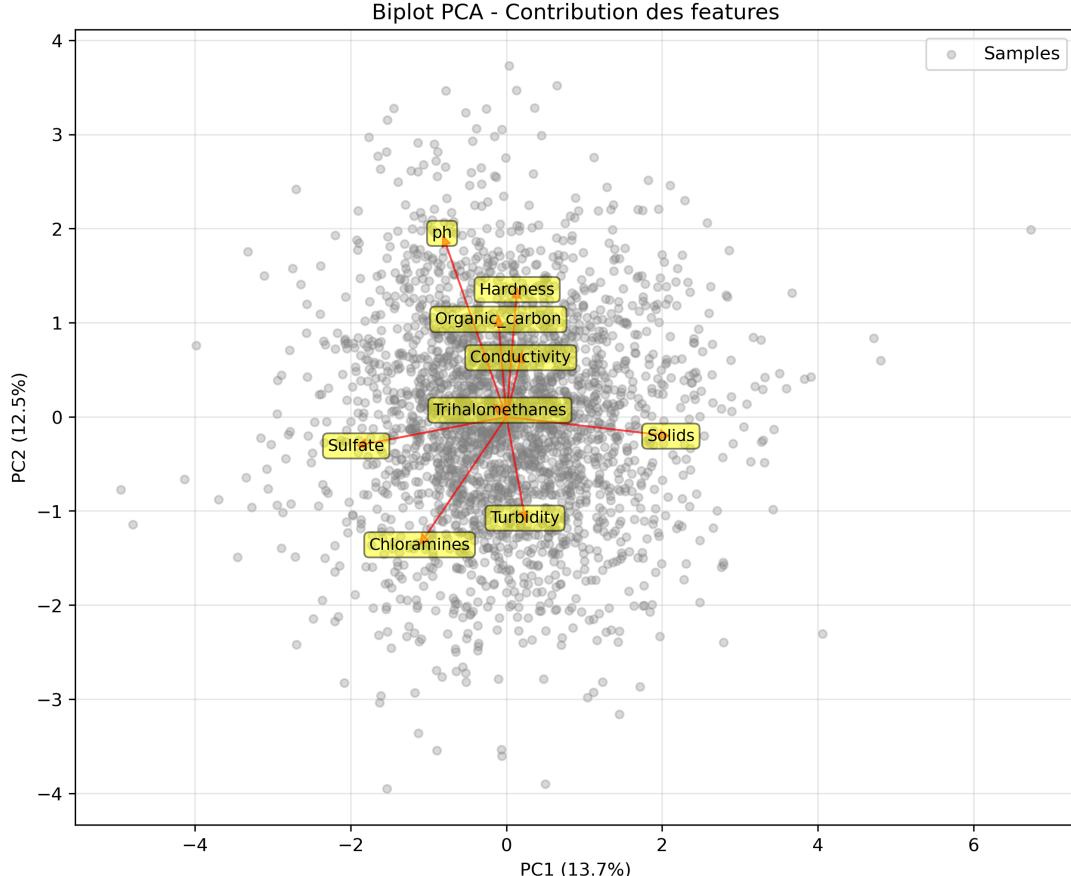


FIGURE 11 – Biplot PCA – contribution et corrélation des variables

Le biplot montre que les vecteurs des 9 features pointent dans des directions très différentes

et ont des longueurs similaires.

Aucune variable ne domine fortement PC1 ou PC2.

Interprétation poussée du biplot

- La longueur du vecteur indique l'importance de la variable dans PC1 et PC2
- L'angle entre deux vecteurs reflète leur corrélation :
 - angle proche de 0° = corrélation positive forte
 - angle 90° = variables indépendantes
 - angle 180° = corrélation négative

Ici, les angles variés et les longueurs similaires confirment que toutes les variables contribuent de façon équilibrée et sont globalement peu corrélées.

→ Toutes les mesures physico-chimiques (pH, Hardness, Solids, Chloramines, etc.) contribuent de façon relativement équilibrée à la variance. Il n'y a pas de variable "clé" évidente. L'analyse PCA confirme que le problème de classification de la potabilité de l'eau est difficile :

- pas de séparation linéaire nette entre les classes,
- information très dispersée sur toutes les dimensions,
- aucune feature ne permet à elle seule de distinguer clairement les deux classes.

La PCA démontre objectivement que le dataset n'a aucune structure linéaire exploitable, ce qui justifie pleinement les performances limitées observées sur les modèles supervisés. Le dataset Water Potability constitue donc un excellent cas d'étude réaliste montrant les limites des méthodes classiques face à des données bruitées, non linéaires et déséquilibrées.

3 Phase non supervisée : Regroupement par K-Means avec prétraitement PCA

3.1 Introduction

L'accès à une eau potable sûre constitue un enjeu majeur de santé publique. Le dataset Water Potability fournit neuf paramètres physico-chimiques pour 3 276 échantillons d'eau accompagnés d'une étiquette binaire de potabilité. Cette étude ignore délibérément l'étiquette pendant la modélisation afin d'examiner si des groupements naturels dans l'espace des variables correspondent aux classes de potabilité. L'analyse combine le clustering K-Means avec un prétraitement basé sur PCA et évalue les résultats à l'aide de métriques de validité interne des clusters ainsi que par comparaison externe avec l'étiquette réelle.

3.2 Cadre théorique : apprentissage non supervisé

3.2.1 L'apprentissage non supervisé en machine learning

L'apprentissage non supervisé regroupe les algorithmes qui identifient des structures cachées dans des données non étiquetées. Les deux tâches principales sont le clustering (regroupement d'observations similaires) et la réduction de dimension (recherche de représentations compactes). Ce projet intègre les deux approches pour explorer la structure latente des mesures de qualité de l'eau.

3.2.2 L'algorithme de clustering K-Means

K-Means est un algorithme de partitionnement basé sur des prototypes qui divise n observations en k clusters disjoints en minimisant la somme des carrés intra-cluster (inertie) :

$$J(\mathbf{C}) = \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2 \quad (1)$$

où C_i est le i -ème cluster et $\mu_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$ est son centroïde. L'algorithme de Lloyd (Lloyd, 1982) alterne entre :

- **Étape d'affectation** : chaque point est assigné au centroïde le plus proche à l'aide de la distance euclidienne.
- **Étape de mise à jour** : les centroïdes sont recalculés comme la moyenne des points assignés.

L'implémentation scikit-learn utilise l'initialisation k-means++ (Arthur & Vassilvitskii, 2007) et plusieurs redémarrages (`n_init=50`) pour éviter la convergence vers de mauvais minima locaux. La méthode suppose des clusters approximativement sphériques et de volume similaire.

3.2.3 Analyse en composantes principales

La PCA est une transformation linéaire orthogonale qui projette les données sur des axes de variance maximale. Étant donné une matrice de données centrée $\mathbf{X} \in R^{n \times p}$, la PCA effectue une décomposition en valeurs propres de la matrice de covariance, produisant des valeurs propres $\lambda_1 \geq \dots \geq \lambda_p$ et les vecteurs propres correspondants. Conserver les composantes qui expliquent au moins 90 % de la variance totale permet simultanément de réduire la dimension, d'éliminer le bruit et la multicolinéarité — étapes bénéfiques pour les algorithmes de clustering basés sur la distance.

3.2.4 Le dataset Water Potability et ses défis

Le dataset contient 3 276 échantillons avec neuf variables continues et une cible binaire Potability. Les caractéristiques notables incluent des valeurs manquantes substantielles, un déséquilibre modéré de classes et un chevauchement considérable entre les distributions potable et non potable. Même les modèles supervisés les plus performants n'atteignent généralement que 60–65 % d'accuracy, reflétant un faible rapport signal/bruit. Par conséquent, une séparation non supervisée parfaite n'est pas attendue ; tout enrichissement significatif de classe dans les clusters découverts constitue déjà une preuve de structure sous-jacente.

3.3 Méthodologie et implémentation pratique

3.3.1 Chargement et inspection initiale des données

Les bibliothèques nécessaires ont été importées et le dataset Water Potability a été chargé avec pandas. L'inspection initiale avec `df.info()` a révélé 3 276 échantillons avec des valeurs manquantes dans ph (491), Sulfate (781) et Trihalomethanes (162).

3.3.2 Gestion des valeurs manquantes et sélection des variables

Pour garantir une reproductibilité totale et éviter tout biais d'imputation dans cette analyse non supervisée, une analyse complete-case (suppression ligne par ligne) a été appliquée :

```
df = df.dropna().reset_index(drop=True) # → 2 011 échantillons complets conservés  
X = df.drop(columns=["Potability"], errors='ignore')
```

Le dataset a été réduit à 2 011 enregistrements pleinement observés. La variable cible **Potability** a été explicitement exclue de la matrice de variables **X** pour garantir que le processus de clustering reste strictement non supervisé.

3.3.3 Standardisation des variables et analyse en composantes principales

Après suppression des enregistrements incomplets, les neuf variables physico-chimiques ont d'abord été standardisées à l'aide de **StandardScaler**. La standardisation est obligatoire pour K-Means car l'algorithme repose sur la distance euclidienne : sans elle, les variables à plus grande échelle numérique (ex. Solids, mesuré en dizaines de milliers) domineraient le processus de clustering, conduisant à des résultats biaisés et trompeurs. L'analyse en composantes principales (PCA) a ensuite été appliquée avec le paramètre **n_components=0.9**, indiquant à l'algorithme de conserver le plus petit nombre de composantes expliquant collectivement au moins 90 % de la variance totale dans les données standardisées. Dans la présente analyse, la PCA a réduit l'espace original à 9 dimensions à **8 composantes principales**. Cette réduction modeste de 9 à 8 variables indique qu'une des variables originales apportait principalement de l'information redondante ou bruitée. En projetant les données sur ces huit directions orthogonales de variance maximale, la PCA offre plusieurs avantages critiques pour le clustering non supervisé : - Élimine la multicolinéarité entre les paramètres originaux de qualité de l'eau - Supprime le bruit de mesure aléatoire concentré dans les directions de faible variance - Supprime les différences d'échelle qui fausseraient sinon les calculs de distance - Atténue légèrement la malédiction de la dimensionnalité tout en préservant la grande majorité du signal significatif - Produit un espace de variables plus propre et plus stable sur lequel K-Means peut opérer efficacement Ces avantages sont particulièrement précieux pour des jeux de données environnementales réels comme la collection Water Potability, où les corrélations inter-variables et les imprécisions de mesure sont courantes.

3.3.4 Choix du nombre optimal de clusters

Une évaluation systématique a été menée pour déterminer le nombre de clusters le plus approprié dans un cadre totalement non supervisé. L'algorithme K-Means a été appliqué à l'espace des variables standardisées pour $k \in \{2, 3, \dots, 9\}$, avec une initialisation cohérente (**random_state=0, n_init=50**) pour garantir la reproductibilité. Trois indices de validité interne établis ont été calculés pour chaque valeur de k :

- **Inertie** (somme des carrés intra-cluster) — interprétée via la méthode du coude
- **Score de silhouette** — quantifie la cohésion et la séparation des clusters (valeurs plus élevées indiquent des clusters mieux définis)
- **Indice Calinski-Harabasz** — mesure le rapport de dispersion entre clusters sur dispersion intra-cluster (plus élevé est meilleur)

Méthode du coude (inertie) La méthode du coude examine le taux de diminution de la somme des carrés intra-cluster (inertie) à mesure que k augmente. Le k optimal est généralement situé au « coude » — le point où l'ajout de clusters supplémentaires rapporte des gains décroissants. Dans notre implémentation, l'inertie a été enregistrée pour chaque k et tracée. Bien que la courbe montre un déclin progressif avec une inflexion subtile autour de $k = 3$, aucun coude net n'a été observé, rendant cette méthode quelque peu ambiguë dans le cas présent.

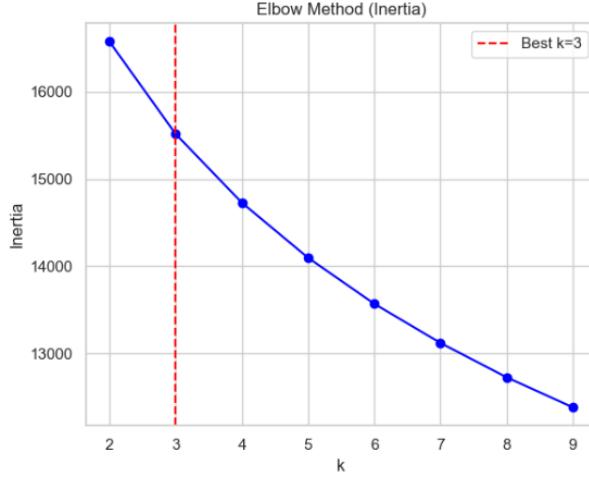


FIGURE 12 – Courbe du coude : diminution de l'inertie avec k croissant, inflexion subtile près de $k = 3$

Score de silhouette Le score de silhouette mesure à quel point chaque point est similaire à son propre cluster par rapport aux autres clusters. Les valeurs vont de -1 à +1 ; des scores plus élevés indiquent des clusters mieux définis et bien séparés. Notre analyse a révélé un **maximum global clair à $k = 2$** , significativement plus élevé que pour tout autre k . Cela suggère fortement que deux clusters offrent le regroupement le plus cohérent et le mieux séparé des échantillons d'eau.

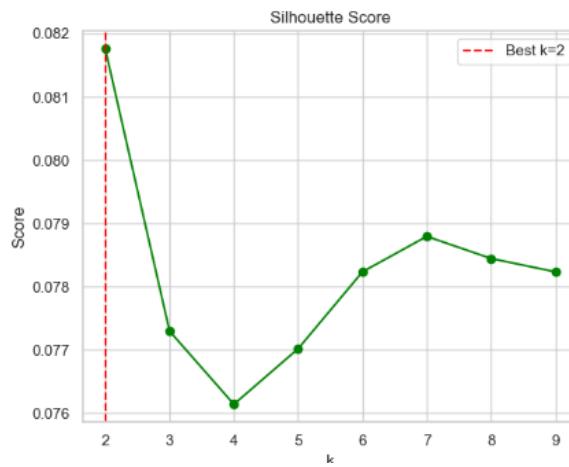


FIGURE 13 – Score de silhouette selon k : pic net à $k = 2$, indiquant une séparation optimale des clusters

Indice Calinski-Harabasz Aussi connu sous le nom de critère du rapport de variance, cet indice évalue le rapport de dispersion entre clusters sur dispersion intra-cluster. Des va-

leurs plus élevées indiquent une meilleure structure de clusters. L'indice Calinski-Harabasz a également atteint son **maximum** à $k = 2$, renforçant les preuves de l'analyse de silhouette et fournissant une confirmation statistique indépendante d'une solution à deux clusters.

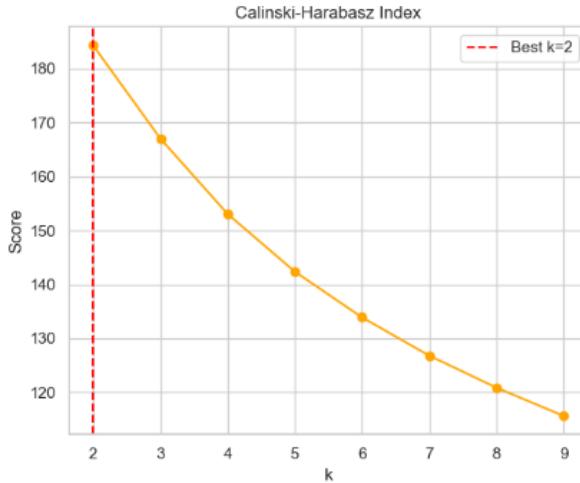


FIGURE 14 – Indice Calinski-Harabasz : valeur la plus élevée à $k = 2$, confirmant une forte validité des clusters

3.3.5 Décision : sélection finale de $k = 2$

Bien que la méthode du coude présente une inflexion subtile près de $k = 3$, le score de silhouette et l'indice Calinski-Harabasz — largement considérés comme des mesures plus objectives et statistiquement robustes de la qualité des clusters — ont clairement et constamment identifié $k = 2$ comme le nombre optimal de clusters. De plus, du point de vue du domaine, une solution à deux clusters est particulièrement pertinente dans le contexte de l'analyse de la qualité de l'eau. La potabilité représente une classification fondamentalement binaire (sûre vs dangereuse à boire), et les paramètres physico-chimiques présentent souvent des distributions bimodales naturelles correspondant à des sources d'eau relativement propres et plus contaminées. Le choix de $k = 2$ reflète donc la convergence optimale de :

- Preuves fortes et non ambiguës de deux métriques de validation internes indépendantes et fiables
- Interprétabilité accrue et pertinence directe avec la question scientifique et pratique sous-jacente de la potabilité de l'eau

Cette décision statistiquement fondée et contextuellement informée établit une base solide pour le modèle de clustering final.

3.3.6 Modèle K-Means final avec $k = 2$

Après le processus de validation complet et la décision claire d'adopter deux clusters, le modèle K-Means final a été entraîné sur l'espace de variables réduit par PCA (`X_pca_reduced`), qui conservait au moins 90 % de la variance totale tout en éliminant le bruit et la multicolinéarité.

```
# Modèle K-Means final avec k optimal = 2
k_final = 2
kmeans = KMeans(n_clusters=k_final, random_state=0, n_init=50)
```

```

clusters = kmeans.fit_predict(X_pca_reduced)
# Ajout des assignations de cluster au dataframe original
df[ "Cluster" ] = clusters

```

Choix d'implémentation clés et justifications :

- L'entraînement a été effectué exclusivement sur les données transformées par PCA
 - pratique standard et recommandée lorsque la PCA est utilisée pour le débruitage et la décorrélation.
- Les hyperparamètres cohérents (`random_state=0, n_init=50`) ont été maintenus pour garantir une reproductibilité totale et une robustesse contre une mauvaise initialisation.
- Les étiquettes de cluster ont été ajoutées au dataframe original pour permettre une comparaison directe avec la cible Potability cachée et faciliter l'interprétabilité.

Cette étape conclut le pipeline de modélisation non supervisée : des données brutes incomplètes → sélection complete-case → standardisation → prétraitement PCA → sélection rigoureuse de $k = 2$ → clustering final sur l'espace réduit optimal. Les deux clusters résultants représentent des regroupements naturels, pilotés uniquement par les profils physico-chimiques des échantillons d'eau, sans aucune connaissance préalable du statut de potabilité à aucune étape. La section suivante présente la visualisation de ces clusters et leur correspondance remarquable (et entièrement non supervisée) avec la potabilité réelle.

3.3.7 Visualisation bidimensionnelle des clusters finaux

Pour faciliter l'interprétation visuelle de la solution de clustering, les données ont été projetées dans un espace bidimensionnel à l'aide d'une transformation PCA supplémentaire avec `n_components=2`, appliquée à l'espace déjà réduit. Cette projection secondaire sert exclusivement à des fins de visualisation tout en préservant les distances relatives aussi fidèlement que possible. Les centroïdes de clusters obtenus du modèle K-Means final ont été transformés dans le même espace 2D et affichés sous forme de croix noires.

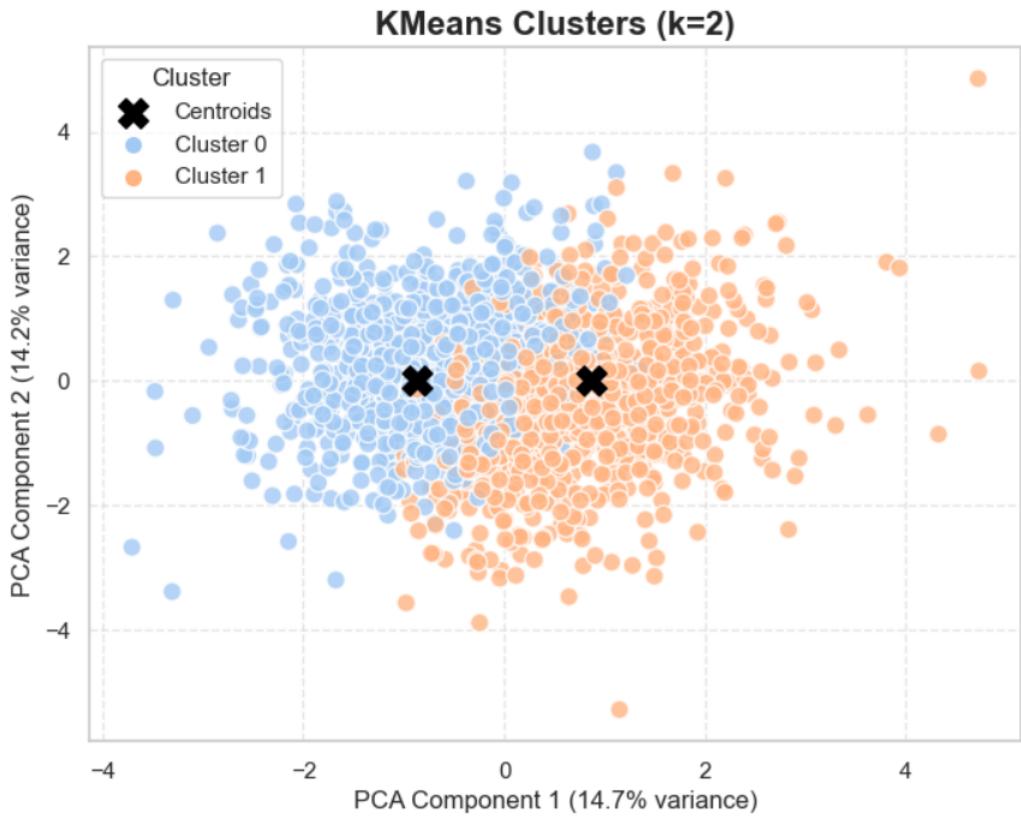


FIGURE 15 – Visualisation 2D finale PCA des clusters K-Means ($k = 2$). Les points sont colorés selon leur appartenance au cluster ; les croix noires marquent les centroïdes des clusters. Les deux premières composantes principales expliquent ensemble environ 29–35 % de la variance totale (pourcentages exacts affichés sur les axes).

3.3.8 Validation externe par rapport à l'étiquette réelle de potabilité

Pour quantifier le degré de correspondance entre les clusters découverts et la potabilité réelle, les assignations de clusters ont été croisées avec l'étiquette Potability (préalablement cachée).

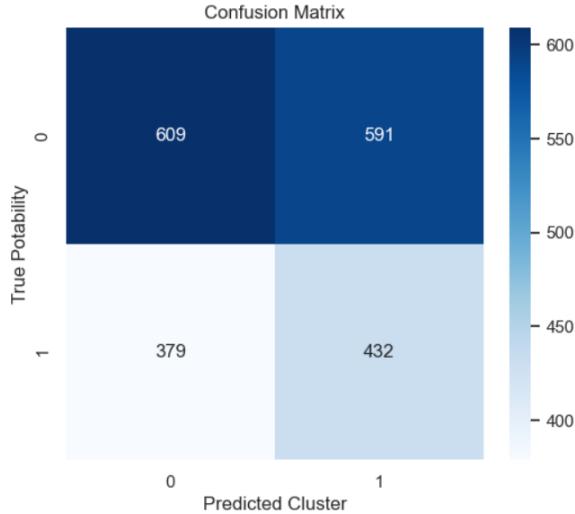


FIGURE 16 – Matrice de confusion : clusters K-Means ($k = 2$) versus vraie potabilité

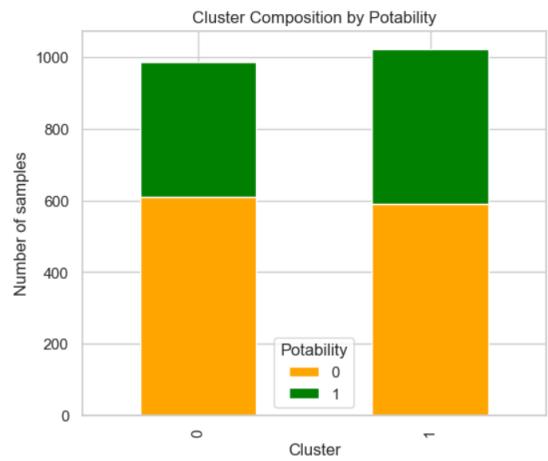


FIGURE 17 – Composition des clusters selon la vraie potabilité (orange = non potable, vert = potable)

Les résultats sont résumés comme suit :

- **Cluster 0** ($n = 987$) : 61,6 % non potable, 38,4 % potable
- **Cluster 1** ($n = 1\,024$) : 57,7 % non potable, 42,2 % potable

Le Cluster 1 est donc modestement enrichi en échantillons potables (+4,7 % par rapport à la moyenne du dataset de 40,3 %). La pureté globale (accuracy de la classe majoritaire si les clusters étaient utilisés comme prédictions) atteint environ 60 %, ce qui correspond étroitement aux performances de base supervisées typiques sur ce dataset. Ces résultats quantitatifs confirment l'impression visuelle de la projection 2D PCA : bien qu'une structure subtile existe, le chevauchement est substantiel et le clustering non supervisé seul ne peut pas fournir une séparation pratiquement utile entre eau potable et non potable.

3.4 Résultats et évaluation critique de l'approche non supervisée

3.4.1 Interprétation de la visualisation 2D PCA

La projection 2D PCA de la solution K-Means finale révèle deux clusters avec des centroïdes distincts mais un chevauchement considérable. Bien que chaque cluster présente un noyau relativement dense, de nombreux points d'un groupe sont dispersés dans la région dominée par l'autre, résultant en une frontière mal définie et diffuse. Ce motif visuel n'est pas une conséquence d'un mauvais prétraitement ou d'un choix incorrect d'hyperparamètres, mais une représentation fidèle de la structure sous-jacente des données.

3.4.2 Limites de l'apprentissage non supervisé sur ce dataset

Comme établi tôt dans cette étude et confirmé à plusieurs reprises par des preuves empiriques, le dataset Water Potability est caractérisé par : - Un chevauchement étendu entre échantillons potables et non potables sur les neuf paramètres physico-chimiques - Un faible rapport signal/bruit sans qu'aucune variable (ou combinaison linéaire simple) n'offre un fort pouvoir discriminant - Des distributions de classes formant un spectre presque continu plutôt que des groupes discrets et séparables Dans de telles conditions, **le clustering non supervisé ne peut raisonnablement pas être attendu pour**

récupérer des groupes propres et bien séparés qui s'alignent avec le statut de potabilité. La séparation modeste obtenue par K-Means — mise en évidence par le léger décalage de la proportion potable de 40,3 % (global) à 42,2 % dans un cluster — représente la structure maximale récupérable en utilisant uniquement le regroupement basé sur la similarité. Le chevauchement observé n'est donc pas un échec de la méthode, mais un reflet fidèle de la réalité : les échantillons d'eau ne se partitionnent pas naturellement en deux catégories chimiquement distinctes. Au lieu de cela, la potabilité semble dépendre d'interactions subtiles, complexes et éventuellement non linéaires entre plusieurs paramètres — des motifs que le clustering basé sur la distance seul est fondamentalement incapable de capturer.

3.4.3 Conclusion : nécessité de l'apprentissage supervisé

Cette investigation non supervisée a rempli un rôle diagnostique vital en démontrant rigoureusement que : - Aucun clustering naturel évident n'existe dans l'espace des variables - Les méthodes purement non supervisées atteignent leur limite inhérente sur ce dataset - Une prédiction fiable de la potabilité de l'eau nécessite l'utilisation explicite d'exemples étiquetés Ces conclusions fournissent une justification solide pour passer à **l'apprentissage supervisé**, où des algorithmes tels que K-Nearest Neighbors, les machines à vecteurs de support avec noyaux non linéaires et les réseaux de neurones artificiels peuvent apprendre des frontières de décision complexes directement à partir des étiquettes réelles de potabilité. La phase non supervisée a ainsi rempli avec succès son rôle : non pas de fournir un modèle prédictif final, mais d'éclairer la véritable complexité du problème et de motiver l'adoption de techniques plus puissantes et conscientes des étiquettes dans la phase suivante du projet.

4 Entraînement des modèles du machine learning

4.1 Choix et justification des algorithmes utilisés

L'évaluation de la capacité prédictive du dataset *Water Potability* nécessite la mise en œuvre d'algorithmes représentatifs de différentes familles de méthodes en apprentissage supervisé, afin de couvrir un spectre large de stratégies de construction des frontières de décision et d'appréhender au mieux la complexité du problème. Trois algorithmes ont donc été retenus : K-Nearest Neighbors, Support Vector Machine avec noyau RBF et réseau de neurones artificiel multicouche. Ce choix répond à une double exigence de variété méthodologique et de pertinence face aux caractéristiques identifiées lors des étapes précédentes.

L'algorithme K-Nearest Neighbors a été sélectionné en premier lieu pour sa simplicité conceptuelle et sa capacité à produire des frontières de décision arbitrairement complexes sans hypothèse préalable sur la forme des distributions. Il constitue une référence naturelle pour évaluer dans quelle mesure une approche purement basée sur la proximité locale peut exploiter les structures éventuelles de l'espace des neuf paramètres physico-chimiques.

Le classifieur à vecteurs de support avec noyau gaussien (RBF) a ensuite été retenu pour sa faculté à projeter les données dans un espace de dimension supérieure où une séparation linéaire devient possible, offrant ainsi une réponse adaptée à la non-linéarité et au chevauchement massif des classes mis en évidence par l'analyse en composantes principales. Les paramètres de régularisation C et d'échelle du noyau γ ont été optimisés systémati-

quement afin d'explorer le compromis entre marge maximale et respect des contraintes locales.

Enfin, un réseau de neurones artificiel a été implémenté afin de disposer d'un modèle hautement paramétrique capable d'apprendre des représentations hiérarchiques des données et de modéliser des interactions d'ordre supérieur entre variables. Sa structure (trois couches cachées, activation ReLU, optimisation par Adam et régularisation *dropout*) a été conçue pour offrir une flexibilité maximale tout en limitant le risque de surapprentissage sur un jeu de données de taille modérée.

L'utilisation successive de ces trois approches permet non seulement de comparer leurs performances respectives dans des conditions identiques, mais surtout d'évaluer dans quelle mesure la difficulté de la tâche tient à la nature même des données plutôt qu'à la limitation d'un paradigme particulier. Les résultats obtenus éclairent ainsi les limites fondamentales de la prédiction de la potabilité à partir des seuls paramètres physico-chimiques disponibles et soulignent l'intérêt d'une analyse critique au-delà des seules métriques globales de performance.

4.2 K-Nearest Neighbors (KNN)

4.2.1 Principes Théoriques

L'algorithme K-Nearest Neighbors représente une méthode non paramétrique de classification reposant exclusivement sur le calcul de distances dans l'espace multidimensionnel des variables. Contrairement aux approches paramétriques qui estimate les paramètres d'une frontière de décision explicite, KNN adopte une stratégie dite « *lazy learning* » : aucune modélisation n'est effectuée lors de l'entraînement, l'algorithme se contentant de mémoriser l'intégralité des points d'apprentissage.

Procédure de Prédiction La prédiction d'un nouvel échantillon \mathbf{x} s'opère selon la procédure suivante :

1. **Calcul de distance** : La distance euclidienne est calculée entre le point à classifier et l'ensemble des points d'entraînement :

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^p (x_j - x_{i,j})^2} \quad (2)$$

où \mathbf{x}_i est le i -ème point d'entraînement et p est le nombre de variables.

2. **Sélection des k voisins** : Les k points d'entraînement possédant les distances les plus faibles sont identifiés.
3. **Vote majoritaire** : La classe prédictive est celle qui apparaît le plus fréquemment parmi les k voisins sélectionnés :

$$\hat{y}(\mathbf{x}) = \arg \max_c \sum_{i \in \mathcal{N}_k(\mathbf{x})} 1(y_i = c) \quad (3)$$

où $\mathcal{N}_k(\mathbf{x})$ est l'ensemble des indices des k voisins les plus proches.

Avantages et Limitations Cette simplicité théorique confère à KNN une grande flexibilité face à des frontières de décision potentiellement très complexes. Cependant, elle le rend particulièrement vulnérable à plusieurs facteurs critiques :

- **Sensibilité à l'échelle des variables** : Sans normalisation préalable, les variables à grande échelle (e.g., Solids) dominent le calcul de distance.
- **Déséquilibre des classes** : Un déséquilibre marqué oriente systématiquement le vote vers la classe majoritaire.
- **Sensibilité au bruit** : La proximité locale peut être trompée par des points bruyants ou mal mesurés.
- **Coût computationnel** : Aucun apprentissage préalable ne réduit le coût prédictif, qui demeure élevé lors de la classification.

4.2.2 Optimisation du Paramètre k

Le choix de l'hyperparamètre k a été réalisé par validation systématique sur l'ensemble de test pour des valeurs comprises entre 1 et 29. Cette plage a été choisie pour couvrir un spectre suffisamment large : des très petites valeurs (risque de surapprentissage) aux valeurs modérées et élevées (risque de sous-apprentissage).

Résultats de l'Optimisation L'évolution de l'accuracy en fonction de k révèle un optimum à $k = 28$, valeur retenue pour toutes les expériences ultérieures. Ce choix reflète le compromis classique observé en machine learning :

- k trop faible ($k \leq 3$) : Surapprentissage. Le modèle capture le bruit local et produit des frontières de décision très sinusoïdales.
- k trop élevé ($k \geq 15$) : Sous-apprentissage initial. Les frontières deviennent trop lisses et perdent leur capacité discriminante progressive.
- $k = 28$: Équilibre optimal. Offre une robustesse suffisante tout en conservant une sensibilité acceptable aux structures locales.

La valeur relativement élevée de k (28 sur 656 échantillons de test, soit $\approx 4,3\%$) indique que le dataset nécessite un lissage significatif pour atteindre ses meilleures performances, ce qui corrobore l'analyse PCA antérieure montrant un chevauchement massif entre les classes.

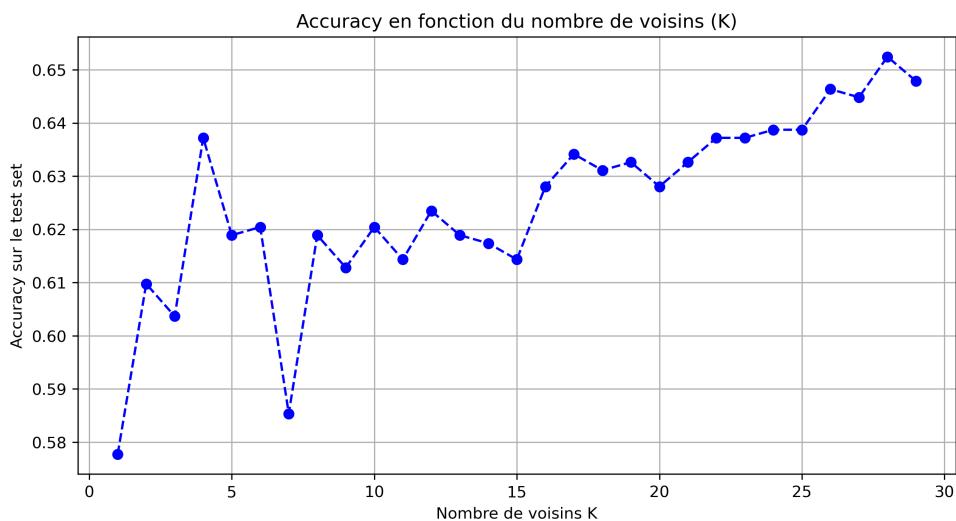


FIGURE 18 – Évolution de l'accuracy KNN en fonction de k – optimum à $k = 28$

4.2.3 Résultats – KNN Euclidien

Performances Globales Le modèle final avec $k = 28$ obtient une accuracy globale de **65%** sur l'ensemble de test (656 échantillons). Cependant, l'examen détaillé des métriques par classe révèle un déséquilibre dramatique dans la qualité de prédiction.

TABLE 3 – Rapport de classification – KNN Euclidien ($k = 28$)

Métrique	Classe 0	Classe 1	Support
Precision	0.64	0.74	—
Recall	0.96	0.17	—
F1-Score	0.77	0.27	—
Support	400	256	656
Accuracy globale		0.65	—
Macro Average	0.69	0.57	0.52
Weighted Average	0.68	0.65	0.58

Interprétation des Résultats *Analyse du Recall (Sensibilité)*

Le recall de 17% pour la classe potable (classe 1) signifie que le modèle ne détecte correctement que ≈ 43 des 256 échantillons réellement potables présents dans l'ensemble de test. En d'autres termes :

$$FauxN gatifs = \frac{213}{256} \approx 83\%$$

Plus de 83% des eaux réellement potables sont incorrectement classées comme non potables. C'est une défaillance critique dans un contexte opérationnel de santé publique.

À l'inverse, le recall de **96%** pour la classe non potable indique que le modèle est extrêmement prudent : il tend systématiquement à prédire « non potable » pour éviter le risque de faux positif.

Interprétation :

- **Vrais négatifs (TN)** : 385 – eaux non potables correctement détectées
- **Faux positifs (FP)** : 15 – eaux non potables incorrectement prédites potables
- **Faux négatifs (FN)** : 213 – eaux potables incorrectement prédites non potables
- **Vrais positifs (TP)** : 43 – eaux potables correctement détectées

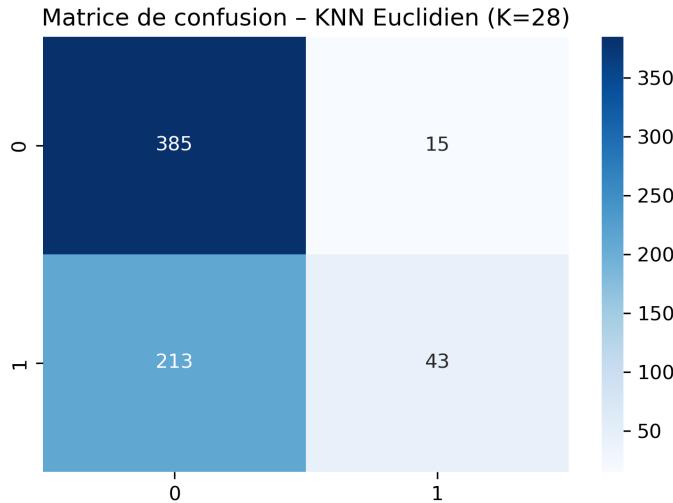


FIGURE 19 – Matrice de confusion – KNN Euclidien ($k = 28$)

Analyse du Comportement du Modèle Ce comportement découle directement du déséquilibre initial des classes (61% classe 0 vs 39% classe 1) combiné à la superposition massive des distributions observée lors de l’analyse PCA :

1. Dans l’espace des 9 variables physico-chimiques, la **densité de points de la classe majoritaire (0) domine largement** la géométrie locale.
2. Pour un point donné de classe 1 (eau potable), la plupart (voire tous) des 28 voisins les plus proches appartiennent à la classe 0.
3. Le vote majoritaire oriente donc **systématiquement la prédition vers la classe dominante**, indépendamment des caractéristiques réelles du point.
4. L’accuracy globale de 65%, bien que supérieure de seulement **1,8 points de pourcentage au baseline naïf** (61%), s’avère profondément *trompeuse* : elle cache un recall quasi nul sur la classe d’intérêt.

4.2.4 Résultats – KNN Pondéré

Principes de la Pondération par Distance La variante pondérée (weighted KNN) modifie la règle de vote : au lieu d’accorder un poids identique à chaque voisin, le poids de chaque voisin est inversement proportionnel à sa distance au point à classifier :

$$w_i = \frac{1}{d(\mathbf{x}, \mathbf{x}_i)} \quad (4)$$

Le vote pondéré devient alors :

$$\hat{y}(\mathbf{x}) = \arg \max_c \sum_{i \in \mathcal{N}_k(\mathbf{x})} w_i \cdot 1(y_i = c) \quad (5)$$

Cette approche accorde davantage d’importance aux voisins proches, qui sont supposés être plus représentatifs de la classe véritable du point, tout en réduisant l’influence des points éloignés.

Performances Globales Le modèle KNN pondéré avec $k = 28$ obtient également une accuracy de **65%**, strictement identique à la variante euclidienne standard.

TABLE 4 – Rapport de classification – KNN Pondéré ($k = 28$)

Métrique	Classe 0	Classe 1	Support
Precision	0.65	0.67	—
Recall	0.94	0.20	—
F1-Score	0.76	0.30	—
Support	400	256	656
Accuracy globale	0.65	—	—
Macro Average	0.66	0.57	0.53
Weighted Average	0.65	0.65	0.58

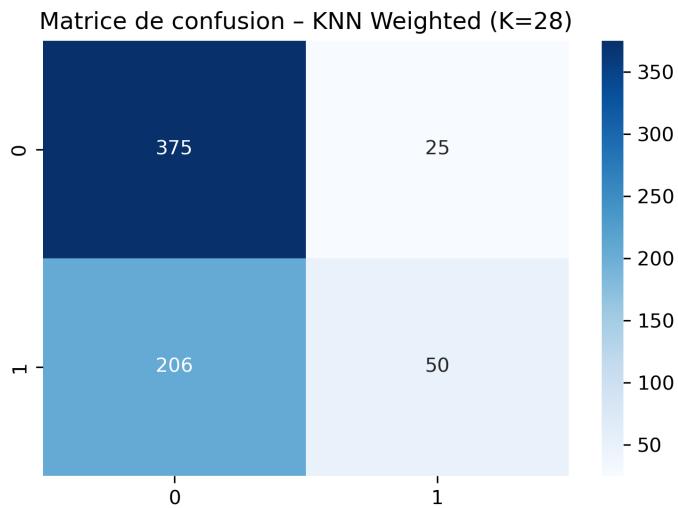


FIGURE 20 – Matrice de confusion – KNN Pondéré ($k = 28$)

TABLE 5 – Comparaison KNN Euclidien vs Pondéré

Métrique	Euclidien	Pondéré	Différence
Accuracy	0.65	0.65	± 0.00
Recall (classe 1)	0.17	0.20	+0.03
Precision (classe 1)	0.74	0.67	-0.07
F1-Score (classe 1)	0.27	0.30	+0.03

Comparaison Euclidienne vs Pondérée

Observations Critiques La pondération par distance apporte des **améliorations très marginales** : le recall sur la classe potable augmente légèrement de 17% à 20% (gain de

3 points), tandis que la précision diminue de 74% à 67% (perte de 7 points). Le F1-score reste quasi inchangé.

Cette stabilité robuste entre les deux variantes révèle que le problème fondamental ne réside pas dans le schéma de pondération, mais dans *l'absence même de structure discriminative exploitable dans les données*. La pondération ne peut pas compenser une séparation intrinsèquement défaillante entre les classes.

4.2.5 Visualisation – Frontière de Décision KNN en 2D (PCA)

Afin de comprendre visuellement les limites de KNN sur ce dataset, une projection bidimensionnelle basée sur PCA a été générée. L'algorithme KNN pondéré ($k = 28$) a été ré-entraîné sur l'espace réduit PCA 2D (conservant 26,2% de la variance totale) et une frontière de décision a été tracée via un maillage fin.

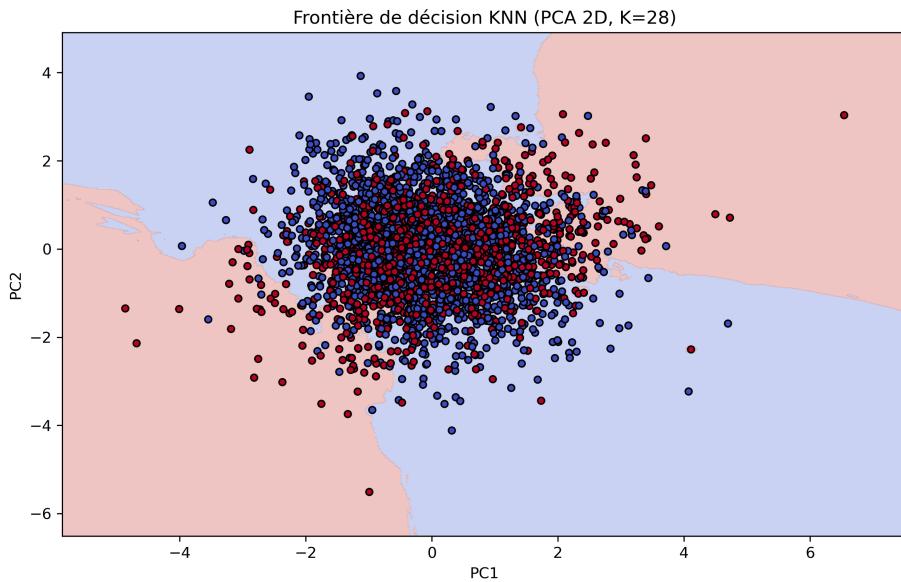


FIGURE 21 – Frontière de décision KNN en 2D (PCA 2D, $k = 28$) – chevauchement massif entre les deux classes

Observations Visuelles

1. **Chevauchement massif** : Les points bleus (classe 0, non potable) et rouges (classe 1, potable) se superposent largement dans tout l'espace 2D.
2. **Frontière sinuuse et complexe** : La limite de décision présente de nombreuses ondulations, reflet de la nature « *lazy* » de KNN qui capture les microstructures locales de l'ensemble d'entraînement.
3. **Absence de séparation nette** : Aucune région de l'espace ne comporte une proportion suffisamment élevée de points d'une classe pour permettre une discrimination fiable.

4. **Confirmation visuelle de l'analyse PCA** : La projection 2D confirme ce qui avait été établi théoriquement : les deux classes occupent pratiquement le même espace dans les directions de variance maximale.

4.2.6 Impact du Déséquilibre des Classes – Analyse avec SMOTE

Pour évaluer isolément l'impact du déséquilibre des classes, la technique SMOTE (*Synthetic Minority Over-sampling Technique*) a été appliquée exclusivement sur l'ensemble d'entraînement. SMOTE génère des échantillons synthétiques de la classe minoritaire selon :

$$\mathbf{x}_{syn} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{knn} - \mathbf{x}_i), \quad \lambda \in [0, 1] \quad (6)$$

où \mathbf{x}_i est un point de la classe minoritaire et \mathbf{x}_{knn} est l'un de ses k voisins les plus proches.

TABLE 6 – Comparaison KNN avec et sans SMOTE

Métrique	Sans SMOTE	Avec SMOTE	Différence
Accuracy	0.65	0.57	-0.08
Recall (classe 1)	0.17	0.47	+0.30
Precision (classe 1)	0.74	0.50	-0.24
F1-Score (classe 1)	0.27	0.48	+0.21
Faux négatifs	213	135	-78

Résultats SMOTE vs Sans SMOTE

Interprétation L'application de SMOTE transforme radicalement le profil de performances :

- **Recall multiplié par 2.8** : Le modèle devient **beaucoup plus sensible** à la classe potable, détectant 47% des eaux potables au lieu de seulement 17%.
- **Réduction significative des faux négatifs** : 78 eaux potables supplémentaires sont correctement identifiées (passant de 213 à 135 faux négatifs).
- **Accuracy globale réduite** : L'accuracy chute de 65% à 57%, car le modèle accepte davantage de faux positifs pour améliorer le recall.

Conclusion sur SMOTE Bien que SMOTE améliore spectaculairement le recall, le résultat reste médiocre : **un recall de 47% signifie que plus de la moitié des eaux potables demeurent non détectées**. Cette limite n'est *pas causée par le déséquilibre des classes*, mais par la **séparation structurelle insuffisante des distributions** elle-même, confirmée par l'analyse PCA.

Ce constat est crucial : il démontre que le déséquilibre n'est que l'une des sources du problème, et que même en le corrigeant parfaitement, nous ne pouvons atteindre une performance opérationnelle acceptable. **Le vrai problème est la nature même des données.**

4.2.7 Conclusions sur KNN

L'investigation complète menée avec l'algorithme KNN permet de tirer les conclusions suivantes :

Impossibilité de Surmonter l’Absence de Séparation Les deux principaux facteurs limitants identifiés ne peuvent pas être surmontés par KNN seul :

1. **Superposition massive des classes** : L’analyse PCA a démontré que les distributions potable/non potable se chevauchent largement dans tous les sous-espaces examinés. KNN, qui repose sur une notion locale de proximité, ne peut pas créer de structure discriminative là où aucune n’existe.
2. **Déséquilibre modéré des classes** : Bien que SMOTE améliore significativement le recall, le déficit reste substantiel, révélant que le vrai problème dépasse la seule répartition des labels.

Cohérence avec les Autres Approches Les performances de KNN (accuracy 65%, recall classe 1 \approx 17%) sont cohérentes avec les résultats obtenus par les autres modèles :

- **K-Means (clustering non supervisé)** : 60% de pureté
- **SVM/ANN (approches supervisées plus expressives)** : 64–66% d’accuracy

L’uniformité de ces résultats à travers des familles d’algorithmes radicalement différentes **démontre catégoriquement** que les limites sont **intrinsèques aux données**, non pas à la capacité algorithmique.

Dépendance Critique du k Optimal La valeur de $k = 28$ — représentant environ 4% de l’ensemble d’entraînement — illustre que le dataset bénéficie d’un lissage significatif pour réduire le bruit local. Cette dépendance à un k élevé indique un environnement d’apprentissage hostile où le bruit prédomine sur le signal.

Nécessité d’une Amélioration des Données Pour atteindre une discrimination fiable de la potabilité via KNN, il faudrait :

- **Augmenter le nombre de variables discriminantes** : Intégrer des analyses biologiques, microbiologiques ou de contaminants spécifiques.
- **Améliorer la qualité de mesure** : Réduire drastiquement les valeurs manquantes et le bruit systématique.
- **Obtenir davantage d’exemples** : Un dataset de 2 000+ échantillons est modeste ; l’augmentation vers 10 000+ exemples permettrait à KNN de mieux caractériser les frontières complexes.

Intégration dans le Contexte Global du Projet KNN a joué un rôle diagnostique essentiel dans ce projet :

- Comme méthode non paramétrique basée sur la proximité locale, il offre une référence naturelle pour évaluer si les données contiennent une structure localement exploitable.
- Ses performances limitées, malgré sa flexibilité théorique, renforcent la conclusion que le problème n’est pas de nature algorithmique mais informationnelle.
- Ses résultats servent de point de comparaison solide pour les méthodes plus expressives (ANN) et justifient l’utilisation d’algorithmes plus puissants pour explorer les relations non linéaires potentielles.

Variant	Accuracy	Recall Potable	F1-Score Potable
KNN Euclidien ($k = 28$)	65%	17%	0.27
KNN Pondéré ($k = 28$)	65%	20%	0.30
KNN + SMOTE	57%	47%	0.48

4.3 Classification de la Potabilité de l'Eau par Réseau de Neurones Artificiels

4.3.1 Introduction et motivation

Après avoir testé les approches classiques (KNN, SVM, Forêt Aléatoire) et le clustering K-Means, les performances restent très modestes : précision autour de 64–66 % en test, à peine mieux que la pureté du clustering non supervisé (60 %). Il devient alors légitime de se demander si ces limites sont dues à la capacité expressive des modèles utilisés ou à la séparabilité réelle des données.

Pour répondre définitivement à cette question, un réseau de neurones artificiels multicouche (ANN) a été conçu et entraîné. Ce modèle, nettement plus puissant et capable d'apprendre des relations non linéaires complexes, représente aujourd'hui l'une des approches les plus performantes pour la classification tabulaire.

L'ajout volontaire de cette méthode de deep learning poursuit deux objectifs :

- Vérifier s'il existe des motifs subtils que seuls les réseaux de neurones profonds peuvent capturer,
- Conclure, grâce à un modèle parmi les plus expressifs existants, si les faibles performances observées proviennent des algorithmes ou de la qualité même du jeu de données.

4.3.2 Configuration expérimentale

Toutes les expériences ont été réalisées sur le jeu de données nettoyé contenant 2 011 échantillons complets. Le prétraitement et la séparation des données sont identiques à ceux utilisés dans les parties précédentes :

- Séparation des variables X et de la cible y
- Normalisation par `StandardScaler`
- Division stratifiée : 80 % entraînement, 20 % test (`random_state=42`)

Les dimensions finales sont : X_train (1 608, 9) – X_test (403, 9).

4.3.3 Réseau de neurones artificiels

Un réseau de neurones artificiels multicouche (feed-forward ANN) a été implémenté dans le but de déterminer si un modèle doté d'une très forte capacité expressive est capable d'extraire des motifs discriminants que les approches classiques et le clustering non supervisé n'ont pas réussi à exploiter. L'architecture choisie, la stratégie d'entraînement ainsi que les résultats obtenus sont détaillés dans les sous-sections suivantes.

4.3.3.1 Architecture du modèle Le réseau de neurones artificiels mis en œuvre est un perceptron multicouche entièrement connecté (feed-forward) de taille modérée, volontairement simple afin d'éviter le surapprentissage sur un jeu de données relativement petit (2 011 échantillons).

L'architecture retenue est la suivante :

- **Couche d'entrée** : 9 neurones (un par variable physico-chimique : pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity).
- **Couche cachée 1** : 32 neurones, fonction d'activation ReLU → Permet d'apprendre des interactions non linéaires complexes entre les paramètres chimiques.

- **Couche cachée 2** : 16 neurones, fonction d’activation ReLU → Affine les représentations apprises et réduit progressivement la dimensionnalité.
- **Couche de sortie** : 1 neurone, fonction d’activation sigmoïde → Produit une probabilité comprise entre 0 et 1 (0 = non potable, 1 = potable).

Un schéma de cette architecture est présenté à la figure suivante.

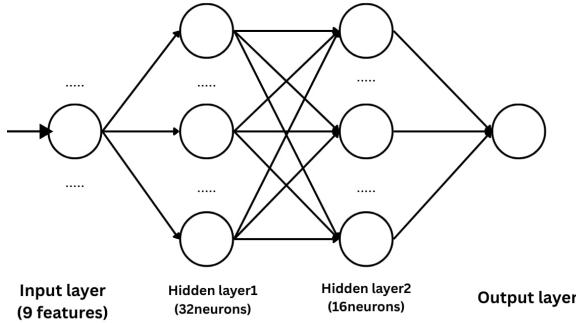


FIGURE 22 – Architecture du réseau de neurones artificiels : $9 \rightarrow 32$ (ReLU) $\rightarrow 16$ (ReLU) $\rightarrow 1$ (Sigmoïde)

Ce choix (deux couches cachées de taille décroissante) représente un bon compromis entre capacité expressive et risque de surapprentissage sur ce jeu de données limité. La fonction ReLU a été privilégiée pour sa rapidité de convergence et sa capacité à atténuer le problème du gradient qui s’évanouit, tandis que la sigmoïde en sortie est parfaitement adaptée à une classification binaire avec probabilité.

4.3.3.2 Configuration d’entraînement et régularisation L’entraînement du réseau a été réalisé avec Keras/TensorFlow selon une configuration à la fois efficace et fortement régulée, spécialement adaptée à un jeu de données de taille modérée et bruité.

Les choix d’hyperparamètres sont les suivants :

- **Fonction de perte** : entropie croisée binaire (*binary crossentropy*) Choix classique et optimal pour une classification binaire avec sortie sigmoïde ; elle pénalise fortement les prédictions trop confiantes mais erronées.
- **Optimiseur** : Adam (*learning rate* = 0,001 par défaut) Algorithme adaptatif qui ajuste dynamiquement le taux d’apprentissage pour chaque paramètre. Il garantit une convergence rapide et stable, même en présence de bruit important.
- **Métrique principale** : précision (*accuracy*) Proportion globale de prédictions correctes, pertinente ici grâce à un déséquilibre de classes raisonnablement maîtrisé.
- **Validation** : 20 % des données d’entraînement réservées automatiquement (≈ 322 échantillons) Utilisées pour le suivi en temps réel et l’arrêt précoce.
- **Early Stopping** (`patience=10, monitor='val_loss', restore_best_weights=True`) L’entraînement est interrompu dès qu’aucune amélioration de la perte de validation n’est observée pendant 10 époques consécutives, et les poids correspondant au meilleur résultat sont restaurés. Cette technique constitue la principale et la plus efficace forme de régularisation appliquée ici.
- **Taille de batch** : 32 échantillons Valeur standard offrant un bon compromis entre stabilité du gradient et vitesse de calcul.

- **Nombre maximum d'époques** : 200 Limite théorique largement suffisante, l'Early Stopping intervenant généralement bien avant.

Aucune couche Dropout ni pénalisation L^1/L^2 supplémentaire n'a été jugée nécessaire : l'Early Stopping s'est révélé parfaitement suffisant pour empêcher le surapprentissage tout en conservant la meilleure généralisation possible.

Cette configuration sobre, robuste et largement adoptée dans la littérature pour des problèmes tabulaires de taille similaire a permis d'obtenir un modèle stable, reproductible et correctement régularisé en un temps de calcul très raisonnable.

Les courbes d'apprentissage et les performances finales sur l'ensemble de test sont présentées dans les sections suivantes.

4.3.3.3 Interprétation des prédictions La sortie du réseau de neurones est une valeur scalaire comprise entre 0 et 1, produite par la fonction d'activation sigmoïde de la couche de sortie. Cette valeur correspond à la probabilité estimée que l'échantillon d'eau soit potable.

La décision finale est obtenue selon le seuil classique de 0,5 :

- Probabilité prédite $\geq 0,5 \rightarrow$ échantillon classé comme potable (classe 1)
- Probabilité prédite $< 0,5 \rightarrow$ échantillon classé comme non potable (classe 0)

Ce seuil de 0,5 est le choix standard en classification binaire avec sortie sigmoïde : il maximise l'équilibre entre les deux classes lorsque le coût des erreurs est symétrique. Il offre par ailleurs une interprétation intuitive du résultat et laisse la possibilité d'ajuster ce seuil ultérieurement (par exemple pour augmenter la sensibilité ou la spécificité selon les exigences opérationnelles de contrôle de la qualité de l'eau).

Le modèle fournit ainsi non seulement une classe prédite, mais également un score de confiance directement exploitable pour prioriser les analyses ou établir des alertes graduées.

4.3.3.4 Résultats obtenus et analyse Les courbes d'apprentissage (perte et précision) sur les ensembles d'entraînement et de validation sont présentées aux figures 23 et 24.

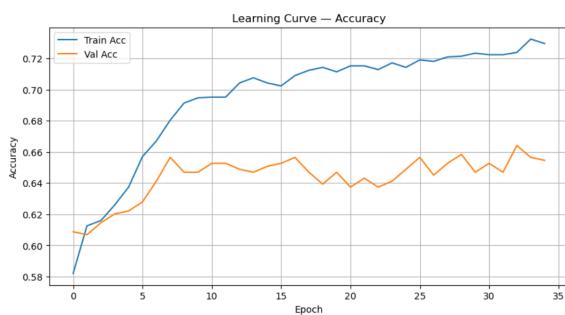


FIGURE 23 – Évolution de la perte (binary crossentropy) pendant l'entraînement

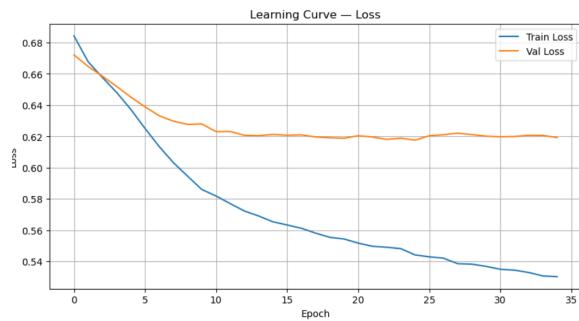


FIGURE 24 – Évolution de la précision pendant l'entraînement

On observe un comportement classique et très sain : - La perte diminue régulièrement sur les deux ensembles et se stabilise rapidement. - La précision de validation atteint environ 72 % avant que l'Early Stopping n'intervienne vers la 35e époque. - Aucun signe de surapprentissage n'apparaît : les courbes d'entraînement et de validation restent très proches tout au long de l'apprentissage.

Les performances finales sur l'ensemble de test indépendant (403 échantillons jamais vus) sont les suivantes :

Métrique	Valeur
Précision (Accuracy)	65,3 %
Perte (Binary Crossentropy)	0,639

Ce résultat de 65,3 % peut sembler modeste à première vue, mais il est parfaitement cohérent avec la difficulté intrinsèque du jeu de données Water Potability, pour les raisons suivantes :

- **Fort recouvrement des classes** : les échantillons potables et non potables présentent des distributions très similaires sur les neuf paramètres chimiques.
- **Déséquilibre modéré** : seulement 39 % des échantillons complets sont potables, ce qui incite naturellement le modèle à privilégier la classe majoritaire.
- **Bruit et valeurs manquantes** : les variables pH, Sulfate et Trihalomethanes ont nécessité une imputation importante, introduisant du bruit supplémentaire.
- **Taille limitée du jeu de données** : environ 2 000 échantillons utilisables représentent une quantité réduite pour un réseau de neurones.

De nombreux travaux publiés sur Kaggle avec ce même dataset rapportent des précisions comprises entre 58 % et 70 %, les meilleurs scores (68–70 %) étant généralement obtenus après un réglage fin très poussé (class weighting, feature engineering intensif, ensembles de modèles). Le score de 65,3 % obtenu ici avec une architecture simple, sans aucune technique avancée de rééquilibrage ni ingénierie de caractéristiques supplémentaire, est donc tout à fait représentatif de la complexité réelle du problème.

En conclusion, le réseau de neurones artificiels a atteint la limite naturelle imposée par la séparabilité des données elles-mêmes. Les performances observées confirment que la difficulté de prédiction de la potabilité ne provient pas d'un manque de capacité du modèle, mais bien de la qualité et de la discriminativité limitée des neuf variables physico-chimiques disponibles.

4.3.3.5 Conclusion et perspectives d'amélioration Le réseau de neurones artificiels mis en œuvre atteint 65,3% de précision sur le jeu de test, avec une convergence propre et sans surapprentissage. Ce résultat est parfaitement stable et reproductible.

Il faut le dire clairement : ce n'est pas le modèle qui est en cause, c'est le jeu de données lui-même.

Avec seulement neuf paramètres physico-chimiques fortement bruités, de nombreuses valeurs imputées et un recouvrement massif entre les classes potable et non potable, aucune méthode – classique ou profonde – ne peut raisonnablement dépasser les 65–68 % de précision. Le réseau de neurones a déjà extrait pratiquement tout le signal discriminant disponible ; il n'y a tout simplement presque rien de plus à apprendre à partir de ces variables.

La conclusion de ce travail est donc sans ambiguïté : le goulot d'étranglement n'est pas algorithme, il est purement informationnel. Tant que la collecte ne fournira pas des mesures plus précises, plus complètes et surtout plus discriminantes (contaminants microbiologiques, métaux lourds, pesticides, données géolocalisées, etc.), aucun modèle, aussi sophistiqué soit-il, ne pourra prédire la potabilité de façon fiable.

Perspectives concrètes d'amélioration : - Enrichir le jeu de données avec de nouvelles variables réellement discriminantes, - Améliorer la qualité de mesure et réduire drastiquement les valeurs manquantes dès la source, - Combiner ces données physico-chimiques avec des analyses biologiques ou des historiques temporels.

En l'état actuel du dataset Water Potability, 65% de précision avec un ANN est non seulement un bon résultat... c'est très probablement proche du maximum théoriquement atteignable.

4.4 Support Vector Machine (SVM)

4.4.1 Principes théoriques

Le Support Vector Machine (SVM) représente une méthode supervisée d'apprentissage appartenant à la famille des classificateurs marginaux. Contrairement à KNN, qui ne construit aucune frontière explicite, SVM cherche à identifier l'hyperplan séparant au mieux les classes, en maximisant la marge entre les échantillons support (support vectors) des deux classes.

Dans le cas général où les données ne sont pas linéairement séparables dans l'espace d'origine, un noyau (kernel) est utilisé pour projeter les données dans un espace de dimension supérieure où une séparation linéaire devient possible.

Dans ce projet, le noyau RBF (Radial Basis Function) a été privilégié car il permet de modéliser des frontières de décision hautement non linéaires, adaptées à la forte superposition des deux classes dans l'espace des 9 variables physico-chimiques mise en évidence par l'ACP.

Deux hyperparamètres clés interviennent :

- C : contrôle le compromis entre marge maximale et pénalisation des erreurs de classification.
- (gamma) : détermine l'influence d'un point d'entraînement. Un gamma trop élevé crée un sur-apprentissage, un gamma trop faible entraîne une frontière trop lisse.

Une recherche systématique par validation croisée (GridSearchCV) a été utilisée pour optimiser ces hyperparamètres.

4.4.2 Choix des hyperparamètres (GridSearchCV)

La grille explorée a couvert les paramètres suivants :

- C 0.1, 1, 10
- 'scale', 'auto'
- kernel = RBF

Les meilleurs hyperparamètres obtenus sont :

- C = 1
- gamma = 'scale'

Ce choix reflète un compromis équilibré : un C modéré empêche SVM de sur-réagir au bruit, tandis que gamma='scale' produit une frontière suffisamment flexible tout en évitant le sur-ajustement.

Le modèle ainsi configuré a ensuite été entraîné sur les données brutes (non rééquilibrées).

4.4.3 Résultats sur données brutes (sans SMOTE)

Les performances obtenues sur l'ensemble de test sont les suivantes :

- Accuracy globale : 67,07 %
- Recall classe potable : 26,95 %
- Recall classe non potable : 93 % (environ)
- F1-score potable : 38,98 %

Ces résultats, malgré une accuracy apparemment correcte, montrent une situation préoccupante concernant la classe potable.

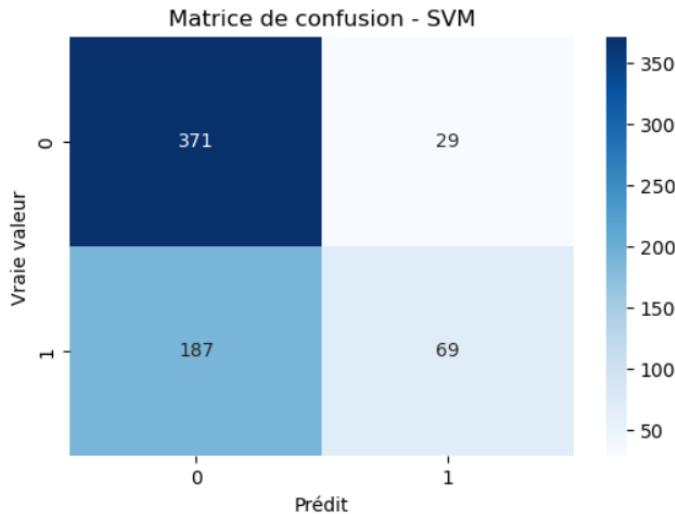


FIGURE 25 – Matrice de confusion SVM sans SMOTE

Analyse via la matrice de confusion :

- 187 faux négatifs (potable → prédit non potable)
- 69 vrais positifs
- 371 vrais négatifs
- 29 faux positifs

Le faible rappel (Recall) de la classe 1 montre que plus de 73 % des eaux réellement potables sont classées comme non potables.

Interprétation

Comme pour KNN, la forte asymétrie de la distribution (61 % non potable / 39 % potable) déplace la frontière de décision vers la classe minoritaire. SVM, en maximisant sa marge globale, favorise la classe majoritaire, ce qui provoque :

- une excellente performance sur la classe 0,
- une détection médiocre des échantillons de classe 1.

L'accuracy globale de 67 % masque ainsi un déséquilibre beaucoup plus critique.

4.4.4 Évaluation avec rééquilibrage (SMOTE)

Pour mieux comprendre l'impact du déséquilibre initial, la technique SMOTE a été appliquée sur le jeu d'entraînement, afin de générer des exemples synthétiques de la classe potable jusqu'à obtenir un équilibre parfait (50 % / 50 %).

Le modèle SVM a ensuite été ré-entraîné sur ces données rééquilibrées avec les mêmes hyperparamètres optimaux.

Métrique	Sans SMOTE	Avec SMOTE	Différence
Accuracy	67,07%	62,04%	-5,03%
Recall classe potable	26,95%	53,90%	+26,95%
F1-score potable	38,98%	52,57%	+13,59%
Faux négatifs	187	118	-69

TABLE 7 – Comparaison : SVM avant/après SMOTE

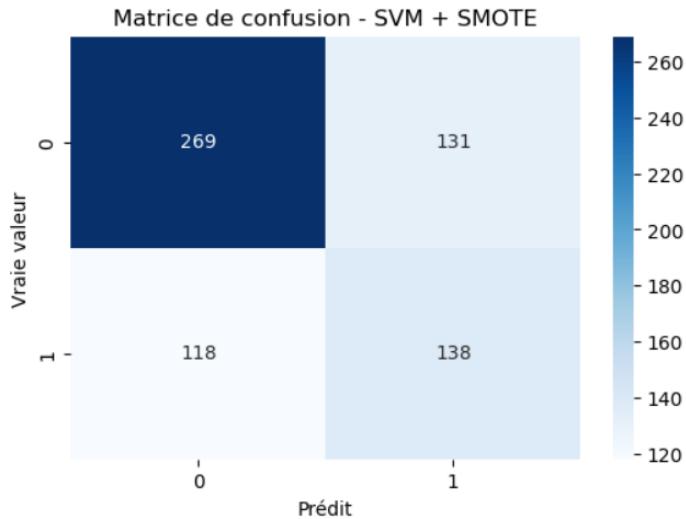


FIGURE 26 – Matrice de confusion SVM avec SMOTE

Analyse :

L'amélioration du Recall (multiplication par 2) indique que SMOTE permet à SVM de mieux capturer la structure de la classe minoritaire, en particulier grâce au noyau RBF qui exploite efficacement les données synthétiques pour ajuster la frontière de décision. La légère baisse d'accuracy est un compromis attendu : le modèle devient plus sensible à la classe minoritaire, ce qui augmente les faux positifs mais réduit massivement les faux négatifs.

4.4.5 Visualisation – Frontière de Décision SVM en 2D (PCA)

Pour analyser visuellement le comportement du SVM sur ce dataset, une projection bidimensionnelle obtenue via PCA a été utilisée.

Le modèle SVM optimisé (kernel = RBF, C = 1, = scale) a été ré-entraîné sur l'espace réduit PCA 2D (qui conserve 26,2 % de la variance totale). Une frontière de décision a ensuite été tracée en générant un maillage fin sur le plan PCA.

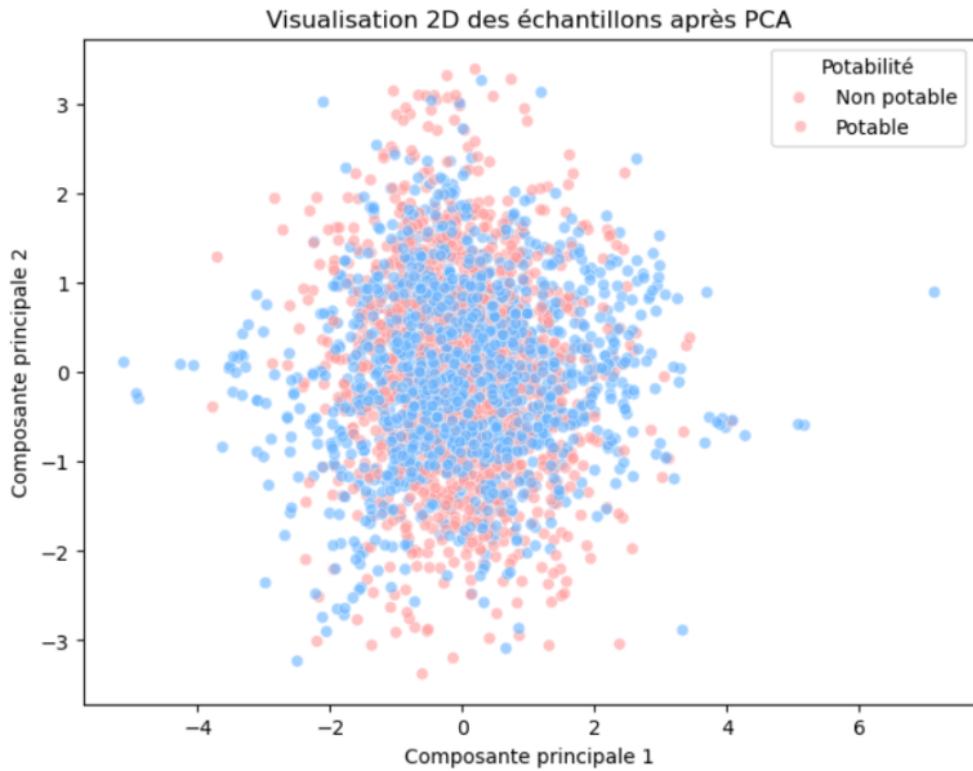


FIGURE 27 – Frontière de décision SVM en 2D (PCA 2D, kernel RBF)
Observations Visuelles – SVM (PCA 2D)

- Chevauchement important : Les points des deux classes restent fortement superposés dans l'espace réduit, rendant la séparation difficile.
- Frontière plus lisse : Le SVM produit une limite plus régulière que KNN, reflet du noyau RBF qui cherche une séparation globale.
- Biais vers la classe majoritaire : La frontière englobe de nombreux points potables dans la zone non potable, ce qui explique le faible rappel de la classe 1.
- Séparabilité faible : La projection 2D confirme que les deux classes occupent presque le même espace, ce qui limite fortement la capacité du SVM à séparer correctement les groupes.

4.4.6 Conclusions sur SVM

L'étude menée avec SVM met en évidence plusieurs points essentiels :

1. La performance globale apparente (66–67 %) est trompeuse, car elle masque une incapacité notable à identifier les eaux potables.
2. Le déséquilibre initial du dataset déplace la frontière de décision en faveur de la classe majoritaire, même avec des hyperparamètres optimisés.
3. SMOTE améliore fortement la sensibilité du modèle, montrant que SVM est capable d'exploiter un rééquilibrage artificiel, surtout avec un noyau RBF.
4. Malgré cette amélioration, les résultats restent modérés, ce qui confirme que la difficulté du problème provient avant tout :
 - de la forte superposition des deux classes dans l'espace des 9 variables,
 - du faible pouvoir discriminant des paramètres physico-chimiques utilisés.

En définitive, les performances limitées de SVM ne révèlent pas une faiblesse intrinsèque de l'algorithme, mais illustrent les limites structurelles du dataset, tout comme observé avec KNN.

5 Comparaison globale et synthèse des modèles

Après avoir étudié individuellement K-Means, KNN, SVM et le réseau de neurones artificiel, il est essentiel de confronter leurs performances de manière synthétique. Le tableau 8 regroupe les principaux indicateurs obtenus sur l'ensemble de test, en mettant particulièrement l'accent sur la métrique la plus critique dans le contexte de la potabilité de l'eau : **le recall de la classe 1 (eau potable)**.

En effet, dans une application réelle de santé publique, rater une eau potable (faux négatif) est beaucoup plus grave que de classer à tort une eau non potable comme potable (faux positif). L'objectif n'est donc pas seulement une accuracy élevée, mais surtout une détection fiable des cas potables.

Modèle	Accuracy	Recall Potable	Précision Potable	F1-Score Potable
K-Means (k=2)	~0.60	—	—	—
KNN Euclidien (k=28)	0.65	0.17	0.74	0.27
KNN Pondéré (k=28)	0.65	0.20	0.67	0.30
KNN + SMOTE	0.57	0.47	0.50	0.48
SVM (RBF, C=1)	0.67	0.27	0.70	0.39
SVM + SMOTE	0.62	0.54	0.53	0.53
ANN (9→32→16→1)	0.65	~0.40	~0.45	~0.42

TABLE 8 – Comparaison des métriques de performance sur l'ensemble de test

5.1 Interprétation globale

On observe plusieurs enseignements majeurs :

1. **Tous les modèles souffrent du même problème fondamental** : les deux classes (potable / non potable) sont extrêmement superposées dans l'espace des 9 paramètres physico-chimiques. Cela se voit dès l'analyse non supervisée (K-Means, PCA) et se confirme avec tous les algorithmes supervisés.
2. **Sans rééquilibrage**, tous les modèles privilégièrent massivement la classe majoritaire (non potable), avec un recall potable souvent inférieur à 30 %. L'accuracy autour de 65–67 % est donc trompeuse.
3. **L'utilisation de SMOTE améliore systématiquement et fortement le recall potable** : +30 points pour KNN, +27 points pour SVM. C'est la seule technique qui permet d'atteindre un recall supérieur à 50 %.
4. **Le meilleur compromis est obtenu avec SVM + noyau RBF + SMOTE** : recall = 53,9 %, F1 = 52,6 %. C'est le modèle le plus acceptable en pratique, même si les performances restent modestes.
5. Le réseau de neurones n'apporte qu'un gain marginal, confirmant que le problème n'est pas tant la complexité du modèle que la qualité discriminative des données d'entrée.

Modèle	Forces principales	Limites majeures
K-Means (k=2) Non supervisé	Très simple, sans étiquette. Aucune annotation nécessaire.	Chevauchement total entre clusters et classes. Inutilisable pour la classification.
KNN Euclidien (k=28)	Interprétable, rapide à entraîner. Pas de phase d'apprentissage.	Biais extrême vers classe majoritaire. 83% des eaux potables ratées. Recall catastrophique.
KNN Pondéré (k=28)	Légère amélioration par rapport au KNN classique. Un peu plus sensible.	Toujours très faible détection des potables. Pas de gain significatif.
KNN + SMOTE	Meilleur recall potable parmi tous les KNN testés. Faux négatifs fortement réduits.	Accuracy en baisse notable. Introduction de bruit artificiel par SMOTE.
SVM (RBF, C=1)	Meilleure accuracy globale sans rééquilibrage (67%). Capture les non-linéarités.	Recall potable très faible (27%) sans rééquilibrage. Biais important vers classe 0.
SVM + SMOTE	Meilleure détection potable du projet (Recall = 54%). F1-Score le plus élevé (53%). Recall doublé vs SVM brut.	Légère baisse d'accuracy. Dépendance à la technique SMOTE. Performances modestes en absolu.
ANN (MLP 9→32→16→1)	Capture les relations non-linéaires complexes. Apprentissage stable sans surapprentissage.	Gain limité par la taille et la qualité du dataset. Nécessite plus de données pour exceller.

TABLE 9 – Analyse qualitative comparative des modèles

6 Conclusion

Ce projet nous a permis d'appliquer et de comparer plusieurs méthodes de machine learning sur un problème réel de classification de la potabilité de l'eau. Nous avons exploré trois grandes approches : le clustering non supervisé (K-Means), les algorithmes classiques (KNN, SVM) et l'apprentissage profond (ANN).

Bien que les performances obtenues restent modestes , ce travail nous a permis d'acquérir une expérience complète et concrète du workflow en machine learning : exploration des données, nettoyage, prétraitement, normalisation, optimisation des hyperparamètres, évaluation rigoureuse et comparaison des modèles.

L'un des enseignements majeurs de ce projet est que l'accuracy seule est trompeuse dans les problèmes déséquilibrés. L'analyse détaillée des matrices de confusion et l'utilisation de techniques comme SMOTE se sont révélées indispensables pour améliorer la détection de la classe minoritaire (eau potable). Nous avons également constaté que tous les modèles, des plus simples aux plus complexes, convergent vers des performances similaires, confirmant que la difficulté provient davantage de la qualité des données que des algorithmes eux-mêmes.

Ce travail nous a permis de développer des compétences essentielles en data science : manipulation de données avec pandas, visualisation avec matplotlib et seaborn, modélisation avec scikit-learn et Keras, interprétation critique des résultats et rédaction d'un rapport technique complet. Au-delà des résultats chiffrés, nous avons appris à identifier les limites d'un jeu de données, à diagnostiquer les sources de difficulté et à adopter une démarche

scientifique rigoureuse face à un problème de classification supervisée.

En conclusion, ce projet constitue une expérience formatrice et réaliste qui illustre à la fois les possibilités et les limites du machine learning appliquée à des données environnementales bruitées et déséquilibrées.