Projets SY09 — La prédiction de la potabilité de l'eau

Elsa Strassia, Victor Demessance, Bartlomiej Grzadziel

09/07/2024

1 Introduction

Ce document présente un résumé de notre travail d'analyse et de traitement des données sur le dataset Water Potability Prediction.

Le jeu de données *Water Potability Prediction* contient près de 2300 entrées décrites par 10 colonnes, toutes quantitatives, dont une étant la colonne de prédiction sur la potabilité de l'eau.
À travers le projet SY09, nous souhaitons explorer et analyser cet ensemble de données d'échantillons d'eau afin de construire un modèle capable de prédire convenablement la potabilité de l'eau à partir de ses composants. Nous commencerons par prétraiter les données pour les nettoyer et les préparer. Ensuite, nous effectuerons la modélisation, la prévision et la classification pour déterminer l'état de potabilité de l'eau échantillonnée.

2 Présentation du jeu de données

L'eau, essentielle à la vie, est souvent contaminée, causant 1,5 million de décès annuels. L'accès à une eau potable sûre est un droit humain crucial pour la santé. Avec l'apprentissage automatique, nous allons analyser et prédire la qualité de l'eau à partir des données fournies.
Notre jeu de données est composé de variables correspondant à divers composants présents dans l'eau, tels que les solides, les chloramines et les sulfates. Certaines colonnes, plus spécifiques, comme le *pH*, représentent des propriétés chimiques de l'eau.

3 Preprocessing

3.1 Valeurs aberrantes et anormales

Lors de l'analyse de certaines données de notre dataset, en particulier celle du *pH* de l'eau, nous avons remarqué que certaines valeurs pourraient être qualifiées d'aberrantes.

En effet, prenons l'exemple de l'eau potable qui devrait avoir un pH compris en moyenne entre *6.5 et 9.5* d'après les normes de l'OMS. Or, comme nous pouvons l'observer sur la *Figure 1*, certaines eaux qualifiées de potables présentent un pH en dehors de cet intervalle, allant parfois aux extrêmes avec de l'eau potable ayant un pH de 1 ou 14. Ces valeurs peuvent être considérées comme anormales et soulèvent un potentiel problème quant à la véracité et à la pertinence de notre variable de classification. Quant aux valeurs manquantes, notre jeu de données en est exempt.

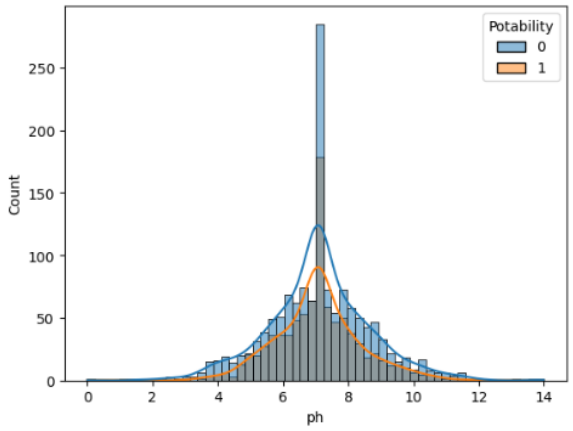

FIGURE 1 – Histogramme de répartition du pH de l'eau en fonction de sa potabilité

3.2 Critique du dataset

Les valeurs aberrantes découvertes lors de l'analyse de la répartition des données de pH en fonction de la potabilité nous poussent à élargir cette analyse à toutes les variables utilisées dans le dataset. Ainsi, nous calculons et affichons la répartition de chaque variable en fonction des deux classes de potabilité. La visualisation de la répartition des chloramines (*Figure n°2*) nous permet de confirmer l'intuition que nous avions.Une eau potable ne devrait pas contenir plus de 8 ppm de chloramines ;

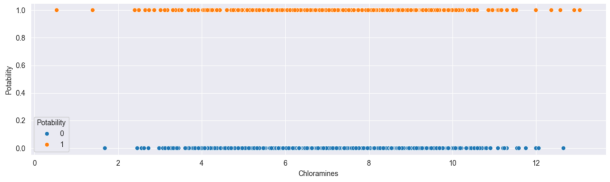

FIGURE 2 – Répartition de la variable chloramines en fonction de la classe de prédiction

or, nous observons ici des prédictions d'eau potable avec des concentrations bien supérieures.

Ce résultat peut se généraliser à l'ensemble du dataset puisque, pour chaque variable, la répartition des données est totalement incohérente lorsqu'on met en relation la classe de l'échantillon avec les normes de potabilité de l'eau.

La solution envisagée serait de créer nous-mêmes une variable de classe qui simulerait le processus de classification manuel des données récupérées. Pour cela, il est nécessaire de prendre en compte les normes de potabilité relatives à chacune des variables de notre dataset, puis de créer une fonction recevant en paramètre chaque échantillon et déterminant sa valeur de potabilité.

3.3 Simulation des données de classes

Les règles de potabilité utilisées dans cette analyse sont basées sur les recommandations de l'Organisation Mondiale de la Santé (OMS) ainsi que sur les réglementations Européennes et Canadiennes. Les critères retenus pour déterminer la potabilité de l'eau sont les suivants :

- *pH* : Entre 6.5 et 9.5
- *Dureté* : Entre 150 et 500 mg/L
- *Solides* : Pas de limite supérieure
- *Chloramines* : $<$ 8 ppm
- *Sulfate* : $<$ 500 mg/L
- *Conductivité* : $<$ 1200 us/cm
- *Carbone organique* : Pas de limite spécifique
- *Trihalométhanes* : $<$ 80 ug/L
- *Turbidité* : $<$ 5 NTU

Après ce travail de preprocessing, les répartitions ne sont désormais plus aberrantes et respectent les règles de potabilité de l'eau courante.

Toutefois, étant donné que notre dataset représente désormais une combinaison linéaire simple de nos variables, il est logique que la plupart de nos algorithmes d'apprentissage obtiennent de bonnes performances. C'est pourquoi nous fournirons toujours les performances de notre modèle sur le dataset original afin d'effectuer une comparaison pertinente.

3.4 Normalisation des données

Pour garantir que les algorithmes d'apprentissage automatique traitent les variables quantitatives de manière équitable, nous avons normalisé ces variables. Nous avons utilisé la méthode de standardisation, qui consiste à ajuster chaque valeur en soustrayant la moyenne et en divisant par l'écart-type, afin de les mettre sur la même échelle.

4 Analyse exploratoire

4.1 Analyse univariée

Sur la base de la visualisation univariée de chaque variable, on peut déduire que la plupart des distributions sont symétriques. De plus, les valeurs mesurées sont relativement centrées autour de la moyenne, avec peu de valeurs extrêmes.

Par ailleurs, la distribution de notre variable de classe reste relativement équilibrée. On constate que la plupart de l'eau dans nos données est impropre à la consommation (*39% d'eau potable contre 61% d'eau non potable*), mais il est tout à fait possible de construire des modèles d'apprentissage pertinents.

4.2 Analyse bivariée

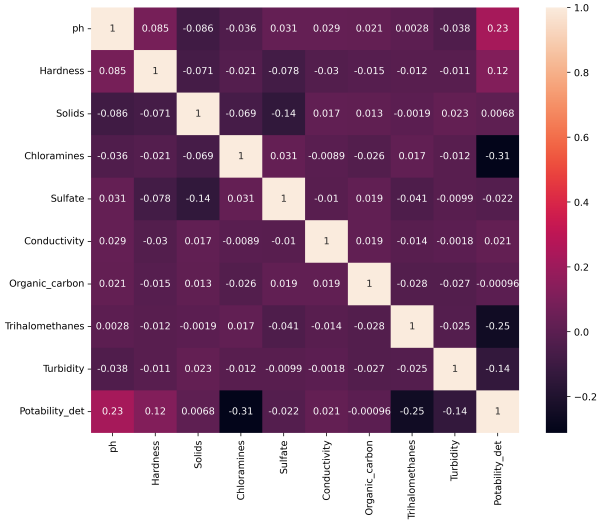

FIGURE 3 – Matrice de corrélation entre les variables

Comme on peut l'observer sur la *Figure n°3*, les variables présentes dans le dataset sont très peu corrélées entre elles.

Cependant, il semble que certaines variables possèdent tout de même une corrélation intéressante avec la variable de classification. On peut relever cela notamment pour le *pH*, les *chloramines*, les *trihalométhanes* et, dans une moindre mesure, la *dureté* et la *turbidité*. Dans le cadre de notre étude, nous pouvons conclure que ces variables *ont une influence relativement importante sur la potabilité de l'eau.*

5 Réduction de dimensionnalité

Dans le but de réduire la dimensionnalité de notre dataset et d'en améliorer la visualisation, nous allons effectuer une analyse en composantes principales.

Les variables que nous utilisons pour l'ACP sont des variables quantitatives continues normalisées *(les 9 variables de notre jeu de données peuvent donc être utilisées).*

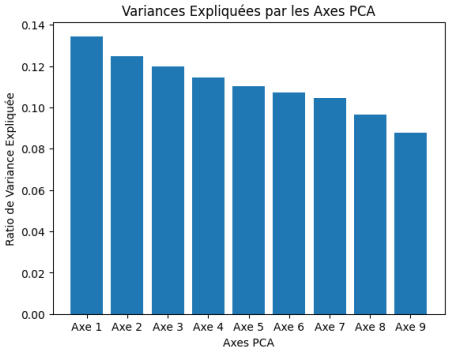

FIGURE 4 – Inerties expliquées des composantes principales

Comme nous l'attendions, la réalisation de l'ACP ne donne pas de résultats très convaincants. En effet, les inerties expliquées par les trois premiers axes sont relativement faibles (*Figure n°4*).

Il est donc difficile de réaliser une réduction de dimension sans perdre des informations importantes, notamment dans le cas d'une réduction significative (dans un espace de dimension visualisable, donc inférieur à 3). C'est pourquoi nous n'allons pas utiliser les résultats de l'ACP pour construire des modèles de prédiction, étant donné que nous perdrions environ 75% de l'informations.

Comme prévu, la projection sur les deux premiers axes principaux ne permet pas de distinguer clairement l'eau potable de l'eau non potable. Les points forment un nuage indistinct, rendant impossible toute délimitation nette entre les deux classes.

6 Prédiction de la potabilité

L'objectif principal de ce jeu de données est de construire un modèle capable de prédire la potabilité de l'eau en se basant sur ses différents composants. Nous avons commencé par une approche d'apprentissage non supervisé en utilisant un modèle basé sur les *K-means*. Cependant, ce modèle n'a pas donné de résultats satisfaisants puisque les clusters trouvés ne correspondaient pas du tout aux clusters réels.

Ensuite, nous avons plus appliquer différentes méthodes d'analyse (LDA, QDA...) pour évaluer l'applicabilité de la théorie bayésienne et des analyses discriminantes à notre dataset. Pour ce faire, nous avons effectué un test de *Shapiro-Wilk* afin de vérifier la normalité de nos variables. Ce test compare les quantiles empiriques de l'échantillon aux quantiles théoriques d'une distribution normale pour tester l'hypothèse nulle de normalité.

Suite à ce test, nous avons identifié que seulement deux caractéristiques suivent une distribution normale : le carbone organique et la turbidité. Néanmoins, leur très faible corrélation, illustrée par la heatmap (*Figure n°3*), laisse présager des défis pour les prédictions.

Trois méthodes d'analyse discriminante ont ensuite été appliquées à ces deux variables : l'Analyse Discriminante Linéaire (*LDA*), l'Analyse Discriminante Quadratique (*QDA*) et le modèle Naive Bayes Gaussien (*NB*).

Les résultats ont montré des modèles moyennement performants en raison de la faible corrélation entre les variables et de la difficulté à former des frontières de décision claires. En conclusion, les analyses discriminantes ne sont pas particulièrement efficaces pour prédire la qualité de l'eau de manière robuste.

Nous ne présenterons donc dans ce rapport que les modèles de classification les plus performants, à savoir le *KNN*, les *arbres de décision* et la *régression logistique*. Tous les résultats ont été obtenus à l'aide d'une validation croisée à 5 plis afin d'assurer une évaluation représentative et robuste.

6.1 Méthode des KNN

Le *KNN* est un algorithme de classification simple mais efficace. Il classe un échantillon en fonction de la classe majoritaire de ses k voisins les plus proches, et son

efficacité dépend fortement du choix du nombre de voisins k. C'est pourquoi, pour notre modèle de *KNN*, nous appliquons une validation croisée à 10 plis pour déterminer le nombre de voisins le plus approprié. Comme le montre la *Figure n°5*, c'est avec 16 voisins les plus proches que nous obtenons le modèle le plus performant.

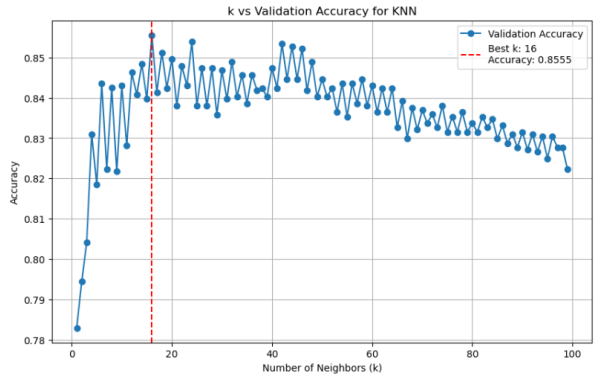

FIGURE 5 – Précision du modèle Knn en fonction de l'hyperparamètre

Après apprentissage, la méthode des *K-nearest neighbors* (*KNN*) a bien fonctionné pour la classification de nos données, puisque nous obtenons un résultat convaincant avec un taux de *85%* de bonnes prédictions en utilisant la méthode de validation croisée et pour $k = 16$ voisins.

La visualisation de la frontière de décision peut se faire via ACP (*Figure n°6*).

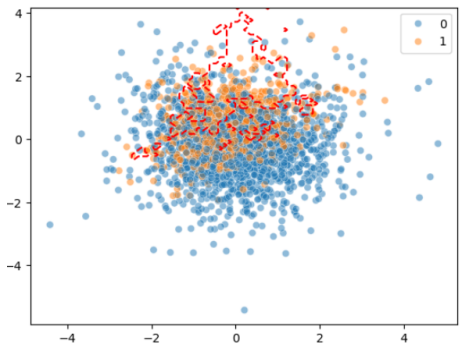

FIGURE 6 – Frontière de décision pour 16nn

Cette illustration ne reflète pas entièrement les résultats obtenus, mais c'est l'une des rares visualisations de nos modèles qui permet de cerner clairement la frontière de décision et qui témoigne visuellement de la performance de notre méthode, puisque la majorité des points sont bien classés.

Quant à notre dataset original, nous obtenons un taux de prédiction de 65%. Comme prévu, la performance est moindre en raison des problèmes mis en lumière précédemment.

6.2 Arbre de décision

L'utilisation des arbres de décision pour construire un bon modèle de classification sur notre dataset est un choix évident, à condition qu'ils soient bien implémentés. En effet, les arbres de décision fonctionnent bien avec des variables non corrélées et des règles de décision claires. Il est donc normal de s'attendre à d'excellents résultats sur notre dataset corrigé, puisqu'il s'agit de différencier l'eau potable de l'eau non potable à partir de critères appliqués à chaque colonne. L'arbre de décision va naturellement trouver ces frontières et produire de bons résultats.

Pour garantir des résultats fiables, nous avons pris en compte les faiblesses des arbres de décision, notamment leur tendance à l'overfitting. Pour résoudre ce problème, nous avons utilisé le bagging, qui consiste à effectuer un échantillonnage aléatoire avec remise des données

Pour renforcer notre approche, nous avons appliqué la technique des forêts aléatoires (*random forests*). En plus de l'échantillonnage aléatoire, les forêts aléatoires ajoutent une couche de randomisation supplémentaire en sélectionnant aléatoirement un sous-ensemble des caractéristiques à chaque nœud de l'arbre. Cela introduit une plus grande diversité parmi les arbres, ce qui conduit à une meilleure performance globale et à une réduction significative du risque d'overfitting.

Les résultats obtenus sont très bons, avec un taux de précision de 99% pour nos prédictions. Sur la *Figure 7*, nous pouvons observer l'importance de chaque variable dans les décisions prises par l'arbre de décision. Comme attendu, le *pH* a le plus grand impact sur les prédictions. Les chloramines et les trihalométhanes jouent également un rôle significatif dans le modèle, influençant fortement les décisions. Ces résultats ne sont pas surprenants, puisque ce sont les trois éléments de la heatmap qui présentent la plus forte corrélation (positive et négative) avec la variable de prédiction *Probability_ det.*

En revanche, pour le dataset original, les résultats sont moins bons, puisqu'il n'y a pas de seuils de potabilité clairs selon les colonnes. Presque toutes les variables peuvent prendre des valeurs aléatoires tout en produisant de l'eau potable. C'est pourquoi nous obtenons une précision plus faible de 68%.

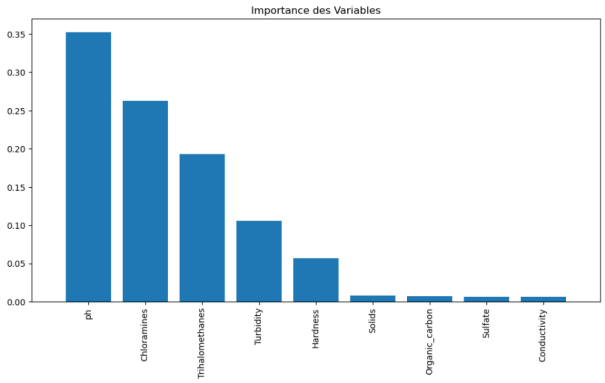

FIGURE 7 – Importance des variables obtenue

6.3 Régression Logistique

Le modèle de régression logistique semble être une approche intéressante pour nos données. En effet, ce modèle permet de prédire la probabilité d'appartenir à une classe binaire, ce qui correspond à notre variable de prédiction. De plus, la régression logistique est particulièrement efficace sur des variables peu corrélées. Pour valider notre hypothèse d'indépendance, nous nous référons à notre matrice de corrélation *(Figure n°3)*. La corrélation entre nos différentes variables étant faible dans la grande majorité des cas, nous pouvons supposer que l'hypothèse d'indépendance est respectée.

Enfin, malgré le taux de corrélation assez faible entre nos variables et la variable explicative (0.23 dans le meilleur des cas entre le ph et la potabiltié), nous allons tout de même valider l'hypothèse de corrélation entre les variables dépendantes et la variable explicative.

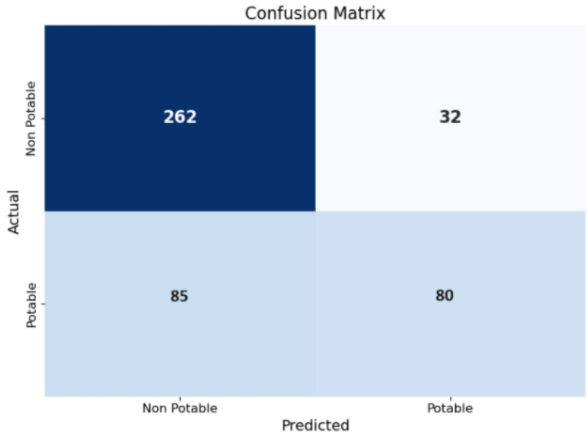

FIGURE 8 – Matrice de confusion du modèle de régression logistique

Grâce à la matrice de confusion *(Figure n°8)*, il est possible de déterminer les performances de l'apprentissage. Le modèle de régression logistique affiche une précision globale de 74,5%, indiquant une majorité de prédictions correctes. La précision pour les échantillons d'eau potable est de 71,4%, démontrant que la plupart des échantillons prévus comme potables sont correctement identifiés. Cependant, le rappel est de 48,5%, révélant que moins de la moitié des échantillons d'eau potable sont correctement détectés. Le F1-score, à 57,6%, offre une mesure équilibrée de la précision et du rappel. Ces résultats suggèrent que, bien que performant pour prédire l'eau non potable, le modèle manque de nombreux cas d'eau potable.

Cela peut être dû au fait que le dataset contient davantage de données correspondant à de l'eau non potable qu'à de l'eau potable, ce qui rend le modèle meilleur pour prédire l'eau non potable que l'eau potable. Pour améliorer notre classifieur, il serait bénéfique d'équilibrer la distribution de notre dataset d'apprentissage (via des techniques de sur/sous-échantillonnage). Cependant, cette approche comporte des risques, notamment celui de l'*overfitting*.

7 Conclusion

Nous sortons de cette analyse de notre dataset avec des sentiments mitigés. En effet, nous sommes confrontés à un dilemme.

En corrigeant les données de notre dataset, nous avons grandement simplifié la tâche de nos modèles de classification pour prédire la potabilité de l'eau. Cela a conduit à des taux de prédiction anormalement élevés (jusqu'à 99% pour les arbres de décision). Cependant, en utilisant le dataset original pour construire ces mêmes classifieurs, nous avons obtenu des taux de prédiction relativement plus faibles, le meilleur étant de 68%. Ce résultat est bien trop bas pour prétendre avoir construit un modèle de prédiction fiable, ce qui pourrait être dangereux quant au sujet de notre dataset.

Ainsi, nous nous trouvons dans l'impossibilité de clôturer cette analyse de manière satisfaisante. Sans un dataset alternatif avec des valeurs différentes et non anormales, il est impossible de vérifier la performance de nos classifieurs. Par conséquent, nous ne pouvons pas affirmer avoir construit un modèle de prédiction efficace de la potabilité de l'eau. Cette analyse nous a toutefois fait prendre conscience de l'importance cruciale de la fiabilité des datasets pour réaliser une analyse précise et pertinente.

8 Annexes

8.1 Tableau des performances des classifieurs

	Dataset corrigé	Dataset non corrigé
Kmeans	60%	60%
Knn	85%	65%
LDA	69%	59%
QDA	80%	65%
NB	80%	65%
Forêt aléatoire	99%	68%
RL	74.5%	57%
Décision de Bayes	88,31%	61,67%

TABLE 1 – Figure 8 : Performances des méthodes de classification

Table des figures

9 Bibliographie

— Organisation mondiale de la Santé. (2017). *Directives de qualité pour l'eau de boisson : Quatrième édition.* Disponible sur : https://www.who.int/fr/publications-detail/9789241549950

— Lenntech. (n.d.). *Normes de l'OMS pour l'eau potable.* Disponible sur : https://www.lenntech.fr/applications/potable/normes/normes-oms-eau-potable.htm

— Lenntech. (n.d.). *Normes OMS et UE pour l'eau potable.* Disponible sur : https://www.lenntech.fr/francais/norme-eau-potable-oms-ue.htm

— Suez. (n.d.). *Normes de qualité des eaux potables.* Disponible sur : https://www.suezwaterhandbook.fr/eau-et-generalites/quelles-eaux-a-traiter-pourquoi/les-eaux-potables-EP/normes-de-qualite

10 Liens

— Github du projet
— Lien du dataset.