

PREVISÃO DA QUALIDADE DA ÁGUA: UMA APLICAÇÃO DO MODELO DE APRENDIZADO SUPERVISIONADO XGBOOST

WATER QUALITY PREDICTION: AN APPLICATION OF THE XGBOOST SUPERVISED LEARNING MODEL

JOSÉ AIRTON AZEVEDO DOS SANTOS

Universidade Tecnológica Federal do Paraná (UTFPR), Medianeira / PR
airton@utfpr.edu.br

ALDINO NORMELIO BRUN POLO

Universidade Tecnológica Federal do Paraná (UTFPR), Medianeira / PR
aldinopolo@alunos.utfpr.edu.br

Resumo: Avaliar a adequação da água potável, recurso essencial para sustentar a vida humana, é vital para gerações atuais e futuras. Sua previsão precisa, por meio de algoritmos de Machine Learning, pode ajudar a reduzir a incidência de doenças transmitidas pela água, tais como diarreia, cólera, entre outras. Neste contexto, este trabalho tem como objetivo avaliar o desempenho, na detecção de água potável, do modelo XGBoost (Xtreme Gradient Boosting). Utilizou-se, para realizar esta avaliação, a base de dados de qualidade da água Water Quality, a técnica resample para balanceamento das classes e a biblioteca de ajuste de hiperparâmetros Optuna. A base de dados possui os seguintes parâmetros: pH, Hardness, Solids, Chloramine, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity e Potability. As métricas Accuracy, Precision, Sensitivity, F1-Score e AUC foram utilizadas para avaliar o desempenho do modelo. Resultados experimentais demonstraram que o modelo XGBoost apresentou muito bom desempenho de classificação, alcançando 92% de acurácia.

Palavras-chave: Classificação. Kaggle. Optuna. Python.

Abstract: Ensure the adequacy of potable water, an essential resource for sustaining human life and vital for current and future generations. Its precise forecast, through Machine Learning algorithms, can help reduce the incidence of diseases transmitted by water, such as diarrhea, cholera, among others. In this context, this work aims to evaluate the performance, in the detection of potable water, of the XGBoost (Xtreme Gradient Boosting) model. To carry out this evaluation, water quality data, a resampling technique for class balancing, and the Optuna hyperparameter tuning library were used to carry out this evaluation. Based on the following parameters: pH, Hardness, Solids, Chloramine, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity and Potability. The metrics Accuracy, Precision, Sensitivity, F1-Score, and AUC are used to evaluate the performance of the model. Experimental results demonstrate that the XGBoost model presents excellent classification performance, reaching 92% accuracy.

Keywords: Classification. Kaggle. Optuna. Python.

Introdução

A água é um dos recursos naturais essenciais, sem a qual não pode existir vida. O fornecimento de água potável, que atenda os padrões de qualidade sanitária, é muito importante para a saúde da população. Ela fornece energia, ajuda a regular a temperatura e a função renal e limpa o corpo. Atua também como um transportador de vitaminas e minerais (POUDEL et al., 2022).

O acesso a água potável é vital para prevenir doenças. Dentre as principais doenças transmitidas pela água tem-se a: Diarreia por *Escherichia coli*, Cólera, Disenteria bacteriana, Hepatite A, Esquistossomose e a Febre tifoide. Recentes descobertas indicam que aproximadamente 3,5 milhões de pessoas perdem suas vidas, a cada ano, devido a doenças relacionadas com a água. Além de problemas relacionados com a saúde, a potabilidade da água também influencia na produtividade econômica e no desenvolvimento sustentável da população. Observa-se que a água, devido à sua importância para a vida, economia e valor estratégico, é um tópico de preocupação mundial (JUNA et al., 2022; PATEL et al., 2023; SINAP, 2024).

Atualmente, técnicas de *Machine Learning* estão sendo muito utilizadas para detectar e monitorar, com precisão e rapidez, a qualidade da água (YURTSEVER; EMEÇ, 2023). Dentre estas técnicas, utilizadas para detectar a qualidade da água, tem-se o algoritmo XGBoost, um dos algoritmos de melhor desempenho em aprendizado supervisionado. Este algoritmo pode ser utilizado para tratar tanto problemas de regressão quanto de classificação (OSMAN et al., 2021).

Vários trabalhos, apresentados na literatura, utilizaram algoritmos de *machine learning* para detectar, por meio da base de dados *Water Quality*, a qualidade da água. Dentre eles pode-se citar os estudos realizados por: Fen et al. (2021) que utilizaram, em seu estudo da qualidade da água, os modelos de aprendizado de máquina *Support vector machine* (SVM), *Gaussian Naive Bayes* (GNB), *K-Nearest Neighbor* (KNN), *Decision tree* (DT) e *Extra Trees Classifier* (ETC). Concluíram que o modelo ETC apresentou a melhor acurácia (86,67%). Chafloque et al. (2021) implementaram, para prever a potabilidade da água, um modelo de rede neural com sete camadas ocultas. O modelo desenvolvido apresentou uma acurácia de aproximadamente 70%. Poudel et al. (2022) utilizaram, para prever a potabilidade da água, os algoritmos LR (*Logistic Regression*), KNN (*K-Nearest Neighbor*), ANN (*Artificial Neural Network*) e RF (*Random Forest*). Concluíram que o modelo com maior acurácia foi o modelo RF (70%). Abraham et al. (2022) implementaram modelos de aprendizado de máquina supervisionados e não supervisionados. Para os modelos supervisionados utilizaram o conjunto de dados *Water Quality*. Utilizaram, neste estudo, os modelos LR (*Logistic Regression*), RF (*Random Forest*), NB (*Naive Bayes*) e KNN (*K-Nearest Neighbor*). O RF, dentre esses modelos, apresentou a melhor acurácia (65%). Preetham e Krfishnan utilizaram, para prever a potabilidade da água, os modelos *Decision Tree*, *Random Forest*, *XGBoost*, *Catboost* e *Adaboost*. O Catboost

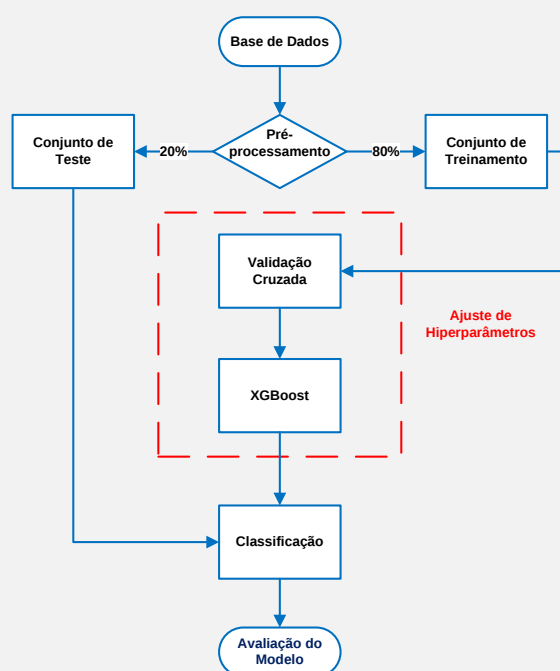
demonstrou, dentre os cinco algoritmos de previsão, um desempenho superior ao atingir uma acurácia de 75%. Já Abuzir e Abuzir (2022) implementaram os algoritmos j48, GNB (*Gaussian Naive Bayes*) e MLP (*MultiLayer Perceptron*) para previsão da qualidade da água. Observaram que o modelo MLP obteve um AUC (*Area under the ROC Curve*) de 66,1%.

Neste contexto, este trabalho tem como objetivo avaliar, no conjunto de dados *Water Quality*, o desempenho, na detecção de água potável, do modelo XGBoost (*Xtreme Gradient Boosting*). Ressalta-se que o conjunto de dados *Water Quality*, utilizado nesta pesquisa, apresenta atributos com associações fracas, dados discrepantes e um grande número de amostras faltantes. Estes três fatores têm um grande impacto nos resultados de previsão da qualidade da água. Portanto, como contribuição deste trabalho, abordagens de pré-processamento adequadas foram implementadas para lidar com estes três fatores.

Metodologia

Basicamente, as etapas deste trabalho incluem: o pré-processamento dos dados, os ajustes dos hiperparâmetros, o processo de classificação e a avaliação do desempenho do modelo (Figura 1).

Figura 1. Fluxograma deste estudo.



Fonte: SANTOS J. A. A. e POLO A. N. B. (2024).

Base de dados

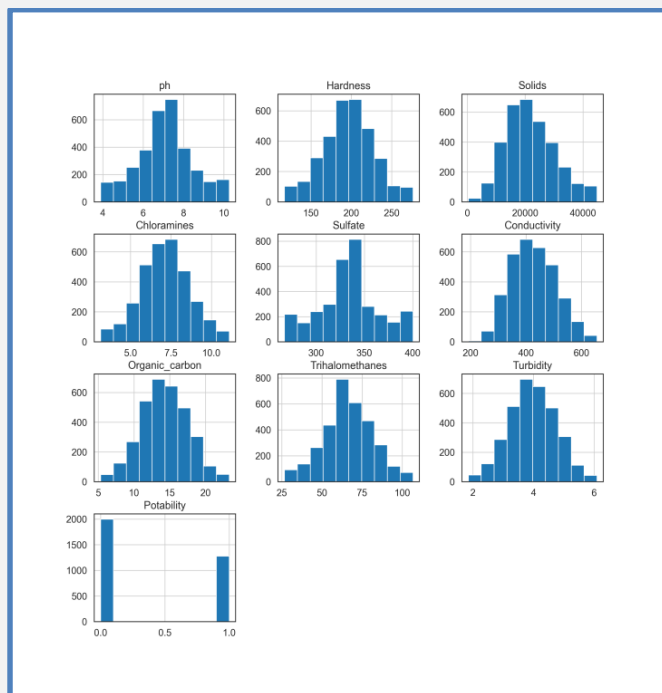
A base de dados da qualidade da água (*Water Quality*), disponível por meio de uma licença CC0: *Public Domain*, foi obtida do repositório Kaggle (KADIVAL, 2020). Esta base de dados, com 3276 observações, é formada pelas seguintes variáveis:

- *pH*: Indica o grau de acidez ou alcalinidade da água, O limite máximo permitido do pH, segundo a Organização Mundial de Saúde (OMS), é de 6.6 a 8.0.
- *Hardness* (Dureza): Reflete o teor de sais de cálcio e magnésio da água, impactando sua dureza bruta.
- *Solids* (Total de Sólidos Dissolvidos): Reflete a mineralização da água, aderindo aos limites da OMS (desejável: 500 mg/l, máximo: 1000 mg/l).
- *Chloramines* (Cloraminas): Resultam da amônia adicionada ao cloro para desinfecção da água, com níveis seguros de até 4 mg/L.
- *Sulfate* (Sulfato): Os sulfatos são substâncias naturais encontradas em minerais, solo e rochas. Sua concentração varia de 3 a 30 mg/L em água doce.
- *Conductivity* (Condutividade): reflete a concentração de íons da água, seguindo o limite de 400 $\mu\text{S}/\text{cm}$ da OMS.
- *Organic_carbon* (Carbono Orgânico): mede o conteúdo de carbono em compostos orgânicos da água, em conformidade com os padrões da *U.S. Environmental Protection Agency* ($< 2\text{mg}/\text{L}$).
- *Trihalomethanes* (Trihalometanos): produtos químicos que podem ser encontrados na água clorada, com níveis dentro do limite seguro de 80 ppm.
- *Turbidity* (Turbidez): Reflete a quantidade de matéria sólida na água, com valores que atendem ao limite recomendado pela OMS (0,98 NTU).
- *Potability* (Potabilidade): Indica se a água é segura para consumo humano, onde 1 significa Potável e 0 significa Não Potável.

As distribuições, apresentadas na Figura 2, revelam que as variáveis, *ph*, *Hardness*, *Chloramine*, *Sulfate*, *Organic_carbon*, *Trihalomethanes* e *Turbidity*, apresentam um padrão de distribuição relativamente uniforme. Enquanto, *Solids* e *Conductivity* têm uma distribuição

Revista Mirante, Anápolis (GO), v. 18, n. 1, p. 17-31, jun. 2025. ISSN 1981-4089
 enviada para a direita. *Potability* é uma variável binária, com apenas dois valores possíveis 0 e 1.

Figura 2. Distribuições dos parâmetros da base de dados.

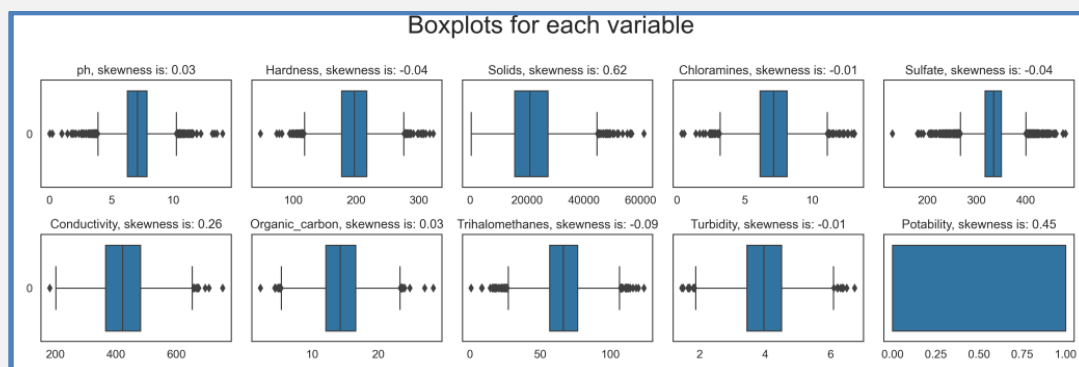


Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Pré-processamento

O pré-processamento, etapa muito importante para a técnica de aprendizado de máquina, impacta diretamente no processo de aprendizagem do modelo. Inicialmente, nesta etapa, observou-se que a base de dados continha valores ausentes nas variáveis *pH* (491), *Sulfate* (781) e *Trihalomethanes* (162). No lugar dos valores ausentes foram atribuídos valores médios de cada uma das variáveis. Observou-se também que a base de dados possuía *outliers* (Figura 3). Neste trabalho os *outliers*, das variáveis *pH*, *Hardness*, *Solids*, *Chloramine*, *Sulfate*, *Conductivity*, *Organic_carbon*, *Trihalomethanes* e *Turbidity*, foram substituídos pelos Limites Inferior e Superior do *boxplot* de cada uma das variáveis. *Outliers* com valores menores que o Limite Inferior foram substituídos pelo Limite Inferior e valores maiores que o limite superior foram substituídos pelo Limite Superior.

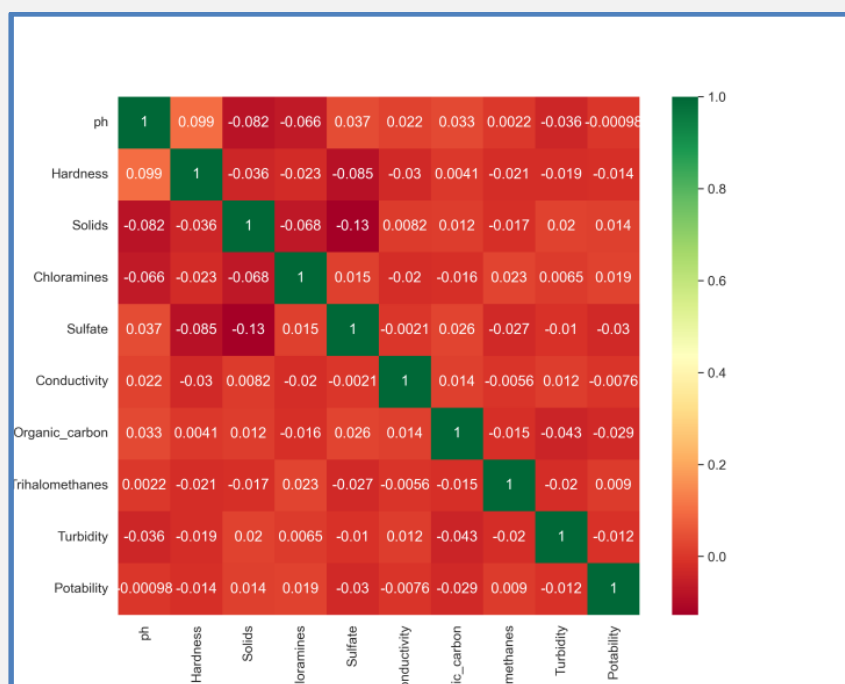
Figura 3. Boxplots com outliers



Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Observou-se também que as correlações das variáveis, *pH*, *Hardness*, *Solids*, *Chloramine*, *Sulfate*, *Conductivity*, *Organic_carbon*, *Trihalomethanes* e *Turbidity* com a variável *Potability*, são muito baixas (Figura 4).

Figura 4. Correlação entre as variáveis.

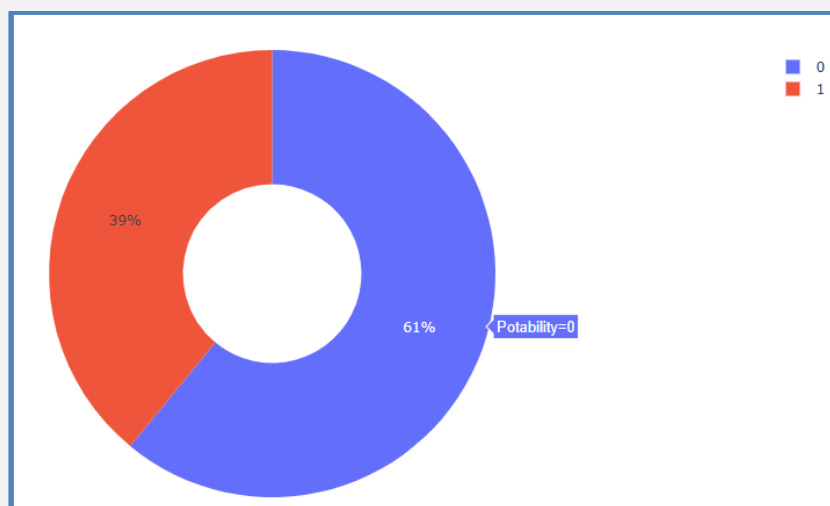


Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Notou-se também que a base de dados possuía distribuição desigual de amostras para variável *Potability* (61% - Não Potável (0) e 39% - Potável (1)) (Figura 5). Observa-se que uma proporção significativa (61%) se enquadra na categoria Não Potável. Dados desbalanceados

Revista Mirante, Anápolis (GO), v. 18, n. 1, p. 17-31, jun. 2025. ISSN 1981-4089 podem impactar na generalização do modelo de classificação (MACEDO; APORETTI, 2024). Utilizou-se, para realizar o balanceamento das classes Potável (1) e Não Potável (0), a função *resample* da biblioteca Scikit-Learn. A função *resample* cria dados sintéticos para classe minoritária até que as classes estejam em equilíbrio.

Figura 5. Distribuição desigual das amostras - *Potability*



Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Nesta etapa a base de dados foi dividida em 80% dos dados para o conjunto de Treinamento e 20% para o conjunto de Teste.

O estágio de pré-processamento consistiu basicamente das seguintes etapas:

- Detecção e processamento dos dados ausentes.
- Detecção e processamento dos *outliers*.
- Normalização dos dados por meio da função *MinMaxScaler*.
- Balanceamento das classes utilizando a função *resample*.
- Divisão da base de dados em um conjunto de Treinamento (80%) e o outro de Teste (20%).

Xtreme Gradient Boosting (XGBoost)

O XGBoost, proposto por Chen e Guestrin em 2016, é um algoritmo, de aprendizado de máquina, conhecido por sua eficiência, velocidade e precisão. É um algoritmo baseado em

árvores de decisão com aumento do gradiente (*gradient boosting*). Fundamentalmente, o XGBoost constrói árvores na sequência, de modo que a árvore subsequente aprenda e reduza os erros da anterior. Foi desenvolvido com o propósito de se obter alto desempenho e velocidade computacional. É um algoritmo de *machine learning* utilizado tanto para regressão quanto para classificação. A equação do algoritmo XGBoost é apresentada na Equação 1 (SINAP, 2024).

$$F(x) = \sum_{m=1}^M \gamma_m b_m(x) \quad (1)$$

Onde $F(x)$ é o modelo de previsão, M é o número de árvores, γ_m é a taxa de aprendizagem de cada árvore e $b_m(x)$ é a previsão de cada árvore.

Métricas

O desempenho do algoritmo XGBoost foi avaliado pelas seguintes métricas (BUDHOLIYA et al., 2022; CHANG et al., 2023; MAULANA et al., 2023):

- *Accuracy*: A *accuracy* é o índice de avaliação mais comum. Em termos gerais, quanto maior a *accuracy*, melhor o classificador.
- *Precision*: indica a proporção de previsões positivas feitas pelo modelo que estão corretas.
- *Sensitivity (Recall)*: indica a proporção de exemplos positivos que foram corretamente identificados pelo modelo.
- *F1-Score*: Indica a precisão de um modelo de classificação binária, calculado pela média ponderada das métricas *Precision* e *Sensitivity*.

Na Tabela 1 a apresentam-se as equações das métricas *Accuracy*, *Precision*, *Sensitivity* e *F1-Score*.

Tabela 1. Equações das métricas de análise de desempenho.

Métricas
$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN}$
$Precision = \frac{(TP)}{(TP + FP)}$
$Sensitivity (Recall) = \frac{(TP)}{(TP + FN)}$
$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$

Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Onde TP: Verdadeiro Positivo; TN: Verdadeiro negativo; FP: Falso Positivo e FN: Falso Negativo.

Utiliza-se também, para medir o desempenho do modelo, a *Confusion Matrix*, que descreve os resultados da previsão da classificação binária, e a *Area under the ROC Curve* (AUC) que ajuda a avaliar a capacidade de um modelo de classificação. A pontuação AUC varia de 0 a 1, onde 1 indica o desempenho perfeito.

Resultados e Discussão

Treinamento

Os valores dos hiperparâmetros do modelo XGBoost foram ajustados por meio da biblioteca Optuna. O Optuna é um *framework Open Source* para otimização de hiperparâmetros. Ele utiliza algoritmos de otimização baseyana para encontrar os melhores hiperparâmetros de um modelo de classificação (LAI et al., 2023). Observa-se também que, neste trabalho, a validação cruzada *K-fold*, com $k=7$, é adotada para avaliar o desempenho do modelo durante a fase de seleção do modelo.

Na Tabela 2 apresentam-se os valores, encontrados pelo Optuna, para os hiperparâmetros *n_estimators*, *learning_rate*, *max_depth*, *min_child_weight*, *subsample* e *colsample_bytree*. Observa-se que os valores de outros hiperparâmetros do modelo foram mantidos com valores padrão. Maiores detalhes sobre os hiperparâmetros do modelo XGBoost podem ser encontrados em Maulana et al. (2023).

Tabela 2. Hiperparâmetros do modelo

Hiperparâmetro	Intervalo de busca	XGBoost
<i>n_estimators</i>	[50,500]	339
<i>learning_rate</i>	[1e ⁻⁵ ,1e ⁻¹]	0.045
<i>max_depth</i>	[1,15]	15
<i>min_child_weight</i>	[1,10]	1
<i>subsample</i>	[0.0,1.0]	0.926
<i>colsample_bytree</i>	[0.0,1.0]	0.857

Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Teste

Na Tabela 3 apresentam-se as acurácias, do modelo XGBoost, com e sem balanceamento das classes. Pode-se observar, dos resultados de acurácias apresentados na tabela, que o modelo XGBoost, com balanceamento das classes, apresentou maior acurácia (92%).

Tabela 3. Acurácias com (C. B.) e sem (S. B.) balanceamento das classes

Métrica	S. B.	C. B.
<i>accuracy</i>	0,79	0.92

Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Na Figura 6 apresenta-se o desempenho, do modelo XGBoost com balanceamento das classes, no Conjunto de Teste (800 observações). Pode-se notar, dos resultados apresentados nesta figura, que o modelo apresentou uma *accuracy* de 92%, com uma *precision* de 90%, uma *sensitivity* de 93% e um *F1-Score* de 92% para classe 0 e uma *precision* de 94%, uma *sensitivity* de 91% e um *F1-Score* de 92% para classe 1, indicando uma boa capacidade de classificar corretamente águas potáveis e não potáveis.

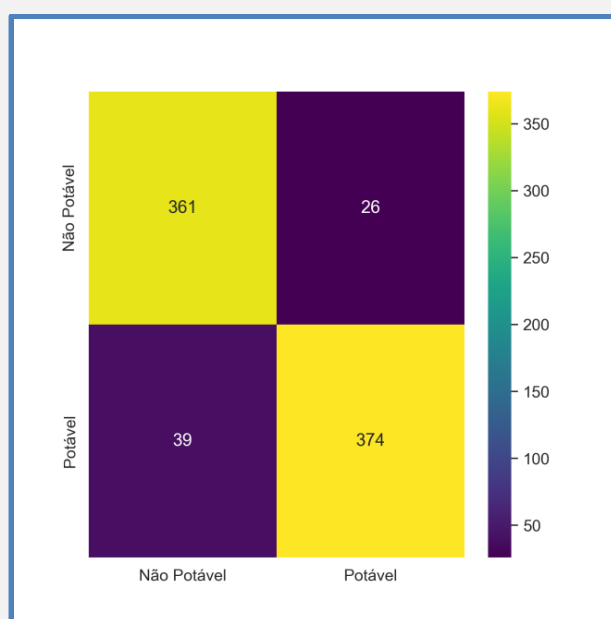
Figura 6. Desempenho do modelo no Conjunto de Teste

	precision	recall	f1-score	support
0	0.90	0.93	0.92	387
1	0.94	0.91	0.92	413
accuracy			0.92	800
macro avg	0.92	0.92	0.92	800
weighted avg	0.92	0.92	0.92	800

Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

A matriz de confusão, do modelo XGBoost (Figura 7), fornece um resumo da capacidade do modelo de diferenciar águas potáveis e não potáveis. Pode-se notar, por meio da matriz do modelo XGBoost, que este modelo obteve-se 26 amostras identificadas erroneamente como Potável e 39 amostras identificadas erroneamente como Não Potável. Portanto, dos 387 casos classificados como Não Potável, o modelo acertou 361 (93%). O modelo, para os classificados como Potável, acertou 374 de 413 casos (91%) (*Recall* – Figura 6).

Figura 7. Matriz de confusão – Resultados

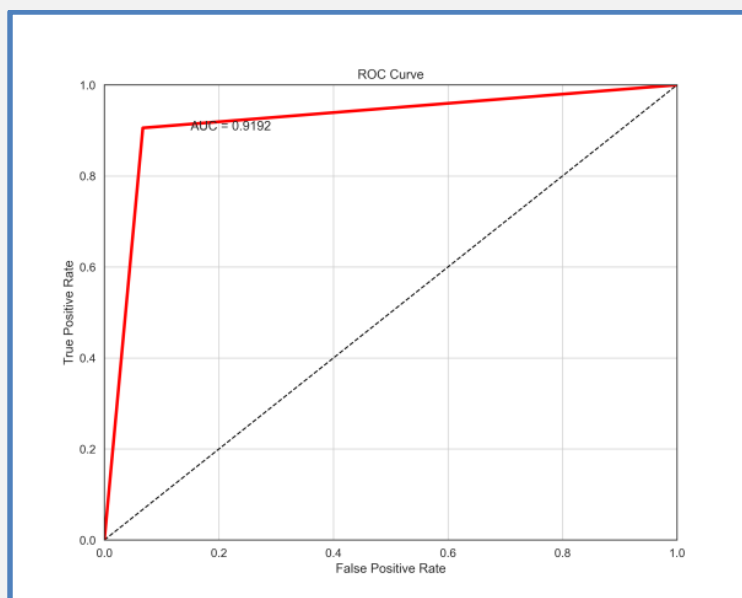


Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

A curva ROC (*Receiver Operating Characteristic Curve*) é apresentada na Figura 8. Pode-se observar pelo valor do indicador AUC (0.9192) que o modelo XGBoost apresentou um bom desempenho. Observa-se que quanto mais perto de 1, o valor de AUC, melhor o desempenho na classificação dos dados.

Ao comparar, na Tabela 4, o modelo, proposto neste trabalho, com estudos anteriores realizados no mesmo conjunto de dados (*Water Quality*), verificou-se que o modelo XGBoost, com o pré-processamento utilizado neste trabalho, apresenta um bom desempenho de classificação.

Figura 8. Curva ROC – XGBoost



Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Tabela 4. Acurácias de trabalhos para detecção da qualidade da água

Modelo	Acurácia	Referência
RF	70%	Poudel et al., 2022
ANN	69%	Poudel et al., 2022
SVM	55%	Kotawale et al., 2024
RF	70%	Kotawale et al., 2024
RF	81%	Laya; Shetty., 2024
KNN	58%	Brian et al., 2025
XGBoost	69%	Ryantoko et al., 2021
XGBoost	75%	Krushna; Sasi., 2024
XGBoost	92%	Esse Estudo

Fonte: SANTOS J. A. A. e POLO A. N. B. (2024)

Conclusões

Neste estudo, foi realizada a previsão, da qualidade da água, utilizando o modelo XGBoost em conjunto com a técnica *resample* da biblioteca Scikit-Learn. Os resultados, obtidos neste trabalho, demonstraram um desempenho robusto do modelo, atingindo uma acurácia de 92%, com métricas, para ambas as classes, equilibradas entre precisão, sensibilidade e *f1-score*.

Especificamente, para a classe Não Potável (Classe 0), o modelo obteve uma precisão de 90%, uma sensibilidade (*recall*) de 93% e um *f1-score* de 92%, indicando que a taxa de falsos negativos foi minimizada, reduzindo os riscos de consumo de água contaminada. Já para a classe Potável (Classe 1), os valores foram ainda mais expressivos, com precisão de 94%, sensibilidade (*recall*) de 91% e *f1-score* de 92%, garantindo uma alta confiabilidade na classificação da água segura para consumo.

A utilização da técnica *resample* mostrou-se eficiente, no balanceamento do conjunto de dados, permitindo que o modelo aprendesse de maneira equitativa entre as classes e evitasse vieses na classificação. Dessa forma, o modelo XGBoost se mostrou uma ferramenta promissora para a previsão da qualidade da água, contribuindo, desde que baseado em dados históricos de amostras coletadas em pontos representativos da realidade estudada, para sistemas de monitoramento e auxiliando na tomada de decisão em políticas de gestão hídrica e segurança sanitária.

Para trabalhos futuros, recomenda-se a aplicação de técnicas explicativas, como SHAP, para compreender melhor a influência das variáveis na previsão, além da comparação com outros algoritmos, como redes neurais artificiais e modelos híbridos, visando o aprimoramento contínuo do desempenho do sistema preditivo.

Referências

ABRAHAM, A.; LIVINGSTON, D.; GUERRA, I.; YANG, J. Exploring the application of machine learning algorithms to water quality analysis. *In: IEEE. 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science. Anais..., Danang: BCD, 2022.*

ABUZIR, S. Y.; ABUZIR, Y. S. Machine learning for water quality classification. **Water Quality Research Journal**, v. 35, n. 3, 2022.

BRIAN, T.; SHOLIKHAH, E. N.; MAULIDHIA, A. N. A.; WIBOWO, S. Application of K-Nearest Neighbor (KNN) Algorithm to Predict Drinking Water Quality. **Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem dan Komputer**, v.5, n. 1, 2025.

BUDHOLIYA, K.; KUMAR, S.; VIVEK, S. An optimized XGBoost based diagnostic system for effective prediction of heart disease. **Journal of King Saud University - Computer and Information Sciences Sharma**. v. 34, n. 7, 2022.

CHAFLOQUE, R. et al. Predictive neural networks model for detection of water quality for human consumption. *In: IEEE. 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN). Anais...*, Lima: CICN, 2021.

CHANG, V.; BAILEY, J.; XU, Q. A.; SUN, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. **Neural Computing and Applications**, v. 35, n. 1, 2023.

CHEN, T.; GUESTRIN, C. XGBoost: **A Scalable Tree Boosting System**. 2016. Disponível em: < <https://arxiv.org/pdf/1603.02754>>. Acesso em: 10 mai. 2024.

FEN, L.; LEI, Z.; TING, C. Study on potability water quality classification based on integrated learning. *In: IEEE. 2021 16th International Conference on Intelligent Systems and Knowledge Engineering. Anais...*, Chengdu: ISKE , 2021.

GUPTA, A.; AGRAWAL, A. An Analysis of Water Quality by Using Supervised Machine Learning Approaches. *In: 8th International Conference on Computing in Engineering and Technology. Anais...*, Patna: ICCET, 2023.

JUNA, A.; UMER, M. SADIQ, S.; KARAMTI, H.; ESHMAWI, A.; MOHAMED, A. ASHRAF, A. Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. **Water**, v. 14, n. 17, 2022.

KADIWAL, A. **Water Quality**. 2020. Disponível em: < <https://www.kaggle.com/datasets/adityakadiwal/water-potability>>. Acesso em: 10 jul. 2024.

KOTHAWALE, A. S.; SAPORETTI, C. M.; DESHPANDE, J. D. Performance Analysis of the portability of water resosurces through smote. **Journal of the Maharaja Sayajirao University of Baroda**. v. 58, n. 1, 2024.

KRUSHNA, V. K.; SASIKALA, D. Comparative Analysis of Machine Learning Models for Water Quality Prediction. *In: Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies. Anais...*, Bhilai: ICAECT, 2024.

LAYA, L.; SHETTY, J. S. Predicting Water Potability: Leveraging Machine Learning Techniques. *In: Second International Conference on Data Science and Information System. Hassan: ICDSIS. Anais...*, 2024.

LAI, J. P.; LIN, Y. L.; LIN, H. C.; SHIH, C. Y., WANG, Y. P. PAI, P. F. Tree-Based Machine Learning Models with Optuna in Predicting Impedance Values for Circuit Analysis. **Micromachines**, v. 14, n. 2, 2023.

MACEDO, B. S.; SAPORETTI, C. M. Uso de técnicas de aprendizado de máquina para predição do tempo de graduação dos discentes da Engenharia de Computação na região sudeste do Brasil. **RBCA**, v. 16, n. 1, 2024.

MAULANA, A.; FAISAL, F. F. NOVIANDY, T. R.; RIZKIA, T.; IDROES, G. M.; IDROES, G. M. TALLEI, T. E. Machine learning approach for diabetes detection using finite tuned XGBoost algorithm. **Infolitika Journal of Data Science**, v. 1, n. 1, 2023.

OSMAN, A. I. A.; AHMED, A. N.; CHOW, M. F. HUANG, Y. F.; EL-SHAFIE, A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. **Ain Shams Engineering Journal**. v. 12, n. 2, 2021.

PATEL, S.; SHAH, K. ; VAGHELA, S.; AGLODIYA, M.; BHATTAD, R. Water potability prediction using machine learning. **Research Square**, v. 1, n. 1, 2023.

POUDEL, D.; SHRESTHAA, D.; GHIMIREA, S. B. A. Comparison of machine learning algorithms in statistically imputed water potability dataset. **Journal of Innovation in Engineering Education**. v. 5, n. 1, 2022.

PREETHAM, M.; KRISHNAN, M. Improving Water Quality Prediction and Monitoring Using Machine Learning Algorithms. *In: 12th International Conference on System Modeling & Advancement in Research Trends*. **Anais...**, Moradabad: SMART, 2023.

SINAP, V. Comparative analysis of machine learning techniques for detecting potability of water. **Journal of Scientific Reports-A**. v. 58, n. 1, 2024.

RIYANTOKO, P. A.; FARRUDIN, T. M. HINDRAYAN, K. M. Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. **Seminar Nasional Sains**, v. 1, n. 1, 2021.

SOLANO, E. S.; DEHGHANIAN, P.; AFFONSO, C. M. Solar Radiation Forecasting Using Machine Learning and Ensemble Feature Selection. **Energies**, v. 15, n. 19, 2022.

YURTSEVER, M.; EMEÇ, M. Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability. **Ege Academic Review**, v. 23, n. 2, 2023.