Avaliação Comparativa de Modelos de Machine Learning para a Predição da Qualidade da Água.*

Jardel Ferreira dos Santos
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo - UNIFESP
São José dos Campos, Brasil
jardel.ferreira@unifesp.br

João Pedro Garlopa de Moraes
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo - UNIFESP
São José dos Campos, Brasil
joao.garlopa@unifesp.br

I. Introdução

Segundo a Organização das Nações Unidas (ONU) [1], em 2022, aproximadamente 2,2 bilhões de pessoas careciam de acesso a serviços de água potável geridos de forma segura. Este cenário tende a se agravar até 2030 devido a fatores como o crescimento populacional, a intensificação das mudanças climáticas e os conflitos armados. A garantia da água potável, no entanto, transcende a questão da sobrevivência: é um direito humano fundamental. Como resposta a este cenário, o Objetivo de Desenvolvimento Sustentável 6 (ODS 6) da ONU, intitulado 'Água Potável e Saneamento', estabelece a meta global de garantir o acesso universal e equitativo à água potável e ao saneamento até 2030.

A tecnologia tem um grande potencial para acelerar o desenvolvimento sustentável. Um relatório da União Internacional de Telecomunicações (UIT) junto com o PNUD mostra que as tecnologias digitais podem ajudar no progresso de cerca de 70% das metas dos Objetivos de Desenvolvimento Sustentável (ODS) [2]. Embora o relatório aborde o panorama de forma abrangente, o ODS 6 — que trata de "Água Potável e Saneamento" — destaca-se como um campo promissor para a aplicação de inovações tecnológicas. A gestão hídrica e do saneamento já vem sendo aprimorada por soluções como sensores baseados em Internet das Coisas (IoT), que permitem o monitoramento em tempo real de redes de distribuição, e ferramentas de análise de grandes volumes de dados (Big Data), utilizadas para prever a demanda e apoiar a tomada de decisões.

No Brasil, a Agência Nacional de Águas e Saneamento Básico (ANA) estabelece as normas técnicas para o monitoramento da qualidade dos corpos d'água. Dentre os diversos indicadores utilizados, o principal é o Índice de Qualidade da Água (IQA) [3].

O Índice de Qualidade da Água (IQA) foi concebido para avaliar a qualidade da água bruta, com foco no seu potencial uso para abastecimento público mediante tratamento. Sua composição baseia-se em nove parâmetros, cujos valores são ponderados por pesos distintos (w_i) que refletem a importância de cada um na qualidade geral da água, conforme detalhado na Tabela 1.

Além de seu peso (w_i), cada parâmetro possui um valor de qualidade (q_i), obtido em função de sua concentração ou medida.

TABLE I
Parâmetros e Pesos do Índice de Qualidade da Água (IQA)

Parâmetro	**Símbolo**	**Peso (w_i)**
Oxigênio Dissolvido	OD	0.17
Coliformes Termotolerantes	CT	0.15
Potencial Hidrogeniônico	pH	0.12
Demanda Bioquímica de Oxigênio	DBO	0.10
Temperatura da Água	T	0.10
Nitrogênio Total	NT	0.10
Fósforo Total	FT	0.10
Turbidez	Turb	0.08
Resíduo Total	RT	0.08

O cálculo do IQA é feito por meio do produtório ponderado dos nove parâmetros, segundo a fórmula apresentada na Equação 1:

$$IQA = \prod_{i=1}^{9} q_i^{w_i} \tag{1}$$

Onde:

- q_i: é a qualidade do i-ésimo parâmetro, um valor de 0 a 100 obtido a partir de curvas de qualidade específicas para cada parâmetro.
- w_i: é o peso correspondente ao i-ésimo parâmetro, um valor que varia de 0 a 1, tal que $\sum_{i=1}^{9} w_i = 1$.

O cálculo apresentado na Equação 1 resulta em um valor numérico único que sintetiza a condição do corpo d'água. A principal aplicação desse resultado é classificar a qualidade da água em categorias como 'Ótima' ou 'Péssima'. Contudo, é fundamental destacar uma particularidade do cenário brasileiro: embora o IQA seja um índice padronizado nacionalmente, as faixas de pontuação que definem cada categoria não são uniformes. Segundo a Agência Nacional de Águas e Saneamento Básico (ANA), diferentes estados adotam seus próprios 'scores' de classificação, o que significa que um mesmo valor de IQA pode corresponder a avaliações distintas dependendo da localidade [3]. Essa heterogeneidade reflete adaptações regionais, mas também impõe um desafio na comparação e gestão integrada dos dados em escala nacional.

Ainda segundo a ANA, esses parâmetros são, predominantemente, indicadores de contaminação por esgoto doméstico. Contudo, o IQA possui limitações significativas: o índice não mensura a presença de substâncias tóxicas, como metais pesados e pesticidas, nem de protozoários patogênicos e outros compostos que podem alterar as propriedades da água, sendo um indicador complementar e não absoluto.

É precisamente diante deste cenário de fragmentação e da necessidade de modernizar o monitoramento que as soluções computacionais surgem como aliadas estratégicas. A Inteligência Artificial (IA), em particular, tem sido amplamente empregada na resolução de problemas complexos. Nesse contexto, os modelos de Machine Learning (ML) se mostram especialmente promissores, pois podem ser treinados com dados históricos para 'aprender' os padrões que caracterizam a qualidade da água, contribuindo assim para a previsão de cenários futuros e para a criação de um método de classificação mais padronizado e robusto.

II. Objetivos

O trabalho tem, como objetivo principal, realizar uma análise comparativa da eficácia de modelos diversos de machine learning, no contexto de classificação de potabilidade de água. Para isso, será realizada uma análise explorátoria em um conjunto de dados públicos sobre qualidade de água e seus critérios. Os algoritmos de KNN, Naive Bayes, Decision Tree, SVM e MLP, então serão treinados para uma avaliação final de performance, que consistirá na comparação de acurácia e tempo de execução entre os modelos para o problema.

III. Trabalhos Relacionados

A literatura apresenta algumas alternativas para a predição da potabilidade de água que se aproveitam de aprendizado de máquina. O estudo de S. Ghoochani et al. [4] propõe uma mesma abordagem que a nossa: a análise de algoritmos diversos de Machine Learning, como SVM e Árvore de Decisão, no contexto de previsão de potabilidade. De forma semelhante, A. T. Ansari et al. [5] realizou um comparativo de desempenho entre os algoritmos KNN, Naive Bayes e SVM, para análise de qualidade de água de forma geral. Os trabalhos deles concluem que, de fato, é possível alcançar uma alta precisão com tais métodos, destacando a superioridade do SVM em relações não-lineares, o que valida essa tentativa de replicar os experimentos com mesma forma.

Além de modelos tradicionais, técnicas mais avançadas já foram empregadas. Y. Im et al. [6] demonstra uma abordagem alternativa aplicando redes neurais em um dataset temporal, prevendo a qualidade da água com Deep Learning e ilustrando seu todo potencial para solução de problemas semelhantes.

Por fim, os trabalhos de H. Wu et al. [7], focado na predição de qualidade do ar, e de Z. Zhang et al. [8], voltado para gestão de taxas de nutrientes em reservatórios para diminuição de poluição, demonstram a importância da otimização de hiperparâmetros em algoritmos de aprendizado de máquina, embora em problemas específicos diferentes do nosso.

IV. Materiais e Métodos

A. O Dataset

O dataset utilizado neste estudo, intitulado 'Water Quality and Potability', foi obtido da plataforma Kaggle. O conjunto de dados foi atualizado por Laksika Tharmalingam e consiste em diversas medições de qualidade da água, com o objetivo de classificar amostras como potáveis ou não potáveis. Cada entrada no dataset representa uma amostra de água, caracterizada por atributos como pH, Dureza, Sólidos, Cloraminas, Sulfato, Condutividade, Carbono Orgânico, Trihalometanos e Turbidez. A variável de interesse é a 'Potabilidade', uma variável binária que indica a adequação da água para consumo.

B. Análise exploratória dos dados

Na etapa inicial, foi feita uma análise exploratória para avaliar a estrutura do conjunto de dados, que totaliza 3.276 amostras. Conforme detalhado na Tabela II, que exibe as informações do DataFrame, observou-se a presença significativa de valores ausentes (nulos) em três colunas específicas. A coluna ph apresentava 491 valores nulos (15% do total), enquanto Sulfate continha 781 (24%) e Trihalomethanes, 162 (5%).

TABLE II
Descrição Estrutural e Tipos de Dados do Conjunto de Dados

#	Coluna	Valores Não Nulos	Tipo de Dado
0	ph	2785 de 3276	float64
1	Hardness	3276 de 3276	float64
2	Solids	3276 de 3276	float64
3	Chloramines	3276 de 3276	float64
4	Sulfate	2495 de 3276	float64
5	Conductivity	3276 de 3276	float64
6	Organic_carbon	3276 de 3276	float64
7	Trihalomethanes	3114 de 3276	float64
8	Turbidity	3276 de 3276	float64
9	Potability	3276 de 3276	int64

A análise da distribuição da variável-alvo, Potability, ilustrada na Figura 1, revelou um claro desbalanceamento entre as classes. O conjunto de dados mostrou-se composto por 61% de amostras classificadas como 'Não Potável' (0) e 39% como 'Potável' (1). Este desequilíbrio é um ponto de atenção, pois pode criar um viés nos modelos de Machine Learning, tornando essencial que a performance seja avaliada com métricas apropriadas para tal cenário.

Para verificar a relação linear entre as variáveis, utilizou-se a matriz de correlação, apresentada como um mapa de calor na Figura 2. O gráfico evidencia que quase todas as correlações são muito fracas ou próximas de zero. A baixa correlação entre as variáveis de entrada é, por um lado, um bom indicativo, pois aponta para a ausência de informações redundantes. Contudo, o ponto mais relevante foi a correlação quase nula entre as variáveis e a Potability. Tal fato sugeriu que o problema não é linear, reforçando que o uso de algoritmos de Machine Learning mais robustos seria benéfico.

A análise da distribuição das variáveis através de boxplots (Figura 3) foi fundamental para a identificação de outliers.

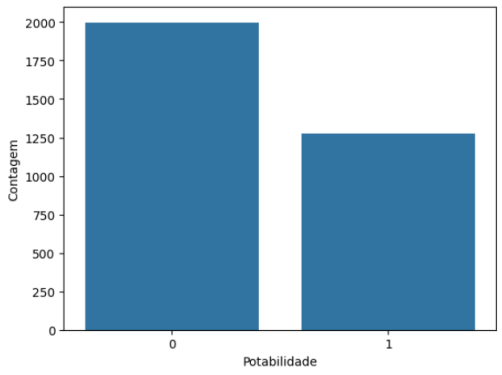

Fig. 1. Distribuição das classes da variável-alvo 'Potability'.

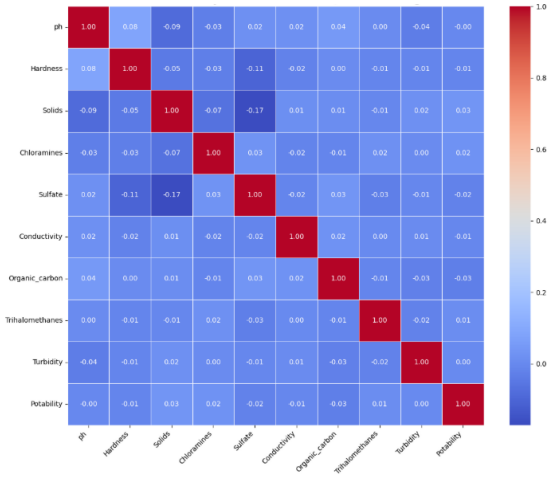

Fig. 2. Matriz de correlação das variáveis de qualidade da água.

Observou-se que variáveis como Solids, Chloramines e Sulfate possuem múltiplos valores atípicos. Este achado sugeriu a imputação de dados, e a necessidade de padronizar a escala dos dados antes de submetê-los aos algoritmos de Machine Learning.

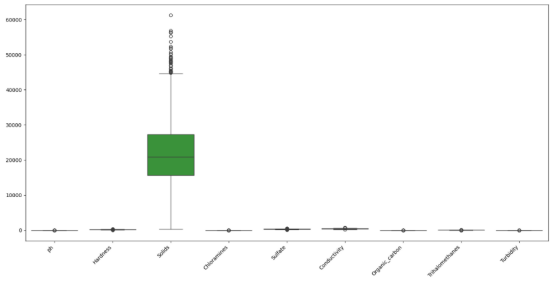

Fig. 3. boxplot de todas as variáveis do dataset.

Recorreu-se à Análise de Componentes Principais (PCA) como forma de visualizar a estrutura dos dados em duas dimensões (Figura 4). O gráfico resultante foi bastante claro: as amostras de água 'Potável' (1) e 'Não Potável' (0) encontravam-se completamente misturadas, sem uma separação evidente. Essa sobreposição serviu como a evidência visual mais forte de que o problema não é de natureza linear, o que validou a decisão de aplicar algoritmos de Machine Learning mais avançados.

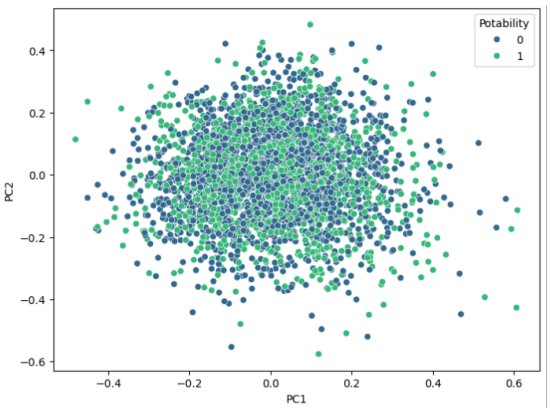

Fig. 4. PCA do dataset.

C. Pré-processamento dos Dados

Após a análise exploratória, a etapa seguinte consistiu no pré-processamento e preparação dos dados para a modelagem. O processo foi dividido em duas fases principais: a imputação de valores ausentes e a normalização das variáveis.

Para os valores nulos identificados em ph, Sulfate e Trihalomethanes, optou-se pelo KNNImputer. A escolha desse método se deu por sua capacidade de preencher os valores faltantes com base na média dos k vizinhos mais próximos, levando em consideração a estrutura dos dados para uma estimativa mais precisa e preservando a variância do dataset. Para este trabalho, foi configurado um número de 5 vizinhos (k=5).

Adicionalmente, como a análise de boxplots já havia revelado, as variáveis possuíam escalas muito distintas. A fim de evitar que os algoritmos fossem enviesados por variáveis de maior magnitude (como Solids), foi necessário normalizar os dados. A técnica aplicada foi o MinMaxScaler, que transforma cada variável para que seus valores se situem em um intervalo comum, entre 0 e 1, passo essencial para muitos algoritmos.

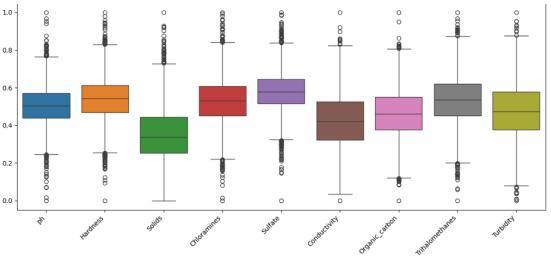

Fig. 5. boxplot após a normalização e imputação dos dados.

O resultado dessas etapas de tratamento pode ser visualizado no novo conjunto de boxplots da Figura 5. Com todas as variáveis agora contidas na mesma escala (de 0 a 1), a comparação direta de suas distribuições tornou-se possível.

Um ponto fundamental é que, embora a escala tenha sido ajustada, a estrutura de distribuição e a presença de outliers foram preservadas, garantindo que a informação original dos dados não fosse distorcida. O dataset encontra-se, então, limpo, completo e normalizado, estando finalmente pronto para a etapa de treinamento e avaliação dos modelos.

V. Algoritmos utilizados

Para o estudo, foi empregado um coletânea de algoritmos de classificação, de estátisticos clássicos até redes neurais

A. Naive Bayes

Algoritmo baseado primariamente no Teorema de Bayes. Parte do presuposto que cada característica físico-química (pH, dureza, etc.) contribui de forma independente para a decisão de potabilidade da água.

B. K-Nearest Neighbors

Se compara um nova amostra com as K amostras mais similares no conjunto de treino, e a classifica com base na classe majoritária desses "vizinhos" mais próximos. O algoritmo acaba sendo sensível à escala do dados.

C. Árvore de Decisão

Cria um fluxograma de perguntas sobre os atributos da água e contrói uma árvore com base nas respostas obtidas, obtendo um veredito final sobra a potabilidade da amostra. Tem a vantagem de não ser uma "caixa preta" de resultados, no sentido que armazena uma linha racional para a escolha da classe da amostra.

D. SVM

O Support Vector Machine (SVM) tenta encontrar uma função de separação entre os dados de água potável e os não potável com a maior margem possível. Para problemas que não é possivel separar linearmente, se utiliza truque de kernel para mapear os dados em dimensões maiores, onde essa separação é mais viável.

E. Multi-Layer Perceptron (MLP)

Uma coleção de camadas de neurônios artificiais que tenta traçar uma linha para separação das duas classe de água, por meio de uma série de transformações algébricas. É mais complexa que um simples neurônio articial, sendo capaz de entender padrões não-lineares dos dados.

VI. Análise experimental

A. Configuração dos algoritmos e do ambiente computacional

Todo experimento foi realizado dentro do ambiente computacional virtual grátis disponibilizado pelo Google Colab, com as seguintes configurações padrões de hardware: CPU Intel Xeon com 2 vCPUs com 13 GB de RAM. O projeto foi totalmente desenvolvido em Python 3 e dependeu de bibliotecas open-source de análise de dados e inteligência artificial, com as principais sendo pandas, seaborn, matplotlib, sklearn, imblearn, numpy, kagglehub, scipy e pyswarms.

B. Critérios de análise

Para garantir um comparativo justo entre os algoritmos, cada algoritmo de classificação foi executada 30 vezes, variando de 0 a 29 a semente aleatorio da função train test split. Tal abordagem foi utilizada para diminuir a variabilidade e viés que poderia ter ao realizar o teste em apenas uma única semente. No final das execuções, foi calculado a média e desvio padrão da acurácia dos resultados e então extraído o melhor parâmetro geral do algoritmo, a fim de se obter uma estimativa mais confiável de perfomance.

A métrica de acurácia foi utilizada com a principal medida de credibilidade dos algoritmos, mas, adicionalmente, foi calculado a matriz de confusão média dos modelos, para uma análise mais detalhada dos classificadores. O tempo total de execução para o ciclo completo com 30 iterações também foi registrado para análise de eficiência. Para algoritmos dependentes de hiperparâmetros, a moda dos melhores valores encontrados foi considerada com a mais consistente.

C. Resultados e discussões

Os dados obtidos durante o experimento foram ordenados conforme acurácia média. Para cada algoritmo, foi registrado seu desvio padrão, melhor hiperparâmetro modal e tempo de execução em segundos. Os valores estão apresentados na Tabela:

TABLE III
Desempenho Comparativo dos Algoritmos (DP: Desvio Padrão).

Alg.	Parâmetro (Moda)	Acurácia Média ± DP	Tempo (s)
MLP	hidden_layer = (50,)	0,681 ± 0,016	623.26
SVM	C = 1.00	0,675 ± 0,014	97.12
KNN	k = 27	0,648 ± 0,015	20.17
DT	max_depth = 6	0,648 ± 0,015	46.07
NB	var_smooth = 1.00	0,632 ± 0,014	13.03

Ao observar a tabela, nota-se que modelo de MLP obteve o melhor desempenho, com acurácia média 0,681, seguido de perto pelo SVM com 0,675, destacando que, para esse tipo de problema, conjunto de dados e pré-processamento abordado, um modelo mais complexos é mais sólido que modelos mais simples como o KNN e Naive Bayes. Este comportamento está de acordo com o teorema de No Free Lunch, que afirma que não existe um único algoritmo superior para todos problemas de machine learning.

Nota-se, também, os valores de desvio padrão foram relativamente baixos (entre 0,014 e 0,016), indicando uma estabilidade entre algoritmos, no que se diz a respeito da forma como os dados de treino e teste foram particionados. Por exemplo, para o MLP, a acurácia média foi de 0,681 com incerteza de $\pm$0,016.

Para uma análise aprofundada, forma geradas as matrizes de confusão média dos algoritmos, o que permite visualizar algumas nuances do comportamento de cada modelo de classificador. No contexto de potabilidade de água, um falso positivo representa o erro mais crítico. O MLP foi o algoritmo que obteve o maio número de acertos de verdadeiros positivos, com

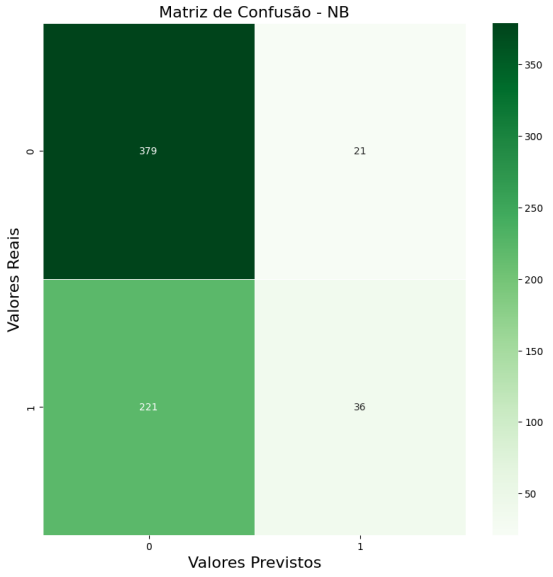

Fig. 6. matriz de confusão do Naive Bayes.

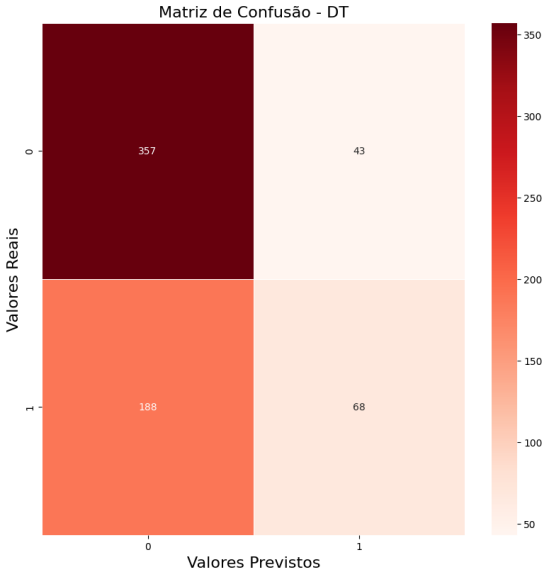

Fig. 8. matriz de confusão da Árvore de Decisão.

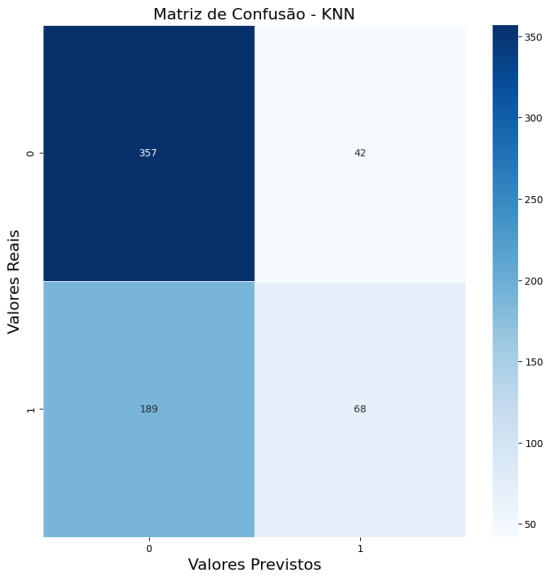

Fig. 7. matriz de confusão do KNN.

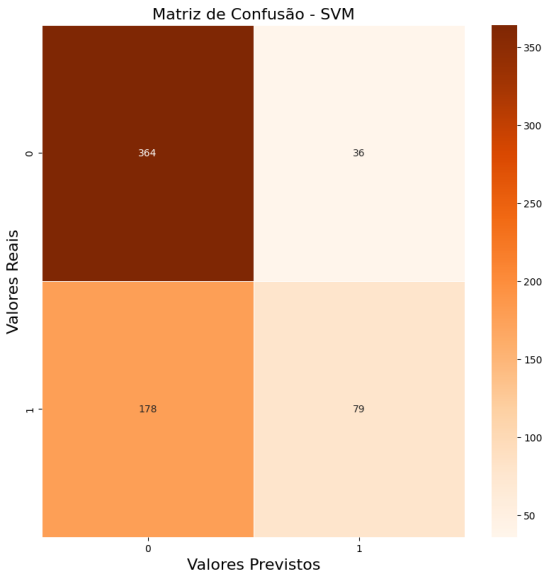

Fig. 9. matriz de confusão do SVM.

média de 90 acertos, sendo o mais confiável para identificar uma água segura para consumo humano. O KNN e a Árvore de decisão pontuaram em média, o mesmo número de verdadeiros negativos e falsos positivos que o MLP, respectivamente 357 e 42, mas com um desempenho inferior no quesito de verdadeiros positivos. O Naive Bayes apresentou o menor número de falsos positivos, com média de 21 erros, que levaria a uma menor utilização de uma água imprópria, porém também obteve o maior número de falsos negativos, 221 erros, que na prática levaria a um descarte de água maior que a dos outros algoritmos.

Outro aspecto observado é a relação entre a performance e tempo de execução. O Naive Bayes foi o algoritmo mais rápido, porém com acurácia média baixa, se tornando pobre para o problema. De outro lado, o MLP e o SVM, embora com as melhores acurácias médias, exigiram um tempo maior para execução. O KNN e a Árvore de Decisão representaram um bom equilíbrio, entregando tempos de execução relativamente baixos com acurácias intermediárias.

Durante a análise, foram notadas algumas limitações que podem estar ligadas a fatores como: sensibilidade aos hiperparâmetros e natureza dos dados. A busca pelos hiperparâmetros foi limitada a uma faixa de valores pré-definida, não considerando potenciais valores ótimos. Além disso, pelo modelo de Árvore de Decisão foi identificado as características de Sulfate e ph como as mais importantes para a análise,

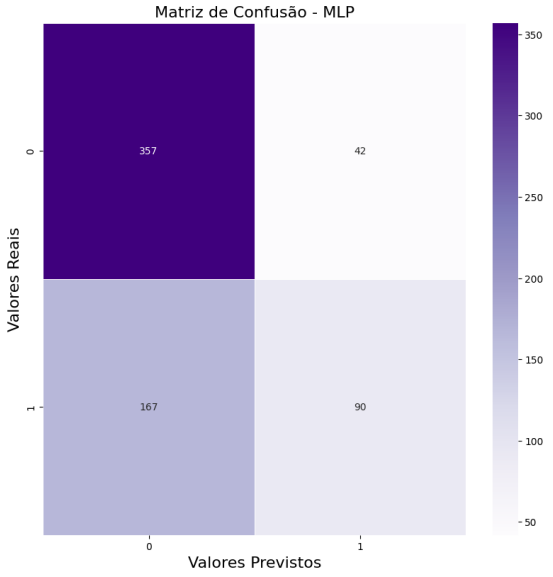

Fig. 10. matriz de confusão do MLP.

indicando que a qualidade da imputação de valores ausentes nessas colunas pode gerar ruído aos algoritmos.

VII. Conclusões

O experimento permitiu comprovar, na prática, a eficiência de diferentes algoritmos de machine learning para o problema de determinação de potabilidade da água a partir de um conjunto de dados públicos. Observou-se que o Multi-Layer Perceptron (MLP) alcançou a maior acurácia média, seguido do pelo Support Vector Machine (SVM). O resultado consolida a capacidade desses modelos de lidar com dados não-lineares. Foi possível, também, notar uma clara relação entre acurácia e tempo de execução, onde os modelos com melhores resultados foram o mais lentos, enquanto os mais rápidos obtiveram os piores.

Para trabalhos futuros, seria interessante aprofundar a otimização dos hiperparâmetros dos modelos de SVM e MLP, por exemplo, com técnicas mais robustas, como algoritmos genéticos ou bioinspirados, podendo extrair resultados ainda melhores. Testes com variações nos métodos de imputação de dados também poderia enriquecer a análise de resultados.

Compreender a aplicação desses modelos é essencial para estudantes das áreas de Computação e Ciência de Dados, já que aprendizado de máquina está presente em inúmeros sistemas industriais, entre eles em sistemas de automação de análise de potabilidade de água. O conhecimento adquirido fornece uma base sólida para o desenvolvimento de dispositivos inteligentes de apoio à decisão, contribuindo para a formação de profissionais preparados para lidar com tecnologias que impactam diretamente o bem-estar social. A realização deste experimento, portanto, não apenas comparou o desempenho de algoritmos, mas também fortaleceu a conexão entre a teoria de inteligência artificial e os desafios de projetos com impacto social.

References

[1] NAÇÕES UNIDAS (ONU). *The Sustainable Development Goals Report 2024*. Nova York: Nações Unidas, 2024. Disponível em: https://unstats.un.org/sdgs/report/2024/. Acesso em: 12 jul. 2025.

[2] PROGRAMA DAS NAÇÕES UNIDAS PARA O DESENVOLVIMENTO (PNUD). "Tecnologias digitais beneficiam diretamente 70% das metas dos ODS, afirmam UIT, PNUD e parceiros". *PNUD Brasil*, 14 set. 2023. Disponível em: https://www.undp.org/pt/brazil/news/tecnologias-digitals-beneficiam-diretamente-70-das-metas-dos-ods-afirmam-uit-pnud-e-parceiros. Acesso em: 12 jul. 2025.

[3] BRASIL. Agência Nacional de Águas e Saneamento Básico (ANA). "Índice de Qualidade das Águas (IQA)". Brasília, DF: ANA. Disponível em: https://www.ana.gov.br/portalpnqa/indicadores-indice-aguas.aspx. Acesso em: 12 jul. 2025.

[4] S. Ghoochani, M. Khorram e N. Nazemi, "Uncovering Top-Tier Machine Learning Classifier for Drinking Water Quality Detection" *Preprints*, 2023, 2023081413. DOI: 10.20944/preprints202308.1413.v1.

[5] A. T. Ansari, N. Nigar, H. M. Faisal e M. K. Shahzad, "AI for clean water: efficient water quality prediction leveraging machine learning" *Water Practice & Technology*, vol. 19, no. 5, pp. 1986-2004, 2024. DOI: 10.2166/wpt.2024.123.

[6] Y. Im, G. Song, J. Lee e M. Cho, "Deep Learning Methods for Predicting Tap-Water Quality Time Series in South Korea," *Water*, vol. 14, no. 22, p. 3766, 2022. DOI: 10.3390/w14223766.

[7] H. Wu, T. Yang, H. Wu, H. Li e Z. Zhou, "Air quality prediction based on Long Short-Term Memory Model with advanced feature selection and hyperparameter optimization" *Water*, vol. 14, no. 22, p. 3766, 2022. DOI:10.3233/JIFS-232308.

[8] Z. Zhang, C. Yang, Q. Qiao, X. Li, F. Wang e C. Li, "Application of Improved Particle Swarm Optimization SVM in Water Quality Evaluation of Ming Cui Lake" *Water*, vol. 14, no. 22, p. 3766, 2022. DOI:10.3390/su15129835.