Predição da Dosagem de Coagulantes em uma Estação de Tratamento de Água por Ferramentas de *Machine Learning*

Sérgio Leonardo Butski Soares Santos, Natan Padoin, Cíntia Soares*

Universidade Federal de Santa Catarina (UFSC), Centro Tecnológico (CTC), Departamento de Engenharia Química e Engenharia de Alimentos (EQA), Programa de Pós-Graduação em Engenharia Química (PósENQ), Laboratório de Materiais e Computação Científica (LabMAC), Florianópolis – SC, Brasil
**cintia.soares@ufsc.br*

RESUMO

Este trabalho investigou a aplicação de diferentes técnicas de *Machine Learning* para a predição da dosagem de coagulante em uma Estação de Tratamento de Água (ETA). Modelos lineares e não lineares, combinados com distintas estratégias de imputação de dados e *feature engineering*, foram avaliados. Os resultados mostraram que o *Extremely Randomized Trees* (ERTrees) sem preenchimento apresentou o maior coeficiente de determinação (R^2 = 0,9337) e os menores erros MSE, RMSE e MAE, superando todas as técnicas de imputação testadas. Apenas na métrica MAPE o método *Miss Forest* (n = 10) obteve desempenho superior. A análise com SHAP reforçou a importância das variáveis derivadas, que tiveram papel decisivo na qualidade das previsões. Apesar das limitações relacionadas à generalização e ao custo computacional, os achados indicam que a integração de modelos preditivos com sistemas de apoio à decisão pode contribuir para maior eficiência operacional e transparência na Engenharia de Sistemas em Processos.

Palavras-chave: aprendizado de máquina; coagulante; estação de tratamento de água; *feature engineering*; SHAP.

1. Introdução

A dosagem adequada de coagulantes em processos de tratamento de água e efluentes é um fator crítico para garantir eficiência operacional, redução de custos e conformidade ambiental. No entanto, a determinação precisa dessa dosagem envolve variáveis complexas e interdependentes, o que torna desafiadora a aplicação de métodos tradicionais de cálculo. Nesse contexto, técnicas de aprendizado de máquina (*Machine Learning* - ML) têm se destacado como ferramentas para modelagem e predição em sistemas de processos, permitindo maior robustez e adaptabilidade frente à variabilidade dos dados (Nasir *et al*., 2022; Ngwenya *et al*., 2025; Nwankwo *et al*., 2024).

O presente trabalho tem como propósito avaliar a aplicação de diferentes algoritmos de *Machine Learning* na predição da dosagem de coagulante, utilizando como estratégia de seleção de atributos a correlação de Pearson. Além disso, diferentes técnicas de imputação de dados foram adotadas e diversas transformações de *Feature Engineering* aplicadas, de modo a explorar relações relevantes entre variáveis e aprimorar o desempenho dos modelos. A abordagem proposta busca oferecer uma análise comparativa entre 14 algoritmos de ML, destacando suas potencialidades e limitações no contexto da engenharia de sistemas em processos (Feng *et al*., 2025; Kim et al., 2024; Ooko *et al*., 2025; Zhao *et al*., 2025).

A motivação central deste estudo é demonstrar a viabilidade de uma metodologia simplificada, que pode servir como base para trabalhos futuros mais abrangentes e detalhados. Dessa forma, pretende-se contribuir para a comunidade de Engenharia de Sistemas em Processos, apresentando resultados iniciais que reforçam o papel do aprendizado de máquina como ferramenta de apoio à tomada de decisão em operações industriais.

2. Materiais e Métodos

2.1. Estação de tratamento de água e coleta de dados

O estudo foi realizado na ETA José Pedro Horstmann, localizada em Palhoça - SC, responsável pelo abastecimento da região metropolitana de Florianópolis - SC. A estação opera em ciclo completo e possui vazão média de 2500 L/s. Entre dezembro de 2019 e abril de 2025, 23144 registros foram coletados, abrangendo parâmetros físico-químicos e operacionais monitorados a cada 2 h, incluindo turbidez, cor, pH, vazão e dosagem de coagulante.

2.2. Preparação dos dados

O pré-processamento dos dados constituiu uma etapa essencial para garantir a qualidade das informações utilizadas na modelagem preditiva.

Tabela 1. Variáveis derivadas e justificativas físico-químicas.

Nome da Relação	Equação	Justificativa
Lags temporais	$Lag_1(n)=PAC_{n-1}$ $Lag_2(n)=PAC_{n-2}$ $Lag_4(n)=PAC_{n-4}$	Considera a inércia operacional e ajustes graduais na dosagem, refletindo decisões passadas que influenciam a próxima.
Mudança de dosagem	$\begin{cases}1, \text{ se } \lvert PAC_{n-1}-PAC_{n-2}\rvert > 2{,}0\\ 0, \text{ caso contrário}\end{cases}$	Detecta alterações abruptas na dosagem, sinalizando eventos relevantes para o processo.
Mudanças de janela	$\begin{cases}1 \text{ em } (n-1, n \text{ e } n+1),\\ \text{se Mudança de dosagem} = 1\\ 0, \text{ caso contrário}\end{cases}$	Suaviza e consolida alertas, evitando ruídos isolados e reforçando a sensibilidade do sistema.
Eficiência do tratamento (cor)	$1-\frac{Cor(AT)}{Cor(AB)}$	Mede a porcentagem de remoção de cor ao longo do processo.
Eficiência do tratamento (turbidez)	$1-\frac{Turbidez(AT)}{Turbidez(AB)}$	Avalia a eficiência na redução da turbidez.
Diferença absoluta de cor	ΔCor_{AB-AD}, ΔCor_{AD-AT}, ΔCor_{AB-AT}	Quantifica a redução de cor em cada fase do tratamento.
Diferença relativa de cor	$\Delta Cor\%_{AB-AT}=\frac{Cor_{(AB)}- Cor_{(AT)}}{Cor_{(AB)}}$	Expressa a redução relativa de cor entre entrada e saída.
Razão cor Cubatão/Pilões	$\frac{Cor_{(Cubatão-AB)}}{Cor_{(Pilões-AB)}}$	Compara a contribuição de cada rio para a carga de cor.
Diferença cor Cubatão–Pilões	$Cor_{(Cubatão-AB)}-Cor_{(Pilões-AB)}$	Identifica qual rio apresenta maior impacto na qualidade da água bruta.
Fonte predominante	Codificação: 0 = Cubatão, 1 = Pilões	Classifica a fonte hídrica com maior influência na cor da água bruta.
Razão cor entrada/saída	$\frac{Cor_{(AB)}}{Cor_{(AT)}}, \frac{Cor_{(AD)}}{Cor_{(AT)}}$	Avalia a eficiência relativa do tratamento em diferentes etapas.
Média da cor das fontes	$\frac{Cor_{(Cubatão-AB)}+Cor_{(Pilões-AB)}}{2}$	Representa a média da qualidade das águas brutas dos dois rios.
Índice combinado cor × turbidez	$Cor_{(AT)}\times Turbidez_{(AT)}$	Indica a qualidade final da água tratada.
Pico de cor (AB)	1 se $Cor_{(AB)}>Q_{90}$; 0 caso contrário	Identifica condições extremas de cor na entrada.
Pico de turbidez (AT)	1 se $Turbidez_{(AT)}>Q_{90}$; 0 caso contrário	Detecta eventos críticos de turbidez na saída.

Primeiramente, a seleção de atributos foi realizada por meio da correlação de Pearson com o objetivo de identificar as variáveis mais relevantes para a predição da dosagem de coagulante. Para reduzir ruídos e assegurar maior robustez estatística, procedeu-se à remoção de *outliers* com o auxílio da ferramenta *Facebook Prophet*, eliminando valores discrepantes não representativos da operação real da ETA. Na sequência, diferentes técnicas de imputação de dados ausentes baseada em interpolação foram aplicadas (23 diferentes imputações), a fim de corrigir falhas de medição e inconsistências no conjunto. A etapa de *Feature Engineering* possibilitou a criação de variáveis derivadas de relações físico-químicas e operacionais (Tabela 1), enriquecendo o conjunto de dados e ampliando sua capacidade explicativa. Por fim, todas as variáveis foram submetidas à padronização com o *StandardScaler*, garantindo uniformidade de escala e integrando esse processo à busca em grade (*GridSearchCV*) para otimização dos modelos de *Machine Learning*.

2.3. Modelagem preditiva e avaliação dos modelos

Um *pipeline* de regressão integrado a uma estratégia de validação cruzada temporal (*TimeSeriesSplit*) com cinco divisões foi desenvolvido, garantindo que a ordem cronológica dos dados fosse preservada e que os modelos fossem avaliados em condições próximas à realidade operacional da ETA. A etapa de otimização de hiperparâmetros foi conduzida por meio do *GridSearchCV*, utilizando como métrica principal o coeficiente de determinação (R^2), de modo a selecionar configurações que maximizassem a capacidade explicativa dos modelos.

Quatorze algoritmos de *Machine Learning* foram utilizados, abrangendo desde métodos lineares clássicos (Regressão Linear e *Stochastic Gradient Descent* (SGD)), passando por *Support Vector Machine* (SVM) e métodos baseados em vizinhança (*K-Nearest Neighbors* (KNN)), até modelos de árvores de decisão e *ensembles* (*Decision Tree, Random Forest, Extremely Randomized Trees* (ERTrees), *Light Gradient Boosting Machine* (LightGBM), *eXtreme Gradient Boosting* (XGBoost) e *Categorical Boosting* (CatBoost)), além de arquiteturas mais complexas de redes neurais (*Multilayer Perceptron* (MLP), *Convolutional Neural Network* (CNN), CNN - *Long Short-Term Memory* (CNN-LSTM) e CNN - *Recurrent Neural Network* (CNN-RNN)). Essa diversidade metodológica permitiu avaliar o impacto de diferentes abordagens sobre a acurácia das previsões e a capacidade de generalização, oferecendo uma visão abrangente do potencial das técnicas de aprendizado de máquina no contexto da engenharia de sistemas em processos.

Os modelos foram avaliados em um conjunto de teste correspondente a 20 % dos dados, respeitando a sequência temporal das observações. Para a análise de desempenho, métricas complementares foram utilizadas: MSE (*Mean Squared Error*), RMSE (*Root Mean Squared Error*), MAE (*Mean Absolute Error*), MAPE (*Mean Absolute Percentage Error*) e R^2 (Coeficiente de Determinação). Essas medidas possibilitaram avaliar tanto a magnitude absoluta e relativa dos erros, quanto a capacidade explicativa dos modelos, permitindo identificar aqueles com maior precisão na predição da dosagem de coagulante e evidenciando as compensações entre complexidade computacional e desempenho preditivo.

3. Resultados

Os resultados obtidos a partir da aplicação de 14 diferentes técnicas de *Machine Learning* foram organizados de forma a destacar os modelos com melhor desempenho na predição da dosagem de coagulante (Tabela 2).

Tabela 2. Comparação dos principais resultados obtidos pelos modelos de ML.

Modelo	Preenchimento	MSE	RMSE	MAE	MAPE	R^2	N.º de Dados
ERTrees	Sem preenchimento	1,2446	1,1156	0,6355	0,0403	0,9337	19549
ERTrees	Média	1,2902	1,1359	0,6486	0,0404	0,9333	20914
ERTrees	Mais Frequente	1,2929	1,1371	0,6428	0,0400	0,9332	20914
ERTrees	Vizinho Próximo	1,2932	1,1372	0,6436	0,0401	0,9332	20914
ERTrees	Akima	1,2965	1,1387	0,6412	0,0399	0,9330	20914
ERTrees	Miss Forest (n = 10)	1,2973	1,1390	0,6412	0,0398	0,9330	20914
ERTrees	Ridge Bayesiana	1,2981	1,1393	0,6447	0,0401	0,9329	20914
ERTrees	Linear	1,2990	1,1398	0,6433	0,0401	0,9329	20914
ERTrees	Pad	1,2997	1,1400	0,6483	0,0404	0,9329	20914
ERTrees	Mediana	1,2997	1,1401	0,6502	0,0406	0,9329	20914

Os dez melhores resultados, ordenados pelo coeficiente de determinação, foram inseridos na Tabela 2 com diferentes combinações de algoritmos e técnicas de imputação de dados. Essa tabela contempla métricas complementares às análises de MSE, RMSE, MAE, MAPE e R^2, que permitem avaliar tanto a magnitude dos erros quanto a capacidade explicativa dos modelos. Analisando-se os valores da Tabela 2, evidencia-se que os dez melhores resultados com base nos valores de R^2 correspondem ao modelo *Extremely Randomized Trees* (*ERTrees*), variando-se apenas o método de preenchimento dos valores ausentes. Isso indica que, para o conjunto de dados da ETA, o *ERTrees* mostrou-se particularmente robusto, superando consistentemente outras arquiteturas, como *LightGBM*, *Random Forest*, regressão linear, SVM e Redes Neurais, que não figuraram entre as dez primeiras posições.

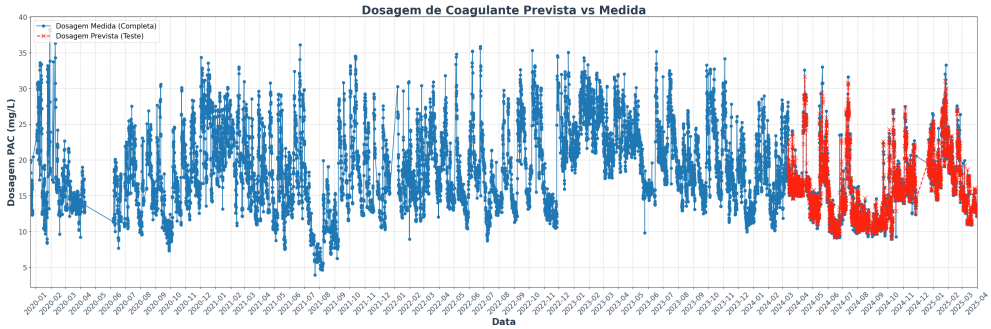
Entre as abordagens com *ERTrees*, a estratégia sem preenchimento (remoção das linhas com *Prophet*) obteve o melhor desempenho geral: maior R^2 (0,9337) e menores valores de MSE (1,2446), RMSE (1,1156) e MAE (0,6355). A única métrica em que essa estratégia não foi a melhor foi a MAPE, onde o menor valor (0,0398) foi alcançado pelo método *Miss Forest* (n = 10), enquanto o modelo sem preenchimento apresentou MAPE de 0,0403. Essa diferença, contudo, é muito pequena, e o conjunto das demais métricas favorece claramente a remoção dos dados ausentes. Cabe destacar que a abordagem sem preenchimento utilizou um número menor de observações (19549) em comparação com os métodos de imputação (20914). Em princípio, uma base de teste menor poderia influenciar as métricas de erro absoluto, mas o fato de R^2, que é uma medida normalizada, também ser superior, indica que a melhoria não decorre apenas da redução amostral, e sim da maior qualidade e consistência dos dados remanescentes. Aparentemente, as linhas removidas continham padrões inconsistentes ou ruidosos, cuja imputação não foi capaz de recuperar adequadamente.

Dentre as estratégias de imputação de dados, observa-se uma competição acirrada entre métodos simples e complexos. O método da Média apresentou o melhor desempenho em termos de variância explicada (R^2 de 0,9304) e erro quadrático (MSE de 1,2902), superando modelos mais sofisticados, como a *Ridge Bayesiana* e o *Miss Forest*. No entanto, a técnica *Miss Forest* (n = 10) destacou-se por registrar o menor erro percentual absoluto (MAPE de 0,0398), sendo a única abordagem a superar o cenário sem preenchimento nesta métrica específica. Esse comportamento sugere que, embora a imputação pela média preserve melhor a estrutura global da distribuição para o algoritmo *ERTrees*, modelos baseados em aprendizado de máquina (como o *Miss Forest*) podem ser mais eficazes na redução de erros relativos em pontos específicos da amostra.

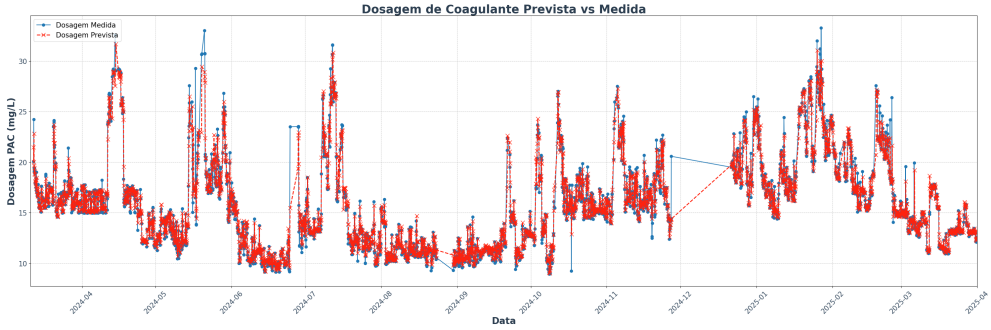
Diante desses resultados, o modelo *ERTrees* sem preenchimento foi selecionado para análise detalhada. A Figura 1 apresenta a comparação entre a dosagem de coagulante prevista por esse modelo e os valores reais.

Figura 1. Dosagem de coagulante prevista vs. medida: (a) série temporal completa; e (b) período de teste.

(a)



(b)

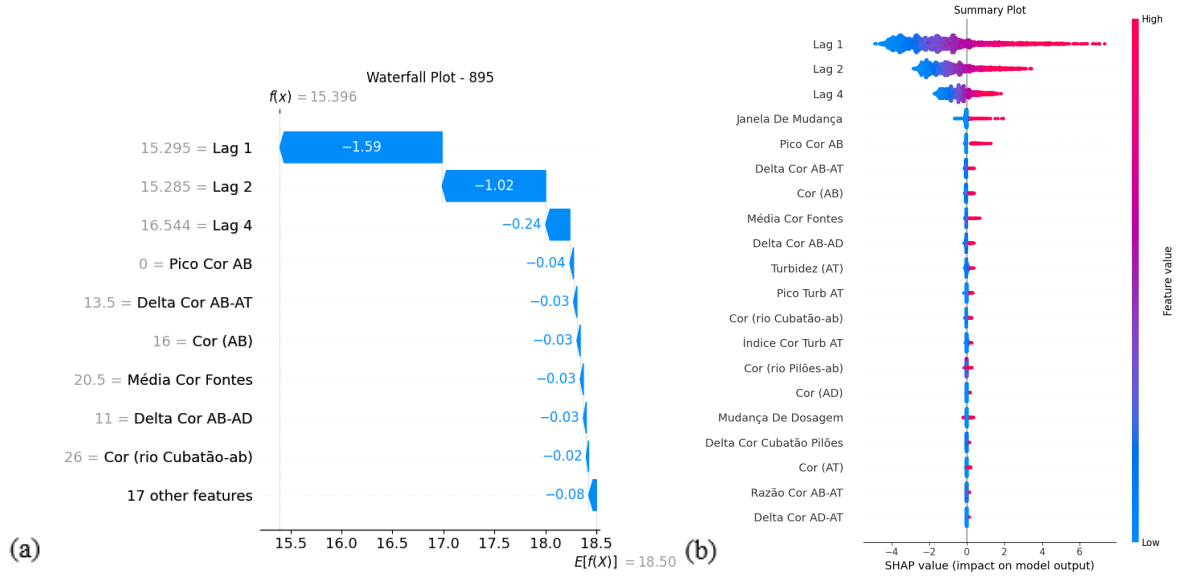


A Figura 1a mostra a série temporal completa, permitindo observar a aderência do modelo às variações de longo prazo, enquanto a Figura 1b destaca apenas o período de teste, facilitando a análise detalhada da capacidade preditiva em condições recentes. Essa dupla visualização possibilita avaliar tanto a robustez geral do modelo quanto sua precisão em intervalos específicos de operação.

A Figura 1 evidencia que o modelo *ERTrees* sem preenchimento conseguiu reproduzir de forma consistente as flutuações da dosagem de PAC, mantendo forte aderência às tendências gerais e às oscilações de curto prazo. Nota-se que, nos períodos de maior variação, as previsões acompanham de perto os valores medidos, ainda que pequenas discrepâncias sejam observadas em picos mais abruptos, o que é esperado em sistemas complexos e sujeitos a fatores externos, como chuvas intensas ou alterações súbitas na qualidade da água bruta. A concordância entre as curvas reforça a robustez do *Extremely Randomized Trees* na configuração sem imputação de dados, demonstrando que essa abordagem não apenas explica grande parte da variabilidade dos dados ($R^2 = 0{,}9337$), mas também oferece previsões confiáveis para aplicação prática na operação da ETA.

A interpretação dos modelos de *Machine Learning* é um aspecto essencial para garantir transparência e confiabilidade em aplicações práticas. Nesse sentido, os gráficos de SHAP (*SHapley Additive exPlanations*), que permitem avaliar a contribuição individual de cada variável para as previsões do modelo, foram utilizados. A Figura 2 apresenta essa análise aplicada ao *ERTrees* sem preenchimento, destacando em (a) as contribuições locais das variáveis em uma instância específica e, em (b), a importância global das variáveis ao longo de todo o conjunto de dados. Essa abordagem possibilita compreender não apenas como o modelo chega a uma determinada previsão, mas também quais fatores exercem maior influência de forma consistente, conectando os resultados estatísticos à lógica físico-química do processo de tratamento de água.

Figura 2. Interpretação do modelo *ERTrees* com SHAP: (a) contribuições das variáveis em instância específica; e (b) importância global das variáveis.



A Figura 2a revela que, para a instância analisada (linha 895 do teste, com valor real de 15,34 mg/L e previsto de 15,40 mg/L), todas as variáveis com contribuição não desprezível atuaram no sentido de reduzir a predição em relação ao valor base do modelo ($E[f(x)] \approx 18{,}50$). As três variáveis defasadas (Lag 1, Lag 2 e Lag 4) foram as que mais contribuíram negativamente, com reduções de 1,59, 1,02 e 0,24, respectivamente. Esse padrão de contribuições exclusivamente negativas reflete uma instância em que todas as características importantes apontam para uma dosagem de coagulante inferior à média, demonstrando a capacidade do modelo de integrar múltiplos sinais redutores. No *summary plot* (Figura 2b), observa-se que as variáveis defasadas (lags temporais) estão entre as mais globalmente importantes, com altos valores de SHAP, indicando que o histórico de dosagens é determinante para as previsões atuais. Variáveis como Janela de Mudança, Pico de Cor AB, Delta de Cor AB-AT, Cor (AB), Média de Cor de Fontes, Delta de Cor AB-AD e Turbidez (AT) também se destacam, sugerindo que flutuações na cor e na turbidez entre diferentes pontos de captação, bem como indicadores derivados de janelas temporais, influenciam a dosagem de coagulante. É importante destacar que a maioria dessas variáveis foi criada no processo de *feature engineering* (defasagens, diferenças, razões, médias móveis, picos), evidenciando o papel

crucial dessa etapa na predição. Esse resultado reforça que a engenharia de atributos não apenas enriquece o conjunto de dados, mas também traduz aspectos físico-químicos e operacionais em indicadores mais interpretáveis e decisivos para o desempenho do modelo.

4. Conclusão

Os resultados obtidos demonstraram que a aplicação de diferentes técnicas de ML, combinadas com estratégias adequadas de *feature engineering* e tratamento de valores ausentes, pode fornecer previsões robustas e precisas da dosagem de coagulante em sistemas de tratamento de água. O modelo *ERTrees* sem preenchimento destacou-se pelo maior coeficiente de determinação (R^2 = 0,9337) e pelos menores erros MSE, RMSE e MAE entre todas as configurações testadas, indicando que a remoção de linhas com dados faltantes foi mais benéfica do que qualquer técnica de imputação avaliada. A única exceção ocorreu na métrica MAPE, na qual o método *Miss Forest* (n = 10) obteve o valor ligeiramente inferior (0,0398 versus 0,0403), evidenciando que diferentes critérios de avaliação podem levar a escolhas distintas. Além disso, a análise com SHAP revelou que as variáveis derivadas criadas no processo de *feature engineering*, especialmente as defasagens temporais (Lag 1, Lag 2 e Lag 4), diferenças entre pontos de captação e indicadores de pico, tiveram papel vital na predição, reforçando a importância dessa etapa para traduzir aspectos físico-químicos e operacionais em indicadores interpretáveis e decisivos. Apesar dos avanços, algumas limitações devem ser reconhecidas, isto é, o estudo foi conduzido com dados de uma única ETA, o que pode restringir a generalização dos resultados para outros contextos. Além disso, fatores externos não modelados, como variações climáticas extremas ou alterações súbitas na qualidade da água bruta, podem impactar a acurácia das previsões. Outro ponto é o custo computacional de modelos mais complexos, que pode limitar sua aplicação em ambientes operacionais com restrições de infraestrutura.

Como perspectivas futuras, recomenda-se ampliar o estudo para diferentes unidades de tratamento, incorporando variáveis adicionais relacionadas às condições ambientais e operacionais. A integração dos modelos preditivos com sistemas de apoio à decisão em tempo real pode potencializar sua aplicabilidade prática, permitindo ajustes mais ágeis e eficientes na dosagem de coagulante. Para a comunidade de Engenharia de Sistemas em Processos, este trabalho reforça a relevância da combinação entre técnicas avançadas de aprendizado de máquina e conhecimento físico-químico, apontando caminhos para soluções mais transparentes, interpretáveis e alinhadas às necessidades operacionais.

Agradecimentos: Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Processos 305472/2025-9 – C.S. e 312247/2022-2 – N.P.) e também pelo apoio financeiro por meio do projeto MAI/DAI, à Companhia Catarinense de Águas e Saneamento (CASAN) pelo fornecimento dos dados operacionais utilizados nas análises e à empresa *Hydroinfo Smart Solution*, parceira tecnológica.

Referências

L. Feng, Y. Zhang, X. Wei, M. Wang, Z. Niu e C. Wang: Interpretable Prediction of Coagulant Dosage in Drinking Water Treatment Plant Based on Automated Machine Learning and SHAP Method, Journal of Water Process Engineering (75), 107925, 2025.

J. Kim, C. Hua, S. Lin, S. Kang, J. Kang e M. Park: Deep Learning-Based Coagulant Dosage Prediction for Extreme Events Leveraging Large-Scale Data, Journal of Water Process Engineering (66), 105934, 2024.

N. Nasir, A. Kansal, O. Alshaltone, F. Barneih, M. Sameer, A. Shanableh e A. Al-Shamma'a: Water Quality Classification Using Machine Learning Algorithms, Journal of Water Process Engineering (48), 102920, 2022.

B. Ngwenya, T. Paepae e P. N. Bokoro: Monitoring Ambient Water Quality Using Machine Learning and IoT: a review and recommendations for advancing SDG indicator 6.3.2, Journal of Water Process Engineering (73), 107664, 2025.

C. O. Nwankwo, E. D. Ugwuanyi, N. N. Ehiobu, E. C. Ani e K. A. Olu-lawal: Advancing Wastewater Treatment Technologies: the role of chemical engineering simulations in environmental sustainability, World Journal of Advanced Research and Reviews (21), 019–031, 2024.

S. O. Ooko, L. Cheptegei e S. M. Karume: Application of Machine Learning for Real-Time Water Quality Monitoring in Developing Countries: a review, Sustainable Futures (10), 100984, 2025.

W. Zhao, Y. Liu, R. Shuang, W. Lu, S. Li, C. Zhao, C. Dou e H. Shu: Regulation of Coagulant Dosage in Water Treatment Based on Explainable Integrated Time-Series Deep Learning Models, Process Safety and Environmental Protection (201), 107613, 2025.