AVALIAÇÃO DO IMPACTO DE VARIÁVEIS HIDROMETEOROLÓGICAS NA PREDIÇÃO DE DOSAGEM DE COAGULANTE EM ESTAÇÕES DE TRATAMENTO DE ÁGUA

Juliana Neves[(1)]
Engenheira Química pela Universidade Federal de Santa Catarina (UFSC), com período sanduíche pela Ecole Supérieure de Chimie Physique Electronique de Lyon (CPE Lyon). Mestranda em Engenharia Química pela UFSC. Consultora na empresa HydroInfo.
Rodrigo Campos de Andrade[(2)]
Engenheiro Civil pela Universidade Federal de Santa Catarina (UFSC). Mestre e Doutor em Engenharia Oceânica pela Universidade Federal do Rio de Janeiro (UFRJ). Sócio e diretor da empresa HydroInfo.
Natan Padoin[(3)]
Professor Adjunto no Departamento de Engenharia Química e Engenharia de Alimentos da Universidade Federal de Santa Catarina (UFSC), em Florianópolis/SC. Desenvolve suas pesquisas no Laboratório de Materiais e Computação Científica (LabMAC/UFSC). Graduado em Engenharia Química (2011), mestre em Engenharia Química (2013) e doutor em Engenharia Química (2016) pela Universidade Federal de Santa Catarina (UFSC).
Cíntia Soares[(4)]
Professora Associada do Departamento de Engenharia Química e Engenharia de Alimentos da Universidade Federal de Santa Catarina (UFSC), em Florianópolis/SC. Desenvolve suas pesquisas no Laboratório de Materiais e Computação Científica, onde atua como supervisora. Graduada em Engenharia Química (1998) pela Universidade Regional de Blumenau (FURB), mestre (2000) e doutora em Engenharia Química (2005) pela Universidade Estadual de Campinas (UNICAMP).

Endereço[(1)]**:** Universidade Federal de Santa Catarina (UFSC), Centro Tecnológico (CTC), Departamento de Engenharia Química e Engenharia de Alimentos (EQA), Laboratório de Materiais e Computação Científica (LabMAC), Campus Universitário Reitor João David Ferreira Lima, s/n – Florianópolis, SC – CEP: 88040-900 – Brasil – Tel: +55 (48) 3721-4070 - e-mail: nevesjulianajn@gmail.com

RESUMO

Eventos de chuva são vistos como um dos principais fatores ambientais determinantes para a antecipação de ajustes na dosagem de coagulante em processos de tratamento de água para abastecimento. Baseado no exposto, o presente estudo teve como objetivo avaliar o desempenho dos modelos *Extreme Gradient Boosting* (XGBoost) e *Long Short-Term Memory* (LSTM) para previsão da dosagem de coagulante em uma ETA ao incorporar dados de qualidade da água, vazão dos rios de captação e registros de precipitação na bacia hidrográfica. Os resultados indicaram que os dados hidrometeorológicos atuaram como ruídos para o desempenho dos modelos. Destacou-se a alta complexidade envolvida no padrão de causalidade do sistema, bem como a subjetividade incutida no critério de controle da dosagem, não seguindo-se um padrão matemático claro. Conclui-se que o aprimoramento da capacidade preditiva dos modelos com a inclusão de dados de chuva e vazão dos rios não é garantido, tendo em vista que a relação entre essas informações e a qualidade da água de captação é multifatorial. A definição de critérios rigorosos de controle de dosagem na operação é fundamental para permitir um adequado desenvolvimento de ferramentas de previsão aplicando modelos de *machine learning*.

PALAVRAS-CHAVE:
Machine learning, tratamento de água, dosagem de coagulante.

INTRODUÇÃO

Nas últimas décadas, iniciativas de digitalização têm se desenvolvido e ganhado destaque no setor do saneamento como um caminho para maior controle e otimização dos processos (Huang *et al*., 2021; Lin *et al*., 2023a; Lowe; Qin; Mao, 2022). Mais recentemente, ferramentas de inteligência artificial, como o *machine learning* (ML), têm demonstrado grande potencial como apoio para a tomada de decisão nas operações, assim como para torná-las mais proativas e resilientes frente a cenários adversos (Raseman *et al*., 2017; Huang *et al*., 2021; Li *et al*., 2021). Algoritmos de ML permitem prever o comportamento de novos dados a partir da observação de um conjunto conhecido de informações e da identificação de relações ocultas entre as variáveis de interesse. Esta capacidade se mostra útil

principalmente para processos em que há grande complexidade de compreensão por métodos convencionais ou quando envolve dinâmicas altamente não-lineares (Huang *et al*., 2021; Li *et al*., 2021).

Estações de tratamento de água (ETAs) são exemplos de sistemas complexos que envolvem o monitoramento de variáveis com relações altamente não-lineares entre si (Lin *et al*., 2023b). Particularmente, em processos convencionais, a coagulação é uma etapa crítica do tratamento em termos de otimização, pois envolve a manutenção de uma dosagem de coagulante ótima que garanta a eficiência do processo e, ao mesmo tempo, seja economicamente viável (Kim *et al*., 2024). No cenário brasileiro, por exemplo, os gastos com produtos químicos em ETAs podem chegar a 57 % do total de despesas anuais da estação (SINISA, 2025). Contudo, é visto que a determinação de uma dosagem ótima é um desafio justamente pela alta complexidade das relações entre as diferentes variáveis de qualidade da água monitoradas (Achite *et al*., 2024; Bressane *et al*., 2023; Kim, Parnichkun, 2017a).

Geralmente, na prática operacional, esta dosagem é determinada a partir de ensaios laboratoriais (*Jar test*). No entanto, realizar estas análises com frequência torna-se custoso tanto em termos econômicos quanto em questão de tempo para sua execução. Devido a este fator do tempo, os resultados obtidos com estes ensaios não necessariamente refletem a condição da operação no momento da tomada de decisão do ajuste e, consequentemente, podem levar a mudanças de dosagem ineficientes. Além disso, estas análises estão sujeitas à avaliação subjetiva dos operadores e apresentam desafios de escala quando resultados de bancada são comparados às respostas no processo real (Achite *et al*., 2023; Boumezbeur *et al*., 2023; Bressane *et al*., 2023; Feng *et al*., 2025; Gomes *et al*., 2015; Haghiri; Daghighi; Moharramzadeh, 2018; Heddam; Bermad; Dechemi, 2012; Jakovljevic; Charlin; Barbeau, 2024; Liu *et al*., 2021; Zhang *et al*., 2013).

Devido às dificuldades de se realizar ensaios laboratoriais com alta frequência, a tomada de decisão acerca do ajuste da dosagem é muitas vezes realizada de forma empírica, tendo por base o conhecimento prático do operador. Normalmente, considera-se os parâmetros de qualidade da água e as condições meteorológicas, sendo então deliberada a necessidade de mudança na dosagem. Entretanto, esta metodologia de decisão é propensa a falhas por erro humano, apesar de ser amplamente relatado pela literatura como um hábito usual em muitas ETAs (Chiavola *et al*., 2023; Feng *et al*., 2025; Kim; Parnichkun, 2017a; Lin *et al*., 2023a, 2023b; Liu *et al*., 2021; Luo *et al*., 2022; Zhang *et al*., 2013).

Neste contexto, técnicas de *machine learning* têm grande potencial de servir como ferramentas de apoio na definição da dosagem ideal para o processo. Diversos casos são retratados pela literatura em que modelos de previsão demonstraram desempenho satisfatório (Achite *et al*., 2024; Ghasemi *et al*., 2022; Haghiri; Daghighi; Moharramzadeh, 2018; Heddam; Bermad; Dechemi, 2012; Jayaweera; Aziz, 2022; Kim; Parnichkun, 2017a, 2017b; Li *et al*., 2023; Lin *et al*., 2023a, 2023b; Narges *et al*., 2021; Pang *et al*., 2025; Wang; Chang; Ma, 2022; Wang; Asefa; Thornburgh, 2022; Wu; Lo, 2008; Zhang *et al*., 2013). Contudo, um dos principais gargalos da utilização desses algoritmos é a sua baixa interpretabilidade, o que dificulta a compreensão das relações físico-químicas estabelecidas pelas arquiteturas dos algoritmos (Schneider *et al*., 2022).

Além disso, outro fator limitante é a sua capacidade de extrapolação, isto é, o seu potencial de representar comportamentos do sistema nunca observados no histórico de dados. A representatividade desses modelos acaba sendo restrita a cenários anteriormente registrados (Quaghebeur, 2023; Torfs *et al*., 2022). Isto restringe a sua aplicação em sistemas em que a avaliação da qualidade de águas naturais é relevante (como em ETAs, por exemplo), pois eventos extremos e adversos podem ocorrer ocasionalmente e alterar significativamente o padrão observado de forma pontual (Suquet *et al*., 2021).

Neste sentido, dados pluviométricos se demonstram de grande relevância para a compreensão dos padrões de variação da qualidade da água. De acordo com a literatura, mesmo em regiões de solo com menor tendência à erosão, eventos de chuvas intensas provocam picos de turbidez, cor e concentração de matéria orgânica em águas naturais, sobretudo devido à lixiviação de partículas e substâncias húmicas (Di Bernardo, Di Bernardo, Centurione Filho, 2002; Libânio, 2010; Lin *et al*., 2023b; Ortiz-Lopez; Bouchard; Rodriguez, 2022; Raseman *et al*., 2017; Zhang, 1996). Além disso, estes eventos também proporcionam um aumento na vazão de rios, o que pode resultar em turbulência no escoamento e gerar maiores forças hidrodinâmicas que podem ressuspender sedimentos decantados (Neves, 2010).

Sabendo-se destas consequências, a precipitação pluviométrica é vista como um dos principais fatores ambientais determinantes para a antecipação de ajustes na dosagem química em processos de tratamento de água para abastecimento. Quando se trata de aplicações de ferramenta de previsão como o *machine learning*, casos práticos têm demonstrado a importância de se considerar este fator para a melhoria da acurácia dos modelos. A exemplo, no estudo desenvolvido por Lin *et al*. (2023a), um extenso conjunto de dados de monitoramento (abrangendo o período de 2011

a 2021) de uma ETA permitiu observar uma correlação direta entre picos de turbidez, demanda de coagulante e a estação chuvosa de verão. Isto evidenciou o efeito da sazonalidade e das condições meteorológicas para a operação.

Neste mesmo estudo, os autores constataram que as previsões de dosagem realizadas pelo modelo de previsão escolhido foram prejudicadas nesses períodos de pico, passando a subestimar a dosagem necessária de coagulante. Por outro lado, nos meses sem condições meteorológicas extremas, o modelo apresentou desempenho satisfatório. Resultados similares foram observados em outros trabalhos da literatura em que modelos de *machine learning* foram utilizados para previsão da dosagem em ETAs (Kim *et al*., 2024; Kim, Parnichkun, 2017a; Wu, Lo, 2008). Essa discrepância destaca que, em cenários de turbidez elevada devido às chuvas, as relações entre os parâmetros operacionais e de qualidade da água tornam-se altamente não-lineares a ponto de comprometer a acurácia de modelos de previsão (Lin *et al*., 2023a).

Considerando esta problemática, o presente trabalho visou avaliar o desempenho de dois modelos distintos de *machine learning* ao incorporar dados de qualidade da água bruta, decantada, vazão dos rios de captação e registros de precipitação pluviométrica na bacia hidrográfica para previsão da dosagem de coagulante em uma ETA. Estes últimos dados foram introduzidos com o objetivo de comparar a acurácia dos modelos ao considerá-los como variáveis de entrada. Os modelos selecionados foram a rede neural *Long Short-Term Memory* (LSTM) e o modelo baseado em árvores de decisão *Extreme Gradient Boosting* (XGBoost). Os resultados foram avaliados e discutidos em termos de comparação de métricas de desempenho ao comparar os dados reais com os previstos. A importância de cada variável foi também analisada de forma a compreender se a introdução dos registros de chuva impacta na acurácia dos modelos.

Por fim, as conclusões do presente estudo visam colaborar com o desenvolvimento de ferramentas de apoio à tomada de decisão em processos, de forma a garantir a conformidade do tratamento com os limites de qualidade exigidos pela legislação. A partir da implementação de uma ferramenta robusta, possibilita-se otimizar o consumo de insumos químicos em ETAs, reduzindo a dependência de decisões puramente empíricas e aprimorando a resiliência da estação frente a eventos climáticos.

METODOLOGIA

Descrição da ETA

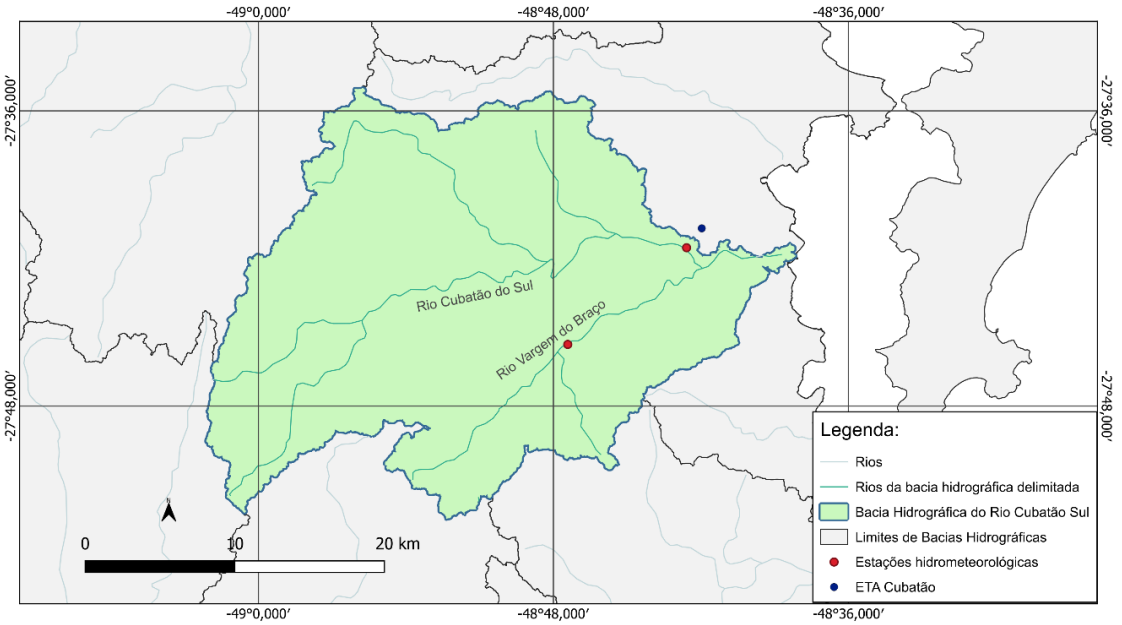
A Estação de Tratamento de Água José Pedro Horstmann (também conhecida como ETA Morro dos Quadros ou ETA Cubatão) é uma estação gerenciada pela Companhia Catarinense de Águas e Saneamento (CASAN) e está localizada na cidade de Palhoça (SC). Possui vazão média de 2500 L/s e sua captação é proveniente do Rio Pilões (ou também nomeado Rio Vargem do Braço) e do Rio Cubatão do Sul, localizados na Bacia Hidrográfica do Rio Cubatão do Sul (Figura 1). Cada rio contribui com cerca de 50 % da vazão bruta total captada.

De acordo com um estudo realizado com dados históricos pluviométricos e fluviométricos destes rios (Ramos, 2023), estes corpos hídricos sofrem com variações sazonais devido a eventos de chuva, sendo o verão caracterizado por elevada precipitação e altas vazões, enquanto no inverno ocorre o oposto. Além disso, os operadores da ETA relatam que no período chuvoso a água bruta captada apresenta maiores variações nos parâmetros de qualidade, ocorrendo, geralmente, diminuição do pH e aumento na cor e na turbidez.

Outro fator importante a pontuar é a presença de atividade antrópica na região da bacia hidrográfica, o que contribui para maior variação na qualidade da água dos rios. Agricultura familiar intensiva, extração mineral de areia e argila e, particularmente no rio Cubatão, o desmatamento da mata ciliar, são exemplos de ações contribuintes para o comprometimento da estabilidade das margens dos rios e a potencialização da erosão e do transporte de sedimentos (CASAN, 2024; Kong *et al*., 2022; Ramos, 2023).

Em termos de operação, a ETA é composta pelas etapas de coagulação/floculação, decantação, filtração, desinfecção, ajuste de pH e fluoretação. A coagulação é proporcionada pela adição do coagulante policloreto de alumínio (PAC) na água bruta e a gestão da dosagem deste produto é realizada de maneira sistêmica, com diversos parâmetros sendo avaliados simultaneamente. Segundo os operadores, além das variáveis de qualidade da água, também são consideradas as condições meteorológicas, deliberando-se a partir disso se haverá necessidade de mudança na dosagem. No período do estudo, ensaios de *Jar test* eram pouco realizados, o que evidencia um controle operacional baseado em empirismo.

Figura 1 – Bacia Hidrográfica do Rio Cubatão do Sul



Fonte: elaborado pelos autores (2026) com o uso do software QGIS a partir de dados de Bases Hidrográficas Ottocodificadas disponibilizadas pela ANA (2013, 2017) e dados de localização (Lopes, 2020; Weschenfelder; Pickbrenner; Pinto, 2016)

Base de dados

Na ETA Cubatão, amostras de água bruta, decantada, filtrada e tratada são coletadas a cada 2 h, sendo analisadas características de cor, turbidez e pH. Especificamente, cor e turbidez são também avaliadas isoladamente nos pontos de captação de cada um dos rios. Próximo a estes locais, a vazão de cada corpo hídrico, bem como dados pluviométricos a partir de estações hidrometeorológicas, também são monitorados. No rio Pilões, a estação (código 84207000) está localizada no município de Santo Amaro da Imperatriz (SC), na latitude 27°45’30’’S e longitude 48°47’25’’O, *datum* SIRGAS 2000. No rio Cubatão, a estação (código 84150100) está localizada no mesmo município, na latitude 27°41’34’’S e longitude 48°42’35’’O.

Para o presente estudo, um histórico de dados coletados no período de dezembro de 2019 a maio de 2025 foi utilizado para o treinamento e teste dos modelos de previsão. Os parâmetros considerados para o estudo e seus respectivos pontos de coleta são elencados na Tabela 1.

Tabela 1 – Parâmetros de monitoramento considerados para o estudo

Ponto de coleta	Parâmetros
Água Bruta (mistura)	Cor Turbidez pH Vazão de entrada
Água Bruta (Rio Cubatão)	Cor Turbidez
Água Bruta (Rio Pilões)	Cor Turbidez
Mistura Rápida	Dosagem de coagulante
Rio Cubatão (estação 84150100)	Precipitação pluviométrica Vazão fluviométrica
Rio Pilões (estação 84207000)	Precipitação pluviométrica Vazão fluviométrica

Fonte: elaborado pelos autores (2026)

Pré-processamento dos dados

Os dados disponibilizados passaram por uma etapa de pré-processamento, que consistiu primeiramente na identificação e remoção de valores ausentes. A remoção ocorreu a partir da exclusão da linha completa de registros de cada ponto em que foi identificada alguma informação em branco.

Uma observação é que, previamente a esta etapa, um conjunto de dados de precipitação pluviométrica, monitorado por uma estação meteorológica próxima da região da ETA, foi utilizado como base para preenchimento de registros ausentes das séries temporais das variáveis de precipitação disponíveis. Isto correspondeu a cerca de 9 % do total de dados do Rio Cubatão e 5 % dos dados do Rio Pilões.

O segundo passo consistiu na aplicação de um filtro preliminar de consistência dos dados, tendo por objetivo eliminar valores fisicamente irreais. Para isso, limites descritos na literatura e o próprio comportamento histórico observado na estação foram adotados como referência para delimitação desse filtro.

A partir disso, procedeu-se à remoção de *outliers* de forma a remover do conjunto de dados possíveis erros ou ruídos. No entanto, considerando-se os dados hidrometeorológicos, variações abruptas podem ser usuais, assim como eventos extremos que ocasionam registros de valores significativamente fora do padrão usual (Bressane *et al*., 2023; Helsel *et al*., 2020). Dessa forma, a remoção destas informações resultaria numa representação incorreta do histórico real observado (Navidi, 2011).

Seguindo este princípio, optou-se por não aplicar a remoção de *outliers* nos dados de precipitação e de vazão dos rios devido aos seus registros não terem apresentado dados que pudessem ser considerados erros evidentes de medição. Além disso, optou-se por esta manutenção de modo a observar o comportamento dos modelos com o uso de dados reais com eventos extremos.
Nas demais variáveis, adotou-se o critério de 3σ (três sigma) aliado à definição do Z-score (ou pontuação Z). Este índice quantifica o distanciamento de um ponto específico em relação à média do conjunto de dados, avaliando em termos de desvios padrão. Ele é expresso pela seguinte equação:

$$Z = \frac{x - \mu}{\sigma} \qquad \text{Equação (1)}$$

onde x representa o valor observado, μ a média da amostra e σ o desvio padrão da amostra

É importante ressaltar que, idealmente, este método serve para identificação de *outliers* em dados com distribuição normal clássica, o que normalmente não é o perfil de distribuição padrão de séries temporais hidrológicas e de qualidade de água (Helsel *et al*., 2020). Para contornar essa limitação, realizou-se um diagnóstico prévio de normalidade para cada variável, empregando-se análises visuais por meio de histogramas e gráficos Q-Q (quantil-quantil). As variáveis que exibiram distribuições com assimetria à direita foram submetidas a uma transformação logarítmica, de forma a aproximá-las do comportamento gaussiano. Isto é retratado na literatura como uma prática comum em dados com este tipo de simetria, permitindo estabilizar sua variância (Helsel *et al*., 2020; Navidi, 2011) Somente após essa transformação, o Z-score foi calculado e o critério de 3σ foi aplicado para a exclusão efetiva dos *outliers*. Por fim, os dados retidos tiveram sua escala original restaurada pela reversão do logaritmo.

Uma etapa posterior foi a implementação de *feature engineering* (ou engenharia de atributos), a qual consiste na extração de características dos dados brutos em novas variáveis a serem fornecidas para os modelos de *machine learning*. Uma das variáveis criadas foi um indicador binário de evento extremo para identificação de picos de turbidez na água bruta. Para isso, adotou-se o percentil 97 como limiar para a classificação, sendo, então, os 3 % dos valores mais elevados de turbidez rotulados como eventos extremos.

Este valor foi definido com base na definição da literatura de eventos incomuns, mas ainda estatisticamente significativos (Helsel *et al*., 2020; Navidi, 2011). Buscou-se selecionar eventos que, embora raros, ocorrem com frequência suficiente para considerá-los um modo operacional da ETA que exige atenção especial. Além disso, considerou-se também um valor aproximado em que os eventos de forte precipitação passam a impactar significativamente a turbidez da água bruta. Para esse objetivo, um algoritmo de classificação baseado em árvore de decisão (*Decision Tree Classifier*) do pacote *Scikit-learn* em Python foi utilizado. Este modelo foi aplicado de forma

a avaliar os dados históricos de precipitação e determinar o valor em que a chuva é forte o suficiente para atingir picos significativos de turbidez.

Neste caso, uma análise de sensibilidade foi realizada e foi identificado que os percentis de 95 e 96 da turbidez da água bruta correspondiam a um limiar de somente 4 mm de precipitação acumulada por hora, o que é considerado uma chuva de intensidade leve (INMET, 1999). De maneira a corresponder a uma precipitação moderada, o percentil 97 foi selecionado, sendo definido um limite crítico de aproximadamente 11 mm/h.

Outras variáveis criadas foram o desvio padrão móvel (assumindo-se o desvio padrão sobre os 6 últimos registros observados) da turbidez e cor da água bruta e decantada e a eficiência de remoção da turbidez a partir da dosagem de coagulante, calculada pela Equação (2):

$$\text{Eficiência} = \frac{\left(\text{Turbidez}_{\text{bruta (t – 1)}} - \text{Turbidez}_{\text{decantada(t)}}\right)}{\text{Turbidez}_{\text{bruta (t – 1)}}} \qquad \text{Equação (2)}$$

O dado de turbidez do tempo anterior foi assumido para este cálculo devido à consideração do tempo de residência, sendo importante para que o efeito da dosagem possa ser avaliado de forma efetiva (Kim *et al*., 2024).

De forma a avaliar o impacto da presença dos dados hidrometeorológicos no desempenho dos modelos de previsão, optou-se por construir três conjuntos distintos de registros:

1) Cenário 1: dados de monitoramento da dosagem de coagulante e da qualidade da água bruta e decantada;
2) Cenário 2: dados de monitoramento da dosagem de coagulante, da qualidade da água bruta e decantada, de precipitação pluviométrica e da vazão dos rios;
3) Cenário 3: dados de monitoramento da dosagem de coagulante, da qualidade da água bruta e decantada, de precipitação pluviométrica e da vazão dos rios, incluindo atraso temporal dos dados hidrometeorológicos.

Neste último cenário, *features* específicas foram criadas para incorporar informações de atraso temporal dos efeitos causados pela chuva e pela vazão dos rios. Para isso, os registros de vazão e precipitação em cada rio medidos 2 h, 12 h, 24 h e 36 h antes do passo de tempo atual foram incluídos, bem como a média móvel desses dados considerando-se os últimos três registros (isto é, os dados medidos 2 h, 4 h e 6 h antes do tempo atual). Esses intervalos de atraso temporal foram definidos a partir da análise de correlação de Spearman entre os dados de chuva e turbidez da água bruta, e chuva e dosagem de coagulante na ETA. O objetivo da inclusão destes dados foi fornecer informações sobre possíveis tendências de variabilidade e possibilitar a identificação de efeitos tardios das condições meteorológicas na qualidade da água captada.

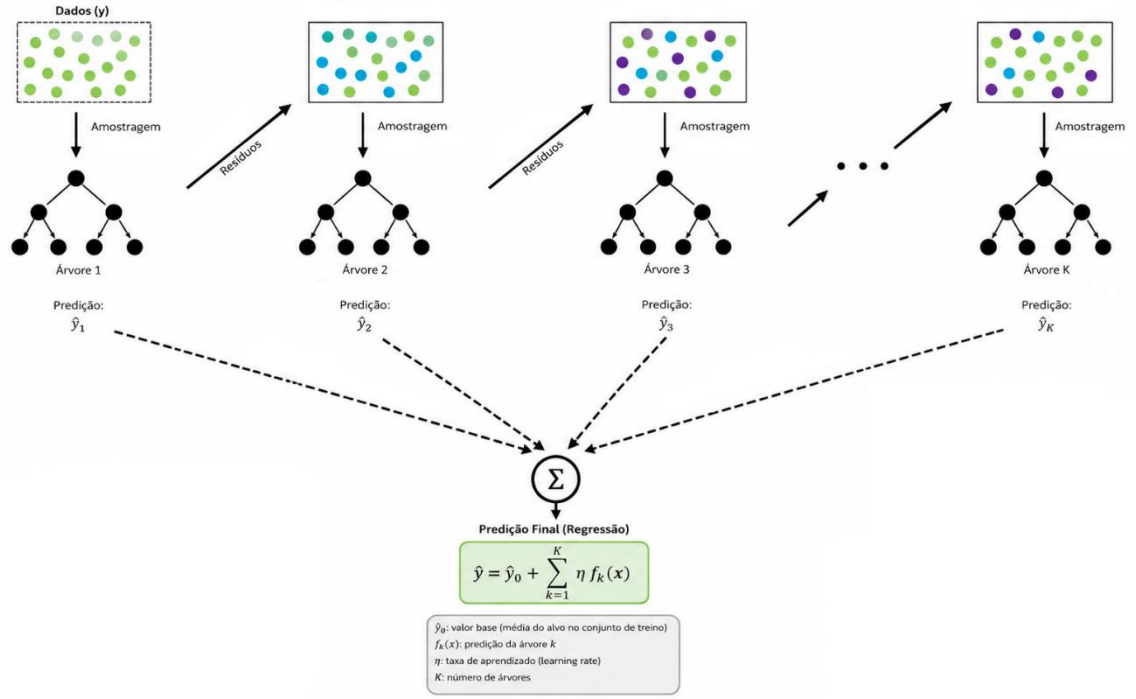
Modelos de *machine learning*, implementação e análise

Duas abordagens diferentes de modelos de *machine learning* foram implementados: o XGBoost e o LSTM. O primeiro é um algoritmo baseado em árvores de decisão que aplica um método sequencial de construção nomeado *boosting*, no qual novos modelos são adicionados iterativamente para corrigir os erros residuais dos preditores anteriores (Figura 2). Essa técnica permite que cada nova árvore concentre seu aprendizado nas regiões do espaço de dados onde o modelo ainda apresenta desempenho insatisfatório (Chen; Guestrin, 2016).

O LSTM é um tipo de arquitetura de rede neural especificamente projetada para dados sequenciais ou com dependência temporal (Jones, Jones, Horsburgh, 2022). O seu principal diferencial reside na estrutura interna, que incorpora dois vetores de memória distintos: o estado oculto, comum a outras redes recorrentes, e o estado da célula, exclusivo dessa arquitetura (Figura 3). Esse arranjo permite que a rede preserve e transporte informações relevantes ao longo de extensos intervalos de tempo, sendo fundamental para tarefas que envolvam a análise de séries temporais e a compreensão de contextos em sequências de dados (Weidman, 2019).

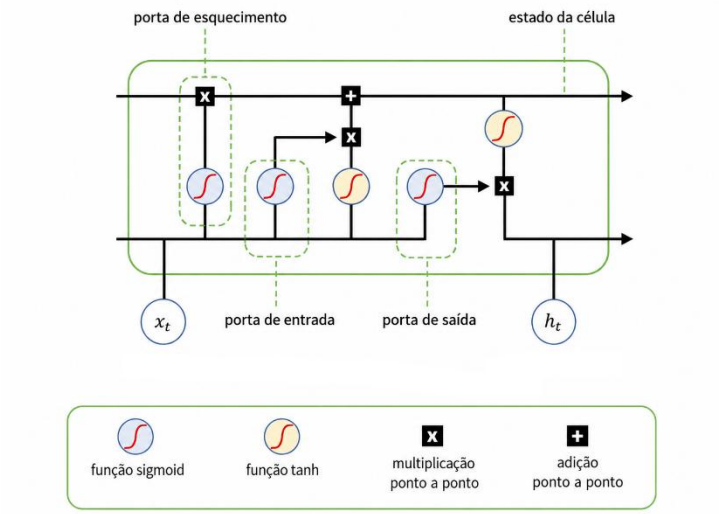
Para o presente estudo, os modelos foram implementados de forma independente, sendo aplicados de forma separada para cada um dos cenários de dados construídos. Cada conjunto de dados de entrada foi dividido em um conjunto de treino e outro de teste em uma proporção de 80 % e 20 %, respectivamente. Esta divisão respeitou a ordem cronológica dos dados devido à importância de se preservar as relações temporais entre os registros. Para o modelo LSTM especificamente, uma etapa prévia de escalonamento dos dados foi necessária de maneira a transformar os valores de cada variável para um intervalo entre 0 e 1.

Figura 2 – Esquema ilustrativo da arquitetura do modelo XGBoost



Fonte: adaptado de Deng *et al*. (2021).

Figura 3 – Esquema ilustrativo da arquitetura do modelo LSTM



Fonte: adaptado de Hansun, Wicaksana e Khaliq (2022).

A seleção dos melhores hiperparâmetros de cada modelo foi realizada através da ferramenta de otimização nomeada Optuna, a qual combina algoritmos de amostragem e de "poda" (*pruning*) de tentativas ruins para tornar este processo mais ágil e preciso (Akiba *et al*., 2019). Para isso, o algoritmo *Tree-structured Parzen Estimator* (TPE) foi selecionado.

Internamente ao ciclo de otimização do Optuna, uma validação cruzada com *TimeSeriesSplit* foi implementada, com maximização do R^2 em no máximo 100 tentativas. Os hiperparâmetros selecionados para otimização e os seus respectivos intervalos em cada modelo são elencados na Tabela 2.

A avaliação dos modelos finais com os melhores hiperparâmetros foi realizada por meio da Raiz do Erro Quadrático Médio (RMSE), Erro Médio Absoluto (MAE), Coeficiente de Determinação (R^2) e Coeficiente de Pearson (r). A

importância das variáveis preditoras foi avaliada através do método *Permutation Feature Importance*, o qual seleciona uma única variável por vez e embaralha aleatoriamente seus valores, mantendo as demais variáveis intactas. Com isso, a queda do R^2 é analisada e quantificada, indicando, assim, as *features* de maior relevância para a manutenção do desempenho.

Tabela 2 – Hiperparâmetros selecionados para otimização e os seus respectivos intervalos em cada modelo

Modelo	Hiperparâmetros	Conceito	Intervalo
XGBoost	n_estimators	Número total de árvores de decisão	100 – 1000
	max_depth	Profundidade máxima de cada árvore	3 – 20
	learning_rate	Passo de iteração	0,01 – 3
	subsample	Fração das observações de treino usadas para construir cada árvore	0,5 – 1,0
	colsample_bytree	Percentual de *features* selecionadas para a construção de cada árvore individual	0,5 – 1,0
	gamma	Valor mínimo de redução de perda	0 – 5
	reg_alpha	Termo de regularização L1 (Lasso) sobre os pesos das folhas	0 – 10
	reg_lambda	Termo de regularização L2 (Ridge) sobre os pesos das folhas	0 – 10
LSTM	units1	Número de neurônios na primeira e segunda camadas	32 – 128
	units2		32 – 128
	dropout1	Taxa de desativação aleatória de neurônios na primeira e segunda camadas durante treinamento	0,1 – 0,5
	dropout2		0,1 – 0,5
	learning_rate	Fator de ajuste do tamanho das correções aplicadas aos pesos da rede	$1e^{-5}$ – $1e^{-2}$ (log)
	batch_size	Número de amostras processadas pela rede antes de atualizar os pesos internos	64, 128, 256, 512

Fonte: elaborado pelos autores (2026)

RESULTADOS E DISCUSSÃO

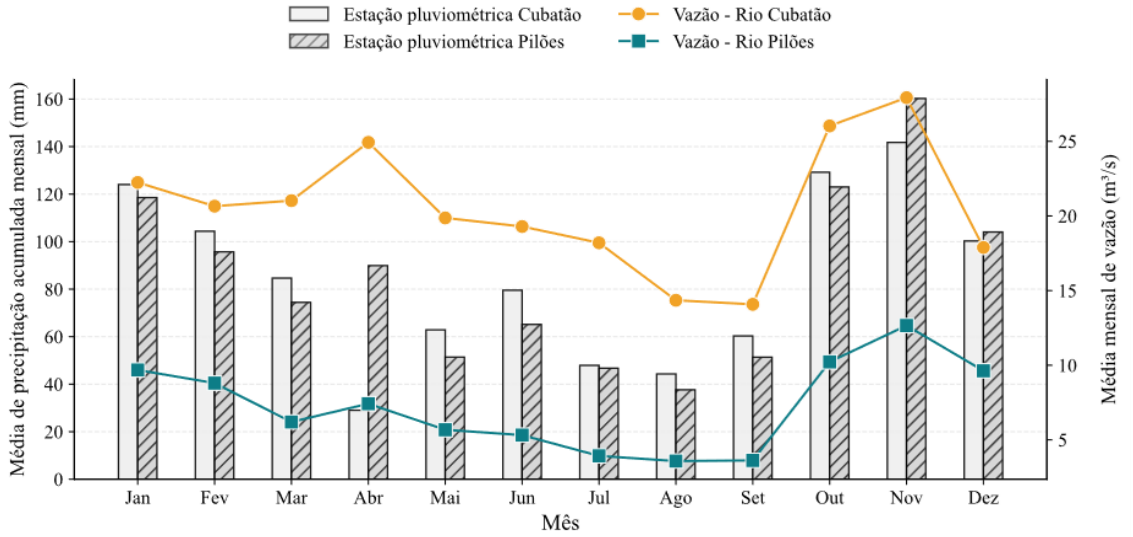
Análise dos dados

O conjunto de dados utilizado para o estudo foi avaliado em termos de médias mensais de modo a verificar padrões de sazonalidade nas características da qualidade da água bruta em função das condições meteorológicas e hidrológicas. Os registros foram avaliados considerando o período de dezembro de 2019 a março de 2025. A Figura 4 ilustra as médias obtidas de precipitação acumulada mensal das estações pluviométricas do rio Cubatão e do rio Pilões, comparando-as às médias de vazão de cada um dos rios no mesmo período. É possível observar que, em geral, as variáveis acompanharam tendências temporais similares, com aumentos e reduções ocorrendo nos mesmos períodos do ano. Entre meados da primavera (outubro e novembro) e durante a maior parte do verão (dezembro a fevereiro) foram registrados os maiores níveis de chuva e, consequentemente, as maiores vazões dos rios. Isto sugere uma correlação positiva entre as variáveis hidrológicas, o que é amplamente corroborado pela literatura (Loch, 2019; Lopes, 2020).

Contudo, observa-se uma pequena discrepância entre o perfil sazonal da vazão do rio Cubatão com a do rio Pilões entre os meses de fevereiro e abril. O primeiro apresenta uma tendência crescente neste período, enquanto no rio Pilões há uma tendência mais estável. Além disso, há um o aumento mais pronunciado da vazão do Rio Cubatão no mês de abril, mesmo com uma considerável redução da precipitação medida na sua região. Por outro lado, a precipitação registrada na estação do rio Pilões foi relativamente elevada. Pode-se inferir que estas diferenças são explicadas por uma possível distribuição irregular das chuvas na bacia. Além disso, como observado anteriormente, o rio Cubatão apresenta atividade antrópica mais intensa em suas margens, além de sofrer de forma mais vasta com o desmatamento da mata ciliar (CASAN, 2021; Loch, 2019). Tendo em vista que estes fatores podem contribuir com alterações no

padrão hidrológico do rio, estas são também possíveis razões para a diferença nas médias deste período (Girardi *et al.*, 2016; Lopes, 2020).

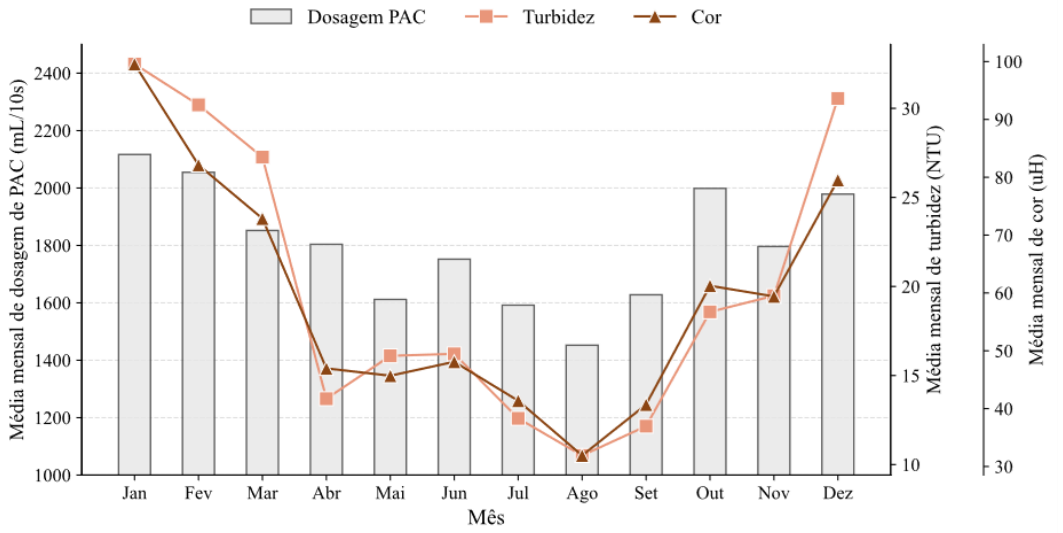
Figura 4 – Perfis de sazonalidade da precipitação acumulada mensal e das vazões dos rios Cubatão e Pilões



Fonte: elaborado pelos autores (2026)

A Figura 5 retrata o perfil sazonal das características da água bruta (turbidez e cor) em comparação ao padrão mensal da dosagem de coagulante aplicada na ETA Cubatão. Novamente, uma associação positiva é observada entre as variáveis, demonstrando que, na prática, de fato a dosagem é aumentada quando a turbidez e a cor da água bruta se elevam no processo. Isto também evidencia que estes parâmetros podem servir como potenciais preditores da dosagem no contexto do estudo.

Figura 5 – Perfis de sazonalidade das médias mensais de dosagem de coagulante PAC, turbidez e cor da água bruta



Fonte: elaborado pelos autores (2026)

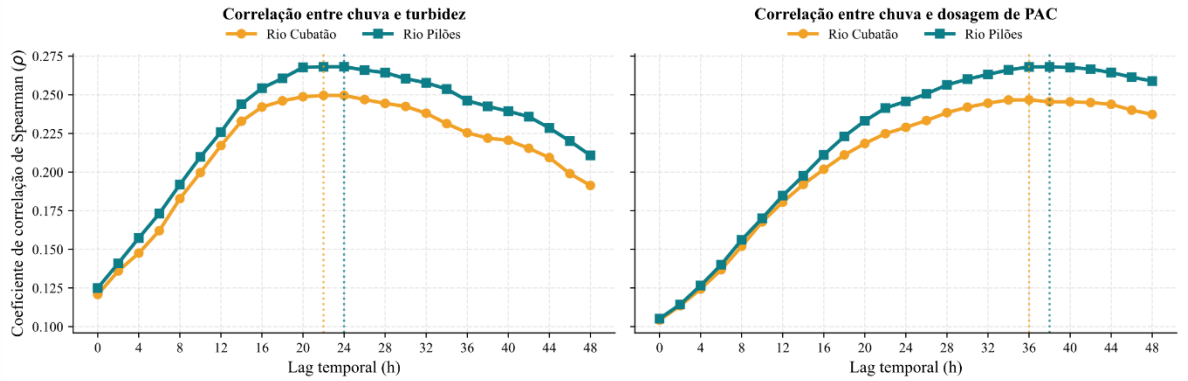
Além disso, a sazonalidade dessas variáveis seguiu um padrão similar ao observado pela precipitação acumulada na região, em que os meses correspondentes a meados da primavera e início do verão apresentaram, de forma geral, um aumento dos valores de dosagem, turbidez e cor. Contudo, neste caso, os meses de dezembro a fevereiro tiveram maiores registros, enquanto a precipitação teve seu maior aumento observado nos meses de outubro e novembro.

Isto pode ser explicado principalmente pela diferença da formação de chuvas em cada um dos períodos. Segundo a literatura, a climatologia da região indica a predominância de chuvas do tipo convectivas nos meses de maior calor (correspondentes ao verão), as quais são mais intensas e apresentam maior volume em um curto intervalo de tempo (Loch, 2019; Lopes, 2020). Com isso, essa precipitação excede a capacidade de infiltração do solo e apresenta maior energia cinética. Como consequência, é observada uma maior erosão do solo com elevado carreamento de partículas para os rios (Santos, Montenegro, 2012; Wischmeier, Smith, 1978), aumentando, assim, os níveis de turbidez e cor dos corpos hídricos da região. Por outro lado, as chuvas que ocorrem no final da primavera, nos meses de outubro e novembro, não costumam ser de grande intensidade como no verão, apesar de ainda poderem ocorrer com frequência. Isto justifica a observação de médias inferiores de turbidez e cor da água neste período.

Contudo, é importante enfatizar que os resultados observados foram avaliados com um histórico de aproximadamente cinco anos, o que pode ser considerado curto e não refletir completamente a climatologia da região. Uma análise realizada por Loch (2019), com um histórico de 20 anos (1997 a 2017) de médias mensais de precipitação na região da bacia hidrográfica do rio Cubatão do Sul, contradiz o exposto pelos dados históricos ilustrados na Figura 4. É visto que os meses de dezembro, janeiro e fevereiro tiveram os maiores acúmulos mensais de chuva, o que deve ser levado em consideração para a análise dos perfis de chuva em associação com a dosagem de coagulante na ETA Cubatão.

Um ponto levantado pela literatura diz respeito ao atraso temporal dos efeitos da precipitação pluviométrica na qualidade da água dos rios (Girardi *et al*., 2016; Loch, 2019). A Figura 6 ilustra resultados de correlação de *Spearman* de forma a compreender possíveis associações existentes entre os dados de chuva e turbidez da água bruta e chuva e dosagem de coagulante. O coeficiente de *Spearman* foi selecionado pois avalia apenas a relação monotônica entre variáveis, sendo relevante para dados hidrológicos que apresentam não-linearidade e assimetrias em sua distribuição (Helsel *et al*., 2020).

Figura 6 – Correlação de Spearman avaliada entre atrasos temporais dos dados de chuva e turbidez da água bruta (esquerdo) e entre atrasos temporais dos dados de chuva e dosagem de PAC



Fonte: elaborado pelos autores (2026)

Analisando-se os gráficos, é possível observar que, em ambos os cenários, há um aumento crescente da correlação com o aumento do atraso temporal, atingindo-se um pico e decrescendo de forma menos pronunciada. No caso da análise entre chuva e turbidez da água bruta, a correlação máxima é observada com atraso de 22 h para o rio Cubatão, com uma correlação de 0,250, e de 24 h para o rio Pilões, com uma correlação de 0,268. Para chuva e dosagem de coagulante, o pico é observado com atraso de 36 h para o rio Cubatão, com uma correlação de 0,247 e 38 h para o rio Pilões, com uma correlação de 0,268. Segundo Cohen (1988), estes valores indicam associações positivas, mas com magnitudes fracas a moderadas, o que para dados hidrológicos é realista tendo em vista a grande quantidade de possíveis interferências, como demais variáveis influenciadoras e erros de medição.

De forma a validar esta correlação, a significância estatística foi calculada para cada ponto de interesse através da medida de p-valor, conforme Helsel *et al.* (2020). Os resultados indicaram a existência de uma correlação de fato estatisticamente significativa, obtendo-se p-valor menor que 0,05 para todos os casos. Isto confirma que, apesar da correlação ser fraca a moderada, os resultados demonstram a existência desta associação.

Desempenho dos modelos de previsão da dosagem de coagulante

Com relação à previsão da dosagem de coagulante com os modelos de *machine learning*, os resultados de métricas obtidos (Tabela 3) se demonstraram em geral abaixo do esperado quando comparados a estudos da literatura que aplicaram os mesmos modelos ou similares (Tabela 4) para previsão do mesmo parâmetro. Dentre os testes realizados no presente estudo, o melhor desempenho observado foi com a aplicação do modelo LSTM com os dados do Cenário 1, em que somente dados de características da água bruta e decantada (bem como as *features* adicionais relacionadas a estas variáveis) foram utilizadas para treinamento e teste. Isso também foi observado com o modelo XGBoost, em que os dados do Cenário 1 mostraram métricas mais satisfatórias do que os demais. Logo, observa-se que a incorporação de dados hidrológicos e meteorológicos da bacia não trouxe benefícios para a melhoria do desempenho dos modelos de previsão.

Isto pode ser explicado pela alta complexidade e não-linearidade da relação entre os eventos de chuva e o ajuste de dosagem de coagulante na ETA (Wu, Lo, 2008; Zhang *et al*., 2013). Na prática, os operadores decidem o valor a ser dosado com base em uma avaliação subjetiva, focando, principalmente, nos parâmetros de qualidade da água, conforme relatado. Variáveis como turbidez e cor são, portanto, variáveis de efeito mais direto nesta decisão (Lin *et al*., 2023b). Além disso, traduzem e quantificam os impactos de eventos meteorológicos ocorridos a montante e que podem afetar o tratamento da água na estação (Chiavola *et al*., 2023; Wu, Lo, 2008). Os dados de precipitação e vazão dos rios estão sujeitos a demais fatores ambientais, por vezes não afetando diretamente a qualidade da água (Lin *et al*., 2023b) a depender, por exemplo, da umidade antecedente do solo ou da distribuição espacial da chuva.

Tabela 3 – Métricas obtidas com modelos de previsão de dosagem de coagulante

Modelo	Cenário	Métrica			
		RMSE	MAE	R^2	r
XGBoost	1	388,01	315,74	0,34	0,67
	2	446,47	369,35	0,18	0,66
	3	504,82	414,76	-0,05	0,63
LSTM	1	367,07	299,29	0,41	0,67
	2	490,84	392,46	0,01	0,59
	3	585,11	464,94	-0,41	0,47

Fonte: elaborado pelos autores (2026).

Tabela 4 – Métricas obtidas em estudos da literatura de aplicação dos modelos de ML para previsão da dosagem de coagulante

Modelo	R^2	r	Referência
XGBoost	0,973	-	(Hu *et al.*, 2023)
XGBoost	-	0,95	(Dunnington *et al.*, 2021)
LSTM	0,62	-	(Lin *et al.*, 2023b)
LSTM	0,70	-	(Wang; Asefa; Thornburgh, 2022)
ANN	0,85	-	(Haghiri *et al*., 2018)
GRU	0,88	0,94	(Lin *et al*., 2023a)
GRU	0,94	-	(Lin *et al*., 2023b)

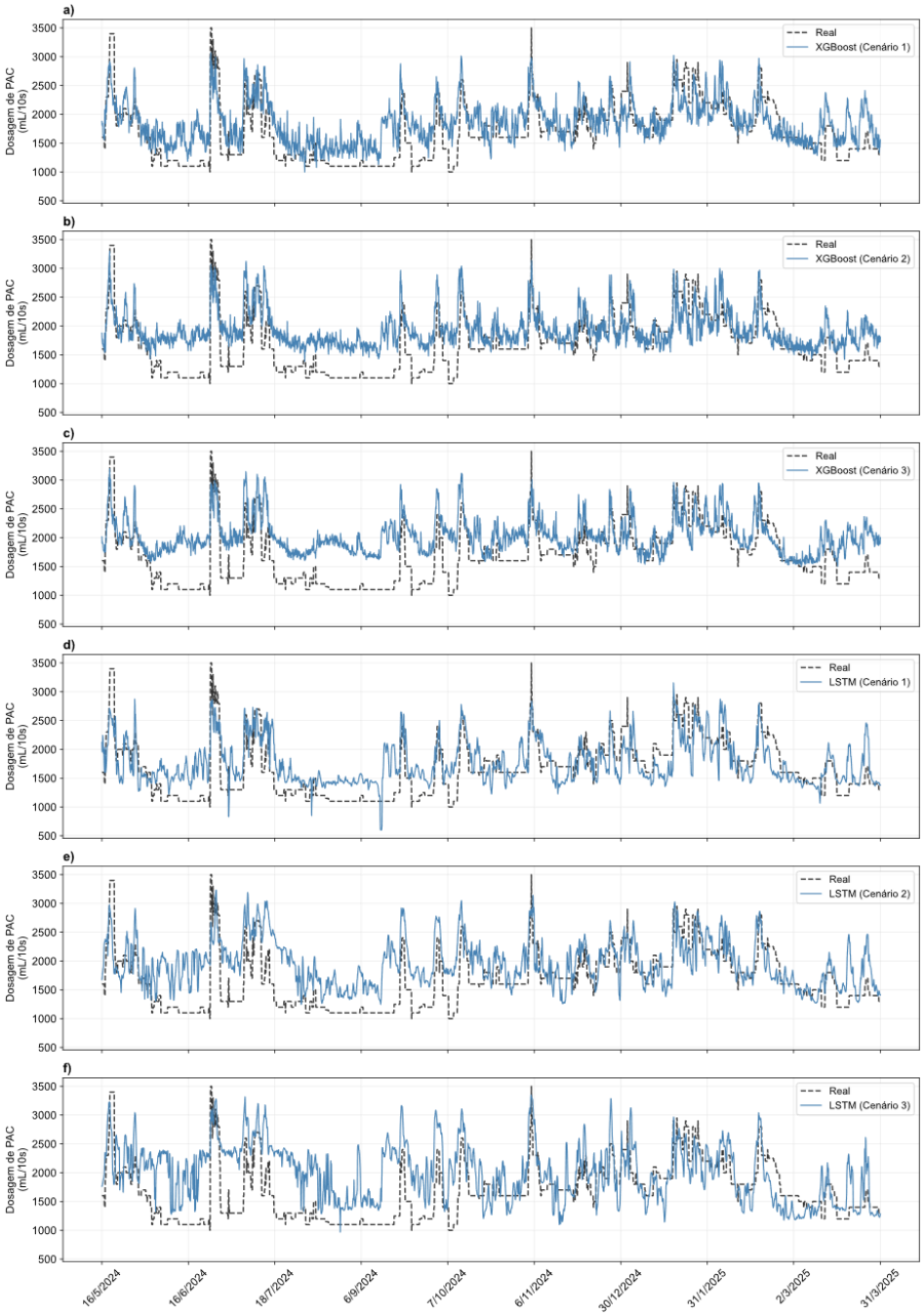
Fonte: elaborado pelos autores (2026).

Com isso, os modelos treinados com os dados dos Cenários 2 e 3 passaram a tentar compreender padrões em eventos de chuva e aumento de vazão dos rios que, por vezes, não resultavam em alterações na qualidade da água ou na decisão do operador. Portanto, conclui-se que a adição destes dados ambientais periféricos diluiu a importância estatística das

variáveis fundamentais do tratamento, indicando que o monitoramento interno da ETA traduz melhor a variância considerada para a tomada de decisão acerca da dosagem de coagulante.

Com relação à comparação visual das séries temporais de dosagem prevista e real, a Figura 7 ilustra os resultados obtidos com cada um dos modelos e cenários. É possível observar que os modelos acompanham o perfil geral dos dados, representando momentos de pico e de queda. Isso justifica o fato de os coeficientes de correlação (r) obtidos serem relativamente satisfatórios, com valores em média em torno de 0,62. Contudo, é evidente que, no período de maio a outubro de 2024, as previsões dos Cenários 2 e 3 tenderam a superestimar significativamente a dosagem, sugerindo valores acima do padrão observado.

Figura 7 – Comparação dos dados previstos e observados da dosagem de coagulante. Resultado das séries temporais obtidas com o modelo XGBoost com os Cenários 1 (a), 2 (b) e 3 (c) e com o modelo LSTM com os Cenários 1 (d), 2 (e) e 3 (f)

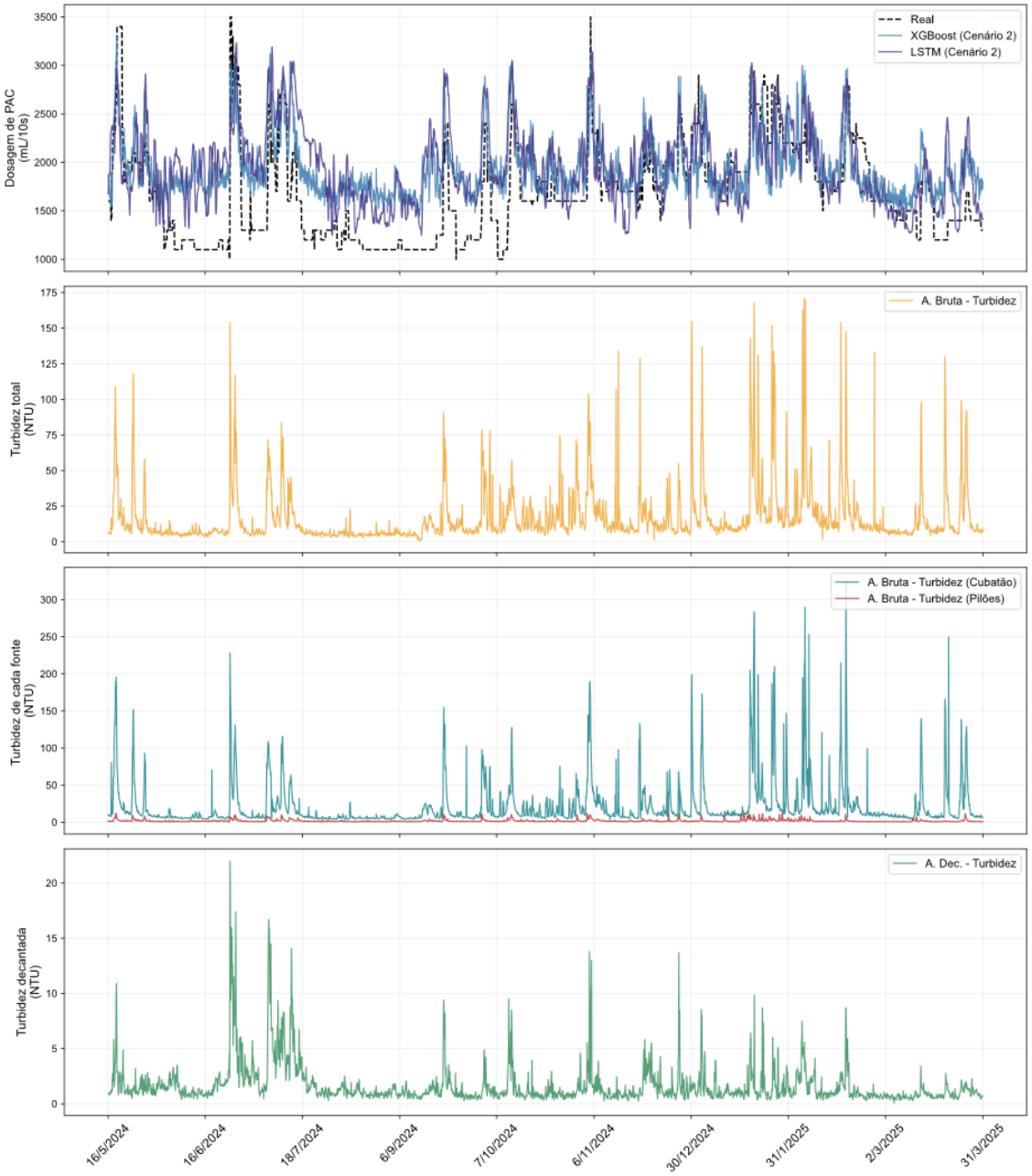


Fonte: elaborado pelos autores (2026)

De modo a compreender as razões para esse comportamento, avaliou-se visualmente as séries temporais das *features* de maior importância para os modelos nestes cenários (tendo por referência os resultados da *permutation feature importance* ilustrados na Figura 10). Em seguida, foram comparados os seus perfis registrados no período dos dados de teste.

Conforme retratado na Figura 8, é possível observar que houve um pico mais pronunciado de turbidez na água decantada durante este período, o que indica que possivelmente o processo sofreu uma degradação em sua eficiência na etapa de coagulação-floculação-decantação. No histórico, a dosagem apresentou registros abaixo do previsto pelos modelos, o que pode ser outra evidência de que o processo não estava operando no padrão anteriormente observado.

Figura 8 – Comparação visual do perfil de dosagem de coagulante prevista e observada com os perfis das séries temporais de turbidez da água bruta (medida na mistura, no rio Cubatão e no rio Pilões) e decantada

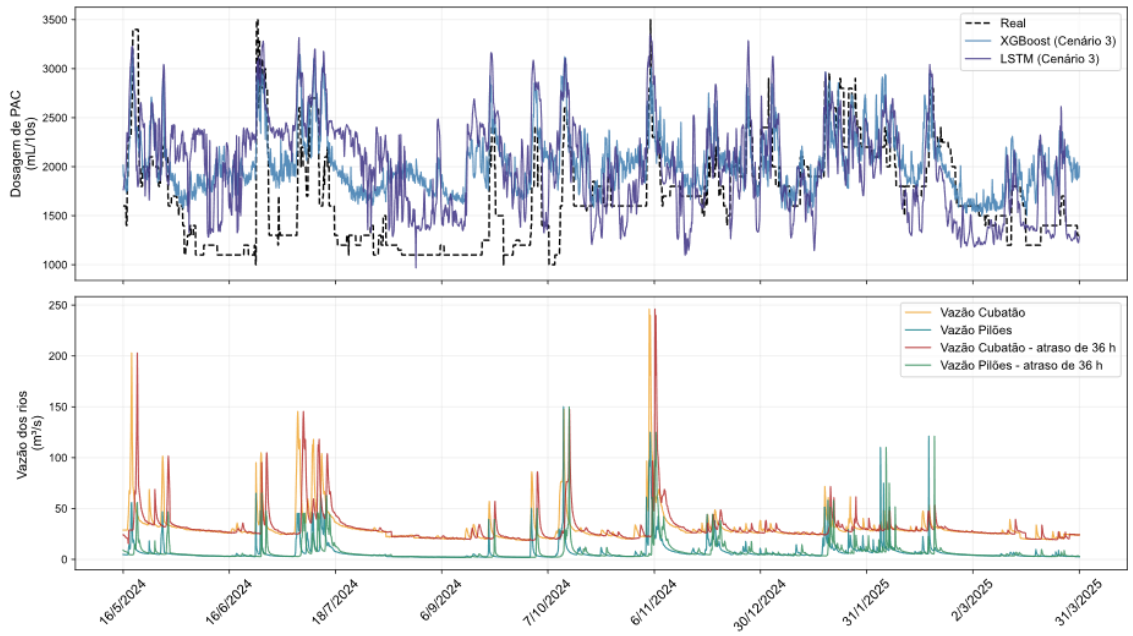


Fonte: elaborado pelos autores (2026)

Além disso, após este evento, observa-se que o período entre meados de julho de 2024 a meados de setembro do mesmo ano, foram observados registros de turbidez sem picos consideráveis, o que coincide também com um período de estabilidade na vazão dos rios. Com isso, a dosagem no processo foi mantida em valores abaixo do padrão médio. Contudo, os modelos novamente superestimaram este parâmetro, possivelmente também pelo fato de o processo ter operado neste período num padrão não observado anteriormente.

Considerando que nos Cenários 2 e 3 foram incluídos os dados de precipitação e vazão dos rios, na Figura 9 são ilustradas e comparadas as séries temporais de vazão, visto que os modelos atribuíram maior importância a essas *features* do que aos registros de chuva.

Figura 9 – Comparação visual do perfil de dosagem de coagulante prevista e observada com os perfis das séries temporais de vazão dos rios Cubatão e Pilões com e sem atraso temporal



Fonte: elaborado pelos autores (2026)

Observa-se que ocorreram elevações consideráveis das vazões, principalmente no período entre junho e julho de 2024, o que pode ter contribuído como um fator de influência na previsão de dosagens superestimadas.

Comparando-se o desempenho dos modelos em termos de estrutura dos algoritmos, no Cenário 1 o modelo LSTM demonstrou capacidade preditiva superior ao XGBoost, conforme mencionado. Isto pode ser explicado pela sua arquitetura voltada ao mapeamento de dependências temporais complexas como, por exemplo, em séries temporais não lineares (Weidman, 2019). No entanto, nos demais cenários, a inclusão de variáveis hidrometeorológicas e de atraso temporal revelou a alta sensibilidade da rede neural a ruídos e ao aumento da dimensionalidade, o que resultou em uma queda significativa do desempenho. As variáveis adicionadas de vazão dos rios e precipitação possuem variações particulares que, como visto na análise de correlação, apresentam uma relação complexa e multivariável com a qualidade da água e a dosagem. Com isso, o modelo buscou encontrar padrões temporais que, por vezes, não eram diretamente relacionados.

Além disso, destaca-se que redes neurais usualmente são mais sensíveis a variáveis que não seguem a tendência de inércia do sistema, o que pode piorar a sua precisão. Como observado pelos gráficos de séries temporais, a dosagem na ETA Cubatão apresentava a tendência de se manter constante por várias horas, dias ou até mesmo semanas, a depender da época do ano. Isso porque os operadores mantêm a dosagem em um valor fixo por longos períodos, enquanto as características da água oscilam constantemente a cada passo de tempo.

Para aplicações de *machine learning*, isto acaba sendo um obstáculo importante uma vez que os modelos captam e representam somente o padrão de decisão dos operadores, impedindo o estabelecimento de relações de fato baseadas nas interações físico-químicas da água. Na prática, como os modelos são alimentados com registros de qualidade da

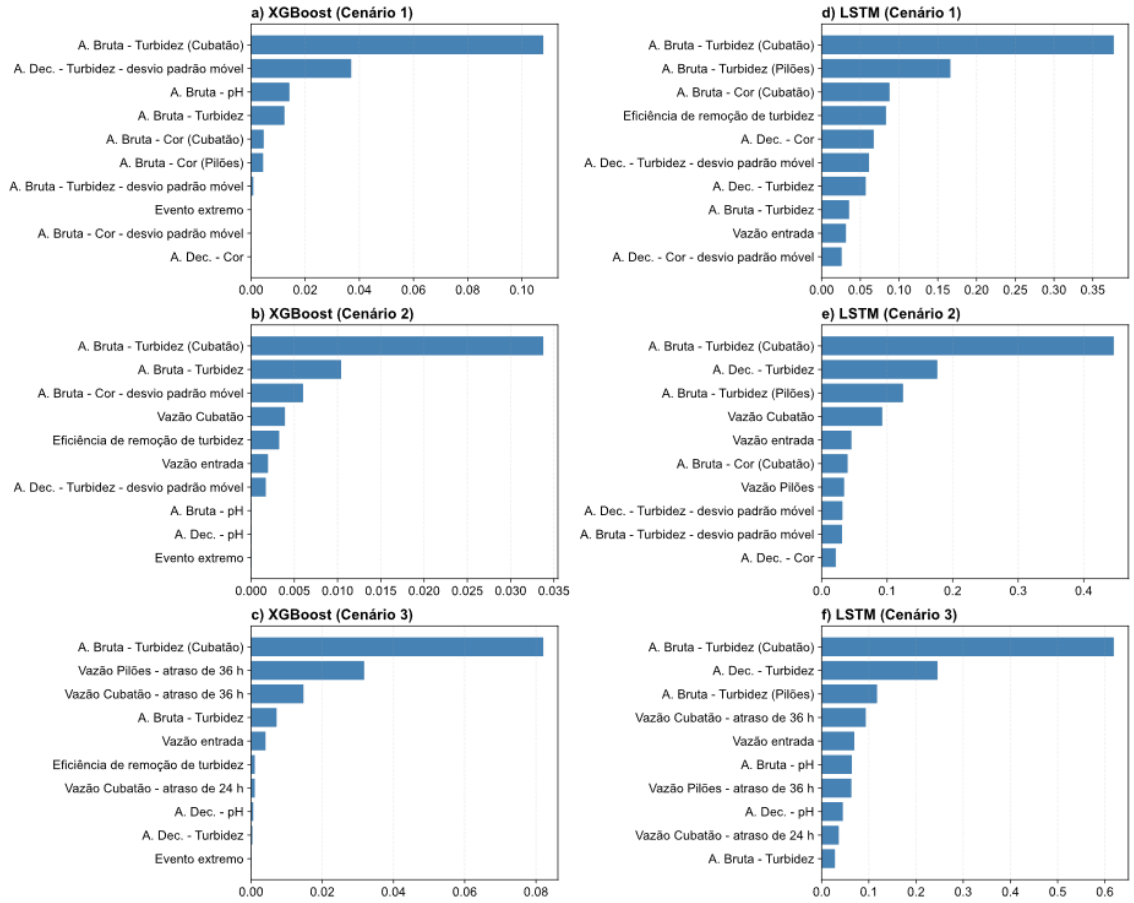
água, eles buscam compreender padrões de variação que não são efetivamente executadas na dosagem, implicando em uma alta penalização.

Com relação ao modelo XGBoost, este demonstrou maior robustez frente ao aumento da complexidade dos dados nos Cenários 2 e 3, tendo-se em vista que sua estrutura permite isolar ruídos através de seleção interna de *features* (Géron, 2021). Isto permitiu obter um maior coeficiente de determinação e menores erros. Contudo, assim como para o modelo LSTM, o Cenário 1 também foi o de melhor desempenho, demonstrando ainda existir uma sensibilidade frente ao aumento da dimensionalidade dos dados e possíveis ruídos.

Análise de importância das variáveis

Por fim, uma última análise avaliou a importância das variáveis preditoras empregando o método de *Permutation Feature Importance*. A Figura 10 ilustra os resultados obtidos para cada um dos cenários e modelos aplicados.

Figura 10 – Resultado da análise de *Permutation Feature Importance* com os modelos XGBoost e LSTM aplicados a todos os cenários de estudo



Fonte: elaborado pelos autores (2026)

É possível observar que a variável turbidez do rio Cubatão foi a de maior importância para todos os casos estudados, sendo seguida por informações de turbidez da água decantada (incluindo desvio padrão móvel) e demais medidas de turbidez da água bruta (isto é, do rio Pilões e da mistura dos dois rios). Segundo relato dos operadores, a turbidez, juntamente com a cor, é de fato o parâmetro de maior atenção por parte da equipe para controle da dosagem. Ademais, assim como observado pelo histórico de registros de turbidez (Figura 8), a água do rio Cubatão é a que mais contribui com o aumento da turbidez da água captada devido à significativa perda da sua mata ciliar e presença de diversas atividades antrópicas na sua região.

Como determinado por parte dos modelos treinados, a água decantada é, igualmente, uma variável relevante para a determinação da dosagem. Na prática, isso também é verdade, pois os operadores avaliam se a dose aplicada está

sendo efetiva para uma adequada coagulação e floculação, o que garante a manutenção de níveis de turbidez em limites aceitáveis pela regulamentação.

Embora não tenham tido um destaque tão pronunciado nos *ranks* de *feature importance*, informações de vazões dos rios apareceram com frequência como uma das mais relevantes nos Cenários 2 e 3 para ambos os modelos. Ressalta-se, principalmente, que no Cenário 3, com a inclusão do atraso temporal das vazões, as *features* que incorporavam este atraso foram elencadas como de maior importância do que as *features* sem atraso, o que corrobora a análise de correlação realizada previamente. Isto evidencia que estas variáveis apresentam potencial de apoio na implementação de ferramentas de previsão de parâmetros operacionais em ETAs.

CONCLUSÕES

O presente estudo buscou investigar se a incorporação de dados de precipitação pluviométrica e vazões dos rios de captação de uma ETA teriam potencial de melhorar o desempenho de modelos de *machine learning* para previsão da dosagem de coagulante. Contudo, os resultados obtidos demonstraram uma queda na acurácia dos modelos quando comparados a cenários de previsão baseados exclusivamente em informações de qualidade da água.

Infere-se que a causa para isso foi, principalmente, a complexidade envolvida no padrão de causalidade do sistema. A precipitação usualmente interfere nas vazões dos rios, bem como na qualidade da água, a depender da sua intensidade. No entanto, esses eventos, por vezes, não implicam necessariamente em alterações significativas que interfiram na operação da ETA. Em situações em que há amortecimento hidrológico, por exemplo, o operador pode não observar um efeito considerável que resulte em um ajuste na dosagem.

Além disso, a própria dinâmica de decisão de controle da dosagem realizada pelos operadores da estação também contribuiu para os resultados. O histórico de dados com dosagens em valores fixos por longos períodos, enquanto as características da água oscilaram constantemente a cada passo de tempo, evidenciou que o critério de controle é subjetivo, não seguindo um padrão matemático claro. Com isso, os modelos de *machine learning* apresentaram dificuldades em representar o padrão de dosagem com acurácia. Isto demonstra que, para o sucesso da implementação de ferramentas deste tipo, é imprescindível primeiramente ter uma base histórica de dados que siga padrões técnicos de garantia da qualidade das informações.

Por fim, conclui-se que o aprimoramento da capacidade preditiva dos modelos com a inclusão de dados de chuva e vazão dos rios não é garantido, tendo em vista que a relação entre essas informações e a qualidade da água de captação é multifatorial. A depender do histórico de dados, das características da bacia, atividades antrópicas na região e demais fatores que influenciam os padrões hidrológicos observados, os modelos de previsão (como o LSTM e o XGBoost testados no presente estudo) podem interpretar estas informações adicionais como ruídos. Ademais, se na prática operacional os critérios de controle da dosagem não forem bem estabelecidos, fatores de subjetividade e empirismo influenciam o padrão de ajuste a ser representado, o que afeta o desempenho dos modelos.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (Processos n.º 305472/2025-9 – C.S. e n.º 312247/2022-2 – N.P.) e ao Programa MAI/DAI pelo apoio financeiro, à HydroInfo e à Companhia Catarinense de Águas e Saneamento (CASAN) pela parceria no projeto e especificamente à CASAN pela disponibilização de dados referentes à ETA de estudo. Além disto, agradecem também ao VLAB (UFSC) pela disponibilização de acesso ao sistema de computação de alto desempenho.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ACHITE, M. *et al*. Modeling and optimization of coagulant dosage in water treatment plants using hybridized random forest model with genetic algorithm optimization. *Environment, Development and Sustainability*, vol. 25, no 10, p. 11189–11207, 2023.

2. ACHITE, M. *et al*. Modeling the optimal dosage of coagulants in water treatment plants using various machine learning models. *Environment, Development and Sustainability*, vol. 26, no 2, p. 3395–3421, 2024.

3. AKIBA, T. *et al*. *Optuna*: A Next-generation Hyperparameter Optimization Framework. 2019. Disponível em: http://arxiv.org/abs/1907.10902.

4. BOUMEZBEUR, H. *et al*. Modelling coagulant dosage in drinking water treatment plant using advance machine learning model: Hybrid extreme learning machine optimized by Bat algorithm. *Environmental Science and Pollution Research*, vol. 30, no 28, p. 72463–72483, 2023.

5. BRESSANE, A. *et al*. A Non-Hybrid Data-Driven Fuzzy Inference System for Coagulant Dosage in Drinking Water Treatment Plant: Machine-Learning for Accurate Real-Time Prediction. *Water* (Switzerland), vol. 15, no 6, 2023.

6. CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. 2016, New York, NY, USA. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. p. 785–794.

7. CHIAVOLA, A. *et al*. A combined experimental-modeling approach for turbidity removal optimization in a coagulation–flocculation unit of a drinking water treatment plant. *Journal of Process Control*, vol. 130, 2023.

8. COHEN, J. *Statistical power analysis for the behavioral sciences*. 2. ed. New York: Lawrence Erlbaum Associates, 1988.

9. DENG, Haowen; ZHOU, Youyou; WANG, Lin; et al. Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC Medical Informatics and Decision Making*, v. 21, n. 1, p. 338, 2021.

10. DI BERNARDO, L.; DI BERNARDO, A.; CENTURIONE FILHO, P. L. *Ensaios de tratabilidade de água e dos resíduos gerados em estações de tratamento de água*. São Carlos: Rima, 2002.

11. FENG, L. *et al*. Interpretable prediction of coagulant dosage in drinking water treatment plant based on automated machine learning and SHAP method. *Journal of Water Process Engineering*, vol. 75, 2025.

12. GÉRON, A. *Mãos à obra*: aprendizado de máquina com Scikit-Learn, Keras e TensorFlow. 2. ed. Rio de Janeiro: Alta Books, 2021.

13. GHASEMI, M. *et al*. Predicting coagulation–flocculation process for turbidity removal from water using graphene oxide: a comparative study on ANN, SVR, ANFIS, and RSM models. *Environmental Science and Pollution Research*, vol. 29, no 48, p. 72839–72852, 2022.

14. GIRARDI, R. *et al*. Alteração da qualidade da água de rios em período chuvoso em bacia hidrográfica com diferentes usos do solo no Sul do Brasil. *Revista Brasileira de Recursos Hidricos*, vol. 21, no 3, p. 514–524, 2016.

15. GOMES, L. S. *et al*. Coagulant Dosage Determination in a Water Treatment Plant Using Dynamic Neural Network Models. *International Journal of Computational Intelligence and Applications*, vol. 14, no 3, 2015.

16. HAGHIRI, S.; DAGHIGHI, A.; MOHARRAMZADEH, S. Optimum coagulant forecasting by modeling jar test experiments using ANNs. *Drinking Water Engineering and Science*, vol. 11, no 1, p. 1–8, 2018.

17. HANSUN, S.; WICAKSANA, A.; KHALIQ, A. Q. M. Multivariate cryptocurrency prediction: comparative analysis of three recurrent neural networks approaches. *Journal of Big Data*, v. 9, n. 1, p. 50, 2022.

18. HEDDAM, S.; BERMAD, A.; DECHEMI, N. ANFIS-based modelling for coagulant dosage in drinking water treatment plant: A case study. *Environmental Monitoring and Assessment*, vol. 184, no 4, p. 1953–1971, 2012.

19. HELSEL, D. R. *et al*. *Statistical Methods in Water Resources*. U.S. Geological Survey, 2020. (Techniques and Methods book 4, chap. A3).

20. HUANG, R. *et al*. *Machine learning in natural and engineered water systems*. Elsevier Ltd, 2021.

21. INMET - Instituto Nacional de Meteorologia. Manual de observações meteorológicas. Ministério da Agricultura e do Abastecimento. Brasília, 3ª edição. 1999.

22. JAKOVLJEVIC, A.; CHARLIN, L.; BARBEAU, B. Applying Recurrent Neural Networks and Blocked Cross-Validation to Model Conventional Drinking Water Treatment Processes. *Water* (Switzerland), vol. 16, no 7, 2024.

23. JAYAWEERA, C. D.; AZIZ, N. An efficient neural network model for aiding the coagulation process of water treatment plants. *Environment, Development and Sustainability*, vol. 24, no 1, p. 1069–1085, 2022.

24. JONES, Amber Spackman; JONES, Tanner Lex; HORSBURGH, Jeffery S. Toward automating post processing of aquatic sensor data. *Environmental Modelling and Software*, v. 151, 2022.

25. KIM, J. *et al*. Optimizing coagulant dosage using deep learning models with large-scale data. *Chemosphere*, vol. 350, 2024.

26. KIM, C. M.; PARNICHKUN, M. MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant. *Journal of Water Supply: Research and Technology - AQUA*, vol. 66, no 1, p. 49–61, 2017a.

27. KIM, C. M.; PARNICHKUN, M. Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system. *Applied Water Science*, vol. 7, no 7, p. 3885–3902, 2017b.

28. KONG, X. *et al*. Reservoir water quality deterioration due to deforestation emphasizes the indirect effects of global change. *Water Research*, vol. 221, 2022.

29. LI, L. *et al*. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal*, vol. 405, 2021.

30. LI, S. *et al*. Ecosystem-inspired model and artificial intelligence predicts pollutant consumption capacity by coagulation in drinking water treatment. *Environmental Chemistry Letters*, vol. 21, no 5, p. 2499–2508, 2023.

31. LIBÂNIO, M. *Fundamentos de qualidade e tratamento de água*. 3. ed. Campinas, SP: Editora Átomo, 2010.

32. LIN, S. *et al*. Coagulant dosage determination using deep learning-based graph attention multivariate time series forecasting model. *Water Research*, vol. 232, 2023a.

33. LIN, S. *et al*. Comparing artificial and deep neural network models for prediction of coagulant amount and settled water turbidity: Lessons learned from big data in water treatment operations. *Journal of Water Process Engineering*, vol. 54, 2023b.

34. LIU, Y. *et al*. An Auto-Adjustable and Time-Consistent Model for Determining Coagulant Dosage Based on Operators' Experience. IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no 9, p. 5614–5625, 2021.

35. LOPES, A. C. Contribuição financeira por usuários de recursos hídricos beneficiados por unidades de conservação: estudo de caso na bacia hidrográfica do rio Vargem do Braço, em Santo Amaro da Imperatriz, Santa Catarina. 2020. - Centro Tecnológico, Universidade Federal de Santa Catarina, 2020.

36. LOWE, M.; QIN, R.; MAO, X. A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. *Water* (Switzerland), vol. 14, no 9, 2022.

37. LUO, H. *et al*. Application of a New Architecture Neural Network in Determination of Flocculant Dosing for Better Controlling Drinking Water Quality. *Water* (Switzerland), vol. 14, no 17, 2022.

38. NARGES, S. *et al*. Prediction of the optimal dosage of coagulants in water treatment plants through developing models based on artificial neural network fuzzy inference system (ANFIS). 2021.

39. NAVIDI, W. *Statistics for Engineers and Scientists*. 3. ed. New York: McGraw-Hill, 2011.

40. NEVES, F. F. *Avaliação da influência de sedimentos em suspensão em populações de Corbicula fluminea (MÜLLER, 1774) (Mollusca, Bivalvia, Corbiculidae)*: análise comparativa em duas microbacias hidrográficas do Estado de São Paulo, Brasil. 2010. 162 p. Tese (Doutorado em Ciência, área: Biologia Comparada) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 2010.

41. ORTIZ-LOPEZ, C.; BOUCHARD, C.; RODRIGUEZ, M. Machine learning models with potential application to predict source water quality for treatment purposes: a critical review. *Taylor and Francis Ltd*., 2022.

42. PANG, H. *et al*. Time series-based machine learning for forecasting multivariate water quality in full-scale drinking water treatment with various reagent dosages. *Water Research*, vol. 268, 2025.

43. QUAGHEBEUR, W. Hybrid models of dynamical systems: neural differential equations, Shapley value analysis and illustrations in water systems. 2023. PhD Thesis - Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium, 2023.

44. RAMOS, K. A. A distribuição das chuvas na bacia do Rio Cubatão do Sul (SC) e seu entorno: reflexos na captação de água para o abastecimento da Grande Florianópolis. 2023. - Centro de Filosofia e Ciências Humanas, Universidade Federal de Santa Catarina, 2023.

45. SANTOS, T. E. M., MONTENEGRO, A. A. A. Erosividade e padrões hidrológicos de precipitação no Agreste Central pernambucano. *Revista Brasileira de Engenharia Agrícola e Ambiental*. v. 16, n. 8, p. 871-880, 2012.

46. SCHNEIDER, M. Y. *et al*. Hybrid modelling of water resource recovery facilities: status and opportunities. *Water Science and Technology*, vol. 85, no 9, p. 2503–2524, 2022.

47. SINISA. Relatório dos Serviços de Abastecimento de Água. Sistema Nacional de Informações em Saneamento Básico - SINISA 2024 (ano de referência 2023). 2025.

48. SUQUET, J. *et al*. Assessing the effect of catchment characteristics to enhanced coagulation in drinking water treatment: RSM models and sensitivity analysis. *Science of the Total Environment*, vol. 799, 2021.

49. TORFS, E. *et al*. The transition of WRRF models to digital twin applications. *Water Science and Technology*, vol. 85, no 10, p. 2840–2853, 2022.

50. WANG, D.; CHANG, X.; MA, K. Predicting flocculant dosage in the drinking water treatment process using Elman neural network. *Environmental Science and Pollution Research*, vol. 29, p. 7014–7024, 2022.

51. WANG, H.; ASEFA, T.; THORNBURGH, J. Integrating water quality and streamflow into prediction of chemical dosage in a drinking water treatment plant using machine learning algorithms. *Water Supply*, vol. 22, no 3, p. 2803–2815, 2022.

52. WEIDMAN, S. Deep learning from scratch: building with Python from first principles. Sebastopol: O'Reilly Media, 2019.

53. WESCHENFELDER, A. B.; PICKBRENNER, K.; PINTO, E. J. de A. Atlas Pluviométrico do Brasil: equações intensidade-duração-frequência (desagregação de precipitações diárias): Município de Santo

Amaro da Imperatriz, Estação Pluviométrica: ETA CASAN MONTANTE, código 02748004. Porto Alegre: [s. d.], 2016.

54. WISCHMEIER, W. H.; SMITH, D. D. *Predicting rainfall erosion losses*: a guide to conservation planning. Washington: U.S. Department of Agriculture, 1978. Agriculture Handbook Nº 537.

55. WU, G. De; LO, S. L. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Engineering Applications of Artificial Intelligence*, vol. 21, no 8, p. 1189–1195, 2008.

56. ZHANG, K. *et al*. Machine learning approaches to predict coagulant dosage in water treatment plants. *International Journal of System Assurance Engineering and Management*, vol. 4, no 2, p. 205–214, 2013.