MODELAGEM DE DADOS PARA PREVENÇÃO DE NÃO CONFORMIDADES DE DBO EM EFLUENTES INDUSTRIAIS: UMA ABORDAGEM PARA SUPORTE À DECISÃO

DATA MODELING FOR THE PREVENTION OF BOD NON-CONFORMITIES IN INDUSTRIAL EFFLUENTS: A DECISION-SUPPORT APPROACH

LUANA NASCIMENTO BARBOSA SOUZA
UNIVERSIDADE FEDERAL DA BAHIA-UFBA

KARLA PATRICIA SANTOS OLIVEIRA RODRIGUEZ ESQUERRE
UNIVERSIDADE FEDERAL DA BAHIA-UFBA

Comunicação:
O XIV SINGEP foi realizado em conjunto com a 14th Conferência Internacional do CIK (CYRUS Institute of Knowledge) e com a SPARK (Society for Promoting Academic Research and Knowledge), em formato híbrido, com sede presencial na UNINOVE - Universidade Nove de Julho, no Brasil.

Agradecimento à orgão de fomento:
As autoras expressam sua gratidão ao Gamma Lab pelo apoio institucional e pelo ambiente propício à colaboração científica. Agradecem também à disciplina de Ciência de Dados na Engenharia, cujos aportes práticos foram fundamentais para a estruturação do pipeline preditivo e a implementação dos modelos orientados a dados. O presente estudo contou com o financiamento da CAPES, mediante concessão de bolsa de mestrado, e vincula-se à Bolsa de Produtividade em Pesquisa do CNPq (Chamada CNPq Nº 18/2024).

MODELAGEM DE DADOS PARA PREVENÇÃO DE NÃO CONFORMIDADES DE DBO EM EFLUENTES INDUSTRIAIS: UMA ABORDAGEM PARA SUPORTE À DECISÃO

Objetivo do estudo
Avaliar e comparar diferentes algoritmos de aprendizado de máquina para a prevenção de não conformidades de DBO em efluente industrial, identificando o modelo mais adequado na detecção de eventos críticos para suporte à decisão na gestão do tratamento.

Relevância/originalidade
Explora a lacuna do tempo analítico da DBO no tratamento de efluentes. Destaca-se por testar estratégias de aprendizado de máquina sob validação temporal, integrando e comparando redução de dimensionalidade por PCA, variação de complexidade algorítmica e reamostragem sintética no suporte à decisão.

Metodologia/abordagem
Avaliaram-se Regressão Logística, Random Forest e Gradient Boosting sob validação cruzada temporal, incorporando redução por PCA, reamostragem via SMOTE e defasagem pelo Tempo de Residência Hidráulica de 48h. A classe crítica de DBO foi definida pelo Percentil 90 da série histórica.

Principais resultados
A Random Forest por validação temporal obteve o melhor equilíbrio sensibilidade-precisão, com 25% de recall na classe alerta . Abordagens com SMOTE e Regressão Logística geraram grande volume de alarmes falsos, enquanto Gradient Boosting e PCA apresentaram baixo desempenho, inviabilizando suas aplicações.

Contribuições teóricas/metodológicas
Demonstra a importância da validação temporal em séries industriais dinâmicas e evidencia que a reamostragem por SMOTE compromete a coerência física em eventos raros. Valida a incorporação do TRH na modelagem e a robustez da Random Forest sem dados sintéticos.

Contribuições sociais/para a gestão
Oferece um sensor virtual complementar para apoio à decisão, permitindo ações preventivas no efluente industrial. Fortalece a governança ambiental ao conter alarmes falsos, preservando a confiabilidade do sistema e mitigando riscos de não conformidade nos corpos hídricos.

Palavras-chave: Tratamento de efluentes, Demanda Bioquímica de Oxigênio, Aprendizado de máquina, Suporte à decisão, Sustentabilidade industrial

DATA MODELING FOR THE PREVENTION OF BOD NON-CONFORMITIES IN INDUSTRIAL EFFLUENTS: A DECISION-SUPPORT APPROACH

Study purpose
Evaluate and compare different machine learning algorithms to prevent BOD non-conformities in industrial effluent, identifying the most suitable model for detecting critical events to support decision-making in treatment management.

Relevance / originality
Explores the BOD analytical time lag in effluent treatment. It stands out by testing machine learning strategies under temporal validation, integrating and comparing dimensionality reduction via PCA, algorithmic complexity variation, and synthetic resampling for decision support.

Methodology / approach
Logistic Regression, Random Forest, and Gradient Boosting were evaluated under time-series cross-validation, incorporating PCA reduction, SMOTE resampling, and a 48-hour Hydraulic Retention Time lag. The critical BOD class was defined by the 90th percentile of the historical series.

Main results
Random Forest with temporal validation achieved the best sensitivity-precision balance, with 25% recall for the alert class. SMOTE and Logistic Regression generated excessive false alarms, whereas Gradient Boosting and PCA showed poor performance, rendering them unfeasible.

Theoretical / methodological contributions
Demonstrates the importance of temporal validation in dynamic industrial time series and shows that SMOTE resampling compromises physical coherence in rare events. Validates HRT incorporation into modeling and the robustness of Random Forest without synthetic data.

Social / management contributions
Offers a complementary virtual sensor for decision support, enabling preventive actions in industrial effluent management. Strengthens environmental governance by containing false alarms, preserving system reliability, and mitigating non-conformity risks in receiving water bodies.

Keywords: Effluent treatment, Biochemical Oxygen Demand, Machine learning, Decision support, Industrial sustainability

MODELAGEM DE DADOS PARA PREVENÇÃO DE NÃO CONFORMIDADES DE DBO EM EFLUENTES INDUSTRIAIS: UMA ABORDAGEM PARA SUPORTE À DECISÃO

1 Introdução

No setor de papel e celulose, o gerenciamento de efluentes exige o controle contínuo do descarte para assegurar a conformidade com as restrições operacionais e ambientais (Bellamoli et al., 2023). O tratamento em lagoas aeradas envolve interações microbiológicas, variações térmicas e dinâmicas biocinéticas não lineares (Oliveira-Esquerre et al., 2004; Sundui et al., 2021; Zamfir et al., 2025; Wang, 2021), condições que reduzem a precisão de modelos fenomenológicos tradicionais. Embora o monitoramento dependa de parâmetros como a Demanda Bioquímica de Oxigênio (DBO), o tempo analítico requerido pelos ensaios laboratoriais dificulta intervenções operacionais imediatas (Chow et al., 2018). Nesse contexto, modelos de aprendizado de máquina atuam como ferramentas de suporte à decisão para estimar cenários de risco antes da obtenção dos resultados analíticos (Wang et al., 2021). Algoritmos baseados em conjuntos de árvores, como Random Forest e Gradient Boosting, incorporam defasagens temporais estruturadas a partir do Tempo de Residência Hidráulica (TRH) para correlacionar a carga afluente à qualidade do efluente final (Bellamoli et al., 2023; Wang et al., 2021; Gulshin e Kuzina, 2024).

Apesar do potencial dessas ferramentas, os episódios de não conformidade constituem eventos minoritários nas bases de dados operacionais, gerando um desbalanceamento de classes que induz os classificadores a priorizarem a condição padrão de operação (Qiu et al., 2024). Embora a técnica SMOTE seja comum para o reequilíbrio de amostras em cenários controlados, sua interpolação sintética pode ignorar a dependência temporal e a inércia física das séries históricas ao conectar registros distantes no tempo (Yang, 2025). Essa geração artificial de dados sem aderência ao comportamento do processo favorece o *overfitting* e a ocorrência de alarmes falsos na rotina industrial. Diante dessas limitações, este estudo avalia a estruturação de um Sistema de Alerta Precoce para eventos de DBO em uma ETAR de papel e celulose. A abordagem compara o desempenho dos algoritmos Random Forest e Histogram-based Gradient Boosting, avaliando o impacto de estratégias para o tratamento do desbalanceamento de classes por meio de um *pipeline* temporal que integra seleção de atributos, engenharia de variáveis via TRH, análise do SMOTE e ajuste de pesos algorítmicos para suporte à tomada de decisão.

O estudo utiliza as informações coletadas na ETAR de uma planta de papel e celulose ao longo de 1.427 dias seguidos, entre setembro de 1996 e julho de 2000, com base nos registros de Oliveira-Esquerre et al. (2004). A estrutura do tratamento reúne decantadores primários em paralelo que alimentam uma lagoa aerada e outras cinco lagoas sem aeração dispostas em sequência. O Tempo de Residência Hidráulica (TRH) da unidade é de dois dias, valor que serviu de guia para a criação das variáveis com atraso temporal (Oliveira-Esquerre et al., 2004). Durante esse intervalo, a vazão média medida foi de 67.358 m³/dia, com metade dos dados situados na faixa de 62.056 a 76.541 m³/dia.

Para a DBO de saída, os valores de média e mediana ficaram em 85 ppm, com pico de 187 ppm e 75% das leituras abaixo de 101 ppm. Por fim, a variável resposta e possíveis preditoras foram avaliadas quanto às suas estatísticas descritivas conforme resumido na Tabela 1.

Tabela 1. Estatísticas descritivas e avaliação de integridade das variáveis físico-químicas e operacionais do sistema de tratamento. Elaboração própria, 2026.

Variável	Descrição	Dados Faltantes (%)	Média Total	Desvio-Padrão	Mínimo	25%	50% (Mediana)	75%	Máximo

FR	Vazão de Entrada em m3/dia	0	67358	11593	4474	62056	66538	76541	97850
BODin	DBO de Entrada em ppm	6,2	245	46	41	219	245	271	449
CODin	DQO de Entrada em ppm	6,2	561	104	136	502	563	621	925
Tin	Temperatura de Entrada em °C	32	45	3,07	28	44	46	47	50
SSin	Sólidos Suspensos de Entrada em ppm	60	149	86	12	90	133	186	591
pHin	pH de Entrada	3,7	7,4	1,2	0,85	6,8	7,1	7,5	12
NAmin	Nitrogênio Amoniacal de Entrada em ppm	53	2,4	1,8	0	1,2	2,3	3,3	20
RF	Chuva em mm	17	4,8	11	0	0	0,19	3,96	175
NNin	Nitrato e Nitrito de Entrada em ppm	80	1,4	0,90	0,03	0,9	1,3	1,8	7,3
Pin	Fósforo de Entrada em ppm	81	1,4	1,0	0,05	0,99	1,2	1,6	14
Condin	Condutividade de Entrada a 25°C em µS/cm	3,9	1530	378	379	1305	1459	1667	5810
Colin	Cor do Efluente de Entrada em Pt-Co	3,6	464	124	41	393	477	532	1317
Pulp	Produção de Celulose em t/dia	7,2	884	159	-48	861	918	963	1112
Pap	Produção de Papel em t/dia	6,4	1042	94	382	1003	1053	1101	1304
BODout	DBO de Saída em ppm	6,0	85	25	16	69	85	101	187
CODout	DQO de Saída em ppm	5,9	315	73	105	281	312	350	865
SSout	Sólidos Suspensos de Saída em ppm	86	52	19	10	39	52	62	130
Colout	Cor do Efluente de Saída em Pt-Co	3,0	418	156	32	364	443	513	1138
Tout	Temperatura de Saída da Lagoa em °C	13	34	2,3	25	33	35	36	41
Condout	Condutividade de Saída a 25°C em uS/cm	3,1	1550	301	222	1352	1490	1662	2890
FRout	Vazão de Saída em m3/dia	0,14	67657	11667	4474	62211	66769	76651	106941

Complementando a caracterização da base de dados, procedeu-se à avaliação das propriedades estatísticas das variáveis por meio da verificação sistemática da assimetria e da curtose. Esse diagnóstico permitiu identificar a dispersão dos dados e a probabilidade de

ocorrência de valores discrepantes na série temporal. Do mesmo modo, analisou-se a correlação entre os preditores contínuos para aferir o nível de colinearidade existente na matriz de dados, viabilizando a eliminação de redundâncias sensoriais que pudessem introduzir ruído ou influenciar negativamente a fase de treinamento dos classificadores.

2 Referencial Teórico

As Estações de Tratamento de Águas Residuárias (ETARs) representam sistemas complexos e não lineares, caracterizados por flutuações constantes na vazão de entrada, na carga orgânica de poluentes e nas condições hidráulicas da planta (Wang, 2021). O monitoramento do desempenho operacional dessas estações depende da determinação de parâmetros indicadores como a Demanda Química de Oxigênio (DQO) e a Demanda Bioquímica de Oxigênio (DBO), sendo a DBO uma variável cuja medição convencional por meio de métodos laboratoriais exige um tempo analítico de 5 dias em média (Dikmen, 2025). Esse atraso pode inviabilizar medidas corretivas a tempo de evitar descartes fora dos padrões regulatórios (Dikmen, 2025).

Para mitigar as limitações do monitoramento analítico reativo, a integração de modelos de aprendizado de máquina surge como uma alternativa viável para prever a qualidade do efluente (Zamfir, 2025). Ao extrair relações matemáticas não lineares das bases de dados, os algoritmos atuam como sensores virtuais de suporte à decisão para prever parâmetros de difícil medição instantânea (Wang, 2021). Na estruturação de sistemas de suporte à decisão, a formulação do problema pode transitar da estimativa contínua para um pipeline de classificação supervisionada focado em categorizar estados de conformidade (Bellamoli, 2023). Essa abordagem permite identificar a iminência de eventos críticos antes que o descarte final de um determinado material ocorra (Zamfir, 2025; Dikmen, 2025; Cheng, 2025).

Processos de tratamento de efluentes são dinâmicos e envolvem múltiplos parâmetros, de modo que a eficácia prática de um classificador dinâmico em ETARs requer consideração da dinâmica hidráulica do sistema, uma vez que, existem defasagens temporais entre as características medidas na entrada e os parâmetros monitorados na saída da planta (Wang, 2021). Dessa forma, variáveis defasadas devem ser incorporadas ao treinamento dos modelos com base no Tempo de Residência Hidráulica (TRH) da lagoa aerada. O alinhamento cronológico das variáveis por meio de atrasos temporais consistentes preserva o efeito de memória do processo biológico (Wang, 2021).

O desenvolvimento de classificadores para o controle de efluentes geralmente enfrenta o desafio do desbalanceamento de dados, uma vez que episódios críticos de não conformidade operacional constituem eventos minoritários no histórico da estação (Zamfir , 2025). Embora técnicas de sobreamostragem sintética como o SMOTE sejam comuns para equilibrar classes, sua aplicação pode introduzir consequências negativas uma vez que, em sistemas de suporte à decisão em ETARs, a síntese artificial de dados minoritários pode reduzir a precisão do classificador (Bellamoli, 2023). A redução de precisão provoca uma elevação de alarmes falsos, o que compromete a confiabilidade operacional e gera custos desnecessários com falsas não conformidades (Wang et al., 2025).

A adoção do percentil 90 para a separação entre a classe normal e a de alerta fundamenta-se em abordagens da literatura, permitindo avaliar o desempenho do modelo de forma independente de padrões de conformidade específicos (López et al., 2012). Conforme discutido em estudos de monitoramento de corpos hídricos (Jung & Krueger, 2020), o estado físico-químico de variáveis de qualidade de água é classificado sistematicamente por meio de métricas como médias ou percentis aplicados sobre escalas regulatórias predefinidas. O uso do percentil 90 permite modelar o comportamento da cauda superior de poluentes, uma vez que

avaliações de conformidade de efluentes requerem estimativas da variabilidade em torno de estatísticas de interesse como os percentis 90 ou 95 (López et al., 2012). A calibração de modelos em limites estatísticos elevados fornece dados estatisticamente consistentes, permitindo que o sistema de alerta mapeie a probabilidade de falhas antes que elas ocorram, o que apoia a tomada de ação proativa no controle de processos (López et al., 2012; Jung & Krueger, 2020).

Para viabilizar essa modelagem sobre percentis e variáveis de interesse, o desempenho de algoritmos é avaliado a partir de estruturas paramétricas e não paramétricas (Jung & Krueger, 2020). Enquanto abordagens paramétricas assumem um formato de distribuição predefinido para a população, os métodos não paramétricos oferecem a flexibilidade de construir fronteiras de decisão sem a necessidade de suposições prévias sobre os dados, embora sua precisão dependa da representatividade da amostragem (Jung & Krueger, 2020). De forma complementar, a Análise de Componentes Principais (PCA) atua no pré-processamento como uma técnica não supervisionada de redução de dimensionalidade (Zamfir et al., 2025; Corominas et al., 2018). Ao projetar atributos de acordo ao grau de correlações em um espaço reduzido, o PCA diminui a redundância informacional e retém a variância do sistema, simplificando a estrutura dos dados para o aprendizado supervisionado posterior (Wang et al., 2025; Corominas et al., 2018).

Nesse contexto do aprendizado supervisionado para classificação em ETARs, os modelos baseados em árvores são empregados pela capacidade de lidar com dados tabulares (Cheng et al., 2024; Wang et al., 2021). Para adequar a estabilidade e a capacidade preditiva dessas árvores de decisão simples, aplica-se o aprendizado por conjunto, dividido principalmente em técnicas de *Bagging* e *Boosting* (Cheng et al., 2024). O *Bagging*, exemplificado pelo algoritmo Random Forest, realiza o treinamento independente e em paralelo de múltiplos estimadores sobre subconjuntos de dados gerados via amostragem de *bootstrap* com reposição, agregando os resultados individuais para reduzir a variância e mitigar o *overfitting* (Cheng et al., 2024; Wang et al., 2021). Por outro lado, a abordagem de *Boosting*, representada por algoritmos como *Gradient Boosting*, *XGBoost* e *LightGBM*, baseia-se no treinamento sequencial de árvores de decisão, em que cada novo modelo é ajustado para corrigir os resíduos das iterações anteriores por meio do gradiente da função de perda (Cheng et al., 2024; Wang et al., 2021; Bellamoli et al., 2023). Essa integração de modelos permite mapear relações não lineares nas variáveis de processo das ETARs com menor tempo de processamento (Cheng et al., 2024; Bellamoli et al., 2023).

Dessa forma, a relação entre ETARs, aprendizado de máquina e inovação tecnológica viabiliza a transição para a gestão digital no saneamento (Ciuccoli, 2025). A aplicação desses algoritmos permite analisar dados, estimar parâmetros de qualidade e otimizar processos, de modo a suportar a tomada de decisão e a detecção de falhas (Zamfir, 2025). Essa abordagem possibilita o desenvolvimento de ferramentas de digitalização, como sensores virtuais e gêmeos digitais, que estimam o comportamento das estações de forma contínua a partir do processamento direto do histórico operacional da planta (Dikmen, 2025). Ao integrar o aprendizado de máquina, as ETARs reduzem a ocorrência de falhas mecânicas, geram economia de recursos e adotam modelos de gestão preventiva de efluentes (Gulshin e Kuzina, 2024). Desse modo, o desenvolvimento dessas tecnologias de dados apoia a conformidade com normas regulatórias e promove a sustentabilidade do setor de tratamento de efluentes (Zamfir, 2025).

3 Metodologia

Os modelos foram construídos seguindo uma estrutura de classificação binária dividida nas etapas de exploração, tratamento de dados, incluindo, interpolação linear para diminuir a quantidade de dados faltantes, adoção de defasagem temporal, definição do alvo preditivo, seleção de atributos, balanceamento de classes e validação temporal.

No tratamento preliminar dos dados, aplicou-se a interpolação linear simples para preencher lacunas contínuas de no máximo dois dias nas variáveis preditoras, fundamentando-se na inércia física do sistema biológico cujo Tempo de Residência Hidráulica (TRH) é de 48 horas.

Os trechos com valores faltantes superiores a esse período foram removidos para evitar distorções estatísticas ou vazamento de informação na validação. Na sequência, efetuou-se a eliminação de inconsistências operacionais: conforme abordado anteriormente, o registro de pH equivalente a 0,85 foi descartado por erro de digitação, ao passo que valores negativos identificados no volume de produção de celulose foram ajustados para zero, refletindo períodos de parada programada para manutenção e suavizando ruídos sem descaracterizar a rotina da planta.

Para a seleção dos atributos preditores, utilizou-se o método Lasso, que realiza a seleção de variáveis e a contração dos coeficientes para simplificar a estrutura do modelo. Além disso, diferentemente das formulações de regressão contínua aplicadas à mesma base por Oliveira-Esquerre et al. (2004), a presente abordagem orienta-se à classificação supervisionada para suporte à decisão em tempo real (Wang et al., 2021; Sundui et al., 2021; Aparna et al., 2024; Bobboi e Dimililer, 2025; Figarola-Figarola et al., 2026).

A variável-alvo (DBO) foi discretizada com base no Percentil 90, adotando o valor de 117 ppm como limiar de corte; leituras iguais ou inferiores foram categorizadas como Classe 0 (operação normal) e valores superiores como Classe 1 (estado de alerta). O emprego do Percentil 90 apoia-se no monitoramento de qualidade hídrica para delimitar o comportamento na cauda superior da distribuição de poluentes (Jung et al., 2020), visto que avaliações de conformidade requerem a estimativa da variabilidade em percentis elevados (López et al., 2012). Essa calibração estatística permite ao sistema mapear antecipadamente a probabilidade de falhas operacionais (Jung et al., 2020).

A definição do limiar pelo Percentil 90 resultou em uma proporção desbalanceada no banco de dados, em que a Classe 1 representa cerca de 10% da amostragem, condição que pode levar os classificadores a priorizarem a acurácia global em detrimento dos falsos negativos (Ghinea et al., 2023; Qiu et al., 2024). Para mitigar esse efeito, analisou-se o comportamento da técnica SMOTE para geração sintética de dados e, de forma complementar, adotou-se a penalização interna nos algoritmos via pesos de classe (*class_weight*), mantendo a integridade da série física da planta.

A escolha das arquiteturas baseou-se nos algoritmos *ensemble* Random Forest e Histogram-based Gradient Boosting, reconhecidos pela capacidade de mapear relações não lineares complexas e auxiliar o gerenciamento ambiental em plantas industriais (Bellamoli et al., 2023; Dikmen et al., 2025; Zamfir et al., 2025). A comparação entre ambas permite avaliar se o ganho em eficiência computacional e estrutura se traduz em maior sensibilidade na detecção de eventos críticos.

Esses ajustes permitem aos modelos analisarem o histórico recente do afluente e a dinâmica de produção dos dois dias anteriores antes de projetar a conformidade no efluente final. A avaliação seguiu o protocolo de validação cruzada temporal em séries cronológicas (TSCV), dividindo o histórico de forma sequencial para impedir o vazamento de dados e simular o desempenho real do sistema em operação.

4 Resultados

A validação dos classificadores foi conduzida considerando a distribuição assimétrica das classes e a aplicação prática do Sistema de Alerta Precoce no monitoramento da DBO efluente. No contexto do controle ambiental de plantas industriais, a ocorrência de falsos negativos implica na não identificação de desvios, ao passo que a elevação indevida de falsos positivos compromete a usabilidade do sensor virtual por desgastar a confiabilidade nas decisões. Assim, a avaliação priorizou a relação entre a capacidade de detecção de eventos e a seletividade das respostas sob a estratégia de validação cruzada temporal (TSCV), adotando o limiar de decisão ajustado em 0,20.

A categorização dos dados, estruturada a partir do Percentil 90 do parâmetro de DBO de saída, resultou na alocação de 90,89% da amostra na Classe 0, associada à operação dentro dos padrões habituais, e 9,11% na Classe 1, representativa das condições de alerta. A distribuição quantitativa desses registros no banco histórico está sintetizada na Tabela 2.

Tabela 2. Distribuição das classes para os modelos propostos. Elaboração própria, 2026

Classe	Descrição do Estado Operacional	Frequência Absoluta	Proporção (%)
0	Normal (Operação Estável)	1047	90,89
1	Alerta (Não-conformidade de DBO)	105	9,11
Total	Conjunto de Dados Global	1152	100

Em modelos de classificação diante de assimetria entre classes a avaliação baseada apenas na acurácia global torna-se insuficiente, visto que classificadores ajustados para priorizar a classe majoritária podem registrar taxas elevadas de acerto geral enquanto falham na identificação dos episódios de alerta. Os indicadores apresentados na Tabela 3 detalham esse comportamento e evidenciam as limitações de cada modelo testado sob tais condições.

Tabela 3. Métricas de classificação da classe de Alerta para os modelos. Elaboração própria, 2026.

Modelo	Precisão	*Recall*	*F1-Score*	*Acurácia*	*AUC-ROC*
Random Forest (TSCV)	0,28	0,25	0,27	0,89	0,667
Random Forest (SMOTE)	0,30	0,12	0,17	0,84	0,7786
HistGradientBoosting	0,18	0,12	0,14	0,89	0.656
Regressão Logística	0,10	0,69	0,17	0,48	0,656
Random Forest com PCA	0,23	0,27	0,25	0,87	0,536

Com base no panorama apresentado na Tabela 3, adotou-se como critério central a análise da sensibilidade e da precisão, avaliando a proporção de verdadeiros positivos tanto em relação aos falsos negativos quanto aos falsos positivos. Os desdobramentos práticos e o comportamento de cada modelo frente à variabilidade dos dados industriais são analisados individualmente nas seções a seguir, estruturadas a partir de três hipóteses de desempenho.

4.1 Reamostragem Sintética e Desempenho Preditivo: Análise da Abordagem SMOTE com Random Forest

Na primeira hipótese avaliada, testou-se a capacidade do SMOTE em conjunto com o Random Forest para aprimorar a identificação da classe crítica mantendo a utilidade prática do sistema. As comparações das matrizes de confusão e das curvas de precisão-sensibilidade para ambos os modelos constam na Figura 1, estando a síntese dos parâmetros de desempenho

detalhada na Tabela 3.

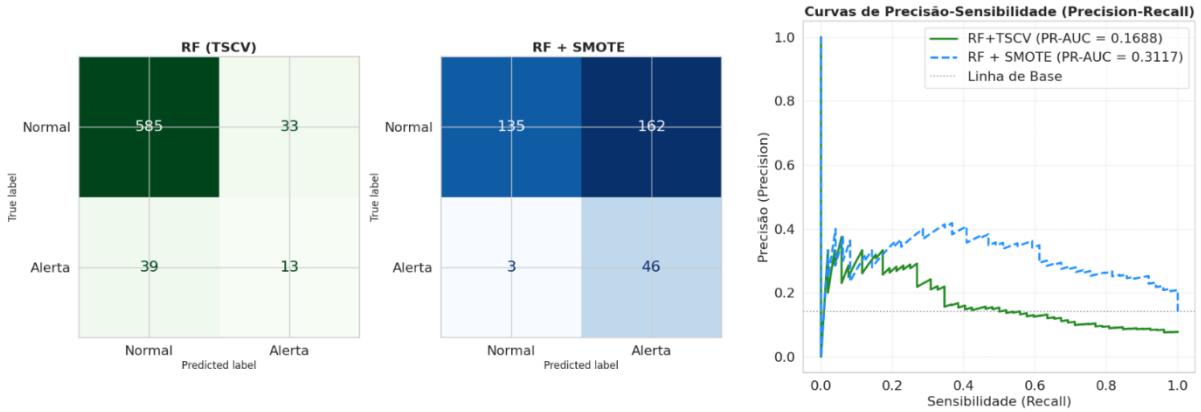


Figura 1. Matrizes de confusão e curvas de precisão-sensibilidade *(Precision-Recall)* para classificar DBO no modelo *Random Forest*: comparação entre as aplicações de TSCV e SMOTE. Elaboração própria, 2026

Ao empregar a reamostragem sintética com SMOTE, o algoritmo Random Forest registrou 162 falsos positivos e 135 verdadeiros negativos, demonstrando uma tendência a classificar a maior parte das condições do processo como situação de risco. Embora essa estratégia tenha elevado o valor da área sob a curva de precisão e sensibilidade (PR-AUC de 0,31) e ampliado a captura dos alertas, o volume de alarmes incorretos compromete a confiabilidade exigida para a aceitação do sensor virtual no acompanhamento contínuo do tratamento. Por outro lado, a configuração sem SMOTE apresentou um perfil mais conservador (PR-AUC de 0,17), identificando 13 eventos de não conformidade e omitindo 39, enquanto manteve 585 verdadeiros negativos e limitou os falsos positivos a 33 registros.

A resposta obtida sem a adição de dados sintéticos apresentou uma relação sinal-ruído mais consistente com o histórico da planta, levando à rejeição da primeira hipótese. Os resultados demonstram que o acréscimo numérico na curva de precisão não gera benefício prático, visto que a geração artificial de amostras altera as dependências temporais em episódios de baixa frequência. Sob a perspectiva da inovação tecnológica no saneamento, a construção de arquiteturas de controle digital depende da preservação da dinâmica física das séries temporais, garantindo que o modelo reflita o comportamento real da instalação sem introduzir distorções sintéticas.

4.2 Impacto da Complexidade Algorítmica: Avaliação do Histogram-based Gradient Boosting

A segunda hipótese avaliou se a adoção da arquitetura *Histogram-based Gradient Boosting* traria ganhos de desempenho na detecção dos episódios de alerta. O pressuposto assumia que o ajuste sequencial característico do *boosting* seria capaz de corrigir as falhas residuais registradas pela Random Forest no mapeamento de oscilações do efluente. Para fundamentar essa comparação, examinaram-se as matrizes de confusão e as curvas ROC de ambas as abordagens (Figura 2) cujas métricas numéricas estão consolidadas na Tabela 3. A análise busca avaliar se o aumento na complexidade do algoritmo melhora a capacidade de detecção de classes, e aumenta a confiabilidade dos sensores virtuais no monitoramento

contínuo em ETARs.

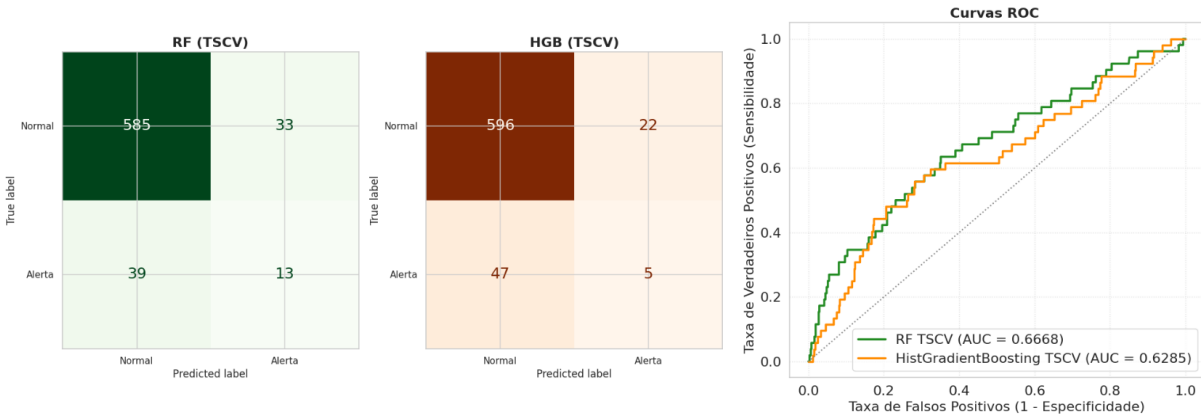


Figura 2. Matrizes de confusão e curvas de característica de operação do receptor (ROC) sob TSCV: análise comparativa de desempenho entre os algoritmos *Random Forest* (*Bagging*) e HistGradientBoosting (*Boosting*)

Os dados obtidos indicaram um desempenho inferior do *Histogram-based Gradient Boosting* quando comparado à *Random Forest*. O algoritmo de *boosting* alcançou um *F1-Score* de 0,14 e AUC-ROC de 0,63, enquanto a Random Forest obteve 0,27 e 0,67 para os mesmos parâmetros. A análise da matriz de confusão mostrou que o *Histogram-based Gradient Boosting* registrou 22 falsos positivos e identificou apenas 5 episódios de alerta. Além disso, as taxas de não detecção (falsos negativos) mantiveram-se elevadas em ambas as abordagens, situando-se em 75% para a Random Forest e 88% para o *Histogram-based Gradient Boosting*. Como a estrutura do *boosting* não superou as limitações de previsão temporal, a segunda hipótese foi rejeitada.

Esse comportamento indica que o aprendizado sequencial do *boosting* concentrou-se nas oscilações locais dos dados sem ampliar a capacidade preditiva da planta. Em contrapartida, o mecanismo de agregação paralela da *Random Forest* manteve maior estabilidade na validação temporal, pois atenua a variância e minimiza o impacto de ruídos nas medições industriais. No entanto, a sensibilidade de 0,25 evidencia que a ferramenta não deve ser utilizada como parâmetro decisivo isolado, devendo atuar como um sistema complementar de suporte às decisões operacionais na planta. Desse modo, o sensor virtual cumpre a função de emitir alertas prévios de tendência para os operadores, funcionando como um recurso complementar de suporte à decisão que não substitui as rotinas de análise laboratorial.

4.3 Avaliação de Estratégias de Simplificação: Modelagem Linear e Redução Dimensional por PCA

A terceira hipótese avaliou duas estratégias de simplificação da modelagem: a aplicação de uma abordagem linear clássica por regressão logística em substituição à *random forest* e a redução de dimensionalidade por Análise de Componentes Principais (PCA), mantendo três componentes, para mitigar a colinearidade e facilitar a separação dos dados. As métricas comparativas desses cenários estão organizadas na Tabela 3.

A Regressão Logística alcançou uma taxa de sensibilidade de 0,69 e capturou um número maior de episódios críticos. Contudo, conforme ilustrado na Figura 3, o algoritmo registrou 335 falsos positivos, demonstrando uma propensão a classificar o processo em estado de risco de forma contínua. Essa frequência excessiva de sinalizações incorretas reduz a viabilidade prática da ferramenta, uma vez que desestabiliza a confiança no sensor virtual para a gestão do efluente.

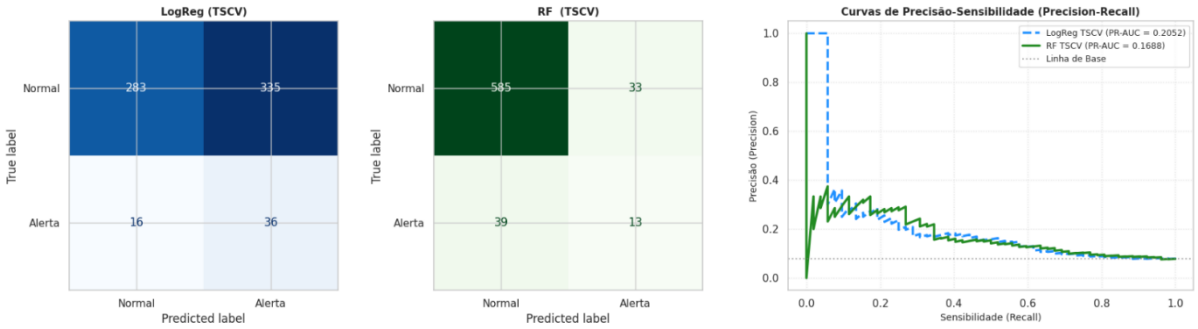


Figura 3. Matrizes de confusão e curvas de precisão-sensibilidade sob TSCV: análise comparativa de desempenho entre os algoritmos Regressão logística (paramétrico) e *Random Forest* (não paramétrico)

Adicionalmente, a área sob a curva de precisão e sensibilidade registrou um valor ligeiramente superior para a Regressão Logística (0,2052) em relação à Random Forest (0,1688). No entanto, a regressão logística apresenta uma precisão de 0,10, visto que a identificação de 36 episódios reais de alerta exigiu a geração de 335 alarmes incorretos. Por outro lado, a aplicação da Random Forest combinada ao PCA alcançou um F1-Score de 0,25 e uma sensibilidade de 0,27. Conforme apresentado na Figura 4, a redução para três componentes principais não trouxe evolução significativa no desempenho em comparação ao modelo sem PCA, além de eliminar a interpretabilidade direta das variáveis operacionais na planta. Com esses resultados, a terceira hipótese foi rejeitada.

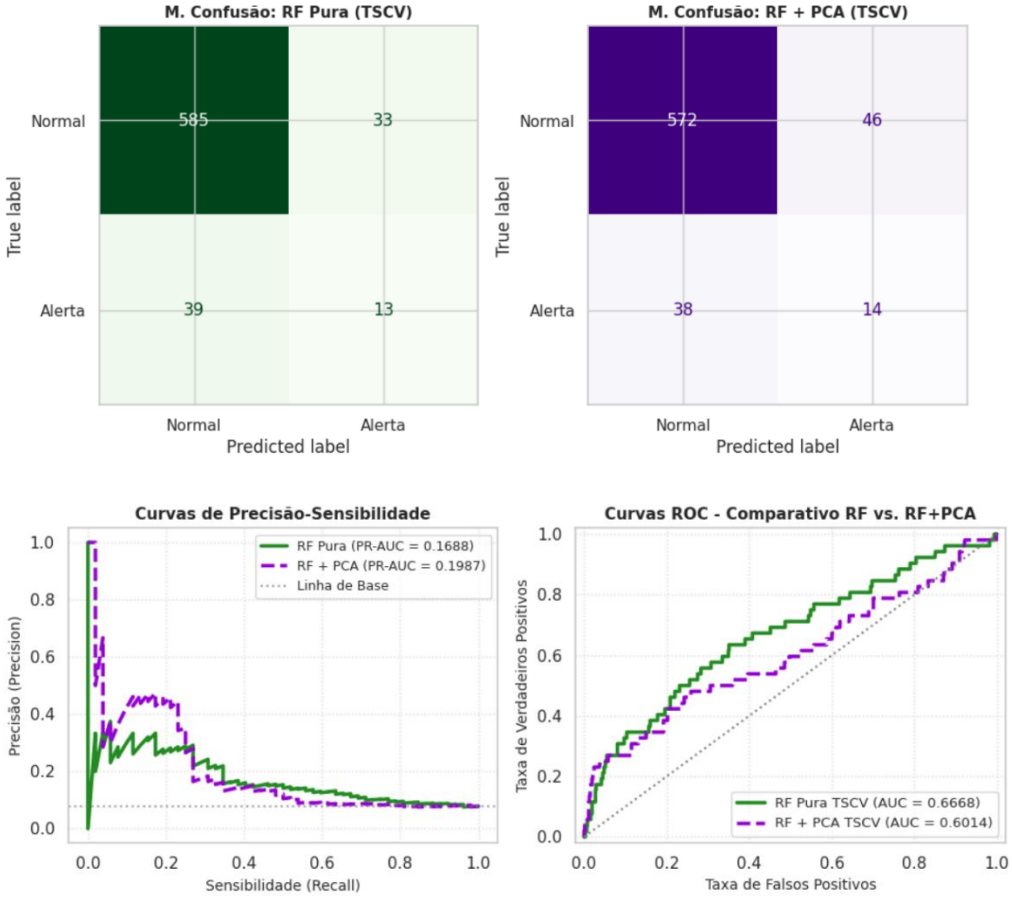


Figura 4. Matrizes de confusão e curvas de precisão-sensibilidade e curvas ROC para análise comparativa de desempenho entre os algoritmos *Random Forest* com e sem PCA

A análise da relação entre precisão e sensibilidade evidencia que a redução dimensional alterou o ponto de equilíbrio do classificador. O acréscimo na sensibilidade, que passou de 0,25 para 0,27 com a identificação de uma amostra adicional de alerta, provocou a queda na precisão de 0,28 para 0,23. Esse movimento elevou a ocorrência de alarmes falsos de 33 para 46 registros e reduziu o indicador AUC-ROC de 0,67 para 0,61.

Por se tratar de uma técnica não supervisionada direcionada à variância global da matriz de dados, a Análise de Componentes Principais tende a concentrar a informação nos eixos primários, reduzindo a representatividade das oscilações específicas da classe minoritária. A partir desses resultados, é possível afirmar que simplificações baseadas em projeções lineares não equacionam a complexidade do monitoramento industrial desbalanceado. No desenvolvimento de sensores virtuais voltados ao saneamento digital, a manutenção das variáveis originais garante a rastreabilidade física dos processos da planta e preserva a estabilidade do sistema ao conter a emissão excessiva de alarmes falsos.

4.4 Síntese dos Resultados, Viabilidade Operacional e Inovação Tecnológica

A síntese dos resultados na Tabela 3 evidencia as limitações da acurácia global como parâmetro de avaliação em matrizes desbalanceadas. Diferentes configurações atingiram índices de acurácia próximos a 80%, mesmo apresentando desempenhos discrepantes no mapeamento dos episódios de alerta, configurando o paradoxo da acurácia. A *Random Forest* sem SMOTE, integrada à validação cruzada temporal (TSCV) e a variáveis defasadas pelo tempo de retenção hidráulico (TRH), estabeleceu o equilíbrio adequado entre a captura de eventos críticos e a contenção de alarmes incorretos. Essa estrutura evitou o excesso de sinalizações registrado na Regressão Logística e na aplicação do SMOTE, superando também o *Histogram-based Gradient Boosting* e a combinação com PCA.

Os achados demonstram que a validação temporal reduz estimativas otimistas e que a geração de dados sintéticos amplia a incidência de falsos alarmes em séries dinâmicas reais. A incorporação do TRH fortaleceu a coerência física do modelo, reforçando a necessidade de conciliar a métrica estatística à viabilidade técnica na planta. Embora a avaliação tenha se restringido a uma única unidade operacional e utilizado uma base de dados histórica antiga, o que exige cautela na extrapolação direta dos valores numéricos, a estrutura metodológica desenvolvida apresenta escalabilidade para outras instalações. Para isso, requer apenas o ajuste das variáveis de entrada e o retreinamento dos algoritmos com dados locais atualizados, permitindo absorver as dinâmicas operacionais vigentes de cada sistema.

Sob a ótica da inovação em gestão e da sustentabilidade organizacional, a contribuição deste trabalho reside na modernização dos processos de tomada de decisão no monitoramento ambiental. A aplicação do sensor virtual como ferramenta preditiva contínua permite antecipar oscilações operacionais e otimizar o tratamento de efluentes industriais. Ao alinhar os parâmetros do processo à modelagem temporal, a pesquisa entrega uma solução aplicável à governança socioambiental. Além disso, ao demonstrar que abordagens sintéticas podem comprometer a assertividade dos alertas, o estudo preserva a confiabilidade do sistema e fortalece a capacidade de resposta preventiva nas organizações.

5 Conclusão

Este estudo avaliou a aplicação de modelos de árvores como base para um Sistema de Alerta Precoce voltado à identificação de eventos críticos de DBO em efluente industrial tratado por lagoas aeradas. O descarte de variáveis com dados ausentes acima de 15% garantiu a

integridade do pipeline, limitando a modelagem a parâmetros operacionais de monitoramento contínuo. A engenharia de atributos baseada nas defasagens de 48 horas do Tempo de Residência Hidráulica incorporou com sucesso a inércia biocinética da lagoa ao modelo. Contudo, o algoritmo SMOTE mostrou-se inviável para a rotina de campo, pois induziu o sistema ao overfitting e elevou os falsos positivos.

O estabelecimento da classe de alerta pelo percentil 90 restringiu o escopo aos 10% de eventos mais extremos da planta, criando um critério de avaliação próprio e independente de legislações ambientais. Essa formulação mostrou-se adequada para avaliar o potencial dos modelos como sensores virtuais orientados ao suporte à decisão.

Os resultados indicaram que a aplicação do SMOTE não foi adequada ao estudo de caso, pois, embora tenha elevado a sensibilidade, produziu número excessivo de falsos positivos, comprometendo a viabilidade operacional do sistema. Também se observou que o aumento de complexidade por meio do Histogram-based Gradient Boosting não resultou em melhor desempenho, apresentando maior taxa de omissão dos eventos críticos. A Regressão Logística gerou quantidade elevada de alarmes falsos, enquanto a combinação de PCA com Random Forest aumentou a sensibilidade penalizando a precisão, gerando também um aumento de falsos alarmes.

A configuração Random Forest (via TSCV) foi a mais adequada para o sistema analizado, o modelo obteve o maior equilíbrio na relação sensibilidade-precisão, conferindo à Random Forest maior confiabilidade para atuar como ferramenta de suporte à decisão em tempo real. Como trabalhos futuros, recomenda-se aplicar valores de SHAP para interpretar os gatilhos físicos dos alarmes e testar o pipeline em outras tipologias de tratamento. Embora esses resultados indiquem desempenho moderado e necessidade de aprimoramento, o modelo mostrou potencial para atuar como sensor virtual complementar, auxiliando a identificação antecipada de condições críticas e apoiando ações preventivas na operação de ETARs industriais.

Referências

Aparna K.G., Swarnalatha R., Changmai M. (2024). Optimizing Wastewater Treatment Plant Operational Efficiency Through Integrating Machine Learning Predictive Models and Advanced Control Strategies. Process Safety and Environmental Protection 35 (1) 148. doi:10.1016/j.psep.2024.05.148.

Bellamoli F., Di Iorio M., Vian M., Melgani F. (2023). Machine learning methods for anomaly classification in wastewater treatment plants. Journal of Environmental Management 35 (1) 118594. doi:10.1016/j.jenvman.2023.118594.

Bobboi I. A., Dimililer K. (2025). Enhancing Explainability of Chemical Oxygen Demand Prediction Models for Industrial Wastewater Treatment in Delfzijl, the Netherlands. International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) 1 (1) 11017189. doi:10.1109/ichora65333.2025.11017189.

Cheng Q., Kim J.-Y., Wang Y., Ren X., Guo Y., Park J.-H., Park S.-G., Lee S.-Y., Zheng G., Wang Y., Lee Y.-J., Hwang M.-H. (2024). Novel Ensemble Learning Approach for Predicting COD and TN: Model Development and Implementation. Water 16 (11) 1561. doi:10.3390/w16111561.

Chow C., Liu J., Li J., Swain N., Reid K., Saint C. (2018). Development of smart data analytics tools to support wastewater treatment plant operation. Chemometrics and Intelligent Laboratory Systems 31 (1) 006. doi:10.1016/j.chemolab.2018.03.006.

Corominas L., Garrido-Baserba M., Villez K., Olsson G., Cortés U., Poch M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling & Software 106 89. doi:10.1016/j.envsoft.2017.11.023.

Ciuccoli N., Fatone F., Sgroi M., Eusebi A.L., Rosati R., Screpanti L., Mancini A., Scaradozzi D. (2025). Forecasting and Early Warning System for Wastewater Treatment Plant Sensors Using Multitask and LSTM Neural Networks: A Simulated and Real-World Case Study. Computers and Chemical Engineering 198 109103. doi:10.1016/j.compchemeng.2025.109103.

Dikmen F., Demir A., Ozkaya B., Raza M., Rasheed J., Asuroglu T., Alsubai S. (2025). AI-driven wastewater management through comparative analysis of feature selection techniques and predictive models. Scientific Reports 26 (1) 07124-0. doi:10.1038/s41598-025-07124-0.

Figarola-Figarola A., Coronado-Apodaca K. G., Buentello-Montoya D. A. (2026). Prediction of chemical oxygen demand in industry effluents using machine learning and IoT: a case study in Tequila, Jalisco. Frontiers in Water 0 (2) 1805143. doi:10.3389/frwa.2026.1805143.

Ghinea L., Miron M., Barbu M. (2023). Semi-Supervised Anomaly Detection of Dissolved Oxygen Sensor in Wastewater Treatment Plants. Sensors 24 (1) 8022. doi:10.3390/s23198022.

Gulshin I., Kuzina O. (2024). Optimization of Wastewater Treatment Through Machine Learning-Enhanced Supervisory Control and Data Acquisition: A Case Study of Granular Sludge Process Stability and Predictive Control. Automation 8 (1) 002. doi:10.3390/automation6010002.

Jung H., Senf C., Jordan P., Krueger T. (2020). Benchmarking inference methods for water quality monitoring and status classification. Environmental Monitoring and Assessment 192 (4) 261. doi:10.1007/s10661-020-8223-4.

López I., Álvarez C., Gil J.L., Revilla J.A. (2012). Does the bathing water classification depend on sampling strategy? A bootstrap approach for bathing water quality assessment, according to Directive 2006/7/EC requirements. Journal of Environmental Management 111 236. doi:10.1016/j.jenvman.2012.08.006.

Oliveira-Esquerre K. P., Seborg D. E., Bruns R. E., Mori M. (2004). Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches. Chemical Engineering Journal 104 (1) 05-11. doi:10.1016/j.cej.2004.05.011.

Oliveira-Esquerre K. P., Seborg D. E., Mori M., Bruns R. E. (2004). Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part II. Nonlinear approaches. Chemical Engineering Journal 104 (1) 06-12. doi:10.1016/j.cej.2004.06.012.

Qiu C., Shao G., Zhang Z., Zhou C., Hou Y., Zhao E., Guo X., Guan X. (2024). Unsupervised Real Time and Early Anomalies Detection Method for Sewer Networks Systems. IEEE Access 8 (1) 3359302. doi:10.1109/access.2024.3359302.

Sundui B., Calderon O. A. R., Abdeldayem O. M., Lazaro-Gil J., Rene E., Sambuu U. (2021). Applications of machine learning algorithms for biological wastewater treatment: Updates and perspectives. Clean Technologies and Environmental Policy 108 (1) 01993-x. doi:10.1007/s10098-020-01993-x.

Wang D., Thunell S., Lindberg U., Jiang L., Trygg J., Mysklind M., Souihi N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. The Science of the Total Environment 165 (1) 147138. doi:10.1016/j.scitotenv.2021.147138.

Wang R., Yu Y., Chen Y., Pan Z., Li X., Tan Z., Zhang J. (2021). Model construction and application for effluent prediction in wastewater treatment plant: Data processing method optimization and process parameters integration. Journal of Environmental Management 69 (1) 114020. doi:10.1016/j.jenvman.2021.114020.

Wang Y.-Q., Zhou H.-B., Luo X.-Q., Deng S.-W., Xu H.-R., Song Y.-P., Chen J.-J., Yin W.-X., Cheng H.-Y., Wang A.-J., Wang H.-C. (2025). Machine Learning-Driven Dynamic Measurement of Environmental Indicators in Multiple Scenes and Multiple Disturbances. Environmental Science & Technology 59 (30) 15877-15889. doi:10.1021/acs.est.5c06126.

Yang, S.; Behzadian, K.; Coleman, C.; Holloway, T. G.; Campos, L. C. Application of AI-based techniques for anomaly management in wastewater treatment plants: A review. Journal of Environmental Management, v. 374, p. 126886, 2025. Disponível em: https://doi.org/10.1016/j.jenvman.2025.126886.

Zamfir F.-S., Carbureanu M., Mihalache S.F. (2025). Application of Machine Learning Models in Optimizing Wastewater Treatment Processes: A Review. Applied Sciences 15 (15) 8360. doi:10.3390/app15158360.