O PAPEL vol. 86, N.o 10, pp. 60 - 68 - OCT 2025

# PREDITOR DE CRISE DA ESTAÇÃO DE TRATAMENTO DE EFLUENTES

Raphael Campos Ribeiro[1], Gabriel Vicente De Pierro[2], Jéssica Milani de Andrade[3], Fagner Machado Rezende[4], Rodrigo Augusto Barella[5]

[1] *Suzano S.A. – Unidade Jacareí/Brasil – Gerência de Recuperação e Utilidades*

[2] *Suzano S.A. – Tribo Digital Industrial – Squad Green Operations*

[3] *Suzano S.A. – Unidade Jacareí/Brasil – Gerência de Meio Ambiente*

[4] *Suzano S.A. – Tecnologia da Informação – Industrial e Florestal*

[5] *Suzano S.A. – Tribo Digital Industrial – Squad Green Operations*

## RESUMO

O arraste de sólidos da estação de tratamento de efluentes para o rio é um problema crítico enfrentado por indústrias de variados segmentos. O presente estudo discorre acerca da aplicação de ciência de dados para controle antecipatório de distúrbios de processo em uma Estação de Tratamento de Efluentes (ETE) do segmento de Celulose e Papel. Com o objetivo de viabilizar a gestão proativa das variáveis de processo da ETE e, por consequência, reduzir o potencial de ocorrências ambientais e de perdas de produção, foi desenvolvido um modelo de aprendizado de máquina para antecipar o estado futuro da estação. O sistema desenvolvido opera em duas etapas: a primeira modela a probabilidade de arraste de sólidos nas próximas 24 horas, enquanto a segunda etapa provê uma visão detalhada de quais variáveis podem estar causando alterações no comportamento atual e futuro da ETE. Essas duas etapas se complementam de maneira que, enquanto uma alerta a operação, a outra ajuda na busca da causa raiz do problema. Ao integrar tecnologia de ponta com um enfoque preventivo, este projeto visa melhorar a eficiência operacional, reduzir os custos de manutenção e promover a sustentabilidade ambiental nas instalações de tratamento de efluentes. A aplicação de Inteligência Artificial (IA), aliada a projetos internos de otimização da estação, permitiu à Suzano Jacareí elevar em mais de 70% a performance da ETE, alcançando em 2023 a marca de melhor performance histórica da estação.

***Palavras-chave:*** *Tratamento de efluentes, monitoramento automatizado, prevenção de impactos ambientais.*

## INTRODUÇÃO

A produção de celulose é um processo intensivo na utilização de água, podendo chegar a mais de 20 m³ por tonelada produzida [1]. O alto consumo, por consequência, gera uma grande quantidade de efluentes que, caso não tratados, podem causar sérios problemas ao meio-ambiente. Neste cenário, a Estação de Tratamento de Efluentes (ETE) é peça fundamental na composição de uma fábrica de celulose, à medida que se faz necessário tratar de maneira eficiente a água utilizada para posteriormente devolvê-la ao rio da qual foi captada, visando minimizar impactos ambientais.

Efluentes industriais provenientes de fábricas de celulose e papel podem conter variados materiais orgânicos tóxicos e não biodegradáveis, incluindo compostos de enxofre, químicos da polpa, ácidos orgânicos, ligninas cloradas, ácidos de resina, fenólicos e ácidos graxos insaturados. Caso a água fosse descarregada diretamente no rio sem tratamento, acarretaria diversos problemas, podendo gerar mudanças na sua temperatura, coloração, e conteúdo de sólidos, além de esgotar o oxigênio dissolvido, e efeitos tóxicos na vida marinha [2].

A Suzano, unidade Jacareí, São Paulo, fábrica integrada de celulose kraft branqueada e papel, gera no seu processo fabril aproximadamente 3.100 m³/h de efluentes líquidos, que são encaminhados para uma Estação de Tratamento de Efluentes (ETE) antes de serem lançados no corpo d'água receptor, o rio Paraíba do Sul. A ETE (**Figura 1**) é constituída por um tanque de neutralização onde é realizada, se necessário, a correção do pH. O tratamento primário é composto por dois decantadores primários que visam remover os sólidos em suspensão. O lodo primário segue para um tanque de mistura de lodo e então, para dois tambores pré-desaguadores (drum pré-thickner) seguido de duas prensas desaguadoras do tipo "screw-press".

O efluente tratado, proveniente dos decantadores, segue para uma estação elevatória equipada com dois conjuntos elevatórios que recalcam os efluentes para a torre de resfriamento abaixando a temperatura dos efluentes de 58 ºC para aproximadamente 36 ºC, compatibilizando ao tratamento biológico mesofílico. O tratamento biológico secundário, constituído por um processo de lodos ativados de duplo estágio, ou seja, é composto por i) um primeiro estágio de aeração constituído por três reatores em paralelo seguidos por três decantadores secundários; ii) um

**Autor correspondente:** Gabriel Vicente De Pierro. Jacareí e 12340-010 Brasil. Fone: +55-27-3270-2122. e-mail: g.vpierro@suzano.com.br

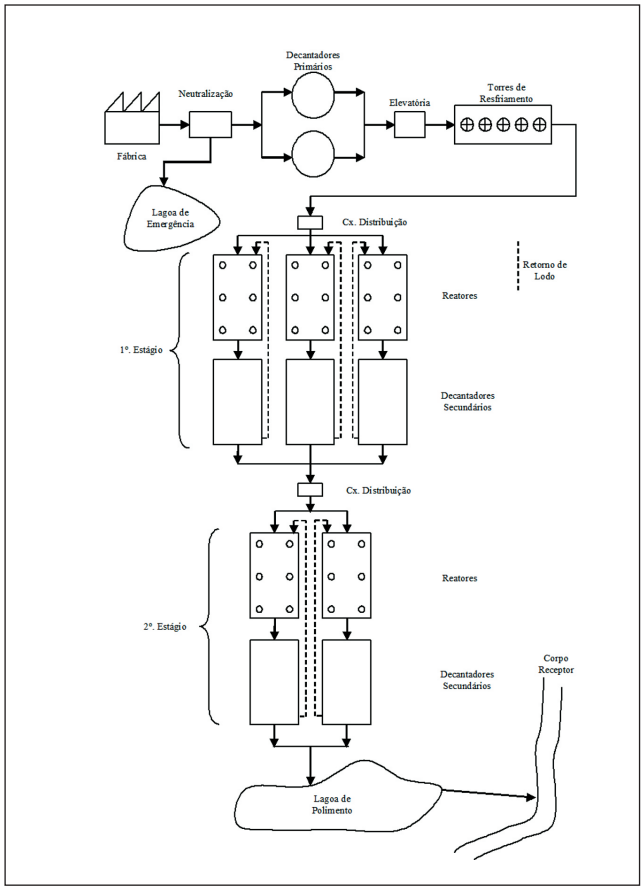


**Figura 1.** Diagrama da ETE

segundo estágio de aeração composto por dois reatores em paralelo seguidos por dois decantadores secundários. Este sistema de duplo estágio de aeração é uma variação dos lodos ativados denominado Processo Attisholz.

Os reatores de aeração, com alguma periodicidade, apresentavam concentração de sólidos inferior às premissas de projeto da estação. Este fato se devia, sobretudo, à perda não intencional de sólidos no efluente sobrenadante dos decantadores, que ocorria na planta com certa frequência devido à má sedimentabilidade do lodo.

Esse "arraste de sólidos", expresso pelos altos valores de IVL e SSD, apresentou como causa provável a proliferação de organismos filamentosos provavelmente do Tipo 0675, Tipo 0041 e Tipo 1701. Tais microrganismos tendem a abrir o floco (open flock) e causar o seu intumescimento, também conhecido como bulking filamentoso. Nestas circunstâncias, o lodo fica muito leve tomando um aspecto nebuloso e que, por não se adensar, acaba sendo carreado com o efluente tratado. A principal causa da proliferação destes organismos, contudo, não era de fácil diagnóstico, justificando a aplicação de soluções de IA à ETE da Suzano Jacareí.

Com o advento de novas e robustas tecnologias digitais, o processo de diagnóstico antecipatório de desvios que causam a proliferação desses organismos filamentosos passou a ser mais ágil, uma vez que viabilizou a análise em tempo real de variáveis de processo correlatas ao *bulking*.

Para essa avaliação, foram aplicadas soluções *data-driven* com abordagens mais tradicionais e mecanísticas, buscando compor o domínio e *expertise* do processo, com metodologias empíricas que podem simplificar e agilizar a modelagem de processos, além de preencher lacunas deixadas por falta ou baixa frequência de sensoriamento.

**1. Trabalhos Correlatos**

O tópico de predição de qualidade e performance em estações de tratamento de efluentes não é novo tanto na indústria em geral quanto na indústria de papel e celulose. De maneira geral, é possível dividir as abordagens em dois grandes campos: mecanísticas, baseadas em modelagem fenomenológica; e empíricas, baseadas em dados. Os modelos mecanísticos, historicamente mais populares, buscam resolver o processo de maneira fenomenológica, desenvolvendo equações de balanço de massa, e estudando outras quantidades conservadas, para todos os compostos envolvidos [3]. Algumas abordagens populares, altamente maduras e presentes em softwares comerciais são as da série ASM [4] e da ADM1 [5]. Esse tipo de abordagem pode se tornar muito complexa à medida que é necessário equacionar diversos processos da planta, sendo preciso um alto grau de caracterização de informações específicas à aplicação e extenso domínio do problema [6].

Em contrapartida, modelos empíricos utilizando aprendizado de máquina vêm se tornando cada vez mais populares como ferramentas para modelagem de tratamento de efluentes [7]. Essas metodologias, baseadas em dados, não requerem a entrada explícita de conhecimento do domínio no sistema, sendo esse conhecimento adquirido por meio das relações dos dados históricos. A *expertise*, no entanto, ainda é valiosa para identificar resultados úteis e errôneos, especialmente considerando as dificuldades de medição e transformação de informação em conhecimento em alguns pontos do processo [8].

Se tratando das técnicas utilizadas, é possível encontrar diversas abordagens, de algoritmos genéticos combinados com *deep belief networks* para mitigar as características dinâmicas e complexas do processo [9], até métodos mais simples utilizando árvores de decisão para prever o influxo na estação de tratamento [10]. Em geral, predominam métodos utilizando redes neurais com diversas estruturas, como *neuro-fuzzy* ou redes neurais recorrentes, na prática sendo combinados com outros e utilizados na forma de *ensembles* [11-14].

Na literatura é possível encontrar também abordagens híbridas, que buscam suplementar as fraquezas de cada método, onde os modelos empíricos buscam simplificar e enriquecer partes da modelagem, e os mecanísticos contribuem com conhecimento específico do processo. Por exemplo, um fenômeno onde não há extensivo conhecimento do domínio pode ser

extrapolado utilizando metodologias *data-driven* e a informação gerada pode ser validada e utilizada para alimentar um próximo modelo mecanístico, provendo informação estrutural do processo a um menor custo [18].

As ferramentas computacionais são evidentemente promissoras para complementar ou substituir as mecanísticas tradicionais por serem mais simples, terem em geral maior capacidade preditiva, e apresentarem erros menores. No entanto, a coleta e curadoria dos dados ainda é um grande desafio enfrentado por esses modelos e pode gravemente limitar quaisquer metodologias *data-driven* [15]. Muitas das abordagens ainda se propõem a usar redes neurais e *deep learning* e obtém bons resultados, apesar desses métodos tenderem a ser intensivos em volume de dados e sofrerem ainda mais com a natureza ruidosa dos conjuntos coletados nas plantas [16].

## MÉTODOS

Seguindo uma metodologia baseada na CRISP-DM [17] (**Figura 2**), o primeiro passo da análise é o entendimento do negócio e dos dados conjuntamente, em uma etapa iterativa de avaliação da qualidade e disponibilidade dos sensores e medidas existentes. O processo como um todo possui mais de 300 indicadores vindo de diversas áreas, não apenas da ETE, mas que podem afetar direta ou indiretamente os resultados. No fim, foram analisados mais de 152 indicadores, entre sensores e medidas de laboratório, sendo 115 utilizados como variáveis na análise de dados final e 21 no modelo selecionado.

A ETE da Suzano Jacareí possui duas medidas fundamentais que são monitoradas para avaliar o arraste de sedimentos para o rio, uma interna, que é medida frequentemente pela operação, e uma externa, na saída para o rio medida em laboratório uma vez por dia, onde de fato é verificado se o arraste ocorreu. Para fins deste trabalho, o "alvo" de predição é a variável de arraste interna por ter uma medição mais frequente, por apresentar mais variações, dando mais insumos para o preditor, e por possuir uma alta correlação com o arraste para o rio.

Outra definição necessária é a janela temporal de previsão, ou seja, quanto tempo antes será predito o alvo. Neste caso, não foi definida uma janela fixa, mas sim feitas várias tentativas para avaliar quando o modelo começava a degradar, pois quanto mais cedo for feita a previsão de arraste, mais tempo há para atuar no problema, porém mais difícil fica a previsão em um futuro mais distante.

Por fim, o problema pode ser tratado como uma regressão ou como uma classificação, ou seja, é possível considerar a predição do valor numérico exato do arraste, por exemplo, o modelo iria indicar que em 24 horas o arraste estaria em 3 mg/L, ou no caso de uma classificação, indicaria uma probabilidade do arraste passar de um determinado valor limite. Ambas as opções foram exploradas, mas dada a natureza ruidosa dos dados e a dificuldade do problema, optou-se pela classificação, considerando que a regressão não seria precisa o suficiente para agregar informação e a classificação binária, aliada à probabilidade do modelo, seriam suficientes para alertar a operação.

**1. Pré-Processamento dos Dados e Engenharia de Atributos**

O processo de exploração dos dados foi iniciado com 152 variáveis (dados de sensores e de laboratório) internas à

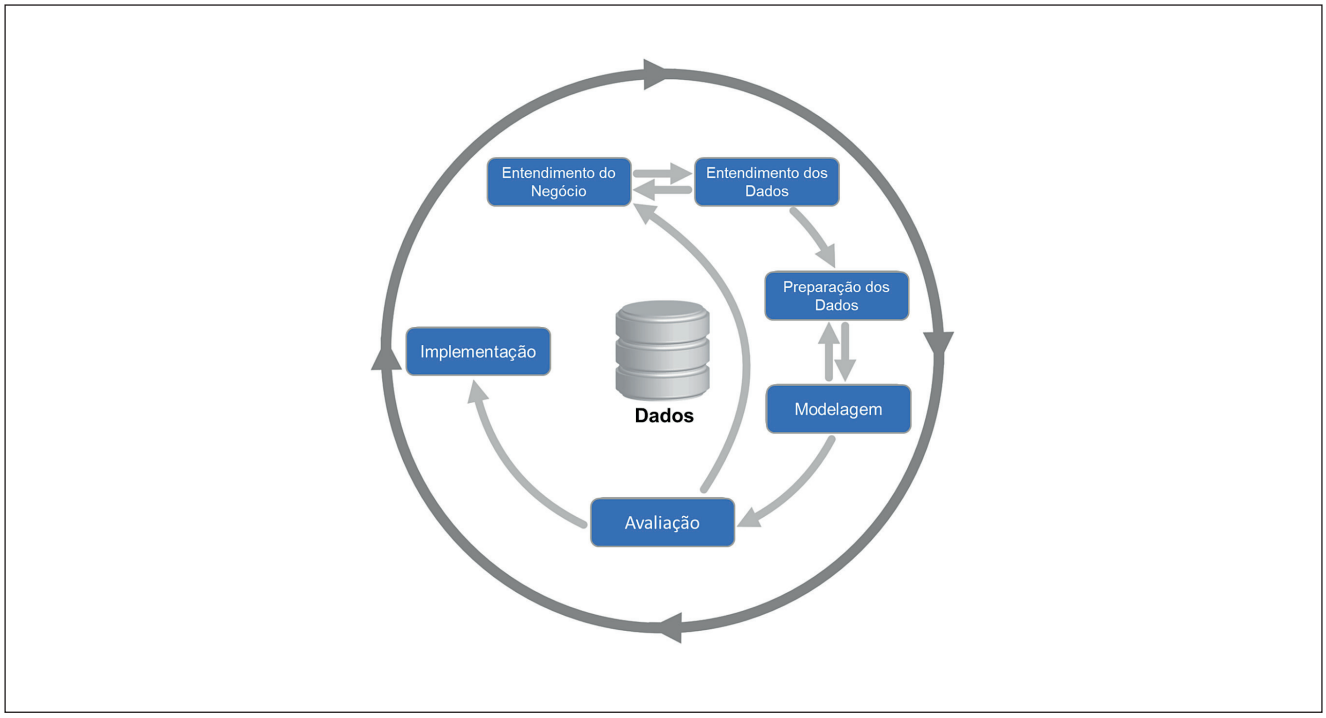


**Figura 2.** Diagrama do processo CRISP-DM, adaptado [17]

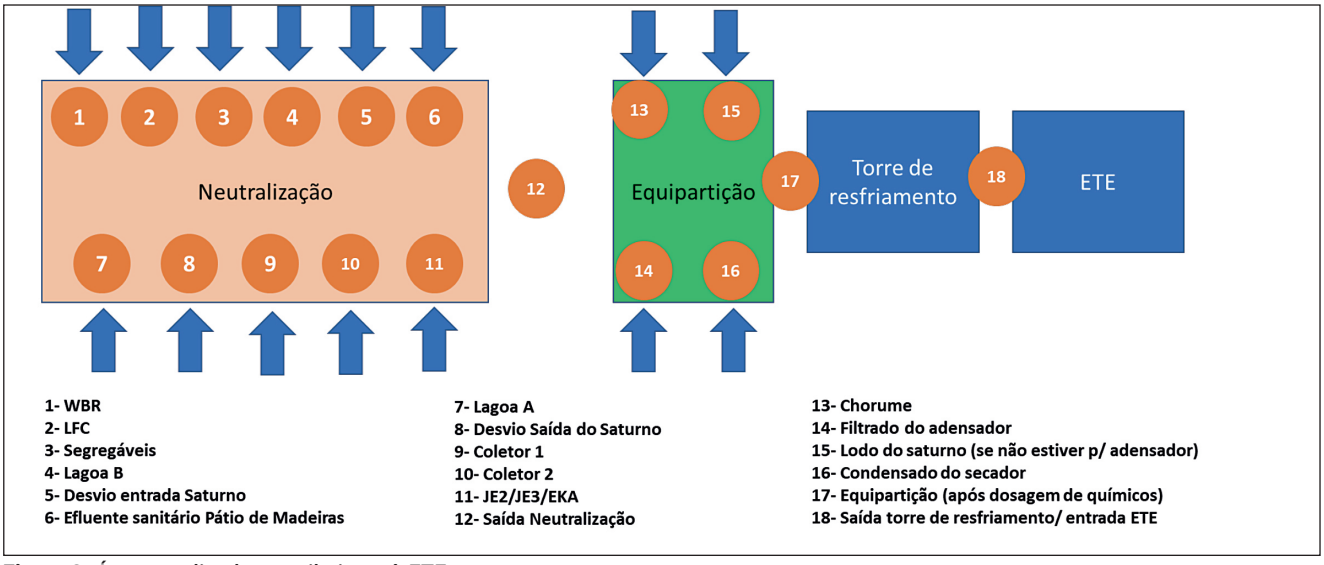


**Figura 3.** Áreas analisada contribuintes à ETE

ETE e provenientes de diversas áreas, como linha de fibras, cozimento, lavagem e evaporação, a **Figura 3**, mostra uma visão geral das principais áreas contribuintes. Dessas 152, aproximadamente metade são variáveis coletadas em tempo real por sensores, e a outra metade variáveis de laboratório onde uma amostra é coletada e analisada manualmente, com periodicidade podendo variar tanto de hora em hora, como semanalmente.

Para consolidar e extrair maiores informações da dinâmica temporal do processo, foram criadas diversas variáveis baseadas em variações no tempo, i.e., diferenças entre o estado atual da variável e janelas fixas de tempo (8, 16, 24 e 72 horas), além de estatísticas descritivas de diversos agrupamentos de variáveis, como médias de demanda de oxigênio em diversos pontos do processo, e média dos sedimentos em vários estágios de coleta.

Em geral, uma grande quantidade de variáveis não leva a melhores modelos, podendo em muitos casos piorar a performance. Assim, a seleção de atributos é fundamental e traz benefícios como: diminuir o tempo de execução do algoritmo, aumentar seu poder preditivo e tornar os resultados mais compreensíveis [19]. Tal seleção foi feita baseada em conhecimento fenomenológico do processo, além de conjuntamente com alguns filtros considerando técnicas como regressão de informação mútua [20], remoção de variáveis com baixa variância, eliminação recursiva de variáveis com baixo poder preditivo [21] e utilização de *scores* proveniente de métodos baseados em árvores que intrinsecamente indicam ganhos de informações nas bifurcações dos galhos e, consequentemente, podem ser usados como intermediários para o poder preditivo [22]. As análises foram feitas levando em conta também as melhores dinâmicas temporais entre as variáveis e o alvo, ou seja, manipulando as relações temporais para verificar os maiores ganhos de informação.

**2. Modelagem**

As tarefas de modelagem e de pré-processamento dos dados são em sua natureza intrínsecas uma à outra [17], e feitas de maneira conjunta e iterativa. As variáveis geradas e as originais foram testadas e refinadas em vários modelos diferentes, desde uma regressão linear simples até redes neurais. Combinado com a parte de seleção e engenharia de atributos, foram feitas tentativas de aumentar o conjunto de eventos de arraste, que é relativamente raro na base, ocorrendo em cerca de 5% dos dias e em geral de maneira sequencial, ou seja, se aconteceu no dia anterior, é muito mais provável que ainda esteja acontecendo. Dada a natureza mais complexa de se manipular amostras de dados temporais, foram tentados apenas métodos mais simples, como podar o começo e o final dos conjuntos, onde não há eventos, ou separar pedaços contínuos da série, mantendo a sequência temporal. No final, concluiu-se que os ganhos não eram significativos e seria suficiente apenas utilizar métricas de avaliação que lidassem bem com classes desbalanceadas.

Para selecionar o modelo final, foi feita uma divisão de dados entre conjunto de treino e de teste com 65% e 35% dos dados respectivamente. Como o modelo é uma série temporal, o conjunto de teste precisou representar um período grande da amostra para ter um número mais relevante de eventos. Também é necessário avaliar que, dada a natureza efêmera de estados do processo, ir muito longe no histórico de dados pode encontrar realidades de fábrica distintas, portanto coletar mais dados no passado não é estritamente benéfico. Neste contexto, o conjunto de treinamento do modelo foi de 1.º de junho de 2021 até 9 de agosto de 2022, enquanto o conjunto de treino foi de 10 de agosto de 2022 até 31 de março de 2023. Em geral, foi observado que modelos mais complexos (como redes neurais), não performavam particularmente melhor que os mais simples, e o ganho não compensaria a perda de transparência com o aumento da complexidade. Os modelos foram avaliados utilizando as métri-

O PAPEL vol. 86, N.o 10, pp. 60 - 68 - OCT 2025

cas de precisão, que é a razão dos casos positivos de arraste pelo total classificado como arraste, revocação, definida pela razão dos casos positivos pelo total de casos classificados como positivos, e F1, que é a média harmónica ponderada dos dois e pela área embaixo da curva de precisão contra revocação, que pode ser utilizada para achar o ponto ideal para a relação de falso positivos e falso negativos ou como métrica de qualidade geral.

**3. Abordagem em Duas Etapas**

O propósito da prevenção de crises não é apenas de caráter preditivo, mas também de auxiliar na análise de causa-raiz do evento. Tendo em vista as imperfeições do modelo e a ocorrência de classificações falso positivas, apontar um possível problema futuro, sem guiar a operação para potenciais causadores pode, por si só, apenas gerar um trabalho adicional de se verificar diversos parâmetros de processo a fim de buscar evidências de anormalidade. Além disso, o modelo criado utiliza um conjunto relativamente pequeno do total de variáveis e, apesar de não ser completamente "caixa-preta", não traz correlações diretas com a variável alvo. Sendo assim, foi criado um segundo modelo, que busca ser mais abrangente em troca de poder preditivo. Este segundo modelo, que na prática é mais uma heurística, utiliza todas as variáveis determinadas como relevantes, neste caso 115, e calcula sua faixa histórica ideal e sua correlação com a variável alvo. A partir destes dois valores, é verificado no período atual como a variável está performando, e sua correlação é adicionada como score a uma fórmula que denominamos como "saúde" da ETE, representada pela Equação (1) seguinte:

$$\begin{cases} \sum_{var} |r_{\text{var}}| & \text{se var} \leq L \text{ ou var} \geq U \\ 0 & \text{se var} \leq L \text{ ou var} \geq U \end{cases} \quad (1)$$

Onde $r_{\text{var}}$ é a correlação da variável com o alvo, e I e U são os limites históricos inferior e superior, respectivamente, calculados com base no efeito de cada variável no alvo, e invertidos nos casos de correlação negativa.

A fórmula de saúde permite verificar de forma mais transparente quais variáveis podem estar afetando, direta ou indiretamente, a ETE e, aliada ao preditor, guia a investigação de possíveis parâmetros de processo que potencialmente podem estar afetando o resultado. O valor final é normalizado entre zero e um para ser apresentado como uma porcentagem ao usuário final.

## RESULTADOS E DISCUSSÃO

Para fins de resultados, na seção 1, deste tópico, foram avaliadas as métricas de desempenho de modelos em diversas granularidades temporais no conjunto de teste, visando selecionar o candidato com mais capacidade preditiva, além de buscar o período ideal no qual a predição traria bom caráter antecipatório sem degradar performance.

Na seção 2, deste tópico, foi avaliado o modelo selecionado já operando em produção, fazendo predições de hora em hora, comparando também com o conjunto de teste para avaliar se as estimativas de erro fora da amostra estavam próximas do realizado. Além de observar as métricas, foram verificadas como as predições do modelo e da fórmula de saúde se comportam em cenários reais e como pode auxiliar a operação.

**1. Comparativos entre Modelos e Tempo de Predição**

Parte do trabalho de modelagem é selecionar variáveis e testar diversos modelos para avaliar os melhores em cada cenário. Assim, foram comparadas diversas abordagens diferentes, de mais complexas, como redes neurais, até algumas mais simples, como regressão logística, para servir como um baseline. Os modelos também foram avaliados em diversos intervalos de tempo para determinar quão antecipatória poderia ser a predição sem degradar sua qualidade.

Como comentado na seção 2, do tópico Métodos, foi criado um conjunto de teste com 35% dos dados buscando validar o erro fora da amostra de treino inicial. O conjunto de teste contemplou o período de dez de agosto de 2022 até o final de março de 2023, totalizando 233 dias e foi utilizado como base comparativa para as métricas dos modelos.

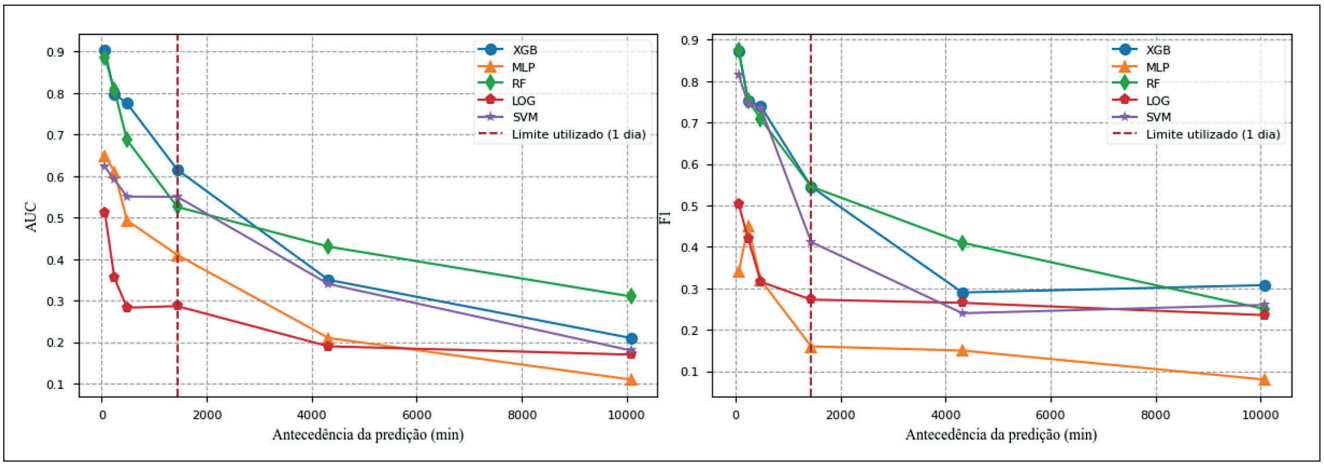


**Figura 4.** Comparativo dos modelos e períodos de predição avaliados

A **Figura 4** mostra o comparativo dos modelos. Os modelos utilizados foram *XGBoost* (XGB), *random forest*, redes neurais (MLP), regressão logística (LOG) e *support vector machines* (SVM). O eixo X mostra a antecedência da predição em minutos. Foram testados intervalos de uma, três e oito horas, além de um, três e sete dias. Na figura são exibidas duas métricas de comparação, área embaixo da curva de precisão de revocacão (AUC) e F1, que foram utilizadas para definir o melhor modelo. Considerando os dados experimentais, o intervalo selecionado foi de um dia, considerando que os modelos ainda apresentavam bons resultados e haveria uma margem de tempo razoável para a operação agir. O modelo selecionado, por performar melhor no período selecionado em ambas as métricas, foi o *XGBoost*.

**2. Modelo Final e Cálculo da Saúde**

Depois da avaliação inicial, o modelo selecionado foi colocado em produção e foi possível comparar o período de teste, que totalizou 233 dias, com o período em produção, segmentado para iniciar depois da parada geral de junho de 2023 e continuando até o final de junho de 2024, totalizando 392 dias. Comparar teste e o modelo operando de forma real permite avaliar se foi estimado corretamente o erro fora da amostra, além de verificar degradações em performance por decorrência de mudanças no processo e problemas na modelagem em selecionar atributos que não estão mais representando o estado atual do processo. Apesar das predições serem horárias, como a predição é avaliada para o dia seguinte e o modelo também foi treinado de tal forma, foram consolidadas as predições em granularidade diária para nivelar as métricas. Ambos os períodos são considerados pelas suas métricas isoladamente.

A Figura 5, acima, mostra o desempenho da predição de probabilidade de arraste do modelo contra os momentos em que o arraste na variável alvo (a interna) passou do limite de 3 mg/L nos períodos de teste (a) e produção (b). A linha pontilhada apresenta o limiar padrão de probabilidade de classificação como evento positivo, 0,5. As métricas apresentadas na Tabela 1 abaixo consideram esse valor, mas na prática o valor em si da probabilidade é usado diretamente, pois a probabilidade pode aumentar sem passar do limiar em ocorrências que o arraste acontece, indicando que há menor confiança do modelo no evento do arraste, mas ainda há alguma.

Seria possível, além de trazer o valor da probabilidade, manipular o valor do limiar de classificação para tender o modelo a dar mais falso positivos, ou mais falso negativos,

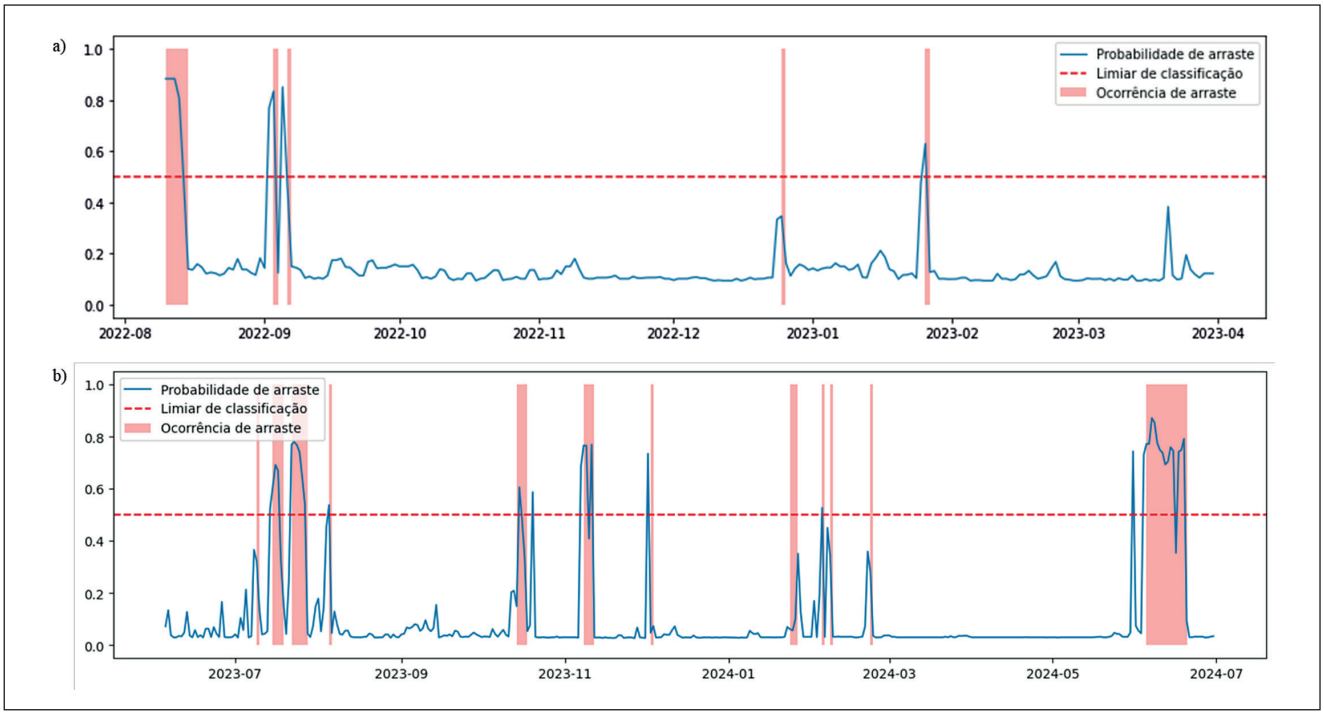


**Figura 5.** Dados de arraste e predição histórica nos períodos de teste e produção

**Tabela 1.** Métricas do modelo em teste e produção

| Período | Classe | Precisão | Revocação | F1 | Suporte |
|---|---|---|---|---|---|
| Teste | Sem arraste | 0.96 | 0.99 | 0.98 | 219 |
| | Arraste | 0.75 | 0.43 | 0.55 | 14 |
| Produção | Sem arraste | 0.91 | 0.98 | 0.95 | 332 |
| | Arraste | 0.83 | 0.48 | 0.61 | 60 |

O PAPEL vol. 86, N.o 10, pp. 60 - 68 - OCT 2025

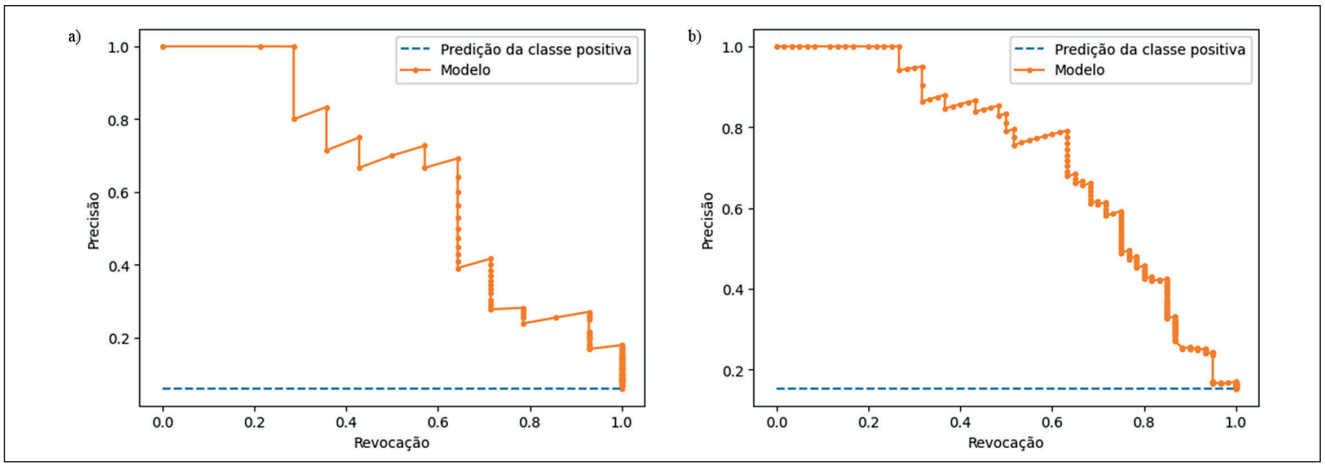


**Figura 6.** Curva de precisão e revocação nos períodos de teste e produção

dependendo da predileção da operação. Mais falso positivos poderiam ocasionar tempo desnecessário em buscar problemas que nunca iriam acontecer e eventualmente os usuários poderiam perder confiança no modelo, enquanto tender a falso negativos poderia deixar passar eventos reais, também acarretando possível perda de segurança na predição. A Figura 6 apresenta a curva de precisão contra revocação, além da comparação com um modelo *naive*, que sempre prevê a classe positiva, sendo a imagem à esquerda (a) do período de teste e à direita (b) do período em produção. A Tabela 2 apresenta a matriz de confusão dos resultados, tradicionalmente utilizada para mostrar os totais de positivos, negativos, falso positivos e falso negativos entre as classes.

Como comentado na seção 3, do tópico Métodos, o modelo de predição sozinho pode não ajudar a encontrar a causa raiz do problema, apenas gerar o alerta. Isso ocorre porque ele utiliza um conjunto pequeno de todas as variáveis e tende a valorizar as que estão mais próximas do final do processo, que apesar de não serem o arresta diretamente, são consequência, e não causa. Portanto, foi criado um segundo indicador chamado de saúde, composto de 115 variáveis selecionadas que busca correlações mais simples com o alvo e indicam de maneira direta possíveis elementos do processo com comportamento anormal que podem estar associados ao evento futuro de arraste. A **Figura 7**, abaixo, mostra a variável de saúde caindo à medida que o arraste interno sobe,

**Tabela 2.** Matriz de confusão

| Período | Classe Predita | | |
|---|---|---|---|
| | Classe Real | Sem arraste | Arraste |
| Teste | Sem arraste | 217 | 2 |
| | Arraste | 8 | 6 |
| Produção | Sem arraste | 325 | 7 |
| | Arraste | 31 | 29 |

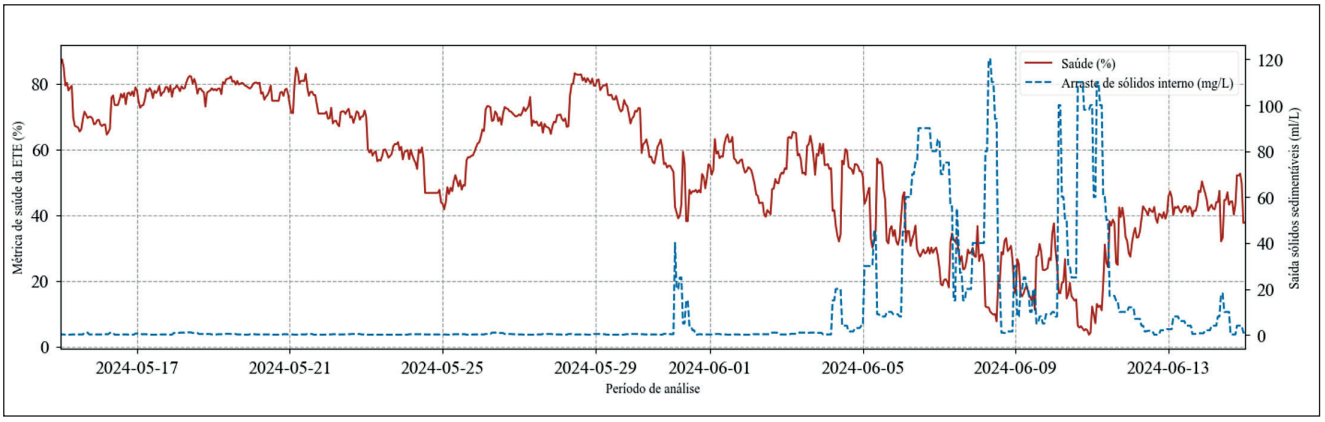


**Figura 7.** Comportamento da métrica de saúde da ETE em eventos de alteração da variável alvo

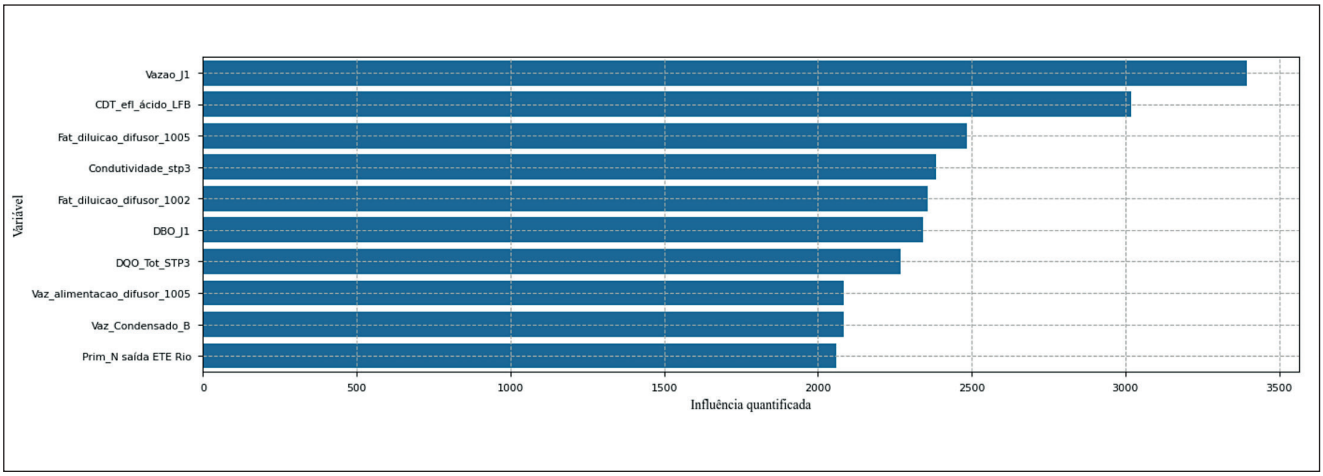


**Figura 8.** As dez principais variáveis apontadas como relevante e fora do normal histórico pela métrica de saúde

enquanto a **Figura 8** identifica as causas da queda, dando informações para se começar a investigar o problema.

Uma vez emitido o alerta via preditor, são disparadas ações para validação do desvio e caso procedentes, executadas as ações corretivas para reestabelecimento dos padrões de processo e redução do nível de risco de arraste.

Na **Figura 9,** temos um exemplo prático de detecção e correção de desvio no mês de julho de 2024 em que foi detectado anomalia tanto em parâmetros internos (oxigênio residual no segundo estágio do sistema biológico) quanto externos à ETE (desvio na qualidade de efluente enviado pela linha de fibras e máquina de papel).

Destaca-se o fato de o monitoramento e detecção de anomalias em tempo real pela própria operação de área/SDCD facilitar o protagonismo e tomada de ação por parte da equipe operacional, promovendo a independência da operação e permitindo uma resposta ágil e eficiente às variáveis do processo.

## 3. Oportunidades Futuras

Quando se tratando de soluções *data-driven*, como o nome implica, mais dados e com maior qualidade ajudariam de maneira geral. Sensoriamento mais amplo, maior frequência de manutenção para garantir confiabilidade e maior taxa de amostragem nos casos laboratoriais mitigariam muitos dos problemas encontrados.

Aliada à dificuldade de sensoriamento, o processo segue em estado constante de mudanças, e dados coletados hoje podem representar distribuições distintas do que o modelo observou no treinamento. Assim, outra possibilidade de melhoria seria criar uma rotina de retreino do modelo para mantê-lo atualizado frente às mudanças de processo que podem prejudicar a performance nos casos em que o conjunto de treino original não represente mais o estado atual da fábrica. Manter uma rotina de retreino, podendo ser automática, por degradação das métricas, ou até mesmo após avaliação manual, pode ajudar a mitigar fatores de degradação da performance e até melhorar os resultados.

Para fins de modelagem, na falta de melhor sensoriamento, podem ser feitos sensores virtuais de variáveis intermediárias, também utilizando aprendizado de máquina, para suprir *gaps* de informação causados por baixa frequência de coleta. Tal trabalho pode ser oneroso pelo número de variáveis, mas usado estrategicamente é capaz de auxiliar nas predições. Ainda na linha dos dados, não foram consideradas as dinâmicas temporais em relação ao tempo de retenção dos processos (e da própria


**Figura 9.** Parâmetros de oxigênio interno e qualidade de efluentes detectados

ETE). Saber que, por exemplo, um determinado pH medido no processo chegará nas lagoas da ETE em três horas pode ser valioso, ao mesmo tempo que complexo, e por isso, no momento, foram feitas algumas abordagens mais simples para tentar inferir dinâmicas temporais, mas no final foram considerados apenas os valores instantâneos. Por fim, sempre é possível adicionar algumas camadas de complexidade e fazer, por exemplo, *ensembles* de modelos, ou modelos híbridos com modelagem fenomenológica e empíricas, mas tais soluções podem onerar em esforços de manutenção, aumentar a dívida técnica da solução e muitas vezes para ganhos marginais, então sua utilização precisa ser avaliada cuidadosamente.

## CONCLUSÕES

A área da inteligência artificial vem fazendo grandes avanços com potencial transformador, não só na indústria de papel em celulose, mas na indústria em geral. No entanto, ainda há grandes desafios dentro do setor para lidar com sistemas altamente complexos que envolvem processos químicos, dinâmicas temporais, instrumentos de coleta de dados ruidosos ou inexistentes e frequentes mudanças no processo. Neste cenário, as soluções de ciência de dados provêm um grande valor no sentido de auxiliar a operação, mas ainda é desafiador fazer predições precisas e dizer quando, como e por que ocorrerá um evento. A solução apresentada agrega como mais uma ferramenta que pode ser utilizada pelos operadores da ETE para antecipar problemas e ter de prontidão algumas possíveis respostas para causas raiz, tendo tanto uma vertente preditiva como uma analítica, mas não dispensa completamente a tomada de decisão e análise de dados e processos feita pelos usuários, sendo mais valiosa quando utilizada conjuntamente com o conhecimento operacional.

## AGRADECIMENTOS

Gostaríamos de expressar nossa sincera gratidão a Ana Paula Kaucs e Anna Eliza Bragança Zóboli, que contribuíram significativamente para o sucesso deste projeto durante seu tempo na empresa. Embora não estejam mais conosco, suas valiosas ideias e esforços deixaram um impacto duradouro. Agradecemos profundamente por suas contribuições e dedicação. ■

## REFERÊNCIAS

Peitz, C. & Schroeder, L. & Xavier, C. Avaliação do tratamento biológico de efluente de fábrica de celulose kraft pela técnica de FT-IR. *O Papel*. v. 80. p. 84-91, 2019.
Furley, T. H., Lombardi, J. B., & Gomes, A. D. S. Principais fontes e impactos da ecotoxicidade de efluentes de celulose e papel. *O Papel*, v. 76 n. 3, p. 51-56, 2015.
Streeter, H. W.; Phelps, E. B. *A Study of the Pollution and Natural Purification of the Ohio River*. Public Heal. Bull. n. 146. U.S. Public Heal. Serv. Washington, DC, USA, 1925.
Henze M., Gujer W., Mino T., van Loosdrecht M.C. *Activated sludge models ASM1, ASM2, ASM2d and ASM3*, 2000.
Batstone D. J., Keller J., Angelidaki I., Kalyuzhnyi S., Pavlostathis S., Rozzi A., Sanders W., Siegrist H., Vavilin V. *The IWA anaerobic digestion model no 1 (ADM1)*. Water Sci. Technol, v. 45, p. 65-73, 2002.
Ocampo-Martinez C. *Model Predictive Control of Wastewater Systems*, 2010. Malviya, A.; Jaspal, D. Artificial Intelligence as an Upcoming Technology in Wastewater Treatment: A Comprehensive Review. *Environ. Technol. Rev*. v. 10, n. 1, p. 177-187, 2021.
Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., & Poch, M. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling and Software*, v. 106, p. 89-103, 2018.
Guoqiang Niu, Xiaohui Yi, Chen Chen, Xiaoyong Li, Donghui Han, Bo Yan, Mingzhi Huang, Guangguo Ying, A novel effluent quality predicting model based on genetic-deep belief network algorithm for cleaner production in a full-scale paper-making wastewater treatment, *Journal of Cleaner Production*, v. 265, 2020.
Zhou, P., Li, Z., Snowling, S. *et al*. A random forest model for inflow prediction at wastewater treatment plants. *Stoch Environ Res Risk Assess*, v. 33, p. 1781-1792, 2019.
Guo, H.; Jeong, K.; Lim, J.; Jo, J.; Kim, Y. M.; Park, J.-p.; Kim, J. H.; Cho, K. H. Prediction of Effluent Concentration in a Wastewater Treatment Plant Using Machine Learning Models. *J. Environ. Sci.*, v. 32, p. 90-101, 2015.
Ly, Q. V.; Truong, V. H.; Ji, B.; Nguyen, X. C.; Cho, K. H.; Ngo, H. H.; Zhang, Z. Exploring Potential Machine Learning Application Based on Big Data for Prediction of Wastewater Quality from Different Full-Scale Wastewater Treatment Plants. *Sci. Total Environ*, v. 832, 2022.
Zaghloul, M. S.; Achari, G. Application of Machine Learning Techniques to Model a Full-Scale Wastewater Treatment Plant with Biological Nutrient Removal. *J. Environ. Chem. Eng*., v. 10 n. 3, 2022.
Nourani, V.; Elkiran, G.; Abba, S. I. Wastewater Treatment Plant Performance Analysis Using Artificial Intelligence - An Ensemble Approach. *Water Sci. Technol*., v. 78 n. 10, p. 2064-2076, 2018.
M. Salomé Duarte, Gilberto Martins, Pedro Oliveira, Bruno Fernandes, Eugénio C. Ferreira, M. Madalena Alves, Frederico Lopes, M. Alcina Pereira, and Paulo Novais. A Review of Computational Modeling in Wastewater Treatment Processes. *ACS ES&T Water*, v. 4, n. 3, p. 784-804, 2024.
Maira Alvi, Damien Batstone, Christian Kazadi Mbamba, Philip Keymer, Tim French, Andrew Ward, Jason Dwyer, Rachel Cardell-Oliver, Deep learning in wastewater treatment: a critical review. *Water Research*, v. 245, 2023.
Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehousing*, v. 22, p. 5-13, 2000.
Schneider M.Y., Quaghebeur W., Borzooei S., Froemelt A., Li F., Saagi R., Wade M. J., Zhu J. J., Torfs E. Hybrid modelling of water resource recovery facilities: Status and opportunities, *Water Sci. Technol*., v. 85, p. 2503-2524, 2022.
Robert Nisbet, Gary Miner, Ken Yale, Chapter 5 – Feature Selection, Handbook of Statistical Analysis and Data Mining Applications, *Academic Press*, p. 83-97, 2018.
A. Kraskov, H. Stogbauer and P. Grassberger, Estimating mutual information. *Phys. Rev*. v. 69, n. 6, 2004.
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. Gene selection for cancer classification using support vector machines, *Mach. Learn*., v. 46 n. 1-3, p. 389-422, 2002.
Houtao Deng and G. Runger. Feature selection via regularized trees, *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Brisbane, QLD, Australia, p. 1-8, 2012.