# APLICAÇÃO DE MODELOS PREDITIVOS DE MACHINE LEARNING PARA PARÂMETROS DE QUALIDADE EM ESTAÇÕES DE TRATAMENTO DE EFLUENTES

**Juliana Neves**[(1)]
Engenheira Química pela Universidade Federal de Santa Catarina (UFSC), com período sanduíche pela École Supérieure de Chimie Physique Électronique de Lyon (CPE Lyon). Consultora da empresa HydroInfo.
**João Vítor Rios Fuck**[(2)]
Engenheiro Químico pela Universidade Federal de Santa Catarina (UFSC). Mestrando em Engenharia Química na Universidade Federal de Santa Catarina (UFSC).
**Maria Alice Prado Cechinel**[(3)]
Engenheira Química pela Universidade do Sul de Santa Catarina (UNISUL). Mestra e Doutora em Engenharia Química pela Universidade Federal de Santa Catarina (UFSC). Pesquisadora na empresa HydroInfo.
**Ricardo Tristão**[(4)]
Engenheiro de Produção pela Universidade do Planalto Catarinense (UNIPLAC). Gerente de Engenharia E&F – Cervejaria Santa Catarina - Ambev.
**Rodrigo Campos de Andrade**[(5)]
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.

**Endereço**[(1)]**:** Rua Emílio Blum, 131, Sala 409, Bloco A – Centro – Florianópolis – CEP:88020-010 – Brasil – Tel: +55 48 3879-6888 – email: jun@hydroinfo.com.br

## RESUMO

Estações de Tratamento de Efluentes (ETEs) apresentam processos bioquímicos complexos de alta variabilidade e difícil previsão. O presente trabalho aborda a utilização de metodologias de *Machine Learning* (ML) para previsão dos valores de saída de nitrogênio total (NT) de uma ETE simulada utilizando *software* WEST da DHI e de demanda química de oxigênio (DQO) da ETE da Cervejaria Santa Catarina - Ambev, de Lages/SC. O estudo aborda uma nova metodologia para lidar com mudanças de cenários de operação das ETEs a fim de evitar a perda de qualidade dos modelos preditivos. Para ambas as estações foram abordados três cenários distintos e avaliada a qualidade das previsões pelos modelos de floresta aleatória (RF), máquina de vetor de suporte (SVM) e *perceptron* multicamadas (MLP). A qualidade das previsões pelo modelo MLP atingiu $R^2$ de 0,72 para previsão de NT de saídas na ETE simulada e o modelo RF melhor se adequou aos dados reais da ETE Ambev, apesar da grande discrepância observada entre os dados reais e previstos. Os resultados obtidos nesse estudo evidenciam a importância da coleta e do armazenamento de dados de qualidade e da necessidade de informação sobre mudanças na operação das ETEs para a performance dos modelos preditivos.

**PALAVRAS-CHAVE:** *Machine Learning*, Tratamento de efluentes, Modelo preditivo.

## INTRODUÇÃO

A otimização das estações de tratamento de efluentes (ETEs) é um desafio constante na busca por melhorias na qualidade e eficiência do processo. Soluções que permitam reduzir custos de operação e ao mesmo tempo melhorar a qualidade do efluente tratado são cada vez mais demandadas por concessionárias e indústrias que necessitam tratar seu efluente para posterior descarte no ambiente ou até mesmo para reutilização em outros processos. Esse movimento de modernização de ETEs e desenvolvimento de novos sensores e técnicas para o monitoramento e controle de parâmetros de qualidade produz um amplo conjunto de dados, geralmente subutilizado pela própria indústria.

Nesse sentido, os procedimentos de modelagem e otimização de processos voltados para sistemas de tratamento de águas e efluentes e que utilizem esses conjuntos de dados coletados têm sido amplamente relatados na literatura nos últimos anos (Al Aani et al., 2019; Ebrahimpour et al., 2008; Fan et al., 2018; Zhao et al., 2020). Uma das tecnologias emergentes que tem sido amplamente utilizada para essa finalidade é o Aprendizado de máquina (ML - *Machine Learning*). Através do treinamento do algoritmo com um grande volume de dados, o ML é capaz de identificar padrões e comportamentos que podem ser usados para aprimorar o funcionamento das ETEs (Tiyasha et

al., 2020). Além disso, os modelos de ML refletem situações reais de reação/processo, em vez de mecanismos formulados com antecedência com base em princípios fundamentais (Ly et al., 2022; Wang et al., 2021).

Dentre as possíveis aplicações de ML em ETEs, os modelos preditivos para parâmetros de qualidade de efluentes são os mais relatados na literatura. São modelos desenvolvidos usando um conjunto de dados históricos e experiências passadas para identificar padrões e tendências e, em seguida, usar esses padrões para prever o valor da variável de interesse no futuro (Yang et al., 2022). Os conjuntos de dados utilizados na construção do algoritmo são categorizados em conjunto de treinamento, teste e validação dos modelos. Vale ressaltar que o número necessário de dados pode variar dependendo da incerteza associada a um determinado parâmetro (Singh et al., 2023). Essa tecnologia tem sido utilizada para previsões de processos complexos, como o processo de lodo ativado (Bagheri et al., 2015), teor de nitrogênio total do efluente (NT) e demanda química de oxigênio em biorreatores de membrana (Almomani, 2020), demanda bioquímica de oxigênio na entrada da estação de tratamento de esgoto (Dogan et al., 2008), índice volumétrico de lodo (Han et al., 2014), entre outros parâmetros. Vários modelos, incluindo Rede Neural Artificial (ANN - *Artificial Neural Network*), K-Vizinhos Mais Próximo (KNN - *K-Nearest Neighbors*), Máquinas de Vetores de Suporte (SVM - *Support Vector Machines*), Memória Longa de Curto Prazo (LSTM - *Long Short-Term Memory)*, Perceptron Multicamadas (MLP - *Multi Layer Perceptron*) e Floresta Aleatória (RF - *Random Forest*), são usados predominantemente para este fim (Singh et al., 2023).

A construção de um banco de dados robusto e unificado é um dos principais desafios para a aplicação bem-sucedida de ML em ETEs. Embora sejam produzidos em quantidade significativa, os dados associados ao funcionamento da ETE e aos parâmetros monitorados são coletados em diferentes intervalos de tempo e a partir de diferentes fontes, sejam através de dados de laboratório ou de sensores instalados ao longo da estação de tratamento, e muitas vezes são anotados em diferentes bases de registro (Newhart et al., 2019). Outra limitação relacionada ao banco de dados está associada à forma de gestão e armazenamento da informação nas ETEs. Alguns sensores são limitados e armazenam apenas informações de alarmes baseados em limites ou anotações de ligado/desligado. Outros parâmetros são anotados com base nos sentidos humanos, ou seja, com base nas observações qualitativas realizadas pelos operadores (Eerikäinen et al., 2020). A unificação e padronização desses dados constitui uma tarefa complexa e desafiadora. Além disso, as variáveis de qualidade do efluente podem mudar repentina ou gradualmente ao longo do tempo e muitas vezes mudam de forma não linear em relação a outras variáveis do processo. ETEs comprometidas com a qualidade do efluente tratado costumam estar em constante atualização do seu processo, seja através da adição de novos sensores, por mudanças na metodologia de análise em laboratório ou inserção de novos equipamentos ao longo da planta. Essas mudanças podem afetar o comportamento dos parâmetros, dificultando a atribuição da causa da mudança entre os eventos de amostragem (Alavi et al., 2022; Newhart et al., 2019) e, consequentemente, dificultando a aplicação de modelos de ML.

O pré-processamento de dados e o *feature engineering* são duas etapas que podem auxiliar na construção de conjuntos de dados mais conveniente para que a "máquina" entenda as relações entre as variáveis do processo e estabeleça modelos robustos de previsão do processo (Bhadeshia et al., 2009). O pré-processamento dos dados inclui as etapas de coleta e limpeza, onde os dados provenientes das diferentes fontes são unificados e problemas como informações redundantes e valores anormais são eliminados. A discretização e normalização de dados, quando necessárias, também fazem parte da etapa de pré-processamento dos dados (Kotsiantis et al., 2007). O *feature engineering* tem como objetivo extrair informações úteis, criar recursos e representar padrões subjacentes a partir de dados brutos para melhorar o desempenho do modelo de aprendizado de máquina. É uma técnica promissora com várias aplicações potenciais por reduzir consideravelmente a dependência de conhecimento especializado no treinamento de modelos e acelerar sua utilização por usuários não especialistas (Cai et al., 2020). Embora o pré-processamento de dados e o *feature engineering* possam mitigar alguns problemas em dados incorretos removendo *outliers*, imputando valores ausentes e transformando variáveis, essas etapas não podem corrigir fundamentalmente dados inerentemente tendenciosos ou imprecisos, ou seja, não substitui boas práticas de coleta de dados.

Nesse estudo, foi avaliado o uso de modelos preditivos de ML para previsão de parâmetros de qualidade de efluentes a partir de dados de entrada e saída obtidos em uma ETE simulada pelo *software* WEST (DHI), uma ferramenta de simulação utilizada em projetos de otimização, operação e a automação de ETEs; e em uma ETE real de uma indústria cervejeira. Três modelos de ML (SVM, RF e MLP) foram testados e o desempenho preditivo de cada modelo foi avaliado por meio de índices de desempenho, Erro Percentual Absoluto Médio (MAPE), Erro Médio Absoluto (MAE), Raiz Quadrada do Erro-Médio (RMSE) e coeficiente de determinação ($R^2$) para dados de teste com intervalo contínuo. Nessa abordagem, os dados de teste são separados seguindo uma sequência ordenada e podem ser usados para testar o desempenho de modelos em situações em que existe uma relação ou tendência contínua nos dados. Eles são particularmente úteis quando se espera que o modelo aprenda a extrapolar além dos dados de treinamento, geralmente para prever valores futuros ou em situações em que a ordem dos dados é

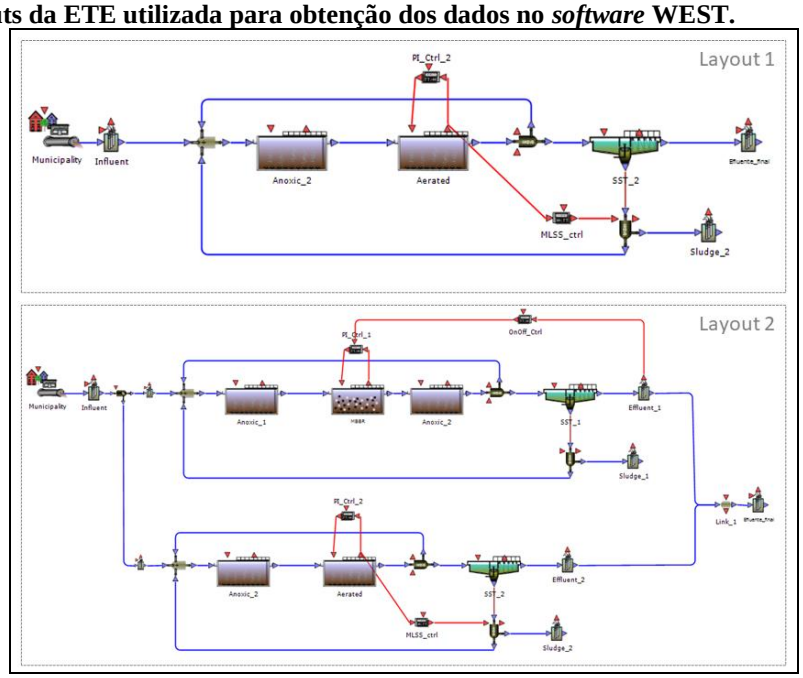
relevante. Este estudo também avaliou os efeitos de modificações na operação da ETE (alterações no layout da planta industrial) no desempenho do modelo preditivo.

## ESTUDO DE CASO E DESCRIÇÃO DO CONJUNTO DE DADOS

Nesse estudo, dois conjuntos de dados foram utilizados para a aplicação dos modelos de ML. O primeiro conjunto de dados foi obtido a partir de dois diferentes layouts de funcionamento de uma Estação de Tratamento de esgoto municipal construída no *software* WEST, uma sofisticada ferramenta de simulação de tratamento de águas residuais, conforme apresentado na Figura 1.

O Layout 1 (Figura 1) foi construído baseado em um caso de teste disponível como amostra no banco de dados do *software* WEST ("*mTestCase_MBBR*") e que é disponibilizado para fins de demonstração e divulgação do *software*. A ETE é composta por dois tanques de lodos ativados com volume constante e mistura ideal e um decantador secundário. Para o primeiro tanque não há fornecimento de oxigênio dissolvido (OD) e para o segundo foi estipulado um *setpoint* fixo de OD em 1,7 mg/L, controlado pelo controlador PI indicado acima do tanque. A simulação dos bioprocessos no *bulk* segue a descrição do modelo de lodos ativados nº1 (*Activated Sludge Model nº1* - ASM1), disponível na biblioteca de modelos *Modelica* no *software* WEST. O modelo adotado para o decantador é baseado numa extensão do modelo de Takács, em que é feita provisão para a estimativa do parâmetro de sedimentação por meio da medição do índice volumétrico de lodo (IVL) (Daigger and Roper, 1985). O sistema ainda conta com um controlador do tipo *on-off* para sólidos suspensos totais (SST), mantendo-se o teor no tanque entre 3500 mg/L e 4800 mg/L. Mais detalhes descritivos sobre os modelos podem ser obtidos em WEST *Models Guide* (DHI, 2022). Os dados de entrada do esgoto bruto foram retirados da base de dados do *software* WEST, o qual apresenta informações de vazão (Q, em $m^3/h$), demanda química de oxigênio (DQO, em mg $O_2$/L), demanda bioquímica de oxigênio (DBO, mg $O_2$/L), sólidos suspensos totais (SST, em mg/L), nitrogênio *Kjeldahl* total (NKT, em mg/L), fósforo total (FT, em mg/L) e sólidos inorgânicos suspensos (SIS, em mg/L), com frequência de 15 minutos. Dentre estas informações, foram aplicados como input para a simulação os dados de Q, DQO, NKT e SST. O resultado da simulação realizada no *software* WEST para o Layout 1 da ETE foi exportado com uma frequência horária de registro equivalente a 31 dias de operação, tanto para os dados de entrada quanto para a saída da ETE, num total de 745 dados para cada variável.

**Figura 1 - Layouts da ETE utilizada para obtenção dos dados no *software* WEST.**



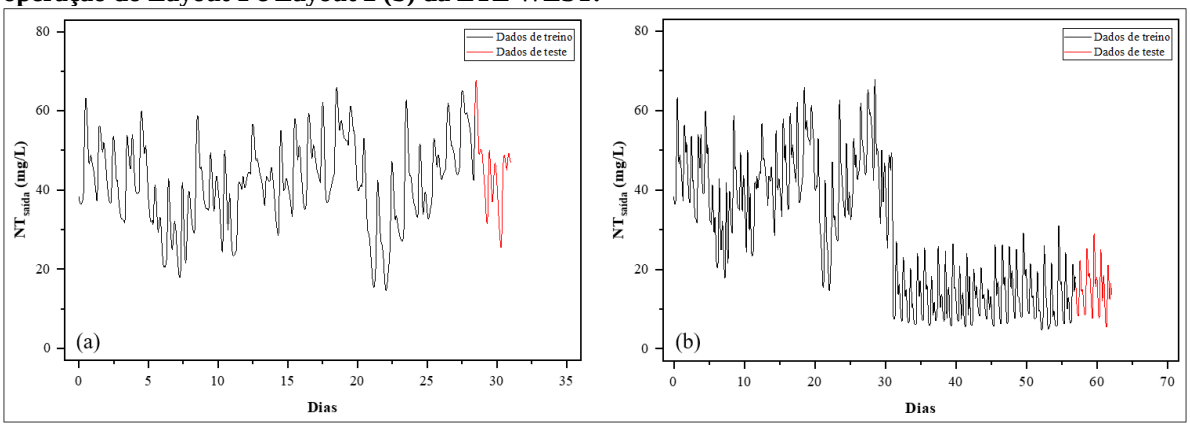

Fonte: Os autores, 2023.

No Layout 2, o fluxograma de operação descrito no Layout 1 foi mantido e, paralelo a ele, foi adicionada uma nova linha de operação. Essa linha é composta por dois tanques de lodos ativados com volume constante e mistura ideal, um reator biológico de leito móvel (MBBR) com volume constante, mistura ideal e biofilme

em modelo unidimensional, e com gradiente de concentração distribuído em 10 camadas homogêneas, além de um decantador secundário. Novamente, os dois tanques são mantidos em condição anóxica, sem fornecimento de OD, e a simulação dos bioprocessos no bulk segue a descrição do modelo ASM1. O mesmo modelo também descreve os processos bioquímicos que ocorrem no bulk líquido e no biofilme do tanque MBBR: o primeiro considera os processos de degradação realizados pela biomassa dispersa; o biofilme é modelado como uma estrutura unidimensional, na qual o gradiente de concentração é simulado através de 10 camadas homogêneas (idealmente misturadas). O reator MBBR ainda conta com um controlador on/off e um controlador PI para OD com um setpoint variável em função do nível de nitrogênio amoniacal (NH) no efluente tratado, variando entre 3 mg $O_2$/L, para uma condição de $NH_{min}$ = 1,5 mg/L, e 4 mg $O_2$/L, para uma condição de $NH_{max}$ = 3,0 mg/L. Mais detalhes descritivos sobre os modelos podem ser obtidos em WEST *Models Guide* (DHI, 2022). Os dados de entrada do esgoto bruto foram os mesmos utilizados na simulação anterior, com os dados de Q, DQO, NKT e SST aplicados como input para a simulação e com a vazão de entrada igualmente dividida entre as linhas. O resultado dessa segunda simulação realizada no software WEST também foi exportado com uma frequência horária de registro equivalente a 31 dias de operação, tanto para os dados de entrada quanto para a saída da ETE, num total de 745 dados para cada variável.

Os dados obtidos a partir das duas simulações foram unificados em uma base de dados contínua, sendo os primeiros dados correspondentes a simulação realizada para o Layout 1 e os dados seguintes correspondentes a simulação realizada para o Layout 2. Dessa forma, a base de dados completa simula 62 dias de operação de uma estação de tratamento de esgoto, havendo nesse período uma alteração na operação da planta com o objetivo de melhorar a qualidade do efluente final. O conjunto de dados obtidos a partir da simulação realizada pelo software WEST apresentava dados de entrada e de saída para as seguintes variáveis de processo: DBO, DQO, nitrogênio amoniacal (NH, em mg/L), nitrato (NO, em mg/L), Q, NKT, NT e SST. A Figura 2 apresenta o perfil de saída do parâmetro NT, utilizado nesse estudo como parâmetro de previsão para a ETE WEST, para os 31 dias de operação do Layout 1 (Figura 2a) e os 62 dias de operação do Layout 1 e Layout 2 (Figura 2b). Ao analisar a figura, é possível observar que há uma diferença significativa entre os valores apresentados para o Layout 1, que variam entre 25,42 mg NT/L e 67,75 mg NT/L, e Layout 2, com valores entre 5,45 mg NT/L e 28,88 mg NT/L. Uma vez que os dados de entrada são os mesmos em ambas as simulações realizadas pelo software WEST, a redução do teor de nitrogênio total na saída pode ser justificada pela inserção da segunda linha de operação no sistema.

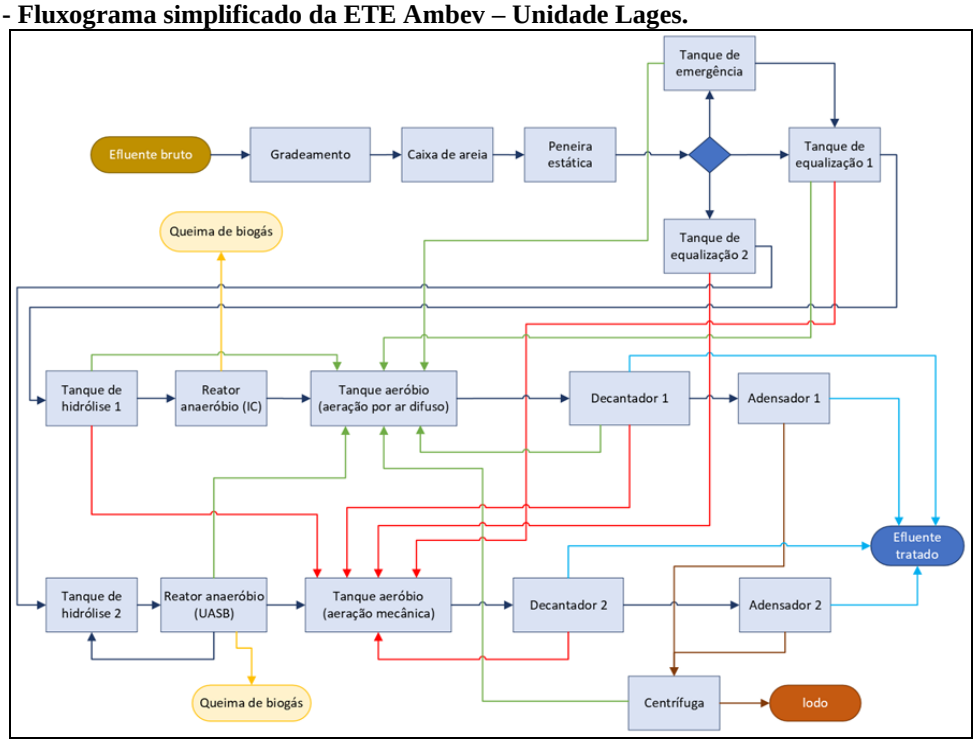
**Figura 2 - Perfil de saída do parâmetro NT para os 31 dias de operação do Layout 1 (a) e os 62 dias de operação do Layout 1 e Layout 2 (b) da ETE WEST.**




Fonte: Os autores, 2023.

O segundo conjunto de dados utilizado nesse estudo foi coletado na estação de tratamento de efluentes industriais da fábrica da Cervejaria Santa Catarina - Ambev, localizada em Lages/SC. A ETE opera continuamente em sistema de turnos e adota um processo de tratamento misto, com processo de tratamento anaeróbio seguido de tratamento aeróbio do efluente. Um fluxograma simplificado do processo é apresentado na Figura 3.

**Figura 3 - Fluxograma simplificado da ETE Ambev – Unidade Lages.**




Fonte: Os autores, 2023.

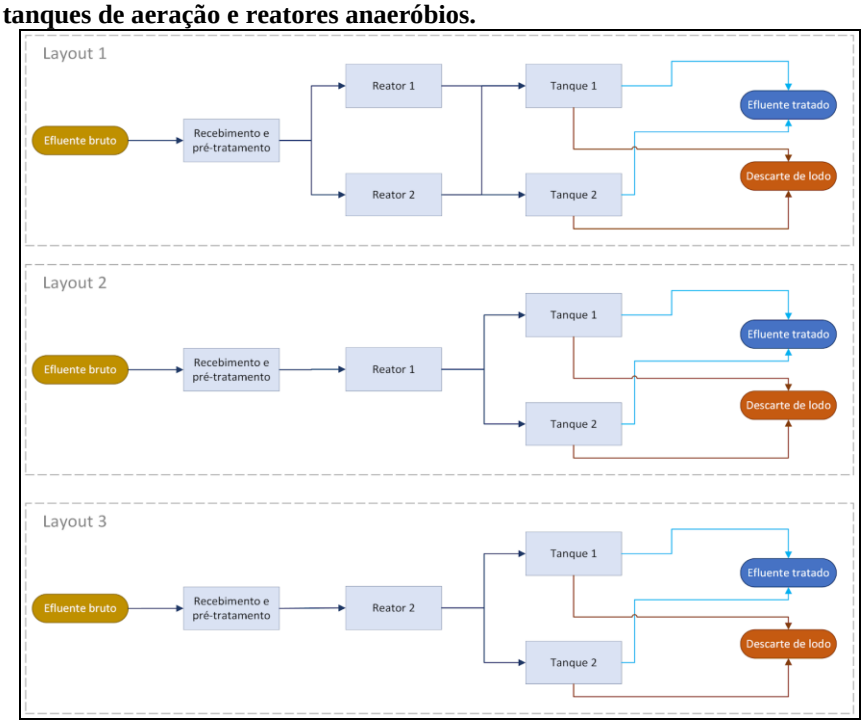
O processo de tratamento da ETE pode ser dividido em 4 macro etapas: recebimento do efluente e pré-tratamento, tratamento anaeróbio, tratamento aeróbio e descarte de lodo e de efluente tratado. A capacidade de tratamento diária da ETE é de 90 m³/h e o padrão de lançamento de poluentes, definido pelas legislações ambientais para emissão de efluentes em vigência, é alcançado no efluente tratado. A ETE possui dois reatores anaeróbios em sua linha de operação, um do tipo reator anaeróbio de fluxo ascendente (UASB), denominado Reator 1, e outro do tipo reator de circulação interna (IC), denominado Reator 2, de 3000 m³ e 700 m³, respectivamente. A etapa de tratamento aeróbio é feita em sistema de lodos ativados. Para essa etapa, a ETE possui dois tanques de aeração: um com capacidade para 2860 m³ de efluente e com aeração mecânica de superfície (denominado Tanque 1) e outro com 3500 m³ e aeração submersa por ar difuso (denominado Tanque 2), além de dois tanques de decantação cilíndricos. O tempo de residência médio do efluente na ETE é equivalente a 6,8 dias, ou seja, o efluente bruto leva, em média, entre 6 e 7 dias para sair da ETE como efluente tratado.

O controle e monitoramento da ETE é realizado pelos operadores responsáveis pelo setor em regime de turnos, utilizando para apoio um programa supervisório específico que fornece informações sobre a operação, tais como vazão e níveis dos tanques, potência de equipamentos, controle de pH e dosagem de reagentes, níveis de oxigenação dos tanques de aeração etc. Em cada etapa do processo, amostras de efluente são coletadas uma vez por turno, totalizando três coletas diárias. A ETE possui um laboratório químico que possibilita a caracterização das amostras coletadas para uma série de parâmetros físico-químicos na própria unidade. Algumas análises de caracterização e monitoramento são terceirizadas em situações de auditoria e fiscalização ambiental. Os indicadores de qualidade do efluente apresentam variabilidade temporal de coleta e preenchimento de dados (por turno, diário, semanal, quinzenal ou mensal).

Para a execução deste trabalho, a equipe da Ambev disponibilizou os dados armazenados entre janeiro de 2021 e janeiro de 2023. Os dados fornecidos possuem duas diferentes fontes: as informações armazenadas em sistema de gestão de processos (LiveMES) e as informações anotadas em planilhas de controle interno da ETE. Nesse estudo, foram utilizados os dados de entrada do efluente bruto e de saída do efluente tratado organizados em uma escala diária, totalizando 752 dias. No período que abrange os dados em estudo, a ETE Ambev realizou melhorias e manutenções em sua planta industrial com o intuito de melhorar a qualidade do efluente final e a operação como um todo. Embora a base de dados não forneça todas as informações sobre essas melhorias e manutenções, as informações sobre o funcionamento dos tanques de aeração e dos reatores

anaeróbios são apresentadas, de forma que foi possível categorizar a base de dados a partir das informações de funcionamento desses equipamentos em três layouts simplificados, conforme apresentado na Figura 4.

**Figura 4 - Fluxograma simplificado da operação da ETE Ambev – Unidade Lages em função da operação dos tanques de aeração e reatores anaeróbios.**

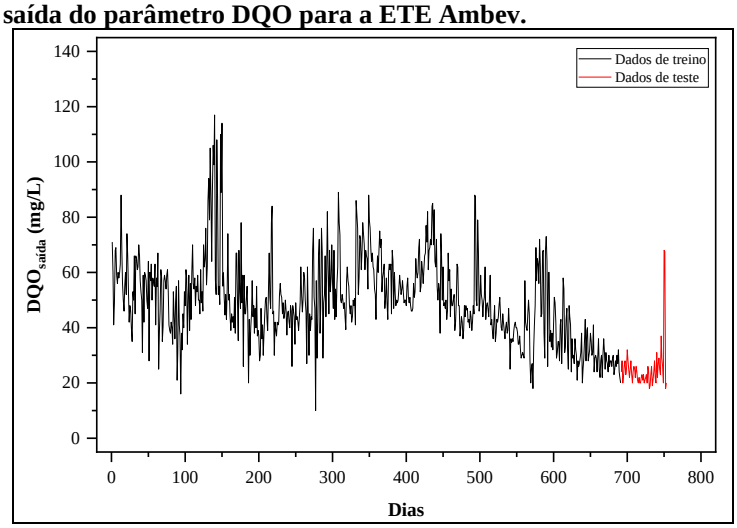



Fonte: Os autores, 2023.

Nesse estudo, foram utilizados os parâmetros Q, DQO, carga orgânica (CO, em mg DQO/L), temperatura (T, em °C) e pH, coletados tanto na entrada da estação quanto em sua saída. Embora a base de dados originalmente fornecida pela empresa apresentasse mais de cem variáveis coletadas, muitos desses dados não eram relevantes para o monitoramento e controle da ETE, tais como dados relacionados ao funcionamento da estação de tratamento de águas (ETA), da unidade industrial ou de monitoramento da qualidade do corpo hídrico receptor. Além disso, um elevado percentual de dados ausentes foi verificado e a periodicidade no preenchimento das informações de alguns parâmetros também não era consistente. Os parâmetros selecionados estão relacionados com a qualidade do efluente e apresentavam boa regularidade de dados, porém é importante destacar que a falta de dados consistentes para alguns parâmetros de qualidade de efluentes, como DBO, nitrogênio total e fósforo, podem afetar consideravelmente o desempenho das próximas etapas.

A Figura 5 ilustra o perfil de saída do parâmetro DQO, que foi definido como parâmetro de previsão para a ETE Ambev por representar um parâmetro crucial de qualidade do efluente. Durante o período de coleta de dados, que abrangeu dois anos, observou-se uma oscilação dos valores de DQO de saída entre o máximo registrado de 117 mg/L e o mínimo de 10 mg/L. No entanto, é perceptível que no último ano houve uma tendência decrescente no valor de DQO. No primeiro ano, a média de DQO de saída foi de 54 mg/L e no segundo ano esse valor médio diminuiu para 42 mg/L. É importante ressaltar que nos últimos 6 meses desse segundo ano, ocorreu uma redução ainda mais significativa, com um valor médio de DQO de saída igual a 32 mg/L. Essa diminuição gradual pode ser atribuída às diversas modificações e melhorias implementadas na operação das Estações de Tratamento de Efluentes (ETE).

**Figura 5 - Perfil de saída do parâmetro DQO para a ETE Ambev.**




Fonte: Os autores, 2023.

## PRÉ-PROCESSAMENTO DOS DADOS

A análise exploratória de dados (EDA) refere-se ao processo de resumir um conjunto de dados usando estatísticas numéricas informativas, tabelas bem construídas ou visualizações gráficas para possibilitar a avaliação da qualidade e fidelidade dos dados, da validade de pressupostos específicos necessários para inferência estatística e das estratégias analíticas e estatísticas apropriadas no lugar das propriedades exibidas pelo conjunto de dados (Tukey, 1977). A EDA foi realizada tanto para a base de dados obtida para a ETE simulada pelo software WEST, a partir desse momento denominada de ETE WEST, quanto para a base de dados obtida para a ETE Ambev. Foram aplicadas análises univariadas e bivariadas com o intuito de identificar o comportamento das variáveis e suas relações, além de detectar a presença de dados problemáticos ou não representativos e auxiliar na tomada de decisão. A construção de gráficos de série temporal para cada uma das variáveis do processo e análises estatísticas foi realizada para obter valores de média, mediana, variância e desvio padrão. As bibliotecas *pandas* (Reback et al., 2021), *pandas-profiling* (YData Labs Inc, 2023) e *sweetviz* (Bertrand, 2022) do Python® foram utilizadas para essas análises (van Rossum and de Boer, 1991). As análises univariadas foram responsáveis pela exclusão de outliers extremos e as análises bivariadas de correlação entre as diferentes variáveis de processo possibilitaram a exclusão de informações altamente correlacionadas e, portanto, redundantes aos modelos de ML, seguindo os princípios da Navalha de Ockham (Baker, 2022).

Nesse estudo, para remoção dos dados anômalos e outliers foi aplicada a condição de 3σ em relação à média (Zhu et al., 2022), sendo o desvio padrão σ determinado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \eta_i^2}{m-1}} \qquad \text{equação (1)}$$

$$\eta_i = X_i - \bar{X} \qquad \text{equação (2)}$$

sendo *m* o número de variáveis na coleta de informações, $X_i$ um dado coletado, $\bar{X}$ é a média das informações e $\eta_i$ é o erro resultante. Se o erro residual de um dado $X_i$ atender ao critério $|\eta_i| > 3\sigma$, isso indica que o dado deve ser descartado do conjunto de dados.

A partir dessas análises foi possível identificar os parâmetros que não trariam informações significativas a qualquer algoritmo de ML estudado em virtude da baixa variância apresentada entre os dados. Também foi possível identificar e eliminar valores outliers e verificar a ausência de dados para as variáveis de processo analisadas. Para minimizar o impacto da exclusão dos outliers e manter a continuidade dos registros da base, utilizou-se a técnicas de imputação de dados de K-vizinhos mais próximos (*KNN Imputation*), conforme apresentado por Troyanskaya et al. (2001).

A *Feature Engineering* foi aplicada em ambas as bases de dados de forma distinta visando fornecer maiores informações aos modelos preditivos de ML. Para a ETE WEST, uma nova variável categórica foi incluída com o objetivo de indicar o modo de operação da ETE. Dessa forma, atribuiu-se o valor de 0 para os primeiros 31 dias de simulação, correspondentes ao Layout 1 (Figura 1) com apenas uma linha de operação, e o valor de 1 para os 31 dias subsequentes, correspondentes ao Layout 2 (Figura 1) com as duas linhas de operação em paralelo. Desta forma, foi possível fornecer aos algoritmos de ML a informação sobre a mudança na operação e a sua relação com os dados de entrada e saída.

Para a ETE Ambev, também foi incluída a nova variável de indicação do modo de operação da ETE adicionando-se a base de dados três novas colunas: a primeira refere-se ao funcionamento do Reator 1, sendo atribuído o valor de 0 aos dias em que este encontrava-se em funcionamento e o valor de 1 para os dias em que estava desativado (Layout 3, Figura 4); a segunda coluna refere-se ao funcionamento do Reator 2, atribuindo o valor de 0 para os dias em que o Reator 2 encontrava-se em funcionamento e 1 para os dias em que estava desativado (Layout 2, Figura 4); e a última coluna refere-se ao funcionamento dos dois reatores simultaneamente, sendo atribuído o valor de 1 para os dias em que ambos reatores encontrava-se em operação (Layout 1, Figura 4) e 0 para os dias em que algum dos dois reatores não estava em operação. Também foram incluídas variáveis que indicam o mês e o dia da semana de cada ocorrência, visando fornecer aos algoritmos de ML informações sobre o comportamento sazonal associado à produção da planta industrial e, consequentemente, do montante de efluente gerado. Estas variáveis foram transformadas em binários (0 ou 1) de acordo com a metodologia de *One-Hot Encoding* (Harrag and Gueliani, 2020). Além disso, uma vez que o tempo de residência médio do efluente na ETE é equivalente a 6,8 dias, foi realizado um retrocesso temporal dos dados de entrada em relação aos dados de saída para alinhar temporalmente as previsões da variável de saída com os dados de entrada que serão fornecidos para treinamento e teste.

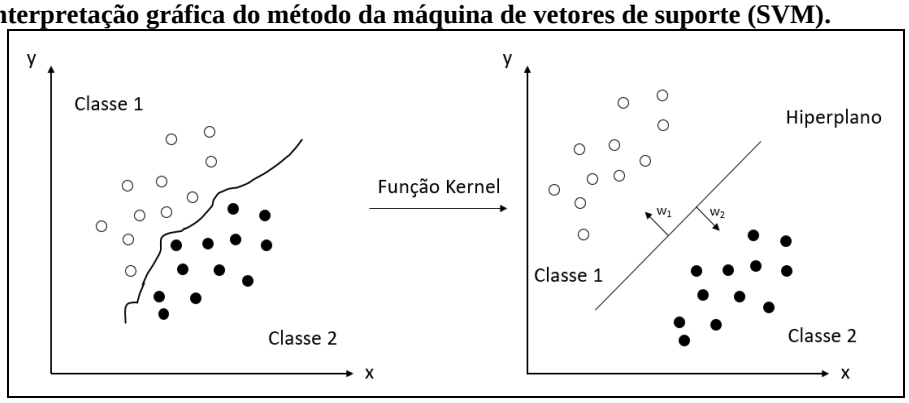
## MODELOS DE MACHINE LEARNING

Neste trabalho, três algoritmos de machine learning foram adotados para estabelecer modelos preditivos da qualidade do efluente: Máquinas de Vetores de Suporte (SVM - *Support Vector Machines*), *Perceptron* Multicamadas (MLP - *Multi Layer Perceptron*) e Floresta Aleatória (RF - *Random Forest*).

SVM são modelos de aprendizado supervisionado inicialmente usados para problemas de classificação e para soluções robustas de regressão (Barnat-Hunek et al., 2021), sendo baseados na minimização de risco estrutural, que reduz o *over-fitting* e aumenta a generalização, minimizando o erro projetado do modelo de aprendizado (Wu and Lin, 2015). O SVM não fornece uma estrutura pré-determinada, pois as contribuições das amostras de treinamento julgam as contribuições dos conjuntos de dados de treinamento. Apenas amostras de dados escolhidas são usadas para o desenvolvimento do modelo final, conhecidas como "vetores de suporte" (Farooq et al., 2021). A Figura 6 representa o processo de modelagem e deslocamento de dados em um espaço dimensional escolhido.

Os dados são representados como um mapa de pontos no espaço e a solução é o hiperplano (pista em 2D, plano em 3D etc.) com o maior gap possível entre duas classes. Cada ponto neste espaço é descrito com vetores de suporte; entretanto, existem algumas situações em que a divisão do conjunto de dados só é possível após o uso das funções kernel. As funções kernel mais utilizadas são a função de base radial (RBF), funções lineares, gaussianas, polinomiais e não lineares (Piri et al., 2015).

**Figura 6 - Interpretação gráfica do método da máquina de vetores de suporte (SVM).**



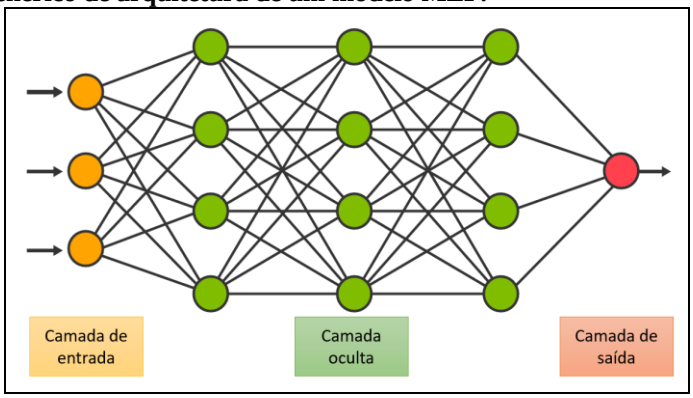

Fonte: Adaptado de Farooq et al. (2021).

Para a construção do modelo SVM, utilizou-se a biblioteca *scikit-learn* (Pedregosa et al., 2011) a partir de sua classe *svm.SVR*, a qual é referente ao modelo de suporte de vetor para regressões com *epsilon*. Foram testados diversos hiperparâmetros para a construção do modelo de forma a melhor ajustar os dados fornecidos. Para todos os cenários estudados, foram executadas diversas configurações de hiperparâmetros selecionados, conforme metodologia de Pesquisa de Grade Exaustiva (*Exhaustive Grid Search*) (Liashchynskyi and Liashchynskyi, 2019). As funções kernel linear, polinomial, RBF e sigmóide foram testadas na construção do modelo. Exclusivamente para as funções polinomiais, foram testadas as funções de segundo a décimo grau, de um em um. O parâmetro de regularização (C) foi testado com os valores de 0,5, 1, 2, 3 e 5. O hiperparâmetro *epsilon*, que especifica a distância para qual erros no treinamento não têm penalidades atribuídas, foi testado com valores de 0,01, 0,05, 0,1, 0,5 e 1. A tolerância para critério de parada foi adotada como $10^{-4}$, de modo a garantir um treinamento mais completo. Os demais hiperparâmetros do modelo de regressão por suporte de vetor foram mantidos conforme os padrões da biblioteca *scikit-learn*. Durante o treinamento utilizando a técnica de Pesquisa de Grade Exaustiva, a validação cruzada com cinco dobras (5-fold) foi implementada. Essa validação consiste na divisão da base de dados em cinco sub-bases, de modo que todas as cinco sub-bases sejam utilizadas uma vez como base de validação e, para as demais quatro execuções do algoritmo, funcionem como bases de treino (Berrar, 2019). A escolha do melhor modelo, após a Pesquisa de Grade Exaustiva de todos os hiperparâmetros acima listados e validação cruzada, foi feita com base no MAPE.

O modelo *Perceptron* Multicamadas (MLP) é um tipo de redes neurais artificiais (ANN) *feedforward* com alto grau de conectividade determinado pelos pesos sinápticos da rede (Lorencin et al., 2020). Isso significa que todos os nós da camada existente estão vinculados à próxima camada. Uma MLP é composta por três ou mais camadas, incluindo uma camada de entrada, uma camada de saída, bem como uma ou mais camadas ocultas (Figura 7). Na camada oculta, cada neurônio artificial contém uma função de ativação, linear ou não linear, e utiliza uma técnica de aprendizado supervisionado (Pal and Mitra, 1992; Yu et al., 2021). O MLP é treinado usando o método de aprendizado supervisionado chamado algoritmo de retropropagação e pode ser dividido em duas fases: na fase *forward*, os pesos sinápticos do MLP são atualizados à medida que o sinal se propaga pela rede. Nesta fase, as mudanças estão confinadas aos potenciais de ativação e saídas dos neurônios na rede. Na fase *backward*, o sinal de erro é produzido como diferença entre a saída gerada e a saída desejada e é, então, propagado para trás até atingir o peso sináptico e ser ajustado (Hao et al., 2023; Lorencin et al., 2020).

Para a construção desse modelo, foi utilizada a biblioteca *scikit-learn* a partir de sua classe *neural_network.MLPRegressor*. Diversos hiperparâmetros foram testados para a construção da rede neural com melhor ajuste aos dados fornecidos. Em todos os cenários, foram executadas as possíveis configurações dos hiperparâmetros selecionados, conforme metodologia de Pesquisa de grade exaustiva. Foram testadas conformações com 2, 3 e 4 camadas ocultas, todas com mesmo número de neurônios variando entre 16, 32, 50 e 100. Os otimizadores de pesos testados foram *Limited-memory Broyden–Fletcher-Goldfarb–Shanno algorithm* (LBFGS) (Zhu et al., 1997), *Stochastic gradient descent algorithm* (SGD) (Bottou, 2010) e *Adaptive learning rate optimization algorithm* (Adam) (Kingma and Ba, 2014). As taxas de aprendizado dos modelos implementadas foram mantidas constantes e seus valores testados na técnica de *Grid Search* foram 0,0001, 0,001 e 0,01. O parâmetro de regularização L2 (alfa) (Schonlau and Zou, 2020) variou entre 0,0001, 0,0005, 0,001, 0,005, 0,01 e 0,05. O número máximo de iterações (épocas) foi testado para 500 e 1000, a fim de aumentar a convergência do modelo. A tolerância definida de $10^{-5}$ é um valor muito baixo para a natureza do processo estudado e, por isso, garantiu-se que todas as épocas foram executadas sem parada precoce de treinamento.

**Figura 7 - Exemplo genérico de arquitetura de um modelo MLP.**

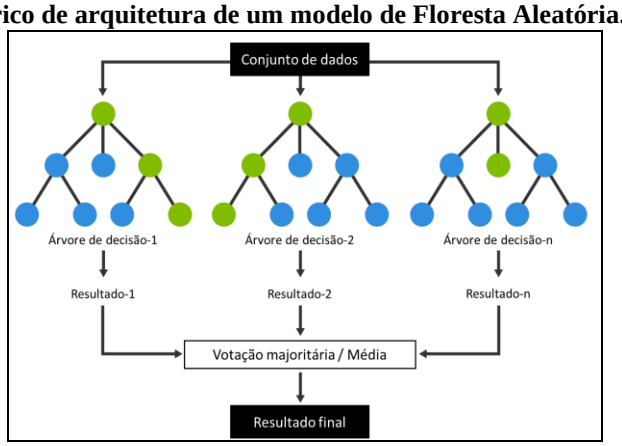



Fonte: Os autores, 2023.

A função de ativação é o elemento chave de um neurônio artificial e, nesse estudo, a função de ativação *Relu* foi usada para todos os neurônios (Fukushima, 1975). O tamanho da batelada utilizado nesse estudo foi de 200 amostras, esse é um importante hiperparâmetro para o treinamento da rede neural, visto que define o número de amostras utilizadas para o treinamento da rede. Com o intuito de possibilitar a reprodutibilidade dos resultados, foi utilizado o valor 42 para o estado aleatório de inicialização de pesos e vieses e para amostragem de bateladas quando o otimizador for o SGD ou o Adam. Demais hiperparâmetros da rede neural foram mantidos os padrões da biblioteca *scikit-learn*. Durante o treinamento utilizando a técnica de Pesquisa de Grade Exaustiva, foi implementada a validação cruzada com cinco dobras (*5-fold*), tal como descrito no modelo anterior. A escolha do melhor modelo após a Pesquisa de Grade Exaustiva de todos os hiperparâmetros acima listados, com validação cruzada, foi feita utilizando o MAPE.

O modelo de Floresta Aleatória, *ou Random Forest* (RF), é um algoritmo de aprendizado supervisionado que combina múltiplas árvores de decisão, todas diferentes umas das outras, para realizar classificação e regressão (Breiman, 2001). As variáveis preditoras são selecionadas e os nós são divididos com base na melhor divisão possível nessas variáveis. Quando *Random Forest* prevê uma nova amostra de acordo com algumas características, cada árvore em *Random Forest* dará seu próprio resultado de classificação e voto, e então a saída total da floresta será a categoria que obtém o maior número de votos (Figura 8).

**Figura 8 - Exemplo genérico de arquitetura de um modelo de Floresta Aleatória.**




Fonte: Os autores, 2023.

No problema de regressão, *Random Forest* gera a saída média de todas as árvores de decisão (Liu et al., 2020). Isso permite que muitos classificadores fracamente correlacionados formem um classificador forte (Rodríguez-Martín et al., 2020). Um dos benefícios da *Random Forest* é que ela permite determinar intuitivamente a importância das variáveis em um problema de regressão ou classificação. Essa importância é calculada medindo a diminuição da impureza em cada nó usado para o particionamento dos dados (Flores and Leiva, 2021). Além disso, a *Random Forest* lida bem com muitas variáveis de entrada, equilibra erros em conjuntos de dados desequilibrados e possui uma boa velocidade e precisão em sua regressão (Liu et al., 2020).

Para a construção do modelo RF, foi utilizada a biblioteca *scikit-learn* a partir de sua classe *ensemble.RandomForestRegressor*. Novamente, diversos hiperparâmetros foram testados na construção do modelo com melhor ajuste aos dados fornecidos. Para todos os cenários estudados, foram executadas todas as combinações possíveis dos hiperparâmetros selecionados, conforme metodologia de Pesquisa de grade exaustiva. Foram testadas conformações com número de árvores na floresta de 1 (caso excepcional de Árvore de Decisão), 3, 5, 10, 100, 200, 400 e 500. A profundidade máxima das árvores foi testada com 2, 3, 5, 10, 15 e 20. Vale destacar que árvores muito rasas têm tendência ao *underfitting*, enquanto há maior variância do modelo em árvores mais profundas (Kohavi and Wolpert, 1996; Neal, 2019). Portanto, a otimização da profundidade das árvores é um importante ponto de estudo. O número mínimo de amostras a serem divididas em um galho da árvore foi adotado como 2, de modo a possibilitar a maior variância possível entre as árvores criadas e possibilitando que cada folha terminal contenha apenas uma amostra. Por sua vez, também foi considerada obrigatória a presença de um valor para cada folha, de modo a não haver folhas sem amostras durante o treinamento. O critério de aprendizado para avaliar a qualidade da divisão foi adotado como o erro absoluto médio, que minimiza a perda L1 usando a mediana de cada nó terminal (Pedregosa et al., 2011). Os demais hiperparâmetros do modelo de floresta aleatória foram mantidos os padrões da biblioteca *scikit-learn*. Assim como para os demais modelos, durante o treinamento utilizando a técnica de Pesquisa de Grade Exaustiva, foi implementada a validação cruzada com cinco dobras (*5-fold*), de modo a garantir maior robustez

ao modelo em virtude da utilização da totalidade da base de treinamento como dados de treinamento e como dados de validação, alternadamente (Buitinck et al., 2013). Para a escolha do melhor modelo, após a Pesquisa de Grade Exaustiva de todos os hiperparâmetros acima listados, com validação cruzada, foi adotada a métrica MAPE.

As métricas de performance Raiz quadrada do erro-médio (RMSE), Erro médio absoluto (MAE), Erro Percentual Absoluto Médio (MAPE) e coeficiente de determinação ($R^2$) foram escolhidas como função de perda em todos os modelos de previsão considerados e são descritos por (Ly et al., 2022):

$$RMSE = \frac{\sum_{i=1}^{N}(Y_i - Y_i')^2}{N} \quad \text{equação (3)}$$

$$MAE = \frac{\sum_{i=1}^{N}|Y_i - Y_i'|}{N} \quad \text{equação (4)}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - Y_i'}{Y_i}\right| \quad \text{equação (5)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - Y_i')^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} \quad \text{equação (6)}$$

onde $N$ é o número de amostras, $Y_i$ e $Y_i'$ são os valores observados e preditos, respectivamente, na saída de uma i-ésima amostra, e $\overline{Y}$ é a média de todos os valores de saída. O RMSE é uma medida do desvio médio dos valores previstos em relação aos valores observados. Quanto menor o valor do RMSE, melhor será a precisão do modelo. No entanto, é importante levar em consideração a escala da variável de resposta ao interpretar o RMSE, pois ele é expresso na mesma unidade da variável de resposta. O MAE é a média das diferenças absolutas entre os valores previstos e os valores observados. Assim como o RMSE, quanto menor o valor do MAE, melhor será a precisão do modelo. O MAE também é expresso na mesma unidade da variável de resposta. O MAPE é uma medida da média das diferenças percentuais absolutas entre os valores previstos e os valores observados, expressa em termos absolutos ou percentuais. Ele fornece uma visão relativa do erro em relação ao tamanho dos valores observados. Normalmente, quanto menor o valor do MAPE, melhor será o desempenho do modelo. No entanto, é importante observar que o MAPE pode ser influenciado por valores próximos a zero, pois resulta em uma divisão por zero. Por fim, $R^2$ avalia quão bem os valores previstos pelo modelo se ajustam aos valores observados. Ele varia de 0 a 1 e quanto mais próximo de 1 for o valor de $R^2$, melhor será o desempenho do modelo. No entanto, o $R^2$ também pode ser negativo se o modelo for pior do que um modelo ingênuo que simplesmente prevê a média dos dados. O $R^2$ pode assumir um valor negativo quando o modelo não está capturando a variação dos dados e é essencialmente inútil para fazer previsões.

## RESULTADOS E DISCUSSÕES

Com relação ao estudo realizado para a ETE WEST, separou-se a base de dados em variáveis independentes (dados de entrada de DBO, DQO, NH, NO, Q, NKT, NT e TSS) e variável dependente (NT de saída). A variável nitrogênio total da saída foi definida como parâmetro de previsão por apresentar a maior variância, (276,60 mg $NT^2/L^2$), do que as demais variáveis de qualidade da saída, como a DQO (75,72 mg $O_2^2/L^2$). A partir dessa definição, os modelos preditivos para o parâmetro NT foram aplicados em três diferentes cenários:

- O cenário 1 considera apenas os dados obtidos a partir da simulação realizada para o Layout 1 da Figura 1. A base de dados utilizada possuía 745 registros em intervalos diários, totalizando 31 dias de operação. Desses registros, as primeiras 683 entradas para todas as variáveis foram definidas como conjunto de dados de treinamento dos modelos para previsão de NT na saída da ETE. As últimas 62 entradas foram definidas como conjunto de dados de teste dos modelos. O conjunto de dados utilizando para teste e treino dos modelos de ML estão graficamente representados na Figura 2a.
- No cenário 2, a base de dados utilizada é composta pela junção dos dados obtidos a partir da simulação realizada para o Layout 1 (Figura 1) aos dados obtidos a partir da simulação realizada para o Layout 2 (Figura 1), totalizando 62 dias de operação (31 dias contínuos de operação do Layout 1 e 31 dias contínuos de operação do Layout 2). O conjunto de dados de treino e teste foi definido mantendo-se a razão teste/treino adotada no cenário 1 (teste/treino = 0,09). Dessa forma, as primeiras 1367 entradas para todas

as variáveis foram definidas como conjunto de dados de treinamento e as últimas 123 entradas foram definidas como conjunto de dados de teste dos modelos para previsão de NT na saída da ETE. O conjunto de dados utilizando para teste e treino dos modelos de ML estão graficamente representados na Figura 2b.

- No cenário 3, por sua vez, manteve-se a base de dados aplicada ao cenário 2, porém com a informação adicional da variável categórica incluída para identificar o modo de operação da ETE. Tal como mencionado anteriormente, essa informação visa facilitar aos modelos o reconhecimento de um novo padrão de operação e a sua relação com os dados de entrada e saída.

Os dados de NT de saída previstos foram comparados com os valores reais de NT de saída do intervalo de teste e a acurácia de cada modelo foi avaliada a partir das métricas de performance (RMSE, MAE, MAPE e $R^2$) apresentadas na Tabela 1. Uma análise gráfica comparativa entre os dados previstos pelos modelos e os dados obtidos pela simulação da ETE WEST também é apresentada na Figura 9.

**Tabela 1 - Comparação de métricas de performance para os modelos preditivos de ML para NT aplicados a ETE WEST.**

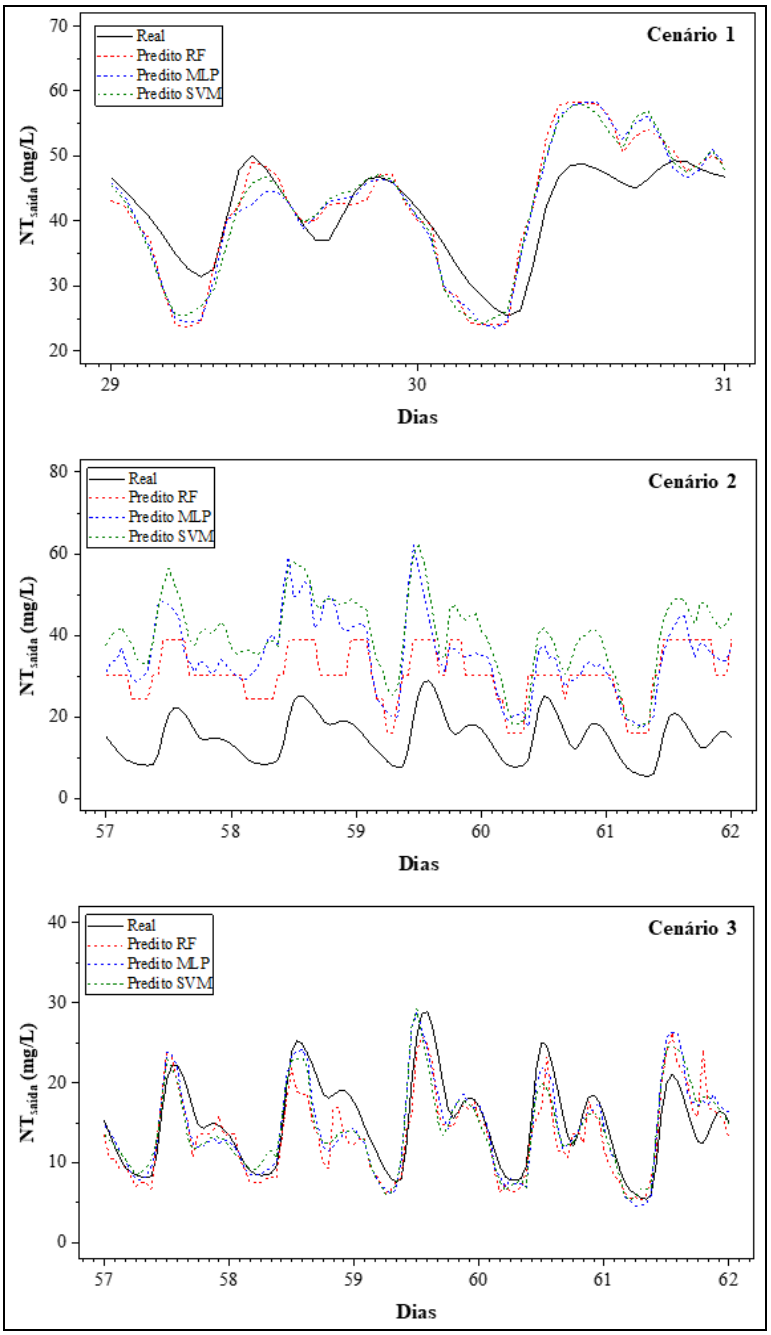
| | **Cenário 1** | | | **Cenário 2** | | | **Cenário 3** | | |
|---|---|---|---|---|---|---|---|---|---|
| | SVM | MLP | RF | SVM | MLP | RF | SVM | MLP | RF |
| RMSE | 5,76 | 5,75 | 5,87 | 25,88 | 21,03 | 16,28 | 3,18 | 2,87 | 3,39 |
| MAE | 4,52 | 4,48 | 4,61 | 24,96 | 19,85 | 15,58 | 2,49 | 2,21 | 2,55 |
| MAPE | 0,11 | 0,11 | 0,11 | 1,85 | 1,5 | 1,2 | 0,16 | 0,14 | 0,16 |
| $R^2$ | 0,6 | 0,6 | 0,58 | -21,44 | -13,82 | -7,89 | 0,66 | 0,72 | 0,61 |

Fonte: Os autores, 2023.

Ao analisar os dados reais de saída para NT, observa-se que há uma diferença significativa entre os valores apresentados para o Layout 1, que variam entre 25,42 mg NT/L e 67,75 mg NT/L, e Layout 2, com valores entre 5,45 mg NT/L e 28,88 mg NT/L. Uma vez que os dados de entrada são os mesmos em ambas as simulações realizadas pelo software WEST, a redução do teor de nitrogênio total na saída pode ser justificada pela inserção da segunda linha de operação no sistema.

Para o cenário 1, observa-se uma elevada semelhança entre as métricas obtidas para as previsões realizadas pelos três modelos empregados, que também é observada na análise visual. Todos os modelos apresentam MAPE de 0,11 (ou 11 %). A análise gráfica também permitiu verificar a presença de um sutil *overshoot* para a previsão realizada pelo modelo RF, isso é, a previsão do modelo RF ultrapassa o resultado real em proporção maior que a apresentada pelos demais modelos. Essa verificação é validade pelo maior valor de RMSE para esse modelo, uma vez que essa métrica de performance penaliza muito mais os valores mais distantes do que as outra métricas. Os resultados das previsões para os modelos SVM e MLP foram semelhantes entre si para esse primeiro cenário, podendo ambos serem aplicados como modelos preditivos para esse sistema. Embora o modelo MLP tenha apresentado métricas de performance ligeiramente melhores, o modelo de MLP apresenta uma menor interpretabilidade (Phillips et al., 2021) e esta falta de entendimento explícito sobre o que ocorre no interior de um modelo caixa preta (Loyola-Gonzalez, 2019) pode desmotivar a sua utilização pelos usuários mais críticos dessas metodologias.

**Figura 9 - Resultados dos modelos preditivos para os dados de NT da ETE EDH WEST usando RF, MLP e SVM.**




Fonte: Os autores, 2023.

Ao analisar os resultados obtidos para o cenário 2, fica evidente a perda de qualidade dos modelos preditivos. Esse comportamento é resultado da base de dados usada para o treinamento, que possui valores de NT de saída mais elevados nas 745 primeiras horas de registro. Mesmo utilizando 683 dados de treinamento com os valores de NT de saída menores na base de dados, os modelos foram induzidos a prever valores mais elevados do que aqueles observados para os dados nas horas finais de treinamento. É possível constatar que, para a base de dados utilizada, os modelos preditivos não foram capazes de identificar sozinhos a mudança no comportamento de saída do parâmetro de qualidade NT, ocasionada pela mudança de operação.

Como todos os modelos previram valores maiores de NT de saída do que os valores simulados para o Layout 2, as métricas de performance apresentaram valores muito maiores do que aqueles obtidos para o cenário 1, indicando uma baixa precisão dos modelos. O modelo de RF foi o que apresentou melhor desempenho para

esse cenário. Por ser um algoritmo do tipo *ensemble*, há a criação de diversos avaliadores - regressores, no presente trabalho - que fazem uma estimativa geral com base em todas as previsões feitas por cada um dos avaliadores que compõe o método *ensemble*. A maioria das árvores de decisão foi criada de modo a prever com qualidade os dados da primeira parte do treinamento. Porém, algumas árvores aproximaram as previsões para se adequarem aos 683 últimos valores da base de dados, com valores mais próximos aos valores do teste. Assim, a previsão final, que é uma média das previsões de todas as árvores de decisão, teve seu valor reduzido em comparação aos demais modelos. Assim, pode-se constatar que o modelo de RF se adequa melhor ao cenário 2 dentre os modelos preditivos estudados, o que é nítido através das métricas e da observação do gráfico de previsões comparadas ao valor real (Figura 9, cenário 2).

Alterações na operação de ETEs visando a melhoria do processo são comuns, o que implica em uma dificuldade adicional considerável para a aplicação de metodologias de ML para a previsão de resultados industriais e implementação de processos baseados em inteligência artificial. Com intuito de mitigar os impactos negativos na previsão dos modelos decorrente de uma alteração de processo, o Cenário 3 foi criado e estudado. Neste cenário, a informação sobre qual layout da ETEs encontrava-se em operação foi adicionada à base de dados usada no modelo a partir de uma nova variável artificial. Observou-se uma melhora considerável das métricas de performance em comparação aos dois cenários anteriores. A exceção está no MAPE, que apresentou valores de 0,14 e 0,16 (ou 14% e 16%), ligeiramente maiores que os valores obtidos para o cenário 1, que foi de 11% para todos os modelos. Esse fato é justificado pois os valores de NT previstos são menores neste cenário ($NT_{médio}$ = 15,00 mg/L) quando comparados aos valores de NT previstos no cenário 1 ($NT_{médio}$ = 44,35 mg/L) e, por esse motivo, são mais sensíveis a erros percentuais. Os três modelos preditivos seguiram satisfatoriamente bem a tendência das ocorrências, como pode ser constatado na Figura 9 - Cenário 3, e apresentam valores de $R^2$ acima de 0,6 para previsão dos dados. Dentre eles, destaca-se o resultado obtido para o modelo MLP, que apresentou as melhores métricas de ajuste. A escolha adequada dos hiperparâmetros pode ter possibilitado uma melhor aproximação da relação entre os dados de entrada e NT de saída, sendo essa uma aplicação que dialoga com o teorema da aproximação universal (Csáji, 2001; Hornik et al., 1989).

Para a ETE Ambev, conforme apresentado anteriormente, todos os dados utilizados na base de dados partiram das anotações realizadas pelos operadores da ETE no sistema de gestão de processos (LiveMES) e das planilhas de controle interno de operação. Por tratar-se de uma ETE real, diversos problemas na obtenção, armazenamento e compartilhamento dos dados podem ocorrer e, por esse motivo, foi necessário realizar um rigoroso pré-processamento dos dados. A base de dados apresentava valores *outliers* de processo e valores sem significado físico para determinados parâmetros. Em situações reais de registro de dados, em especial quando a anotação é feita manualmente por diferentes operadores, não é surpreendente a presença de inconsistências como essas, que muitas vezes são decorrentes de erros de medição e de anotação humana. A exclusão de dados, realizada por meio da análise univariada descrita na seção 3, correspondeu a 0,76% da base de dados inicial, equivalente a 69 dos 9120 dados registrados na base inicial. Para todas as informações faltantes, realizou-se a imputação pela técnica KNN.

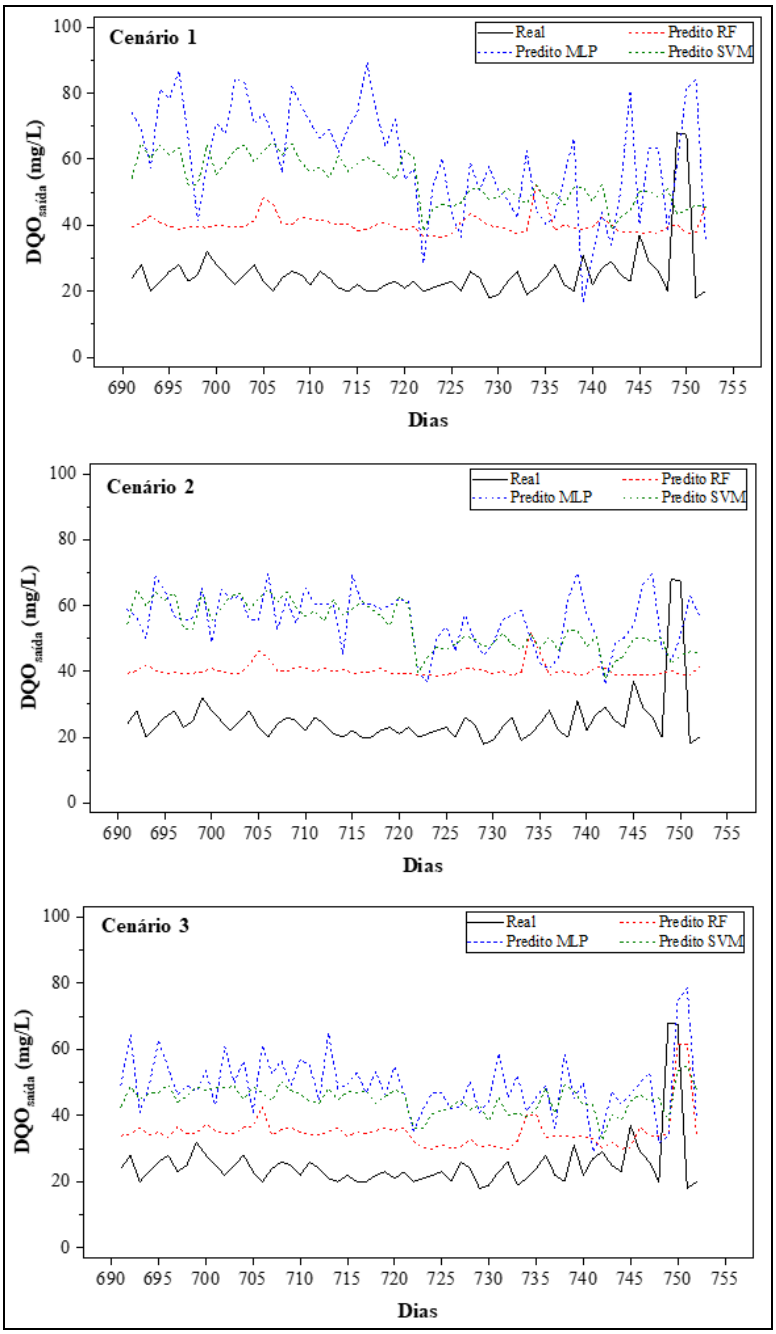
A base de dados completa foi, então, separada em variáveis independentes (dados de entrada de Q, DQO, CO, T e pH) e variável dependente (DQO de saída). Dos 752 registros diários para cada variável, os primeiros 690 dias foram definidos como conjunto de dados de treinamento dos modelos para previsão de DQO na saída da ETE e os últimos 62 dias foram definidos como conjunto de dados de teste dos modelos, conforme apresentado na Figura 5. Três diferentes cenários foram aplicados para os modelos preditivos para o parâmetro DQO:

- O cenário 1 considera como variáveis independentes os dados de entrada da ETE de Q, DQO, CO, T e pH para a previsão de DQO na saída da ETE;
- No cenário 2, as informações referentes à sazonalidade e ao modo de operação da ETE, definida com base na operação dos Reatores 1 e 2, foram adicionadas à base de dados utilizada pelo modelo preditivo como variáveis independentes adicionais;
- Para o cenário 3, os dados de Q, DQO, CO, T e pH de saída da ETE do dia anterior foram adicionados como variáveis independentes para a predição da DQO de saída. Uma vez que o tempo de residência da ETE é de até 7 dias, é provável que o efluente de saída de um determinado dia *i* tenha características muito similares ao efluente de saída do dia *i-1*. A inclusão das informações sobre a qualidade do efluente do dia anterior pode, portanto, auxiliar na performance dos modelos.

Os dados de DQO de saída previstos foram comparados com os valores reais de DQO de saída do intervalo de teste e a acurácia de cada modelo foi avaliada a partir das métricas de performance (RMSE, MAE, MAPE e

R²) apresentadas na Tabela 2. Uma análise gráfica comparativa entre os dados previstos pelos modelos e os dados obtidos pela simulação da ETE Ambev também é apresentada na Figura 10.

**Figura 10: Resultados dos modelos preditivos para os dados de DQO da ETE Ambev usando RF, MLP e SVM.**




Fonte: Os autores, 2023.

**Tabela 2 - Comparação de métricas de performance para os modelos preditivos de ML para DQO aplicados a ETE Ambev.**

| | Cenário 1 | | | Cenário 2 | | | Cenário 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | SVM | MLP | RF | SVM | MLP | RF | SVM | MLP | RF |
| RMSE | 30,86 | 39,47 | 17,77 | 30,95 | 33,00 | 17,66 | 21,74 | 26,83 | 13,09 |
| MAE | 29,94 | 35,84 | 16,97 | 30,03 | 31,80 | 16,98 | 21,10 | 25,07 | 11,37 |
| MAPE | 1,28 | 1,56 | 0,73 | 1,29 | 1,37 | 0,73 | 0,91 | 1,08 | 0,50 |
| R² | -11,97 | -20,21 | -3,30 | -12,04 | -13,83 | -3,24 | -5,44 | -8,80 | -1,33 |

No cenário 1, é possível observar que os valores de DQO de saída previstos pelos três modelos preditivos ficaram consideravelmente distantes dos valores reais. Entre os motivos para tal resultado, destaca-se a utilização dos poucos parâmetros de processo coletados diariamente como previsores dos modelos de ML. Parâmetros como sólidos suspensos totais, nitrogênio total e fósforo total são relatados como importantes para a previsão de DQO (Arismendy et al., 2021; Fernandez de Canete et al., 2016; Sharafati et al., 2020; Szeląg et al., 2020; Xu et al., 2022) e, por não serem computados diariamente pela ETE Ambev, não puderam ser utilizados neste estudo. A ausência dessas informações é um aspecto de destaque. Ao dispor de um conjunto abrangente de dados registrados para as variáveis de processo da ETE que apresentam uma forte correlação com a variável objetivo (nesse caso, a DQO), é provável que uma análise mais aprofundada e resultados mais precisos para os modelos preditivos possam ser alcançados. Com uma quantidade significativa de parâmetros à disposição, torna-se factível conduzir um estudo completo sobre a relação e importância de cada variável para a previsão de parâmetros de qualidade em ETEs. Isso permite uma análise abrangente e minuciosa sobre o impacto de cada variável no processo de previsão e no processo de tratamento do efluente, o que pode ser útil para a tomada de decisão sobre a operação da ETE.

Para o cenário 2, ao qual três novas variáveis indicadoras dos diferentes layouts de operação (Figura 4) foram adicionadas aos dados de treinamento, foi possível verificar uma sutil melhoria na performance dos modelos, tal como observado na ETE WEST. Os valores previstos para a DQO de saída estão mais próximos dos valores reais em comparação ao primeiro cenário, porém todos os modelos previram valores acima dos valores reais. Isso pode ser resultado de outras alterações na operação da ETE não relacionadas ao funcionamento dos reatores anaeróbios, mas relacionadas a outras melhorias na operação que não foram consideradas para o presente trabalho, como instalação de novos sensores ou mudanças em lotes de produtos químicos.

Visando mitigar a falta de outros parâmetros previsores e de informações relacionadas às alterações de operação, ao cenário 3 foram incluídos como previsores da DQO de saída os demais valores do efluente tratado da ETE observados no dia anterior. Foi possível observar que há uma melhora na performance dos modelos em comparação aos demais cenários, uma vez que um maior número de informações sobre os parâmetros de qualidade foi dado aos algoritmos de ML. Todavia, ainda há um grande potencial de estudo visando a melhoria nas previsões desses modelos.

Em todos os cenários aplicados a ETE Ambev, o modelo de RF foi o que apresentou as melhores métricas de performance e que conseguiu prever valores mais próximos aos valores reais. A explicação para esta ocorrência é idêntica à fornecida nos resultados das previsões na base simulada. Apesar dos erros consideráveis observados para as previsões feita pelo modelo RF, é possível verificar que as tendências de comportamento da variável são representadas pelo modelo preditivo RF e a tendência apresentada no maior pico foi prevista com um leve atraso. Devido à grande discrepância entre as previsões e observações reais, os modelos preditivos não poderiam ser implementados para auxiliar na tomada de decisões ou em outros processos envolvendo tecnologias de inteligência artificial para esse caso. Contudo, há espaço para melhoria dos modelos preditivos a partir da inclusão de mais informações a respeito da operação e do processo.

## CONCLUSÃO

Os modelos preditivos SVM, MLP e RF foram capazes de prever os parâmetros de qualidade de efluentes a partir de dados de entrada e saída obtidos em uma ETE simulada utilizando o software WEST da DHI e da ETE da Ambev, de Lages/SC. Para a estação simulada, os melhores resultados foram obtidos no cenário 3, que possuía mais informações disponíveis para o aprendizado de máquinas a partir do uso de *feature engineering*. O modelo de melhor desempenho foi o MLP, cuja fundamentação é sólida e pode ter seu resultado explicado pela teoria da aproximação universal.

Para a ETE real da Ambev, os resultados obtidos para os modelos preditivos apresentaram melhoria visível decorrente da aplicação da *feature engineering* proposta. O resultado mais favorável foi alcançado no cenário que contempla o maior número de variáveis preditivas disponíveis para o aprendizado da máquina (cenário 3), o que reitera a importância de fornecer um conjunto abrangente de informações de qualidade aos modelos de aprendizado de máquina. Isso inclui dados relacionados tanto ao modo de operação quanto às variáveis de processo que estão associadas ao parâmetro de saída que se deseja prever. Tal constatação destaca a relevância de incluir um amplo espectro de informações relevantes nos modelos de aprendizado de máquina a fim de obter resultados superiores. Apesar dos erros consideráveis de previsão obtidos para essa ETE, o modelo RF apresentou as melhores métricas de performance e as tendências de comportamento das variáveis de saída foram representadas pelo modelo preditivo.

Com relação às constantes alterações verificadas em operações de ETEs, ainda que a aplicação da técnica de *feature engineering* apresentada nesse estudo melhore os resultados dos modelos preditivos, a depender do impacto e da frequência de alterações de processo, os resultados das previsões podem não ter a qualidade necessária para a sua utilização com confiabilidade. Além disso, destaca-se a necessidade de haver a medição e anotação do maior número de parâmetros de qualidade do efluente, e com maior frequência, para que haja uma melhoria na precisão dos modelos.

Por fim, pode-se concluir que, para um conjunto de dados coletados em uma ETE, em condições controladas e sem mudanças abruptas em seu estado de operação, os modelos de ML são capazes de predizer os valores e tendências de variáveis da saída, sendo alimentados apenas com informações referentes à entrada. Ao ocorrer uma grande alteração no funcionamento da ETE, os modelos preditivos têm sua precisão severamente prejudicada. A inserção de novas variáveis por meio da técnica de *feature engineering*, informando aos modelos os novos cenários, permite que os modelos se adaptem rapidamente à nova operação, precisando de um número menor de registros referentes ao novo processo, podendo ainda aproveitar os dados antigos que ainda trazem informação relevante sobre a relação entre as variáveis de entrada e a variável objetivo.

## AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Pesquisa e Desenvolvimento – CNPq e ao Ministério da Ciência, Tecnologia e Inovações pelo apoio financeiro e pelas bolsas de fomento tecnológico (Processos nº 350509/2022-0, 350508/2022-4, 424532/2021-2), à Hydroinfo, DHI e LABMAC/UFSC pelas contribuições técnicas, à AMBEV Unidade Lages pela disponibilização dos dados de monitoramento e operação da ETE e aos demais colaboradores do projeto que contribuíram de alguma forma nesse estudo.

## REFERÊNCIAS BIBLIOGRÁFICAS

1. AL AANI, S. et al. *Can machine language and artificial intelligence revolutionize process automation for water treatment and desalination? Desalination*, v. 458, p. 84–96, 2019.

2. ALAVI, J. et al. *A new insight for real-time wastewater quality prediction using hybridized kernel-based extreme learning machines with advanced optimization algorithms. Environmental Science and Pollution Research*, v. 29, n. 14, p. 20496–20516, 2022.

3. ALMOMANI, F. *Prediction the performance of multistage moving bed biological process using artificial neural network (ANN). Science of The Total Environment*, v. 744, p. 140854, 2020.

4. ARISMENDY, L. et al. *A prescriptive intelligent system for an industrial wastewater treatment process: Analyzing pH as a first approach. Sustainability)*, v. 13, n. 8, p. 4311, 2021.

5. BAGHERI, M. et al. *Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. Process Safety and Environmental Protection*, v. 95, p. 12–25, 2015.

6. BAKER, A. *Simplicity.* In: ZALTA, E. N. (Ed.). *The Stanford Encyclopedia of Philosophy.* Summer 2022. Stanford, CA: Metaphysics Research Lab, Stanford University, 2022.

7. BARNAT-HUNEK, D. et al. *An integrated texture analysis and machine learning approach for durability assessment of lightweight cement composites with hydrophobic coatings modified by nanocellulose. Measurement,* v. 179, p. 109538, 2021.

8. BERRAR, D. *Cross-Validation.* In: RANGANATHAN, S. et al. (Eds.). *Encyclopedia of Bioinformatics and Computational Biology.* Oxford: Elsevier, p. 542–545, 2019.

9. BERTRAND, F. Sweetviz 2.1.4. Disponível em: https://pypi.org/project/sweetviz/. Acesso em: 22/05/2023.

10. BHADESHIA, H. K. D. H. et al. *Performance of neural networks in materials science. Materials Science and Technology*, v. 25, n. 4, p. 504–510, 2009.

11. BOTTOU, L. *Large-Scale Machine Learning with Stochastic Gradient Descent.* In: LECHEVALLIER, Y.; SAPORTA, G. *Proceedings of COMPSTAT'2010.* Heidelberg: Physica-Verlag HD, p. 177–186, 2010.

12. BREIMAN, L. *Random Forests. Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

13. BUITINCK, L. et al. *API design for machine learning software: experiences from the scikit-learn project. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases.* Praga, República Tcheca. 2013.

14. CAI, J. et al. *Machine learning-driven new material discovery. Nanoscale Advances*, v. 2, n. 8, p. 3115–3130, 2020.

15. CSÁJI, B. C. Approximation with Artificial Neural Networks. Hungria. Faculty of Sciences, Eötvös Loránd University, 2001.

16. DAIGGER, G. T.; ROPER, R. E. *The Relationship between SVI and Activated Sludge Settling Characteristics. Journal of the Water Pollution Control Federation*, v. 57, n. 8, p. 1–23, 1985.

17. DHI. WEST Models Guide. Disponível em: https://manuals.mikepoweredbydhi.help/latest/Cities/TornadoModels/index.htm. Acesso em: 22/05/2023.

18. DOGAN, E. et al. *Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand. Environmental Progress*, v. 27, n. 4, p. 439–446, 2008.

19. EBRAHIMPOUR, A. et al. *A modeling study by response surface methodology and artificial neural network on culture parameters optimization for thermostable lipase production from a newly isolated thermophilic Geobacillus sp. strain ARM*. BMC Biotechnology, v. 8, n. 1, p. 96, 2008.

20. EERIKÄINEN, S. et al. *Data analytics in control and operation of municipal wastewater treatment plants: qualitative analysis of needs and barriers. Water Science and Technology*, v. 82, n. 12, p. 2681–2690, 2020.

21. FAN, M. et al. *A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. Chemosphere*, v. 200, p. 330–343, 2018.

22. FAROOQ, F. et al. *A Comparative Study for the Prediction of the Compressive Strength of Self-Compacting Concrete Modified with Fly Ash. Materials*, v. 14, n. 17, p. 4934, 2021.

23. FERNANDEZ DE CANETE, J. et al. *Soft-sensing estimation of plant effluent concentrations in a biological wastewater treatment plant using an optimal neural network. Expert Systems with Applications*, v. 63, p. 8–19, 2016.

24. FLORES, V.; LEIVA, C. *Comparative Study on Supervised Machine Learning Algorithms for Copper Recovery Quality Prediction in a Leaching Process. Sensors*, v. 21, n. 6, p. 2119, 2021.

25. FUKUSHIMA, K. *Cognitron: A self-organizing multilayered neural network. Biological Cybernetics*, v. 20, n. 3, p. 121–136, 1975.

26. HAN, H.-G.; WANG, L.-D.; QIAO, J.-F. *Hierarchical extreme learning machine for feedforward neural network. Neurocomputing*, v. 128, p. 128–135, 2014.

27. HAO, X. et al. *Machine Learning Models for Predicting Adverse Pregnancy Outcomes in Pregnant Women with Systemic Lupus Erythematosus. Diagnostics*, v. 13, n. 4, p. 612, 2023.

28. HARRAG, F.; GUELIANI, S. *Event Extraction Based on Deep Learning in Food Hazard Arabic Texts. International Journal of Advanced Computer Science and Applications*, v. 11, n. 8, 2020.

29. HORNIK, K.; STINCHCOMBE, M.; WHITE, H. *Multilayer feedforward networks are universal approximators. Neural Networks*, v. 2, n. 5, p. 359–366, 1989.

30. KINGMA, D. P.; BA, J. *Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.* 2015.

31. KOHAVI, R.; WOLPERT, D. *Bias plus Variance Decomposition for Zero-One Loss Functions. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996.

32. KOTSIANTIS, S. B.; KANELLOPOULOS, D.; PINTELAS, P. E. Data preprocessing for supervised leaning. International Journal of Computer and Information Engineering, v. 1, n. 12, p. 1–7, 2007.

33. LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. Disponível em: http://arxiv.org/abs/1912.06059. Acesso em: 22/05/2023.

34. LIU, L. et al. *Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. Computer Methods and Programs in Biomedicine*, v. 196, p. 105624, 2020.

35. LORENCIN, I. et al. *Using multi-layer perceptron with Laplacian edge detector for bladder cancer diagnosis. Artificial Intelligence in Medicine*, v. 102, p. 101746, 2020.

36. LOYOLA-GONZALEZ, O. *Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access*, v. 7, p. 154096–154113, 2019.

37. LY, Q. V. et al. *Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. Science of The Total Environment*, v. 832, p. 154930, 2022.

38. NEAL, B. On the Bias-Variance Tradeoff: Textbooks Need an Update. Disponível em: https://arxiv.org/abs/1912.08286. Acesso em: 22/05/2023.

39. NEWHART, K. B. et al. *Data-driven performance analyses of wastewater treatment plants: A review. Water Research*, v. 157, p. 498–513, 2019.

40. PAL, S. K.; MITRA, S. *Multilayer perceptron, fuzzy sets, and classification. IEEE Transactions on Neural Networks*, v. 3, n. 5, p. 683–697, 1992.

41. PEDREGOSA, F. et al. *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

42. PHILLIPS, P. J. et al. Four Principles of Explainable Artificial Intelligence. Disponível em: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf. Acesso em: 22/05/2023.

43. PIRI, J. et al. *Prediction of the solar radiation on the Earth using support vector regression technique. Infrared Physics & Technology*, v. 68, p. 179–185, 2015.

44. REBACK, J. et al. pandas-dev/pandas: Pandas 1.2.3. Disponível em: https://zenodo.org/record/4572994. Acesso em: 22/05/2023.

45. RODRÍGUEZ-MARTÍN, M. et al. Predictive Models for the Characterization of Internal Defects in Additive Materials from Active Thermography Sequences Supported by Machine Learning Methods. Sensors, v. 20, n. 14, p. 3982, 2020.

46. SCHONLAU, M.; ZOU, R. Y. *The random forest algorithm for statistical learning. The Stata Journal*, v. 20, n. 1, p. 3–29, 2020.

47. SHARAFATI, A.; ASADOLLAH, S. B. H. S.; HOSSEINZADEH, M. *The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. Process Safety and Environmental Protection*, v. 140, p. 68–78, 2020.

48. SINGH, N. K. et al. *Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems. Bioresource Technology*, v. 369, p. 128486, 2023.

49. SZELĄG, B. et al. *Soft Sensor Application in Identification of the Activated Sludge Bulking Considering the Technological and Economical Aspects of Smart Systems Functioning. Sensors*, v. 20, n. 7, p. 1941, 2020.

50. TIYASHA; TUNG, T. M.; YASEEN, Z. M. *A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology*, v. 585, p. 124670, 2020.

51. TROYANSKAYA, O. et al. *Missing value estimation methods for DNA microarrays. Bioinformatics*, v. 17, n. 6, p. 520–525, 2001.

52. TUKEY, J. W. Exploratory Data Analysis. Pearson Education Inc., 1977.

53. VAN ROSSUM, G.; DE BOER, J. *Interactively testing remote servers using the Python programming language. CWI Quarterly*, v. 4, n. 4, p. 283–304, 1991.

54. WANG, D. et al. *A machine learning framework to improve effluent quality control in wastewater treatment plants. Science of The Total Environment*, v. 784, p. 147138, 2021.

55. WU, M.-C.; LIN, G.-F. *An Hourly Streamflow Forecasting Model Coupled with an Enforced Learning Strategy. Water*, v. 7, n. 11, p. 5876–5895, 2015.

56. XU, Y. et al. *Data-driven prediction of neutralizer pH and valve position towards precise control of chemical dosage in a wastewater treatment plant. Journal of Cleaner Production*, v. 348, p. 131360, 2022.

57. YANG, Y. et al. *Prediction of effluent quality in a wastewater treatment plant by dynamic neural network modeling. Process Safety and Environmental Protection*, v. 158, p. 515–524, 2022.

58. YDATA LABS INC. pandas-profiling 3.6.6. Disponível em: <https://pypi.org/project/pandas-profiling/#files>. Acesso em: 22/05/2023.

59. YU, J. et al. *Structural features modeling of substituted hydroxyapatite nanopowders as bone fillers via machine learning. Ceramics International*, v. 47, n. 7, Part A, p. 9034–9047, 2021.

60. ZHAO, L. et al. *Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. Process Safety and Environmental Protection*, v. 133, p. 169–182, 2020.

61. ZHU, C. et al. *Algorithm 778: L-BFGS-B. ACM Transactions on Mathematical Software*, v. 23, n. 4, p. 550–560, 1997.

62. ZHU, J.; JIANG, Z.; FENG, L. *Improved neural network with least square support vector machine for wastewater treatment process. Chemosphere*, v. 308, p. 136116, 2022.