



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

JOÃO VÍTOR RIOS FUCK

**DESENVOLVIMENTO DE MÉTODO PARA OTIMIZAÇÃO DA OPERAÇÃO
DE ESTAÇÕES DE TRATAMENTO DE EFLUENTE COM PREDIÇÃO EM TEMPO
REAL POR MODELOS DE APRENDIZAGEM DE MÁQUINA**

Florianópolis

2026

João Vítor Rios Fuck

**Desenvolvimento de Método para Otimização da Operação de Estações de
Tratamento de Efluente com Predição em Tempo Real por Modelos de
Aprendizagem de Máquina**

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Mestre em Engenharia Química.

Orientador: Prof. Natan Padoin, Dr.
Coorientadores: Prof.^a Cíntia Soares, Dr.^a
Nicolas Spogis, Dr.

Florianópolis

2026

Fuck, João Vítor Rios
DESENVOLVIMENTO DE MÉTODO PARA OTIMIZAÇÃO DA OPERAÇÃO
DE ESTAÇÕES DE TRATAMENTO DE EFLUENTE COM PREDIÇÃO EM
TEMPO REAL POR MODELOS DE APRENDIZAGEM DE MÁQUINA / João
Vítor Rios Fuck ; orientador, Natan Padoin, coorientadora,
Cíntia Soares, coorientador, Nicolas Spogis, 2026.
138 p.

Dissertação (mestrado) - Universidade Federal de Santa
Catarina, Centro Tecnológico, Programa de Pós-Graduação em
Engenharia Química, Florianópolis, 2026.

Inclui referências.

1. Engenharia Química. 2. Aprendizado de Máquina. 3.
Tratamento de Efluentes. 4. Modelagem Preditiva. I.
Padoin, Natan. II. Soares, Cíntia. III. Spogis, Nicolas
IV. Universidade Federal de Santa Catarina. Programa de
Pós-Graduação em Engenharia Química. V. Título.

João Vítor Rios Fuck

Desenvolvimento de Método para Otimização da Operação de Estações de Tratamento de Efluente com Predição em Tempo Real por Modelos de Aprendizagem de Máquina

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 5 de dezembro de 2025, pela banca examinadora composta pelos seguintes membros:

Prof. Maurício Bezerra de Souza Júnior, Dr.
Universidade Federal do Rio de Janeiro (UFRJ)

Prof.^a Vivien Rossbach, Dr.^a
Universidade Federal de Santa Catarina (UFSC)

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Engenharia Química.



Documento assinado digitalmente
Agenor Furigo Junior
Data: 20/02/2026 11:51:16-0300
CPF: ***.789.438-**
Verifique as assinaturas em <https://v.ufsc.br>

Coordenação do Programa de Pós-Graduação



Documento assinado digitalmente
Natan Padoin
Data: 19/02/2026 15:45:12-0300
CPF: ***.583.829-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Natan Padoin, Dr.
Orientador

Florianópolis, 2026

Dedico este trabalho aos meus pais e irmão que sempre me incentivaram nos estudos.

AGRADECIMENTOS

À Dr.^a Maria Alice Prado Cechinel, pela dedicação incansável e pela parceria leal em cada etapa deste percurso, cujo empenho foi decisivo para a consolidação do projeto que ora se materializa nesta dissertação.

Ao Rodrigo Campos de Andrade, pela idealização, condução e apoio contínuos ao longo do desenvolvimento do projeto.

A Juliana Neves e a Felipe Bublitz Bier, pelo suporte técnico prestado durante a execução deste trabalho.

Aos Profs. Natan Padoin, Cíntia Soares e Nicolas Spogis, pela orientação no mestrado e pelo imprescindível suporte para a conclusão desta etapa, mesmo diante das dificuldades e contratemplos enfrentados.

Ao Ricardo e à equipe da AMBEV – Lages, pela receptividade exemplar, pela valiosa troca de ideias e pela disponibilização de dados que fundamentaram parte substancial dos resultados aqui apresentados.

E ao CNPq, pela concessão da bolsa que viabilizou a realização desta pesquisa.

RESUMO

O tratamento de efluentes envolve processos complexos, altamente não lineares e influenciados por variáveis físico-químicas, biológicas e operacionais, cuja interação dificulta decisões baseadas apenas na experiência. A crescente demanda por eficiência, redução de custos e atendimento a padrões ambientais mais rigorosos tem impulsionado o uso de métodos de Inteligência Artificial (IA) aplicados à modelagem e otimização de Estações de Tratamento de Efluentes (ETEs). Neste contexto, esta dissertação desenvolve e valida uma metodologia de aprendizado de máquina voltada à previsão de parâmetros de qualidade do efluente tratado, visando subsidiar a operação de ETEs e constituir a base de um futuro Sistema de Apoio à Decisão (SAD). A metodologia inclui: (i) diagnóstico da qualidade e consistência dos dados; (ii) pré-processamento rigoroso, com padronização, tratamento de *outliers*, unificação de rótulos e imputação; (iii) construção, treinamento e validação de modelos (RF, SVM, MLP e LSTM); e (iv) aplicação de técnicas de interpretabilidade (*Permutation Importance* e PDP) para elucidar relações e avaliar a robustez dos modelos. A abordagem é aplicada a três estudos de caso complementares: uma ETE industrial com dados reais (AMBEV – Lages/SC); uma ETE municipal simulada no *software* WEST (dados controlados); e uma ETE municipal (Umbilo) com dados reais e simulados, incorporando séries temporais de alta frequência. Os resultados mostram que, em ambientes com dados de boa qualidade (simulações WEST), os modelos preditivos alcançam elevada acurácia (R^2 até 0,99), especialmente MLP e LSTM, capazes de capturar defasagens temporais e padrões dinâmicos. Em contrapartida, no cenário industrial real (ETE AMBEV), a presença de dados inconsistentes, variabilidade na frequência de coleta, lacunas em parâmetros essenciais e ausência de monitoramento contínuo prejudicaram o desempenho dos modelos, revelando que a maturidade dos dados é o fator mais determinante para a viabilidade de aplicações de IA em ETEs. A engenharia de atributos mostrou-se crucial para representar o contexto operacional não explícito nos dados brutos, permitindo incorporar informações sobre mudanças estruturais, condições dos reatores, efeitos de *carryover* e eventos de manutenção. A inclusão dessas variáveis elevou o desempenho dos modelos, demonstrando que a representação do estado operacional é tão importante quanto os parâmetros físico-químicos tradicionais. As análises de interpretabilidade permitiram identificar as variáveis mais influentes em cada cenário, fornecendo aos operadores uma visão clara sobre os fatores que determinam a qualidade do efluente tratado. De forma geral, o estudo demonstra que: (i) a qualidade dos dados é o principal limitante da aplicação de ML em ETEs; (ii) modelos complexos não superam problemas de dados escassos ou inconsistentes; (iii) a engenharia de atributos é essencial para capturar o contexto operacional; e (iv) a interpretabilidade é indispensável para a adoção prática de soluções de IA em ambientes industriais. Conclui-se que, embora os algoritmos avaliados tenham mostrado alto potencial, a implementação plena de soluções baseadas em IA depende de melhorias estruturais nos sistemas de coleta e gerenciamento de dados. A metodologia proposta representa um avanço rumo ao desenvolvimento de um SAD para otimização e controle inteligente de ETEs, com potencial de aplicação em diferentes escalas e configurações de tratamento.

Palavras-chave: Aprendizado de Máquina; Tratamento de Efluentes; Modelagem Preditiva.

ABSTRACT

Wastewater treatment involves complex and highly nonlinear processes influenced by physicochemical, biological, and operational variables, whose interactions make decision-making difficult when based solely on operator experience. The growing demand for efficiency, cost reduction, and compliance with increasingly stringent environmental standards has driven the use of Artificial Intelligence (AI) methods applied to the modeling and optimization of Wastewater Treatment Plants (WWTPs). In this context, this dissertation develops and validates a machine learning methodology aimed at predicting treated effluent quality parameters, with the goal of supporting WWTP operation and serving as the foundation for a future Decision Support System (DSS). The methodology includes: (i) diagnosing data quality and consistency; (ii) rigorous preprocessing, including standardization, outlier treatment, label harmonization, and data imputation; (iii) construction, training, and validation of predictive models (RF, SVM, MLP, and LSTM); and (iv) application of interpretability techniques (Permutation Importance and PDP) to elucidate relationships and assess model robustness. The approach is applied to three complementary case studies: an industrial WWTP with real data (AMBEV – Lages/SC); a municipal WWTP simulated in the WEST software (controlled datasets); and a municipal WWTP (Umbilo) with real and simulated data, including high-frequency time series. The results show that in environments with high-quality data (WEST simulations), predictive models achieve high accuracy (R^2 up to 0.99), particularly MLP and LSTM networks, which effectively capture temporal lags and dynamic input patterns. In contrast, in the real industrial setting (AMBEV WWTP), inconsistent data, variability in sampling frequency, missing essential parameters, and the absence of continuous monitoring significantly impaired model performance, revealing that data maturity is the most decisive factor for the feasibility of AI applications in WWTPs. Feature engineering proved crucial for representing the operational context not explicitly present in raw datasets, enabling the incorporation of information on structural changes, reactor operating conditions, carryover effects, and maintenance events. Including these additional variables significantly improved model performance, demonstrating that representing the operational state is as important as traditional physicochemical parameters. Interpretability analyses allowed the identification of the most influential variables in each scenario, providing operators with a clear understanding of the factors that determine treated effluent quality. Overall, the study demonstrates that: (i) data quality is the main limiting factor for ML applications in WWTPs; (ii) complex models do not overcome issues related to scarce or inconsistent data; (iii) feature engineering is essential for capturing operational context; and (iv) interpretability is indispensable for the practical adoption of AI solutions in industrial environments. It is concluded that, although the evaluated algorithms show strong potential, full implementation of AI-based solutions depends on structural improvements in data collection and management systems. The proposed methodology represents a significant step toward the development of a DSS for intelligent optimization and control of WWTPs, with potential applicability at different scales and treatment configurations.

Keywords: Machine Learning; Effluent Treatment; Predictive Modeling.

LISTA DE FIGURAS

Figura 1 – Fluxograma simplificado da ETE da AMBEV – Unidade Lages/SC	30
Figura 2 – Esquema representativo dos reatores UASB (a) e IC (b).....	33
Figura 3 – <i>Layouts</i> da ETE utilizada para obtenção dos dados no <i>software</i> WEST..	44
Figura 4 – Perfil de saída do parâmetro NT para os 31 dias de operação do <i>Layout</i> 1 (a) e os 62 dias de operação do <i>Layout</i> 1 e <i>Layout</i> 2 (b) da ETE WEST	47
Figura 5 – Fluxograma simplificado da operação da ETE AMBEV (unidade Lages/SC) em função da operação dos tanques de aeração e reatores anaeróbios	49
Figura 6 – Perfil de saída do parâmetro DQO para a ETE AMBEV	50
Figura 7 – Interpretação gráfica do método SVM.....	53
Figura 8 – Exemplo genérico de arquitetura de um modelo MLP.....	54
Figura 9 – Exemplo genérico de arquitetura de um modelo de Floresta Aleatória	57
Figura 10 – Resultados dos modelos preditivos para os dados de NT _{saída} da ETE WEST usando RF, MLP e SVM: Cenários (a) 1, (b) 2 e (c) 3	63
Figura 11 – Resultados dos modelos preditivos para os dados de DQO _{saída} da ETE AMBEV usando RF, MLP e SVM: Cenários (a) 1, (b) 2 e (c) 3	68
Figura 12 – Matriz de correlação de Pearson para as variáveis de entrada da ETE WEST	71
Figura 13 – Importância das variáveis de entrada na previsão de NT _{saída} nos cenários operacionais 1 e 3 da ETE WEST	72
Figura 14 – Gráficos de dependência parcial mostrando a média dos efeitos do influente DQO, DBO e SST (canto superior esquerdo), o efeito da vazão do influente (canto superior direito), a média dos efeitos do influente NKT, NT e NH (canto inferior esquerdo) e indicação do modo de funcionamento da ETE (canto inferior direito) na previsão do NT _{saída} em escala padronizada	74
Figura 15 – <i>Layouts</i> da ETE Umbilo utilizados para obtenção de dados no <i>software</i> WEST: (a) <i>layout</i> original disponível no banco de dados do <i>software</i> e (b) <i>layout</i> modificado excluindo a entrada de sobrenadante	83
Figura 16 – Perfil de entrada e saída do parâmetro DQO para os 366 dias de operação da ETE Umbilo	86

Figura 17 – Matriz de correlação de Pearson para as variáveis de entrada da ETE Umbilo	91
Figura 18 – Resultados dos modelos preditivos para $DQO_{saída}$ utilizando os dados simulados pelo <i>software</i> WEST em base diária para todos os modelos testados (a) para o modelo com melhor resultado (MLP) (b)	95
Figura 19 – Importância das variáveis de entrada para os modelos preditivos de $DQO_{saída}$ utilizando o banco de dados simulado diariamente pelo <i>software</i> WEST (a) e gráficos de dependência parcial ilustrando os efeitos das variáveis de entrada Q (canto superior esquerdo), NH (canto superior direito), SST (canto inferior esquerdo) e NT (canto inferior direito) na predição de $DQO_{saída}$ em escala padronizada (b) para o modelo MLP ..	97
Figura 20 – Resultados dos modelos preditivos para $DQO_{saída}$ utilizando os dados simulados pelo <i>software</i> WEST em base de teste horária para todos os modelos testados (a) para os modelos com melhor resultados (C_LSTM e S_LSTM) (b).....	100
Figura 21 – Importância das variáveis de entrada para os modelos preditivos de $DQO_{saída}$ utilizando a base de dados horária obtida pelo <i>software</i> WEST (a) e gráficos de dependência parcial ilustrando os efeitos das variáveis de entrada Q, NH, SST e NT na predição de $DQO_{saída}$ em escala padronizada (b) para os modelos C_LSTM e S_LSTM.....	103
Figura 22 – Resultados dos modelos preditivos para $DQO_{saída}$ utilizando a base de dados reais da ETE Umbilo (a) e para o modelo com melhor resultado (SVM) (b).....	104
Figura 23 – Importância das variáveis de entrada para os modelos preditivos de $DQO_{saída}$ utilizando a base de dados real da ETE Umbilo (a) e gráficos de dependência parcial ilustrando os efeitos das variáveis de entrada Q, NH, SST e NT na predição de $DQO_{saída}$ em escala padronizada (b) para o modelo SVM.....	108

LISTA DE TABELAS

Tabela 1 – Intervalo de valores observados para características físico-químicas relevantes do efluente bruto produzido pela cervejaria no mês de abril de 2022.....	24
Tabela 2 – Comparação de métricas de performance para os modelos preditivos de ML para NT _{saída} aplicados a ETE WEST	61
Tabela 3 – Comparação de métricas de performance para os modelos preditivos de ML para DQO _{saída} aplicados a ETE AMBEV	67
Tabela 4 – Resumo estatístico dos parâmetros de entrada da ETE Umbilo durante os 366 dias de operação	92
Tabela 5 – Comparação de métricas de desempenho para modelos preditivos de ML para DQO _{saída} aplicados a dados simulados diariamente pelo <i>software</i> WEST	94
Tabela 6 – Comparação de métricas de desempenho para modelos preditivos de ML para DQO _{saída} aplicados a dados simulados pelo <i>software</i> WEST em base de teste horária.....	99
Tabela 7 – Comparação de métricas de desempenho para modelos preditivos de ML para DQO _{saída} aplicados a dados reais coletados na ETE Umbilo.....	104

LISTA DE ABREVIATURAS E SIGLAS

ANN	<i>Artificial Neural Networks</i> (Redes Neurais Artificiais)
ASM1	<i>Activated Sludge Model</i> nº 1 (Modelo de Lodos Ativados nº 1)
C_LSTM	Estrutura LSTM com Grid Search CV
CO	Carga Orgânica
DBO	Demanda Bioquímica de Oxigênio
DQO	Demanda Química de Oxigênio
EDA	Análise Exploratória de Dados
ETE	Estação de Tratamento de Efluentes
FI	<i>Feature Importance</i>
FO	Fosfato
FT	Fósforo Total
IA	Inteligência Artificial
IC	Reator de Circulação Interna
IVL	Índice Volumétrico do Lodo
KNN	<i>K-Nearest Neighbors</i>
LIME	<i>Local Interpretable Model-agnostic Explanations</i>
LSTM	<i>Long Short-Term Memory</i>
MAE	<i>Mean Absolute Error</i> (Erro Médio Absoluto)
MAPE	<i>Mean Absolute Percentage Error</i> (Erro Percentual Absoluto Médio)
MBBR	Reator Biológico de Leito Móvel
ML	<i>Machine Learning</i> (Aprendizado de Máquina)
MLP	<i>Multi-Layer Perceptron</i>
MPC	<i>Model Predictive Control</i> (Modelo de Controle Preditivo)
NH	Nitrogênio Amoniacal
NKT	Nitrogênio Kjeldahl Total
NO	Nitrato
NT	Nitrogênio Total
OD	Oxigênio Dissolvido
PDP	<i>Partial Dependence Plots</i>
PI	<i>Permutation Importance</i>
Q	Vazão
R ²	Coeficiente de determinação

RBF	Função de Base Radial
RF	<i>Random Forest</i>
RMSE	<i>Root Mean Square Error</i> (Raiz do Erro Quadrático Médio)
S_LSTM	Estrutura LSTM Simples
SAD	Sistema de Apoio à Decisão
SGD	<i>Stochastic Gradient Descent Algorithm</i>
SHAP	<i>Shapley Additive Explanations</i>
SST	Sólidos Suspensos Totais
SVM	<i>Support Vector Machine</i>
T	Temperatura
UASB	<i>Upflow Anaerobic Sludge Blanket</i> (reator anaeróbio de manta de lodo de fluxo ascendente)
WEST	Software de simulação de ETEs (MIKE <i>Powered by</i> DHI).
XAI	<i>Explainable AI</i>
XML	<i>Explainable Machine Learning</i>

SUMÁRIO

1.	INTRODUÇÃO.....	15
1.1	ESTAÇÕES DE TRATAMENTO DE EFLUENTES	17
1.2	O PROJETO	19
1.3	OBJETIVO	20
1.3.1	Objetivo geral	20
1.3.2	Objetivos específicos	20
1.4	ORGANIZAÇÃO DO TRABALHO	21
2.	BASE TEÓRICA E CONTEXTUALIZAÇÃO	23
2.1	EFLUENTE DE CERVEJARIA.....	23
2.2	MODELOS DE INTELIGÊNCIA ARTIFICIAL	25
2.2.1	Principais modelos de aprendizado de máquina aplicados a ETEs.....	25
2.2.2	Desafios recorrentes na implementação de modelos preditivos	26
2.2.3	A importância da interpretabilidade (XAI) em processos industriais ...	27
2.2.4	Lacuna do conhecimento e justificativa do trabalho.....	27
3.	ESTAÇÃO DE TRATAMENTO DE EFLUENTES DA AMBEV (LAGES/SC)	
		29
3.1	DESCRIÇÃO E CARACTERIZAÇÃO	29
3.2	ESTAÇÃO DE TRATAMENTO DE EFLUENTES INDUSTRIAIS - AMBEV	30
3.3	IDENTIFICAÇÃO DOS OBJETOS DE OTIMIZAÇÃO DA ETE AMBEV	36
3.4	DIAGNÓSTICO DOS DADOS DE CONTROLE E MONITORAMENTO DA ETE AMBEV	37
4.	DESENVOLVIMENTO DE MODELOS PREDITIVOS: VALIDAÇÃO EM AMBIENTE CONTROLADO E DESAFIOS EM CENÁRIO INDUSTRIAL...	41
4.1	JUSTIFICATIVA DA PROPOSTA DE ESTUDO	41
4.2	ESTUDO DE CASO E DESCRIÇÃO DO CONJUNTO DE DADOS	43
4.3	PRÉ-PROCESSAMENTO DOS DADOS	50
4.4	MODELOS DE <i>MACHINE LEARNING</i>	52
4.5	INTERPRETABILIDADE DE MODELOS - <i>FEATURE IMPORTANCE</i>	59
4.6	RESULTADOS OBTIDOS PARA A ESTAÇÃO DE TRATAMENTO DE ESGOTO SIMULADA (ETE WEST).....	60
4.7	RESULTADOS OBTIDOS PARA A ESTAÇÃO DE TRATAMENTO DE EFLUENTES INDUSTRIAIS (ETE AMBEV)	66

4.8	RESULTADOS OBTIDOS PARA A INTERPRETABILIDADE DE MODELOS DE ML - <i>FEATURE IMPORTANCE</i> (FI).....	71
4.9	CONCLUSÕES.....	76
5.	APROFUNDAMENTO DA MODELAGEM PREDITIVA: ANÁLISE COM DADOS DE ALTA FREQUÊNCIA E REDES NEURAIS RECORRENTES.	79
5.1	JUSTIFICATIVA DA PROPOSTA DE ESTUDO	79
5.2	DESCRIÇÃO DO ESTUDO DE CASO E COLETA DE DADOS	83
5.3	MODELOS DE <i>MACHINE LEARNING</i>	87
5.4	RESULTADOS OBTIDOS PELA ANÁLISE EXPLORATÓRIA DE DADOS ..	90
5.5	RESULTADOS OBTIDOS PARA A PREVISÃO DIÁRIA USANDO DADOS GERADOS PELA SIMULAÇÃO WEST	93
5.6	RESULTADOS OBTIDOS PARA A PREVISÃO HORÁRIA USANDO DADOS GERADOS PELA SIMULAÇÃO WEST	99
5.7	RESULTADOS OBTIDOS PARA A PREVISÃO USANDO DADOS DE MEDIÇÕES REAIS DA ETE UMBILO	104
5.8	CONCLUSÕES.....	109
6.	DISCUSSÃO GERAL	111
6.1	O IMPACTO DA REALIDADE DOS DADOS	111
6.2	ENGENHARIA DE ATRIBUTOS COMO CHAVE PARA A CAPTURA DO CONTEXTO OPERACIONAL	112
6.3	ESCOLHA DO ALGORITMO ADEQUADO À NATUREZA DOS DADOS ..	113
6.4	IMPLICAÇÕES PARA A OTIMIZAÇÃO E PARA O CONTROLE DE PROCESSOS	114
6.5	CONSIDERAÇÕES FINAIS	114
	REFERÊNCIAS	118

1. INTRODUÇÃO

O tratamento de efluentes é um processo complexo afetado por diversos fatores químicos, físicos e microbiológicos, além de fenômenos naturais e atividades antrópicas externas à estação de tratamento de efluentes (ETE). A complexidade desses fatores resulta em muitas incertezas e exige que os operadores realizem controles operacionais apropriados no sistema (LOOS *et al.*, 2013; ZHAO *et al.*, 2020).

Atender aos padrões de lançamento de efluentes exige elevados custos de operação das ETEs, associados ao consumo de produtos químicos e energia, bem como aos custos com pessoal. A pressão da sociedade por soluções ambientalmente sustentáveis e economicamente viáveis vem se refletindo em padrões de lançamento cada vez mais restritivos, exigindo a busca por operações mais eficientes. Neste sentido, soluções que permitam reduzir custos e, ao mesmo tempo, melhorar a qualidade do efluente tratado serão 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.

Os procedimentos de modelagem e otimização de processos em sistemas de tratamento de águas são importantes para aumentar sua eficiência sem aumentar os custos de operação e 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). Os avanços na velocidade dos computadores e a redução significativa de seus custos têm favorecido a aplicação de controladores preditivos de modelos para processos de interesse comercial (SELVARAJ *et al.*, 1995). Contudo, os métodos convencionais para modelagem e otimização de processos de remoção de poluentes requerem uma ampla gama de experimentos para serem realizados, exigindo longos períodos de testes e alto custo (SINGH *et al.*, 2010; TAK *et al.*, 2015). Além de não serem capazes de revelar a influência das interações entre as variáveis independentes (FAN *et al.*, 2018), os modelos convencionais podem incluir certos desvios entre modelos teóricos e observações experimentais devido ao conhecimento insuficiente sobre a complexidade do sistema e por não considerar as questões associadas ao local de operação onde se encontram, prejudicando assim o modelo proposto. Portanto, soluções mais eficientes, como modelos baseados em inteligência artificial (IA), são desejáveis para esse fim.

As ferramentas de IA fornecem melhor controle de qualidade para as plantas industriais, aumentando a sua produtividade, minimizando erros de operação, otimizando os custos de produção e melhorando a eficiência geral do processo (AL AANI *et al.*, 2019). As principais ferramentas de IA incluem sistemas baseados em redes neurais artificiais (*artificial neural networks* - ANNs), algoritmo genético (*genetic algorithm* - GA), máquina de vetor de suporte (*support vector machine* - SVM), árvore de decisão (*decision tree* - DT), floresta aleatória (*random forest* - RF), árvore de regressão impulsionada (*boosted regression tree* - BRT), recozimento simulado (*simulated annealing* - SA), simulação de Monte Carlo (*Monte Carlo simulation* - MCS), otimização de enxame de partículas (*particle swarm optimization* - PSO), algoritmo imunológico (*immune algorithm* - IAg), algoritmo de colônia de formigas (*ant colony algorithm* - ACA) e algoritmo competitivo imperialista (*imperialist competitive algorithm* - ICA). As ferramentas de IA também podem ser combinadas com o design experimental para melhorar a previsão da solução ótima (BAGHERI *et al.*, 2015).

Embora inúmeros métodos de modelagem e otimização tenham sido desenvolvidos e testados visando a melhoria da eficiência no tratamento de efluentes, ainda existem desafios a serem enfrentados relacionados à comparação objetiva das técnicas, a falta de diretrizes, a validação limitada em escala real, a falta de métodos colaborativos e as opções limitadas para a otimização ativa do conteúdo e da qualidade da informação dos dados (COROMINAS *et al.*, 2018). Neste sentido, há espaço para o desenvolvimento de novos métodos que consigam contornar as limitações acima listadas.

Neste contexto, esta dissertação de mestrado tem como objetivo principal o desenvolvimento de um algoritmo de computador capaz de otimizar o processo de tratamento em uma ETE com base nos dados de operação desta. Esses dados já são comumente gerados pelas ETEs como parte do controle supervisionado dos processos que buscam garantir que o efluente final atenda aos padrões da legislação vigente. Desta forma, toda ETE possui, desde o início de sua operação, um conjunto de dados relativos ao tratamento, tais como as características do efluente bruto recebido, fluxos e características do efluente entre cada etapa do processo, produtos químicos adicionados, consumo de energia e as características do efluente final. Quanto maior a quantidade e qualidade dos dados da operação da ETE, maior será a capacidade do algoritmo de realizar as correlações necessárias entre as variáveis operacionais para alcançar a otimização do processo.

Pretende-se, assim, obter uma solução automatizada baseada em Inteligência Artificial que otimize a operação de ETEs combinando redução de custos e eficiência no tratamento. A solução envolve um conjunto de rotinas de programação que evoluem para um Sistema de Apoio à Decisão (SAD) para operações de ETEs. Entre o público-alvo da solução que se pretende desenvolver destacam-se concessionárias de saneamento públicas e privadas, indústrias químicas em geral, incluindo alimentos, bebidas, celulose, têxtil, petroquímica, entre outras, e empresas de consultoria e prestação de serviços na área de tratamento de efluentes.

1.1 ESTAÇÕES DE TRATAMENTO DE EFLUENTES

Estações de Tratamento de Efluentes (ETEs) são sistemas complexos responsáveis por tratar de modo eficiente o efluente bruto de residências e/ou indústrias para que seja descartado no corpo hídrico receptor dentro dos padrões estabelecidos pela legislação, exercendo um papel importante para a redução da poluição e melhora na qualidade do ambiente local.

A operação das ETEs, contudo, requer o uso de tecnologias e insumos que trazem consigo custos de operação e manutenção. Uma análise dos custos operacionais de ETEs localizadas no Sudeste brasileiro realizada por Von Sperling (VON SPERLING, 2016) mostra que os gastos com adição de produtos químicos e energia podem superar 40% dos custos para operação e manutenção de uma ETE a depender do processo de tratamento executado. Os gastos com pessoal, entre mão de obra própria e terceirizada, também respondem por uma parcela significativa dos custos, podendo superar 80% do total (BROSTEL; NEDER; SOUZA, 2001). Além disso, as características físico-químicas do efluente recebido variam em escala horária, diária e sazonal, o que implica em um controle periódico dos parâmetros de entrada e saída da ETE e adaptações de processo para o seu bom funcionamento.

Especificamente no tratamento de efluentes, a busca por uma operação custo-eficiente muitas vezes ocorre de forma manual por meio de tentativa e erro, observando-se como cada ajuste realizado na operação (por exemplo, o aumento na dosagem de um produto químico ou a redução do tempo de aeração) afeta os custos e a característica final do efluente tratado. Além de custosa e morosa, a adoção deste tipo de técnica não permite compreender de forma clara a relação entre as diversas variáveis da operação e a qualidade do efluente tratado, tendo em vista a

complexidade de uma ETE. Neste sentido, a Inteligência Artificial tem surgido como uma ferramenta poderosa e prática para reduzir a complexidade e as incertezas associadas ao tratamento de efluentes.

Utilizando os dados da operação de uma ETE e através de técnicas de IA, é possível desenvolver soluções capazes de modelar os processos de uma ETE e investigar como as variáveis operacionais influenciam a qualidade do efluente tratado. Wang *et al.* (2021), por exemplo, aplicaram métodos de Aprendizado de Máquina (*machine learning* - ML) para analisar os dados da ETE Umea, na Suécia. Os autores perceberam, entre outras descobertas, que a temperatura do efluente bruto é a variável que mais influencia as concentrações de Sólidos Suspensos Totais (SST) e fosfato (FO) no efluente tratado. As conclusões do estudo podem facilitar o desenvolvimento de estratégias de controle mais sofisticadas que podem melhorar significativamente a precisão do controle e reduzir os custos da operação.

Quando aplicada em conjunto com métodos de Controle Preditivo baseado em Modelo (*Model Predictive Control* – MPC), a IA possibilita prever em tempo real, a partir das características previstas do efluente bruto, as medidas operacionais necessárias para alcançar os objetivos em termos de eficiência de remoção e custos. Um método MPC baseado em ML aplicado por Bernardelli *et al.* (2020) na ETE do município de Modena, na Itália, mostrou que o algoritmo foi capaz de prever os picos de Nitrogênio Total do efluente com antecedência suficiente para adaptar a vazão de ar no tanque de aeração e garantir o atendimento aos padrões de lançamento do efluente juntamente com economia de energia.

Outras referências encontradas na literatura, tais como Mulas *et al.* (2015) e O'Brien *et al.* (2011), reforçam que tais métodos são uma alternativa viável para balancear as metas divergentes do tratamento de efluente – redução de custos e melhora na qualidade do efluente – visando otimizar a operação de uma ETE. Soluções deste tipo podem ser a chave para uma operação economicamente mais rentável e com menor impacto ambiental, trazendo benefícios para concessionárias de saneamento e indústrias, consumidores dos serviços e o meio ambiente.

O caráter inovador do método proposto neste trabalho pode ser evidenciado sob diferentes perspectivas de potencial de impacto social:

1. Científico: a solução proposta incorpora novos conhecimentos e auxilia na definição de uma metodologia mais apropriada ao cenário local considerando

os processos de tratamento comumente utilizados no Brasil e a qualidade dos dados operacionais.

2. Tecnológico: as conclusões do estudo podem resultar em mudanças estruturais na forma de operar as ETEs. Muitas ETEs no país possuem baixo controle de seus processos, o que resulta em dados escassos e pouco conhecimento sobre a relação entre as variáveis do processo e as características do efluente tratado. A metodologia proposta tem potencial para melhorar a eficiência e reduzir o custo de operação das ETEs ao promover um maior controle das operações e volume de dados.
3. Econômico: a melhoria da eficiência de ETEs implica na redução de custos operacionais, em especial no que se refere à adição de produtos químicos, consumo de energia e até em gastos com pessoal, tendo em vista o caráter autônomo do modelo. A partir do resultado obtido pelo método proposto, é possível ampliar a aplicação da ferramenta para processos de reuso de águas, beneficiando a indústria ou concessionária ao reduzir o consumo de águas da rede de distribuição ou fontes naturais, além de ampliar a disponibilidade hídrica para outras demandas.
4. Socioambiental: a melhor qualidade do efluente tratado é uma consequência direta da solução proposta, trazendo menor impacto aos corpos hídricos receptores e ao abastecimento humano e dessedentação de animais.

1.2 O PROJETO

O presente trabalho é fruto de um projeto CNPq - Programa RHAE (Recursos Humanos em Áreas Estratégicas) -, gerido pela Hydroinformatics Solutions Ltda, ou simplesmente, HydroInfo, empresa de consultoria nas áreas de hidráulica, recursos hídricos, oceanografia e engenharia ambiental, localizada em Florianópolis-SC, em parceria com a Universidade Federal de Santa Catarina (UFSC) – especificamente através do Laboratório de Materiais de Computação Científica (LabMAC) do Departamento de Engenharia Química e Engenharia de Alimentos desta instituição.

A HydroInfo tem realizado projetos no Brasil e em outros países nas áreas de portos, costas, geração de energia, saneamento, recursos hídricos, entre outras. A consultoria especializada oferecida pela HydroInfo envolve investigações de campo, análises de dados e estudos de modelagem matemática.

Os dados de operação da ETE foram fornecidos pela Cervejaria AMBEV, localizada no município de Lages-SC, e parceira do projeto. A cervejaria permitiu a divulgação de todos os dados contidos neste trabalho e nos que o antecederam, sem restrições quanto à Lei Geral de Proteção de Dados. A partir dos dados da operação da ETE o modelo de otimização foi desenvolvido e validado.

1.3 OBJETIVO

1.3.1 Objetivo geral

Avaliar se modelos de aprendizado de máquina conseguem prever com precisão parâmetros de qualidade do efluente (como DQO/DBO, NT e SST) de uma ETE a partir das variáveis de entrada e do contexto operacional, considerando defasagens temporais e efeitos do tempo de residência hidráulico, de modo a apoiar decisões operacionais.

1.3.2 Objetivos específicos

Para alcançar o objetivo geral, esta dissertação pauta-se nos seguintes objetivos específicos:

- Avaliar a capacidade preditiva de diferentes arquiteturas de Aprendizado de Máquina (RF, SVM, MLP e LSTM) na estimativa de parâmetros de qualidade (DQO, DBO, NT e SST), considerando distintas granularidades temporais — abrangendo horizontes de curto prazo (escala horária) e médio prazo (escala diária) — de modo a capturar tanto a dinâmica rápida de controle quanto as tendências operacionais consolidadas e respeitando o tempo de residência hidráulico.
- Verificar se a importância das variáveis de entrada – calculada por Permutation Importance e PDP – resulta em rankings consistentes entre os modelos e entre diferentes horizontes de previsão.

- Investigar se a inclusão de atributos derivados do modo de operação (*layout*, estado da planta, setpoints, eventos operacionais) melhora o desempenho preditivo dos modelos e em que magnitude esse ganho ocorre.

1.4 ORGANIZAÇÃO DO TRABALHO

A presente dissertação está estruturada de forma a conduzir o leitor desde o contexto teórico e operacional das Estações de Tratamento de Efluentes (ETEs) até o desenvolvimento, validação e discussão abrangente dos modelos de aprendizado de máquina propostos. A organização do trabalho reflete a natureza incremental da pesquisa, composta por duas grandes etapas investigativas, cada uma delas associada a um artigo científico publicado durante o projeto.

O Capítulo 1 apresenta a introdução geral ao tema, justificando a relevância da aplicação de métodos de Inteligência Artificial (IA) no tratamento de efluentes, contextualizando o projeto institucional que deu origem ao estudo e definindo os objetivos geral e específicos da pesquisa.

O Capítulo 2 reúne a base teórica e o estado da arte relacionados à área de estudo. São revisados conceitos fundamentais sobre efluentes industriais e municipais, técnicas de aprendizado de máquina comumente aplicadas a ETEs, desafios de implementação, lacunas do conhecimento e a importância da interpretabilidade (XAI) em processos industriais.

O Capítulo 3 descreve detalhadamente a Estação de Tratamento de Efluentes da AMBEV (Lages/SC), incluindo sua configuração físico-operacional, as operações unitárias envolvidas, os parâmetros monitorados, os objetos de otimização de interesse industrial e o diagnóstico abrangente do banco de dados fornecido pela empresa. Esta caracterização fundamenta a avaliação posterior do desempenho dos modelos em ambiente industrial real.

O Capítulo 4 corresponde à primeira etapa investigativa do trabalho. Nela, é desenvolvida e validada uma metodologia completa de predição de parâmetros de qualidade do efluente, aplicada simultaneamente a um ambiente controlado (simulação no software WEST) e ao cenário industrial da AMBEV. Essa etapa evidencia o contraste entre dados ideais e dados reais, permitindo identificar

limitações, desafios e requisitos mínimos para viabilizar aplicações de ML em ETEs. Os resultados apresentados neste capítulo originaram o artigo Predicting effluent quality parameters for wastewater treatment plant: A machine learning-based methodology, publicado na revista Chemosphere (FUCK et al., 2024; DOI: 10.1016/j.chemosphere.2024.141472).

O Capítulo 5 constitui a segunda etapa investigativa, na qual a análise é aprofundada com o uso de uma terceira ETE (Umbilo), envolvendo dados reais e simulados em diferentes frequências (diária e horária). São incorporadas arquiteturas recorrentes, como modelos LSTM, de modo a investigar a capacidade dos algoritmos em capturar dependências temporais de longo prazo. Essa etapa reforça a discussão sobre a maturidade dos dados, a engenharia de atributos e a interpretabilidade dos modelos. Os resultados deste capítulo embasaram o artigo Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis, publicado no Journal of Water Process Engineering (CECHINEL et al., 2024; DOI: 10.1016/j.jwpe.2023.104758).

O Capítulo 6 discute os achados das duas etapas combinadas, sintetizando os impactos da qualidade dos dados, da engenharia de atributos, da escolha dos algoritmos e da interpretabilidade para aplicações práticas em ETEs. Este capítulo integra os resultados dos diferentes cenários e propõe uma visão consolidada sobre a viabilidade de implementação de soluções baseadas em IA. Além disso, apresenta as conclusões gerais da pesquisa, destacando suas contribuições científicas, tecnológicas e operacionais para o desenvolvimento de métodos preditivos em ETEs.

2. BASE TEÓRICA E CONTEXTUALIZAÇÃO

Durante o desenvolvimento desta dissertação, foi conduzida uma investigação científica aprofundada sobre modelos de Inteligência Artificial aplicados a processos de otimização e controle de Estações de Tratamento de Efluentes (ETEs). Esse levantamento contemplou estudos publicados em bases de dados reconhecidas, como Scielo, Periódicos CAPES e Web of Science. Paralelamente, realizou-se uma revisão teórica abrangente sobre o funcionamento de ETEs, fundamentada tanto na literatura clássica quanto em trabalhos científicos recentes disponibilizados em plataformas digitais.

Essa etapa permitiu identificar e compreender os modelos de Inteligência Artificial mais adequados ao problema de pesquisa, bem como aprofundar o entendimento sobre a operação de ETEs e os parâmetros relevantes para atividades de monitoramento e controle. A revisão da literatura foi conduzida de forma contínua ao longo do trabalho, de modo a subsidiar discussões, orientar decisões metodológicas e apoiar a definição de estratégias ao longo das demais etapas da pesquisa.

2.1 EFLUENTE DE CERVEJARIA

O processo que envolve a produção de cerveja exige um consumo elevado de água e, conseqüentemente, gera um volume grande de efluentes. Em 1993, BRAILE e CAVALCANTI estimavam, com base em estudos americanos, que para cada 1 L de cerveja produzida eram gerados entre 13 L e 17 L de efluente. Institutos alemães, por sua vez, estimavam um volume ainda maior, variando entre 17 L e 24 L de efluente para cada 1 L de cerveja produzido. Trabalhos recentes mostram uma diminuição nesse índice. Santos (2005) descreve a geração de 3 L a 6 L de efluente por litro de cerveja produzida. Simate *et al.* (2011) relatam que a produção de 1 L de cerveja gera entre 3 L e 10 L de efluente residual.

Os efluentes de cervejaria são caracterizados por apresentarem uma alta demanda química de oxigênio (DQO) proveniente dos componentes orgânicos açúcares, amido solúvel, etanol, ácidos graxos voláteis etc. Geralmente sua temperatura varia entre 25 °C e 38 °C. Os níveis de pH variam entre 2 e 12 e são influenciados pela quantidade e tipo de produtos químicos usados na limpeza e

sanitização, que podem incluir compostos clorados, ácido nítrico e fosfórico e hidróxido de sódio (SIMATE *et al.*, 2011).

Em termos quantitativos, destacam-se as concentrações de Demanda Química de Oxigênio (DQO) entre 2.000 e 6.000 mg/L e de Demanda Bioquímica de Oxigênio (DBO) entre 1.200 e 3.600 mg/L. O efluente apresenta ainda alto teor de Sólidos Suspensos Totais (SST), na faixa de 2.900 a 3.000 mg/L, além da presença de nutrientes como fosfatos (10 a 50 mg/L) e nitrogênio (25 a 80 mg/L). Tais cargas são provenientes, sobretudo, do manuseio da matéria-prima e da levedura presente no efluente (BRITO *et al.*, 2007; SIMATE *et al.*, 2011).

A Tabela 1 apresenta os valores referentes ao mês de abril de 2022 das características físico-químicas do efluente bruto produzido pela cervejaria em estudo.

Tabela 1 – Intervalo de valores observados para características físico-químicas relevantes do efluente bruto produzido pela cervejaria no mês de abril de 2022

Parâmetro	Valores
pH	5,82 – 12,73
Temperatura (°C)	16,8 – 37,0
Demanda química de oxigênio - DQO (mg/L)	3100 – 7980
Demanda biológica de oxigênio - DBO (mg/L)	1292 – 1992
Relação DQO:DBO	3,00 – 4,08
Carga orgânica (kg DQO/dia)	7139,3 – 20902,8
Sólidos suspensos totais (mg/L)	5744 – 12114

Fonte: elaborado pelo autor.

Observa-se que as concentrações de DQO e SST obtidas na ETE em estudo superam os limites superiores das faixas típicas relatadas pela revisão bibliográfica. Tal fato evidencia que o efluente bruto desta unidade específica possui características de alta carga, devido à incorporação de correntes de processo ricas em resíduos sólidos, como levedura residual que, na literatura revisada, são frequentemente segregadas antes do envio à ETE.

O efluente bruto produzido pela cervejaria é caracterizado por grandes variações nos parâmetros de entrada, levando a uma importante variabilidade no

controle e monitoramento do sistema de tratamento. Caso a variabilidade não seja detectada em tempo hábil, pode afetar negativamente o processo de tratamento, sendo necessária a execução de ações corretivas que, em geral, demandam custos adicionais. Para evitar esse problema, se faz necessário saber como as características físico-químicas do efluente de entrada e como o funcionamento de cada etapa da ETE influenciam na qualidade do efluente final.

2.2 MODELOS DE INTELIGÊNCIA ARTIFICIAL

A crescente complexidade operacional e as rigorosas exigências ambientais impostas às Estações de Tratamento de Efluentes (ETEs) têm impulsionado a busca por estratégias de controle e otimização mais inteligentes e proativas (ZHAO *et al.*, 2020). Neste contexto, as técnicas de Aprendizado de Máquina (*Machine Learning*, em inglês; abreviado como ML) emergiram como uma ferramenta promissora, capaz de modelar as relações não-lineares intrínsecas aos processos biológicos e físico-químicos do tratamento de efluentes e revolucionar sua automação (AL AANI *et al.*, 2019).

2.2.1 Principais modelos de aprendizado de máquina aplicados a ETEs

A literatura científica reporta uma vasta gama de algoritmos de ML aplicados à previsão de parâmetros e à otimização de ETEs. Dentre os mais consolidados, destacam-se as Redes Neurais Artificiais (ANN), as Máquinas de Vetores de Suporte (SVM) e os métodos baseados em árvores de decisão, como o *Random Forest* (RF) (SINGH *et al.*, 2023).

As Redes Neurais Artificiais, em suas diversas arquiteturas como o *Multi-Layer Perceptron* (MLP), são frequentemente empregadas devido à sua capacidade universal de aproximação de funções complexas. Sua aplicação abrange desde a previsão da Demanda Química de Oxigênio (DQO) e Demanda Bioquímica de Oxigênio (DBO) (ALMOMANI, 2020; DOGAN *et al.*, 2008) até a modelagem do comportamento de lodos ativados. Trabalhos recentes demonstram, inclusive, a aplicação bem-sucedida de arquiteturas mais avançadas, como as redes de Memória de Curto e Longo Prazo (LSTM), para capturar a dinâmica temporal de séries de dados operacionais, sendo particularmente eficazes em prever picos de carga com

antecedência. Bernardelli *et al.* (2020), por exemplo, utilizaram um modelo preditivo baseado em ML para antecipar picos de Nitrogênio Total, permitindo o ajuste proativo da aeração e resultando em economia de energia em uma ETE na Itália.

Os modelos de *Random Forest* (RF) e *Support Vector Machines* (SVM) também apresentam desempenho robusto. O RF, por ser um método *ensemble* que combina múltiplas árvores de decisão, é notabilizado por sua resiliência a ruídos e a dados ausentes, além de fornecer métricas intrínsecas de importância de variáveis (BREIMAN, 2001). Já o SVM se destaca pela sua eficácia em problemas de classificação e regressão, mesmo com conjuntos de dados de dimensionalidade elevada, sendo uma alternativa poderosa para a modelagem de processos em ETEs, com estudos recentes demonstrando sua alta performance na otimização de sistemas avançados de tratamento (WAQAS *et al.*, 2022).

2.2.2 Desafios recorrentes na implementação de modelos preditivos

Apesar do sucesso reportado em diversos estudos, a transposição dos modelos de ML do ambiente acadêmico para a operação rotineira de uma ETE industrial é permeada por desafios significativos. O principal gargalo reside na qualidade e na disponibilidade dos dados operacionais (COROMINAS *et al.*, 2018; NEWHART *et al.*, 2019). A realidade industrial frequentemente envolve a compilação de dados em formatos heterogêneos, a presença de *outliers* decorrentes de falhas de sensores ou erros de registro manual, a falta de padronização de rótulos e, crucialmente, um elevado percentual de dados ausentes com periodicidade de coleta inconsistente (EERIKÄINEN *et al.*, 2020).

Esta lacuna entre os dados ideais e os dados reais compromete severamente a capacidade de treinamento e a generalização dos modelos. Muitas ETEs, embora gerem um grande volume de informações, não possuem um sistema de monitoramento contínuo para todos os parâmetros de relevância preditiva, como as concentrações de nitrogênio e fósforo, o que limita fundamentalmente a acurácia de qualquer modelo preditivo (NEWHART *et al.*, 2019). Portanto, a etapa de pré-processamento, diagnóstico e tratamento dos dados assume um papel tão ou mais importante que a própria escolha do algoritmo de ML.

2.2.3 A importância da interpretabilidade (XAI) em processos industriais

Um terceiro pilar para a aplicação bem-sucedida de ML em ETEs é a interpretabilidade dos modelos, um campo conhecido como Inteligência Artificial Explicável (em inglês, *Explainable AI*; abreviada como XAI). Muitos dos algoritmos de melhor desempenho, como as redes neurais profundas, operam como "caixas-pretas", o que gera ceticismo e dificulta sua adoção por parte dos operadores e engenheiros de processo, visto que a própria noção de interpretabilidade é complexa e multifacetada (LIPTON, 2018). Em um ambiente industrial, onde decisões operacionais podem ter impactos financeiros e ambientais significativos, é imperativo que os gestores possam compreender como e por que um modelo chega a uma determinada predição ou recomendação (MOLNAR, 2020).

Técnicas de interpretabilidade, como a Análise de Importância de Permutação (em inglês, *Permutation Importance*) e os Gráficos de Dependência Parcial (PDPs), são ferramentas essenciais para elucidar as relações que o modelo aprendeu entre as variáveis de entrada e os parâmetros de saída. A aplicação de tais técnicas permitiu, por exemplo, identificar que a temperatura do efluente bruto era a variável mais influente para as concentrações de Sólidos Suspensos Totais e fosfato, um *insight* valioso para o desenvolvimento de estratégias de controle mais sofisticadas (WANG *et al.*, 2021). A interpretabilidade, portanto, não é apenas um requisito acadêmico, mas uma condição necessária para construir a confiança do usuário no sistema e promover a colaboração efetiva entre a inteligência artificial e a equipe humana (RIBEIRO; SINGH; GUESTRIN, 2016).

2.2.4 Lacuna do conhecimento e justificativa do trabalho

A análise da literatura revela que, apesar dos avanços significativos na aplicação de algoritmos de ML para a modelagem de ETEs, persistem lacunas importantes a serem preenchidas. Embora muitos trabalhos validem modelos em cenários específicos, revisões críticas da área apontam que poucas pesquisas realizam uma comparação sistemática do desempenho desses algoritmos quando submetidos a fontes de dados de natureza distinta – como dados industriais ruidosos e incompletos *versus* dados simulados e controlados –, destacando a limitada validação dos modelos em condições reais e a falta de diretrizes claras para sua

aplicação (COROMINAS *et al.*, 2018; NEWHART *et al.*, 2019). Adicionalmente, a necessidade de transformar dados brutos em conhecimento útil sugere que a exploração do impacto da engenharia de atributos (em inglês, *feature engineering*) como ferramenta para capturar o contexto operacional da planta – tal como mudanças estruturais no processo ou paradas de equipamentos – ainda é um campo com grande potencial investigativo (COROMINAS *et al.*, 2018).

É precisamente nesta intersecção que a presente dissertação se insere. O trabalho propõe-se a endereçar esta lacuna ao desenvolver e validar uma metodologia de predição de ponta a ponta, desde o diagnóstico rigoroso dos dados até a avaliação da interpretabilidade, aplicando-a a diferentes casos de estudo. Ao contrastar os resultados obtidos em um cenário industrial real (ETE AMBEV) com cenários simulados e de alta fidelidade (ETEs WEST e Umbilo), bem como ao investigar o efeito da incorporação de informações operacionais, busca-se construir um entendimento mais profundo sobre os pré-requisitos práticos e os reais limites para a implementação bem-sucedida de soluções de Inteligência Artificial na otimização de Estações de Tratamento de Efluentes.

3. ESTAÇÃO DE TRATAMENTO DE EFLUENTES DA AMBEV (LAGES/SC)

3.1 DESCRIÇÃO E CARACTERIZAÇÃO

A unidade da Cervejaria AMBEV de Lages/SC possui uma estação de tratamento de efluentes (ETE) com capacidade para processamento de 90 m³/h. O efluente recebido pela ETE é composto por todas as correntes de descarte de líquidos geradas pela cervejaria, incluindo a água proveniente da lavagem das garrafas e das tinas de fermentação, da limpeza de equipamentos e maquinários, dos despejos que ocorrem por produtos fora das especificações pré-estabelecidas pelas empresas e do esgoto sanitário da unidade fabril. As águas pluviais coletadas na unidade fabril são analisadas para controle e descartadas em lagoa específica, não sendo enviadas para a ETE.

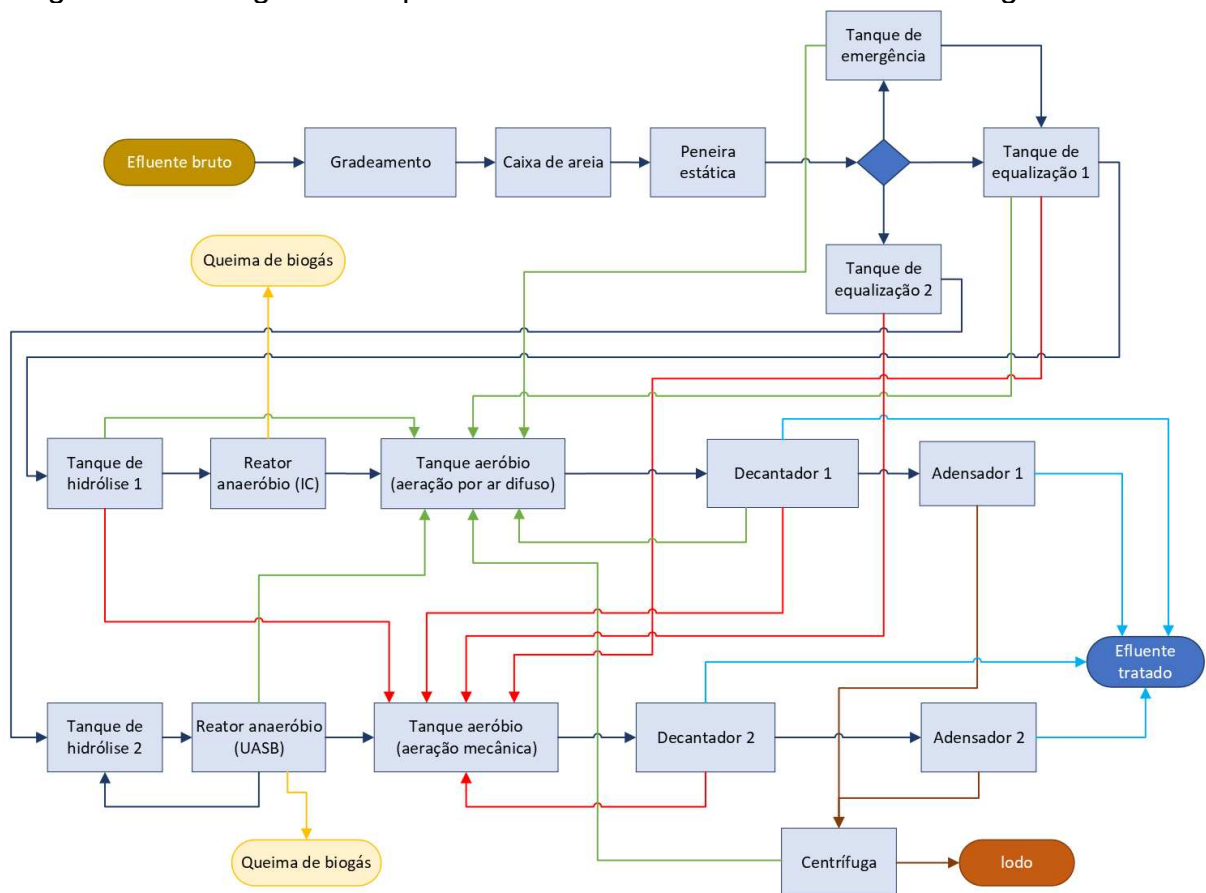
De acordo com informações coletadas com os operadores da ETE, o volume e a qualidade do efluente variam conforme o nível de produção, com a operação de manutenção na unidade fabril e com as condições climáticas. Em períodos de alta produção de cerveja ou quando há o descarte de algum lote em não conformidade, o índice de carga orgânica a ser tratado é maior. Quando há manutenção e limpeza de tanques na unidade fabril, o efluente bruto terá a presença de agentes químicos e biocidas, que podem ser prejudiciais ao bom funcionamento da ETE. Por fim, alterações bruscas de temperatura ou elevados volumes pluviométricos também alteram significativamente a eficiência de tratamento.

A despeito da complexidade da matriz efluente supracitada, a infraestrutura de monitoramento da ETE apresenta restrições instrumentais que dificultam o controle em tempo real. Observa-se a ausência de analisadores *on-line* para quantificação contínua de parâmetros críticos, o que impõe uma dependência de coletas manuais pontuais. Agrava-se este cenário pela inerente morosidade das metodologias analíticas convencionais, notadamente a determinação da Demanda Bioquímica de Oxigênio (\$DBO_5\$), que requer um período de incubação de cinco dias. Esse hiato temporal entre a amostragem e a obtenção do resultado analítico inviabiliza ações corretivas imediatas frente a oscilações de carga orgânica, evidenciando a necessidade de ferramentas preditivas capazes de inferir tais parâmetros com maior celeridade.

3.2 ESTAÇÃO DE TRATAMENTO DE EFLUENTES INDUSTRIAIS - AMBEV

A estação de tratamento de efluentes industriais (ETE) da fábrica da AMBEV, unidade Lages, 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 1.

Figura 1 – Fluxograma simplificado da ETE da AMBEV – Unidade Lages/SC



Fonte: elaborado pelo autor.

O processo de tratamento da ETE pode ser dividido em 4 macro etapas: (i) recebimento do efluente e pré-tratamento; (ii) tratamento anaeróbio; (iii) tratamento aeróbio; e (iv) descarte de lodo e de efluente tratado.

A primeira etapa consiste no recebimento do efluente bruto da produção, conjuntamente com o efluente sanitário. O controle da qualidade do afluente da ETE é muito importante para o bom funcionamento do tratamento biológico, que é muito

sensível a variações bruscas do meio e à presença de alguns compostos químicos. A primeira avaliação do efluente bruto feita pelos operadores da ETE é visual, realizada através das câmeras de controle e no ato da coleta da amostra. Essa avaliação permite verificar a presença de coloração atípica e/ou quantidade excessiva de particulados e espuma, que caracterizam não conformidades relevantes nos padrões de entrada.

Ao identificar essas alterações, o efluente bruto pode ser desviado para o tanque de emergência após passar pelo pré-tratamento, preservando a qualidade do tratamento biológico. Segundo informações coletadas na ETE, efluentes com alta carga orgânica ou que apresentem resíduos de biocidas são encaminhados para o tanque de emergência, que possui capacidade de 1.100 m³ e também é utilizado em situações de transbordamentos. O efluente armazenado no tanque de emergência retorna ao processo aos poucos, de forma a manter os parâmetros de entradas adequados ao sistema.

A etapa de pré-tratamento consiste no conjunto de operações unitárias responsáveis pela separação de particulados mais grosseiros presentes no efluente. Ao entrar na ETE, o efluente passa pelo gradeamento, operação unitária responsável pela remoção de materiais grosseiros do efluente que podem danificar ou obstruir equipamentos subsequentes, reduzir a confiabilidade e a eficiência do processo ou contaminar tubulações (TCHOBANOGLIOUS *et al.*, 2016). Na sequência, o efluente passa por uma caixa de areia, responsável pela remoção de sólidos pesados ainda presentes no efluente após a etapa de gradeamento e que possuem velocidade de sedimentação superior aos sólidos orgânicos de menor massa específica. Essa etapa é necessária para reduzir a formação de depósitos pesados em tanques de aeração, digestores aeróbios, adutoras, canais e condutos; diminuir a frequência de limpeza de equipamentos causada pela acumulação excessiva de sólidos; e evitar a abrasão de partes móveis dos equipamentos (TCHOBANOGLIOUS *et al.*, 2016).

O efluente resultante da caixa de areia não apresenta sólidos pesados, porém pode conter resíduos sólidos da produção de cerveja, como o bagaço de malte, que não é separado nas etapas anteriores por ser muito leve e de menor diâmetro. Para isso, o efluente é submetido ao peneiramento em 3 peneiras estáticas parabólicas fixas, onde esses sólidos leves e pequenos são retidos. A matéria orgânica retida nessa etapa é coletada em containers e encaminhada para descarte. O efluente resultante isento de particulados é, então, encaminhado aos tanques de equalização.

O tanque de equalização tem como função principal evitar problemas operacionais causados pela variação de vazão e de carga, fornecendo valores aproximadamente constante para as etapas posteriores de tratamento (VON SPERLING, 2017). Com isso, elimina-se ou minimiza-se as cargas de choque no tratamento biológico e melhora-se a utilização dos equipamentos de aeração, com redução da capacidade do soprador (TCHOBANOGLIOUS *et al.*, 2016). O efluente equalizado é encaminhado a dois tanques de hidrólise, que correspondem ao início do tratamento biológico do efluente.

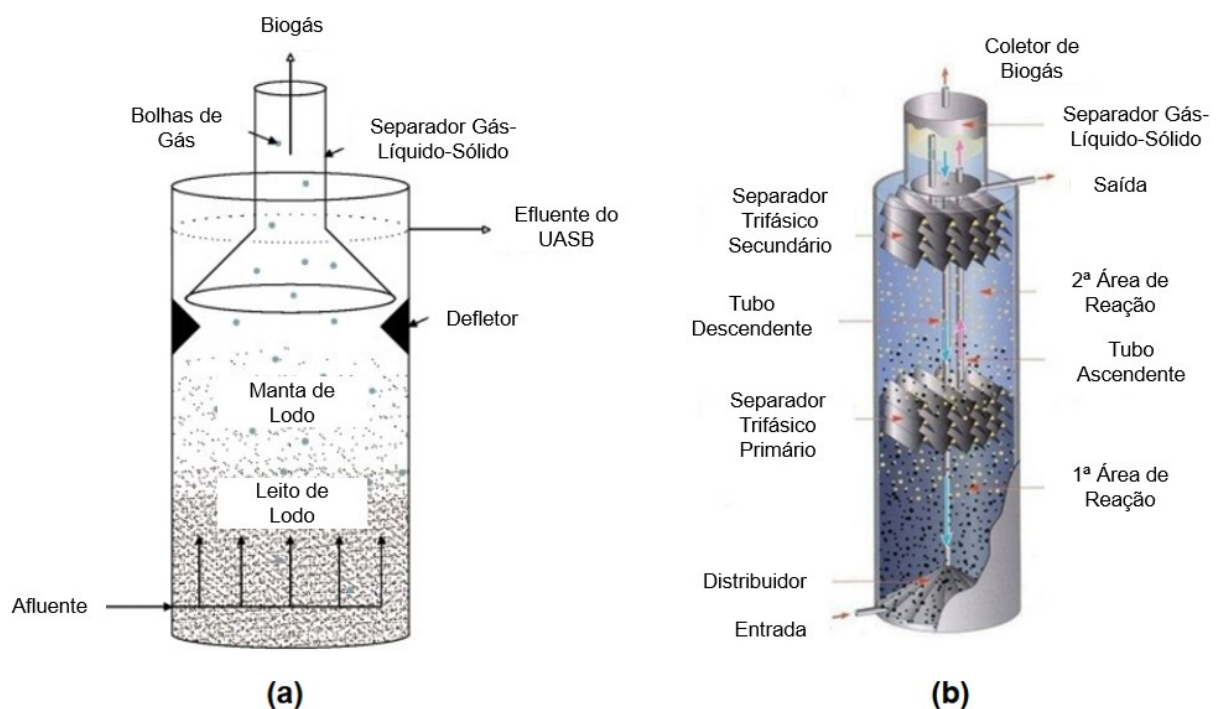
O tratamento biológico tem como função reduzir a quantidade de matéria orgânica presente no efluente a partir da ação de microrganismos capazes de oxidá-la em produtos aceitáveis. Pode ser conduzido em sistemas aeróbios (com a presença de oxigênio) ou em sistemas anaeróbios (em ausência de oxigênio). Além de oxidar os constituintes biodegradáveis, o tratamento biológico pode capturar ou incorporar sólidos suspensos e coloidais em flocos biológicos ou biofilmes, e remover outros compostos e nutrientes, como nitrogênio e fósforo (TCHOBANOGLIOUS *et al.*, 2016; VON SPERLING, 2017).

A hidrólise constitui a primeira etapa de tratamento biológico anaeróbio. Nela, o material particulado de alto peso molecular é convertido a compostos solúveis de menor peso molecular por enzimas extracelulares (ou exoenzimas), que são mais facilmente assimilados pelos microrganismos presentes nas etapas seguintes (HOYOS, 2016; ROCHA, 2003). Nos tanques de hidrólise da ETE também é feito o ajuste do pH do efluente por adição de NaOH ou HCl. A dosagem dos reagentes é feita de forma automatizada para que o pH na entrada dos reatores anaeróbios esteja entre 6,5 e 7,5.

A ETE possui dois reatores anaeróbios em sua linha de operação, um do tipo UASB (nesta dissertação denominado Reator 1) e outro do tipo IC (denominado Reator 2), de 3.000 m³ e 700 m³, respectivamente. O principal objetivo dos reatores anaeróbios é a decomposição de matéria orgânica e a consequente redução no teor de DBO. O reator anaeróbio de fluxo ascendente e manta de lodo, conhecido como UASB (em inglês, *Upflow Anaerobic Sludge Blanket*) é usado para o tratamento de uma variedade de águas residuais e possui eficiência de remoção de matéria orgânica de até 90% (BRITO *et al.*, 2007; SIMATE *et al.*, 2011). Um esquema representativo dos reatores UASB e IC é apresentado na Figura 2.

Quando em funcionamento estacionário, o reator possui um leito de lodo muito denso que pode ser de natureza granular ou floculante com altas propriedades de decantação e que se concentra na região mais ao fundo do reator. Acima do leito de lodo denso, há uma zona de manta de lodo com partículas com menor velocidade de sedimentação. As reações biológicas ocorrem nessas duas regiões (CHONG *et al.*, 2012; LATIF *et al.*, 2011). O efluente é inserido no reator em fluxo ascendente, passando inicialmente pelo leito de lodo denso, onde ocorre a decomposição anaeróbica da matéria orgânica com uma alta taxa de atividade microbológica. Os compostos orgânicos solúveis presentes no efluente são convertidos em biogás, uma mistura gasosa composta principalmente por CH_4 e CO_2 . O efluente segue o fluxo ascendente e é coletado por calhas laterais. Já o biogás produzido pela decomposição anaeróbica é separado do efluente pelo separador trifásico. Os defletores de gás impedem o arraste de lodo pelas bolhas de gás deslizando os sólidos de volta à zona de reação. A turbulência natural, causada pela vazão de efluente e pela produção de biogás, proporciona o contato entre a biomassa presente no reator e os compostos a serem oxidados e, por isso, o reator não necessita de um sistema de mistura adicional (CHONG *et al.*, 2012).

Figura 2 – Esquema representativo dos reatores UASB (a) e IC (b)



Fonte: Chong *et al.* (2012); Cui, Zhou e Zhang (2011).

O reator de circulação interna, ou reator IC, consiste basicamente na instalação de dois reatores UASB sobrepostos, tal como apresentado na Figura 2b. Por apresentar dois estágios de separação, o reator IC possui alta relação altura-diâmetro, proporcionando uma economia de espaço (WANG *et al.*, 2018). O princípio de funcionamento do reator IC é muito semelhante ao do reator UASB. O efluente é inserido de forma ascendente no reator UASB inferior, onde a maior parte do biogás é removida. Com a redução da turbulência, o segundo reator é mais efetivo para separar o lodo anaeróbio do efluente ascendente (BRITO *et al.*, 2007).

O biogás produzido pelos reatores anaeróbios é altamente energético e, atualmente, é queimado em *flare* atmosférico, convertendo o CH₄ em CO₂, menos prejudicial ao ambiente. Há em andamento um projeto de aproveitamento energético do biogás visando o aquecimento de caldeiras. Com relação ao lodo, caso haja crescimento celular acima do ideal para o projeto, o excesso é purgado do sistema e encaminhado para desidratação e estocagem, ou enviado a outra unidade de tratamento de efluentes. Por fim, embora a eficiência de remoção de matéria orgânica dos reatores UASB e IC seja significativa, usualmente se faz necessário um pós-tratamento ao efluente para atender os padrões de lançamento. No caso da ETE em estudo, ao sair dos reatores anaeróbios, o efluente é encaminhado a dois tanques de lodos ativados para o tratamento aeróbio.

A segunda etapa do tratamento biológico realizado na ETE corresponde ao tratamento aeróbio em sistema de lodos ativados. O processo de lodos ativados tem por objetivo a conversão de matéria orgânica e/ou outros compostos em gases e tecido celular por microrganismos aeróbios que são mantidos em suspensão (TCHOBANOGLIOUS *et al.*, 2016). Para essa etapa, a ETE possui dois tanques de aeração, um com capacidade para 2.860 m³ de efluente e com aeração mecânica de superfície (nesta dissertação denominado Tanque 1) e outro com 3.500 m³ e aeração submersa por ar difuso (denominado Tanque 2), além de dois tanques de decantação cilíndricos.

O efluente proveniente dos reatores anaeróbios é bombeado aos tanques de aeração, onde a matéria orgânica e os flocos microbianos são mantidos em suspensão homogênea pela ação da aeração. Essa mesma etapa promove a oxigenação do meio, permitindo que a concentração de O₂ seja mantida entre 1,0 mg/L e 3,0 mg/L. A remoção da matéria orgânica ocorre por reações bioquímicas de

oxidação microbiana, consumindo oxigênio e convertendo-a em CO₂, água e novas células (crescimento de microrganismos).

O sistema de lodos ativados possui duas modalidades de operação: a aeração convencional e a aeração prolongada, sendo essa última a adotada na ETE. Nessa modalidade, o lodo permanece no tanque por um período mais longo (entre 18 e 30 dias) e o tempo de detenção hidráulica do efluente também é elevado (entre 16 h e 24 h), promovendo uma maior proliferação de bactérias presentes no efluente (VON SPERLING, 2017). Como a carga alimentada no sistema permanece constante, haverá menor disponibilidade de alimento para as bactérias, mantendo a relação alimento/microrganismos (A/M) entre 0,07 e 0,15. Com isso, após consumir toda a matéria orgânica disponível, a biomassa passa a utilizar a matéria orgânica de suas células para efetuar a metabolização. A combinação dos tratamentos anaeróbio e aeróbio proporciona obter eficiências de remoção de DQO superiores a 98 % (SIMATE *et al.*, 2011).

Conforme observado na Figura 1, os tanques de aeração podem receber cargas de diferentes pontos da ETE, o que possibilita operar a estação de forma contínua mesmo em situações atípicas de cargas de entrada ou de rotinas operacionais e manutenção de equipamentos. O tanque com aeração por ar difuso pode receber carga proveniente dos tanques de equalização, emergência e hidrólise; da saída dos reatores UASB e IC; do efluente separado do lodo na etapa de centrifugação; e de parte do lodo coletado nos decantadores. O tanque com agitação mecânica pode receber carga proveniente dos tanques de equalização e de hidrólise, da saída do reator UASB e parte do lodo coletada nos decantadores.

Após sair dos tanques de aeração, o efluente é enviado aos decantadores, ou clarificadores, onde ocorre a separação do lodo aeróbio e do efluente tratado. O decantador tem por finalidade clarificar o efluente, tornando-o límpido e com baixa concentração de sólidos suspensos e turbidez, e espessar o lodo gerado para a etapa posterior de tratamento. A separação das partículas ocorre por ação gravitacional e a eficiência de decantação está intimamente associada à característica de sedimentabilidade da biomassa, tal como o índice volumétrico do lodo (IVL). O efluente tratado separado nessa etapa é direcionado à calha de descarte para o corpo receptor.

Como a produção de lodo em sistemas de lodos ativados é alta, convencionou-se recircular parte do lodo separado na etapa de decantação aos tanques de aeração. O

lodo recirculado possui microrganismos ativos, o que aumenta a concentração da biomassa no tanque e auxilia na estabilidade da matéria orgânica (VON SPERLING, 2017). A parte do lodo não recirculada ao tanque de aeração é encaminhada para a etapa posterior de adensamento. Essa etapa é necessária para aumentar o teor de sólidos antes da etapa de centrifugação a partir da remoção de parte da água que constitui o lodo. A água separada no adensamento é direcionada à calha de descarte para o corpo receptor e o lodo parcialmente desidratado é encaminhado para centrifugação.

A centrifugação é a última etapa de acondicionamento do lodo gerado no processo de tratamento aeróbio. É um processo de separação forçada sólido/líquido por ação centrífuga que ocorre em dois estágios: no primeiro, conhecido como clarificação, as partículas sólidas do lodo se depositam a uma velocidade muito maior do que por gravidade; no segundo, ocorre a compactação e o lodo perde parte de sua água sob a ação prolongada da centrifugação (GONÇALVES; LUDUVICE; SPERLING, 2007). É comum, nessa etapa, a adição de solução de polímero para aumentar a eficiência de separação sólido-líquido a partir do aumento do peso e tamanho dos sólidos presentes no resíduo (LIBÂNIO, 2010). O lodo centrifugado é condicionado em *bags* para posterior descarte e o líquido removido nessa etapa retorna ao tanque de aeração por ar difuso para tratamento.

3.3 IDENTIFICAÇÃO DOS OBJETOS DE OTIMIZAÇÃO DA ETE AMBEV

Durante o projeto, foi realizada visita de campo à ETE da cervejaria AMBEV, localizada em Lages/SC, de modo a conhecer os detalhes do funcionamento da ETE. Na primeira visita, realizada no dia 30/5/2022, foram coletadas informações sobre o fluxograma do processo, necessárias para identificar as operações unitárias envolvidas, as correntes de fluxo e os detalhes de funcionamento dos equipamentos presentes na ETE. Também foram coletadas informações qualitativas e quantitativas sobre o efluente, sobre os pontos de amostragem e sobre o sistema de controle da operação da ETE.

Durante a visita, foi possível acompanhar a operação de coleta das amostras do efluente e identificar os diferentes pontos de amostragem da ETE. As amostras coletadas são caracterizadas para uma série de parâmetros, dependendo do ponto de coleta. Também foi possível compreender a rotina e a frequência de preenchimento

dos dados de monitoramento e controle da ETE no sistema interno da AMBEV. Esses dados compõem o banco de dados utilizado neste estudo.

Além disso, foram realizadas diversas reuniões com os responsáveis pela operação da ETE para identificar os principais problemas e desafios vivenciados durante o seu funcionamento, além de obter informações sobre a sazonalidade do efluente e sobre os métodos de análise e caracterização para melhor compreensão do histórico de coleta dos dados. Nessas reuniões, a Cervejaria AMBEV indicou os possíveis objetos de otimização de interesse da indústria:

- a) Redução do consumo de produtos químicos nas etapas de hidrólise e adensamento/centrifugação de lodo.
- b) Redução do consumo de energia pelos equipamentos responsáveis pela aeração dos tanques de lodos ativados.
- c) Redução do custo associado à destinação do lodo removido dos tanques de aeração, ou seja, a redução da produção de lodo.
- d) Aumento na produção de lodo anaeróbico no reator UASB.

As etapas posteriores desse estudo foram realizadas com base nos objetos de otimização e nos dados de monitoramento e controle da ETE fornecidos pela AMBEV.

3.4 DIAGNÓSTICO DOS DADOS DE CONTROLE E MONITORAMENTO DA ETE AMBEV

O controle e monitoramento de diversos parâmetros de qualidade em estações de tratamento de efluentes possibilita avaliar o desempenho do processo ao longo do tempo, detectar quaisquer eventos especiais ocorridos e, assim, controlar a qualidade do efluente final (AVELLA *et al.*, 2011). Contudo, uma enorme quantidade de dados é registrada para esse fim e técnicas adequadas são necessárias para utilização a obtenção de informações úteis a partir do banco de dados.

O controle e o monitoramento da ETE são realizados pelos operadores responsáveis pelo setor, em regime de turnos, por meio de um programa supervisor específico que fornece informações sobre a operação, tais como vazões, níveis dos tanques, desempenho dos equipamentos, controle de pH, dosagem de produtos químicos e níveis de oxigênio nos tanques de aeração, entre outros parâmetros. Em

cada etapa do processo, amostras de efluentes são coletadas uma vez por turno, totalizando três coletas por dia. A ETE dispõe de um laboratório de análises químicas que permite a caracterização das amostras recolhidas para diversos parâmetros físico-químicos. Algumas análises de caracterização e monitoramento são terceirizadas em situações de auditoria e fiscalização ambiental. Os indicadores de qualidade das águas residuais são recolhidos e registados com variabilidade temporal (por turno, diário, semanal, quinzenal ou mensal). As informações obtidas são armazenadas em sistema de gestão de processos (LiveMES) e em planilhas de controle interno.

Para a execução deste projeto, a AMBEV disponibilizou os dados adquiridos e armazenados a partir do monitoramento e controle dos parâmetros da ETE entre janeiro de 2020 e janeiro de 2023. Os dados disponibilizados têm duas fontes: as informações armazenadas no sistema de gestão de processos (LiveMES) e as informações registradas nas planilhas de controle interno da ETE. Este estudo utilizou dados de medições diárias de afluentes e efluentes abrangendo 752 dias. De posse dessas informações, um diagnóstico global da qualidade do banco de dados foi realizado, identificando uma série de inconsistências:

- a) Os dados estavam compilados em formatos diferentes entre as planilhas, de forma que não era possível unificar todos os dados em um único arquivo sem um tratamento prévio.
- b) Os arquivos apresentavam dados que não eram relevantes para o monitoramento e controle da ETE, tais como dados relacionados ao funcionamento da ETA e da unidade industrial.
- c) As unidades de medida dos itens de processo listados nas planilhas e alguns valores de especificação de projeto para determinação de não conformidades não foram fornecidas nos arquivos exportados.
- d) Os arquivos apresentavam siglas e abreviações cuja identificação do significado não foi possível de forma imediata.
- e) Alguns itens de processo apresentavam diferentes rótulos em diferentes períodos de coleta de dados e entre as planilhas, havendo a necessidade de padronização dos rótulos.

- f) A presença de dados discrepantes (ou dados *outliers*), sem conexão com o estado real do processo e/ou inconsistentes com a operação da ETE, foi verificada para maior parte dos itens de processo.
- g) Parte dos dados calculados nas planilhas de controle também é registrada na plataforma MES, gerando duplicidade de dados.
- h) Os itens de processo apresentam variabilidade temporal de coleta e preenchimento de dados (turno, diário, semanal, quinzenal, mensal). Contudo, a periodicidade no preenchimento não é consistente.
- i) Um elevado percentual de dados ausentes foi verificado para alguns dos itens de processo, o que pode afetar consideravelmente o desempenho das próximas etapas.

Além do diagnóstico global, foi avaliado se o conjunto de dados fornecidos pela AMBEV possuía dados suficientes e de qualidade para serem aplicados aos modelos de IA visando a otimização dos objetos elencados na seção 3.3. As seguintes inconsistências foram identificadas:

- a) Não havia informações disponíveis suficientes sobre os consumos de produtos químicos, tal como a quantidade de produto usada em cada operação (e não somente o consumo total na ETE) e os parâmetros de controle e monitoramento associados ao consumo (pH, acidez, alcalinidade, potencial redox, etc).
- b) Informações sobre os parâmetros operacionais dos tanques de aeração não eram registradas em tempo real, o que se considerava fundamental para aprimorar a precisão dos modelos para implementação de MPC. As informações de controle dos parâmetros eram registradas apenas durante os turnos, ou seja, três vezes ao dia, o que poderia resultar em atrasos na resposta do modelo. Para abordar essa questão, foi proposto à AMBEV a instalação de sensores de monitoramento para aquisição e registro contínuo de dados de OD, pH, T, DQO e vazão em tempo real. No entanto, essa sugestão não foi implementada pela empresa parceira.
- c) A disponibilidade de informações relacionadas ao desempenho do reator UASB era insuficiente para atender ao objetivo de otimização identificado. Para suprir essa necessidade, era vital conduzir análises diárias dos

parâmetros SV/ST do lodo anaeróbio, bem como avaliar, no mínimo, a massa de lodo anaeróbio e a DBO. Essa solicitação foi estendida à empresa parceira, porém não obteve atendimento.

A partir dessa avaliação, decidiu-se por explorar diferentes abordagens de pesquisa com o objetivo de criar modelos preditivos de ML para a qualidade do efluente da AMBEV e analisar a capacidade desses modelos de se ajustarem a conjuntos de dados de entrada de outras ETEs. Os resultados alcançados a partir desses estudos serão detalhados nos capítulos subsequentes.

4. DESENVOLVIMENTO DE MODELOS PREDITIVOS: VALIDAÇÃO EM AMBIENTE CONTROLADO E DESAFIOS EM CENÁRIO INDUSTRIAL

4.1 JUSTIFICATIVA DA PROPOSTA DE ESTUDO

Dentre as possíveis aplicações de ML em ETEs, os modelos preditivos de parâmetros de qualidade de efluentes são os mais relatados na literatura. Esses modelos são desenvolvidos usando um conjunto de dados históricos e experiência para identificar padrões e tendências e depois usar esses padrões para prever o valor futuro da variável de interesse (YANG *et al.*, 2022). Essa tecnologia tem sido aplicada a previsões de processos complexos, como processo de lodo ativado (BAGHERI *et al.*, 2015), conteúdo de nitrogênio total (NT) e demanda química de oxigênio (DQO) em biorreatores de membrana (ALMOMANI, 2020), demanda bioquímica influente de oxigênio (DBO) em estações de tratamento de esgoto (DOGAN *et al.*, 2008), índice volumétrico de lodo (IVL) (HAN; WANG; QIAO, 2014), e outros parâmetros. Vários modelos, incluindo *K-Nearest Neighbors* (KNN), *Support Vector Machines* (SVM), *Random Forest* (RF) e diferentes estruturas de Rede Neural Artificial (ANN), como *Multi-Layer Perceptron* (MLP) e *Long Short-Term Memory* (LSTM), são usados para esse fim (SINGH *et al.*, 2023).

Os conjuntos de dados utilizados na construção do algoritmo são categorizados em dados de treinamento, teste e validação dos modelos. Deve-se notar que o número de dados necessários pode variar dependendo da incerteza associada a um determinado parâmetro (SINGH *et al.*, 2023). Construir um banco de dados robusto e consistente é um dos principais desafios para a aplicação bem-sucedida de ML em ETEs. Embora produzidos em quantidades significativas, os dados relativos ao funcionamento da ETE e demais parâmetros monitorizados são recolhidos em diferentes intervalos de tempo e provenientes de outras fontes, quer através de análises laboratoriais, quer através de sensores instalados na central. Eles são frequentemente registrados em vários bancos de dados (NEWHART *et al.*, 2019). Outra limitação relacionada ao banco de dados é a forma como as informações são gerenciadas e armazenadas nas ETEs. Alguns sensores são limitados e armazenam apenas informações de alarme com base em limites ou anotações de ativação/desativação. Outros parâmetros são registrados com base nos sentidos humanos, ou seja, com base em observações qualitativas feitas pelos operadores

(EERIKÄINEN *et al.*, 2020). Agrupar e padronizar esses dados é 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. As ETEs comprometidas com a qualidade do efluente tratado atualizam frequentemente e constantemente o seu processo através da implementação de novos sensores, alterações na metodologia de análise laboratorial ou instalação de novos equipamentos em toda a central. Essas mudanças podem afetar o comportamento dos parâmetros, dificultando a atribuição da causa das diferenças entre os eventos de amostragem (ALAVI *et al.*, 2022; NEWHART *et al.*, 2019) e, conseqüentemente, complicando a aplicação de modelos de ML.

O pré-processamento de dados e a engenharia de recursos são duas etapas que podem ajudar a criar conjuntos de dados mais convenientes para a “máquina” entender as relações entre as variáveis do processo e construir modelos robustos de previsão de processos (BHADESHIA *et al.*, 2009). O pré-processamento de dados inclui coleta e limpeza, onde os dados de diferentes fontes são agrupados e problemas como informações redundantes e *outliers* são removidos. A discretização e normalização dos dados, se necessária, também fazem parte da etapa de pré-processamento dos dados (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2007).

A engenharia de recursos visa extrair informações úteis de dados brutos, criar recursos e representar padrões subjacentes para melhorar o desempenho de modelos de aprendizado de máquina. É uma técnica promissora com diversas aplicações potenciais, pois reduz significativamente a dependência do conhecimento especializado no treinamento de modelos e acelera seu uso por usuários não especialistas (CAI *et al.*, 2020). Embora o pré-processamento de dados e a engenharia de recursos possam mitigar alguns problemas com dados insuficientes, removendo valores discrepantes, inserindo valores ausentes e transformando variáveis, essas etapas não podem corrigir fundamentalmente dados inerentemente tendenciosos ou imprecisos, ou seja, não substituem boas práticas de coleta de dados.

Outra ferramenta essencial para a compreensão dos resultados produzidos pelos modelos de ML, especialmente para modelos de classificação e regressão, é a *Feature Importance (FI)*. Uma variável é necessária para a previsão do modelo se o erro do modelo aumentar quando os valores da variável forem embaralhados ou

alterados. Isso indica que o modelo depende da variável avaliada para apresentar melhores previsões. Da mesma forma, uma variável é considerada “sem importância” se o erro do modelo quase não muda ao embaralhar ou alterar seus valores. Isso indica que o modelo ML ignorou ou atribuiu menos importância a essa variável durante o processo de predição (MOLNAR, 2020). Esta análise permite maior confiabilidade dos modelos de ML por parte dos operadores humanos, visto que os usuários podem verificar se as decisões tomadas pelos modelos são consistentes com a realidade operacional ou se é necessário reconsiderar as decisões tomadas por eles (LIPTON, 2018; RIBEIRO; SINGH; GUESTRIN, 2016). Alguns exemplos de ferramentas usadas na análise de importância de recursos incluem os Gráficos de Dependência Parcial (PDPs), a Importância por Permutação (PI), o método de Explicações Locais Agnósticas ao Modelo (LIME) e as Explicações Aditivas de Shapley (SHAP).

Com base nessas informações, decidiu-se avaliar o uso de modelos de ML para prever parâmetros de qualidade do efluente a partir de dados de entrada e saída obtidos em duas ETEs distintas: uma simulada por meio da ferramenta de modelagem de qualidade da água WEST (MIKE Powered by DHI Group, 2023) e outra referente à ETE AMBEV.

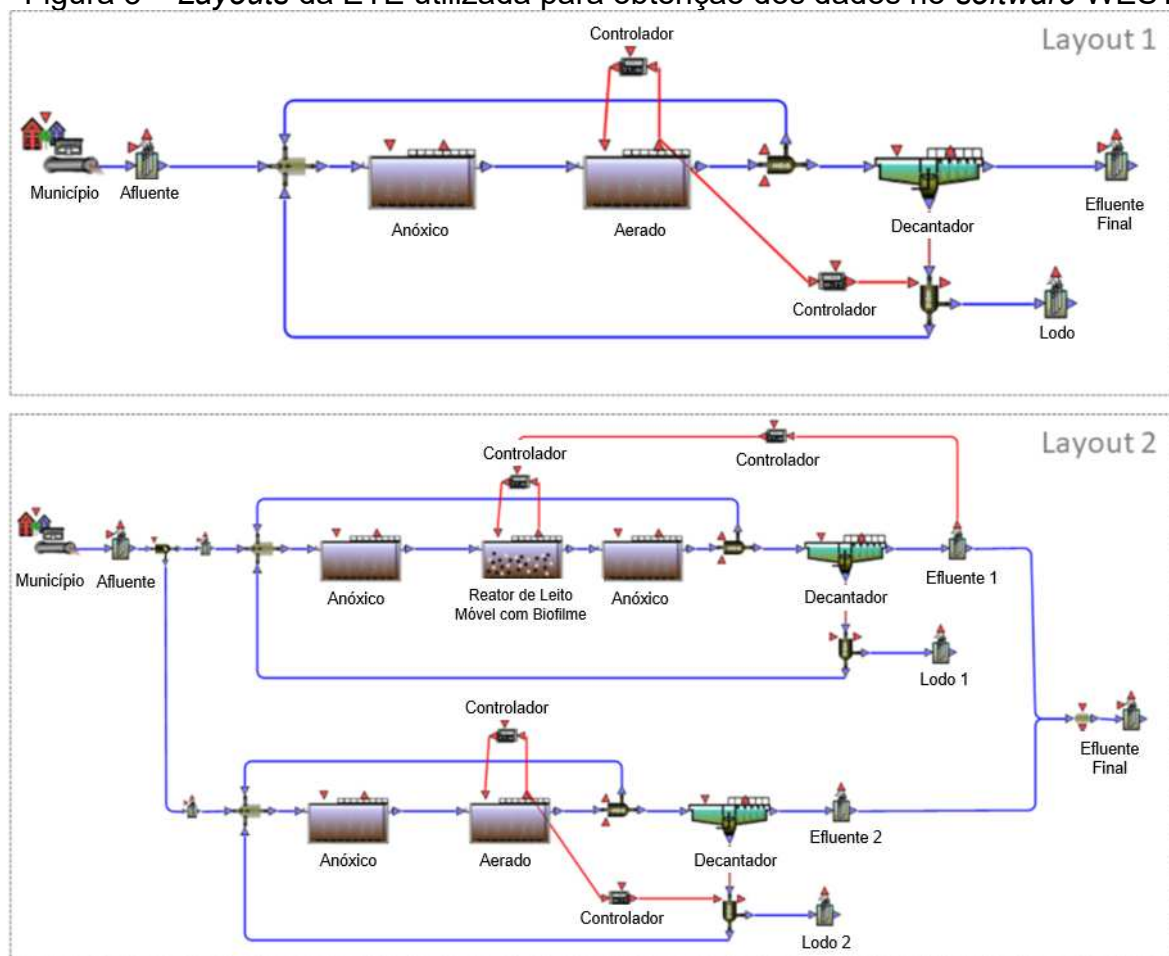
4.2 ESTUDO DE CASO E DESCRIÇÃO DO CONJUNTO DE DADOS

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 3.

O *Layout* 1 (Figura 3) 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 por 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 nº 1 (*Activated Sludge Model nº 1 - ASM1*), disponível na biblioteca de modelos *Modelica* no *software*

WEST. O modelo adotado para o decantador é baseado em uma extensão o modelo proposto por Takács, Patry e Nolasco (1991), em que é feita provisão para a estimativa do parâmetro de sedimentação por meio da medição do índice volumétrico de lodo (IVL) (DAIGGER; 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 3.500 mg/L e 4.800 mg/L. Mais detalhes descritivos sobre os modelos podem ser obtidos em *WEST Models Guide* (DHI, 2022).

Figura 3 – *Layouts* da ETE utilizada para obtenção dos dados no *software* WEST



Fonte: elaborado pelo autor.

Os dados de entrada do esgoto bruto foram retirados da base de dados do *software* WEST, que apresenta informações de vazão (Q , em m^3/h), demanda química de oxigênio (DQO, em $\text{mg O}_2/\text{L}$), demanda bioquímica de oxigênio (DBO, em $\text{mg O}_2/\text{L}$), 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 min. Dentre estas informações, foram aplicados como *input* para a simulação os dados de Q, DQO, NKT e SST.

Não obstante a disponibilidade de dados com discretização de 15 minutos, optou-se pela consolidação dos dados em base horária para a alimentação do modelo. Esta decisão metodológica justifica-se pela busca de maior verossimilhança com a prática operacional industrial, visto que, em cenários reais, o monitoramento e as tomadas de decisão de controle raramente ocorrem em intervalos sub-horários. Adicionalmente, essa abordagem considera a inércia hidráulica e biológica inerente ao sistema de tratamento, permitindo a mitigação de ruídos estocásticos de alta frequência que não impactam o desempenho macroscópico do processo.

Consequentemente, os resultados da simulação realizada no software WEST para o Layout 1 da ETE foram gerados e exportados com frequência horária, abrangendo um horizonte de 31 dias de operação. Tal procedimento resultou em um conjunto de 745 registros para as variáveis de entrada e saída da ETE.

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), também com volume constante e mistura ideal, modelado com biofilme unidimensional e 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; por sua vez, 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 O₂/L, para uma condição de NH_{min} = 1,5 mg/L, e 4 mg O₂/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).

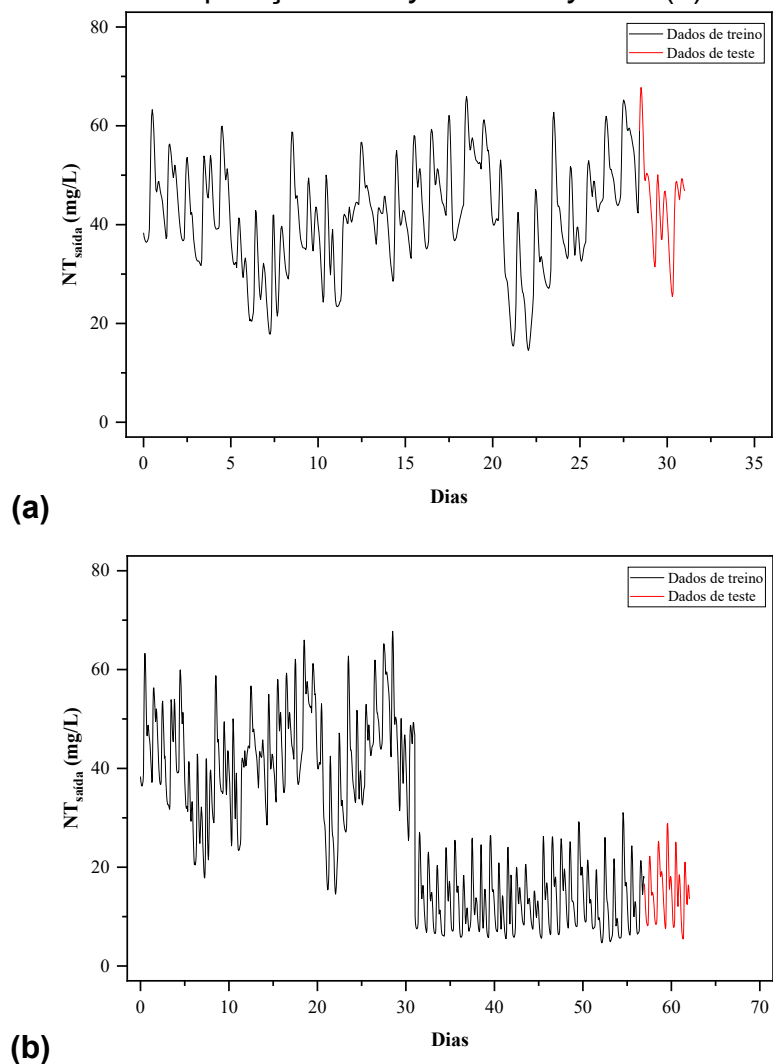
Os dados de entrada do esgoto bruto foram os mesmos utilizados na simulação anterior, com os dados de Q, DQO, NKT e SST aplicados como *input* para a

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 à 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 4 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 4a) e os 62 dias de operação do *Layout* 1 e *Layout* 2 (Figura 4b).

Figura 4 – 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: elaborado pelo autor.

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 o *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.

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, tal como já apresentado na Figura 1. 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.

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 5.

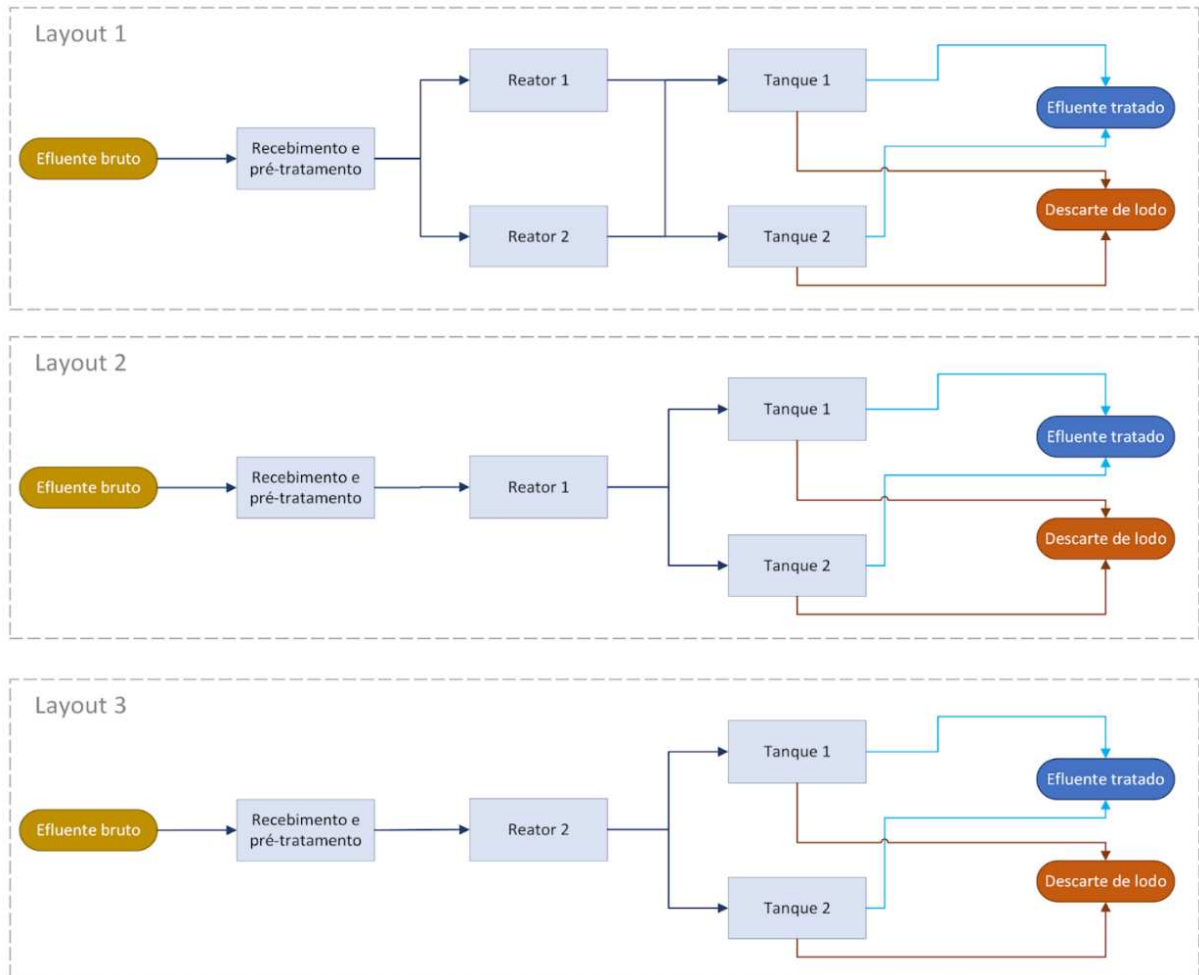
Nesse estudo, foram utilizados os parâmetros Q, DQO, carga orgânica (CO, em mg DQO/L), temperatura (T, em °C) e pH, coletados tanto na entrada da estação quanto em sua saída. Embora a base de dados originalmente fornecida pela empresa apresentasse mais de cem variáveis coletadas, muitos desses dados 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 6 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

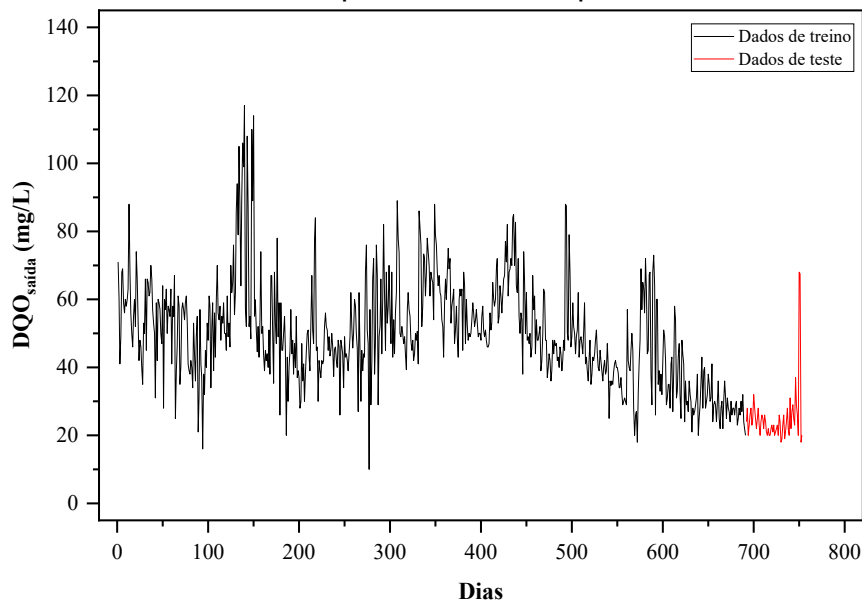
diminuição gradual pode ser atribuída às diversas modificações e melhorias implementadas na operação das Estações de Tratamento de Efluentes (ETE).

Figura 5 – Fluxograma simplificado da operação da ETE AMBEV (unidade Lages/SC) em função da operação dos tanques de aeração e reatores anaeróbios



Fonte: elaborado pelo autor.

Figura 6 – Perfil de saída do parâmetro DQO para a ETE AMBEV



Fonte: elaborado pelo autor.

4.3 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; 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 (BAKER, 2022).

Neste estudo, a remoção de dados anômalos e outliers foi realizada aplicando-se o critério de 3σ em relação à média (Zhu; Jiang; Feng, 2022), em que o desvio padrão σ foi determinado pelas Equações (1) e (2).

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

$$\eta_i = X_i - \bar{X} \quad (2)$$

sendo m o número de variáveis na coleta de informações, X_i um dado coletado, $\bar{X}$ é a média das informações e η_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 técnica 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* com apenas uma linha de operação, e o valor de 1 para os 31 dias subsequentes, correspondentes ao *Layout 2* 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 uma nova variável indicadora do modo de operação da estação, adicionando-se à base de dados três novas colunas. A primeira refere-se ao funcionamento do Reator 1, atribuindo valor 0 nos dias em que estava em operação e valor 1 nos dias em que estava desativado. A segunda coluna refere-se ao funcionamento do Reator 2, adotando a mesma lógica: valor 0 para dias de operação e 1 para dias de desativação. Por fim, a terceira coluna indica o funcionamento simultâneo dos dois reatores, recebendo valor 1 quando ambos estavam em operação e valor 0 quando pelo menos um deles não estava operando.

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; 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.

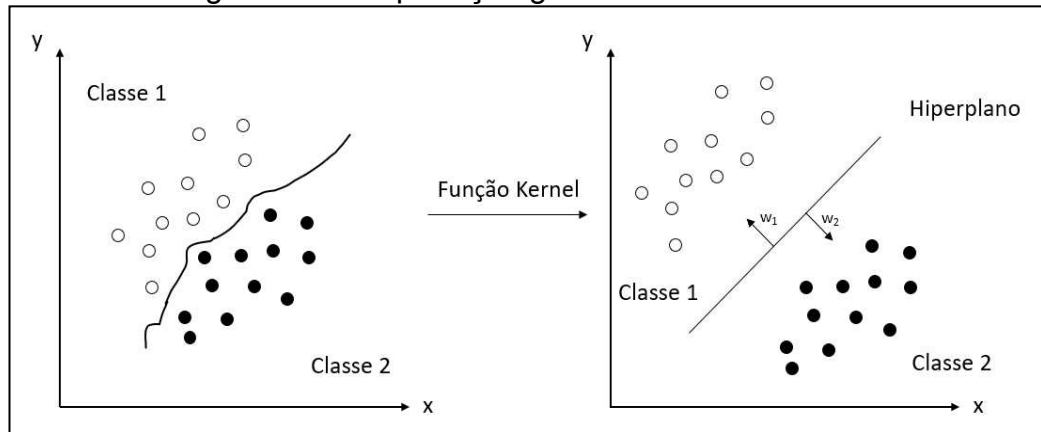
4.4 MODELOS DE MACHINE LEARNING

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 (BARNATHUNEK *et al.*, 2021), sendo baseados na minimização de risco estrutural, que reduz o *overfitting* e aumenta a generalização, minimizando o erro projetado do modelo de aprendizado (WU; 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 7 representa o processo de modelagem e deslocamento de dados em um espaço dimensional escolhido.

Figura 7 – Interpretação gráfica do método SVM



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

Os dados são representados como um conjunto de pontos em um espaço de características, e a solução consiste em encontrar o hiperplano (reta em 2D, plano em 3D, etc.) que maximize o *gap* – ou margem – entre as duas classes. Cada ponto contribui para essa separação por meio dos vetores de suporte. Contudo, há situações em que a separação linear não é possível no espaço original, sendo necessário utilizar funções kernel para transformar os dados em um espaço de maior dimensionalidade. Entre as funções kernel mais empregadas estão a de base radial (RBF), linear, gaussiana, polinomial e outras funções não lineares (PIRI *et al.*, 2015).

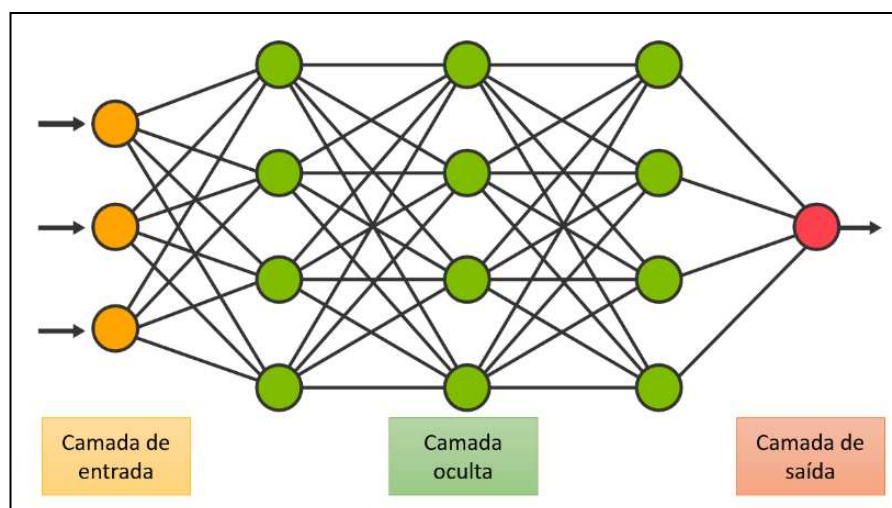
Para a construção do modelo SVM, utilizou-se a biblioteca *scikit-learn* (PEDREGOSA *et al.*, 2011) a partir de sua classe *svm.SVR*, a qual é referente ao modelo de suporte de vetor para regressões com *epsilon*. Foram testados diversos hiperparâmetros para a construção do modelo de forma a melhor ajustar 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; 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 a 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 8).

Figura 8 – Exemplo genérico de arquitetura de um modelo MLP



Fonte: elaborado pelo autor.

Na camada oculta, cada neurônio artificial contém uma função de ativação, linear ou não linear, e utiliza uma técnica de aprendizado supervisionado (PAL; 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; 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; 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 1.000, 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.

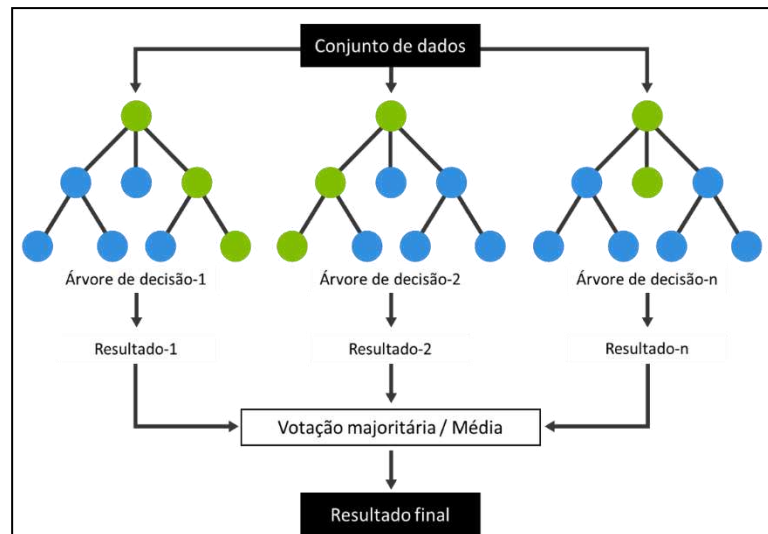
A função de ativação é o elemento chave de um neurônio artificial e, nesse estudo, a função de ativação ReLU foi usada para todos os neurônios (FUKUSHIMA, 1975). O tamanho da batelada utilizado nesse estudo foi de 200 amostras. Esse é um importante hiperparâmetro para o treinamento da rede neural, visto 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 RF 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 9).

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 RF é a possibilidade de determinar, de forma intuitiva, a importância das variáveis em problemas de regressão ou classificação. Essa importância é calculada a partir da redução da impureza em cada nó utilizado no particionamento dos dados (FLORES; LEIVA, 2021). Além disso, a RF 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).

Figura 9 – Exemplo genérico de arquitetura de um modelo de Floresta Aleatória



Fonte: elaborado pelo autor.

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; 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 (BUTINCK *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 do erro quadrático médio (RMSE), erro médio absoluto (MAE), erro percentual absoluto médio (MAPE) e coeficiente de determinação (R^2), apresentadas nas Equações (3)-(6), foram escolhidas como função de perda em todos os modelos de previsão considerados e são descritos por Ly *et al.* (2022).

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

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

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

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

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 $\bar{Y}$ 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.

4.5 INTERPRETABILIDADE DE MODELOS - *FEATURE IMPORTANCE*

Neste estudo, foram utilizadas técnicas de *Partial Dependence Plots* (PDPs) e *Permutation Importance* (PI) para determinar a importância das variáveis de entrada na previsão das variáveis de saída e para entender como ocorre a relação entre cada variável utilizada no modelo preditivo.

A *Permutation Importance* (PI) é uma técnica simples e relativamente confiável que mede diretamente a importância de cada variável, observando o efeito na precisão do modelo ao embaralhar aleatoriamente cada variável preditora (BREIMAN, 2001; MOLNAR, 2020). Uma das vantagens desta técnica é que ela é independente do modelo e pode ser calculada quantas vezes forem necessárias com diferentes permutações para obter uma estimativa razoável da relevância real de cada variável. A construção algorítmica dos modelos PI foi realizada utilizando a função *permutation_importance* contida no módulo de inspeção da biblioteca *scikit-learn* (PEDREGOSA *et al.*, 2011; SCIKIT-LEARN DEVELOPERS, 2023a). Para cada variável foram realizadas 30 réplicas para garantir maior cobertura das permutações do modelo e mitigar os efeitos de mudanças mais significativas. Todas as avaliações foram realizadas utilizando os modelos previamente treinados e dados de entrada e

saída normalizados. Os valores de importância atribuídos às variáveis foram dados em escala percentual para que a soma da significância de todas as variáveis para cada um dos modelos de ML seja igual a 100%.

Os gráficos de dependência parcial, outra técnica de FI utilizada neste estudo, permitem avaliar o impacto de uma determinada variável de entrada nas previsões de um modelo através da construção e análise de gráficos de dependência parcial obtidos em situações em que os efeitos das outras variáveis são controlados (MOLNAR, 2020). Esses gráficos de dependência parcial podem ser usados para visualizar e analisar a interação entre a resposta alvo e todas as variáveis de entrada (HASTIE; FRIEDMAN; TIBSHIRANI, 2001). O gráfico também fornece informações sobre as relações entre as variáveis, indicando se a dependência parcial é linear, monotônica ou mais complexa (JOHNSON *et al.*, 2022). Os modelos PDP foram construídos usando o método *from_estimator* da classe *PartialDependenceDisplay* no módulo de inspeção da biblioteca *scikit-learn* (SCIKIT-LEARN DEVELOPERS, 2023b).

4.6 RESULTADOS OBTIDOS PARA A ESTAÇÃO DE TRATAMENTO DE ESGOTO SIMULADA (ETE WEST)

Para o estudo realizado na ETE WEST, o banco de dados foi organizado em variáveis independentes (DBO, DQO, NH, NO, Q, NKT, NT e SST do afluente) e uma variável dependente (NT do efluente). A concentração de nitrogênio total do efluente (NT_{saída}) foi definida como parâmetro de predição por apresentar a maior variância (276,60 mg²/L²) que as demais variáveis de qualidade do efluente, como DQO (75,72 mg²/L²). Com base nesta definição, os modelos preditivos de concentração de NT_{saída} foram aplicados em três cenários distintos:

- a) O Cenário 1 considera apenas os dados da simulação realizada para o *Layout* 1 da Figura 3. A base de dados utilizada teve 745 medições horárias, totalizando 31 dias de operação. A partir desses registros, as primeiras 683 entradas para todas as variáveis foram definidas como conjunto de dados de treinamento dos modelos de predição de NT_{saída}. As últimas 62 entradas foram definidas como conjunto de dados de teste para os modelos. O conjunto de

dados utilizado para testar e treinar os modelos de ML é mostrado graficamente na Figura 4a.

- b) No Cenário 2, a base de dados utilizada consiste na combinação dos dados obtidos na simulação realizada para o *Layout 1* (Figura 3) e os dados obtidos na simulação realizada para o *Layout 2* (Figura 3), perfazendo um total de 62 dias de operação (31 dias consecutivos de operação do *Layout 1* e 31 dias consecutivos de operação do *Layout 2*). O conjunto de dados de treinamento e teste foi definido mantendo a relação teste/treinamento adotada no Cenário 1 (teste/treinamento = 0,09). Assim, as primeiras 1.367 entradas para todas as variáveis foram definidas como o conjunto de dados de treinamento, e as últimas 123 entradas como o conjunto de dados de teste para os modelos de previsão de $NT_{saída}$. O conjunto de dados usado para testar e treinar os modelos de ML é mostrado graficamente na Figura 4b.
- c) No Cenário 3, foi mantida a base de dados utilizada no Cenário 2, mas com a informação adicional da variável categórica para identificar o modo de operação da ETE. Conforme mencionado acima, esta informação tem como objetivo ajudar os modelos a reconhecerem um novo padrão operacional e sua relação com os dados de entrada e saída.

Os dados de $NT_{saída}$ previstos foram comparados com os valores reais de $NT_{saída}$ do intervalo de teste, e a precisão de cada modelo foi avaliada a partir do desempenho apresentado na Tabela 2. Uma análise gráfica comparativa entre os dados previstos pelos modelos e os dados obtidos pela simulação da ETE WETS também é mostrado na Figura 10.

Tabela 2 – Comparação de métricas de performance para os modelos preditivos de ML para $NT_{saída}$ aplicados a ETE WEST

(continua)

	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

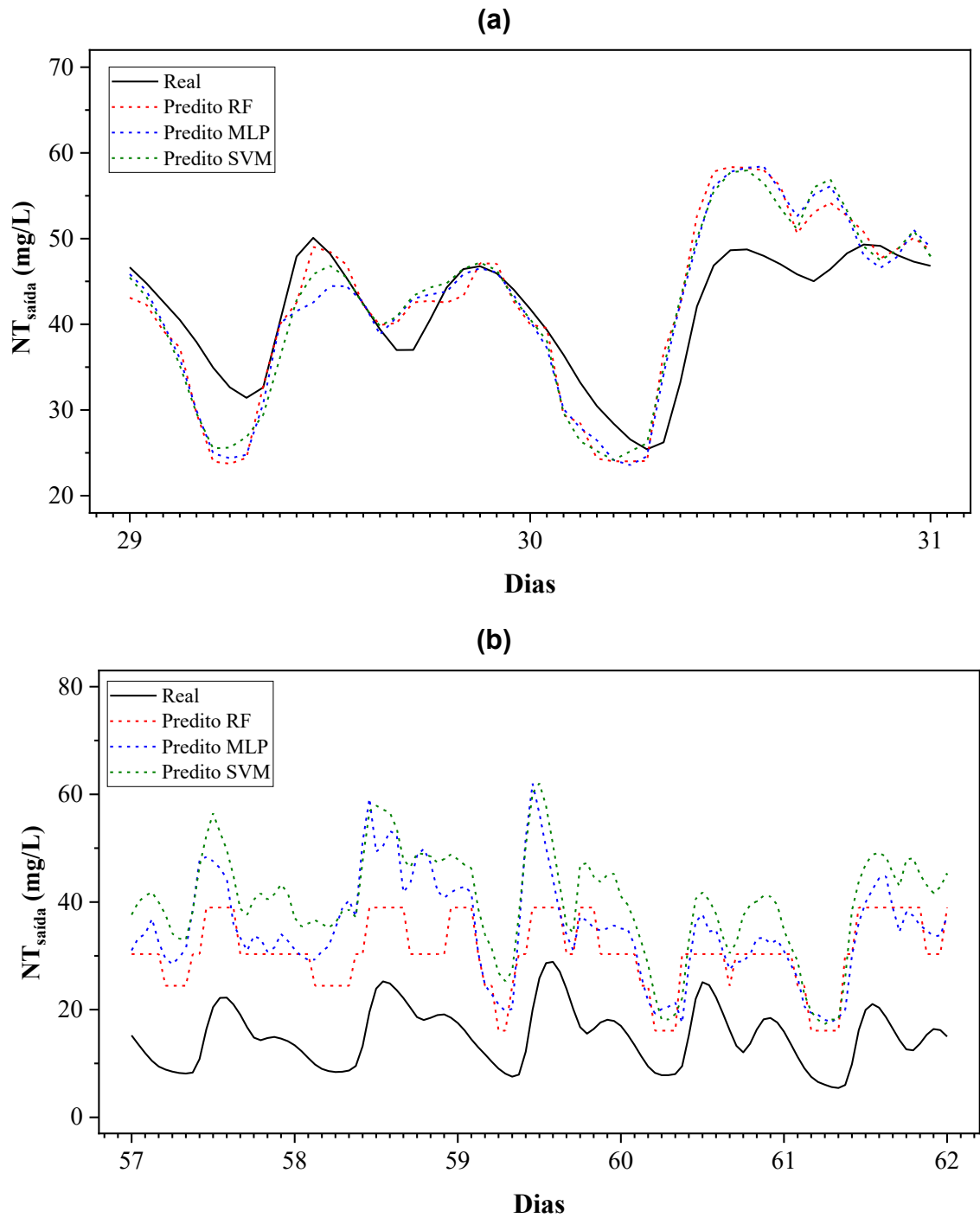
Tabela 2 - Comparação de métricas de performance para os modelos preditivos de ML para NT_{saída} aplicados a ETE WEST
(conclusão)

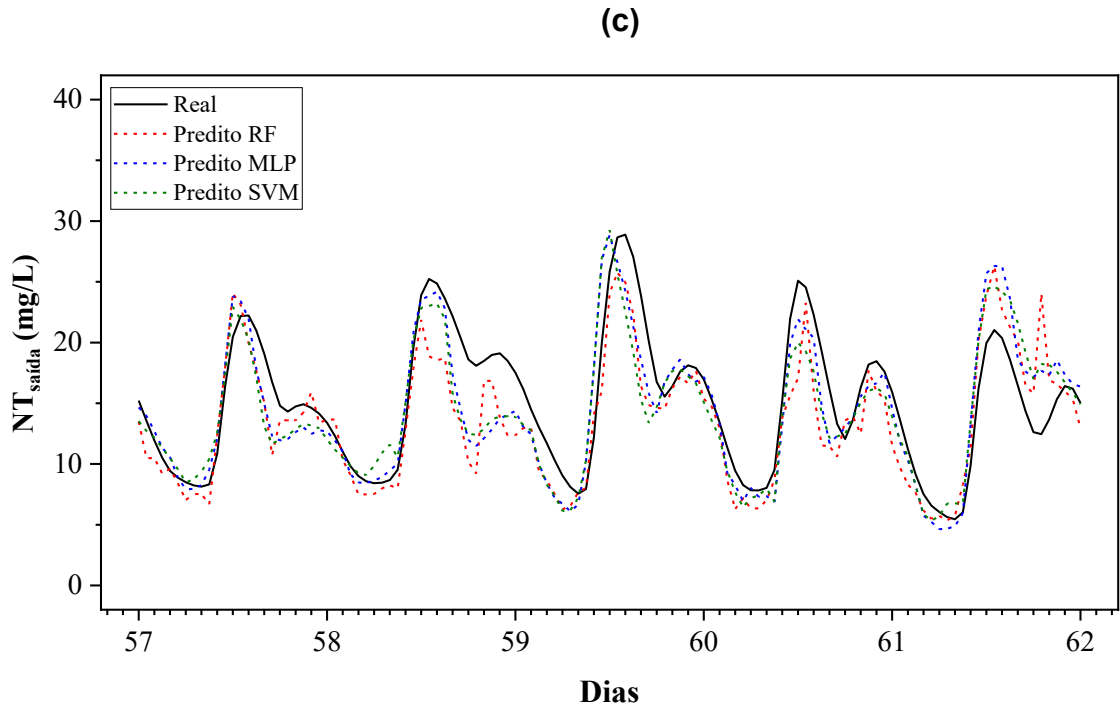
MAE	4,52	4,48	4,61	24,96	19,85	15,58	2,49	2,21	2,55
MAPE	0,11	0,11	0,11	1,85	1,5	1,2	0,16	0,14	0,16
R ²	0,6	0,6	0,58	-21,44	-13,82	-7,89	0,66	0,72	0,61

Fonte: elaborado pelo autor.

Para o Cenário 1, existe uma grande similaridade entre as métricas obtidas para as previsões dos três modelos, o que também é observado 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 feita pelo modelo RF, ou seja, as estimativas feitas pelo modelo RF excedem o resultado real em uma proporção mais significativa do que a produzida pelos demais modelos. Isto é confirmado pelo maior valor de RMSE para este modelo, pois essa métrica de desempenho penaliza valores mais distantes do que as outras métricas. Os resultados de predição para os modelos SVM e MLP foram semelhantes para este primeiro cenário e ambos podem ser utilizados como modelos de predição para este sistema. Embora o modelo MLP tenha apresentado métricas de desempenho ligeiramente melhores, o modelo SVM tem menor interpretabilidade (PHILLIPS *et al.*, 2021) e esta falta de compreensão explícita do que está acontecendo dentro de um modelo de caixa preta (LOYOLA-GONZALEZ, 2019) pode desencorajar seu uso pelos usuários mais críticos desses métodos.

Figura 10 – Resultados dos modelos preditivos para os dados de $NT_{saída}$ da ETE WEST usando RF, MLP e SVM: Cenários (a) 1, (b) 2 e (c) 3





Fonte: elaborado pelo autor.

Ao analisar os resultados obtidos para o Cenário 2, fica evidente a perda de qualidade dos modelos de previsão. Isso se deve ao banco de dados utilizado para treinamento, que apresenta maiores valores de NT_{saída} do efluente nas primeiras 745 horas de registro. Mesmo utilizando os 683 dados de treinamento com os menores valores de NT_{saída} do efluente na base de dados, os modelos foram induzidos a prever valores superiores aos observados para os dados nas últimas horas de treinamento. Para a base de dados utilizada, os modelos preditivos não conseguiram identificar a mudança no comportamento de saída da concentração de NT_{saída} do efluente causada pela diferença de operação.

Como todos os modelos previram valores de NT_{saída} do efluente superiores aos valores simulados para o *Layout 2*, as métricas de desempenho apresentaram resultados muito superiores aos obtidos para o Cenário 1, indicando uma baixa precisão dos modelos. O modelo RF teve melhor desempenho neste cenário. Por se tratar de um algoritmo *ensemble*, vários avaliadores - regressores neste trabalho - são criados para fazer uma estimativa global com base em todas as previsões feitas por cada avaliador que compõem o método *ensemble*. A maioria das árvores de decisão foram desenvolvidas para prever com qualidade os dados da primeira parte do treinamento. No entanto, algumas árvores aproximaram as previsões dos últimos

683 valores do banco de dados, com valores mais próximos dos valores de teste. Este procedimento reduziu o valor da previsão final, que é uma média das previsões de todas as árvores de decisão, em comparação com os demais modelos. Assim, o modelo RF é mais adequado ao Cenário 2 entre os modelos de predição estudados, o que fica claro pelas métricas e pela observação do gráfico de predição comparado ao valor real (Figura 10, Cenário 2).

Geralmente são esperadas mudanças nas operações das ETEs para melhorar a eficiência dos processos, o que representa um desafio adicional significativo na aplicação de metodologias de ML para prever resultados de qualidade de efluentes e implementar processos baseados em inteligência artificial. Para mitigar o impacto negativo de uma mudança de processo na previsão dos modelos, o Cenário 3 foi criado e investigado. Neste cenário, a informação de qual *layout* da ETE estava em funcionamento foi adicionada ao banco de dados utilizado no modelo a partir de uma nova variável artificial.

Houve uma melhoria significativa nas métricas de desempenho em comparação aos dois cenários anteriores. A exceção é o MAPE, que apresentou valores de 0,14 e 0,16 (ou 14% e 16%), ligeiramente superiores aos obtidos para o Cenário 1, que foi de 11% para todos os modelos. A razão para este resultado pode ser explicada pelo fato de que os valores de $NT_{saída}$ previstos neste cenário ($NT_{médio} = 15,00$ mg/L) são inferiores aos valores de $NT_{saída}$ previstos no Cenário 1 ($NT_{médio} = 44,35$ mg/L) e, portanto, são mais sensíveis a erros percentuais. Os três modelos de predição acompanharam satisfatoriamente a tendência dos eventos, conforme Figura 10 - Cenário 3, e apresentaram valores de R^2 acima de 0,6 para predição dos dados. O resultado do modelo MLP se destaca, oferecendo as métricas mais adequadas. A escolha adequada dos hiperparâmetros pode ter permitido uma melhor aproximação da relação entre os dados NT de entrada e os dados NT de saída. Esta aplicação está alinhada com o Teorema da Aproximação Universal (CSÁJI, 2001b; HORNIK; STINCHCOMBE; WHITE, 1989; LU *et al.*, 2017).

4.7 RESULTADOS OBTIDOS PARA A ESTAÇÃO DE TRATAMENTO DE EFLUENTES INDUSTRIAIS (ETE AMBEV)

Para a ETE AMBEV, conforme descrito anteriormente, todas as informações utilizadas no banco de dados foram provenientes de anotações dos operadores da ETE no sistema de gerenciamento de processos (LiveMES) e de planilhas de controle interno das operações. Como se trata de uma ETE real, vários problemas podem surgir na anotação, armazenamento e partilha de dados. Portanto, foi necessário realizar um pré-processamento rigoroso dos dados. O banco de dados continha valores discrepantes e valores sem significado físico para parâmetros específicos. Em situações reais de anotação de dados, especialmente quando o registro é feito manualmente por diferentes operadores, não é surpreendente encontrar tais inconsistências, muitas vezes devido a erros de medição e erros humanos. Utilizando a análise univariada, a exclusão de dados correspondeu a 0,76% da base de dados original, equivalente a 69 dos 9.120 dados registrados na base de dados inicial. Todas as informações faltantes foram imputadas usando a técnica KNN.

O banco de dados completo foi, então, separado em variáveis independentes (dados de Q, DQO, CO, T e pH do afluente) e variáveis dependentes ($DQO_{saída}$). A partir dos 752 registros diários de cada variável, os primeiros 690 dias foram definidos como conjunto de dados de treinamento dos modelos de previsão de concentração de DQO em efluentes ($DQO_{saída}$), e os últimos 62 dias como conjunto de dados de teste dos modelos, conforme a Figura 6. Três cenários diferentes foram aplicados aos modelos de previsão para $DQO_{saída}$:

- a) O Cenário 1 considera os dados de Q, DQO, CO, T e pH do afluente para esta ETE como variáveis independentes para prever a $DQO_{saída}$.
- b) No Cenário 2, as informações referentes à variação sazonal e ao modo de operação da ETE (definido com base na operação dos Reatores 1 e 2) foram adicionadas ao banco de dados utilizado pelo modelo preditivo como variáveis independentes adicionais.
- c) Para o Cenário 3, os dados de Q, DQO, CO, T e pH efluentes da ETE do dia anterior foram adicionados como variáveis independentes para prever a DQO do efluente. Como o tempo de residência da ETE é de até 7 dias, é provável

que o efluente de um determinado i -ésimo dia tenha características muito semelhantes ao efluente do $i-1$ dia. A inclusão de informação sobre a qualidade do efluente do dia anterior pode, portanto, melhorar o desempenho dos modelos.

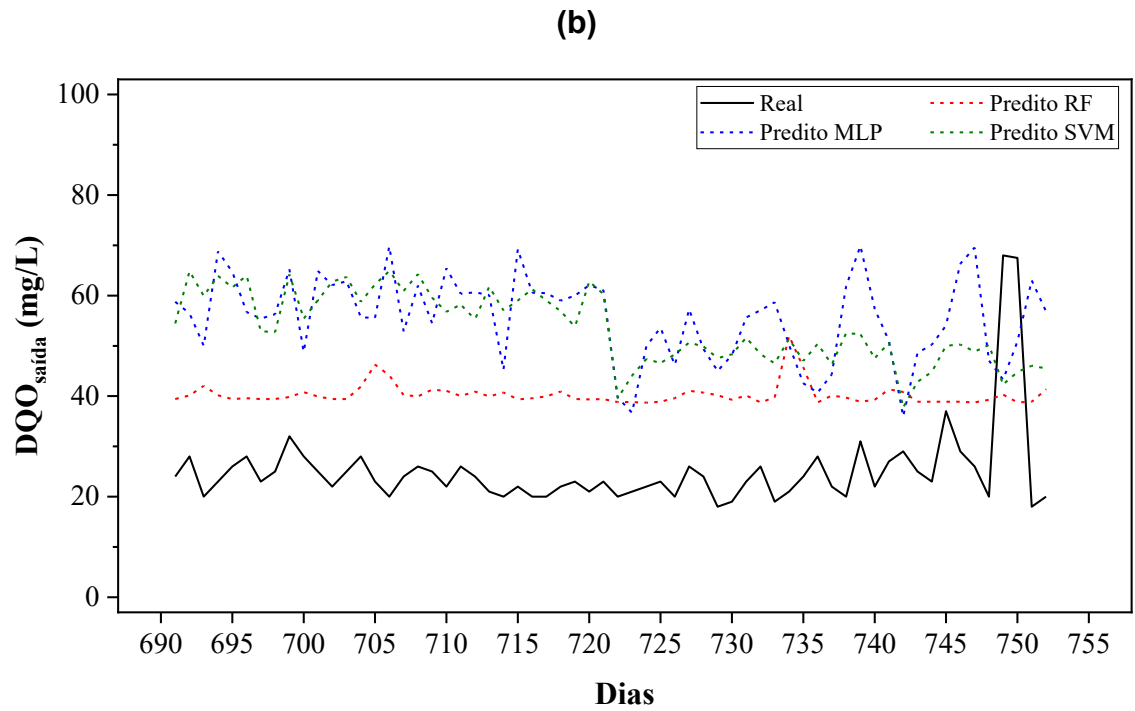
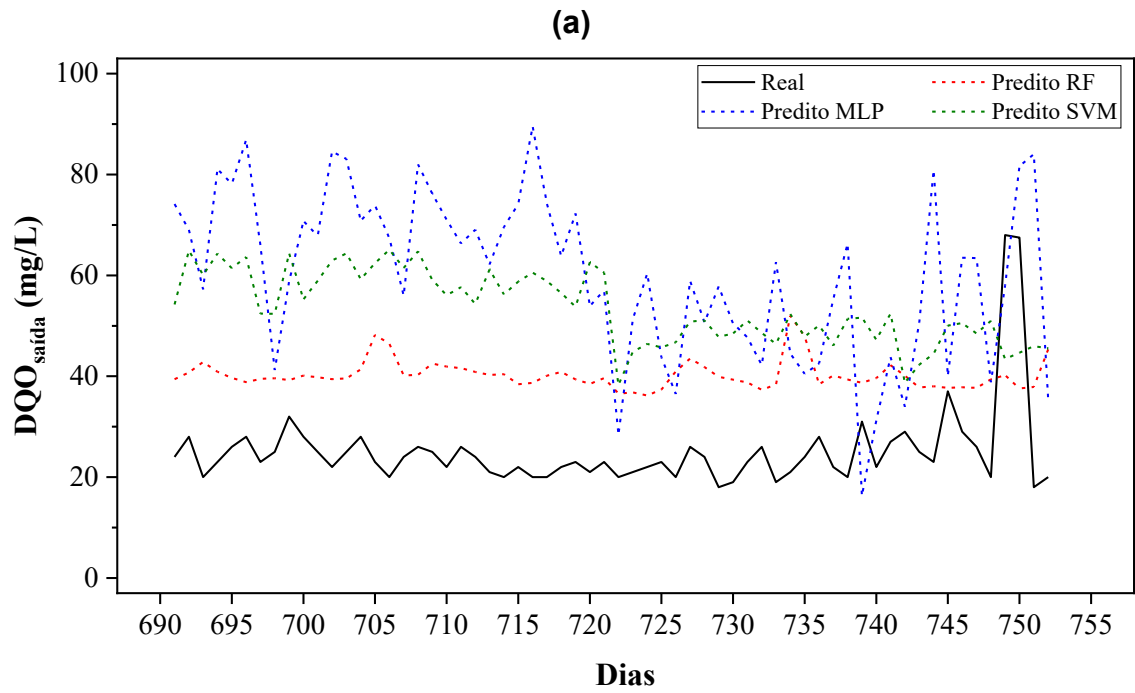
Os dados de $DQO_{saída}$ previstos foram comparados com os valores reais de $DQO_{saída}$ do efluente do intervalo de teste, e a precisão de cada modelo foi avaliada usando as métricas de desempenho apresentadas na Tabela 3. Uma análise gráfica comparativa entre os dados previstos pelos modelos e os dados medidos em a ETE AMBEV também é apresentada na Figura 11.

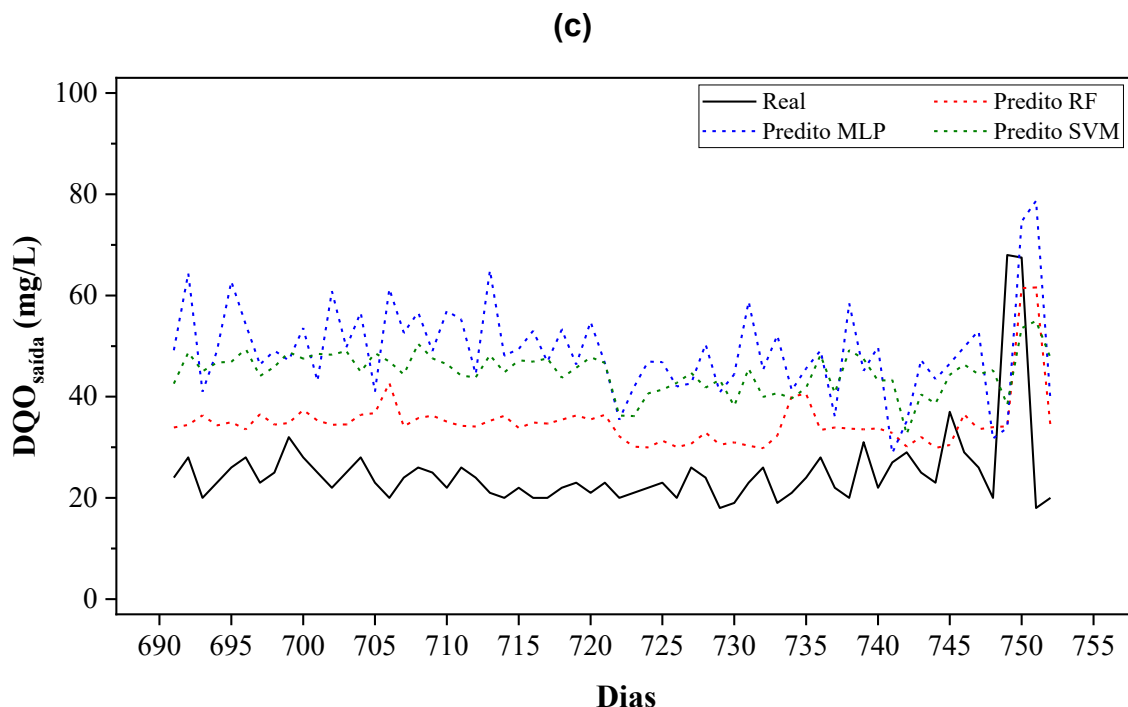
Tabela 3 – Comparação de métricas de performance para os modelos preditivos de ML para $DQO_{saída}$ aplicados a ETE AMBEV

	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
R ²	-11,97	-20,21	-3,30	-12,04	-13,83	-3,24	-5,44	-8,80	-1,33

Fonte: elaborado pelo autor.

Figura 11 – Resultados dos modelos preditivos para os dados de DQO_{saída} da ETE AMBEV usando RF, MLP e SVM: Cenários (a) 1, (b) 2 e (c) 3





Fonte: elaborado pelo autor.

No Cenário 1 pode-se observar que os valores de DQO do efluente previstos pelos três modelos de ML foram significativamente diferentes dos valores reais. As razões para este resultado incluem o uso de uma quantidade insuficiente de parâmetros de processo que são coletados diariamente como preditores para os modelos de ML. Parâmetros como sólidos suspensos totais, nitrogênio total e fósforo total são considerados importantes para prever a concentração de DQO (ARISMENDY *et al.*, 2021; FERNANDEZ DE CANETE *et al.*, 2016; SHARAFATI; ASADOLLAH; HOSSEINZADEH, 2020; SZELAĞ *et al.*, 2020; XU *et al.*, 2022b). No entanto, como esses parâmetros não são medidos diariamente pelos operadores da estação, eles não puderam ser utilizados neste estudo. A ausência dessas informações é um ponto relevante. Caso houvesse um conjunto de dados mais abrangente, incluindo variáveis de processo com forte correlação com a variável-alvo (neste caso, DQO_{saída}), seria possível obter análises mais aprofundadas e resultados mais precisos. Com um número mais significativo de parâmetros disponíveis, seria possível compreender melhor a relação e a importância de cada variável para prever os parâmetros de qualidade da ETE. Isto permitiria uma análise abrangente e aprofundada do impacto de cada variável na previsão dos parâmetros do processo e

das características dos efluentes, o que poderia ser útil para os processos de tomada de decisão relativos à operação da estação.

Para o cenário 2, onde três novas variáveis indicando diferentes *layouts* de plantas (Figura 5) foram adicionadas aos dados de treinamento, foi observada uma melhoria sutil no desempenho dos modelos, como observado na ETE WEST. Os valores previstos para DQO dos efluentes estão mais próximos dos valores reais em comparação com o primeiro cenário, mas todos os modelos previram valores acima dos níveis reais. Isso pode ser resultado de outras alterações na operação da planta que não estão diretamente relacionadas aos reatores anaeróbios, mas a melhorias operacionais não consideradas neste estudo, como a instalação de novos sensores ou a utilização de produtos químicos de diferentes fontes.

Para compensar a falta de outros parâmetros preditores e informações relacionadas às mudanças operacionais, o Cenário 3 incluiu outras medições do efluente tratado do dia anterior para prever a DQO do efluente. Foi possível observar uma melhoria no desempenho dos modelos em comparação com outros cenários, uma vez que mais informações sobre os parâmetros de qualidade foram fornecidas aos algoritmos de ML. No entanto, ainda existe um grande potencial para novos estudos que melhorem as previsões destes modelos.

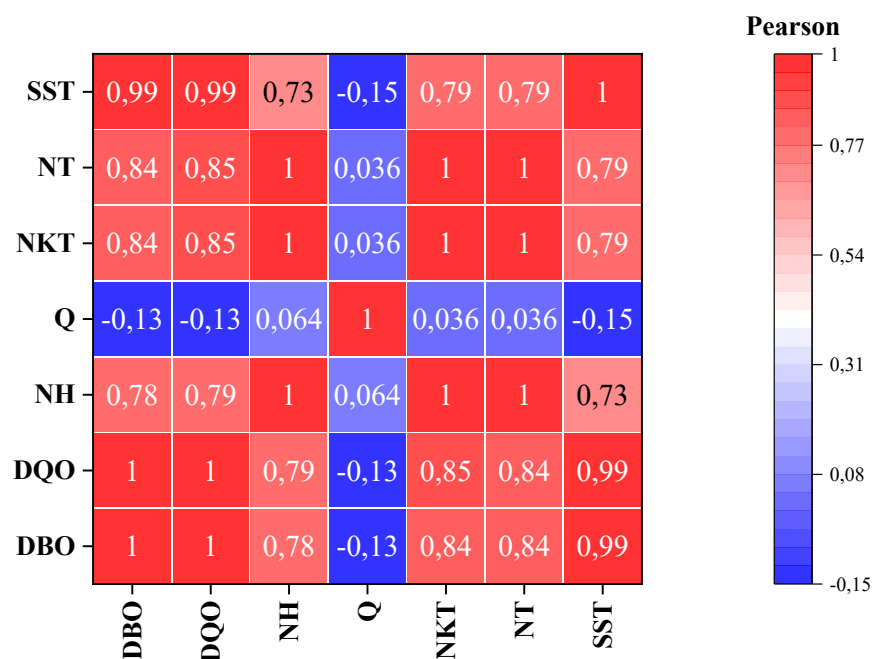
Em todos os cenários aplicados à ETE AMBEV, o modelo RF foi o que apresentou as melhores métricas de desempenho e foi capaz de prever valores mais próximos dos reais. As mesmas razões explicadas para o caso da ETE WEST podem ser aplicadas a esta. Apesar dos erros consideráveis observados nas previsões do modelo RF, é possível verificar que as tendências de comportamento da variável foram capturadas pelo modelo, incluindo o pico mais elevado, embora com um ligeiro atraso. Devido à grande discrepância entre as previsões e as observações reais, os modelos preditivos não puderam ser utilizados para apoiar a tomada de decisões ou outros processos que utilizem tecnologias de inteligência artificial neste caso. No entanto, existem oportunidades para melhorar os modelos preditivos e os seus resultados se mais informações sobre a operação e o processo puderem ser fornecidas.

4.8 RESULTADOS OBTIDOS PARA A INTERPRETABILIDADE DE MODELOS DE ML - *FEATURE IMPORTANCE* (FI)

Considerando os resultados insatisfatórios obtidos com os modelos preditivos implementados na ETE AMBEV e considerando que esses modelos preditivos não poderiam ser utilizados para apoiar a tomada de decisão ou outros processos que envolvessem tecnologias de inteligência artificial para este caso, optou-se por não realizar a análise de *Feature Importance* (FI) para este caso. A aplicação de tal análise a modelos de ML com previsões desfavoráveis pode levar a simplificações indevidas e a conclusões errôneas, dada a inadequação dos algoritmos propostos para compreender eficazmente as variações comportamentais inerentes a este sistema de tratamento de águas residuais. Dado o desempenho superior destes modelos aplicados aos dados da ETE WEST, optou-se por focar os estudos de FI exclusivamente neste caso.

Uma análise de correlação de Pearson entre as variáveis independentes foi realizada para ambos os cenários utilizando a biblioteca Python pandas (REBACK *et al.*, 2021) para avaliar a relação linear e quantificar o grau de associação entre duas variáveis (Figura 12).

Figura 12 – Matriz de correlação de Pearson para as variáveis de entrada da ETE WEST

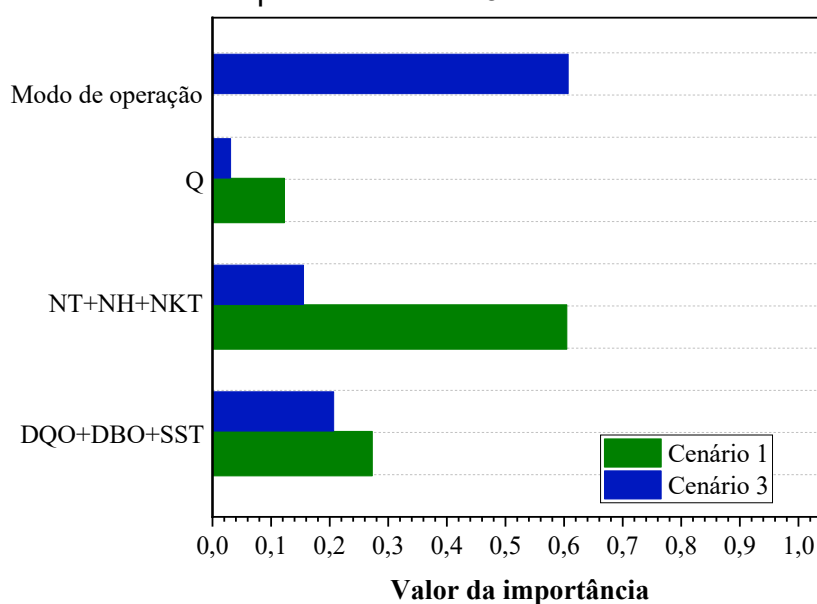


Fonte: elaborado pelo autor.

Devido à alta correlação observada entre algumas das variáveis independentes utilizadas no modelo, ou seja, sua multicolinearidade, as importâncias atribuídas às variáveis DBO, DQO e SST foram agregadas e apresentadas em conjunto para a análise de importância de permutação. O mesmo procedimento foi aplicado para a importância das variáveis NH, NKT e NT. O procedimento de agregação de importância foi necessário porque essas variáveis apresentavam correlações de Pearson superiores a 0,99, o que poderia comprometer a estimativa da natureza das relações entre as variáveis (FARRAR; GLAUBER, 1967; KALNINS, 2018). A variável de entrada NO, por outro lado, apresentou correlação zero com todas as variáveis. Isto se deve à natureza dos dados de entrada utilizados no modelo, que são iguais e constantes ao longo do período analisado.

A Figura 13 apresenta a relação entre a importância das variáveis independentes e a resposta da variável dependente ($NT_{saída}$) para os diferentes cenários de operação da ETE WEST avaliados neste estudo. A resposta da importância das variáveis foi avaliada apenas para o modelo MLP, que apresentou os melhores resultados de previsão entre os três modelos de ML propostos.

Figura 13 – Importância das variáveis de entrada na previsão de $NT_{saída}$ nos cenários operacionais 1 e 3 da ETE WEST



Fonte: elaborado pelo autor.

A importância, representada pelas barras horizontais na Figura 13, mostra o efeito na precisão do modelo de embaralhar aleatoriamente o valor de uma variável preditora (SCIKIT-LEARN DEVELOPERS, 2023c). Para o Cenário 1 pode-se observar que os parâmetros de entrada relacionados ao teor de nitrogênio influente (NT, NH e NKT) se destacam em relação aos demais parâmetros. Percebeu-se que esses parâmetros têm aproximadamente 2,2 vezes mais influência na previsão do NT do efluente do que os parâmetros (DQO, DBO e SST) que possuem a segunda maior importância de permutação. Por fim, a vazão afluente teve o menor valor de importância entre as variáveis investigadas. Esses resultados sugerem que as variáveis (NT, NH e NKT) são os fatores dominantes na previsão do NT do efluente, enquanto as demais características têm menor influência no desempenho do modelo.

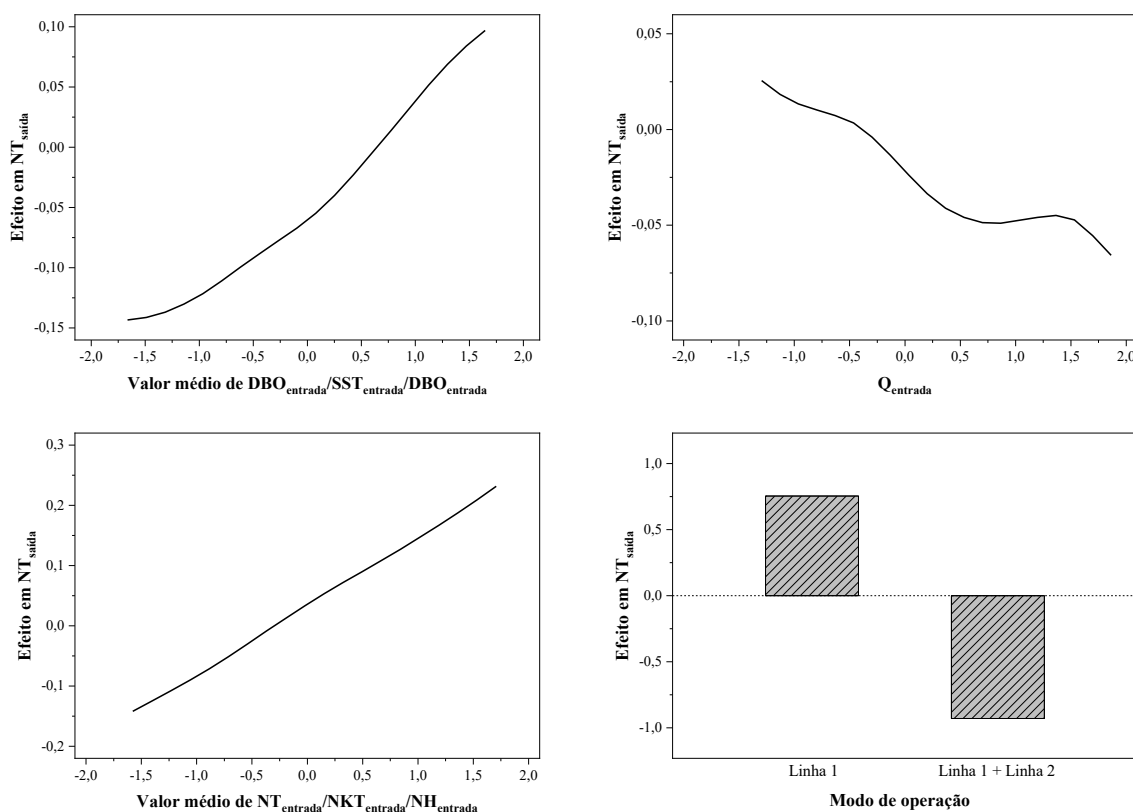
De acordo com a *Environmental Protection Agency* (EPA, 2009) o nitrogênio na água é geralmente medido em três formas: amônia, nitratos e nitritos. O nitrogênio total é, portanto, a soma das concentrações de nitrato, nitrito, amônia e nitrogênio orgânico em uma determinada amostra. Deve-se notar que o nitrogênio Kjeldahl total é a soma das concentrações de amônia e nitrogênio orgânico, mas não inclui nitrato ou nitrito.

O nitrogênio presente nas águas residuais municipais está geralmente na forma de nitrogênio orgânico e/ou amônia. Durante o processo de tratamento de água nas ETEs, o nitrogênio orgânico é convertido em suas formas inorgânicas (nitrito e nitrato) por um processo químico-biológico denominado nitrificação (EPA, 2023). A nitrificação ocorre basicamente em duas etapas: a conversão da amônia em nitrito seguida pela oxidação do nitrito (NO_2^-) em nitrato (NO_3^-). Isso é conhecido como nitritação e depois nitratação. Os nitratos, por sua vez, são reduzidos a nitrogênio gasoso por desnitrificação. Em resumo, o conteúdo afluente de nitrogênio é convertido em suas frações inorgânicas ao longo dos processos de tratamento de águas residuais, resultando em baixas concentrações de nitrogênio no lodo e no efluente tratado (HIET WONG; BARTON; BARFORD, 2003). É, portanto, compreensível que o teor de nitrogênio – seja na forma de NKT ou NH – seja um parâmetro de significativa importância para a previsão do modelo no Cenário 1. Vale ressaltar que, embora tenha sido considerado no estudo, o NO influente considerado na ETE simulada foi praticamente insignificante. Devido a este fato, os afluentes NKT e NT ficaram com valores muito semelhantes, o que justifica valores de correlação semelhantes.

O nível de matéria orgânica nas águas residuais também é um fator importante para o processo de nitrificação-desnitrificação, o que confirma a importância demonstrada pela análise de importância. Os compostos orgânicos atuam como doadores de elétrons nas reações de desnitrificação. Estequiometricamente, são necessários 3 g a 5 g de DQO para cada grama de NO_3^- reduzido (HENZE *et al.*, 2001). Portanto, conhecer o teor de matéria orgânica nas águas residuárias é essencial para avaliar se haverá uma conversão efetiva da espécie em nitrogênio gasoso (HIET WONG; BARTON; BARFORD, 2003; HOLMES; DANG; SMITH, 2019).

No Cenário 3, a informação sobre o modo de operação da ETE é definitivamente a variável mais importante. Assim, a previsão do modelo é mais precisa neste cenário com mais dias de operação. A interpretação dos dados do PI pode ser complementada pela análise dos gráficos de dependência parcial mostrados na Figura 14.

Figura 14 – Gráficos de dependência parcial mostrando a média dos efeitos do influente DQO, DBO e SST (canto superior esquerdo), o efeito da vazão do influente (canto superior direito), a média dos efeitos do influente NKT, NT e NH (canto inferior esquerdo) e indicação do modo de funcionamento da ETE (canto inferior direito) na previsão do $\text{NT}_{\text{saída}}$ em escala padronizada



Fonte: elaborado pelo autor.

Para entender melhor como analisar este gráfico, o eixo X representa os diferentes valores possíveis da variável de entrada, que neste estudo são apresentados de forma normalizada. O eixo Y representa a média da saída do modelo para a variável, dadas as diferentes instâncias ou valores dessa variável de entrada. O gráfico exibe uma linha reta ou curva que mostra como a produção média varia conforme a variável de entrada muda. Esta linha representa a relação média entre a variável de entrada e a saída, mantendo todas as outras variáveis constantes. Uma linha horizontal indicaria que a variável de entrada não tem efeito sobre a saída, enquanto uma inclinação positiva ou negativa indicaria um efeito positivo ou negativo sobre a saída. Pela leitura do gráfico é possível observar como a variável de entrada afeta a saída do modelo. Se a linha tiver inclinação ascendente, significa que à medida que a variável de entrada aumenta, a produção tende a aumentar. Se for descendente, significa que à medida que a variável de entrada aumenta, a produção tende a diminuir.

Pode-se observar que o PDP indica que valores mais baixos de entrada de DQO, DBO e SST levam a valores mais baixos de $NT_{saída}$, com uma aparente relação exponencial. Embora o valor mais apropriado para a relação entre o teor de carbono e nitrogênio dependa da natureza da fonte externa de carbono, um aumento na relação DQO:N pode afetar a etapa de nitrificação, reduzindo a eficiência de remoção de nitrogênio (ZOU *et al.*, 2012). Por sua vez, observa-se uma relação linear no PDP obtido para as espécies de nitrogênio $NT_{entrada}$, $NKT_{entrada}$ e $NH_{entrada}$, indicando que valores mais baixos desses parâmetros induzem uma menor predição de $NT_{saída}$, o que concorda com o resultado apresentado na análise de PI.

Para a vazão, o PDP implica que valores mais baixos de Q mostram um maior teor de nitrogênio no efluente. Contudo, os PDPs não devem ser analisados isoladamente, mas em conjunto com outros parâmetros. Vazões mais altas podem implicar níveis mais elevados das demais variáveis de qualidade do afluente, o que contribui para aumentar o $NT_{saída}$ previsto. Por fim, em relação ao modo de operação, a barra positiva obtida para a operação da ETE com apenas uma linha (Figura 3, *Layout 1*) indica que o valor do $NT_{saída}$ deverá ser, em média, 0,753 superior ao valor médio do $NT_{saída}$ da base de treinamento após a padronização. Quando as duas linhas estão em operação (Figura 3, *Layout 2*), o valor previsto de $NT_{saída}$ tende a ser em média 0,928 menor.

A informação apresentada sobre as variáveis operacionais que afetam os níveis de NT do efluente, em conjunto com análises convencionais (químicas, bioquímicas e

hidromecânicas), pode ajudar os operadores de ETEs a compreender de forma mais eficiente e fiável as relações intrínsecas que influenciam processos complexos de tratamento de águas residuais. Isto pode ser usado para determinar se os parâmetros operacionais atuais são adequados ou se são necessárias medidas preventivas para evitar a deterioração da qualidade do efluente.

4.9 CONCLUSÕES

Os modelos de predição SVM, MLP e RF foram capazes de prever parâmetros de qualidade de efluentes a partir de dados de afluentes e efluentes obtidos de um modelo simulado de ETE WEST e da ETE AMBEV. Para a planta simulada, os melhores resultados foram obtidos no Cenário 3, onde mais informações estavam disponíveis para aprendizagem de máquina utilizando engenharia de recursos. O modelo com melhor desempenho foi o MLP, cuja fundamentação teórica é sólida e pode ser explicada pela teoria da aproximação universal. Para a ETE AMBEV, os resultados obtidos para os modelos preditivos mostraram uma melhoria visível resultante da aplicação da engenharia de recursos proposta. O resultado mais favorável foi obtido no cenário que considera o maior número de variáveis preditivas disponíveis para o modelo ML (Cenário 3), confirmando a importância de fornecer um conjunto de dados abrangente para os modelos ML. Isto inclui dados relacionados ao modo de operação e às variáveis de processo associadas ao parâmetro de saída que está sendo previsto.

Essa etapa do estudo desenvolvido no projeto destaca a importância de incluir uma ampla gama de informações relevantes nos modelos de ML para obter resultados superiores. Apesar dos erros de previsão significativos obtidos para a ETE AMBEV, o modelo RF apresentou as melhores métricas de desempenho e as tendências no comportamento das variáveis de saída foram representadas pelo modelo preditivo.

Em relação às constantes mudanças observadas nas operações das ETEs, embora a aplicação da técnica de engenharia de recursos apresentada neste estudo melhore os resultados dos modelos preditivos – dependendo do impacto e da frequência das mudanças no processo –, os resultados das previsões podem não ser da qualidade necessários para seu uso confiável. Além disso, enfatiza-se a

necessidade de medir e registrar o maior número possível de parâmetros de qualidade do efluente, com maior frequência, para melhorar a precisão dos modelos.

Para um conjunto de dados contendo medições de águas residuais, sob condições controladas e sem mudanças abruptas na operação do processo, os modelos de ML podem prever os valores e tendências das variáveis de saída quando são alimentados apenas com informações sobre a entrada. Quando há uma mudança significativa na operação da planta, a precisão dos modelos preditivos fica gravemente comprometida. A implementação de novas variáveis por meio de técnicas de engenharia de atributos, que informam os modelos de ML sobre os novos cenários operacionais, permite uma rápida adaptação dos modelos à nova condição de operação. Isso reduz a necessidade de registros adicionais referentes ao novo processo e ainda possibilita o aproveitamento dos dados antigos, que continuam fornecendo informações relevantes sobre a relação entre as variáveis de entrada e a variável de destino.

Ainda, a análise de importância dos PDPs e PI mostrou que dentre os parâmetros operacionais, NT, NH e NKT são os mais influentes na previsão do $NT_{saída}$, seguidos por DBO, DQO e SST. A vazão influente foi o parâmetro de menor importância dentre os parâmetros apresentados. No entanto, a indicação do modo de operação da ETE é a variável mais crucial para a precisão da previsão do modelo ML. Uma análise mais aprofundada destas variáveis operacionais pode fornecer informações valiosas aos operadores de ETE, permitindo-lhes compreender melhor os processos de tratamento e tomar medidas preventivas para manter a qualidade dos efluentes. Isto contribui para uma gestão mais eficiente e confiável do sistema de tratamento de águas residuais.

Por fim, os resultados contrastantes obtidos neste capítulo evidenciam um ponto nevrálgico na aplicação de *Machine Learning* em ambientes industriais: a metodologia é robusta em cenários de dados controlados, mas sua eficácia é severamente comprometida pela qualidade e pela granularidade dos dados operacionais reais. A aplicação na ETE AMBEV revelou que a ausência de variáveis preditivas chave e a baixa frequência de amostragem constituem o principal gargalo, mais até do que a escolha do algoritmo em si. Diante desta constatação, a investigação avança para uma segunda etapa, detalhada no capítulo subsequente, que visa responder a uma questão fundamental: qual o verdadeiro potencial preditivo

da metodologia quando aplicada a um conjunto de dados que, embora representativo de uma ETE real, possui a completude e a frequência necessárias para um treinamento robusto? Para tal, a análise será aprofundada com a introdução de um novo estudo de caso, a exploração de dados com maior frequência temporal e a inclusão de modelos de séries temporais.

5. APROFUNDAMENTO DA MODELAGEM PREDITIVA: ANÁLISE COM DADOS DE ALTA FREQUÊNCIA E REDES NEURAIS RECORRENTES

5.1 JUSTIFICATIVA DA PROPOSTA DE ESTUDO

A etapa anterior, detalhada no Capítulo 4, demonstrou a viabilidade da aplicação de modelos de *Machine Learning* para a predição de parâmetros de qualidade de efluentes em um ambiente simulado e controlado (ETE WEST). Contudo, o estudo também expôs a fragilidade desta abordagem frente a um cenário industrial com dados operacionais desafiadores (ETE AMBEV), onde a escassez de preditores relevantes e a inconsistência dos registros inviabilizaram a obtenção de modelos preditivos precisos.

Tal disparidade suscita a necessidade de validar a metodologia em um cenário intermediário: um que combine a complexidade e a variabilidade de um processo real com a integridade e a riqueza de um conjunto de dados bem estruturado. Portanto, o presente capítulo avança a investigação ao endereçar diretamente as limitações previamente identificadas. Para tanto, esta nova fase do estudo se propõe a:

1. Validar os modelos com um conjunto de dados mais robusto: Utilizando os dados da ETE Umbilo, um caso de estudo baseado em medições reais e validado em ambiente de simulação, busca-se avaliar o desempenho dos algoritmos em um cenário de alta fidelidade e sem as lacunas de dados encontradas anteriormente.
2. Explorar uma maior granularidade temporal: Considerando que os processos em uma ETE possuem dinâmicas rápidas, a análise será expandida para incluir predições em base horária, superando a limitação das predições diárias do estudo anterior e aproximando a aplicação de um cenário de controle em tempo real.
3. Incorporar modelos específicos para séries temporais: Dada a natureza sequencial dos dados de alta frequência, introduz-se o modelo *Long Short-Term Memory* (LSTM), uma arquitetura de rede neural recorrente

especialmente projetada para capturar dependências temporais de longo prazo, comparando sua performance com os algoritmos previamente testados.

Desta forma, este estudo não apenas busca replicar os resultados anteriores em um novo contexto, mas sim aprofundar a análise, construindo um entendimento mais completo sobre a interação entre a qualidade dos dados, a frequência de amostragem e a arquitetura do modelo para a predição fidedigna da qualidade de efluentes.

Além disso, a gestão diária da operação de estações de tratamento de efluentes (ETE) depende normalmente de instrumentos de monitorização *online* ou *offline* dos principais índices de qualidade, aliados ao conhecimento e experiência profissional dos operadores, o que é arriscado em face aos padrões de qualidade de efluentes cada vez mais rigorosos (XIE *et al.*, 2021). *Softwares* de simulação de tratamento de águas residuais, como a ferramenta de modelagem de qualidade da água WEST (MIKE *Powered by* DHI), foram desenvolvidos integrando vários modelos dinâmicos baseados em processos que descrevem matematicamente tratamentos físicos, químicos e biológicos. Este *software* tem apresentado desempenho satisfatório na simulação de diferentes cenários e configurações de ETEs (BROUCKAERT *et al.*, 2022; ELDUAYEN-ECHAVE *et al.*, 2021; MUOIO *et al.*, 2019; MURLA *et al.*, 2016; WU *et al.*, 2016).

A aplicação de modelos de aprendizado de máquina (ML) para prever parâmetros de ETE é um tema de interesse crescente na literatura científica. Vários modelos de ML têm sido utilizados com sucesso em ETEs para prever parâmetros como demanda bioquímica de oxigênio (DBO) (SINGH *et al.*, 2023), demanda química de oxigênio (DQO) (ZHAO *et al.*, 2022), nitratos e fosfatos (JAFAR *et al.*, 2022), nitrogênio total (BAGHERZADEH *et al.*, 2021), pH (XU *et al.*, 2022b), sólidos suspensos totais (SST) (WANG *et al.*, 2022), etc. Dentre esses parâmetros, o DQO desempenha um papel crucial na otimização do desempenho das ETEs, pois é um indicador fundamental para avaliar a sua eficiência e o controle automático de seus sistemas biológicos (TIYASHA; TUNG; YASEEN, 2020). No entanto, a medição direta da concentração de DQO frequentemente requer equipamentos caros e complexos, tornando o monitoramento suscetível a limitações operacionais, incluindo atrasos na obtenção dos resultados (ALAVI *et al.*, 2022; WANG *et al.*, 2019). A previsão em tempo real da DQO influente é uma abordagem promissora para enfrentar esses desafios, permitindo o ajuste

apropriado da aeração e das quantidades químicas. A correlação de diferentes parâmetros de qualidade com valores de DQO influentes em águas residuais oferece uma solução econômica, confiável e contínua para estimar DQO influente. Esta facilitação da operação eficaz permite a implementação no mundo real.

Diversos modelos de ML são utilizados para previsão de parâmetros de ETEs (SINGH *et al.*, 2023). Cada modelo tem prós e contras, e a escolha do modelo adequado depende da situação específica e dos dados disponíveis. Nesse sentido, validar modelos de ML com dados reais é crucial para o desenvolvimento de modelos mais robustos e confiáveis para prever parâmetros de ETE.

A validação de modelo computacional é o processo de formulação e fundamentação de afirmações explícitas sobre a aplicabilidade e precisão dos resultados computacionais relativos aos propósitos pretendidos do modelo e ao sistema natural que ele representa (DEE, 1995). Em outras palavras, a validação do modelo é a prova de que o modelo possui uma faixa de precisão satisfatória dentro do seu domínio de aplicabilidade, consistente com o uso pretendido (SARGENT, 2013).

Embora diversos métodos de modelagem e otimização tenham sido desenvolvidos e testados para aumentar a eficiência do tratamento de efluentes, ainda persistem vários desafios. Entre eles estão a dificuldade de comparar técnicas de forma objetiva, a ausência de diretrizes claras, a validação limitada em condições reais de operação, a carência de abordagens colaborativas e as opções restritas para otimizar ativamente o conteúdo e a qualidade dos dados (COROMINAS *et al.*, 2018). Portanto, existe uma oportunidade para criar abordagens que possam lidar com as restrições descritas anteriormente.

Várias técnicas podem ser usadas para verificar e validar modelos. Este estudo centra-se em duas destas técnicas: verificação por comparação com outros modelos e validação por comparação com dados históricos. Na verificação por comparação com outros modelos, os resultados de saída são comparados com aqueles obtidos em modelos previamente validados (modelos analíticos ou numéricos). Na validação por comparação com dados históricos, um conjunto de dados coletados de um sistema real é dividido em dois grupos: um grupo de dados é usado para construir o modelo; ao mesmo tempo, o outro é reservado para comparação com os resultados de saída para determinar se o modelo se comporta corretamente, ou seja, se replica o comportamento do sistema real (SARGENT, 2013). No entanto, deve-se notar que, uma vez que cada

modelo é validado utilizando um conjunto de dados original, a sua fiabilidade pode ser limitada a estas situações específicas (ANDREIDES; DOLEJŠ; BARTÁČEK, 2022).

A confiabilidade dos modelos de ML também está relacionada ao entendimento de como os modelos tomam decisões ou previsões, considerando que os usuários precisam ter certeza de que o modelo terá um bom desempenho em dados reais de acordo com as métricas de interesse (RIBEIRO; SINGH; GUESTRIN, 2016). Embora uma variedade de modelos de ML tenha sido desenvolvida para prever parâmetros associados à operação de ETEs, há ceticismo entre os pesquisadores devido à natureza de caixa preta das abordagens de ML (GUPTA *et al.*, 2023).

Explainable Machine Learning (XML) visa fornecer explicações claras e compreensíveis das decisões tomadas pelos modelos de ML. Para algoritmos de ML considerados “caixas pretas”, como redes neurais profundas e florestas aleatórias, o XML pode trazer transparência, confiança e responsabilidade aos sistemas de ML, permitindo que os usuários entendam e questionem as decisões tomadas pelos modelos (LIPTON, 2018; MOLNAR, 2020; RIBEIRO; SINGH; GUESTRIN, 2016). O uso de ML explicável (ou interpretável) é recomendado para todas as aplicações, incluindo ETEs, e outros estudos demonstraram os benefícios do ML integrado com várias métricas de explicabilidade (CHOI *et al.*, 2023; CINAR; CINAR; KUČHTA, 2022; LONG *et al.*, 2021; WANG; HUNTINGTON; SCOWN, 2021). *Partial Dependence Plots* (PDPs), *Permutation Importance*, *Local Interpretable Model-agnostic Explanations* (LIME) e *Shapley Additive Explanations* (SHAP) são algumas das ferramentas utilizadas na área de interpretabilidade de ML, fornecendo *insights* sobre como as variáveis de entrada estão influenciando os modelos preditivos.

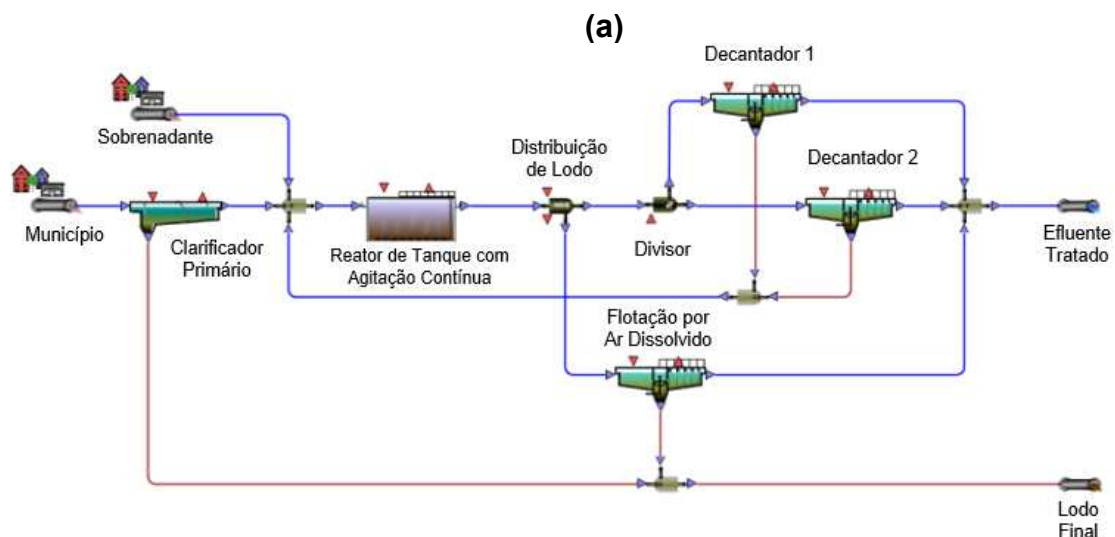
Com base nessas informações, optou-se por construir, validar e compreender modelos de ML (*Random Forest* - RF, *Long Short-Term Memory* - LSTM, *Support Vector Machine* - SVM e *Multi-Layer Perceptron* - MLP) para prever medições de DOQ de saída em uma ETE. A base de dados combina informações coletadas em uma ETE real com dados gerados por simulações através do *software* WEST (MIKE *Powered by* DHI Group, 2023). A precisão do modelo ML foi medida comparando seus resultados com dados reais e com um modelo calibrado de ETE disponível no WEST.

5.2 DESCRIÇÃO DO ESTUDO DE CASO E COLETA DE DADOS

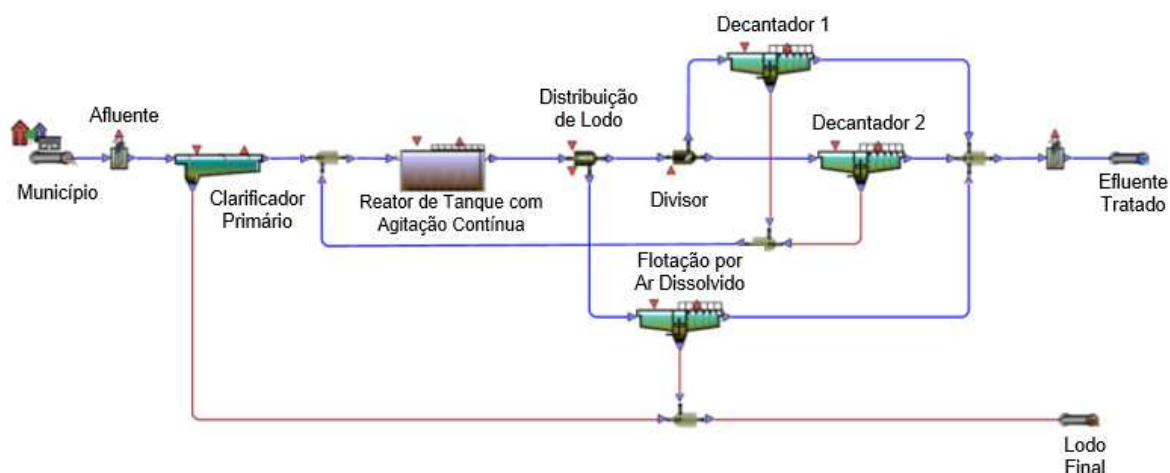
Neste estudo, o conjunto de dados utilizado para treinamento dos modelos ML foi obtido de um arquivo de projeto disponível na base de dados do WEST (DHI, 2022) designado como “Umbilo”, que foi desenvolvido para simular a ETE Umbilo, localizada na África do Sul. Este projeto fez parte dos resultados obtidos no desenvolvimento dos estudos conduzidos por Poulsen *et al.* (2005). O arquivo do projeto é descrito como um modelo calibrado, ou seja, um modelo com valores de parâmetros que foram ajustados para representar o processo real. É importante destacar que os dados utilizados neste estudo foram obtidos diretamente de uma ETE real, conforme detalhado por Poulsen *et al.* (2005). Isto significa que esses dados são precisos e intrinsecamente representativos das condições reais da estação de tratamento da qual foram coletados. Portanto, pode-se afirmar que o conjunto de dados captura fielmente as características e variações encontradas no ambiente real, contribuindo para a confiabilidade e relevância das informações de entrada em nossa análise.

Para simplificar a composição de entrada da ETE, uma entrada de dados adicional (sobrenadante) foi excluída do *layout* do modelo simulado, que incluía uma baixa vazão afluente ($Q = 125 \text{ m}^3/\text{d}$) concentrada em nitrogênio amoniacal ($\text{NH} = 135 \text{ g/m}^3$), fósforo ($\text{PO} = 14 \text{ g/m}^3$) e alcalinidade ($\text{ALK} = 3.250 \text{ g/m}^3$) (Figura 15). Esta exclusão não teve efeito significativo nos resultados da simulação.

Figura 15 – *Layouts* da ETE Umbilo utilizados para obtenção de dados no *software* WEST: (a) *layout* original disponível no banco de dados do *software* e (b) *layout* modificado excluindo a entrada de sobrenadante



(b)



Fonte: elaborado pelo autor.

A ETE Umbilo é composta por um clarificador primário, um separador de fases ideal, um reator de lamas ativadas com volume constante e mistura ideal e três clarificadores secundários. A simulação do granel no reator segue a descrição do modelo de lodo ativado *ASM2dModTemp*, disponível na biblioteca de modelos do WEST. O modelo adotado para o clarificador é baseado em uma extensão o modelo proposto por Takács, Patry e Nolasco (1991), que fornece uma estimativa do parâmetro de sedimentação por meio da medição do índice volumétrico do lodo (IVL), conforme proposto por Daigger e Roper (1985). Mais detalhes descritivos sobre esses modelos podem ser encontrados no WEST Models Guide (DHI, 2022).

Os dados reais do afluente medidos na ETE Umbilo foram fornecidos em formato fracionado implementado para importação na interface WEST. Esses dados fracionados foram convertidos para um formato não fracionado devido à relevância dos dados não fracionados para a rotina de monitoramento das ETEs. Os dados fracionados podem não ter aplicação prática para o controle operacional usual como os dados não fracionados. Além disso, os dados fracionados não costumam ser analisados nas rotinas laboratoriais.

O fracionamento utilizado neste estudo difere daquele apresentado por Poulsen *et al.* (2005), uma vez que esses autores concentraram seus esforços na modelagem com o *software* WEST para simular a ETE Umbilo, contribuindo para que a área metropolitana de Durban avançasse em direção à meta de modelar suas estações de tratamento de águas residuais e reduzir a carga orgânica e de nutrientes descarregada no rio Umbilo. No presente estudo, o *layout* Umbilo construído no

projeto WEST está disponível como amostra na versão utilizada. Os dados gerados pela simulação foram as informações aplicadas para a investigação dos modelos de ML. Os resultados da calibração obtidos por Poulsen *et al.* (2005) foram preservados como os valores originais e as frações de entrada foram mantidas iguais no modelo. Além disso, o fracionamento aplicado foi o mesmo tanto para os valores de entrada quanto de saída, mantendo assim um padrão de interpretação dos dados do modelo. Dessa forma, como o fracionamento real foi semelhante ao fracionamento padrão aplicado no WEST, esta etapa não teve impacto significativo nos resultados do estudo.

A partir daí, foi criado um novo banco de dados contendo vazão afluente diária (Q , em m^3/h), demanda química de oxigênio (DQO, em $mg\ O_2/L$), demanda bioquímica de oxigênio (DBO, em $mg\ O_2/L$), sólidos suspensos totais (SST, em mg/L), nitrogênio Kjeldahl total (NKT, em mg/L), fósforo total (FT, em mg/L), fosfato (FO, em mg/L), nitrogênio total (NT, em mg/L) e nitrogênio amoniacal (NH, em mg/L). A simulação WEST da ETE Umbilo gerou dados de afluentes e efluentes considerando frequências horárias e diárias. Essas variáveis foram escolhidas para o estudo por serem as mais analisadas nas ETEs para caracterizar a qualidade dos afluentes e efluentes. Além disso, são de grande importância para definir a entrada do modelo *ASM2dModTemp* de acordo com seus requisitos de dados e verificar a qualidade da calibração do modelo.

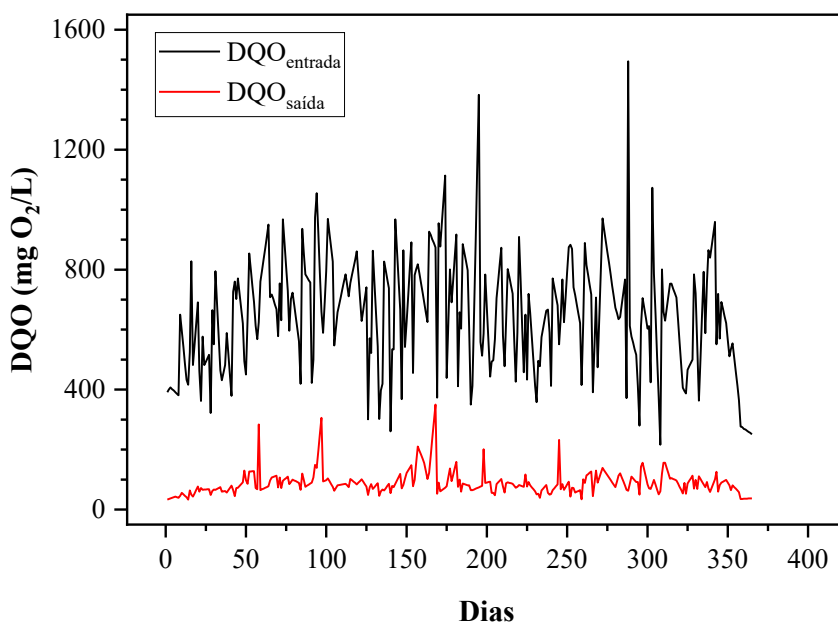
Quanto à frequência de monitoramento, optou-se pelo uso de dados diários devido à disponibilidade de medições reais realizadas nesse intervalo, o que permite uma comparação mais adequada entre os dados simulados e os dados observados. A frequência horária também foi escolhida devido à necessidade de grande volume de dados para alimentar os modelos de aprendizado de máquina. Como foi assumido o mesmo período de simulação para ambos os casos, a caracterização dos perfis de afluentes e efluentes de hora em hora fornece maiores informações sobre o processo do que diariamente. Como resultado, foram obtidos 365 valores de dados diários e 8.761 valores de dados horários por variável.

Conforme mencionado anteriormente, a DQO é um parâmetro representativo da eficiência do tratamento. Por esta razão, a DQO na saída da ETE ($DQO_{saída}$) foi definida como o parâmetro de previsão neste estudo. Os dados afluentes dos demais

parâmetros citados acima foram utilizados como entrada para os modelos de predição.

A Figura 16 apresenta perfis de DQO afluentes e efluentes, que foram utilizados como variáveis preditivas para os modelos ML neste estudo. Conforme mostrado na figura, uma diferença marcante entre os valores do afluente e do efluente é evidente ao longo do período escolhido, variando de 1.494 mg O₂/L a 216 mg O₂/L para os dados do afluente e de 350 mg O₂/L e 33 mg O₂/L para os dados do efluente. A variação significativa entre os valores de afluentes e efluentes deve-se à diminuição do teor de matéria orgânica nas águas residuais, promovida pela ETE. Normalmente, a matéria orgânica é decomposta através de processos biológicos, que são presentes na planta de tratamento. Como resultado, o efluente diminuiu significativamente os níveis de poluentes orgânicos em comparação com o afluente. Os processos biológicos removem efetivamente a maior parte da matéria orgânica das águas residuais.

Figura 16 – Perfil de entrada e saída do parâmetro DQO para os 366 dias de operação da ETE Umbilo



Fonte: elaborado pelo autor.

5.3 MODELOS DE MACHINE LEARNING

Nesta fase do estudo, além dos três algoritmos de ML já apresentados no capítulo anterior, o modelo de memória de curto longo prazo (LSTM, sigla em inglês para *long short-term memory*) também foi adotado para construir os modelos preditivos para DQO_{saída} na ETE Umbilo. Esses modelos foram escolhidos por sua capacidade de prever parâmetros-chave essenciais ao funcionamento das ETEs, incluindo DQO (HOSSAIN *et al.*, 2022; KIM; OH, 2021; LIU; CHEN, 2023; LY *et al.*, 2022; SALLES *et al.*, 2022; WAQAS *et al.*, 2022; XU *et al.*, 2022a; ZHANG *et al.*, 2023; ZHOU *et al.*, 2023). A diversidade desses modelos é fundamental no estudo de processos complexos, como os de uma ETE, pois cada abordagem apresenta características, vantagens e limitações específicas. Isso resulta em diferentes desempenhos potenciais, permitindo identificar qual modelo representa com maior precisão os dados disponíveis na base analisada.

Outro fator a considerar na escolha desses modelos é o seu custo computacional. Deve-se notar que modelos como o LSTM, devido ao seu *design* complexo e processamento de dados orientado a sequências, podem exigir mais recursos computacionais do que modelos com arquitetura mais simples. Esta avaliação considerou o compromisso entre a complexidade do modelo e a qualidade da previsão no contexto específico deste estudo.

A configuração experimental utilizada para o desenvolvimento dos modelos de aprendizado de máquina foi implementada no editor de código-fonte *Visual Studio Code* (VS Code, Microsoft), onde os *scripts* foram desenvolvidos em Python utilizando o ambiente *Jupyter Notebook*. Conforme descrito anteriormente, os dados de entrada dos modelos foram obtidos a partir das simulações realizadas no *software* WEST.

O modelo LSTM é uma variação de redes neurais recorrentes (RNNs) que foram projetadas para lidar com problemas de dependências de longo prazo em sequências de dados (GUANG *et al.*, 2019). O algoritmo pode capturar dependências de longo prazo em sequências de dados. Eles possuem células de memória que podem armazenar informações por longos períodos e mecanismos de controle para decidir quais informações reter e esquecer. O LSTM é amplamente utilizado em tarefas de

processamento de sequências, como previsão de séries temporais e detecção de anomalias (PISA *et al.*, 2021).

Para construir o modelo LSTM em escala horária, inicialmente foi criada uma variável a partir do deslocamento de um *timestep* dos dados a serem previstos. A seguir, os dados foram transformados em tensores com 10 intervalos de tempo. Assim, o banco de dados do preditor (X) possui uma dimensão de (8.750, 10, 5), sendo a primeira dimensão referente ao número de amostras no banco de dados, a segunda dimensão representando o número de *timesteps* utilizados no modelo, e a última dimensão referente às variáveis utilizadas para a regressão. A base de dados foi separada em treinamento, validação e teste, sendo as duas bases anteriores do mesmo tamanho (719 itens).

Utilizando *Keras* (DEVELOPERS, 2023; TENSORFLOW CORE, 2023), foram construídas duas estruturas LSTM: a primeira, S_LSTM, apresenta uma rede simples e fixa, com a primeira camada oculta do tipo LSTM com 64 neurônios, seguida de duas camadas densas, cada uma com 16 neurônios, função de ativação ReLU e taxa de *dropout* de 10%, finalizando com a camada de saída com um nó e ativação linear. O modelo foi construído utilizando a função de perda *Mean Squared Error*, o otimizador Adam com taxa de aprendizado 0,0001 e a métrica de avaliação RMSE.

Além disso, o modelo foi treinado por 500 épocas utilizando a base de treinamento e validado utilizando a base de validação. Para este estudo, uma variável de entrada adicional (valor DQO_{saída} do intervalo de tempo anterior) foi adicionada ao banco de dados de entrada. Os dados de treinamento foram utilizados no formato (7.312, 10, 5), onde as dimensões representam o número de amostras, intervalos de tempo e variáveis reportadas ao algoritmo, respectivamente. Nessa estrutura, utilizou-se exclusivamente para a validação do modelo, durante sua construção, um conjunto de dados com o mesmo número de amostras do conjunto de teste (719 dados).

A segunda estrutura LSTM, denominada C_LSTM, foi criada com a implementação da técnica *Grid Search CV*. Nesta configuração, foram fixadas duas camadas ocultas: a primeira consiste em uma camada LSTM, enquanto a segunda é uma camada densa e totalmente conectada. Os hiperparâmetros foram exaustivamente testados para essas camadas para encontrar os valores ideais que melhor descrevem a relação entre os dados de DQO de entrada e saída. Os valores

dos neurônios testados independentemente para cada camada foram 16, 32, 64 e 128; a taxa de abandono para cada uma das duas camadas ocultas foi definida como idêntica e os valores testados foram 0, 10% ou 20%. A taxa de aprendizado do otimizador Adam foi testada como 0,001, 0,01 ou 0,1. Para criar o modelo preditivo, o algoritmo foi treinado por 100 épocas, com tamanho de lote 10. A validação cruzada com 5 vezes foi usada para identificar a maior correlação entre os dados de DQO de entrada e saída em todo o conjunto de dados de treinamento. A melhor estrutura foi obtida com 128 neurônios na camada LSTM, 32 neurônios na camada densa, *dropout* de 10% e taxa de aprendizado de 0,01.

Para prever a $DQO_{saída}$ utilizando os dados reais coletados na ETE Umbilo, foi desenvolvido outro modelo LSTM com uma arquitetura de 64 neurônios do tipo LSTM na primeira camada, seguida de três camadas ocultas de 16 neurônios cada, utilizando uma taxa de *dropout* de 10%. A compilação do modelo foi feita utilizando o otimizador Adam e a função de perda foi o *Mean Absolute Error*. O treinamento do modelo foi realizado por 100 épocas com tamanho de lote de 50. O conjunto de treinamento consistiu em 335 amostras, cada uma com registro diário e quatro variáveis.

O conjunto de validação continha 30 amostras no mesmo formato de registro diário e quatro variáveis. Essa configuração permitiu que o modelo se adaptasse aos dados de treinamento ao longo das épocas para minimizar o erro absoluto médio, resultando em previsões de DQO mais precisas com base nas características diárias do conjunto de dados reais.

O conjunto de treinamento foi composto por 335 amostras e o conjunto de validação continha 30 amostras, com um único registro diário e quatro variáveis. Essa configuração permitiu que o modelo se adaptasse aos dados de treinamento ao longo das épocas para minimizar o erro absoluto médio, resultando em previsões de DQO mais precisas com base nas características diárias do conjunto de dados reais.

As métricas de desempenho foram selecionadas como função de perda em todos os modelos de previsão analisados: erro quadrático médio (RMSE), erro médio absoluto (MAE), erro percentual médio absoluto (MAPE) e coeficiente de correlação (R^2) – descritas no capítulo anterior.

A interpretabilidade dos modelos de ML foi realizada a partir dos Gráficos de Dependência Parcial (PDPs) e a Importância de Permutação (PI), utilizados para

verificar a significância das variáveis de entrada para prever a DQO de saída e compreender a conexão entre cada variável de entrada e a variável alvo do modelo preditivo com melhor desempenho.

A importância do modelo LSTM foi calculada conforme metodologia sugerida por Altmann (2010). O método de importância de permutação (PI), desenvolvido por Altmann, é uma heurística que tenta mitigar qualquer viés nas estimativas de importância de recursos. O procedimento PI envolve a realização de um teste de permutação no qual os valores de resposta (ou pontuações) são permutados aleatoriamente. Esta permutação visa avaliar a importância aleatória de cada característica, criando assim uma distribuição de importância aleatória que permite avaliar a significância das estimativas de importância observadas. Além disso, o PI fornece valores p de significância associados a cada recurso, indicando quão relevante é a importância de cada recurso em comparação com a importância aleatória.

Os valores dos tensores (amostra e *timestep*) foram permutados, preservando as informações da variável. É importante considerar que esta análise considera no máximo 2 características para 1 alvo. Isso se deve à incapacidade do ser humano de visualizar gráficos com mais de 3 dimensões. Como os tensores são multidimensionais, isso pode ser um gargalo para a execução de PDPs usando dados de tensores (MOLNAR, 2020). O modelo treinado foi então usado para prever os valores reorganizados, enquanto o erro quadrático médio foi calculado e comparado com a métrica original sem rearranjo. O aumento do erro de previsão foi utilizado para determinar a importância de cada variável. Todos os valores de importância atribuídos foram apresentados em escala percentual para melhorar a interpretabilidade e comparabilidade dos resultados. Valores percentuais mais elevados indicam uma contribuição mais significativa, enquanto valores mais baixos indicam uma contribuição menos relevante.

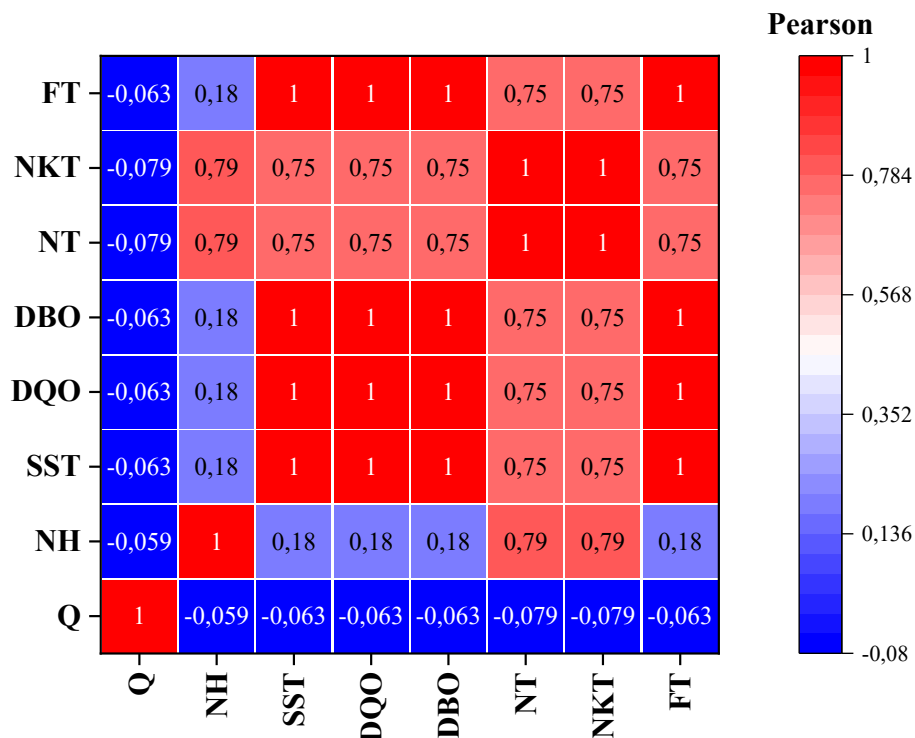
5.4 RESULTADOS OBTIDOS PELA ANÁLISE EXPLORATÓRIA DE DADOS

O banco de dados utilizado neste estudo teve dados de entrada para 9 variáveis: Q, NH, NKT, NT, SST, DBO, DQO, FT e FO. Realizou-se uma análise exploratória dos dados para identificar e excluir variáveis cujo valor permanecia

constante ao longo da série temporal. A variável de concentração de fosfato (FO) apresentava essa característica e, portanto, foi removida do conjunto de dados.

Para avaliar e quantificar o grau de associação entre duas variáveis, realizou-se uma análise de correlação de Pearson utilizando a biblioteca Python pandas (REBACK *et al.*, 2021). O resultado é apresentado na Figura 17.

Figura 17 – Matriz de correlação de Pearson para as variáveis de entrada da ETE Umbilo



Fonte: elaborado pelo autor.

Foi possível observar que algumas variáveis apresentaram multicolinearidade, com correlações de Pearson superiores a 0,95. Utilizar uma base de dados com esse perfil pode constituir um problema na estimativa das relações entre essas variáveis (FARRAR; GLAUBER, 1967; KALNINS, 2018). A multicolinearidade é crítica na construção de modelos de ML, pois pode afetar a estabilidade, interpretação e generalização do modelo (ZAGHLOUL *et al.*, 2021). Além disso, quando duas ou mais variáveis independentes são altamente correlacionadas, podem conter informações muito semelhantes sobre o resultado da variável dependente. Isso pode levar a informações redundantes no modelo, tornando menos claro quais variáveis contribuem para a predição (CHAN *et al.*, 2022).

Com base nas informações obtidas pela análise de correlação de Pearson, apenas as variáveis de entrada Q, NH, SST e NT e a saída de DQO foram mantidas no banco de dados utilizado para construção do algoritmo de ML. Os dados de entrada Q, NH, SST e NT foram variáveis independentes e a saída de DQO foi a variável dependente. A escolha do SST como variável independente em vez de DBO e DQO baseia-se na alta correlação observada entre essas variáveis e no fato de que a análise laboratorial para medir a concentração de SST ser geralmente considerada menos complexa, exigindo equipamentos menos especializados e reagentes específicos em comparação para análise de DQO. A mesma razão se aplica à preferência pelo NT sobre o NKT. A Tabela 4 apresenta algumas métricas relacionadas às variáveis utilizadas no presente trabalho, incluindo os três diferentes cenários de DQO de efluentes utilizados neste estudo.

Tabela 4 – Resumo estatístico dos parâmetros de entrada da ETE Umbilo durante os 366 dias de operação

Variável	Máximo	Mínimo	Média	Mediana
Q_{entrada} (m³/d)	13259,9	3183,4	8003 ± 1813	7879
NH_{entrada} (mg/L)	68,9	8,0	25 ± 6	24,0
SST_{entrada} (mg/L)	1033,6	150,2	453 ± 129	457,0
NT_{entrada} (mg/L)	88,6	22,2	47 ± 10	46,9
DQO_{saída} (mg/L)¹	350,0	33,0	89 ± 40	83,0
DQO_{saída} (mg/L)²	272,7	33,2	95 ± 39	85,9
DQO_{saída} (mg/L)³	291,7	33,1	95 ± 39	84,8

Fonte: ¹Valores reais obtidos da ETE Umbilo (POULSEN; LAURIDSEN, 2005); ²Valores diários simulados a partir do modelo calibrado criado no *software* WEST; ³Valores horários simulados a partir do modelo calibrado elaborado no *software* WEST.

Ao avaliar os valores de DQO_{saída}, que correspondem ao parâmetro objetivo do algoritmo de ML desenvolvido neste estudo, observa-se que os três conjuntos de dados utilizados para treinar o modelo apresentam um ligeiro desvio entre eles. Os valores médios e medianos estão próximos entre si, indicando uma distribuição

aproximadamente simétrica, ou seja, os valores dos dados tendem a ser equilibrados em torno de um ponto central, a mediana.

Ao comparar as medições reais de $DQO_{saída}$ com dados simulados gerados pelo WEST (horária e diariamente), observa-se uma semelhança, indicando que o modelo pode prever tendências e médias de concentração de DQO. Este resultado era esperado, uma vez que a maioria dos parâmetros de remoção de DQO são determinados experimentalmente (POULSEN; LAURIDSEN, 2005). No entanto, a limitação do modelo de simulação WEST reside nas flutuações reais do efluente, que não podem ser simuladas com precisão. Isso foi observado ao avaliar o valor máximo de $DQO_{saída}$, que é maior no conjunto de dados reais (350 mg O_2/L) do que nos valores simulados (272,7 mg O_2/L e 291,7 mg O_2/L). Esta disparidade indica que a simulação não conseguiu prever o desvio deste estado normal de funcionamento da ETE. Esta dificuldade também é um desafio para os modelos de ML, uma vez que são concebidos para aprender padrões e comportamentos a partir de dados históricos fornecidos durante o treino. No entanto, dados calibrados e validados obtidos a partir de modelos de simulação WEST fornecem um banco de dados confiável para apoiar o treinamento de algoritmos de ML em situações em que medições reais de águas residuais são escassas ou ausentes.

5.5 RESULTADOS OBTIDOS PARA A PREVISÃO DIÁRIA USANDO DADOS GERADOS PELA SIMULAÇÃO WEST

Os dados diários de $DQO_{saída}$ previstos pelos modelos RF, SVM e MLP foram comparados com os valores de $DQO_{saída}$ obtidos pela simulação WEST. A precisão de cada modelo foi avaliada utilizando as métricas de desempenho RMSE, MAE, MAPE e R^2 , apresentadas na Tabela 5. Uma análise gráfica comparativa entre os dados previstos pelos modelos ML e os dados simulados gerados pelo WEST é mostrada na Figura 18a.

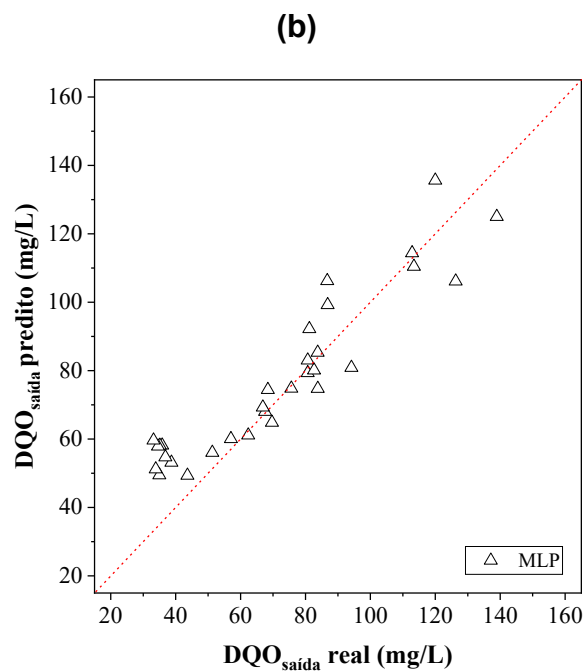
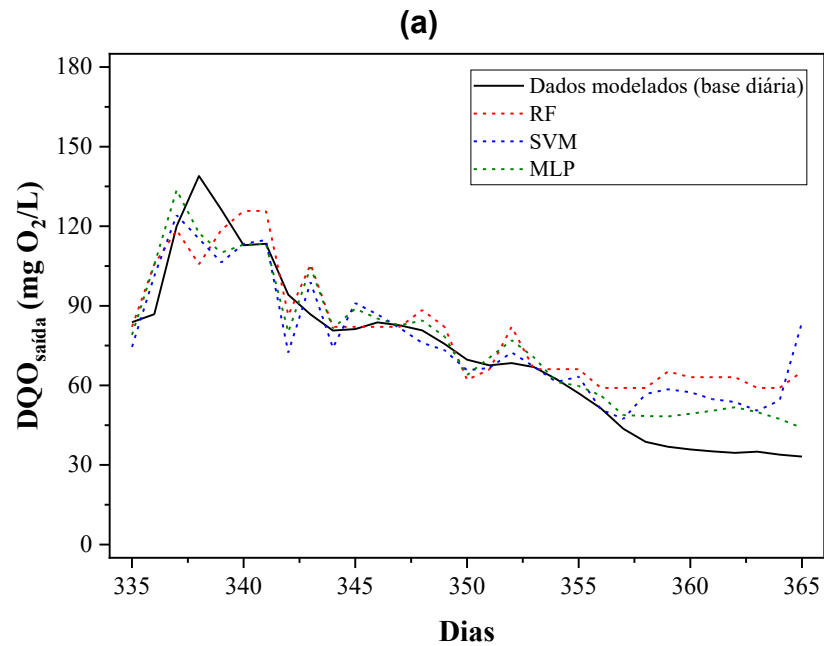
Tabela 5 – Comparação de métricas de desempenho para modelos preditivos de ML para DQO_{saída} aplicados a dados simulados diariamente pelo *software* WEST

Métricas	RF	SVM	MLP
RMSE	16,85	15,44	10,66
MAE	13,12	11,03	8,53
MAPE	0,27	0,23	0,16
R ²	0,67	0,73	0,87

Fonte: elaborado pelo autor.

O teste de Tukey indica que os modelos são estatisticamente diferentes entre si. Com base nos índices estatísticos que avaliam a capacidade preditiva dos modelos, constatou-se que o modelo MLP foi o mais eficiente dentre os modelos preditivos testados, apresentando os menores valores de RMSE, MAE e MAPE e o maior valor de R², o que é consistente com a teoria da aproximação universal (CSÁJI, 2001b; HORNIK; STINCHCOMBE; WHITE, 1989; LU *et al.*, 2017). Os hiperparâmetros selecionados após a busca exaustiva com validação cruzada do modelo *scikit-learn MLPRegressor* foram alfa 0,05, tamanhos de camada oculta (16, 16, 16), taxa de aprendizagem 0,0001, max_iter 500, solucionador Adam e tolerância 10⁻⁵. Porém, vale ressaltar que a análise visual das previsões de cada modelo de ML mostra comportamento semelhante entre eles.

Figura 18 – Resultados dos modelos preditivos para $DQO_{saída}$ utilizando os dados simulados pelo *software* WEST em base diária para todos os modelos testados (a) para o modelo com melhor resultado (MLP) (b)



Fonte: elaborado pelo autor.

Os desvios mais significativos dos valores de referência foram observados nos últimos 9 dias do período de teste. Em particular, os modelos preveem valores superiores aos de referência nestes dias específicos. Essa disparidade de previsão

ocorre porque o conjunto de dados de treinamento continha valores superiores aos observados nesse período. Consequentemente, os modelos precisavam estar adequadamente preparados para lidar com essas situações incomuns, resultando em previsões mais altas do que os valores iniciais nestes últimos dias de teste.

Outros estudos mostraram resultados promissores quanto ao desempenho do MLP para prever o comportamento dos parâmetros de qualidade em ETEs municipais e industriais (LIU; CHEN, 2023; ZHOU *et al.*, 2023). Modelos baseados em RNAs têm sido frequentemente preferidos devido às suas diversas capacidades, incluindo o tamanho relativo do campo, adequação para resolver problemas reais não lineares, eficiência em termos de tempo e custo, potencial para computação paralela e completa adaptabilidade no caso de uso (SHAMSHIRBAND *et al.*, 2019). A Figura 18b apresenta a comparação entre o valor real do DQO_{saída} e o valor previsto pelo modelo MLP. Embora apresente tendência linear, observam-se desvios importantes entre os valores reais e previstos.

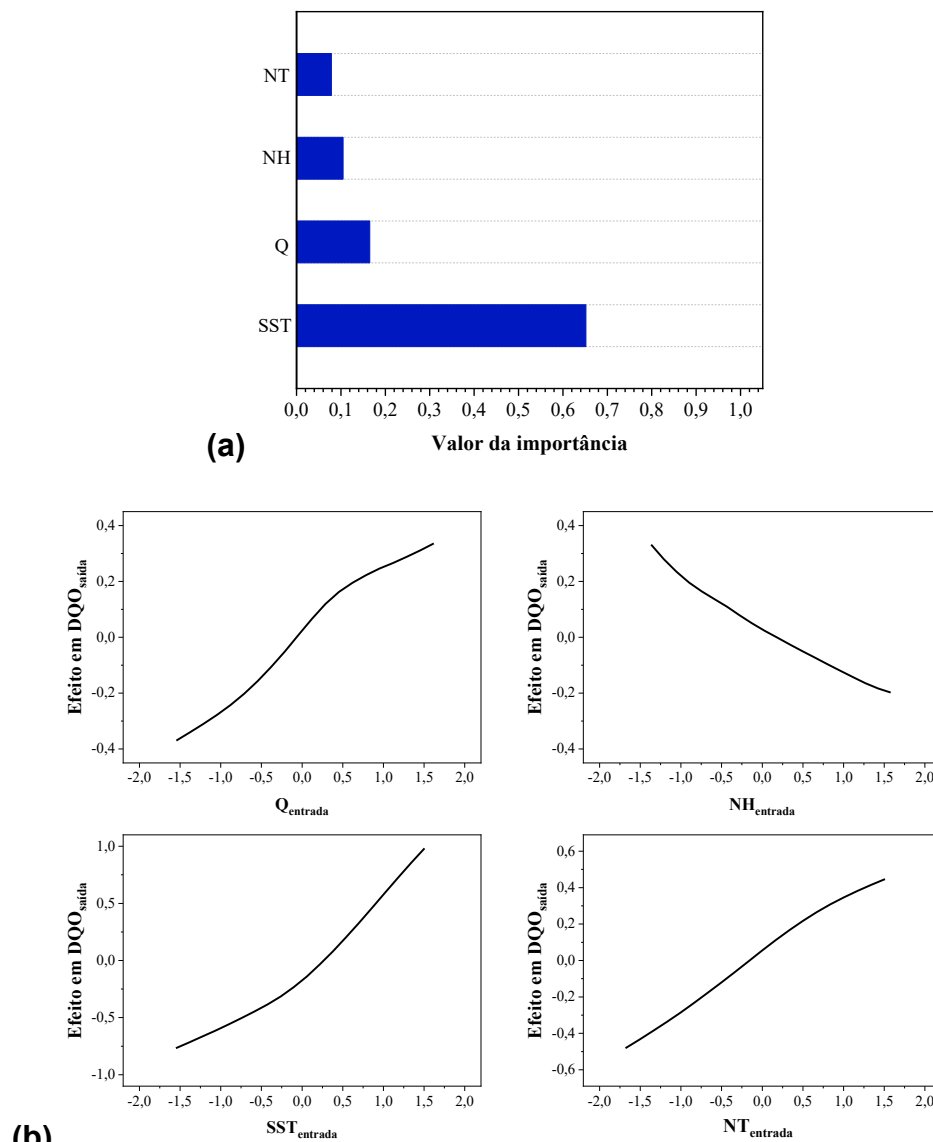
A análise de importância dos parâmetros foi empregada para apoiar a interpretação dos resultados da modelagem em ML e auxiliar na identificação do modelo de melhor desempenho (MLP). A Figura 19a mostra a relação de importância das variáveis independentes (dados influentes de NT, NH, Q e SST) para prever a variável dependente (DQO_{saída}). Essa importância representa o impacto na precisão do modelo quando os valores das variáveis preditoras são embaralhados aleatoriamente (SCIKIT-LEARN DEVELOPERS, 2023c).

Como pode ser observado, a concentração de SST se destaca em relação às demais variáveis. A influência do SST é aproximadamente 2,4 vezes maior do que a influência da vazão afluente, que tem a segunda maior importância de permutação. As variáveis relacionadas ao teor de nitrogênio influente (NT e NH) apresentam os menores valores de importância entre as variáveis estudadas.

Do ponto de vista físico-químico dos processos de tratamento de águas residuais, é bem conhecido que existe uma relação bem estabelecida entre SST e concentração de DQO. Essa relação não é direta porque o SST inclui sólidos suspensos orgânicos e inorgânicos (FONDRIEST ENVIRONMENTAL INC., 2014), enquanto a DQO representa o oxigênio necessário para decompor quimicamente a matéria orgânica (BURTON *et al.*, 2013). Alguns estudos sugerem fortes correlações entre SST e DQO (GIOKAS *et al.*, 2002). No entanto, esta correlação varia dependendo das

características das águas residuais e das fontes de poluição. SST e DQO podem ter uma correlação positiva em alguns casos, especialmente se os sólidos suspensos forem orgânicos. Em outros casos, a correlação pode ser mais fraca ou prejudicial se o SST consistir principalmente em partículas inorgânicas. Vale ressaltar que SST, DQO e DBO apresentaram multicolinearidade (Figura 17), indicando que as três variáveis estão altamente correlacionadas entre si em comparação com DQO_{saída}.

Figura 19 – Importância das variáveis de entrada para os modelos preditivos de DQO_{saída} utilizando o banco de dados simulado diariamente pelo *software* WEST (a) e gráficos de dependência parcial ilustrando os efeitos das variáveis de entrada Q (canto superior esquerdo), NH (canto superior direito), SST (canto inferior esquerdo) e NT (canto inferior direito) na predição de DQO_{saída} em escala padronizada (b) para o modelo MLP



Fonte: elaborado pelo autor.

A interpretação do PI pode ser enriquecida pela análise do PDP, conforme mostrado na Figura 19b. O PDP indica que menores Q, NT e SST influentes resultam em menor DQO_{saída}, com uma relação quase linear. Fluxos influentes mais elevados podem implicar níveis mais elevados de DQO e outras variáveis de qualidade do influente, contribuindo para um aumento da DQO prevista.

Para entender melhor como analisar este gráfico, o eixo X representa os diferentes valores possíveis da variável de entrada, que neste estudo são apresentados de forma normalizada. O eixo Y representa a média da saída do modelo da variável, dadas as diferentes instâncias ou valores da variável de entrada. O gráfico exibe uma linha reta ou curva que mostra como a predição média varia conforme a variável de entrada muda. Esta linha representa a relação média entre a variável de entrada e a saída, mantendo todas as outras variáveis constantes. Uma linha horizontal indicaria que a variável de entrada não afeta a saída, enquanto uma inclinação positiva ou negativa indicaria um efeito positivo ou negativo na saída. Pela leitura do gráfico é possível observar como a variável de entrada afeta a saída do modelo. Se a linha tiver inclinação ascendente, significa que à medida que a variável de entrada aumenta, a previsão tende a aumentar. Se for descendente, significa que à medida que a variável de entrada aumenta, a previsão tende a diminuir.

Diferentes comportamentos são observados para as espécies de nitrogênio (NH e NT). A concentração total de nitrogênio nas águas residuais é a soma de amônia e nitrogênio orgânico (também conhecido como nitrogênio Kjeldahl total – NKT), nitritos e nitratos (ENVIRONMENTAL PROTECTION AGENCY, 2009). O teor de matéria orgânica presente nas águas residuárias também é um fator essencial para o processo de nitrificação-desnitrificação, indicando que são necessários 3 g a 5 g de DQO para cada grama de NO_3^- reduzido (HENZE *et al.*, 2001), com compostos orgânicos atuando como doadores de elétrons na reação de desnitrificação. Isto indica que quanto maior o teor de NH nas águas residuárias brutas, maior será a formação de nitrato nos reatores (considerando a nitrificação efetiva) e maior será a necessidade de fornecimento de matéria orgânica para promover a desnitrificação. Em relação à concentração de NT, o aumento do seu valor não indica necessariamente que haverá a mesma necessidade, uma vez que também contém compostos orgânicos de nitrogênio. Se um aumento na concentração de NH causar um valor de NT mais alto,

a relação de dependência torna-se a mesma que para NH. Porém, se um valor de NT mais elevado for devido a um aumento na fração orgânica, isso contribuirá para o aumento da DQO no efluente tratado.

Após análise dos resultados dos modelos de ML treinados na base de dados diária gerada pela simulação WEST, constatou-se que eram consistentes com a realidade operacional da ETE. Esta análise é fundamental para obter uma compreensão mais fidedigna das relações identificadas pelos modelos, além de apontar a necessidade de possíveis correções adicionais.

5.6 RESULTADOS OBTIDOS PARA A PREVISÃO HORÁRIA USANDO DADOS GERADOS PELA SIMULAÇÃO WEST

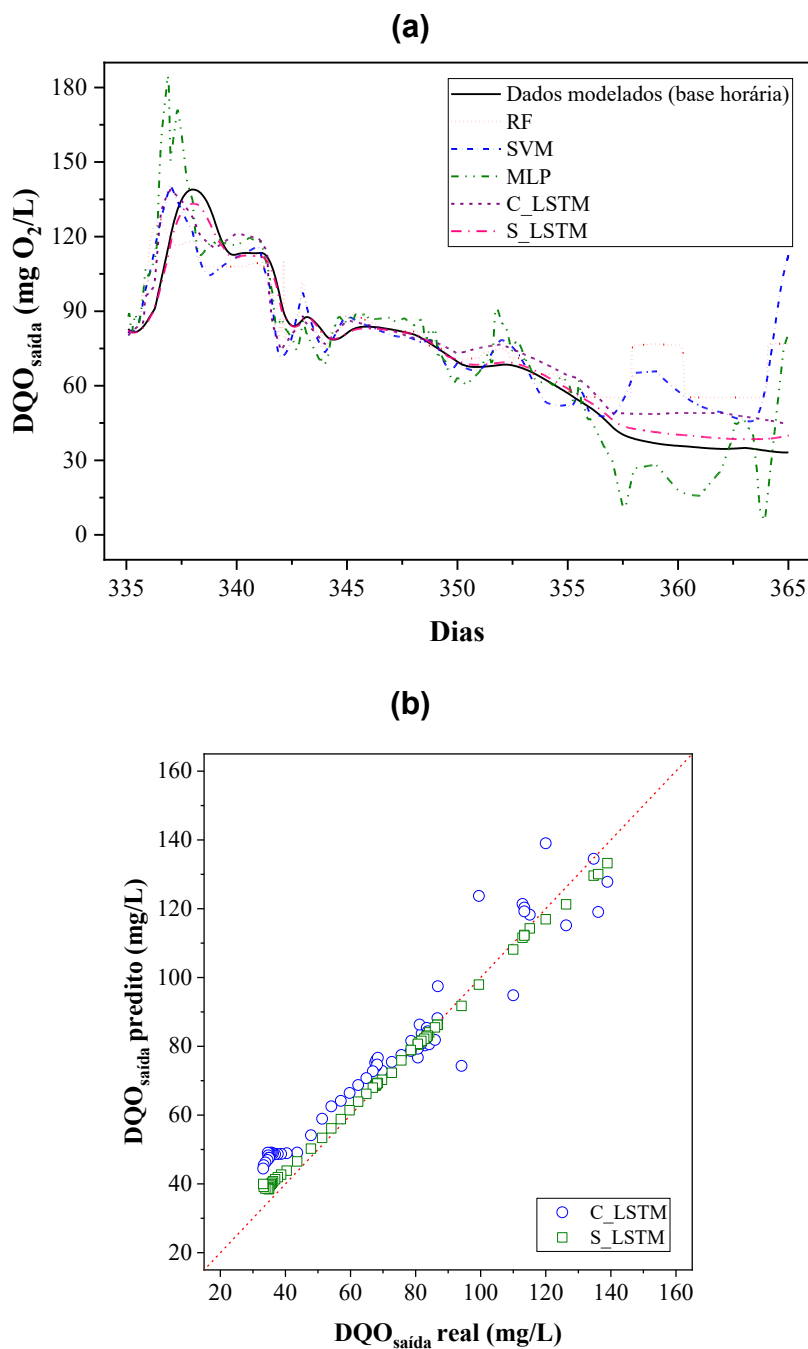
As métricas das previsões horárias de $DQO_{saída}$ feitas pelos modelos ML são apresentadas na Tabela 6. A análise gráfica comparativa entre os dados previstos pelos modelos e os dados obtidos pela simulação WEST é apresentada na Figura 20a.

Tabela 6 – Comparação de métricas de desempenho para modelos preditivos de ML para $DQO_{saída}$ aplicados a dados simulados pelo *software* WEST em base de teste horária

Métricas	RF	SVM	MLP	S_LSTM	C_LSTM
RMSE	17,77	16,59	14,85	2,86	9,51
MAE	12,28	10,98	10,45	2,18	7,83
MAPE	0,26	0,23	0,19	0,04	0,150
R ²	0,64	0,69	0,75	0,99	0,90

Fonte: elaborada pelo autor.

Figura 20 – Resultados dos modelos preditivos para $DQO_{saída}$ utilizando os dados simulados pelo *software* WEST em base de teste horária para todos os modelos testados (a) para os modelos com melhor resultados (C_LSTM e S_LSTM) (b)



Fonte: elaborado pelo autor.

O desempenho superior dos modelos LSTM é evidente, com os menores valores de RMSE, MAE e MAPE e os maiores valores de R^2 . Além disso, os modelos são estatisticamente diferentes entre si de acordo com o teste de Tukey. A estrutura otimizada por *Grid Search* para o C_LSTM (que foi o melhor modelo

preditivo estruturado apenas com variáveis de entrada como preditores) foi composta por ativação de ReLU, *dropout* de 10%, taxa de aprendizagem de 0,01, 128 unidades na camada LSTM e 32 na camada densa subsequente. A análise visual das previsões de cada modelo ML mostra um comportamento semelhante entre eles, com os desvios mais significativos entre os valores previstos e os valores de referência observados nos últimos 10 dias do período de teste.

Embora computacionalmente mais simples, a estrutura S_LSTM apresentou melhores resultados de predição DQO_{saída}. Isso indica que a qualidade dos dados alimentados no modelo impacta mais a previsão final do que uma arquitetura refinada como C_LSTM. No estudo de Zhang *et al.* (2023), menciona-se que a precisão de detecção do modelo LSTM pode ser melhorada com a inclusão de indicadores de entrada que contribuam significativamente para os indicadores de saída, o que corrobora os resultados obtidos neste estudo. Por outro lado, observam também que a exclusão de indicadores de entrada que têm relevância limitada para os indicadores de saída pode reduzir o custo de aquisição de dados sem causar uma perda significativa na precisão da detecção. O ajuste consistente do modelo é observado na Figura 20b, cujo comportamento linear é confirmado.

Os LSTMs têm sido descritos na literatura como uma alternativa interessante para prever variáveis em ETES devido à sua capacidade de compreender padrões e tendências em séries temporais complexas, tornando-os valiosos para a previsão de parâmetros essenciais (LY *et al.*, 2022; SALLES *et al.*, 2022; XU *et al.*, 2022a; ZHANG *et al.*, 2023). Esse recurso só é possível porque o neurônio LSTM contém três portas adicionais (porta de esquecimento, porta de entrada e porta de saída), e as conexões entre esses nós permitem que a informação permaneça (ZHOU *et al.*, 2022). No período do conjunto de dados onde são verificados os desvios mais significativos entre os valores previstos e observados, os modelos MLP, SVM e RF apresentaram o pior desempenho, possivelmente relacionado à limitada capacidade de extrapolação (ARIMOTO; PRASAD; GIFFORD, 2005).

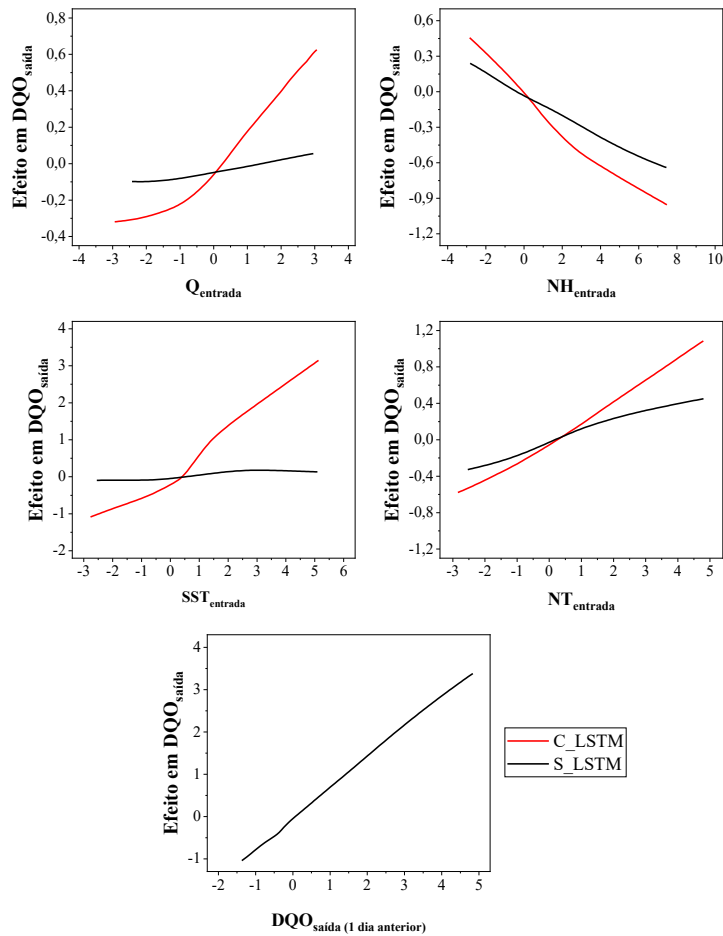
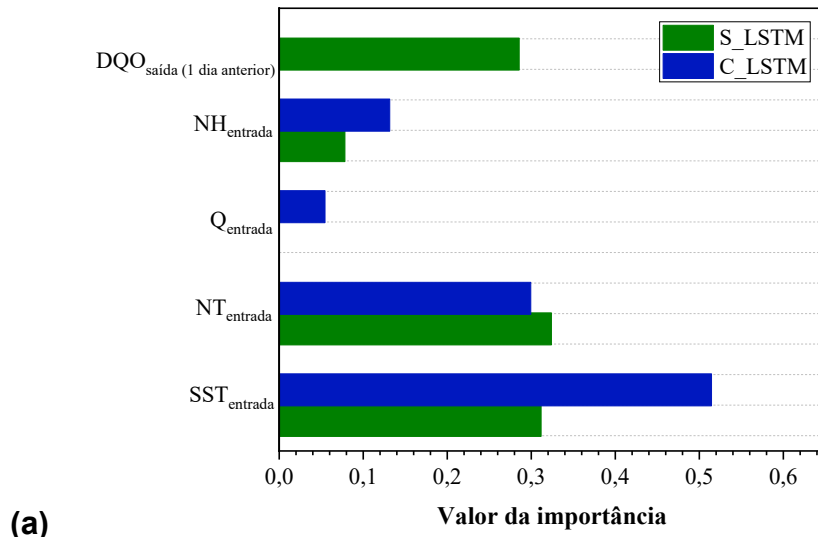
Embora os modelos LSTM apresentem desempenho superior em muitos cenários de análise de dados sequenciais, sua complexidade e requisitos computacionais associados são aspectos críticos a serem considerados. A implementação e operação desses modelos podem exigir recursos computacionais significativos, principalmente devido à sua arquitetura de rede neural recorrente

(MAOR *et al.*, 2019). Portanto, ao usar modelos LSTM, é essencial avaliar cuidadosamente a capacidade computacional disponível e considerar os recursos necessários para treinar, ajustar e aplicar efetivamente esses modelos em cenários do mundo real.

A Figura 21a apresenta a importância das variáveis independentes na previsão da $DQO_{saída}$ utilizando os modelos S_LSTM e C_LSTM. Assim como observado para os dados diários, SST é a variável mais relevante para a predição da DQO no modelo C_LSTM, seguido por NT, NH e Q. No entanto, quando a variável associada ao passo de tempo anterior foi incluída como entrada no modelo, o padrão de importância das variáveis se alterou. Isso resultou em uma distribuição mais equilibrada dos pesos, com valores muito próximos para DQO, NT e SST. Além disso, a importância atribuída à vazão afluente tornou-se praticamente nula nesse cenário. Esse comportamento é corroborado pela análise PDP apresentada na Figura 21b.

Observa-se que para o modelo C_LSTM (linhas vermelhas), o aumento nos valores de Q resulta em um efeito maior na predição de $DQO_{saída}$, bem como no aumento dos valores de NT e SST, que também têm um efeito crescente na predição de $DQO_{saída}$. Além disso, o aumento do valor de NH provoca um efeito negativo na DQO. Este perfil de efeito é consistente com o observado no conjunto de dados diários. Ao aplicar o modelo S_LSTM com valores variados de Q, NT e SST, observa-se um efeito insignificante no valor $DQO_{saída}$ previsto. Mais uma vez, a concentração influente de NH mostra um efeito negativo, diminuindo o valor de $DQO_{saída}$ à medida que seu valor aumenta. Porém, é importante destacar que a nova variável da série temporal ($DQO_{saída}$ do dia anterior) tem um efeito positivo muito significativo e o aumento do seu valor tem uma relação linear quase proporcional com o aumento do efeito preditivo da variável $DQO_{saída}$.

Figura 21 – Importância das variáveis de entrada para os modelos preditivos de $DQO_{saída}$ utilizando a base de dados horária obtida pelo *software* WEST (a) e gráficos de dependência parcial ilustrando os efeitos das variáveis de entrada Q, NH, SST e NT na predição de $DQO_{saída}$ em escala padronizada (b) para os modelos C_LSTM e S_LSTM



Fonte: elaborado pelo autor.

5.7 RESULTADOS OBTIDOS PARA A PREVISÃO USANDO DADOS DE MEDIÇÕES REAIS DA ETE UMBILO

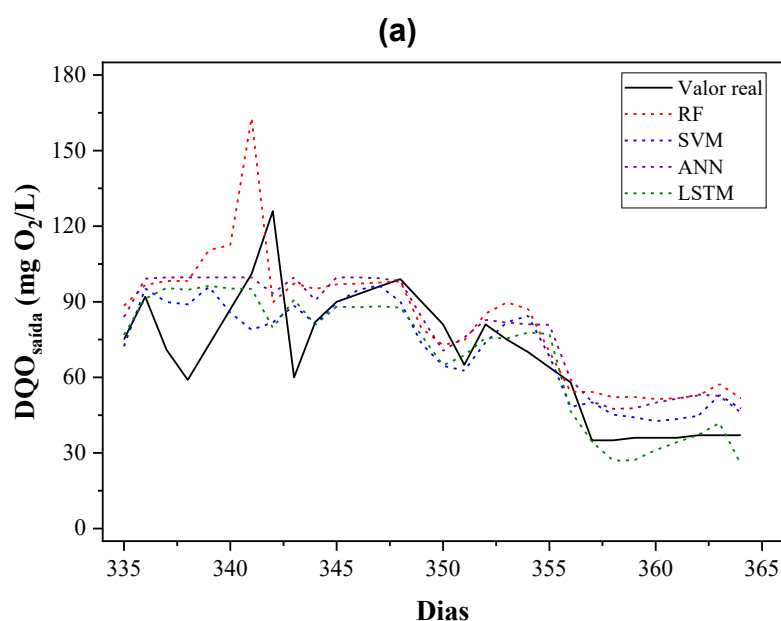
As métricas de previsões de $DQO_{saída}$ utilizando dados de medições reais da ETE Umbilo (POULSEN; LAURIDSEN, 2005) estão resumidas na Tabela 7. Neste caso, também são apresentados os resultados obtidos com o modelo WEST, considerando os últimos 30 dias de dados simulados e medidos, a partir da linha de base de 365 dias. Além disso, a Figura 22a mostra a análise gráfica comparativa entre os dados previstos e os dados reais de medição.

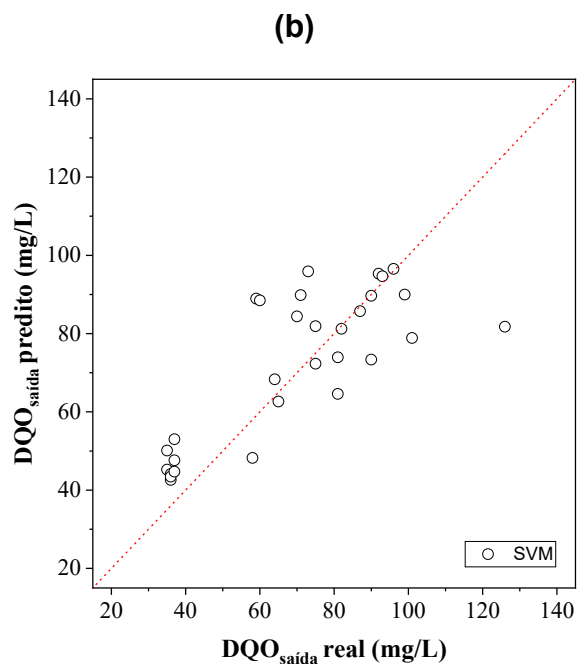
Tabela 7 – Comparação de métricas de desempenho para modelos preditivos de ML para $DQO_{saída}$ aplicados a dados reais coletados na ETE Umbilo

Métricas	RF	SVM	ANN	LSTM	WEST
RMSE	21,97	15,31	16,94	16,30	22,94
MAE	17,30	11,53	13,41	12,04	14,53
MAPE	0,29	0,19	0,23	0,19	0,20
R ²	0,18	0,60	0,51	0,55	0,11

Fonte: elaborado pelo autor.

Figura 22 – Resultados dos modelos preditivos para $DQO_{saída}$ utilizando a base de dados reais da ETE Umbilo (a) e para o modelo com melhor resultado (SVM) (b)





Fonte: elaborado pelo autor.

Os modelos baseados em redes neurais (RNA e LSTM) apresentaram resultados de predição satisfatórios. Não há diferença estatisticamente significativa entre esses dois modelos segundo o teste de Tukey. O conjunto de dados para construir o modelo LSTM foi dividido em 305 dias para treinamento, 30 dias para validação e 30 dias para teste. O algoritmo do modelo LSTM foi implementado com 64 neurônios do tipo LSTM, seguidos por três camadas ocultas de 16 neurônios cada, com *dropout* de 10%. O otimizador Adam foi utilizado para treinar o modelo e o Erro Médio Absoluto (MAE) foi escolhido como função de perda. O treinamento do modelo foi realizado em 100 épocas com um tamanho de lote de 50.

O modelo de RNA foi construído com duas camadas ocultas, cada uma contendo 8 neurônios, após busca em grade com validação cruzada de 5 vezes. Um *dropout* de 0,5 foi usado para evitar *overfitting*. Além disso, foram aplicadas restrições de peso com limite de 3,0 para cada neurônio. Para o otimizador Adam, foi utilizada uma taxa de aprendizado de 0,1. Essas escolhas de hiperparâmetros foram feitas para otimizar o desempenho e generalização do modelo de rede neural artificial (RNA) para a tarefa.

Métricas de previsão apropriadas também foram fornecidas pelo modelo SVM. O teste de Tukey não mostra diferença estatisticamente significativa entre os modelos SVM, RNA e LSTM. O modelo SVM foi construído com *kernel* polinomial, grau 4, tolerância 10^{-4} , valor C de 0,5 e valor *epsilon* de 0,5. Como pode ser observado na

Figura 22a, não foram verificados *underfitting* e *overfitting* no modelo desenvolvido. O *underfitting* ocorre quando a rede não é treinada adequadamente e os padrões nos dados de treinamento não são totalmente capturados. Isto ocorre antes do ponto de convergência e, neste cenário, o desempenho da generalização não é significativamente diferente do desempenho nos dados de treinamento. O *overfitting*, por outro lado, ocorre após o ponto de convergência quando a rede é sobreajustada. Neste cenário, o erro continua a diminuir durante o treinamento enquanto o erro nos dados de teste aumenta. O *overfitting* é caracterizado por baixo viés e alta variância (KHAN *et al.*, 2020). Alguns casos que levam ao *overfitting* incluem dados de treinamento insuficientes, exemplos insuficientes para cobrir adequadamente todos os valores possíveis dos dados de entrada e grandes quantidades de informações sem sentido ou ruído nos dados de treinamento (HAMED; AB AZIZ; YAAKUB, 2023).

Embora esses problemas não tenham sido identificados no modelo desenvolvido neste estudo, é essencial avaliar cuidadosamente o ajuste do modelo aos dados e utilizar técnicas adequadas para evitar *underfitting* e *overfitting* para garantir modelos robustos e adaptáveis e obter resultados confiáveis em diferentes situações do mundo real.

Como o teste estatístico não mostrou diferenças significativas entre os modelos SVM, RNA e LSTM, o custo computacional foi considerado como critério para seleção do modelo mais adequado. Neste contexto, o modelo SVM se destaca por apresentar a melhor eficiência em termos de tempo de treinamento durante a fase de modelagem. Conforme observado para a base de dados diária obtida pelo *software* WEST, o comportamento linear dos dados de DQO_{saída} obtidos pelo modelo SVM é observado na Figura 22b, porém com importantes desvios de centralidade.

O modelo SVM é amplamente conhecido por suas aplicações bem-sucedidas em vários campos, incluindo classificação, regressão e detecção de anomalias. Sua capacidade de encontrar hiperplanos de separação ideais mesmo em espaços de alta dimensão o torna uma escolha valiosa para problemas de aprendizado de máquina (ANDRADE CRUZ *et al.*, 2022). SVMs são algoritmos de aprendizagem supervisionada que requerem dados de treinamento rotulados para fazer previsões sobre dados novos e não vistos (ONU; MBOHWA; PRADHAN, 2023). Vários pesquisadores têm utilizado o algoritmo SVM para prever parâmetros relacionados ao tratamento de águas (HOSSAIN *et al.*, 2022; KIM; OH, 2021; WAQAS *et al.*, 2022).

Em comparação com abordagens de aprendizagem superficial, os modelos de aprendizagem profunda, como o modelo SVM, oferecem capacidades mais avançadas para identificar características não lineares em variáveis de águas residuais (ANSARI; OTHMAN; EL-SHAFIE, 2020; CHENG *et al.*, 2020; NIU *et al.*, 2020; Yaqub *et al.*, 2020).

SVMs apresentam diversas vantagens sobre outros algoritmos de aprendizado de máquina: 1) lidam com conjuntos de dados de alta dimensão, como aqueles encontrados em genômica, processamento de linguagem natural e visão computacional (OMAR *et al.*, 2022); 2) ao escolher adequadamente o *kernel* e os parâmetros de regularização, é possível controlar o *overfitting* no SVM (GUO *et al.*, 2021), tornando-o útil mesmo em pequenos conjuntos de dados; 3) são resistentes a dados ruidosos e discrepantes devido à ênfase na maximização da margem de separação (ONU; MBOHWA; PRADHAN, 2023); e 4) podem ser estendidos para lidar com classes desequilibradas, atribuindo pesos diferentes às classes ou utilizando técnicas de amostragem (TAO *et al.*, 2019).

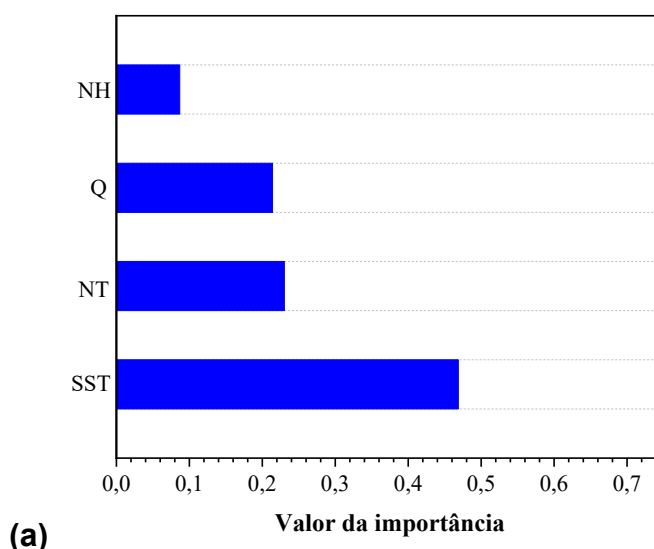
Contudo, algumas limitações do modelo devem ser mencionadas: 1) a capacidade preditiva do SVM é altamente dependente da função do *kernel*. Um *kernel* mal escolhido pode levar a resultados ruins (CAO; XIN; YUAN, 2016); e 2) para grandes conjuntos de dados, treinar um SVM pode ser computacionalmente caro e demorado (CERVANTES *et al.*, 2020). Portanto, deve-se ressaltar que a escolha do modelo SVM deve ser baseada na natureza do problema, nos dados disponíveis e nos objetivos do projeto.

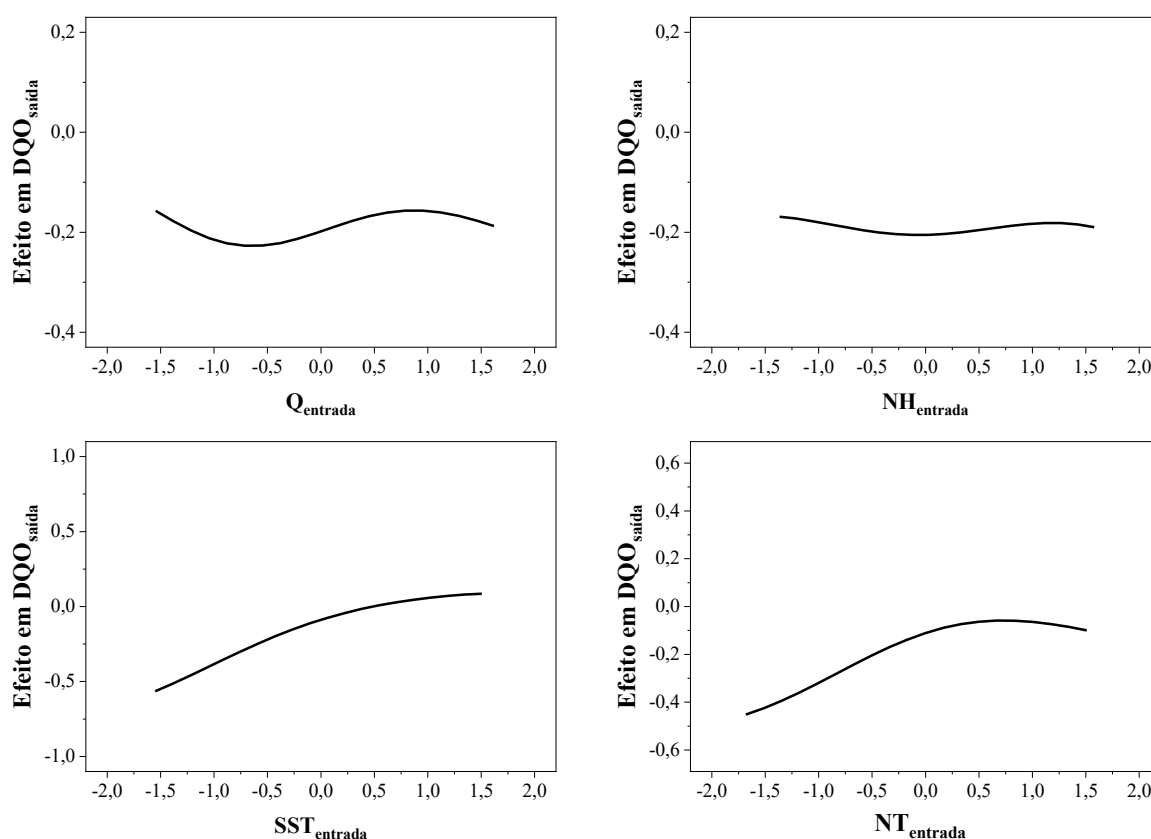
Embora os modelos construídos possam prever efetivamente a DQO_{saída}, os dados foram medidos diariamente para este estudo de caso. Portanto, os modelos ML foram aplicados apenas para prever mudanças diárias na qualidade do efluente em um curto período (12 meses). Para obter uma previsão mais robusta, estudos futuros exigiriam a recalibração e revalidação do modelo com um conjunto de dados mais amplo. Além disso, outras variáveis de entrada poderiam ser consideradas em futuros trabalhos de modelagem. É importante notar também que os modelos ML produziram resultados que excederam a precisão de previsão dos modelos mecanísticos aplicados ao *software* de simulação de processos. Isto destaca a eficácia das ferramentas de inteligência artificial utilizadas neste estudo para prever DQO_{saída}.

A análise PI para este caso (conforme mostrado na Figura 23a) confirma que o SST é o mais significativo para prever $DQO_{saída}$, seguido por NT, Q e NH. A análise PDP (mostrada na Figura 23b) também confirma este resultado, demonstrando que os dados de SST de entrada impactam significativamente a previsão de DQO do efluente. Baixos valores de SST de entrada resultam em valores de $DQO_{saída}$ abaixo da média. Além disso, taxas de fluxo de entrada mais altas estão associadas a valores mais elevados de $DQO_{saída}$. Ambas as conclusões são consistentes com o funcionamento fenomenológico de uma ETE e validam a consistência do modelo.

No entanto, em contraste com situações anteriores, os dados de entrada de NT e NH tiveram uma influência ligeiramente menor na previsão de $DQO_{saída}$. Uma hipótese para explicar este último resultado é a baixa eficiência da desnitrificação no processo real comparado ao processo simulado, possivelmente devido à ausência de um ambiente anóxico adequado. Isto implica um menor consumo de matéria orgânica neste processo, reduzindo, assim, a influência na concentração de DQO do efluente. Porém, também é possível considerar que os modelos de ML possuem natureza estocástica. Portanto, a aleatoriedade pode desempenhar um papel essencial na determinação da importância da variável. Para este e os demais casos, recomenda-se uma análise mais detalhada, incluindo simulação em série, para avaliar uma média dos resultados obtidos para as análises de importância e PDP.

Figura 23 – Importância das variáveis de entrada para os modelos preditivos de $DQO_{saída}$ utilizando a base de dados real da ETE Umbilo (a) e gráficos de dependência parcial ilustrando os efeitos das variáveis de entrada Q, NH, SST e NT na previsão de $DQO_{saída}$ em escala padronizada (b) para o modelo SVM





(b)

Fonte: elaborado pelo autor.

5.8 CONCLUSÕES

Os resultados apresentados demonstram de forma convincente a viabilidade e eficácia do uso de algoritmos de ML para prever DQO_{saída} em uma ETE. A análise exploratória dos dados selecionou cuidadosamente variáveis essenciais, eliminando a multicolinearidade e garantindo um conjunto de dados adequado dentre aqueles disponíveis para o estudo de treinamento dos modelos. Foi possível construir e avaliar com sucesso modelos de ML, incluindo *Support Vector Machine* (SVM), *Random Forest* (RF), Redes Neurais Artificiais (ANN) e *Long Short-Term Memory* (LSTM), todos aplicados para prever a concentração de DQO em águas residuais tratadas.

Os resultados mostraram que as arquiteturas de redes neurais tiveram desempenho superior em métricas de avaliação como RMSE, MAE, MAPE e R^2 . Em particular, o modelo LSTM destacou-se como o mais eficaz para dados horários, demonstrando a capacidade das redes neurais de replicar padrões presentes em

dados de séries temporais. A inclusão de variáveis temporais antecedentes na estrutura do LSTM aumentou seu poder preditivo. A análise da importância das variáveis confirmou que o teor de Sólidos Suspensos Totais (SST) prevaleceu como o parâmetro mais influente na predição da DQO, seguido pela concentração de Nitrogênio Total (NT), vazão afluente (Q) e Nitrogênio Amoniacoal (NH) em importância menor. Tais resultados refletem a natureza complexa dos processos de tratamento de águas residuais, onde fatores como a composição do afluente e a vazão influenciam significativamente a qualidade do efluente tratado.

Apesar dos resultados encorajadores obtidos por estes modelos, é crucial enfatizar a necessidade de realizar investigações futuras visando a melhoria contínua das previsões, como o desempenho de simulações em série, para verificar a média dos resultados obtidos para a importância e análises de PDP. Além disso, melhorar o conjunto de dados e refinar os hiperparâmetros do modelo poderia contribuir para previsões ainda mais precisas e robustas. Aplicar estes modelos sob diferentes condições operacionais e compará-los com outras abordagens de modelagem fenomenológica também poderia fornecer uma compreensão mais holística de suas capacidades. Esta pesquisa expande as possibilidades para outras aplicações de ML na otimização e controle de processos ambientais.

6. DISCUSSÃO GERAL

Esta dissertação investigou o desenvolvimento e a validação de modelos de aprendizado de máquina para a predição de parâmetros de qualidade de efluentes, visando a futura otimização de Estações de Tratamento de Efluentes (ETEs). Os estudos de caso, detalhados nos capítulos anteriores, abrangeram cenários distintos: uma ETE industrial com dados operacionais reais (AMBEV), uma ETE municipal simulada (WEST) e uma ETE municipal com dados reais e simulados (Umbilo). A análise conjunta destes casos permite extrair conclusões transversais sobre os desafios, as melhores práticas e o real potencial da aplicação de Inteligência Artificial neste domínio. Esta discussão sintetiza os achados, comparando os resultados sob a ótica da qualidade dos dados, da engenharia de atributos e do desempenho dos algoritmos

O primeiro estudo (Capítulo 4) revelou que, embora os modelos RF, SVM e MLP tenham tido sucesso em prever o NT em uma ETE simulada, especialmente ao incorporar informações sobre mudanças operacionais, sua aplicação em uma ETE industrial (AMBEV) foi severamente limitada pela qualidade e pela falta de variáveis preditivas chave. O segundo estudo (Capítulo 5), focado na predição de DQO com dados mais completos de outra ETE (Umbilo), demonstrou a superioridade de modelos de redes neurais, com destaque para o LSTM em dados horários, e validou a consistência dos modelos com dados reais.

6.1 O IMPACTO DA REALIDADE DOS DADOS

O contraste mais significativo observado nos resultados reside no desempenho dos modelos quando aplicados a diferentes fontes de dados. Enquanto os modelos alcançaram alta performance nos cenários das ETEs WEST e Umbilo, com coeficientes de determinação (R^2) chegando a 0,87 e 0,99, respectivamente, a aplicação na ETE AMBEV resultou em métricas insatisfatórias, com valores de R^2 consistentemente negativos. Esta disparidade não pode ser atribuída apenas aos algoritmos, mas fundamentalmente à qualidade e à natureza dos dados de entrada.

O diagnóstico realizado nos dados da ETE AMBEV revelou uma série de desafios operacionais comuns em ambientes industriais, como a compilação de dados em formatos distintos, a presença de *outliers*, a falta de padronização de rótulos e um elevado percentual de dados ausentes e com periodicidade inconsistente. Mais criticamente, foi identificada a ausência de medições diárias para parâmetros de alta relevância preditiva, como DBO, nitrogênio total e fósforo. Em contrapartida, os estudos de caso de sucesso (WEST e Umbilo) foram construídos sobre bases de dados completas, que incluíam variáveis como Nitrogênio Kjeldahl Total (NKT), Sólidos Suspensos Totais (SST) e Nitrogênio Amoniacal (NH) que, na análise de interpretabilidade, mostraram-se fundamentais para a predição precisa dos parâmetros de saída das ETEs.

Este comparativo evidencia uma lição fundamental: a viabilidade da implementação de modelos de ML em ETEs está condicionada a uma avaliação prévia da maturidade e prontidão dos dados. O sucesso não reside apenas na sofisticação do modelo, mas na existência de uma infraestrutura de coleta de dados que garanta a disponibilidade, consistência e relevância das variáveis monitoradas.

6.2 ENGENHARIA DE ATRIBUTOS COMO CHAVE PARA A CAPTURA DO CONTEXTO OPERACIONAL

Os resultados deste estudo demonstram inequivocamente que o desempenho dos modelos preditivos é drasticamente aprimorado quando se fornece, além das variáveis físico-químicas, informações sobre o contexto operacional da planta. Duas estratégias distintas de engenharia de atributos foram validadas com sucesso.

A primeira, aplicada no estudo da ETE WEST (Capítulo 4), consistiu na criação de uma variável categórica para representar uma mudança estrutural no processo: a transição de uma para duas linhas de operação. A inclusão deste único atributo transformou os modelos, elevando o R^2 do MLP de -13,82 para 0,72 e permitindo que o algoritmo distinguísse corretamente entre os dois regimes operacionais.

A segunda estratégia, explorada no estudo da ETE Umbilo com dados horários (Capítulo 5), foi a inclusão de uma variável temporal, utilizando o valor da

DQO de saída do *timestep* anterior como um preditor para o valor atual. Esta abordagem, que captura a inércia e a "memória" do sistema, foi crucial para o desempenho superior do modelo S_LSTM, que alcançou um R^2 de 0,99.

A eficácia de ambas as estratégias sugere que um sistema de monitoramento inteligente para ETEs deve ir além das medições de sensores. É imperativo que o "diário de bordo" da planta – registros de manutenções, paradas de reatores, mudanças de insumos, etc. – seja digitalizado e estruturado para servir como atributos de entrada para os modelos.

6.3 ESCOLHA DO ALGORITMO ADEQUADO À NATUREZA DOS DADOS

O presente estudo avaliou um portfólio de algoritmos (RF, SVM, MLP e LSTM), permitindo uma análise comparativa sobre sua adequação a diferentes cenários.

Redes Neurais (MLP e LSTM): As arquiteturas de redes neurais se destacaram, corroborando a teoria da aproximação universal. O modelo MLP demonstrou ser altamente eficaz para prever dados com frequência diária. Contudo, para dados com maior granularidade temporal (horária), a arquitetura LSTM foi inequivocamente superior, dada sua capacidade intrínseca de modelar dependências de longo prazo em séries temporais.

Random Forest (RF): Consistentemente, o modelo RF apresentou o melhor desempenho entre os algoritmos testados nos cenários mais desafiadores, como na base de dados da ETE AMBEV e no Cenário 2 da ETE WEST. Por ser um método *ensemble*, sua robustez a dados ruidosos e possivelmente incompletos o torna uma opção valiosa para uma primeira abordagem em ETEs que ainda estão em fase de aprimoramento de seus sistemas de coleta de dados.

A escolha do modelo, portanto, não deve ser arbitrária, mas sim uma decisão técnica baseada na frequência, qualidade e complexidade temporal dos dados disponíveis.

6.4 IMPLICAÇÕES PARA A OTIMIZAÇÃO E PARA O CONTROLE DE PROCESSOS

Os estudos de predição aqui realizados constituem um passo fundamental e indispensável para atingir a otimização da operação de ETEs. Um sistema de otimização, seja um Controle Preditivo Baseado em Modelo (MPC) ou um Sistema de Apoio à Decisão (SAD), necessita de predições acuradas da qualidade do efluente para poder tomar decisões ótimas sobre variáveis de manipulação (por exemplo, dosagem de químicos e vazão de aeração).

Desta forma, este trabalho contribui ao definir os pré-requisitos para a construção de um módulo preditivo confiável, demonstrando que a qualidade dos dados e a incorporação de contexto operacional são mais críticas do que a própria escolha do algoritmo. Ao validar a eficácia dos modelos em prever o comportamento futuro do sistema a partir de dados de entrada, abre-se o caminho para o acoplamento destes modelos a um algoritmo de otimização que possa, em tempo real, testar diferentes cenários operacionais e sugerir o que leva à redução de custos e à melhoria da eficiência, conforme almejado nos objetivos deste projeto.

6.5 CONSIDERAÇÕES FINAIS

O presente estudo debruçou-se sobre o desenvolvimento e a validação de modelos de aprendizado de máquina (*machine learning*) com o objetivo de prever, com acurácia e antecedência, parâmetros de qualidade de efluentes. Tal esforço representa um passo fundamental e indispensável para a futura otimização da operação de ETEs, almejando o equilíbrio entre a eficiência de tratamento e a sustentabilidade econômica dos processos. Ao longo de duas etapas investigativas distintas, que abrangeram cenários industriais reais, dados simulados e diferentes granularidades temporais, o trabalho logrou êxito em não apenas construir modelos preditivos, mas também em delinear os pré-requisitos essenciais para a sua implementação bem-sucedida no chão de fábrica.

É importante ressaltar que os resultados apresentados não são válidos apenas para o conjunto de dados coletados na ETE Umbilo ou na ETE AMBEV e utilizados para o treinamento dos modelos de ML. As características operacionais, a composição dos resíduos e as condições ambientais podem variar

significativamente entre as diferentes estações de tratamento de águas residuais, de modo que é crucial reconhecer que as conclusões obtidas neste estudo são exclusivas para cada estação de tratamento.

No entanto, isto não deve ser visto como um obstáculo à aplicação de modelos de ML para prever parâmetros críticos para o funcionamento de ETEs, mas sim como uma oportunidade para destacar a importância da coleta de dados e reciclagem de modelos na sua aplicação a diferentes cenários. Os modelos de ML podem ser adaptados com sucesso a outras ETEs, desde que tenham sido feitos esforços suficientes para adaptar e treinar os modelos de acordo com as características específicas de cada ambiente.

Uma implicação prática relevante deste estudo é o notável uso de modelos preditivos no controle em tempo real de ETEs. Pesquisas anteriores, como o estudo realizado por Wang *et al.* (2021) na ETE de Umea, na Suécia, e por Bernardelli *et al.* (2020) na ETE de Modena, na Itália, demonstram com sucesso como a IA pode prever e adaptar as operações da ETE em tempo real, resultando em economia de recursos e conformidade com as normas ambientais. Ao analisar continuamente os dados operacionais e as condições do sistema, os modelos de ML podem antecipar mudanças nas características dos efluentes brutos, nas condições climáticas e em outros fatores que afetam a eficiência do tratamento.

Para reduzir os custos operacionais, os modelos podem ajustar automaticamente as configurações do processo, como dosagem de reagentes, fluxo de ar ou outros parâmetros operacionais, de forma a minimizar o consumo de recursos como energia e produtos químicos, sem comprometer a qualidade do tratamento. Além disso, o controle preditivo pode ser direcionado para outros objetivos, como maximizar a eficiência da remoção de poluentes ou garantir que a ETE atenda a padrões ambientais rigorosos. Esta flexibilidade na adaptação de modelos de ML permite que as ETEs atinjam vários objetivos, permitindo uma gestão de operações mais eficaz e econômica.

A principal conclusão deste estudo reside na constatação inequívoca de que a maturidade e a qualidade dos dados operacionais são mais críticas do que a própria sofisticação do algoritmo de *Machine Learning*. A disparidade de desempenho observada entre os modelos aplicados aos dados controlados e completos das ETEs WEST e Umbilo (com R^2 atingindo 0,99) e o desempenho

insatisfatório na ETE industrial AMBEV (com R^2 consistentemente negativos) ilustra que a viabilidade da Inteligência Artificial em ETEs está condicionada a uma infraestrutura de coleta e gerenciamento de dados que garanta consistência, frequência e, acima de tudo, a disponibilidade de variáveis de alta relevância preditiva.

Visando a superação das barreiras identificadas para a implementação de sistemas inteligentes em escala real, recomenda-se a modernização da infraestrutura de monitoramento através da instalação de sensores *online*. A automação da coleta de dados mitigaria a variabilidade inerente à amostragem manual e garantiria a padronização temporal das medições — assegurando que as variáveis de entrada e saída sejam registradas em janelas de tempo síncronas. Tal investimento permitiria a aquisição de séries temporais de alta frequência para parâmetros críticos (como SST, NT e DQO), viabilizando a aplicação das arquiteturas de Redes Neurais Recorrentes (LSTM) que demonstraram, neste estudo, superioridade no aprendizado de dinâmicas temporais complexas. Desta forma, pode-se afirmar que a qualidade dos dados é o alicerce da predição.

Demonstrou-se que o desempenho dos modelos preditivos é drasticamente aprimorado quando se fornece, para além das variáveis físico-químicas, informações sobre o contexto operacional da planta. A criação de atributos categóricos para representar mudanças estruturais no processo (como a transição entre regimes operacionais na ETE WEST) foi capaz de transformar um modelo ineficaz em um modelo robusto, elevando seu coeficiente de determinação de -13,82 para 0,72. De forma análoga, a inclusão de variáveis temporais defasadas — utilizando o estado do sistema no instante anterior como preditor do estado atual — foi fator determinante para o sucesso do modelo LSTM em dados horários, alcançando um R^2 de 0,99. Tais resultados validam que a digitalização do "diário de bordo" operacional é um passo imperativo para a construção de modelos verdadeiramente inteligentes e que a engenharia de atributos é uma importante ferramenta de conhecimento para modelos aplicados às ETEs.

A avaliação comparativa de diferentes algoritmos permitiu traçar um panorama sobre sua aplicabilidade. As arquiteturas de redes neurais (MLP e LSTM) confirmaram sua superioridade, com o modelo MLP mostrando-se altamente eficaz para dados de frequência diária e o LSTM sendo inequivocamente a melhor escolha

para dados de alta frequência (horária), dada sua capacidade intrínseca de modelar dependências sequenciais de replicar padrões presentes em dados de séries temporais. Por outro lado, o modelo *Random Forest* (RF) revelou-se o mais resiliente em cenários com dados ruidosos e incompletos, emergindo como uma valiosa opção para uma abordagem inicial em ETEs que ainda aprimoram seus sistemas de monitoramento. Assim, sugere-se que, para trabalhos futuros, seja dada atenção à adequação do algoritmo à natureza dos dados.

REFERÊNCIAS

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, maio 2019.

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, 5 mar. 2022.

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.

ALTMANN, A. *et al.* Permutation importance: a corrected feature importance measure. **Bioinformatics**, v. 26, n. 10, p. 1340–1347, 15 maio 2010.

ANDRADE CRUZ, I. *et al.* Application of machine learning in anaerobic digestion: Perspectives and challenges. **Bioresource Technology**, v. 345, p. 126433, 1 fev. 2022.

ANDREIDES, M.; DOLEJŠ, P.; BARTÁČEK, J. The prediction of WWTP influent characteristics: Good practices and challenges. **Journal of Water Process Engineering**, v. 49, p. 103009, 1 out. 2022.

ANSARI, M.; OTHMAN, F.; EL-SHAFIE, A. Optimized fuzzy inference system to enhance prediction accuracy for influent characteristics of a sewage treatment plant. **Science of The Total Environment**, v. 722, p. 137878, 20 jun. 2020.

ARIMOTO, R.; PRASAD, M. A.; GIFFORD, E. M. Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular

descriptors. **Journal of biomolecular screening**, v. 10, n. 3, p. 197–205, abr. 2005.

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

AVELLA, A. C. *et al.* A combined approach for a better understanding of wastewater treatment plants operation: Statistical analysis of monitoring database and sludge physico-chemical characterization. *Water Research*, v. 45, n. 3, p. 981–992, jan. 2011.

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, maio 2015.

BAGHERZADEH, F. *et al.* Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. **Journal of Water Process Engineering**, v. 41, p. 102033, 1 jun. 2021.

BAKER, A. Simplicity. Em: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Summer 202 ed. Stanford, CA: Metaphysics Research Lab, Stanford University, 2022.

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, jul. 2021.

BERNARDELLI, A. *et al.* Real-time model predictive control of a wastewater treatment plant based on machine learning. **Water Science and Technology**, v. 81, n. 11, p. 2391–2400, 1 jun. 2020.

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

BERTRAND, F. **Sweetviz 2.1.4**. Disponível em:
<<https://pypi.org/project/sweetviz/>>. Acesso em: 21 maio. 2023.

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, 1 abr. 2009.

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

BRAILE, P. M.; CAVALCANTI, J. E. W. A. **Manual De Tratamento de Águas Residuárias Industriais**. São Paulo, SP: CETESB, 1993.

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

BRITO, A. G. *et al.* Brewery and Winery Wastewater Treatment: Some Focal Points of Design and Operation. Em: OREOPOULOU, V.; RUSS, W. (Eds.). **Utilization of By-Products and Treatment of Waste in the Food Industry**. Boston, MA: Springer US, 2007. p. 109–131.

BROSTEL, R. C.; NEDER, K. D.; SOUZA, M. A. A. **Análise comparativa do desempenho de estações de tratamento de esgotos do Distrito Federal**. 21o Congresso Brasileiro de Engenharia Sanitária e Ambiental. **Anais...** João Pessoa/PB: 2001.

BROUCKAERT, B. *et al.* Using plant data to estimate biodegradable COD fractions – case study kwaMashu WWTP. **Water Science and Technology**, v. 86, n. 9, p. 2045–2058, 1 nov. 2022.

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. **Anais...**Praga, República Tcheca: 2013. Disponível em: <<https://arxiv.org/abs/1309.0238>>. Acesso em: 22 maio. 2023

BURTON, F. L. *et al.* **Wastewater Engineering: Treatment and Resource Recovery**. 5. ed. [s.l.] McGraw Hill Education, New York, NY, 2013.

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

CAO, H.; XIN, Y.; YUAN, Q. Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. **Bioresource Technology**, v. 202, p. 158–164, 1 fev. 2016.

CECHINEL, Maria Alice Prado *et al.* Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: a plant-level analysis. **Journal Of Water Process Engineering**, [S.L.], v. 58, p. 104758, fev. 2024. Elsevier BV. <http://dx.doi.org/10.1016/j.jwpe.2023.104758-104773>.

CERVANTES, J. *et al.* A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, v. 408, p. 189–215, 30 set. 2020.

CHAN, J. Y.-L. *et al.* Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. **Mathematics**, v. 10, n. 8, p. 1283, 12 abr. 2022.

CHENG, T. *et al.* Forecasting of wastewater treatment plant key features using deep learning-based models: A case study. **IEEE Access**, v. 8, p. 184475–184485, 2020.

CHOI, S. *et al.* Proactive Prediction of Total Volatile Fatty Acids Concentration in Multiple Full-Scale Food Waste Anaerobic Digestion Systems Using Substrate Characteristics with Machine Learning and Feature Analysis. **Waste and Biomass Valorization**, v. 14, n. 2, p. 593–608, 1 fev. 2023.

CHONG, S. *et al.* The performance enhancements of upflow anaerobic sludge blanket (UASB) reactors for domestic sludge treatment – A State-of-the-art review. **Water Research**, v. 46, n. 11, p. 3434–3470, jul. 2012.

CINAR, S. Ö.; CINAR, S.; KUÇTA, K. Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process. **Fermentation** **2022, Vol. 8, Page 65**, v. 8, n. 2, p. 65, 30 jan. 2022.

COROMINAS, LL. *et al.* Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. **Environmental Modelling & Software**, v. 106, p. 89–103, ago. 2018.

CSÁJI, B. C. **Approximation with Artificial Neural Networks**. [s.l.] Eötvös Loránd University, 2001a.

CSÁJI, B. C. **Approximation with Artificial Neural Networks**. MSc thesis—Hungria: Eötvös Loránd University, jun. 2001b.

CUI, P.; ZHOU, X.; ZHANG, Y. The Feasibility Study of Cotton Pulp Wastewater Treatment with IC Anaerobic Reactor. **Procedia Environmental Sciences**, v. 11, p. 686–692, 2011.

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.

DEE, D. P. A pragmatic approach to model validation. Em: [s.l: s.n.]. p. **1–13**.

DEVELOPERS, T. **TensorFlow**. Disponível em: <<https://zenodo.org/record/7916447>>. Acesso em: 16 jul. 2023.

DHI. **WEST Models Guide**. Disponível em: <<http://bit.ly/3QrpRI0>>. Acesso em: 21 maio. 2023.

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, 1 dez. 2008.

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, 23 dez. 2008.

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, 6 jul. 2020.

ELDUAYEN-ECHAVE, B. *et al.* Inclusion of shear rate effects in the kinetics of a discretized population balance model: Application to struvite precipitation. **Water Research**, v. 200, p. 117242, 15 jul. 2021.

ENVIRONMENTAL PROTECTION AGENCY. **Total Nitrogen**. Disponível em: <<https://bit.ly/46FhAiN>>. Acesso em: 15 jul. 2023.

ENVIRONMENTAL PROTECTION AGENCY. **Denitrifying Filters - Wastewater Management Fact Sheet**. Disponível em: <<https://bit.ly/3s5DWv6>>. Acesso em: 15 jul. 2023.

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, jun. 2018.

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, 30 ago. 2021.

FARRAR, D. E.; GLAUBER, R. R. Multicollinearity in Regression Analysis: The Problem Revisited. **The Review of Economics and Statistics**, v. 49, n. 1, p. 92, fev. 1967.

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.

FLORES, V.; LEIVA, C. A Comparative Study on Supervised Machine Learning Algorithms for Copper Recovery Quality Prediction in a Leaching Process. **Sensors**, v. 21, n. 6, p. 2119, 17 mar. 2021.

FONDRIEST ENVIRONMENTAL INC. **Turbidity, Total Suspended Solids & Water Clarity**. Disponível em: <<https://www.fondriest.com/environmental-measurements/parameters/water-quality/turbidity-total-suspended-solids-water-clarity/>>. Acesso em: 8 ago. 2023.

FUCK, João Vitor Rios *et al.* Predicting effluent quality parameters for wastewater treatment plant: a machine learning-based methodology. **Chemosphere**, [S.L.], v. 352, p. 141472, mar. 2024. Elsevier BV.
<http://dx.doi.org/10.1016/j.chemosphere.2024.141472>–141491.

FUKUSHIMA, K. Cognitron: A self-organizing multilayered neural network.

Biological Cybernetics, v. 20, n. 3–4, p. 121–136, 1975.

GIOKAS, D. L. *et al.* Systematic analysis of the operational response of activated sludge process to variable wastewater flows. A case study. **Clean Technologies and Environmental Policy** **2002 4:3**, v. 4, n. 3, p. 183–190, 1 nov. 2002.

GONÇALVES, R. F.; LUDUVICE, M.; SPERLING, M. VON. Sludge thickening and dewatering. Em: ANDREOLI, C. V.; SPERLING, M. VON; FERNANDES, F. (Eds.). **Sludge Treatment and Disposal**. Londres, Reino Unido: IWA Publishing, 2007. v. 6p. 76–119.

GUANG, H. *et al.* **Chemical Oxygen Demand Soft-Measurement Method via Long Short-Term Memory Network**. 2019 Chinese Automation Congress (CAC). **Anais...IEEE**, nov. 2019.

GUO, H. NAN *et al.* Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: A review. **Bioresource Technology**, v. 319, p. 124114, 1 jan. 2021.

GUPTA, R. *et al.* Review of explainable machine learning for anaerobic digestion. **Bioresource Technology**, v. 369, p. 128468, 1 fev. 2023.

HAMED, S. K.; AB AZIZ, M. J.; YAAKUB, M. R. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. **Heliyon**, v. 9, n. 10, p. e20382, 1 out. 2023.

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

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, 7 fev. 2023.

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

HASTIE, T.; FRIEDMAN, J.; TIBSHIRANI, R. **The Elements of Statistical Learning**. New York, NY: Springer New York, 2001.

HENZE, M. *et al.* **Wastewater treatment: biological and chemical processes**. 3. ed. Berlin : Springer Berlin Heidelberg, 2001. v. 1

HIET WONG, C.; BARTON, G. W.; BARFORD, J. P. The nitrogen cycle and its application in wastewater treatment. Em: MARA, D.; HORAN, N. (Eds.). **Handbook of Water and Wastewater Microbiology**. 1. ed. London: Elsevier, 2003. v. 1p. 427–439.

HOLMES, D. E.; DANG, Y.; SMITH, J. A. Nitrogen cycling during wastewater treatment. Em: **Advances in Applied Microbiology**. [s.l.] Academic Press, 2019. v. 106p. 113–192.

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

HOSSAIN, S. M. Z. *et al.* Soft-computing modeling and multiresponse optimization for nutrient removal process from municipal wastewater using microalgae. **Journal of Water Process Engineering**, v. 45, p. 102490, 1 fev. 2022.

HOYOS, N. L. M. **Análise comparativa do desempenho de reator anaeróbio híbrido e reator de manto de lodo de fluxo ascendente (UASB) aplicados ao tratamento de esgoto sanitário**. Dissertação (Mestrado em Recursos Hídricos e

Saneamento Ambiental)—Porto Alegre: Universidade Federal do Rio Grande do Sul, 2016.

JAFAR, R. *et al.* Predicting Effluent Quality in Full-Scale Wastewater Treatment Plants Using Shallow and Deep Artificial Neural Networks. **Sustainability**, v. 14, n. 23, p. 15598, 23 nov. 2022.

JOHNSON, P. M. *et al.* Using machine learning to examine freight network spatial vulnerabilities to disasters: A new take on partial dependence plots. **Transportation Research Interdisciplinary Perspectives**, v. 14, p. 100617, jun. 2022.

NEVES, J. *et al.* **Aplicação de modelos preditivos de machine learning para parâmetros de qualidade em estações de tratamento de efluentes**. 34^o Congresso Nacional de Saneamento e Meio Ambiente da Associação dos Engenheiros – AESabesp – FENASAN. **Anais...**São Paulo: SABESP, 2023. Disponível em: <http://evolvedoc.com.br/aesabesp/detalhes-4911_aplicacao-de-modelos-preditivos-de-machine-learning-para-parametros-de-qualidade-em-estacoes-de-tratamento-de-efluentes>. Acesso em: 11 dez. 2023

KALNINS, A. Multicollinearity: How common factors cause Type 1 errors in multivariate regression. **Strategic Management Journal**, v. 39, n. 8, p. 2362–2385, 1 ago. 2018.

KHAN, W. A. *et al.* Machine learning facilitated business intelligence (Part II): Neural networks optimization techniques and applications. **Industrial Management and Data Systems**, v. 120, n. 1, p. 128–163, 10 jan. 2020.

KIM, Y.; OH, S. Machine-learning insights into nitrate-reducing communities in a full-scale municipal wastewater treatment plant. **Journal of Environmental Management**, v. 300, p. 113795, 15 dez. 2021.

KINGMA, D. P.; BA, J. Adam: A Method for Stochastic Optimization. **3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings**, 22 dez. 2014.

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. Anais...: ICML'96.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996.

KOTSIANTIS, S. B.; KANELLOPOULOS, D.; PINTELAS, P. E. Data Preprocessing for Supervised Learning. 28 dez. 2007.

LATIF, M. A. *et al.* Integrated application of upflow anaerobic sludge blanket reactor for the treatment of wastewaters. **Water Research**, v. 45, n. 16, p. 4683–4699, out. 2011.

LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. 12 dez. 2019.

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

LIPTON, Z. C. The Mythos of Model Interpretability. **Queue**, v. 16, n. 3, p. 31–57, jun. 2018.

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, nov. 2020.

LIU, Y. Z.; CHEN, Z. Prediction of biochemical oxygen demand with genetic algorithm-based support vector regression. **Water Quality Research Journal**, v. 58, n. 2, p. 87–98, 1 maio 2023.

LONG, F. *et al.* Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. **Water Research**, v. 199, p. 117182, 1 jul. 2021.

LOOS, R. *et al.* EU-wide monitoring survey on emerging polar organic contaminants in wastewater treatment plant effluents. **Water Research**, v. 47, n. 17, p. 6475–6487, nov. 2013.

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

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.

Lu, Zhou *et al.* The Expressive Power of Neural Networks: A View from the Width. **Advances in Neural Information Processing Systems**, vol. 30, set. 2017.

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, ago. 2022.

MAOR, G. *et al.* An FPGA implementation of stochastic computing-based LSTM. **Proceedings - 2019 IEEE International Conference on Computer Design, ICCD 2019**, p. 38–46, 1 nov. 2019.

MOLNAR, C. **Interpretable Machine Learning**. Disponível em: <<https://christophm.github.io/interpretable-ml-book/>>. Acesso em: 2 jul. 2023.

MULAS, M. *et al.* Predictive control of an activated sludge process: An application to the Viikinmäki wastewater treatment plant. **Journal of Process Control**, v. 35, p. 89–100, nov. 2015.

MUOIO, R. *et al.* Optimization of a large industrial wastewater treatment plant using a modeling approach: A case study. **Journal of Environmental Management**, v. 249, p. 109436, 1 nov. 2019.

MURLA, D. *et al.* Coordinated management of combined sewer overflows by means of environmental decision support systems. **Science of The Total Environment**, v. 550, p. 256–264, 15 abr. 2016.

NEAL, B. On the Bias-Variance Tradeoff: Textbooks Need an Update. 17 dez. 2019.

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

NIU, G. *et al.* 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, p. 121787, 20 ago. 2020.

O'BRIEN, M. *et al.* Model predictive control of an activated sludge process: A case study. **Control Engineering Practice**, v. 19, n. 1, p. 54–61, jan. 2011.

OMAR, M. *et al.* Review of Machine Learning Applications to the Modeling and Design Optimization of Switched Reluctance Motors. **IEEE Access**, v. 10, p. 130444–130468, 2022.

ONU, P.; MBOHWA, C.; PRADHAN, A. An analysis of the application of machine learning techniques in anaerobic digestion. **2023 International Conference on Control, Automation and Diagnosis, ICCAD 2023**, 2023.

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

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

PHILLIPS, P. J. *et al.* **Four principles of explainable artificial intelligence**. Gaithersburg, 29 set. 2021. Disponível em:
<<https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf>>

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, jan. 2015.

PISA, I. *et al.* Transfer learning in wastewater treatment plant control design: From conventional to long short-term memory-based controllers. **Sensors**, v. 21, n. 18, p. 6315, 21 set. 2021.

POULSEN, J.; LAURIDSEN, C. L. **Modelling of the new works at Umbilo Sewage Purification Works with the WEST-program**. Report—Aalborg: Aalborg University, 2005.

REBACK, J. *et al.* **pandas-dev/pandas: Pandas 1.2.3**. Disponível em:
<<https://zenodo.org/record/4572994>>. Acesso em: 21 maio. 2023.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **“Why Should I Trust You?”** Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais...**New York, NY, USA: ACM, 13 ago. 2016. Disponível em: <<https://dl.acm.org/doi/10.1145/2939672.2939778>>

ROCHA, M. A. G. DA. **Estudos da atividade metanogênica de lodos e da biodegradabilidade anaeróbia de efluentes de indústrias de bebidas**.

Dissertação (Mestrado em Engenharia Civil) — Recife: Universidade Federal de Pernambuco, 2003.

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, 17 jul. 2020.

SALLES, R. *et al.* Prediction of Key Variables in Wastewater Treatment Plants Using Machine Learning Models. **Proceedings of the International Joint Conference on Neural Networks**, v. 2022- July, 2022.

SANTOS, M. S. DOS. **Cervejas e Refrigerantes**. Disponível em: <<https://bit.ly/46INNWy>>. Acesso em: 29 jun. 2022.

SARGENT, R. G. Verification and validation of simulation models. **Journal of Simulation**, v. 7, n. 1, p. 12–24, 19 fev. 2013.

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

SCIKIT-LEARN DEVELOPERS. **sklearn.inspection.permutation_importance — scikit-learn 1.3.0 documentation**. Disponível em: <<https://bit.ly/46SAp1U>>. Acesso em: 2 jul. 2023a.

SCIKIT-LEARN DEVELOPERS. **sklearn.inspection.PartialDependenceDisplay — scikit-learn 1.3.0 documentation**. Disponível em: <<https://bit.ly/3SeVTSr>>. Acesso em: 2 jul. 2023b.

SCIKIT-LEARN DEVELOPERS. **Permutation feature importance — scikit-learn 1.3.0 documentation**. Disponível em: <<https://bit.ly/45OilEZ>>. Acesso em: 16 jul. 2023c.

SELVARAJ, R. *et al.* Neural networks for the identification of MSF desalination plants. **Desalination**, v. 101, n. 2, p. 185–193, abr. 1995.

SHAMSHIRBAND, S. *et al.* Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters.

Engineering Applications of Computational Fluid Mechanics, v. 13, n. 1, p. 91–101, 1 jan. 2019.

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, ago. 2020.

SIMATE, G. S. *et al.* The treatment of brewery wastewater for reuse: State of the art. **Desalination**, v. 273, n. 2–3, p. 235–247, jun. 2011.

SINGH, K. P. *et al.* Experimental design and response surface modeling for optimization of Rhodamine B removal from water by magnetic nanocomposite. **Chemical Engineering Journal**, v. 165, n. 1, p. 151–160, nov. 2010.

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, fev. 2023.

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, 30 mar. 2020.

TAKACS, I.; PATRY, G.G.; NOLASCO, D. A dynamic model of the clarification-thickening process. **Water Research**, v. 25, n. 10, p. 1263-1271, out. 1991.

TAK, B. *et al.* Optimization of color and COD removal from livestock wastewater by electrocoagulation process: Application of Box–Behnken design (BBD). **Journal of Industrial and Engineering Chemistry**, v. 28, p. 307–315, ago. 2015.

TAO, X. *et al.* Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. **Information Sciences**, v. 487, p. 31–56, 1 jun. 2019.

TCHOBANOGLIOUS, G. *et al.* **Tratamento de efluentes e recuperação de recursos**. 5a. ed. Porto Alegre: AMGH Editora LTDA., 2016.

TENSORFLOW CORE. **Keras: The high-level API for TensorFlow**. Disponível em: <<https://www.tensorflow.org/guide/keras>>. Acesso em: 16 jul. 2023.

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.

TROYANSKAYA, O. *et al.* Missing value estimation methods for DNA microarrays . **Bioinformatics**, v. 17, n. 6, p. 520–525, 1 jun. 2001.

TUKEY, J. W. **Exploratory Data Analysis**. London: Pearson Education Inc., 1977.

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.

VON SPERLING, M. **Urban Wastewater Treatment in Brazil**. Disponível em: <<https://bit.ly/3Qt25Vy>>. Acesso em: 17 maio. 2022.

VON SPERLING, M. **Introdução à qualidade das águas e ao tratamento de esgotos**. 4 ed. Belo Horizonte: Editora UFMG, 2017.

VON WANGENHEIM, Aldo. **Deep Learning**: aprendizado por transferência e ajuste fino. Aprendizado por Transferência e Ajuste Fino. 2018. Disponível em: <https://lapix.ufsc.br/ensino/visao/visao-computacionaldeep-learning/deep-learningaprendizado-por-transferencia-e-ajuste-fino/>. Acesso em: 07 out. 2025.

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, ago. 2021.

WANG, D. *et al.* Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. **Journal of Environmental Management**, v. 301, p. 113941, 1 jan. 2022.

WANG, T. *et al.* Insights into sludge granulation during anaerobic treatment of high-strength leachate via a full-scale IC reactor with external circulation system. **Journal of Environmental Sciences**, v. 64, p. 227–234, fev. 2018.

WANG, Y.; HUNTINGTON, T.; SCOWN, C. D. Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste. **ACS Sustainable Chemistry and Engineering**, v. 9, n. 38, p. 12990–13000, 27 set. 2021.

WANG, Z. *et al.* A deep learning based dynamic COD prediction model for urban sewage. **Environmental Science: Water Research & Technology**, v. 5, n. 12, p. 2210–2218, 2019.

WAQAS, S. *et al.* SVM and ANN Modelling Approach for the Optimization of Membrane Permeability of a Membrane Rotating Biological Contactor for Wastewater Treatment. **Membranes**, v. 12, n. 9, p. 821, 23 ago. 2022.

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, 28 out. 2015.

WU, X. *et al.* Simulation and optimization of a coking wastewater biological treatment process by activated sludge models (ASM). **Journal of Environmental Management**, v. 165, p. 235–242, 1 jan. 2016.

XIE, Y. D. *et al.* Towards the formulation of rural sewage discharge standards in China. **Science of The Total Environment**, v. 759, p. 143533, 10 mar. 2021.

XU, Y. *et al.* Artificial Intelligence-Assisted Prediction of Effluent Phosphorus in a Full-Scale Wastewater Treatment Plant with Missing Phosphorus Input and Removal Data. **ACS ES and T Water**, 2022a.

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, maio 2022b.

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.

YAQUB, M. *et al.* Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network. **Journal of Water Process Engineering**, v. 37, p. 101388, 1 out. 2020.

YDATA LABS INC. **pandas-profiling** 3.6.6. Disponível em: <<https://pypi.org/project/pandas-profiling/#files>>. Acesso em: 21 maio. 2023.

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

ZAGHLOUL, M. S. *et al.* Development of an ensemble of machine learning algorithms to model aerobic granular sludge reactors. **Water Research**, v. 189, p. 116657, fev. 2021.

ZHANG, Y. *et al.* Deep learning based data-driven model for detecting time-delay water quality indicators of wastewater treatment plant influent. **Chemical Engineering Journal**, v. 467, p. 143483, 1 jul. 2023.

ZHAO, G. YAO *et al.* Machine learning estimation of biodegradable organic matter concentrations in municipal wastewater. **Journal of Environmental Management**, v. 323, p. 116191, 1 dez. 2022.

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, jan. 2020.

ZHOU, P. *et al.* Multi-step ahead prediction of hourly influent characteristics for wastewater treatment plants: a case study from North America. **Environmental Monitoring and Assessment**, v. 194, n. 5, p. 1–14, 1 maio 2022.

ZHOU, P. *et al.* Online machine learning for stream wastewater influent flow rate prediction under unprecedented emergencies. **Frontiers of Environmental Science & Engineering**, v. 17, n. 12, p. 152, 20 dez. 2023.

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

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.

ZOU, J. *et al.* Nitrogen removal and biofilm structure affected by COD/NH₄⁺–N in a biofilter with porous sludge-ceramsite. **Separation and Purification Technology**, v. 94, p. 9–15, jun. 2012.