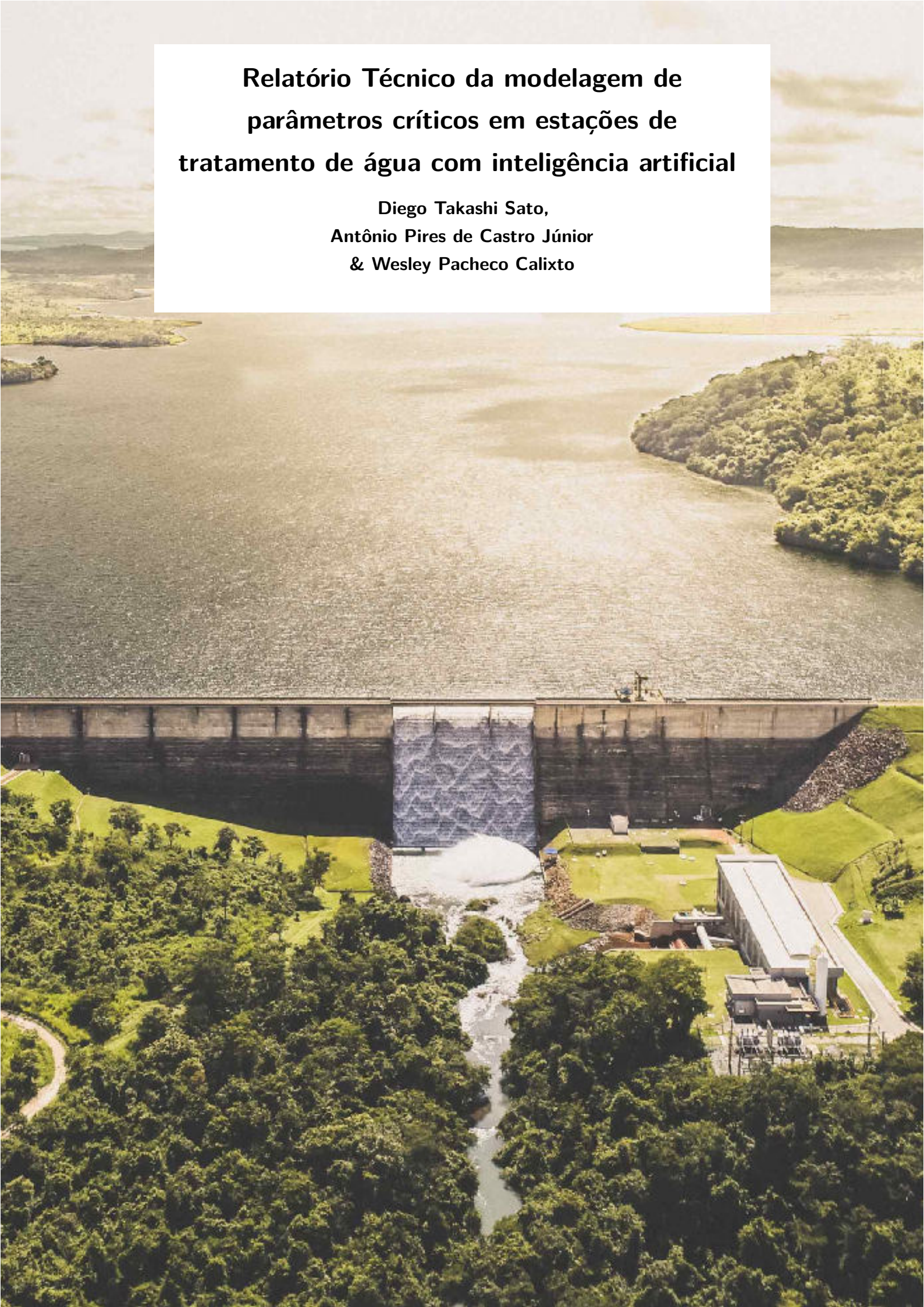


Relatório Técnico da modelagem de parâmetros críticos em estações de tratamento de água com inteligência artificial

**Diego Takashi Sato,
Antônio Pires de Castro Júnior
& Wesley Pacheco Calixto**



Diego Takashi Sato, Antônio Pires de Castro Júnior & Wesley Pacheco Calixto

Relatório Técnico da modelagem de parâmetros críticos em estações de tratamento de água com inteligência artificial

Relatório Técnico em conformidade com as
normas ABNT e apresentado ao Mestrado
Profissional em Tecnologia, Gestão e Susten-
tabilidade como Produto Técnico.

Instituto Federal de Educação, Ciência e Tecnologia de Goiás – IFG
Programa de Pós-Graduação Profissional em Tecnologia, Gestão e Sustentabilidade
Campus Goiânia

Brasil
Goiânia, 20 de fevereiro de 2025

Agradecimentos

Agradecemos sinceramente a todos que, de alguma forma, contribuíram para a realização deste trabalho. Reconhecemos o apoio das instituições envolvidas, especialmente à Saneamento de Goiás S/A, pela disponibilização dos dados indispensáveis ao desenvolvimento deste estudo. Estendemos nossa gratidão ao Instituto Federal de Goiás (IFG) pela oportunidade de aprimorar nossos conhecimentos e enriquecer nossa experiência como pesquisadores. A cada indivíduo que, direta ou indiretamente, colaborou para a concretização deste projeto, registramos nosso agradecimento.

Resumo

Este relatório detalha as etapas para a aplicação de modelos de regressão utilizando inteligência artificial, com o objetivo de estimar as dosagens de compostos químicos, principalmente derivados de cloro e flúor, em estação de tratamento de água. São descritos os critérios para a seleção da estação de tratamento de água, bem como os parâmetros necessários para a construção dos modelos, utilizando algoritmos como floresta randômica, XGBoost e Redes Neurais Artificiais *Multilayer Perceptron*, em suas configurações padrão. Propostas de segmentação dos dados para treinamento, validação interna e validação externa são apresentadas. Adicionalmente, o relatório aborda ajustes de otimização dos hiperparâmetros dos modelos preditivos utilizados, indicando melhorias no desempenho. Em todos os casos, são fornecidos parâmetros para avaliar a assertividade, como erro médio absoluto, coeficiente de determinação e raiz do erro médio quadrático. Por fim, a validação é realizada por meio da análise de desempenho dos modelos propostos.

Palavras-chaves: Modelagem de sistemas, Tratamento de água, Inteligência artificial, Otimização de processos, Sustentabilidade.

Lista de ilustrações

Figura 1 – Fluxograma da metodologia proposta.	33
Figura 2 – Fluxo hipotético de requisitos da Estação de Tratamento de Água (ETA) para inclusão no estudo	34
Figura 3 – Mapa georreferenciado da captação e ETA de Brazabranes.	35
Figura 4 – Estação de tratamento de água de Brazabranes, Goiás/Brasil.	36
Figura 5 – Equipamentos, ambientes e processo: (a) medidor de turbidez em NTU, (b) tanque de solubilização de cloreto férrico em batelada, (c) floco-decantador, (d) tela de retenção de sólidos flutuantes/grosseiros na saída do floco-decantador, antes de entrar no filtro de fluxo ascendente, (e) calha de coleta de água superficial no filtro-decantador e (f) medidor automatizado de cloro, flúor, M_T e M_{pH}	37
Figura 6 – Comparação de métodos para preenchimento de dados ausentes.	39
Figura 7 – Divisão dos dados tratados para validação do modelo proposto.	41
Figura 8 – Valores reais do sistema e previsões do modelo com amostra de 2022: (a) cloro sem otimização de hiperparâmetros, (b) cloro com otimização de hiperparâmetros, (c) flúor sem otimização de hiperparâmetros e (d) flúor com otimização de hiperparâmetros.	47
Figura 9 – Valores reais do sistema e previsões do modelo com amostra de 2024: (a) cloro sem otimização de hiperparâmetros, (b) cloro com otimização de hiperparâmetros, (c) flúor sem otimização de hiperparâmetros e (d) flúor com otimização de hiperparâmetros.	48
Figura 10 – Distribuição dos erros dos modelos preditivos para: (a) cloro e (b) flúor.	51
Figura 11 – Resíduos dos modelos preditivos ao longo das observações: (a) XGBoost sem otimização cloro, (b) XGBoost otimizado cloro, (c) RNA-MLP sem otimização flúor e (d) RNA-MLP otimizado flúor.	52
Figura 12 – Dispersão com valores reais $\times$ previstos: (a) RNA-MLP sem otimização cloro, (b) RNA-MLP otimizado cloro, (c) Floresta Randômica (FR) sem otimização flúor e (d) FR otimizado flúor.	53

Lista de tabelas

Tabela 1 – Correlação entre algumas variáveis observadas e a dosagem de cloro e flúor.	38
Tabela 2 – Critérios de avaliação dos métodos usados para preencher dados ausentes.	39
Tabela 3 – Estatísticas descritivas dos parâmetros observados.	40
Tabela 4 – Valores de hiperparâmetros e avaliações para os métodos RNA-MLP, FR e XGBoost.	44
Tabela 5 – Valores otimizados de hiperparâmetros e avaliações para os métodos RNA-MLP, FR e XGBoost.	45
Tabela 6 – Validação externa comparando os resultados do modelo com os valores reais coletados no sistema.	47
Tabela 7 – Validação externa comparando os resultados dos modelos otimizados com os valores reais coletados no sistema.	48
Tabela 8 – Resultados das métricas de avaliação para os modelos preditivos de cloro e flúor.	49

Lista de abreviaturas e siglas

k -FCV	<i>k-fold cross-validation</i>
R^2	Coeficiente de determinação
ADN	Ácido desoxirribonucleico
AHA	Ácido haloacéticos
AM	Aprendizado de máquina
CRL	Cloro residual livre
CSV	<i>Comma-separeted values</i>
EMA	Erro médio absoluto
EMQ	Erro médio quadrático
ETA	Estação de tratamento de água
FR	Floresta Randômica
GBC	<i>Gradient Boosting Classifier</i>
GBT	<i>Gradient Boost Tree</i>
IA	Inteligência artificial
ICF	Índice de coliformes fecais
ICT	Índice de coliformes totais
IQA	Índice de Qualidade da Água
MLP	<i>Multi-Layer Perceptron</i>
NTU	<i>Nephelometric Turbidity Unit</i>
OMS	Organização Mundial de Saúde
ONU	Organização das Nações Unidas
POA	Processo de oxidação avançado
REMQ	Raiz do erro médio quadrático

RLM	Regressão linear múltipla
RNA-MLP	Redes Neurais Artificiais - <i>Multi Layer Perceptron</i>
RNA	Redes neurais artificiais
RNP	Redes neurais profundas
RVS	Regressão de vetor de suporte
SODIS	Desinfecção solar da água
STL	<i>Seasonal-Trend decomposition procedure based on Loess</i>
SVM	Máquinas de vetor de suporte
THM	Trihalometanos
UV	Ultravioleta
XGBoost	<i>Extreme Gradient Boosting</i>

Lista de símbolos

α	Paciência
γ	Ganho mínimo
Δ_{min}	Melhoria mínima na validação
ϵ	Número máximo de épocas
η	Taxa de aprendizado
μ	Média
ρ	Correlação
σ	Desvio padrão
c_{mf}	Crescimento mínimo de folhas
C_{ST}	Características amostradas aleatoriamente em cada árvore durante o treinamento
D_{Cl}	Dosagem de cloro
D_F	Dosagem de flúor
D_M	Profundidade máxima das árvores de decisão durante o treinamento
D_S	Dosagem de sulfato
M_{Cl}	Medições de Cl
M_{cw}	Peso mínimo de instância
M_F	Medições de F
M_{pH}	Medições de pH
M_T	Temperatura da água tratada
n_a	Número máximo de árvores
N_{RA}	Nível do reservatório apoiado
N_{rc}	Número de variáveis escolhidas aleatoriamente
N_{RE}	Nível do reservatório elevado

N_{rn}	Número de rodadas do algoritmo de treinamento
pH	Potencial hidrogeniônico
S_{sa}	Observações amostradas aleatoriamente em cada árvore durante o treinamento
S_z	Neurônios na camada oculta
T_{ab}	Turbidez da água bruta
t_{maf}	Tamanho mínimo da amostra por folha
T	Temperatura
V_{max}	Valor máximo
V_{min}	Valor mínimo
V_{ab}	Vazão de água bruta
V_{at}	Vazão de água tratada
W_D	Decaimento de peso

Sumário

I	PREPARAÇÃO DA PESQUISA	19
1	INTRODUÇÃO	21
II	BASES TEÓRICAS	23
2	BASES TEÓRICAS	25
2.1	Fundamentos do processo de cloração	25
2.2	Fundamentos do processo de fluoretação	27
2.3	Ciência de dados e aprendizado de máquina	28
2.4	Técnicas de tratamento de dados em séries temporais	30
III	APLICAÇÃO DO MÉTODO	31
3	APLICAÇÃO DO MÉTODO	33
3.1	Contextualização	33
3.2	Seleção da ETA adequada	33
3.3	Pré-processamento de dados	36
3.3.1	Utilização de séries temporais	36
3.3.2	Tratamento de valores atípicos e lacunas	38
3.3.3	Análise estatística descritiva	39
3.3.4	Validação da operação da ETA em conformidade com as normas	39
3.4	Aplicação de algoritmos preditivos	40
3.4.1	Segmentação dos dados	40
3.4.2	Métricas de avaliação dos modelos	41
3.4.3	Construção dos modelos sem ajuste de hiperparâmetros	42
3.5	Otimização de hiperparâmetros	44
3.5.1	Construção dos modelos com otimização de hiperparâmetros	45
3.6	Validação externa dos modelos	46
3.6.1	Avaliação dos modelos preditivos	46
3.6.2	Avaliação estatística dos modelos preditivos	49
4	CONCLUSÃO	55
	REFERÊNCIAS	57

APÊNDICES	63
APÊNDICE A – SOLICITAÇÃO DOS DADOS FUNDADO NA LEI DE ACESSO À INFORMAÇÃO	65
APÊNDICE B – CÓDIGO DO TRATAMENTO DE DADOS	79
APÊNDICE C – CÓDIGO DA CONSTRUÇÃO DOS MODELOS DE IA	83
APÊNDICE D – CÓDIGO DE VALIDAÇÃO DOS MODELOS	87
APÊNDICE E – CÓDIGO PARA ANÁLISE ESTATÍSTICA DOS MODELOS	91
Índice	95

Parte I

Preparação da Pesquisa

1 INTRODUÇÃO

Este relatório foi elaborado utilizando $\text{\LaTeX}$, especificamente a classe `abntex2` e o pacote `abntex2cite`. Estas ferramentas foram selecionadas devido à sua conformidade com a norma ABNTNBR10719:2015, que define os padrões para a apresentação de relatórios técnico-científicos. Tal escolha possibilita a entrega de material padronizado e didático, facilitando sua reutilização em outros trabalhos acadêmicos. Este relatório integra a dissertação intitulada **Modelagem, simulação e otimização de processos em estação de tratamento de água utilizando inteligência artificial**, defendida no âmbito do Programa de Pós-Graduação *Stricto Sensu* em Tecnologia, Gestão e Sustentabilidade do Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG).

A água tratada é fundamental para a saúde pública, sendo submetida a processos físico-químicos destinados à remoção de impurezas, microrganismos e poluentes, garantindo conformidade com os padrões regulatórios definidos pela Declaração Universal dos Direitos da Água. A desinfecção, um dos principais objetivos do tratamento, utiliza métodos como a cloração para eliminar agentes patogênicos e impurezas. Contudo, a formação de subprodutos prejudiciais à saúde, como os trihalometanos (THM) e ácido haloacéticos (AHA) ([HRUDEY, 2009](#)), demanda a implementação de processos complementares, como aeração e uso de carvão ativado ([OCAMPO-RODRÍGUEZ et al., 2022](#)). Adicionalmente, o controle rigoroso da dosagem de cloro é indispensável para evitar efeitos citotóxicos ([SLAUGHTER et al., 2019](#)). A fluoretação da água, introduzida devido aos benefícios à saúde bucal, reduziu significativamente os casos de cárie ([MURRAY; World Health Organization et al., 1986](#)), mas também levantou preocupações sobre a fluorose dentária, exigindo um equilíbrio criterioso para otimizar os benefícios e mitigar efeitos adversos ([NEWBRUN, 1989](#)).

Diversos estudos indicam avanços no uso de tecnologias e modelos preditivos para otimizar o desempenho das ETA. [Zhang et al. \(2019\)](#) e [Solaimany-Aminabad, Maleki e Hadi \(2013\)](#) empregam modelos para prever e gerenciar condições de estresse e qualidade da água, utilizando redes neurais artificiais (RNA) para simular comportamentos não lineares. [Pinto et al. \(2009\)](#) aplicam mineração de dados e algoritmos, como *K*-Means, para solucionar problemas específicos, como altos níveis de magnésio e turbidez. [Librantz, Santos e Dias \(2018\)](#) utilizam RNA para otimizar a dosagem de cloro, enquanto [Khan et al. \(2021\)](#) empregam algoritmos de inteligência artificial (IA) na predição e classificação do Índice de Qualidade da Água (IQA). [Wang et al. \(2021\)](#) demonstram a resiliência de modelos preditivos na otimização de processos industriais, e [More e Wolkersdorfer \(2023\)](#) ressaltam o uso de *multi-layer perceptron* (MLP), RNA, *deep neural network* (DNN),

FR e *Gradient Boost Tree* (GBT) para prever parâmetros críticos no tratamento de águas subterrâneas. Estes estudos apresentam a importância das tecnologias avançadas na melhoria da eficiência, redução de custos e gestão sustentável nas ETA.

O objetivo principal deste relato é apresentar o modelo computacional preditivo de regressão supervisionado, guiado por séries temporais, que possibilite a otimização dos processos de cloração e fluoretação em determinada estação de tratamento de água superficial. Adicionalmente, os objetivos específicos incluem: i) alcançar a redução dos custos operacionais e da qualidade da água, ii) obter valores de dosagens de produtos químicos críticos para o tratamento de água e seguros para a saúde da população, iii) aprimorar o processo para garantir a qualidade do produto final e iv) contribuir para o alcance do Objetivo de Desenvolvimento Sustentável n.º 6 (ODS 6), estabelecido na Agenda 2030 pela Organização das Nações Unidas (ONU).

Este relatório enfatiza a relevância de modelos computacionais preditivos, com base em séries temporais, para otimizar processos em estações de tratamento de água, como cloração e fluoretação. Estes modelos oferecem soluções ágeis e de baixo custo operacional, em contraste com alterações físicas na planta industrial. Focada em aplicações práticas, especialmente em países em desenvolvimento, a pesquisa destaca o compromisso com a sustentabilidade ambiental e a gestão eficiente de recursos hídricos.

A aplicação do método proposto utilizou dados reais da ETA do município de Brazabrantes, operada pela Saneamento de Goiás S/A (Saneago). Para assegurar a eficiência dos modelos de regressão, torna-se necessária a coleta sistemática das variáveis mais relevantes, estruturada em séries temporais com intervalos regulares. Isto exige que a ETA disponha de dispositivos automatizados para coleta e armazenamento contínuos dos dados, garantindo qualidade e consistência das informações utilizadas no modelo. Além disto, a metodologia proposta pode ser adaptada para a predição de outros parâmetros de interesse em ETA.

Parte II

Bases Teóricas

2 Bases teóricas

2.1 Fundamentos do processo de cloração

Dentro do conjunto de procedimentos realizados na estação de tratamento de água, a desinfecção é uma das fases mais importantes. A remoção de agentes patogênicos da água destinada ao abastecimento da população é o elemento principal para assegurar a preservação da saúde (HOFKES, 1981; NAMOR, 2007). De acordo com Ocampo-Rodríguez et al. (2022), nos últimos anos têm sido empregados diversos métodos para a desinfecção da água, agrupados predominantemente em quatro categorias: i) métodos químicos que envolvem o uso de oxidantes potentes, ii) métodos térmicos como a pasteurização, iii) saturação utilizando íons de metais não ferrosos como prata, cobre e propriedades bactericidas e iv) métodos físicos que incluem a aplicação de radiação ultravioleta, ultrassom ou radiações ionizantes (NEGOITESCU; TOKAR, 2012). Além destes, existem outros métodos, como o uso de ozônio (FAUST; ALY, 1998), e a ultrafiltração utilizando membranas (LAINÉ; VIAL; MOULART, 2000). A desinfecção de águas destinadas ao consumo humano busca reduzir significativamente a presença de microrganismos, em vez de atingir esterilização completa (FAUST; ALY, 1998).

Segundo Miklos et al. (2018), os processos de oxidação avançados (POA) podem ser categorizados em cinco grupos principais: i) com base em ozônio, ii) com base em radiação ultravioleta, iii) processo avançado de oxidação eletroquímica, iv) processo avançado de oxidação catalítica e v) processo avançado de oxidação física. Todos estes métodos envolvem a geração *in loco* do agente oxidante, que então reage com os contaminantes. Esta característica de produção *in loco* demanda consumo elevado de energia e utiliza ampla gama de processos físico-químicos para a remoção eficiente de matéria orgânica, minimizando a formação de derivados de subprodutos do cloro. Em todos os POA, o objetivo é gerar o oxidante *in loco* e promover a reação com os contaminantes alvo (KIM; ZOH, 2016).

O estudo de Ocampo-Rodríguez et al. (2022) descreve o uso de cloro e seus derivados para a desinfecção da água destinada ao abastecimento é amplamente empregado em escala global, presente em mais de dois terços dos sistemas de tratamento de água do planeta. Esta preferência advém de seu custo financeiro e facilidade de manuseio (CHHETRI et al., 2018). A eficiência deste método de tratamento pode ser avaliada pela redução ou eliminação da presença de microrganismos, principalmente *Escherichia coli* e *Enterococcus* spp, que servem como indicadores de contaminação fecal. Entretanto, os desinfetantes à base de cloro e seus derivados reagem com a matéria orgânica e inorgânica presentes na água, gerando derivados de subprodutos do cloro que são prejudiciais à saúde e desafiadores de

eliminar devido à sua diversidade química, física e biológica. Estas substâncias, associadas a doenças como o câncer, são principalmente produzidas pela reação do cloro com materiais orgânicos. Em águas com elevada concentração de compostos orgânicos, formam-se nitritos, cloratos, fenacetina, acetanilida, sulfanilamida, nitrobenzeno e quinonas, substâncias especialmente nocivas para crianças, com potencial de causar metemoglobinemia e/ou sulfemoglobinemia, condição conhecida como síndrome do bebê azul (FERNÍCOLA; AZEVEDO, 1981; LUKENS, 1987; SADEQ et al., 2008; JOHNSON, 2019). A remoção destes subprodutos pode ser alcançada por meio de processos como osmose inversa, troca iônica e adsorção. No entanto, o mecanismo mais comumente empregado é a prefiltração com carvão ativado e trocadores iônicos magnéticos (ZOU; TANG, 2019).

Para Ocampo-Rodríguez et al. (2022), o mecanismo de desinfecção microbiológica do cloro consiste em ocasionar danos às membranas cito-celulares, interrompendo as funções celulares e afetando o ácido desoxirribonucleico (ADN), além de ser absorvido pelos biofilmes e eliminar as bactérias interiores (SILVA et al., 2020). Entre as várias formas de oxidantes à base de cloro, o HClO destaca-se por sua capacidade de degradar proteínas e destruir as enzimas envolvidas no metabolismo de carboidratos, oxidando grupos sulfídricos. Por sua vez, o dióxido de cloro ClO₂ atua de maneira seletiva, permeabilizando as membranas celulares e interrompendo a síntese proteica, favorecendo a perda de nutrientes vitais e, conseqüentemente, levando à morte celular. A escolha pelo uso de cloro é justificada pela ampla eficiência na inativação de patógenos, na proteção da saúde humana e pelo seu efeito residual (DESTIANI; TEMPLETON, 2019).

A pesquisa conduzida por Slaughter et al. (2019) investiga o hipoclorito de sódio quando dissolvido em água, observando maior presença de HClO em relação ao íon hipoclorito, sendo este equilíbrio dependente do *pH* da solução. Quando em contato com mucosas, como as da boca, esôfago, estômago e trato respiratório, o HClO se converte rapidamente em ácido clorídrico e íons de oxigênio, resultando na morte de tecidos e induzindo intoxicação por citotoxicidade. A ingestão de hipoclorito de sódio em concentrações diluídas raramente é clinicamente relevante, embora possa causar desconfortos digestivos, como dores de estômago e vômitos (RHEE; BEAUMONT, 1990). Na ausência de antídoto específico é necessário o tratamento primário focado na redução dos sintomas observados. Diante destes aspectos, destaca-se a importância de utilizar o cloro como oxidante na quantidade adequada, visto que quantidade insuficiente não atende ao propósito de sanitização, enquanto em excesso pode resultar em envenenamento (MARTINEZ; LONG, 1995).

O processo de cloração, fundamental para garantir a qualidade microbiológica da água, apresenta variáveis críticas que precisam ser controladas de forma precisa para evitar riscos à saúde pública e desperdícios operacionais. Assim, torna-se relevante adotar ferramentas capazes de prever e otimizar os níveis de dosagem, permitindo a operação

segura e eficiente.

2.2 Fundamentos do processo de fluoretação

O flúor é frequentemente adicionado à água destinada ao consumo humano como medida de saúde pública para prevenir a cárie dentária. A prática de controlar a adição de flúor à água está estabelecida em vários países ([KAUR et al., 2021](#)). O primeiro indício da relação entre flúor e saúde dental surge em 1911, quando o dentista McKay observa que crianças em Colorado Springs/EUA, têm dentes manchados de marrom, enquanto as crianças das cidades vizinhas não têm estas manchas. McKay suspeita que a causa das manchas esteja na água que as crianças consomem. Em 1916, McKay e Black confirmam a hipótese ao analisarem a água de abastecimento de Colorado Springs, encontrando níveis elevados de flúor. Em 1928, McKay sugere que, além de causar as manchas, o flúor também pode proteger os dentes contra cáries ([MURRAY; World Health Organization et al., 1986](#)). Em 1931, o químico Petrey corrobora a ideia ao observar que as crianças com dentes mais manchados são as que têm menos cáries, pois bebem água com maior teor de flúor ([MULLEN, 2005](#)).

Desde então, vários estudos têm surgido, associando as doses seguras de flúor para a promoção da saúde bucal da população. Esta iniciativa é reconhecida e incentivada pela Organização Mundial de Saúde (OMS) como medida terapêutica cientificamente comprovada, eficiente e segura. A recomendação da fluoretação da água é apresentada na 25ª Assembleia Mundial de Saúde, em 1975, e aprovada por 148 membros ([MULLEN, 2005](#); [NEWBRUN, 1989](#)). Atualmente, vários países incorporam a fluoretação das águas destinadas ao abastecimento humano, incluindo Estados Unidos, Brasil, Austrália, Canadá, Espanha, Coreia do Sul e Nova Zelândia ([MULLANE et al., 2016](#); [GARBIN et al., 2017](#)). A prática da fluoretação é mantida, geralmente, com o objetivo de assegurar a conformidade por meio do heterocontrole exercido por entidades/instituições reguladoras e pela vigilância em saúde pública. Entretanto, a responsabilidade primária recai sobre os operadores dos sistemas de abastecimento público ([PERES; FERNANDES; PERES, 2004](#); [CHARROIS, 2010](#)).

O estudo de [Roy e Dass \(2013\)](#), realizado na Índia, evidencia diversos problemas associados ao consumo excessivo de flúor, frequentemente decorrentes do uso de fontes subterrâneas sem tratamento adequado e da ingestão de fontes alternativas, como chás. De modo geral, a exposição crônica ao flúor causa efeitos prejudiciais, incluindo deformações severas e permanentes nos ossos e articulações (osteoclerose). Ressalta-se que a fluorose é uma condição irreversível e não possui tratamento.

O estudo conduzido por [Levy et al. \(2023\)](#) compara a incidência de cárie em jovens adultos não expostos ao abastecimento de água fluoretada e que não passaram por

tratamentos dentais, com outros jovens adultos expostos à água fluoretada, seja com ou sem tratamentos dentais associados. Os autores observam correlação significativa entre a fluoretação obrigatória da água para consumo humano e a redução da necessidade de tratamentos dentários. De acordo com [Charrois \(2010\)](#), a melhoria das condições de higiene bucal e dentária na população tem levado à diminuição da adição de flúor na água potável em diversos países, mesmo sem considerar as mudanças nos hábitos alimentares. A OMS alerta que concentrações de flúor superiores a 1,5mg por litro podem levar ao desenvolvimento de fluorose dentária. A fluoretação da água de abastecimento público é medida de saúde pública de baixo custo, ampla abrangência populacional, segura e eficiente no controle e redução da doença cárie ([PERES; FERNANDES; PERES, 2004](#); [NEWBRUN, 1989](#); [CHARROIS, 2010](#)).

A fluoretação, embora eficiente na prevenção de doenças bucais, exige controle rigoroso para evitar concentrações inadequadas, que podem ser prejudiciais à saúde. Desta forma, o uso de técnicas avançadas de modelagem e predição é uma alternativa promissora para assegurar a dosagem correta e reforçar a eficiência operacional das estações de tratamento de água.

2.3 Ciência de dados e aprendizado de máquina

A ciência de dados é uma área multidisciplinar dedicada à extração de conhecimento e percepções a partir de conjuntos de dados com múltiplos atributos e interdependências ([AL-HAIJA, 2022](#)). Esta disciplina integra técnicas estatísticas, computacionais e de visualização ([BLEI; SMYTH, 2017](#)), abrangendo etapas que incluem a coleta, o pré-processamento e a análise de grandes volumes de dados provenientes de diversas fontes. Profissionais da área empregam ferramentas como Python e R, bem como bibliotecas especializadas, tais como `scikit-learn`, `caret` e `scipy`, em conjunto com métodos estatísticos avançados. Estes recursos permitem identificar padrões, tendências e correlações nos dados, promovendo decisões fundamentadas em evidências ([RIEPL et al., 2010](#)).

O aprendizado de máquina (AM) é a subárea da IA voltada ao desenvolvimento de algoritmos capazes de identificar padrões em dados de forma autônoma e aprimorar seu desempenho com o processamento de novos dados ([FEDORKO; KRÁL'; FEDORKO, 2022](#)). Em algoritmos de aprendizado supervisionado, os modelos são treinados com conjuntos de dados previamente rotulados, permitindo a generalização de padrões e a realização de previsões sobre novos dados ([GHAHRAMANI, 2015](#)). Por outro lado, algoritmos de aprendizado não supervisionado analisam dados não rotulados para identificar padrões intrínsecos, enquanto o aprendizado por reforço envolve agentes que tomam decisões sequenciais em ambientes dinâmicos com base em recompensas ou penalidades recebidas ([MORALES; ESCALANTE, 2022](#)). Estas abordagens são amplamente aplicadas

em tarefas como classificação, agrupamento, regressão e otimização, de acordo com suas especificidades (JOLLY, 2018).

Com a crescente disponibilidade de conjuntos de dados massivos e o avanço das técnicas de AM, incluindo redes neurais profundas, a capacidade de criar modelos preditivos precisos tem se destacado. Dentro do AM, as RNA, especialmente as MLP, são importantes. As MLP são arquiteturas de redes neurais profundas (CHANDRA et al., 2021), compostas por camadas de neurônios organizados em múltiplas camadas, incluindo: entrada, ocultas e saída (DAYHOFF, 1990). Cada conexão entre neurônios possui um peso associado e funções de ativação não lineares são aplicadas para introduzir não linearidades. Esta flexibilidade faz das MLP componentes fundamentais em diversas aplicações de AM. Apesar de sua eficiência, o treinamento de MLP pode ser desafiador em conjuntos de dados extensos, exigindo técnicas avançadas para otimizar seu desempenho em cenários práticos (LASHARI et al., 2019).

A aplicação de Redes Neurais Artificiais - *Multi Layer Perceptron* (RNA-MLP) na modelagem de tratamento e qualidade da água representa avanço considerável no campo da engenharia ambiental (ALIASHRAFI et al., 2021). As RNA-MLP são arquiteturas capazes de capturar relações difíceis em conjuntos de dados hidroambientais. Ao integrar dados de parâmetros de tratamento, características hidrológicas e indicadores de qualidade da água na RNA-MLP, é possível desenvolver modelos preditivos precisos para otimizar processos de tratamento, prever concentrações de poluentes e antecipar mudanças na qualidade da água (TAHERSHAMSI; MENHAJ; AHMADIAN, 2006; GOURMELON et al., 2010). Esta abordagem avançada oferece ferramenta valiosa, permitindo a tomada de decisão eficiente e proativa no monitoramento e controle dos sistemas de tratamento de água (JAYAWEERA; AZIZ, 2018; GHOBADI et al., 2022; SAFEER et al., 2022).

Em aprendizado de máquina, uma técnica amplamente utilizada devido à sua versatilidade em trabalhar tanto com regressão quanto com classificação é a FR (BREIMAN, 2001). Nesta abordagem, várias árvores de decisão são combinadas durante o treinamento, e cada árvore na floresta opera de forma independente. A combinação de suas previsões é empregada para aprimorar a precisão e o desempenho do modelo. No contexto do aprendizado supervisionado, a FR é frequentemente escolhida por ser método de propósito geral que não exige parametrização (JENA et al., 2022). Os dados de treinamento são subdivididos em subconjuntos menores, organizados em árvores, e os melhores são selecionados para formar novos agrupamentos. Este procedimento é realizado exaustivamente, reduzindo erros e minimizando a influência de ruídos e valores atípicos. A adoção desta técnica é apropriada ao lidar com conjuntos de dados não lineares (JENKINS; KOWALSKI; ALVARENGA, 2018).

O *Extreme Gradient Boosting* (XGBoost) também é técnica de aprendizado de máquina da categoria de métodos ensemble, usado em classificação e regressão. Utilizando

gradient boosting, o algoritmo constrói sequencialmente árvores para corrigir erros residuais, otimizando a função de perda. Esta abordagem, guiada por dados, é amplamente adotada na ciência de dados, proporcionando resultados eficientes (CHEN; GUESTRIN, 2016). O XGBoost utiliza paralelismo e otimização de recursos de memória, dividindo o conjunto de treinamento em árvores e ajustando os dados até que os cruzamentos resultem no menor valor da função objetivo. Ao final do treinamento, as pontuações das árvores restantes são calculadas nas folhas, gerando o modelo do valor predito (LI et al., 2022). Em sistemas de tratamento de água, caracterizados por não linearidade dura, técnicas de aprendizado de máquina, como FR (BREIMAN, 2001) e XGBoost (MO; LI; LI, 2023; MORE; WOLKERSDORFER, 2023), são apropriadas.

2.4 Técnicas de tratamento de dados em séries temporais

A elevada complexidade dos sistemas de tratamento de água decorre das interações não lineares entre variáveis químicas, físicas e operacionais (BULHOES et al., 2020). Estas interações tornam desafiador o monitoramento contínuo e preciso, especialmente quando há lacunas em séries temporais devido a falhas de medição ou interrupções na coleta de dados. Estas lacunas podem comprometer a eficiência dos modelos preditivos, exigindo o uso de métodos robustos de preenchimento e análise (WU et al., 2022). Entre as estratégias disponíveis, destacam-se as abordagens *ensemble*, que combinam múltiplos algoritmos para obter previsões mais precisas e resilientes (RODRIGUEZ et al., 2021; CHOI; LIM; JI, 2023).

A interpolação/extrapolação polinomial por funções spline (PFS) é uma das técnica frequentemente aplicada para preencher dados ausentes em séries temporais, sendo comum o uso de funções spline. No entanto, estas funções podem apresentar o efeito de Runge durante extrapolações, o que leva a oscilações nas extremidades da série e compromete a precisão das previsões. Por esta razão, recomenda-se a utilização de polinômios de grau menor para minimizar este problema (FORNBERG; ZUEV, 2007; CHEN et al., 2014). Uma alternativa utilizada é a decomposição de tendência sazonal em séries temporais *Seasonal-Trend decomposition procedure based on Loess* (STL) (CLEVELAND et al., 1990), devido às vantagens em relação aos métodos tradicionais que realizam previsões diretas (HUO et al., 2019; SEBASTIAN; GAO; XING, 2020; ZHAO et al., 2023). O processo de decomposição consiste em dividir os dados em três componentes principais: crescimento, sazonalidade e erro residual. Após esta separação, os dados são tratados em segmentos menores, o que reduz a complexidade computacional e o erro envolvido. Por fim, as variáveis interpoladas/extrapoladas individualmente são recombinadas por meio do processo inverso de decomposição, resultando em previsões com menor margem de erro (OUYANG; JABLOUN; RAVIER, 2023).

Parte III

Aplicação do Método

3 Aplicação do Método

3.1 Contextualização

A aplicação de técnicas de IA nos dados obtidos em ETA permite desenvolver um modelo representativo do sistema, possibilitando análises avançadas. No presente estudo, adotou-se a abordagem de pesquisa hipotético-dedutiva (LAWSON, 2000), com foco no desenvolvimento de um modelo que simula o comportamento da ETA. A partir desse modelo, é possível realizar simulações para prever o desempenho do sistema e analisar variáveis previamente definidas. Este método viabiliza a otimização da operação sem a necessidade de intervenções físicas na planta. O processo modelado abrange desde a captação superficial da água até a sua distribuição como água potável. A Figura 1 apresenta o fluxograma da metodologia desenvolvida.

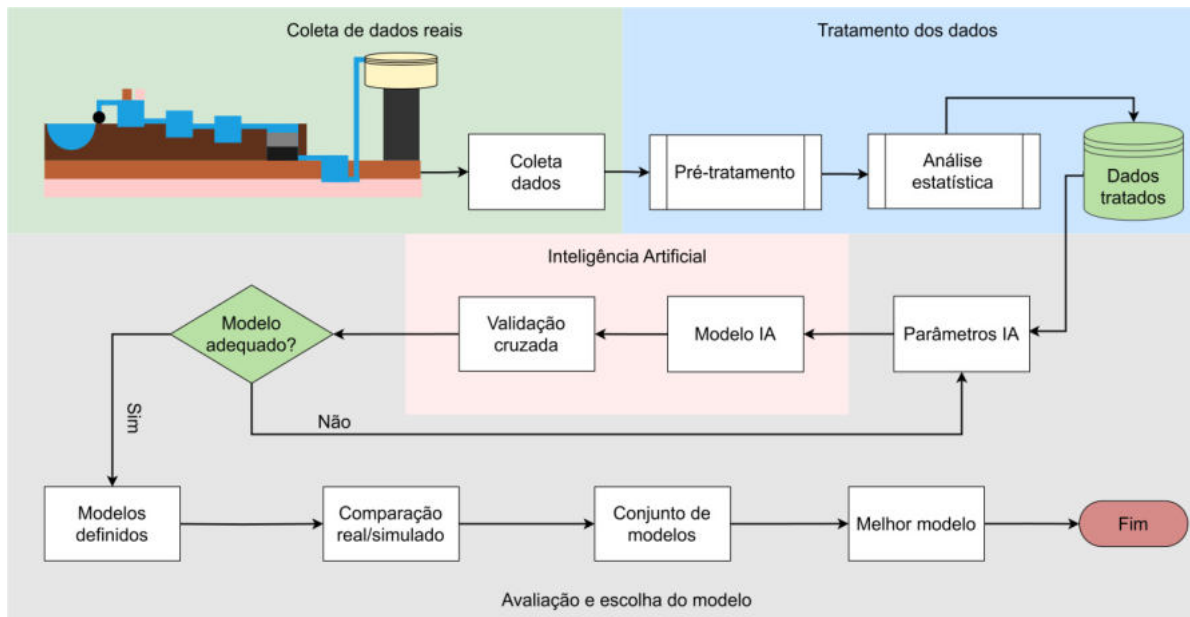


Figura 1 – Fluxograma da metodologia proposta.

3.2 Seleção da ETA adequada

Para se enquadrar no escopo desta pesquisa, a ETA deve apresentar as seguintes características fundamentais: i) possuir captação superficial de água e ii) adotar um sistema de tratamento convencional, que inclua os processos de floculação, decantação, filtração, cloração e fluoretação. Além disso, é essencial que a ETA realize de forma sistemática a medição e o registro dos parâmetros relevantes, garantindo a retroalimentação desses dados para o ajuste das dosagens de produtos químicos, seja de maneira manual ou automatizada.

Esta proposta de metodologia define as características para uma ETA se enquadrar nos critérios deste estudo, conforme ilustrado na Figura 2.

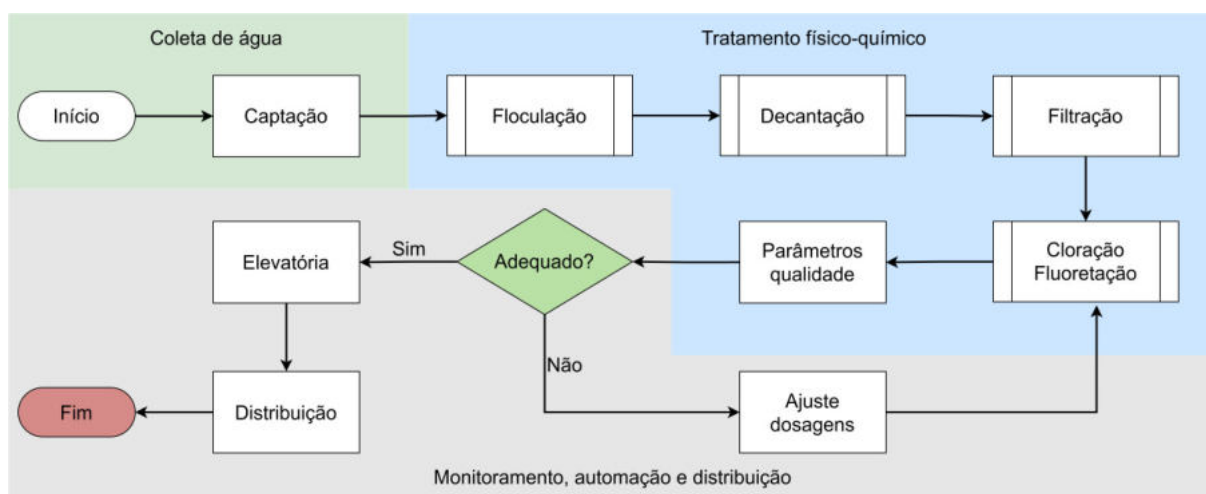


Figura 2 – Fluxo hipotético de requisitos da ETA para inclusão no estudo

A captação superficial permite que a ETA se beneficie da desinfecção solar da água (SODIS), processo capaz de inativar ampla gama de microrganismos patogênicos, associando a radiação ultravioleta (UV) e temperatura. No entanto, este método não é suficiente para tratar elevados volumes de água, necessitando de processos auxiliares, como a oxidação por meio da cloração. Adicionalmente, captações subterrâneas apresentam maior probabilidade de contaminação química e biológica, exigindo processos desinfetantes mais rigorosos e maior número de análises. Assim, por estes motivos, não são objeto desta pesquisa.

Dentre as ETA pesquisadas, foi selecionada a estação localizada no município de Brazabrantes, Goiás, Brasil. Operada pela Saneago, esta unidade atende aos critérios estabelecidos e apresenta a vantagem da proximidade, situando-se a aproximadamente 40km de Goiânia, Goiás, Brasil, onde os autores conduzem os estudos. Adicionalmente, optou-se por selecionar a ETA em município de menor porte, considerando que aproximadamente 80% das estações operadas pela Saneago possuem características semelhantes. As informações utilizadas foram obtidas com base na Lei de Acesso à Informação (PERRITT, 2001), como apresentado no Apêndice A.

O Município de Brazabrantes têm população de 3.981 *hab*, e pertence à Região Metropolitana de Goiânia, Capital do Estado de Goiás. Aproximadamente 67% da população é urbana e o Município de Brazabrantes está geograficamente localizado na bacia do Paraná, e integra o Comitê de Bacia Hidrográfica do Rio Meia Ponte, na qual a Cidade de Brazabrantes representa $\approx 1\%$ do território do município. A Cidade de Brazabrantes possui índice de atendimento urbano de água de 93,55%, e índice de hidrometração¹ de

¹ Processo de instalação de medidores/registadores de água para mensurar o consumo e emitir faturas com base nestes registros (VEIGA; KALBUSCH; HENNING, 2023).

97,67%, o que reflete realidade acima da média nacional brasileira, que é de atendimento urbano de 93,46% e hidrometração de 91,27%. A captação é realizada pelo Sistema Ribeirão Cachoeira, afluente do Rio Meia Ponte, situado no estado de Goiás/Brasil. A Saneago detém a outorga de 20L/s, embora a capacidade produtiva da planta instalada seja de 14L/s. A Figura 3 apresenta, na parte inferior, o ponto de captação de água nas coordenadas geográficas 16°26'57.3"S e 49°22'15.9"W. Na parte superior, encontra-se a ETA localizada a aproximadamente 450m de distância da captação e em local mais alto na proximidade da cidade.



Figura 3 – Mapa georreferenciado da captação e ETA de Brazabrantes.

A planta industrial da ETA da Cidade de Brazabrantes encontra-se segmentada em várias etapas, incluindo: i) captação superficial do manancial, ii) aferição da turbidez, iii) dosagem proporcional de coagulante de acordo com a turbidez medida, iv) decantação em tanque floccodecantador, v) filtragem para remoção de partículas restantes, vi) fluoretação e cloração em tanques de contato e vii) aferição de T , pH , níveis de cloro e flúor a jusante para ajustes nas dosagens. Todo este processo culmina no bombeamento para reservatório elevado e distribuição à população. O diagrama operacional da ETA da Cidade de Brazabrantes é apresentado na Figura 4, nos quais os retângulos identificados na cor laranja com uma linha vertical indicam cálculo e registro dos valores das variáveis. Os produtos químicos empregados são hipoclorito de sódio para cloração, fluorsilicato de sódio para fluoretação e cloreto férrico para floccodecantação.

A Figura 5 apresenta alguns equipamentos, ambientes e indícios de processo realizados na ETA da Cidade de Brazabrantes. A operação desta ETA consiste em captação de manancial superficial, na qual o V_{ab} é medido e aduzido ao sistema de tratamento. A primeira etapa do tratamento é a adequação da turbidez da água bruta T_{ab} , através da aferição, Figura 5(a), e correção com coagulante, Figura 5(b). Os dados são registrados e a água segue para o floco-decantador, Figura 5(c), no qual a cinética é reduzida e os

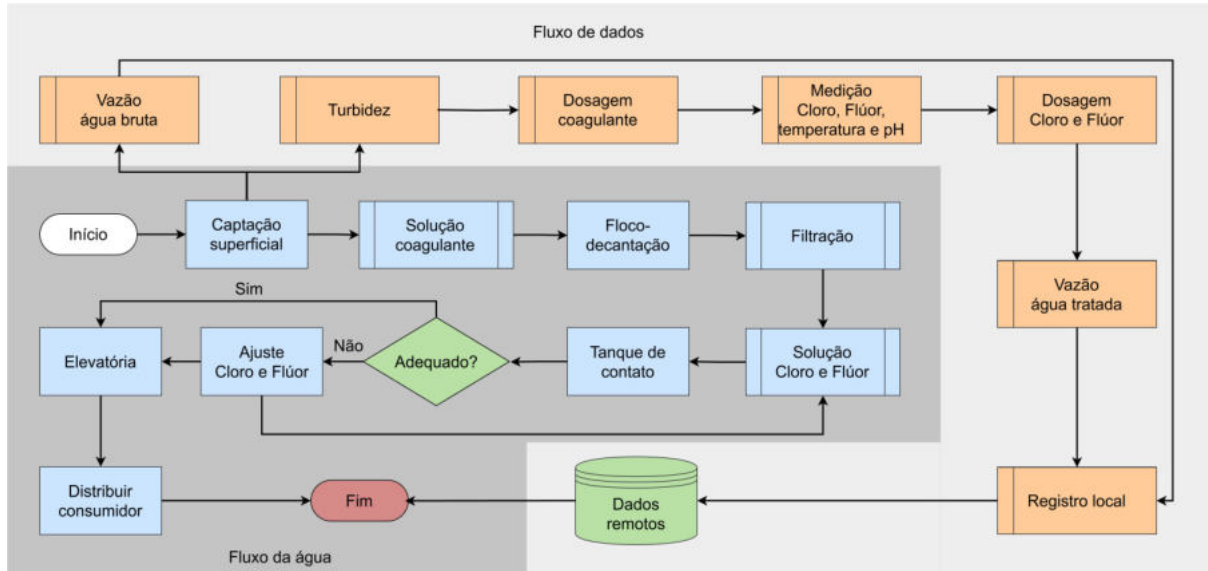


Figura 4 – Estação de tratamento de água de Brazabrantes, Goiás/Brasil.

flocos sedimentam por gravidade. A água superficial é direcionada para o filtro de fluxo ascendente, Figura 5(d) e Figura 5(e) para remoção de partículas excedentes e correção de sabor e odor utilizando carvão. Na sequência é enviada ao tanque de contato, no qual é realizada a dosagem de cloro, flúor e medido M_T e M_{pH} , Figura 5(f). Caso os parâmetros estejam adequados, a água é enviada por recalque para o reservatório elevado, no qual é distribuída por gravidade para a população. Os dados de monitoramento destas etapas são armazenados localmente e posteriormente enviados ao repositório remoto.

3.3 Pré-processamento de dados

3.3.1 Utilização de séries temporais

Após a seleção da ETA, os dados devem ser obtidos no formato de *dataframe*. As informações requeridas para a avaliação incluem: vazão de água bruta V_{ab} [m^3/min], turbidez da água bruta T_{ab} [NTU], dosagem de cloro D_{Cl} [mg/L], dosagem de flúor D_F [mg/L], concentração de cloro na água tratada M_{Cl} [mg/L], concentração de flúor na água tratada M_F [mg/L], temperatura da água tratada M_T [$^{\circ}C$], pH da água tratada M_{pH} e vazão de água tratada V_{at} [m^3/min]. Os dados armazenados são séries temporais, permitindo a reprodução de características sazonais do sistema e garantindo que o intervalo entre as coletas seja padronizado. Esta padronização previne a simulação de comportamentos anômalos que não representem o funcionamento atual do sistema.

Após a coleta, realiza-se o pré-processamento, assegurando a consistência das informações e a coerência das medições. Em caso de lacunas nos dados, aplica-se um método de regressão polinomial ou de decomposição de tendência sazonal em séries temporais, conforme o desempenho avaliado por métricas específicas (DURRLEMAN;

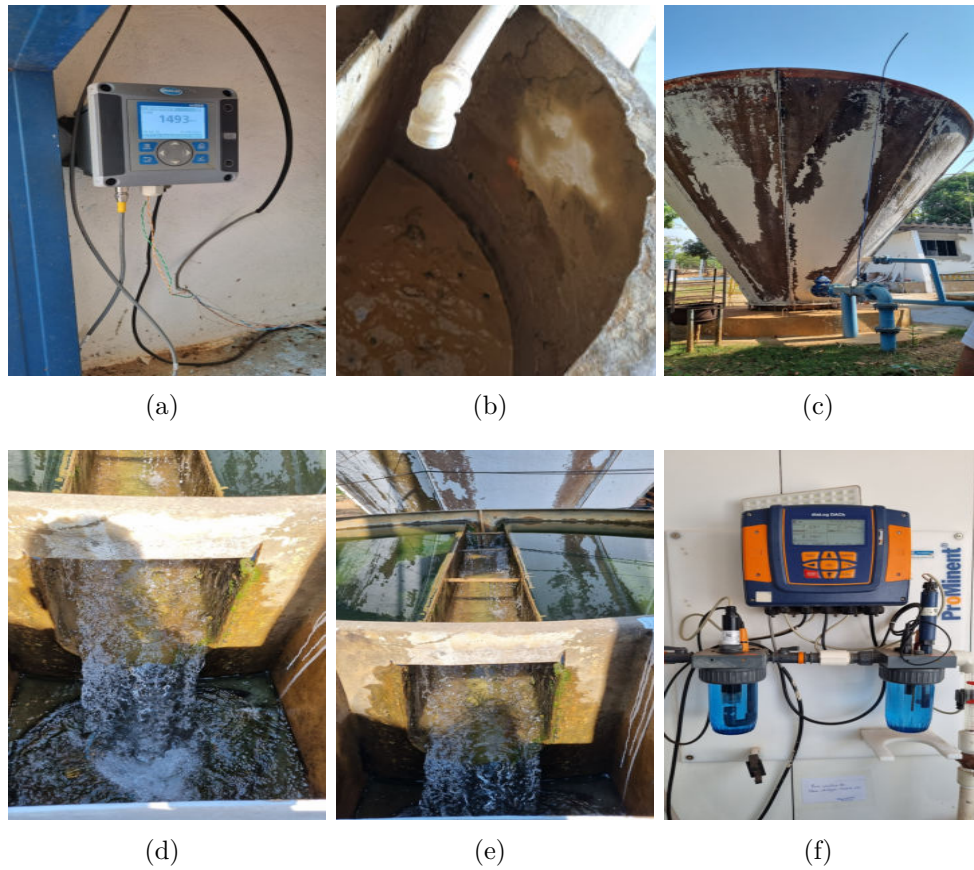


Figura 5 – Equipamentos, ambientes e processo: (a) medidor de turbidez em NTU, (b) tanque de solubilização de cloreto férrico em batelada, (c) flocodecantador, (d) tela de retenção de sólidos flutuantes/grosseiros na saída do flocodecantador, antes de entrar no filtro de fluxo ascendente, (e) calha de coleta de água superficial no filtro-decantador e (f) medidor automatizado de cloro, flúor, M_T e M_{pH} .

SIMON, 1989; WEGMAN; WRIGHT, 1983; CLEVELAND et al., 1990). As variáveis monitoradas, como D_{Cl} e D_F , T_{ab} e V_{ab} , estão inter-relacionadas no contexto do modelo, influenciando diretamente as decisões de dosagem e a qualidade final da água tratada. A turbidez, por exemplo, afeta a quantidade de coagulante necessária e, indiretamente, a demanda de D_{Cl} para desinfecção. Além disto, parâmetros como M_{pH} e M_T influenciam a estabilidade da solução de cloro e flúor, afetando sua eficiência ao longo da distribuição. Desta forma, o modelo deve considerar estas interações para prever com precisão os ajustes necessários, buscando minimizar erros de predição e garantir a estabilidade dos níveis de concentração recomendados por normas.

Todas as informações registradas na ETA da Cidade de Brazabrantes do período de 28/outubro/2022 até 23/julho/2024 foram disponibilizadas, incluindo: i) turbidez da água bruta T_{ab} , ii) nível do reservatório apoiado N_{RA} e elevado N_{RE} , iii) informações elétricas de corrente, tensão, potência e frequência da bomba dosadora de coagulante, iv) dosagem de coagulante D_S , v) nível do reservatório do subdistrito de Deuslândia, Goiás, vi) informações elétricas de corrente, tensão, potência e frequência da bomba dosadora

de flúor, vii) dosagem de flúor D_F , viii) vazão de água bruta V_{ab} , ix) volume histórico acumulativo de água bruta, x) vazão de água tratada V_{at} , xi) volume histórico de água tratada, xii) informações elétricas de corrente, tensão, potência e frequência da bomba dosadora de cloro, xiii) dosagem de cloro D_{Cl} , e xiv) medições de pH , M_{pH} , flúor M_F , cloro M_{Cl} e temperatura T da água tratada M_T . Destes dados, foram utilizados os que apresentaram maior correlação com a dosagem de cloro e flúor, como disposto na Tabela 1.

Tabela 1 – Correlação entre algumas variáveis observadas e a dosagem de cloro e flúor.

	M_{Cl}	M_F	M_{pH}	M_T	V_{at}	V_{ab}	D_{Cl}	D_F
M_{Cl}	1,00							
M_F	$-1,00 \times 10^{-1}$	1,00						
M_{pH}	$5,00 \times 10^{-2}$	$2,20 \times 10^{-1}$	1,00					
M_T	$-2,00 \times 10^{-2}$	$2,10 \times 10^{-1}$	$-1,60 \times 10^{-1}$	1,00				
V_{at}	$4,80 \times 10^{-1}$	$-2,40 \times 10^{-1}$	$3,00 \times 10^{-2}$	0,00	1,00			
V_{ab}	$4,60 \times 10^{-1}$	$-2,40 \times 10^{-1}$	$3,00 \times 10^{-2}$	0,00	$9,20 \times 10^{-1}$	1,00		
D_{Cl}	$4,30 \times 10^{-1}$	$-1,80 \times 10^{-1}$	$4,00 \times 10^{-2}$	$-2,00 \times 10^{-2}$	$8,90 \times 10^{-1}$	$8,20 \times 10^{-1}$	1,00	
D_F	$4,30 \times 10^{-1}$	$-1,80 \times 10^{-1}$	$4,00 \times 10^{-2}$	$-2,00 \times 10^{-2}$	$8,90 \times 10^{-1}$	$8,20 \times 10^{-1}$	$9,80 \times 10^{-1}$	1,00

As variáveis dispostas na Tabela 1 foram as utilizadas na construção do modelo. Inicialmente, o conjunto de dados fornecidos continha 599.386 linhas de valores históricos do período analisado, com aproximadamente 81Mb de dados estruturados no formato de banco de dados relacional, que foram exportados para o formato de *dataframe* em *.csv. Estes representavam as medições históricas de diversas variáveis que permitiam calcular o IQA.

3.3.2 Tratamento de valores atípicos e lacunas

Com base no comportamento do sistema e na coerência das informações fornecidas pelos técnicos, foram descartados os valores atípicos. O critério de descarte considerou valores de $2 \leq M_{pH} < 8$, $0 \leq D_{Cl} < 5mg/L$, $0 \leq D_F < 3mg/L$, $10^\circ C \leq M_T < 50^\circ C$, $0 \leq V_{ab} < 1m^3/min$ e $0 \leq V_{at} < 1m^3/min$, $0 \leq M_{Cl} < 50mg/L$ e $0 \leq M_F < 10mg/L$.

Adicionalmente, foi observado que os dados tinham periodicidade de coleta de 15 minutos, e na data de 11 de maio de 2023, às 14h42min, a periodicidade foi alterada para cada minuto. Devido a esta alteração, se mostrou necessário ajustar a base do tempo de amostragem de 15 minutos para 1 minuto. Para o preenchimento dos dados ausentes, ou lacunas, foram aplicadas as técnicas polinomial por funções spline (PFS) e decomposição sazonal (STL). Com o objetivo de avaliar a precisão dos dois métodos, foi selecionado um trecho de dados conhecidos, no qual 20% das observações foram suprimidas. Utilizando os 80% restantes, ambas as técnicas foram empregadas para estimar os valores ausentes.

A Tabela 2 dispõe os resultados da análise com as respectivas métricas de avaliação, enquanto a Figura 6 apresenta o comparativo gráfico referente ao parâmetro M_{Cl} . No gráfico, a linha amarela representa os valores medidos pelo sistema por meio do equipamento apresentado na Figura 5(f), a linha vermelha indica o preenchimento das lacunas utilizando interpolação spline, e a linha azul apresenta o preenchimento utilizando STL. O eixo das

abscissas corresponde às datas no formato mês, dia, ano, hora e minuto, com intervalos de um minuto, enquanto o eixo das ordenadas representa os valores medidos de cloro M_{Cl} .

Tabela 2 – Critérios de avaliação dos métodos usados para preencher dados ausentes.

Parâmetro	REMQ PFS	REMQ STL	R^2 PFS	R^2 STL	EMA PFS	EMA STL
D_{Cl}	$1,04 \times 10^{-1}$	$9,11 \times 10^{-2}$	$9,96 \times 10^{-1}$	$9,97 \times 10^{-1}$	$8,08 \times 10^{-3}$	$5,57 \times 10^{-3}$
M_{Cl}	$2,27 \times 10^{-2}$	$1,84 \times 10^{-2}$	$9,98 \times 10^{-1}$	$9,99 \times 10^{-1}$	$1,34 \times 10^{-3}$	$1,16 \times 10^{-3}$
M_{pH}	$1,97 \times 10^{-2}$	$1,47 \times 10^{-2}$	$9,98 \times 10^{-1}$	$9,99 \times 10^{-1}$	$1,47 \times 10^{-3}$	$1,09 \times 10^{-3}$
M_T	$7,34 \times 10^{-2}$	$6,54 \times 10^{-2}$	$9,99 \times 10^{-1}$	$9,99 \times 10^{-1}$	$4,75 \times 10^{-3}$	$3,39 \times 10^{-3}$
M_F	$2,34 \times 10^{-2}$	$2,02 \times 10^{-2}$	$9,96 \times 10^{-1}$	$9,97 \times 10^{-1}$	$2,71 \times 10^{-3}$	$2,12 \times 10^{-3}$
D_F	$6,11 \times 10^{-2}$	$5,35 \times 10^{-2}$	$9,96 \times 10^{-1}$	$9,97 \times 10^{-1}$	$4,78 \times 10^{-3}$	$3,29 \times 10^{-3}$
V_{ab}	$2,74 \times 10^{-2}$	$2,27 \times 10^{-2}$	$9,95 \times 10^{-1}$	$9,97 \times 10^{-1}$	$4,90 \times 10^{-3}$	$3,94 \times 10^{-3}$
V_{at}	$2,69 \times 10^{-2}$	$2,29 \times 10^{-2}$	$9,95 \times 10^{-1}$	$9,96 \times 10^{-1}$	$5,31 \times 10^{-3}$	$3,68 \times 10^{-3}$
T_{ab}	$6,86 \times 10^{-5}$	$5,07 \times 10^{-5}$	$9,99 \times 10^{-1}$	$9,99 \times 10^{-1}$	$6,26 \times 10^{-8}$	$5,32 \times 10^{-8}$

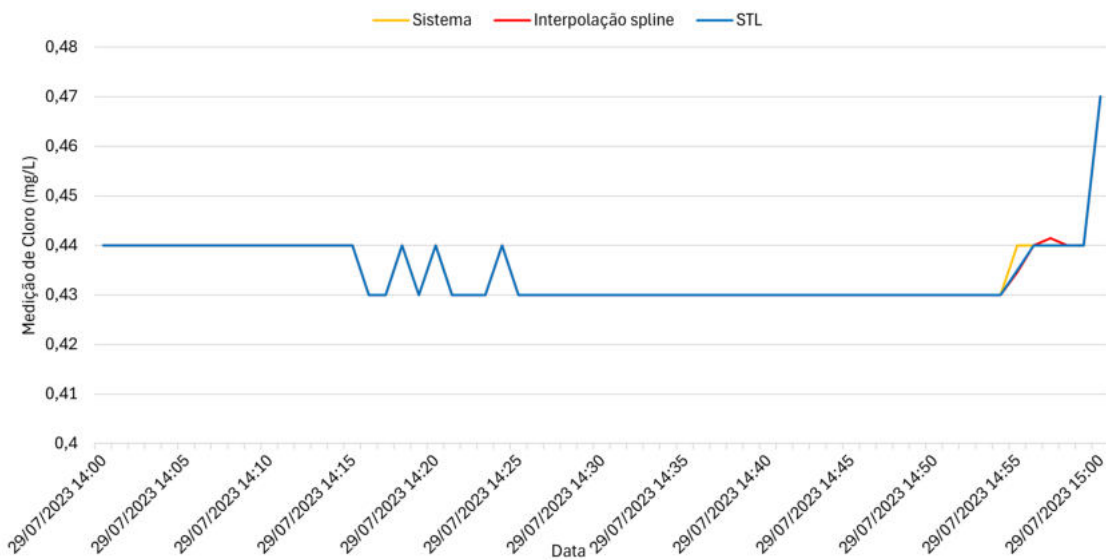


Figura 6 – Comparação de métodos para preenchimento de dados ausentes.

Após a aplicação dos métodos para preenchimento das lacunas, o conjunto de dados passou a ter 909.250 linhas de valores históricos no período compreendido entre 31 de outubro de 2022 e 23 de julho de 2024.

3.3.3 Análise estatística descritiva

Antes de aplicar os dados nos modelos de IA, realiza-se análise utilizando a estatística descritiva na tentativa de compreender e resumir a natureza dos dados. Avalia-se as médias μ , o desvio padrão σ e os valores máximos V_{max} e mínimos V_{min} dos parâmetros para obtenção do comportamento típico. Os parâmetros estatísticos do conjunto de dados são dispostos na Tabela 3.

3.3.4 Validação da operação da ETA em conformidade com as normas

Os valores de cloração e fluoretação do sistema devem estar em conformidade com as recomendações internacionais de potabilidade da água (FAWELL et al., 2006; World

Tabela 3 – Estatísticas descritivas dos parâmetros observados.

Parâmetro	μ	σ	V_{max}	V_{min}
D_{Cl}	1,45	1,61	4,82	0,00
M_{Cl}	$6,20 \times 10^{-1}$	$4,93 \times 10^{-1}$	$2,83 \times 10^1$	$1,00 \times 10^{-2}$
M_{pH}	6,62	$4,04 \times 10^{-1}$	7,93	2,67
M_T	$2,58 \times 10^1$	2,64	$4,51 \times 10^1$	$1,72 \times 10^1$
M_F	$9,07 \times 10^{-1}$	$3,81 \times 10^{-1}$	9,14	0,00
D_F	$8,60 \times 10^{-1}$	$9,50 \times 10^{-1}$	2,82	0,00
V_{ab}	$4,00 \times 10^{-1}$	$3,92 \times 10^{-1}$	1,00	0,00
V_{at}	$3,65 \times 10^{-1}$	$3,84 \times 10^{-1}$	1,00	0,00
T_{ab}	$1,31 \times 10^2$	$1,55 \times 10^2$	$3,00 \times 10^3$	2,00

[Health Organization, 2022](#)). Para águas superficiais sem fonte natural de flúor, o valor recomendado é de $1.5mg/L$, sendo indicada a fluoretação artificial na faixa de $0.5mg/L$ a $1.0mg/L$. A desinfecção com compostos à base de cloro, como hipocloritos, possui o valor recomendado de $5mg/L$. Embora a concentração de cloro possa ser reduzida a zero após o tratamento, recomenda-se manter valor residual entre $0.2mg/L$ e $1.0mg/L$, garantindo proteção contra recontaminação e controle do crescimento de microrganismos na rede de distribuição. Esta observação é fundamental para a aplicabilidade do trabalho, pois, caso o sistema opere fora dos limites recomendados, os modelos de IA treinados com estes dados poderão gerar previsões igualmente imprecisas.

A partir das amostras obtidas da ETA, foi realizada uma análise estatística para comparar os valores dos parâmetros com os limites recomendados internacionalmente ([World Health Organization, 2002](#)). Do total de 909.250 registros do sistema, 908.054 atenderam aos critérios de cloração, correspondendo a $\approx 99\%$ dos dados. Em relação à fluoretação, 866.621 registros estavam dentro dos limites recomendados, representando $\approx 95\%$ do total. Assim, considera-se que a operação desta ETA segue os padrões internacionais para desinfecção com hipocloritos e suplementação artificial de flúor, sendo possível reproduzir seu comportamento com confiança.

3.4 Aplicação de algoritmos preditivos

3.4.1 Segmentação dos dados

Dado o caráter de séries temporais dos dados analisados, após o processo de tratamento e avaliação estatística, a metodologia propõe a segmentação das séries em intervalos contínuos e ininterruptos com duração mínima de um mês. Essa segmentação é imprescindível para garantir a identificação e a análise de padrões sazonais do sistema. Os dados restantes são destinados ao **treinamento**, assegurando a segmentação adequada e a independência das amostras. O conjunto de **validação interna** deve compreender, no mínimo, 20% do total de dados, e, caso um mês isolado não atenda a esse critério, adicionam-se meses subsequentes até alcançar a proporção mínima exigida.

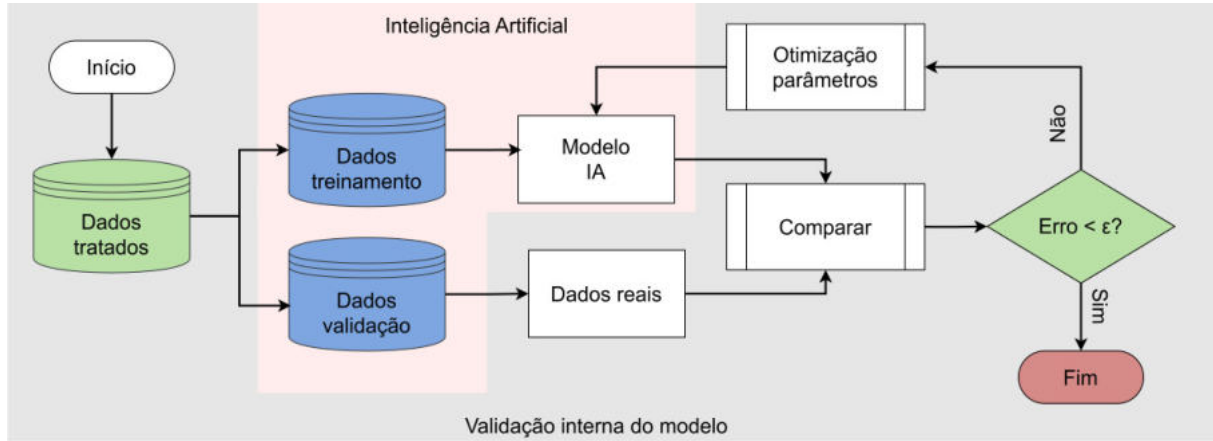


Figura 7 – Divisão dos dados tratados para validação do modelo proposto.

Para a construção do modelo e validação interna, os dados foram segmentados entre 01 de janeiro de 2023 e 30 de março de 2023 (89 dias) para **testes e validação interna**, totalizando 131.041 registros, o que corresponde a $\approx 25\%$ dos dados. Os registros do período de 01 de abril de 2023 a 31 de dezembro de 2023 (275 dias) foram utilizados para o **treinamento dos modelos**, totalizando 393.121 registros, equivalente a $\approx 75\%$ dos dados. Para a validação da solução (**validação externa**), foram utilizados os dados do período de 31 de outubro de 2022 a 31 de dezembro de 2022 (62 dias), totalizando 129.601 registros, denominados amostra de 2022 (A22), e os dados de 01 de abril de 2024 a 12 de abril de 2024 (12 dias), totalizando 25.084 registros, denominados amostra de 2024 (A24).

3.4.2 Métricas de avaliação dos modelos

Para a construção de modelos utilizando AM geralmente são empregadas as seguintes métricas de avaliação: i) raiz do erro médio quadrático (REMQ) que calcula a média dos quadrados das diferenças entre os valores observados e os valores previstos, ii) erro médio absoluto (EMA) que calcula a média das diferenças absolutas entre os valores observados e os valores previstos e iii) coeficiente de determinação (R^2) que avalia o modelo de regressão observando se este ajusta aos dados observados, representando a proporção da variância da variável dependente explicada pelos regressores. Na escolha do método de avaliação, é necessário considerar a natureza do problema, especialmente em casos de regressão de séries históricas. O REMQ é a métrica comum para problemas de regressão, penalizando erros maiores, enquanto o EMA é menos sensível a valores discrepantes em comparação com o REMQ. Além destas métricas, a validação cruzada é aplicada para avaliar o desempenho. No FR e RNA-MLP, utiliza-se o *k-fold cross-validation* (*k*-FCV), que divide o conjunto de dados em *k* partições, treina o modelo em *k* – 1 partições e avalia o desempenho na partição restante. Este processo é repetido *k* vezes, com cada partição sendo usada como conjunto de teste única vez.

Os critérios de convergência desempenham um papel fundamental na definição do momento ideal para encerrar o treinamento dos algoritmos, equilibrando a prevenção de sobreajuste e a eficiência computacional. No caso do algoritmo FR, os critérios incluem o número máximo de árvores (n_a) e o crescimento mínimo de folhas (c_{mf}). Para o algoritmo XGBoost, os critérios abrangem n_a , o tamanho mínimo da amostra por folha (t_{maf}) e o ganho mínimo (γ), que corresponde à redução mínima necessária na função de perda para permitir uma nova divisão. Por sua vez, no algoritmo RNA-MLP, os critérios de convergência consideram o número máximo de épocas (ϵ), a melhoria mínima na validação (Δ_{min}) e a paciência (α), definida como o número de épocas consecutivas sem melhoria admissíveis antes da interrupção do treinamento. O ajuste adequado destes critérios é realizado com base nas características intrínsecas do conjunto de dados e nas especificidades do problema em estudo, garantindo eficiência e eficácia no desempenho dos modelos gerados.

Os critérios de convergência são importantes para determinar o momento em que os algoritmos devem encerrar o treinamento, visando evitar o sobreajuste ou garantir eficiência computacional. No contexto do algoritmo FR, os critérios de convergência incluem o número máximo de árvores n_a e o crescimento mínimo de folhas c_{mf} . Para o algoritmo XGBoost, os critérios compreendem o n_a , o tamanho mínimo da amostra por folha t_{maf} e o ganho mínimo γ , que define a quantidade mínima de redução na função de perda necessária para realizar nova divisão. Por fim, para o algoritmo RNA-MLP, os critérios envolvem o número máximo de épocas ϵ , a melhoria mínima na validação Δ_{min} e a paciência α , indicando o número de épocas sem melhoria que podem ocorrer antes de interromper o treinamento. O ajuste nos valores destes critérios é realizado com base nas características específicas do conjunto de dados e nos requisitos do problema. A seleção apropriada dos critérios de convergência contribui para a eficiência dos modelos treinados.

3.4.3 Construção dos modelos sem ajuste de hiperparâmetros

Os modelos preditivos devem ser desenvolvidos utilizando algoritmos amplamente aplicados em estudos semelhantes e reconhecidos por seu desempenho superior, como FR, XGBoost e RNA-MLP, incluindo variações na arquitetura de camadas entre outros parâmetros. O objetivo central consiste na otimização das dosagens de cloro e flúor no sistema, conforme ilustrado no fluxo de dados tratados na Figura 7, que sintetiza a metodologia proposta. Os modelos foram construídos na linguagem R utilizando o pacote `caret`. O método de validação cruzada k -folds foi empregado com $k = 10$. A função `seed` foi fixada com valor constante ao longo de todo o processo, garantindo a reprodutibilidade dos geradores de números pseudoaleatórios e viabilizando a comparação entre execuções.

No modelo desenvolvido com o algoritmo RNA-MLP, o processamento inicia-se nas camadas de entrada, onde os neurônios recebem as variáveis do conjunto de dados como entradas. A propagação direta ocorre ao longo das camadas ocultas, nas quais cada

neurônio realiza uma combinação linear ponderada das entradas, seguida pela aplicação de uma função de ativação não linear. As camadas ocultas são responsáveis por capturar padrões complexos e não lineares, culminando na camada de saída, que fornece a previsão final do modelo. O cálculo do erro, obtido pela comparação entre os valores previstos e os valores reais, aciona o processo de retropropagação, ajustando iterativamente os pesos sinápticos para minimizar a função de perda e melhorar o desempenho do modelo.

No treinamento do modelo RNA-MLP com 1-camada oculta, para cada cenário o modelo foi treinado com 1, 3 e 5 neurônios na camada oculta S_z , e foi aplicado o decaimento de peso W_D no intervalo de $[0; 10^{-4}]$ para regularização e evitar o sobreajuste. O menor valor de REMQ foi obtido no modelo com $S_z = 5$, enquanto o menor valor de EMA e o maior valor de R^2 foram observados no modelo com $S_z = 3$, ambos com $W_D = 0$. Para a fluoretação, os menores valores de REMQ e EMA, e o maior valor de R^2 foram obtidos com o modelo com $S_z = 5$ e $W_D = 0$. Os modelos de RNA-MLP com 2-camadas e 3-camadas, sem otimização de hiperparâmetros, apresentaram os mesmos resultados que o modelo de RNA-MLP com 1-camada, não sendo desejáveis devido à maior necessidade de recursos computacionais.

No modelo desenvolvido utilizando o algoritmo FR, inicia-se com o conjunto de dados de treinamento, realizando amostragem aleatória com reposição para construir diversas árvores de decisão. Cada árvore é treinada em conjunto de dados diferente e a previsão final é determinada pela votação ou média das previsões individuais das árvores. Este processo diversificado reduz o sobreajuste.

No treinamento do modelo FR, o modelo foi treinado com diferentes números de variáveis amostradas aleatoriamente como candidatas em cada divisão, sendo $N_{rc} = [2, 4, 7]$. O modelo com $N_{rc} = 2$ apresentou o menor valor de REMQ e o maior valor de R^2 , enquanto o modelo com $N_{rc} = 4$ obteve o menor valor de EMA para cloração e fluoretação.

Para o modelo desenvolvido utilizando o algoritmo XGBoost, inicia-se com a árvore base, calcula os resíduos entre as previsões atuais e os valores reais, construindo árvores adicionais para corrigir deficiências nas previsões anteriores de maneira iterativa. As previsões das árvores são combinadas e o modelo é atualizado para minimizar a função de perda global. Com técnicas de regularização para evitar sobreajuste, o XGBoost produz previsões finais ponderadas pelo desempenho das árvores individuais.

No treinamento do modelo XGBoost, foram utilizadas as mesmas quantidades de amostras do modelo FR, e foi mantido o parâmetro $\gamma = 0$ para evitar a aplicação de penalidades no controle de complexidade do modelo. Embora esta configuração possa levar ao sobreajuste, não foi estabelecida a redução mínima de perda necessária para realizar divisões adicionais em um nó da árvore. Além disto, foi definido o peso mínimo de instância $M_{cw} = 1$, indicando que a quantidade mínima total de instâncias necessárias em um nó para realizar nova divisão seja sempre 1. Esta configuração controla a profundidade

mínima da árvore, evitando que nós com número pequeno de instâncias sejam divididos, tornando o algoritmo menos conservador. Os menores valores de REMQ e EMA e o maior valor de R^2 foram obtidos com $N_{rn} = 150$ (número de rodadas), $D_M = 3$ (profundidade máxima das árvores), $\eta = 0.4$ (taxa de aprendizado) e $C_{ST} = 0.8$ (proporção de colunas amostradas). A proporção de observações (linhas) amostradas em cada árvore apresentou melhor desempenho com $S_{sa} = 0.75$ para a cloração e $S_{sa} = 1.00$ para a fluoretação. A Tabela 4 dispõe os valores dos hiperparâmetros e as métricas de convergência da validação interna para os algoritmos avaliados.

Tabela 4 – Valores de hiperparâmetros e avaliações para os métodos RNA-MLP, FR e XGBoost.

Hiperparâmetros (RNA-MLP)				Cloração			Fluoretação		
S_z	–	–	W_D	REMQ	R^2	EMA	REMQ	R^2	EMA
3	–	–	0,00	$5,31 \times 10^{-1}$	$8,98 \times 10^{-1}$	$2,05 \times 10^{-1}$	$3,18 \times 10^{-1}$	$8,96 \times 10^{-1}$	$1,39 \times 10^{-1}$
5	–	–	0,00	$5,26 \times 10^{-1}$	$8,97 \times 10^{-1}$	$2,06 \times 10^{-1}$	$2,94 \times 10^{-1}$	$9,09 \times 10^{-1}$	$1,28 \times 10^{-1}$
Hiperparâmetros (FR)				Cloração			Fluoretação		
N_{rc}	–	–	–	REMQ	R^2	EMA	REMQ	R^2	EMA
2	–	–	–	$8,46 \times 10^{-2}$	$9,97 \times 10^{-1}$	$1,18 \times 10^{-2}$	$5,05 \times 10^{-2}$	$9,97 \times 10^{-1}$	$7,14 \times 10^{-3}$
4	–	–	–	$8,81 \times 10^{-2}$	$9,96 \times 10^{-1}$	$1,08 \times 10^{-2}$	$5,22 \times 10^{-2}$	$9,96 \times 10^{-1}$	$6,32 \times 10^{-3}$
Hiperparâmetros (XGBoost)				Cloração			Fluoretação		
η	D_M	C_{ST}	S_{sa}	REMQ	R^2	EMA	REMQ	R^2	EMA
0,4	3	0,8	0,75	$2,78 \times 10^{-1}$	$9,70 \times 10^{-1}$	$1,12 \times 10^{-1}$	$1,72 \times 10^{-1}$	$9,66 \times 10^{-1}$	$7,34 \times 10^{-2}$
0,4	3	0,8	1,00	$2,81 \times 10^{-1}$	$9,69 \times 10^{-1}$	$1,15 \times 10^{-1}$	$1,72 \times 10^{-1}$	$9,67 \times 10^{-1}$	$7,19 \times 10^{-2}$

3.5 Otimização de hiperparâmetros

Nos algoritmos de aprendizado de máquina, os hiperparâmetros, definidos pelo usuário antes do treinamento do modelo, não são ajustados automaticamente pela inteligência artificial. Esses hiperparâmetros regulam aspectos do treinamento, como a complexidade do modelo, a taxa de aprendizado e a regularização. A escolha adequada desses parâmetros influencia diretamente o desempenho do modelo. Para defini-los, podem ser empregados quatro métodos de otimização: i) otimização bayesiana (BO), que aplica técnicas estatísticas para modelar a relação entre os hiperparâmetros e a métrica de avaliação; ii) algoritmos genéticos (AG), abordagem evolutiva que manipula populações de combinações de hiperparâmetros; iii) pesquisa em grade, que realiza uma busca exaustiva com base em uma lista predefinida de hiperparâmetros e critérios de desempenho; e iv) pesquisa aleatória, que testa combinações de hiperparâmetros de forma aleatória para identificar o melhor ajuste.

A aplicação inicial dos algoritmos foi realizada sem ajustes, não incorporando métodos de seleção de hiperparâmetros. Nos modelos otimizados, foi adotada a pesquisa aleatória para seleção de hiperparâmetros, devido à ausência de conhecimento prévio sobre o espaço de busca e ao elevado custo computacional das alternativas exaustivas. Após a amostragem inicial, ajustes adicionais foram realizados em torno dos valores ótimos

identificados, utilizando a pesquisa em grade para refinar a configuração de cada modelo treinado.

3.5.1 Construção dos modelos com otimização de hiperparâmetros

O modelo RNA-MLP foi treinado com três camadas ocultas e otimização de hiperparâmetros, configurado com 1 a 15 neurônios em cada camada oculta S_1 , S_2 e S_3 e aplicado o decaimento de peso W_D no intervalo $[0; 10^{-6}]$ para regularização e prevenção de sobreajuste. A Tabela 5 dispõe os melhores valores de avaliação para a cloração após a validação interna, incluindo as métricas REMQ, R^2 e EMA. O menor REMQ, o menor EMA e o maior R^2 foram obtidos com a configuração $S_1 = 11$, $S_2 = 4$, $S_3 = 8$ e $W_D = 10^{-5}$. Para a fluoretação, o menor REMQ e o maior R^2 foram alcançados com a configuração $S_1 = 10$, $S_2 = 12$, $S_3 = 2$ e $W_D = 10^{-6}$, enquanto o menor EMA foi obtido com $S_1 = 10$, $S_2 = 11$, $S_3 = 2$ e $W_D = 10^{-6}$.

O modelo FR foi treinado com diferentes quantidades de variáveis amostradas aleatoriamente como candidatas em cada divisão, sendo N_{rc} variando no intervalo $[1; 10]$. Os resultados da validação interna para cloração e fluoretação estão dispostos na Tabela 5. O modelo com $N_{rc} = 4$ registrou o menor valor de REMQ e o maior valor de R^2 para cloração e fluoretação, enquanto o modelo com $N_{rc} = 5$ obteve o menor valor de EMA em ambos os casos.

No treinamento do modelo XGBoost após a otimização dos hiperparâmetros, o parâmetro $\gamma = 0$ foi mantido e o peso mínimo de instância foi definido como $M_{cw} = 1$. A Tabela 5 dispõe os valores dos hiperparâmetros e as métricas de convergência otimizadas obtidas na validação interna para o modelo XGBoost. Os menores valores de REMQ e EMA foram alcançados com número de rodadas $N_{rn} = 300$, profundidade máxima $D_M = 8$, taxa de aprendizado $\eta = 0.2$, proporção de colunas amostradas $C_{ST} = 0.8$ e proporção de instâncias amostradas $S_{sa} = 0.75$, tanto para a cloração quanto para a fluoretação. O maior valor de R^2 em ambos os casos foi obtido com $N_{rn} = 150$ e $D_M = 6$, mantendo os demais hiperparâmetros inalterados.

Tabela 5 – Valores otimizados de hiperparâmetros e avaliações para os métodos RNA-MLP, FR e XGBoost.

Hiperparâmetros (RNA-MLP)				Cloração			Fluoretação		
S_1	S_2	S_3	W_D	REMQ	R^2	EMA	REMQ	R^2	EMA
11	4	8	10^{-5}	$1,39 \times 10^{-1}$	$9,94 \times 10^{-1}$	$4,12 \times 10^{-2}$	$1,29 \times 10^{-1}$	$9,83 \times 10^{-1}$	$3,51 \times 10^{-2}$
10	11	2	10^{-6}	$3,02 \times 10^{-1}$	$9,60 \times 10^{-1}$	$1,02 \times 10^{-1}$	$8,99 \times 10^{-2}$	$9,92 \times 10^{-1}$	$1,79 \times 10^{-2}$
10	12	2	10^{-6}	$3,01 \times 10^{-1}$	$9,61 \times 10^{-1}$	$1,04 \times 10^{-1}$	$8,45 \times 10^{-2}$	$9,93 \times 10^{-1}$	$1,89 \times 10^{-2}$
Hiperparâmetros (FR)				Cloração			Fluoretação		
N_{rc}				REMQ	R^2	EMA	REMQ	R^2	EMA
4	-	-	-	$7,38 \times 10^{-2}$	$9,98 \times 10^{-1}$	$7,93 \times 10^{-3}$	$5,41 \times 10^{-2}$	$9,97 \times 10^{-1}$	$5,56 \times 10^{-3}$
5	-	-	-	$7,42 \times 10^{-2}$	$9,98 \times 10^{-1}$	$7,63 \times 10^{-3}$	$5,43 \times 10^{-2}$	$9,97 \times 10^{-1}$	$5,46 \times 10^{-3}$
Hiperparâmetros (XGBoost)				Cloração			Fluoretação		
η	D_M	C_{ST}	S_{sa}	REMQ	R^2	EMA	REMQ	R^2	EMA
0,2	6	0,8	0,75	$8,86 \times 10^{-2}$	$9,97 \times 10^{-1}$	$1,58 \times 10^{-2}$	$8,32 \times 10^{-2}$	$9,98 \times 10^{-1}$	$1,40 \times 10^{-2}$
0,2	8	0,8	0,75	$5,61 \times 10^{-2}$	$9,97 \times 10^{-1}$	$9,31 \times 10^{-3}$	$5,42 \times 10^{-2}$	$9,97 \times 10^{-1}$	$8,87 \times 10^{-3}$

3.6 Validação externa dos modelos

3.6.1 Avaliação dos modelos preditivos

A utilização de dados independentes, não empregados durante o treinamento de modelos de AM, é particularmente relevante no contexto de séries temporais, em que a dependência temporal entre as observações pode introduzir desafios adicionais. Nestes casos, os dados de validação devem ser selecionados respeitando a ordem cronológica, prevenindo o vazamento de informações futuras para o treinamento, evitando *overfit* e assegurando a avaliação realista da capacidade preditiva do modelo. Este procedimento é importante para avaliar a capacidade do modelo de generalizar para períodos ainda não observados, garantindo previsões resilientes frente a variações sazonais, tendências e eventos inesperados. Adicionalmente, ao trabalhar com séries temporais, a validação com dados independentes permite verificar a adequação do modelo na captura de padrões complexos de evolução temporal, assegurando sua aplicabilidade em cenários reais, nas quais decisões com base no tempo são críticas.

Foram realizadas validações externas dos modelos construídos sem ajustes nos hiperparâmetros, com os resultados das métricas aplicadas apresentados na Tabela 6. Os dados mostram que, para a cloração, o método de predição mais eficaz foi o algoritmo de FR, enquanto, para a fluoretação, o melhor desempenho foi alcançado pelo RNA-MLP na amostra A22 e pelo XGBoost na amostra A24. Os testes de validação externa com hiperparâmetros otimizados foram realizados utilizando as mesmas amostras A22 e A24. Os resultados das métricas aplicadas nesta validação estão dispostas na Tabela 7. Para a cloração, na amostra A22, o menor valor de REMQ e o maior valor de R^2 foram obtidos pelo modelo FR, enquanto o RNA-MLP apresentou o menor valor de EMA. Na amostra A24, os menores valores de REMQ e EMA também foram alcançados pelo modelo FR, mas o maior valor de R^2 foi obtido pelo RNA-MLP. Para a fluoretação, na amostra A22, os menores valores de REMQ e EMA, bem como o maior valor de R^2 , foram obtidos pelo modelo RNA-MLP. Na amostra A24, todos os melhores valores (REMQ, EMA e R^2) foram alcançados pelo modelo XGBoost. Os comportamentos das dosagens de cloro e flúor ao longo de 24 horas para a amostra A22 estão ilustrados nas Figuras 8(a), 8(b), 8(c) e 8(d), onde os valores reais aferidos no sistema são comparados com os valores previstos pelos modelos. De forma similar, as Figuras 9(a), 9(b), 9(c) e 9(d) apresentam os resultados para a amostra A24.

Ao comparar os resultados da Tabela 7 que mostra estatísticas dos modelos com hiperparâmetros otimizados, com os dados da Tabela 6, que apresenta estatísticas dos modelos sem otimização, observa-se que o ajuste dos hiperparâmetros aprimorou o desempenho dos modelos com base nas métricas R^2 e EMA. Para a cloração, após a otimização, os valores de REMQ de todos os modelos foram superiores quando as previsões foram

Tabela 6 – Validação externa comparando os resultados do modelo com os valores reais coletados no sistema.

Sample	Modelos	Cloração			Fluoretação		
		REMQ	R^2	EMA	REMQ	R^2	EMA
2022	RNA-MLP	$4,21 \times 10^{-1}$	$9,64 \times 10^{-1}$	$2,31 \times 10^{-1}$	$2,07 \times 10^{-1}$	$9,60 \times 10^{-1}$	$1,04 \times 10^{-1}$
2022	FR	$3,10 \times 10^{-1}$	$9,87 \times 10^{-1}$	$2,47 \times 10^{-1}$	$3,37 \times 10^{-1}$	$9,54 \times 10^{-1}$	$2,40 \times 10^{-1}$
2022	XGBoost	$8,11 \times 10^{-1}$	$8,14 \times 10^{-1}$	$5,90 \times 10^{-1}$	$2,51 \times 10^{-1}$	$9,54 \times 10^{-1}$	$1,61 \times 10^{-1}$
2024	RNA-MLP	$3,25 \times 10^{-1}$	$8,54 \times 10^{-1}$	$1,66 \times 10^{-1}$	$2,72 \times 10^{-1}$	$6,50 \times 10^{-1}$	$1,35 \times 10^{-1}$
2024	FR	$2,68 \times 10^{-1}$	$8,47 \times 10^{-1}$	$1,42 \times 10^{-1}$	$1,75 \times 10^{-1}$	$8,65 \times 10^{-1}$	$9,88 \times 10^{-2}$
2024	XGBoost	$3,05 \times 10^{-1}$	$8,12 \times 10^{-1}$	$1,62 \times 10^{-1}$	$1,60 \times 10^{-1}$	$8,71 \times 10^{-1}$	$8,61 \times 10^{-2}$

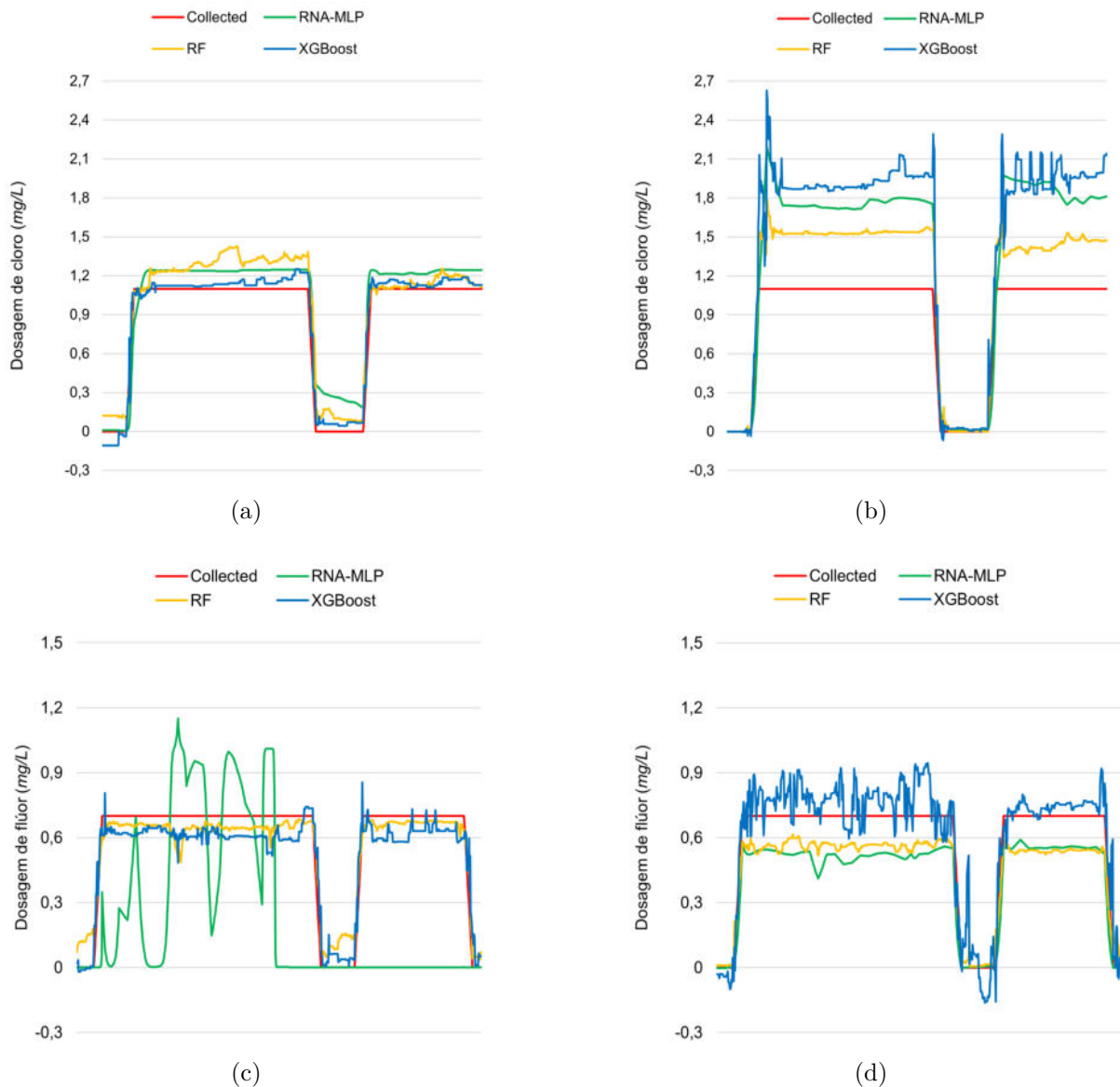


Figura 8 – Valores reais do sistema e previsões do modelo com amostra de 2022: (a) cloro sem otimização de hiperparâmetros, (b) cloro com otimização de hiperparâmetros, (c) flúor sem otimização de hiperparâmetros e (d) flúor com otimização de hiperparâmetros.

comparadas com os dados das amostras A22 e A24, como disposto na Tabela 5. Para a fluoretação, com exceção do modelo RNA-MLP, foi observado o mesmo comportamento em

Tabela 7 – Validação externa comparando os resultados dos modelos otimizados com os valores reais coletados no sistema.

Sample	Modelos	Cloração			Fluoretação		
		REMQ	R^2	EMA	REMQ	R^2	EMA
2022	RNA-MLP	$1,90 \times 10^{-1}$	$9,89 \times 10^{-1}$	$6,37 \times 10^{-2}$	$6,11 \times 10^{-2}$	$9,96 \times 10^{-1}$	$1,39 \times 10^{-2}$
2022	FR	$1,49 \times 10^{-1}$	$9,93 \times 10^{-1}$	$3,93 \times 10^{-2}$	$1,28 \times 10^{-1}$	$9,85 \times 10^{-1}$	$3,17 \times 10^{-2}$
2022	XGBoost	$1,87 \times 10^{-1}$	$9,89 \times 10^{-1}$	$5,76 \times 10^{-2}$	$1,07 \times 10^{-1}$	$9,89 \times 10^{-1}$	$3,85 \times 10^{-2}$
2024	RNA-MLP	$4,17 \times 10^{-1}$	$9,44 \times 10^{-1}$	$9,64 \times 10^{-2}$	$2,47 \times 10^{-1}$	$9,44 \times 10^{-1}$	$4,40 \times 10^{-2}$
2024	FR	$4,03 \times 10^{-1}$	$9,48 \times 10^{-1}$	$7,79 \times 10^{-2}$	$2,45 \times 10^{-1}$	$9,45 \times 10^{-1}$	$5,39 \times 10^{-2}$
2024	XGBoost	$4,11 \times 10^{-1}$	$9,46 \times 10^{-1}$	$9,43 \times 10^{-2}$	$2,38 \times 10^{-1}$	$9,48 \times 10^{-1}$	$6,05 \times 10^{-2}$

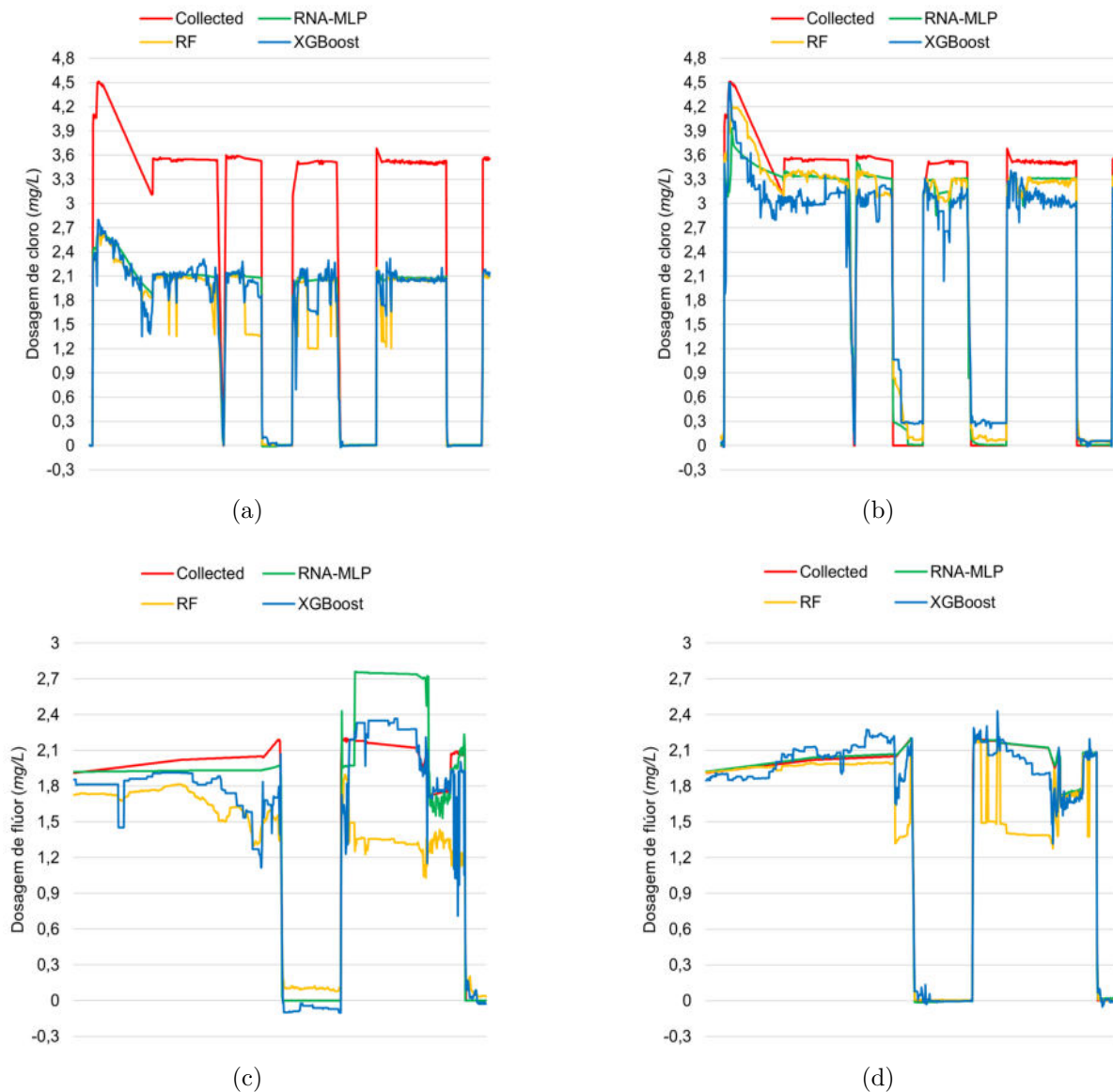


Figura 9 – Valores reais do sistema e previsões do modelo com amostra de 2024: (a) cloro sem otimização de hiperparâmetros, (b) cloro com otimização de hiperparâmetros, (c) flúor sem otimização de hiperparâmetros e (d) flúor com otimização de hiperparâmetros.

relação ao REMQ. Embora este resultado possa indicar sobreajuste, a validação externa

demonstra que as previsões mantêm conformidade com os valores aferidos durante a operação do sistema.

3.6.2 Avaliação estatística dos modelos preditivos

A análise estatística é conduzida para avaliar o desempenho dos modelos preditivos RNA-MLP, FR e XGBoost na estimativa das dosagens de cloro e flúor em ETA, estruturando-se em etapas que abrangem a definição de métricas de avaliação, a aplicação de testes de hipótese e a execução de análises visuais. Três métricas principais são empregadas para mensurar a precisão e a estabilidade dos modelos: o REMQ, o R^2 e o EMA. O teste de Mann-Whitney U, escolhido por sua robustez frente a distribuições não normais, é utilizado para verificar diferenças estatisticamente significativas entre os erros das previsões e a referência ideal de erro nulo. A representatividade e a redução de vieses são asseguradas por meio de extrações aleatórias de subamostras de tamanho N em cada rodada de análise. Os p -valores são interpretados com base no nível de significância $\alpha = 0.05$, sendo valores abaixo deste limiar indicativos de diferenças estatisticamente significativas, demonstrando que os erros não são atribuíveis ao acaso.

O REMQ é a medida que quantifica o desvio médio das previsões, penalizando mais intensamente erros de maior magnitude. O coeficiente R^2 representa a proporção da variabilidade explicada pelo modelo, sendo $R^2 \approx 1$ indicativos de melhor desempenho. O EMA fornece medida direta do erro médio das previsões em relação aos valores reais. O p -valor avalia a significância estatística, sendo $p\text{-valor} < 0.05$ indicativos de diferenças significativas entre as previsões e os valores reais. A Tabela 8 dispõe os valores de REMQ, R^2 , EMA e p -valor para cada modelo e parâmetro avaliado neste estudo, destacando os desempenhos individuais de cada abordagem.

Tabela 8 – Resultados das métricas de avaliação para os modelos preditivos de cloro e flúor.

Parâmetro	Modelo	REMQ	R^2	EMA	p -valor
Cloro	RNA-MLP	0,93	0,44	0,68	$1,11 \times 10^{-8}$
	FR	0,99	0,38	0,68	$5,33 \times 10^{-29}$
	XGBoost	0,96	0,41	0,65	$3,40 \times 10^{-28}$
	RNA-MLP Otimizado	0,66	0,72	0,56	$9,99 \times 10^{-5}$
	FR Otimizado	0,40	0,90	0,36	$1,07 \times 10^{-10}$
	XGBoost Otimizado	0,79	0,60	0,70	$1,11 \times 10^{-8}$
Flúor	RNA-MLP	0,51	0,45	0,44	$3,17 \times 10^{-225}$
	FR	0,30	0,81	0,21	$3,38 \times 10^{-232}$
	XGBoost	0,25	0,87	0,16	$2,37 \times 10^{-163}$
	RNA-MLP Otimizado	0,18	0,93	0,13	$7,17 \times 10^{-71}$
	FR Otimizado	0,19	0,92	0,13	$3,70 \times 10^{-207}$
	XGBoost Otimizado	0,15	0,95	0,09	$7,32 \times 10^{-18}$

De forma geral, todos os modelos apresentaram melhorias após o ajuste dos hiperparâmetros. No caso das previsões para cloro, o modelo FR otimizado destacou-se, atingindo os seguintes resultados: REMQ = 0,40, R^2 = 0,91, EMA = 0,36 e p -

valor = $1,07 \times 10^{-10}$. Para o flúor, o modelo XGBoost otimizado apresentou o melhor desempenho, com REMQ = 0,15, $R^2 = 0,95$, EMA = 0,09 e p -valor = $7,32 \times 10^{-18}$. Ao avaliar o modelo único para previsão simultânea de ambos os parâmetros, o XGBoost otimizado foi o mais adequado, atendendo aos seguintes critérios: i) os menores valores de REMQ para ambos os parâmetros, ii) elevado R^2 , indicando alta capacidade de explicação da variância dos dados, iii) resíduos sem padrões de tendência, corroborando a ausência de sobreajuste e iv) consistência das predições em diferentes períodos e subamostras. Estes fatores consolidam o XGBoost otimizado como o modelo mais robusto e confiável para aplicação no contexto avaliado.

Para aprofundar a análise e avaliar o desempenho dos modelos preditivos, são elaborados gráficos complementares que auxiliam na investigação da distribuição dos erros e do comportamento preditivo. Dentre estes, destacam-se os *boxplots*, que representam visualmente a dispersão dos erros cometidos pelos modelos. No *boxplot*, a caixa central ilustra o intervalo interquartil, correspondente aos valores entre o primeiro e o terceiro quartil, indicando onde se concentra a maior parte dos erros. A linha horizontal no interior da caixa representa a mediana, fornecendo a indicação do erro típico do modelo. Os extremos dos bigodes e os pontos fora destes destacam valores atípicos ou discrepantes, fornecendo informações relevantes sobre a estabilidade do modelo. Assim, estes gráficos oferecem uma análise visual intuitiva e comparativa entre diferentes modelos e cenários, contribuindo para a interpretação dos resultados.

Neste contexto foram elaboradas a Figura 10(a) e a Figura 10(b). Observa-se que os modelos otimizados apresentam menor dispersão, sugerindo previsões mais precisas. No caso do cloro, os modelos otimizados apresentam caixas mais compactas e linhas de alcance menores, indicando distribuição de erros mais concentrada. O modelo FR otimizado destaca-se por apresentar a mediana próxima de zero e número reduzido de valores atípicos, sugerindo maior consistência nas previsões. Para o flúor, o modelo XGBoost otimizado demonstra menor dispersão e valores próximos de zero, reforçando sua precisão e estabilidade.

Gráficos de resíduos são empregados para identificar padrões ou desvios sistemáticos nas diferenças entre os valores reais e os previstos pelos modelos. Estes gráficos permitem verificar se os erros estão distribuídos de forma aleatória ou se apresentam tendências específicas, o que pode indicar problemas no ajuste do modelo, como subajuste (*underfit*) ou sobreajuste (*overfit*). Adicionalmente, gráficos de dispersão são gerados para avaliar a proximidade entre os valores previstos e os reais. Nestes gráficos, a linha ideal $p = r$, que representa a igualdade entre os valores previstos e reais, é utilizada como referência. Quanto mais próximas as observações estiverem desta linha, maior será a precisão do modelo analisado.

Análises visuais dos resíduos foram realizadas para identificar padrões ou desvios

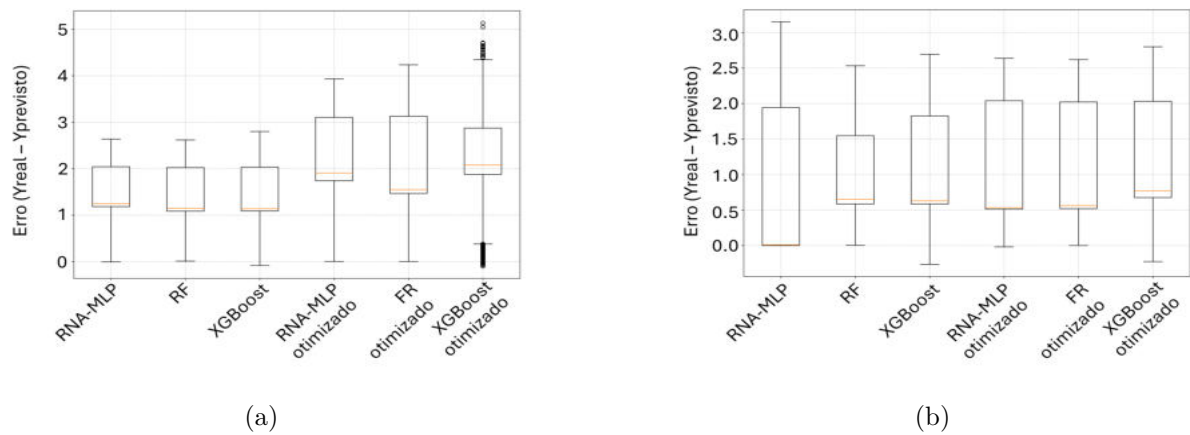


Figura 10 – Distribuição dos erros dos modelos preditivos para: (a) cloro e (b) flúor.

sistemáticos nas diferenças entre os valores reais e previstos pelos modelos. A Figura 11 exibe as previsões ao longo das observações no eixo das abscissas, enquanto o eixo das ordenadas apresenta a magnitude dos erros (diferença entre valores reais e previstos). Os pontos azuis representam os erros individuais de cada observação, e a linha horizontal vermelha indica o erro nulo, correspondente a previsões exatas.

A análise indica que o modelo otimizado XGBoost apresenta menor dispersão dos resíduos em torno da linha de erro nulo, apresentando maior precisão. A amplitude reduzida dos erros e a menor quantidade de pontos distantes da linha vermelha sugerem menor ocorrência de erros extremos. Por outro lado, o modelo otimizado RNA-MLP demonstra redução na variabilidade dos erros em comparação à sua versão não otimizada, embora sua dispersão permaneça maior que a do XGBoost, indicando menor consistência nas previsões. Estes resultados destacam a superioridade do modelo XGBoost otimizado em termos de precisão e estabilidade na estimativa das dosagens.

A seleção do modelo mais adequado baseia-se em múltiplos critérios, incluindo o desempenho nas métricas avaliadas, a consistência das previsões em diferentes períodos e subamostras, além da ausência de tendências ou padrões nos resíduos. Também se considera a capacidade do modelo de prever simultaneamente as dosagens de cloro e flúor, o que constitui um diferencial relevante em termos de flexibilidade e aplicabilidade prática. Estes critérios asseguram que o modelo escolhido seja resiliente, confiável e apropriado para os objetivos estabelecidos no contexto operacional das ETA. A análise visual dos resíduos demonstra que os modelos otimizados apresentam maior concentração de erros próximos de zero, evidenciando sua precisão preditiva. A distribuição aleatória dos erros, sem padrões sistemáticos, indica que os modelos não são tendenciosos e apresentam comportamento consistente em suas estimativas.

Para complementar a análise, foi avaliada a dispersão entre os valores previstos e reais, como apresentado na Figura 12, no qual, cada ponto azul representa uma previsão em relação ao valor real correspondente, enquanto a linha vermelha indica os valores ideais,

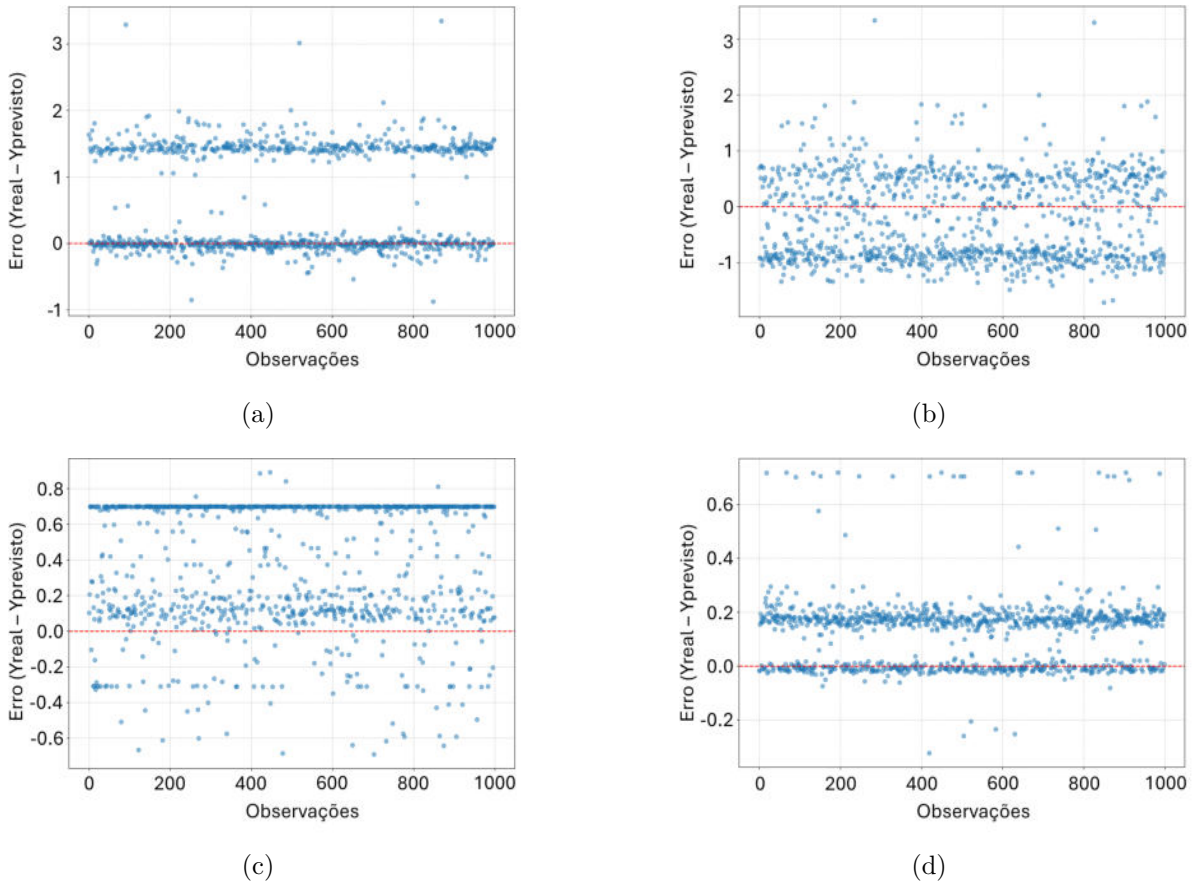


Figura 11 – Resíduos dos modelos preditivos ao longo das observações: (a) XGBoost sem otimização cloro, (b) XGBoost otimizado cloro, (c) RNA-MLP sem otimização flúor e (d) RNA-MLP otimizado flúor.

onde $p = r$ (valor previsto igual ao real). A proximidade dos pontos à linha vermelha reflete a precisão do modelo, sendo mais evidente nos modelos otimizados, especialmente no XGBoost, que demonstrou desempenho superior em termos de alinhamento com os valores reais.

A análise visual dos resíduos demonstra que os modelos otimizados apresentam maior concentração de erros próximos de zero, reforçando sua precisão preditiva. A distribuição aleatória dos resíduos, sem padrões sistemáticos, indica ausência de viés e comportamento adequado do modelo. Adicionalmente, a menor ocorrência de valores atípicos nos modelos XGBoost indica maior estabilidade e melhor capacidade de generalização, corroborando a inexistência de sobreajuste nas predições. Para apresentar a proximidade entre os valores previstos e reais, foi analisada a dispersão, como apresentado na Figura 12, no qual, cada ponto azul representa uma previsão em relação ao valor real correspondente, enquanto a linha vermelha indica os valores ideais, onde $p = r$.

Para o modelo RNA-MLP, observa-se uma dispersão considerável dos pontos ao redor da linha ideal, caracterizada por uma tendência de subestimação ou superestimação em diferentes intervalos de valores, o que sugere uma menor precisão nas predições. Mesmo

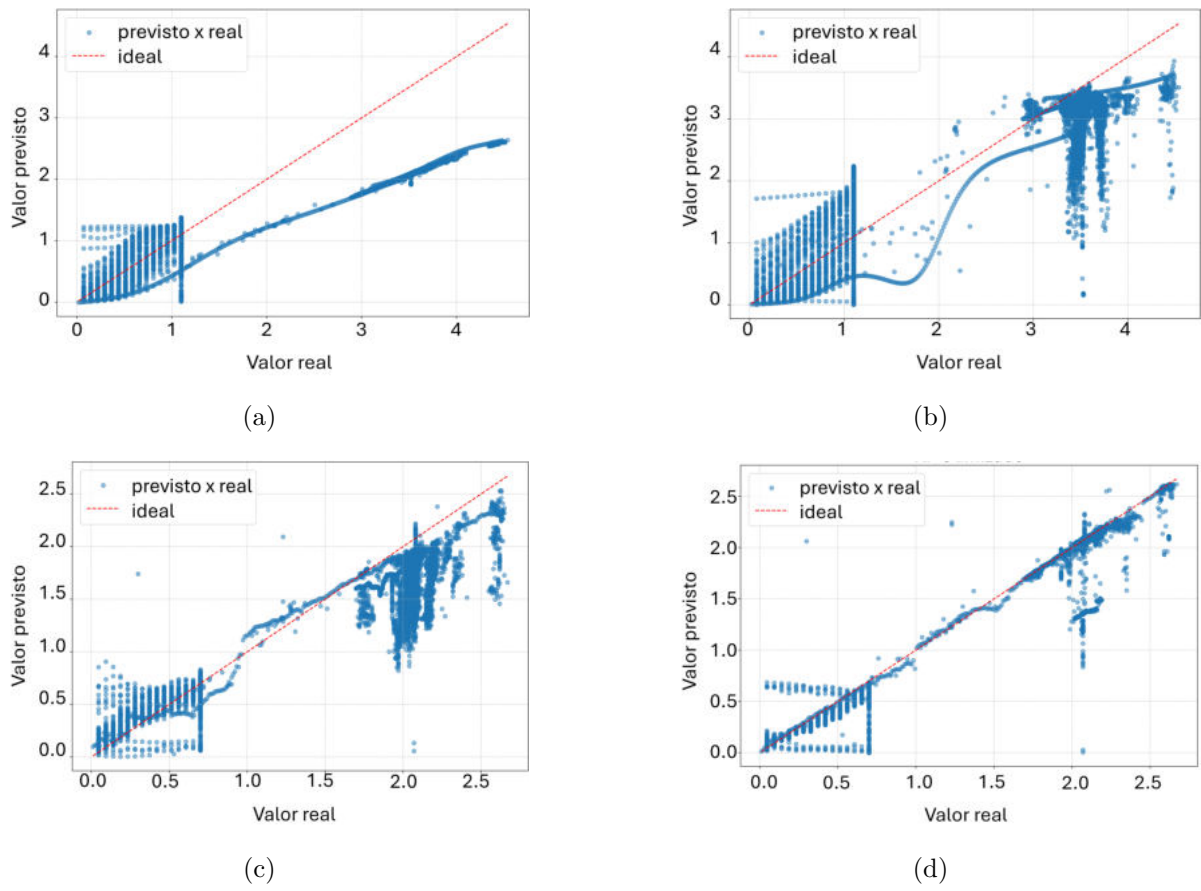


Figura 12 – Dispersão com valores reais $\times$ previstos: (a) RNA-MLP sem otimização cloro, (b) RNA-MLP otimizado cloro, (c) FR sem otimização flúor e (d) FR otimizado flúor.

após a otimização, o RNA-MLP ainda apresenta padrões não lineares nos resíduos e maior concentração de pontos afastados da linha vermelha, evidenciando dificuldades na captura exata dos valores reais. Por outro lado, o modelo FR demonstra uma distribuição mais próxima da linha ideal, com menor dispersão. No entanto, a presença de alguns pontos isolados e distantes aponta desafios específicos em prever certos valores com alta precisão. Após a otimização, o FR exibe uma distribuição mais uniforme e concentrada em torno da linha ideal, com significativa redução dos desvios e maior proximidade dos pontos à linha vermelha, indicando melhorias relevantes na capacidade de generalização e previsibilidade. A ausência de agrupamentos significativos de pontos distantes reforça que os modelos otimizados não exibem indícios de sobreajuste ou subajuste.

4 Conclusão

Este relatório avaliou a eficiência de modelos de inteligência artificial na previsão das dosagens de cloro e flúor em Estação de Tratamento de Água (ETA), com o objetivo de aprimorar os processos de tratamento de água e aumentar a resiliência operacional. A metodologia, aplicada na ETA de Brazabrantes, Goiás, Brasil, incluiu a coleta, tratamento e segmentação de dados em séries temporais, seguida pela aplicação de três modelos preditivos: Redes Neurais Artificiais - *Multi Layer Perceptron* (RNA-MLP), Floresta Randômica (FR) e *Extreme Gradient Boosting* (XGBoost). Os modelos foram testados em duas configurações: com e sem ajustes de hiperparâmetros, sendo avaliados com base em critérios estatísticos como raiz do erro médio quadrático (REMQU), coeficiente de determinação (R^2), erro médio absoluto (EMA) e p -valor.

Os resultados indicaram a melhoria dos modelos otimizados, com destaque para o FR otimizado na previsão das dosagens de cloro, devido à sua alta capacidade de generalização e menores valores de erro. Por outro lado, o XGBoost otimizado demonstrou maior eficiência na previsão de flúor e na modelagem conjunta dos parâmetros, apresentando maior precisão, estabilidade e consistência em diferentes períodos e condições. A ausência de padrões sistemáticos nos resíduos reforça a resiliência dos modelos otimizados, garantindo previsões confiáveis para o manejo das variabilidades nas condições da água tratada. A organização didática deste relatório oferece um recurso valioso para futuras pesquisas e implementações práticas em estações de tratamento de água, ampliando o alcance da inteligência artificial como ferramenta estratégica na gestão de recursos hídricos.

Referências

- AL-HAIJA, Q. A. Exploration of tools for data science. *Data Science with Semantic Technologies: Theory, Practice, and Application*, Wiley Online Library, p. 31–69, 2022. Citado na página 28.
- ALIASHRAFI, A. et al. A review of data-driven modelling in drinking water treatment. *Reviews in Environmental Science and Bio/Technology*, v. 20, n. 4, p. 985–1009, 2021. ISSN 1572-9826. Citado na página 29.
- BLEI, D. M.; SMYTH, P. Science and data science. *Proceedings of the National Academy of Sciences*, v. 114, n. 33, p. 8689–8692, 2017. Citado na página 28.
- BREIMAN, L. Random forests. *Machine Learning*, Netherlands, v. 45, p. 5–32, October 2001. Citado 2 vezes nas páginas 29 e 30.
- BULHOES, J. S. et al. Indirect prediction system for variables that have gaps in their time series. *Chaos, Solitons & Fractals*, Elsevier, v. 131, p. 109509, 2020. Citado na página 30.
- CHANDRA, S. S. et al. Deep neural network architectures for spectrum sensing using signal processing features. In: IEEE. *2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics*. DC, USA, 2021. p. 129–134. Citado na página 29.
- CHARROIS, J. W. Private drinking water supplies: challenges for public health. *Canadian Medical Association Journal*, Canadian Medical Association, v. 182, n. 10, p. 1061–1064, 2010. Citado 2 vezes nas páginas 27 e 28.
- CHEN, D. et al. Solving the problem of runge phenomenon by pseudoinverse cubic spline. In: IEEE. *2014 IEEE 17th International Conference on Computational Science and Engineering*. Chengdu, China, 2014. p. 1226–1231. Citado na página 30.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: . New York, NY, USA: Association for Computing Machinery, 2016. p. 785–794. Citado na página 30.
- CHHETRI, R. K. et al. Treatment of arctic wastewater by chemical coagulation, uv and peracetic acid disinfection. *Environmental Science and Pollution Research*, Springer, v. 25, p. 32851–32859, 2018. Citado na página 25.
- CHOI, J.; LIM, K. J.; JI, B. Robust imputation method with context-aware voting ensemble model for management of water-quality data. *Water Research*, Elsevier BV, v. 243, p. 120369, 9 2023. ISSN 0043-1354. Citado na página 30.
- CLEVELAND, R. B. et al. Stl: A seasonal-trend decomposition. *J. off. Stat*, v. 6, n. 1, p. 3–73, 1990. Citado 2 vezes nas páginas 30 e 37.
- DAYHOFF, J. E. *Neural network architectures: an introduction*. New York, NY: Van Nostrand Reinhold Co., 1990. Citado na página 29.

- DESTIANI, R.; TEMPLETON, M. Chlorination and ultraviolet disinfection of antibiotic-resistant bacteria and antibiotic resistance genes in drinking water. *AIMS Environmental Science*, v. 6, n. 3, 2019. Citado na página 26.
- DURRLEMAN, S.; SIMON, R. Flexible regression models with cubic splines. *Statistics in Medicine*, v. 8, n. 5, p. 551–561, 1989. Citado na página 37.
- FAUST, S. M.; ALY, O. M. *Chemistry of water treatment*. 2nd. ed. Boca Raton, FL: Taylor and Francis Group, 1998. 485–565 p. Citado na página 25.
- FAWELL, J. et al. *Fluoride in drinking-water*. London, UK: World Health Organization, 2006. v. 1. Citado 2 vezes nas páginas 39 e 40.
- FEDORKO, R.; KRÁL', Š.; FEDORKO, I. Artificial intelligence and machine learning in the context of e-commerce: A literature review. *Communication and Intelligent Systems: Proceedings of ICCIS 2021*, Springer, p. 1067–1082, 2022. Citado na página 28.
- FERNÍCOLA, N. G. G. d.; AZEVEDO, F. A. d. Metemoglobinemia e nitrato nas águas. *Revista de Saúde Pública*, Faculdade de Saúde Pública da Universidade de São Paulo, v. 15, n. 2, p. 242–248, Apr 1981. Citado na página 26.
- FORNBERG, B.; ZUEV, J. The runge phenomenon and spatially variable shape parameters in rbf interpolation. *Computers & Mathematics with Applications*, Elsevier, v. 54, n. 3, p. 379–398, 2007. Citado na página 30.
- GARBIN, C. A. S. et al. Fluoridation of public water supply: bioethical, legal and political approach. *Revista Bioética*, SciELO Brasil, v. 25, p. 328–337, 2017. Citado na página 27.
- GHAHRAMANI, Z. Probabilistic machine learning and artificial intelligence. *Nature*, Nature Publishing Group UK London, v. 521, n. 7553, p. 452–459, 2015. Citado na página 28.
- GHOBADI, A. et al. Groundwater quality modeling using a novel hybrid data-intelligence model based on gray wolf optimization algorithm and multi-layer perceptron artificial neural network: a case study in asadabad plain, hamedan, iran. *Environmental Science and Pollution Research*, Springer, p. 1–15, 2022. Citado na página 29.
- GOURMELON, M. et al. Safe management of shellfish and harvest waters. WHO, IWA Publishing, London, p. 297, 2010. Citado na página 29.
- HOFKES, E. *Small community water supplies*. Netherlands: International Reference Centre for Community Water Supply and Sanitation, 1981. Citado na página 25.
- HRUDEY, S. E. Chlorination disinfection by-products, public health risk tradeoffs and me. *Water research*, Elsevier, v. 43, n. 8, p. 2057–2092, 2009. Citado na página 21.
- HUO, Y. et al. Long-term span traffic prediction model based on stl decomposition and lstm. In: IEEE. *2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. Matsue, Japan, 2019. p. 1–4. Citado na página 30.
- JAYAWEERA, C.; AZIZ, N. Development and comparison of extreme learning machine and multi-layer perceptron neural network models for predicting optimum coagulant dosage for water treatment. In: IOP PUBLISHING. *Journal of physics: conference series*. United Kingdom, 2018. v. 1123, n. 1, p. 012032. Citado na página 29.

- JENA, P. K. et al. Surface water quality assessment by Random Forest. *Water Practice and Technology*, v. 18, n. 1, p. 201–214, 12 2022. Citado na página 29.
- JENKINS, J. M.; KOWALSKI, M.; ALVARENGA, E. F. S. Predictive modelling of water losses using random forests on weather covariates. *Water Supply*, v. 18, n. 6, p. 2180–2187, 02 2018. Citado na página 29.
- JOHNSON, S. F. Methemoglobinemia: Infants at risk. *Current problems in pediatric and adolescent health care*, Elsevier, v. 49, n. 3, p. 57–67, 2019. Citado na página 26.
- JOLLY, K. *Machine learning with scikit-learn quick start guide: classification, regression, and clustering techniques in Python*. Birmingham, Reino Unido: Packt Publishing Ltd, 2018. Citado na página 29.
- KAUR, A. et al. Health impacts due to fluoride contamination in water: current scenario. In: *Contamination of Water*. New Delhi, India: Academic Press, 2021. p. 65–84. ISBN 978-0-12-824058-8. Citado na página 27.
- KHAN, M. et al. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University - Computer and Information Sciences*, v. 34, 06 2021. Citado na página 21.
- KIM, M.-K.; ZOH, K.-D. Occurrence and removals of micropollutants in water environment. *Environmental Engineering Research*, Korean Society of Environmental Engineers, v. 21, n. 4, p. 319–332, 2016. Citado na página 25.
- LAINÉ, J. M.; VIAL, D.; MOULART, P. Status after 10 years of operation - overview of uf technology today. *Desalination*, v. 131, p. 17–25, 2000. Citado na página 25.
- LASHARI, S. e. Z. et al. Drilling performance monitoring and optimization: a data-driven approach. *Journal of Petroleum Exploration and Production Technology*, Springer, v. 9, p. 2747–2756, 2019. Citado na página 29.
- LAWSON, A. E. The generality of hypothetico-deductive reasoning: Making scientific thinking explicit. *The American Biology Teacher*, JSTOR, v. 62, n. 7, p. 482–495, 2000. Citado na página 33.
- LEVY, D. H. et al. Association of nationwide water fluoridation, changes in dental care legislation, and caries-related treatment needs: A 9-year record-based cross-sectional study. *Journal of Dentistry*, v. 134, p. 104550, 2023. ISSN 0300-5712. Citado na página 27.
- LI, X. et al. Identification of mine mixed water inrush source based on genetic algorithm and xgboost algorithm: A case study of huangyuchuan mine. *Water*, v. 14, n. 14, 2022. Citado na página 30.
- LIBRANTZ, A. F.; SANTOS, F. C. R. d.; DIAS, C. G. Artificial neural networks to control chlorine dosing in a water treatment plant. *Acta Scientiarum. Technology*, v. 40, n. 1, p. e37275, set. 2018. ISSN 1807-8664. Citado na página 21.
- LUKENS, J. N. The legacy of well-water methemoglobinemia. *Jama*, American Medical Association, v. 257, n. 20, p. 2793–2795, 1987. Citado na página 26.

- MARTINEZ, T. T.; LONG, C. Explosion risk from swimming pool chlorinators and review of chlorine toxicity. *Journal of Toxicology: Clinical Toxicology*, Taylor & Francis, v. 33, n. 4, p. 349–354, 1995. Citado na página 26.
- MIKLOS, D. B. et al. Evaluation of advanced oxidation processes for water and wastewater treatment - a critical review. *WATER RESEARCH*, PERGAMON-ELSEVIER SCIENCE LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, ENGLAND, v. 139, p. 118–131, AUG 1 2018. ISSN 0043-1354. Citado na página 25.
- MO, T.; LI, S.; LI, G. An interpretable machine learning model for predicting cavity water depth and cavity length based on xgboost-shap. *JOURNAL OF HYDROINFORMATICS*, v. 25, n. 4, p. 1488–1500, JUL 2023. Citado na página 30.
- MORALES, E. F.; ESCALANTE, H. J. A brief introduction to supervised, unsupervised, and reinforcement learning. In: *Biosignal processing and classification using computational learning and intelligence*. Puebla, Mexico: Elsevier, 2022. p. 111–129. Citado na página 28.
- MORE, K. S.; WOLKERSDORFER, C. Application of machine learning algorithms for nonlinear system forecasting through analytics — a case study with mining influenced water data. *Water Resources and Industry*, v. 29, p. 100209, 2023. ISSN 2212-3717. Citado 2 vezes nas páginas 21 e 30.
- MULLANE, D. O. et al. Fluoride and oral health. *Community dental health*, Dennis Barber & The British Association for the Study of Community Dentistry, v. 33, n. 2, p. 69–99, 2016. Citado na página 27.
- MULLEN, J. History of water fluoridation. *British Dental Journal*, Nature Publishing Group, v. 199, n. 7, p. 1–4, 2005. Citado na página 27.
- MURRAY, J. J.; World Health Organization et al. *Appropriate use of fluorides for human health*. Geneve, Switzerland: World Health Organization, 1986. Citado 2 vezes nas páginas 21 e 27.
- NAMOR, A. Danil de. Water purification: from ancient civilization to the xxi century. *Water Science and Technology: Water Supply*, IWA Publishing, v. 7, n. 1, p. 33–39, 2007. Citado na página 25.
- NEGOITESCU, A.; TOKAR, A. Water disinfection and the implementation of modern technologies at potable water treatment station. *Revista de Chimie*, v. 63, n. 10, p. 1079–1089, 2012. Citado na página 25.
- NEWBRUN, E. Effectiveness of water fluoridation. *Journal of Public Health Dentistry*, Wiley Online Library, v. 49, n. 5, p. 279–289, 1989. Citado 3 vezes nas páginas 21, 27 e 28.
- OCAMPO-RODRÍGUEZ, D. B. et al. Water disinfection: a review of conventional and advanced treatments with chlorine and peracetic acid. *Ingeniería del agua*, IWA Publishing, v. 26, n. 3, p. 185–204, 2022. Citado 3 vezes nas páginas 21, 25 e 26.
- OUYANG, Z.; JABLOUN, M.; RAVIER, P. Stlformer: exploit stl decomposition and rank correlation for time series forecasting. In: IEEE. *2023 31st European Signal Processing Conference (EUSIPCO)*. Helsinki, Finland, 2023. p. 1405–1409. Citado na página 30.

- PERES, M. A.; FERNANDES, L. S.; PERES, K. G. Inequality of water fluoridation in southern brazil—the inverse equity hypothesis revisited. *Social Science & Medicine*, Elsevier, v. 58, n. 6, p. 1181–1189, 2004. Citado 2 vezes nas páginas 27 e 28.
- PERRITT, H. H. *Law and the information superhighway*. New York, NY, USA: Wolters Kluwer, 2001. Citado na página 34.
- PINTO, A. et al. Optimizing water treatment systems using artificial intelligence based tools. *WIT Transactions on Ecology and the Environment*, Southampton, Hampshire, UK, v. 125, p. 185–194, 2009. ISSN 1743-3541. Citado na página 21.
- RHEE, F. V.; BEAUMONT, D. Gastric stricture complicating oral ingestion of bleach. *British journal of clinical practice*, Medicom International, v. 44, n. 12, p. 681–682, 1990. Citado na página 26.
- RIEPL, D. et al. Making informed decisions—a collaborative and knowledge based iwrn planning exercise in wadishueib, jordan. *Proceedings IWRM*, 2010. Citado na página 28.
- RODRIGUEZ, R. et al. Water-quality data imputation with high percentage of missing values: A machine learning approach. *Sustainability*, 2021. Citado na página 30.
- ROY, S.; DASS, G. Fluoride contamination in drinking water—a review. *Resour. Environ*, v. 3, n. 3, p. 53–58, 2013. Citado na página 27.
- SADEQ, M. et al. Drinking water nitrate and prevalence of methemoglobinemia among infants and children aged 1–7 years in moroccan areas. *International journal of hygiene and environmental health*, Elsevier, v. 211, n. 5-6, p. 546–554, 2008. Citado na página 26.
- SAFEER, S. et al. A review of artificial intelligence in water purification and wastewater treatment: Recent advancements. *Journal of Water Process Engineering*, Elsevier, v. 49, p. 102974, 2022. Citado na página 29.
- SEBASTIAN, K.; GAO, H.; XING, X. Utilizing an ensemble stl decomposition and gru model for base station traffic forecasting. In: IEEE. *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. Chiang Mai, Thailand, 2020. p. 314–319. Citado na página 30.
- SILVA, W. P. da et al. Peracetic acid: Structural elucidation for applications in wastewater treatment. *Water Research*, Elsevier, v. 168, p. 115143, 2020. Citado na página 26.
- SLAUGHTER, R. J. et al. The clinical toxicology of sodium hypochlorite. *CLINICAL TOXICOLOGY*, v. 57, n. 5, p. 303–311, MAY 4 2019. ISSN 1556-3650. Citado 2 vezes nas páginas 21 e 26.
- SOLAIMANY-AMINABAD, M.; MALEKI, A.; HADI, M. Application of artificial neural network (ann) for the prediction of water treatment plant influent characteristics. *Journal of advances in environmental health research*, Kurdistan University of Medical Sciences, v. 1, n. 2, p. 89–100, 2013. Citado na página 21.
- TAHERSHAMSI, A.; MENHAJ, M.; AHMADIAN, R. Sediment loads prediction using multilayer feedforward neural networks. *Amirkabir*, v. 16, n. 63, p. 103–110, 2006. Citado na página 29.

- VEIGA, F. da; KALBUSCH, A.; HENNING, E. Drivers of urban water consumption in brazil: a countrywide, cross-sectional study. *Urban Water Journal*, Taylor & Francis, v. 20, n. 10, p. 1462–1470, 2023. Citado na página 34.
- WANG, D. et al. Water plant optimization control system based on machine learning. *DESALINATION AND WATER TREATMENT*, v. 222, p. 168–181, MAY 2021. ISSN 1944-3994. Citado na página 21.
- WEGMAN, E. J.; WRIGHT, I. W. Splines in statistics. *Journal of the American Statistical Association*, v. 78, n. 382, p. 351, 1983. Citado na página 37.
- World Health Organization. *Guidelines for drinking-water quality*. Geneve, Switzerland: World Health Organization, 2002. Citado na página 40.
- World Health Organization. *Guidelines for drinking-water quality: fourth edition incorporating the first and second addenda*. Geneve, Switzerland: World Health Organization, 2022. v. 4. Citado 2 vezes nas páginas 39 e 40.
- WU, R. et al. Data Imputation for Multivariate Time Series Sensor Data With Large Gaps of Missing Data. *IEEE Sensors Journal*, Institute of Electrical and Electronics Engineers (IEEE), v. 22, n. 11, p. 10671–10683, jun 1 2022. ISSN 1530-437X. Citado na página 30.
- ZHANG, Y. et al. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water research*, Elsevier, v. 164, p. 114888, 2019. Citado na página 21.
- ZHAO, L. et al. Prediction of soil moisture using bigru-lstm model with stl decomposition in qinghai–tibet plateau. *PeerJ*, PeerJ Inc., v. 11, p. e15851, 2023. Citado na página 30.
- ZOU, H.; TANG, H. Comparison of different bacteria inactivation by a novel continuous-flow ultrasound/chlorination water treatment system in a pilot scale. *Water*, MDPI, v. 11, n. 2, p. 258, 2019. Citado na página 26.

Apêndices

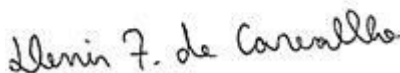
APÊNDICE A – Solicitação dos dados fundado na Lei de Acesso à Informação

Termo de Autuação de Processo

Número: 14539/2022

Nesta data, às 17:35, procedemos a autuação dos documentos com protocolo nº para formação do processo nº 14634/2022, que se inicia com o sequencial de páginas.

Goiânia, 13 de setembro de 2022



DENIS FERREIRA DE CARVALHO - .

G-GDI - GERÊNCIA DE GESTÃO DOCUMENTAL E INFORMAÇÃO em 13/09/2022 17:35:33

Tipo MEMORANDO

Número 38302/2022

Data 13/09/2022

De SUPERINTENDÊNCIA DE OUVIDORIA GERAL

Para GERÊNCIA DE GESTÃO DOCUMENTAL E INFORMAÇÃO - G-GDI - G0384

Assunto SERVIÇO DE INFORMAÇÃO AO CIDADÃO - SIC - RO 4730/2022

Solicito a abertura de Processo Administrativo, em atendimento a solicitação constante no **Registro de Ocorrência – RO nº 4730/2022**. Após autuação, devolver a esta unidade de Ouvidoria.

Interessado: **DIEGO TAKASHI SATO - RO 4730/2022**

Destinatário: **DIPRO**

Origem: **OGE – LAI**

Localidade: **GOIÂNIA/GO**

Assunto: **SERVIÇO DE INFORMAÇÃO AO CIDADÃO - SIC**

Atenciosamente,



STELLA MARCIA O MODES LINO - SUPERINTENDENTE A2
SUOGE - SUPERINTENDÊNCIA DE OUVIDORIA GERAL em 13/09/2022 16:56:35

Registro de Ocorrência Nº 4730/2022

Data:	13/09/2022 16:40:42	Status:	ABERTO	Situação:	
Origem:	5 - OGE	Tipo:	8 - INFORMAÇÃO	Contato:	8 - INTERNET
Motivo:	SIC-SERVIÇO DE INFORMAÇÃO AO CIDADÃO				
Desdobrament	EM ANÁLISE				
Local:	-				
Nº OGE:	2022091315255063			Data OGE:	13/09/2022
Nº AGR:	2022091315255063			Data AGR:	13/09/2022
Processo:	14634/2022				

Solicitante

CPF:		Nome:	DIEGO TAKASHI SATO
Telefones:			

Identificação do Cliente

Conta:	
Nome:	
Cidade:	GOIANIA
Endereço:	

Ocorrência

Descrição:
Texto transcrito na íntegra, conforme manifestação registrada na OGE: Em atendimento às orientações do processo n.º 238830/2022, estou registrando o pedido de informações conforme solicitação em anexo.

Encaminhamento(s)

Nenhum encaminhamento para este registro de ocorrência

Fechamento

Data:	
Descrição:	

Registro(s) de Atendimento (RA´s)

Nenhum RA para este registro de ocorrência

Anexo do Registro

Data	Nome
13/09/2022	ANEXO_RO_13-09-2022-045713005.pdf

Data	Nome
------	------

Anexo do Fechamento

Nenhum anexo

SOLICITAÇÃO DE AUTORIZAÇÃO DE USO DE DADOS

À Diretoria de Gestão Corporativa – DICOR.

Eu, Diego Takashi Sato, matrícula funcional [REDACTED] venho requerer autorização para acesso e utilização de dados dos sistemas Supervisório Web e Supervisório, gerenciados pela V0143 - SUPERV. DE PESQUISA, DESENVOLVIMENTO E AUTOMAÇÃO (P-SDA), para fins de pesquisa no Programa de Pós Graduação em Tecnologia, Gestão e Sustentabilidade do IFG Câmpus Goiânia.

O objetivo é desenvolver pesquisa para melhorias que serão cedidas sem ônus à Saneago. Comprometo-me em caráter irrevogável, manter o sigilo e a confidencialidade em relação aos dados base/banco de dados. Garanto que as informações a serem coletadas, descritas acima, serão exclusivamente para realização do presente projeto de pesquisa.

Goiânia, 13 de setembro de 2022.



Documento assinado digitalmente

DIEGO TAKASHI SATO

Data: 13/09/2022 13:32:35-0300

Verifique em <https://verificador.iti.br>

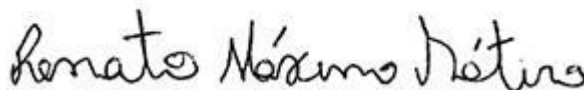
Tipo DESPACHO**Número** 48800/2022**Data** 14/09/2022**Processo** 14634/2022**Destinatário** DIRETORIA DE PRODUCAO. - DIPRO - D6000**Assunto** SERVIÇO DE INFORMAÇÃO AO CIDADÃO - SIC - RO 4730/2022

Encaminhamos o processo para conhecimento e providências, conforme solicitação registrada no **Registro de Ocorrência n.º 4730/2022, páginas 4 e 5.**

Solicitamos o envio a unidade de Ouvidoria até o dia **23/09/2022**, para o devido fechamento, a fim de evitar os transtornos oriundos da devolução fora do prazo.

Assim como, se for necessário, a Unidade supracitada poderá verificar junto a Assessoria Jurídica desta Diretoria, em atenção a Lei Estadual 18.025/2013, se existe algum impedimento no fornecimento da informação solicitada. **Caso exista, a informação deverá ser encaminhada a esta UO de Ouvidoria concluída, com as informações sigilosas já retiradas ou tarjadas, principalmente as informações pessoais, em atendimento a Lei n.º 13.709/2018.**

Atenciosamente,



RENATO MAXIMO SATIRO - ADMINISTRADOR

SUOGE - SUPERINTENDÊNCIA DE OUVIDORIA GERAL em 14/09/2022 10:06:16



STELLA MARCIA O MODES LINO - SUPERINTENDENTE A2

SUOGE - SUPERINTENDÊNCIA DE OUVIDORIA GERAL em 14/09/2022 09:46:37

Tipo DESPACHO**Número** 49233/2022**Data** 15/09/2022**Processo** 14634/2022**Destinatário** SUPER. DE OUVIDORIA GERAL - SUOGE - S0079**Assunto** SERVIÇO DE INFORMAÇÃO AO CIDADÃO - SIC -

Em atendimento ao Despacho 48800/2022 (pág. 6) e a solicitação contida no Registro de Ocorrência N° 4730/2022 (págs. 3/5), esta Diretoria de Produção **autoriza** o uso dos dados pelo funcionário Diego Takashi Sato, matrícula [REDACTED] dos sistemas Supervisório Web e Supervisório para fins de pesquisa no Programa de Pós-Graduação em Tecnologia, Gestão e Sustentabilidade do IFG Campus Goiânia.

Atenciosamente,



DANIEL VIEIRA CARNEIRO DE MENDONCA - ASSESSOR DIRETORIA B1

DIPRO - DIRETORIA DE PRODUCAO. em 15/09/2022 17:16:33

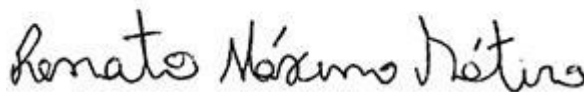
Tipo TERMO DE RESPOSTA DA OUVIDORIA**Número** 172/2022**Data** 16/09/2022**Dados do Requerimento de Informação****Protocolo** 2022091315255063**Data** 16/09/2022**Resumo** SERVIÇO DE INFORMAÇÃO AO CIDADÃO - SIC - RO 4730/2022

Prezado Senhor,

A Saneago informa que, segue anexo o Despacho nº 49233/2022 - DIPRO, o qual encaminha resposta acerca do pedido efetuado.

A Saneago agradece o contato e coloca-se à disposição.

Atenciosamente,



RENATO MAXIMO SATIRO - ADMINISTRADOR

SUOGE - SUPERINTENDÊNCIA DE OUVIDORIA GERAL em 16/09/2022 14:46:28



STELLA MARCIA O MODES LINO - SUPERINTENDENTE A2

SUOGE - SUPERINTENDÊNCIA DE OUVIDORIA GERAL em 16/09/2022 14:24:48

Importante: no caso de indeferimento de acesso à informação ou caso a considere insatisfatória, poderá ser interposto recurso através do Sistema de Ouvidoria (<http://www.cge.go.gov.br/ouvidoria/Consulta.php>) no prazo de 10 (dez) dias, contados a partir da ciência da resposta, conforme disposto no art. 20 da Lei nº 18.025/2013.

Número RO: 4730/2022

Data: 16/09/2022

Dados do Requerimento de Informação

Protocolo: 2022091315255063

Data: 13/09/2022

Solicitante: DIEGO TAKASHI SATO

Resumo: INFORMAÇÕES DIVERSAS/TÉC/ACADÊMICAS

Manifestação:

Texto transcrito na íntegra, conforme manifestação registrada na OGE:

Em atendimento às orientações do processo n.º 238830/2022, estou registrando o pedido de informações conforme solicitação em anexo.

Teor da Resposta:

A Saneago agradece o contato e informa que, segue anexo Termo de Resposta n.º 172/2022, referente ao seu Pedido de Acesso a Informação – LAI.

Atenciosamente,

Ouvidoria da Saneago.



ALBINO FERREIRA JUNIOR
GER. DE TRANSPARÊNCIA E INF.



STELLA MARCIA O MODES LINO
SUPER. DE OUVIDORIA GERAL

Importante: Esta é uma mensagem enviada automaticamente. Não responder este e-mail. No caso de insatisfação da resposta, poderá contestá-la, através dos canais de atendimento da Ouvidoria (www.saneago.com.br, em Ouvidoria ou, 0800 645 0117), no prazo de 10 (dez) dias, contados a partir da data deste Termo.

Saneamento de Goiás – Saneago – Av. Fued José Sebba, nº 1.245, Jardim Goiás, Goiânia/GO, CEP: 74.805-100 – Telefone: 3243-3276

Formulário
FR04.2024

Revisão
00

Data
26/10/2017

Cód. TTD
019.01

Doc. Normativo
====

UO Responsável
SUOGE

Página
1 de 1



Sistema de Ouvidoria do Estado de Goiás

Sexta-Feira, 16 de Setembro de 2022.

Bem vindo(a): [Renato Maximo Satiro](#)

Cadastro

Manifestações

Relatórios

Configurações

Sair

Manifestação

Localização

Detalhes da manifestação

Módulo de análise e edição

**Manifestação Finalizada****2022.0913.152550-63**Tipo: **L.A.I.**Canal: **- Origem Externa**

Descrição do fato

Em atendimento às orientações do processo n.º 238830/2022, estou registrando o pedido de informações conforme solicitação em anexo.

Orgão que fará a verificação: **Saneamento do Estado de Goiás S/A**

Município do Fato: **GOIANIA**

Local do Fato: **[REDACTED]**

[Editar manifestação](#)[Complementação](#)

Teste Beta - IA

Avaliação de intenção:

Avaliação de tipologia:

Assunto/Subassunto

Assunto: Saneamento Básico/Água

Subassunto: Informações Acadêmicas/ Diversas/Técnicas

Anexos



Solicitante

Tipo manif.: **Quero me identificar sem restrição**

Nome:

CPF:

Fone:

Município:

Endereço:

Bairro:

Complemento:

Título de Eleitor:

Email:

UF: **9**

Número:

CEP:

Opções

[Enviar a Ouvidoria Geral](#)[Prorrogar](#)[Criar alerta](#)[Imprimir Tudo](#)[Imprimir Manifestação](#)

Dia a dia

[Chat interno](#)

Lista de Manuais

- [Ficha Cadastral Para o Sistema de Gestão De Ouvidoria](#)
- [Exclusivo LAI - Modelo Termo Resposta LAI](#)
- [Config Scanner](#)
- [Digitalização](#)
- [Manual Ouvidoria GO](#)



Lista de usuários



Classificações

Protocolo
fornecido:



Histórico desta manifestação

Dias de Vida: 3

Manifestação
cadastrada para uma
unidade.



Não registrado

13/09/2022

0 dia

Resp. Preliminar
Enviado para Setor
Responsável



Renato

13/09/2022

3 dias

Sob análise do órgão
responsável.

SANEAGO



Renato

13/09/2022

0 dia

Manifestação
finalizada pela
unidade.



Renato

16/09/2022

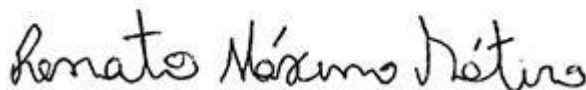
0 dia



Tipo DESPACHO**Número** 49474/2022**Data** 16/09/2022**Processo** 14634/2022**Destinatário** SUPERVISÃO DE GESTÃO ARQUIVÍSTICA - G-SGA - V3156**Assunto** ARQUIVAMENTO - RO 4730/2022

Após finalização do Registro de Ocorrência, encaminhamos o processo para arquivamento.

Atenciosamente,



RENATO MAXIMO SATIRO - ADMINISTRADOR

SUOGE - SUPERINTENDÊNCIA DE OUVIDORIA GERAL em 16/09/2022 15:21:55

APÊNDICE B – Código do tratamento de dados

```
#####  
# Script: Tratamento dos dados ETA  
# Author: Diego Takashi Sato (diegotakashi@saneago.com.br)  
# Date Created: 2024-09-30  
# R version 4.2.3 (2023-03-15 ucrt)  
#####
```

```
#Define o diretorio raiz, onde todos os dados estarao
setwd("/home/diegotakashi/")
```

```
# importar dados csv -----
consolidado <- read.csv("consolidado.csv",
                        header = TRUE, sep = ",", dec = ".")
```

```
#Para ver quantas amostras tem no conjunto de treinamento
nrow(consolidado)
```

```
### Rodar paralelamente
library(doParallel)
cl <- makePSOCKcluster(40)
registerDoParallel(cl, cores=40)
### no final encerrar o cluster
```

```
#####  
library(forecast)  
library(zoo)  
library(Metrics)  
library(dplyr)
```

[illegible]

```
consolidado$dacb_cloro)

#estatística descritiva
summary(consolidado)

#Definicao do intervalo de dados
dados_teste <- subset(consolidado,
                      consolidado$data >= as.POSIXct("2023-04-01_00:00") &
                      consolidado$data <= as.POSIXct("2023-12-31_00:00") )

# calculando 20% das amostras e guardando os indices
n <- nrow(dados_teste)
indice_nas <- sample(1:n, size = round(0.2 * n))

# omissao de 20% das amostras
dados_teste$dacb_cloro_with_nas[indice_nas] <- NA

#interpolacao linear
dados_teste$dacb_cloro_linear <- na.approx(dados_teste$dacb_cloro,
                                           na.rm = FALSE)

#interpolacao spline
dados_teste$dacb_cloro_spline <- na.spline(
    dados_teste$dacb_cloro_with_nas)

#interpolacao STL
dados_teste$dacb_cloro_stl <- na.interp(
    dados_teste$dacb_cloro_with_nas)

write.csv(dados_teste, file = "dados_tratados.csv",
          row.names = FALSE)

#criterios de convergencia
print("SPLINE")
print("MAE_stl:")
mae(dados_teste$dacb_cloro, dados_teste$dacb_cloro_stl)
print("MAE_spline:")
mae(dados_teste$dacb_cloro, dados_teste$dacb_cloro_spline)
print("RMSE_stl:")
```

```
rmse(dados_teste$dacb_cloro, dados_teste$dacb_cloro_stl)
print("RMSE_spline:")
rmse(dados_teste$dacb_cloro, dados_teste$dacb_cloro_spline)
print("Rsquared_stl:")
cor(dados_teste$dacb_cloro, dados_teste$dacb_cloro_stl)^2
print("Rsquared_spline:")
cor(dados_teste$dacb_cloro, dados_teste$dacb_cloro_spline)^2

stopCluster(cl)
```


APÊNDICE C – Código da construção dos modelos de IA

```
#####
# Script: Modelos de IA ETA
# Author: Diego Takashi Sato (diegotakashi@saneago.com.br)
# Date Created: 2023-09-09
# R version 4.2.3 (2023-03-15 ucrt)
#####

#Define o diretorio raiz, onde todos os dados estarao
setwd("/home/diegotakashi/")

# importar dados csv
dados <- read.csv("brazabrantres/dados.csv",
                  header = TRUE, sep = ",", dec = ".")

### Como os modelos sao pesados, e o R nativamente eh single
# thread, podemos melhorar a execucao para rodar paralelamente
library(doParallel)
# aqui definimos quantos nucleos do processados vamos usar
# lembrando que nao pode colocar 100% pois existe o sistema
# operacional e outras atividades do servidor
# sugiro manter pelo menos 2 nucleos para estas tarefas
cl <- makePSOCKcluster(40)
# aqui iniciamos o cluster
registerDoParallel(cl)
### no final encerrar o cluster com
# stopCluster(cl)

summary(dados)

# importar os datasets
# para series temporais, e possivel filtrar os dados com base
# nas datas/intervalos
treinamento <- subset(dados,
```

```

      dados$data >= as.POSIXct("2023-04-01_00:00") &
      dados$data <= as.POSIXct("2023-12-31_23:59"))
teste <- subset(dados,
      dados$data >= as.POSIXct("2023-01-01_00:00") &
      dados$data <= as.POSIXct("2023-03-30_23:59"))

# eliminando dados nao utilizados no treinamento
treinamento$data <- NULL

# ajustando formato da data para uso em graficos e predicoes
teste$data <- as.POSIXct(teste$data, format="%Y-%m-%d_%H:%M")

summary(teste)
summary(treinamento)

#####
## CLORO CLORO CLORO CLORO CLORO CLORO CLORO CLORO CLORO CLORO ##
#####

library(caret)
#-----
# hiperparametros comuns a todos os modelos
# validacao cruzada, k-fold = 10
control <- trainControl(method = "CV", number = 10,
      allowParallel = TRUE)
# definindo numero pseudoaleatorio
set.seed(100)
#-----
# RNA-MLP
# Rede neural com 1 camada utilizando 'weight decay'
# ele reduz o learning rate a cada iteracao para melhorar
# o treinamento e diminuir overfitting

tune_mat = expand.grid(size = c(2,3,4,5,6,7,8,9,10,11,12,13,14,15),
      decay = c(0.1, 0.01, 0.001, 0.0001, 0.00001))
modelo_mlpWeightDecay <- caret::train(dosagem_cloro ~ .,
      data = treinamento,
      method = "mlpWeightDecay",
      preProc = c("center", "scale"),

```

```

                                trControl = control,
                                tuneGrid = tune_mat,
                                metric="RMSE")

print(modelo_mlpWeightDecay)
saveRDS(modelo_mlpWeightDecay,
        "modelo_otimizado_mlpweightdecay_cloro_rmse.rds")
#-----
# RNP-MLP
# Uma rede neural com tres layers
set.seed(100)
tune_mat = expand.grid(layer1 = c(2,3,4,5,6,7,8,9,10,11,12,13,14,15),
                        layer2 = c(2,3,4,5,6,7,8,9,10,11,12,13,14,15),
                        layer3 = c(2,3,4,5,6,7,8,9,10,11,12,13,14,15),
                        decay = c(0.1, 0.01, 0.001, 0.0001, 0.00001))
modelo_mlpWeightDecayML3 <- caret::train(dosagem_cloro ~ .,
                                data = treinamento,
                                method = "mlpWeightDecayML",
                                preProc = c("center", "scale"),
                                trControl = control,
                                tuneGrid = tune_mat,
                                metric="RMSE")

print(modelo_mlpWeightDecayML3)
saveRDS(modelo_mlpWeightDecayML3,
        "modelo_otimizado_mlpweightdecayml3_cloro_rmse.rds")

#-----
# Random Forest
# Cada arvore precisa iniciar e terminar no mesmo processador
# Portanto, colocar muitos processadores para este modelo causa
# travamento e encerramento inesperado do programa
# Por sondagem e ajuste, identifiquei que o valor maximo de
# processadores eh 50% do servidor
cl <- makePSOCKcluster(20)
registerDoParallel(cl)

library(randomForest)
tune_mat <- expand.grid(.mtry = c(1 : 10))
modelo_rf <- caret::train(dosagem_cloro ~ .,
                        data = treinamento,

```

```

        method = "rf",
        trControl = control,
        tuneGrid = tune_mat,
        metric="RMSE")

print(modelo_rf)
saveRDS(modelo_rf, "modelo_otimizado_randomforest_cloro_rmse.rds")

#-----
# XGBoost CLORO
# ajustando novamente o cluster para usar mais processadores
cl <- makePSOCKcluster(40)
registerDoParallel(cl)

library(xgboost)
tune_mat = expand.grid(nrounds = seq(from = 50, to = 300, by = 50),
                      eta = c(0.2, 0.3, 0.4),
                      max_depth = c(2, 3, 4),
                      gamma = c(0, 0.1, 0.5, 0.7, 1.0),
                      colsample_bytree = c(0.6, 0.8, 1.0),
                      min_child_weight = c(1, 2, 3),
                      subsample = c(0.5, 0.75, 1.0))

modelo_xgboost <- caret::train(dosagem_cloro ~ .,
                              data = treinamento,
                              method = "xgbTree",
                              trControl = control,
                              tuneGrid = tune_mat,
                              metric = "RMSE")

modelo_xgboost$bestTune

print(modelo_xgboost)
saveRDS(modelo_xgboost, "modelo_otimizado_xgboost_cloro_rmse.rds")

#Fazer o mesmo para fluor

```

APÊNDICE D – Código de validação dos modelos

```
#####
# Script: Valida modelos ETA
# Author: Diego Takashi Sato (diegotakashi@saneago.com.br)
# Date Created: 2024-01-09
# R version 4.2.3 (2023-03-15 ucrt)
#####

#-----
# Carregando modelos - usar quando ja foram executados os modelos
modelo_mlpWeightDecay <- readRDS(
  "modelo_otimizado_mlpweightdecay_cloro_rmse.rds")
modelo_mlpWeightDecayML3 <- readRDS(
  "modelo_otimizado_mlpweightdecayml3_cloro_rmse.rds")
modelo_rf <- readRDS("modelo_otimizado_randomforest_cloro_rmse.rds")
modelo_xgboost <- readRDS("modelo_otimizado_xgboost_cloro_rmse.rds")
#-----
# Comparando resultados dos modelos
comparacao <- resamples(list(Random_Forest=modelo_rf,
                             XGBoost=modelo_xgboost,
                             Rede_Neural=modelo_mlpWeightDecay,
                             Deep_Learning=modelo_mlpWeightDecayML3))
summary(comparacao)
#-----
# aqui calculamos o valor predito pelos modelos, utilizando o conjunto
# de testes reservado para esta finalidade
library(Metrics)
teste$estimado_mlpWeightDecay <- predict(modelo_mlpWeightDecay,
                                         teste, type="raw")
teste$estimado_mlpWeightDecayML3 <- predict(modelo_mlpWeightDecayML3,
                                           teste, type="raw")
teste$estimado_rf <- predict(modelo_rf, teste, type="raw")
teste$estimado_xgboost <- predict(modelo_xgboost, teste, type="raw")
```

```

# aqui sao plotados os graficos dos valores reais e valores preditos
# pelos modelos
library(ggplot2)
library(reshape2)
library(scales)
melted <- melt(teste, id.vars = "dt_leitura", measure.vars = c(
  "dosagem_cloro", "estimado_mlpWeightDecay",
  "estimado_mlpWeightDecayML3", "estimado_rf",
  "estimado_xgboost"))

# aqui voce define o periodo de amostragem do grafico
lims <- as.POSIXct(strptime(c("2024-04-06 15:00",
  "2024-04-06 16:00"), format = "%Y-%m-%d %H:%M"))

# aqui temos uma infinidade de opcoes para construçao dos graficos
graph <- ggplot(melted, aes(x = dt_leitura, y = value,
  colour = variable)) +
  geom_line() +
  scale_x_datetime(limits = lims, breaks = date_breaks("5 day"),
  labels = date_format("%d/%m/%Y")) +
  scale_y_continuous(n.breaks = 9, limits = c(0, 2.25)) +
  labels=c("XGBoost", "RF", "RNA_MLP", "Sistema_real"))+
  scale_color_manual( values=c("blue", "orange", "green", "red")) +
  xlab("Datas") +
  ylab("Dosagem")

#aqui plota o grafico na memoria
graph
#aqui salva o grafico plotado
ggsave("20241021-valida_cloracao.png")

#-----
# Critérios de avaliação do melhor modelo
print("RMSE_mlp, mlp-ml3, rf, xgboost")
rmse(teste$dosagem_cloro, teste$estimado_mlpWeightDecay)
rmse(teste$dosagem_cloro, teste$estimado_mlpWeightDecayML3)
rmse(teste$dosagem_cloro, teste$estimado_rf)
rmse(teste$dosagem_cloro, teste$estimado_xgboost)
print("Rsquared_mlp, mlp-ml3, rf, xgboost")

```

```
cor(teste$dosagem_cloro, teste$estimado_mlpWeightDecay)^2
cor(teste$dosagem_cloro, teste$estimado_mlpWeightDecayML3)^2
cor(teste$dosagem_cloro, teste$estimado_rf)^2
cor(teste$dosagem_cloro, teste$estimado_xgboost)^2
print("MAE_ _mlp, _mlp-ml3, _rf, _xgboost")
mae(teste$dosagem_cloro, teste$estimado_mlpWeightDecay)
mae(teste$dosagem_cloro, teste$estimado_mlpWeightDecayML3)
mae(teste$dosagem_cloro, teste$estimado_rf)
mae(teste$dosagem_cloro, teste$estimado_xgboost)
```


APÊNDICE E – Código para análise estatística dos modelos

```
#####
# Script: Modelos de IA ETA
# Author: Diego Takashi Sato (diegotakashi@saneago.com.br)
# Date Created: 2025-01-14
# R version 4.2.3 (2023-03-15 ucrt)
#####

library(data.table)
library(dplyr)
library(ggplot2)
library(stringr)

# Funcao para carregar e corrigir virgulas nos numeros
carregar_corrigir_arquivo <- function(caminho) {
  df <- fread(caminho)
  df <- df %>% mutate(across(everything(), ~ as.numeric(
    gsub(",", ".", .))))
  cat("Linhas em", basename(caminho), ":", nrow(df), "\n")
  return(df)
}

# Funcoes de metricas de avaliacao
calcular_rmse <- function(erros) {
  sqrt(mean(erros^2, na.rm = TRUE))
}

calcular_r2 <- function(y_real, y_prev) {
  ss_res <- sum((y_real - y_prev)^2, na.rm = TRUE)
  ss_tot <- sum((y_real - mean(y_real, na.rm = TRUE))^2,
    na.rm = TRUE)
  1 - (ss_res / ss_tot)
}
```

```
calcular_mae <- function(erros) {  
  mean(abs(erros), na.rm = TRUE)  
}  
  
# Funcao para realizar os testes de hipotese e gerar graficos  
realizar_teste_hipotese <- function(df, tipo, diretorio_saida,  
                                     tamanho_subamostra = 1000){  
  col_real <- paste0(tipo, "_sist")  
  colunas_previstas <- names(df)[str_detect(names(df), "mlp|rf|xgb")]  
  resultados <- list()  
  
  # Criar pasta para salvar graficos  
  dir_imagens <- file.path(diretorio_saida, paste0("graficos_", tipo))  
  if (!dir.exists(dir_imagens)) dir.create(dir_imagens)  
  
  for (col_prev in colunas_previstas) {  
    y_real <- df[[col_real]]  
    y_prev <- df[[col_prev]]  
    erros <- y_real - y_prev  
  
    if (sum(!is.na(erros)) > tamanho_subamostra) {  
      idx <- sample(which(!is.na(erros)), tamanho_subamostra,  
                    replace = FALSE)  
      erros <- erros[idx]  
      y_real <- y_real[idx]  
      y_prev <- y_prev[idx]  
    }  
  
    # Calculo das metricas  
    rmse <- calcular_rmse(erros)  
    r2 <- calcular_r2(y_real, y_prev)  
    mae <- calcular_mae(erros)  
  
    # Teste de Mann-Whitney U  
    p_valor <- wilcox.test(erros, mu = 0)$p.value  
  
    resultados[[col_prev]] <- data.frame(  
      Tipo = toupper(tipo),  
      Modelo = col_prev,
```

```

    RMSE = rmse,
    R2 = r2,
    MAE = mae,
    'p-valor (teste)' = p_valor
  )

  # Grafico de histograma de residuos
  g <- ggplot(data = data.frame(Erro = erros), aes(x = Erro)) +
    geom_histogram(bins = 30, fill = "skyblue", alpha = 0.7) +
    geom_vline(xintercept = 0, color = "red", linetype = "dashed") +
    labs(title = paste("Histograma_de_Residuos_", toupper(tipo),
      "(", col_prev, ")"),
      x = "Erro_(Y_real_-Y_previsto)", y = "Frequencia") +
    theme_minimal()
  ggsave(file.path(dir_imagens, paste0(col_prev, "_histograma.png")),
    plot = g)

  # Grafico de dispersao dos residuos
  g <- ggplot(data = data.frame(Observacao = seq_along(erros),
    Erro = erros),
    aes(x = Observacao, y = Erro)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0,
    color = "red", linetype = "dashed") +
    labs(title = paste("Residuos_ao_longo_das_observacoes_",
      toupper(tipo), "(", col_prev, ")"),
      x = "Observacoes", y = "Erro") +
    theme_minimal()
  ggsave(file.path(dir_imagens, paste0(col_prev, "_dispersao.png")),
    plot = g)
}

return(do.call(rbind, resultados))
}

# Caminho base
caminho_base <- "/home/diegotakashi"
df_cloro <- carregar_corrigir_arquivo(file.path(caminho_base,
  "cloro.csv"))

```

```
df_fluor <- carregar_corrigir_arquivo(file.path(caminho_base,
                                                "fluor.csv"))

# Adicionar coluna de sequencia de coleta
df_cloro <- df_cloro %>% mutate(coleta_seq = seq_len(nrow(df_cloro)))
df_fluor <- df_fluor %>% mutate(coleta_seq = seq_len(nrow(df_fluor)))

# Diretorio de saida para os graficos
diretorio_saida <- caminho_base

# Realizar os testes e salvar graficos
resultados_cloro <- realizar_teste_hipotese(df_cloro, "cl",
                                             diretorio_saida)
resultados_fluor <- realizar_teste_hipotese(df_fluor, "fl",
                                             diretorio_saida)

# Consolidar os resultados e salvar em CSV
resultados_completos <- bind_rows(resultados_cloro, resultados_fluor)
caminho_resultado <- file.path(caminho_base, "teste_hipotese.csv")
fwrite(resultados_completos, caminho_resultado)
```

Índice

coeficiente de determinação, 41, 43–46, 49,
50, 55

erro médio absoluto, 41, 43–46, 49, 50, 55
estação de tratamento de água, 21, 22,
33–37, 40, 49, 51, 55

floresta randômica, 22, 29, 30, 41–43, 45,
46, 49, 50, 53, 55

inteligência artificial, 1, 3, 7, 21, 28, 33,
40, 44, 55

modelagem, 1, 3, 7, 21, 28, 29, 55

modelos, 7, 21, 22, 28–30, 39–44, 46, 48–
53, 55

otimização, 7, 21, 22, 29, 30, 33, 42–48,
52, 53

raiz do erro médio quadrático, 41, 43–46,
48–50, 55

RNA-MLP, 29, 41–43, 45–47, 49, 51–53,
55

Saneago, 5, 22, 34, 35

XGBoost, 29, 30, 42–46, 49–52, 55