UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Larissa de Lima e Silva

USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E CLASSIFICAÇÃO DO ÍNDICE DA QUALIDADE DA ÁGUA: UM CASO DE ESTUDO NO RIO PARAIBUNA, JUIZ DE FORA, BRASIL

Juiz de Fora
2025

Larissa de Lima e Silva

USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E CLASSIFICAÇÃO DO ÍNDICE DA QUALIDADE DA ÁGUA: UM CASO DE ESTUDO NO RIO PARAIBUNA, JUIZ DE FORA, BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional

Orientadora: Doutora Priscila Vanessa Zabala Capriles Goliatt

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Silva, Larissa de Lima e.
USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E CLASSIFICAÇÃO DO ÍNDICE DA QUALIDADE DA ÁGUA: UM CASO DE ESTUDO NO RIO PARAIBUNA, JUIZ DE FORA, BRASIL / Larissa de Lima e Silva. – 2025.
151 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Modelagem Computacional, 2025.

1. Índice Qualidade da Água. 2. Inteligência Artificial. 3. Modelagem Computacional. 4. Aprendizagem de Máquina. I. Goliatt, Priscila, orient. II. Doutora III. Barros, Nathan, coorient. IV. Doutor.

Larissa de Lima e Silva

USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E CLASSIFICAÇÃO DO ÍNDICE DA QUALIDADE DA ÁGUA: UM CASO DE ESTUDO NO RIO PARAIBUNA, JUIZ DE FORA, BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 09 de setembro de 2025.

BANCA EXAMINADORA

Prof.ª Dr.ª. Priscila Vanessa Zabala Capriles Goliatt - Orientadora
Universidade Federal de Juiz de Fora

Prof. Dr. Nathan Oliveira Barros
Universidade Federal de Juiz de Fora

Prof. Dr. Eduardo Krempser da Silva
Fundação Oswaldo Cruz

Juiz de Fora, 27/08/2025.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 29/09/2025, às 19:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Nathan Oliveira Barros**, **Professor(a)**, em 30/09/2025, às 11:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Krempser da Silva**, **Usuário Externo**, em 30/09/2025, às 16:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2580776** e o código CRC **8A15A134**.

Em memória à minha amada mãe, Janice de Lima e Silva, que acreditou em mim desde o momento em que nasci.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pois como está escrito em Isaías 25:1: "Senhor, tu és o meu Deus; eu te exaltarei e louvarei o teu nome, pois com grande perfeição tens feito maravilhas, coisas há muito planejadas com fidelidade e firmeza."

À Nossa Senhora de Fátima, por estar comigo nos momentos em que precisei de colo de mãe e por me acolher em silêncio quando o coração se fazia frágil.

Agradeço também a mim mesma, por ter sido resiliente diante da pressão e dos eventos externos que atravessaram meu caminho. Pela força de vontade, pela coragem e pela perseverança em não desistir dos meus sonhos.

Aos meus pais, por me proporcionarem as condições para sonhar em ser estudante e pesquisadora, e por sempre me incentivarem a buscar mais. À minha mãe, que dedicou sua vida para garantir que eu tivesse qualidade de estudo, e ao meu pai, por todos os sacrifícios feitos em nome da minha educação.

À minha irmã, Maria Eduarda, pelo colo, pela amizade, pelo companheirismo e pelo amor que sempre me sustentaram. Às minhas tias, especialmente Tia Ju e Tia Janete, que foram como segundas mães, sempre presentes com carinho, incentivo e apoio em todas as etapas da minha vida.

Ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, pela oportunidade, pelo acolhimento e pelo espaço para o desenvolvimento desta pesquisa. Aos professores e funcionários, em especial aos professores Bernardo Martins, Bárbara Quintela e Flávia Bastos, e à Maíra Macário, por todo o suporte essencial ao longo do curso. Agradeço também aos colegas e amigos do PPGMC, Thiago Esterci, Ana Clara, Fabrício e Lara, pela ajuda nas disciplinas e pela compreensão durante minha ausência nos primeiros meses de 2023, período em que enfrentei o luto pela perda de minha mãe.

Ao Prof. Leonardo Goliatt, que me incentivou e encorajou a ingressar no mestrado. Foi ele quem, com suas palavras firmes, desviou minha rota e me fez acreditar que este caminho era possível e necessário na minha vida acadêmica.

Aos amigos que sempre estiveram ao meu lado e me amaram mesmo nos momentos em que eu não merecia. À Luciana, minha companheira desde o início da faculdade em 2018, por compreender meus surtos e me apoiar nos momentos difíceis. Ao Thiago Luiz, meu melhor amigo, que me conhece melhor do que eu mesma e foi essencial nessa trajetória.

Ao meu gatinho Toddy, pela parceria diária, por me mostrar a hora de descansar deitando sobre o computador, e por trazer alegria e leveza ao ambiente de trabalho.

À minha orientadora, Prof. Priscila Capriles, pelos ensinamentos que ultrapassaram

a esfera acadêmica, contribuindo também para minha formação pessoal. Agradeço pelas cobranças e exigências que me impulsionaram a buscar sempre o melhor de mim, mesmo quando os desafios pareciam maiores do que eu podia suportar.

Ao Prof. Nathan Barros, pela generosidade em compartilhar dados do rio Paraibuna, sem os quais este trabalho não teria sido possível, e pela colaboração em momentos cruciais para a pesquisa.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização desta dissertação.

O que sabemos é uma gota, o que ignoramos é um oceano. (Isaac Newton)

RESUMO

Este estudo investiga o uso de algoritmos de aprendizagem de máquina para a previsão e classificação do índice de qualidade da água (IQA) no Rio Paraibuna, em Juiz de Fora, Brasil. A problemática reside na necessidade de métodos mais rápidos e precisos para monitorar a qualidade da água em tempo real, superando limitações dos métodos tradicionais que demandam tempo e recursos. Justifica-se pela importância de fornecer dados confiáveis e atualizados para a gestão ambiental e a proteção da saúde pública, oferecendo uma alternativa eficiente para a tomada de decisões e políticas de preservação ambiental. O objetivo geral é desenvolver e aplicar um modelo de aprendizagem de máquina que preveja e classifique o IQA, utilizando dados históricos, enquanto os objetivos específicos incluem avaliar a qualidade dos dados, selecionar algoritmos adequados, validar o modelo e viabilizar sua replicação em outros rios da região. As hipóteses levantadas sugerem que algoritmos de aprendizagem de máquina serão eficazes para prever o IQA com precisão, oferecendo uma resposta mais ágil que métodos convencionais e permitindo intervenções ambientais mais eficazes.

Palavras-chave: Qualidade da água, algoritmos de aprendizagem de máquina, inteligência artificial, regressão, classificação, impacto ambiental, recursos hídricos

ABSTRACT

This study investigates the use of machine learning algorithms for the prediction and classification of the water quality index (WQI) in the Paraibuna River, in Juiz de Fora, Brazil. The problem lies in the need for faster and more accurate methods to monitor water quality in real time, overcoming limitations of traditional methods that require time and resources. It is justified by the importance of providing reliable and up-to-date data for environmental management and the protection of public health, offering an efficient alternative for decision-making and environmental preservation policies. The overall objective is to develop and apply a machine learning model that predicts and classifies the WQI, using historical and real-time data, while the specific objectives include evaluating the quality of the data, selecting appropriate algorithms, validating the model, and enabling its replication in other rivers in the region. The hypotheses raised suggest that machine learning algorithms will be effective in predicting WQI accurately, offering a more agile response than conventional methods and allowing for more effective environmental interventions.

Keywords: Water quality, machine learning algorithms, artificial intelligence, regression, classification, environmental impact, water resources.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Rede Neuronal Artificial (*Artificial Neural Network*)
API	Ídicie de Gravidade Específica do Petróleo (*American Petroleum Institute Gravity*)
BPNN	Rede Neuronal de Retropropagação (*Backpropagation Neural Network*)
ELM	*Extreme Learning Machine*
FNN	Rede Neuronal Funcional (*Functional Neural Network*)
FBHP	Pressão de Fundo de Poço (*Flowing Bottom-hole Pressure*)
GFR	Taxa de Fluxo de Gás (*Gas Flow Rate*)
GMDH	Método de Agrupamento de Dados (*Group Method of Data Handling*)
GRNN	Rede Neuronal de Regressão Geral (*General Regression Neural Network*)
IC	Inteligência Computacional
ID	Diâmetro Interno do Tubo (*Internal Diameter of Pipe*)
KNN	k-Vizinhos mais Próximos (*k-Nearest Neighbors*)
LSSVM	Máquina de Vetores de Suporte por Mínimos Quadrados (*Least Squares Support Vector Machine*)
LSTM	Modelo de Memória de Longo Prazo (*Long Short-Term Memory*)
MARS	*Multivariate Adaptive Regression Splines*
MAE	Erro Médio Absoluto (*Mean Absolute Error*)
MAPE	Erro Percentual Médio Absoluto (*Mean Absolute Percentage Error*)
ML	Aprendizagem de Máquina (*Machine Learning*)
MSE	Erro Quadrático Médio (*Mean Squared Error*)
OFR	Taxa de Fluco de Óleo (*Oil Flow Rate*)
PGS	Programação Genética Simbólica
PSO	Otimização por Enxame de Partículas (*Particle Swarm Optimization*)
PTC	*Probabilistic Tree-Creation*
R	Coeficiente de correlação
R^2	Coeficiente de determinação
RF	Floresta Aleatória (*Random Forest*)
RMSE	Erro Quadrático Médio Residual (*Root Mean Squared Error*)
RBFNN	Rede Neuronal de Função de Base Radial (*Radial Basis Function Neural Network*)
SVR	*Support Vector Regression*
TOC	Teor de Carbono Orgânico (*Total Organic Carbon*)
XGB	*Extreme Gradient Boosting*
WBHT	Temperatura na cabeça do poço (*Wellbore Head Temperature*)
WFR	Taxa de Fluxo de Água (*Water Flow Rate*)
WHP	Pressão na Cabeça do Poço (*Wellhead Pressure*)
WPD	Produção Diária de Água *Water Production Rate*

LISTA DE SÍMBOLOS

$\cup$	União
$\in$	Pertence
$\sum$	Somatório
F	Conjunto de funções
f	Função
$\mathcal{T}$	Conjunto de terminais
τ	Terminal
q_τ	Probabilidade de escolher um terminal $t \in T$
q_f	Probabilidade de escolher uma função $f \in F$
ρ	Probabilidade de escolher um não-terminal
E_{tree}	Tamanho esperado de uma arvore
b_n	Aridade não-terminal
S	Tamanho máximo de uma arvore
w_s	Probabilidade associada a cada arvore com s variando de 1 ate S
i	Individuo
t	Geração
$a(i,t)$	Aptidão ajustada do individuo i na geração t
$s(i,t)$	Aptidão padronizada do indivíduo i na geração t
$n(i,t)$	Aptidão normalizada do indivíduo i na geração t
$\hat{y}_i$	Saídas gera pelo programa
y_i	Saídas corretas desejadas
P	População
p	Programa associado a uma população P
f_p	Aptidão associada a um programa p

SUMÁRIO

1 INTRODUÇÃO

O monitoramento da qualidade da água é um fator crucial para a manutenção da saúde pública e da sustentabilidade ambiental, especialmente em áreas urbanas onde a demanda por água potável é elevada e os riscos de contaminação aumentam devido à poluição industrial, agrícola e doméstica. [Ministério da Saúde, 2025]

No Brasil, os desafios relacionados ao monitoramento da qualidade da água são particularmente críticos devido à ampla diversidade de fontes de poluição e à variabilidade climática que impactam diretamente os rios e corpos hídricos. Um dos aspectos mais alarmantes é a inadequação no tratamento de esgoto. De acordo com a Agência Nacional de Águas [Agência Nacional de Águas, 2013], cerca de 70% do esgoto gerado no país é despejado diretamente no meio ambiente sem qualquer tratamento adequado.

Essa precariedade no saneamento básico gera consequências graves para a saúde pública. O Brasil registrou mais de 344 mil internações por doenças relacionadas ao saneamento ambiental inadequado em 2024 [Agência Brasil, 2025]. Neste ano, foram registrados 11.544 óbitos por doenças relacionadas ao saneamento ambiental. Esses números evidenciam a necessidade urgente de políticas públicas voltadas à gestão sustentável da água e à ampliação do acesso a sistemas de saneamento básico.

A combinação de fatores como a insuficiência de infraestrutura de saneamento, o crescimento populacional desordenado e as mudanças climáticas intensificam a degradação dos recursos hídricos, tornando o monitoramento contínuo e eficiente da qualidade da água essencial para mitigar impactos ambientais e proteger a saúde da população.

Neste sentido, o Rio Paraibuna, situado em Juiz de Fora, Minas Gerais, é um recurso hídrico essencial para a área. Aproximadamente metade da água utilizada na cidade é oriunda desse rio, principalmente através da Represa de Chapéu d'Uvas, que desempenha um papel vital na gestão do fluxo e no fornecimento de água. Isso significa que cerca de 270 mil habitantes, em uma população total de 540.756, conforme o Censo de 2022 do IBGE, têm seu abastecimento garantido. Além disso, o rio é de grande importância para a indústria local, com aproximadamente 1.000 fábricas localizadas em sua bacia, sendo que 83% delas têm potencial para causar poluição, o que ressalta a urgência de um monitoramento rigoroso da qualidade da água [Ministério do Meio Ambiente *et al.*, 1999] Na área de energia, o Paraibuna responde por cerca de 12,5% do consumo de eletricidade em Juiz de Fora, através de usinas que estão instaladas ao longo do seu curso. [Instituto Brasileiro de Geografia e Estatística - IBGE, 2022], [da Silva, 2021].

A problemática deste estudo reside na crescente necessidade de métodos eficientes e precisos para monitorar a qualidade da água em tempo real. Métodos tradicionais de análise são frequentemente limitados pela disponibilidade de recursos e pela demora nos resultados, comprometendo a rápida tomada de decisões em situações de risco [Uddin

et al., 2021b].

Por outro lado, os testes rápidos, como as fitas indicadoras de pH, cloro, dureza e alcalinidade, são frequentemente utilizados em ambientes como piscinas, aquários e fontes de água, proporcionando conveniência e resultados imediatos. Contudo, esses métodos carecem da precisão e da abrangência necessárias para assegurar que a água é própria para consumo humano, já que não identificam contaminantes microbiológicos ou substâncias químicas mais complexas. Segundo a Portaria GM/MS nº 888/2021, toda água destinada ao consumo humano deve cumprir um padrão de potabilidade estabelecido por parâmetros específicos e ser sujeita a controle e supervisão adequados.

O uso de algoritmos de aprendizagem de máquina representa uma alternativa metodológica para superar essas limitações, fornecendo previsões baseadas em dados históricos e em informações coletadas provenientes de sensores automatizados ou sistemas de monitoramento contínuo. Ainda assim, sua aplicação no monitoramento ambiental de rios brasileiros é recente, destacando a importância de pesquisas que avaliem a aplicabilidade e precisão desses métodos em contextos locais. [Uddin *et al.*, 2023b], [Azrour *et al.*, 2022]

A justificativa para este estudo baseia-se na importância de garantir o acesso a dados precisos e atualizados sobre a qualidade da água, fator essencial para a proteção da saúde pública e para a gestão sustentável dos recursos hídricos. Com a aplicação de algoritmos de aprendizagem de máquina, espera-se aprimorar a precisão e a rapidez das análises de qualidade da água, possibilitando o desenvolvimento de sistemas preditivos que suportem decisões ambientais e políticas públicas mais eficazes. Este estudo contribuirá com dados relevantes para a comunidade científica e gestores ambientais, ao oferecer um modelo que pode ser replicado e ajustado para outras bacias hidrográficas.

1.1 Justificativa e Abordagem Metodológica

O desenvolvimento de um modelo preditivo robusto e eficaz requer uma série de etapas, incluindo a limpeza e preparação dos dados, a escolha de variáveis preditivas, a seleção do algoritmo mais adequado e a validação do modelo. Para garantir a robustez e a aplicabilidade dos resultados, serão testados diferentes algoritmos de aprendizagem supervisionada, como árvores de decisão, redes neuronais e máquinas de vetor de suporte, cada um com características específicas que poderão oferecer um entendimento melhor sobre a dinâmica da qualidade da água.

A escolha do melhor algoritmo será baseada na precisão das previsões e na capacidade de processamento dos dados em tempo hábil. Adicionalmente, este estudo enfatiza a necessidade de uma abordagem interdisciplinar, combinando conhecimentos de ciência ambiental e ciência de dados. O uso de aprendizagem de máquina no contexto da qualidade da água exige uma compreensão aprofundada dos indicadores ambientais, bem como dos fatores que influenciam a variação desses indicadores Zhu *et al.* [2022] Ao desenvolver um

modelo preditivo, é essencial considerar não apenas as variáveis de poluição direta, mas também as mudanças sazonais e o impacto das variações climáticas na qualidade da água.

O monitoramento da qualidade da água é um pilar essencial para a preservação ambiental e da saúde pública, especialmente em regiões urbanizadas onde a demanda por água potável é elevada e os riscos de contaminação são intensificados pela poluição industrial, agrícola e doméstica [Ministério da Saúde, 2025]. No Brasil, esse desafio é ainda mais acentuado devido à grande diversidade de fontes de poluição e à variabilidade climática que afeta os corpos hídricos. [Medeiros *et al.*, 2017] O Rio Paraibuna, em Juiz de Fora, Minas Gerais, é um exemplo emblemático dessa complexidade. Este rio desempenha um papel crucial no abastecimento hídrico da região, ao mesmo tempo em que enfrenta mudanças ambientais que exigem estratégias de monitoramento contínuo e eficaz.

1.2 Métodos Convencionais e Abordagens Baseadas em aprendizagem de Máquina

Os métodos tradicionais de análise da qualidade da água, frequentemente baseados em protocolos rígidos e procedimentos manuais, desempenham um papel fundamental na obtenção de dados confiáveis e padronizados. No entanto, essas abordagens apresentam limitações quanto à velocidade, custo e capacidade de processamento de grandes volumes de dados, dificultando a identificação ágil de situações de risco e a implementação imediata de medidas corretivas.

Nesse contexto, as técnicas baseadas em aprendizagem de máquina não substituem as práticas convencionais de coleta e análise, mas as complementam e potencializam, ao permitir uma interpretação mais abrangente e preditiva dos dados obtidos. Essas ferramentas possibilitam a integração de múltiplas variáveis, o reconhecimento de padrões complexos e a geração de previsões com base em séries temporais, ampliando a capacidade de resposta a eventos críticos e mudanças ambientais.

O método tradicional da Fundação Nacional de Saneamento (NSF), criado nos Estados Unidos na década de 1970, foi um marco para padronizar a avaliação da qualidade da água, oferecendo considerações claras e confiáveis. Sua longevidade pode ser atribuída à eficácia inicial em garantir uniformidade nos resultados, sendo amplamente utilizado como referência global. No entanto, a dificuldade desse modelo, desenvolvido para atender a uma realidade industrial e tecnológica da época, tornou-se uma limitação em tempos de avanços rápidos em ciência e tecnologia [Uddin *et al.*, 2021a].

A principal crítica a esse método reside na sua incapacidade de incorporar dados em tempo real e variáveis dinâmicas que afetam a qualidade da água, como mudanças climáticas e sazonais. Além disso, sua aplicação em contextos diversos, como o brasileiro, apresenta desafios adicionais devido à complexidade das bacias hidrográficas locais e à heterogeneidade das fontes de poluição [Kumar *et al.*, 2024], [Uddin *et al.*, 2023a], [Lee *et al.*, 2024]

Este estudo é justificado pela necessidade de métodos mais eficazes e modernos para monitorar a qualidade da água, especialmente em rios estratégicos como o Paraibuna. A utilização de aprendizagem de máquina permite superar as limitações dos métodos tradicionais, fornecendo uma abordagem mais robusta e acessível para gestores ambientais.

O presente trabalho busca não apenas demonstrar a eficácia dos algoritmos de aprendizagem de máquina na previsão do índice da qualidade da água do Rio Paraibuna, mas também contribuir para a ciência ambiental por meio do desenvolvimento de um modelo preditivo replicável em outras regiões do Brasil. Essa abordagem está alinhada às diretrizes da Política Nacional de Vigilância da Qualidade da Água para Consumo Humano [Ministério da Saúde, 2025], coordenada pelo Ministério da Saúde, e, em âmbito estadual, ao Instituto Mineiro de Gestão das Águas [Instituto Mineiro de Gestão das Águas, 2024], que realiza avaliações periódicas da qualidade das águas superficiais, fornecendo dados estratégicos para subsidiar políticas de gestão ambiental em Minas Gerais. Assim, o modelo, ao ser replicado, tem potencial para fortalecer sistemas de vigilância da água em diversas regiões, apoiando decisões mais informadas e promovendo a gestão sustentável dos recursos hídricos.

1.3 Visão Geral da Dissertação

Esta dissertação está estruturada de forma a proporcionar uma compreensão gradual e aprofundada sobre o uso de algoritmos de aprendizagem de máquina na previsão e classificação do Índice de Qualidade da Água (IQA) do Rio Paraibuna. O Capítulo 1 apresenta a introdução do tema, abordando a problemática, os objetivos e a justificativa para a realização deste estudo.

No Capítulo 1.4, são detalhados os objetivos geral e específicos do trabalho, assim como os modelos de aprendizagem de máquina empregados e os parâmetros ambientais utilizados. Em seguida, o Capítulo 2 traz uma revisão bibliográfica sobre os principais trabalhos e conceitos relacionados à qualidade da água e às técnicas de aprendizagem de máquina aplicadas a contextos ambientais.

O Capítulo 3 descreve a metodologia adotada, incluindo o pré-processamento dos dados, as técnicas de inferência de valores faltantes, os critérios de seleção e validação dos algoritmos, bem como a abordagem de otimização multiobjetivo.

O Capítulo 4 apresenta os resultados obtidos com os diferentes modelos aplicados, tanto para regressão quanto para classificação, com ênfase na comparação entre os métodos e na análise da fronteira de Pareto. Finalmente, o Capítulo 5 discute as conclusões do estudo e propõe direções para trabalhos futuros, como detalhado no Capítulo 6.

1.4 Objetivo

Desenvolver e aplicar um modelo de aprendizagem de máquina para prever e classificar o índice de qualidade da água do Rio Paraibuna, em Juiz de Fora.

1.4.1 Objetivos Específicos

Entre os objetivos específicos, estão:

1. Avaliar a qualidade dos dados disponíveis sobre o Rio Paraibuna para uso em modelos de aprendizagem de máquina;
2. Selecionar e testar diferentes algoritmos de aprendizagem de máquina para identificar o que melhor se adapta aos dados e apresenta maior precisão;
3. Comparar os resultados obtidos pelo modelo preditivo com dados reais para validar a eficácia do modelo; e
4. Disponibilizar o modelo para uso futuro em análises de qualidade da água em outros rios da região.

A intenção é que este modelo seja capaz de processar dados históricos, permitindo uma análise preditiva que seja útil para gestores ambientais e políticas públicas. Esses objetivos visam garantir que o modelo proposto não apenas funcione de forma eficaz no contexto específico do Rio Paraibuna, mas também possa ser adaptado para outros estudos de qualidade de água.

1.5 Hipóteses

As hipóteses formuladas para este estudo são:

1. Algoritmos de aprendizagem de máquina podem prever com precisão o índice de qualidade da água do Rio Paraibuna a partir de dados históricos e atuais, auxiliando na detecção precoce de variações nos níveis de poluição;
2. Modelos preditivos baseados em aprendizagem de máquina apresentarão maior precisão e menor tempo de resposta em relação aos métodos convencionais de monitoramento de qualidade de água;
3. A aplicação do modelo preditivo permitirá intervenções mais eficazes e direcionadas para a gestão da qualidade da água, reduzindo os impactos ambientais e protegendo a saúde pública.

2 REVISÃO BIBLIOGRÁFICA

2.1 Revisão da Literatura

A revisão da literatura recente evidencia o crescente interesse no uso de algoritmos de aprendizagem de máquina para análise e predição em contextos ambientais. Com a capacidade de lidar com grandes volumes de dados e fornecer previsões de alta precisão, esses métodos têm se mostrado promissores em estudos relacionados à qualidade da água e à gestão de recursos hídricos. Diversas abordagens foram exploradas, desde algoritmos clássicos, como regressão linear e redes neuronais artificiais, até técnicas avançadas baseadas em ensembles, como o XGBoost e o Random Forest. Esta seção apresenta uma análise crítica dos principais estudos sobre o tema, destacando os métodos, os resultados obtidos e as implicações para a gestão ambiental.

Estudos recentes têm explorado o uso de algoritmos de aprendizagem de máquina para melhorar a análise e predição de dados ambientais. Por exemplo, no estudo conduzido por Kouadri (2021) [Kouadri *et al.*, 2021], foi avaliada a qualidade das águas subterrâneas na região de Illizi, sudeste da Argélia, utilizando oito algoritmos de inteligência artificial, incluindo regressão multilinear (MLR), floresta aleatória (RF), árvore M5P (M5P), subespaço aleatório (RSS), regressão aditiva (AR), rede neuronal artificial (RNA, também conhecida como ANN, Artificial Neuronal Network), vetor de suporte regressão (SVR) e regressão linear ponderada localmente (LWLR). Entre esses, os algoritmos MLR e RF apresentaram os melhores resultados, destacando-se por suas métricas de desempenho robustas.

De forma semelhante, Mourade Azrour et al. (2021) [Azrour *et al.*, 2022] realizaram um estudo utilizando uma entrada composta por quatro parâmetros principais: temperatura da água, pH, turbidez e coliformes. Nesse contexto, foram avaliados algoritmos como MLR, Gradient Boosting e Regressão Lasso. Além disso, foram aplicados métodos de classificação, como SVM, Decision Tree e ANN, sendo o ANN o que apresentou melhor desempenho. Outros estudos também corroboram a eficiência do ANN, como observado em referências adicionais ([Sakizadeh, 2016], [Othman *et al.*, 2020]).

Contudo, é importante destacar que alguns trabalhos apontam para a superioridade do SVM em relação às redes neuronais artificiais em determinados contextos. Em um estudo [Haghiabi *et al.*, 2018], foi realizada uma comparação direta entre RNA e SVM para a predição do IQA, revelando que o SVM foi o modelo mais preciso com base nos índices de erro. Outro exemplo [Mohammadpour *et al.*, 2015] também demonstrou a superioridade do SVM em relação a métodos como retropropagação de alimentação direta (FFBP) e função de base radial (RBF).

Além disso, algoritmos como Extra Tree Regression (ETR) têm-se mostrado promissores. Em uma análise específica [Asadollah *et al.*, 2021] , o ETR superou os algoritmos

Support Vector Regression (SVR) e Decision Tree Regression (DTR), evidenciando que não existe um modelo de IA universalmente superior. A escolha do algoritmo ideal depende das características do conjunto de dados e dos ajustes realizados nos parâmetros dos modelos.

Por fim, existem os métodos que combinam múltiplos modelos para melhorar a precisão e a robustez das previsões. Essa abordagem permite que erros individuais de modelos simples se compensem, resultando em um desempenho superior no conjunto. O *XGBoost*, por exemplo, apresentou resultados excelentes em diferentes estudos [Uddin *et al.*, 2022], [Uddin *et al.*, 2023c], e em [Lu & Ma, 2020]. Da mesma forma, o Random Forest (RF) continua sendo amplamente utilizado, com resultados positivos observados em diversas análises [Uddin *et al.*, 2022], [Lu & Ma, 2020]; e em [Lap *et al.*, 2023]

Esses avanços destacam o papel fundamental das tecnologias de aprendizagem de máquina na modernização do monitoramento ambiental, permitindo análises mais precisas e fundamentadas para a gestão sustentável dos recursos hídricos.

A próxima seção apresenta os principais modelos de aprendizagem de máquina utilizados nesta pesquisa, detalhando seus princípios, aplicações e justificativas para a escolha no contexto da previsão da qualidade da água do Rio Paraibuna.

2.2 Modelos de aprendizagem de Máquina (ML)

Os modelos de Aprendizado de Máquina (do inglês, Machine Learning, ML) foram utilizados para avaliar a qualidade da água do Rio Paraibuna por meio de variáveis físico-químicas, biológicas e ambientais, como temperatura da água, pH, turbidez e concentração de coliformes, com o objetivo de identificar padrões e influências dessas variáveis na poluição do rio. Para tal, diversos algoritmos de aprendizagem supervisionados foram explorados, passando por uma análise abrangente dos dados e fornecendo modelos preditivos capazes de detectar alterações nos períodos da água em tempo real ou em períodos futuros. A escolha do modelo depende da natureza dos dados e dos objetivos específicos da pesquisa.

Inicialmente, a regressão linear foi escolhida por sua simplicidade e capacidade de fornecer uma visão clara da relação linear entre as variáveis. Já as Máquinas de Vetores de Suporte foram empregadas para tratar dados mais complexos e não lineares, proporcionando um desempenho robusto ao lidar com variações e interações mais complexas nos dados. Esses modelos permitiram entender como os parâmetros ambientais influenciam a qualidade da água ao longo do rio e como diferentes variáveis interagem entre si.

Além disso, modelos de Árvores de Decisão e Random Forest foram aplicados para classificação, como a determinação da qualidade da água em diferentes classes (boa, moderada, ruim). Estes modelos são de identificação de padrões de decisão baseados em características específicas dos dados, como pH, turbidez, temperatura da água, entre outros. As Árvores de Decisão permitem entender as regras de decisão que dividem os dados em categorias, enquanto o Random Forest, uma técnica baseada em múltiplas árvores de decisão, foi utilizada para reduzir o risco de $overfitting$, fornecendo uma abordagem mais robusta para a previsão da qualidade da água.

2.2.1 Regressão Linear

A equação da regressão linear para um problema univariado é expressa da seguinte forma:

$$y = mx + b \tag{2.2.1}$$

onde y é a variável dependente, x é a variável independente (a entrada), m é o coeficiente angular e b é o termo de interceptação.

Para problemas multivariados, considerando várias variáveis independentes, a equação se generaliza para:

$$y = b + m_1x_1 + m_2x_2 + \ldots + m_nx_n \tag{2.2.2}$$

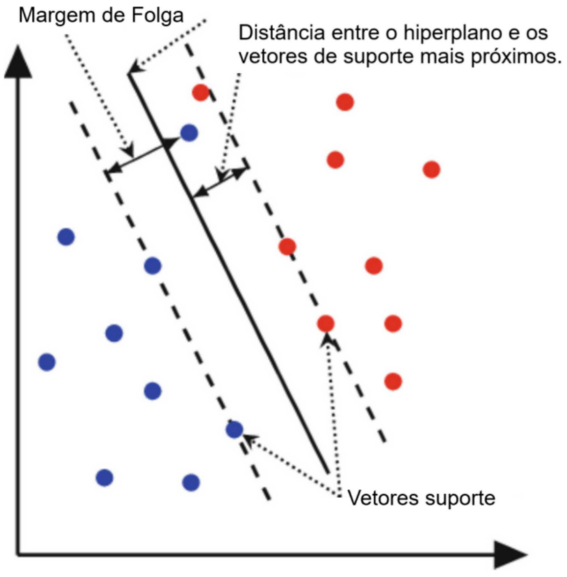
onde y é a variável dependente, $x_1, x_2, \ldots, x_n$ são as variáveis independentes, b é o termo de interceptação, e $m_1, m_2, \ldots, m_n$ são os coeficientes angulares associados a cada variável independente.

O objetivo durante o treinamento de um modelo de regressão linear é ajustar os valores dos coeficientes (m e b) de modo que a diferença entre as previsões do modelo e os valores reais seja minimizada.

2.2.2 Máquina Vetor Suporte (SVM)

As máquinas vetor suporte foram introduzidas por Vapnik e Cortes em 1995 [Cortes & Vapnik, 1995] com teorias baseadas em aprendizagem estatístico com intuito de resolver problemas de classificação de padrões. As SVMs se tornaram populares devido a sua capacidade de lidar com dados não-lineares e de alta dimensionalidade. O princípio fundamental por trás das SVMs é encontrar o hiperplano de separação ótimo que maximiza a margem entre diferentes classes. Isso é alcançado através da identificação de vetores de suporte, que são instâncias de dados críticas para a definição do hiperplano de decisão. [Cristianini & Shawe-Taylor, 2000]. Uma ilustração do algoritmo é exibida na figura 2.1.

Figura 2.1 – Representação do algoritmo de Máquina de Vetores de Suporte (SVM)



Fonte: Adaptado de Wang *et al.* [2024].

Neste trabalho aplicaremos uma variação da SVM, o método de regressão para as máquinas suporte, SVR. Apresentaremos a formulação matemática de forma resumida abaixo, porém todos os detalhes estão disponíveis na literatura [Vapnik *et al.*, 1997]

Suponha que tenhamos um conjunto de dados de treinamento $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$, onde $\mathbf{x}_i$ é um vetor de características de entrada e y_i é o valor alvo correspondente. O objetivo do SVR é aprender uma função $f(\mathbf{x})$ que aproxime os valores y_i o melhor possível.

O modelo SVR assume uma função de decisão da forma:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{2.2.3}$$

onde $\mathbf{w}$ é o vetor de pesos, $\mathbf{x}$ é o vetor de características de entrada e b é o termo de viés (bias).

O SVR introduz uma restrição de margem para garantir uma generalização eficaz. Define-se uma margem de tolerância ϵ e busca-se minimizar a função objetivo sujeita à condição de que as predições $f(\mathbf{x}_i)$ devem estar dentro de uma faixa de ϵ do valor real y_i. A formulação desta restrição é dada por:

$$-\epsilon \leq y_i - f(\mathbf{x}_i) \leq \epsilon \tag{2.2.4}$$

A função objetivo a ser minimizada é composta por duas partes: a minimização da norma do vetor de pesos $\|\mathbf{w}\|^2$ para evitar overfitting e a penalização por violações da margem ϵ. A função objetivo do SVR é expressa como:

$$\min_{\mathbf{w},b} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right) \tag{2.2.5}$$

sujeito às restrições:

$$y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \leq \epsilon + \xi_i \tag{2.2.6}$$

$$(\mathbf{w} \cdot \mathbf{x}_i + b) - y_i \leq \epsilon + \xi_i^* \tag{2.2.7}$$

$$\xi_i, \xi_i^* \geq 0 \tag{2.2.8}$$

onde C é o parâmetro de regularização e ξ_i, ξ_i^* são variáveis de folga que representam a violação da margem para cada ponto de treinamento.

O otimizador tenta encontrar os parâmetros $\mathbf{w}$ e b que minimizam essa função objetivo, ajustando a reta de regressão de maneira a manter as predições dentro da margem de tolerância ϵ.

2.2.3 K-vizinhos Mais Próximos

O conceito de "vizinhos mais próximos", que fundamenta o algoritmo *k-nearest neighbors* (KNN), foi inicialmente proposto por Cover et al. em 1967.

O KNN é um método supervisionado de aprendizagem de máquina baseado na ideia de identificar e classificar pontos vizinhos com características similares. Ele possui ampla aplicação em problemas de regressão, classificação e inferência de dados faltantes, sendo um método simples, eficiente e de fácil implementação.

O algoritmo clássico do KNN define um parâmetro k, que representa o número de vizinhos mais próximos a serem considerados para a classificação de um ponto. Para cada ponto do conjunto de dados, calcula-se a distância em relação aos demais pontos, de modo que os k vizinhos mais próximos sejam identificados e classificados.

As três métricas de distância mais comuns para determinar os vizinhos mais próximos são:

- Distância Euclidiana:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.2.9}$$

- Distância de Manhattan:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.2.10}$$

- Distância de Minkowski, que generaliza as anteriores:

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^q\right)^{\frac{1}{q}} \tag{2.2.11}$$

Onde $x = (x_1, x_2, \ldots, x_n)$ e $y = (y_1, y_2, \ldots, y_n)$ representam dois pontos no espaço de características. Para $q = 1$, obtém-se a distância de Manhattan, e para $q = 2$, a distância Euclidiana.

Após o cálculo das distâncias, os vizinhos mais próximos são selecionados. A forma mais comum de classificação é a votação majoritária, em que a classe com maior número de vizinhos determina a rotulação do ponto.

Outra abordagem considera a atribuição de pesos aos vizinhos mais próximos, onde a influência de cada vizinho é inversamente proporcional à sua distância. Além disso, algumas variantes do método utilizam funções exponenciais para ponderação, atribuindo maior relevância aos pontos mais próximos. Diferentes versões do KNN foram propostas para aprimorar sua eficiência e aplicabilidade:

Algumas variantes do KNN clássico são do tipo KNN adaptativo, cujo foco reside na seleção do valor ótimo para k em cada ponto do conjunto considerado [Sun & Huang, 2010]. Um exemplo é o KNN localmente adaptativo com discriminação de classe, que determina que a quantidade e a distribuição dos vizinhos das classes majoritária principal e secundária devem ser utilizadas para definir um valor ideal de k na vizinhança k de um determinado ponto [Hastie & Tibshirani, 1995] . Outro exemplo é o KNN com ajuste de pesos, que atribui pesos a cada um dos pontos do conjunto de dados considerado. Existem ainda outras variações com abordagens semelhantes [Zuo *et al.*, 2008].

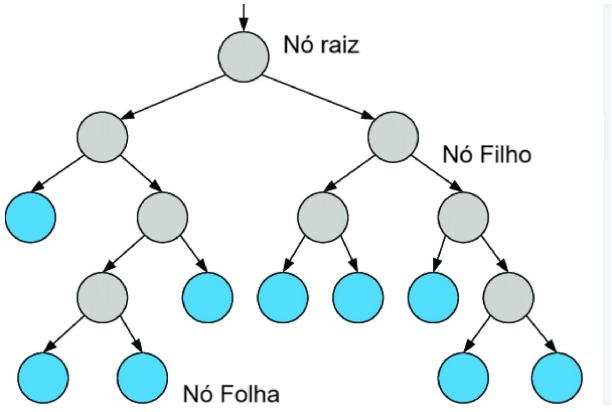
2.2.4 Árvore de Decisão

Uma árvore é definida como um conceito amplamente utilizado nos algoritmos dentro da computação, onde uma estrutura de dados é organizada com nós internos ou terminais e um nó raiz. Dentro da aprendizagem de máquina, as árvores são organizadas com base em um atributo, onde cada nó folha representa um rótulo de classe, para problemas de classificação, e um valor predito para casos de problemas de regressão. [Osei-Bryson, 2004]

A construção da árvore é baseada em um particionamento recursivo dos dados em nós, de forma a separar os conjuntos de amostras em subgrupos. Primeiramente todas as amostras determinam a estrutura da árvore; após isso o algoritmo separa os dados em ramificações, de modo que minimize a soma dos desvios quadrados da média nas partes separadas. O processo se repete recursivamente, aplicado a novos nós até que se chegue no nó folha. [Osei-Bryson, 2004] e [Xu *et al.*, 2005]

Os nós de decisão (os nós folha) contêm as condições utilizadas para segmentar os dados com base em um atributo específico, enquanto os ramos representam os diferentes valores possíveis que um atributo pode assumir, determinando o caminho a ser seguido dentro da árvore. Os nós folha, por sua vez, correspondem às saídas finais do modelo, podendo indicar uma classe predita em problemas de classificação ou um valor numérico em casos de regressão.

Figura 2.2 – Exemplo de uma Árvore de Decisão



Fonte: Adaptado de [Mohanty & Gao, 2024].

A figura 2.2 ilustra o algoritmo das árvores de decisão. Cada nó interno representa uma condição sobre uma variável, enquanto os nós folha indicam a saída prevista para um determinado conjunto de características.

Dentre as principais vantagens das árvores de decisão, destacam-se a simplicidade e a interpretabilidade. Como sua estrutura é intuitiva e de fácil compreensão, esses modelos são frequentemente aplicados a diversas áreas do conhecimento. Outra vantagem

é sua capacidade de modelar relações não lineares entre variáveis, garantindo uma boa flexibilidade na segmentação dos dados.

Por outro lado, uma das principais limitações das árvores de decisão é sua propensão ao sobreajuste (overfitting), o que pode resultar em modelos excessivamente adaptados aos dados de treinamento e com baixa capacidade de generalização para novos dados. Além disso, em problemas que envolvem grandes volumes de informação, a estrutura da árvore pode se tornar complexa e de difícil interpretação. Outra desvantagem é a instabilidade do modelo, pois pequenas variações nos dados podem levar à geração de árvores consideravelmente diferentes.

Em resumo, as árvores de decisão representam um dos principais modelos de aprendizagem de máquina, combinando interpretabilidade e eficiência computacional. No entanto, para garantir um bom desempenho preditivo e evitar sobreajuste, é comum o uso de técnicas complementares, como poda da árvore (pruning), ajuste de hiperparâmetros e a aplicação de métodos baseados em conjuntos (ensemble methods), como Random Forest e Gradient Boosting.

2.2.5 Floresta Aleatória

A Floresta Aleatória, (do inglês, Random Forest), é um algoritmo de aprendizagem de máquina introduzido por Breiman em 2001 [Breiman, 2001b] que combina várias árvores de decisão para tomar decisões mais robustas. Cada árvore é construída de forma independente usando uma amostra aleatória dos dados e considerando apenas um subconjunto aleatório das características em cada ponto de decisão.

A construção da floresta depende de dois fatores na regressão aleatória: o número de árvores (k) a serem plantadas na floresta e o número de variáveis (m) especificadas em cada nó para o crescimento da árvore. [Kouadri *et al.*, 2021]

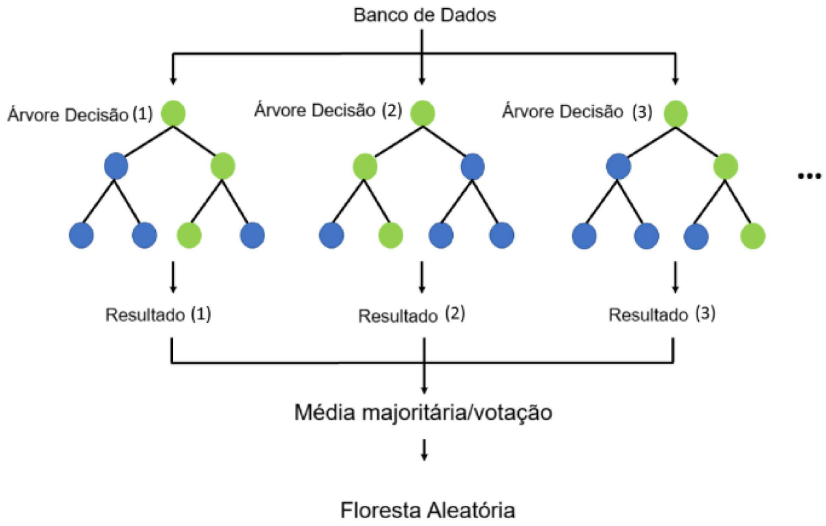
A predição final de um modelo de Floresta Aleatória (Random Forest) pode ser representada matematicamente pela Equação 2.2.12.

$$y = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{2.2.12}$$

onde y é a predição final, T é o número total de árvores, e $h_t(x)$ representa a predição feita pela árvore t para uma entrada x [Harrison, 2019].

Durante a fase de amostragem, um subconjunto aleatório dos dados de treinamento é selecionado para cada árvore por meio da técnica de bagging, com reposição. Em seguida, para cada divisão de nó dentro da árvore, um subconjunto aleatório de variáveis é escolhido. Esse processo garante uma menor correlação entre as árvores, aspecto fundamental para um bom desempenho do modelo.

Figura 2.3 – Ilustração do algoritmo Floresta Aleatória.



Fonte: Adaptado de [Li *et al.*, 2024].

Na etapa de predição, cada árvore contribui com uma previsão para um novo dado, e a estimativa final do modelo é obtida por meio da média das previsões no caso de problemas de regressão, ou por votação no caso de classificação.

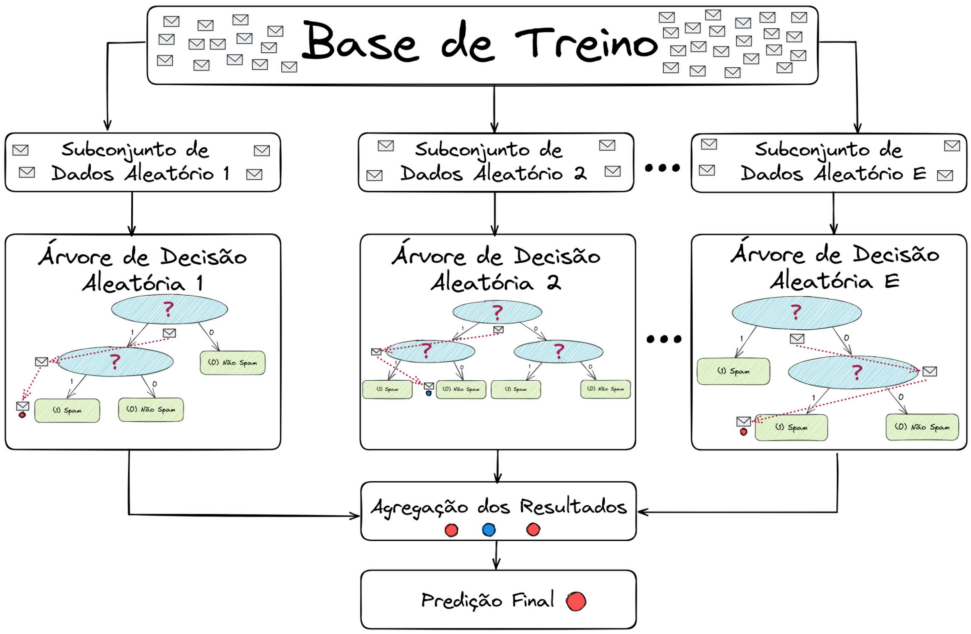
Entre as principais vantagens das florestas aleatórias estão a capacidade de processar grandes volumes de dados, lidar com valores atípicos e ruídos, além de reduzir a variância e minimizar o risco de sobreajuste. Além disso, em um modelo de RF, é possível medir a similaridade entre amostras dentro do conjunto de treinamento, proporcionando informações valiosas sobre a estrutura dos dados [Breiman, 2001a], [Alnuaimi & Albaldawi, 2024].

2.2.6 Árvore de Decisão Extra (Extra Trees)

O algoritmo de Árvore de Decisão Extra, proposto por Geurts *et al.* [2006], é uma variação da floresta aleatória que introduz maior aleatoriedade na construção das árvores. Enquanto o RF seleciona a melhor divisão entre um subconjunto aleatório de características, as árvores de decisão escolhem não apenas as variáveis de forma aleatória, mas também os pontos de divisão (*thresholds*) em cada nó.

A figura 2.4 ilustra o algoritmo das árvores de decisão extra. A base de treino é dividida em subconjuntos aleatórios, utilizados para gerar múltiplas árvores de decisão com divisões estocásticas. Os resultados dessas árvores são então agregados para produzir uma predição final, característica dos métodos de ensemble.

Figura 2.4 – Diagrama ilustrativo do funcionamento do algoritmo Extra Trees.



Fonte: Adaptado de Lopes [2023].

Considere um conjunto de dados de treinamento

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \tag{2.2.13}$$

em que $x_i \in R^p$ representa o vetor de atributos e y_i o valor-alvo, contínuo ou categórico. O algoritmo Extra Trees constrói um conjunto de M árvores $\{T_m\}_{m=1}^{M}$, seguindo o procedimento resumido a seguir.

Em cada nó da árvore, seleciona-se aleatoriamente um subconjunto de variáveis $F \subset \{1, 2, \ldots, p\}$. Para cada variável $X_j \in F$, um ponto de divisão t_j é escolhido aleatoriamente dentro do intervalo observado dessa variável. A partição do nó é determinada pelo par (X_{j^*}, t_{j^*}) que maximiza uma função de ganho de informação G:

$$(X_{j^*}, t_{j^*}) = \arg\max_{j \in F} G(X_j, t_j). \tag{2.2.14}$$

As árvores são expandidas até atingir uma profundidade máxima ou um critério de parada definido. A predição final do ensemble depende do tipo de problema.

Para regressão, calcula-se a média das saídas das árvores:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x) \tag{2.2.15}$$

enquanto, para classificação, utiliza-se a votação majoritária:

$$\hat{y} = \arg\max_{c} \sum_{m=1}^{M} \mathbf{1}\{T_m(x) = c\}, \tag{2.2.16}$$

em que $\mathbf{1}\{\cdot\}$ é a função indicadora.

Essa estratégia aumenta a robustez a ruídos e reduz a variância do modelo, apresentando desempenho competitivo com o Random Forest, porém com maior eficiência computacional [Geurts *et al.*, 2006], [Wehenkel *et al.*, 2006].

2.2.7 Perceptron Multicamadas

O perceptron multicamadas (*Multilayer Perceptron* – MLP) é uma categoria de rede neuronal artificial (ANN) cuja origem está associada ao estudo pioneiro de Rosenblatt [1958], no qual foi apresentado o perceptron simples. Esse modelo inicial foi concebido para o reconhecimento e processamento de padrões, porém sua aplicação se restringia a problemas linearmente separáveis. Anos depois, em 1986, Rumelhart *et al.* [1986] desenvolveram o algoritmo de retropropagação, o que impulsionou significativamente as investigações acerca do MLP e ampliou seu uso em diferentes áreas.

O MLP pode ser considerado uma versão aprimorada do perceptron simples, pois sua estrutura, composta por diversas camadas, permite a aprendizagem de padrões mais complexos e a resolução de problemas não lineares. A introdução do treinamento baseado em retropropagação foi um marco essencial para a evolução das redes neuronal, tornando possível sua aplicação em domínios como classificação, regressão e reconhecimento de padrões.

A configuração do Perceptron Multicamadas (MLP) estabelece uma relação não linear entre um vetor de entrada e um vetor de saída, conforme as Equações 2.2.17 e 2.2.18.

$$x = [x_1, x_2, x_3] \tag{2.2.17}$$

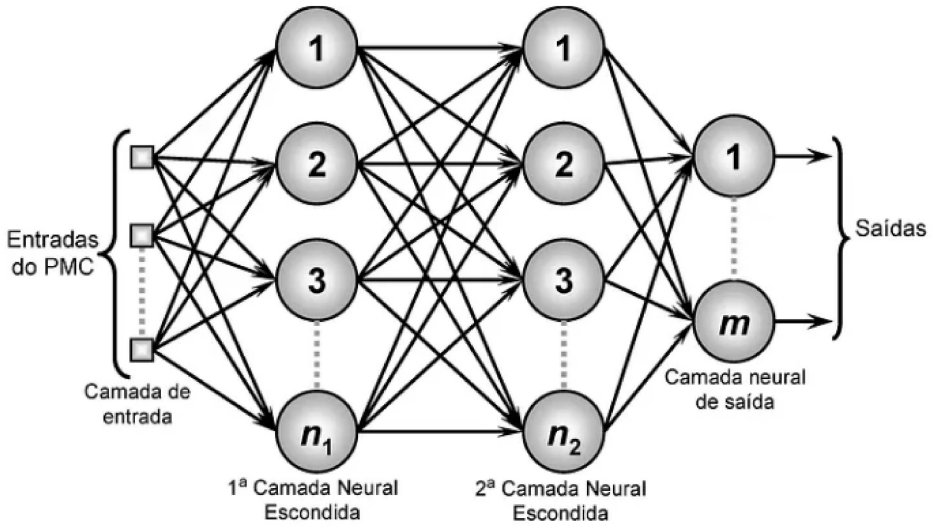
$$y = [y_1, y_2] \tag{2.2.18}$$

por meio de um conjunto de neurônios interligados, denominados nós. Esses nós são conectados por pesos e transmitem sinais de saída calculados a partir da soma ponderada das entradas e da aplicação de uma função de ativação, também conhecida como função de propagação direta (*feedforward*).

Esse processo é conduzido por uma função de transferência não linear, que desempenha um papel fundamental na capacidade do MLP de modelar relações altamente complexas. Essa propriedade decorre da combinação sucessiva de múltiplas funções de transferência, sendo a função sigmoide logística uma das mais empregadas, em razão da simplicidade no cálculo de sua derivada.

A arquitetura típica de um Perceptron Multicamadas é ilustrada na Figura 2.8, composta por uma camada de entrada, duas camadas ocultas e uma camada de saída. Essa estrutura permite a modelagem de relações não lineares entre variáveis de entrada e saída, sendo amplamente utilizada em tarefas de classificação e regressão.

Figura 2.5 – Arquitetura de um Perceptron Multicamadas (MLP)



Fonte: Adaptado de Ensina.AI [2023].

O processo de treinamento é realizado em duas fases principais: a propagação direta e a retropropagação de erros. No estágio de propagação direta, as informações iniciais transitam por diferentes camadas da rede, interagindo com os neurônios e suas sinapses com pesos. A saída gerada neste processo é comparada aos valores previstos, sendo analisada através de uma função de perda que mede o erro do modelo.

Durante a fase de retropropagação, a rede neuronal modifica seus pesos com o objetivo de reduzir o erro total. Esse processo ocorre de forma repetitiva, utilizando métodos de otimização, como o gradiente descendente, que se dedica a localizar o mínimo global da função de erro. No início, os pesos são definidos com valores pequenos e aleatórios. Em seguida, calcula-se o gradiente da função de erro para identificar a direção na qual os pesos devem ser ajustados, minimizando o erro a cada iteração. Se a superfície de erro for suficientemente suave, espera-se que os pesos se aproximem de um ótimo global, levando a um modelo mais eficaz.

O treinamento de uma rede MLP tem início com a definição inicial dos pesos. Posteriormente, os dados de entrada são inseridos no modelo, que processa a informação através das camadas até produzir um resultado. Esse resultado é comparado com os valores reais para determinar o erro. O erro calculado é então retroalimentado na rede, modificando os pesos para diminuir a diferença entre o resultado previsto e o desejado. Esse processo de ajustes é repetido continuamente, utilizando novos dados de entrada, até

que o erro alcance um nível aceitável. Esse procedimento cíclico possibilita que a rede neuronal identifique padrões.

2.2.8 Gradient Boosting

O Modelo de Gradient Boost é um algoritmo de aprendizagem de máquina usado para classificação e regressão. O termo gradiente no nome se refere aos gradientes negativos também definidos como erros residuais, que são as diferenças entre os valores reais e valores previstos [Tanha *et al.*, 2020]. A etapa inicial é construir uma árvore simples de decisão, uma "árvore fraca", e calcular o gradiente. A partir disso, novas árvores são geradas, porém aprendendo padrões dos erros calculados nas etapas anteriores. O objetivo principal é minimizar os resíduos. A definição matematica foi realizada por (Friedman, 2001)[Friedman, 2001], onde:

$$(\beta_m, a_m) = \underset{\beta,a}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \beta h(x_i, a)) \tag{2.2.19}$$

Aqui, β_m e a_m são os parâmetros que minimizam a soma ponderada da função Ψ, que corresponde função de perda de erro quadrático médio, para todas as amostras, considerando o modelo $F_{m-1}(x)$ e a adição de um termo $\beta h(x, a)$.

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \tag{2.2.20}$$

O modelo $F_m(x)$ é atualizado com a adição de um termo ponderado $\beta_m h(x; a_m)$.

$$a_m = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{N} [-g_m(x_i) - \beta h(x_i; a)]^2 \tag{2.2.21}$$

O parâmetro a_m é escolhido minimizando a soma dos quadrados dos resíduos entre $-g_m(x_i)$ e $-\beta h(x_i; a)$, onde $g_m(x_i)$ é o gradiente do modelo em relação às predições anteriores.

$$\rho_m = \underset{\rho}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}(x_i) + \rho h(x_i, a_m)\right) \tag{2.2.22}$$

O parâmetro ρ_m é escolhido minimizando a soma dos quadrados dos erros para todas as amostras, considerando o modelo anterior $F_{m-1}(x)$ e a adição de um termo $\rho h(x, a_m)$.

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m) \tag{2.2.23}$$

O modelo $F_m(x)$ é atualizado com a adição de um termo ponderado $\rho_m h(x; a_m)$.

2.2.9 Extreme Gradient Boosting

O XGBoost é uma versão aprimorada do Gradient Boosting, que incorpora técnicas eficientes para o controle do sobreajuste (do inglês, *overfitting*), a otimização das divisões e

o tratamento de valores ausentes na fase de treinamento [Tanha *et al.*, 2020]. O modelo foi proposto por Chen (2016) [Chen & Guestrin, 2016] e introduz a definição de uma função de perda associada a um termo de regularização, conforme expresso a seguir:

$$\text{Obj}(\Theta) = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.2.24}$$

A primeira parte $\sum_{i=1}^{N} L(y_i, F(x_i))$ representa a função de perda,e $\sum_{k=1}^{K} \Omega(f_k)$ é a regularização dos modelos base (árvores de decisão). A regularização (Ω) ajuda a evitar o *overfitting*.

Os pesos das folhas para cada iteração são calculadas com

$$w_j^* = -\frac{\sum_i g_i}{\sum_i h_i + \lambda} \tag{2.2.25}$$

Calculado como a soma dos gradientes (g_i) dividida pela soma das hessianas (h_i) mais um termo de regularização (λ).

Por fim, o critério de seleção da árvore é usado para escolher a melhor árvore em cada iteração:

$$L(q) = -\frac{1}{2}\frac{\left(\sum_i g_i\right)^2}{\sum_i h_i + \lambda} + \gamma T \tag{2.2.26}$$

onde o primeiro termo é uma penalidade ao quadrado dos gradientes normalizados pela soma das hessianas mais o termo de regularização (λ), segundo termo (γT) é uma penalidade pela complexidade da árvore (T é o número de folhas na árvore). Isso ajuda a evitar árvores muito complexas.

Para otimizar o modelo ensemble apresentado na Equação (3.20), os métodos tradicionais não podem ser utilizados dentro do espaço euclidiano. Desta forma, é utilizado $\hat{y}^{(t)}$ deve representar a predição da i-ésima instância na iteração dada por t, de forma que o modelo seja treinado de forma gulosa com a inclusão de f_t para a minimização da função objetivo .

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.2.27}$$

À Equação (3.21), o valor ótimo de f_t é incorporado em $f^{(t)}$ para aprimorar o modelo. Como resultado, pode-se empregar a aproximação de segunda ordem para otimização. Nessa expressão, g_i e h_i correspondem, respectivamente, às estatísticas de gradiente de primeira e segunda ordem dentro da função de perda.

$$L^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.2.28}$$

onde $g_i = \frac{\partial}{\partial \hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ e $h_i = \frac{\partial^2}{\partial \hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$.

Removendo os termos constantes, a Equação (3.22) pode ser reescrita de maneira simplificada, resultando na Equação (3.23).

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.2.29}$$

Como mostra a Equação (3.24), Ω pode ser expandido na Equação (3.23) e o termo $I_j = \{i|q(x_i) = j\}$ pode ser inserido como o conjunto de instâncias da folha dada por j.

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{2.2.30}$$

Usando o valor ótimo do j-ésimo peso de uma folha j, dado por w_j^*, apresentado na Equação (3.19). Obtemos o $\tilde{L}^{(t)}$ ótimo correspondente:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{2.2.31}$$

A eficiência de uma determinada estrutura q de árvore pode ser avaliada com base na Equação correspondente, que atua como uma métrica de pontuação (score), análoga à medida de impureza utilizada na análise de árvores de decisão. Devido ao grande número de configurações possíveis para q, enumerar todas as alternativas é inviável. Como solução, pode-se adotar um algoritmo guloso que inicia com uma única folha e, progressivamente, insere ramos na estrutura da árvore. Sejam I_L e I_R os subconjuntos de instâncias correspondentes aos nós esquerdo e direito após a divisão, respectivamente. Considerando que $I = I_L \cup I_R$, a Equação (4.18) representa a redução da perda decorrente dessa divisão.

$$L_{\text{split}} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{2.2.32}$$

Em aplicações práticas, o termo L_{split} da Equação (4.18) é utilizado para a avaliação dos candidatos resultantes da divisão. Um dos desafios fundamentais na abordagem com aprendizagem de árvores consiste justamente na identificação da divisão ótima L_{split}.

2.2.10 Light Gradient Boosting Machine (LightGBM)

O Light Gradient Boosting Machine (LightGBM), proposto por Ke *et al.* [2017], é uma implementação otimizada do algoritmo Gradient Boosting, projetada para maior eficiência computacional em grandes conjuntos de dados. O LightGBM apresenta duas inovações principais: o *Gradient-based One-Side Sampling* (GOSS) e o *Exclusive Feature Bundling* (EFB), que reduzem o custo de treinamento e melhoram o desempenho preditivo.

Gradient-based One-Side Sampling (GOSS)

O GOSS tem como objetivo acelerar o treinamento priorizando instâncias com maiores gradientes, que contribuem mais para a atualização do modelo. Considere um conjunto de dados

$$D = \{(x_i, y_i, g_i, h_i)\}_{i=1}^{N}, \tag{2.2.33}$$

onde x_i representa o vetor de atributos, y_i o rótulo (classe ou valor contínuo), g_i o gradiente da função de perda em relação à predição do modelo para a instância i, e h_i o Hessiano correspondente. O algoritmo segue as seguintes etapas:

1. Ordenam-se as instâncias pelo valor absoluto do gradiente $|g_i|$; 2. Selecionam-se as $a \times 100\%$ instâncias com maiores gradientes (mais informativas); 3. Amostra-se aleatoriamente $b \times 100\%$ das instâncias restantes com menores gradientes; 4. Durante o cálculo do ganho de informação, os gradientes das instâncias com menores valores são escalados por um fator de correção

$$\lambda = \frac{1-a}{b} \tag{2.2.34}$$

garantindo que o modelo mantenha uma estimativa de gradiente não enviesada.

A função de perda genérica utilizada no *boosting* é dada por:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) \tag{2.2.35}$$

onde $l(y_i, \hat{y}_i)$ representa a função de perda (como erro quadrático ou *log-loss*), e $\hat{y}_i$ é a predição atual do modelo.

Os gradientes e hessianos são calculados como:

$$g_i = \frac{\partial \mathcal{L}}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 \mathcal{L}}{\partial \hat{y}_i^2} \tag{2.2.36}$$

O ganho de informação G para um candidato a ponto de divisão é dado por:

$$G = \frac{1}{2}\left(\frac{\left(\sum_{i\in L} g_i\right)^2}{\sum_{i\in L} h_i + \lambda} + \frac{\left(\sum_{i\in R} g_i\right)^2}{\sum_{i\in R} h_i + \lambda} - \frac{\left(\sum_{i\in D} g_i\right)^2}{\sum_{i\in D} h_i + \lambda}\right) \tag{2.2.37}$$

em que L e R representam os conjuntos de amostras enviadas para os ramos esquerdo e direito, respectivamente.

Exclusive Feature Bundling (EFB)

O EFB é utilizado para reduzir a dimensionalidade em conjuntos de dados esparsos, principalmente quando variáveis categóricas codificadas em *one-hot* são mutuamente

exclusivas. Essa técnica constrói um grafo de conflitos entre variáveis, em que cada nó representa uma característica, e cada aresta ponderada representa a ocorrência simultânea de valores não nulos.

Formalmente, seja f_i um atributo e Z_i o conjunto de instâncias em que $f_i \neq 0$. Dois atributos f_i e f_j podem ser agrupados se:

$$Z_i \cap Z_j = \emptyset \quad \text{ou} \quad |Z_i \cap Z_j| < c \tag{2.2.38}$$

onde c é o limite de colisão aceitável. Variáveis são agrupadas iterativamente em *bundles*, minimizando conflitos internos e reduzindo o número total de atributos processados [Tanha *et al.*, 2020, Thai, 2022].

2.2.11 *CatBoosting*

O *CatBoost* é um algoritmo de aprendizagem de máquina baseado no método de *gradient boosting* sobre árvores de decisão. Diferentemente de abordagens tradicionais, ele utiliza uma técnica denominada *Ordered Boosting*, que busca minimizar o viés introduzido durante o treinamento sequencial das árvores. Essa estratégia é particularmente eficaz para lidar com atributos categóricos, principal diferencial do algoritmo, permitindo seu tratamento de forma nativa, sem necessidade de codificações extensas como *one-hot* [Dorogush *et al.*, 2018].

No Ordered Boosting, os dados de treinamento são aleatoriamente embaralhados e, para cada permutação P_k, as árvores são construídas considerando apenas as instâncias anteriores àquela que se deseja predizer. Seja o conjunto de dados de treinamento

$$D = \{(x_i, y_i)\}_{i=1}^{N} \tag{2.2.39}$$

onde x_i representa o vetor de atributos e y_i o rótulo associado. Para cada árvore T_t e para uma instância x_i, a atualização da predição é dada por:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot T_t \left(x_i \mid \{(x_j, y_j) : j < i \text{ em } P_t\} \right) \tag{2.2.40}$$

onde η é a taxa de aprendizagem e P_t é a permutação utilizada na iteração t. Essa construção garante que a predição de x_i não seja influenciada por sua própria informação, reduzindo o risco de *overfitting* e o viés de treinamento [Prokhorenkova *et al.*, 2018].

Outro diferencial do CatBoost é o tratamento nativo de variáveis categóricas. Em vez de utilizar codificações como *one-hot*, ele aplica uma codificação estatística incremental que evita vazamento de informação. Para uma variável categórica c, a codificação para a instância i é dada por:

$$C_i = \frac{\sum_{j<i} \mathbf{1}\{c_j = c_i\} \cdot y_j + a}{\sum_{j<i} \mathbf{1}\{c_j = c_i\} + a} \tag{2.2.41}$$

onde a é um parâmetro de suavização, e a soma considera apenas instâncias anteriores a i na permutação, garantindo imparcialidade na estimativa.

Como outros algoritmos baseados em *boosting*, o CatBoost minimiza uma função de perda genérica $\mathcal{L}$, que para classificação binária pode ser escrita como:

$$\mathcal{L} = -\sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{2.2.42}$$

Os gradientes e hessianos são calculados para cada instância i como:

$$g_i = \frac{\partial \mathcal{L}}{\partial \hat{y}_i}, \qquad h_i = \frac{\partial^2 \mathcal{L}}{\partial \hat{y}_i^2}. \tag{2.2.43}$$

Esses valores são utilizados para a construção sequencial das árvores de decisão que compõem o *ensemble* final.

Outro aspecto relevante é que o CatBoost fornece ao usuário uma série de parâmetros ajustáveis, que permitem o controle sobre a profundidade das árvores, a taxa de aprendizagem e outras configurações que influenciam diretamente o desempenho do modelo. A documentação oficial do algoritmo [Yandex, 2024] oferece diretrizes detalhadas para a escolha desses hiperparâmetros, facilitando sua aplicação em diversos contextos analíticos. Em conjunto, essas características fazem do CatBoost uma ferramenta poderosa para tarefas supervisionadas, especialmente em cenários com grande presença de dados categóricos.

2.3 Bacia Hidrográfica

Para compreender adequadamente o conceito de bacia hidrográfica, é necessário, primeiramente, distinguir os termos rio e bacia. O rio é um curso natural de água que escoa continuamente por um canal, normalmente em direção a um corpo hídrico maior, como um lago, mar ou oceano. Já a bacia hidrográfica constitui uma unidade territorial delimitada topograficamente, na qual todas as águas provenientes da precipitação pluviométrica convergem, superficial ou subterraneamente, para um ponto comum de exutório, geralmente situado no leito de um rio principal [Tucci, 2001].

Dessa forma, a bacia hidrográfica não se limita apenas ao rio principal, mas abrange também todos os seus afluentes e a área de drenagem correspondente, incluindo o relevo, o solo, a vegetação e as intervenções antrópicas. Trata-se, portanto, de uma unidade fundamental para o planejamento e a gestão dos recursos hídricos, uma vez que representa o espaço físico onde ocorrem os processos hidrológicos que determinam o fluxo das águas [de Oliveira, 2010].

No caso da bacia hidrográfica do Rio Paraíba do Sul, esta abrange uma vasta rede de drenagem composta por diversos corpos hídricos, entre os quais se destaca o Rio Paraibuna, um de seus principais afluentes. O Rio Paraibuna, portanto, é um curso d'água secundário que contribui com parte significativa da vazão do Rio Paraíba do Sul, o qual exerce o papel de rio principal dentro da bacia.

A relação entre o Rio Paraíba do Sul, o Rio Paraibuna e a bacia hidrográfica pode ser assim sintetizada:

- O Rio Paraibuna constitui um afluente da bacia hidrográfica do Rio Paraíba do Sul, contribuindo para sua formação e manutenção;
- As águas provenientes da precipitação que ocorrem nas áreas compreendidas pela bacia são coletadas e direcionadas por meio de uma rede de drenagem composta por diversos rios, como o Paraibuna, até atingirem o leito do Paraíba do Sul;
- A bacia hidrográfica do Rio Paraíba do Sul corresponde, portanto, à totalidade da área drenada por este rio e seus tributários, até seu deságue no oceano Atlântico [de Águas, 2020].

Em síntese, o Rio Paraibuna é parte integrante do sistema hidrográfico que compõe a bacia do Paraíba do Sul [Comitê de Integração da Bacia Hidrográfica do Rio Paraíba do Sul (CEIVAP), 2020]. A compreensão dessa interdependência entre os rios e a bacia como um todo é fundamental para a análise dos processos de gestão ambiental que ocorrem no território.

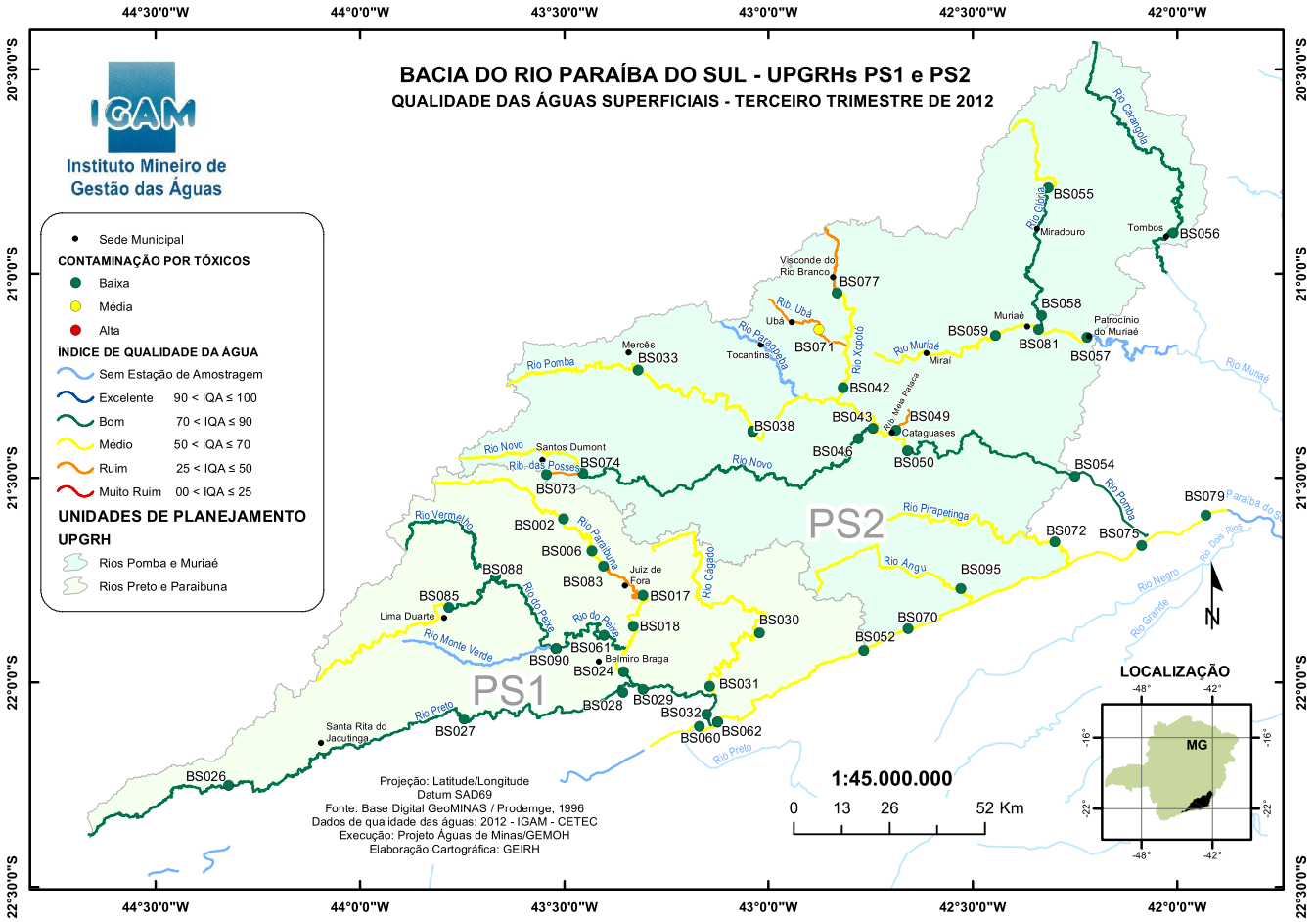
A Figura 2.6 apresenta a delimitação da Bacia do Rio Paranaíba Sul. É possível observar a distribuição dos pontos de monitoramento da qualidade das águas superficiais, classificados em diferentes níveis de qualidade (boa, regular, ruim e péssima), conforme dados do Instituto Mineiro de Gestão das Águas (2014).

2.4 Área de Estudo

O rio Paraibuna origina-se a 1.200 metros de altitude na serra da Mantiqueira, no município de Antônio Carlos no estado de Minas Gerais, e atravessa cerca de 37 municípios como Santos Dumont, Ewbanck da Câmara, Matias Barbosa, Simão Pereira, Belmiro Braga, Santana do Deserto, Chiador, Juiz de Fora, entre outros. No total, são aproximadamente 166 quilômetros até desaguar no Rio Paraíba do Sul, no município de Três Rios, no Rio de Janeiro, a 250 metros de altitude, com uma vazão de 200 m^3/s. [Araújo *et al.*, 2009]

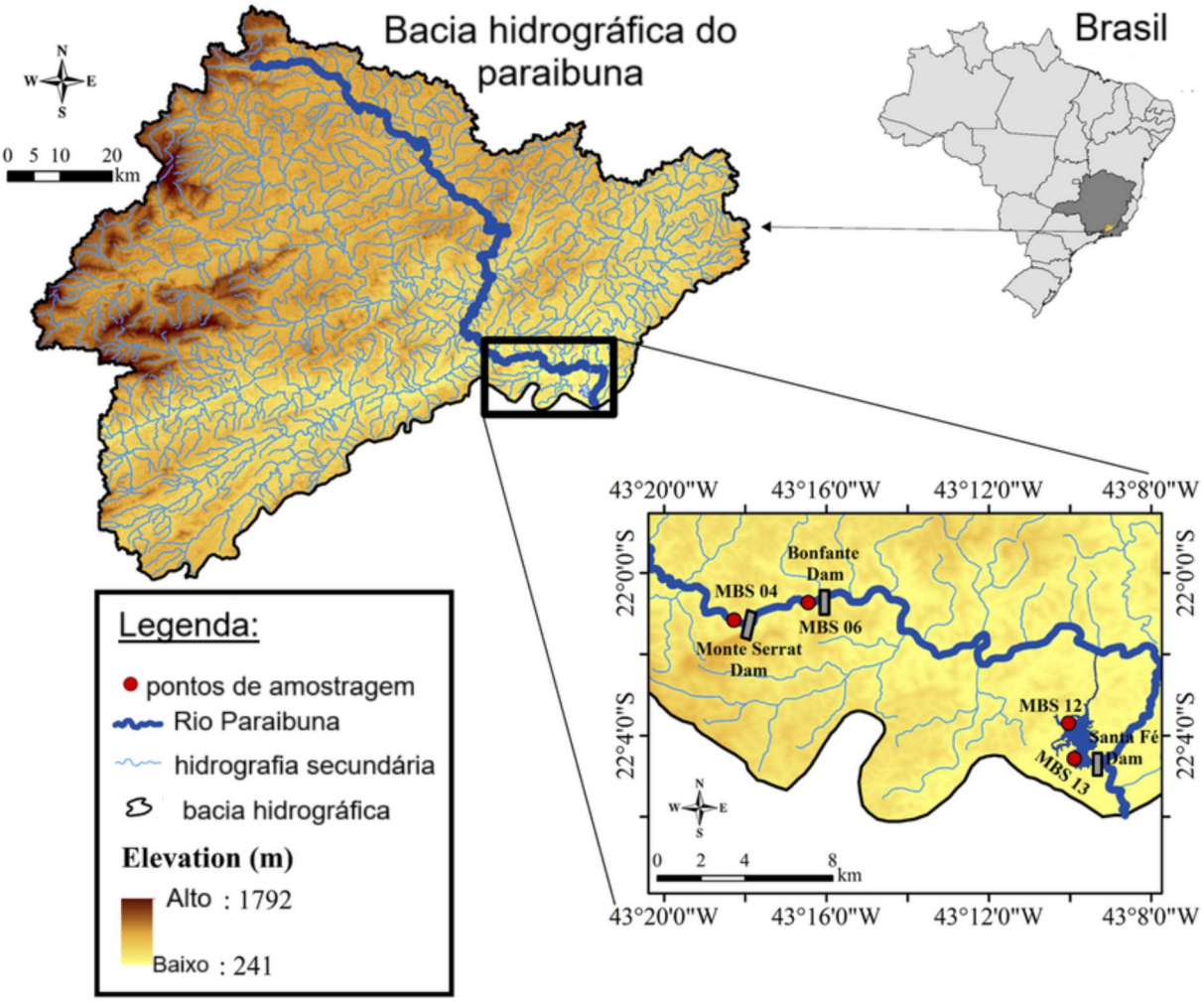
A Figura 2.7 apresenta a localização do rio Paraibuna e a hidrografia secundária. O mapa inserido detalha os níveis de elevação e a localização dos pontos de amostragem, próximos aos reservatórios Bonfante, Monte Serrat e Santa Fé, que são utilizados como referência neste estudo.

Figura 2.6 – Bacia Rio Paraíba do Sul, retirado do relatório Avaliação da Qualidade das Águas Superficiais de Minas Gerais



Fonte: Instituto Mineiro de Gestão das Águas [2024].

Figura 2.7 – Mapa de localização dos reservatórios Monte Serrat, Bonfante e Santa Fé ao longo do rio Paraibuna.



Fonte: Adaptado de [da Silva Resende *et al.*, 2022].
Nota: Os pontos vermelhos representam as estações de amostragem: MBS04 — Monte Serrat, MBS06 — Bonfante, MBS12 — Santa Fé I e MBS13 — Santa Fé II.

Seu curso é sinuoso, com águas geralmente em baixa velocidade e estreito, porém volumoso. A declividade média é diversificada, no trecho da cidade de Juiz de Fora é da ordem de 1,0 m/km, porém em Matias Barbosa até o encontro com o Rio Paraíba do Sul é da ordem de 5,0m/km e nos 4km iniciais atinge valores máximos da ordem de 70m/km. [Orlando, 2006]

Uma das características marcantes do Rio Paraibuna é o seu papel na história. Os povos indígenas utilizavam as águas para alimentação e transporte muito antes da chegada dos colonos europeus. Assim, denominaram Parayuna, com PARA(água), HYB(rio) e UNA(preta), decorrente do fato do rio correr sobre formação rochosa. A margem do rio testemunhou a atividade humana durante séculos, desde a exploração colonial até a mineração e o desenvolvimento agrícola em Minas Gerais.

Atualmente, o Rio Paraibuna desempenha um papel vital na região onde está

localizado, não apenas em termos de abastecimento de água para as comunidades locais, mas também em termos de ecologia e biodiversidade. Até hoje 12,5% do abastecimento energético de Juiz de Fora é mantido pelas usinas do Rio Paraibuna, conhecidas como Joasal, paciência e Marmelos.

No entanto, com o desenvolvimento do município, o Rio Paraibuna enfrenta desafios significativos relacionados à poluição e degradação ambiental devido ao desenvolvimento humano na região. A descarga de esgoto e resíduos industriais no rio representa uma ameaça à qualidade da água e à saúde dos ecossistemas aquáticos. Segundo a Cesama, em 2019 foi possível tratar apenas 4,77% de todo o esgoto do município, sendo o restante 95,33%a lançado no rio. Cerca de 1.100 litros de esgoto por segundo atualmente.

2.5 Parâmetros

Os parâmetros físicos incluem turbidez, temperatura da água e sólidos totais dissolvidos, que indicam alterações no aspecto visual e nas condições físicas da água. Por exemplo, a turbidez elevada pode ser causada pela erosão do solo ou pela descarga de resíduos, afetando a amplitude da luz, essencial para os processos fotossintéticos. A temperatura da água também é determinante, influenciando a solubilidade de gases como o oxigênio e a vida aquática em geral. ([Sperling, 2002])

Parâmetros químicos englobam concentrações de oxigênio distribuídas, pH, demanda bioquímica de oxigênio (DBO) e nutrientes como nitrogênio e fósforo. A concentração de oxigênio disponível, por exemplo, é um indicador crítico, pois níveis baixos podem sinalizar poluição orgânica e comprometer ecossistemas aquáticos. O pH, por sua vez, influencia a disponibilidade de nutrientes e a toxicidade de substâncias químicas na água, tornando-se um parâmetro relevante para determinar a habitabilidade para espécies aquáticas e o uso humano. [APHA, AWWA, WEF, 2017]

Os parâmetros biológicos incluem indicadores como a presença de coliformes fecais [Instituto Mineiro de Gestão das Águas, 2024], utilizados para avaliar contaminações por esgoto doméstico, resíduos de origem animal e outras fontes de poluição orgânica, representando um importante indicativo da qualidade sanitária da água.

Tabela 2.1 – **Relação dos parâmetros analisados nas campanhas. Parâmetros comuns a todos os pontos. Adaptado de Instituto Mineiro de Gestão das Águas [2024]**

• Temperatura do ar • Temperatura da água • Condutividade • pH • Turbidez • Oxigênio Dissolvido • Sólidos Totais Dissolvidos • Sólidos em Suspensão • Sólidos Totais • Demanda Química de Oxigênio • Alcalinidade • Nitrogênio Amoniacal • Nitrito • Nitrato • Ortofosfato • Fósforo Total • Silicatos • Clorofila • Demanda Bioquímica de Oxigênio • Cianetos • Cloretos • *E. coli* • Coliformes Totais	• Cor • Cromo Hexavalente • Cromo Trivalente • Dureza • Fenóis • Mercúrio • Óleos e Graxas • Alumínio Dissolvido • Cádmio Total • Cobre Dissolvido • Cromo Total • Ferro Solúvel • Ferro Total • Manganês • Níquel • Zinco • Riqueza Bentônica Total • Densidade Bentônica Total • Riqueza Fitoplanctônica Total • Densidade Fitoplanctônica Total • Densidade de Cianobactérias • Riqueza de Macrófitas • Biomassa de Macrófitas • Riqueza Zooplanctônica Total • Densidade Zooplanctônica Total • IQA

Na tabela 2.1 são apresentados os parâmetros de qualidade de água analisados entre setembro de 2011 e março de 2023, abrangendo os pontos de monitoramento CP01 e MBS01 a MBS11. A descrição dos pontos sobre sua coordenada e localidade estão no quadro 2.2. A quantidade mínima de parâmetros observada em uma coleta foi 36, e a máxima chegou a 49. Não houve um padrão rígido na frequência ou na estrutura das coletas, mas percebe-se que a intenção era realizar campanhas em pelo menos uma estação chuvosa e outra de estiagem ao longo do ano. Ressalta-se que nem todos os parâmetros foram monitorados em todas as estações, havendo variações conforme a campanha e o ponto de coleta.

Quadro 2.2 – Pontos de Medição

Pontos	**Corpo Hídrico**	**Local**	**Coordenadas (UTM)**
MBS01	Rio Paraibuna	Fazenda Cambuí, a montante da confluência com o Rio Preto	672305, 7566038
MBS02	Rio Preto	A montante da confluência com o rio Paraibuna	672641, 7565317
MBS03	Rio Paraibuna	A montante da ponte no centro do distrito de Monte Serrat	674377, 7564581
MBS04	Rio Paraibuna	Interior do reservatório da PCH Monte Serrat, próximo ao Areal Monte Serrat	675030, 7564004
MBS05	Rio Paraibuna	Próximo à régua de medição de volume, após a água turbinada pela PCH Monte Serrat	675739, 7563931
MBS06	Rio Paraibuna	Interior do reservatório da PCH Bonfante	678497, 7564780
MBS07	Rio Paraibuna	TVR da PCH Bonfante, debaixo da ponte da BR-040	678652, 7565050
MBS08	Rio Paraibuna	Após restituição da PCH Bonfante e a montante da cidade de LevyGasparian	686607, 7562241
MBS09	Rio Paraibuna	A jusante da cidade de LevyGasparian	689688, 7562883
MBS10	Rio Paraibuna	Próximo ao barramento que desvia água para o Canal que alimenta o reservatório continental da PCH Santa Fé	690323, 7562446
MBS11	Rio Paraibuna	Cerca de 1km a montante da foz do rio Cágado, a jusante do barramento da PCH Santa-Fé	691974, 7562427
MBS12	Reservatório continental	Ponto localizado próximo à entrada de água no sistema	689966, 7559471
MBS13	Reservatório continental	Ponto localizado na margem oposta ao ponto MSB12, na margem direita da barragem principal	689334, 7558702
MBS14	Rio Paraibuna	Cerca de 1Km a montante do canal de fuga da casa de força, TVR a jusante da Casa de Força, ponto localizado no TVR	691419, 7558441
MBS15	Rio Paraibuna	Cerca de 1 Km a jusante do canal de fuga da PCH Santa Fé	690902, 7556594
CP01	Córrego Palmira	Margem esquerda do Rio Paraibuna	678578, 7565515

No contexto brasileiro, a Resolução Conama nº 357/2005, estabelece diretrizes para os padrões de qualidade da água, definindo os limites aceitáveis para cada parâmetro de acordo com o uso pretendido. Esses padrões são indispensáveis para subsidiar a gestão dos recursos hídricos e a preservação ambiental.

Diante do volume crescente de dados ambientais disponíveis, a utilização de algoritmos de aprendizagem de máquina apresenta-se como uma ferramenta promissora para avaliar e prever a variação desses parâmetros. Modelos baseados em aprendizagem de máquina podem identificar padrões e correlacionar múltiplos fatores, permitindo análises preditivas que otimizam os processos de monitoramento e tomada de decisão.

No caso específico do Rio Paraibuna, a variabilidade das parâmetros mencionadas reflete os desafios de gestão ambiental enfrentados pela região. O estudo e a modelagem desses dados podem contribuir para um sistema de monitoramento mais eficiente, oferecendo uma base sólida para a gestão sustentável da qualidade da água e mitigando os impactos ambientais na bacia hidrográfica.

2.5.1 Temperatura do Ar

No ar existem átomos e moléculas que se movem em todas as direções e se lançam livremente, colidindo uns com os outros. Alguns átomos e moléculas se movem mais rápidos que os outros, e essa diferença contabilizamos como a temperatura. Fisicamente, a temperatura mede o nível de agitação entre as moléculas, onde temperaturas mais altas correspondem a velocidades médias mais rápidas. Por exemplo, dentro de um balão flexível há um volume de ar em torno do seu tamanho. Se o ar do interior é aquecido, as moléculas se movem mais rápido, porém essas também se afastariam, e o ar se torna mais denso. De maneira contrária, se o ar é esfriado, as moléculas desacelaram e o ar se torna mais denso. [Halliday *et al.*, 2018]

Na atmosfera percebe-se que as variações na temperatura do ar ocorrem tanto durante um dia quanto ao longo de um ano inteiro, e essas variações estão associadas aos movimentos da terra de rotação e de translação, a radiação que chega e que sai de cada camada atmosférica, a latitude, a altitude e a proximidade com corpos d'água e as circulações oceânicas. [Wallace & Hobbs, 2006]

A medição da temperatura do ar é realizada com termômetros. Quando aumentam de temperatura, se dilatam e aumentam de volume, mas, quando perdem a temperatura se contraem. Para medir as variações temporais das temperaturas do ar em determinado dia, os meteorologistas utilizam termômetros de máxima, de mercúrio, para a mais elevada temperatura e termômetros de mínima,de álcool para a mais baixa. [Halliday *et al.*, 2018]

Quando a temperatura do ar aumenta, isso significa menos perda de calor, causando consequentemente um aumento temperatura da água e esta, por sua vez, afeta outras variáveis de qualidade da água [Morrill *et al.*, 2001]. O aumento da temperatura da

água agrava os problemas de poluição e afeta a saúde de muitas espécies aquáticas. [Environmental Protection Agency (EPA), N/A], [International Institute for Sustainable Development (IISD), 2018]. No trabalho [Guzzo *et al.*, 2017] é demonstrado como as mudanças na temperatura do ar a longo prazo estão reduzindo as taxas de crescimento e o tamanho dos peixes de água fria, como a truta do lago. Já em [Ozaki *et al.*, 2003] o aumento da temperatura do ar resultou no aumento da demanda biológica de oxigênio e de sólidos suspensos, e na diminuição do oxigênio dissolvido em quase todos os rios.

Além disso, o aquecimento das temperaturas do ar tem sido associado à intensificação do ciclo hidrológico, por exemplo, os níveis de precipitação, que por sua vez afetam o escorrimento da água superficial e mudanças na hidrologia que influenciam a capacidade térmica e o balanço energético dos corpos d'água. [Paul *et al.*, 2019] As chuvas intensas e tempestades podem resultar em transporte sedimentos, poluentes e nutrientes que contribuem diretamente valores dos parâmetros como sólidos suspensos, turbidez e cor.

2.5.2 Temperatura da Água

A temperatura da água, assim como descrito na seção da temperatura do ar, é uma medida física da energia térmica das moléculas em um corpo de água, em outras palavras, indica quanto quente ou fria a água está. Para a medição também é usado termômetros que são colocados dentro da água geralmente medidos em graus Celsius (C) ou Fahrenheit (F). Muitos parâmetros precisam de medição paralela com a temperatura da água: pH, condutividade elétrica, DBO e oxigênio dissolvido. [Halliday *et al.*, 2018]

Pequenas flutuações da temperatura da água podem afetar a densidade, o oxigênio dissolvido, viscosidade, e o ecossistema aquático, desta forma é bastante importante de ser monitorada. A temperatura da água pode variar devio a radiação solar ou ação humana como despejos industriais e águas de refriamento de máquinas. [Dutra, 2014]. O aumento da temperatura natural das águas reduz a solubilidade dos gases como oxigênio e dióxido de carbono. A falta do oxigênio afeta a respiração dos organismos e acelera a decomposição da matéria orgânica por microorganismos, com a possibilidade de proliferação acelerada de algas e macrófitas aquáticas. [Esteves, 1998]

Os organismos aquáticos possuem um intervalo de temperatura ideal para sua alimentação, migração crescimento e reprodução que variam nas fases adultas e jovem. Por exemplo, as tilápias resistem a temperaturas acima de 35C, mas não resistem a exposição prolongada em temperaturas abaixo de 10C. Já as trutas vivem em águas mais frias, sendo o ideal para essa espécie temperaturas entre 10 e 20^{o}C. De modo geral, poucas espécies resistem a altas temperaturas (acima de 35C), pois nesse nível há a diminuição dos OD no meio, e aumento da taxa respiratória que afetam o metabolismo dos peixeis e diminuem a afinidade da hemoglobina. Em baixas temperaturas, também há uma mortalidade dos

peixes provocados pelo à diminuição da produção do muco protetor da pele, facilitando o ataque de parasitas. Por fim, variações bruscas na temperatura são extressantes para os peixes, com possibilidade de morte já que por serem organismos de sangue frio não têm a capacidade de regular a temperatura do corpo e necessitam de um tempo de adaptação quando há alterações na temperatura do ambiente.

Além disso, segundo [Environmental Protection Agency (EPA), N/A] o aumento da temperatura da água pode resultar no aumento de patógenos e espécies invasoras, o aumento nas concentrações de poluentes, como amônia e pentaclorofenol, devido à sua resposta química a temperaturas mais altas, aumento de taxas de evapotranspiração dos corpos d'água. (Maurrem Ramon Vieira) Acrescenta também diminuição da densidade e da viscosidade da água, para temperaturas acima de 4°C, facilitando a sedimentação de materiais em suspensão, evasão de substâncias orgânicas voláteis podendo causar maus odores, evasão de gases tóxicos como H_2S e coagulação de proteínas que constituem a matéria viva.

2.5.3 Cor

A cor da água é medida em unidades de cor (uC) e é causada principalmente por sólidos dissolvidos. A decomposição da matéria orgânica, como ácidos húmicos e fúlvicos, ferro, manganês e rejeitos industriais (tinturarias, tecelagem, produção de papel), contribui para essa coloração. Embora não ofereça riscos diretos à saúde, a cor da água levanta dúvidas sobre sua qualidade. A coloração resulta da redução da intensidade da luz devido à presença de sólidos dissolvidos, especialmente coloides orgânicos e inorgânicos. [Sperling, 2002], [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

Medida de coloração na água, medida em uC (unidades de cor) com forma de constituinte principal sólidos dissolvidos origem de decomposição da matéria organica principalemente vegetais, ácidos húmicos e fúlvicos e ferro e manganês. ou residuos industriais (tinturarias, tecelagem, produção de papel). Não possui risco direto a saude mas consumidores questionam confiabilidade. alem disso a coloração da água com matéria orgânica dissolvida responsável pela cor pode gerar produtos potencialmente cancerigenos [Sperling, 2002]

A cor de uma amostra de água está associada ao grau de redução de intensidade que a luz sofre ao atravessá-la (e esta redução dá-se por absorção de parte da radiação eletromagnética), devido à presença de sólidos dissolvidos, principalmente material em estado coloidal orgânico e inorgânico. Dentre os coloides orgânicos, podem ser mencionados os ácidos húmico e fúlvico, substâncias naturais resultantes da decomposição parcial de compostos orgânicos presentes em folhas, dentre outros substratos. Os esgotos domésticos se caracterizam por apresentarem predominantemente matéria orgânica em estado coloidal, além de diversos efluentes industriais, que contêm taninos (efluentes de curtumes, por

exemplo), anilinas (efluentes de indústrias têxteis, indústrias de pigmentos etc.), lignina e celulose (efluentes de indústrias de celulose e papel, da madeira etc.). [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

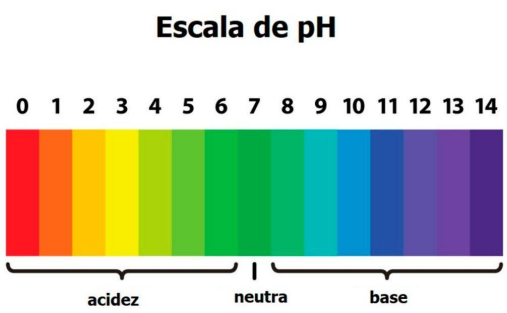
Há também compostos inorgânicos capazes de causar cor na água. Os principais são os óxidos de ferro e manganês, que são abundantes em diversos tipos de solo. Alguns outros metais presentes em efluentes industriais conferem-lhes cor, mas, em geral, íons dissolvidos pouco ou quase nada interferem na passagem da luz. O problema maior de cor na água é, em geral, o estético, já que causa um efeito repulsivo na população. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

Além dos fatores mencionados, é importante destacar que a coloração da água, embora não seja diretamente nociva à saúde humana, pode gerar subprodutos potencialmente perigosos quando submetidos a tratamentos elétricos de potabilização, como a cloração. Substâncias orgânicas dissolvidas, como ácidos húmicos e fúlvicos, podem reagir com o cloro, formando compostos organoclorados, como trihalometanos (THMs), que são conhecidos por seu potencial carcinogênico [Von Sperling, 2007]. Desta forma, a avaliação da cor da água transcende uma questão meramente estética, sendo também uma preocupação relevante do ponto de vista da saúde pública.

2.5.4 Potencial Hidrogeniônico

O potencial hidrogêniônico, comumente conhecido como pH, foi proposto por Soren Peder Lauritz Sorensen em 1909 para determinar concentrações muito pequenas de íons hidrogênio em soluções aquosas, utilizando a função logarítmica $pH = -log[H^+]$. Mais tarde, a definição foi atualizada por Sorensen em colaboração com Linderstrom-Lang, na atividade dos íons hidrogênio $pH = -log(a_{H^+}) = -\log\left(\frac{m_H \gamma_H}{m^\circ}\right)$ onde a_H é a atividade relativa (em base de molaridade), γ_H é o coeficiente de atividade molar do íon hidrogênio H^+ na molaridade m_H, e m° é a molaridade padrão. [Andrade, 2018]

Figura 2.8 – Escala pH.



Fonte: [Normas ABNT, 2025].

Em geral, o pH é medido pelo equipamento pHmetro que determina a medida de

forma precisa e quantitativa em uma escala que vai de 0 a 14, ou também pode ser medido por métodos menos precisos como fitas medidoras de pH que mudam de cor em diferentes faixas na presença de soluções ácidas e básicas. Quando o valor de pH é 7, defini-se a solução como neutra. Já valores acima de 7 são definidos como básicos e por fim, valores inferiores a 7 são consideradospara soluções ácidas. [Sörensen & Palitzsch, 1910]

As medidas de pH podem fornecer informações sobre a qualidade da água. Geralmente, a presença de ácidos fortes na água, muitas vezes resultantes de despejos industriais contendo ácidos como ácido sulfúrico ou ácido clorídrico, pode diminuir significativamente o pH da água. A presença de substâncias alcalinas na água, como carbonatos e bicarbonatos, pode elevar o pH, tornando-a alcalina. Isso também pode ser um indicativo de poluição industrial, especialmente se os despejos contiverem produtos químicos alcalinos, como hidróxido de sódio (soda cáustica). [Dutra, 2014]

Muitos organismos aquáticos são sensíveis ao pH da água em que vivem. De acordo com a a CETESB [1] com a diminuição do pH, os peixes os peixes apresentarão uma maior freqüência respiratória, passando a abocanhar o ar na superfície; em pH extremamente baixo, têm morte imediata. Já o aumento do pH desencadeia a formação de óxido de cálcio que provoca corrosão do epitélio branquial e das nadadeiras, levando os peixes à morte.

O Ministério da Saúde determina no Art. 39, da Portaria Nº 2.914/2011 [2] que, no sistema de distribuição, o pH da água seja mantido na faixa de 6,0 a 9,5.

O monitoramento do pH em corpos d'água não apenas fornece informações sobre sua qualidade, mas também serve como um indicador crucial de condições ambientais que podem afetar ecossistemas aquáticos e a saúde humana. Valores de pH fora da faixa recomendada pelo Ministério da Saúde (6,0 a 9,5) podem indicar a presença de substâncias industriais ou domésticas, ou até mesmo desequilíbrios naturais. Por exemplo, a acidificação de rios e lagos devido à presença de fortes ácidos, como o ácido sulfúrico, pode comprometer a sobrevivência de espécies aquáticas sensíveis, enquanto um pH fraco básico, muitas vezes causado por despejos de refrigerante cáustico, pode agravar a corrosão das infraestruturas e toxicidade para os organismos vivos. [Eiben & Smith, 2015]

Além disso, o controle do pH é essencial para a eficiência dos processos de tratamento de água. Um pH inadequado pode reduzir a eficácia de etapas como a coagulação e a infecção, além de aumentar a probabilidade de formação de sub [Dutra, 2014] enfatizam que o pH também influencia a biodisponibilidade de metais pesados, como chumbo e mercúrio, ampliando os riscos de bioacumulação na cadeia alimentar. Dessa forma, a inclusão de tecnologias avançadas, como algoritmos de aprendizagem de máquina, pode melhorar a análise contínua do pH, correlacionando seus valores a outras variáveis ambientais e permitindo a identificação precoce de anomalias, garantindo a preservação ambiental e a

[1] https://cetesb.sp.gov.br/mortandade-peixes/alteracoes-fisicas-e-quimicas/ph/
[2] https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2914_12_12_2011.html

saúde pública.

2.5.5 Condutividade Elétrica

A condutividade se caracteriza pela facilidade ou dificuldade de passagem de eletricidade na água pela presença de substâncias dissolvidas, que se dissociam em compostos iônicos. [de Saint-Ange Comnène Carloni, 2014], [Estevam *et al.*, 2019], [Nascimento, 2017] Este parâmetro não especifica a quantidade e quais íons estão presentes em uma amostra, mas pode ser utilizado para reconhecimento de impacto ambiental e usado como indicador da qualidade da água [Dutra, 2014], [Piratoba *et al.*, 2017]

Informações sobre alterações que ocorram na água sobre qualquer tipo de efluente tais como esgoto, resíduos industriais, agrotóxicos podem ser identificadas pelo aumento da condutividade elétrica A medição é conduzida por meio de um equipamento conhecido como condutivímetro, que conecta uma célula a um medidor de corrente alternada em 1000 Hz. O aparelho é calibrado para um determinado valor da constante da célula utilizando uma solução de cloreto de potássio com concentração previamente estabelecida. O medidor possui um sensor de temperatura que ajusta automaticamente as medições de condutividade para a calibração do equipamento.A unidade de medida da condutividade é $\mu S/cm$ (microsiemens por centímetro). [Siqueira *et al.*, 2018]

Os compostos iônicos que permitem a passagem da eletricidade são, na maioria, sódio, magnésio, cálcio, ferro, alumínio e amônio classificados como compostos catiônicos, pois perdem elétrons na camada de valência, e os compostos ânionicos, como cloretos, sulfatos, nitratos e fosfatos que possuem elétrons livres na camada de valência. Em contrapartida, materiais orgânicos como o óleos, graxas, álcool, fenóis não possuem a capacidade de conduzir a eletricidade. [Nascimento, 2017]

De acordo [Tundisi & Tundisi, 2008] com a distibuição da vida aquática é relacionada com a concentração iônica e na colonização de ambientes com diferentes condutividades, pois essa estimula diferentes métodos de regulação e tolerância para grupos de animais e plantas aquáticas. Além disso, influencia na colonização de espécies invasoras porque elas dependem da disponibilidade dos cátions e ânions na água.

Tanto a temperatura da água quanto a precipitação afetam diretamente a condutividade elétrica. Em época de estiagem, rios e lagos evaporam a água, a temperatura é elevada e consequentemente aumento de concentração de sais na água que influenciam diretamente no aumento da condutividade. Em contra partida, os períodos de chuva diminuem a condutividade devido a diluição e ao transporte dos íons para outras áreas. [Piratoba *et al.*, 2017], [Arcos *et al.*, 2022], [Silva Júnior *et al.*, 1999]. Outro fator importante de citar é o pH, visto que, quanto mais ácido ou básico for o pH, mais íons existem. Em faixas extremas (pH$>$ 9 e pH$<$5), os valores de condutividade são devidos apenas às altas concentrações de poucos íons em solução, dentre os quais os mais frequentes são H^+ e o

OH^- [Dutra, 2014].

Além dos fatores já mencionados, a condutividade elétrica também pode ser influenciada pelas atividades humanas, como a descarga de efluentes industriais e agrícolas, que aumenta a concentração de íons distribuídos na água. Por exemplo, o uso intensivo de fertilizantes ricos em nitratos e fosfatos em áreas agrícolas pode levar ao escoamento desses compostos para corpos d'água, aumentando significativamente a condutividade.[Eiben & Smith, 2015]. Da mesma forma, despejos industriais contendo cloretos e sulfatos podem alterar o equilíbrio iônico da água, contribuindo para um aumento nos valores de condutividade elétrica. Esses impactos não apenas modificaram a qualidade da água, mas também afetaram os ecossistemas aquáticos, alterando a composição e a distribuição de espécies.

A análise da condutividade elétrica é, portanto, uma ferramenta crucial para o monitoramento ambiental, permitindo detectar alterações causadas por atividades antropogênicas e eventos naturais. Além de ser um indicador da qualidade da água, a condutividade elétrica também tem aplicações práticas em processos de tratamento de água. Por exemplo, valores elevados podem exigir ajustes nos processos de coagulação e filtração para evitar a deposição de sais nas membranas e em outros equipamentos. Nesse contexto, o uso de tecnologias avançadas, como algoritmos de aprendizagem de máquina, pode oferecer insights valiosos, correlacionando mudanças na condutividade elétrica com outras variáveis ambientais e fontes de poluição, contribuindo para uma gestão mais eficaz e sustentável dos recursos hídricos. [Kumar *et al.*, 2023]

2.5.6 Turbidez

O termo turbidez se refere transparência da água, indica presença de partículas que não se dissolvem na mesma como minerais, partículas orgânicas ou microorganismos provindos do processo natural de erosão ou despejo de efluente doméstico e industrial. Águas com alta turbidez são turvas e opacas, podendo ser um sinal de construção, mineração, agricultura ou sinal de poluição. Alta turbidez também afeta a vida aquática pois reduz a penetração da luz solar e diminui a eficiência do tratamento químico e físico da água. [Nascimento, 2017]

A medição da turbidez é em termos ópticos, com relação ao grau de interferência à passagem da luz através de um líquido. A luz é espalhada e absorvida e não transmitida em linha reta através da amostra. [Rocha, 2019] A unidade de medida é (NTU), unidades de turbidez nefelométrica, e considerada parâmetro que indica a qualidade das águas para abastecimento público (CETESB, 2016). O Ministério da Saúde determina na portaria Nº 2.914/2011 o valor máximo permitido como 0,5 unidades de turbidez para água filtrada por filtração rápida e 1,0 unidade de turbidez para água filtrada com filtração lenta, e para qualquer amostra pontual para reservatórios e rede no sistema de distribuição, 5,0

unidades de turbidez.

Figura 2.9 – Ilustração da Turbidez.

Fonte: [Digital Water, 2024].

Penedo *et al.* [2023] indicaram em seu estudo que a precipitação possui forte influência na turbidez, pois as chuvas acarretam aumento de velocidade e volume de água nos corpos hídricos que afetam a vazão, desagregando e carregando partículas de argila, silte, areia, fragmentos de rocha e óxidos metálicos do solo para água. A maioria rios brasileiros possuem naturalmente águas turvas em decorrência de de altos índices pluviométricos em acréscimo ao uso agrícula inadequado e lançamento de esgotos domésticos. [Ministério da Saúde, 2006]

Em Marcó *et al.* [2004] os autores abordam como a turbidez pode dificultar os processos de desinfecção,e consequentemente aumentam a proliferação bacteriana. Além disso, metais pesados e produtos químicos podem se ligar nas partículas da água, bem como produtos químicos e radioativos, pois as partículas em suspensão podem transportar esses materiais. Já

Desta forma fica evidente a importância dos processos de tratamento da água. Para realizar essa tarefa, é necessário um conjunto de operações e processos a fim de remover o máximo a turbidez. Geralmente, o processo é iniciado com a coagulação que neutraliza as cargas elétricas nas partículas suspensas, facilitando sua aglomeração e remoção. O vazamento é a última barreira para a remoção de partículas em suspensão. Finalmente, um processo de desinfecção é realizado destruindo os microrganismos patogênicos ainda presentes na água. [Martínez-Orjuela *et al.*, 2020]

2.5.7 Oxigênio Dissolvido

O oxigênio dissolvido (OD) é a quantidade de oxigênio molecular O_2 presente na água provindo da atmosfera ou como subproduto da fotossíntese realizada pelas plantas aquáticas. É um dos parâmetros mais essenciais para investigação do ambiente aquático visto que os organismos aeróbicos precisam do oxigênio para sobreviver e quando ausente

provoca a liberação de substâncias que geram odor, sabor e aspecto indesejável por organismos anaeróbicos.[Wetzel, 2000]

Em acréscimo, o OD pode ser indicador de poluição na água por esgoto doméstico ou despejo industrial. Quando os resíduos são jogados no meio aquático, os microorganismos como bactérias e fungos realizam a digestão da matéria orgânica, respirando o oxigênio e competindo com os demais organismos. Desta forma, os níves de OD se reduzem ou chegam até a zero, causando a morte dos peixes. [Braga, 2005], [Von Sperling, 2007]

A medição de OD pode ser realizada por diferentes técnicas a depender da precisão almejada, a frequência e disponibilidade dos equipamentos. O método tradicional é o de Winkler que consiste em adicionar uma solução alcalina de sulfato manganoso e de iodeto alcalino de potássio em uma amostra de água para a formação do hidróxido de manganês. Na mistura, se a cor do precipitado é marrom, indica que o processo de oxidação ocorreu, sendo possível mensurar a concentração de OD equivalente a concentração de iodo presente na amostra. Caso o precipitado permaneça na cor branca, a concentração de OD é ausente. [Vargas *et al.*, 2007], [JÚNIOR *et al.*, 2019]

Outro método para determinar de OD é na medição de corrente eétrica pela redução eletroquímica de oxigênio $O_2 \rightarrow OH^-$ da amostra por meio de um equipamento denominado oxímetro. É baseado na célula de Clark (proposto por Leland Clark) composta por dois eletrodos (cátodo e ânodo) imersos em um eletrólito com uma membrana que separa ambos. Na polarização dos eletrodos, a concentração de oxigênio é proporcional pela corrente elétrica que passa entre eles. [CPRM – Serviço Geológico do Brasil, Diretoria de Hidrologia e Gestão Territorial – DHT, Superintendência Regional de Belo Horizonte – SUREG-BH, Gerência de Hidrologia e Gestão Territorial – GEHITE, Laboratório de Sedimentometria e Qualidade das Águas – LSQA, 2007],[JÚNIOR *et al.*, 2019], [Mendonça *et al.*, 2020], [Mendes *et al.*, 2021]

Por fim, existem os métodos óticos com sensores de fibra óptica de oxigênio no qual algumas substâncias que reagem com a amostra desenvolvem alguma alteração na luminescência. O composto quando é excitado, emite uma luz luminescente vermelha com intensidade ou tempo de vida como respota a concetração de OD presente. Essa abordagem não requer uso de eletrodos de referência, não consome oxigênio e apresenta precisão similar ao método de Winkler. [Ferreira, 2007], [Pereira, 2016], [Mendes *et al.*, 2021]

Níveis de OD são afetados pela tempratura da água, força iônica e sólidos dissolvidos porque a solubilidade do oxigênio diminui a medida que esses parâmetros aumentam e assim, a diminuição de OD na água. [Wetzel, 2000] Além disso, mudanças no fluxo de um canal em um rio como a diminuição da sua velocidade e o aumento da profundidade alteram a concetração de OD. Águas rápidas trocam mais oxigênio do ar com a água que aumentam os níveis de OD enquanto em velocidades mais baixas essa troca é reduzida.

Aumentos da profundidade podem levar acúmulo de matéria orgânica no fundo do rio onde as bactérias consumiram oxigênio. [EPA United States Environmental Protection, 2024]

Pela resolução da CONAMA 357/2005 [3] o valor para preservação aquática é de 5,0 mg/L, porém há espécies que conseguem aguentar até 3,0 mg/L, como exemplo a carpa, e outras necessitam de maiores valores como as trutas em torno de 8,0 mg/L. O valor crítico ocorro com valores de OD menores que 2mg/L apresentam condições perigosas onde ocorre o fenômeno da hipoxia (quando as células não conseguem respirar). [CETESB, 2024a]

2.5.8 Demanda Bioquímica de Oxigênio

Na seção anterior foi visto como a poluição em meio aquático pode contribuir para a diminuição do OD presente na água, pois os microorganismos consomem oxigênio durante a estabilização da matéria orgânica. O objetivo então é inferir o quanto de força de um potencial poluidor de um efluente causaria na medição do consumo de oxigênio, ou seja uma quantificação indireta do potencial de impacto e não na medição direta do impacto em si. [Von Sperling, 2007]

Desta forma definimos a Demanda Bioquímica de Oxigênio (DBO) como a quantidade de oxigênio dissolvida na água necessária para a decomposição da matéria orgânica. A DBO serve como forma de conhecimento para determinar o impacto do despejo de esgoto na concentração de oxigênio do corpo de água receptor. Altos valores de matéria orgânica podem induzir esgotamento total do oxigênio na água, interfirindo no equilíbrio da vida aquática. [Braga, 2005]

A determinação da DBO é realizada com determinada quantidade de amostra de esgoto, colocando-a em um frasco com água de diluição e alguns nutriente. Da solução inicial determina-se a concentração inicial de OD. O frasco é então colocado a uma incubadora a uma temperatura de 20 oC. e sem presença de luz. Após determinado período de tempo mede-se novamente o OD e realiza a diferença da concentração inicial e final de OD que corresponde a DBO. Por convenção as medidas de DBO são realizadas por ensaios que duram cinco dias, já que a estabilização completa da matéria orgânica exige um tempo maior. [KAMIYAMA, 1988]

Os valores da DBO variam a depender na população local, como consumo de água, condições socioeconômicas da comunidade, hábitos de higiene e condições do clima, pois temperaturas elevadas geram maior consumo de água interferindo na DBO. [Silva, 2022] Em esgotos domésticos o valor fica em torno de 300 mg/L [Braga, 2005], indústrias alimentícias possuem geralmente valores mais elevados da DBO devido à alta concentração de matéria orgânica biodegradável nos seus efluentes.

De acordo com a Resolução 357/2005 da CONAMA [Conselho Nacional do Meio

[3] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

Ambiente, 2005] [4] é permitido o valor de até 3mg/L O_2 de DBO para águas de classe 1, 5mg/L O_2 para classe 2 e 10mg/L O_2 para classe 3. A classe 1 corresponde ao consumo humano e contato primário, a classe 2 para pesca e contato secundário, e classe 3 para navegação e paisagem. E na Resolução 430/2011 da CONAMA estabelece no Art. 21 "máximo de 120 mg/L, sendo que este limite somente poderá ser ultrapassado no caso de efluente de sistema de tratamento com eficiência de remoção mínima de 60% de DBO, ou mediante estudo de autodepuração do corpo hídrico que comprove atendimento às metas do enquadramento do corpo receptor"

2.5.9 Demanda Química de Oxigênio

A demanda química de oxigênio (DQO) mede o quanto de oxigênio é consumido para oxidar a matéria orgânica presente na água por meio de reações químicas sem o uso de microorganismos como ocorre na medida da DBO. Além disso, a DQO é um método mais rápido, geralmente com duração de um pouco mais de duas horas. Sperling [2002]

A Demanda Química de Oxigênio (DQO) é amplamente utilizada como um parâmetro essencial para avaliar a qualidade da água, especialmente em corpos hídricos que recebem contribuições significativas de efluentes industriais e domésticos. Ao medir a quantidade de oxigênio necessária para oxidar quimicamente a matéria orgânica e inorgânica presente, o DQO fornece uma estimativa abrangente do nível de poluição em um dado corpo d'água. Diferente da Demanda Bioquímica de Oxigênio (DBO), que depende de processos biológicos realizados por microorganismos, a DQO utiliza agentes químicos oxidantes, como o dicromato de potássio, em meio ácido. Essa característica permite que a análise seja concluída em um tempo limitado, geralmente em cerca de duas a três horas, o que torna o método especialmente útil para aplicações que impedem a rapidez na obtenção de resultados. CETESB - Companhia Ambiental do Estado de São Paulo [2018], Companhia Ambiental do Estado de São Paulo (CETESB) [2016]

Além de sua eficiência temporal, o DQO possui a vantagem de ser um método mais abrangente, pois oxida uma gama maior de compostos, incluindo substâncias que não são facilmente degradadas por microorganismos. Isso torna o DQO particularmente relevante para o monitoramento de águas contaminadas por compostos orgânicos recalcitrantes, como hidrocarbonetos e alguns tipos de pesticidas. No entanto, o DQO também apresenta limitações, como a impossibilidade de distinguir entre fontes de poluição orgânica e inorgânica. CARMO [2021] Por esse motivo, a interpretação do DQO em conjunto com outros parâmetros, como DBO e sólidos distribuídos totais, é crucial para fornecer uma avaliação mais precisa da qualidade da água. O uso de tecnologias analíticas avançadas, combinadas com ferramentas de aprendizagem de máquina, pode potencializar a análise

[4] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

dos dados de DQO, permitindo correlacionar seus valores com fontes específicas de poluição e prever impactos ambientais, contribuindo para uma gestão hídrica mais eficaz.

2.5.10 Sólidos totais, Sólidos suspensos totais e Sólidos dissolvidos totais

Qualquer substância que permanece como resíduo após os processos de evaporação, secagem ou calcinação da amostra a uma determinada temperatura é considerada como sólida. Esses resíduos podem ser classificados como sólidos totais (ST), englobando sólidos em suspensão, dissolvidos, fixos e voláteis. Eles incluem materiais orgânicos e inorgânicos, como matéria vegetal e animal em decomposição, resíduos industriais e de esgoto doméstico, microplásticos, entre outros. [Ramos, 2021], [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

Os sólidos dissolvidos totais (SDT) são os que dissolvem na água como sais minerais, metais e outros compostos químicos. Já os sólidos suspensos totais (SST)são os que não dissolvem, como argila, areia, matéria orgânica e inorgânica. A divisão entre os dois é através de um filtro de tamanho específico, as partículas que passam através do filtro são classificadas como SDT e as de maiores dimensões retidas pelo filtro são classificadas como SST. [Von Sperling, 2007] Os SST contribuem para o transporte de bactérias, nutrientes, pesticidas e metais, uma vez que os aglomerados de partículas facilitam o transporte de poluentes e microrganismos, além de contribuir para aumento da turbidez. Já os SDT podem contribuir para aumento da demanda química e bioquímica de oxigênio nas águas causando baixa de oxigênio dissolvido e afetando a vida aquática. [Barreto *et al.*, 2014], [Ramos, 2021]

A resolução CONAMA 357/2005 [Conselho Nacional do Meio Ambiente, 2005] [5] determina que o padrão aceitável para as águas de classe 1,2 e 3 são resíduos sólidos objetáveis: virtualmente ausentes e para sólidos dissolvidos totais 500 mg/L.

Qualquer alteração nos níveis de sólidos distribuídos totais (SDT) e sólidos suspensos totais (SST) pode indicar efeitos significativos na qualidade da água. O aumento nos níveis de SST, por exemplo, pode ser consequência de atividades humanas, como a construção civil, a mineração e o manejo inadequado do solo em áreas agrícolas, que intensificam o carregamento de partículas para os corpos hídricos. Esse processo pode causar o assoreamento de rios e lagos, além de dificuldades a penetração da luz solar, o que interfere diretamente na fotossíntese e, consequentemente, em nossos ecossistemas aquáticos. Além disso, altos níveis de SST estão associados ao aumento da turbidez, dificultando a captação e o tratamento da água para consumo humano. [Adjovu *et al.*, 2023]

Os SDT, por sua vez, são particularmente preocupantes em regiões onde há desperdício de efluentes industriais ou domésticos sem o devido tratamento. A presença

[5] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

elevada de compostos químicos dissolvidos, como nitratos, fosfatos e metais pesados, pode causar eutrofização, resultando em regularidade de algas e conseqüentemente irrelevantes na qualidade da água [Smith, 2003]. Além disso, águas com alta concentração de SDT são menos potáveis e podem representar riscos à saúde humana, principalmente quando contêm metais tóxicos como chumbo e mercúrio [Zamora-Ledezma *et al.*, 2021]. Monitorar e controlar os níveis de SDT e SST, portanto, é essencial para garantir a preservação dos corpos hídricos e o equilíbrio dos ecossistemas aquáticos, conforme orientado pelas normas reguladoras, como a Resolução CONAMA 357/2005. [6]

2.5.11 Alcalinidade

A alcalinidade é uma medida do conjunto total de substâncias presentes na água que são capazes de neutralizar os ácidos, como os bicarbonatos, carbonatos e hidróxidos, sendo expressa em mg/L de CaCO3. Essa propriedade está diretamente relacionada ao pH, pois a capacidade de uma solução de resistir a variações de acidez depende de sua alcalinidade. Em uma solução com baixa alcalinidade, a adição de ácidos pode reduzir significativamente o pH. Em contrapartida, uma solução com alta alcalinidade apresenta maior capacidade de neutralização de ácidos, diminuindo as variações de pH mesmo diante da presença de ácidos fracos. [Companhia de Desenvolvimento e Ação Regional (CAR), 2019]

A concentração total de alcalinidade pode ser medida através de volumetria, permitindo assim a quantificação da soma da alcalinidade originada por todos os íons presentes em uma amostra de água, sendo expressa em mg/L de $CaCO_3$. Na maioria das vezes, a presença de bicarbonatos é responsável pela alcalinidade da água devido à ação do dióxido de carbono dissolvido em águas que entraram em contato com rochas calcárias (calcita = $CaCO_3$). [de Holanda Cavalcante, 2012]

A presença de alcalinidade, embora não seja um parâmetro de potabilidade, pode influenciar diretamente o sabor da água, sendo que concentrações podem levar à contaminação da água pelos consumidores. Apesar de não ser utilizada para classificação de águas naturais ou para tratamento de esgoto, a alcalinidade é um parâmetro essencial em processos específicos em estações de tratamento de água potável e águas residuais. Ela auxilia no controle da corrosividade da água e na otimização de processos como a coagulação, uma vez que um nível inadequado de alcalinidade pode interferir na eficiência dos produtos químicos usados no tratamento. [Rojas & Rocha, 2009]

Além disso, a alcalinidade desempenha um papel crítico no equilíbrio ecológico dos corpos d'água. Em ecossistemas aquáticos, ela atua como um amortecedor, estabilizando o pH e protegendo os aquários de flutuações extremas que poderiam ser acessíveis à vida. Por

[6] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

esse motivo, o monitoramento da alcalinidade é relevante em programas de conservação ambiental e na avaliação de impactos causados por gastos de efluentes industriais ou agrícolas que alteram o equilíbrio químico dos cursos d'água. [de Holanda Cavalcante, 2012]

2.5.12 Dureza

A dureza da água é a concentração de íons de cálcio (Ca^{2+}) e magnésio (Mg^{2+}) em sua maior forma, e outros metais como ferro (Fe^{2+})/(Fe^{2+}), manganês (Mn^{2+}), estrôncio (Sr^{2+}), zinco (Zn^{2+}) e alumínio (Al^{3+}) em menores quantidades. Assim, a dureza indica a concentração de cátions metálicos presentes na água. Os principais compostos que tornam dureza às águas são bicarbonato de cálcio, bicarbonato de magnésio, sulfato de cálcio e sulfato de magnésio. [Nolasco *et al.*, 2020]

A origem da dureza pode ser natural, provinda da dissolução de rochas calcárias ou outros minerais contendo cálcio e magnésio, bem como ação antropológica com despejos industriais.A dureza é conhecida por impedir a formação de espumas, como a do sabão, devido à formação de compostos insolúveis com os íons de cálcio e magnésio. [Nolasco *et al.*, 2020]

Tanto o cálcio quanto o magnésio são importantes para os seres humanos e para as espécies aquáticas. No corpo humano, o cálcio é fundamental para a estrutura óssea, a circulação coronária e várias funções celulares. Para os organismos aquáticos, esses minerais são igualmente importantes, por exemplo, para o fitoplâncton, o cálcio é essencial aos animais pois participa de vários processos biológicos, tais como construção óssea, coagulação sanguínea, e muitas outras funções celulares. No entanto, a presença de dureza pode causar sabor desagradável e pode ter efeitos laxativos bem como pode provocar depósitos em encanamentos. [de Holanda Cavalcante, 2012], [Companhia de Desenvolvimento e Ação Regional, 2019]

A dureza da água é medida geralmente em mg/L de carbonato de cálcio (mg/L de CaCO3). A Portaria 2.914 do Ministério da Saúde sobre Potabilidade da Água admite um valor de dureza (até 500 mg/L de CaCO3).

A presença de dureza na água apresenta implicações ambientais e econômicas. Por exemplo, águas duras são menos eficazes na interação com sabões e detergentes, levando a um maior consumo desses produtos e, consequentemente, a um aumento na produção de efluentes consumidores [Gotoh *et al.*, 2016]. Além disso, em sistemas industriais e de distribuição de água, a dureza pode ocasionar incrustações em tubulações, aquecedores e caldeiras, reduzindo a eficiência energética e aumentando os custos de manutenção [Contreras-Ramírez & Nieves-Rivas, 2023]. Por outro lado, em ecossistemas aquáticos, níveis adequados de cálcio e magnésio são essenciais para o desenvolvimento de organismos aquáticos, incluindo moluscos, crustáceos e algas, que dependem desses minerais para

processos como a formação de conchas e esqueletos. Desta forma, o controle e monitoramento da dureza são fundamentais não apenas para garantir a qualidade da água potável, mas também para preservar o equilíbrio ambiental e melhorar os recursos em sistemas industriais e domésticos. [Oliveira-Filho *et al.*, 2014]

2.5.13 Fósforo e Ortofosfato

O fósforo é um nutriente imprencindível para os seres vivos como as plantas e animais pois participa dos processos de metabolismo dos seres vivos, no material genético, transferencia de energia e suporte estrutural dos organismos provindos de membranas e ossos. [Ruttenberg, 2003]

No entanto, a grande quantidade dessa substância pode causar o processo de eutrofização, processo de proliferação desenfreada de algas, que limita o oxigênio afetando a vida das espécies aquáticas. [Klein & Agne, 2013], [Parron *et al.*, 2011] Altas concentrações de fósforo também são consideradas fatais para os seres humanos na faixa de valores superiores a concentração de 100mg. [da Mata Ribeiro, 2006]

A presença de fósforo na água provém de processos naturais de rochas e solo e decomposição da matéria orgânica bem como descarga de esgostos,efluentes industriais, detergentes, fertilizantes e pesticidas. Relacionado a forma dessa substância, é encontrado de maneira orgânica, inorgânica e a forma mais comum são os fosfatos solúveis, classificados como ortofosfatos. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016] [Parron *et al.*, 2011],

A medida do fósforo total é em mg/L e utilizada como um dos parâmetros para calcular o IQA. A resolução do CONAMA 357/2005 [7] estabelece como limite o valor de 0,020 – 0,025 mg/L para águas de classe 1, 0,030 – 0,050 mg/L para classe 2 e 0,050 – 0,075 mg/L e 0,050 – 0,075 para classe 3.

O fósforo desempenha um papel essencial nos ecossistemas aquáticos, sendo um dos principais nutrientes responsáveis pelo crescimento de organismos fotossintetizantes, como algas e plantas aquáticas. No entanto, a presença de fósforo em excesso, muitas vezes proveniente de atividades humanas, como o uso intensivo de fertilizantes e o descarte inadequado de efluentes domésticos e industriais, pode desestabilizar os ecossistemas[Esteves, 1998] . A eutrofização, aparentemente causada pelo aumento excessivo de nutrientes, resulta em um crescimento descontrolado de algas, que, ao morrerem e se decomporem, consomem o oxigênio distribuído na água. Isso leva à formação de zonas hipóxicas ou anóxicas, afetando qualidades de biodiversidade aquática e comprometendo a qualidade da água para uso humano e animal.[Smith, 2003]

[7] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

2.5.14 Ferro

O ferro está presente na formas insolúvel Fe^{+3} em uma grande quantidade de tipos de solos. Já na água subterrânea, pode se apresentar na forma solúvel Fe^{+2}. Quando exposto ao oxigênio o ferro pode ser oxidado, foramando precipitados e conferir cor à água, variando de tons amarelos a marrons, dependendo das condições específicas do ambiente e da concentração de ferro presente. Em determinadas quantidades também pode influenciar no sabor e no cheiro. [Companhia de Desenvolvimento e Ação Regional, 2019]

A reação do minério com o dióxido de carbono na água produz carbonato ferroso solúvel, que é comum em poços com altos teores de ferro. Este é o motivo pelo qual o ferro é frequentemente encontrado nas águas subterrâneas. Nas águas superficiais, o nível de ferro aumenta nas estações chuvosas devido ao carreamento de solos e a ocorrência de processos de erosão das margens, ou por contribuição industrial, devido ao fato de várias indústrias metalúrgicas removerem ferrugem de peças usando banho ácido. [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

O ferro não é considerado um tóxico, mas causa vários problemas a saúde como anemia, anorexia, sensibilidade óssea, prisão de ventre, distúrbios digestivos, tontura, fadiga, problemas de crescimento, irritabilidade e inflamação da língua em sua carência no organismo humano. O excesso de nutrientes pode causar anorexia, tontura, cansaço e dores de cabeça. Para o sistema de abastecimento público de água, adicionando sabor e cor à água, como citado, causando manchas em roupas e utensílios de limpeza. [Companhia de Desenvolvimento e Ação Regional, 2019]

O Padrão de portabilidade é de concentração limite de 0,3 mg/L na Portaria nº 2914/11 do Ministério da Saúde. É também padrão de emissão de esgotos e de classificação das águas naturais. [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

2.5.15 Cloreto

As águas naturais possuem ions resultantes da dissolução de minerais, como por exemplo, o cloreto de sódio ($NaCl$). Os cloretos se referem a à determinação da concentração de íons cloreto (Cl-) presentes na água [Sperling, 2002]. Os esgotos sanitários são as principais fontes de cloreto, dado que cada pessoa escreta cerca de 4g de cloreto por dia através da urina, o que resulta em concentrações maiores de 15mg/L no escoamento de efluentes.[CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

Valores altos de concentração de cloreto podem conter efeito laxativo e sabor "salgado"a depender do cátion associado. Por exemplo, se o cátion for o sódio (Na+) o sabor pode ser perceptível com águas contendo 250 mg/L Cl-. Caso seja cálcio ou magnésio, o sabor pode ser detectado apenas com concentrações acima de 1000 mg/L. Como o cloreto de sódio é presente na alimentação humana, a sua presença em concetração elevada pode ser indicativa de poluição por efluentes domésticos ou industriais. [ADAM &

de OLIVEIRA, 2024]

A regulamentação para regras nas águas superficiais no Brasil é imposta pela Resolução CONAMA nº 357/2005 [8] , que fixa o limite máximo de 250 mg/L para as classes 1, 2 e 3 de águas doces. Este valor visa proteger tanto a saúde pública quanto os ecossistemas aquáticos, considerando o uso múltiplo das águas, como abastecimento humano, segurança e biodiversidade. No entanto, em situações de poluição por efluentes domésticos e industriais, os níveis de cloreto podem exceder esse limite, comprometendo a qualidade da água e exigindo disposições para o controle de fontes de contaminação.

2.5.16 Fenóis

Um fenol é uma função orgânica do tipo ROH, onde R é um grupo benzênico e OH é um ou mais grupos hidroxila ligados a um anel aromático. [de Morais Cavalcante, 2016] Em efluentes industriais, podem ser encontrados mais de um tipo de poluente fenólico, pela produção de plásticos, tintas, fármacos, pesticidas, antioxidantes, papel e também nas indústrias do aço e petroquímica. [da Silva Castro, 2013]

Os fenóis são tóxicos para o seres humanos, aos organismos aquáticos e microorganismos e podem trazer problemas para o meio ambiente por reatividade e dificuldade de biodegradação. Um valor de 40 mg/L inibe a nitrificação, e concentrações em torno de 50 a 200 mg/L trazem inibição da atividade microbiana. Valores de 100 a 200 mg/L de fenóis também provocam inibição na digestão anaeróbia. Em acréscimo, a presença de fenóis altera o sabor e odor da água. Por essas razões legislação restringe fortemente a presença de fenóis. [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

De acordo com Artigo 19-A do Decreto Estadual n.º 8.468/76, 0,001 mg/l para águas de classe 1 e 2 e até 1 mg/l para classe 3. O índice de fenóis constitui também padrão de emissão de esgotos diretamente no corpo receptor, sendo estipulado o limite de 0,5 mg/L pela Legislação Federal (Artigo 16 da Resolução n.º 430/11 do CONAMA). [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]. [9]

Os fenóis são amplamente utilizados na indústria e, consequentemente, representam uma das principais classes de poluentes encontradas em corpos d'água. Sua presença nos efluentes está relacionada às atividades como fabricação de resinas sintéticas, adesivos, combustíveis, e até mesmo tratamentos de madeira [Cunha & de Aguiar, 2014]. Devido à sua toxicidade e baixa biodegradabilidade, os fenóis podem se acumular no ambiente, comprometendo a qualidade das águas e afetando ecossistemas aquáticos [Costa *et al.*, 2008]. Além disso, sua reatividade química contribui para reações que formam compostos ainda mais tóxicos, ampliando os impactos ambientais e sanitários. [CETESB, 2024b]

[8] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

[9] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

Do ponto de vista ambiental, os fenóis alteram significativamente a qualidade da água. Pequenas concentrações são suficientes para provocar alterações no sabor e no odor, tornando-a imprópria para o consumo humano. Nos ecossistemas aquáticos, os fenóis prejudicam o equilíbrio microbiológico, inibindo a nitrificação um processo essencial no ciclo do nitrogênio e dificultando a digestão anaeróbia em sistemas de tratamento de esgoto. A toxicidade para os organismos aquáticos também exige cadeias alimentares e de biodiversidade, destacando a necessidade de um controle rigoroso sobre a liberação desses compostos no ambiente. [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

A regulamentação de fenóis na água é severa devido aos seus conflitos. O Decreto Estadual n. 8.468/76 estabelece limites rigorosos, como 0,001 mg/L para águas de classe 1 e 2, e até 1 mg/L para classe 3, minimizando os riscos à saúde humana e ao meio ambiente. No caso de efluentes industriais, a Resolução CONAMA n. 430/11 estipula um limite máximo de 0,5 mg/L para emissão direta em corpos receptores. Essas regulamentações buscam não apenas proteger os recursos hídricos, mas também garantir que as atividades econômicas sejam realizadas de forma sustentável e com o menor impacto ambiental possível. A adoção de tecnologias de tratamento e monitoramento contínuo dos níveis de fenóis é fundamental para atender aos padrões estabelecidos e preservar a qualidade da água. [CETESB - Companhia Ambiental do Estado de São Paulo, 2018]

2.5.17 Cromo

O cromo é um metal caracterizado por sua dureza e alta resistência a corrosão, muito utilizado em ligas metálicas como no aço inoxidável. É presente na forma de oxidação cromo trivalente (Cr (III)) que é essencial para o metabolismo humano, pois regula o metabolismo de glicose, proteínas e gorduras e o cromo hexavalente (Cr (VI)), que é tóxico e pode causar vários tipos de câncer e danos ao DNA. A US Food Nutrition [Board *et al.*, 2000] recomenda a ingestão de 35 mg/dia para adultos masculinos e 25mg/dia para o sexo feminino.[Mataveli *et al.*, 2018], [Greenwood & Earnshaw, 1997]

Sua presença na água provém de ações naturais por lixiviação ou antropogênicas pelas descargas de efluentes complexos e substâncias químicas em bacias de drenagem. Na vida aquática, a toxicidade varia com a espécie, temperatura, pH, valência, OD. Causa problemas de corrosão das mucosas problemas respiratórios e modificações hematológicas. Uma concentração de 0,05 mg/L já acarreta morte nos peixes, sendo a recomendação de cromo ser abaixo de 0,03 mg/L. [Sampaio, 2003]

De acordo com a resolução CONAMA 357/2005 [Conselho Nacional do Meio Ambiente, 2005] [10] , os valores máximo permitidos são 0,05 mg/L Cr para águas de classe 1,2 e 3 águas doces, e 1,1 mg/L Cr para classe 2 água salinas e salobras. O valor de

[10] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

lançamento de efluentes é também de 0,5 mg/L.

2.5.18 Zinco

O zinco é um metal abundante na terra, utilizado na indústria automobilística, construção civil, e fabricação de eletrodomésticos. Esse metal desempenha um papel importante na produção de ligas metálicas resistentes à corrosão e galvanização de produtos de ferro e aço. Por essa razão, é encontrado nos efluentes industriais bem como nos produtos descartados como pilhas, baterias e eletrônicos. [Companhia Ambiental do Estado de São Paulo (CETESB), 2013b], [Gianotti, 1986]

Esse elemento é necessário para o organismo humano em pequenas quantidades e também para o crescimento de plantas e animais por participar na divisão celular expressão genética, processos fisiológicos como crescimento e desenvolvimento. Porém para os peixes e organismos aquáticos pode provocar mudanças na morfologia e fisiologia, obstrução das guelras, crescimento e maturação retardada e morte. [Sampaio, 2003]

De acordo com a resolução CONAMA 357/2005 [Conselho Nacional do Meio Ambiente, 2005] [11] , os valores máximos permitidos (VM) de zinco em águas doces são de 0,18 mg/L para classes 1 e 2, e de 5 mg/L para classe 3. Para águas salinas, os valores máximos permitidos são de 0,09 mg/L para classe 1 e de 0,12 µg/L para classe 2. Em águas salobras, os valores máximos permitidos são de 0,09 mg/L para classe 1 e de 0,12 µg/L para classe 2.

2.5.19 Cádmio

O cádmio é um metal branco, brilhante e maleável com propriedades físicas e químicas semelhantes ao zinco, aparecendo com frequência na forma de sulfetos. Sua fonte principal são rochas sedimentares, a erosão dessas rochas pode resultar no transporte de cádmio para os rios e, eventualmente, para os oceanos. Em acréscimo, o cádmio pode alcançar os sistemas aquáticos através do intemperismo e erosão do solo, descargas diretas, operações industriais como vazamentos de aterros, esgoto e outras atividades industriais. [Almeida, 2012]

O cádmio reúne características importantes relacionadas à atuação como elemento tóxico. É conhecido por ser bioacumulável e persistente no meio ambiente, causando uma série de efeitos adversos tanto no homem quanto na vida aquática. Quando exposto a organismos aquáticos, o cádmio pode resultar na redução da taxa de natação, diminuição de atividades enzimáticas e alterações histológicas em diferentes órgãos e espécies. [Alkimin, 2016]

[11] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

O cádmio, apesar de ser um elemento naturalmente presente no ambiente, é amplamente reconhecido como um dos metais pesados mais perigosos devido à sua toxicidade elevada e ao impacto duradouro no meio ambiente e na saúde humana. Além das fontes naturais, atividades antropogênicas como mineração, galvanização, fabricação de baterias, pigmentos e plásticos aumentam significativamente para sua liberação em sistemas aquáticos. Uma vez introduzido no ambiente, o cádmio pode aderir a partículas de sedimentos ou permanecer dissolvido na água, aumentando sua biodisponibilidade para os organismos aquáticos. [Almeida, 2012]

No corpo humano, o cádmio pode se acumular em órgãos como rins e fígado, causando danos graves a longo prazo. Exposições prolongadas a níveis elevados podem levar a disfunções renais, problemas ósseos, hipertensão e doenças respiratórias, além de serem classificadas como carcinogênicas pela Agência Internacional de Pesquisa sobre o Câncer (IARC). Para as populações que consomem água contaminada ou alimentos provenientes de áreas poluídas, os riscos são ainda mais elevados, especialmente em regiões próximas às indústrias que utilizam o metal em seus processos. Esses impactos reforçam a necessidade de controles rigorosos em atividades industriais e no tratamento de efluentes. [Alkimin, 2016]

A norma da conama [Conselho Nacional do Meio Ambiente, 2005] [12] para águas de classe 1, que são destinadas ao consumo humano sem tratamento, o cádmio total não deve exceder 0,001 mg/L. Já para águas de classe 3, destinadas à navegação e à harmonia paisagística, o limite é de 0,01 mg/L. Em ambientes aquáticos salinos, os valores são mais restritivos, com um máximo de 0,005 mg/L para classe 1 e 0,04 mg/L para classe 2. Em águas salobras, também há limites rigorosos, com 0,005 mg/L para classe 1 e 0,04 mg/L para classe 2. Para lançamentos de efluentes, o cádmio total permitido é de até 0,2 mg/L.

2.5.20 Níquel

O níquel (Ni) é um metal duro, dúctil e maleável, muito utilizado com outros metais como ferro, cobre, cromo e zinco na fabricação de peças e moedas, joias, baterias, entre outros. A maior aplicação industrial está na produção de aço inoxidável. Possui diversas etapas de oxidação, porém o mais frequente é o Ni2+, que tem a capacidade de formar vários complexos. Forma compostos inorgânicos solúveis, como os hidróxidos, sulfatos, cloretos e nitratos, e insolúveis, como os óxidos e sulfetos. [CETESB – Companhia Ambiental do Estado de São Paulo, 2021], [Almeida, 2012]

O escoamento superficial dos solos, deposições atmosféricas e intemperismos naturais contribuem para a presença de níquel na água. Os efluentes domésticos e industriais também contribuem para a presença de níquel na água. O níquel pode formar compostos

[12] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

solúveis como cloretos, nitratos e sulfatos, que têm uma coloração verde distinta em soluções aquosas; óxidos e sulfetos, por outro lado, são pouco solúveis. [Almeida, 2012]

O níquel metálico apresenta riscos significativos à saúde humana, sendo associado à carcinogenicidade em diversos estudos toxicológicos. A exposição ao níquel pode afetar o sistema nervoso, cardiovascular e respiratório, além de provocar reações dermatológicas, especialmente em trabalhadores de setores como siderurgia e refinarias, onde o contato com compostos de níquel é mais intenso. [Gonzalez, 2016] Concentrações naturais em águas superficiais podem chegar a cerca de 0,1 mg/L, mas áreas de mineração podem ter concentrações superiores a 11,0 mg/L. Além de afetar os nervos, o coração e o sistema respiratório, exposição elevada pode causar dermatite em pessoas sensíveis.[Costa, 2018]

O grau de toxicidade dos compostos de níquel depende de sua solubilidade em água; os compostos solúveis geralmente são mais tóxicos, enquanto os compostos insolúveis têm maior probabilidade de causar cancerígenos no local de deposição. Pneumonite intersticial difusa, edema cerebral, hipertermia, tosse, tontura, mal-estar, vômitos, náuseas, pulso rápido e colapso são sintomas da intoxicação por níquel que afeta os pulmões e o sistema gastrointestinal. [Almeida, 2012]

A legislação ambiental brasileira, por meio da Resolução CONAMA n.º 357/2005 [13] e da Resolução CONAMA n.º 430/2011, estabelece limites rigorosos para a concentração de níquel em águas doces, salinas e salobras, com valores máximos permitidos de 0,025 mg /L para águas de consumo e até 2,0 mg/L para efluentes. Esses limites buscam minimizar os impactos ambientais e proteger a saúde pública. No entanto, a eficiência dessas regulamentações depende da fiscalização adequada e do cumprimento pelas indústrias de práticas de gestão sustentável. A implementação de tecnologias de tratamento de água e pacotes, juntamente com programas de monitoramento ambiental, é essencial para mitigar os efeitos do trabalho pesado em ecossistemas aquáticos e na saúde humana.

2.5.21 Manganês

O manganês está presente em várias indústrias, pois participa na produção do aço, das ligas metálicas, e usado em vários produtos como vidros, fertilizantes, vernizes, suplementos veterinários e oxidantes para limpeza. Mesmo tão presente a sua concentração raramente ultrapassam 1,0 mg/L em águas naturais, sendo mais comuns em concentrações de 0,2 mg/L ou menos. [Companhia de Desenvolvimento e Ação Regional, 2019]

A alta presença de manganês na água, assim como o ferro, pode resultar em várias consequências indesejáveis. Causa manchas em tecidos, roupas e utensílios sanitários além de conferir um sabor desagradável e "metálico"à água, o que é indesejável para consumo humano. Em processos industriais, como na fabricação de papel, tecidos, tinturarias e

[13] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

cervejarias, o manganês pode interferir e comprometer a qualidade dos produtos finais. [Moruzzi & Reali, 2012]

O manganês pode também levar à formação de depósitos e incrustações em encanamentos e equipamentos, reduzindo a eficiência dos sistemas. Além de proporcionar condições favoráveis para o crescimento de bactérias ferruginosas nocivas, representando um risco adicional à saúde pública. [Moruzzi & Reali, 2012]

Além dos efeitos diretos à saúde, o manganês presente em altas concentrações na água pode influenciar significativamente os ecossistemas aquáticos. Este metal pode se acumular em aquários, alterando a cadeia alimentar e afetando a biodiversidade. Em condições de alta concentração (por exemplo, 1,5 a 2,9 mg/L), o manganês pode se acumular nos tecidos de organismos aquáticos, induzindo efeitos genotóxicos e bioquímicos, como demonstrado em tilápias *Oreochromis niloticus* [Coppo *et al.*, 2018].

No Brasil, a legislação ambiental e sanitária estabelece limites rigorosos para a concentração de manganês em diferentes classes de água. De acordo com a Resolução CONAMA n. 357/2005 [14] , o limite máximo para águas de classe 1 é de 0,1 mg/L, enquanto a Portaria nº 2914/11 do Ministério da Saúde reforça este valor como padrão para consumo humano. Esses limites são específicos para proteger tanto a saúde pública quanto o meio ambiente, mas sua eficácia depende de ações contínuas de monitoramento e fiscalização, bem como do compromisso das indústrias e do governo em investir em tecnologias para o controle da poluição e do tratamento.

2.5.22 Nitrogênio Amoniacal

O nitrogênio amoniacal é a quantidade de nitrogênio presente na forma de amônia (NH_3) e amônio (NH_4^+) em uma amostra de água, solo ou outro meio ambiental. A amônia é produzida em fertilizantes comerciais, bem como nas aplicações industriais, como fonte de hidrogênio em acabamentos metálicos e também usos na indústria química, como produção de produtos farmacêuticos e corantes. Pode-se citar também a amônia presente no processamento de petróleo bruto e na proteção contra corrosão, e na indústria de mineração para extração de metais. [U.S. Environmental Protection Agency, 2013]

A amônia pode entrar na água através de efluentes das cidades, escoamento da agricultura, fixação de nitrogênio e excreção de resíduos nitrogenados por animais. Sua presença impacta diretamente nos sistemas de aquicultura, como tanques de criação de peixes, incubadoras e tanques de transporte. Muitas espécies de peixes não suportam concentrações acima de 5 mg/L. [Companhia de Desenvolvimento e Ação Regional, 2019], [Companhia Ambiental do Estado de São Paulo, 2024] Para evitar concentrações inaceitavelmente altas de amônia nas águas superficiais, muitos efluentes precisam passar por um

[14] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

tratamento extensivo. [Lin *et al.*, 2023]

Sua toxicidade aumenta em função do aumento de pH. Em pH e temperatura baixos, a amônia combina-se com a água para formar íons amônio ($NH4^+$) e íons hidróxido (OH^-). Os íons de amônio não são tóxicos e não causam problemas aos organismos vivos, mas a forma não ionizada é tóxica. Em valores de pH acima de 9, a amônia não iônica torna-se a forma predominante no meio e pode atravessar a membrana celular mais rapidamente à medida que o valor do pH aumenta. [Companhia Ambiental do Estado de São Paulo, 2024]

Pela Resolução nº 357/2005 [Conselho Nacional do Meio Ambiente, 2005], do Conselho Nacional do Meio Ambiente (CONAMA), o nitrogênio amoniacal é considerado um parâmetro tanto para a classificação das águas naturais quanto para os padrões de emissão de esgotos. Para águas doces de Classe 1, os limites são 3,7 mg/L N para pH = 7,5, 2,0 mg/L N para 7,5 < pH <= 8,0, 1,0 mg/L N para 8,0 < pH < 8,5, e 0,5 mg/L N para pH > 8,5. Para águas doces de Classe 3, os limites são 13,3 mg/L N para pH = 7,5, 5,6 mg/L N para 7,5 < pH <= 8,0, 2,2 mg/L N para 8,0 < pH <= 8,5, e 1,0 mg/L N para pH > 8,5. Em águas salinas e salobras de Classe 1, o limite é 0,40 mg/L N, e para águas salinas de Classe 2, o limite é 0,70 mg/L N. Para o lançamento de efluentes, o limite é 20,0 mg/L N.

2.5.23 Nitrito

Os nitritos (NO_2^-) se formam a partir da oxidação da amônia, e a transformação de íons amônio a nitrito, representa uma fase intermediária do processo de conversão da amônia em nitrato.[da Costa e Silva Crespim, 2017], [Torres, 2011].Por se tratar de uma fase transitória, os nitritos geralmente não são encontrados em concentrações significativas, exceto em ambientes redutores, onde o nitrato, que é o estado de oxidação mais estável, não é predominante. Assim, a presença simultânea de nitrito e amônia na água pode ser um forte indicativo de poluição orgânica recente, refletindo a entrada de matéria orgânica fresca e sua decomposição inicial .[Companhia de Desenvolvimento e Ação Regional, 2019]

O monitoramento dos níveis de nitrito na água é especialmente relevante devido aos impactos diretos na saúde humana e nos ecossistemas aquáticos. A ingestão de nitrito pela água pode causar efeitos tóxicos imediatos, sendo mais pronunciados do que aqueles provocados pelo nitrato. Em humanos, a ingestão de nitrito pode desencadear a metemoglobinemia, uma condição caracterizada pela redução da capacidade do sangue de transportar oxigênio para os tecidos, podendo afetar consumidores de todas as faixas etárias. [Alaburda & Nishihara, 1998]

No Brasil, a Resolução CONAMA 357/2005 [15] estabelece limites específicos para a

[15] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

concentração de nitrito em diferentes classes de água. Para águas doces das classes 1, 2 e 3, o limite máximo permitido é de 1,0 mg/L de nitrogênio. Em águas salinas, os limites são ainda mais restritivos: até 0,07 mg/L de nitrogênio para classe 1 e 0,20 mg/L para classe 2. Para águas salobras, os valores permitidos seguem os mesmos padrões estabelecidos para águas salinas.

No contexto do uso de algoritmos de aprendizagem de máquina para previsão e classificação do índice de qualidade da água, o nitrito desempenha um papel essencial como parâmetro de entrada. Sua variabilidade pode ser analisada para identificar padrões associados à poluição orgânica recente e prever alterações na qualidade da água. [Jafar *et al.*, 2023], [Chen *et al.*, 2024].

Modelos preditivos baseados em aprendizagem de máquina têm o potencial de avaliar rapidamente concentrações de nitrito, amônia e nitrato, contribuindo para a identificação de áreas críticas de poluição e facilitando a tomada de decisão em tempo real para a gestão sustentável de recursos hídricos. [Peixoto *et al.*, 2023]

2.5.24 Nitrato

A formação do íon nitrato (NO_3^-) ocorre a partir da oxidação do íon nitrito. Uma vez liberado no solo, o nitrato pode ser absorvido e metabolizado por vegetais ou ser retirado do solo pela ação da água. [da Costa e Silva Crespim, 2017]

O nitrato pode alcançar as águas através do uso dos fertilizantes e da erosão de depósitos naturais ao descarte de águas residuais e à oxidação bacteriana de resíduos nitrogenados provenientes das excreções humanas e animais, incluindo sistemas sépticos, indicando potencial contaminação por efluentes domésticos. [Companhia de Desenvolvimento e Ação Regional, 2019]

A ingestão do nitrato por meio da água está ligada a dois impactos negativos na saúde: o desenvolvimento de metemoglobinemia, principalmente em crianças, e a possível criação de substâncias carcinogênicas como nitrosaminas e nitrosamidas. [Alaburda & Nishihara, 1998]

Além disso, a presença de nitratos na água pode afetar as variedades do ecossistema aquático, especialmente em ambientes de águas doces [Gomez Isaza *et al.*, 2020]. Em altas concentrações, os nitratos podem estimular a eutrofização, um processo no qual a água recebe um excesso de nutrientes, especialmente concentrações de nitrogênio e fósforo, que favorecem o crescimento excessivo de algas. Esses fenômenos podem resultar na diminuição da concentração de oxigênio na água, criando zonas de "morte"onde a vida aquática não consegue sobreviver. Isso compromete a biodiversidade aquática e a qualidade da água para o consumo humano. [Smith, 2003]

O impacto do nitrato na saúde humana, especialmente em águas potáveis, é uma preocupação crescente. A metemoglobinemia, comumente conhecida como "síndrome do

bebê azul", ocorre quando os níveis de nitrato ingeridos são convertidos em nitrito no organismo, o que interfere na capacidade do sangue de transporte de oxigênio, levando à falta de ar e, em casos graves, ao risco de morte. Crianças menores de seis meses são particularmente vulneráveis a essa condição, já que seu sistema digestivo ainda não é capaz de reduzir eficientemente o nitrito. Além disso, a formação de compostos como as nitrosaminas, que são potencialmente cancerígenos, ocorre quando o nitrato interage com aminas presentes em alimentos ou no sistema digestivo. [Dovidauskas *et al.*, 2019]

A regulamentação do teor de nitrato na água potável é crucial para proteger a saúde pública. No Brasil, a Portaria nº 2914/11 do Ministério da Saúde estabelece limites para a concentração de nitrato em águas destinadas ao consumo humano, com o valor máximo permitido sendo 10 mg/L, equivalente a 45 mg/L de nitrato, expresso como NO3 . Esses limites visam garantir que a ingestão de nitrato por meio da água não represente riscos à saúde. No entanto, o controle eficaz da poluição por nitratos exige um esforço conjunto entre setores como agricultura, indústria e saneamento, além de estratégias para promover a educação ambiental e o uso responsável de fertilizantes e produtos químicos.

2.5.25 Silicatos

Os silicatos são minerais compostos de silício e oxigênio, que são combinados com outros elementos como alumínio, ferro, magnésio, cálcio, sódio e potássio. A estrutura é de de tetraedros de sílica (SiO_4), que podem se organizar de várias maneiras dependendo das condições em que são encontrados, obtendo assim uma variedade de minerais com diferentes propriedades químicas e físicas [Klein & Dutrow, 2012]. Exemplos mais comuns de silicatos são os quartzo, feldspatos, micas e os piroxênios.

De acordo com [Braga *et al.*, 2020], a presença de sílica em águas potáveis pode contribuir para reduzir a taxa de mortalidade devido a doenças coronárias e problemas de coração na população além de desempenhar um papel importante no controle da disponibilidade biológica tóxica do alumínio, aumenta a eficácia hidroterápica e melhora a capacidade de compreensão fisiológica dos alimentos.

Outro trabalho [Tilman *et al.*, 1982] indica o silício como um recurso limitante para o fitoplâncton, juntamente com fósforo e nitrogênio. Portanto, a presença de silicatos é benéfica para o crescimento do fitoplâncton que requer silício, como as diatomáceas. Não excluindo o fato que a sua concentração em excesso pode promover um desbalanceamento nas cadeias alimentares aquáticas, favorecendo o crescimento excessivo de diatomáceas em detrimento de outros tipos de fitoplâncton, o que pode afetar negativamente a biodiversidade aquática e alterar a dinâmica dos ecossistema

A sílica não é considerada um contaminante, ainda que alguns fornecedores de água optam por retirar a sílica pois ela pode interferir alguns processos de tratamento de água, como a coagulação e a floculação, reduzindo a eficiência desses processos. Outros serviços

removem a sílica para evitar incrustações em equipamentos domésticos, como aquecedores de água. [Edzwald, 2011]

2.5.26 Clorofila

A clorofila é é um pigmento encontrado em plantas, algas e algumas bactérias que apresenta como característica mais marcante sua coloração verde. Ele é frequentemente utilizada como indicadora de biomassa fitoplanctônica, indicando o crescimento de algas e cianobactérias devido ao enriquecimento por nutrientes principalmente nitrogênio e fosforo. [Wetzel, 2000]

A medição de clorofila ajuda a monitorar a qualidade da água. Altas concentrações de clorofila podem indicar eutrofização, um processo que pode levar à proliferação excessiva de algas e consequente depleção de oxigênio na água, afetando negativamente a vida aquática. [de Saneamento Ambiental, 2014]

A resolução Conama 357/2005 [Conselho Nacional do Meio Ambiente, 2005] estabelece os valores de clorofila a 10 μg/L classe 1 águas doces, clorofila a: até 30 μg/L; classe 2, Clorofila a 60 μg/L classe 3.

A clorofila, além de sua função essencial na fotossíntese, também desempenha um papel crucial como indicador ecológico em ambientes aquáticos. Sua presença nas águas, especialmente em concentrações elevadas, reflete o nível de produtividade biológica, ou seja, a quantidade de organismos fotossintetizantes presentes, como fitoplânctons e algas. Esses organismos são fundamentais para o equilíbrio ecológico, de base para a cadeia alimentar aquática. A medição da clorofila, portanto, não só ajuda a avaliar a qualidade da água, mas também fornece dados sobre a saúde dos ecossistemas aquáticos. Concentrações excessivas de clorofila, no entanto, podem sinalizar que há uma disponibilidade excessiva de nutrientes, como o nitrogênio e o fósforo, frequentemente provenientes de efluentes domésticos e agrícolas.[Nabout *et al.*, 2022], [Wetzel, 2000]

A relação entre altas concentrações de clorofila e a eutrofização é um dos principais desafios para a gestão da qualidade da água em ambientes aquáticos. A eutrofização ocorre quando há um aumento excessivo de nutrientes na água, especialmente engenharia e fósforo, que estimulam o crescimento descontrolado de algas e fitoplâncton [Nabout *et al.*, 2022]. Embora o aumento da biomassa fitoplanctônica não seja necessariamente prejudicial, sua concentração consome grandes quantidades de oxigênio dissolvido, o que pode resultar em zonas de baixa oxigenação, conhecidas como zonas mortas, onde a fauna aquática não consegue sobreviver. Em casos extremos, isso pode levar à perda de biodiversidade e à degradação dos ecossistemas aquáticos.[Smith, 2003]

A regulamentação da concentração de clorofila nas águas é uma ferramenta importante para a gestão da qualidade hídrica e o controle da eutrofização. A resolução CONAMA 357/2005 define limites para a clorofila nas diferentes classes de águas, mudando

para garantir a manutenção de um ambiente aquático saudável. Para as águas de classe 1, destinadas ao consumo humano e à conservação da biodiversidade, o limite é de 10 µg/L de clorofila a, diminuindo que a concentração de águas com esse nível de biodiversidade são adequadas para essas barbatanas. Já para as águas de classe 3, com finalidade de lazer e navegação, o limite pode ser mais elevado, de até 60 µg/L, refletindo as diferentes exigências de qualidade conforme o uso da água. Dessa forma, o monitoramento da clorofila, em conjunto com outras parâmetros de qualidade, pode fornecer informações essenciais para prevenir problemas relacionados à poluição e à perda de qualidade das águas.

2.5.27 Cianetos

Os cianetos são substâncias químicas que contêm o grupo ciano ($-\mathrm{C} \equiv \mathrm{N}$). Nas amostras de água, a sua medida pode indicar poluição por atividades industriais ou agrícolas ou ao descarte inadequado de produtos químicos contendo cianeto. Além de sua origem natural, os compostos cianogênicos também podem ser sintetizados por alguns fungos, bactérias e algas e são encontrados em diversos alimentos e plantas.Extração de ouro e prata, na limpeza de metais, na produção de fibras sintéticas, corantes, pigmento também utilizam os cianetos.A maioria das vezes, quando o cianeto de hidrogênio (HCN) está presente na água, sua parte evapora. Além disso, as bactérias podem transformá-los em substâncias menos tóxicas, como zooplâncton e fitoplâncton, ou podem formar complexos com metais, como o ferro. [Schneider, 2009], [ATSDR - Agency for Toxic Substances and Disease Registry, 2006]

Várias formas do cianeto prejudicam a vida aquática, terrestre e aérea porque impedem o transporte de oxigênio no metabolismo. Ao entrar no organismo por via oral, inalação ou ingestão, o cianeto se espalha rapidamente, afetando processos vitais. O HCN molecular é mais venenoso do que o íon CN-.O cianeto não se acumula em organismos expostos a concentrações baixas devido a mecanismos de detoxificação. A temperatura, o teor de oxigênio dissolvido e a concentração de minerais na solução são fatores que o influenciam na toxicidade dos cianetos para os peixes. A proporção de HCN não dissociado aumenta com o pH. A ação letal de um aumento de 10 oC na temperatura da água é duplicada ou triplicada. [Schneider, 2009]

Para águas subterrâneas, o valor máximo permitido (VMP) para consumo humano é de 70 μg/L e para recreação é de 100 μg/L, conforme a Resolução CONAMA 396/2008. Em água doce, os valores de referência são 0,005 mg/L para a classe 1 e 2 e 0,022 mg/L para a classe 3, conforme a Resolução CONAMA 357/2005. Para águas salinas e salobras, a concentração máxima permitida é de 0,001 mg/L para as classes 1 e 2, também de acordo com a Resolução CONAMA 357/2005. [CETESB, 2022]

2.5.28 *Escherichia Coli*

As *Escherichia Coli* são bactérias que habitam naturalmente o intestino das pessoas e de organismos de sangue quente, desempenhando funções benéficas no intestino, auxiliando na digestão e competindo com bactérias nocivas. No entanto, há alguns tipos de *E. coli* que são nocivos, sendo responsável por quadros infecciosos que levam a elevados custos sociais e econômicos na medicina humana e animal. [Russo & Johnson, 2000] [Chen *et al.*, 2003], [Abdallah *et al.*, 2011].

A *E. coli* é amplamente conhecida como um indicador confiável da qualidade microbiológica da água, principalmente quando se trata de contaminação fecal. A sua presença em amostras de água está normalmente associada à possibilidade de contaminação por fezes humanas ou animais, o que representa uma ameaça significativa à saúde pública. Por ser uma bactéria encontrada no trato digestivo de humanos e animais de sangue quente, a detecção de *E. coli* na água pode indicar a presença de outros patógenos transmitidos pelas fezes, como vírus e protozoários. Portanto, é muito importante verificar regularmente a presença de *E. coli* nas fontes de água para avaliar a segurança do abastecimento de água e garantir a proteção da saúde humana [Cabral, 2010]

Em Brasil [2011] encontra-se a Portaria nº 2914 de 12 de dezembro de 2011, emitida pelo Ministério da Saúde do Brasil, representa um marco regulatório importante no controle e vigilância da qualidade da água para consumo humano. Esta portaria, que modificou a Portaria 518/2004, estabelece os procedimentos e padrões de potabilidade da água. No artigo 27, é determinado que a água potável deve atender ao padrão microbiológico estabelecido. Conforme disposto no anexo I da referida portaria, a presença de *Escherichia coli* não é permitida em 100 mL de água, sendo este um parâmetro crucial para avaliar a segurança da água para consumo humano.

2.5.29 Coliformes

Os coliformes totais são microrganismos presentes no trato intestinal de todos os mamíferos, incluindo humanos, e são essenciais como marcadores de contaminação fecal. Esses microrganismos pertencem a diversos gêneros da família *Enterobacteriaceae* e são gram-negativos, com bastonetes não formadores de esporos, comumente encontrados no solo e nas fezes de animais de sangue quente. Eles têm a capacidade de fermentar a lactose e podem ser encontrados em águas com alta matéria orgânica, como efluentes industriais, ou em solo e vegetais em decomposição. [Carnaúba *et al.*, 2021]

A Portaria nº 518/2004 define coliformes totais e coliformes termotolerantes. Os coliformes totais são bacilos gram-negativos, aeróbios ou anaeróbios facultativos, não formadores de esporos e oxidase-negativos, que fermentam a lactose produzindo ácido, gás e aldeído, pertencendo principalmente aos gêneros *Escherichia*, *Citrobacter*, *Klebsiella* e *Enterobacter*. Os coliformes termotolerantes, um subgrupo dos coliformes totais, fermentam

a lactose, com *Escherichia coli* como principal representante, originária exclusivamente de fezes. A *Escherichia coli* é considerada o indicador mais específico de contaminação fecal recente devido à sua capacidade de fermentar lactose e manitol, produzir ácido e gás, indol do triptofano, ser oxidase negativa, não hidrolisar ureia e realizar atividades das enzimas β-galactosidase e β-glucoronidase. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

Na avaliação da qualidade de águas naturais, os coliformes totais têm valor sanitário limitado, sendo mais aplicáveis à água tratada, onde sua presença pode indicar falhas no tratamento, contaminação pós-tratamento ou excesso de nutrientes em reservatórios e redes de distribuição. O grupo dos coliformes inclui bactérias não exclusivamente de origem fecal, podendo estar presentes em plantas, água e solo, e são capazes de se reproduzir na água, especialmente em ambientes tropicais (OMS, 1995). [Ministério da Saúde, 2006]

De acordo com a Portaria de Consolidação nº 5, de 28 de setembro de 2017, que regulamenta os padrões de potabilidade da água para consumo humano no Brasil, os coliformes totais devem estar ausentes em 100 mL. Quanto aos coliformes totais, os sistemas ou soluções alternativas coletivas que abastecem menos de 20.000 habitantes podem ter uma amostra positiva entre as amostras examinadas no mês. Para sistemas ou soluções alternativas coletivas que abastecem a partir de 20.000 habitantes, deve haver ausência de coliformes totais em 100 mL em 95% das amostras examinadas no mês.

Já a Resolução CONAMA nº 274 de 2000, Art. 2º, as águas doces, salobras e salinas destinadas à balneabilidade (recreação de contato primário) serão avaliadas nas categorias própria e imprópria. As águas consideradas próprias poderão ser subdivididas nas seguintes categorias: Excelente: quando em 80% ou mais de um conjunto de amostras obtidas em cada uma das cinco semanas anteriores, colhidas no mesmo local, houver no máximo 250 coliformes fecais (termotolerantes) ou 200 *Escherichia coli* ou 25 enterococos por 100 mililitros; Muito Boa: quando em 80% ou mais de um conjunto de amostras obtidas em cada uma das cinco semanas anteriores, colhidas no mesmo local, houver no máximo 500 coliformes fecais (termotolerantes) ou 400 Escherichia coli ou 50 enterococos por 100 mililitros; Satisfatória: quando em 80% ou mais de um conjunto de amostras obtidas em cada uma das cinco semanas anteriores, colhidas no mesmo local, houver no máximo 1.000 coliformes fecais (termotolerantes) ou 800 *Escherichia coli* ou 100 enterococos por 100 mililitros.

2.5.30 Mercúrio

O mercúrio (Hg), de acordo com o Programa Ambiental das Nações Unidas (PNUMA), é um elemento natural extremamente tóxico para os humanos e tem um ciclo de poluição extenso que passa pelo ar, água, sedimentos, solo e seres vivos. [PIRES *et al.*, 2023] Esse metal pode ser encontrado no ambiente em uma variedade de formas, como

gases, metal e compostos orgânicos e inorgânicos. O mercúrio tem um grande potencial de causar danos ao meio ambiente, especialmente na forma de metilmercúrio no sistema aquático, devido à sua biomagnificação e bioacumulação ao longo da cadeia trófica.

O mercúrio é amplamente utilizado no Brasil para a extração de prata, bem como para a produção de cloro eletrolítico, equipamentos elétricos, amalgamas e compostos de mercúrio. Pode ser encontrado tanto na água superficial quanto subterrânea, geralmente em forma inorgânica, com concentrações que podem aumentar devido a atividades antropogênicas, na indústrias, processos de mineração e fundição, efluentes de estações de tratamento de esgoto e indústrias de tintas. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

A toxicidade do mercúrio no corpo humano é devido à sua capacidade de bioacumulação nos organismos aquáticos. Além de ser o metal com a maior toxicidade, o mercúrio é o único metal capaz de sofrer biomagnificação em quase todas as cadeias alimentares. Sua concentração aumenta com o nível trófico de uma espécie, o que resulta em uma exposição ambiental significativa para consumidores de alto nível trófico, como os humanos, na cadeia alimentar. [Lacerda & Malm, 2008]

Isso significa que a presença de mercúrio no meio ambiente representa uma ameaça significativa à saúde pública. Doses de de três a trinta gramas podem ser extremamente venenoso para seres humanos. É cumulativo e pode causar lesões cerebrais. Náuseas, vômitos, dores abdominais, diarréia, lesões nos ossos e morte são sinais de intoxicação aguda. Isso pode ser fatal em 10 dias. A intoxicação persistente afeta os rins, as glândulas salivares e as funções psicológicas e psicomotoras. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

A resolução Conama 357/2005 [Conselho Nacional do Meio Ambiente, 2005] [16] estabelece os valores de até 0,0002 mg/L para a Classe 1, de 0,0002 mg/L a 0,002 mg/L para a Classe 2, e acima de 0,002 mg/L até 0,01 mg/L para a Classe 3, que é aplicada especificamente a lançamentos de efluentes. Em águas salinas e salobras, os limites são definidos como até 0,0002 mg/L para a Classe 1 e de 0,0002 mg/L até 0,0018 mg/L para a Classe 2.

2.5.31 Óleos e Graxas

Os óleos e gorduras são substâncias orgânicas que podem ter origem em fontes minerais, vegetais ou animais. Grande parte desses compostos é formada por hidrocarbonetos, gorduras e ésteres. Os principais responsáveis pela presença de óleos e gorduras em ambientes aquáticos são os resíduos industriais, o esgoto doméstico, os efluentes de oficinas de automóveis, postos de gasolina, além da poluição em estradas e ruas. [Companhia

[16] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

Ambiental do Estado de São Paulo (CETESB), 2016]

São medidos utilizando método analítico com amostra e uso de um solvente específico e que não se evapora quando o solvente é aquecido a 100 C. Os compostos solúveis em n-hexano abrangem ácidos graxos, gorduras de origem animal e vegetal, sabões, graxas, óleos de origem vegetal, ceras e óleos minerais. Além disso, a avaliação do parâmetro também pode ser realizada por meio da análise do material solúvel em hexano (HSM). [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

A Resolução CONAMA 430/2011, para o lançamento direto de efluentes provenientes de sistemas de tratamento de esgotos sanitários, devem ser obedecidas condições e padrões específicos, incluindo a concentração de substâncias solúveis em hexano (óleos e graxas) que não deve exceder 100 mg/L. Já a Resolução CONAMA 357/2005 estabelece que para as classes 1, 2 e 3 de corpos d'água, a presença de óleos e graxas deve ser virtualmente ausente.

2.5.32 Alumínio

O alumínio é um metal bastante abundante na crosta terreste, encontrado naturalmente no solo, água, e no ar, mas também pode pode ser redistribuído ou movido por atividades naturais ou humanas. [Cleto, 2008] O alumínio e seus sais são usados no tratamento da água para fins como aditivos alimentares, telhas, latas, papel alumínio e indústria farmacêutica. Por meio da suspensão da poeira do solo e da combustão do carvão, o alumínio pode entrar na atmosfera como material particulado. O pH, a temperatura e a presença de fluoretos, sulfatos, matéria orgânica e outros ligantes na água determinam as formas em que o metal pode estar presente.

Na água potável, as concentrações de alumínio podem variar dependendo da quantidade presente na fonte e do uso de coagulantes à base de alumínio no processo de tratamento. Alimentos podem conter alumínio tanto de fontes naturais quanto de aditivos, como ocorre em batatas, espinafre e chá. Com exceção de algumas ervas e folhas de chá, a maioria dos alimentos não processados contém menos de 5 mg/kg de alumínio.[Companhia Ambiental do Estado de São Paulo (CETESB), 2013a]

A concentração de alumínio deve ser maior em profundidades onde o pH é menor e pode haver anaerobiose. À medida que a estação das chuvas se aproxima, o teor de alumínio no corpo de água como um todo diminui. Isso ocorre se a estratificação e a consequente anaerobiose não forem significativas. O período de chuvas e a alta turbidez estão ligados ao aumento da concentração de alumínio. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

A Organização Mundial da Saúde e a Agência de Proteção Ambiental dos Estados Unidos – USEPA, em investigações realizadas e em andamento, sugerem uma influência do alumínio na etiologia de doenças neurodegenerativas, tais como, o Mal de Parkinson e

o Mal de Alzheimer. Deficiências nutricionais crônicas de cálcio e magnésio possivelmente aumentam a absorção do alumínio, resultando em sua deposição nos neurônios, o que interfere na estrutura dessas células e nas funções cerebrais. [Figueiredo *et al.*, 2004]

2.5.33 Cobre

O cobre é encontrado em fontes ambientais como minas de cobre bem como em fluentes de estações de tratamento de esgotos, uso de compostos de cobre como algicidas aquáticos, escoamento superficial e contaminação da água subterrânea por uso agrícola de cobre. Também é presente naturalmente em todas as plantas e animais, sendo um nutriente essencial. A ingestão diária desse metal varia entre adultos. entre 0,9 e 2,2 miligramas, com estimativas de crianças de 0,6 a 0,8 miligramas. A falta de cobre na dieta de animais pode causar anemia, problemas nervosos e diarreias. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016] , [Pereira, 2017].

A principal fonte de contaminação por cobre, embora existam na natureza, são as ações que são antropogênicas. Esse metal é um dos componentes químicos mais frequentemente usados nas áreas, principalmente construção civil, eletrodeposição, agricultura, automóveis e de equipamentos elétricos, que representa cerca de cinquenta por cento do consumo total de metal. As áreas de mineração e metalurgia são as principais responsáveis pela maioria das emissões de efluentes que contêm íons de cobre. [Pereira, 2017]

Concentrações superiores a 2,5 mg/L adicionam sabor amargo à água e concentrações superiores a 1 mg/L causam coloração em louças e limpeza. As doses elevadas de cobre são extremamente nocivas para os peixes, muito mais do que para os humanos. Trutas, carpas, bagres, peixes vermelhos de aquários ornamentais e outros peixes podem morrer em concentrações de 0,5 mg/L. Os microorganismos são mortais em concentrações superiores a 1,0 mg/L. De acordo com a Portaria 2914/11 do Ministério da Saúde, o padrão de potabilidade do cobre é de 2 mg/L. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

Já pela Resolução CONAMA 357/2005 [Conselho Nacional do Meio Ambiente, 2005] [17] , tem-se valores especificados conforme a classe e o tipo de ambiente aquático. Para águas doces de Classe 1 e 2, o limite é de 0,009 mg/L, enquanto para águas doces de Classe 3, é de 0,013 mg/L. Em águas salinas, o limite é mais restritivo, com 0,005 mg/L para a Classe 1 e 7,8 para a Classe 2, bem como para águas salobras. Para o lançamento direto de efluentes, a concentração máxima permitida de cobre é de 1,0 mg/L.

[17] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

2.5.34 Cianobactérias

As cianobactérias são microrganismos procariontes que realizam fotossíntese e sintetizam clorofila a. Possuem um papel similar as algas unicelulares, mas não podem ser classificadas como essas algas, nem como como bactérias comuns. Elas possuem características celulares típicas dos procariontes, como a ausência de membrana nuclear, e um sistema fotossintético que não está organizado em cloroplastos. Contém os pigmentos ficobilinas, ficocianina que a leva à cor azulada característica ou ficoeritrina, de cor vermelha, nem sempre presente. [Mann, 2002b], [Siqueira & Oliveira-Filho, 2023]

São capazes de crescer e sobreviver a mais diferentes meios, como em solos e rochas, participando em processos funcionais do ecossistema e ciclagem de nutrientes.Contudo, a maioria é encontrada em ecossistemas aquáticos. Aumentos nas concentrações de nitrogênio e fósforo podem levar a uma multiplicação excessiva dessas bactérias, resultando em florações. Essas florações estão associadas ao processo de eutrofização, processo já descrito nas seções anteriores [Wetzel, 2000].

Diversos tipos de espécies de cianobactérias detêm a propriedade de gerar toxinas denominadas por cianotoxinas. [Azevedo, 1998] Em estudos sistemáticos, cerca de 25% a 70% das florações de cianobactérias mostraram ser potencialmente tóxicas. [Siqueira & Oliveira-Filho, 2023]São uma fonte significativa produtos nocivos produzidos por esses microrganismos. Algumas dessas toxinas são caracterizadas por seus efeitos rápido, levando à morte por parada respiratória em poucos minutos após a exposição.

As cianotoxinas podem afetar o organismo humano, altas doses podem causar morte por hemorragia hepática ou insuficiência hepática, e a exposição crônica a baixas doses a longo prazo pode promover o crescimento de tumores. [Mann, 2002a] Além disso, as toxinas podem afetar o ecossistema aquático. Por exemplo, as cianotoxinas podem influenciar a estrutura do zooplâncton, especialmente quando as cianobactérias dominam o fitoplâncton. [Azevedo, 1998] Essas toxinas podem se acumular nos tecidos dos peixes ao longo do tempo, e a exposição contínua pode levar a níveis tóxicos que afetam múltiplos órgãos, como fígado, coração, rim, brânquias.

A Portaria GM/MS Nº 888, de 4 de maio de 2021 [da Saúde, 2021], do Ministério da Saúde estabelece que o monitoramento de cianotoxinas pode substituir o de clorofila-a, desde que a contagem de células de cianobactérias no ponto de captação não ultrapasse 10.000 células/mL. Se essa contagem exceder 20.000 células/mL, é necessário realizar análises semanais para microcistinas, saxitoxinas e cilindrospermopsinas. Os valores máximos permitidos para essas cianotoxinas são 1,0 μg/L para cilindrospermopsinas e microcistinas (MCYST-LR), e 3,0 μg/L para saxitoxinas. A frequência de monitoramento das cianobactérias varia conforme a contagem, sendo trimestral para até 10.000 células/mL e semanal para contagens superiores a esse valor.

2.5.35 Macrófitas

As macrófitas são plantas que podem ser vistas a olho nu, com partes que realizam fotossíntese e forma permanente ou sazonal, totalmente ou parcialmente submersas, ou até mesmo flutuando na água. [Pompêo, 2008], [Rgang & Gastal Jr, 1996]Elas são essenciais para diversas espécies aquáticas, incluindo os zooplâncton, perifíton, peixes e sapos, oferecendo habitat e refúgio. Além disso, contribuem para manutenção de fósforo e a produção de carbono orgânico. Essas funções afetam diretamente a hidrologia e a dinâmica dos sedimentos em ecossistemas de água doce, afetando o fluxo de água. [Diniz *et al.*, 2005]

Podem ser usados como marcadores de qualidade da água, pois desempenham várias funções importantes, incluindo acumular biomassa, acelerar o ciclo de nutrientes, afetar a química da água, servir como substrato para algas e sustentar a cadeia de detritos e herbivoria. Desta maneira, as macrófitas possuem papel importante na na estruturação e dinâmica dos ecossistemas aquáticos. [Hegel & Melo, 2016]

Porém, a constante poluição pode ocasionar a reprodução desenfreada das macrófitas que é prejudicial como aumento das taxas de trocas de gás da água com a atmosfera, redução da biodiversidade,impedimentos das atividades recreativas como pesca, recreação, navegação e podendo até ser prejudicial a saúde, uma vez que há a formação de ambiente propícios à reprodução de vetores de doenças de veiculação hídrica. [Xavier *et al.*, 2021]

Na base de dados disponibilizada há duas medidas que envolvem coleta de amostras, identificação e quantificação para a comunidade de macrófitas. A primeira é a riqueza total de macrófitas refere-se ao número total de espécies diferentes de macrófitas encontradas em uma amostra ou em uma área específica. A unidade $taxamst - 1$ se refere ao número de táxons (espécies, gêneros ou famílias) por unidade de área. Já a segunda medida biomassa de macrófitas refere-se à quantidade de massa vegetal presente em uma unidade de área, geralmente expressa em gramas por metro quadrado (g/m^2).

A Resolução CONAMA 357/2005 [18] estabelece critérios para a classificação das águas e padrões de qualidade relacionados à proteção dos corpos d'água e ecossistemas aquáticos, e a Portaria GM/MS nº 888/2021 do Ministério da Saúde estabelece padrões de potabilidade da água para consumo humano, incluindo limites para diversos parâmetros de qualidade da água. No entanto, não há nenhuma disposição normativa brasileira a respeito da quantificação da riqueza ou biomassa de macrofitas em corpos d'água, ficando assim com abordagem em diretrizes técnicas e manuais por órgãos ambientais e institutos de pesquisa locais.

[18] Disponível em: `https://conama.mma.gov.br/option=com_sisconama&task=arquivo.download&id=450`

2.5.36 Comunidade Bentônica

Os macroinvertebrados bentônicos desempenham papéis importantes na manutenção dos processos ecológicos nos riachos [Linares *et al.*, 2018]. Participam ativamente na degradação da matéria orgânica que se acumula nos cursos de água, desempenhando um papel de fundamental importância na ciclagem de nutrientes e nas cadeias tróficas aquáticas [Castro *et al.*, 2019]. Assim, o estudo da composição e estrutura das comunidades bentônicas permite avaliar a integridade dos ecossistemas de água doce como um todo [Stoddard *et al.*, 2006].

Os macroinvertebrados bentônicos são organismos que habitam o fundo dos corpos d'água, como rios, riachos e lagos, e possuem uma grande diversidade de formas e funções ecológicas. Esses organismos, que incluem insetos aquáticos, moluscos, crustáceos, entre outros, desempenham funções adicionais em ecossistemas aquáticos. Além de sua participação na manipulação da matéria orgânica, os macroinvertebrados bentônicos também ajudam na aeração do solo e no processo de reciclagem de nutrientes. Ao se alimentarem de matéria orgânica, eles transformam resíduos em nutrientes que são reabsorvidos pelas plantas aquáticas e outros organismos, promovendo a saúde e o equilíbrio do ecossistema [Linares *et al.*, 2018].

A composição e estrutura das comunidades bentônicas refletem diretamente as condições ambientais e a qualidade da água. Como bioindicadores, esses organismos podem fornecer informações úteis sobre a saúde do ecossistema aquático. Alterações no número e na diversidade de macroinvertebrados bentônicos podem indicar estresses ambientais, como poluição, mudanças na temperatura da água, redução de oxigênio ou alteração nos fluxos de água. A ausência ou ausência de espécies sensíveis pode ser um indicativo de manipulação ambiental, enquanto a diversidade de formas e funções dentro da comunidade é um reflexo da estabilidade ecológica do habitat aquático [Castro *et al.*, 2019].

Estudos sobre comunidades bentônicas, como os realizados por Stoddard *et al.* [2006], marcaram a importância desses organismos no monitoramento da qualidade da água. Ao analisar a diversidade e a abundância dessas comunidades, é possível identificar padrões de eliminação e de recuperação ambiental em ecossistemas aquáticos. Isso torna os macroinvertebrados bentônicos uma ferramenta essencial no planejamento e na implementação de políticas de conservação e manejo de recursos hídricos. A monitorização dessas comunidades, portanto, não oferece apenas uma compreensão profunda da saúde dos ambientes aquáticos, mas também permite intervenções mais específicas na proteção e restauração desses ecossistemas ecológicos.

2.5.37 Comunidade Fitoplanctônica

O fitoplâncton é composto por organismos que são organizados em colônias e filamentos ou que são formados por uma única célula, como as algas microscópicas.

Embora alguns possam ter estruturas de locomoção, como flagelos, a turbulência, as correntes e a densidade controlam os movimentos da coluna d'água. A clorofila e outros pigmentos adicionais são fotoautotróficos e alimentam a maioria destes organismos. [CRUZ, 2004] São constituídos principalmente por algas: clorofíceas, diatomáceas, euglenofíceas, crisofíceas, dinofíceas e xantofíceas e cianobactérias. [Companhia Ambiental do Estado de São Paulo (CETESB), 2016]

É importante destacar que as algas fitoplanctônicas são essenciais para avaliar as condições ambientais porque estão presentes na água doce. Desde o início do século passado, muitos estudos sobre algas como indicadores da qualidade da água foram realizados em várias partes do mundo. [Gentil *et al.*, 2008]. Os fitoplânctons são uns dos principais parâmetros na avaliação ambiental nos programas de monitoramento nos últimos anos.É uma comunidade que mostra o estado trófico. Além disso, a presença de espécies resistentes ao cobre em reservatórios usados para abastecimento pode indicar a poluição por pesticidas ou metais tóxicos.

2.5.38 Comunidade Zooplanctônica

A comunidade zooplanctônica é composta por organismos microscópicos que vivem em suspensão na coluna d'água, sendo formada principalmente por protozoários, rotíferos, cladóceros e copépodes, os grupos dominantes em ambientes de água doce. Esses organismos exercem papel fundamental na regulação da comunidade fitoplanctônica, uma vez que utilizam o fitoplâncton como principal fonte de alimento, além de participarem ativamente da reciclagem de nutrientes. Dessa forma, o zooplâncton contribui para a manutenção do equilíbrio ecológico dos ecossistemas aquáticos e integra a base da cadeia alimentar de diversos peixes [Esteves, 1998, Wetzel, 2000].

Além de atuar como consumidores primários, o zooplâncton também influencia a transparência da água, a dinâmica de nutrientes e a estrutura da comunidade microbiana. Sua presença e abundância são sensíveis a variações ambientais, como temperatura, disponibilidade de alimento, predadores e alterações hidrológicas, sendo considerados bioindicadores da qualidade da água. [De-Carli *et al.*, 2018]

3 MATERIAL E MÉTODOS

3.1 Cálculo do Índice de Qualidade da Água

O Índice de Qualidade da Água (IQA) foi calculado de acordo com a metodologia apresentada no site oficial do órgão [Instituto Mineiro de Gestão das Águas, (IGAM), 2025], referente ao monitoramento das águas superficiais de Minas Gerais. O IQA corresponde a um valor que varia de 0 a 100 e é calculado com base em nove parâmetros, ponderados de acordo com sua importância relativa: oxigênio dissolvido (OD), coliformes fecais, pH, nitratos (NO_3), fosfatos (PO_4), demanda bioquímica de oxigênio (DBO), turbidez, temperatura da água e resíduos totais. O índice é obtido pelo produtório ponderado das qualidades individuais de cada parâmetro, conforme a equação 3.1.1:

$$IQA = \prod_{i=1}^{n} q_i^{w_i} \tag{3.1.1}$$

onde n é o número de parâmetros considerados, q_i é a qualidade do parâmetro obtida a partir da curva média específica de qualidade, e w_i é o peso associado ao parâmetro.

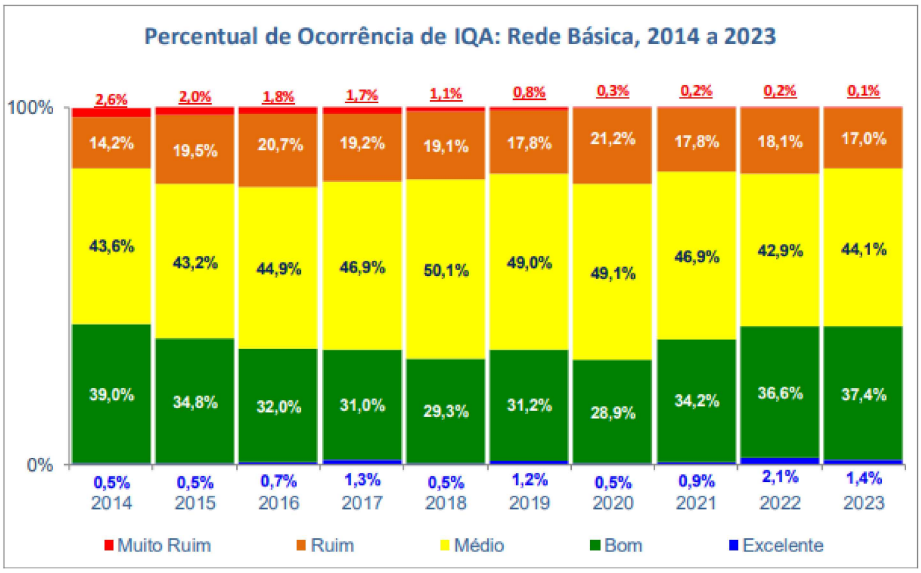
A tabela 3.1 apresenta os pesos atribuídos a cada parâmetro, conforme a metodologia do IGAM.

Tabela 3.1 – Pesos dos parâmetros utilizados no cálculo do IQA [Instituto Mineiro de Gestão das Águas, (IGAM), 2025]

Parâmetro	Peso (w_i)
Oxigênio dissolvido – OD (% OD)	0,17
Coliformes fecais (NMP/100 mL)	0,15
pH	0,12
Nitratos (mg/L NO_3)	0,10
Fosfatos (mg/L PO_4)	0,10
Demanda Bioquímica de Oxigênio – DBO (mg/L)	0,10
Turbidez (UNT)	0,08
Resíduos totais (mg/L)	0,08
Variação da temperatura da água (°C)	0,10

As equações obtidas para o Sistema de Cálculo de Qualidade da Água, bem como as curvas médias de qualidade dos nove parâmetros individuais, também podem ser consultadas em [SEMAD & II, 2005].

Figura 3.2 – Percentual do IQA, retirado do Relatório IGAM 2014 a 2023.



Fonte: [Instituto Mineiro de Gestão das Águas, 2024].

Figura 3.1 – Classificação do Índice de Qualidade das Águas – IQA

Nível de Qualidade	Faixa
Excelente	90 < IQA ≤ 100
Bom	70 < IQA ≤ 90
Médio	50 < IQA ≤ 70
Ruim	25 < IQA ≤ 50
Muito Ruim	0 ≤ IQA ≤ 25

Fonte: [Instituto Mineiro de Gestão das Águas, (IGAM), 2025].

Verificando os percentuais de variação das faixas de IQA entre os anos de 2022 e 2023, observou-se melhoria da qualidade das águas em oito bacias hidrográficas do estado de Minas Gerais: Itapemerim, Itanhém, Jequitinhonha, Pardo, Mucuri, Doce, Paraíba do Sul e São Francisco.

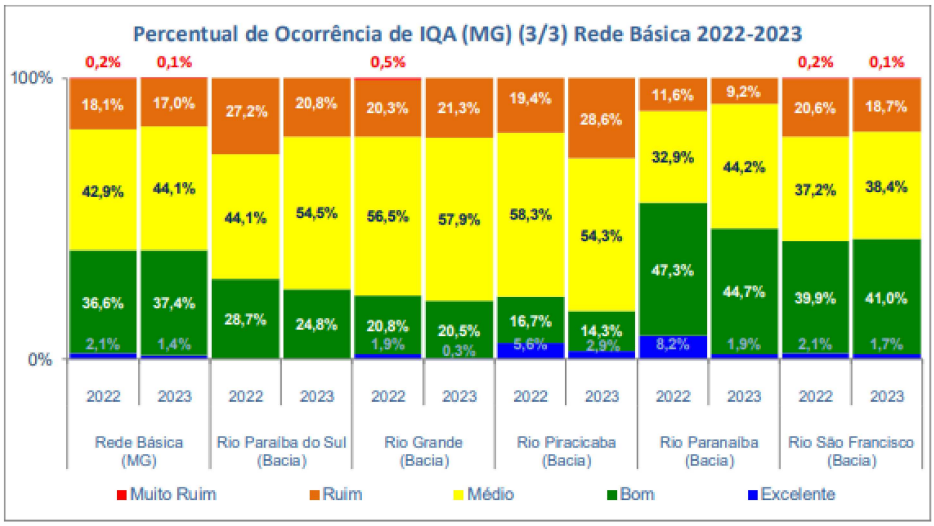


Figura 3.3 – Percentual do IQA, retirado do Relatório IGAM de 2022 a 2023 [Instituto Mineiro de Gestão das Águas, 2024]

3.2 Ferramentas Utilizadas

As análises foram realizadas utilizando a linguagem Python e R, além de bibliotecas especializadas como Scikit-learn, Pandas, NumPy, Matplotlib e outras. O desenvolvimento e a validação dos modelos foram realizados em ambientes computacionais como o Google Colaboratory, proporcionando flexibilidade e facilidade na execução e compartilhamento dos códigos.

O conjunto de dados foi dividido em subconjuntos de treinamento e teste. A função *train-test-split* da biblioteca Scikit-learn foi utilizada para gerar um subconjunto de teste com 20% dos dados, enquanto os 80% restantes foram usados para treinamento.

Para avaliar de forma abrangente o desempenho dos modelos, foi adotada a técnica de validação cruzada do tipo *ShuffleSplit*, que realiza divisões aleatórias do conjunto de dados em subconjuntos de treinamento e teste a cada iteração. Em cada divisão, 80% dos dados foram utilizados para o treinamento e 20% para o teste, garantindo que todas as amostras participassem de ambas as etapas em diferentes rodadas.

Foram realizadas dez divisões (*splits*), e as métricas de desempenho apresentadas correspondem à média dos resultados obtidos nessas repetições. Nesse procedimento, não foi utilizado um conjunto de teste final independente, sendo toda a base de dados incorporada nas diferentes divisões geradas pelo ShuffleSplit, devido à limitação do tamanho do banco de dados e à baixa variabilidade dos dados disponíveis, fatores que tornariam inviável a separação de um conjunto de teste completamente independente.

3.3 Seleção de Parâmetros

- **S1: Medições com Tiras Reagentes**

Figura 3.4 – Exemplo de tiras reagentes utilizadas para análise de parâmetros de qualidade da água.

Fonte: Aliexpress (2025).

Para a técnica de medição com tiras reagentes, os parâmetros considerados incluem pH, alcalinidade ($CaCO_3$ mg L^{-1}), nitrito (mg L^{-1}), nitrato (mg L^{-1}), cromo hexavalente (mg L^{-1}), cromo trivalente (mg L^{-1}), dureza ($CaCO_3$ mg L^{-1}), mercúrio (mg L^{-1}), cobre dissolvido (mg L^{-1}), cromo total (mg L^{-1}), ferro solúvel (mg L^{-1}) e ferro total (mg L^{-1}). Esses parâmetros são medidos por meio das tiras de teste, como ilustrado na figura 3.4 que permitem avaliar visualmente as concentrações dessas substâncias na água.

Para este trabalho, foram considerados apenas os parâmetros presentes tanto nas tiras quanto no banco de dados utilizado, garantindo a compatibilidade entre as medições e as análises realizadas.

- **S2: Índice de Qualidade da Água (IQA)**

Para a técnica do índice de qualidade da água (IQA), os parâmetros selecionados incluem oxigênio dissolvido (O.D., mg L^{-1}), pH, coliformes totais (NMP 100 mL^{-1}), demanda bioquímica de oxigênio (DBO, mg L^{-1}), nitrato (mg L^{-1}), fósforo total (mg L^{-1}), turbidez (NTU) e sólidos totais (mg L^{-1}). Esses parâmetros são essenciais para uma avaliação abrangente da qualidade da água e são utilizados no cálculo do IQA, que reflete a saúde geral dos corpos hídricos.

- **S3: Modelo de Gradient Boosting**

No modelo de Gradient Boosting, após o treinamento, foi realizada a análise de importância das variáveis, com base nos pesos internos atribuídos pelo algoritmo. As variáveis que apresentaram maior contribuição para a predição do IQA foram: coliformes totais (NMP 100 mL^{-1}), turbidez (NTU), *E. coli* (NMP 100 mL^{-1}), sólidos suspensos (mg L^{-1}), oxigênio dissolvido (O.D., mg L^{-1}), alumínio dissolvido (mg L^{-1}), alcalinidade ($CaCO_3$ mg L^{-1}) e pH. Essas variáveis foram consideradas as mais influentes, mas o modelo foi aplicado integralmente utilizando todos os parâmetros disponíveis, permitindo avaliar sua capacidade preditiva completa.

- **S4: Modelo de Floresta Aleatória (*Random Forest*)**

De forma análoga, a análise de importância das variáveis no modelo de Floresta Aleatória (do inglês, Random Forest, RF) indicou como mais relevantes: coliformes totais (NMP 100 mL^{-1}), turbidez (NTU), *E. coli* (NMP 100 mL^{-1}), sólidos suspensos (mg L^{-1}), oxigênio dissolvido (O.D., mg L^{-1}), alumínio dissolvido (mg L^{-1}), cobre dissolvido (mg L^{-1}) e pH. Essas variáveis apresentaram maior peso nas decisões internas das árvores, refletindo maior influência na predição do IQA. Assim como no Gradient Boosting, o modelo foi treinado e avaliado utilizando a totalidade dos parâmetros, garantindo uma análise abrangente do desempenho do algoritmo.

- **S5: Todos parâmetros**

De modo geral, a análise utilizou todos os parâmetros disponíveis (físicos, químicos e biológicos) presentes no conjunto de dados, totalizando 48 parâmetros. Ao aproveitar toda essa gama de variáveis, foi possível avaliar de forma abrangente a qualidade da água por meio de diversas técnicas. Essa abordagem holística garante uma avaliação completa dos fatores que afetam a qualidade da água.

3.4 Pré-processamento do Banco de Dados

O sucesso dos algoritmos de aprendizagem de máquina depende da qualidade dos dados utilizados. Assim, as etapas de pré-processamento foram conduzidas para garantir a integridade e a consistência dos dados. Inicialmente, os dados brutos do IQA foram avaliados quanto à presença de valores ausentes, outliers e inconsistências. Os valores faltantes foram inferidos utilizando técnicas considerando a relevância de cada parâmetro ambiental. Os detalhes podem ser vistos na Seção **3.5.1**.

Os dados foram normalizados utilizando a técnica *Min-Max Scaling*, reescalando todas as variáveis para o intervalo $[0, 1]$. Essa etapa é essencial para evitar que variáveis com

magnitudes superiores dominem os cálculos realizados pelos algoritmos de aprendizagem de máquina, garantindo que todas as características contribuam de forma equilibrada para os modelos. A normalização foi aplicada após a inferência dos valores faltantes, de forma que os dados completos fossem reescalados de maneira consistente. Essa transformação foi realizada por meio da função `MinMaxScaler` da biblioteca `scikit-learn` em Python.

3.5 Valores Faltantes

Durante o processo de limpeza e preparação dos dados, identificou-se uma dificuldade na remoção das linhas com valores faltantes. Ao excluir uma linha com valor ausente em determinada variável, frequentemente outra variável apresentava ausência em uma linha diferente, o que acarretava na perda de uma quantidade considerável de dados. Como a base de dados possui apenas 737 registros, essa perda se tornava ainda mais relevante, prejudicando a análise. Por essa razão, optou-se por não remover diretamente as linhas com valores faltantes, mas sim considerar abordagens alternativas, como técnicas de inferência ou métodos que toleram dados ausentes.

Em particular, observou-se que aproximadamente 30% dos valores do Índice de Qualidade da Água (IQA), correspondentes a 227 registros, estavam ausentes. A quantidade de ausentes para cada parâmetro estão no quadro 3.2. Para evitar a perda dessas observações, optou-se por preencher os valores faltantes do IQA utilizando a fórmula de cálculo descrita na seção anterior, a partir das variáveis físico-químicas e biológicas disponíveis. Dessa forma, buscou-se preservar o tamanho e a representatividade da amostra.

Para as demais variáveis, consideraram-se abordagens alternativas, como técnicas de inferência estatística, de modo a minimizar o impacto da incompletude sobre o desempenho dos algoritmos.

Parâmetro	Valores Faltantes	Parâmetro	Valores Faltantes
IQA	227	Manganês	4
Cobre dis.	163	Riq. Tot. zooplanctônica	4
Óleos e graxas	100	D.Q.O.	4
Clorofila a	100	pH	4
Alumínio dis.	68	Turbidez	4
Fósf. total	67	O.D.	4
Cromo tot.	57	Sól. tot.	3
Cianetos	53	Sól. tot. dis.	3
Cromo triv.	52	Temp. da água	3
Níquel	52	Sól. susp.	3
Cádmio tot.	52	Nitrito	3
Mercúrio	52	Dureza (CaCO3)	3
Cromo hexa.	52	Ferro tot.	3
Zinco	36	Silicatos	3
Temp. do ar	35	Ortofosfato	3
Riq. tot. bentônica	32	Alcalinidade (CaCO3)	3
Densi. tot. bentônica	11	Nitr. amoniacal	3
E. Coli	6	Nitrato	3
Fenóis	6	D.B.O.	3
Condutividade	4	Riq. tot. macrófitas	3
Densi. Tot. zooplanctônica	4	Densi. tot. cianobactérias	3
Cor	4	Ferro sol.	3
Cloretos	4	Riq. tot. fitoplanctônica	3
Coliformes tot.	4	Densi. tot. fitoplanctônica	3

Quadro 3.2 – Parâmetros e Valores Faltantes

3.5.1 Inferência de Valores Faltantes pelo Qui-Quadrado

A inferência de valores faltantes é um passo crucial no pré-processamento de dados. No contexto deste trabalho, utilizamos o teste do qui-quadrado para estimar e preencher valores ausentes.

O qui-quadrado (χ^2) é uma medida estatística que avalia a diferença entre as frequências observadas e as frequências esperadas em um conjunto de dados categorizado. A fórmula para o teste do qui-quadrado é dada por:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.5.1}$$

onde O_i são as frequências observadas e E_i são as frequências esperadas. Este teste é aplicado para cada variável com valores faltantes, permitindo-nos inferir valores com base nas relações observadas.

3.5.2 Normalização por Min-Max

A normalização por Min-Max é uma técnica comum para escalar os valores de um conjunto de dados para um intervalo específico, geralmente $(0, 1)$. Isso é feito através da seguinte fórmula:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{3.5.2}$$

onde X é o valor original, $\min(X)$ é o valor mínimo da variável e $\max(X)$ é o valor máximo da variável.

Esta técnica garante que todas as variáveis permaneçam na mesma escala, e trata-se de um passo essencial para os algoritmos de inteligência artificial utilizados neste estudo.

3.5.3 Copulas

Introduzido por Sklar em 1959 [Sklar, 1959], o conceito de cópulas desempenha um papel fundamental na estatística multivariada ao possibilitar a modelagem da dependência entre variáveis univariadas. Quando as distribuições marginais individuais são conhecidas, as cópulas permitem estabelecer uma conexão com distribuições conjuntas multivariadas [Nelsen, 2007], criando um estrutura conceitual consistente para análise.

Seja um vetor aleatório de dimensão n, representado por $X = (X_1, ..., X_n)$. O Teorema de Sklar [Sklar, 1959] afirma que existe uma função de distribuição conjunta H, associada às distribuições marginais $F_{X1}, ..., F_{Xn}$. Assim, é possível definir uma cópula n-dimensional $C : [0, 1]^d \to [0, 1]$, que vincula a distribuição conjunta às suas respectivas marginais, conforme ilustrado na Equação (3.5.3):

$$H(x_1, x_2, ..., x_n) = C(F_{X1}(x_1), F_{X2}(x_2), ..., F_{Xn}(x_n)) \tag{3.5.3}$$

Esse teorema demonstra que, para qualquer função de distribuição conjunta F cujas marginais sejam $F_{X1}, ..., F_{Xn}$, existe uma cópula C que satisfaz a relação acima. Quando F é contínua, essa cópula é única [Nelsen, 2007].

Essas ferramentas são amplamente utilizadas na geração de dados sintéticos [Silva, 2020], pois, além de preservar as distribuições marginais das variáveis, asseguram que as interdependências entre elas sejam mantidas. Esse aspecto é essencial para garantir que os dados gerados mantenham a integridade da estrutura multivariada.

Ao contrário dos métodos tradicionais, as cópulas oferecem vantagens na inferência de dados ausentes, mesmo em cenários onde as dependências entre variáveis são não lineares ou complexas. Ademais, sua invariância às transformações das distribuições marginais proporciona maior estabilidade na inferência estatística, especialmente em situações onde há diferença significativa nas escalas das variáveis [Nelsen, 2007].

- **Métricas de Avaliação**

Diversas métricas foram utilizadas para avaliar o desempenho dos modelos. Estas incluem:

- **Erro Quadrático Médio (MSE)**

O Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*) quantifica a média dos quadrados das diferenças entre os resultados reais observados e os resultados previstos pelo modelo. É calculado da seguinte forma:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.5.4}$$

onde y_i representa os valores reais, $\hat{y}_i$ representa os valores previstos, e n é o número de observações.

- **Raiz do Erro Quadrático Médio (RMSE)**

A Raiz do Erro Quadrático Médio (RMSE, do inglês *Root Mean Squared Error*) é a raiz quadrada da média dos erros quadráticos. Ela indica o desvio padrão dos resíduos, fornecendo uma noção de quão bem as previsões do modelo se ajustam aos dados reais. É definida como:

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{3.5.5}$$

- **Erro Absoluto Médio (MAE)**

O Erro Absoluto Médio (MAE, do inglês *Mean Absolute Error*) mede a magnitude média dos erros em um conjunto de previsões, sem considerar a direção dos erros. É a média das diferenças absolutas entre a previsão e a observação real sobre a amostra de teste. É dado por:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.5.6}$$

- **Coeficiente de Determinação (R^2)**

O Coeficiente de Determinação (R^2) representa a proporção da variância na variável dependente que é previsível a partir das variáveis independentes. É calculado como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5.7}$$

onde $\bar{y}$ é a média dos valores reais.

- **Precisão**

A precisão mede a proporção de identificações verdadeiramente positivas entre todas as identificações positivas feitas pelo modelo. É dada por:

$$\text{Precisão} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}} \tag{3.5.8}$$

- **Pontuação F1**

A pontuação F1 média é obtida calculando essa pontuação para cada classe e, em seguida, tirando a média (macro ou ponderada) entre elas. A fórmula da pontuação F1 é:

$$F1 = 2 \cdot \frac{\text{precisão} \cdot \text{recall}}{\text{precisão} + \text{recall}} \tag{3.5.9}$$

- **Recall**

O recall mede a capacidade do modelo em identificar corretamente todos os exemplos relevantes (verdadeiros positivos) entre todos os exemplos que realmente pertencem a uma determinada classe (verdadeiros positivos + falsos negativos).

$$\text{Recall} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}} \tag{3.5.10}$$

- **Melhor Precisão**

A melhor precisão indica a maior taxa de precisão entre todas as classes. Esse valor pode destacar o desempenho máximo alcançado em alguma classe específica, mesmo que a média geral seja inferior.

- **Tempo Total de Execução**

O tempo total de execução mede o tempo necessário para treinar o modelo e realizar as previsões. Essa métrica é essencial para compreender a eficiência computacional dos modelos.

3.6 Seleção e Justificativa dos Algoritmos

Para atender ao objetivo de prever e classificar o Índice de Qualidade da Água (IQA) do Rio Paraibuna, foram selecionados algoritmos de aprendizagem de máquina exclusivamente reconhecidos pela robustez e aplicabilidade em análises ambientais e preditivas. Os modelos considerados neste estudo incluem: Regressão Linear, Regressão Logística, Máquinas de Vetor de Suporte (SVM), Árvores de Decisão, Extra Trees, Floresta Aleatória, Gradient Boosting, Extreme Gradient Boosting (XGBoost), LightGBM, Redes Neurais Artificiais (MLP) e K-Vizinhos Mais Próximos (KNN).

A escolha dos algoritmos foi fundamentada em sua capacidade de lidar com dados ambientais que apresentam características heterogêneas, como variabilidade temporal, não linearidade e a coexistência de variáveis categóricas e contínuas. Esses algoritmos possuem propriedades específicas que os tornam adequados para modelar o IQA, permitindo uma análise comparativa de desempenho. Por exemplo, enquanto a Regressão Linear serve como um modelo base para identificação de padrões lineares, o Random Forest e o XGBoost são particularmente eficazes na captura de relações não lineares, complexos e interações entre variáveis. O SVM , por sua vez, é recomendado para conjuntos de dados com alta dimensionalidade e permite o uso de kernels para resolver problemas não lineares. O MLP permite modelar relações complexas e não lineares por meio de camadas ocultas; e KNN é útil para padrões locais e proximidade entre amostras. Essa combinação permite uma análise comparativa de desempenho e adequação à variabilidade temporal, heterogeneidade e coexistência de variáveis contínuas e categóricas nos dados ambientai

3.7 Validação e Aplicação do Modelo

Após o treinamento e a validação, os resultados dos modelos foram comparados com os dados reais do IQA para garantir a precisão da precisão. Também foi realizada uma análise de sensibilidade para identificar quais variáveis têm maior impacto no índice de qualidade da água. Esses resultados serão fundamentais para apoiar gestores ambientais na priorização de ações e na mitigação de impactos ambientais na bacia do Rio Paraibuna.

4 Resultados

Esta seção apresenta os resultados obtidos a partir da aplicação dos diferentes métodos de inferência e modelagem de aprendizagem de máquina aos conjuntos de dados deste estudo. Inicialmente, analisou-se a distribuição das categorias do Índice de Qualidade da Água (IQA) nos dados originais e após os procedimentos de inferência e geração de dados sintéticos. Em seguida, são apresentados os resultados de desempenho dos modelos de regressão e classificação para cada um dos métodos testados.

4.1 Estatísticas

Com o objetivo de compreender o comportamento das variáveis ambientais e biológicas que compõem o conjunto amostral realizou-se uma análise exploratória dos dados. O Quadro 4.1 apresenta as estatísticas descritivas dessas variáveis, incluindo medidas de tendência central (média e mediana), dispersão (desvio padrão) e amplitude (valores mínimo e máximo).

A partir da análise estatística e dos resultados obtidos pelos modelos de inferência e predição, observou-se que determinados parâmetros apresentaram maior relevância para a modelagem do IQA. Essa constatação foi reforçada pela análise de importância das variáveis realizada nos modelos de Gradient Boosting e Floresta Aleatória, em que os parâmetros turbidez, oxigênio dissolvido, pH, sólidos suspensos e coliformes totais apresentaram os maiores pesos na predição do índice.

Por sua vez, parâmetros como cádmio, mercúrio, fenóis e óleos e graxas, embora relevantes do ponto de vista ambiental, apresentaram baixa frequência de medição e menor contribuição preditiva, sendo mais indicados para campanhas específicas de monitoramento de poluentes industriais do que para o monitoramento rotineiro da qualidade da água.

Quadro 4.1 – Estatísticas descritivas das variáveis ambientais

Estatística	Temp. ar	Temp. água	Cond.	pH	Turbidez	O.D.	Sól. Totais	Ortofosfato	Fósf. Total	Silicatos
Count	702	734	733	733	733	733	734	734	670	734
Mean	25,18	23,52	50,58	7,34	41,50	7,52	53,19	0,028	0,078	9,25
Std	3,68	2,65	16,47	0,71	50,63	1,74	50,44	0,020	0,083	4,31
Min	2	14,07	0	4,15	1,2	3,14	1,5	0	0	0
25%	23,2	21,66	40	6,98	12	6,45	19	0,017	0,047	6,91
50%	25	23,73	49	7,37	22,2	7,32	42	0,024	0,066	9,89
75%	27,5	25,3	61	7,7	46,9	8,18	65	0,035	0,087	11,10
Max	39,9	32,7	109,3	10,1	544	15,61	350	0,204	1,59	80,05
Estatística	Riq. Bentônica	Dens. Bentônica	Riq. Fitoplanctônica	Dens. Fitoplanctônica	Dens. Cianobactérias	Riq. Macrófitas	Biom. Macrófitas	Riq. Zooplanctônica	Dens. Zooplanctônica	IQA
Count	705	726	734	734	734	734	735	733	733	510
Mean	3,63	51,16	14,44	1139,07	5708,39	0,86	8,20	9,05	28,68	70,34
Std	4,69	128,66	7,71	2164,47	35550,75	1,45	91,55	5,78	126,53	8,69
Min	0	0	0	0	0	0	0	0	0	4
25%	1	0	9	54	0	0	0	5	1,2	65
50%	2	8	14	519,45	119	0	0	8	3,07	72
75%	5	40	18	1213,15	3671,22	1	0	12	9	76
Max	49	1216,67	51,66	25047,51	673981,18	7	2300	40	1651,51	90

4.2 Distribuição das categorias do IQA

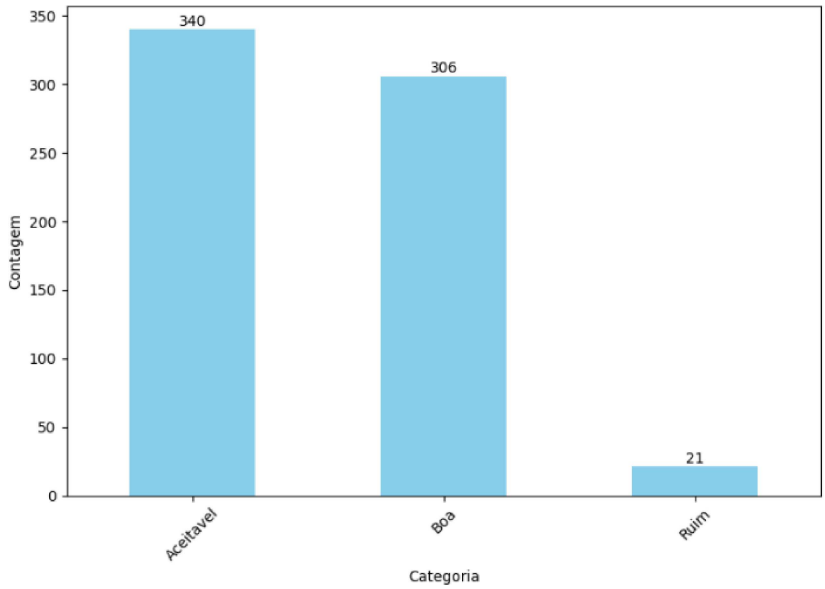
O Índice de Qualidade da Água (IQA), em sua formulação original, contempla cinco categorias de classificação. Contudo, no conjunto de dados analisado, apenas três delas estavam presentes: Boa, Aceitável e Ruim. Inicialmente, a base continha 737 registros, mas, após o processo de limpeza, que envolveu a remoção de valores atípicos, linhas em branco, campos com espaços ou preenchimentos incorretos, o total reduziu-se para 668 amostras válidas.

A distribuição no conjunto real dos dados revelou um desnível significativo: a categoria *Ruim* foi drasticamente sub-representada, com apenas 21 amostras, em comparação com *Aceitável* (340 amostras) e *Boa* (306 amostras), totalizando 667 amostras. Essa diferença pode sugerir que o processo de inferência favoreceu as categorias que tinham um maior número de amostras originalmente, o que pode prejudicar a eficácia dos modelos na previsão de classes menos representativas, além de comprometer a integridade estatística das análises inferenciais.

Para mitigar esse desbalanceamento, foi implementada a geração de dados sintéticos utilizando cópulas, conforme descrito na seção anterior. A adição sintética resultou em um total de 967 amostras, com uma distribuição mais uniforme entre as categorias: *Boa* (306 amostras), *Aceitável* (340 amostras) e *Ruim* (321 amostras). Essa abordagem contribuiu para a criação de um conjunto de dados mais equilibrado e representativo, favorecendo o desempenho e a estabilidade dos modelos de previsão do IQA.

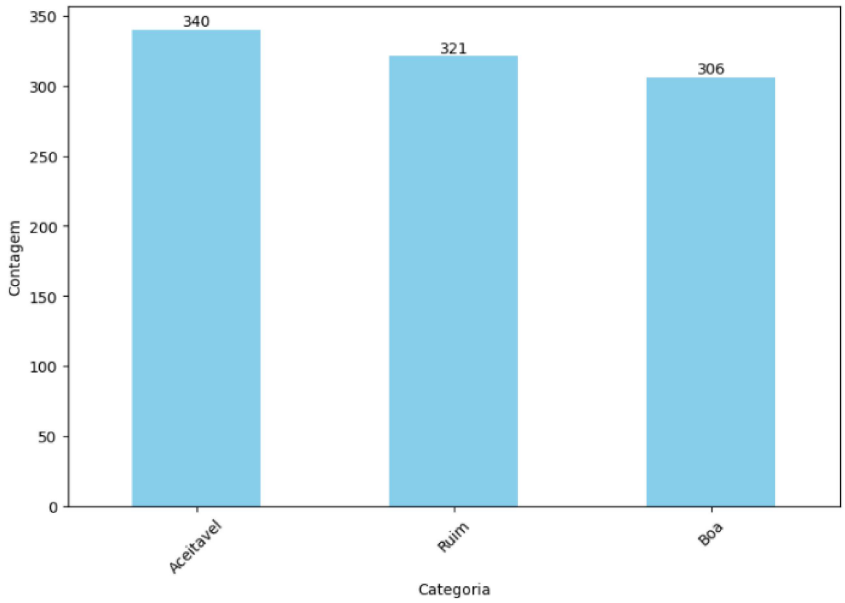
Esses achados podem ser visualizados nas Figuras 4.1 e 4.2, que apresentam dois gráficos. O primeiro, referente à base de dados original, evidencia o desequilíbrio entre as categorias, com a classe Ruim significativamente sub-representada. Já o segundo gráfico, correspondente à inferência sintética, mostra uma distribuição mais uniforme entre as três categorias. Essa visualização reforça a importância de selecionar métodos apropriados para o tratamento de dados sintéticos, considerando tanto o equilíbrio entre as classes quanto a manutenção da representatividade original das amostras.

Figura 4.1 – Contagem de Categorias IQA na base original - 667 amostras



Fonte: elaborado pela própria autora.

Figura 4.2 – Contagem de Categorias IQA - Inferência Sintética - 967 amostras



Fonte: elaborado pela própria autora.

A etapa de inferência de dados constituiu-se como o principal diferencial metodológico deste estudo. A aplicação das técnicas de inferência (qui-quadrado e cópulas) permitiu reconstruir a estrutura estatística dos dados ausentes e ampliar a representatividade amostral. Assim, os resultados apresentados a seguir refletem diretamente os ganhos advindos do processo de inferência, evidenciando como o tratamento adequado dos dados influencia o desempenho e a estabilidade dos modelos.

4.3 Resultados de Regressão

Esta seção apresenta os resultados da análise dos modelos de aprendizagem de máquina aplicados aos diferentes conjuntos de dados. Para garantir a consistência dos resultados e reduzir o risco de *overfitting*, foi empregada a técnica de validação cruzada do tipo *ShuffleSplit*, que realiza divisões aleatórias do conjunto em subconjuntos de treinamento e teste. Em cada iteração, 80% das amostras foram destinadas ao treinamento e 20% ao teste, assegurando que todas as instâncias participassem de ambas as etapas em diferentes rodadas. Foram realizadas dez divisões (*splits*), e as métricas apresentadas correspondem à média dos resultados obtidos nessas repetições.

Os experimentos foram conduzidos considerando três abordagens distintas de base de dados:

1. M1: a base real com inferência de dados faltantes por meio da distribuição do qui-quadrado;
2. M2: a base real com inferência de dados faltantes a partir de amostras sintéticas; e
3. M3: uma base expandida, composta pela base real acrescida de dados sintéticos gerados.

Os dados sintéticos foram produzidos antes da etapa de modelagem, utilizando o método de cópulas (ver Seção **3.5.3**), de modo a replicar as distribuições originais das variáveis reais, mantendo suas relações multivariadas. Essa estratégia foi empregada com o objetivo de aumentar a representatividade amostral e mitigar o desbalanceamento entre categorias observadas na base original.

O foco da análise foi duplo: avaliar o desempenho geral dos modelos de regressão e verificar os ganhos potenciais decorrentes do uso de dados sintéticos. Dessa forma, os resultados apresentados nas seções subsequentes refletem tanto a eficácia das técnicas de inferência e geração de dados quanto o comportamento preditivo dos algoritmos sob diferentes condições de amostragem.

No contexto da inferência qui-quadrado (M1), os modelos Extra Trees e Gradient Boosting apresentaram desempenho superior em termos de MSE e RMSE, destacando-se na redução de erros em comparação com abordagens mais simples, como Regressão Linear e Árvore de Decisão, como exibido no quadro 4.3. Ao considerar a inferência sintética (M2), o desempenho geral também melhorou, com o Light Gradient Boosting permanecendo como o modelo mais eficaz, demonstrando as menores médias de MSE e RMSE. Detalhes no quadro 4.4.

Por fim, quando amostras sintéticas foram adicionadas (M3), os modelos Gradient Boosting e CatBoost exibiram um desempenho robusto. No entanto, o Light Gradient

Boosting consolidou sua posição de liderança ao registrar as melhores métricas de MSE e MAE, como apresentado no quadro 4.5

Essas conclusões são confirmadas pelos quadros 1, 2 e 3, que mostram uma melhoria consistente nas métricas de desempenho com a adoção de métodos mais avançados e uma redução geral dos erros nos modelos mais complexos. A análise dos resultados sugere uma vantagem clara na utilização das técnicas de inferência para obter resultados mais precisos nos nossos conjuntos de dados.

De modo geral, observa-se que o desempenho dos modelos evoluiu de forma consistente à medida que a complexidade da inferência aumentou. A abordagem M1, baseada na inferência por qui-quadrado, apresentou ganhos iniciais; M2, com inferência sintética, houve redução significativa nos erros de previsão (MSE e RMSE); e M3, ao integrar dados reais e sintéticos, consolidou o melhor desempenho com aumento do coeficiente de determinação (R^2). Esse comportamento demonstra empiricamente o potencial das técnicas de inferência na melhoria da qualidade dos dados e reforça o diferencial metodológico do presente trabalho, que vai além da simples aplicação de modelos de aprendizagem de máquina, propondo uma estratégia de pré-processamento baseada em inferência estatística.

A tabela 4.2 apresenta uma legenda auxiliar que sintetiza as denominações utilizadas ao longo desta seção. Essa tabela tem o objetivo de facilitar a interpretação dos resultados apresentados nas tabelas subsequentes, permitindo uma rápida identificação das diferentes abordagens de modelagem (M1, M2 e M3) e dos conjuntos de simulação (S1 a S5).

Tabela 4.2 – Descrição resumida das abordagens de modelagem (M) e dos conjuntos de simulação (S).

Código	**Descrição**
M1	Base real com inferência de dados faltantes por meio da distribuição do qui-quadrado.
M2	Base real com inferência de dados faltantes a partir de amostras sintéticas.
M3	Base expandida composta pela base real acrescida de dados sintéticos gerados.
S1	Medições com Tiras Reagentes.
S2	Índice de Qualidade da Água (IQA).
S3	Modelo de Gradient Boosting.
S4	Modelo de Floresta Aleatória (*Random Forest*).
S5	Todos os parâmetros.

4.3.1 Regressão

Quadro 4.3 – Resultados para inferência qui-quadrado (M1), com 10 execuções.

Método	**Seleção**	**MSE**	**RMSE**	**MAE**	**R2**	**Tempo total (s)**
Regressão Linear	S3	0.0188 ± 0.0030	0.1368 ± 0.0112	0.1079 ± 0.0070	0.4224 ± 0.0739	0.0803
Árvore de decisão	S5	0.0135 ± 0.0022	0.1157 ± 0.0095	0.0763 ± 0.0053	0.5843 ± 0.0757	0.1867
K-Vizinhos Mais Próximos	S3	0.0123 ± 0.0026	0.1104 ± 0.0121	0.0790 ± 0.0067	0.6197 ± 0.0790	0.0923
Máquina de vetores de suporte	S5	0.0105 ± 0.0021	0.1018 ± 0.0104	0.0782 ± 0.0069	0.6760 ± 0.0688	0.1605
Floresta aleatória	S5	0.0071 ± 0.0008	0.0840 ± 0.0050	0.0574 ± 0.0029	0.7808 ± 0.0315	12.3424
Árvores extras	S5	0.0066 ± 0.0013	0.0807 ± 0.0081	0.0562 ± 0.0026	0.7975 ± 0.0369	5.1888
Gradient Boosting	S5	0.0065 ± 0.0010	0.0807 ± 0.0064	0.0553 ± 0.0033	0.7978 ± 0.0333	4.6239
CatBoost	S5	0.00636 ± 0.00117	0.0794 ± 0.00726	0.0560 ± 0.00314	0.8043 ± 0.0336	10.2211
Extreme Gradient Boosting	S5	0.0061 ± 0.0012	0.0777 ± 0.0073	0.0521 ± 0.0032	0.8146 ± 0.0302	2.6501
Light Gradient Boosting	**S5**	**0.00597 ± 0.00117**	**0.07688 ± 0.00737**	**0.05026 ± 0.00246**	**0.8180 ± 0.0320**	**1.2140**

Quadro 4.4 – Resultados para inferência sintética (M2), com 10 execuções.

Método	**Seleção**	**MSE**	**RMSE**	**MAE**	**R2**	**Tempo total (s)**
Regressão Linear	S4	0.0188 ± 0.0029	0.1367 ± 0.0108	0.1071 ± 0.0065	0.4231 ± 0.0675	0.0529
Árvore de decisão	S5	0.0132 ± 0.0021	0.1145 ± 0.0090	0.0767 ± 0.0058	0.5934 ± 0.0637	0.1910
K-Vizinhos Mais Próximos	S3	0.0126 ± 0.0028	0.1115 ± 0.0127	0.0796 ± 0.0074	0.6117 ± 0.0839	0.0578
Máquina de vetores de suporte	S5	0.0108 ± 0.0022	0.1034 ± 0.0111	0.0793 ± 0.0071	0.6664 ± 0.0714	0.2758
Floresta aleatória	S4	0.0068 ± 0.0010	0.0820 ± 0.0062	0.0547 ± 0.0034	0.7909 ± 0.0371	3.9031
Árvores extras	S5	0.0064 ± 0.0014	0.0797 ± 0.0089	0.0555 ± 0.0037	0.8026 ± 0.0385	5.2750
CatBoost	S5	0.00623 ± 0.00108	0.0787 ± 0.00682	0.0549 ± 0.00342	0.8083 ± 0.0304	5.9623
Gradient Boosting	S5	0.0060 ± 0.0011	0.0770 ± 0.0076	0.0521 ± 0.0028	0.8160 ± 0.0328	4.7089
Extreme Gradient Boosting	S5	0.00583 ± 0.00121	0.07594 ± 0.00783	0.05025 ± 0.00313	0.8225 ± 0.0328	2.1401
Light Gradient Boosting	**S5**	**0.00576 ± 0.00106**	**0.07558 ± 0.00682**	**0.04892 ± 0.00331**	**0.8243 ± 0.0294**	**1.8547**

Quadro 4.5 – Resultados da adição de amostras sintéticas (M3), com 10 execuções.

Método	**Seleção**	**MSE**	**RMSE**	**MAE**	**R2**	**Tempo total (s)**
Regressão Linear	S4	0.0341 ± 0.0031	0.1844 ± 0.0087	0.1520 ± 0.0084	0.3093 ± 0.0259	0.0422
K-Vizinhos Mais Próximo	S4	0.0200 ± 0.0033	0.1408 ± 0.0117	0.1025 ± 0.0071	0.5947 ± 0.0600	0.0624
Máquina de vetores de suporte	S5	0.0162 ± 0.0017	0.1272 ± 0.0065	0.0974 ± 0.0049	0.6702 ± 0.0299	0.3602
Árvore de decisão	S5	0.0104 ± 0.0021	0.1017 ± 0.0103	0.0662 ± 0.0046	0.7889 ± 0.0328	0.3195
Floresta aleatória	S5	0.00587 ± 0.00166	0.0759 ± 0.0105	0.0511 ± 0.0052	0.8813 ± 0.0295	20.4596
Árvores extras	S5	0.0052 ± 0.0013	0.0713 ± 0.0087	0.0483 ± 0.0040	0.8955 ± 0.0218	7.6141
CatBoost	S5	0.0050 ± 0.0012	0.0700 ± 0.0081	0.0486 ± 0.0034	0.8994 ± 0.0210	10.2867
Gradient Boosting	S5	0.0048 ± 0.0014	0.0685 ± 0.0096	0.0463 ± 0.0043	0.9033 ± 0.0252	9.7162
XGBoost	S5	0.0037 ± 0.0007	0.0601 ± 0.0061	0.0430 ± 0.0033	0.9274 ± 0.0151	4.0084
LightGBM	**S5**	**0.0036 ± 0.00076**	**0.0600 ± 0.00623**	**0.0415 ± 0.00222**	**0.9276 ± 0.0155**	**2.4224**

4.3.2 Classificação

Quadro 4.6 – Resultados da classificação para inferência qui-quadrado (M1), com 10 execuções

Modelo	**Seleção**	**Precisão média**	**Pontuação F1 média**	**Precisão média**	**Recall médio**	**Melhor precisão**	**Tempo médio de execução (s)**
Regressão Logística	$S5$	0.760 ± 0.040	0.752 ± 0.039	0.766 ± 0.038	0.760 ± 0.040	0.836	0.036 ± 0.004
Máquina de vetores de suporte	$S5$	0.766 ± 0.039	0.761 ± 0.039	0.773 ± 0.037	0.766 ± 0.039	0.828	0.015 ± 0.001
Árvore de decisão	$S5$	0.790 ± 0.040	0.790 ± 0.041	0.792 ± 0.041	0.790 ± 0.040	0.873	0.017 ± 0.002
Multi-Layer Perceptron	$S5$	0.803 ± 0.028	0.800 ± 0.029	0.806 ± 0.029	0.803 ± 0.028	0.836	1.123 ± 0.470
Gradient Boosting	$S5$	0.851 ± 0.030	0.847 ± 0.029	0.852 ± 0.027	0.851 ± 0.030	0.896	1.686 ± 0.233
Extreme Gradient Boosting	$S5$	0.855 ± 0.030	0.852 ± 0.030	0.854 ± 0.032	0.855 ± 0.030	0.896	0.184 ± 0.082
Floresta aleatória	$S4$	0.857 ± 0.030	0.851 ± 0.029	0.860 ± 0.031	0.857 ± 0.030	0.933	0.209 ± 0.007
Árvore Extra	$S4$	0.862 ± 0.021	0.858 ± 0.022	0.859 ± 0.025	0.862 ± 0.021	0.910	0.152 ± 0.006
Light Gradient Boosting	$S5$	0.866 ± 0.027	0.859 ± 0.028	0.867 ± 0.028	0.866 ± 0.027	0.896	23.399 ± 1.288
CatBoosting	$S5$	**0.869 ± 0.026**	**0.863 ± 0.025**	**0.871 ± 0.027**	**0.869 ± 0.026**	**0.910**	**364.401 ± 49.962**

Quadro 4.7 – Resultados da classificação para inferência sintética (M2), com 10 execuções.

Modelo	**Seleção**	**Precisão média**	**Pontuação F1 média**	**Precisão média**	**Recall médio**	**Melhor precisão**	**Tempo médio de execução (s)**
Regressão Logística	$S5$	0.747 ± 0.038	0.738 ± 0.036	0.751 ± 0.039	0.747 ± 0.038	0.821 ± 0.0	0.036 ± 0.002
Máquina de vetores de suporte	$S5$	0.749 ± 0.039	0.746 ± 0.039	0.755 ± 0.038	0.749 ± 0.039	0.828	0.021 ± 0.006
Árvore de decisão	$S3$	0.788 ± 0.044	0.787 ± 0.044	0.790 ± 0.043	0.788 ± 0.044	0.843	0.006 ± 0.001
Multi Layer Perceptron	$S5$	0.803 ± 0.027	0.801 ± 0.029	0.807 ± 0.026	0.803 ± 0.027	0.836	0.830 ± 0.246
Gradient Boosting	$S5$	0.851 ± 0.029	0.847 ± 0.029	0.849 ± 0.028	0.851 ± 0.029	0.888	1.740 ± 0.254
Extreme Gradient Boosting	$S5$	0.855 ± 0.036	0.849 ± 0.035	0.855 ± 0.038	0.855 ± 0.036	0.910	0.393 ± 0.019
Floresta aleatória	$S4$	0.857 ± 0.017	0.849 ± 0.020	0.859 ± 0.014	0.857 ± 0.017	0.881	0.217 ± 0.007
Árvore Extra	$S4$	0.861 ± 0.016	0.857 ± 0.018	0.856 ± 0.020	0.861 ± 0.016	0.896	0.152 ± 0.007
Light Gradient Boosting	$S5$	0.870 ± 0.036	0.865 ± 0.035	0.873 ± 0.036	0.870 ± 0.036	0.910	55.203 ± 1.369
CatBoosting	**S5**	**0.878 ± 0.026**	**0.873 ± 0.027**	**0.880 ± 0.028**	**0.878 ± 0.026**	**0.925**	**340.054 ± 47.762**

Quadro 4.8 – Resultados da classificação para adição de amostras sintéticas (M3), com 10 execuções

Modelo	**Seleção**	**Precisão média**	**Pontuação F1 média**	**Precisão média**	**Recall médio**	**Melhor precisão**	**Tempo médio de execução (s)**
Regressão Logística	$S5$	0.655 ± 0.032	0.656 ± 0.031	0.667 ± 0.026	0.655 0.660 ± 0.034	0.716	0.052 ± 0.003
Máquina de vetores de suporte	$S5$	0.660 ± 0.034	0.662 ± 0.033	0.677 ± 0.030	0.660 ± 0.034	0.716	0.052 ± 0.003
Multi Layer Perceptron	$S5$	0.787 ± 0.016	0.787 ± 0.016	0.792 ± 0.016	0.787 ± 0.016	0.809	1.488 ± 0.569
Árvore de decisão	$S5$	0.860 ± 0.021	0.860 ± 0.022	0.864 ± 0.022	0.860 ± 0.021	7	0.020 ± 0.002
Floresta aleatória	$S5$	0.885 ± 0.016	0.886 ± 0.015	0.891 ± 0.015	0.885 ± 0.016	0.907	0.349 ± 0.010
Árvore Extra	$S5$	0.889 ± 0.016	0.890 ± 0.016	0.893 ± 0.017	0.889 ± 0.016	0.918	0.334 ± 0.023
Gradient Boosting	$S5$	0.894 ± 0.016	0.895 ± 0.016	0.898 ± 0.015	0.894 ± 0.016	0.923	3.042 ± 0.348
Extreme Gradient Boosting	$S5$	0.894 ± 0.021	0.895 ± 0.020	0.898 ± 0.020	0.894 ± 0.021	0.923	0.631 ± 0.020
Light Gradient Boosting	$S5$	0.900 ± 0.015	0.901 ± 0.014	0.904 ± 0.014	0.900 ± 0.015	0.918	94.481 ± 3.090
CatBoosting	**S5**	**0.903 ± 0.012**	**0.903 ± 0.012**	**0.906 ± 0.012**	**0.903 ± 0.012**	**0.923**	744.368 ± **97.757**

4.4 Resultados de Classificação

Entre os três conjuntos de dados deste estudo (M1, M2, M3), foram reveladas discrepâncias entre os modelos examinados, com ênfase particular no procedimento M3, que se destacou por lidar com informações balanceadas, um elemento crucial para o desempenho adequado dos modelos de aprendizagem de máquina.

No procedimento M1, a Precisão Média variou de 0.760 ± 0.040 para Regressão Logística até 0.869 ± 0.026 para CatBoosting, sendo este último o que alcançou a maior precisão de 0.910. No M2, as médias de precisão foram de 0.747 ± 0.038 para Regressão Logística até 0.878 ± 0.026 para CatBoosting, com uma precisão máxima de 0.925, conforme mostrado no quadro 4.6.

Por outro lado, no M3, onde os dados foram balanceados por amostras sintéticas, os resultados foram ainda melhores, com Precisão Média variando entre 0.655 ± 0.032 para Regressão Logística e 0.894 ± 0.016 para Gradient Boosting, destacando o impacto positivo do balanceamento das informações (Quadro 4.7).

O CatBoosting demonstrou desempenho consistente em todos os procedimentos; no entanto, o uso de dados balanceados no M3 permitiu que modelos como Gradient Boosting e Random Forest alcançassem seus melhores resultados (Quadro 4.8). Isso ressalta a relevância do balanceamento da informação para aprimorar a precisão e a eficácia dos modelos de aprendizagem de máquina.

4.5 Matrizes

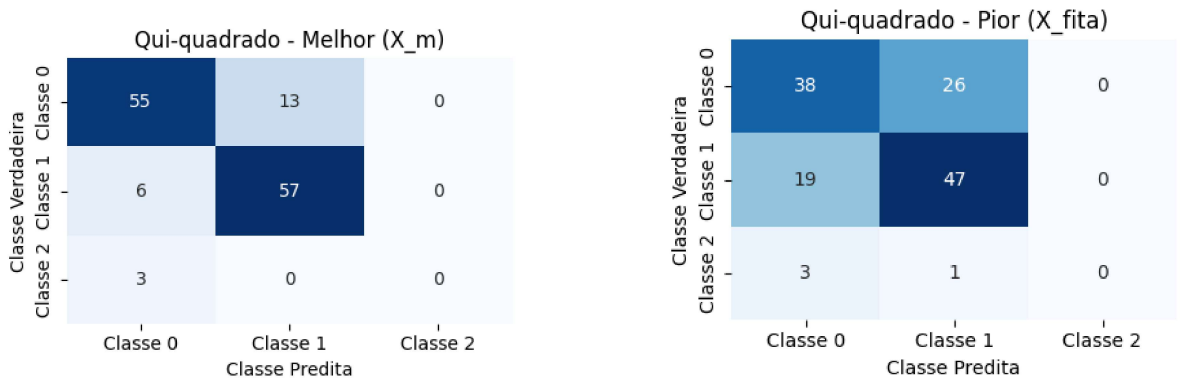
Para os algoritmos de classificação, foram geradas as matrizes de confusão correspondentes a cada conjunto de dados, incluindo *S5*, *S3*, *S4*, *S2* e *S1*. Optamos por exibir as melhores e piores matrizes com base na soma dos valores fora da diagonal principal, ou seja, os falsos positivos e falsos negativos. Na classificação proposta, a classe 0 é considerada a classe boa, a classe 1 é a classe média e a classe 2 é a classe ruim.

Através da análise das matrizes de confusão, foi possível observar que, com a inferência utilizando o teste qui-quadrado, os algoritmos apresentaram dificuldades em acertar a classe 2 (ruim), devido à escassez de dados dessa classe. Como resultado, a classe 2 foi frequentemente classificada como classe 0. No entanto, ao adicionar amostras sintéticas, os algoritmos melhoraram significativamente sua capacidade de acerto. Por exemplo, na matriz de confusão do algoritmo de Árvore de Decisão, o melhor modelo acertou 64 casos e errou apenas 5.

Adicionalmente, pode-se observar que as piores matrizes de confusão foram obtidas com o conjunto de dados *S1*. Por outro lado, o algoritmo Light Boosting, um dos melhores em termos de desempenho, obteve resultados notáveis. Na melhor matriz, esse algoritmo acertou 57 classificações da classe 2 (ruim), acertou 66 da classe 0 e errou 8, além de ter

acertado 55 da classe 1 e errando 7. As demais matrizes seguem a mesma lógica de análise, com a tendência de melhoria observada à medida que se adicionam amostras sintéticas, refletindo a importância dessa abordagem para o desempenho dos modelos.

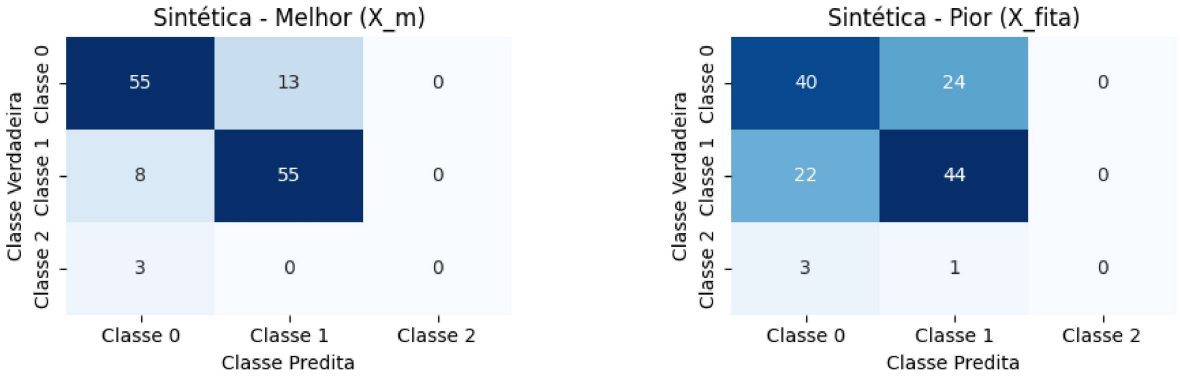
Figura 4.3 – Regressão Logística - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

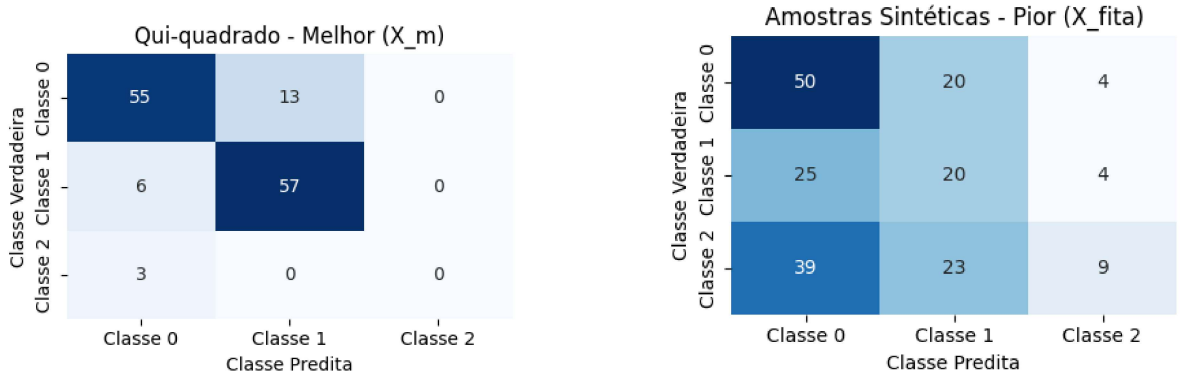
Figura 4.4 – Regressão Logística - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

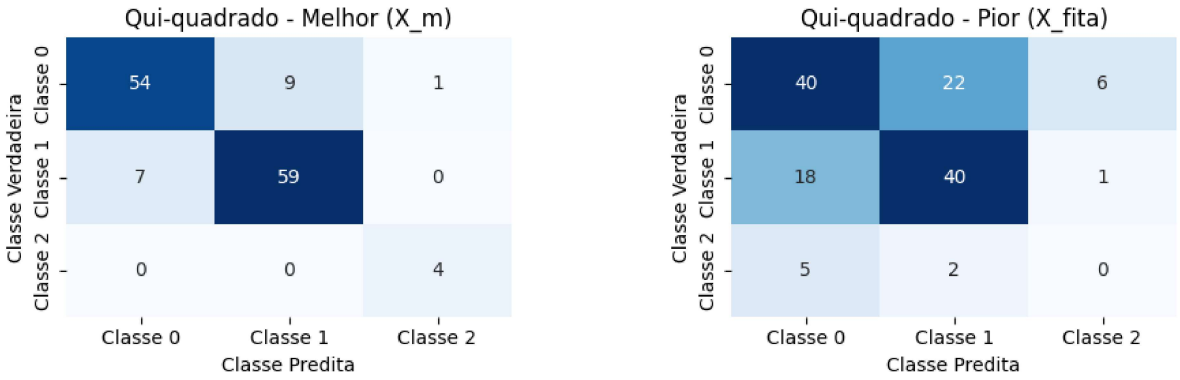
Figura 4.5 – Regressão Logística - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

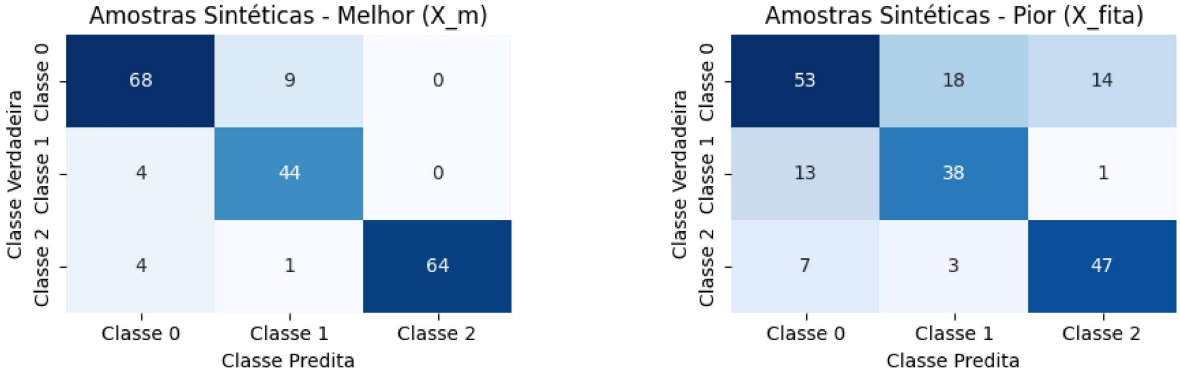
Figura 4.6 – Árvore Decisão - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

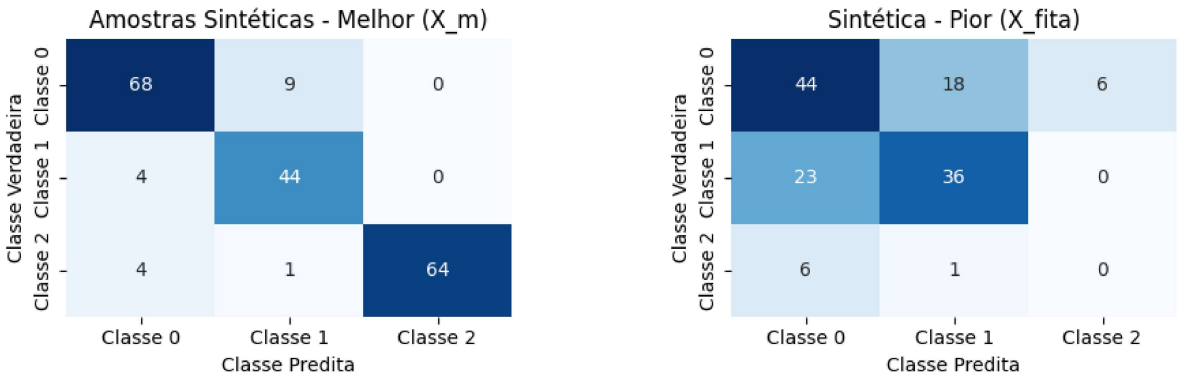
Figura 4.7 – Árvore Decisão - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

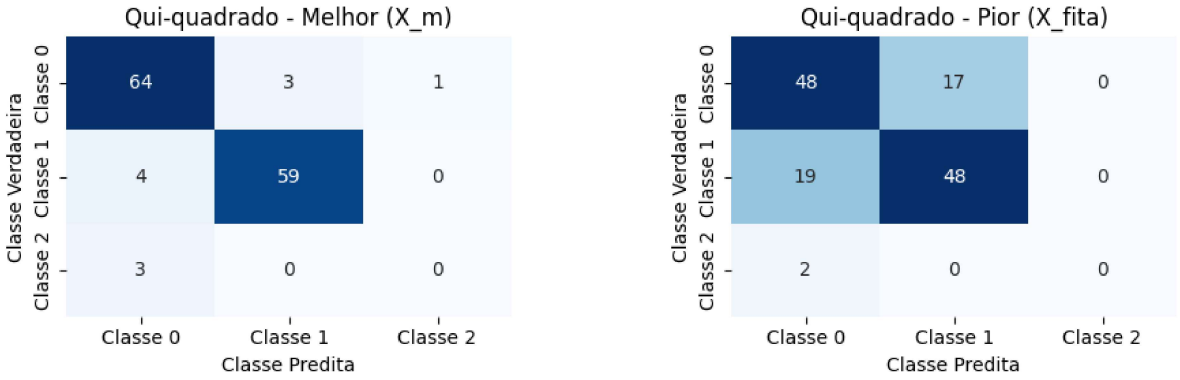
Figura 4.8 – Árvore Decisão - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

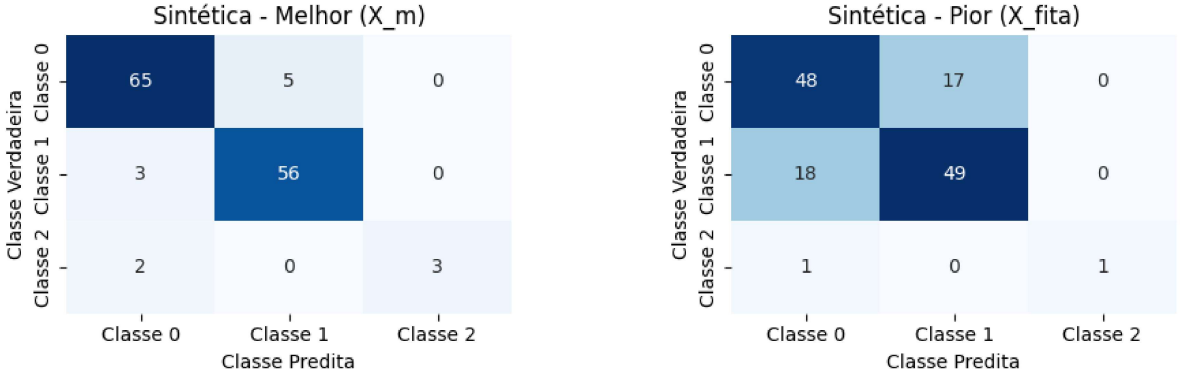
Figura 4.9 – CatBoosting- Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

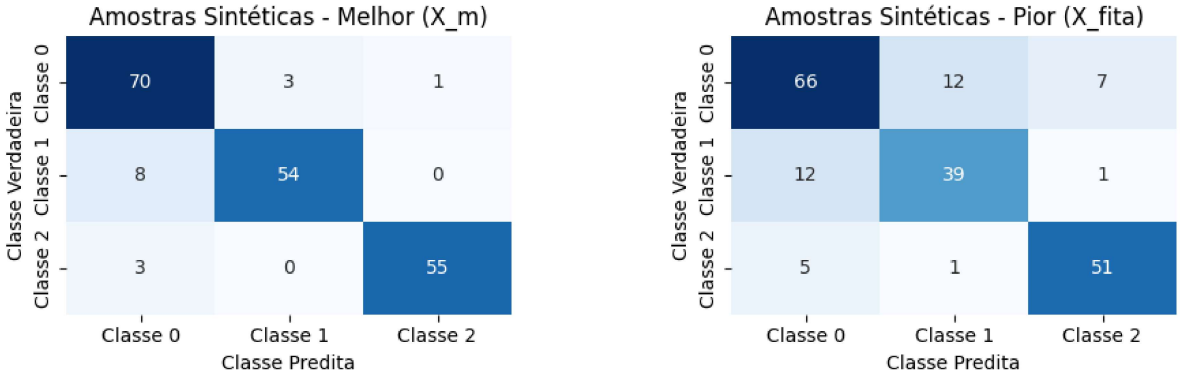
Figura 4.10 – CatBoosting - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

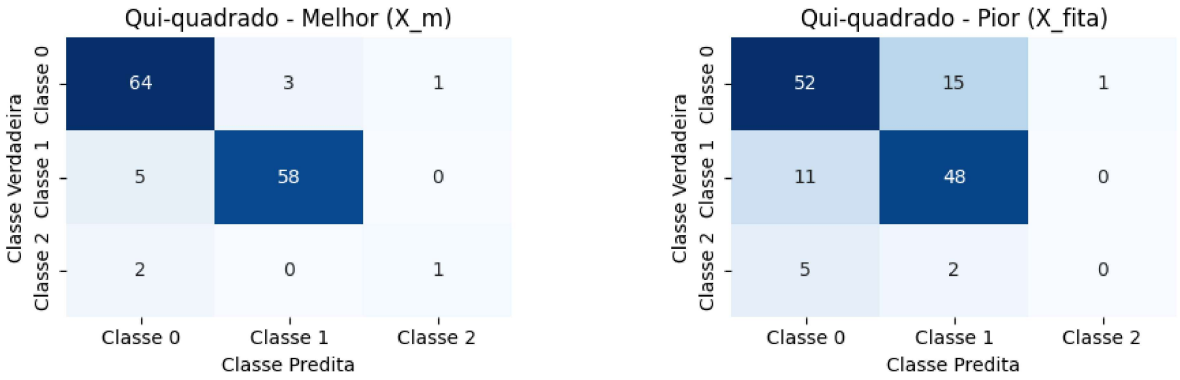
Figura 4.11 – CatBoosting - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

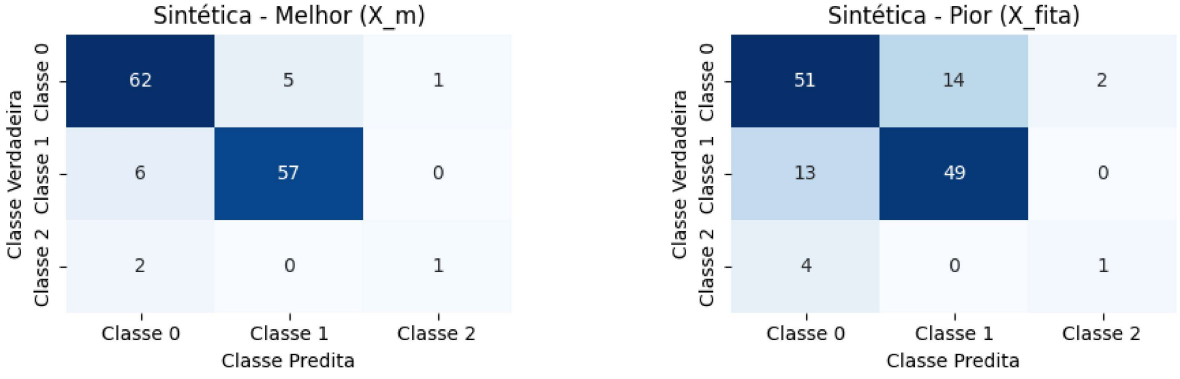
Figura 4.12 – Árvore Extra - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

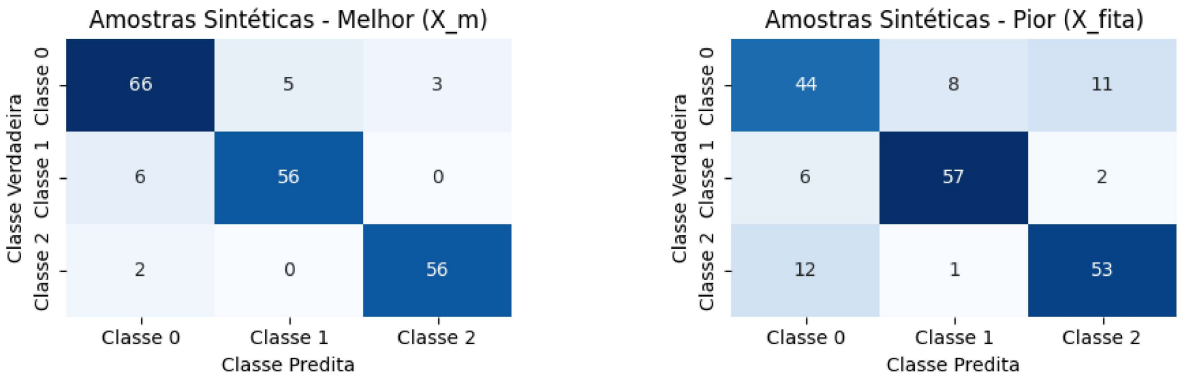
Figura 4.13 – Árvore Extra - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

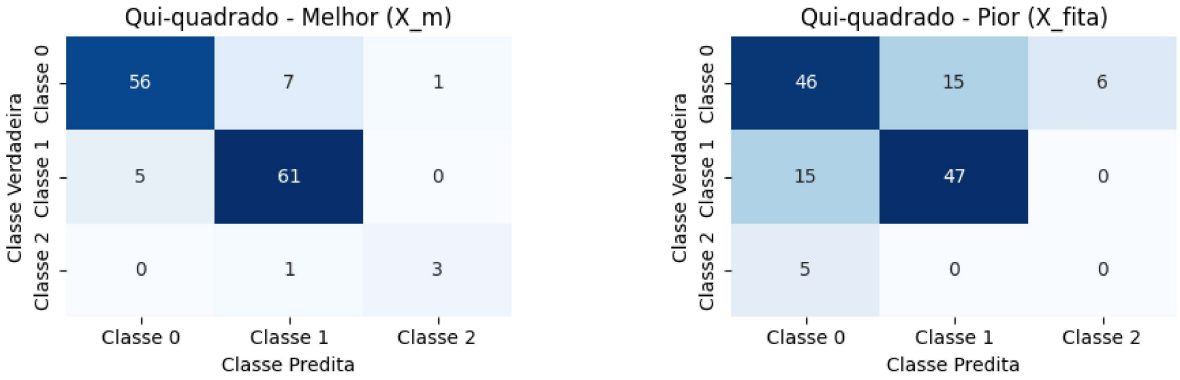
Figura 4.14 – Árvore Extra - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

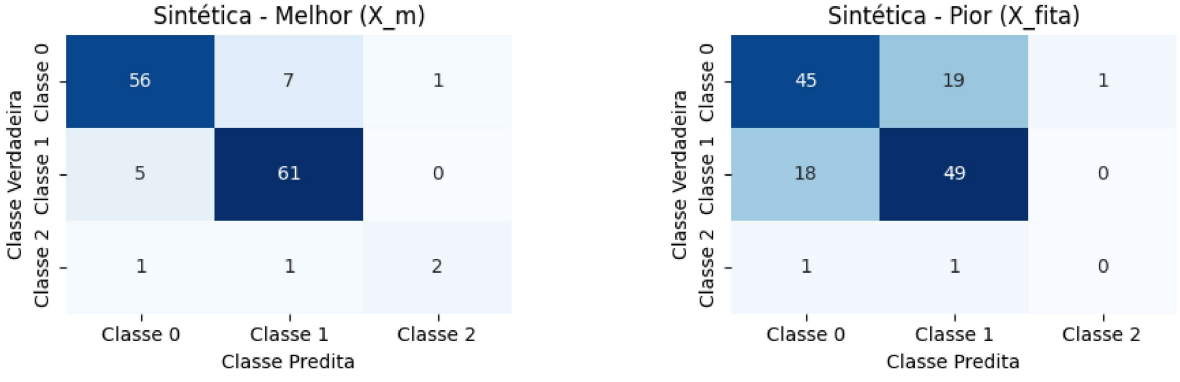
Figura 4.15 – Gradient Boosting - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

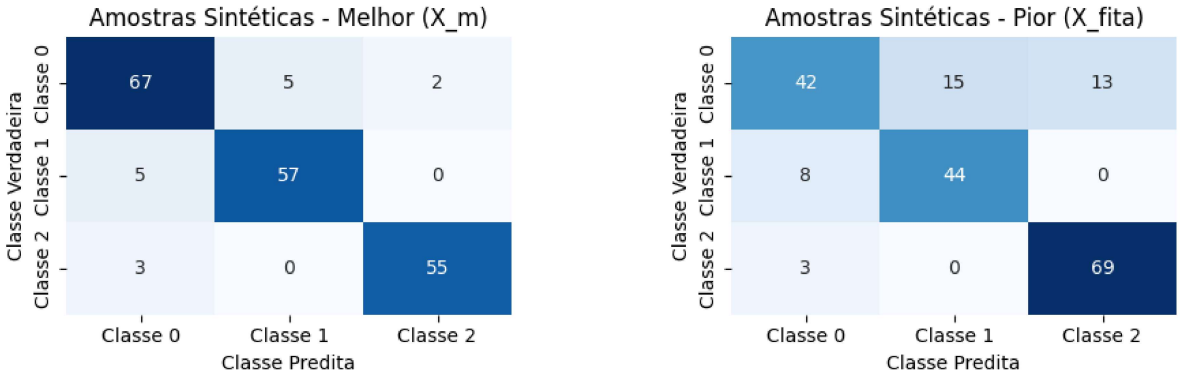
Figura 4.16 – Gradient Boosting - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

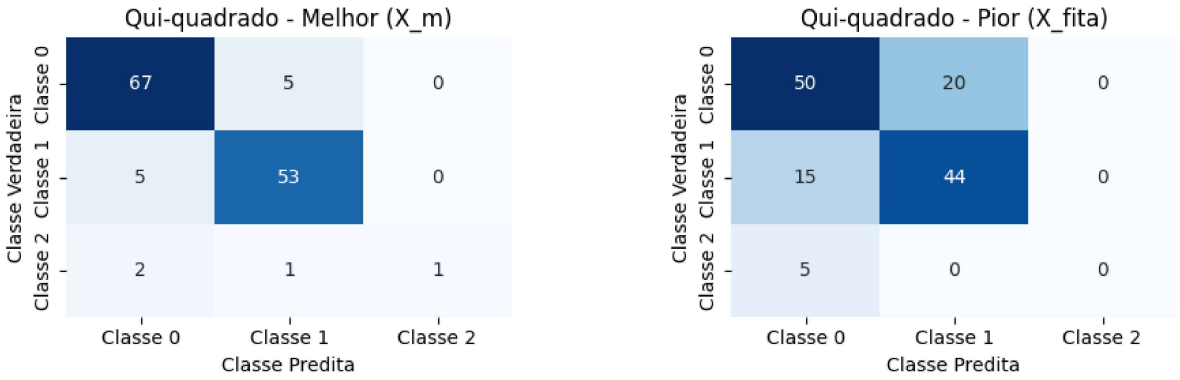
Figura 4.17 – Gradient Boosting - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

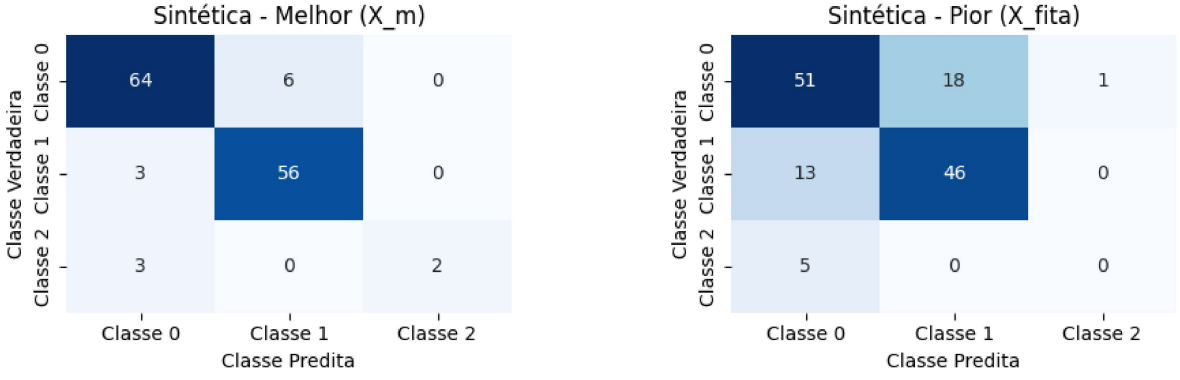
Figura 4.18 – Light Boosting - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

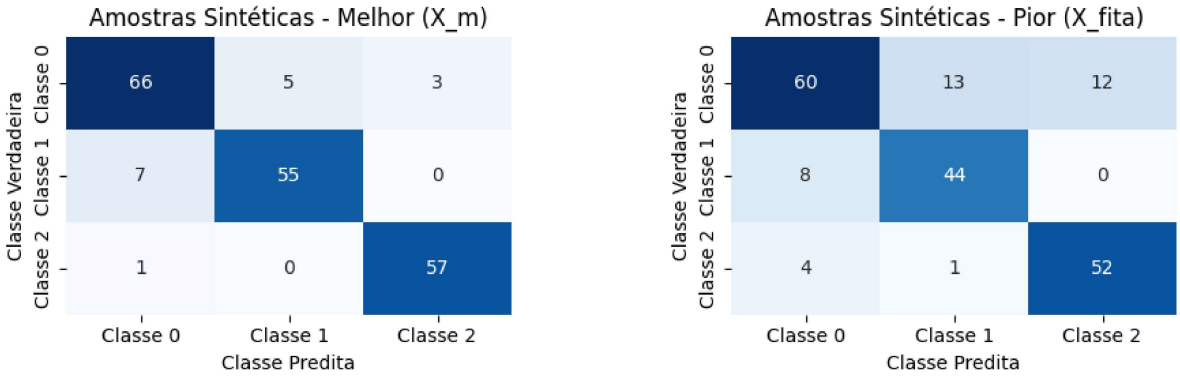
Figura 4.19 – Light Boosting - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

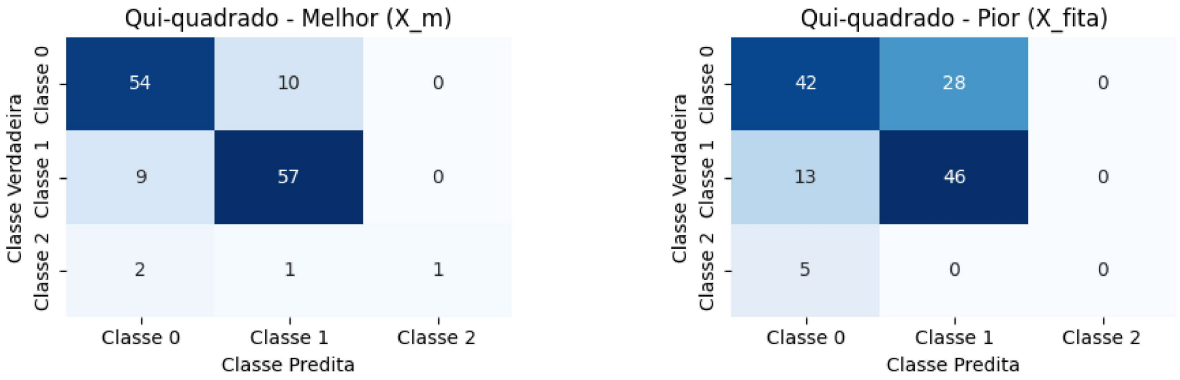
Figura 4.20 – Light Boosting - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

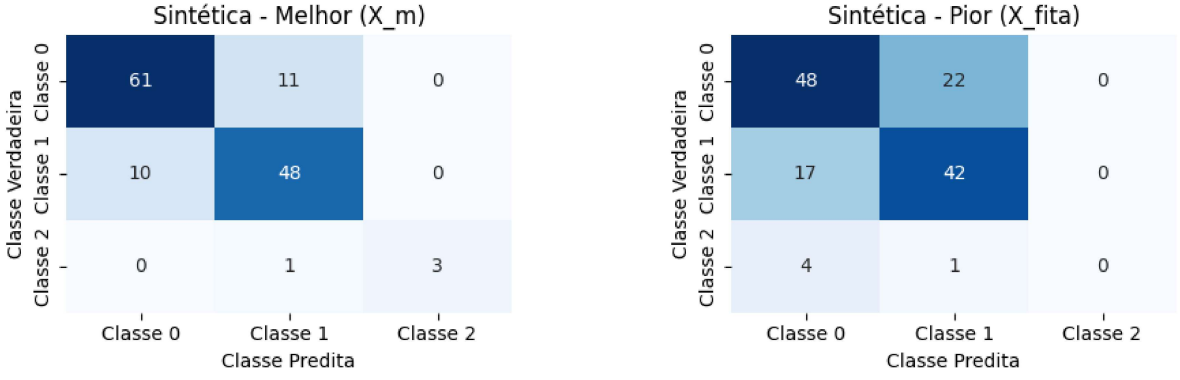
Figura 4.21 – Multi Layer Perceptron - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

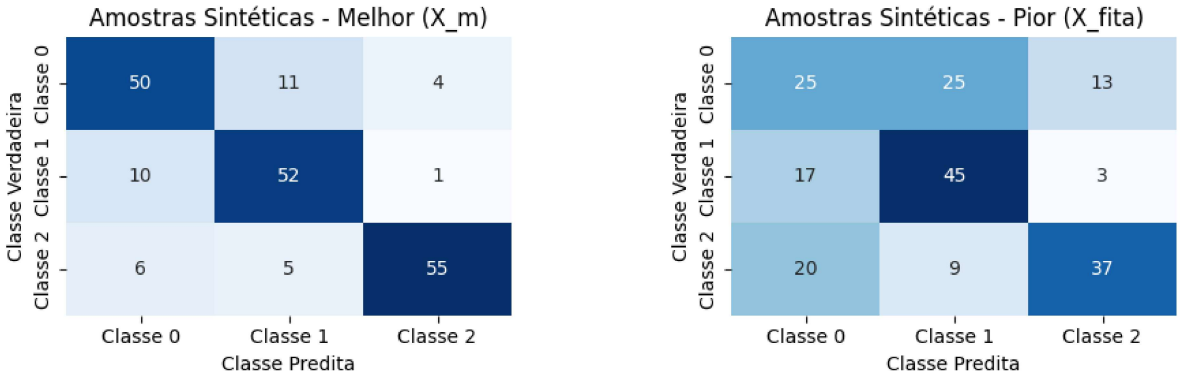
Figura 4.22 – Multi Layer Perceptron - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

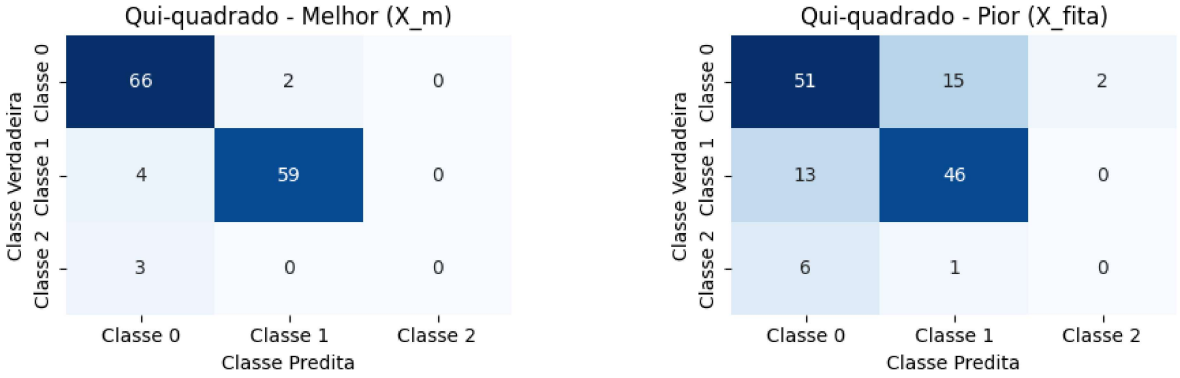
Figura 4.23 – Multi Layer Perceptron - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

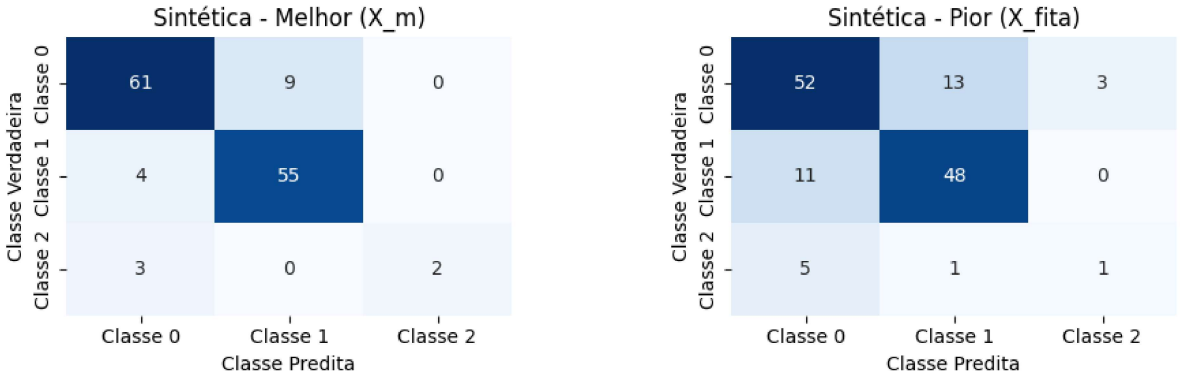
Figura 4.24 – Floresta Aleatória - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

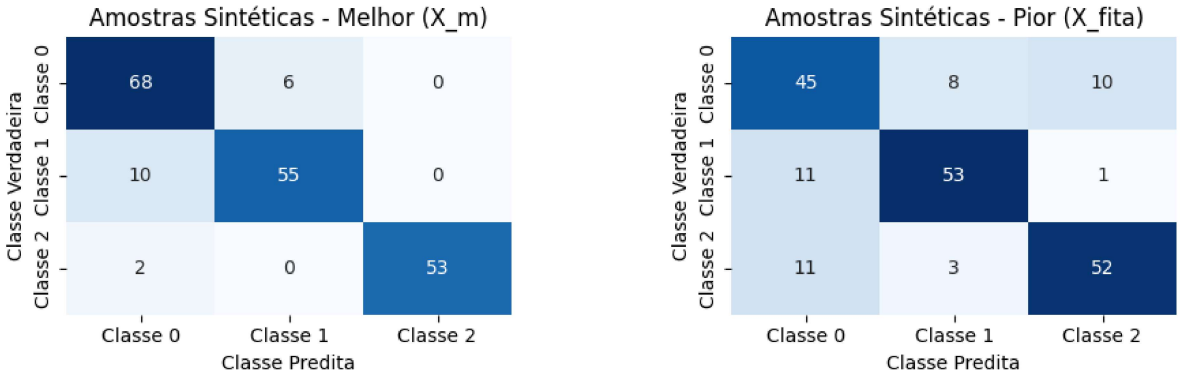
Figura 4.25 – Floresta Aleatória - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

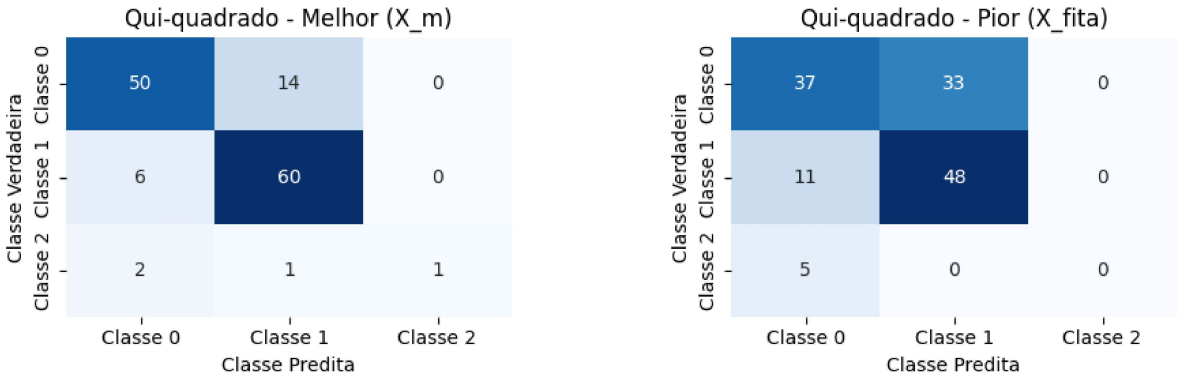
Figura 4.26 – Floresta Aleatória - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

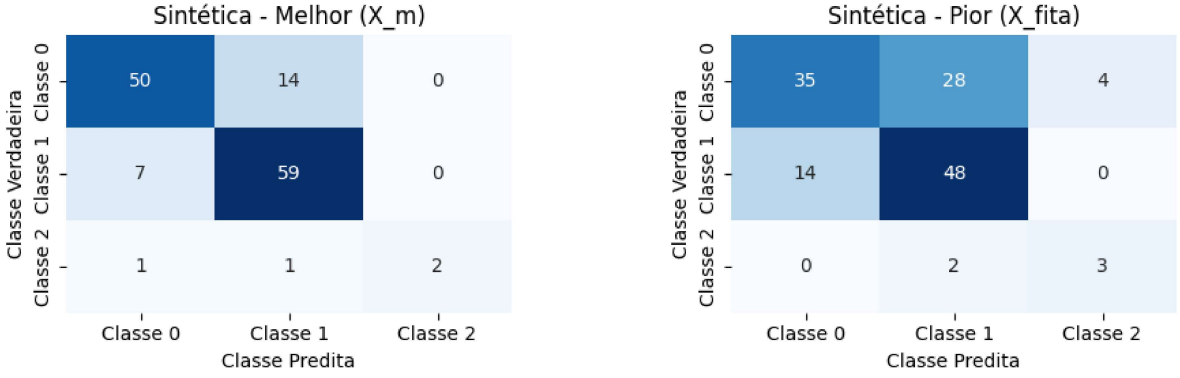
Figura 4.27 – Máquina Suporte - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

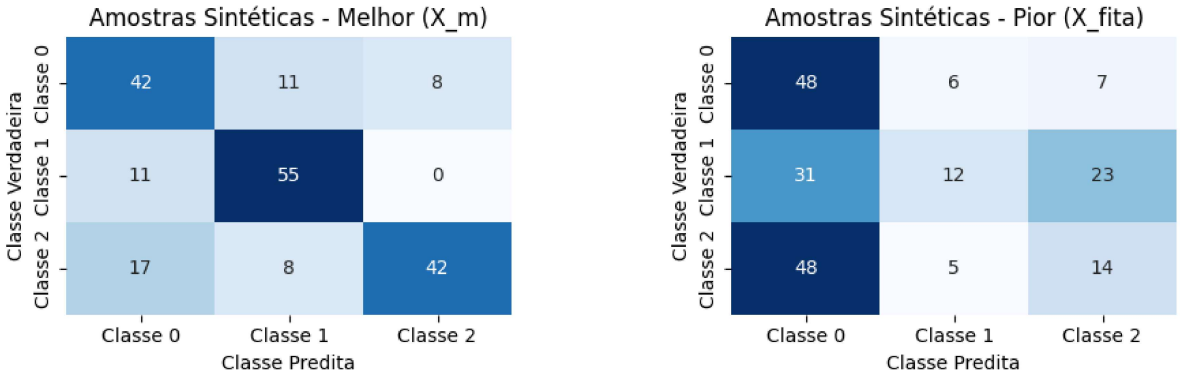
Figura 4.28 – Máquina Suporte - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

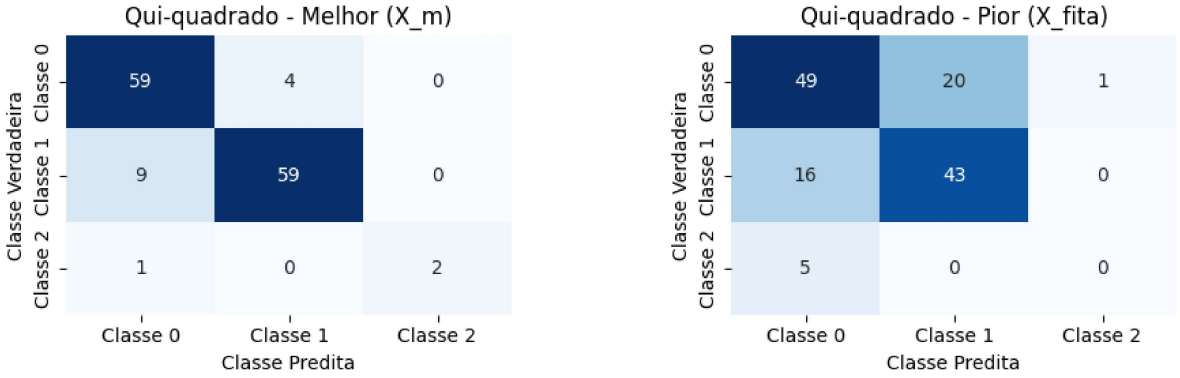
Figura 4.29 – Máquina Suporte - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5) (b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

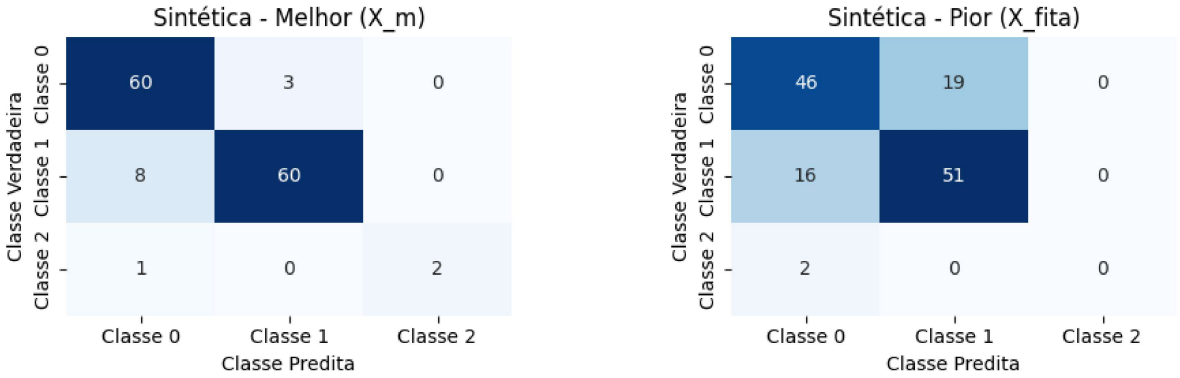
Figura 4.30 – Extreme Gradient Boosting - Matrizes de confusão com inferência Qui-quadrado.



(a) Qui-quadrado - Melhor (S5) (b) Qui-quadrado - Pior (S1)

Fonte: elaborado pela própria autora.

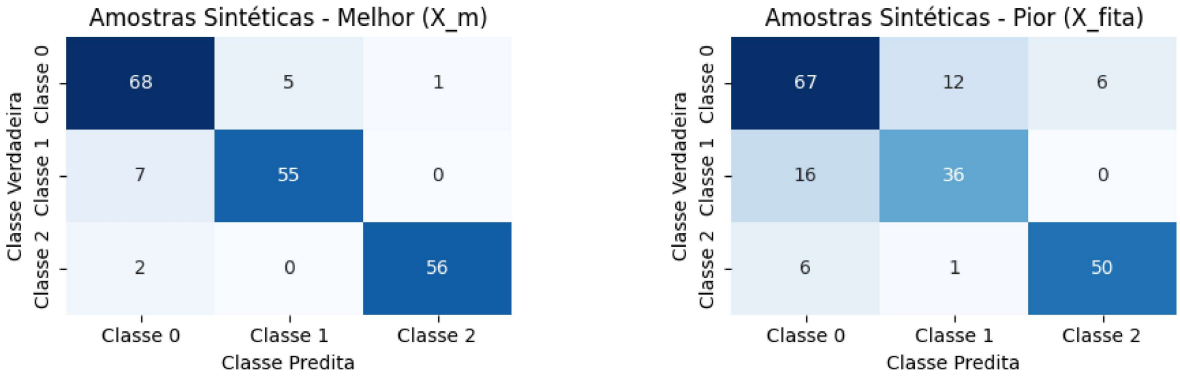
Figura 4.31 – Extreme Gradient Boosting - Matrizes de confusão com inferência Sintética.



(a) Sintética - Melhor (S5) (b) Sintética - Pior (S1)

Fonte: elaborado pela própria autora.

Figura 4.32 – Extreme Gradient Boosting - Matrizes de confusão com adição de amostras sintéticas.



(a) Amostras Sintéticas - Melhor (S5)

(b) Amostras Sintéticas - Pior (S1)

Fonte: elaborado pela própria autora.

5 CONCLUSÃO

A pesquisa apresentada nesta dissertação teve como objetivo explorar o uso de algoritmos de aprendizagem de máquina para prever e classificar o Índice de Qualidade da Água (IQA) do Rio Paraibuna, em Juiz de Fora, Brasil. A importância desse estudo se baseia na crescente necessidade de monitoramento e avaliação da qualidade da água, principalmente em regiões urbanas em desenvolvimento, onde as pressões humanas sobre os corpos hídricos são significativas. O uso de modelos preditivos baseados em dados históricos e variáveis ambientais tem se mostrado uma ferramenta promissora para a gestão sustentável da água, possibilitando intervenções mais eficazes e rápidas.

Através do pré-processamento de dados, foi possível selecionar as variáveis mais relevantes para a análise da qualidade da água, como pH, turbidez, temperatura da água, coliformes e concentrações de poluentes. Essa etapa foi crucial para garantir a qualidade dos dados utilizados nos modelos de aprendizagem de máquina. As variáveis selecionadas foram implementadas em diferentes algoritmos de aprendizagem supervisionados com o intuito de prever a qualidade da água e identificar padrões que indicassem possíveis riscos à saúde pública e ao ecossistema.

Os resultados obtidos mostram que os modelos de aprendizado de máquina tiveram desempenho variável em função da abordagem utilizada e do tratamento dos dados. Nas três abordagens analisadas inferência qui-quadrado (M1), inferência sintética (M2) e adição ou balanceamento de amostras sintéticas (M3). Os modelos baseados em *ensemble learning* apresentaram os melhores resultados em termos de MSE, RMSE, MAE, R^2 Precisão Média. Modelos como Extra Trees, Gradient Boosting, Light Gradient Boosting e CatBoosting alcançaram maior precisão média e menor erro, enquanto modelos individuais como Regressão Linear, SVM e Árvores de Decisão apresentaram desempenho inferior.

A aplicação dos modelos de aprendizagem de máquina permitiu uma análise mais aprofundada das relações entre as variáveis ambientais e o IQA, destacando a importância de parâmetros como temperatura da água, pH e sobrecarga na determinação da qualidade da água. Além disso, foi possível observar a influência de variáveis como turbidez e concentrações de metais pesados, como o níquel e o manganês, nas classificações do IQA, refletindo a complexidade da biodiversidade aquática e os desafios enfrentados pelas águas do Rio Paraibuna em termos de poluição.

Um aspecto importante desta pesquisa foi a aplicação da técnica de inferência como etapa de pré-processamento dos dados, proporcionando uma representação mais completa das diferentes categorias de qualidade da água. De modo geral, observou-se que o desempenho dos modelos evoluiu à medida que a complexidade da inferência aumentou. A abordagem M3, que integrou dados reais e sintéticos, apresentou o melhor desempenho, com aumento do coeficiente de determinação (R^2) e redução significativa dos erros de

previsão (MSE e RMSE). Esses resultados evidenciam o potencial das técnicas de inferência na melhoria da qualidade dos dados e reforçam o diferencial metodológico deste trabalho.

Apesar das contribuições positivas dessa abordagem, a pesquisa também apresentou algumas limitações. A qualidade dos dados históricos é fundamental para o sucesso dos modelos, e em algumas situações, a disponibilidade de dados completos e consistentes foi um desafio. Além disso, a complexidade das interações entre variáveis ambientais pode exigir modelos ainda mais sofisticados e uma maior coleta de dados para aprimorar a precisão das especificidades.

Além disso, os resultados desta pesquisa podem ser aplicados em outras regiões com características semelhantes, onde o uso de dados ambientais e modelos de aprendizagem de máquina poderiam contribuir para a melhoria da gestão hídrica e da qualidade dos corpos d'água. O potencial de expandir essa abordagem para outras bacias hidrográficas do Brasil, especialmente aquelas que enfrentam problemas de poluição devido à urbanização, é significativo.

Dessa forma, os objetivos específicos desta pesquisa foram alcançados. A avaliação da qualidade dos dados do Rio Paraibuna envolveu etapas de tratamento, limpeza e seleção de variáveis, garantindo a consistência das informações utilizadas nos modelos. Foram testados diferentes algoritmos de aprendizagem de máquina, entre os quais *Gradient Boosting*, *Light Gradient Boosting* e *CatBoosting* apresentaram melhor desempenho. Os resultados foram comparados com dados reais e validados por meio de métricas adequadas e validação cruzada, assegurando a confiabilidade das previsões. O modelo desenvolvido mostrou-se replicável e aplicável a outros contextos, podendo apoiar o monitoramento da qualidade da água em diferentes corpos hídricos.

É importante destacar que o uso de algoritmos de aprendizagem de máquina, embora promissor, não substitui as abordagens tradicionais de monitoramento da qualidade da água, como análises laboratoriais. Em vez disso, ele deve ser visto como uma ferramenta complementar que pode ajudar a melhorar os processos de monitoramento, tornando-os mais ágeis e permitindo uma resposta rápida diante das mudanças nos parâmetros da água. O desenvolvimento de novos algoritmos e a melhoria da qualidade dos dados podem ampliar ainda mais as capacidades da inteligência artificial na gestão da qualidade da água.

É essencial que os modelos de aprendizagem de máquina sejam continuamente atualizados com novos dados e intervalos, garantindo que eles possam refletir as mudanças ambientais e os impactos das atividades humanas. A adaptação constante dos modelos de aprendizagem de máquina às condições locais é crucial para a manutenção da sua precisão e eficácia ao longo do tempo.

A cooperação entre investigadores, órgãos ambientais, empresas e comunidades locais é fundamental para o sucesso de iniciativas como esta investigação. A implementação

de sistemas de monitoramento automatizados e preditivos, alimentados por modelos de aprendizagem de máquina, pode contribuir para a construção de uma gestão integrada e sustentável dos recursos hídricos.

A educação e conscientização da população também são fatores chave para a melhoria da qualidade da água. Com a utilização de modelos preditivos, é possível gerar alertas e comunicar de forma mais eficiente sobre a qualidade da água, permitindo que uma população se torne mais engajada na preservação e conservação dos corpos hídricos locais. Para futuras pesquisas, recomenda-se a expansão do conjunto de dados utilizados, com a inclusão de mais variáveis que possam influenciar a qualidade da água, como os efeitos das mudanças climáticas.

6 Trabalhos Futuros

Para o aprimoramento desta pesquisa, diversos caminhos promissores podem ser explorados. Um dos principais objetivos é a ampliação da base de dados, com a realização de web scraping em fontes governamentais oficiais, como portais de dados ambientais, visando obter um conjunto mais robusto e com maior representatividade das classes. Além de dados do rio Paraibuna, pretende-se expandir o escopo para incluir informações de toda a bacia do rio São Francisco, proporcionando uma análise mais abrangente e generalizável.

No que diz respeito à análise estatística, planeja-se aprimorar os testes de distribuição dos dados com a aplicação de gráficos QQ-plots e a execução de testes de normalidade mais abrangentes, incluindo os testes de Anderson-Darling, Kolmogorov-Smirnov e Shapiro-Wilk. Para a comparação entre grupos, além dos testes já utilizados, pretende-se empregar também o teste de Dunn (para comparações múltiplas pós-hoc) e o teste de Mann-Whitney (para comparações bivariadas não paramétricas), enriquecendo a análise inferencial.

Outra proposta relevante é a realização de um Grid Search com validação cruzada, possibilitando a identificação automática dos melhores hiperparâmetros para cada algoritmo testado, o que tende a melhorar consideravelmente a performance dos modelos. Complementarmente, pretende-se investigar a aplicação de algoritmos de otimização, como algoritmos genéticos ou swarm intelligence, com o intuito de buscar combinações ideais de parâmetros e variáveis.

Além das propostas de aprimoramento, esta pesquisa já gerou contribuições para a comunidade científica. O trabalho foi apresentado no 10^{th} International Conference on Information and Communication Technology (ICICT 2025), realizado entre os dias 18 e 21 de fevereiro de 2025, em Londres, Reino Unido, por meio da plataforma digital Zoom. O artigo intitulado *Comparison of Machine Learning Algorithms in Water Quality Index Prediction: A Case Study in Juiz de Fora, Brazil* foi incluído nos anais do congresso, publicado pela Springer Nature, evidenciando o reconhecimento da pesquisa em âmbito internacional.

REFERÊNCIAS

Kamal Said Abdallah, Yang Cao, & Dian-Jun Wei. Epidemiologic investigation of extra-intestinal pathogenic e. coli (expec) based on pcr phylogenetic group and fimh single nucleotide polymorphisms (snps) in china. *International Journal of Molecular Epidemiology and Genetics*, 2(4):339, 2011.

Isadora Padilha ADAM & Cristiane da Silva Paula de OLIVEIRA. Análise de parâmetros de potabilidade da água para consumo humano obtida de bebedouros. *Visão Acadêmica*, 24(2), 2024.

Godson Ebenezer Adjovu, Haroon Stephen, David James, & Sajjad Ahmad. Measurement of total dissolved solids and total suspended solids in water systems: A review of the issues, conventional, and remote sensing techniques. *Remote Sensing*, 15(14), 2023. ISSN 2072-4292. doi: 10.3390/rs15143534. URL `https://www.mdpi.com/2072-4292/15/14/3534`.

Agência Brasil. Falta de saneamento provocou mais de 340 mil internações em 2024. *Agência Brasil*, mar 2025. `https://agenciabrasil.ebc.com.br/saude/noticia/2025-03/falta-de-saneamento-provocou-mais-de-340-mil-internacoes-em-2024`.

Agência Nacional de Águas. *Atlas do Saneamento: Abastecimento Urbano de Água*. Agência Nacional de Águas (ANA), Brasília, 2013. URL `https://www.gov.br/ana/pt-br/assuntos/saneamento/atlas`.

Janete Alaburda & Linda Nishihara. Presença de compostos de nitrogênio em águas de poços. *Revista de Saúde Pública*, 32:160–165, 1998.

Gilberto Dias de Alkimin. Toxicidade de cádmio e zinco em danio rerio: comparação entre valores permitidos em legislação para proteção da vida aquática e a potencial atuação como interferentes endócrinos. Dissertação de mestrado em ciências ambientais, Universidade Estadual Paulista "Júlio de Mesquita Filho", Sorocaba, 2016. Área de Concentração: Diagnóstico, Tratamento e Recuperação Ambiental. Orientadora: Profa. Dra. Renata Fracácio Francisco.

Débora da Silva Almeida. Determinação de metais nas águas superficiais da bacia do rio são joão, março 2012. Trabalho apresentado para obtenção do título de Químico com Atribuições Tecnológicas.

A. Alnuaimi & T. Albaldawi. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97, 2024.

João Carlos de Andrade. Química analítica básica: os conceitos acido-base e a escala de ph. *Revista Chemkeys*, (1):1–6, set. 2018. doi: 10.20396/chemkeys.v0i1.9642. URL `https://econtents.bc.unicamp.br/inpec/index.php/chemkeys/article/view/9642`.

APHA, AWWA, WEF. *Standard Methods for the Examination of Water and Wastewater*. American Public Health Association, Washington, 23 edition, 2017. Métodos para análise de coliformes termotolerantes.

João Paulo Araújo, Tatiana Castro, Pedro Machado, & Ricardo Zaidan. Determinação do rio principal de bacias hidrográficas: o caso do rio paraibuna. *Periódico Eletrônico Fórum Ambiental da Alta Paulista*, 5:131–148, 05 2009.

A.N. Arcos, A.R.T. Vital, M.A. Rebelo, L.E.S. Silva, C.C.R. Oliveira, A. Lopes, S.J.F. Ferreira, & M.L. da Silva. Monitoramento da qualidade da água da precipitação na área urbana de manaus, amazonas. *Congresso Brasileiro de Qualidade em Química*, novembro 2022. Disponível em: `https://www.abq.org.br/cbq/2022/trabalhos/5/775-579.html`.

Seyed Babak Haji Seyed Asadollah, Ahmad Sharafati, Davide Motta, & Zaher Mundher Yaseen. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of environmental chemical engineering*, 9(1): 104599, 2021.

ATSDR - Agency for Toxic Substances and Disease Registry. Toxicological profile for cyanide. `http://www.atsdr.cdc.gov/toxprofiles/tp8.html`, 2006. Acesso em: julho 2024.

Sandra M. F. O. A. Azevedo. Toxinas de cianobactérias: Causas e consequências para a saúde pública. *Medicina On line - Revista Virtual de Medicina*, 1(3), 1998.

Mourade Azrour, Jamal Mabrouki, Ghizlane Fattah, Azedine Guezzaz, & Faissal Aziz. Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, 8(2):2793–2801, 2022.

Luciano Vieira Barreto, Micael de Souza Fraga, Flávia Mariani Barros, Felizardo Adenilson Rocha, Jhones da Silva Amorim, Stênio Rocha de Carvalho, Paulo Bonomo, & Danilo Paulúcio da Silva. Relationship between stream flow and water quality in a river section. *Ambiente e Agua-An Interdisciplinary Journal of Applied Science*, 9(1): 118–129, 2014.

Nutrition Board, Subcommittee on Upper Reference Levels of Nutrients, Standing Committee on the Scientific Evaluation of Dietary Reference Intakes, its Panel on Folate, Other B Vitamins, & Choline. Dietary reference intakes for thiamin, riboflavin, niacin, vitamin b6, folate, vitamin b12, pantothenic acid, biotin, and choline. 2000.

Benedito Braga. *Introdução à Engenharia Ambiental*. Prentice Hall, 2 edition, 2005.

Erika de Almeida Sampaio Braga, Marisete Dantas de Aquino, Carlos Márcio Soares Rocha, & Luzia Suerlange Araújo dos Santos Mendes. Presença de sílica em águas subterrâneas e possíveis benefícios para a saúde. *Águas Subterrâneas*, 34(2), 2020.

Brasil. Portaria nº 2914, de 12 de dezembro de 2011. Diário Oficial da União, Brasília, DF, 2011.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001a.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001b.

João PS Cabral. Water microbiology. bacterial pathogens and water. *International Journal of Environmental Research and Public Health*, 7(10):3657–3703, 2010. doi: 10.3390/ijerph7103657.

JENNIFFER UNFER DO CARMO. *Uma revisão crítica sobre os métodos analíticos para a determinação da demanda química de oxigênio (DQO)*. PhD thesis, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2021.

Raphael Ferreira Carnaúba, José Vicente Ferreira Neto, Luiz Carlos Sarmento Fernandez, RKLV Carnaúba, & Thiago José Matos Rocha. Análise dos parâmetros de coliformes totais e fecais em areia de praias urbanas de maceió, alagoas, brasil analysis of total and fecal coliform parameters in sand on urban beaches in maceió, alagoas, brazil. *Brazilian Journal of Development*, 7(12):115825–115848, 2021.

Diego M. P. Castro, Marcos Callisto, Marden S. Linares, Déborah R. O. Silva, Juliana S. França, Diego R. Macedo, Débora R. Carvalho, Paulo S. Pompeu, & Kele R. Firmiano. Abordagens ecológicas. In *Bases Conceituais para Conservação e Manejo de Bacias Hidrográficas*, volume 7 of *Série Peixe Vivo*, pages 63–130. Companhia Energética de Minas Gerais, Belo Horizonte, 2019. doi: 10.17648/bacias-hidrograficas-364.

CETESB. Cianetos. `https://cetesb.sp.gov.br/laboratorios/wp-content/uploads/sites/24/2022/02/Cianetos.pdf`, 2022. Acesso em: [Data de Acesso].

CETESB. Oxigênio dissolvido. `https://cetesb.sp.gov.br/mortandade-peixes/alteracoes-fisicas-e-quimicas/oxigenio-dissolvido/#:~:text=%C3%81guas%20com%20temperaturas%20mais%20baixas,oxig%C3%AAnio%20dissolvido%20apresenta%20menor%20solubilidade`, 2024a. Accessed: 2024-05-17.

CETESB. Fenol - contaminantes em alterações físicas e químicas, 2024b. URL `https://cetesb.sp.gov.br/mortandade-peixes/alteracoes-fisicas-e-quimicas/contaminantes/fenol/`. Acessado em: 19 de julho de 2025.

CETESB - Companhia Ambiental do Estado de São Paulo. Fundamentos do controle de poluição das Águas. Escola Superior da CETESB, Gestão do Conhecimento Ambiental, Conformidade Ambiental com Requisitos Técnicos e Legais, Pós-Graduação Lato Sensu, novembro 2018. URL `https://cetesb.sp.gov.br/veicular/wp-content/uploads/sites/33/2018/11/Apostila-Fundamentos-do-Controle-de-Poluicao-das-Aguas-T3.pdf`.

CETESB – Companhia Ambiental do Estado de São Paulo. Níquel e seus compostos. `https://cetesb.sp.gov.br/laboratorios/wp-content/uploads/sites/24/2021/05/NiÌĄquel.pdf`, 2021. Acesso em: 19 jul. 2025.

Fudi Chen, Tianlong Qiu, Jianping Xu, Jiawei Zhang, Yishuai Du, Yan Duan, Yihao Zeng, Li Zhou, Jianming Sun, & Ming Sun. Rapid real-time prediction techniques for ammonia and nitrite in high-density shrimp farming in recirculating aquaculture systems. *Fishes*, 9(10), 2024. ISSN 2410-3888. doi: 10.3390/fishes9100386. URL `https://www.mdpi.com/2410-3888/9/10/386`.

Tianqi Chen & Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Yvette MM Chen, Patrick J Wright, Chee-Seong Lee, & Glenn F Browning. Uropathogenic virulence factors in isolates of escherichia coli from clinical cases of canine pyometra and feces of healthy bitches. *Veterinary microbiology*, 94(1):57–69, 2003.

Catarina Isabel Terenas Pinto Cleto. O alumínio na água de consumo humano. Master's thesis, Universidade da Beira Interior (Portugal), 2008.

Comitê de Integração da Bacia Hidrográfica do Rio Paraíba do Sul (CEIVAP). Relatório de situação da bacia do rio paraíba do sul. Technical report, CEIVAP, 2020. URL https://www.ceivap.org.br/conteudo/relsituacao2020.pdf. Acesso em: 20 out. 2025.

Companhia Ambiental do Estado de São Paulo. Amônia, 2024. URL https://cetesb.sp.gov.br/mortandade-peixes/alteracoes-fisicas-e-quimicas/contaminantes/amonia/. Acesso em: 14 jun. 2024.

Companhia Ambiental do Estado de São Paulo (CETESB). Ficha de informação toxicológica: Zinco. Technical report, 2013a. URL https://cetesb.sp.gov.br/wp-content/uploads/sites/24/2013/11/Aluminio.pdf.

Companhia Ambiental do Estado de São Paulo (CETESB). Ficha de informação toxicológica: Zinco. Technical report, 2013b. URL https://cetesb.sp.gov.br/laboratorios/wp-content/uploads/sites/24/2013/11/Zinco.pdf.

Companhia Ambiental do Estado de São Paulo (CETESB). *Qualidade das Águas Interiores no Estado de São Paulo*. CETESB, 2016. URL https://cetesb.sp.gov.br/wp-content/uploads/sites/12/2018/03/Apendice-E-Significado-Ambiental-e-Sanitario-das-Variaveis-de-Qualidade-2016.pdf.

Companhia de Desenvolvimento e Ação Regional. ParÂmetros do processo de osmose inversa e de qualidade da Água em sistemas de dessalinizaÇÃo. Anexo 3 Manutenção C, 2019. URL https://www.car.ba.gov.br/sites/default/files/2019-07/Anexo_3_Manutencao_C.pdf. Governo do Estado da Bahia.

Companhia de Desenvolvimento e Ação Regional (CAR). Manutenção de sistemas de dessalinização – anexo 3. https://www.car.ba.gov.br/sites/default/files/2019-07/Anexo_3_Manutencao_C_0.pdf, 2019.

Conselho Nacional do Meio Ambiente. Resolução conama nº 357/2005, March 2005. URL https://www.mma.gov.br/port/conama/res/res05/res35705.pdf. Dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento, bem como estabelece as condições e padrões de lançamento de efluentes.

Jesús Miguel Contreras-Ramírez & Jesús Javier Nieves-Rivas. Incrustaciones en los sistemas de abastecimiento de agua potable: Formación y métodos de inhibición. *Base de la Ciencia*, 8(2):49–67, 2023. doi: 10.33936/revbasdelaciencia.v8i2.5851. URL https://www.researchgate.net/publication/374065534. Publicado em 31 de agosto de 2023.

G. C. Coppo, L. S. Passos, T. O. M. Lopes, *et al.* Genotoxic, biochemical and bioconcentration effects of manganese on oreochromis niloticus (cichlidae). *Ecotoxicology*, 27:1150–1160, 2018. doi: 10.1007/s10646-018-1970-0. URL https://doi.org/10.1007/s10646-018-1970-0.

Corinna Cortes & Vladimir Vapnik. Support-vector networks. *Machine learning*, 20: 273–297, 1995.

Carla Regina Costa, Paulo Olivi, Clarice MR Botta, & Evaldo LG Espindola. A toxicidade em ambientes aquáticos: discussão e métodos de avaliação. *Química nova*, 31: 1820–1830, 2008.

Laís Azevedo da Costa. Análise das concentrações de metais pesados na água e sedimento do rio sergipe (se). `https://ri.ufs.br/handle/riufs/11052`, 2018. Trabalho de Conclusão de Curso (Bacharelado em Geologia) – Universidade Federal de Sergipe, São Cristóvão, SE. Acesso em: 19 jul. 2025.

CPRM – Serviço Geológico do Brasil, Diretoria de Hidrologia e Gestão Territorial – DHT, Superintendência Regional de Belo Horizonte – SUREG-BH, Gerência de Hidrologia e Gestão Territorial – GEHITE, Laboratório de Sedimentometria e Qualidade das Águas – LSQA. *Medição in loco: Temperatura, pH, Condutividade Elétrica e Oxigênio Dissolvido*. Organizado por Magda Cristina Ferreira Pinto, 2007. Versão maio 2007.

Nello Cristianini & John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

URF CRUZ. Caracterização da comunidade fitoplanctônica no trecho final dos rios piraquê-açu e piraquê-mirim. aracruz-es. *Monografia de graduação–Departamento de Ecologia e Recursos Naturais, Universidade Federal do Espírito Santo, Vitória*, 2004.

Felipe S Cunha & Alcino P de Aguiar. Métodos para remoção de derivados fenólicos de efluentes aquosos. *Revista Virtual de Química*, 6(4):844–865, 2014.

Renata da Costa e Silva Crespim. Qualidade das Águas subterrâneas rasas: Estudo de caso no distrito de icoaraci - pa. Dissertação de mestrado, Universidade Federal do Pará, Instituto de Geociências, Programa de Pós-Graduação em Recursos Hídricos, Belém, Junho 2017. Orientador: Prof. Dr. Paulo Pontes Araújo.

Cristina Maria Carvalho da Mata Ribeiro. *Estabelecimento de uma rotina laboratorial para análise química de sedimentos e sua aplicação a sedimentos continentais do Minho (NW Portugal) : contribuição para a reconstituição paleoambiental da região*. PhD thesis, Universidade do Minho, February 2006. URL `https://repositorium.sdum.uminho.pt/bitstream/1822/7381/4/4-Elementos.pdf`.

Ministério da Saúde. Portaria gm/ms nº 888, de 4 de maio de 2021, 2021. URL `https://www.gov.br/saude/pt-br/assuntos/saude-publica/saneamento-e-qualidade-da-agua/portaria-gm-ms-no-888-de-4-de-maio-de-2021`. Acesso em [31 de julho de 2024].

Fernanda Faria da Silva. Metais pesados e qualidade da água da bacia do rio paraibuna. Dissertação de mestrado, Universidade Federal de Juiz de Fora, Juiz de Fora, 2021. URL `https://repositorio.ufjf.br/jspui/handle/ufjf/4317`. Acesso em: 05 ago. 2025.

Luis Fernando Durão da Silva Castro. Descontaminação de águas poluídas com compostos fenólicos utilizando discos de c18. Dissertação de mestrado, Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, dezembro 2013. Dissertação para obtenção do Grau de Mestre em Engenharia do Ambiente.

Nathália da Silva Resende, Juliana Barreto Oliveira dos Santos, Iollanda Ivanov Pereira Josué, Nathan Oliveira Barros, & Simone Jaqueline Cardoso. Comparing spatio-temporal dynamics of functional and taxonomic diversity of phytoplankton community in tropical

cascading reservoirs. *Frontiers in Environmental Science*, 10:903180, 2022. doi: 10.3389/fenvs.2022.903180. URL https://www.frontiersin.org/articles/10.3389/fenvs.2022.903180/full.

Bruno P De-Carli, Felícia P de Albuquerque, Viviane Moschini-Carlos, & Marcelo Pompêo. Comunidade zooplanctônica e sua relação com a qualidade da água em reservatórios do estado de são paulo. *Iheringia. Série Zoologia*, 108:e2018013, 2018.

Davi de Holanda Cavalcante. Relação dureza/alcalinidade da água e seus efeitos sobre a qualidade da água, do solo e desempenho zootécnico de juvenis de tilápia do nilo, oreochromis niloticus, mantidos em condições laboratoriais, fevereiro 2012. Orientador: Prof. Dr. Marcelo V.C. Sá.

Paula Romyne de Morais Cavalcante. Remoção de fenol de efluentes aquosos utilizando floculação iônica. Dissertação de mestrado, Universidade Federal do Rio Grande do Norte, Natal, RN, janeiro 2016. Centro de Tecnologia – CT, Departamento de Engenharia Química, Programa de Pós-Graduação em Engenharia Química.

José Carlos de Oliveira. *Gestão Ambiental de Bacias Hidrográficas*. Oficina de Textos, São Paulo, 2010.

Isabella Louise Bodin de Saint-Ange Comnène Carloni. Monitoramento da qualidade da Água dos chuveiros das praias de ipanema e leblon. Relatório de iniciação científica, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) - Departamento de Engenharia Química, 2014. Orientador: José Marcus de Oliveira Godoy, Coorientador: Daniela Soluri.

Companhia de Tecnologia de Saneamento Ambiental. Determinação de clorofila ae feofitina a: método espectrofotométrico. diário oficial do estado de são paulo–caderno executivo i, v. 124 (71) de 15/04/14, poder executivo, seção i, 53-55, 2014.

Agência Nacional de Águas. Bacia do rio paraíba do sul. https://www.gov.br/ana/pt-br, 2020. Acesso em: 7 abr. 2025.

Digital Water. Turbidez da Água: o que é, causas, impactos e como medir, 2024. URL https://www.digitalwater.com.br/turbidez-da-agua/. Acessado em: 23 abr. 2025.

Célia R. Diniz, Beatriz S. O. de Ceballos, José E. de L. Barbosa, & Annemarie Konig. Uso de macrófitas aquáticas como solução ecológica para melhoria da qualidade de água. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 9:226–230, 2005. ISSN 1415-4366. doi: 10.1590/1807-1929/agriambi.v9nsupp226-230. URL https://doi.org/10.1590/1807-1929/agriambi.v9nsupp226-230.

Anna Veronika Dorogush, Vasily Ershov, & Andrey Gulin. Catboost: gradient boosting with categorical features support. In *Workshop on ML Systems at NeurIPS*, 2018. URL https://catboost.ai. Yandex.

Sergio Dovidauskas, Isaura Akemi Okada, Maria Helena Iha, & Álvaro Gennari Cavallini. Concentrações de nitrato em águas de abastecimento público de 88 municípios da rede regional de atenção à saúde 13 do estado de são paulo, brasil. *Revista do Instituto Adolfo Lutz*, 78:e1765, 2019.

Wander Clay Pereira Dutra. Modelagem dos parâmetros de qualidade de Água em trecho urbanizado do rio paraibuna em juiz de fora (mg). Master's thesis, Universidade Federal de Juiz de Fora, Curso de Engenharia Sanitária e Ambiental, Juiz de Fora, 2014. Linha de Pesquisa: Hidráulica e Saneamento, Orientador: José Homero Pinheiro Soares.

James K. Edzwald. *Water Quality Treatment: A Handbook on Drinking Water*. McGraw-Hill Professional, 2011. ISBN 9780071630115.

Agoston E Eiben & James E Smith. *Introduction to evolutionary computing*. Springer, 2015.

Ensina.AI. Rede neural perceptron multicamadas, 2023. URL `https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9`. Acesso em: 5 ago. 2025.

Environmental Protection Agency (EPA). The effect on water resources, N/A. URL `https://cfpub.epa.gov/watertrain/moduleFrame.cfm?parent_object_id=2456&object_id=2459#:~:text=The%20Effect%20on%20Water%20Resources,habitats%20will%20be%20negatively%20affected.`

EPA United States Environmental Protection. Causal analysis/diagnosis decision information system (caddis): Dissolved oxygen. `https://www.epa.gov/caddis/dissolved-oxygen#:~:text=In%20addition%2C%20DO%20levels%20are,of%20DO%20in%20the%20water`, 2024. Accessed: 2024-05-17.

Marcelo Estevam, Adriano Willian da Silva, Frederico Fonseca da Silva, *et al.* Physical analysis of entry water in the agro-industry system of tannery in the city of maringá-paraná. *Ciência E Natura*, 41, 2019.

Francisco de Assis Esteves. *Fundamentos de Limnologia*. Interciência, Rio de Janeiro, 3 edition, 1998. ISBN 978-85-7193-223-6.

Marcos Aparecido Chaves Ferreira. *Desenvolvimento de Sensores de Oxigênio Dissolvido Utilizando Métodos Eletroquímicos e Ópticos para Monitoramento em Tempo Real da Qualidade da Água*. PhD thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, 2007. Tese de Doutorado em Engenharia Elétrica, orientador: Prof. Dr. Antônio Carlos Seabra.

Gesivaldo Jesus Alves de Figueiredo *et al.* Avaliação da presença de alumínio na água do sistema de abastecimento público da cidade de joão pessoa e grande joão pessoa no estado da paraíba e os possíveis riscos para a saúde da população, 2004.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Regina Célia Gentil, Andréa Tucci, & Célia Leite Sant'Anna. Dinâmica da comunidade fitoplanctônica e aspectos sanitários de um lago urbano eutrófico em são paulo, sp. *Hoehnea*, 35(2):265–280, Apr 2008. ISSN 2236-8906. doi: 10.1590/S2236-89062008000200008. URL `https://doi.org/10.1590/S2236-89062008000200008`.

Pierre Geurts, Damien Ernst, & Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

Eloisa Pozzi Gianotti. Contaminação das águas pelo zinco: a dureza da água como um fator de modificação da toxicidade do zinco a peixes / water contamination by zinc: hardness of water as a modification factor of zinc toxicity to fishes. *Revista DAE*, 145, 1986. Edição nº: 145. Ano: 1986.

D. F. Gomez Isaza, R. L. Cramp, & C. E. Franklin. Living in polluted waters: A meta-analysis of the effects of nitrate and interactions with other environmental stressors on freshwater taxa. *Environmental Pollution*, 261:114091, 2020. ISSN 0269-7491. doi: 10.1016/j.envpol.2020.114091. URL `https://doi.org/10.1016/j.envpol.2020.114091`. Epub 2020 Feb 4.

Karina Regina Gonzalez. Toxicologia do níquel. *Revista Intertox de Toxicologia, Risco Ambiental e Sociedade*, 9(2), jun. 2016. doi: 10.22280/revintervol9ed2.242. URL `http://autores.revistarevinter.com.br/index.php?journal=toxicologia&page=article&op=view&path[]=242`.

Kazuhiko Gotoh, Koji Horibe, Yan Mei, & Takashi Tsujisaka. Effects of water hardness on textile detergency performance in aqueous cleaning systems. *Journal of Oleo Science*, 65(2):123–133, 2016. doi: 10.5650/jos.ess15168. URL `https://doi.org/10.5650/jos.ess15168`. Epub 2016 Jan 15.

Norman N. Greenwood & Alan Earnshaw. *Chemistry of the Elements*. Butterworth-Heinemann, 2nd edition, 1997. ISBN 978-0750633659.

Matthew M. Guzzo, Paul J. Blanchfield, & Michael D. Rennie. Behavioral responses to annual temperature variation alter the dominant energy pathway, growth, and condition of a cold-water predator. *Proceedings of the National Academy of Sciences*, 114(37): 9912–9917, 2017. doi: 10.1073/pnas.1702584114. URL `https://doi.org/10.1073/pnas.1702584114`.

Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, & Abbas Parsaie. Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1):3–13, 2018.

D. Halliday, R. Resnick, & J. Walker. *Fundamentals of Physics*. Wiley, New York, 10 edition, 2018. Teoria cinética dos gases.

M. Harrison. *Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python*. Novatec Editora, 2019. ISBN 9788575228180.

Trevor Hastie & Robert Tibshirani. Discriminant adaptive nearest neighbor classification and regression. *Advances in neural information processing systems*, 8, 1995.

Carla Grasiele Zanin Hegel & Evanisa Fátima Reginato Quevedo Melo. Macrófitas aquáticas como bioindicadoras da qualidade da Água dos arroios da rppn maragato. *Revista Brasileira de Biociências*, 9(3):673–693, 2016. doi: 10.17765/2176-9168.2016v9n3p673-693.

Instituto Brasileiro de Geografia e Estatística - IBGE. Juiz de fora (mg) – cidades e estados. `https://www.ibge.gov.br/cidades-e-estados/mg/juiz-de-fora.html`, 2022. Acesso em: 05 ago. 2025.

Instituto Mineiro de Gestão das Águas. *Avaliação da Qualidade das Águas Superficiais de Minas Gerais em 2023: Resumo Executivo Anual*. Instituto Mineiro de Gestão das Águas (IGAM), Belo Horizonte, 2024. 244 p.: il. Vários colaboradores.

Instituto Mineiro de Gestão das Águas, (IGAM). Índice de qualidade das Águas – iqa, 2025. URL https://igam.mg.gov.br/w/indice-de-qualidade-das-aguas-iqa.

International Institute for Sustainable Development (IISD). Temperature and water quality, August 2018. URL https://www.iisd.org/ela/blog/temperature-quality-fresh-water/#:~:text=Warmer%20air%20temperatures%20mean%20less,quality%20and%20all%20resident%20biota.

Raed Jafar, Adel Awad, Iyad Hatem, Kamel Jafar, Edmond Awad, & Isam Shahrour. Multiple linear regression and machine learning for predicting the drinking water quality index in al-seine lake. *Smart Cities*, 6(5):2807–2827, 2023. ISSN 2624-6511. doi: 10.3390/smartcities6050126. URL https://www.mdpi.com/2624-6511/6/5/126.

ARISTON SILVA MELO JÚNIOR, Gabriel Reginaldo Santos, Gustavo Santos Silva, Robson Camilo Costa Melo, & Thalis Almeida Jesus. Monitoramento da concentração de oxigênio dissolvido (od) em lagoas de estabilização. *INOVAE-Journal of Engineering, Architecture and Technology Innovation (ISSN 2357-7797)*, 7(1):128–146, 2019.

HISSASHI KAMIYAMA. A complexidade do dbo. *Revista DAE [online]*, 48, 1988.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, & Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Claudia Klein & Sandra Aparecida Antonini Agne. Phosphorus: From the nutrient to pollutant! *Revista Eletrônica em Gestão, Educação e Tecnologia Ambiental*, 8(8): 1713–1721, Jan. 2013. doi: 10.5902/223611706430. URL https://periodicos.ufsm.br/reget/article/view/6430.

Cornelis Klein & Barbara W. Dutrow. *The Manual of Mineral Science*. Wiley, 23rd edition, 2012. ISBN 978-1118211522.

Saber Kouadri, Ahmed Elbeltagi, Abu Reza Md Towfiqul Islam, & Samir Kateb. Performance of machine learning methods in predicting water quality index based on irregular data set: application on illizi region (algerian southeast). *Applied Water Science*, 11(12):190, 2021.

Deepak Kumar, Vijay Kumar Singh, Salwan Ali Abed, Vinod Kumar Tripathi, Shivam Gupta, Nadhir Al-Ansari, Dinesh Kumar Vishwakarma, Ahmed Z Dewidar, Ahmed A Al-Othman, & Mohamed A Mattar. Multi-ahead electrical conductivity forecasting of surface water based on machine learning algorithms. *Applied Water Science*, 13(10):192, 2023.

Dheeraj Kumar, Rakesh Kumar, Madhuben Sharma, Amit Awasthi, & Manish Kumar. Global water quality indices: Development, implications, and limitations. *Total Environment Advances*, 9:200095, 2024. ISSN 2950-3957. doi: https://doi.org/10.1016/j.teadva.2023.200095. URL https://www.sciencedirect.com/science/article/pii/S2950395723000176.

Luiz Drude de Lacerda & Olaf Malm. Contaminação por mercúrio em ecossistemas aquáticos: uma análise das áreas críticas. *Estudos Avançados*, 22(63):173–190, 2008. ISSN 0103-4014. doi: 10.1590/S0103-40142008000200011. URL https://doi.org/10.1590/S0103-40142008000200011.

Bui Quoc Lap, Huu Du Nguyen, Phi Thi Hang, Nguyen Quang Phi, Vinh Truong Hoang, Pham Gia Linh, Bui Thi Thanh Hang, *et al.* Predicting water quality index (wqi) by feature selection and machine learning: a case study of an kim hai irrigation system. *Ecological Informatics*, 74:101991, 2023.

Sangung Lee, Bu Geon Jo, Jaeyeon Lim, Jong Mun Lee, & Young Do Kim. Assessment of climate change impacts on hydrology using an integrated water quality index. *Hydrology*, 11(11), 2024. ISSN 2306-5338. doi: 10.3390/hydrology11110178. URL https://www.mdpi.com/2306-5338/11/11/178.

Y. Li, J. Smith, & R. Kumar. Data driven machine learning prognostics of buckling failure modes in ballasted railway track. *SN Applied Sciences*, 6(4):1234, 2024. doi: 10.1007/s42452-024-05885-3. URL https://link.springer.com/article/10.1007/s42452-024-05885-3.

Wang Lin, Huimin Luo, Jingyi Wu, Tien-Chieh Hung, Beibei Cao, Xiangli Liu, Jifeng Yang, & Pinhong Yang. A review of the emerging risks of acute ammonia nitrogen toxicity to aquatic decapod crustaceans. *Water*, 15(1), 2023. ISSN 2073-4441. doi: 10.3390/w15010027. URL https://www.mdpi.com/2073-4441/15/1/27.

Marden Seabra Linares, Marcos Callisto, & João Carlos Marques. Compliance of secondary production and eco-exergy as indicators of benthic macroinvertebrates assemblages' response to canopy cover conditions in neotropical headwater streams. *Science of the Total Environment*, 613:1543–1550, 2018.

André Lopes. Extra-trees: Árvores extremamente aleatórias, 2023. URL https://brains.dev/2023/extra-trees-arvores-extremamente-aleatorias/. Acesso em: 5 ago. 2025.

Hongfang Lu & Xin Ma. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249:126169, 2020.

N. Mann. Cyanotoxins. In B. Whitton & M. Potts, editors, *The Ecology of Cyanobacteria*, chapter 22, pages 613–622. Springer, Netherlands, 2002a. doi: 10.1007/0-306-46855-7_14.

N. Mann. Detecting the environment. In B. Whitton & M. Potts, editors, *The Ecology of Cyanobacteria*, chapter 1, pages 1–13. Springer, Netherlands, 2002b. doi: 10.1007/0-306-46855-7_14.

Leandro Marcó, Ricardo Azario, Celia Metzler, M d Garcia, L Marcó, & R Azario. La turbidez como indicador básico de calidad de aguas potabilizadas a partir de fuentes superficiales. propuestas a propósito del estudio del sistema de potabilización y distribución en la ciudad de concepción del uruguay (entre ríos, argentina). *Higiene y Sanidad Ambiental*, 4(11):4–72, 2004.

MR Martínez-Orjuela, JY Mendoza-Coronado, BE Medrano-Solís, LM Gómez-Torres, & CA Zafra-Mejía. Evaluation of turbidity as a parameter indicator of treatment in a drinking water treatment plant. *Revista UIS Ingenierías*, 19(1):15–24, 2020.

Lidiane Raquel Verola Mataveli, Márcia Liane Buzzo, Maria de Fátima Henriques Carvalho, Luciana Juncioni de ARAUZ, & Guilherme Augusto Verola Mataveli. Avaliação dos níveis de cromo total em águas para consumo humano. *Revista do Instituto Adolfo Lutz*, 77:1–11, 2018.

Gerson Medeiros, Ana Tresmondi, Brigida Queiroz, Felipe Fengler, André Rosa, Joziane Fialho, Renata Lopes, Caio Negro, Leandro Santos, & Admilson Ribeiro. Water quality, pollutant loads, and multivariate analysis of the effects of sewage discharges into urban streams of southeast brazil. *Energy, Ecology and Environment*, 2, 06 2017. doi: 10.1007/s40974-017-0062-y.

Thiago Augusto Mendes, Fernanda Caroline Romanielo Alves, Diandra Ferreira, Daniel Mendes, & Renata Medici Frayne Cuba. Avaliação de diferentes técnicas de medição do oxigênio dissolvido para o saneamento básico. *Fronteira: Journal of Social, Technological and Environmental Science*, 10(1):406–426, 2021.

Jean Karlo Acosta Mendonça, Débora Farina Gonçalves, & Fernanda Monteiro Rigue. Experimento para determinação semiquantitativa de oxigênio dissolvido em água doce. *Revista Sítio Novo*, 4(1):53–61, 2020.

Ministério da Saúde. *Vigilância e Controle da Qualidade da Água para Consumo Humano*. Ministério da Saúde, Brasília, DF, Brasil, 2006. Secretaria de Vigilância em Saúde, Coordenação-Geral de Vigilância em Saúde Ambiental.

Brasil Ministério da Saúde. Sistema de informação de vigilância da qualidade da Água para consumo humano – vigiagua. `https://www.gov.br/saude/pt-br/composicao/seidigi/demas/situacao-de-saude/vigiagua`, 2025. Acesso em: 13 out. 2025.

Brasil Ministério do Meio Ambiente, Secretaria de Recursos Hídricos, Comitê para Integração da Bacia Hidrográfica do Rio Paraíba do Sul, UNESCO, Banco Mundial, & Governo do Japão. Projeto inicial de gerenciamento dos recursos hídricos da bacia do rio paraíba do sul: Carta consulta à cofiex (minuta). Technical report, Laboratório de Hidrologia – COPPE/UFRJ, Rio de Janeiro, julho 1999. Acesso em: 13 out. 2025.

Reza Mohammadpour, Syafiq Shaharuddin, Chun Kiat Chang, Nor Azazi Zakaria, Aminuddin Ab Ghani, & Ngai Weng Chan. Prediction of water quality index in constructed wetlands using support vector machine. *Environmental Science and Pollution Research*, 22:6208–6219, 2015.

Adyasha Mohanty & Grace Gao. A survey of machine learning techniques for improving global navigation satellite systems. *EURASIP Journal on Advances in Signal Processing*, 2024(73), 2024. doi: 10.1186/s13634-024-01167-7. URL `https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-024-01167-7`.

J. C. Morrill, R. C. Bales, & M. H. Conklin. The Relationship Between Air Temperature and Stream Temperature. In *AGU Spring Meeting Abstracts*, volume 2001, pages H42A–09, May 2001.

Rodrigo Braga Moruzzi & Marco Antonio Penalva Reali. Oxidação e remoção de ferro e manganês em águas para fins de abastecimento público ou industrial: uma abordagem geral. *Revista de Engenharia e Tecnologia*, pages 29–43, 2012.

João Carlos Nabout, Ana Clara Maciel David, Jéssica Fagundes Felipe, Karine Borges Machado, Laurence Carvalho, & Hélida Ferreira da Cunha. As pessoas podem detectar a perda de qualidade da água? um experimento de campo para avaliar a correlação entre a percepção visual e o grau de eutrofização da água. *Acta Limnologica Brasiliensia*, 34:e8, 2022.

Gerson Flôres Nascimento. Construção e mapeamento de Índice de qualidade de Águas subterrâneas em porto velho. Master's thesis, Universidade Federal do Pará, Belém, Brasil, 2017.

R. B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2007. ISBN 9780387286785.

Glauco Maciel Nolasco, Ednilton Moreira Gama, Bruna Morais Reis, Ana Clara Pereira Reis, Fernando José Santana Gomes, & Roberta Pereira Matos. Análise da alcalinidade, cloretos, dureza, temperatura e condutividade em amostras de água do município de almenara/mg. *Recital-Revista de Educação, Ciência e Tecnologia de Almenara/MG*, 2(2): 52–64, 2020.

Normas ABNT. ph e água – definição, diagrama de acidez e base, e mais, 2025. URL `https://www.normasabnt.org/ph-e-agua/`. Acesso em 23 de abril de 2025.

EC Oliveira-Filho, NR Caixeta, NCS Simplício, SR Sousa, TP Aragão, & DHF Muniz. Implications of water hardness in ecotoxicological assessments for water quality regulatory purposes: a case study with the aquatic snail biomphalaria glabrata (say, 1818). *Brazilian Journal of Biology*, 74(1):175–180, 2014.

Paulo Henrique Kingma Orlando. Produção do espaço e gestão hídrica na bacia do rio paraibuna (mg-rj): uma análise crítica. 2006.

Kweku-Muata Osei-Bryson. Evaluation of decision trees: a multi-criteria approach. *Computers & Operations Research*, 31(11):1933–1945, 2004.

Faridah Othman, ME Alaaeldin, Mohammed Seyam, Ali Najah Ahmed, Fang Yenn Teo, Chow Ming Fai, Haitham Abdulmohsin Afan, Mohsen Sherif, Ahmed Sefelnasr, & Ahmed El-Shafie. Efficient river water quality index prediction considering minimal number of inputs variables. *Engineering Applications of Computational Fluid Mechanics*, 14(1): 751–763, 2020.

Noriatsu Ozaki, Takehiko Fukushima, Hideo Harasawa, Toshiharu Kojiri, Katsunori Kawashima, & Miyuki Ono. Statistical analyses on the effects of air temperature fluctuations on river water qualities. *Hydrological Processes*, 17(14):2837–2853, 2003. doi: https://doi.org/10.1002/hyp.1437. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.1437`.

Lucilia Maria Parron, Daphne Heloisa de Freitas Muniz, & Clauida Mara Pereira. *Manual de procedimentos de amostragem e análise físico-química de água*. Embrapa Florestas, Curitiba, Brazil, 1 edition, December 2011. ISBN 1980-3958. Versão digital.

M. J. Paul, R. Coffey, J. Stamp, & T. Johnson. A review of water quality responses to air temperature and precipitation changes 1: Flow, water temperature, saltwater intrusion. *Journal of the American Water Resources Association*, 55(4):824, 2019. doi: 10.1111/1752-1688.12710.

Luis Otávio Miranda Peixoto, Bárbara Alves de Lima, Camila de Carvalho Almeida, Cristóvão Vicente Scapulatempo Fernandes, Jorge Antonio Silva Centeno, & Júlio César Rodrigues de Azevedo. Data imputation of water quality parameters through feed-forward neural networks. *RBRH*, 28:e14, 2023.

Pedro Henrique Silva Penedo, Giovanni de Oliveira Garcia, Roberto Avelino Cecílio, Sidney Sara Zanetti, & Mariza Pereira de Oliveira Roza. RelaÇÃo entre precipitaÇÃo e turbidez em cursos d'Água no espÍrito santo. *Caminhos de Geografia*, 24(91):132–152, fev. 2023. doi: 10.14393/RCG249161901. URL `https://seer.ufu.br/index.php/caminhosdegeografia/article/view/61901`.

Jéssyca Emanuella Saraiva Pereira. Biossorção de cobre em solução aquosa utilizando os pós das folhas do cajueiro (anacardium occidentale l.) e da carnaúba (copernicia prunifera). Master's thesis, Brasil, 2017.

Munique Rodrigues Pereira. Uso da luminescência como ferramenta de detecção de oxigênio dissolvido para verificação da qualidade de Água bruta superficial. Master's thesis, Universidade Federal do Rio Grande do Sul, Instituto de Química, Porto Alegre, 2016. Dissertação de Mestrado, orientadora: Profa. Dra. Leandra Franciscato Campo.

Alba Rocio Aguilar Piratoba, Hebe Morganne Campos Ribeiro, Gundisalvo Piratoba Morales, & Wanderson Gonçalves e Gonçalves. Caracterização de parâmetros de qualidade da água na área portuária de barcarena, pa, brasil. *Revista Ambiente & Água*, 12(3):435–456, 2017.

Alina Criane de Oliveira PIRES *et al.* Concentração de mercúrio e selênio na plataforma costeira do sudeste do brasil. 2023.

Marcelo Pompêo. Monitoramento e manejo de macrófitas aquáticas. *Oecologia brasiliensis*, 12(3):5, 2008.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, & Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648, 2018.

Railson de Oliveira Ramos. *Desenvolvimento de um multianalisador automático baseado em visão de máquina e processamento de vídeo para determinação sequencial do teor de sólidos suspensos totais e sedimentáveis em águas residuárias.* Tese de doutorado, Universidade Federal da Paraíba, Centro de Ciências Exatas e da Natureza, Departamento de Química, Programa de Pós-Graduação em Química, João Pessoa, PB, Brasil, setembro 2021.

B. E. Rgang & C. V. de S. Gastal Jr. *Macrófitas Aquáticas da Planície Costeira do RS.* Editora/Instituição, Porto Alegre, 1 edition, 1996.

Paulo Sérgio Gonçalves Rocha. Análise da influência da turbidez em resultados de amostra de Água subterrânea. Master's thesis, COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO ESCOLA SUPERIOR DA CETESB, São Paulo, Brasil, 2019.

N. E. Rojas & F. A. Rocha. Influência da alcalinidade no crescimento de larvas de tilápia do nilo. In *Anais do Congresso Brasileiro de Engenharia de Pesca*, 2009. URL `https://pdfs.semanticscholar.org/ed1b/7b63f8649267195e728446c179b62b51ec3a.pdf`.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

David E. Rumelhart, Geoffrey E. Hinton, & Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Thomas A Russo & James R Johnson. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of escherichia coli: Expec. *The Journal of infectious diseases*, 181(5):1753–1754, 2000.

K.C. Ruttenberg. The global phosphorus cycle. In Heinrich D. Holland & Karl K. Turekian, editors, *Treatise on Geochemistry*, chapter 8.13, pages 585–643. Elsevier, 2003. doi: 10.1016/B0-08-043751-6/08153-6.

Mohamad Sakizadeh. Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*, 2:1–9, 2016.

Antônio Carlos Silva Sampaio. Metais pesados na Água e sedimentos dos rios da bacia do alto paraguai. Master's thesis, Universidade Federal de Mato Grosso do Sul, Centro de Ciências Exatas e Tecnologia, Programa de Pós-Graduação em Tecnologias Ambientais, Ministério da Educação, dezembro 2003. Dissertação de Mestrado.

Ismael Luís Schneider. Modo alternativo de tratamento de efluentes com presença de cianeto, 2009.

SEMAD & UCEMG / PNMA II. Sistema de cálculo da qualidade da água (scqa) – estabelecimento das equações do índice de qualidade das águas (iqa). relatório i, 2005.

Larissa Rodrigues da Silva. Estudo da importância da demanda bioquímica de oxigênio (dbo) para o controle de qualidade de efluentes. Monografia, Universidade Federal de Alagoas, Maceió, 2022. Trabalho de Conclusão de Curso (TCC).

T. A. Silva. Estudo do desempenho da combinação de preditores baseados em cópulas e máquinas de vetor de suporte para séries temporais Úteis ao desenvolvimento sustentável. Master's thesis, Universidade Federal Rural de Pernambuco, 2020.

Luiz Gonzaga de Albuquerque Silva Júnior, Hans Raj Gheyi, & José Francismar de Medeiros. Chemical composition of water in the cristalline region of northeast brazil. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 3:11–17, 1999.

Danilo Barbosa Siqueira & Eduardo Cyrino Oliveira-Filho. Cianobactérias de água doce e saúde pública: uma revisão. *Universitas Ciências da Saúde*, 3(1):109–127, 2023. doi: 10.5102/ucs.v3i1.549. Monografia apresentada para conclusão do curso de Biomedicina no UniCEUB.

Ruben Cruz Siqueira *et al.* Avaliação do ph e condutividade em águas superficiais na barragem de rejeitos em minas do camaquã. 2018.

A. Sklar. Distribution functions of n dimensions and margins. *Institute of Statistics at the University of Paris*, pages 229–231, 1959.

Val H. Smith. Eutrophication of freshwater and coastal marine ecosystems: a global problem. *Environmental Science and Pollution Research*, 10(2):126–139, 2003. ISSN 1614-7499. doi: 10.1065/espr2002.12.142. URL `https://doi.org/10.1065/espr2002.12.142`.

Søren Peter Lauritz Sörensen & Sven Palitzsch. *Sur un indicateur nouveau, a-naphtolphtaléine, ayant un virage au voisinage du point neutre; Sur le mesurage de la concentration en ions hydrogène de l'eau de mer*. Hagerup, 1910.

Marcos Von Sperling. *Princípios do Tratamento Biológico de Águas Residuárias*. Departamento de Engenharia Sanitária e Ambiental, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil, 3 edition, 2002. ISBN 978-8585173773.

John L. Stoddard *et al.* Setting expectations for the ecological condition of streams: The concept of reference condition. *Ecological Applications*, 16(4):1267–1277, 2006. doi: 10.1890/05-0745.

Shiliang Sun & Rongqing Huang. An adaptive k-nearest neighbor algorithm. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 1, pages 91–94, 2010. doi: 10.1109/FSKD.2010.5569740.

Jafar Tanha, Yousef Abdi, Negin Samadi, Nazila Razzaghi, & Mohammad Asadpour. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7:1–47, 2020.

Huu-Tai Thai. Machine learning for structural engineering: A state-of-the-art review. In *Structures*, volume 38, pages 448–491. Elsevier, 2022.

David Tilman, Susan S. Kilham, & Peter Kilham. Phytoplankton community ecology: the role of limiting nutrients. *Annual Review of Ecology and Systematics*, 13(1):349–372, 1982.

Rafael Melo Torres. Remoção biológica de nitrato em Água de abastecimento humano utilizando o endocarpo de coco como fonte de carbono. Dissertação de mestrado, Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Sanitária, Natal, 2011.

Carlos E. M. Tucci. *Gestão de Recursos Hídricos*. Editora da UFRGS, Porto Alegre, 2001.

José Galizia Tundisi & Takako Matsumura Tundisi. *Limnologia*. Oficina de Textos, 2008.

Md. Galal Uddin, Stephen Nash, & Agnieszka I. Olbert. A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, 122:107218, 2021a. ISSN 1470-160X. doi: https://doi.org/10.1016/j.ecolind.2020.107218. URL `https://www.sciencedirect.com/science/article/pii/S1470160X20311572`.

Md Galal Uddin, Stephen Nash, & Agnieszka I Olbert. A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, 122:107218, 2021b.

Md Galal Uddin, Stephen Nash, Mir Talas Mahammad Diganta, Azizur Rahman, & Agnieszka I Olbert. Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*, 321:115923, 2022.

Md Galal Uddin, Stephen Nash, Azizur Rahman, & Agnieszka I. Olbert. A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. *Water Research*, 229:119422, 2023a. ISSN 0043-1354. doi: https://doi.org/10.1016/j.watres.2022.119422. URL `https://www.sciencedirect.com/science/article/pii/S0043135422013677`.

Md Galal Uddin, Stephen Nash, Azizur Rahman, & Agnieszka I. Olbert. Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, 169:808–828, 2023b. ISSN 0957-5820. doi: https://doi.org/10.1016/j.psep.2022.11.073. URL https://www.sciencedirect.com/science/article/pii/S0957582022010473.

Md Galal Uddin, Stephen Nash, Azizur Rahman, & Agnieszka I Olbert. Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, 169:808–828, 2023c.

U.S. Environmental Protection Agency. Aquatic life ambient water quality criteria for ammonia – freshwater. Technical Report EPA-822-R-13-001, U.S. Environmental Protection Agency, Washington, DC, April 2013. URL https://www.epa.gov/sites/default/files/2015-08/documents/aquatic-life-ambient-water-quality-criteria-for-ammonia-freshwater-2013.pdf.

V. Vapnik, S. Golowich, & A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems (NIPS)*, volume 9, pages 281–287, 1997.

A. V. Vargas, E. S. Vieira, & K. B. Nunes. Determinação rápida do oxigênio dissolvido em sistemas aquosos. *Química Analítica*, 2007. 2007d.

Marcos Von Sperling. *Wastewater characteristics, treatment and disposal*. IWA publishing, 2007.

J. M. Wallace & P. V. Hobbs. *Atmospheric Science: An Introductory Survey*. Academic Press, San Diego, 2 edition, 2006.

Junjie Wang, Jiawei Xu, Rui Liu, Zhifeng Li, Jun Liu, & Xiaojiang Wei. A novel support vector machine optimization algorithm based on hybrid sparrow search algorithm. *EURASIP Journal on Advances in Signal Processing*, 2024(1):1–21, 2024. doi: 10.1186/s13634-024-01167-7. URL https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-024-01167-7.

Louis Wehenkel, Damien Ernst, & Pierre Geurts. Ensembles of extremely randomized trees and some generic applications. In *Proc.Robust Methods for Power System State Estimation and Load Forecasting*, RTE, France, 2006.

Robert G. Wetzel. *Limnology: Lake and River Ecosystems*. Springer Science+Business Media New York, New York, NY, 3 edition, 2000. Originally published by Springer-Verlag New York, Inc. in 2000. Softcover reprint of the hardcover 3rd edition 2000. © 2000, 1991 Springer Science+Business Media New York.

Juliana de Oliveira Xavier, Mônica de Cássia Souza Campos, Sylvia Therese Meyer Ribeiro, & Helen Regina Mota. *Macrófitas Aquáticas: Caracterização e Importância em Reservatórios Hidrelétricos*. Companhia Energética de Minas Gerais – Cemig, Belo Horizonte, 1 edition, 2021. ISBN 978-85-87929-85-3. Copyright: Companhia Energética de Minas Gerais – Cemig; Presidência – DPR: Reynaldo Passanezi Filho; Diretoria Adjunta de Estratégia, Meio Ambiente, Inovação e Gabinete da Presidência – DEP: Maurício Dall'Agnese; Gerência de Gestão Ambiental – DEP/GA: Rafael Augusto Fiorine.

Min Xu, Pakorn Watanachaturaporn, Pramod K Varshney, & Manoj K Arora. Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97(3):322–336, 2005.

Yandex. Catboost documentation. `https://catboost.ai/en/docs`, 2024. Acessado em abril de 2025.

Camilo Zamora-Ledezma, Daniela Negrete-Bolagay, Freddy Figueroa, Ezequiel Zamora-Ledezma, Ming Ni, Frank Alexis, & Victor H. Guerrero. Heavy metal water pollution: A fresh look about hazards, novel and conventional remediation methods. *Environmental Technology Innovation*, 22:101504, 2021. ISSN 2352-1864. doi: https://doi.org/10.1016/j.eti.2021.101504. URL `https://www.sciencedirect.com/science/article/pii/S2352186421001528`.

Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, & Lin Ye. A review of the application of machine learning in water quality evaluation. *Eco-Environment Health*, 1(2):107–116, 2022. ISSN 2772-9850. doi: https://doi.org/10.1016/j.eehl.2022.06.001. URL `https://www.sciencedirect.com/science/article/pii/S2772985022000163`.

Wangmeng Zuo, David Zhang, & Kuanquan Wang. On kernel difference-weighted k-nearest neighbor classification. *Pattern Analysis and Applications*, 11(3):247–257, 2008.

APÊNDICE A –

Artigo publicado - Apêndice

Comparison of Machine Learning Algorithms in Water Quality Index Prediction: A Case Study in Juiz de Fora, Brazil

Larissa de Lima[1] and Priscila Capriles[1]

Federal University of Juiz de Fora, UFJF, BRA, MG. 36036-900
larissa.lima@estudante.ufjf.br, priscila.capriles@ufjf.br

Abstract. This paper explores the use of machine learning (ML) with various physical, chemical, and biological parameter combinations to predict water quality, focusing on the Water Quality Index (WQI). We assess the performance of several regression algorithms across five different data combinations and examine the impact of inference and class balancing techniques on model outcomes. Our analysis reveals that LightGBM achieved the highest accuracy in WQI regression at 93%. This research introduces a novel approach to calculating WQI by automating the traditional manual and complex parameter collection and calculation process. By streamlining water quality monitoring, our ML-based method offers a more efficient and innovative solution. Additionally, the study provides practical insights into handling data scarcity and using statistical inference for skewed sampling distributions.

Keywords: water quality, machine learning algorithms, artificial intelligence, regression, classification, environmental impact, water resources

1 Introduction

Water is indispensable for human life and ecosystems. The United Nations (UN) addresses its sustainable management in the 2030 Agenda [4]. The World Health Organization (WHO) estimates 3.5 million annual deaths from waterborne diseases such as cholera, dysentery, and hepatitis A, highlighting the critical importance of water quality monitoring [3].

The Water Quality Index (WQI) is a widely used metric that synthesizes parameters such as dissolved oxygen, temperature, turbidity, and pH into a single value between 0 and 100 [5]. Its structure involves four steps: parameter selection, sub-index calculation, weight assignment, and aggregation function [6].

Despite its adoption, WQI has limitations, such as dependence on local data and challenges when specific parameters are unavailable [5]. The subjectivity in weight assignment and the complexity of formulas can hinder its applicability [7], [8]. These challenges indicate the need for more robust and adaptable methodologies.

In this context, machine learning algorithms can provide a more comprehensive solution. These algorithms can handle large amounts of data and identify

complex patterns that are not noticeable via conventional methods. They can consider several parameters and adjust their models dynamically, overcoming the difficulties related to data scarcity or subjectivity in assigning weights. For example, in the study "Efficient Water Quality Prediction Using Supervised Machine Learning" by Umair Ahmed, et al. (2022) [2], machine learning algorithms were used to enhance water quality monitoring. The researchers achieved an impressive 85% accuracy in identifying the Water Quality Index (WQI) using only four variables: temperature, turbidity, pH, and total dissolved solids. This demonstrates that even with limited input data, machine learning can achieve accurate and efficient water quality predictions, making real-time detection systems more reliable and easier to implement.

In the real world, we often encounter imbalanced datasets and a lack of measurements of some samples during water collection. This research also analyzes how different types of datasets influence algorithm performance in imbalanced data scenarios. We apply statistical techniques such as chi-square tests to infer missing measurements to verify whether the observed distributions correspond to the expected theoretical distributions. This allows us to compare the observed frequencies in different categories with those predicted if the data followed binomial, regular, or another specific distribution. In this way, we evaluate not only the impact of the datasets but also the adequacy of the statistical models used.

In general, gradient-boosted decision trees (GBDTs) are often considered more resilient to imbalanced settings because of their focus on particularly challenging examples. However, we compared the results with the following algorithms: linear regression (LR), support vector regression (SVR), decision tree (DT), k-nearest-neighbors (KNN), extra trees (ET), gradient boosting (GB), random forest (RF), Extreme Gradient Boosting (XGB), LightBoost (LGBM), and CatBoosting (CAT).

2 Related Works

Kouadri et al. (2021) [9] analysed groundwater quality in the Illizi area, Algeria, using eight artificial intelligence algorithms, with multilinear regression (MLR) and random forest (RF) showing the best performance. MLR achieved the highest accuracy in the first scenario, with an $R = 1$ and extremely low error metrics such as MAE = 1.4572×10^{-8} and RMSE = 2.1418×10^{-8}. In the second scenario, RF demonstrated the lowest error rates, with $R = 0.9984$ and MAE = 1.9942, among others. Similarly, Mourade Azrour et al. (2021) [8] used parameters like temperature, pH, turbidity, and coliforms to evaluate models such as MLR, gradient boosting, and lasso regression. In their study, artificial neural networks (ANNs) outperformed other models like support vector machines (SVMs) and decision trees (DTs).

However, two studies have shown that SVMs can outperform ANNs. Haghiabi et al. (2018) [12] and Mohammadpour et al. (2015) [19] demonstrated that SVMs achieved higher accuracy than ANNs and other models such as feedforward propagation (FFBP) and radial basis function (RBF). This variation in

performance indicates that no single AI model is universally optimal, as model performance depends on the specific case study and parameters. Ensemble-based methods, including XGBoost and Random Forest (RF), have been highlighted as highly effective in studies by Uddin et al. (2022) [14], Uddin et al. (2023) [16], and Lap et al. (2023) [18].

3 Experiments and Datasets

3.1 Datasets

The dataset includes 737 samples from 16 points along the Paraibuna River, Brazil, collected between September 2011 and March 2023. Forty-eight parameters covering physical, chemical and biological aspects were analyzed, such as water temperature, pH, turbidity, and dissolved oxygen. The sampling stations include the Monte Serrat and Bonfante reservoirs (Fig. 1).

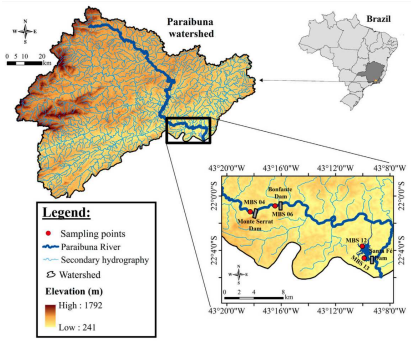


Fig. 1. Figure taken from [20]. Map showing the location of the Monte Serrat, Bonfante, and Santa Fé reservoirs along the Paraibuna River. Red points represent sampling stations: MBS04—Monte Serrat, MBS06—Bonfante, MBS12—Santa Fé I, and MBS13—Santa Fé II.

3.2 Data Preprocessing

Min-Max Normalization Min-max normalization is a common technique for scaling the values of a data set to a specific range, usually $(0, 1)$. This is done via the following formula:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

where X is the original value, $\min(X)$ is the minimum value of the variable, and $\max(X)$ is the maximum value of the variable.

This technique ensures that all variables remain on the same scale and is an essential step for the artificial intelligence algorithms used in this study.

4 Methodology

4.1 Calculation of the Water Quality Index

The Water Quality Index (WQI) was calculated using the methodology established by the Minas Gerais Water Management Institute (IGAM) for the "Águas de Minas" Project, which monitors surface waters in Minas Gerais. This method, adapted from the National Sanitation Foundation Water Quality Index (NSF-WQI) [1], assigns a value between 0 and 100 based on nine key parameters: dissolved oxygen (DO), thermotolerant coliforms, pH, nitrates (NO_3), phosphates (PT), biochemical oxygen demand (BOD), turbidity, temperature, and total solids. Each parameter contributes to the WQI calculation with a specific weight: DO has a weight of 0.17, thermotolerant coliforms 0.15, pH 0.12, BOD 0.10, nitrates 0.10, phosphates 0.10, turbidity 0.08, and total solids 0.08. The WQI is calculated as the weighted product of these water quality parameters, as outlined in equation 2 [17].

$$IQA = \sum_{i=1}^{n} (q_i^{w_i}) \tag{2}$$

where n is the number of parameters considered, q_i is the value of the parameter obtained through the specific average quality curve, and w_i is the weight associated with the parameter.

The equations obtained for the Water Quality Calculation System, as well as the average quality curves of the nine individual parameters, are presented in [17]

4.2 Parameter Selection

Strip-Based Measurements For the strip-based measurement technique, the parameters considered include pH, alkalinity ($CaCO_3$ mg L^{-1}), nitrite (mg L^{-1}), nitrate (mg L^{-1}), hexavalent chromium (mg L^{-1}), trivalent chromium (mg L^{-1}), hardness ($CaCO_3$ mg L^{-1}), mercury (mg L^{-1}), dissolved copper (mg L^{-1}), total chromium (mg L^{-1}), soluble iron (mg L^{-1}), and total iron (mg L^{-1}). These parameters are measured using test strips, which visually assess the levels of these substances in the water.

Water Quality Index (WQI) For the water quality index (WQI) technique, the selected parameters include dissolved oxygen (O.D., mg L^{-1}), pH, total coliforms (NMP 100 mL^{-1}), biochemical oxygen demand (B.O.D., mg L^{-1}), nitrate (mg L^{-1}), total phosphorus (mg L^{-1}), turbidity (NTU), and total solids (mg L^{-1}).

These parameters are essential for a comprehensive assessment of water quality and are used to calculate the WQI, which reflects the overall health of water bodies.

Gradient Boosting Model For parameter selection in the gradient boosting model, the algorithm identified the following variables as having the most significant importance: total coliforms (NMP 100 mL^{-1}), turbidity (NTU), *E. coli* (NMP 100 mL^{-1}), suspended solids (mg L^{-1}), dissolved oxygen (O.D., mg L^{-1}), dissolved aluminum (mg L^{-1}), alkalinity ($CaCO_3$ mg L^{-1}), and pH. These parameters were deemed the most influential in predicting outcomes using the Gradient Boosting algorithm.

Random Forest Model In the random forest model, the selected parameters are total coliforms (NMP 100 mL^{-1}), turbidity (NTU), *E. coli* (NMP 100 mL^{-1}), suspended solids (mg L^{-1}), dissolved oxygen (O.D., mg L^{-1}), dissolved aluminum (mg L^{-1}), dissolved copper (mg L^{-1}), and pH. These parameters were considered the most relevant for the Random Forest model's performance.

All Techniques Combined Overall, the analysis utilized all available parameters—physical, chemical, and biological—present in the dataset, encompassing a total of 48 parameters. By leveraging the full range of these parameters, we were able to comprehensively evaluate water quality through various techniques. This holistic approach ensures a thorough assessment of the factors affecting water quality.

4.3 Statistical Techniques

Chi-Square Missing Value Imputation (M1) Missing value imputation is a crucial step in data preprocessing. This work uses the chi-square test to estimate and fill in missing values.

The chi-square (χ^2) is a statistical measure that evaluates the difference between observed and expected frequencies in a categorized dataset. The formula for the chi-square test is as follow:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

where O_i represents the observed frequency and E_i are the expected frequency. This test is applied to each variable with missing values, allowing us to infer and attribute values on the basis of the observed relationships.

Missing Value Imputation with Copulas (M2) Given an n-dimensional random vector $X = (X_1, ..., X_n)$, Sklar's Theorem [23] defines that there exists a joint distribution function H with marginal functions $F_{x_1}, ..., F_{x_n}$. Consequently, there is an n-dimensional copula given by $C : [0, 1]^d \to [0, 1]$, which links the joint distribution function to its marginals, as shown in Equation 4 [**?**]:

$$H(x_1, x_2, ..., x_n) = C(F_{x_1}(x_1), F_{x_2}(x_2), ..., F_{x_n}(x_n)) \tag{4}$$

Sklar's theorem demonstrates that for any joint distribution F with marginals $F_{x_1}, ..., F_{x_n}$, there exists a copula C that satisfies this equation. If F is continuous, C is unique [25].

Copulas are particularly useful for missing value imputation, as they allow for the generation of synthetic data that accounts not only for the marginal distributions of the variables but also for their interdependencies. This helps preserve the multivariate structure of the data, which is crucial for maintaining coherence in the imputed dataset [25]. This method imputes only the missing values in the original dataset, thus preserving the integrity of the existing observations.

Sample Addition with Copulas (M3) In addition to imputing missing values, copulas can also be used to generate new samples in a dataset. By fitting marginal distributions to each variable, it is possible to model the data univariately and identify the distribution that best represents their behavior. By using various types of distributions, such as Beta, Gamma, Gaussian Kernel-Density Estimate (GaussianKDE), Gaussian (Normal), and Truncated Gaussian, one can accurately capture the data variation. This allows for the generation of new samples that preserve the statistical characteristics of the original data.

This process goes beyond simple imputation by adding new observations to the dataset, which can be beneficial in applications such as synthetic data simulation or the modeling of random variables in contexts where increasing the data volume is necessary while maintaining its original distribution.

4.4 Evaluation Metrics

Various metrics were utilized to assess the performance of the models. These include:

Mean Squared Error (MSE) Mean Squared Error (MSE) quantifies the average squared difference between the observed actual outcomes and the outcomes predicted by the model. It is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of observations.

Root Mean Squared Error (RMSE) Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors. It indicates the standard deviation of the residuals, giving insight into how well the model's predictions match the actual data. It is defined as:

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{6}$$

Mean Absolute Error (MAE) Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions without considering their direction. It is the average of the absolute differences between prediction and actual observation over the test sample. It is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

Coefficient of Determination (R^2) The Coefficient of Determination (R^2) represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

where $\bar{y}$ is the mean of the actual values.

Total Execution Time The total execution time measures the time taken to train the model and make predictions. This metric is crucial for understanding the computational efficiency of the models.

5 Results

This section presents the analysis results of the machine learning models used on our data. To ensure the consistency of results and avoid overfitting, we used cross-validation and data-splitting strategies.

First, the dataset was split into training and testing subsets. The *train-test-split* function from the Scikit-learn library was used to generate a testing subset with 20% of the data, with the remaining 80% being used for training. To comprehensively evaluate the performance of the models, we adopted the cross-validation technique. We chose to use ShuffleSplit cross-validation, which allows the generation of multiple random splits of the dataset into training and testing sets, reserving 20% of the data for testing in each split. We chose to perform ten splits to obtain a more consistent average of the performance metrics.

Table 1. Results for chi-square imputation (M1), with 10 runs.

Method	**Model**	**MSE**	**RMSE**	**MAE**	**R2**	**Total Time (s)**
Regressão Linear	X_model	0.0188 ± 0.0030	0.1368 ± 0.0112	0.1079 ± 0.0070	0.4224 ± 0.0739	0.0803
Decision Tree	X_m	0.0135 ± 0.0022	0.1157 ± 0.0095	0.0763 ± 0.0053	0.5843 ± 0.0757	0.1867
K-Nearest Neighbors	X_model	0.0123 ± 0.0026	0.1104 ± 0.0121	0.0790 ± 0.0067	0.6197 ± 0.0790	0.0923
Support Vector Regression	X_m	0.0105 ± 0.0021	0.1018 ± 0.0104	0.0782 ± 0.0069	0.6760 ± 0.0688	0.1605
Random Forest	X_m	0.0071 ± 0.0008	0.0840 ± 0.0050	0.0574 ± 0.0029	0.7808 ± 0.0315	12.3424
Extra Trees	X_m	0.0066 ± 0.0013	0.0807 ± 0.0081	0.0562 ± 0.0026	0.7975 ± 0.0369	5.1888
Gradient Boosting	X_m	0.0065 ± 0.0010	0.0807 ± 0.0064	0.0553 ± 0.0033	0.7978 ± 0.0333	4.6239
CatBoost	X_m	0.00636 ± 0.00117	0.0794 ± 0.00726	0.0560 ± 0.00314	0.8043 ± 0.0336	10.2211
Extreme Gradient Boosting	X_m	0.0061 ± 0.0012	0.0777 ± 0.0073	0.0521 ± 0.0032	0.8146 ± 0.0302	2.6501
Light Gradient Boosting	**X_m**	**0.00597 ± 0.00117**	**0.07688 ± 0.00737**	**0.05026 ± 0.00246**	**0.8180 ± 0.0320**	**1.2140**

Table 2. Results for synthetic imputation (M2), with 10 runs.

Method	Model	MSE	RMSE	MAE	R2	Total Time (s)
Regressão Linear	X_model2	0.0188 ± 0.0029	0.1367 ± 0.0108	0.1071 ± 0.0065	0.4231 ± 0.0675	0.0529
Decision Tree	X_m	0.0132 ± 0.0021	0.1145 ± 0.0090	0.0767 ± 0.0058	0.5934 ± 0.0637	0.1910
K-Nearest Neighbors	X_model	0.0126 ± 0.0028	0.1115 ± 0.0127	0.0796 ± 0.0074	0.6117 ± 0.0839	0.0578
Support Vector Regression	X_m	0.0108 ± 0.0022	0.1034 ± 0.0111	0.0793 ± 0.0071	0.6664 ± 0.0714	0.2758
Random Forest	X_model2	0.0068 ± 0.0010	0.0820 ± 0.0062	0.0547 ± 0.0034	0.7909 ± 0.0371	3.9031
Extra Trees	X_m	0.0064 ± 0.0014	0.0797 ± 0.0089	0.0555 ± 0.0037	0.8026 ± 0.0385	5.2750
CatBoost	X_m	0.00623 ± 0.00108	0.0787 ± 0.00682	0.0549 ± 0.00342	0.8083 ± 0.0304	5.9623
Gradient Boosting	X_m	0.0060 ± 0.0011	0.0770 ± 0.0076	0.0521 ± 0.0028	0.8160 ± 0.0328	4.7089
Extreme Gradient Boosting	X_m	0.00583 ± 0.00121	0.07594 ± 0.00783	0.05025 ± 0.00313	0.8225 ± 0.0328	2.1401
Light Gradient Boosting	**X_m**	**0.00576 ± 0.00106**	**0.07558 ± 0.00682**	**0.04892 ± 0.00331**	**0.8243 ± 0.0294**	**1.8547**

Table 3. Results of adding synthetic samples (M2.1), with 10 runs.

Method	Model	MSE	RMSE	MAE	R2	Total Time (s)
Linear Regression	X_model2	0.0341 ± 0.0031	0.1844 ± 0.0087	0.1520 ± 0.0084	0.3093 ± 0.0259	0.0422
K-Nearest Neighbors	X_model2	0.0200 ± 0.0033	0.1408 ± 0.0117	0.1025 ± 0.0071	0.5947 ± 0.0600	0.0624
Support Vector Regression	X_m	0.0162 ± 0.0017	0.1272 ± 0.0065	0.0974 ± 0.0049	0.6702 ± 0.0299	0.3602
Decision Tree	X_m	0.0104 ± 0.0021	0.1017 ± 0.0103	0.0662 ± 0.0046	0.7889 ± 0.0328	0.3195
Random Forest	X_m	0.00587 ± 0.00166	0.0759 ± 0.0105	0.0511 ± 0.0052	0.8813 ± 0.0295	20.4596
Extra Trees	X_m	0.0052 ± 0.0013	0.0713 ± 0.0087	0.0483 ± 0.0040	0.8955 ± 0.0218	7.6141
CatBoost	X_m	0.0050 ± 0.0012	0.0700 ± 0.0081	0.0486 ± 0.0034	0.8994 ± 0.0210	10.2867
Gradient Boosting	X_m	0.0048 ± 0.0014	0.0685 ± 0.0096	0.0463 ± 0.0043	0.9033 ± 0.0252	9.7162
XGBoost	X_m	0.0037 ± 0.0007	0.0601 ± 0.0061	0.0430 ± 0.0033	0.9274 ± 0.0151	4.0084
LightGBM	**X_m**	**0.0036 ± 0.00076**	**0.0600 ± 0.00623**	**0.0415 ± 0.00222**	**0.9276 ± 0.0155**	**2.4224**

6 Conclusion

The research presented in this study highlights the importance of integrating advanced machine learning algorithms with various imputation techniques to enhance water quality assessment. Traditional methods like the Water Quality Index (WQI), while widely used, have inherent limitations, including sensitivity to missing data and subjective weighting of parameters. These challenges underscore the need for more robust methodologies capable of handling imbalanced datasets and missing measurements, which are common in real-world scenarios.

By applying chi-square imputation (M1), synthetic imputation (M2), and synthetic sample augmentation (M3), we demonstrated that more advanced imputation methods and machine learning models, such as Extra Trees, Gradient Boosting, and Light Gradient Boosting, consistently outperformed simpler models like Linear Regression and Decision Tree. The results showed a clear trend: more sophisticated imputation strategies and models led to significant reductions in error metrics, particularly MSE and RMSE, and enhanced prediction accuracy.

Light Gradient Boosting emerged as the most effective algorithm across all approaches, proving its robustness in handling complex datasets with missing information. The addition of synthetic samples further improved the performance of Gradient Boosting and CatBoost, although Light Gradient Boosting remained the most reliable method overall.

This study confirms that machine learning algorithms, when combined with effective imputation techniques, provide a powerful solution to the limitations of traditional water quality assessment methods. The application of these advanced approaches not only improves the accuracy of predictions but also makes water

quality monitoring more adaptable to real-world data challenges, such as imbalanced and incomplete datasets. Future research could explore the integration of additional variables and real-time monitoring systems to further refine the predictive models and support more proactive water management strategies.

7 Data availability statement

The base is not openly available, but can be requested from the authors with the necessary justifications.

References

1. National Sanitation Foundation (NSF), *Water Quality Index (WQI): A Tool for Communication and Water Quality Management*, 1970. Available at: https://www.nsf.org
2. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J., *Efficient Water Quality Prediction Using Supervised Machine Learning*, *Water*, **11**(11), 2210, 2019. https://www.mdpi.com/2073-4441/11/11/2210 doi10.3390/w11112210
3. Ahmed, J., Wong, L. P., Chua, Y. P., Channa, N., Mahar, R. B., Yasmin, A., VanDerslice, J. A., & Garn, J. V. (2020). Quantitative microbial risk assessment of drinking water quality to predict the risk of waterborne diseases in primary-school children. *International Journal of Environmental Research and Public Health*, **17**(8), 2774. https://www.mdpi.com/1660-4601/17/8/2774. https://doi.org/10.3390/ijerph17082774
4. Silva, E. R. A. da C. (2018). *Agenda 2030: ODS-Metas nacionais dos objetivos de desenvolvimento sustentável*. Instituto de Pesquisa Econômica Aplicada (Ipea).
5. CETESB - Companhia de Tecnologia de Saneamento Ambiental. (2006). *Índices de qualidade das águas interiores do Estado de São Paulo, 2006*. http://www.cetesb.sp.gov.br/Agua/rios/indice.asp.
6. Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, **122**, 107218. Elsevier.
7. Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, **169**, 808-828. https://www.sciencedirect.com/science/article/pii/S0957582022010473. https://doi.org/10.1016/j.psep.2022.11.073
8. Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, **8**(2), 2793-2801. Springer.
9. Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science*, 11(12), 190. Springer.
10. Sakizadeh, M. (2016). Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*, 2, 1–9. Springer.

11. Othman, F., Alaaeldin, M. E., Seyam, M., Ahmed, A. N., Teo, F. Y., Ming Fai, C., Afan, H. A., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2020). Efficient river water quality index prediction considering minimal number of inputs variables. *Engineering Applications of Computational Fluid Mechanics*, 14(1), 751–763. Taylor & Francis.
12. Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1), 3–13. IWA Publishing.
13. Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering*, 9(1), 104599. Elsevier.
14. Uddin, M. G., Nash, S., Diganta, M. T. M., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*, 321, 115923. Elsevier.
15. Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249, 126169. Elsevier.
16. Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, 169, 808–828. Elsevier.
17. SEMAD and UCEMG / PNMA II, *Sistema de cálculo da qualidade da água (SCQA) – Estabelecimento das equações do índice de qualidade das águas (IQA). Relatório I*, Secretaria de Estado de Meio Ambiente e Desenvolvimento / Universidade do Estado de Minas Gerais, Minas Gerais, Brasil, 2005.
18. Lap, B. Q., Du Nguyen, H., Hang, P. T., Phi, N. Q., Hoang, V. T., Linh, P. G., Hang, B. T. T., & others. (2023). Predicting water quality index (WQI) by feature selection and machine learning: a case study of An Kim Hai irrigation system. *Ecological Informatics*, 74, 101991. Elsevier.
19. Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., Ghani, A. A., & Chan, N. W. (2015). Prediction of water quality index in constructed wetlands using support vector machine. *Environmental Science and Pollution Research*, **22**, 6208–6219.
20. Resende, Nathália da Silva, Santos, Juliana Barreto Oliveira dos, Josué, Iollanda Ivanov Pereira, Barros, Nathan Oliveira, and Cardoso, Simone Jaqueline. *Comparing Spatio-Temporal Dynamics of Functional and Taxonomic Diversity of Phytoplankton Community in Tropical Cascading Reservoirs*. Frontiers in Environmental Science, vol. 10, 2022. URL: https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2022.903180. https://doi.org/10.3389/fenvs.2022.903180.
21. CHERUBINI, U.; LUCIANO, E.; VECCHIATO, W.; Copula Methods in Finance. John Wiley & Sons Ltd 2004. ISBN: 9781118673331.
22. GENEST, C.; FAVRE, A. C.; Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. Journal of Hydrologic Engineering. pp. 347-368, vol. 12, n. 4, 2007.
23. SKLAR, A. Distribution functions of n dimensions and margins. Institute of Statistics at the University of Paris. pp. 229-231, 1959.
24. SILVA, T. A. Estudo do desempenho da combinação de preditores baseados em cópulas e máquinas de vetor de suporte para séries temporais úteis ao desenvolvimento sustentável. 2020. Dissertação (Programa de Pós-Graduação em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco, Recife.
25. NELSEN, R. B. An Introduction to Copulas. Springer Series in Statistics. Springer New York, 2007. ISBN: 9780387286785.