

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

WILSON DOS SANTOS JUNIOR

**PREDIÇÃO DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA) NA REGIÃO
METROPOLITANA DO VALE DO PARAÍBA E LITORAL NORTE POR MEIO DE
TÉCNICAS DE APRENDIZAGEM DE MÁQUINA.**

Guaratinguetá

2026

Wilson dos Santos Junior

**PREDIÇÃO DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA) NA REGIÃO
METROPOLITANA DO VALE DO PARAÍBA E LITORAL NORTE POR MEIO DE
TÉCNICAS DE APRENDIZAGEM DE MÁQUINA. :**

Dissertação apresentada à Faculdade de Engenharia e Ciências do Campus de Guaratinguetá, Universidade Estadual Paulista, sendo a etapa de Defesa como parte dos requisitos para obtenção do título de Mestre em Engenharia.

Orientador: Prof. Dra. Marcela Aparecida
Guerreiro Machado

Coorientador: Prof. Dr. Fabrício Maciel
Gomes

Guaratinguetá

2026

S237p	Santos Junior, Wilson dos Predição do índice de qualidade da água (IQA) na região metropolitana do Vale do Paraíba e Litoral Norte por meio de técnicas de aprendizagem de máquina / Wilson dos Santos Junior - Guaratinguetá, 2026. 76 f : il. Bibliografia: f. 70-75 Dissertação (Mestrado) – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2026. Orientadora: Prof^a. Dr^a. Marcela A. Guerreiro Machado Coorientador: Prof. Dr. Fabrício Maciel Gomes 1. Água - Qualidade. 2. Aprendizado do computador. 3. Desenvolvimento sustentável. I. Título. CDU 628.1(043)
-------	---

Luciana Máximo
Bibliotecária/CRB-8 3595

IMPACTO POTENCIAL DESTA PESQUISA

O impacto esperado da pesquisa fundamenta-se na convergência entre inteligência artificial e gestão hídrica sustentável. No âmbito científico, técnico e inovador, o estudo inova ao demonstrar nos testes realizados que a predição individualizada de parâmetros físico-químicos e biológicos para reconstrução do IQA supera a predição direta, viabilizando a integração com sistemas alinhados à Indústria 4.0. No aspecto econômico, a metodologia reduz custos operacionais de amostragem e oferece uma alternativa acessível em linguagem R, democratizando ferramentas preditivas para organizações com restrições orçamentárias. Alinhada ao desenvolvimento sustentável e ao Objetivo de Desenvolvimento Sustentável nº 6 da ONU, a pesquisa fortalece o monitoramento e a preservação dos recursos hídricos. Em termos de inserção local, regional e nacional, a solução atende inicialmente à Região Metropolitana do Vale do Paraíba e Litoral Norte, possuindo capacidade de generalização que viabiliza sua expansão para bacias hidrográficas em todo o território nacional e potencial de internacionalização. Por fim, o trabalho gera relevante impacto social, educacional e cultural ao promover a proteção da saúde pública e a segurança hídrica por meio do aprendizado de máquina, disseminando uma cultura orientada a dados no setor.

POTENTIAL IMPACT OF THIS RESEARCH

The expected impact of the research is grounded in the convergence between artificial intelligence and sustainable water management. In the scientific, technical, and innovative scope, the study innovates by demonstrating in conducted tests that the individualized prediction of physicochemical and biological parameters for WQI (Water Quality Index) reconstruction outperforms direct prediction, enabling integration with systems aligned with Industry 4.0. From an economic perspective, the methodology reduces sampling operational costs and offers an accessible alternative in the R programming language, democratizing predictive tools for organizations with budget constraints. Aligned with sustainable development and UN Sustainable Development Goal 6 (SDG 6), the research strengthens the monitoring and preservation of water resources. In terms of local, regional, and national reach, the solution initially addresses the Metropolitan Region of Vale do Paraíba and Litoral Norte, possessing a generalization capability that enables its expansion to river basins across the national territory, along with potential for internationalization. Finally, the work generates a significant social, educational, and cultural impact by promoting public health protection and water security through machine learning, fostering a data-driven culture in the sector.

WILSON DOS SANTOS JUNIOR

Predição do Índice de Qualidade da Água (IQA) na Região Metropolitana do Vale do Paraíba e Litoral Norte por Meio de Técnicas de Aprendizado de Máquina.

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Faculdade de Engenharia e Ciências de Guaratinguetá, para obtenção do título de Mestre em Engenharia

Data da defesa: 18/06/2026

Banca Examinadora:

Documento assinado digitalmente
 **MARCELA APARECIDA GUERREIRO MACHADO DI**
Data: 07/07/2026 16:59:29-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. MARCELA APARECIDA GUERREIRO MACHADO DE FREITAS

Orientador - UNESP

Documento assinado digitalmente
 **PALOMA MARIA SILVA ROCHA RIZOL**
Data: 08/07/2026 17:33:03-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. PALOMA MARIA SILVA ROCHA RIZOL

UNESP

Documento assinado digitalmente
 **NILO ANTONIO DE SOUZA SAMPAIO**
Data: 30/06/2026 20:58:49-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. NILO ANTONIO DE SOUZA SAMPAIO

UERJ

*“Mas vocês devem ser fortes e não desanimar, pois o trabalho de vocês será recompensado.”
(2 Crônicas 15:7)*

RESUMO

A preservação e o monitoramento da qualidade dos recursos hídricos representam desafios fundamentais para o desenvolvimento sustentável, alinhando-se diretamente às metas estabelecidas pelo Objetivo de Desenvolvimento Sustentável nº 6 da Organização das Nações Unidas. Diante da complexidade dos ecossistemas aquáticos e da frequente dificuldade na obtenção de dados completos, o que muitas vezes inviabiliza o cálculo tradicional do índice, este trabalho investiga a aplicação de técnicas de Aprendizagem de Máquina para a predição do Índice de Qualidade da Água (IQA) na Região Metropolitana do Vale do Paraíba e Litoral Norte (RMVALE-LN). A pesquisa utiliza uma base de dados robusta, compreendendo o período de 2013 a 2023, com informações provenientes da CETESB e do sistema InfoÁgua, processadas integralmente em linguagem R. O foco metodológico reside na utilização do algoritmo de Floresta Aleatória (*Random Forest*) para superar a limitação de dados incompletos, comparando duas abordagens distintas: a predição direta do valor final do IQA e, diferentemente das abordagens convencionais, a predição individual dos parâmetros físico-químicos e biológicos do índice para sua posterior reconstrução matemática.

Os resultados obtidos demonstram que a abordagem de predição dos parâmetros individuais para o posterior cálculo do IQA apresenta um desempenho superior e mais consistente do que a tentativa de prever o índice de forma direta, mostrando-se uma estratégia eficaz para lidar com lacunas de monitoramento. A eficácia dos modelos foi validada por meio de métricas para tarefas de regressão, com destaque para o Erro Percentual Absoluto Médio (MAPE), que em diversos cenários manteve-se abaixo de 10%, conferindo às previsões o status de “altamente precisas” conforme a literatura especializada. Além da precisão estatística, o estudo revela uma importante capacidade de generalização dos modelos, sugerindo que padrões identificados em determinados pontos de monitoramento podem ser aplicados a áreas correlatas, dispensando assim a necessidade de criação de modelos individualizados para cada localidade específica. Essa capacidade preditiva não apenas soluciona o problema de dados ausentes, mas também fundamenta a viabilidade da redução de parâmetros medidos em campo, permitindo um monitoramento mais ágil e de menor custo, sem perda de confiabilidade técnica.

PALAVRAS-CHAVE: Aprendizagem de Máquina; IQA; Floresta Aleatória; Redução de parâmetros; RMVALE-LN.

ABSTRACT

The preservation and monitoring of water resource quality represent fundamental challenges for sustainable development, aligning directly with the goals established by the United Nations Sustainable Development Goal 6. Given the complexity of aquatic ecosystems and the frequent difficulty in obtaining complete data—which often prevents the traditional calculation of the index—this study investigates the application of Machine Learning techniques to predict the Water Quality Index (WQI) in the Metropolitan Region of Vale do Paraíba and Litoral Norte (RMVALE-LN). The research utilizes a robust database covering the period from 2013 to 2023, with data from CETESB and the InfoÁgua system, fully processed using the R language. The methodological focus lies in using the Random Forest algorithm to overcome the limitation of incomplete data, comparing two distinct approaches: the direct prediction of the final WQI value and, unlike conventional approaches, the individual prediction of the index's physicochemical and biological parameters for its subsequent mathematical reconstruction.

The results demonstrate that the approach of predicting individual parameters for the subsequent calculation of the WQI yields superior and more consistent performance compared to attempting to predict the index directly, proving to be an effective strategy for dealing with monitoring gaps. The effectiveness of the models was validated using regression metrics, with emphasis on the Mean Absolute Percentage Error (MAPE), which remained below 10% in several scenarios, granting the predictions a "highly accurate" status according to specialized literature. In addition to statistical accuracy, the study reveals a significant generalization capability of the models, suggesting that patterns identified at certain monitoring points can be applied to correlated areas, thereby eliminating the need to create individualized models for each specific location. This predictive capability not only solves the problem of missing data but also supports the feasibility of reducing parameters measured in the field, enabling faster and lower-cost monitoring without compromising technical reliability.

KEYWORDS: Machine Learning; WQI; Random Forest; Parameter reduction; RMVALE-LN.

LISTA DE ILUSTRAÇÕES

Figura 1	Informações gerais das buscas na base Scopus.	16
Figura 2	Produção científica anual.	17
Figura 3	Comparação entre o número de publicações brasileiras e indianas.	18
Figura 4	Hierarquia de aprendizado estendida.	19
Figura 5	Diferenciação entre duas classes por meio de margens de suporte.	22
Figura 6	Aplicação da Função Kernel para separação linear em maior dimensão.	22
Figura 7	Diagrama esquemático do algoritmo de Floresta Aleatória.	23
Figura 8	Cinco estágios de um modelo de IQA.	26
Figura 9	Curvas Médias de Variação de Qualidade das Águas.	30
Figura 10	Proporção de corpos d'água com boa qualidade da água ambiente, 2017–2020 (%)	32
Figura 11	Infográfico ONU - ODS.	33
Figura 12	Classificação da pesquisa.	34
Figura 13	Localização da RMVALE-LN.	35
Figura 14	Fluxograma da metodologia do trabalho.	37
Figura 15	Dados IQA das cidades por sazonalidade.	40
Figura 16	Localização do ponto SANT00100.	41
Figura 17	Histograma dos valores de IQA em Jambeiro.	41
Figura 18	Dispersão do IQA em Cruzeiro.	42
Figura 19	Localização em Cruzeiro dos pontos JACU 02900 e PONT 04950.	43
Figura 20	Pontos com melhor IQA na estação seca.	44
Figura 21	Pontos com melhor IQA na estação úmida.	44
Figura 22	Fluxo para tratamento de dados direcionado ao modelo de FA.	46
Figura 23	Métricas para predição dos parâmetros excluídos - PARB02310.	48
Figura 24	IQA predito versus IQA calculado por parâmetros preditos - PARB02310.	48
Figura 25	Métricas do IQA predito versus IQA calculado por parâmetros preditos - PARB02310.	49
Figura 26	Métricas para predição dos parâmetros excluídos - SANT00100.	50
Figura 27	IQA predito versus IQA calculado por parâmetros preditos - SANT00100.	50
Figura 28	Métricas do IQA predito versus IQA calculado por parâmetros preditos - SANT00100.	51
Figura 29	Métricas para predição dos parâmetros excluídos - UNNA02800.	52
Figura 30	IQA predito versus IQA calculado por parâmetros preditos - UNNA02800.	52
Figura 31	Métricas do IQA predito versus IQA calculado por parâmetros preditos - UNNA02800.	53
Figura 32	Métricas para predição dos parâmetros excluídos - GUAT02800.	54

Figura 33	IQA predito versus IQA calculado por parâmetros preditos - GUAT02800.	54
Figura 34	Métricas do IQA predito versus IQA calculado por parâmetros preditos - GUAT02800.	55
Figura 35	Métricas para predição dos parâmetros excluídos - ALIM02950.	56
Figura 36	IQA predito versus IQA calculado por parâmetros preditos - ALIM02950.	56
Figura 37	Métricas do IQA predito versus IQA calculado por parâmetros preditos - ALIM02950.	57
Figura 38	Métricas para predição dos parâmetros excluídos - TAVE02950.	57
Figura 39	IQA predito versus IQA calculado por parâmetros preditos - TAVE02950.	58
Figura 40	Métricas do IQA predito versus IQA calculado por parâmetros preditos - TAVE02950.	58
Figura 41	Métricas do IQA predito versus IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: ALIM02950.	59
Figura 42	Métricas do IQA predito versus IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: GUAT02800.	59
Figura 43	Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: PARB02310.	59
Figura 44	Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: SANT00100.	60
Figura 45	Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: TAVE02950.	60
Figura 46	Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: UNNA02800.	60
Figura 47	Fluxo de execução para predição do IQA com utilização iterativa de valores preditos.	61
Figura 48	Métricas de desempenho para predição do IQA com utilização iterativa de valores preditos.	62
Figura 49	Predição do IQA com utilização iterativa de valores preditos - Modelo: PARB02900, Dados: PARB02900.	63
Figura 50	Predição do IQA com utilização iterativa de valores preditos - Modelo: PARB02900, Dados: demais pontos.	64
Figura 51	Fluxograma do código VB <i>script</i>	66
Figura 52	<i>Software</i> supervisorio (Aveva Edge) em execução integrado à linguagem R.	67

LISTA DE TABELAS

Tabela 1 – Termos de pesquisa utilizados no Scopus.	16
Tabela 2 – Critérios qualitativos do MAPE.	26
Tabela 3 – Parâmetros de Qualidade da Água do IQA e respectivos pesos.	31
Tabela 4 – Classificação de Qualidade da Água.	31
Tabela 5 – Descrição das variáveis do conjunto de dados.	35
Tabela 6 – Relação dos pontos de monitoramento selecionados por município.	45
Tabela 7 – Tabela de variáveis excluídas por teste de IQA.	46
Tabela 8 – Dados temporais de Oxigênio Dissolvido formatados para o modelo de FA.	47
Tabela 9 – Parâmetros de qualidade da água e IQA para validação do ponto PARB02310.	47
Tabela 10 – Parâmetros de qualidade da água e IQA para validação do ponto SANT00100.	49
Tabela 11 – Parâmetros de qualidade da água e IQA para validação do ponto UNNA02800.	51
Tabela 12 – Parâmetros de qualidade da água e IQA para validação do ponto GUAT02800.	53
Tabela 13 – Parâmetros de qualidade da água e IQA para validação do ponto ALIM02950.	55
Tabela 14 – Parâmetros de qualidade da água e IQA para validação do ponto TAVE02950.	56

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvores de Decisão
AM	Aprendizagem de Máquina
ANA	Agência Nacional de Águas e Saneamento Básico
ANS	Aprendizagem Não Supervisionada
AR	Aprendizagem por Reforço
AS	Aprendizagem Supervisionada
CETESB	Companhia Ambiental do Estado de São Paulo
DBO	Demanda Bioquímica de Oxigênio
DAEE	Departamento de Águas e Energia Elétrica
FA	Floresta Aleatória
FN	Falso Negativo
FP	Falso Positivo
IQA	Índice de Qualidade da Água
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
ML	<i>Machine Learning</i>
MSE	Erro Quadrático Médio
MVS	Máquinas de Vetores de Suporte
NFS	<i>National Sanitation Foundation</i>
N-total	Nitrogênio Total
OD	Oxigênio Dissolvido
ODS	Objetivos de Desenvolvimento Sustentável
ONU	Organização das Nações Unidas
pH	Potencial Hidrogeniônico

P-total	Fósforo Total
R ²	Coeficiente de Determinação
RL	<i>Reinforcement Learning</i>
RMSE	Raiz do Erro Quadrático Médio
RMVALE-LN	Região Metropolitana do Vale do Paraíba e Litoral Norte
RP	Regressão Polinomial
SCADA	<i>Supervisory Control and Data Acquisition</i>
TFP	Taxa de Falsos Positivos
TPV	Taxa de Positivos Verdadeiros
TVP	Taxa de Verdadeiros Positivos
UGRHI	Unidades de Gerenciamento de Recursos Hídricos
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1	INTRODUÇÃO	14
1.1	CONTEXTUALIZAÇÃO DA PESQUISA	14
1.2	QUESTÕES DE PESQUISA	15
1.3	JUSTIFICATIVA E CONTRIBUIÇÕES CIENTÍFICAS E TECNOLÓGICAS	16
1.4	DELIMITAÇÃO DA PESQUISA	18
1.5	OBJETIVO GERAL	18
1.6	OBJETIVOS ESPECÍFICOS	18
2	FUNDAMENTAÇÃO TEÓRICA	19
2.1	APRENDIZAGEM DE MÁQUINA	19
2.1.1	APRENDIZAGEM POR REFORÇO	20
2.1.2	APRENDIZAGEM NÃO SUPERVISIONADA	20
2.1.3	APRENDIZAGEM SUPERVISIONADA	20
2.1.3.1	Regressão Polinomial	20
2.1.3.2	Máquina de Vetores de Suporte	21
2.1.3.3	Floresta Aleatória	23
2.2	MÉTRICAS DE AVALIAÇÃO DE MODELOS DE Aprendizagem de Má- quina (AM)	23
2.2.1	Métricas para Tarefas de Regressão	23
2.2.1.1	Erro Médio Absoluto	24
2.2.1.2	Erro Quadrático Médio	24
2.2.1.3	Raiz do Erro Quadrático Médio	25
2.2.1.4	Coefficiente de Determinação R^2	25
2.2.1.5	Erro Percentual Absoluto Médio	25
2.3	ÍNDICE DE QUALIDADE DA ÁGUA.	26
2.3.1	Parâmetros Físico-Químicos, Biológicos e Microbiológicos.	27
2.3.2	Metodologia para Cálculo do IQA.	29
2.3.3	Classificação e Interpretação dos Resultados do IQA.	31
2.3.4	Importância do IQA para o Desenvolvimento Sustentável.	31
3	MÉTODO	34
3.1	Materiais	34
3.2	Métodos e Procedimentos	36
4	RESULTADOS OBTIDOS	40
4.1	Análise Descritiva dos Dados	40

4.1.1	Análise dos dados de Jambeiro	41
4.1.2	Análise dos dados de Cruzeiro	42
4.1.3	Comparativo das análises descritivas	43
4.2	Seleção, Manipulação e Aplicação dos dados de Índice de Qualidade da Água (IQA) ao modelo de AM	44
4.2.1	Manipulação dos dados para treinamento de modelos de Floresta Aleatória	45
4.3	Execução dos modelos Floresta Aleatória (FA) e comparação dos resultados.	47
4.3.1	Aplicação dos dados do ponto de monitoramento PARB02310 no modelo FA	47
4.3.2	Aplicação dos dados do ponto de monitoramento SANT00100 no modelo FA	48
4.3.3	Aplicação dos dados do ponto de monitoramento UNNA02800 no modelo FA	50
4.3.4	Aplicação dos dados do ponto de monitoramento GUAT02800 no modelo FA	52
4.3.5	Aplicação dos dados do ponto de monitoramento ALIM02950 no modelo FA	54
4.3.6	Aplicação dos dados do ponto de monitoramento TAVE02950 no modelo FA	55
4.3.7	Aplicação de dados de diversos pontos de monitoramento aos modelos do ponto PARB02900	58
4.3.8	Utilização Iterativa de Valores Preditos em Horizontes de Longo Prazo .	61
4.4	Aplicação dos modelos gerados em uma simulação de ambiente industrial de tratamento de água	65
4.4.1	Integração da Linguagem R com Sistema SCADA	65
5	CONCLUSÃO	68
	REFERÊNCIAS	70
	DADOS CURRICULARES	76

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO DA PESQUISA

Em setembro de 2015, a Organização das Nações Unidas (ONU) lançou uma agenda com 17 objetivos e 169 metas a serem atingidas até o ano de 2030, conhecidos como os Objetivos de Desenvolvimento Sustentável (ODS). Com o intuito de erradicar a pobreza, proteger o meio ambiente e o clima, e garantir que todas as pessoas tenham condições dignas de vida, esses 17 objetivos podem ser divididos em quatro dimensões principais: Social, Ambiental, Econômica e Institucional. Em seu objetivo número 6, aborda-se especificamente a disponibilidade de água potável e saneamento para todos (ONU, 2024a) (ONU, 2024c).

A dependência da água para a sustentação da vida humana, associada ao crescente índice de poluição em todo o planeta, faz da análise da qualidade desse recurso um tema de fundamental importância na atualidade (Azrour *et al.*, 2022). O IQA é uma das técnicas mais utilizadas para descrever a qualidade dos corpos hídricos, baseando-se em fatores físicos, químicos e biológicos combinados em um único valor (Chidiac *et al.*, 2023). Assim, esse conjunto de parâmetros é integrado para produzir um indicador sintético de qualidade, o que reduz expressivamente o volume de dados e as complexidades matemáticas (Aljanabi *et al.*, 2021); entretanto, a ausência de um único parâmetro inviabiliza o cálculo final do índice (Pereira *et al.*, 2013).

A alta frequência na aquisição de dados por sistemas automatizados de coleta e eventuais falhas nos sensores podem gerar dados faltantes (Badrudeen *et al.*, 2026). Da mesma forma, a inviabilidade de acesso aos pontos de amostragem — seja por dificuldades geográficas ou por fatores excepcionais, como a pandemia de COVID-19 — também resulta em lacunas na base de dados, impossibilitando o cálculo do IQA. Enquanto os métodos convencionais recorrem à interpolação para preencher essas lacunas, o trabalho de (Wang, J. *et al.*, 2025) apresenta a proposta de utilizar modelos preditivos para estimar esses dados ausentes de forma mais dinâmica. Diversos estudos, como os de (Azrour *et al.*, 2022), (Krishnan *et al.*, 2024) e (Lin *et al.*, 2023), apontam vantagens significativas — como a diminuição do número de parâmetros medidos e a redução da complexidade operacional — ao predizer o valor do índice de qualidade da água por meio de técnicas de AM.

A AM, termo oriundo da expressão em língua inglesa *Machine Learning*, consiste no processo de fornecer dados a um sistema computacional para que este identifique padrões e gere novos conhecimentos. Atualmente, esse conceito define a capacidade de evolução autônoma e aprendizado dos computadores (Kannadhasan; Venusamy; Nagarajan, 2023). Segundo (Kayikci; Khoshgoftaar, 2024), as técnicas de AM buscam mimetizar os mecanismos de aprendizado do cérebro humano. O forte desenvolvimento dessa área, nos últimos anos, impulsionou o surgimento de variadas soluções voltadas a facilitar ou aprimorar as atividades humanas. Desse modo, os métodos de AM encontram aplicações em setores diversos, como o comércio, as

ciências e a tecnologia, tornando os processos de tomada de decisão consideravelmente mais assertivos (Jordan; Mitchell, 2015). Porém, esse avanço não ocorreu de forma linear, visto que houve períodos de menor interesse e desaquecimento ao longo da trajetória de evolução do AM (Tursunalieva *et al.*, 2024).

Em (Jordan; Mitchell, 2015), justifica-se que esse crescimento acentuado foi impulsionado pelo surgimento de novos algoritmos e teorias de aprendizagem, aliado à expansão contínua da disponibilidade de dados e à redução dos custos computacionais. As técnicas de AM permitem que os computadores processem grandes volumes de informação para identificar padrões, realizar classificações e reconhecer estruturas, sem que tenham sido explicitamente programados para cada uma dessas tarefas. Em outros termos, a AM viabiliza a inferência de modelos probabilísticos e comportamentais fundamentados diretamente nos dados observados (Kayikci; Khoshgoftaar, 2024).

Atualmente, com os avanços tecnológicos, estações de monitoramento situadas em múltiplos pontos de medição geram fluxos contínuos e massivos de informação. Esse tipo de conjunto de dados — caracterizado por sua grande escala, diversidade e complexidade — é rotulado como *Big Data* (Weese *et al.*, 2016), e grande parte dos métodos aplicados à sua análise provém diretamente de subcampos da AM.

Dessa forma, o presente trabalho tem como objetivo analisar a qualidade da água na Região Metropolitana do Vale do Paraíba e Litoral Norte (RMVALE-LN) entre os anos de 2013 e 2023 por meio do uso de AM para a predição do índice. Busca-se verificar a viabilidade dessa predição a partir da redução do número de variáveis independentes, bem como avaliar a eficácia de prever os parâmetros individualmente para a posterior reconstrução matemática do IQA. Para o desenvolvimento dos modelos preditivos, optou-se pelo algoritmo de FA, dada a sua capacidade de modelar relações não lineares e sua elevada robustez a ruídos (Breiman, 2001), demonstrando excelente adaptação à complexidade de conjuntos de dados ambientais heterogêneos (Bouchraki *et al.*, 2025). Além disso, a FA lida eficazmente com a alta dimensionalidade dos dados e com problemas de multicolinearidade, caracterizando-se como uma técnica veloz e altamente insensível ao sobreajuste (*overfitting*) (Belgiu; Drăguț, 2016).

1.2 QUESTÕES DE PESQUISA

- Como se distribuem os valores do IQA na RMVALE-LN entre os anos de 2013 e 2023?
- É eficiente utilizar os modelos de AM para prever os valores do índice de qualidade da água?
- É eficiente prever o valor do IQA excluindo-se alguns parâmetros utilizados pelo modelo de AM?
- É eficiente calcular o IQA combinando valores de parâmetros medidos e preditos?

- É eficiente utilizar valores preditos por AM de forma iterativa para a previsão de cenários futuros de longo prazo?

1.3 JUSTIFICATIVA E CONTRIBUIÇÕES CIENTÍFICAS E TECNOLÓGICAS

Para fundamentar a escolha do tema de pesquisa e demonstrar a existência de uma lacuna na literatura, realizou-se uma busca inicial na base de dados *Scopus*, conforme apresentado na Tabela 1. A ausência de resultados na linha 3 indica essa lacuna científica.

Tabela 1 – Termos de pesquisa utilizados no Scopus.

Linha	Termos de Pesquisa	Resultados
1	(TITLE ("water quality index") AND TITLE ("Predictions"))	87
2	(TITLE ("water quality index") AND TITLE ("machine learning"))	82
3	(TITLE ("water quality index") AND TITLE ("Predictions") AND TITLE-ABS-KEY ("Reducing parameters"))	0

Fonte: Autor.

Para uma análise aprofundada dos dados levantados, utilizou-se o *software* R em conjunto com a biblioteca *Bibliometrix*. Os dados analisados correspondem aos resultados das buscas das linhas 1 e 2, os quais foram consolidados e tratados no ambiente R, sendo posteriormente exibidos por meio do pacote *Shiny* e da referida biblioteca. Como resultado, gerou-se uma base de dados com documentos publicados entre os anos de 2002 e 2025, apresentando uma taxa de crescimento anual de 6,21%, composta por 147 documentos e 402 autores. Um resumo dessa base pode ser visto na Figura 1.

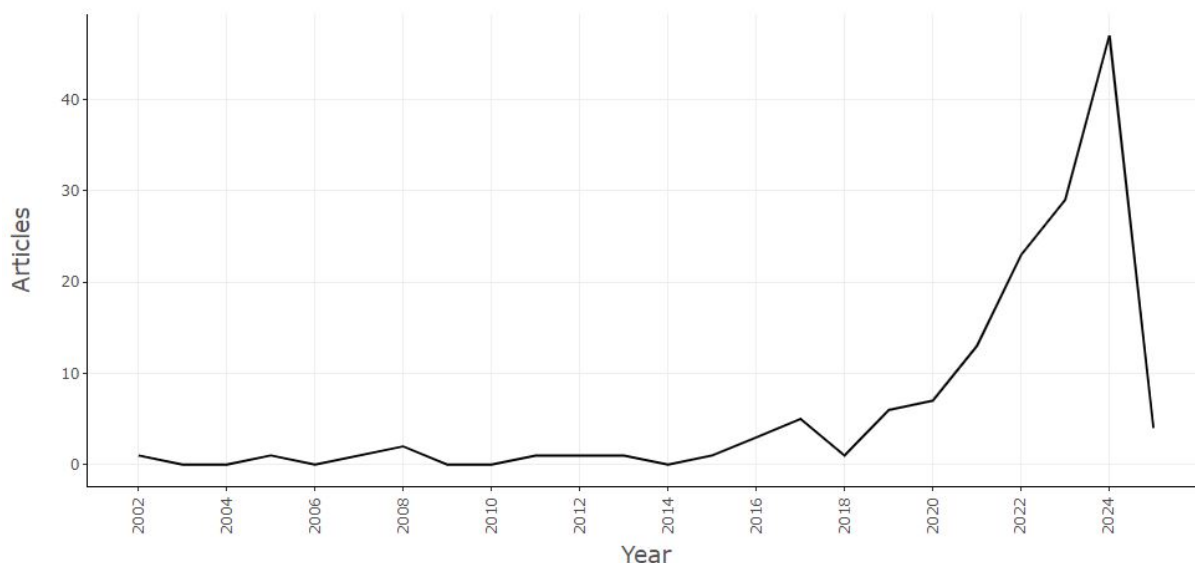
Figura 1 – Informações gerais das buscas na base Scopus.



Fonte: Autor.

Dentre as opções de visualização disponibilizadas pelo *Bibliometrix*, o gráfico de produção científica anual evidencia a contemporaneidade do tema pesquisado. Na Figura 2, é possível observar uma ascensão mais acentuada da curva de publicações a partir do ano de 2020.

Figura 2 – Produção científica anual.



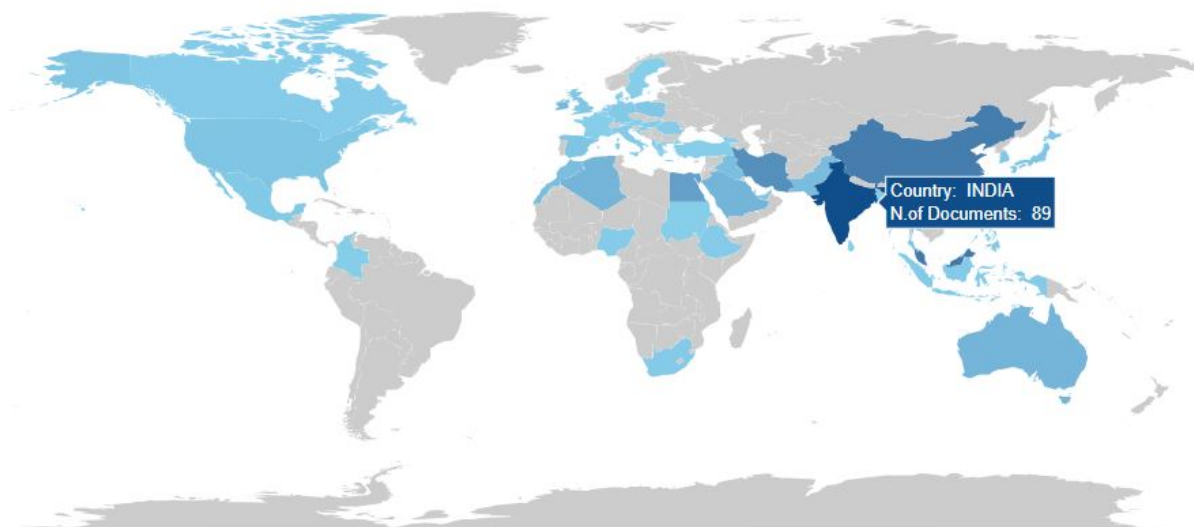
Fonte: Autor.

Como outro fator que justifica a relevância desta pesquisa, destaca-se a escassez de publicações brasileiras sobre o tema: dentre os artigos retornados na busca estruturada na base de dados da *Scopus*, nenhum documento foi localizado. Esse cenário mostra-se muito inferior quando comparado ao de outras nações, como a Índia, que ultrapassa 80 publicações, conforme demonstrado na Figura 3. Por fim, a relevância deste estudo fundamenta-se em lacunas críticas identificadas tanto na literatura quanto na gestão regional:

- **Inexistência de modelos:** Não há modelos de AM aplicados especificamente para a predição da qualidade da água na RMVALE-LN;
- **Inovação Regional:** Ausência de estudos que integrem dados da CETESB, técnicas de AM e estratégias de redução de parâmetros para essa região;
- **Escassez Operacional:** Poucos estudos focam na operacionalização do monitoramento dos rios do Vale do Paraíba;
- **Variabilidade Hidrográfica:** A variabilidade regional é pouco explorada em modelos preditivos;
- **Monitoramento Contínuo:** Falta de soluções práticas para o acompanhamento contínuo da qualidade hídrica regional.

Dessa forma, diante de todos os fatores apresentados, justifica-se a relevância deste trabalho.

Figura 3 – Comparação entre o número de publicações brasileiras e indianas.



Fonte: Autor.

1.4 DELIMITAÇÃO DA PESQUISA

Esta pesquisa limita-se a analisar os dados de medições da CETESB referentes à Região Metropolitana do Vale do Paraíba e Litoral Norte do Estado de São Paulo, compreendendo o período entre os anos de 2013 e 2023.

1.5 OBJETIVO GERAL

O objetivo geral deste trabalho é avaliar o desempenho de modelos de AM na predição do IQA na RMVALE-LN.

1.6 OBJETIVOS ESPECÍFICOS

- Investigar o impacto da redução de parâmetros do IQA no desempenho dos modelos;
- Verificar a viabilidade do cálculo do IQA combinando parâmetros medidos e preditos;
- Avaliar a capacidade de generalização dos modelos em diferentes pontos da RMVALE-LN;
- Analisar o desempenho da utilização iterativa de valores preditos para estimativas de longo prazo.

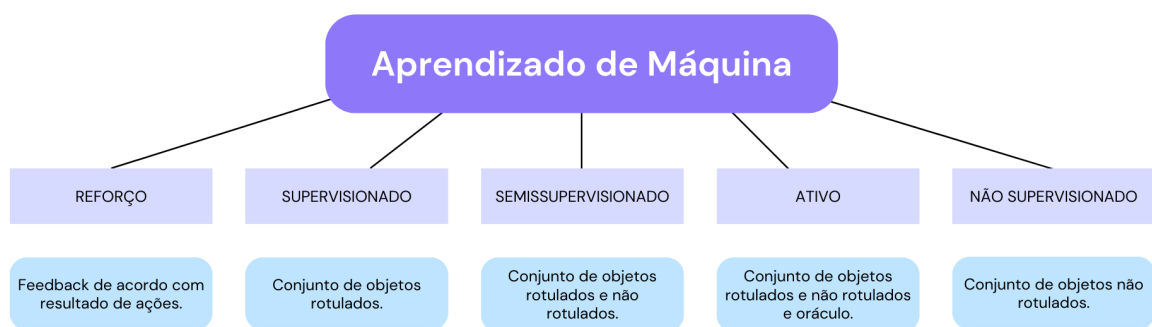
2 FUNDAMENTAÇÃO TEÓRICA

2.1 APRENDIZAGEM DE MÁQUINA

Segundo (Devipriya; Kalpana, 2019), o termo em língua inglesa *Machine Learning* foi cunhado por Arthur Samuel em 1950, tendo como tradução livre para o português a terminologia Aprendizagem de Máquina (AM). A AM envolve o estudo de como programas computacionais podem melhorar seu desempenho com o tempo por meio de experiências. Para tanto, esse campo engloba conceitos de diversas áreas, tais como ciência da computação, estatística, ciência de dados e inteligência artificial (Jordan; Mitchell, 2015).

Segundo (Sun *et al.*, 2024), a AM pode ser dividida em três categorias principais: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. A primeira aprende a partir de dados rotulados e é normalmente utilizada para tarefas de classificação e regressão; a segunda utiliza dados não rotulados, sendo aplicada em tarefas de agrupamento e redução de dimensionalidade; por fim, a terceira atua na maximização de recompensas, sendo frequentemente utilizada em tarefas de otimização e tomada de decisão. Em complemento a essa divisão, (Faceli *et al.*, 2021) acrescentam mais duas categorias: a aprendizagem semissupervisionada e a aprendizagem ativa. A abordagem semissupervisionada é utilizada quando os dados não possuem rótulos, mas conhecem-se algumas de suas restrições; já a aprendizagem ativa seleciona iterativamente os dados a serem rotulados e seus respectivos rótulos, sendo possível consultar um especialista (oráculo). A Figura 4 ilustra a estrutura dessa hierarquia estendida.

Figura 4 – Hierarquia de aprendizado estendida.



Fonte: Adaptado de (Faceli *et al.*, 2021).

A seguir, serão apresentadas, conforme proposto por (Sun *et al.*, 2024), a aprendizagem por reforço e a aprendizagem não supervisionada. Adicionalmente, a aprendizagem supervisionada será detalhada com maior profundidade, juntamente com o modelo selecionado, por constituir a abordagem utilizada nesta pesquisa.

2.1.1 APRENDIZAGEM POR REFORÇO

A Aprendizagem por Reforço (AR) é centrada em um objetivo definido. Nela, o modelo interage com um ambiente desconhecido e descobre progressivamente a sua dinâmica por meio de um processo de tentativa e erro, no qual o resultado de cada ação é interpretado como uma recompensa ou uma punição do meio. Essa sistemática é semelhante à de uma criança em seu processo inicial de aprendizado, no qual realiza ações e analisa se os resultados foram positivos ou negativos (Naeem; Rizvi; Coronato, 2020).

A AR é comumente empregada na resolução de problemas de controle em que a tomada de decisão representa um cenário desafiador (Amin, 2024).

2.1.2 APRENDIZAGEM NÃO SUPERVISIONADA

A Aprendizagem Não Supervisionada (ANS) visa identificar padrões específicos que representem a estrutura estatística dos dados de entrada, não possuindo, portanto, um resultado-alvo predefinido (Nerbonne, 2000). Dessa forma, os dados utilizados não são rotulados; são apresentados dados reais ao modelo para que este identifique os padrões ocultos neles contidos. Um exemplo prático consiste na classificação do perfil de personalidade de um indivíduo com base em dados de seu comportamento em redes sociais (Portugal; Alencar; Cowan, 2018).

2.1.3 APRENDIZAGEM SUPERVISIONADA

O termo “supervisionada” é utilizado porque se simula a presença de um supervisor externo quando um valor é predito com base em novos dados de entrada (Faceli *et al.*, 2021). Ao contrário da ANS, na Aprendizagem Supervisionada (AS) os dados utilizados são previamente rotulados, o que significa que tanto os dados de entrada quanto as suas saídas correspondentes são conhecidos. O modelo desenvolvido busca capturar e mapear o relacionamento identificado entre essas variáveis de entrada e saída (Ajourloo *et al.*, 2024). Essa configuração se consolida como a forma mais comum de AM (Lecun; Bengio; Hinton, 2015). A AS lida essencialmente com problemas de classificação e de regressão: no primeiro caso, prediz-se o valor de uma variável categórica; no segundo, estima-se um valor numérico, contínuo ou discreto. A maioria dos algoritmos de AS está disponível para ambas as finalidades (Jovel; Greiner, 2021).

A seguir, serão apresentados alguns dos principais algoritmos de AS.

2.1.3.1 Regressão Polinomial

A Regressão Polinomial (RP) consiste em uma variação da Regressão Linear Múltipla (Dikbaş, 2024). Devido à sua capacidade de modelar problemas complexos que envolvem múltiplas variáveis independentes em uma relação não linear, essa técnica é amplamente aplicada em problemas de otimização de processos produtivos, gerenciamento de sistemas de energia e análise de dados de sobrevivência (Park; Kwon, R.; Kwon, G., 2024). Segundo (Belany

et al., 2024), a RP é uma forma de regressão na qual tanto a variável independente quanto a dependente são modeladas de acordo com um polinômio de grau n , podendo ser representada pela Equação 1:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \dots + \beta_p \cdot X^p + \epsilon \quad (1)$$

Onde:

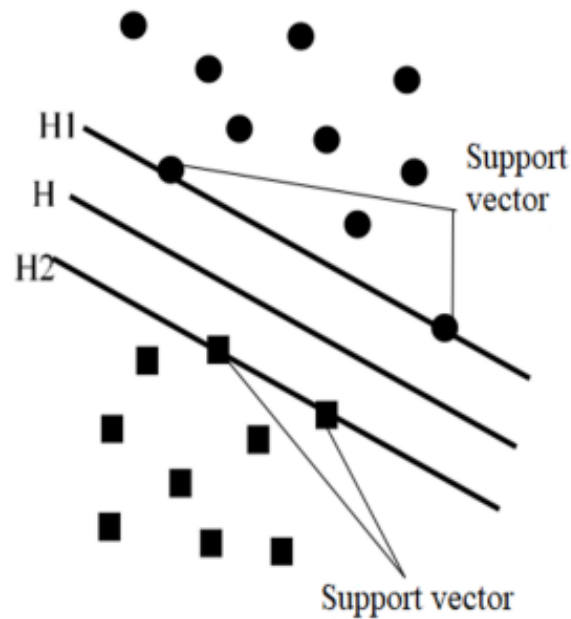
- Y representa a variável predita;
- $X \dots X^p$ representam os múltiplos das variáveis independentes;
- $\beta_0 \dots \beta_p$ correspondem aos coeficientes de regressão;
- ϵ representa o termo de erro.

2.1.3.2 Máquina de Vetores de Suporte

A Máquinas de Vetores de Suporte (MVS) engloba algoritmos de AM utilizados tanto em problemas de regressão quanto em problemas de classificação (Guido *et al.*, 2024). O modelo de MVS fundamenta-se em dois princípios essenciais: o hiperplano de classificação ótimo e a aplicação de uma função *Kernel* (Yusof *et al.*, 2023).

A Figura 5 ilustra como a MVS realiza a diferenciação entre duas classes distintas, representadas graficamente por círculos e quadrados. A linha H define o hiperplano de classificação, enquanto H_1 e H_2 consistem em linhas paralelas à linha de classificação que cruzam os pontos mais próximos de cada classe. O espaço compreendido entre H_1 e H_2 é denominado margem de classificação (Yusof *et al.*, 2023).

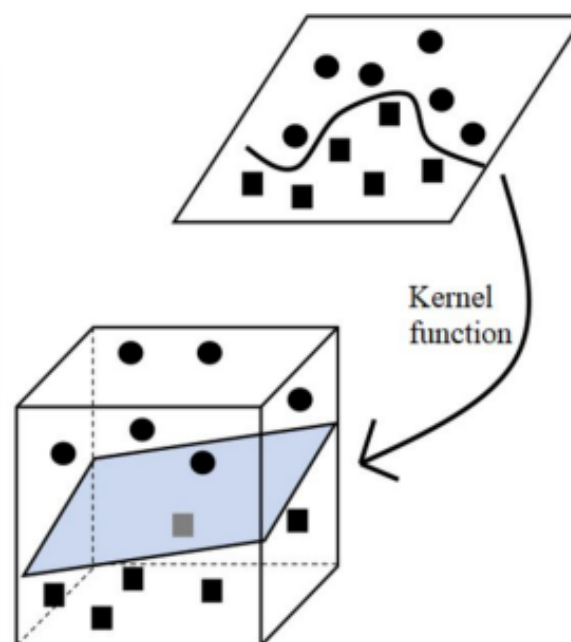
Figura 5 – Diferenciação entre duas classes por meio de margens de suporte.



Fonte: (Yusof *et al.*, 2023).

Para realizar a classificação de dados que não são separáveis por uma reta, a MVS utiliza a função *Kernel* para mapear e projetar os dados em um espaço de maior dimensão, conforme exibido na Figura 6, tornando viável a sua separação linear (Yusof *et al.*, 2023). As funções *Kernel* aplicadas podem ser do tipo linear, polinomial, de base radial (RBF) ou sigmoide (Kaul; Raina, 2022).

Figura 6 – Aplicação da Função Kernel para separação linear em maior dimensão.



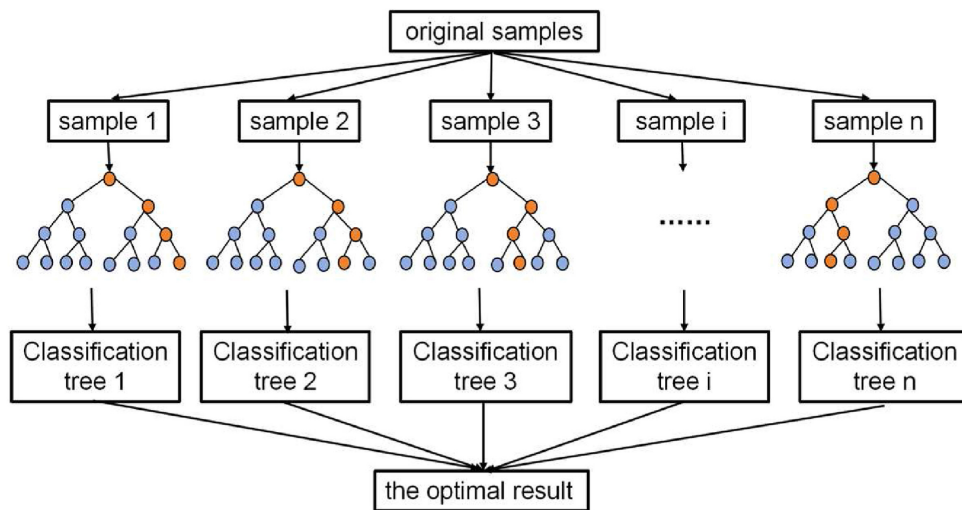
Fonte: (Yusof *et al.*, 2023).

2.1.3.3 Floresta Aleatória

Em (Borup *et al.*, 2023), essa técnica é descrita como uma abordagem baseada no princípio de “dividir e conquistar”, uma vez que inicializa frações amostrais dos dados, implementa uma árvore de decisão para cada fração e, ao final, agrega todos os resultados obtidos. Desse modo, o conceito-chave do algoritmo de FA consiste em construir um conjunto de árvores de decisão (floresta) e selecionar a previsão estruturada pela maioria dos votos em tarefas de classificação, ou calcular a média das previsões em tarefas de regressão (Choudhury; Mondal; Sarkar, 2024).

A FA apresenta alta adaptabilidade, sofre pouca interferência de ruídos e dispensa a necessidade de dados rigorosamente padronizados (Wang, F. *et al.*, 2024). A Figura 7 apresenta o diagrama esquemático desse algoritmo.

Figura 7 – Diagrama esquemático do algoritmo de Floresta Aleatória.



Fonte: (Wang, F. *et al.*, 2024).

2.2 MÉTRICAS DE AVALIAÇÃO DE MODELOS DE AM

Com o crescente uso de algoritmos de AM em diversos setores, a disponibilidade de métodos para avaliar com precisão tais modelos torna-se cada vez mais importante, principalmente em aplicações reais (Naidu; Zuva; Sibanda, 2023). Serão apresentadas as principais métricas utilizadas em AM para a avaliação de tarefas de classificação e de regressão.

2.2.1 Métricas para Tarefas de Regressão

As métricas de desempenho são cruciais na análise de modelos preditivos em tarefas de regressão. Uma métrica de desempenho constitui uma estrutura lógica e matemática desenvolvida para verificar o quão próximo o valor predito está do valor real observado (Plevris *et al.*, 2022).

O erro de viés é expresso pela diferença entre o valor predito e o valor real, podendo assumir valores positivos, negativos ou nulos. A Equação 2 define matematicamente o erro de viés:

$$e_i = p_i - r_i \quad (2)$$

Onde:

- i representa o índice do dado observado, variando de 1 a n ;
- e_i representa o erro de viés na posição i ;
- p_i representa o valor predito na posição i ;
- r_i representa o valor real na posição i .

2.2.1.1 Erro Médio Absoluto

O Erro Médio Absoluto (Erro Absoluto Médio (MAE)), cujo acrônimo deriva do termo em língua inglesa *Mean Absolute Error*, é determinado pela média dos valores absolutos dos erros de viés, ou seja, a diferença absoluta entre o valor predito e o valor real (Miller *et al.*, 2024):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |p_i - r_i| \quad (3)$$

O MAE apresenta a vantagem de ter seu resultado expresso na mesma unidade de medida das variáveis real e predita, o que simplifica a sua interpretação prática (Miller *et al.*, 2024). Sob essa ótica, quanto menor for o valor do MAE, melhor será o desempenho do modelo (Handelman *et al.*, 2019).

2.2.1.2 Erro Quadrático Médio

O Erro Quadrático Médio (Erro Quadrático Médio (MSE)), que tem sua origem no termo em inglês *Mean Squared Error*, é calculado a partir da média dos quadrados dos erros entre os valores reais e preditos, consolidando-se como uma métrica bastante popular para a avaliação de modelos de regressão (Plevris *et al.*, 2022). A Equação 4 representa o cálculo do MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2 \quad (4)$$

Uma desvantagem significativa do MSE é a sua baixa robustez em relação a valores discrepantes (*outliers*). Se uma única amostra apresentar um erro consideravelmente maior que as demais, o quadrado desse valor penalizará o índice de forma desproporcional, distorcendo o valor final do MSE e tornando a métrica altamente sensível a essas anomalias (Plevris *et al.*, 2022). Assim como ocorre com o MAE, quanto menor for o valor do MSE, melhor será a performance do modelo (Rainio; Teuho; Klén, 2024).

2.2.1.3 Raiz do Erro Quadrático Médio

A Raiz do Erro Quadrático Médio (Raiz do Erro Quadrático Médio (RMSE)), cujo acrônimo provém do termo em inglês *Root Mean Squared Error*, é outra métrica amplamente utilizada para mensurar a diferença média entre os valores preditos e reais (Miller *et al.*, 2024):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_i - r_i)^2}{n}} \quad (5)$$

A partir da Equação 5, observa-se que a RMSE corresponde à raiz quadrada do MSE. Diferentemente deste, a RMSE apresenta o seu resultado na mesma unidade da variável real (Plevris *et al.*, 2022), facilitando a interpretação dos resultados. Alinhada às métricas anteriores, quanto menor o valor da RMSE, superior será a performance do modelo.

2.2.1.4 Coeficiente de Determinação R^2

O Coeficiente de Determinação (R^2), fornece uma medida da proporção da variância da variável alvo que é explicada e capturada pelo modelo preditivo (Miller *et al.*, 2024), sendo expresso pela Equação 6:

$$R^2 = 1 - \frac{\sum_{i=1}^n (r_i - p_i)^2}{\sum_{i=1}^n (r_i - \bar{r})^2} \quad (6)$$

Onde:

- r_i representa o valor real na posição i ;
- p_i representa o valor predito na posição i ;
- $\bar{r}$ representa a média aritmética dos valores reais.

Ao contrário do MAE e da RMSE, o R^2 é uma métrica adimensional, ou seja, não compartilha a unidade de medida dos dados amostrais. Seus resultados variam de zero a um, em que 0 indica 0% de explicação da variância da variável alvo por parte do modelo e 1 representa 100% de explicação. Essa característica simplifica o seu uso na comparação direta entre diferentes modelos (Miller *et al.*, 2024).

Por esse motivo, inversamente às demais métricas apresentadas, quanto maior for o valor do R^2 , melhor será o modelo (Handelman *et al.*, 2019). Contudo, um valor de R^2 excessivamente elevado exige cautela por parte do pesquisador, tornando-se necessário avaliar a ocorrência de problemas de sobreajuste (*overfitting*) (Miller *et al.*, 2024).

2.2.1.5 Erro Percentual Absoluto Médio

O Erro Percentual Absoluto Médio (MAPE), cuja sigla provém do termo em inglês *Mean Absolute Percentage Error*, é uma métrica comumente utilizada para comparar o desempenho

de previsões entre escalas ou conjuntos de dados distintos, visto que avalia o erro em termos relativos (Lewis, 1982).

O MAPE é calculado em conformidade com a Equação 7:

$$\text{MAPE} = \frac{1}{n} \sum_{t=0}^{n-1} \frac{|e_t|}{d_t} \times 100 \quad (7)$$

De modo geral, a capacidade de previsão do modelo pode ser classificada qualitativamente a partir dos resultados obtidos para o MAPE, conforme os critérios estabelecidos na Tabela 2.

Tabela 2 – Critérios qualitativos do MAPE.

MAPE (%)	Capacidade de Previsão
< 10	Previsão altamente precisa (HAP)
10 – 20	Boa previsão (GPR)
20 – 50	Previsão razoável (RP)
> 50	Previsão imprecisa (IPR)

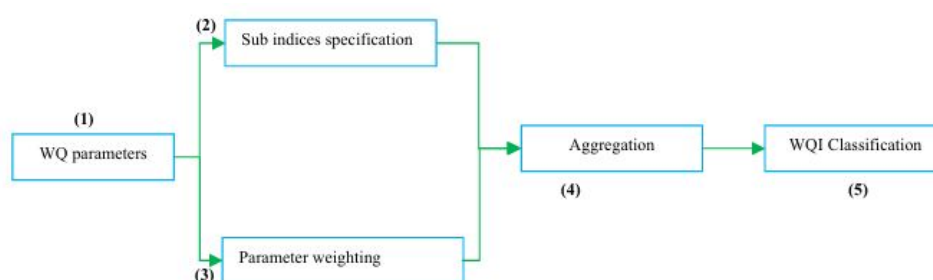
Fonte: (Lewis, 1982).

2.3 ÍNDICE DE QUALIDADE DA ÁGUA.

A água é um recurso vital para o planeta, estando na base de toda a vida existente. A poluição das águas até certo nível é tolerável, porem acima de um ponto podem causar desastres ambientais, como também impactar a saúde humana (Nair; Vijaya, 2022). Desta forma agencias como *National Sanitation Foundation* (NFS), Agência Nacional de Águas e Saneamento Básico (ANA), Companhia Ambiental do Estado de São Paulo (CETESB), entre outras, trabalharam para desenvolver um índice e avaliar a qualidade da água.

O IQA traduz um grande número de parâmetros individuais de qualidade da água em um valor único, tornando sua interpretação clara e simples (Azrour *et al.*, 2022). Os modelo de IQA consistem, principalmente, de cinco estágios, parâmetros de qualidade da água, especificação de subíndices, ponderação dos parâmetros, agregação e classificação do IQA (Manna; Biswas, 2023). O estrutura deste modelo é apresentada na Figura 8 a seguir.

Figura 8 – Cinco estágios de um modelo de IQA.



Fonte: (Manna; Biswas, 2023).

Onde, segundo (Manna; Biswas, 2023), temos:

- **(1) *WQ parameters*** - Seleção dos parâmetros de qualidade da água que serão utilizados no modelo para análise;
- **(2) *Sub indices specification*** - A geração de subíndices converte o valor do parâmetro em um valor numérico sem unidade;
- **(3) *Parameter weighing*** - A definição do peso de cada parâmetro é realizada com base na sua relevância na avaliação da qualidade da água;
- **(4) *Aggregation*** - É o cálculo do valor do IQA, os valores obtidos nos passos 2 e 3 são combinados gerando um único resultado;
- **(5) *WQI classification*** - O valor do IQA, gerado no passo 4, é classificado seguindo uma escala de classificação, para assim se definir o nível de qualidade da água.

Várias metodologias, para o cálculo do IQA são proposta pela literatura, e uma das mais utilizadas é a *National Sanitation Foundation Water Quality Index* (NSFWQI) (Azrour *et al.*, 2022).

EM 1975, a CETESB desenvolveu o seu IQA, adaptado do NSFWQI (CETESB, 2024b). Com o objetivo de avaliar a qualidade de água, para abastecimento público após tratamento, hoje o IQA é o principal índice adotado no Brasil (ANA, 2024). O IQA é calculado a partir de nove parâmetros de seus respectivos pesos, que foram definidos a partir de sua importância global na qualidade da água (ANA, 2024).

2.3.1 Parâmetros Físico-Químicos, Biológicos e Microbiológicos.

Estes parâmetros representam medições de características físico-químicos, biológicos e microbiológicos da água. A seguir teremos uma breve descrição destes parâmetros segundo (ANA, 2024).

- **Oxigênio Dissolvido (OD)** - É um indicador vital para a vida aquática, em águas limpas apresentam valores superiores a 5mg/L. Em águas poluídas apresentam baixo valores, já que o processo de decomposição orgânica consome o oxigênio dissolvido na água. Além da fotossíntese, as condições físicas dos corpos d'água também interferem na introdução do oxigênio na água;
- **Demanda Bioquímica de Oxigênio (DBO)** - A Demanda Bioquímica de Oxigênio mede a quantidade de oxigênio necessária para oxidar a matéria orgânica presente na água. A $DBO_{5,20}$ é a quantidade de oxigênio consumido durante 5 dias em uma temperatura de 20°C. Um alto valor de DBO significa uma grande carga orgânica na água, normalmente oriunda de despejo de esgoto doméstico no corpo d'água;

- **Potencial Hidrogeniônico (pH)** - O pH afeta o metabolismo dos seres aquáticos, deve ter seus valores entre 6 e 9. A alteração de seus valores pode interferir no efeito de substâncias tóxicas que estejam presentes na água;
- **Nitrogênio Total (N-total)** - Nos corpos d'água o nitrogênio pode ocorrer nas formas de nitrogênio orgânico, amoniacal, nitrito e nitrato. Os nitratos são tóxicos aos seres humanos, e em altas concentrações podem causar doenças chegando a ser letal para crianças.

Por serem nutrientes nos processos biológicos, seu acúmulo em corpos d'água propiciam o crescimento desequilibrado de algas, processo conhecido como eutrofização, que é prejudicial a vida aquática, ao abastecimento e a recreação. As fontes de nitrogênio para os corpos d'água são variadas, estando entre elas o lançamento de esgoto sanitário, efluente industrial, drenagem de áreas agrícolas e deposição atmosférica pelas águas das chuvas.

- **Fósforo Total (P-total)** - O fósforo, assim como o nitrogênio, um importante nutriente para os processos biológicos, e em excesso pode causar a eutrofização das águas. Entre as fontes de fósforo destacam-se os esgotos domésticos, a drenagem pluvial de áreas agrícolas e os efluentes industriais, em especial das indústrias de fertilizantes, alimentícias, laticínios, frigoríficos e abatedouros.
- **Temperatura** - A temperatura influencia vários parâmetros físico-químicos da água, organismos aquáticos são afetados por temperaturas fora de seus limites de tolerância causando impactos sobre seu crescimento e reprodução. Os corpos d'água apresentam variações de temperatura ao longo do dia e do ano, entretanto o lançamento de efluentes com altas temperaturas pode causar impacto significativo.
- **Coliformes Termotolerantes** - As bactérias coliformes termotolerantes ocorrem no trato intestinal de animais de sangue quente. Mesmo não causando doenças, são indicadores de contaminação por esgoto doméstico, e em grande concentração indicam a possibilidade da existência de micro-organismos patogênicos que são responsáveis pela transmissão de doenças.
- **Turbidez** - A turbidez indica o grau de atenuação que um feixe de luz sofre ao atravessar a água. Isso ocorre pela absorção e espalhamento da luz causada pelos sólidos em suspensão, uma alta turbidez faz com que uma quantidade maior de produtos químicos sejam utilizados para o tratamento de águas, aumentando assim os custos. A principal fonte de turbidez é a erosão dos solos, que se intensificam no período chuvoso.
- **Resíduo Total** - O resíduo total é a matéria que permanece após a evaporação. Quando os resíduos sólidos se depositam nos leitos dos corpos d'água podem causar seu assoreamento, aumentando o risco de enchente, impactando na navegação e na vida aquática.

2.3.2 Metodologia para Cálculo do IQA.

Em 1975 a CETESB iniciou o desenvolvimento do IQA, este ocorreu por meio de uma pesquisa junto à especialistas em qualidade de água, que foram responsáveis em definir quais variáveis seriam utilizadas e qual o peso relativo que cada uma teria no cálculo. Das trinta e cinco variáveis, propostas inicialmente, nove foram selecionadas, e cada especialista propôs uma curva de variação da condição de cada parâmetro, bem como seu peso. Tais curvas e seus pesos foram sintetizados em uma curva média para cada parâmetro (CETESB, 2024c), resultando na Figura 9.

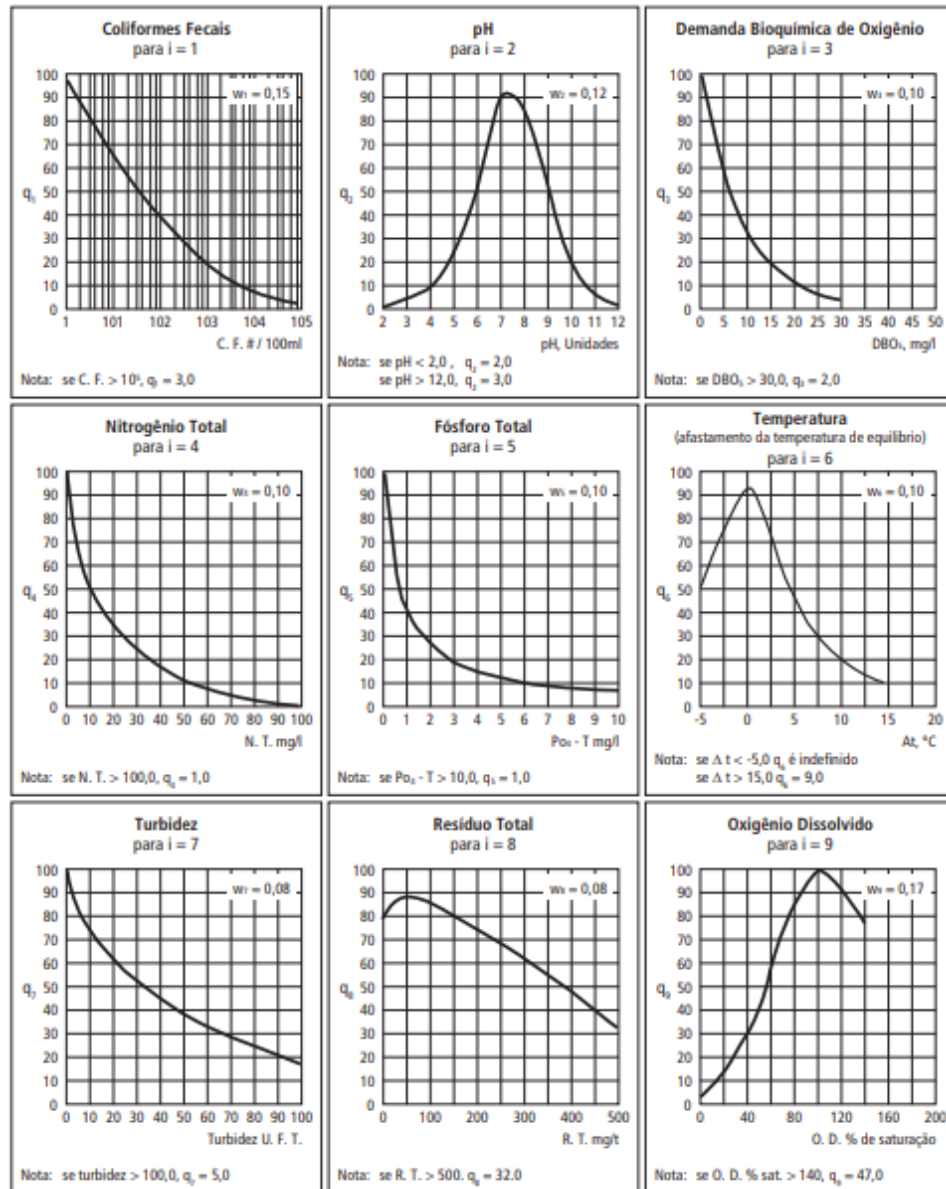
O IQA é calculado a partir do produtório ponderado dos parâmetros que compõem o índice (CETESB, 2024c). Tendo como fórmula a Equação 8.

$$IQA = \prod_{i=1}^n q_i^{w_i} \quad (8)$$

Onde:

- q_i - representa a qualidade individual de cada parâmetro (varia de 0 a 100),
- w_i - é o peso atribuído a cada parâmetro (varia de 0 a 1),
- n - é o número total de parâmetros.

Figura 9 – Curvas Médias de Variação de Qualidade das Águas.



Fonte: (CETESB, 2024c)

Os pesos de cada parâmetro são definidos de acordo com a relevância do parâmetro na qualidade da água (Chidiac *et al.*, 2023). Quanto maior o peso, maior será o impacto deste indicador no valor do índice. A maioria dos modelos de IQA adota valores desiguais para a ponderação, de forma que o somatório de todos os pesos é igual a um (Uddin; Nash; Olbert, 2021), conforme a Equação 9. A Tabela 3 apresenta os pesos utilizados no cálculo do IQA adotado pela CETESB.

$$\sum_{i=1}^n w_i = 1 \quad (9)$$

Tabela 3 – Parâmetros de Qualidade da Água do IQA e respectivos pesos.

Parâmetro de Qualidade da Água	Peso (w_i)
Oxigênio dissolvido	0,17
Coliformes termotolerantes	0,15
Potencial hidrogeniônico (pH)	0,12
Demanda Bioquímica de Oxigênio (DBO _{5,20})	0,10
Temperatura da água	0,10
Nitrogênio total	0,10
Fósforo total	0,10
Turbidez	0,08
Resíduo total	0,08
Total	1,00

Fonte: Adaptado de (ANA, 2024).

2.3.3 Classificação e Interpretação dos Resultados do IQA.

Após o cálculo do IQA, é possível classificar a qualidade da água bruta, comparando o valor do IQA com uma tabela de classificação que varia de 0 à 100 (CETESB, 2024c). (Manna; Biswas, 2023) alertam para o fato de diferentes modelos possuírem diferentes faixas de classificação, dificultando assim uma classificação realista. No Brasil, os estados adotam faixas diferentes para classificar a qualidade da água(ANA, 2024). A Tabela 4 apresenta os valores utilizados por cada estado do país.

Tabela 4 – Classificação de Qualidade da Água.

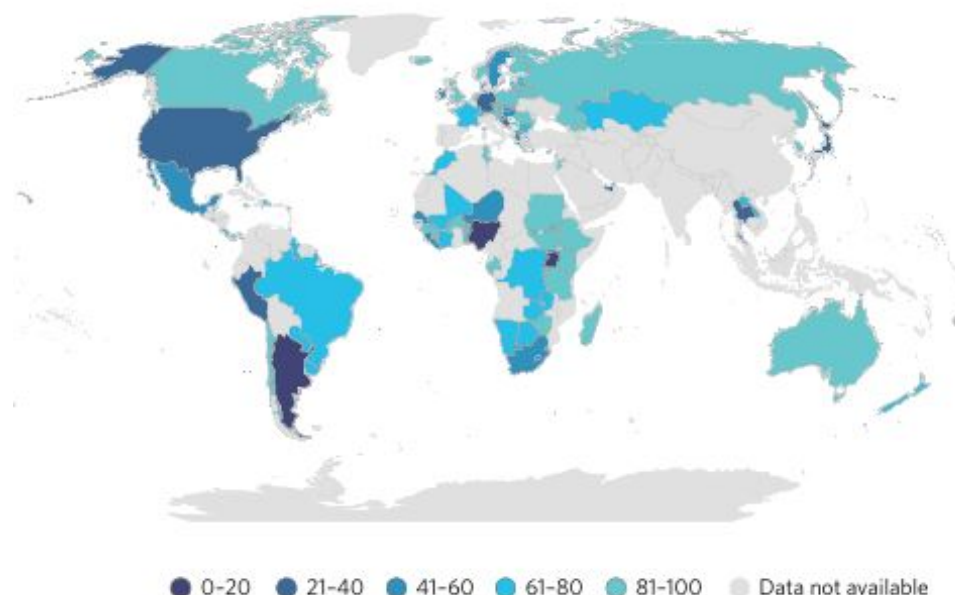
Estados: AL, MG, MT, PR, RJ, RN, RS	Estados: BA, CE, ES, GO, MS, PB, PE, SP	Classificação da Qualidade da Água
$90 < \text{IQA} \leq 100$	$79 < \text{IQA} \leq 100$	Ótima
$70 < \text{IQA} \leq 90$	$51 < \text{IQA} \leq 79$	Boa
$50 < \text{IQA} \leq 70$	$36 < \text{IQA} \leq 51$	Razoável
$25 < \text{IQA} \leq 50$	$19 < \text{IQA} \leq 36$	Ruim
$\text{IQA} \leq 25$	$\text{IQA} \leq 19$	Péssima

Fonte: Adaptado de (ANA, 2024)

2.3.4 Importância do IQA para o Desenvolvimento Sustentável.

Metade da população mundial deve sofrer com a escassez de água, por pelo menos um mês no ano, até 2030. E este fato tende a pior com a intensificação dos efeitos do aquecimento global (ONU, 2024d). A qualidade da água está diretamente relacionada com o desenvolvimento socioeconômico e os objetivos de desenvolvimento sustentável 3 e 6 (Alexakis, 2021). Onde o

Figura 10 – Proporção de corpos d'água com boa qualidade da água ambiente, 2017–2020 (%).



Fonte: (ONU, 2024e).

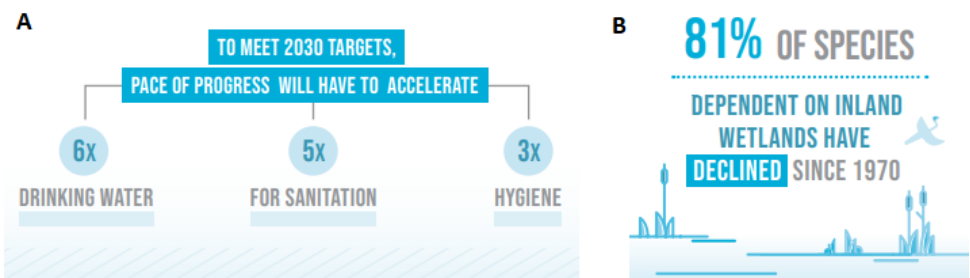
objetivo 3 visa garantir vidas saudáveis e promover o bem-estar para todos e o objetivo 6 visa garantir a água limpa e disponível para todos (ONU, 2024d).

O relatório sobre os objetivos de desenvolvimento sustentável de 2023, relata que faltam dados estatísticos em muitos países, que apresentaram relatórios fracos. Isso representa um risco para mais de 3 bilhões de pessoas que estão expostas a água doce com qualidade desconhecida. Em contrapartida, países com um sistema de monitoramento robusto apresentam uma tendência positiva em seus resultados. A agricultura e o esgoto não tratado são as grandes ameaças a qualidade da água, sendo necessário melhorar as práticas agrícolas e incrementar o tratamento de esgoto em áreas altamente populosas (ONU, 2024e).

Assim, os modelos para mensurar a qualidade da água são essenciais para o monitoramento da poluição dos corpos d'água (Aldhyani *et al.*, 2020). A Figura 10 mostra a proporção percentual de corpos d'água com boa qualidade.

De forma que, ter um vigoroso sistema de medição do IQA para nortear ações e políticas públicas se faz fundamental, tanto do ponto de vista ambiental como no aspecto de saúde e bem estar público. A Figura 11 - B, indica uma diminuição de 81% do número de espécies de regiões úmidas ou alagadas desde 1970. Enquanto na Figura 11 - A, que os esforços atualmente empregados para se atingir as metas do objetivo 6 dos ODS precisam ser acelerados em até 6 vezes para que a meta seja atingida.

Figura 11 – Infográfico ONU - ODS.

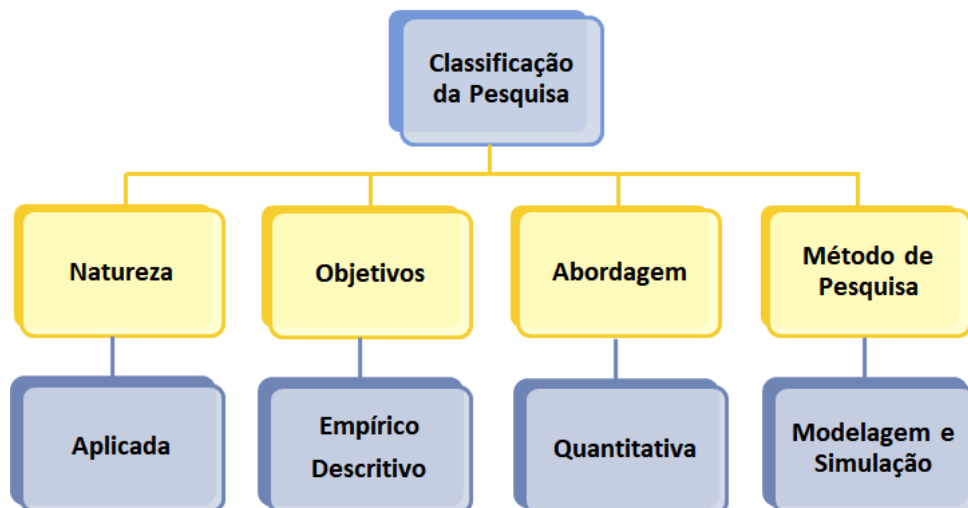


Fonte: (ONU, 2024b).

3 MÉTODO

Segundo (Prodanov; Freitas, 2013), este trabalho pode ser classificado como de natureza aplicada, com objetivo empírico-descritivo. Conforme aponta (Cauchick-Miguel *et al.*, 2012), o modelo a ser desenvolvido visa estabelecer as relações causais entre as variáveis, levando a uma melhor compreensão desse processo real. A abordagem adotada é quantitativa, e os métodos de pesquisa utilizados serão a Modelagem e a Simulação, visto que o modelo será desenvolvido em linguagem computacional, empregando técnicas experimentais (simulação) para encontrar o valor numérico de uma propriedade do sistema em questão (Cauchick-Miguel *et al.*, 2012). A Figura 12 ilustra a classificação desta pesquisa.

Figura 12 – Classificação da pesquisa.



Fonte: Autor.

3.1 MATERIAIS

Software:

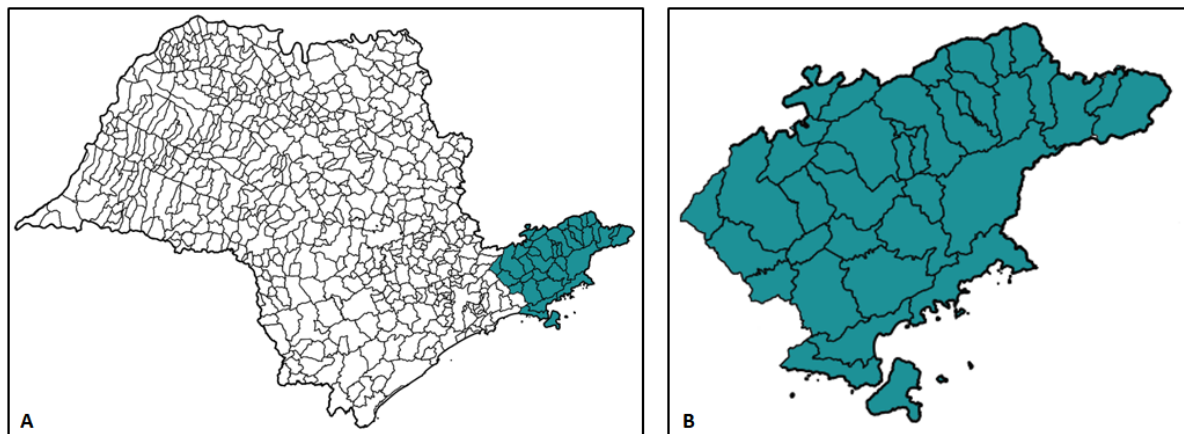
Para o desenvolvimento do trabalho, será utilizado o *software* R. Segundo (R Foundation, 2024), o R é um *software* de código aberto, caracterizado como uma “linguagem e ambiente para a computação estatística e gráfica”, que provê uma grande variedade de técnicas metodológicas. Todas as análises estatísticas, a implementação dos modelos de AM e as eventuais interfaces que se façam necessárias serão desenvolvidas nesse ambiente.

Área Geográfica:

A RMVALE-LN, localizada no estado de São Paulo, conta com 39 municípios, estende-se por uma área de 16.180,94 km² (AGEMVALE, 2021) e possui uma população de 2.505.723 habitantes, o que corresponde a 1,23% da população nacional (SEADE, 2022). A Figura 13

apresenta a RMVALE-LN, na qual o painel A ilustra a sua localização dentro do estado de São Paulo e o painel B destaca a divisão municipal da região.

Figura 13 – Localização da RMVALE-LN.



Fonte: Adaptado de (GeoSEADE, 2024).

Banco de Dados:

A RMVALE-LN engloba três Unidades de Gerenciamento de Recursos Hídricos (UGRHI) da CETESB, sendo elas: a unidade 01 - Mantiqueira, a 02 - Paraíba do Sul e a 03 - Litoral Norte. Os dados de 68 pontos de monitoramento distribuídos entre essas unidades foram obtidos por meio do portal Infoáguas e dos relatórios anuais de qualidade da água, ambos disponibilizados pela CETESB, abrangendo o período de 2013 a 2023.

Após a compilação de todos os dados, foram selecionadas 12 variáveis que indicam local, data, os nove parâmetros de qualidade da água utilizados no IQA da CETESB e o valor final do índice. A Tabela 5 apresenta um resumo do conjunto de dados utilizado.

Tabela 5 – Descrição das variáveis do conjunto de dados.

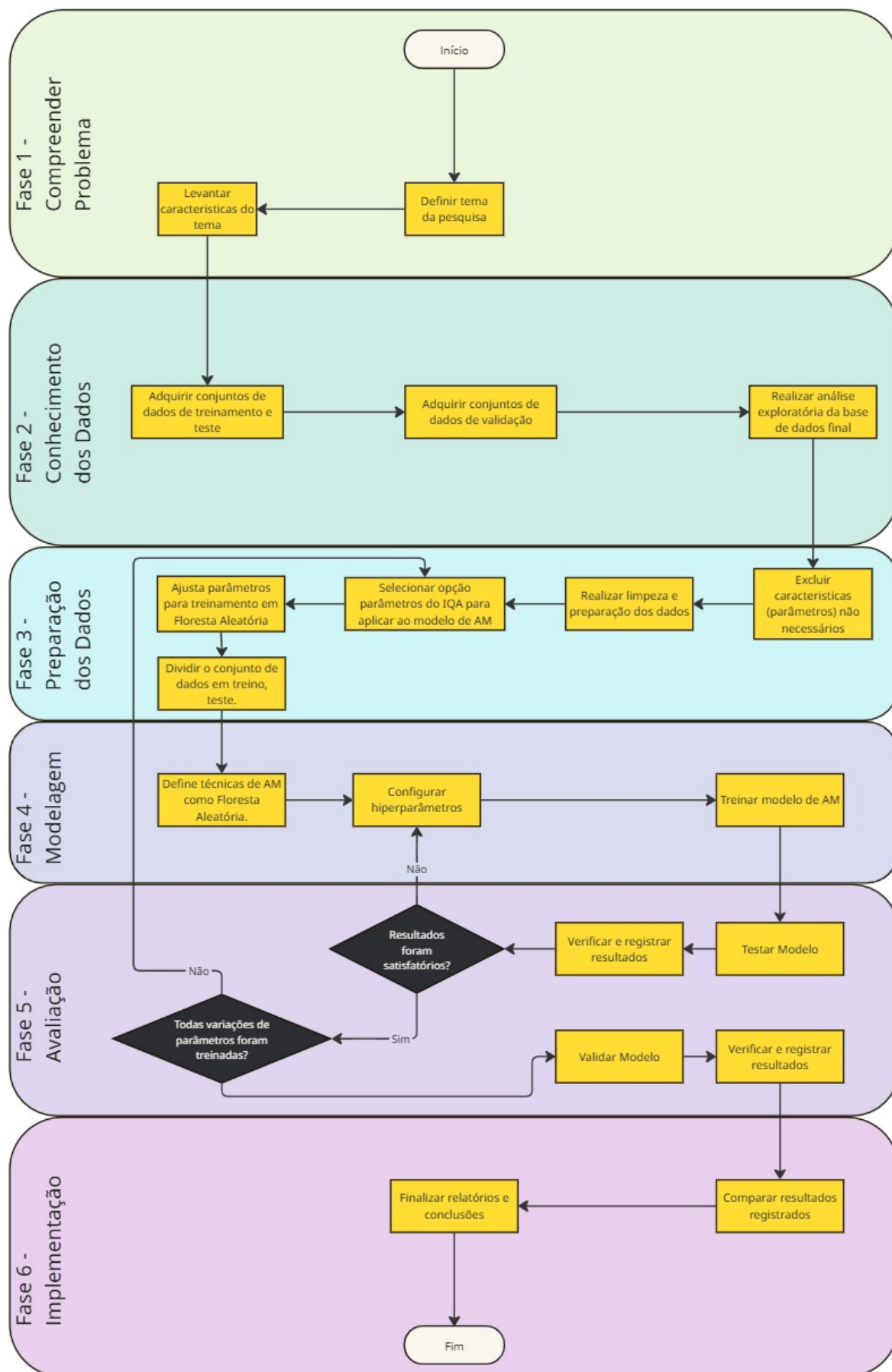
Variável	Descrição
Data	Data de realização da medição
O2_Dis	Valor do Oxigênio Dissolvido
Coli_TermTol	Valor de Coliformes Termotolerantes
pH	Valor do pH da água
DBO_(5, 20)	Demanda Bioquímica de Oxigênio (5 dias, 20 °C)
N2_Tot	Nitrogênio Total (Soma de compostos nitrogenados)
Temp_H2O	Temperatura da água
P_Tot	Valor de Fósforo Total
Turb	Valor da Turbidez
Sol_Tot	Valor de Sólidos Totais
IQA	Valor do IQA calculado (Variável Alvo)

Fonte: Autor.

3.2 MÉTODOS E PROCEDIMENTOS

Para a metodologia deste trabalho, será aplicada uma adaptação da estrutura do *Cross-Industry Standard Process for Data Mining* (CRISP-DM), visto que mesmo projetos de *data mining* em pequena escala se beneficiam de seu uso (Rivo *et al.*, 2012). Outra vantagem é que o CRISP-DM apresenta fases de desenvolvimento nas quais a sequência de execução não é rígida (Chapman *et al.*, 2000). Segundo (Garcia-Rios *et al.*, 2023), essas fases são flexíveis e facilmente customizáveis, o que facilita a adaptação às necessidades desta pesquisa. A Figura 14 apresenta o fluxograma de execução deste trabalho. A seguir, serão detalhadas as seis fases da metodologia CRISP-DM, bem como cada ação executada no fluxo de atividades.

Figura 14 – Fluxograma da metodologia do trabalho.

Fonte: Adaptado de (Chapman *et al.*, 2000).

Fase 1 - Compreender o Problema:

Nessa fase, compreende-se o problema em análise, realizando-se o levantamento de requisitos e definindo-se os objetivos do projeto (Chapman *et al.*, 2000). As ações do fluxograma “Definir tema da pesquisa” e “Levantar características do tema” compreendem os capítulos de Introdução e Fundamentação Teórica deste trabalho. Nelas, determinam-se as questões de pesquisa, os objetivos, as delimitações e os parâmetros que serão utilizados.

Fase 2 - Conhecimento dos Dados:

A Fase 2 compreende a ação de aquisição dos dados, seguida de uma análise inicial que visa à familiarização com o conjunto de dados, à detecção de problemas de qualidade e à obtenção de *insights* iniciais (Chapman *et al.*, 2000). O ciclo do fluxograma destinado à aquisição do conjunto de dados é encarregado de consolidar a base de dados, porém os arquivos adquiridos não são ideais para o uso e precisam ser manipulados. A primeira etapa desta manipulação é unir todos os arquivos em um só, após isto é necessário alterar a estrutura dos dados, já que estes possuem os parâmetros como linhas e tem apenas uma coluna de valor. Assim, a segunda etapa consiste em criar um arquivo onde os parâmetros são colunas e o arquivo é organizado por ordem da data de coleta e ponto de medição. A ação “Realizar análise exploratória da base de dados final” corresponde a esse estudo inicial focado em compreender melhor os dados, identificando previamente as ações necessárias para a próxima etapa.

Fase 3 - Preparação dos Dados:

Esta etapa abrange a construção do conjunto de dados final, envolvendo atividades de seleção, transformação e limpeza de dados, que podem ser executadas multiplamente e sem uma ordem definida (Chapman *et al.*, 2000). Esta fase consiste em padronizar alguns parâmetros, nela é calculada o $DBO_{5,20}$ em função do Carbono Orgânico Total (COT) e Coliforme Termotolerante em função do *Escherichia coli*, os nitratos são totalizados e parâmetros não necessários são excluídos. Nela também são excluídos os dados faltantes, 9.35% da base foi excluída por presença de dados faltantes nos parâmetros que compõe o IQA, assim ao final desta etapa a base de dados pronta para o cálculo do IQA. Por fim é calculado o IQA para os valores adquiridos, e assim finalizar a base de dados utilizado na pesquisa.

Ao final dessa etapa, selecionam-se os parâmetros do conjunto final que serão utilizados em cada teste. Esses dados são reestruturados na etapa “Ajustar parâmetro para treinamento em Floresta Aleatória”, a qual é detalhadamente descrita na Figura 22. Nesse ponto, os dados históricos são organizados cronologicamente, empregando-se uma abordagem de janela deslizante (*Sliding Window*) para a representação temporal das variáveis. Esse procedimento é executado para manipular os dados na geração dos modelos de predição direta do IQA, bem como nos nove modelos dos parâmetros individuais. Por fim, esse conjunto de dados é dividido para dar início ao treinamento e ao teste dos modelos.

Fase 4 - Modelagem:

Nesta etapa, a técnica de modelagem FA é selecionada e seus hiperparâmetros são ajustados para valores ótimos. Como algumas técnicas apresentam necessidades específicas sobre o formato dos dados, muitas vezes torna-se necessário retornar à Fase 3 (Chapman *et al.*, 2000). As ações desse bloco selecionam, configuram e treinam os modelos de FA, sendo executadas de forma iterativa de acordo com os resultados obtidos na fase subsequente de avaliação. Foram utilizadas técnicas de auto-tune para obter o ajuste dos hiperparâmetros, ressaltando que não foi detectado *overfitting* observando os resultados obtidos pelos modelo com os dados de validação.

Fase 5 - Avaliação:

Nessa fase, o modelo criado é avaliado antes de seguir para a etapa de implantação, sendo comum a necessidade de reavaliar passos anteriores (Chapman *et al.*, 2000). Na Fase 5, as ações se concentram em testar o modelo, aplicar as métricas de avaliação e registrar os desempenhos para posterior comparação. Essa atividade é executada de forma cíclica juntamente com as Fases 3 e 4, visando testar novos conjuntos de dados ou otimizar os resultados do modelo de FA.

Concluída a etapa de treinamento, os modelos foram submetidos ao conjunto de dados de validação, composto pelos registros dos anos de 2024 e 2025. A exclusão desse período da base de treino visou prevenir o vazamento temporal de informações e assegurar uma avaliação fidedigna da capacidade preditiva em horizontes futuros. Os desempenhos obtidos foram devidamente registrados para subsidiar as análises comparativas subsequentes.

Fase 6 - Implementação:

Mesmo após a validação do modelo, seus resultados precisam ser organizados e apresentados. Essa etapa pode ser tão simples quanto a geração de relatórios ou envolver atividades de integração bem mais complexas (Chapman *et al.*, 2000). As ações dessa fase envolvem a comparação dos resultados obtidos na fase de avaliação. A partir da interpretação dessas comparações, o relatório é finalizado e a conclusão do trabalho é estruturada.

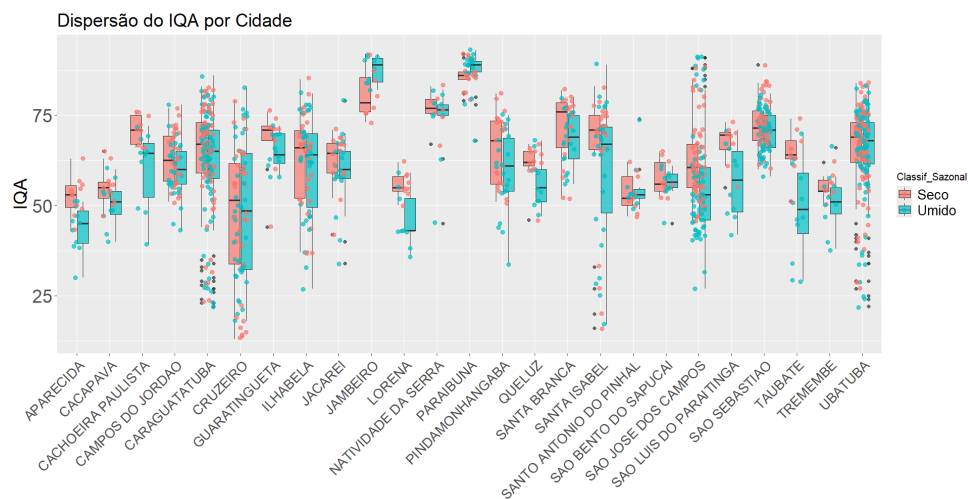
4 RESULTADOS OBTIDOS

Neste capítulo, serão apresentados os resultados obtidos com a pesquisa, iniciando pela análise descritiva dos dados, juntamente com o seu tratamento e manipulação. Nesse escopo, foram criadas colunas de classificação para auxiliar na melhor compreensão das informações. Na sequência, apresentam-se os ensaios utilizando o modelo de FA para realizar a predição dos valores de IQA.

4.1 ANÁLISE DESCRITIVA DOS DADOS

Iniciando a análise dos dados, foi adicionada uma coluna classificando o período sazonal das medições em “Seco”, para medições realizadas nos meses de abril a setembro (Outono-Inverno), e em “Úmido”, para medições realizadas nos meses de outubro a março (Primavera-Verão). Essa divisão justifica-se porque, segundo (Santos, 2020), percebem-se nitidamente dois períodos no comportamento do ciclo de chuvas no estado de São Paulo: uma estação chuvosa, de outubro a março, e outra estação seca, de abril a setembro. A Figura 15 apresenta o gráfico do tipo *boxplot*, sobreposto pela dispersão dos pontos de IQA de cada cidade, separados pela sazonalidade (seca e úmida), auxiliando na visualização geral da distribuição dos dados de IQA de cada localidade.

Figura 15 – Dados IQA das cidades por sazonalidade.



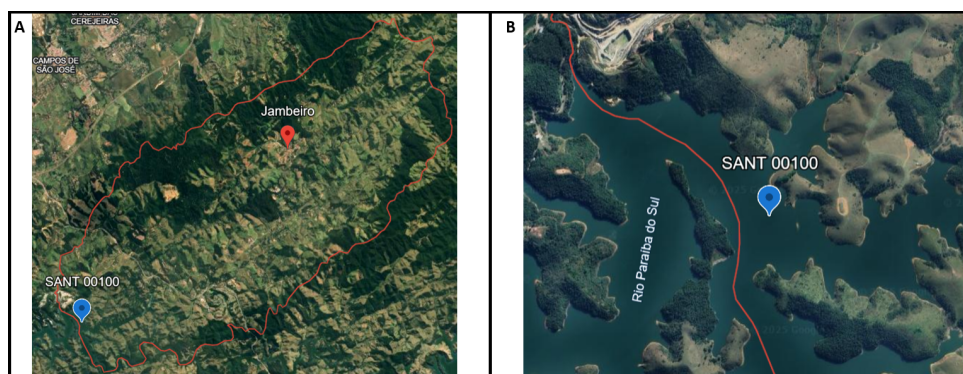
Fonte: Autor.

Para aprofundar o estudo sobre os dados, foram selecionadas duas cidades com base nos resultados do gráfico da Figura 15: Jambeiro, que apresenta um dos melhores valores de IQA, e Cruzeiro, que apresenta os valores mais baixos de IQA.

4.1.1 Análise dos dados de Jambeiro

Segundo (IBGE, 2023b), Jambeiro é uma cidade de 184,413 km², com uma população de 6.397 habitantes e IDH de 0,756. O único ponto de amostragem da cidade está localizado na represa de Santa Branca, “no meio do corpo central, na junção dos braços Capivari e Paraibuna” (CETESB, 2024a). A Figura 16 indica a localização desse ponto de monitoramento, apresentando em “A” a sua localização em referência aos limites do município e em “B” a sua localização exata na represa de Santa Branca.

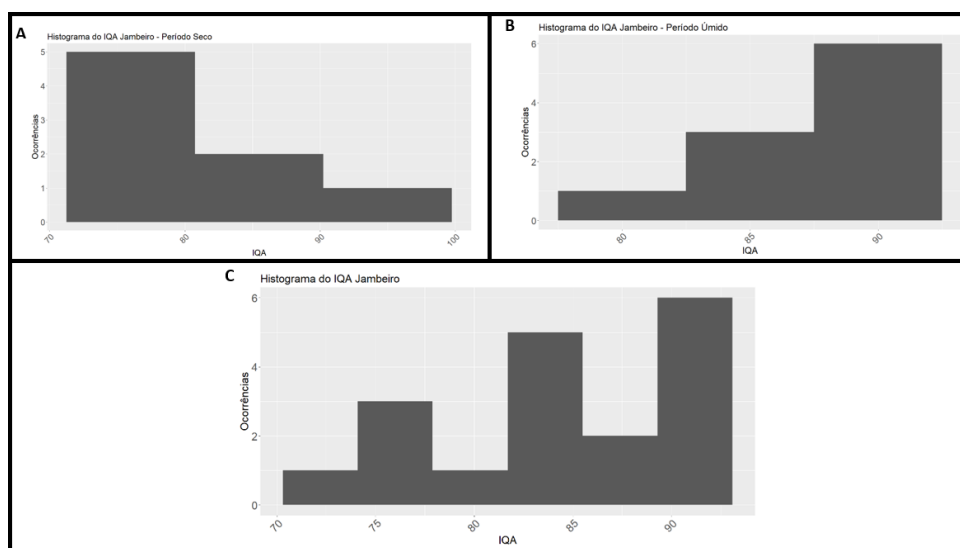
Figura 16 – Localização do ponto SANT00100.



Fonte: Autor.

Os histogramas apresentados na Figura 17 exibem em “A” os valores de IQA das estações secas, em “B” os valores das estações úmidas e em “C” a distribuição geral dos valores de IQA. É possível perceber uma maior concentração de valores mais altos de IQA durante o período úmido e, de forma complementar, uma concentração de valores mais baixos na estação seca.

Figura 17 – Histograma dos valores de IQA em Jambeiro.

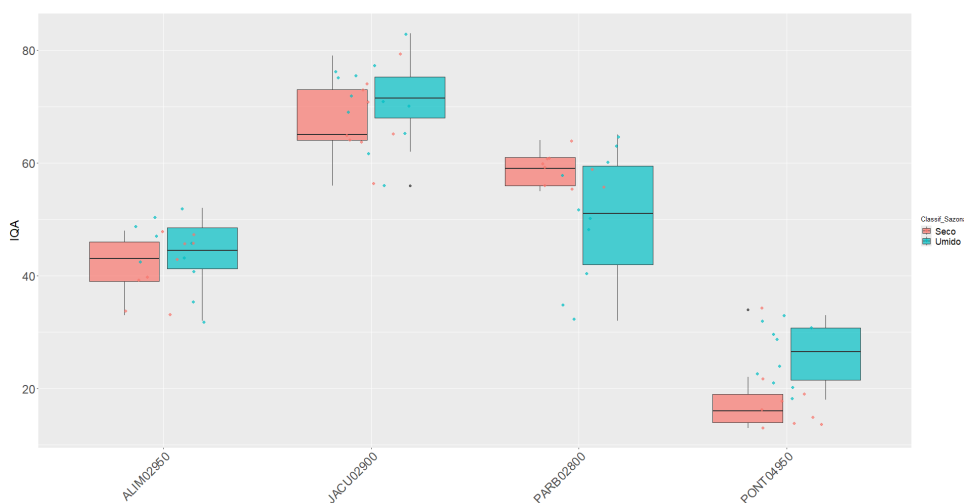


Fonte: Autor.

4.1.2 Análise dos dados de Cruzeiro

Segundo (IBGE, 2023a), Cruzeiro é uma cidade de 305,699 km², com uma população de 74.961 habitantes e IDH de 0,788. A cidade de Cruzeiro apresenta quatro pontos de medição com grande variação de IQA, conforme indicado na Figura 15. A Figura 18 apresenta o gráfico *boxplot* dos pontos de coleta de Cruzeiro nas estações seca e úmida. Nela, é possível observar que a variabilidade interna de cada ponto não é tão expressiva; ela se torna maior quando a comparação ocorre entre os diferentes pontos de medição.

Figura 18 – Dispersão do IQA em Cruzeiro.



Fonte: Autor.

Outro ponto a observar é que, exceto no ponto de medição do rio Paraíba do Sul (PARB 02800), os demais pontos seguem a variação já observada no ponto de monitoramento de Jambeiro, ou seja, exibem um IQA melhor na estação úmida do que na estação seca. Avaliado o posicionamento geográfico, o ponto de melhor IQA da cidade, JACU 02900, localiza-se na “Ponte sobre o rio Jacu, na Rodovia Júlio Fortes, na divisa de Cruzeiro e Lavrinhas” (CETESB, 2024a). Por outro lado, o ponto com os piores valores de IQA é o PONT 04950, localizado no Córrego do Pontilhão, “na sua foz, na ponte junto à linha do trem, em Cruzeiro” (CETESB, 2024a).

Na Figura 19, apresenta-se em “A” a localização da cidade de Cruzeiro. Em “B”, destaca-se o ponto de monitoramento JACU 02900, que registrou os melhores valores de IQA do município e caracteriza-se por estar em uma região rural, com poucas construções no entorno e com mata ciliar já degradada. Em “C”, exibe-se o ponto de medição PONT 04950, que obteve os piores valores de IQA de Cruzeiro, caracterizado por estar inserido em área urbana, com densidade de construções ao longo de seu curso e praticamente desprovido de mata ciliar.

Figura 19 – Localização em Cruzeiro dos pontos JACU 02900 e PONT 04950.



Fonte: Autor.

4.1.3 Comparativo das análises descritivas

As análises já realizadas permitem identificar uma característica comum entre os pontos de medição: a variação do IQA em função da sazonalidade. Alguns pontos apresentam IQA melhor na estação seca, enquanto outros obtêm melhores índices na estação úmida. Para realizar a separação dos dados nesses dois grupos, foi feita a comparação entre as medianas de cada ponto nas estações seca e úmida. A estação que registrasse a maior mediana determinaria o agrupamento daquele ponto de monitoramento. Dessa forma, foram gerados os gráficos da Figura 20, que mostra os pontos com melhora do IQA na estação seca, e da Figura 21, que ilustra as localidades com melhora do índice na estação úmida.

A maioria dos pontos está concentrada no grupo que possui melhor IQA na estação seca, incluindo todos os pontos de monitoramento do rio Paraíba do Sul. (Prado *et al.*, 2021) afirma que, na estação úmida, aumenta o aporte de eletrólitos, nitrato, ferro total e bactérias nos corpos d'água, sendo esses níveis mais elevados nas regiões mais antropizadas. Tal aporte de componentes químicos e biológicos causa um impacto negativo no valor do IQA. Sendo assim, áreas mais urbanizadas tendem a apresentar esse perfil de variação do IQA.

Nesta seção, os dados serão aplicados aos modelos de AM, visando alcançar os objetivos desta pesquisa. Estabeleceu-se o critério de selecionar um ponto de medição para cada município.

Cidades que contavam com mais de um ponto de coleta de água tiveram a definição do ponto estabelecida pela mediana das médias de IQA. A Tabela 6 define essa relação de município, ponto de coleta e média do IQA.

Tabela 6 – Relação dos pontos de monitoramento selecionados por município.

Município	Cod. Pont	Média IQA	Município	Cod. Pont	Média IQA
APARECIDA	PARB02600	49,30	PINDAMONHANGABA	PARB02530	56,66
CAÇAPAVA	PARB02400	53,05	QUELUZ	PARB02900	61,51
CACHOEIRA PAULISTA	PQTE02800	66,07	SANTA BRANCA	PARB02050	68,55
CAMPOS DO JORDÃO	SAGU02050	52,85	SANTA ISABEL	JAGJ00200	71,49
CARAGUATATUBA	COCA02900	64,06	SANTO ANTÔNIO DO PINHAL	PRAT02400	52,19
CRUZEIRO	ALIM02950	40,83	SÃO BENTO DO SAPUCAÍ	SAMI02200	58,02
GUARATINGUETÁ	GUAT02800	67,16	SÃO LUÍS DO PARAÍTINGA	PTIN00850	63,45
ILHABELA	NSRA02900	61,98	SÃO JOSÉ DOS CAMPOS	PARB02310	59,16
JACAREÍ	PTEI02900	58,62	SÃO SEBASTIÃO	RUNA02950	67,58
JAMBEIRO	SANT00100	83,78	TAUBATÉ	UNNA02800	58,38
LORENA	PARB02700	52,25	TREMEMBÉ	PARB02490	58,62
NATIVIDADE DA SERRA	PUNA00800	77,37	UBATUBA	TAVE02950	67,38
PARAIBUNA	IUNA00950	85,43			

Fonte: Autor.

Dessa forma, estabelece-se uma regra para a seleção dos dados para o treinamento e verifica-se se os modelos gerados por pontos selecionados pela mediana da média dos IQAs apresentam resultados satisfatórios para prever o IQA dos demais pontos de medição do município ou de outras estações. Por falta de dados no período de validação, os pontos COCA02900, NSRA02900 e RUNA02950 foram excluídos desta etapa. Com base na relação apresentada na Tabela 6, foram selecionados seis municípios distribuídos pelas sub-regiões da RMVALE-LN para o treinamento e a realização das análises comparativas. As localidades e seus respectivos pontos de monitoramento abrangem: São José dos Campos (PARB02310), Jambuí (SANT00100), Taubaté (UNNA02800), Guaratinguetá (GUAT02800), Cruzeiro (ALIM02950) e Ubatuba (TAVE02950).

Para validar a viabilidade de predição do IQA omitindo certos parâmetros, serão realizados sete testes, com a exclusão de até três variáveis (as três de maior peso no índice), conforme detalhado na Tabela 7.

Na sequência, esses parâmetros excluídos serão substituídos pela predição de seus respectivos valores, e o IQA será calculado com esse novo conjunto de dados. Assim, será possível verificar o desempenho desse método para estimar o valor do IQA, bem como identificar qual abordagem é mais assertiva na ausência de parâmetros. Para a validação dos métodos propostos, serão utilizados os dados de medição referentes aos anos de 2024 e 2025.

4.2.1 Manipulação dos dados para treinamento de modelos de Floresta Aleatória

Para o treinamento dos modelos de AM, será utilizada a Floresta Aleatória. Apesar de ter alta adaptabilidade, a FA não trabalha nativamente com dados de séries temporais. Para a obtenção de

Tabela 7 – Tabela de variáveis excluídas por teste de IQA.

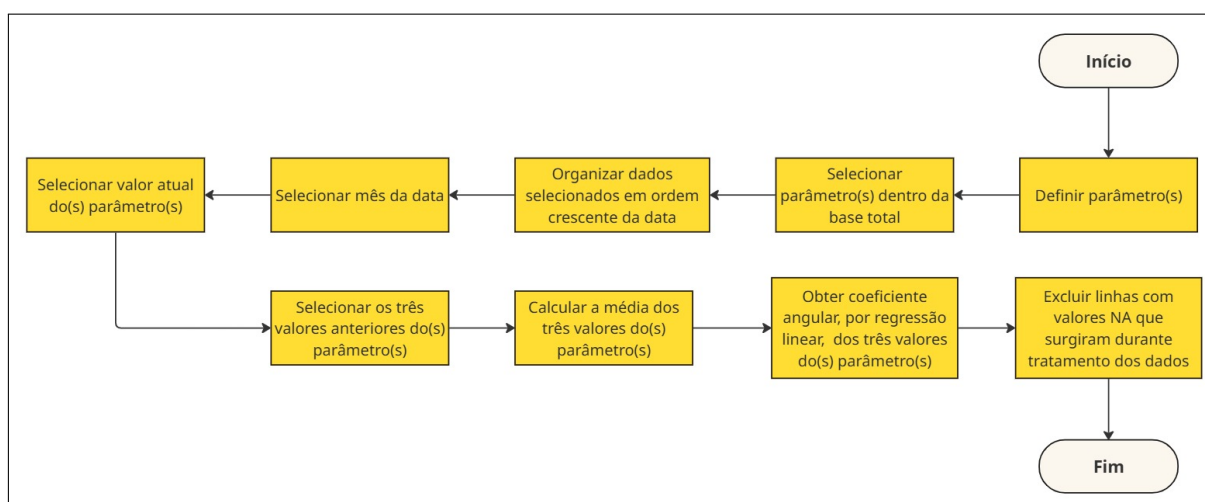
Id Teste	Variável Excluída	Variável Excluída	Variável Excluída
IQA_1	Oxigênio Dissolvido	-	-
IQA_2	Coliformes Termotolerantes	-	-
IQA_3	pH	-	-
IQA_4	Oxigênio Dissolvido	Coliformes Termotolerantes	-
IQA_5	Oxigênio Dissolvido	pH	-
IQA_6	Coliformes Termotolerantes	pH	-
IQA_7	Oxigênio Dissolvido	Coliformes Termotolerantes	pH

Fonte: Autor.

melhores resultados, é necessário estruturar esses dados e criar métricas auxiliares. A abordagem de janela deslizante (*Sliding Window*) atua ajudando os modelos a capturar padrões temporais e a se adaptarem de forma mais eficaz à dinâmica de variação dos dados (Ndungi; Stanislavovich, 2025).

Essa manipulação é executada conforme o fluxo apresentado na Figura 22. O procedimento organiza a base em ordem cronológica crescente e, em seguida, seleciona como colunas a data, o mês correspondente, o valor atual, três valores defasados temporalmente (t_{-1} , t_{-2} , t_{-3}), a média desses três valores e o coeficiente angular (slope) obtido pela regressão linear dos três registros anteriores. A Tabela 8 exemplifica o resultado dessa estruturação para o treinamento focado no parâmetro Oxigênio Dissolvido.

Figura 22 – Fluxo para tratamento de dados direcionado ao modelo de FA.



Fonte: Autor.

Por fim, antes de os dados serem disponibilizados para a rotina de treinamento, a coluna “Data Coleta” é desconsiderada, a coluna com o valor atual do parâmetro é indicada como parâmetro alvo e as demais colunas são indicadas como parâmetros de treinamento.

Tabela 8 – Dados temporais de Oxigênio Dissolvido formatados para o modelo de FA.

Data Coleta	Oxigênio Dissolvido	Mês	Oxigênio Dissolvido t_1	Oxigênio Dissolvido t_2	Oxigênio Dissolvido t_3	Oxigênio Dissolvido media3	Oxigênio Dissolvido slope3
07/08/2013	5,2	8	5,2	4,3	3,9	4,90	0,45
09/10/2013	4,9	10	5,2	5,2	4,3	5,10	-0,15
04/12/2013	3,6	12	4,9	5,2	5,2	4,57	-0,8
05/02/2014	3	2	3,6	4,9	5,2	3,83	-0,95
09/04/2014	3,2	4	3	3,6	4,9	3,27	-0,2
11/06/2014	4,9	6	3,2	3	3,6	3,70	0,95
20/08/2014	4,2	8	4,9	3,2	3,0	4,10	0,5

Fonte: Autor.

4.3 EXECUÇÃO DOS MODELOS FA E COMPARAÇÃO DOS RESULTADOS.

4.3.1 Aplicação dos dados do ponto de monitoramento PARB02310 no modelo FA

Para a validação das previsões realizadas com os dados do ponto de coleta PARB02310, foram utilizados os dados da Tabela 9, referentes aos anos de 2024 e 2025.

Tabela 9 – Parâmetros de qualidade da água e IQA para validação do ponto PARB02310.

Data Coleta	Oxigênio Dissolvido	Colif. Termotol.	pH	DBO 5, 20	Nitrogênio Total	Temp. da Água	Fósforo Total	Turbidez	Sólido Total	IQA
31/01/2024	4,7	1700	6,9	5,81	4,68	25	0,06	32	58	
22/05/2024	5,0	2625	6,5	4,62	1,32	22	0,05	10	60	
07/08/2024	6,0	2175	6,8	5,59	1,42	22	0,11	14	62	
30/10/2024	4,9	9250	6,3	4,72	1,29	25	0,05	17	55	
29/01/2025	4,7	2625	6,2	7,76	1,63	25	0,08	74	51	
07/05/2025	5,7	5000	6,6	3,20	1,50	21	0,05	16	60	
06/08/2025	6,8	1500	6,4	2,35	1,35	20	0,04	8	67	
29/10/2025	6,0	1975	6,8	2,81	1,84	23	0,02	15	66	

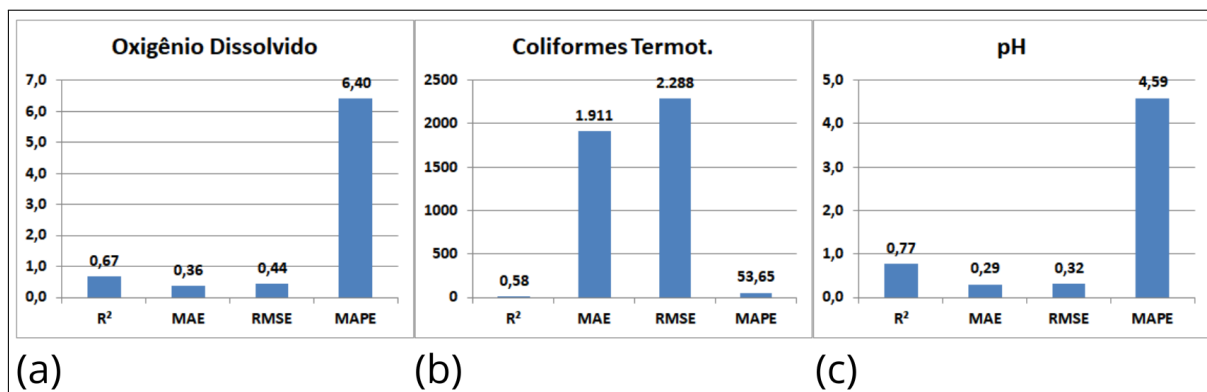
Fonte: Autor.

Após a manipulação dos dados para a utilização no modelo de FA, as três primeiras linhas foram excluídas, restando os dados de 30/10/2024 a 29/10/2025. A previsão dos parâmetros oxigênio dissolvido, pH e coliformes termotolerantes para o cálculo do IQA, após a exclusão desses parâmetros, apresentou as métricas demonstradas na Figura 23 em (a), (b) e (c), respectivamente.

A Figura 24 (a) apresenta o resultado da previsão direta dos valores de IQA, enquanto em (b) é exibido o resultado do cálculo do IQA ao substituir os valores dos parâmetros excluídos por seus valores preditos, conforme apresentado na Tabela 7.

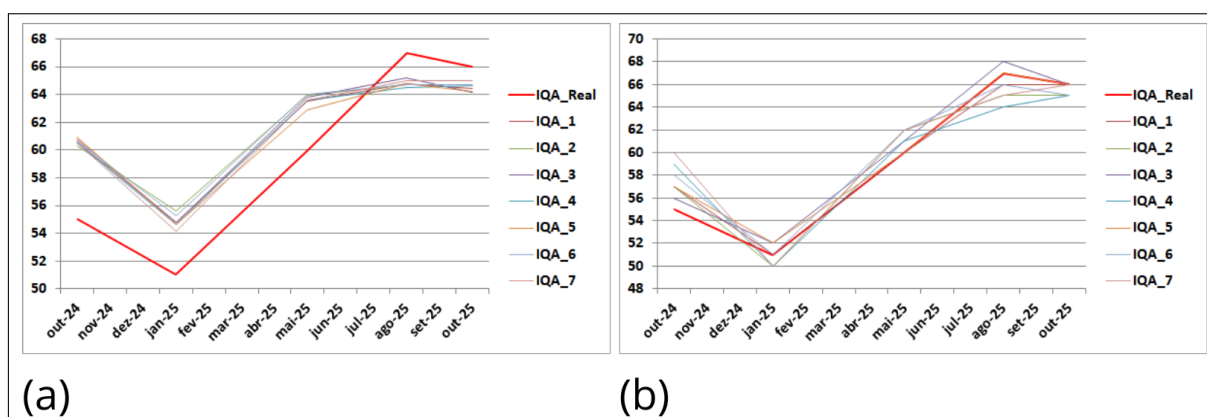
É possível observar, na Figura 24 (a), que a previsão direta do IQA apresenta linhas

Figura 23 – Métricas para predição dos parâmetros excluídos - PARB02310.



Fonte: Autor.

Figura 24 – IQA predito versus IQA calculado por parâmetros preditos - PARB02310.



Fonte: Autor.

ligeiramente mais espaçadas do que no gráfico da Figura 24 (b) do IQA calculado com parâmetros preditos, no qual algumas linhas se sobrepõem em determinados períodos. Isso indica uma maior repetibilidade entre os resultados a cada teste deste segundo método.

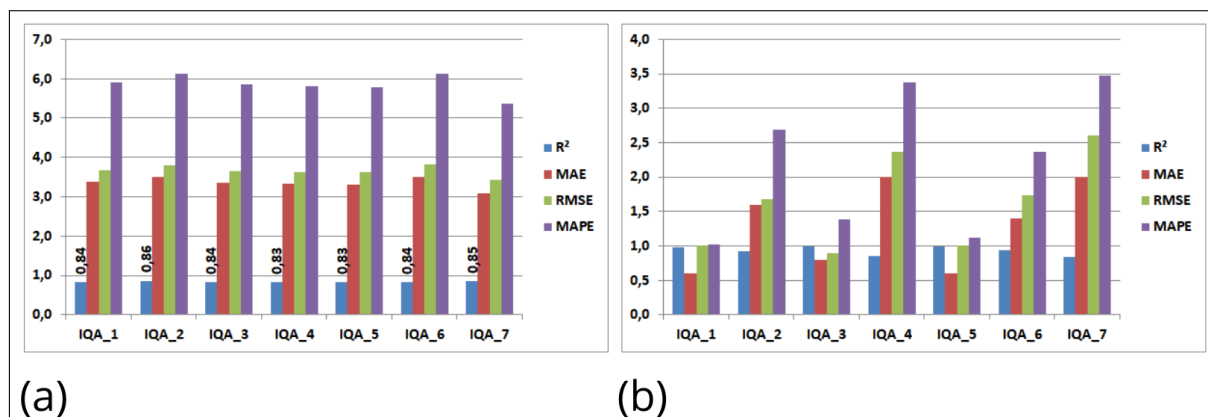
O melhor desempenho do método que utiliza o IQA calculado com parâmetros preditos pode ser constatado na Figura 25. Observa-se que suas métricas em (b) correspondem a cerca da metade dos valores obtidos em (a), referentes ao primeiro método testado.

Tomando a Tabela 2 como referência, o valor de predição de coliformes termotolerantes pode ser considerado "impreciso". Apesar disso, a aplicação de suas predições no cálculo do IQA demonstrou ter um impacto reduzido, visto que os valores de MAPE do segundo método mantiveram-se abaixo de 3,5% em todos os testes, sendo classificados como uma previsão altamente precisa.

4.3.2 Aplicação dos dados do ponto de monitoramento SANT00100 no modelo FA

Para a validação das predições realizadas com os dados do ponto de coleta SANT00100, foram utilizados os dados da Tabela 10, referentes aos anos de 2024 e 2025.

Figura 25 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - PARB02310.



Fonte: Autor.

Tabela 10 – Parâmetros de qualidade da água e IQA para validação do ponto SANT00100.

Data Coleta	Oxigênio Dissolvido	Colif. Termotol.	pH	DBO 5, 20	Nitrogênio Total	Temp. da Água	Fósforo Total	Turbidez	Sólido Total	IQA
19/02/2024	7,0	2,5	6,9	4,72	1,07	26	0,02	5	87	
06/05/2024	7,7	1,25	6,6	4,20	1,07	26	0,02	5	88	
14/08/2024	7,7	1,25	6,6	2,44	1,14	18	0,02	5	89	
04/11/2024	8,0	1,25	7,6	2,81	1,10	24	0,02	5	91	
12/02/2025	8,3	1,25	7,9	3,49	1,17	30	0,02	5	89	
27/05/2025	7,0	1,25	6,7	1,99	1,19	23	0,02	5	90	
13/08/2025	7,8	5,0	6,6	2,35	1,07	20	0,02	5	87	
03/11/2025	8,0	12,5	6,6	3,59	1,16	23	0,02	5	84	

Fonte: Autor.

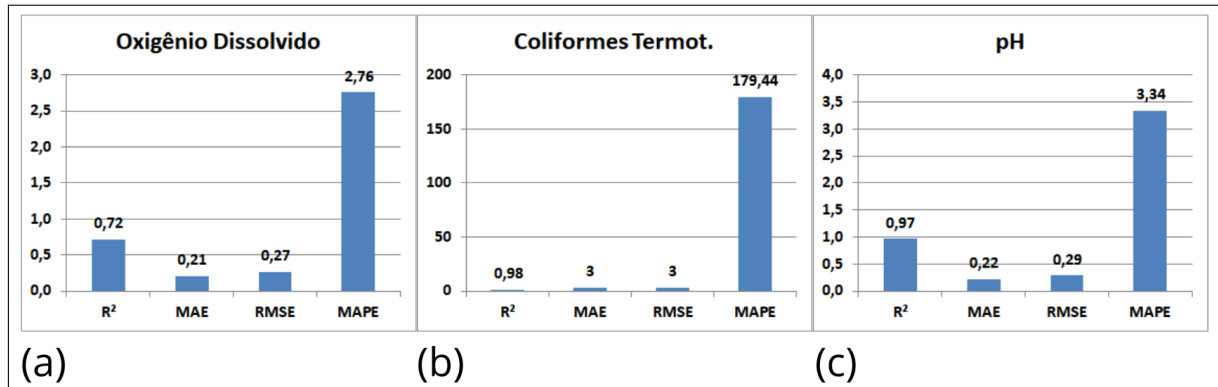
Após a manipulação dos dados para a utilização no modelo de FA, as três primeiras linhas foram excluídas, restando os dados de 04/11/2024 a 03/11/2025. A predição dos parâmetros oxigênio dissolvido, pH e coliformes termotolerantes para o cálculo do IQA, após a exclusão desses parâmetros, apresentou as métricas demonstradas na Figura 26 em (a), (b) e (c), respectivamente.

A Figura 27 (a) apresenta o resultado da predição direta dos valores de IQA, enquanto em (b) é exibido o resultado do cálculo do IQA substituindo os valores dos parâmetros excluídos por seus valores preditos, conforme apresentado na Tabela 7.

É possível observar, na Figura 27 (a), que a predição direta do IQA apresenta linhas ligeiramente mais espaçadas do que no gráfico da Figura 27 (b) do IQA calculado com parâmetros preditos, no qual algumas linhas se sobrepõem em determinados períodos. Isso indica uma maior repetibilidade entre os resultados a cada teste deste segundo método.

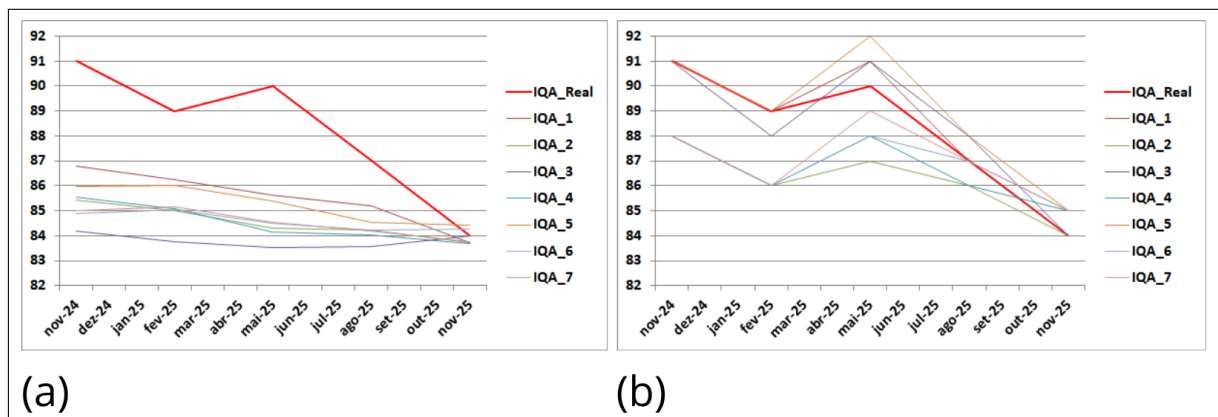
O melhor desempenho do método que utiliza o IQA calculado com parâmetros preditos pode ser constatado na Figura 28. Observa-se que suas métricas em (b) correspondem a cerca da

Figura 26 – Métricas para predição dos parâmetros excluídos - SANT00100.



Fonte: Autor.

Figura 27 – IQA predito versus IQA calculado por parâmetros preditos - SANT00100.



Fonte: Autor.

metade do valor de (a), referente ao primeiro método testado.

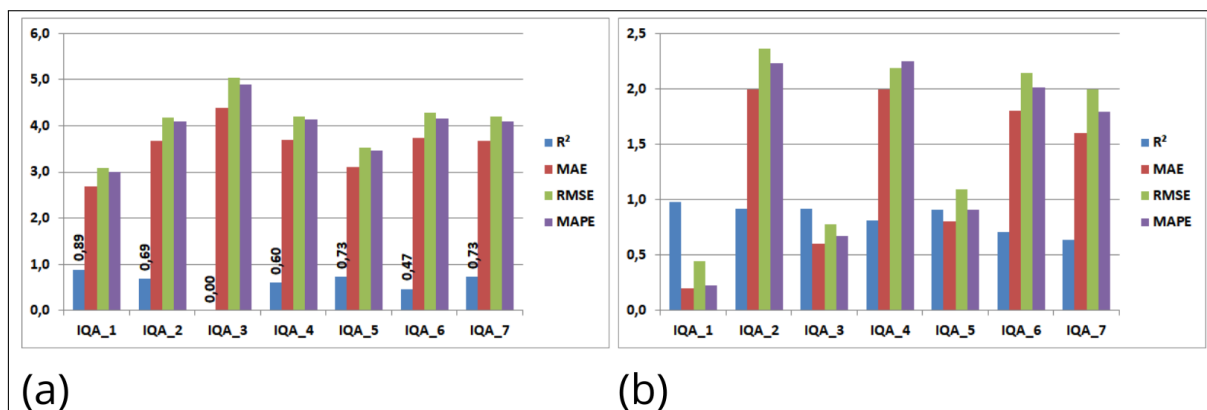
Tomando a Tabela 2 como referência, a predição de coliformes termotolerantes pode ser classificada como "imprecisa". No entanto, se avaliados em conjunto os valores de R^2 , MAE e RMSE, percebe-se que o modelo explica 98% da variância dos dados e possui um erro médio de 3 unidades — um valor baixo para uma variável com escala ampla. Desse modo, mesmo com um MAPE elevado (em torno de 200%), o seu impacto numérico no cálculo do IQA é desprezível, permitindo que os valores de MAPE do segundo método fiquem abaixo de 2,5%, o que configura uma previsão altamente precisa.

4.3.3 Aplicação dos dados do ponto de monitoramento UNNA02800 no modelo FA

Para a validação das predições realizadas com os dados do ponto de coleta UNNA02800, foram utilizados os dados da Tabela 11, referentes aos anos de 2024 e 2025.

Após a manipulação dos dados para a utilização no modelo de FA, excluem-se as três primeiras linhas, restando os dados de 18/11/2024 a 05/11/2025. A predição dos parâmetros oxigênio dissolvido, pH e coliformes termotolerantes para o cálculo do IQA, após a exclusão desses

Figura 28 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - SANT00100.



Fonte: Autor.

Tabela 11 – Parâmetros de qualidade da água e IQA para validação do ponto UNNA02800.

Data Coleta	Oxigênio Dissolvido	Colif. Termotol.	pH	DBO 5, 20	Nitrogênio Total	Temp. da Água	Fósforo Total	Turbidez	Sólido Total	IQA
21/02/2024	5,9	121250	5,8	67,04	1,20	20	0,02	1840	25	
15/05/2024	7,2	6750	7,7	3,20	1,33	20	0,04	60	58	
21/08/2024	7,9	975	7,8	2,81	3,57	18	0,02	51	65	
18/11/2024	6,2	10625	7,5	8,71	1,26	23	0,02	488	44	
19/02/2025	6,1	14250	7,4	10,07	1,38	25	0,02	404	42	
21/05/2025	7,2	7250	6,9	2,72	1,44	17	0,02	5	63	
20/08/2025	9,0	1075	6,7	1,48	4,36	17	0,02	5	71	
05/11/2025	7,8	3125	7,2	2,72	1,32	23	0,03	63	62	

Fonte: Autor.

parâmetros, apresentou as métricas demonstradas na Figura 29 em (a), (b) e (c), respectivamente.

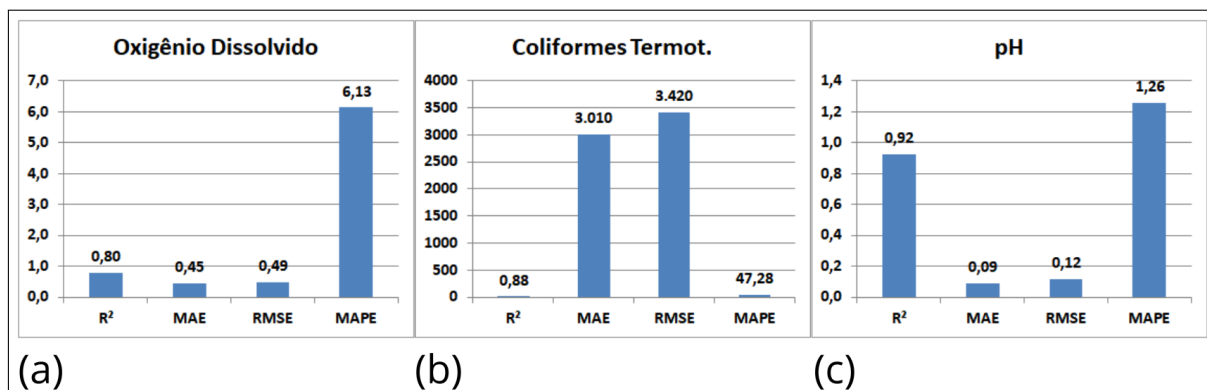
A Figura 30 (a) apresenta o resultado da predição direta dos valores de IQA, enquanto em (b) tem-se o resultado do cálculo do IQA substituindo os valores dos parâmetros excluídos por seus valores preditos, conforme apresentado na Tabela 7.

É possível observar, na Figura 30 (a), que a predição direta do IQA apresenta linhas ligeiramente mais deslocadas do valor real do que no gráfico da Figura 30 (b) do IQA calculado com parâmetros preditos, no qual algumas linhas se sobrepõem em certos períodos, inclusive aproximando-se do valor real. Isso indica uma maior repetibilidade entre os resultados a cada teste deste segundo método.

O melhor desempenho do método que utiliza o IQA calculado com parâmetros preditos pode ser constatado na Figura 31. Observa-se que suas métricas em (b) são menores que os valores de (a), referentes ao primeiro método testado.

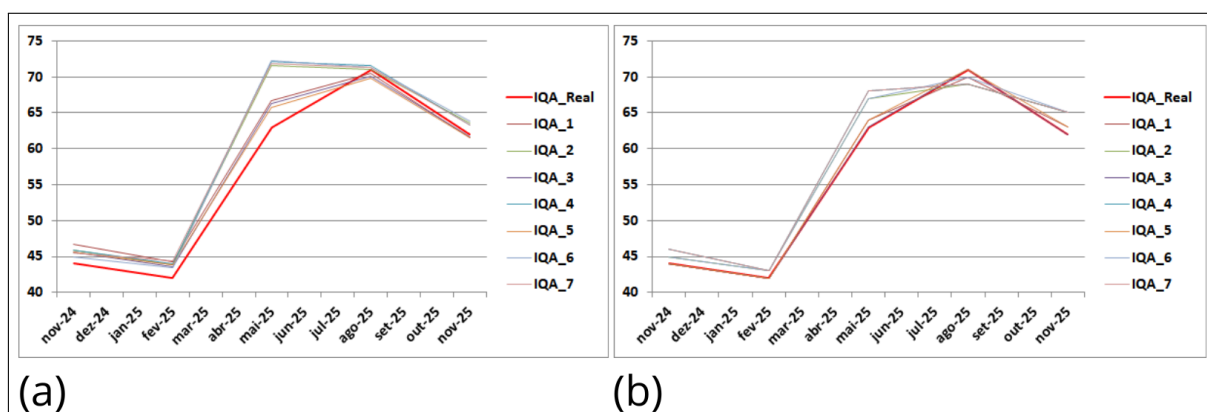
Tomando a Tabela 2 como referência, o valor de predição de coliformes termotolerantes

Figura 29 – Métricas para predição dos parâmetros excluídos - UNNA02800.



Fonte: Autor.

Figura 30 – IQA predito versus IQA calculado por parâmetros preditos - UNNA02800.



Fonte: Autor.

pode ser considerado "razoável". Ainda assim, a aplicação de suas predições no cálculo do IQA demonstrou ter um impacto reduzido, já que os valores de MAPE do segundo método mantiveram-se abaixo de 4,5% em todos os testes, configurando previsões altamente precisas.

4.3.4 Aplicação dos dados do ponto de monitoramento GUAT02800 no modelo FA

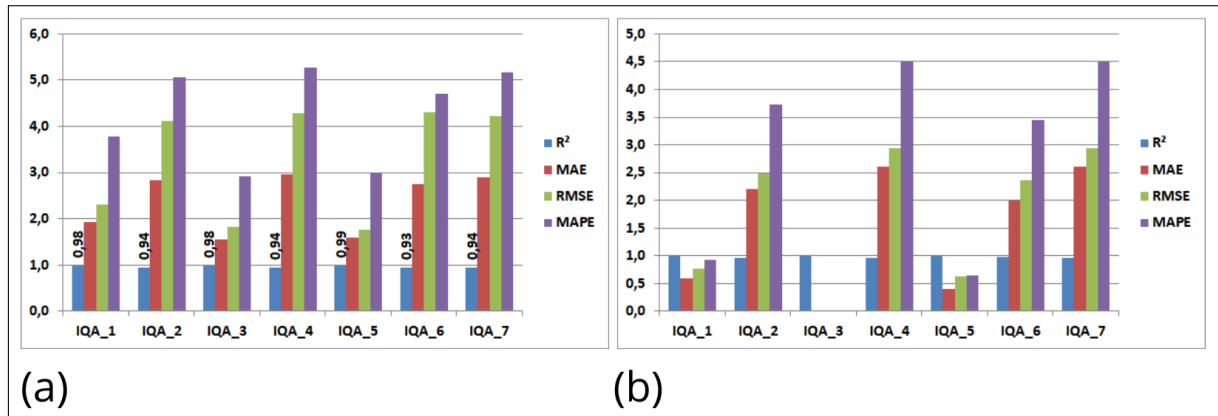
Para a validação das predições realizadas com os dados do ponto de coleta GUAT02800, foram utilizados os dados da Tabela 12, referentes aos anos de 2024 e 2025.

Após a manipulação dos dados para a utilização no modelo de FA, as três primeiras linhas foram excluídas, restando os dados de 18/11/2024 a 05/11/2025. A predição dos parâmetros oxigênio dissolvido, pH e coliformes termotolerantes para o cálculo do IQA, após a exclusão desses parâmetros, apresentou as métricas demonstradas na Figura 32 em (a), (b) e (c), respectivamente.

A Figura 33 (a) apresenta o resultado da predição direta dos valores de IQA, enquanto em (b) é exibido o resultado do cálculo do IQA substituindo os valores dos parâmetros excluídos por seus valores preditos, conforme apresentado na Tabela 7.

É possível observar, na Figura 33 (a), que a predição direta do IQA apresenta linhas

Figura 31 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - UNNA02800.



Fonte: Autor.

Tabela 12 – Parâmetros de qualidade da água e IQA para validação do ponto GUAT02800.

Data Coleta	Oxigênio Dissolvido	Colif. Termotol.	pH	DBO 5, 20	Nitrogênio Total	Temp. da Água	Fósforo Total	Turbidez	Sólido Total	IQA
21/02/2024	5,6	10125	6,5	11,33	3,65	20	0,05	101	50	41
15/05/2024	6,8	575	7,1	4,1	4,9	20	0,04	12	50	68
21/08/2024	7	625	7,3	1,73	1,11	19	0,02	7	50	73
18/11/2024	6,2	1200	7,1	5,59	1,18	22	0,04	29	50	64
19/02/2025	6,2	600	7	3	1,13	25	0,02	7	68	72
21/05/2025	7,2	1000	7,2	2,63	1,15	17	0,03	6	50	70
20/08/2025	8,4	320	6,4	1,65	1,07	18	0,02	5	50	75
05/11/2025	7,2	1225	6,7	3,49	1,73	23	0,09	25	94	66

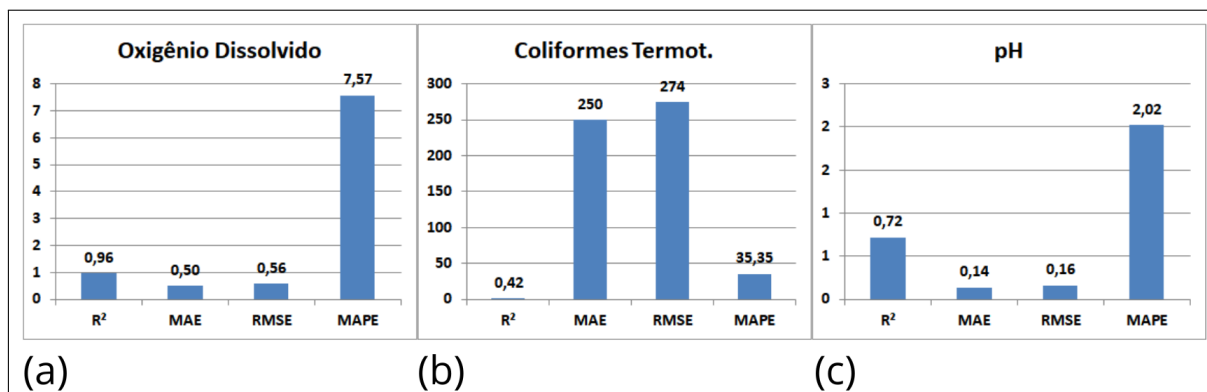
Fonte: Autor.

ligeiramente mais espaçadas do que no gráfico da Figura 33 (b) do IQA calculado com parâmetros preditos, no qual algumas linhas se sobrepõem em determinados períodos. Isso indica uma maior repetibilidade entre os resultados a cada teste deste segundo método.

O melhor desempenho do método que utiliza o IQA calculado com parâmetros preditos pode ser constatado na Figura 34. Observa-se que suas métricas em (b) correspondem a cerca da metade do valor de (a), correspondente ao primeiro método testado.

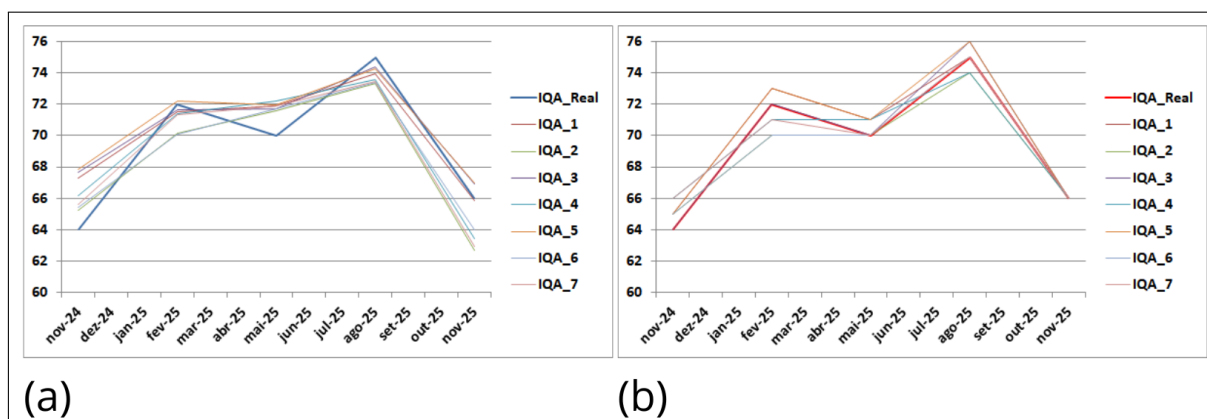
Tomando a Tabela 2 como referência, o valor de predição de coliformes termotolerantes pode ser classificado como "razoável". Ainda assim, a aplicação de suas predições no cálculo do IQA demonstrou ter um impacto reduzido, visto que os valores de MAPE do segundo método mantiveram-se abaixo de 1,5% em todos os testes, o que configura uma previsão altamente precisa.

Figura 32 – Métricas para predição dos parâmetros excluídos - GUAT02800.



Fonte: Autor.

Figura 33 – IQA predito versus IQA calculado por parâmetros preditos - GUAT02800.



Fonte: Autor.

4.3.5 Aplicação dos dados do ponto de monitoramento ALIM02950 no modelo FA

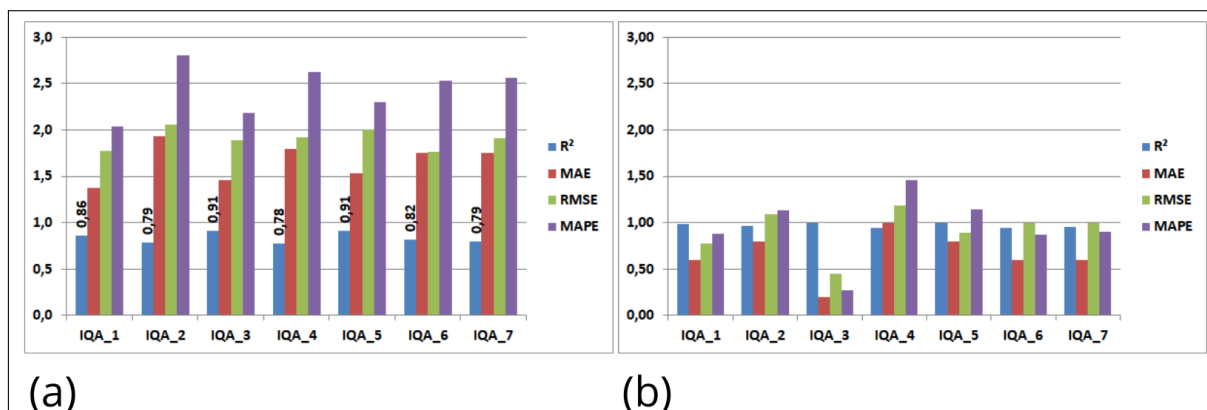
Para a validação das predições realizadas com os dados do ponto de coleta ALIM02950, foram utilizados os dados da Tabela 13, referentes aos anos de 2024 e 2025.

Após a manipulação dos dados para a utilização no modelo de FA, as três primeiras linhas foram excluídas, restando os dados de 18/11/2024 a 05/11/2025. A predição dos parâmetros oxigênio dissolvido, pH e coliformes termotolerantes para o cálculo do IQA, após a exclusão desses parâmetros, apresentou as métricas demonstradas na Figura 35 em (a), (b) e (c), respectivamente.

A Figura 36 (a) apresenta o resultado da predição direta dos valores de IQA, enquanto em (b) tem-se o resultado do cálculo do IQA substituindo os valores dos parâmetros excluídos por seus valores preditos, conforme apresentado na Tabela 7.

É possível observar, na Figura 36 (a), que a predição direta do IQA apresenta linhas bem deslocadas do valor real. Já no gráfico da Figura 36 (b), referente ao IQA calculado com parâmetros preditos, algumas linhas se sobrepõem em determinados períodos, mostrando-se bem concentradas em torno do valor real. Isso indica uma maior acurácia nos resultados a cada teste deste segundo método.

Figura 34 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - GUAT02800.



Fonte: Autor.

Tabela 13 – Parâmetros de qualidade da água e IQA para validação do ponto ALIM02950.

Data Coleta	Oxigênio Dissolvido	Colif. Termotol.	pH	DBO 5, 20	Nitrogênio Total	Temp. da Água	Fósforo Total	Turbidez	Sólido Total	IQA
21/02/2024	4,4	95000	6,2	5,48	1,26	23	0,10	18	45	
15/05/2024	3,5	25000	6,8	4,41	2,21	21	0,20	9	45	
21/08/2024	2,8	116250	6,9	5,26	2,51	18	0,26	12	37	
18/11/2024	5,0	262500	6,9	5,15	1,98	24	0,15	43	44	
19/02/2025	4,8	142500	6,7	24,72	2,63	25	0,13	9	37	
21/05/2025	4,9	40000	6,6	3,69	3,33	18	0,24	8	47	
20/08/2025	2,4	78750	6,8	6,82	5,18	18	0,30	9	36	
05/11/2025	1,9	262500	6,8	6,14	5,09	23	0,29	11	34	

Fonte: Autor.

O melhor desempenho do método que utiliza o IQA calculado com parâmetros preditos pode ser constatado na Figura 37. Observa-se que suas métricas em (b) correspondem a cerca da metade do valor obtido em (a), que representa o primeiro método testado.

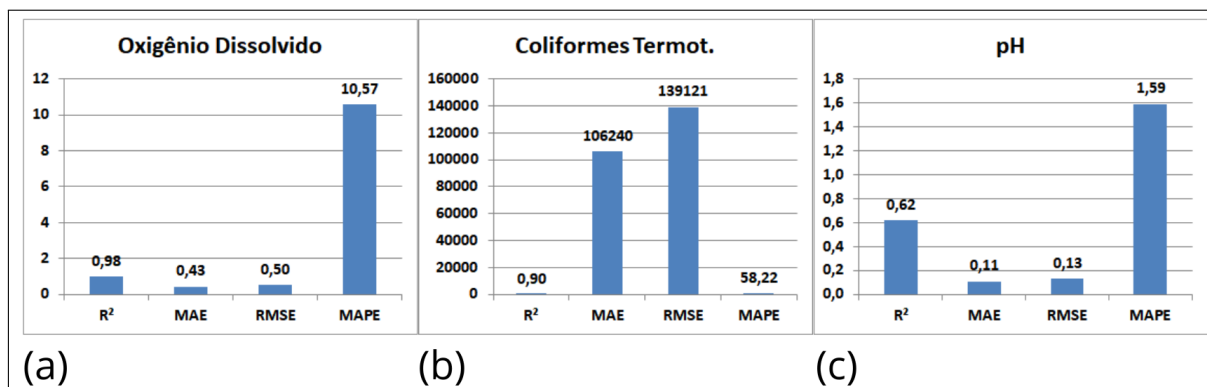
Tomando a Tabela 2 como referência, o valor de predição de coliformes termotolerantes pode ser considerado "impreciso". Apesar disso, a aplicação de suas predições no cálculo do IQA demonstrou ter um impacto reduzido, já que os valores de MAPE do segundo método mantiveram-se abaixo de 4 em todos os testes, o que é classificado como uma previsão altamente precisa.

4.3.6 Aplicação dos dados do ponto de monitoramento TAVE02950 no modelo FA

Para a validação das predições realizadas com os dados do ponto de coleta TAVE02950, foram utilizados os dados da Tabela 14, referentes aos anos de 2024 e 2025.

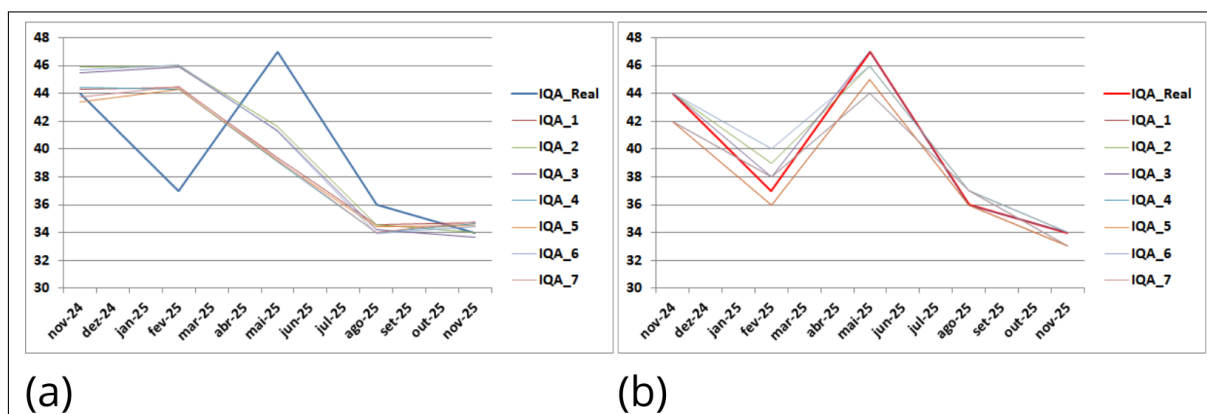
Após a manipulação dos dados para a utilização no modelo de FA, as três primeiras linhas

Figura 35 – Métricas para predição dos parâmetros excluídos - ALIM02950.



Fonte: Autor.

Figura 36 – IQA predito versus IQA calculado por parâmetros preditos - ALIM02950.



Fonte: Autor.

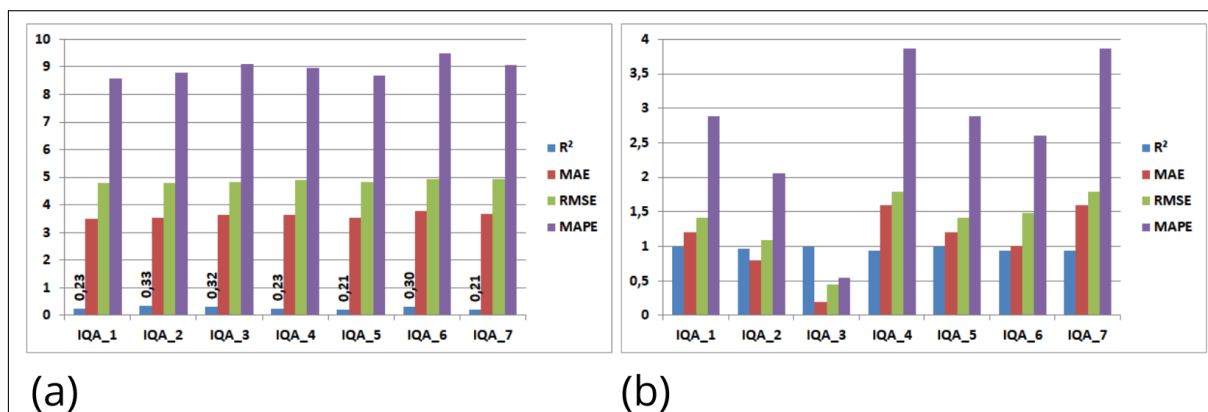
Tabela 14 – Parâmetros de qualidade da água e IQA para validação do ponto TAVE02950.

Data Coleta	Oxigênio Dissolvido	Colif. Termotol.	pH	DBO 5, 20	Nitrogênio Total	Temp. da Água	Fósforo Total	Turbidez	Sólido Total	IQA
13/03/2024	5,6	11625	6,6	3,49	1,30	24	0,06	8	57	
05/06/2024	5,4	12000	6,3	3,39	2,00	20	0,06	18	53	
11/09/2024	7,0	6250	6,6	2,26	1,97	21	0,10	5	62	
06/11/2024	3,9	275000	6,5	3,39	2,07	25	0,11	10	43	
19/03/2025	5,4	9875	6,7	7,29	1,48	24	0,12	109	43	
11/06/2025	8,4	6375	6,7	2,35	1,42	19	0,04	5	64	
10/09/2025	6,9	1625	6,9	3,00	1,72	20	0,02	5	68	
17/11/2025	6,1	9000	6,5	2,81	1,51	23	0,10	5	59	

Fonte: Autor.

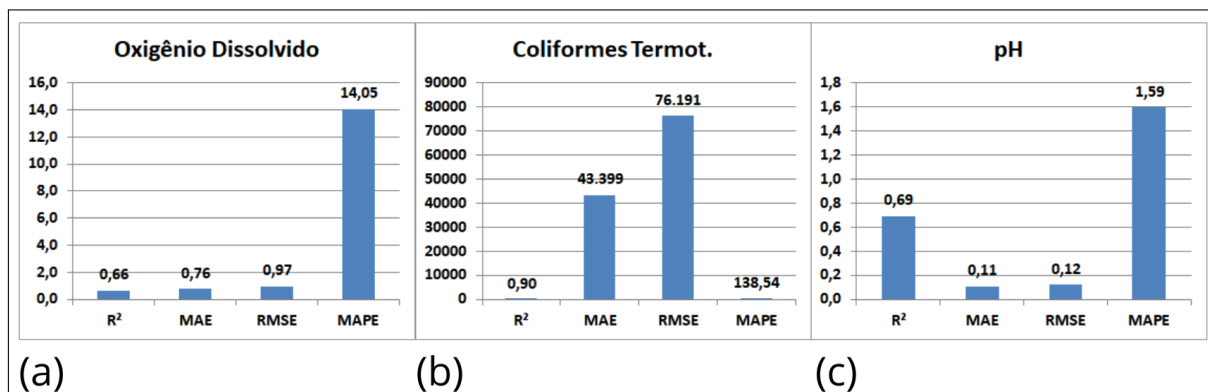
foram excluídas, restando os dados de 06/11/2024 a 17/11/2025. A predição dos parâmetros oxigênio dissolvido, pH e coliformes termotolerantes para o cálculo do IQA, após a exclusão desses parâmetros, apresentou as métricas demonstradas na Figura 38 em (a), (b) e (c), respectivamente.

Figura 37 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - ALIM02950.



Fonte: Autor.

Figura 38 – Métricas para predição dos parâmetros excluídos - TAVE02950.



Fonte: Autor.

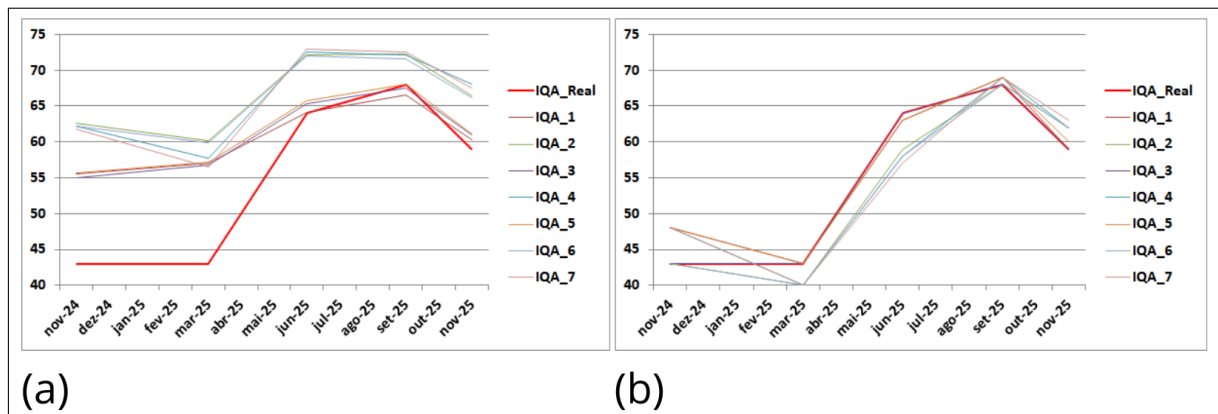
A Figura 39 (a) apresenta o resultado da predição direta dos valores de IQA, enquanto em (b) tem-se o resultado do cálculo do IQA substituindo os valores dos parâmetros excluídos por seus valores preditos, conforme apresentado na Tabela 7.

É possível observar, na Figura 39 (a), que a predição direta do IQA apresenta linhas ligeiramente mais espaçadas do que no gráfico da Figura 39 (b) do IQA calculado com parâmetros preditos, no qual algumas linhas se sobrepõem em determinados períodos. Isso indica uma maior repetibilidade entre os resultados a cada teste deste segundo método.

O melhor desempenho do método que utiliza o IQA calculado com parâmetros preditos pode ser constatado na Figura 40. Observa-se que suas métricas em (b) são cerca de três vezes menores do que os valores obtidos em (a), relativos ao primeiro método testado.

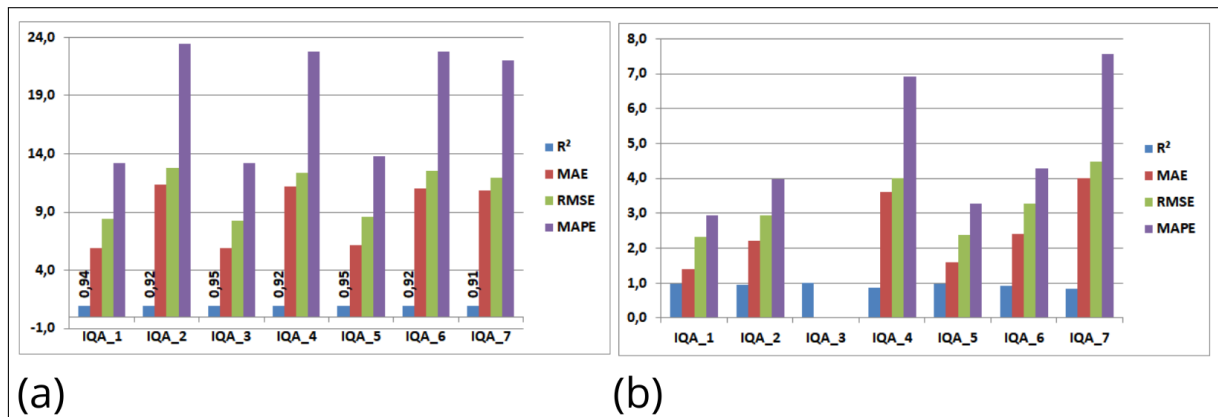
Tomando a Tabela 2 como referência, o valor de predição de coliformes termotolerantes pode ser classificado como "impreciso". Apesar disso, a aplicação de suas predições no cálculo do IQA demonstrou ter um impacto reduzido, já que os valores de MAPE do segundo método mantiveram-se abaixo de 8 em todos os testes, qualificando-se como previsões altamente precisas.

Figura 39 – IQA predito versus IQA calculado por parâmetros preditos - TAVE02950.



Fonte: Autor.

Figura 40 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - TAVE02950.



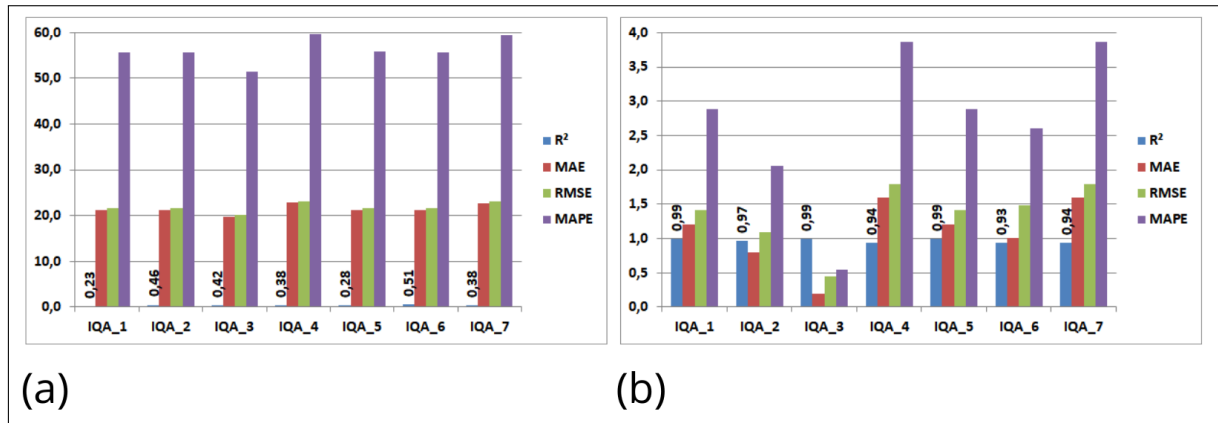
Fonte: Autor.

4.3.7 Aplicação de dados de diversos pontos de monitoramento aos modelos do ponto PARB02900

Com o objetivo de verificar a efetividade em utilizar um modelo de um ponto de monitoramento aplicando a ele dados de outros pontos, selecionou-se o ponto correspondente à mediana dos locais listados na Tabela 6. Assim, para validar essa hipótese, foram utilizados os modelos referentes ao ponto PARB02900 e os dados de medição dos pontos ALIM02950, GUAT02800, PARB02310, SANT00100, TAVE02950 e UNNA02800.

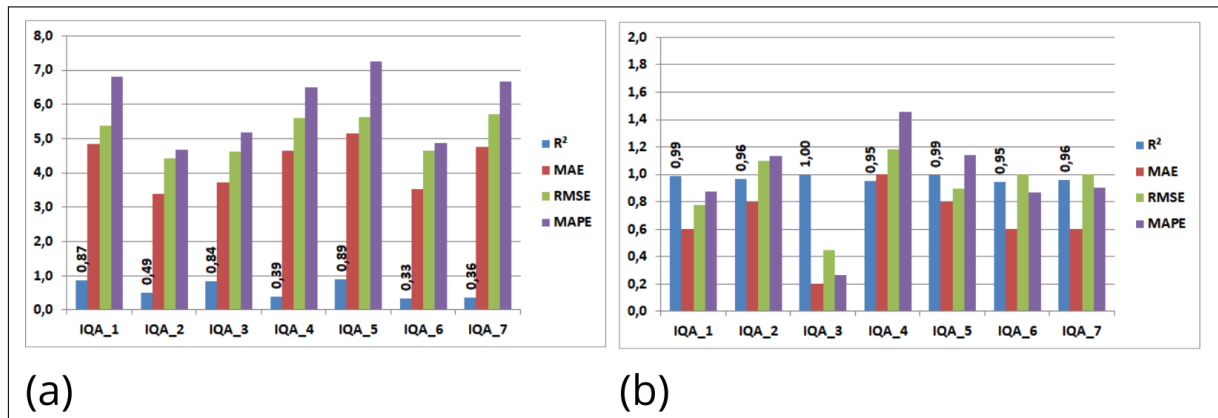
A Figura 41, a Figura 42, a Figura 43, a Figura 44, a Figura 45 e a Figura 46 apresentam a comparação das métricas do IQA predito (em [a]) com o IQA calculado por parâmetros preditos (em [b]). É perceptível a queda de desempenho na predição direta do IQA (a), que atinge na Figura 41 (a) um MAPE próximo a 60%, o que representa uma previsão imprecisa de acordo com a Tabela 2. Por outro lado, o cálculo do IQA com parâmetros preditos não sofreu grandes alterações, apresentando um MAPE abaixo de 10% em todos os casos testados, o que corresponde a uma previsão altamente precisa conforme os critérios da Tabela 2.

Figura 41 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: ALIM02950.



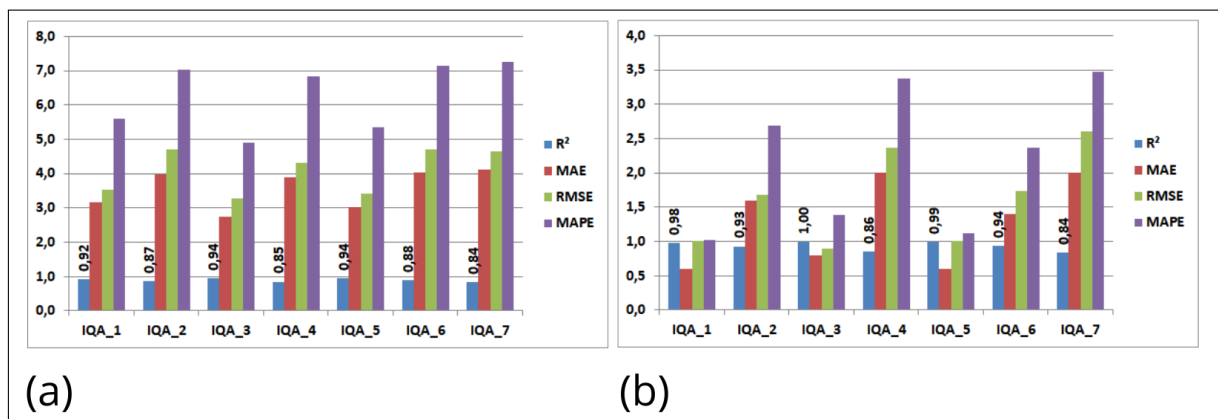
Fonte: Autor.

Figura 42 – Métricas do IQA predito versus IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: GUAT02800.



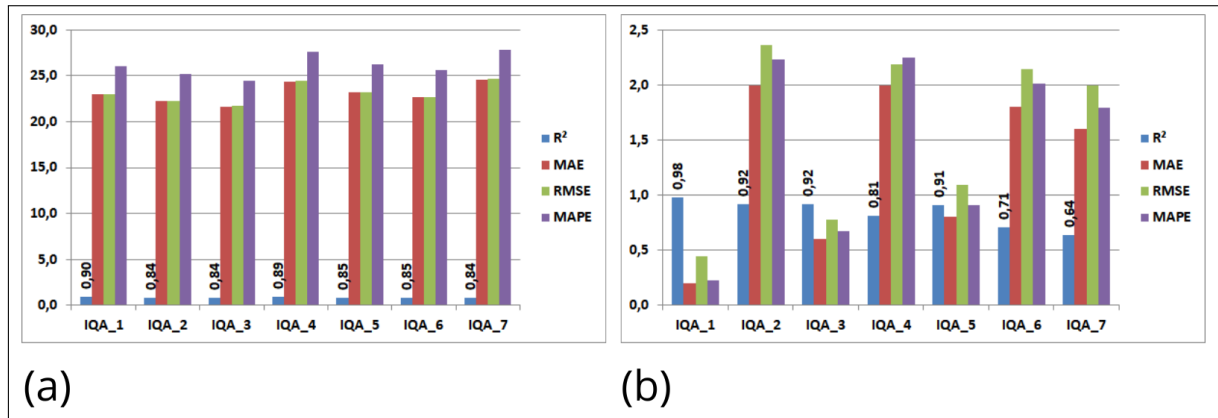
Fonte: Autor.

Figura 43 – Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: PARB02310.



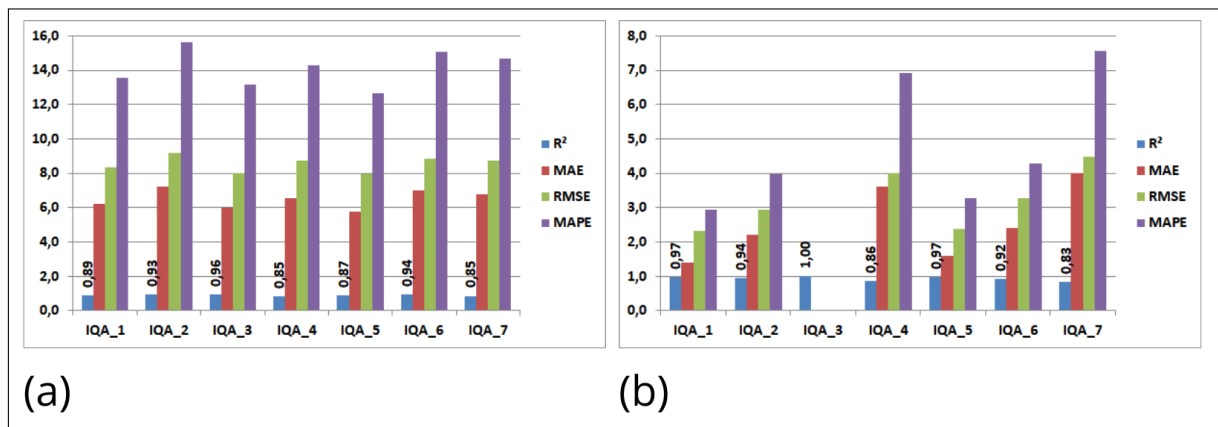
Fonte: Autor

Figura 44 – Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: SANT00100.



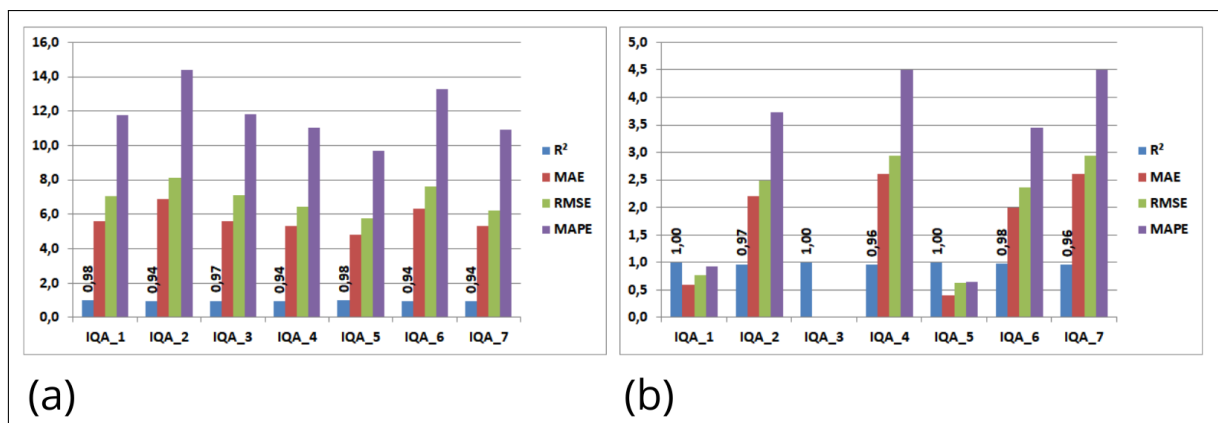
Fonte: Autor

Figura 45 – Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: TAVE02950.



Fonte: Autor

Figura 46 – Métricas do IQA predito x IQA calculado por parâmetros preditos - Modelo: PARB02900, Dados: UNNA02800.

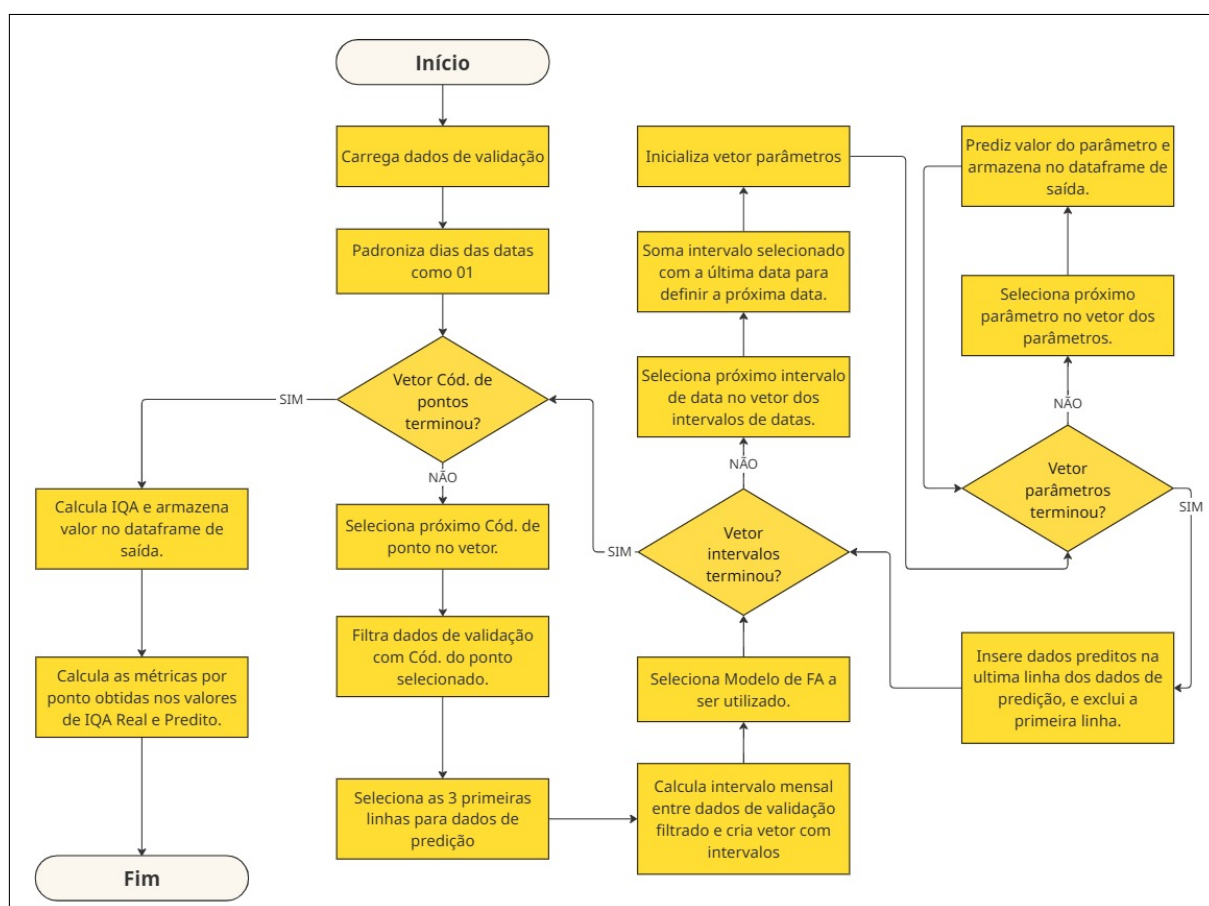


Fonte: Autor

4.3.8 Utilização Iterativa de Valores Preditos em Horizontes de Longo Prazo

Até aqui, foram testadas questões de pesquisa que envolviam a predição do IQA com parâmetros faltantes; desta forma, parte dos parâmetros utilizados era real. Nesta seção, iremos abordar a eficácia da predição de valores futuros do IQA. Será utilizado o segundo método apresentado neste trabalho — o cálculo do IQA por meio da predição de seus parâmetros —, porém, desta vez, todos os parâmetros serão preditos, e os valores estimados irão alimentar o modelo na predição do próximo intervalo, e assim sucessivamente. O modelo utilizado será o do ponto PARB02900 e os dados, além do PARB02900, contemplam os pontos ALIM02950, GUAT02800, PARB02310, SANT00100, TAVE02950 e UNNA02800. A Figura 47 descreve o fluxo de execução para a obtenção dos valores futuros de IQA.

Figura 47 – Fluxo de execução para predição do IQA com utilização iterativa de valores preditos.



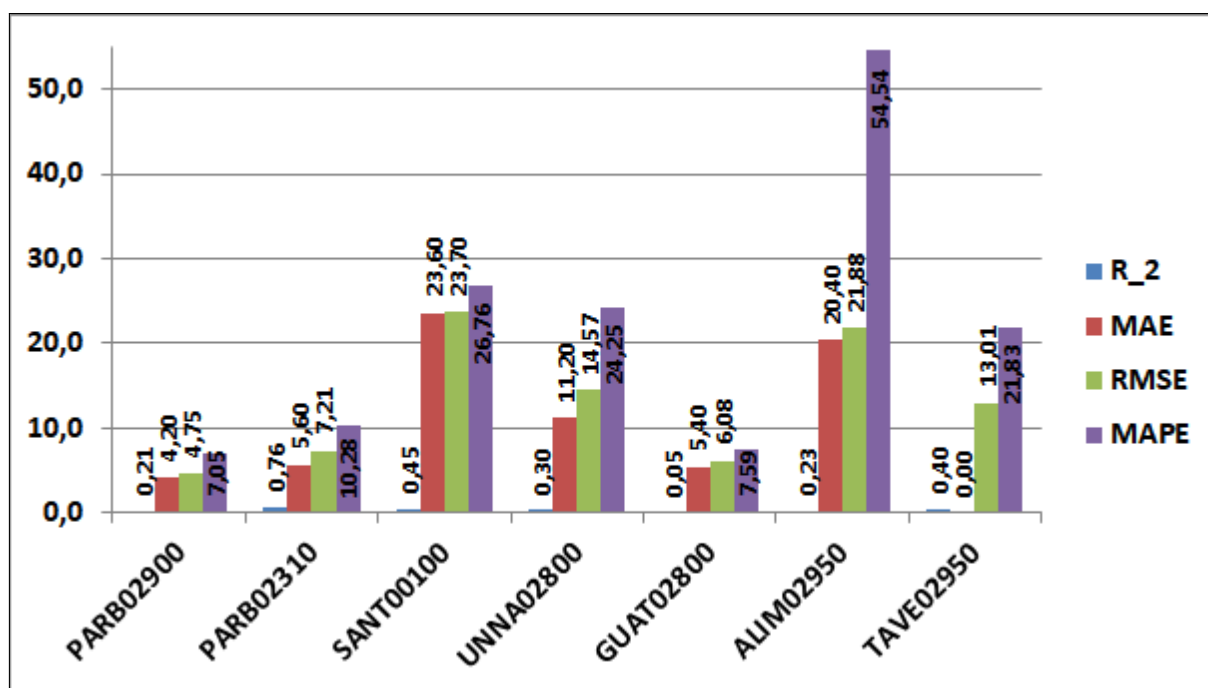
Fonte: Autor

A Figura 48 apresenta as métricas obtidas em cada teste de predição utilizando o modelo do ponto PARB02900 com os dados dos pontos indicados no gráfico. Conforme a Figura 20 e a Figura 21 indicam, o ponto PARB02900 está no conjunto dos pontos em que o IQA melhora no período seco, bem como os pontos GUAT02800, PARB02310, TAVE02950 e UNNA02800. Os dois piores resultados indicados na Figura 48 são os dos pontos ALIM02950 e SANT00100, justamente os que fazem parte do conjunto em que o IQA melhora na estação úmida. Essa

característica de sazonalidade do IQA pode ser um indicador de viabilidade para a utilização de dados de um ponto aplicados ao modelo de outro ponto.

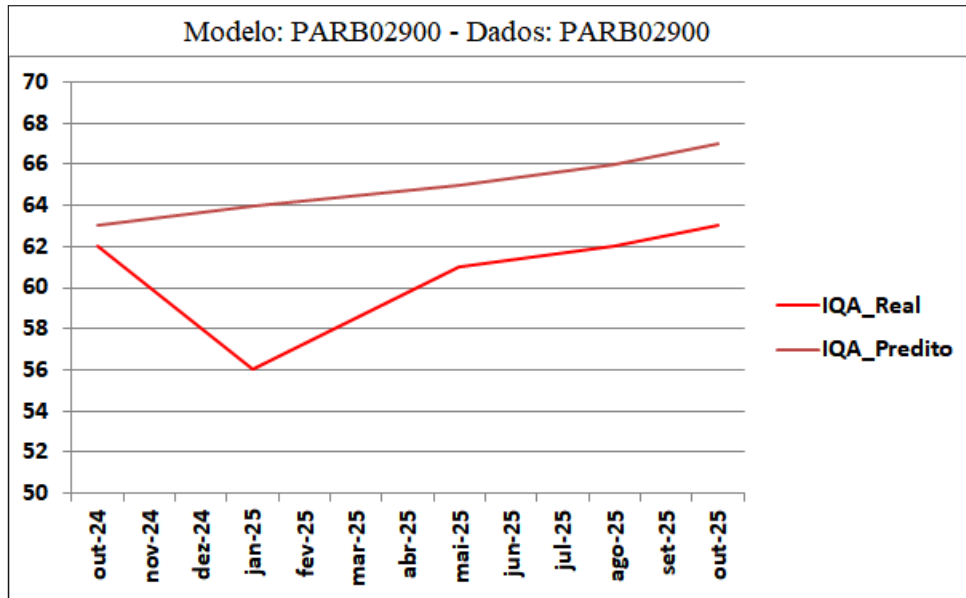
A Figura 49 mostra a comparação entre o valor do IQA real e do IQA predito para o teste com dados e modelo do ponto PARB02900, o melhor resultado obtido, como é possível constatar no resultado das métricas da Figura 48. A Figura 50 apresenta a comparação entre o IQA real e o IQA predito dos demais pontos testados.

Figura 48 – Métricas de desempenho para predição do IQA com utilização iterativa de valores preditos.



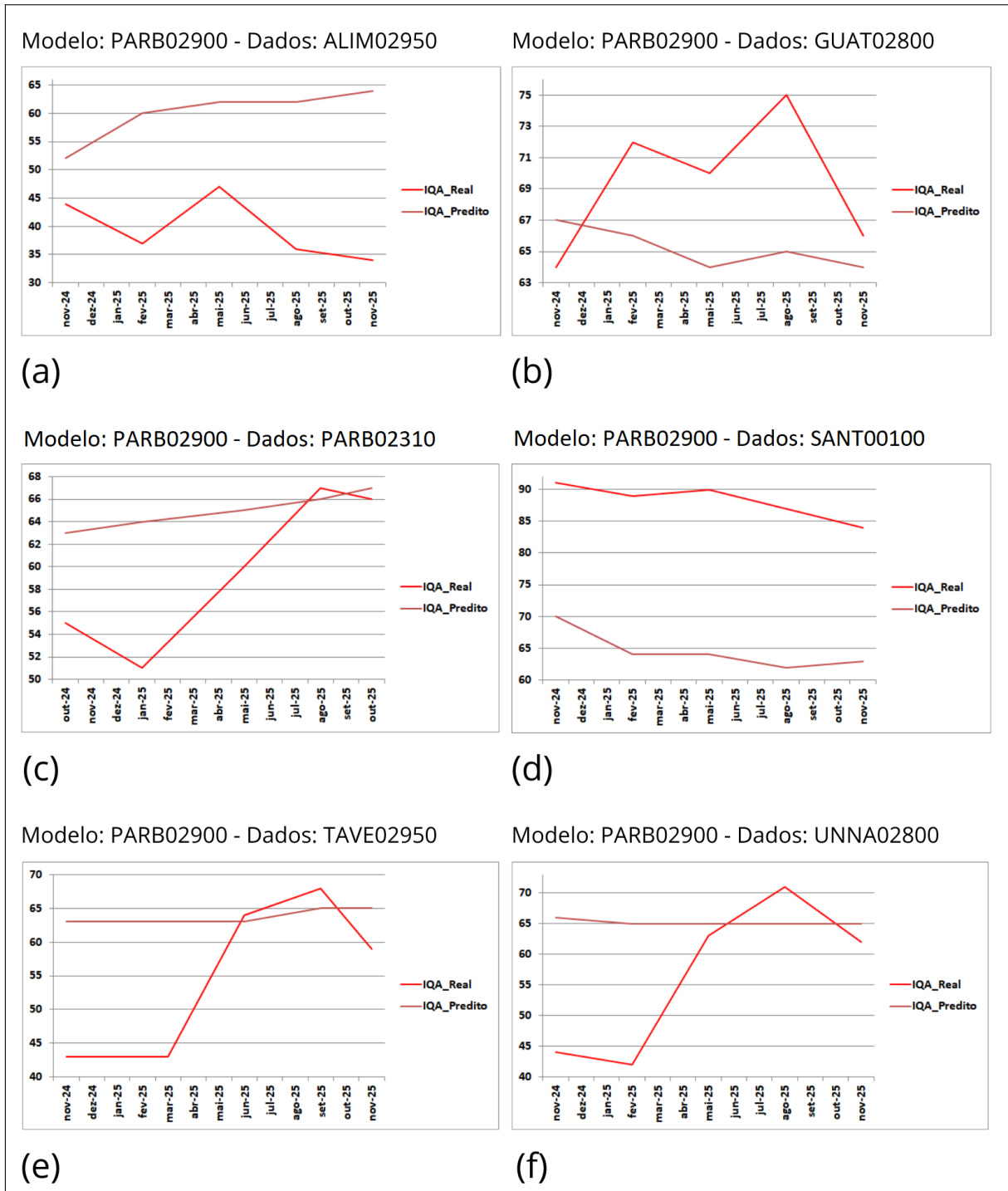
Fonte: Autor

Figura 49 – Predição do IQA com utilização iterativa de valores preditos - Modelo: PARB02900, Dados: PARB02900.



Fonte: Autor

Figura 50 – Predição do IQA com utilização iterativa de valores preditos - Modelo: PARB02900, Dados: demais pontos.



Fonte: Autor

4.4 APLICAÇÃO DOS MODELOS GERADOS EM UMA SIMULAÇÃO DE AMBIENTE INDUSTRIAL DE TRATAMENTO DE ÁGUA

Visando inserir os frutos desta pesquisa em um ambiente industrial de controle e monitoramento do tratamento de água, desenvolveu-se em ambiente virtual a simulação de um sistema *Supervisory Control and Data Acquisition* (SCADA), com o intuito de reproduzir a estrutura de monitoramento de qualidade de uma grande empresa do setor. A arquitetura da solução foi desenhada para garantir a interoperabilidade com sistemas industriais de monitoramento SCADA. Através do *software* Aveva Edge, o modelo de Floresta Aleatória atua como um sensor virtual, provendo estimativas em tempo real que auxiliam na tomada de decisão operacional, reduzindo custos com análises laboratoriais frequentes e aumentando a confiabilidade do monitoramento contínuo. Para isso, utilizou-se a linguagem R para a execução dos modelos desenvolvidos nesta pesquisa. Desta forma, integra-se ao sistema SCADA uma funcionalidade de AM que, naturalmente, não é nativa a este.

4.4.1 Integração da Linguagem R com Sistema SCADA

No *software* AVEVA™ Edge, foi desenvolvida uma aplicação para simular o monitoramento da qualidade da água na Região Metropolitana do Vale do Paraíba e Litoral Norte. A interface é estruturada em duas telas principais: a "Tela do Mapa", que apresenta a delimitação geográfica da região e redireciona o usuário ao clicar em um município, e a "Tela de Dados". Esta última é subdividida em três seções estratégicas: medições em tempo real, valores preditos e botões de comando e navegação.

As medições previamente armazenadas em formato eletrônico (.xlsx) são carregadas pelo sistema, servindo de entrada para que novos prognósticos sejam gerados por meio de dois *scripts* desenvolvidos em linguagem R. O gerenciamento e a requisição dessas rotinas ocorrem na plataforma SCADA (*Supervisory Control and Data Acquisition*), a qual também atua como receptora dos resultados processados.

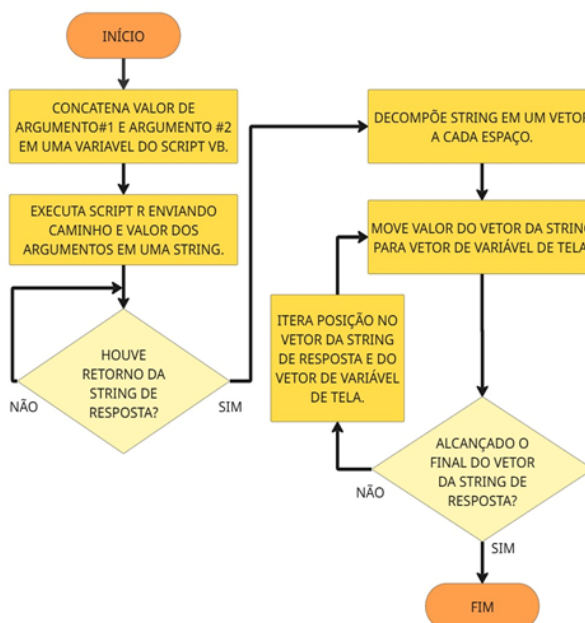
Para viabilizar essa integração, desenvolveram-se duas funções em *Visual Basic* (VB) no ambiente do SCADA, fundamentadas na metodologia proposta por (Pavan, 2016), com adaptações na sintaxe e na manipulação das estruturas de dados. O primeiro *script* em VB é responsável pela leitura dos dados da planilha e encaminha um único argumento à rotina subsequente; o segundo realiza a execução do modelo preditivo associado aos parâmetros do Índice de Qualidade da Água (IQA), transmitindo dois argumentos.

Cada rotina em VB é associada de forma unívoca a um *script* em R equivalente. O acoplamento entre as ferramentas é estabelecido a partir da instanciação de um objeto `WScript.Shell`, o qual concede ao sistema supervisor a capacidade de executar comandos diretamente no nível do sistema operacional, mimetizando a interface de linha de comando (*prompt*). O fluxo de dados bidirecional ocorre por meio de cadeias de caracteres (*strings*), o que demanda etapas explícitas de conversão de tipos, tanto na recepção dos argumentos pelo ambiente R quanto no

tratamento dos valores de retorno no ambiente VB.

A rotina implementada em linguagem *Visual Basic* (VB) segue o fluxo metodológico delineado na Figura 51, sendo responsável pela estruturação da cadeia de caracteres (*string*) de envio e pelo subsequente tratamento da resposta recebida. Este processo de tratamento mitiga a presença de ruídos de formatação e caracteres espúrios no fluxo de dados, viabilizando a extração precisa das variáveis de interesse.

Figura 51 – Fluxograma do código VB *script*.



Fonte: Autor.

Adicionalmente, a Figura 52 ilustra a interface computacional desenvolvida, a qual exhibe os valores projetados pelo modelo de AM executado no ambiente estatístico R. Observa-se que o valor predito para o IQA é devidamente apresentado na interseção entre a coluna “Predição” e a linha “IQA”, omitindo-se, nesta etapa, a classificação qualitativa da água para o usuário final.

Figura 52 – *Software* supervisorio (Aveva Edge) em execução integrado à linguagem R.

	2013-02-20	Medições		Predição
		2013-04-24	2013-06-12	2013/08/xx
Oxigênio Dissolvido	6.0	8.0	8.2	8.4
DBO (5,20)	5.0	6.0	3.0	4.4
Coliformes Termotolerantes	47500.0	800.0	280.0	1111.9
pH	6.4	7.1	7.0	7.1
Temperatura	22.0	19.0	20.0	19.3
Nitrogênio Total	1.0	0.7	0.7	0.7
Fósforo Total	0.034	0.021	0.020	0.021
Turbidez	198.0	15.0	9.0	4.8
Resíduo Total	174.0	74.0	50.0	50.8
IQA	42	69	76	71

Carregar Dados
Executar Predição
Mapa SIMPLIF.LIN

Fonte: Autor

5 CONCLUSÃO

A conclusão deste estudo confirma que a utilização de técnicas de Aprendizagem de Máquina, especificamente o algoritmo de Floresta Aleatória, representa uma solução tecnológica viável e de alta precisão para o monitoramento da qualidade da água na Região Metropolitana do Vale do Paraíba e Litoral Norte, principalmente quando há parâmetros ausentes. A pesquisa indica que, na situação de dados faltantes, a abordagem metodológica de predição individual dos parâmetros físico-químicos e biológicos para o cálculo subsequente do IQA é aparentemente superior à predição direta do valor final do índice. Os resultados obtidos neste caso apresentam maior repetibilidade e consistência, alcançando níveis de erro (MAPE) inferiores a 10%, o que classifica as previsões como altamente precisas segundo a literatura especializada.

O objetivo geral foi plenamente atingido ao se desenvolver uma ferramenta capaz de prever as condições de qualidade hídrica da região com base em um histórico de dados da CETESB entre 2013 e 2023. Além disso, a capacidade de generalização dos modelos revelou que padrões identificados em determinados pontos de monitoramento podem ser transpostos para outras áreas correlatas sem perda expressiva de acurácia, o que otimiza recursos e simplifica a implementação do sistema em larga escala. A integração bem-sucedida dos modelos gerados em um ambiente de simulação industrial via software supervisor SCADA reforça a aplicabilidade prática deste trabalho como um suporte à tomada de decisão e ao controle operacional de sistemas de tratamento. A viabilidade desse mecanismo recursivo, isento de ônus financeiro adicional, configura-se como um fator crítico para organizações caracterizadas por restrições orçamentárias. Sistemas industriais dotados de ferramentas de *analytics* alinhadas às diretrizes da Indústria 4.0 — tais como aprendizado de máquina, mineração de dados e demais tecnologias voltadas à conversão de dados em conhecimento estratégico — demandam, historicamente, elevado investimento de capital. Consequentemente, o custo de aquisição e implementação dessas plataformas proprietárias de alto valor agregado transcende a capacidade de investimento de uma parcela significativa de pequenas e médias empresas, tornando soluções alternativas economicamente estratégicas.

Diante do cenário global que exige a aceleração das metas do Objetivo de Desenvolvimento Sustentável nº 6 da ONU, este trabalho oferece propostas para futuras pesquisas que possam expandir o horizonte aqui estabelecido. Uma oportunidade promissora reside na inclusão de variáveis climáticas e de pluviosidade, visando compreender como eventos extremos impactam a dinâmica dos poluentes, bem como refinar o indicador de sazonalidade do IQA e a forma como ele correlaciona corpos de água, permitindo a melhor generalização dos modelos. Adicionalmente, aprofundar por meio de testes estatísticos de comparação a análise de qual método (predição dos parâmetros para posterior cálculo do IQA ou predição direta do IQA) apresentam melhor desempenho. E Também o emprego da utilização iterativa de valores preditos abre caminho para estudos de séries temporais de longo prazo com outras abordagens, visando reduzir o erro

acumulado a cada ciclo.

Por fim, a conexão direta desses modelos com sensores conectados à *Internet of Things* (IoT) poderá permitir um monitoramento preditivo totalmente automatizado, consolidando a inteligência artificial como ferramenta estratégica na preservação ambiental e na saúde pública.

REFERÊNCIAS

- AGEMVALE. **A RMVALE-LN em números**. 2021. Disponível em: <https://agemvale.sp.gov.br/a-rmvale-ln-em-numeros/>. Acesso em: 4 nov. 2024.
- AJORLOO, Sedighe *et al.* A systematic review of machine learning methods in software testing. **Applied Soft Computing**, Amsterdam, v. 162, p. 111805, 2024.
- ALDHYANI, Theyazn H. H *et al.* Water Quality Prediction Using Artificial Intelligence Algorithms. **Applied Bionics and Biomechanics**, London, v. 2020, p. 6659314, 2020.
- ALEXAKIS, Dimitrios E. Linking dpsir model and water quality indices to achieve sustainable development goals in groundwater resources. **Hydrology**, Basel, v. 8, n. 2, p. 90, 2021.
- ALJANABI, Zahraa Zahraw *et al.* A brief review of water quality indices and their applications. *In: Bristol. Conference paper*. v. 779, p. 012088. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85109758752&doi=10.1088%2f1755-1315%2f779%2f1%2f012088&partnerID=40&md5=9827fe0cf216723b735ee58e03c6b977>. Acesso em: 5 abr. 2024.
- AMIN, Samina. **An introduction to reinforcement learning and its application in various domains**. Hershey: IGI Global, 2024. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189606487&doi=10.4018%2f979-8-3693-1738-9.ch001&partnerID=40&md5=fa3ca276b2cc7f59bb7a91dd65bdc4f4>. Acesso em: 6 mar. 2025.
- ANA. **Indicadores de qualidade: índice de qualidade das águas (IQA)**. 2024. Disponível em: <https://portalpnqa.ana.gov.br/indicadores-indice-aguas.aspx>. Acesso em: 30 out. 2024.
- AZROUR, Mourade *et al.* Machine learning algorithms for efficient water quality prediction. **Modeling Earth Systems and Environment**, Cham, v. 8, n. 2, p. 2793–2801, 2022.
- BADRUDEEN, A.T. *et al.* A critical review of statistical, signal processing and machine learning methods for continuous and high-frequency water quality data improvement. **Ecological Informatics**, Amsterdam, v. 94, p. 103619, 2026.
- BELANY, Pavol *et al.* A Comparative Analysis of Polynomial Regression and Artificial Neural Networks for Prediction of Lighting Consumption. **Buildings**, Basel, v. 14, n. 6, p. 1712, 2024.
- BELGIU, Mariana; DRĂGUT, Lucian. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 114, p. 24–31, 2016.

BORUP, Daniel *et al.* Targeting predictors in random forest regression. **International Journal of Forecasting**, Amsterdam, v. 39, n. 2, p. 841–868, 2023.

BOUCHRAKI, Faiza *et al.* Water Quality Assessment Using the Random Forest Classification Model. **Water Environment Research**, Hoboken, v. 97, n. 10, e70197, 2025.

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

CAUCHICK-MIGUEL, Paulo Augusto *et al.* **Metodologia de pesquisa em engenharia de produção e gestão de operações**. 2. ed. Rio de Janeiro: Elsevier, 2012.

CETESB. **Aspectos descritivos e quantitativos das redes de monitoramento – 2023:**

Apêndice A. 2024. Disponível em:

<https://cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2024/09/Apendice-A-Aspectos-Descritivos-e-Quantitativos-das-Redes-de-Monitoramento-2023.pdf>. Acesso em: 15 fev. 2025.

CETESB. **IQA – Índice de Qualidade das Águas**. 2024. Disponível em:

<https://www.cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2013/11/02.pdf>. Acesso em: 30 out. 2024.

CETESB. **Relatório de qualidade das águas superficiais: Apêndice C**. 2024. Disponível em:

<https://cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2013/11/Apendice-C-Indices-de-Qualidade-das-Aguas.pdf>. Acesso em: 31 out. 2024.

CHAPMAN, Peter *et al.* **CRISP-DM 1.0: step-by-step data mining guide**. Chicago: SPSS Inc., 2000.

CHIDIAC, Sandra *et al.* A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives. **Reviews in Environmental Science and Bio/Technology**, Dordrecht, v. 22, n. 2, p. 349–395, 2023.

CHOUDHURY, Arghya; MONDAL, Arpita; SARKAR, Subhadeep. Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: a comparative study and review of random forest, AdaBoost, XGBoost and LightGBM frameworks. **European Physical Journal: Special Topics**, Berlin, v. 233, n. 15-16, p. 2425–2463, 2024.

DEVIPRIYA, B.; KALPANA, Y. A survey on recommender system based on machine learning concepts and semantic analysis. **Journal of Advanced Research in Dynamical and Control Systems**, v. 11, 7 Special Issue, p. 901–908, 2019.

DIKBAŞ, Fatih. Forecasting extreme precipitations by using polynomial regression. **Idojaras**, Budapest, v. 128, n. 3, p. 379–398, 2024.

FACELI, Katti *et al.* **Inteligência artificial: uma abordagem de aprendizado de máquina**. 2. ed. Rio de Janeiro: LTC, 2021.

GARCIA-RIOS, Victor *et al.* Predictive machine learning applying cross industry standard process for data mining for the diagnosis of diabetes mellitus type 2. **IAES International Journal of Artificial Intelligence**, Yogyakarta, v. 12, n. 4, p. 1713–1726, 2023.

GEOSEADE. **Mapa interativo - GeoSEADE**. 2024. Disponível em: <https://portalgeo.seade.gov.br/i3geo/interface/osm.php>. Acesso em: 3 nov. 2024.

GUIDO, Rosita *et al.* An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. **Information (Switzerland)**, Basel, v. 15, n. 4, p. 235, 2024.

HANDELMAN, Guy S. *et al.* Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. **American Journal of Roentgenology**, Leesburg, v. 212, n. 1, p. 38–43, 2019.

IBGE. **IBGE - Cruzeiro-SP**. 2023. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/cruzeiro.html>. Acesso em: 25 fev. 2025.

IBGE. **IBGE - Jambeiro-SP**. 2023. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/jambeiro.html>. Acesso em: 15 fev. 2025.

JORDAN, M.I.; MITCHELL, T.M. Machine learning: Trends, perspectives, and prospects. **Science**, Washington, v. 349, n. 6245, p. 255–260, 2015.

JOVEL, Juan; GREINER, Russell. An Introduction to Machine Learning Approaches for Biomedical Research. **Frontiers in Medicine**, Lausanne, v. 8, p. 771607, 2021.

KANNADHASAN, S.; VENUSAMY, Kanagaraj; NAGARAJAN, R. **Recent Trends in Machine Learning Techniques, Challenges and Opportunities**. Boca Raton: CRC Press, 2023. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152315930&doi=10.1201%2f9781003298335-6&partnerID=40&md5=de5d58f07696cb0fb6ac574bbd291a13>. Acesso em: 6 nov. 2024.

KAUL, Ajay; RAINA, Sneha. Support vector machine versus convolutional neural network for hyperspectral image classification: A systematic review. **Concurrency and Computation: Practice and Experience**, Chichester, v. 34, n. 15, e6945, 2022.

KAYIKCI, Safak; KHOSHGOFTAAR, Taghi M. Blockchain meets machine learning: a survey. **Journal of Big Data**, London, v. 11, n. 1, 2024.

KRISHNAN, Reshmy *et al.* Prediction and Analysis of Water Quality Using Machine Learning Techniques. **Lecture Notes in Networks and Systems**, Cham, v. 960, p. 148–158, 2024.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, London, v. 521, n. 7553, p. 436–444, 2015.

LEWIS, Colin D. **Industrial and business forecasting methods**. London: Butterworth & Co Ltd, 1982.

LIN, Junan *et al.* Temporal Prediction of Coastal Water Quality Based on Environmental Factors with Machine Learning. **Journal of Marine Science and Engineering**, Basel, v. 11, n. 8, p. 1608, 2023.

MANNA, Atanu; BISWAS, Debasish. Assessment of Drinking Water Quality Using Water Quality Index: A Review. **Water Conservation Science and Engineering**, Dordrecht, v. 8, n. 1, p. 8, 2023.

MILLER, Catriona *et al.* A review of model evaluation metrics for machine learning in genetics and genomics. **Frontiers in Bioinformatics**, Lausanne, v. 4, p. 1457619, 2024.

NAEEM, Muddasar; RIZVI, Syed Tahir Hussain; CORONATO, Antonio. A Gentle Introduction to Reinforcement Learning and its Application in Different Fields. **IEEE Access**, Piscataway, v. 8, p. 209320–209344, 2020.

NAIDU, Gireen; ZUVA, Tranos; SIBANDA, Elias Mmbongeni. A Review of Evaluation Metrics in Machine Learning Algorithms. **Lecture Notes in Networks and Systems**, Cham, v. 724, p. 15–25, 2023.

NAIR, Jitha P; VIJAYA, M.S. River Water Quality Prediction and index classification using Machine Learning. In: Bristol. **Conference paper**. v. 2325, p. 012011. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138296720&doi=10.1088%2f1742-6596%2f2325%2f1%2f012011&partnerID=40&md5=11a66eb1979d6ba17f3fac9e36852ada>. Acesso em: 17 dez. 2024.

NDUNGI, Rebecca; STANISLAVOVICH, Lekanov Ivan. Improving Time Series Forecasting by Applying the Sliding Window Approach. **Lecture Notes in Networks and Systems**, Cham, v. 1531, p. 294–302, 2025.

NERBONNE, John. The MIT encyclopedia of the cognitive sciences. **Computational Linguistics**, Cambridge, v. 26, n. 3, p. 463–467, 2000.

ONU. **Centro de informações da ONU Rio de Janeiro**. 2024. Disponível em: <https://www.un.org/pt/rio>. Acesso em: 1 set. 2024.

ONU. **Ensure availability and sustainable management of water and sanitation for all**. 2024. Disponível em: https://www.un.org/sustainabledevelopment/wp-content/uploads/2023/08/2309739_E_SDG_2023_infographics-6-6.pdf. Acesso em: 3 nov. 2024.

ONU. **Estratégia ODS**. 2024. Disponível em: <https://www.estrategiaods.org.br/conheca-os-ods/>. Acesso em: 1 set. 2024.

ONU. **Goal 6**: Ensure access to water and sanitation for all. 2024. Disponível em: <https://www.un.org/sustainabledevelopment/water-and-sanitation/>. Acesso em: 2 nov. 2024.

ONU. **The sustainable development goals report - 2023**. 2024. Disponível em: <https://unstats.un.org/sdgs/report/2023/The-Sustainable-Development-Goals-Report-2023.pdf>. Acesso em: 2 nov. 2024.

PARK, Sohee; KWON, Ryeonggu; KWON, Gihwon. Survivability Prediction of Open Source Software with Polynomial Regression. **Applied Sciences (Switzerland)**, Basel, v. 14, n. 7, p. 2812, 2024.

PAVAN. **How to call R script from VB script**. 2016. Disponível em: <https://sharepointwalk.wordpress.com/2016/06/01/how-to-call-r-script-from-vb-script/>. Acesso em: 15 jun. 2023.

PEREIRA, Rejane Félix *et al.* Análise de sensibilidade dos parâmetros do Índice de Qualidade de Água (IQA). In: Bento Gonçalves. ANAIS do XX Simpósio Brasileiro de Recursos Hídricos. [S. l.: s. n.], 2013. Disponível em: <https://anais.abrhidro.org.br/job.php?Job=999>. Acesso em: 20 abr. 2025.

PLEVRIS, Vagelis *et al.* Investigation of performance metrics in regression analysis and machine learning-based prediction models. In: Barcelona. **Conference paper**. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146939623&doi=10.23967%2feccomas.2022.155&partnerID=40&md5=33f2f5139f7bb8757b92c8886024cd37>. Acesso em: 23 nov. 2025.

PORTUGAL, Ivens; ALENCAR, Paulo; COWAN, Donald. The use of machine learning algorithms in recommender systems: A systematic review. **Expert Systems with Applications**, Oxford, v. 97, p. 205–227, 2018.

PRADO, Joici Rodrigues Silva *et al.* Variação da qualidade da água em relação à sazonalidade e ao uso da terra no sul da região amazônica. **Caderno Prudentino de Geografia**, Presidente Prudente, v. 2, n. 43, p. 159–184, 2021.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Novo Hamburgo: Feevale, 2013.

R FOUNDATION. **R: A language and environment for statistical computing**. Vienna, 2024. Disponível em: <https://www.r-project.org/>. Acesso em: 4 nov. 2024.

RAINIO, Oona; TEUHO, Jarmo; KLÉN, Riku. Evaluation metrics and statistical tests for machine learning. **Scientific Reports**, London, v. 14, n. 1, p. 6086, 2024.

RIVO, Eduardo *et al.* Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. **Clinical and Translational Oncology**, Barcelona, v. 14, n. 1, p. 73–79, 2012.

SANTOS, Maria Juraci Zani dos. Mudanças climáticas no estado de São Paulo. **Geografia**, Rio Claro, v. 21, n. 2, 2020. ISSN 1983-8700.

SEADE. **Censo SEADE 2022**. 2022. Disponível em: <https://censo2022.seade.gov.br/>. Acesso em: 4 nov. 2024.

SUN, Xiaohao *et al.* Perspective: Machine Learning in Design for 3D/4D Printing. **Journal of Applied Mechanics, Transactions ASME**, New York, v. 91, n. 3, p. 030801, 2024.

TURSUNALIEVA, Ainura *et al.* Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. **Applied Sciences (Switzerland)**, Basel, v. 14, n. 2, p. 496, 2024.

UDDIN, Md. Galal; NASH, Stephen; OLBERT, Agnieszka I. A review of water quality index models and their use for assessing surface water quality. **Ecological Indicators**, Amsterdam, v. 122, p. 107218, 2021.

WANG, Fuzeng *et al.* Ground visibility prediction using tree-based and random-forest machine learning algorithm: Comparative study based on atmospheric pollution and atmospheric boundary layer data. **Atmospheric Pollution Research**, Amsterdam, v. 15, n. 11, p. 102270, 2024.

WANG, Jihao *et al.* Real-Time Variable-Weight Aquaculture Water Quality Index Evaluation Method Based on Adaptive Sliding Window. **IEEE Sensors Journal**, Piscataway, v. 25, n. 10, p. 17293–17308, 2025.

WEESE, Maria *et al.* Statistical learning methods applied to process monitoring: An overview and perspective. **Journal of Quality Technology**, Milwaukee, v. 48, n. 1, p. 4–27, 2016.

YUSOF, Mohamed Khatif Tawaf Mohamed *et al.* A review of the application of support vector machines in landslide susceptibility mapping. **Disaster Advances**, Indore, v. 16, n. 11, p. 71–83, 2023.

DADOS CURRICULARES

IDENTIFICAÇÃO	
Nome	WILSON DOS SANTOS JUNIOR
Data de Nascimento	14/10/1986
Nacionalidade	brasileira
Nome em citações bibliográficas	Junior, Wilson Santos Junior, W. S.
Currículo Lattes	https://lattes.cnpq.br/0085225322072905
ORCID	https://orcid.org/0009-0001-3225-764X
FORMAÇÃO ACADÊMICA	
12/2018	Engenharia Elétrica, Engenheiro Centro Universitário Salesiano de São Paulo
09/2011	Tecnologia em Informática, Tecnólogo Centro Estadual de Educação Tecnológica Paula Souza (FATEC)
PRODUÇÃO BIBLIOGRÁFICA	
Junior W. S., Freitas M. A. G. M., Gomes F. M. and Carvalho E. F., “Integração da Linguagem R ao Sistema Scada para Predição do Índice de Qualidade da Água,” In 16th IEEE International Conference on Industry Applications (INDUSCON), 2025, São Sebastião: IEEE, 2025. p. 202-209.	
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	
16TH IEEE INTERNATIONAL CONFERENCE ON INDUSTRY APPLICATIONS (INDUSCON), 2025, (São Sebastião). Integração da Linguagem R ao Sistema Scada para Predição do Índice de Qualidade da Água. 2025. (Apresentação de Artigo).	
IX WORKSHOP DA PÓS-GRADUAÇÃO EM ENGENHARIA., 2025, (Guaratinguetá). Pôster: Incorporando as Funcionalidades de Aprendizagem de Máquina da Linguagem R em Um Sistema SCADA, Predizendo Variáveis de Rotação e Temperatura de Um Processo Industrial. 2025. (Apresentação de Pôster).	